Clinical history and diagnostic performance of cone beam computed tomography and digital periapical radiography in endodontic disease detection

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Kehn Enn Yapp
BDSc DCD (Endo)

Faculty of Medicine and Health
University of Sydney
February 2023

STATEMENT OF ORIGINALITY

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Kehn Enn Yapp
February 2023

ABSTRACT

Aims: 1) To assess and compare the diagnostic performance of digital periapical (PA) radiography and cone beam computed tomography (CBCT) in endodontic disease detection. 2) To assess the effect of clinical history on the interpretation of endodontic disease in dental cone beam computed tomography. 3) To assess the effect of clinical history on incidental abnormality detection and diagnostic confidence in endodontic CBCT imaging.

Methods: A free-response, factorial study design was used to account for changes in the independent variables: case type (disease or non-disease), case severity (subtle, moderate or obvious), reader type (high or low level of clinical experience and monthly CBCT reading volume) and reading modality (digital periapical radiography, cone beam computed tomography with and without clinical history). Readers interpreted 60 PA and 60 CBCT images divided into five categories: diseased – subtle, diseased – moderate, diseased – obvious, non-diseased – subtle, and non-diseased – obvious. The 60 CBCT images were read twice over two reading sessions using a balanced design, once with access to clinical history and once without, where 30 in each session included history. Readers were not told about the double reading nature of the CBCT component of the study and were informed that all 120 CBCT cases were different imaging modalities. Lesion localisation fraction, specificity, false positive marks, diagnostic confidence (rating 2, 3 or 4) of correctly localised incidental abnormalities and the weighted alternative free-response receiver operating characteristic (wAFROC1) figure of merit were calculated.

Results: CBCT had greater specificity than PA in the obvious non-diseased cases ($p=0.01$), but this was not significantly different in the subtle non-diseased category. wAFROC1 values were higher for PA than CBCT in the subtle diseased ($p=0.02$) and moderate diseased ($p=0.01$) groups with no significant difference between in the obvious diseased group. CBCT had higher mean false positives than PA ($p<0.05$) in subtle diseased cases. Mean lesion localisation fraction in the moderate diseased group was higher in PA than CBCT ($p=0.003$). No relationships were found between years of clinical experience and all diagnostic performance measures, except for in the obvious diseased group, where increasing experience was associated with fewer mean false positive marks in CBCT than PA ($p=0.04$).
Clinical history had no significant effect on specificity and false positive rates in non-diseased cases ($p>0.05$), but improved lesion localisation in subtle and obvious diseased cases ($p<0.01$). wAFROC1 values were higher with clinical history for subtle ($p<0.001$) and obvious ($p=0.006$) diseased categories. Performance with clinical history did not vary across readers' years of experience and reading volume in the non-diseased categories. Readers with fewer ($p=0.03$) and moderate ($p=0.008$)

years of experience and low (p=0.002) CBCT reading volume demonstrated better lesion localisation in subtle diseased cases when clinical history was available.
Clinical history increased the detection of incidental abnormalities in non-diseased subtle cases (p=0.04). Reader experience and monthly CBCT reading volume did not affect incidental abnormality detection. The highest confidence rating was most often used in each case type when clinical history was available. For this rating, history had significantly greater lesion localisations in subtle diseased (p=0.03) and non-diseased images (p=0.02).

Conclusions: Periapical radiography performed better than CBCT in the detection of endodontic disease. CBCT had greater diagnostic performance than PA radiography in non-diseased images. Clinical history improved the interpretation of CBCT images with disease without significantly affecting the interpretation of images without disease. Clinical history improved the detection of incidental endodontic abnormalities only in non-diseased subtle CBCT images. Reader confidence in correctly identified abnormalities was higher when clinical history was available for images with subtle disease and non-disease, but was not associated with an improvement in diagnostic performance. Clinical experience did not impact upon the accuracy of interpretation of both PA radiography and CBCT. Clinical history improved CBCT diagnostic performance in less and moderately experienced readers and low volume readers.

PREFACE

This thesis contains seven chapters. They are organised as follows:

- Chapter 1 is an introduction to the thesis, providing background information on diagnostic performance of cone beam computed tomography and digital periapical radiography in endodontic disease detection. It highlights the relevance of clinical history as a bias in image interpretation.
- Chapter 2 is a systematic review comparing the diagnostic performance of cone beam computed tomography and digital periapical radiography in endodontic disease detection.
- Chapter 3 is a systematic review analysing the effect of clinical history on diagnostic image interpretation.
- Chapter 4 is a factorial study using the free-response paradigm to compare diagnostic performance of cone beam computed tomography and digital periapical radiography in endodontic disease detection.
- Chapter 5 is a crossover study analysing the effect of clinical history on diagnostic performance of cone beam computed tomography in endodontic disease detection.
- Chapter 6 is a study on the effect of clinical history on incidental abnormality detection and diagnostic confidence in cone beam computed tomography.
- Chapter 7 summarises the findings, relevance and limitations of the thesis and directions for future studies.

A bridging section prefaces chapters 4-6, outlining the relevance of the study in relation to the existing literature.

AUTHOR ATTRIBUTION STATEMENT

The following peer reviewed journal papers have been published and are included in this thesis:

- Yapp KE, Brennan P, Ekpo E. Endodontic disease detection: digital periapical radiography versus cone-beam computed tomography—a systematic review. Journal of Medical Imaging 2021, 8(4):041205. https://doi.org/10.1117/1.JMI.8.4.041205
- Yapp KE, Brennan P, Ekpo E. The Effect of Clinical History on Diagnostic Imaging Interpretation – A Systematic Review. Academic Radiology 2022, 29(2):255-266. https://doi.org/10.1016/j.acra.2020.10.021
- Yapp KE, Suleiman M, Brennan P, Ekpo E. Periapical radiography versus cone beam computed tomography in endodontic disease detection: a free-response, factorial study. Journal of Endodontics 2023, 49(4):419-429. https://doi.org/10.1016/j.joen.2023.02.001
- Yapp KE, Brennan P, Ekpo E. The effect of clinical history on diagnostic performance of endodontic cone beam computed tomography interpretation. Clinical Radiology 2023, 78(5):e433-e441. https://doi.org/10.1016/j.crad.2022.12.005

In all included papers, I am the corresponding author.

I provided substantial contribution to the study design, literature review, data collection, data analysis, writing of the manuscript, discussion of the results and conclusion of the paper, for all included papers.

Kehn Enn Yapp
February 2023

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help from the following people. I wish to thank:

- My supervisors, Dr Ernest Ekpo and Professor Patrick Brennan, for their guidance and teaching.
- My family and friends for their continued support.
- Drs David Barnard and Eric Herbranson for their help with reviewing the image test set.
- All the participants who contributed by undertaking several hours of image interpretation.
- Drs Dale Jung and Viraj Vora for providing imaging data.
- Drs John Khademi, John Hatton and Craig Rhodes for their help in collecting pilot data.
- Dr Mo'ayyad Suleiman for assistance with data collection.

TABLE OF CONTENTS

CHAPTER 1

1.1 Introduction

Endodontic disease involving the pulpal and periradicular tissues is a widespread disease entity, with prevalence ranging from 7% to 86% (Persoon and Özok 2017). An estimated 22 million endodontic procedures are performed annually in the United States of America (2007). Endodontically-related pain is very common and estimated to affect 88% of people (Kakoei et al. 2013). This problem is therefore significant in affecting one's quality of life (Locker and Quiñonez 2011).

Diagnosis and treatment planning of endodontic disease is performed with the aid of diagnostic imaging, namely periapical (PA) radiography as the main modality (Berman and Rotstein 2016). One of the first studies to evaluate performance of conventional (analogue) PA radiographs used cadavers, finding that simulated pathologic changes in cancellous bone were viewed as imaging radiolucencies if there was cortical bone perforation (Bender and Seltzer 1961). A subsequent study reported that PA radiography had good diagnostic performance in endodontic disease detection (Reit and Gröndahl 1983). A limitation of PA imaging is reader perception and interpretation, where six readers with varying levels of endodontic clinical experience agreed with each other's PA radiograph interpretation less than half the time (Goldman et al. 1972). Diagnostic interpretation and performance of PA imaging has also been shown to have a wide variation, with sensitivity ranging from 27 to 60% (Kanagasingam et al. 2017, Kruse et al. 2019).

The evolution of diagnostic imaging in endodontics has embraced three dimensional imaging, in particular cone beam computed tomography (CBCT). The use of CBCT has seen a recent increase in popularity in both number of imaging studies ordered (Brown and Monsour 2015) and machines installed (Lam et al. 2021). As the use of CBCT has increased, more studies evaluating its diagnostic performance have been published; however, many limitations exist. The validation of the index test, using a reference standard, has not been used in CBCT studies (Estrela et al. 2008), data on non-diseased patients are unavailable because test-sets only contained images with disease (Patel et al. 2012), and the severity of disease or non-diseased has not been accounted for (Abella et al. 2014). It is important to understand and improve diagnostic imaging performance because errors can lead to incorrect teeth being treated, disease progressing by being untreated and a lack of confidence in the clinician.

Diagnostic performance of imaging modalities requires evaluation at multiple levels, each focusing on different types of outcomes. A model of efficacy has described these levels ranging from technical and diagnostic, to therapeutic and societal (Fryback and Thornbury 1991). The value of diagnostic imaging can also be classified into medical (impact on treatment decisions), planning (ability for patients to make better life decisions) and psychic (effect on patients' sense of self) (Lee et al. 2010).

Although there are other factors to consider with these outcomes and values, regarding at risk population, anticipated clinical impact and economic impact (Gazelle et al. 2011), transparent data on diagnostic accuracy and performance are still required (Cohen et al. 2016) to assess the diagnostic ability of imaging modalities.

1.2 Digital periapical radiography

Periapical (PA) radiography is the main imaging modality and an integral part of endodontic diagnosis and treatment planning (Berman and Rotstein 2016). The images are generated using an X-ray tube, when a steam of electrons is accelerated from the cathode to the anode by a potential difference between the electrodes. When these electrons interact with the anode material, some of the electrons are converted into X-ray photons (White and Pharoah 2013). X-ray photons are generated with a heterogenous spectrum comprising of photons with different wavelengths. These photons are filtered to remove low energy photons and harden the beam to improve its quality. The final X-ray beam emanating from the X-ray tubes is collimated to restrict the size of the X-ray beam to the region under examination and limit the volume of tissue irradiated.

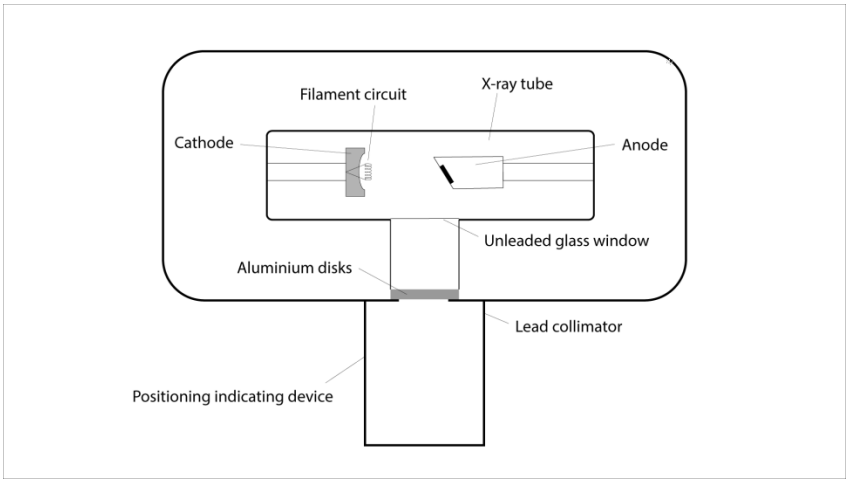


Figure 1.1: X-ray tube internal components.

The X-ray beam is attenuated as it passes through matter, before reaching a detector, which converts the incident photons to an electrical signal or charge, depending on whether an indirect or direct conversion detector is used. The information contained in the charge is sampled allowing the computer to store the information as raw data and distribute these data into picture elements (pixels). The computer then quantises (assigns a value to) these pixels so that the information can be distinguished by different shades of grey. The sampling level determines the spatial distribution of information from the X-ray beam and controls the spatial resolution of the image. The quantisation level determines the number of bits per pixel (bit depth) and controls the intensity resolution of the

image. The resultant digital image can be post-processed as may be necessary. Examples of digital PA radiographs are shown in Figure 1.2.

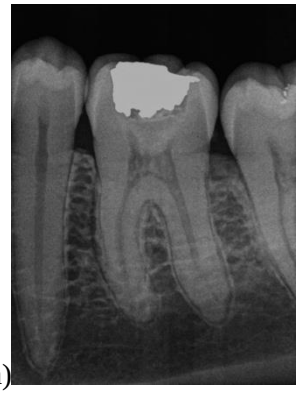
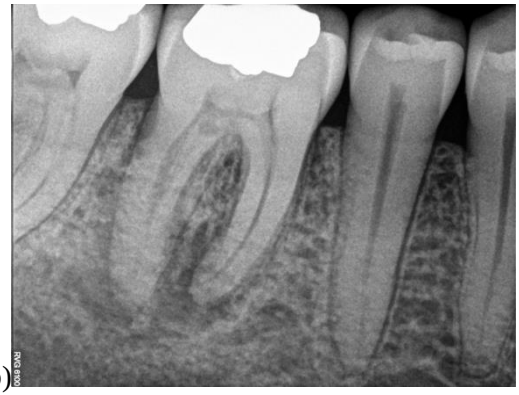
(a)  (b) 

Figure 1.2: (a) non-diseased tooth, (b) endodontic disease with area of lucency around root tip

1.3 Cone beam computed tomography

Cone beam computed tomography (CBCT) is a three dimensional imaging modality that has been recently adopted in dentistry. It has some similarities with PA radiography in that the X-ray beam is projected towards a digital sensor, however in addition to taking multiple images, the larger sensor detects the primary and scattered x-ray photons.

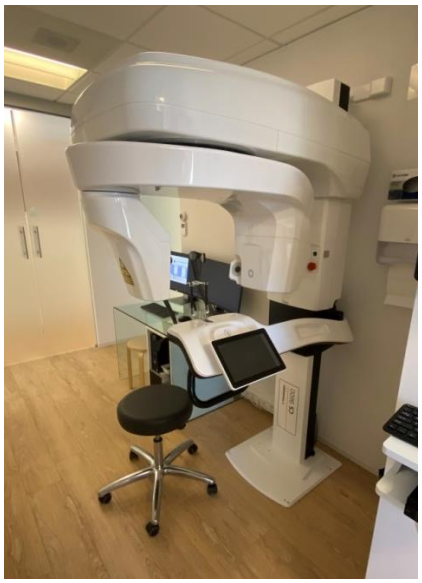


Figure 1.3: dental cone beam computed tomography machine.

From there, a series of transformations occur which form the reconstructive process. The three key steps are the Radon transform, Fourier transform and projection-slice theorem (Khademi 2017). These complex processes allow the reader to view a three dimensional reconstruction of the acquired object

and use viewing software to manipulate not only the window/level settings, but also the different slices in the captured field of view.

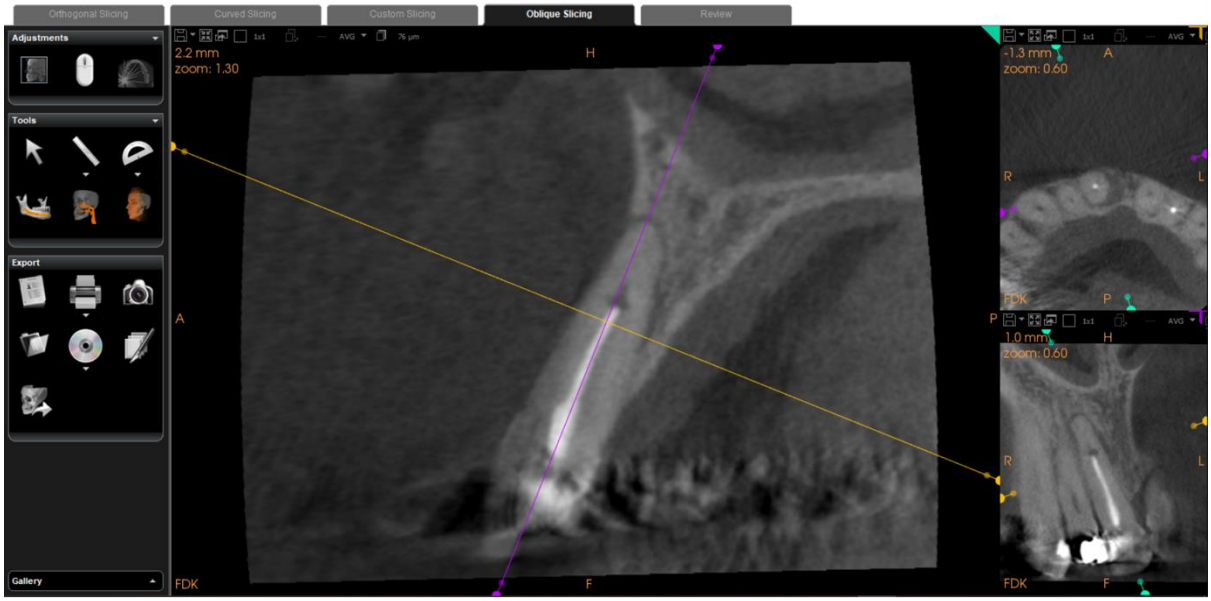


Figure 1.4: cone beam computed tomography images displayed using viewing software.

The introduction of CBCT into the clinical evaluation of endodontic disease has led to several studies investigating its diagnostic performance compared to PA radiography. The performance metrics comparing these two modalities have included sensitivity, specificity and area under the receiver operating characteristic (ROC) curve, which measure diagnostic efficacy at the case level. Sensitivity for PA radiography and CBCT imaging have ranged from 27 – 60% and 80 – 89% respectively (Pope et al. 2014, Kanagasingam et al. 2017, Kruse et al. 2019). In the same studies, specificity for PA radiography and CBCT ranged from 99% and 26 – 100% respectively. The area under the ROC curve for PA radiography and CBCT imaging was 0.629 and 0.943 respectively.

The limitations of these studies include measurement not at the abnormality level, instead of at the case level. Without data on the location of abnormalities, errors can be disguised as correct calls (Bunch et al. 1977). In particular, two errors are present – a location-level false positive and false negative occur on the same image. This is significant because treatment can occur at the wrong location and the missed abnormality can grow and increase in severity. This is overcome by conducting a study in the free-response paradigm (Chakraborty and Berbaum 2004). This mimics the clinical task by allowing the reader to mark as many abnormalities as they wish. Every abnormality is assigned a confidence level, creating a mark-rating pair (Chakraborty 2011). This method awards correct calls and penalises incorrect decisions.

Another limitation is the lack of a valid reference standard for the index test (Cohen et al. 2016), which does not provide a "truth value" for disease presence or absence. Some studies have compared the agreement between CBCT and PA reporting (Abella et al. 2012) and used cadaver findings (Kruse et al. 2019) as the assigned reference. A way to provide a valid reference standard is to use a Delphi panel, where a panel use examination and follow up clinical information to collectively determine a consensus (Stheeman et al. 1995).

The patient population in many of these studies had high disease prevalence, with several with 100% prevalence (Lofthag-Hansen et al. 2007, Lennon et al. 2011, Patel et al. 2012). This gives data on diseased patients only, with no information on how the imaging modalities perform in a non-diseased population. Conversely one study had 8% disease prevalence (Pope et al. 2014), meaning there was very little information on how PA radiography and CBCT imaging performed in diseased patients. To address these limitations of insufficient data on diseased or non-diseased patients, they should be individually sampled to not only have separate diseased and non-diseased groups, but also degrees of severity in each category. This is done using a factorial study design (Khademi 1994), which accounts for changes in the independent variable. Other relevant factors can be controlled, such as imaging modality and reader type.

Reader numbers in studies comparing CBCT and PA imaging have been low with a limited range of experience and training. Many included two to five readers (Pope et al. 2014, Kanagasingam et al. 2017, Kruse et al. 2019), who were mainly endodontists or oral and maxillofacial radiologists. To be able to generalise data to a broader population, a range of reader types should be included (Hanley 1989), with larger sample sizes to increase power. Diagnostic performance of PA radiography and CBCT imaging is poorly understood because of these methodological limitations in the published literature. To provide a proper understanding of these imaging modalities, my thesis will aim to overcome these limitations and undertake a study comparing diagnostic performance of PA radiography and CBCT using a rigorously planned method.

1.4 Clinical history bias in image interpretation

Clinical history in a diagnostic context includes patient information and medical history, reasons for presentation to a clinician, patient concerns, reports of symptoms, examination findings, test results and any other relevant information pertaining to the entire examination process. This data is used to determine the need for further testing (Bickley and Szilagyi 2017). History is important in helping a clinician make a diagnosis and it can be argued that a diagnosis cannot and should not be made without clinical history. In diagnostic imaging, the rationale behind clinical history accompanying imaging requests is to ensure a request is justified and appropriate (Maizlin and Somers 2019).

However, there is argument for image interpretation without reading clinical history because of the interpretive bias from history, where this information should only be incorporated into decision making after an unbiased read (Griscom 2002).

Accessing patient information prior to diagnostic image interpretation has been discussed ever since the 1960s (Schreiber 1963). Cognitive biases affecting image interpretation subsequent to having knowledge of clinical history are numerous, including anchoring bias – the tendency to focus on salient features in the initial presentation without adjusting this initial impression in light of later information (Croskerry 2003). Another is attribution bias, when a diagnostic conclusion depends on how the clinical information is presented (Busby et al. 2018), as is framing bias – a tendency to be influenced by how a problem is presented (Itri and Patel 2018). Clinical history is a form of bias that affects image interpretation; its effect should be evaluated because of the potential errors that can occur.

Various effects of clinical history on image interpretation and diagnostic performance have been documented, such as history having a significant increase in abnormality detection (Doubilet and Herman 1981), no difference to having no history in diagnostic performance (Cooperstein et al. 1990, Good et al. 1990), a mixed effect on accuracy and an increased tendency to influence questionable images as being positive for disease when correlated with clinical findings (Eldevik et al. 1982). A systematic review from 2004 on the accuracy of not just imaging, but all diagnostic tests, interpreted with and without clinical information found that when measured at the case level, test accuracy improved with clinical information (Loy and Irwig 2004). Due to the measurement tools included, they may not have been able to account for relevant interpretation errors (Bunch et al. 1977). Newer methods of data collection and analysis were not included in this review; these allow for a more rigorous observer performance assessment that account for location sensitivity and multiple abnormalities per case (Starr et al. 1975, Chakraborty and Berbaum 2004). Some studies using these newer methods have shown that history reduced diagnostic performance (Dhingsa et al. 2004) and increased false positive calls without improving diagnostic performance (Littlefair et al. 2016b). Given the lack of recent data encompassing these newer data analysis methods, an updated review was required.

1.5 Clinical history in image interpretation

Studies comparing the effect of clinical history on image interpretation have had variations in study design, including using a localising clinical history (Berbaum et al. 1988a), specific reporting instructions (Robinson et al. 2016) and providing prior images or previous radiological reports to readers (Aideyan et al. 1995). Others used a tentative diagnosis (Berbaum et al. 1986), an expected

disease prevalence (Littlefair et al. 2016b), a sham history (Elmore et al. 1997) and two contrived history scenarios – a "defendants read" and "expert witness read" (Littlefair et al. 2016a). To adequately compare the effect of clinical history on interpretation, readers would need to read the same images twice, ie. once with and once without clinical history. This crossover design was not used in one study (Eldevik et al. 1982).

To control reader bias, study design would need to limit any risk of the memory effect. This would require enough time between sessions, so readers would not remember reading the same images from a previous session. Some studies reported reading sessions were separated by "several months" or more (Berbaum et al. 1986, Berbaum et al. 1988b, Elmore et al. 1997, Tudor et al. 1997, Millet et al. 2013, Soh et al. 2014), others had 2–4 weeks (McNeil et al. 1983, Dhingsa et al. 2004, Houssami et al. 2005, Roelofs et al. 2007, Cereser et al. 2010), another 1–3 days (Littlefair et al. 2016b) and others had no time separation (Houssami et al. 2004) or it was not disclosed (Cooperstein et al. 1990, Good et al. 1990). It has been reported that a mean time gap of seven weeks between reading sessions is sufficient, where recognition accuracy is similar to chance (Evans et al. 2016).Another bias is the possibility of changes in reader performance and training between sessions. To account for this, a balanced design should be used – where images both with and without history are read at each session. Only a few studies incorporated this into their design (Berbaum et al. 1986, Berbaum et al. 1988b, Elmore et al. 1997).

Imaging studies evaluating the effect of clinical history on image interpretation have been conducted in the sensitivity and specificity paradigm (Kinnunen et al. 1987, Elmore et al. 1997, Tudor et al. 1997, Quekel et al. 2001, Cereser et al. 2010, Millet et al. 2013) and the ROC paradigm – measuring the relationship between true positive fraction and false positive fraction (McNeil et al. 1983, Berbaum et al. 1986, Berbaum et al. 1988b, Cooperstein et al. 1990, Good et al. 1990, Song et al. 1992, Ehara and Katsuragawa 1999, Houssami et al. 2004, Houssami et al. 2005, Baek et al. 2009, Filippone et al. 2009, Saba et al. 2019), all using abnormality measurement at the case level. A few studies used data analysis measured at image location level, such as the Localised Receiver Operating Characteristic (LROC) paradigm (Roelofs et al. 2007) and the region of interest (ROI) paradigm (Soh et al. 2014). The free response paradigm, which measures interpretation at the abnormality level, was used in two studies (Dhingsa et al. 2004, Littlefair et al. 2016b). Without measurement at the abnormality level, multiple abnormalities per image cannot be accounted for, the clinical task is not replicated and errors can be disguised as correct calls (Bunch et al. 1977).

Disease prevalence in studies comparing image interpretation with and without history have varied from less than 20% (Elmore et al. 1997, Soh et al. 2014), 50–62% (Berbaum et al. 1986, Berbaum et al. 1988b, Song et al. 1992, Tudor et al. 1997) and 100% (Dhingsa et al. 2004). This wide range in

disease prevalence does not necessarily incorporate the entire spectrum of disease presentation and may impact external validity by not being representative of the broader population. When diseased and non-diseased cases are analysed separately, this allows for sampling into the complete spectrum of disease and non-disease presentation (subtle, moderate and obvious) and including all case types into the analysis, such as with a factorial study design (Khademi 1994).

1.6 Deficiencies in the literature

The published literature demonstrates significant methodological limitations in studies assessing the effect of clinical history on image interpretation. These limitations were that published studies:

- did not account for the reader's clinical task, which limits external validity of the findings from these studies
- performed case-based analyses which only considers readers' ability to classify images as normal (non-diseased) or abnormal (diseased) without allowing for measurement of multiple abnormalities per image. There is a paucity of data on studies incorporating detection tasks or abnormality-based analysis into observer performance in endodontics
- were based on older observer performance methodologies, which only reward correct decisions but do not penalise incorrect abnormality locations
- did not account for memory bias by allowing an adequate interval between reading sessions while assessing the impact of clinical history on diagnostic performance
- did not provide a balanced reading design, which does not account for changes in reader ability between sessions, and
- did not include the entire spectrum of disease and non-disease severity in the study population, limiting the external validity of the collected data.

1.7 Thesis objectives

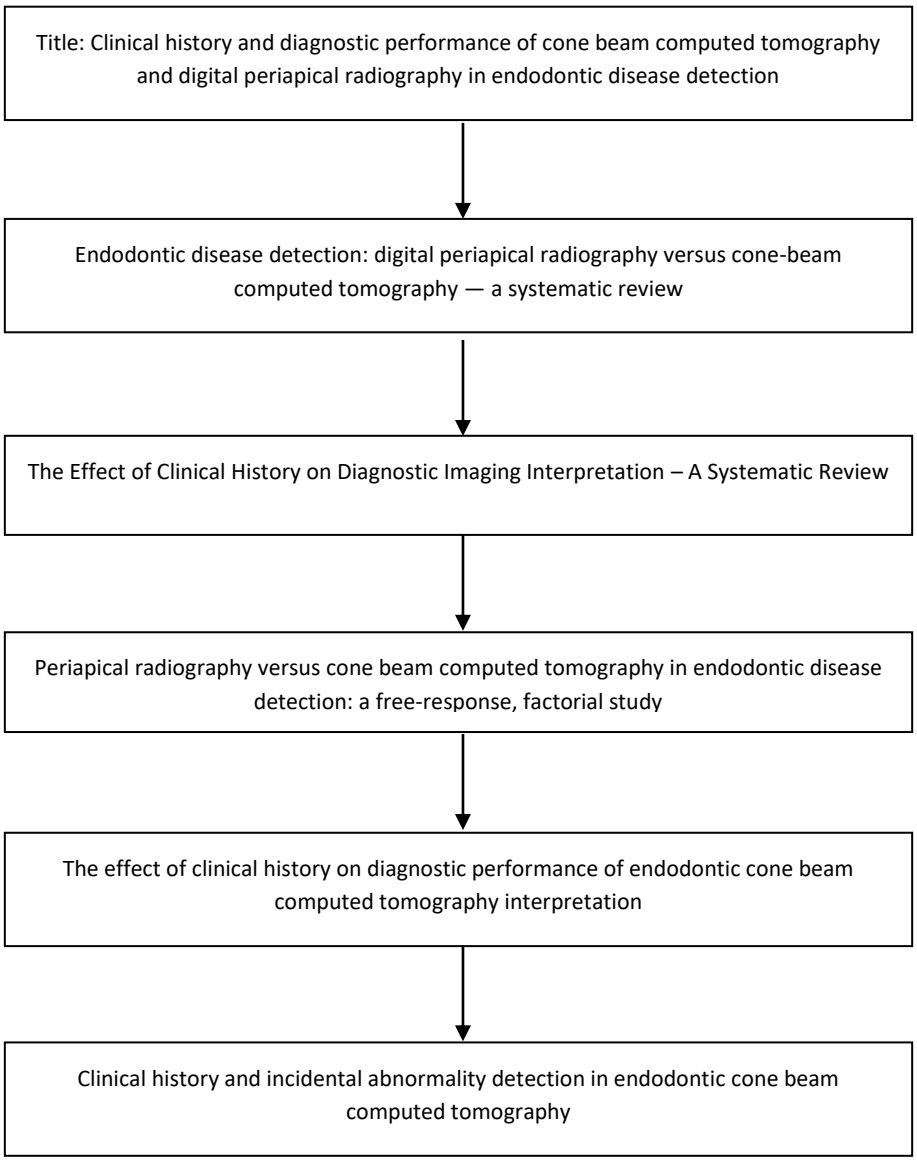
This thesis aims to provide experimental data on the diagnostic performance of CBCT and digital PA radiography in endodontic disease detection. It also aims to examine the effect of clinical history on CBCT image interpretation, incidental abnormality detection, and diagnostic confidence in endodontic CBCT imaging.

The objectives are:

1. To undertake a factorial study using the free-response paradigm to compare the diagnostic performance of CBCT and digital PA radiography in endodontic disease detection.

2. To perform a factorial, crossover study to examine the effect of clinical history on diagnostic performance of CBCT in endodontic disease detection.
3. To investigate the effect of clinical history on incidental abnormality detection and diagnostic confidence of CBCT in endodontic disease detection.

1.8 Thesis structure



References

American Dental Association (2007). 2005-2006 Survey of Dental Services Rendered. Chicago.
Abella, F., et al. (2012). "Evaluating the periapical status of teeth with irreversible pulpitis by using cone-beam computed tomography scanning and periapical radiographs." Journal of Endodontics **38**(12): 1588-1591.

Abella, F., et al. (2014). "An evaluation of the periapical status of teeth with necrotic pulps using periapical radiography and cone- beam computed tomography." Int Endod J **47**(4): 387-396.
Aideyan, U. O., et al. (1995). "Influence of Prior Radiologic Information on the Interpretation of Radiographic Examinations." Academic Radiology **2**(3): 205-208.
Baek, S. E., et al. (2009). "Effect of clinical information on diagnostic performance in breast sonography." J Ultrasound Med **28**(10): 1349-1356.
Bender, I. B. and S. Seltzer (1961). "Roentgenographic and direct observation of experimental lesions in bone: I." J Am Dent Assoc **62**(2): 152-160.
Berbaum, K. S., et al. (1988a). "Impact of clinical history on fracture detection with radiography." Radiology **168**(2): 507-511.
Berbaum, K. S., et al. (1986). "Tentative Diagnoses Facilitate the Detection of Diverse Lesions in Chest Radiographs." Investigative Radiology **21**(7): 532-539.
Berbaum, K. S., et al. (1988b). "Influence of Clinical History upon Detection of Nodules and Other Lesions." Investigative Radiology **23**(1): 48-55.
Berman, L. and I. Rotstein (2016). Diagnosis. Pathways of the Pulp. K. Hargreaves and L. Berman, Elsevier**:** 2-32.
Bickley, L. S. and P. G. Szilagyi (2017). Bates' Guide to Physical Examination and History Taking. Philadelphia, Wolters Kluwer.
Brown, L. F. and P. Monsour (2015). "The growth of Medicare rebatable cone beam computed tomography and panoramic radiography in Australia." Aust Dent J **60**(4): 511-519.
Bunch, P., et al. (1977). A Free Response Approach To The Measurement And Characterization Of Radiographic Observer Performance, SPIE.
Busby, L. P., et al. (2018). "Bias in Radiology: The How and Why of Misses and Misinterpretations." Radiographics **38**(1): 236-247.
Cereser, L., et al. (2010). "Impact of clinical data on chest radiography sensitivity in detecting pulmonary abnormalities in immunocompromised patients with suspected pneumonia." Radiol Med **115**(2): 205-214.
Chakraborty, D. P. (2011). "New developments in observer performance methodology in medical imaging." Seminars in Nuclear Medicine **41**(6): 401-418.
Chakraborty, D. P. and K. S. Berbaum (2004). "Observer studies involving detection and localization: Modeling, analysis, and validation." Medical Physics **31**(8): 2313-2330.
Cohen, J. F., et al. (2016). "STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration." BMJ open **6**(11): e012799-e012799.
Cooperstein, L. A., et al. (1990). "The effect of clinical history on chest radiograph interpretations in a PACS environment." Investigative Radiology **25**(6): 670-674.
Croskerry, P. (2003). "The importance of cognitive errors in diagnosis and strategies to minimize them." Acad Med **78**(8): 775-780.
Dhingsa, R., et al. (2004). "Prostate cancer localization with endorectal MR imaging and MR spectroscopic imaging: effect of clinical data on reader accuracy." Radiology **230**(1): 215-220.
Doubilet, P. and P. G. Herman (1981). "Interpretation of radiographs: effect of clinical history." American Journal of Roentgenology **137**(5): 1055-1058.
Ehara, S. and S. Katsuragawa (1999). "Influence of clinical information on the detection of wrist fractures in children." Tohoku J Exp Med **189**(2): 147-153.
Eldevik, O. P., et al. (1982). "The effect of clinical bias on the interpretation of myelography and spinal computed tomography." Radiology **145**(1): 85-89.
Elmore, J. G., et al. (1997). "The Impact of Clinical History on Mammographic Interpretations." JAMA **277**(1): 49-52.
Estrela, C., et al. (2008). "Accuracy of Cone Beam Computed Tomography and Panoramic and Periapical Radiography for Detection of Apical Periodontitis." J Endod **34**(3): 273-279.
Evans, K. K., et al. (2016). "Radiologists remember mountains better than radiographs, or do they?" J Med Imaging (Bellingham) **3**(1): 011005.
Filippone, A., et al. (2009). "Effect of Clinical History in Focal Liver Lesion Detection and Classification on 4-Detector Row Computed Tomography and Gadoxetic Acid Enhanced MR Imaging in Oncologic Patients." Journal of Computer Assisted Tomography **33**(6): 851-857.

Fryback, D. G. and J. R. Thornbury (1991). "The Efficacy of Diagnostic Imaging." Medical Decision Making **11**(2): 88-94.
Gazelle, G. S., et al. (2011). "A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research." Radiology **261**(3): 692-698.
Goldman, M., et al. (1972). "Endodontic success--who's reading the radiograph?" Oral Surg Oral Med Oral Pathol **33**(3): 432-437.
Good, B. C., et al. (1990). "Does Knowledge of the Clinical History Affect the Accuracy of Chest Radiograph Interpretation?" American Journal of Roentgenology **154**(4): 709-712.
Griscom, N. T. (2002). "A Suggestion: Look at the Images First, Before You Read the History." Radiology **223**(1): 9-10.
Hanley, J. A. (1989). "Receiver operating characteristic (ROC) methodology: the state of the art." Crit Rev Diagn Imaging **29**(3): 307-335.
Houssami, N., et al. (2004). "The influence of clinical information on the accuracy of diagnostic mammography." Breast Cancer Res Treat **85**(3): 223-228.
Houssami, N., et al. (2005). "The influence of knowledge of mammography findings on the accuracy of breast ultrasound in symptomatic women." Breast J **11**(3): 167-172.
Itri, J. N. and S. H. Patel (2018). "Heuristics and Cognitive Error in Medical Imaging." American Journal of Roentgenology **210**(5): 1097-1105.
Kakoei, S., et al. (2013). "Prevalence of toothache and associated factors: a population-based study in southeast iran." Iranian Endod J **8**(3): 123-128.
Kanagasingam, S., et al. (2017). "Diagnostic accuracy of periapical radiography and cone beam computed tomography in detecting apical periodontitis using histopathological findings as a reference standard." Int Endod J **50**(5): 417-426.
Khademi, J. (1994). The Effects of Digitization on Diagnostic Performance in Endodontics and Periodontics: an ROC study Master Thesis, The University of Iowa.
Khademi, J. A. (2017). Advanced CBCT for Endodontics: Technical Considerations, Perception, and Decision-Making. Hanover Park, IL, USA, Quintessence Publishing Co, Inc.
Kinnunen, J., et al. (1987). "The effects of clinical data on radiological diagnostic performance in middle-face trauma." Rofo **147**(2): 143-146.
Kruse, C., et al. (2019). "Diagnostic accuracy of cone beam computed tomography used for assessment of apical periodontitis: an ex vivo histopathological study on human cadavers." Int Endod J **52**(4): 439-450.
Lam, M., et al. (2021). "Current trends in the adoption and education of cone beam computed tomography and panoramic radiography machines across Australia." Dentomaxillofac Radiol **50**(5): 20200380.
Lee, D. W., et al. (2010). "Understanding the Medical and Nonmedical Value of Diagnostic Testing." Value in Health **13**(2): 310-314.
Lennon, S., et al. (2011). "Diagnostic accuracy of limited-volume cone-beam computed tomography in the detection of periapical bone loss: 360degree scans versus 180degree scans." International Endodontic Journal **44**(12): 1118-1127.
Littlefair, S., et al. (2016a). "Outcomes Knowledge May Bias Radiological Decision-making." Academic Radiology **23**(6): 760-767.
Littlefair, S., et al. (2016b). "Increasing Prevalence Expectation in Thoracic Radiology Leads to Overcall." Academic Radiology **23**(3): 284-289.
Locker, D. and C. Quiñonez (2011). "To what extent do oral disorders compromise the quality of life?" Community Dent Oral Epidemiol **39**(1): 3-11.
Lofthag-Hansen, S., et al. (2007). "Limited cone-beam CT and intraoral radiography for the diagnosis of periapical pathology." Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontics **103**(1): 114-119.
Loy, C. T. and L. Irwig (2004). "Accuracy of Diagnostic Tests Read With and Without Clinical Information - A Systematic Review." JAMA **292**(13): 1602-1609.
Maizlin, N. N. and S. Somers (2019). "The Role of Clinical History Collected by Diagnostic Imaging Staff in Interpreting of Imaging Examinations." Journal of Medical Imaging and Radiation Sciences **50**(1): 31-35.

McNeil, B. J., et al. (1983). "Paired receiver operating characteristic curves and the effect of history on radiographic interpretation - CT of the head as a case study." Radiology **149**(1): 75-77.
Millet, I., et al. (2013). "Acute Abdominal Pain in Elderly Patients: Effect of Radiologist Awareness of Clinicobiologic Information on CT Accuracy." American Journal of Roentgenology **201**(6): 1171-1179.
Patel, S., et al. (2012). "The detection of periapical pathosis using periapical radiography and cone beam computed tomography - Part 1: pre-operative status." Int Endod J **45**(8): 702-710.
Persoon, I. F. and A. R. Özok (2017). "Definitions and Epidemiology of Endodontic Infections." Current oral health reports **4**(4): 278-285.
Pope, O., et al. (2014). "A Comparative Investigation of Cone-beam Computed Tomography and Periapical Radiography in the Diagnosis of a Healthy Periapex." J Endod **40**(3): 360-365.
Quekel, L. G. B. A., et al. (2001). "Detection of lung cancer on the chest radiograph: impact of previous films, clinical information, double reading, and dual reading." Journal of Clinical Epidemiology **54**(11): 1146-1150.
Reit, C. and H. G. Gröndahl (1983). "Application of statistical decision theory to radiographic diagnosis of endodontically treated teeth." Scand J Dent Res **91**(3): 213-218.
Robinson, J. W., et al. (2016). "Reporting instructions significantly impact false positive rates when reading chest radiographs." European radiology **26**(10): 3654-3659.
Roelofs, A. J., et al. (2007). "Importance of Comparison of Current and Prior Mammograms in Breast Cancer Screening." Radiology **242**(1): 70-77.
Saba, L., et al. (2019). "Does the clinical information play a role in the magnetic resonance diagnostic confidence analysis of ovarian and deep endometriosis?" British Journal of Radiology **92**(1096).
Schreiber, M. H. (1963). "The Clinical History as a Factor in Roentgenogram Interpretation." JAMA **185**(5): 399-401.
Soh, B. P., et al. (2014). "Mammography test sets: Reading location and prior images do not affect group performance." Clinical Radiology **69**(4): 397-402.
Song, K. S., et al. (1992). "Impact of clinical history on film interpretation." Yonsei Med J **33**(2): 168-172.
Starr, S. J., et al. (1975). "Visual Detection and Localization of Radiographic Images." Radiology **116**(3): 533-538.
Stheeman, S. E., et al. (1995). "Use of the Delphi technique to develop standards for quality assessment in diagnostic radiology." Community Dent Health **12**(4): 194-199.
Tudor, G. R., et al. (1997). "An assessment of inter-observer agreement and accuracy when reporting plain radiographs." Clinical Radiology **52**(3): 235-238.
White, S. C. and M. J. Pharoah (2013). Oral Radiology : Principles and Interpretation. St Louis, UNITED STATES, Elsevier.

CHAPTER 2

2.1 Publication

This is the first of two systematic reviews in this thesis.

This chapter is a systematic review comparing the diagnostic performance of CBCT and digital PA radiography in endodontic disease detection. It is published in the Journal of Medical Imaging as:

Yapp KE, Brennan P, Ekpo E. Endodontic disease detection: digital periapical radiography versus cone-beam computed tomography—a systematic review. Journal of Medical Imaging 2021, 8(4):041205.
https://doi.org/10.1117/1.JMI.8.4.041205

Endodontic disease detection: digital periapical radiography versus cone-beam computed tomography—a systematic review

Kehn E. Yapp,* **Patrick Brennan, and Ernest Ekpo**
The University of Sydney, School of Health Sciences, Medical Image Optimisation and Perception Group, Discipline of Medical Imaging Science, Faculty of Medicine and Health, Australia

Abstract

Purpose: To assess the comparative diagnostic performance of digital periapical (PA) radiography and cone-beam computed tomography (CBCT) imaging on endodontic disease detection and to provide study methodology and design recommendations for future studies comparing the diagnostic performance of imaging modalities on endodontic disease detection.

Approach: A search of the Medline, Embase, Scopus, Web of Science, and the Cochrane Central Register of Controlled Trials databases was conducted. Studies that compared the performance of CBCT to digital PA radiography for detecting endodontic disease had an independent reference standard determining the presence of endodontic disease and conducted data analysis including either sensitivity, specificity, receiver operating characteristic (ROC) analysis or free response operating characteristic analysis were included. Of the 20,530 identified studies, only 3 fulfilled the inclusion criteria.

Results: Most studies assessed for eligibility were excluded due to limitations and biases in study design 15 of 18 studies had no reference standard. Only one retrospective clinical study reported on the diagnostic performance of CBCT and showed a sensitivity of 86% and specificity of 26%. Two cadaver studies reported sensitivity ranging from 60% to 100%, specificity ranging from 79% to 100%, and an area under the ROC curve of 0.943 for CBCT. The reported sensitivity for digital PA radiography ranged from 27% to 60%, specificity was 99%, and the area under the ROC curve was 0.629.

Conclusions: There is a lack of quality evidence and insufficient data to compare diagnostic performance of digital PA and CBCT imaging. This emphasizes the need for well-designed studies to inform clinicians about the relative diagnostic performance of these imaging modalities.

 [DOI: 10.1117/1.JMI.8.4.041205]

Keywords: cone-beam computed tomography; diagnostic performance; periapical radiography.

Paper 20321SSR received Dec. 3, 2020; accepted for publication Jan. 28, 2021; published online Feb. 24, 2021.

1 Introduction

Endodontic disease prevalence has been reported to range from 7% to 86%,[1] and it is estimated that 22 million endodontic procedures are performed annually in the United States of America.[2] Prior to these procedures, dental imaging is required not only for diagnostic, but also for medico-legal and treatment planning purposes.[3] Diagnosis of dental and endodontic abnormalities follows a Bayesian approach just like in medicine patient history and examination data are gathered to generate pre-test odds (prior probability) of a disease being present. This is multiplied by the weight of new testing information (likelihood ratio) that generates post-test odds (posterior probability) of the disease being present.[4] Dental imaging has historically used intra-oral and extra-oral diagnostic radiographs, with an early cadaver study showing the limitations of radiographs in

*Address all correspondence to Kehn E. Yapp, keyapptdo@gmail.com

showing simulated pathologic changes in cancellous bone; these radiolucent changes could be detected radiographically only if there was cortical bone perforation.[5] A later clinical study showed that periapical (PA) radiography had high diagnostic value in endodontic disease detection.[6]

Medical imaging is constantly evolving; three-dimensional (3D) cone-beam computed tomography (CBCT) has been recently introduced into the clinical dental setting and is gaining popularity.[7] Newer imaging modalities can be considered fit for purpose if their diagnostic performance is comparable to, or better than, current modalities. Diagnostic efficacy analysis is therefore needed to establish the diagnostic ability of these new imaging tools.[8] Since the introduction of CBCT into the clinical evaluation of endodontic diseases, several studies have attempted to investigate its diagnostic efficacy compared to two-dimensional PA radiography. They showed differences in the diagnostic performance of these modalities; however, these studies differ by design and results. For example, a majority of the studies was based on imaging examination records of PA, panoramic, and CBCT imaging and included patients with different presentations of endodontic infections or patients referred to a specialist endodontic practice for endodontic treatment, with a consensus opinion of a panel being used to establish the presence of disease.[9] Second, most of these studies did not perform an evaluation of CBCT using an established independent reference standard.[10] Third, many of the studies were based on conventional PA radiography and either assessed the agreement between CBCT and PA reporting or PA and panoramic image reporting. Most utilized only images with disease, which limited the calculation of diagnostic performance metrics such as specificity and false positive rates. Some of the published studies used CBCT as a "reference standard" to assess the sensitivity of PA imaging. These differences in methodologies and result frameworks emphasize the need for a review of the literature to understand the diagnostic efficacy of CBCT relative to PA radiography.

Previous systematic reviews comparing CBCT and PA radiography in endodontic disease detection[11–15] were mostly based on plain film PA radiography, included studies that assessed the agreement between both imaging modalities or used cadaver findings as a reference standard. Some used an artificial reference standard: "mechanically or chemically induced lesions," which does not establish the truth about endodontic disease presence or absence.[12,14] Two of these reviews, which focused on the diagnostic efficacy of CBCT and PA radiography using a hierarchical model, reported that the diagnostic efficacy was unclear[13] and that human CBCT studies using a histological reference standard were needed.[12] A meta-analysis of *ex-vivo* studies with artificial apical periodontitis found CBCT imaging had a greater area under the receiver operating characteristic (ROC) curve than PA radiography.[14] A more recent review showed that the odds ratio of CBCT detecting endodontic disease was double that for PA radiography[15]. However, these reviews have some limitations: the diagnostic performance of digital PA and CBCT imaging were not compared directly,[11–13,15] *ex-vivo* studies were included, which limit the external validity of these findings[11,12,14] and therapeutic efficacy rather than diagnostic accuracy was evaluated.[12,13] Therefore, the comparative diagnostic performance of these imaging modalities in the endodontic domain is poorly understood.

In addition, digitization has improved the quality of radiological images and allowed post-processing of acquired images to suit different diagnostic tasks. In dentistry, digitization of the imaging process has been shown to improve image quality, which may optimize the detection of dental caries and assessment of bony anomalies.[16] Thus, a review of the literature on the diagnostic performance of CBCT relative to digital PA radiography in the digital era will provide informed choices of imaging options for patients and clinicians requesting dental imaging. This review aims to assess the comparative diagnostic performance of digital PA and CBCT imaging on endodontic disease detection and to provide study methodology and design recommendations for future studies comparing the diagnostic performance of imaging modalities on endodontic disease detection.

2 Methods

2.1 *Databases and Search Strategy*

The literature search was conducted based on the Preferred Reporting Items for a Systematic Review and Meta-Analysis of Diagnostic Testing Studies (PRISMA-DTA) statement.[17]

Medline, Embase, Scopus, Web of Science, and the Cochrane Central Register of Controlled Trials databases were searched for relevant articles published from database inception to January 12, 2021. Google Scholar was also used to search for relevant articles and the reference lists of published articles were manually screened to identify additional publications. Search terms were combined with "OR" and included the following main terms: "Cone beam computed tomography" OR "cone beam" OR "periapical radiography" OR "periapical" OR "endodontics" OR "pulp disease" OR "apical periodontitis" OR "periapical disease" OR "periapical lesion" OR "endodontic pathosis" OR "apical pathology" OR "apical radiolucency" OR "receiver operating characteristic" OR "free response."

2.2 *Eligibility Criteria*

Inclusion and exclusion criteria were based on the population, intervention, comparator, and outcome (PICO) elements (Table 1). The clinical research question we sought to address was, in permanent human teeth, does CBCT have greater diagnostic performance in endodontic disease detection than PA radiography? Studies were included if they: compared the performance of CBCT to digital PA radiography for detecting endodontic disease, included humans with permanent teeth, had an independent reference standard determining the presence of endodontic disease, conducted data analysis including at least one of the following outcomes: sensitivity, specificity, ROC analysis, or free response operating characteristic (FROC) analysis, and were published in English. Studies were excluded if they did not meet these inclusion criteria. Literature reviews, conference papers, letters to editors, and posters were also excluded. Initial triage of the abstracts was performed by two authors (K.Y. and E.E.). Disagreements were resolved by objectively evaluating the inclusion and exclusion criteria and establishing a consensus.

2.3 *Quality Assessment*

Quality assessment was performed by two authors (K.Y. and E.E.) using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool.[18] It consists of four main domains: patient selection, index test, reference standard, and flow and timing of the index tests and reference standard. The QUADAS-2 tool is mainly recommended for judging the risk of bias and the applicability of original diagnostic accuracy studies. A weighted kappa was used to assess the agreement between the two assessors. Kappa was interpreted as follows: <0.20 = poor; 0.21 to 0.40 = fair; 0.41 to 0.60 = moderate; 0.61 to 0.80 = substantial; and 0.81 to 0.99 = almost perfect.[19] Any discrepancies in the quality assessment were discussed and resolved through consensus.

2.4 *Data Extraction Process*

Data were extracted in two phases. First, the authors determined the study characteristics (e.g., study design, reported outcome measures, provision of clinical history, and recruitment method for patients and readers), population characteristics (e.g., sample size, disease prevalence, and distribution of disease severity), reader characteristics (observer clinical experience, CBCT experience, and qualifications), and interpretation protocol. Second, the diagnostic performance of

Table 1 The PICO method regarding inclusion and exclusion criteria.

Element	Characteristics
Population	Permanent human teeth
Intervention	CBCT imaging
Comparator	Digital PA radiography
Outcome	Diagnostic performance in endodontic disease detection: ROC curve analysis, FROC analysis, sensitivity, and specificity

PA imaging was compared to CBCT. The performance metrics analyzed were number and location of detected abnormalities, ROC curve construction, relationship between true positive fraction (TPF) at a given false positive fraction (FPF), area under the ROC curve, FROC analysis, and diagnostic accuracy measures such as jackknife FROC figure of merit, sensitivity, and specificity. All authors reviewed the full text articles and any discrepancies regarding data analysis or interpretation were resolved by objectively evaluating the reported findings and establishing a consensus.

3 Results

3.1 *Identification of Included Studies*

The search strategy identified a total of 20,530 studies. After the screening of titles and abstracts, 18 studies were selected for full-text reading (Fig. 1). Only three studies fulfilled the inclusion criteria.[20–22] One study used clinical information[20] and two used data from cadavers[21,22] as the reference standard to assess the diagnostic performance of CBCT and digital PA radiography (Table 2). Fifteen studies that were identified to have examined disease detection using CBCT and digital PA radiography but were excluded are summarized in Table 3. These studies were

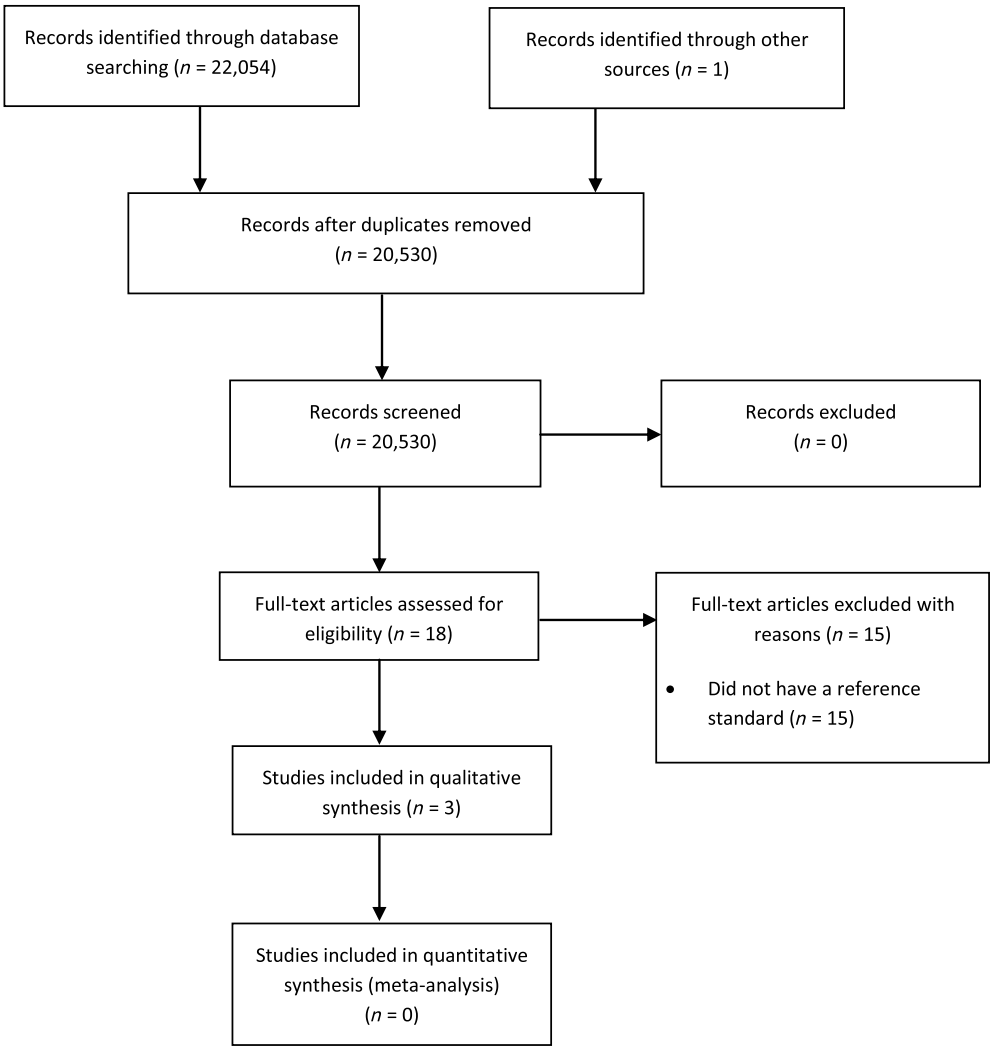


Fig. 1 A Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement flowchart of the search and selection strategy.

Table 2 Characteristics of included studies.

Study	Sample size (teeth)	Participant characteristics	Disease prevalence	Index test	Reference standard	Readers	Disease severity distribution	Performance metrics provided
Pope et al.[20]	200	Retrospective record analysis from endodontic practices	7.8% (14/180)	Diameter of PA radiolucency using modified CBCT-PA index (PAI) scale	Clinical information	Two: one endodontist and one endodontic postgraduate student	No	CBCT sensitivity (86%) and specificity (26% to 80%)
Kanagasingam et al.[21]	67	Cadavers	Not provided	Presence of a "PA lesion"	Histopathology findings	Five endodontists	No	Sensitivity, specificity, and area under the ROC curve. CBCT (89%, 100%, and 0.943), PA (27%, 99%, and 0.629)
Kruse et al.[22]	222	Cadavers	Not provided	Presence of apical periodontitis using a 5-point rating scale	Histopathology findings	Three: two endodontists and one oral radiologist	No	CBCT and PA sensitivity (80%, 60%) and CBCT specificity (79%)

Table 3 Characteristics of excluded studies.

Study	Sample size (teeth)	Disease prevalence	Index test	Reference standard	Readers	Performance metrics provided
Lofthag-Hansen et al [23]	46	100%	Consensus report of a PA lesion	No	Three oral and maxillofacial radiologists	None
Estrela et al [10]	1508	100%	PAI score	No	Three "calibrated examiners"	None
Low et al [24]	74	100%	Consensus report of a PA lesion	No	Two one oral radiologist and one endodontist	None
Lennon et al [25]	10	100%	Presence of artificial bone lesions using a 5-point rating scale	No	Ten two endodontists two dental radiologists and six postgraduate endodontic students	None
Abella et al [26]	138	100%	Consensus report of an "apical periodontitis lesion"	No	Two endodontists	Number of "lesions" seen on PA and CBCT
Patel et al [9]	151	100%	Consensus report of an apical periodontitis lesion	No	Two endodontists	Number of lesions seen on PA and CBCT
Abella et al [27]	161	100%	Consensus report of an apical periodontitis lesion	No	Two endodontists	Number of lesions seen on PA and CBCT
Venskutonis et al [28]	35	Not provided	Consensus report of a PA lesion	No	Two endodontists	None
Bornstein et al [29]	58	100%	Report of either "cyst" or "granuloma"	No	Four two oral surgeons and two oral surgery residents	Number of radiographic reports designated as granuloma or cyst
Davies et al [30]	100	100%	Consensus report of a PA lesion	No	Two endodontists	Number of roots with a PA lesion detected
Weissman et al [31]	67	Not provided	Presence of apical radiolucency	No	Three two endodontists and one oral and maxillofacial radiologist	Number of lesions seen on PA and CBCT
Davies et al [32]	98	Not provided	Consensus report on the change in PA status at review	No	Two endodontists	Healing or non-healing category
Beacham et al [33]	18 imaging studies	Not provided	Report on the location of any finding considered "notable or important"	No	Nine four endodontists and five endodontic residents	Number of radiographic findings assigned by an "expert reviewer" that were identified by the observer
Kruse et al [34]	74	Not provided	Consensus score determining the level of healing and the treatment plan	No	Three two endodontists and one oral radiologist	Change in treatment plan based on CBCT report
Chang et al [35]	68 imaging studies	Not provided	Presence of a PA lesion	No	Two one endodontist and one oral and maxillofacial radiologist	Number of lesions seen on PA and CBCT

Table 4 QUADAS-2 tool results.

	Risk of bias				Applicability concerns		
Study	Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Pope et al.[20]	High	Low	Low	Uncertain	Low	High	Low
Kanagasingam et al.[21]	Uncertain	Low	Uncertain	Uncertain	Uncertain	High	Uncertain
Kruse et al.[22]	Uncertain	Low	Uncertain	Uncertain	Uncertain	High	Uncertain

excluded for the following reasons: none had an independent reference standard, either did not provide information about disease prevalence or used only images with disease limiting the assessment of other diagnostic performance metrics, assessed agreement between CBCT and PA radiography reports, or did not provide diagnostic performance metrics. None of these studies accounted for case difficulty and severity of diseased or non-diseased patients.

3.2 *Quality Assessment*

The three included studies had a different risk of bias and a range of applicability concerns regarding patient selection and reference standard. For all three, the risk of bias about the index test was low and risk for flow and timing was uncertain. All studies had high applicability concerns about the index test. The quality assessment results are summarized in Table 4. Inter-reader agreement between the two assessors of quality showed a weighted kappa of $k = 0.92$, 95% CI: 0.767 to 1.000.

3.3 *Diagnostic Performance of Periapical Radiography versus Cone-Beam Computed Tomography*

When clinical information was used as the reference standard and disease was considered to be having a PA radiolucency with diameter >0.5 mm, CBCT had a sensitivity of 86% (12/14).[20] When non-disease was considered to be an intact PA bone structure, specificity was 26% (43/166). If the threshold for non-disease was PA radiolucency with diameter no greater than 1 mm, CBCT specificity was 80% (133/166). Although both modalities were compared with different methods of analysis, no data were reported on the diagnostic performance of PA radiography.

For cadaver studies, sensitivity of CBCT for detecting endodontic disease ranged from a mean of 89%[21] to 80% (66/83) when individual roots were calculated with no figures provided for teeth.[22] Sensitivity of PA radiography was reported to range between a mean of 27%[21] and 60% (134/223) for individual roots.[22] Specificity of CBCT varied from 79%[22] to 100%[21] when individual roots were calculated. PA radiography had a reported specificity mean of 99%.[21] When ROC data were given, the area under the curve values were 0.629 for PA radiography and 0.943 for CBCT.[21] Data on the relationship between true and false positive fractions the sensitivity at a given specificity, and vice versa, were not provided.

The three included studies each had different index tests and displayed methodological heterogeneity in reporting measures. The cadaver histology results used the presence of inflammation as the reference standard; however, the relationship to disease in humans was not shown. Due to the methodological heterogeneity, meta-analysis was not performed.

4 Discussion

The analysis shows that there is a lack of high-level evidence, with notable uncertainty about the study quality and bias, regarding the diagnostic performance of digital PA radiography and CBCT for endodontic disease detection. The purpose of evaluating diagnostic performance

is to determine diagnostic accuracy efficacy[8] so that the truth in yielding an abnormal or normal diagnosis can be ascertained. Evaluation at this level is clinically relevant as it forms part of the framework in a model of understanding decision making.[8] Without the truth value of the test being evaluated, diagnostic performance is unknown.[36] Only one study provided sensitivity, specificity, and area under the ROC curve metrics for both modalities,[21] using cadaver samples. Of the identified studies, there were issues with study design that limit the external validity of the available data. A major finding in studies that have examined endodontic disease detection using CBCT and PA radiography was that they were designed with the aim of observers identifying radiographic findings, such as PA radiolucency, which may not always be pathognomonic for disease.[37] In contrast, endodontic disease such as irreversible pulpitis can occur in the absence of PA radiolucency[26] and the presence or absence of a PA radiolucency as a surrogate measure for disease has not been shown to be a relevant or valid proxy. Instead of using these reported test indices, observers should be rating their confidence in the presence of an abnormality.[38]

Population sampling across studies on CBCT and PA radiography has been skewed to contain only diseased cases. The only retrospective clinical study that fulfilled inclusion criteria had a very low disease prevalence,[20] with cadaver studies focusing on roots, not teeth, having unknown disease prevalence.[21,22] The main limitation of this skewed sampling strategy is that indices of test accuracy calculated in one patient group cannot be generalized to other groups if they show different clinical spectra.[39] It should be noted that the rationale for assessing performance of a diagnostic system using a sample of cases, observers, and readings is to provide an estimate of how the imaging system would perform "on the average" in those similar cases and observers and readings that were not studied.[40] Therefore, it is important that a diagnostic test performance study encompasses cases with a wide distribution of clinical features and includes a broad range of patients both with and without the disease.[41] Such inclusion criteria provide opportunities to assess other diagnostic metrics, considering the variability in population characteristics and disease conditions encountered in the clinical setting. The exclusion of patients with a specific condition or high prevalence may influence observer interaction with the images and lead to inflated diagnostic accuracy estimates,[36] particularly when cases with presenting diagnostic difficulty are excluded.[18] Low-quality studies with a non-representative sample have a tendency to overestimate the diagnostic performance of a test.[42] A test-set containing a wide distribution of cases with different levels of difficulty is needed to provide representation of the variation in the clinical setting,[43] where diagnostic performance decreases as disease findings become more subtle.[44] All three included studies did not report on their population case spectrum and their ability to extrapolate findings to the broader population is unknown. Only one study reported on sensitivity and specificity of both modalities using histopathology findings as the reference standard[21] and found that sensitivity was higher for CBCT compared to PA radiography (89% and 27%, respectively), with no significant change in specificity (100% and 99%, respectively). An animal study[45] with a similar study design also found CBCT had higher sensitivity (91%) than PA imaging (77%) with no difference in specificity (both 100%). Because the disease severity in both studies was not reported, it is unknown which case types of disease or non-disease these results apply to. Future studies should include cases with a range of severity in both diseased and non-diseased patients.[44]

Intrinsic human limitations can influence the diagnostic performance of imaging modalities and there are variations in the human ability to interpret radiological images. Diagnostic accuracy efficacy is not just a function of the image, it is a joint function of the images and of an observer.[8] Reader variability has been shown in previous endodontic studies on PA radiography[46,47] and CBCT.[48] Therefore, studies assessing diagnostic image performance should include a significant number of readers. The number of observers in the identified studies had a tendency to be low and most ranged from 2 to 5, with two studies having nine and ten readers, respectively. Given that every case was read once, or a consensus report was used, the number of total opinions used to establish the performance of CBCT relative to PA radiography was low. The excluded studies also suffered from low observer numbers. In other radiology domains, certain factors such as training[49] and number of annual cases read[50] have been shown to be associated with diagnostic performance but no study has explored how these factors affect diagnostic performance in digital PA radiography and CBCT interpretation. No information was provided on reader experience and expertise on diagnostic performance, which is of clinical

relevance; diagnostic accuracy has been shown to increase with reader experience.[51] Therefore, future studies should account for variation in reader characteristics that affect diagnostic performance. Importantly, endodontic CBCT images are interpreted by dentists and radiologists.[7] A comparison of the diagnostic performance of these professionals and the factors that impact their performance would help provide informed strategies for improving diagnostic efficacy of dental imaging interpretation.

Across the literature on endodontic disease imaging, there is a lack of an independent and valid reference standard for assessing the performance of CBCT or PA radiography. The reference standard is needed to establish the truth about disease presence or absence and to measure sensitivity and specificity of these imaging technologies;[52] without it, the true test results are unknown.[36] Biopsy has been used as a reference standard in medicine;[50] however, for endodontic disease, histology results do not have the same level of dichotomy. Inflammatory cells in the PA tissues have been shown to be present for healed teeth[53] and while histological findings are independent, the presence of inflammation does not necessarily indicate disease presence and has not been shown to be a valid reference standard for disease. Furthermore, the use of cadaver histology is limited due to the lack of clinical evidence to corroborate histological findings in endodontics. An example of a valid and independent reference standard in radiology studies is the Delphi panel, where examination and follow-up clinical information is given to a consensus panel who collectively determine the presence or absence of disease.[54] It should be emphasized that these panelists are not involved in the reporting of images in the study. A Delphi panel approach should be used as a reference standard for future dental diagnostic imaging studies.

Observer performance measurement in the identified studies demonstrated significant limitations. The assigned observer task was to report on a type of radiographic finding without indicating its location. This is inconsistent with previous teachings from medical radiology that an observer's task is to not only to detect but also to locate the abnormality.[55] When the decision task involves more than just a determination of whether the patient is diseased or non-diseased, the bivariate ROC method has significant limitations in assessing diagnostic efficacy.[56] To inform treatment interventions on the correct tooth, the exact location of the disease is required. Therefore, inclusion of location information in dental imaging studies is important. Without location assignment on images, errors can be disguised as correct calls.[57] This is overcome by reporting using the free response paradigm; on an image, the reader reports a "mark" a region suspicious for abnormality, and assigns a "rating" the corresponding confidence level.[38] This search paradigm accounts for ambiguities that can occur unnoticed in the ROC paradigm, such as when a location-level false positive and a location-level false negative occur on the same image.[57] In this situation, ROC analysis provides an image-level true positive for the wrong reasons; an incorrect abnormality location was reported and an abnormality missed. Without a free response analysis, these potentially significant errors are overlooked. For this reason, future dental diagnostic imaging studies should use the free response paradigm.

This review has highlighted the limitations of the current literature on assessing diagnostic performance of dental imaging modalities, identified methodological issues, and provided examples of study designs to address these limitations. The lack of sensitivity, specificity, and relationship between TPF and FPF data in published studies emphasize the need for further studies to establish the diagnostic efficacy of CBCT relative to digital PA radiography. Only three studies were included, which further highlights the need for properly designed studies comparing digital PA and CBCT imaging in endodontic disease detection. In particular, future studies need to overcome the limitations of the existing studies and avoid repeating the errors previously made, in order to provide valid and relevant data that can help improve clinical decision making. Without further research, the comparative performance of these endodontic imaging modalities and the factors that influence their diagnostic efficacy cannot be determined.

5 Conclusion

There is a lack of evidence to establish the diagnostic performance of digital PA radiography relative to CBCT in endodontic disease detection. Well-designed studies are required in order to inform clinicians about the diagnostic performance of commonly used digital imaging modalities

in detection of endodontic disease. These should reflect the task in clinical practice, use a valid reference standard, allow for measuring multiple abnormalities per image, include localization of abnormalities, reward correct and penalize incorrect abnormality locations, and encompass the entire spectrum of disease and non-disease severity present within the study population.

Disclosures

None.

References

1. I. F. Persoon and A. R. Özok, "Definitions and epidemiology of endodontic infections," *Curr. Oral Health Rep.* **4**(4), 278 285 (2017).
2. "2005 2006 survey of dental services rendered," American Dental Association (2007).
3. M. Alrahabi, M. S. Zafar, and N. Adanir, "Aspects of clinical malpractice in endodontics," *Eur. J. Dent.* **13**(3), 450 458 (2019).
4. C. J. Gill, L. Sabin, and C. H. Schmid, "Why clinicians are natural Bayesians," *BMJ* **330**(7499), 1080 1083 (2005).
5. I. B. Bender and S. Seltzer, "Roentgenographic and direct observation of experimental lesions in bone. I," *J. Am. Dent. Assoc.* **62**(2), 152 160 (1961).
6. C. Reit and H. G. Gröndahl, "Application of statistical decision theory to radiographic diagnosis of endodontically treated teeth," *Scand. J. Dent. Res.* **91**(3), 213 218 (1983).
7. L. F. Brown and P. Monsour, "The growth of Medicare rebatable cone beam computed tomography and panoramic radiography in Australia," *Aust. Dent. J.* **60**(4), 511 519 (2015).
8. D. G. Fryback and J. R. Thornbury, "The efficacy of diagnostic imaging," *Med. Decis. Making* **11**(2), 88 94 (1991).
9. S. Patel et al., "The detection of periapical pathosis using periapical radiography and cone beam computed tomography Part 1: pre-operative status," *Int. Endod. J.* **45**(8), 702 710 (2012).
10. C. Estrela et al., "Accuracy of cone beam computed tomography and panoramic and periapical radiography for detection of apical periodontitis," *J. Endod.* **34**(3), 273 279 (2008).
11. A. Petersson et al., "Radiological diagnosis of periapical bone tissue lesions in endodontics: a systematic review," *Int. Endod. J.* **45**(9), 783 801 (2012).
12. C. Kruse et al., "Cone beam computed tomography and periapical lesions: a systematic review analysing studies on diagnostic efficacy by a hierarchical model," *Int. Endod. J.* **48**(9), 815 828 (2015).
13. E. Rosen et al., "The diagnostic efficacy of cone-beam computed tomography in endodontics: a systematic review and analysis by a hierarchical model of efficacy," *J. Endod.* **41**(7), 1008 1014 (2015).
14. K. L. Dutra et al., "Diagnostic accuracy of cone-beam computed tomography and conventional radiography on apical periodontitis: a systematic review and meta-analysis," *J. Endod.* **42**(3), 356 364 (2016).
15. A. Aminoshariae, J. C. Kulild, and A. Syed, "Cone-beam computed tomography compared with intraoral radiographic lesions in endodontic outcome studies: a systematic review," *J. Endod.* **44**(11), 1626 1631 (2018).
16. P. F. van der Stelt, "Better imaging: the advantages of digital radiography," *J. Am. Dent. Assoc.* **139**, S7 S13 (2008).
17. M. D. F. McInnes et al., "Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement," *JAMA* **319**(4), 388 396 (2018).
18. P. F. Whiting et al., "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies," *Ann. Intern Med.* **155**(8), 529 36 (2011).
19. A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," *Fam. Med.* **37**(5), 360 363 (2005).

20. O. Pope, C. Sathorn, and P. Parashos, "A comparative investigation of cone-beam computed tomography and periapical radiography in the diagnosis of a healthy periapex," *J. Endod.* **40**(3), 360 365 (2014).
21. S. Kanagasingam et al., "Diagnostic accuracy of periapical radiography and cone beam computed tomography in detecting apical periodontitis using histopathological findings as a reference standard," *Int. Endod. J.* **50**(5), 417 426 (2017).
22. C. Kruse et al., "Diagnostic accuracy of cone beam computed tomography used for assessment of apical periodontitis: an *ex vivo* histopathological study on human cadavers," *Int. Endod. J.* **52**(4), 439 450 (2019).
23. S. Lofthag-Hansen et al., "Limited cone-beam CT and intraoral radiography for the diagnosis of periapical pathology," *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* **103**(1), 114 119 (2007).
24. K. M. T. Low et al., "Comparison of periapical radiography and limited cone-beam tomography in posterior maxillary teeth referred for apical surgery," *J. Endod.* **34**(5), 557 562 (2008).
25. S. Lennon et al., "Diagnostic accuracy of limited-volume cone-beam computed tomography in the detection of periapical bone loss: 360 degree scans versus 180 degree scans," *Int. Endod. J.* **44**(12), 1118 1127 (2011).
26. F. Abella et al., "Evaluating the periapical status of teeth with irreversible pulpitis by using cone-beam computed tomography scanning and periapical radiographs," *J. Endod.* **38**(12), 1588 1591 (2012).
27. F. Abella et al., "An evaluation of the periapical status of teeth with necrotic pulps using periapical radiography and cone- beam computed tomography," *Int. Endod. J.* **47**(4), 387 396 (2014).
28. T. Venskutonis et al., "Accuracy of digital radiography and cone beam computed tomography on periapical radiolucency detection in endodontically treated teeth," *J. Oral Maxillofac. Res.* **5**(2), e1 e1 (2014).
29. M. M. Bornstein et al., "Comparison between radiographic (2-dimensional and 3-dimensional) and histologic findings of periapical lesions treated with apical surgery," *J. Endod.* **41**(6), 804 811 (2015).
30. A. Davies et al., "The detection of periapical pathoses in root filled teeth using single and parallax periapical radiographs versus cone beam computed tomography a clinical study," *Int. Endod. J.* **48**(6), 582 592 (2015).
31. J. Weissman et al., "Association between the presence of apical periodontitis and clinical symptoms in endodontic patients using cone-beam computed tomography and periapical radiographs," *J. Endod.* **41**(11), 1824 1829 (2015).
32. A. Davies et al., "The detection of periapical pathoses using digital periapical radiography and cone beam computed tomography in endodontically retreated teeth part 2: a 1 year post-treatment follow-up," *Int. Endod. J.* **49**(7), 623 635 (2016).
33. J. T. Beacham et al., "Accuracy of cone-beam computed tomographic image interpretation by endodontists and endodontic residents," *J. Endod.* **44**(4), 571 575 (2018).
34. C. Kruse et al., "Impact of cone beam computed tomography on periapical assessment and treatment planning five to eleven years after surgical endodontic retreatment," *Int. Endod. J.* **51**(7), 729 737 (2018).
35. L. Chang et al., "Periradicular lesions in cancellous bone can be detected radiographically," *J. Endod.* **46**(4), 496 501 (2020).
36. J. F. Cohen et al., "STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration," *BMJ Open* **6**(11), e012799 (2016).
37. O. Molven et al., "Periapical changes following root-canal treatment observed 20 27 years postoperatively," *Int. Endod. J.* **35**(9), 784 790 (2002).
38. D. P. Chakraborty, "A brief history of free-response receiver operating characteristic paradigm data analysis," *Acad. Radiol.* **20**(7), 915 919 (2013).
39. I. A. Scott, P. B. Greenberg, and P. J. Poole, "Cautionary tales in the clinical interpretation of studies of diagnostic tests," *Intern Med. J.* **38**(2), 120 129 (2008).
40. J. A. Hanley, "Receiver operating characteristic (ROC) methodology: the state of the art," *Crit. Rev. Diagn. Imaging* **29**(3), 307 335 (1989).

41. D. F. Ransohoff and A. R. Feinstein, "Problems of spectrum and bias in evaluating the efficacy of diagnostic tests," *N. Engl. J. Med.* **299**(17), 926 930 (1978).
42. J. G. Lijmer et al., "Empirical evidence of design-related bias in studies of diagnostic tests," *JAMA* **282**(11), 1061 1066 (1999).
43. J. T. Philbrick, R. I. Horwitz, and A. R. Feinstein, "Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias," *Am. J. Cardiol.* **46**(5), 807 812 (1980).
44. J. Khademi, "The effects of digitization on diagnostic performance in endodontics and periodontics: an ROC study," The University of Iowa, Iowa (1994).
45. F.W.G. de Paula-Silva et al., "Accuracy of periapical radiography and cone-beam computed tomography scans in diagnosing apical periodontitis using histopathological findings as a gold standard," *J. Endod.* **35**(7), 1009 1012 (2009).
46. M. Goldman, A. H. Pearson, and N. Darzenta, "Endodontic success who's reading the radiograph?" *Oral. Surg. Oral Med. Oral Pathol.* **33**(3), 432 437 (1972).
47. D. D. Antrim, "Reading the radiograph: a comparison of viewing techniques," *J. Endod.* **9**(11), 502 505 (1983).
48. J. M. Parker et al., "Cone-beam computed tomography uses in clinical endodontics: observer variability in detecting periapical lesions," *J. Endod.* **43**(2), 184 187 (2017).
49. C. Buissink et al., "The influence of experience and training in a group of novice observers: a jackknife alternative free-response receiver operating characteristic analysis," *Radiography* **20**(4), 300 305 (2014).
50. M. Rawashdeh et al., "Markers of good performance in mammography depend on number of annual readings," *Radiology* **269**(1), 61 67 (2013).
51. P. Kasprowski, K. Harezlak, and S. Kasprowska, "Development of diagnostic performance and visual processing in different types of radiological expertise," in *Proc. ACM Symp. Eye Tracking Res. and Appl.*, Association for Computing Machinery, Warsaw, Poland, p. 40 (2018).
52. C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.* **8**(4), 283 298 (1978).
53. A. Khayat, "Histological observations of periradicular healing following root canal treatment," *Aust. Endod. J.* **31**(3), 101 105 (2005).
54. S. E. Stheeman et al., "Use of the Delphi technique to develop standards for quality assessment in diagnostic radiology," *Commun. Dent. Health* **12**(4), 194 199 (1995).
55. C. E. Metz, "ROC analysis in medical imaging: a tutorial review of the literature," *Radiol. Phys. Technol.* **1**(1), 2 12 (2008).
56. D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: modeling, analysis, and validation," *Med. Phys.* **31**(8), 2313 2330 (2004).
57. P. Bunch et al., "A free response approach to the measurement and characterization of radiographic observer performance," *Proc. SPIE* **0127**, 124 135 (1977).

Biographies of the authors are not available.

CHAPTER 3

3.1 Publication

The first systematic review focused on the diagnostic performance of CBCT and digital PA radiography in endodontic disease detection. Because this thesis also examined the effect of clinical history on CBCT interpretation, this second systematic review explores the literature on the effect of clinical history on interpretation of all types of diagnostic imaging.

This chapter is a systematic review on the effect of clinical history on diagnostic image interpretation. It is published in Academic Radiology as:

Yapp KE, Brennan P, Ekpo E. The Effect of Clinical History on Diagnostic Imaging Interpretation – A Systematic Review. Academic Radiology 2022, 29(2):255-266.
https://doi.org/10.1016/j.acra.2020.10.021

The Effect of Clinical History on Diagnostic Imaging Interpretation – A Systematic Review

Kehn E Yapp, BDSc, DCD (Endo), Patrick Brennan, PhD, Ernest Ekpo, PhD

Rationale and Objectives: To provide updated information on the effect of clinical history on diagnostic image interpretation and to provide study methodology and design recommendations for future studies assessing the effect of clinical history on diagnostic image performance.

Materials and Methods: A literature search of Medline, Embase, Scopus, Web of Science, and the Cochrane Central Register of Controlled Trials (CENTRAL) databases was conducted from database inception to July 21, 2020. Studies comparing diagnostic imaging performance with and without clinical history, using observers reading images under both conditions that used an independent reference standard were included.

Results: Twenty-two studies met the inclusion criteria, with 15 showing clinical history improved diagnostic performance. One study reported a decrease in diagnostic performance with clinical history and the remaining six studies found no significant change in performance. Two studies used the free response paradigm with both reporting clinical history increased location sensitivity, decreased specificity and had no overall change in diagnostic performance. The disease spectrum of included cases was largely unreported and a balanced reading design was not used in 19 studies.

Conclusion: Most published studies found that clinical history improved diagnostic performance. More recent studies accounting for abnormality location and multiple abnormalities showed an increase in false positives and no significant change in overall diagnostic performance with clinical history.

Key Words: Clinical history; interpretive bias; free response paradigm; diagnostic performance; imaging interpretation.

INTRODUCTION

In medicine, the diagnostic reasoning process begins with taking a patient's clinical history which, when there is clinical uncertainty, determines the need for additional testing (1). Clinical history accompanying a request for diagnostic imaging is considered the norm in clinical practice and is required to ensure the request is justified and appropriate (2). Having access to a patient's clinical information prior to diagnostic image interpretation has been a debated topic since the 1960s (3). This may be due to the cognitive biases that available history has on subsequent image interpretation, such as anchoring bias – the tendency to focus on salient features in the initial presentation without adjusting this initial impression in light of later information (4).

Clinical history can also be considered a form of attribution bias (5), when a diagnostic conclusion depends on how the clinical information is presented, or framing bias (6), a tendency to be influenced by how a problem is presented. The availability of clinical history prior to diagnostic image interpretation has been reported to vary from one third of radiography, ultrasound, and computed tomography cases (7) to 53% of ultrasound and plain radiography requests (8). In the clinical context, clinical history is important because it establishes the prior probability (pretest odds) of a disease being present and helps to determine if a patient is more suggestive or less suggestive of having a disease (9).

The effect of clinical history on diagnostic efficacy and observer performance has also shown variations, where having history prior to interpreting chest radiographs has led to a significant increase in abnormality detection (10), no difference in diagnostic performance to having no history (11,12) and a mixed effect on accuracy (13). Reading chest radiographs with a specific preconception or search task has been shown to increase false positives compared to without a preconception or task (14). Access to clinical history has been shown to increase the tendency for observers to interpret questionable myelograms or computed tomography images

Acad Radiol 2022; 29:255–266

From the Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging Science, School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Australia. Received October 1, 2020; revised October 21, 2020; accepted October 21, 2020. **Address correspondence to:** K.Y. e mail: keyapptdo@gmail.com

https://doi.org/10.1016/j.acra.2020.10.021

as positive for disease when there was a correlation with clinical findings (15). These variations suggest that the effect of patient history prior to diagnostic image interpretation on diagnostic performance is unclear.

The most recent systematic review on the impact of clinical history on diagnostic performance was published in 2004 (16); this review reported on the accuracy of all diagnostic tests read with and without clinical information using sensitivity, specificity and receiver operating characteristic (ROC) analysis metrics. Although clinical information was found to improve test reading accuracy (16), the measurement tools used may not have been able to capture relevant interpretation errors (17) nor determine the effect on diagnostic accuracy based on the type of clinical information provided or assigned interpretive task. Newer methods not considered in the previous review have been implemented to assess diagnostic performance. These methods include the free response (17), Localized Receiver Operating Characteristic (LROC) (18), and Region of Interest (ROI) (19) paradigms and allow for a more robust observer performance assessment accounting for multiple abnormalities per case (20) and location sensitivity (18).

Studies on the effect of clinical history have variations in study design including differences in localizing clinical history (21), availability of prior images or the presence of previous radiological reports (22) and specific reporting instructions (23). Whilst having this type of information available prior to test interpretation may not necessarily reflect normal clinical practice, it is also a potential source of bias and the specific effect on test performance is unclear. The uncertainty and variation in results across studies that have explored the effect of clinical history on diagnostic performance emphasize the need for a review of the literature on this topic. Therefore, this review aims to examine the effect of clinical history on diagnostic imaging interpretation and diagnostic performance.

METHODS

Databases and search strategy

The literature search was conducted based on the Preferred Reporting Items for a Systematic Review and Meta Analysis of Diagnostic Testing Studies (PRISMA DTA) Statement (24). Medline, Embase, Scopus, Web of Science and the Cochrane Central Register of Controlled Trials (CENTRAL) databases were searched for relevant articles published from database inception to July 21, 2020. Google Scholar was also used to search for relevant articles and the reference lists of published articles were manually screened to identify additional publications not identified on database search. Search terms were combined with "OR" and included the following main terms: "Clinical history" OR "Clinical information" OR "Patient history" OR "Patient information" OR "Diagnosis" OR "Diagnostic imaging" OR "Diagnostic test" OR "Radiography" OR "Radiographic imaging" OR "Radiology" OR "Radiologic study" OR "Localizing prompt" OR "Localizing history" OR "Reporting instruction" OR "Receiver operating characteristic" OR "Free response" OR "Free response operating characteristic. "

TABLE 1. The PICO Method Regarding Inclusion and Exclusion Criteria

Element	Characteristics
Population	Diagnostic imaging studies
Intervention	Clinical information
Comparator	No clinical information
Outcome	Diagnostic performance in diagnostic imaging: receiver operating characteristic (ROC) curve analysis, free response operating characteristic (FROC) analysis, sensitivity and specificity

Eligibility criteria

Inclusion and exclusion criteria were based on the Population, Intervention, Comparator, and Outcome elements (Table 1). The review sought to answer the following clinical research question: in diagnostic imaging studies, does having access to clinical history improve diagnostic performance? Thus, studies were included if they: compared diagnostic imaging performance both with and without clinical history, had observers read images under both conditions (crossover design), used an independent reference standard to validate the target condition, conducted data analysis including at least one of the following outcomes: sensitivity, specificity, ROC analysis, or free response operating characteristic (FROC) analysis and were published in English. Studies were excluded if they did not meet these inclusion criteria. Literature reviews, conference papers, letters to editors and posters were also excluded.

Quality assessment

Quality assessment to judge the risk of bias and applicability of original diagnostic accuracy studies was performed by two authors (KY and EE) using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS 2) tool (25). It consists of four main domains: patient selection, index test, reference standard, and flow and timing of the index tests and reference standard. A weighted Kappa was used to assess the agreement between the two assessors. Kappa was interpreted as follows: <0.20 = Poor; $0.21-0.40$ = Fair; $0.41-0.60$ = Moderate; $0.61-0.80$ = Substantial; and $0.81-0.99$ = Almost perfect (26). Any discrepancies in the quality assessment were discussed and resolved through consensus.

Data extraction process

Data were extracted in two phases. First, the authors determined the study characteristics (e.g., study design, reported outcome measures, included imaging modalities, provision of clinical history type, and recruitment method for readers), population characteristics (e.g., sample size,

disease prevalence and disease severity distribution), reader characteristics (observer experience and qualifications), and interpretation protocol. The performance metrics analyzed were number and location of detected abnormalities, ROC curve construction, relationship between true positive fraction at a given false positive fraction, area under the ROC curve, FROC analysis, and diagnostic accuracy measures such as Jackknife Alternative Free Response Receiver Operating Characteristic (JAFROC) figure of merit, sensitivity, and specificity.

RESULTS

Identification of included studies

The search strategy identified a total of 85,733 studies. After the screening of titles and abstracts, 42 studies were selected for full text reading. Twelve studies were excluded because no reference standard was used to assess observer performance (3,10,21,27–35) and eight studies were excluded because of no crossover design (23,36–42). The reference standard is imperative in determining and validating the true presence or absence of the target condition or disease that is being detected; without it, the true test results are unknown (43). A crossover design of readers interpreting images twice, both with and without history, addresses baseline imbalance of the history and nonhistory groups, thereby limiting the potential bias of between group differences. Subsequently, 22 studies fulfilled the inclusion criteria (Fig 1): 6 used sensitivity and specificity to examine the effect of clinical history (44–49), 12 used the ROC paradigm (12,50–60), 1 used the LROC paradigm (61), 1 used the ROI paradigm (62), and 2 studies used the free response paradigm (63,64) (Table 2).

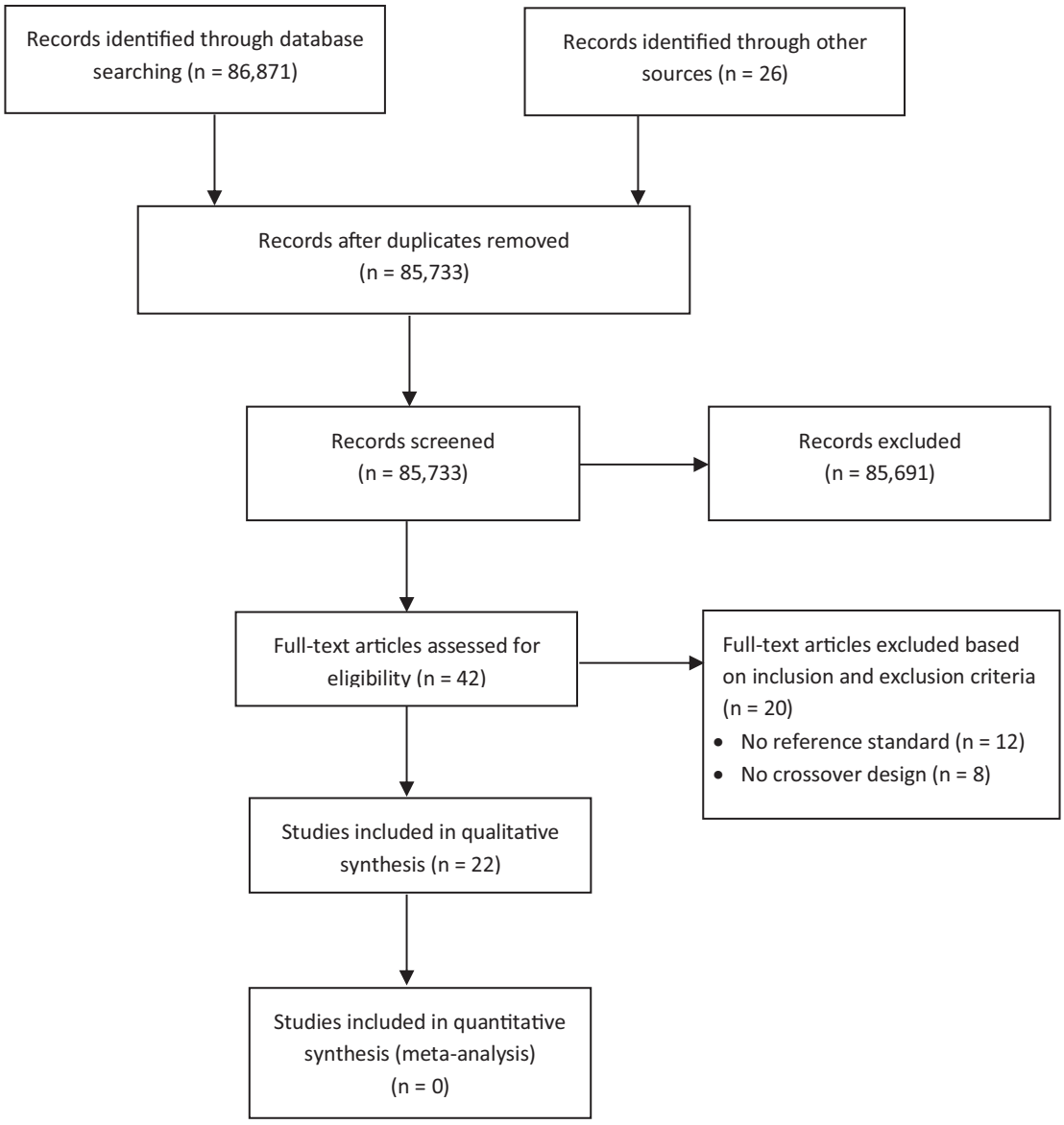


Figure 1. PRISMA flow diagram.

TABLE 2. Characteristics of Included Studies

Study	Imaging Type	Index Test	Reference Standard	Disease Prevalence	Readers	History Type	Balanced Design	Time Between Reading sessions	Abnormality Location Marking	Performance Metrics
McNeil et al (50)	Computed tomography of the head	Five point rating scale	Biopsy, surgery, angiography, pneumoencephalography or clinical follow up data	39% (35/89)	Four radiologists	Clinical information at study order	No	Two weeks	No	Area under the ROC curve 0.944 without history, 0.977 with
Berbaum et al (51)	Chest radiographs	Five point rating scale	Follow up studies, surgery, laboratory tests and autopsy reports	58% (25/43)	Six experienced radiologists	Tentative diagnosis by a prompt that is correct and abnormals and plausible for normals, prompt suggesting a category that is plausible but incorrect for both abnormals and normals	Yes	Several months	Yes	Jackknife estimates of area under the ROC curve 0.791 without a prompt, 0.880 with
Kinnunen et al (44)	Middle third face radiographs	15 diagnostic categories	Surgical verification and follow up clinical information	43% (43/100)	Two radiology specialists in bone radiology, 3 specialists in general radiology, 2 senior and 2 junior residents	Clinical data from the initial referral	No	Four months	No	Sensitivity and specificity 76% and 91% respectively without clinical data, 83% and 85% respectively with
Berbaum et al 1988 (52)	Chest radiographs	Five point rating scale	Follow up studies, surgery, laboratory tests and autopsy reports	59% (26/44)	Six 3rd and 4th year radiology residents	Tentative diagnosis by a prompt that is correct and abnormals and plausible for normals	Yes	Several months	No	ROC curve and area under the ROC curve 0.800 without history and 0.890 with a diagnostic prompt for nodule cases, 0.688 without history and 0.889 with a diagnostic prompt for non nodule cases

(*continued on next page*)

TABLE 2. (Continued)

Study	Imaging Type	Index Test	Reference Standard	Disease Prevalence	Readers	History Type	Balanced Design	Time Between Reading sessions	Abnormality Location Marking	Performance Metrics
Cooperstein et al (53)	Chest radiographs	Five point rating scale	Surgical, follow up imaging or clinical status	68% (168/247)	Five board certified radiologists	Clinical information and medical history	Unclear	Unknown	No	Area under the ROC curve 0.87 without history and 0.87 with
Good et al (12)	Chest radiographs	Five point rating scale	Surgical, follow up imaging or clinical status	68% (168/247)	Four board certified radiologists	Clinical information and medical history	Unclear	Unknown	No	Area under the ROC curve mean 0.895 without history and 0.898 with for interstitial disease, 0.873 without history and 0.855 with for nodules, 0.948 without history and 0.930 with for pneumothorax
Song et al (54)	Chest, abdomen and bone radiographs	Six point rating scale	Clinical, laboratory, surgical and imaging findings	50% (54/109)	Six residents and 2 board certified radiologists	History from radiological request paper or patient chart with modified definitive diagnostic information	No	One month	No	ROC curve and area under the ROC curve 0.75 without history and 0.84 with
Elmore et al (45)	Mammograms	Descriptive observations and four point rating scale for diagnostic interpretations	Clinical or histopathology follow up	18% (18/100)	Ten radiologists	Clinical data at time of mammogram, five cases had a sham history	Yes	Five months	Yes	Median sensitivity and specificity 67% and 93% respectively without history, 72% and 94% respectively with
Tudor et al (49)	Chest, abdominal and musculoskeletal radiographs	Five point rating scale	A “known diagnosis”	62% (31/50)	Five consultant radiologists	Clinical information from request card	No	Five months	No	ROC curve, sensitivity and specificity 80% and 72% respectively without history, 83% and 76% respectively with

(*continued on next page*)

TABLE 2. (Continued)

Study	Imaging Type	Index Test	Reference Standard	Disease Prevalence	Readers	History Type	Balanced Design	Time Between Reading sessions	Abnormality Location Marking	Performance Metrics
Ehara and Katsuragawa (55)	Wrist radiographs	Confidence level on a 5cm line	Clinical data that "followed the process of healing"	50% (20/40)	Three radiologists and 2 orthopedic surgeons	"Point of tenderness was notified to the readers" for all fracture and half of control cases	No	Two months	No	ROC curve and area under the ROC curve 0.928 without history and 0.972 with
Quekel et al (46)	Chest radiographs	"Normal" or "pathological"	Consensus of 3 co authors and histopathology	30% (30/100)	Ten radiologists, 2 chest physicians, 2 radiology residents	Standardised clinical information, ie "suspicion of malignancy" and previous radiographs	No	Four months	No	Sensitivity and specificity 33% and 92% respectively without history, 31% and 93% respectively with
Dhingsa et al (63)	Endorectal magnetic resonance imaging	Five point rating scale and location of prostate cancer	Biopsy	100% (37/37)	Two radiologists	Clinical data from patient record	No	Two weeks	Yes	AFROC curve and area under the AFROC curve mean 0.59 without clinical data and 0.47 with
Houssami et al (56)	Mammograms	Five point rating scale	Histological report	50% (240/480)	Two radiologists	Type and site of symptoms	No	None	No	ROC curve, sensitivity, specificity and area under the ROC curve mean A_z 0.821 without history and 0.838 with
Houssami et al (57)	Breast ultrasound	Five point rating scale	Histological report	50% (240/480)	Two radiologists	Accompanying mammogram*	No	2 4 weeks	No	ROC curve, sensitivity, specificity and area under the ROC curve mean A_z 0.868 without history and 0.900 with
Roelofs et al (61)	Mammograms	Five point rating scale	Follow up mammogram	50% (80/160)	Twelve radiologists	Prior mammogram*	No	Four weeks	Yes	LROC curve, Lesion Localized Fraction scores mean 0.19 without prior mammograms and 0.26 with

(*continued on next page*)

TABLE 2. (Continued)

Study	Imaging Type	Index Test	Reference Standard	Disease Prevalence	Readers	History Type	Balanced Design	Time Between Reading sessions	Abnormality Location Marking	Performance Metrics
Baek et al (58)	Breast sonography	Five point rating scale	Histopathologically proven cancer or benign lesion with follow up	52% (78/150)	Three radiologists	Age, palpability and history	No	Eight weeks	No	Sensitivity, specificity and area under the ROC curve 0.893 without clinical information and 0.931 with
Filippone et al (59)	Multidetector row computed tomography and magnetic resonance imaging of the liver	Four point rating scale	Surgically and histologically proven focal liver lesions	71% (87/122)	Two radiologists	Patient related information and history of laboratory diseases	No	Three months	No	Sensitivity, specificity and area under the ROC curve for MDCT: mean 0.88 without history and 0.94 with, for MRI: mean 0.93 without history and 0.97 with
Cereser et al (47)	Chest radiographs	Four point rating scale, appearance and location of parenchymal abnormalities	Computed tomography within 3 days	64% (41/64)	Two radiologists	Patient file history	No	15 days	Yes	Sensitivity and specificity 48.7% and 70.0% respectively without history, 42.7% and 93.5% respectively with
Millet et al (48)	Computed tomography of the abdomen	Categorisation of various diseases	Consensus between a radiologist and emergency physician	100% (333/333)	Two radiologists	Complete emergency medical report	No	Four months	No	Sensitivity 85.3% without history and 87.4% with
Soh et al (62)	Mammograms	Four point rating scale	Pathology and recall data	15% (30/200)	Five radiologists	Prior images*	No	Four months	Yes	Sensitivity, specificity and region of interest (ROI) figure of merit 0.85 without prior images and 0.88 with

(*continued on next page*)

TABLE 2. (Continued)

Study	Imaging Type	Index Test	Reference Standard	Disease Prevalence	Readers	History Type	Balanced Design	Time Between Reading sessions	Abnormality Location Marking	Performance Metrics
Littlefair et al (64)	Chest radiographs	Five point rating scale	Computed tomography, histology or clinical follow up	21% (10/47)	33 radiologists	Low or high expected prevalence	No	1 3 days	Yes	Location sensitivity, specificity, and JAFROC values median 0.65 without history vs 0.57 with for low expected prevalence, 0.60 without history vs 0.64 for high expected prevalence
Saba et al (60)	Magnetic resonance imaging for endometriosis detection	Five point rating scale	Surgery results	Unknown	Four radiologists	Patient clinical record	No	Three months	Yes	ROC curve, sensitivity, specificity and area under the ROC curve mean 0.814 without history and 0.849 with

* Although different images were not strictly defined as a clinical history, the embedded patient history from the images was deemed relevant and contributes to the interpretive process in a similar manner to the clinical history.

TABLE 3. QUADAS-2 Tool Results

Study	Risk of Bias				Applicability Concerns		
	Patient Selection	Index test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
McNeil et al (50)	Low	Low	Low	Low	Low	Low	Low
Berbaum et al (51)	Low	Low	Low	Low	Low	Low	Low
Kinnunen et al (44)	Low	Low	Low	Low	Low	Low	Low
Berbaum et al (52)	Low	Low	Low	Low	Low	Low	Low
Cooperstein et al (53)	Low	Low	Low	Low	Low	Uncertain	Low
Good et al (12)	Low	Low	Low	Low	Low	Uncertain	Low
Song et al (54)	Uncertain	Low	Low	Low	Uncertain	Low	Low
Elmore et al (45)	Low	Low	Low	Low	Low	Low	Low
Tudor et al (49)	Uncertain	Low	Uncertain	Low	Uncertain	Low	Uncertain
Ehara and Katsuragawa (55)	Low	Low	Low	Low	Low	Low	Low
Quekel et al (46)	Low	Low	Low	Low	Low	Low	Low
Dhingsa et al (63)	Low	Low	Low	Low	Low	Low	Low
Houssami et al (56)	Low	Low	Low	Low	Low	Uncertain	Low
Houssami et al (57)	Low	Low	Low	Low	Low	Low	Low
Roelofs et al (61)	Low	Low	Uncertain	Low	Low	Low	Uncertain
Baek et al (58)	Low	Low	Low	Low	Low	Low	Low
Filippone et al (59)	Low	Low	Low	Low	Low	Low	Low
Cereser et al (47)	Low	Low	Low	Low	Low	Low	Low
Millet et al (48)	Low	Low	Uncertain	Low	Low	Low	Uncertain
Soh et al (62)	Low	Low	Low	Low	Low	Low	Low
Littlefair et al (64)	Low	Low	Low	Low	Uncertain	Uncertain	Low
Saba et al (60)	Low	Low	Low	Low	Low	Low	Low

Quality assessment

The included studies generally had a low risk of bias for most categories. Fourteen studies had a low risk of bias for all seven categories (44–47,50–52,55,57–60,62,63). Three studies had an uncertain risk of bias in one category (applicability concerns, index test) and low risk in the other six (12,53,56) arising from potential memory bias due to reading sessions held with insufficient or unknown time between them. The remaining five studies had an uncertain risk of bias in two categories and low risk in five (48,54,61,64) and an uncertain risk in four categories and low risk in three (49). None of the studies had a high risk of bias in any category. The quality assessment results are summarized in Table 3. A balanced design, where readers read images both with and without history at each session, was used in only three studies (45,51,52). Inter reader agreement between the two assessors of quality showed a weighted Kappa of k = 1.000, 95% CI: 1.000–1.000.

Diagnostic performance of having clinical history versus no clinical history

The studies that used the sensitivity and specificity paradigm showed that clinical history had a mixed effect on sensitivity and a small mixed effect on specificity (44–46,48,49) with only one study showing that history was associated with an improvement in specificity from 70.0% to 93.5% (47). When the ROC paradigm was used, most included studies found clinical history improved diagnostic performance, with the area under the ROC curve increasing in 10 studies (50–52,54–60), a mixed effect on area when multiple diseases were included (12) and no effect in a study on chest radiographs (53).

The study using the LROC paradigm showed history improved reader performance; higher Lesion Localized Fraction scores were obtained when clinical history from prior mammograms was available (61). A similar study using the ROI paradigm reported an increase in diagnostic performance and showed a higher figure of merit with clinical history from prior mammograms (62).

Two studies using the free response paradigm showed that in contrast, history reduced or had no significant effect on diagnostic performance. For prostate cancer localization, the area under the alternative free response ROC (AFROC) curve was smaller with history due to increased false positive findings (63). In chest radiographs, JAFROC values were lower when history with a low expected disease prevalence was available but higher when history with a high expected disease prevalence was given (64). This study also showed that specificity was significantly lower when a high expected disease prevalence history was given compared to a low expected prevalence. Both free response studies reported an increase in false positive calls with more clinical history. A summary of diagnostic performance is in Table 2.

DISCUSSION

Published evidence demonstrates diagnostic performance improvement with clinical history (16). Fifteen (45,48–52,54–62) of the 22 included studies in the present study showed that clinical history improved diagnostic performance for all reported metrics and only one study (63) showed clinical history decreased diagnostic performance. Whilst having clinical history prior to a subsequent examination can improve diagnostic accuracy (65), the underlying mechanism on diagnostic image interpretation can be explained by prior probability before test interpretation (66). The clinical history, in a similar way to disease prevalence, establishes and updates the pretest probability – an assessment of how likely a patient has a specific disease (67). These data are combined with the information from the imaging to form the posterior (post test) probability and is a likely explanation for greater diagnostic performance when imaging is read with clinical history. Most of the evidence for improvement in diagnostic performance is based on the ROC paradigm, which limits every case to a single assigned rating and does not take into account abnormality location nor multiple abnormalities (20).

Developments on ROC methodology include the LROC paradigm (18), which uses a single confidence rating and mark for the most suspicious region and the ROI paradigm (68), a further development of LROC using a rating for each specified region per case. Whilst two included studies used the LROC (61) and ROI (62) paradigms demonstrated improved diagnostic performance with clinical history, it should be noted that these methodologies have some limitations: not accounting for multiple abnormalities, imposing a reading paradigm fundamentally different to that in clinical practice (20) and introduction of artificial and arbitrary boundaries into the interpretation process – where each ROI cannot account for more than one abnormality, or for an abnormality occupying two adjoining ROIs (69).

In contrast, studies using the free response observer performance paradigm showed contradictory results. The FROC paradigm involves readers marking every region suspicious for abnormality with a corresponding confidence rating (19); in the analysis, correct decisions are rewarded and incorrect ones penalized (70). This paradigm accounts for unnoticed ambiguities in the ROC paradigm – i.e., when a location level false positive and false negative occur on the same image (17). Here, an image level true positive occurs for the wrong reasons – an incorrect abnormality location being reported and an abnormality missed. This is clinically significant, as treatment can occur at the wrong location and/or the missed abnormality may grow and increase in severity (71). One of two included FROC studies found that clinical history increased prostate cancer nodule detection with an associated increase in false positive calls; overall history had a nonsignificant decrease in reader accuracy (63). The second included study using the free response paradigm reported similar results; information from clinical history had more false positive calls and no significant improvement in diagnostic performance (64). This is relevant because the ROC, LROC, or ROI paradigms use a case based analysis with only one rating or a fixed number of ratings per case. Studies using these methods can detect a performance change at the case level, but may fail to detect changes at the abnormality level because they do not consider multiple abnormalities per ROI or case. The increased false positive rate may be explained by a framing effect or confirmation bias, where the clinical history may have influenced readers' decision making (72).

Different amounts of history have been analyzed in other diagnostic imaging studies using a free response analysis. When a brief clinical history was compared to a more detailed history that included the location of a tumor in chest radiographs, more clinical information biased the diagnostic performance of radiologists (41). This was able to show that with more clinical history, location sensitivity in nodule detection increased and specificity decreased within the lobe of interest. This increase in false positives was also seen when reporting instructions were compared in a similar study (23) where observers were given an unframed (report all abnormalities) and framed (report on lung nodules) task for two readings of the same image test set. A framing task can be considered an increased suspicion of abnormality, similar to having clinical history. The framed task had a significantly lower JAFROC figure of merit and specificity compared to the unframed task, with location sensitivity being unaffected. The increased false positive rate was seen in both nodule free and nodule containing images. This study (23) included the entire range of lung nodule subtlety, which allows generalization of the results to a broader population.

All these more recent free response studies indicate that location sensitivity increases with an interpretive bias, such as clinical history, and is accompanied with an increase in false positives compared to interpreting the images without bias. This notable finding is undetectable using the ROC paradigm and has not been reported in a previous review that showed clinical information improved test accuracy (16). The ROC paradigm works for case level analysis, where readers assign a rating of their confidence in abnormality presence for the entire case, without necessarily stating what informed their decision. This paradigm may be influenced by clinical history and could explain why ROC studies show improved diagnostic performance with clinical history. In contrast, free response studies involve analysis at abnormality level, accounting for a reading scenario with a random number of signals with localization (17). If an observer determines presence of any abnormality, all must be correctly localized otherwise lower diagnostic performance values are obtained if abnormalities are incorrectly marked. It has been stated that the reading task should be validated by being similar to that used in clinical practice (20) and the free response paradigm is a realistic reflection of a clinical situation. Clinical history can also affect free response studies and it could also explain why these studies show increased location sensitivity, false positives and no overall change in diagnostic performance. It is apparent that clinical history does affect diagnostic image

interpretation and the reading paradigm implemented would largely determine the type of effect. In order to provide gen eralizable results, studies comparing the effect of clinical his tory on diagnostic image interpretation should reflect clinical practice and use a method that: measures multiple abnormali ties per image, only rewards marks if the location is correctly identified and penalizes incorrect locations and missed abnor malities – even if another in the same image was located.

The time gap between the first reading session and subse quent image interpretations ranged from no time to five months. Timing between experimental sessions is relevant because if held too close together, the second viewing could be influenced by observer recall from the earlier observation (73). The ideal time gap between diagnostic image reading sessions in the specific context of interpreting image datasets with and without history has not been studied; however, a mean of 7 weeks has been reported to be sufficient because recognition accuracy was similar to chance (74). Future stud ies should take this into account and allow for enough time between sessions to limit memory bias.

Some limitations and inconsistencies were observed in the studies reviewed. Nineteen of 22 included studies did not have a balanced design. These studies may be subject to bias from any reader changes, reporting practice or training that may have occurred between reading sessions (16). Whilst a balanced design can account for reader changes between ses sions, having images with and without history in the same session can create a demand characteristic (75). The effect of demand characteristic in a diagnostic imaging reading study is outside the scope of this review, however future studies should account for intrinsic reader changes between reading sessions by using a balanced design.

A previous review highlighted the type of history being provided – real or constructed – as having an effect on test reading accuracy, with constructed clinical history having a larger improvement of diagnostic performance (16). Only two included studies (51,52) used constructed history and the specific effect in comparison to real history is difficult to determine. Nevertheless, the reading task should be similar to that found in clinical practice (20) and a real clinical history should be used in future studies, unless a particular framing or reading task is to be studied (23).

Limitations of this review include that the analysis was restricted to studies published in English. Secondly, limited data on disease severity or case difficulty and differences in study design made the comparison of results difficult. How ever, our review captured all observer performance evalua tion paradigms that were not considered in previous reviews; thus, it provides a more detailed understanding of the impact of clinical history on diagnostic performance.

CONCLUSION

Whilst most published studies show that clinical history improved diagnostic performance, more recent studies accounting for location and multiple abnormalities show increased sensitivity, decreased specificity and no significant change in overall diagnostic performance with clinical history. Further studies are required to better understand the effect of clinical history on diagnostic imaging performance; these should reflect the task in clinical practice, allow for measuring multiple abnormalities per image, reward correct and penalize incorrect abnormality locations, account for memory bias and allow adequate time between reading sessions, provide a bal anced reading design and include the entire spectrum of disease and nondisease severity in the study population.

REFERENCES

1. Bickley LS, Szilagyi PG. Bates' Guide to Physical Examination and His tory Taking. 12th ed. Philadelphia: Wolters Kluwer, 2017.
2. Maizlin NN, Somers S. The role of clinical history collected by diagnostic imaging staff in interpreting of imaging examinations. J Med Imaging Radiat Sci 2019; 50(1):31 35.
3. Schreiber MH. The clinical history as a factor in roentgenogram interpre tation. JAMA 1963; 185(5):399 401.
4. Croskerry P. The importance of cognitive errors in diagnosis and strate gies to minimize them. Acad Med 2003; 78(8):775 780.
5. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. Radiographics 2018; 38(1):236 247.
6. Itri JN, Patel SH. Heuristics and cognitive error in medical imaging. Am J Roentgenol 2018; 210(5):1097 1105.
7. Hanna T, Rohatgi S, Shekhani H, et al. Clinical information available dur ing emergency department imaging order entry and radiologist interpre tation. Emerg Radiol 2017; 24(4):361 367.
8. Maizlin N, Somers S. The role of clinical history collected by diagnostic imaging staff in interpreting of imaging examinations. J Med Imaging Radiat Sci 2019; 50(1):31 35.
9. Gill CJ, Sabin L, Schmid CH. Why clinicians are natural Bayesians. research Clinical, editor. Why clinicians are natural Bayesians. BMJ 2005; 330(7499):1080 1083.
10. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. Am J Roentgenol 1981; 137(5):1055 1058.
11. Cooperstein LA, Good BC, Eelkema EA, et al. The effect of clinical history on chest radiograph interpretations in a PACS environment. Invest Radiol 1990; 25(6):670 674.
12. Good BC, Cooperstein LA, DeMarino GB, et al. Does knowledge of the clinical history affect the accuracy of chest radiograph interpretation? Am J Roentgenol 1990; 154(4):709 712.
13. Cereser L, Zuiani C, Graziani G, et al. Impact of clinical data on chest radiography sensitivity in detecting pulmonary abnormalities in immuno compromised patients with suspected pneumonia. Radiol Med (Torino) 2009; 115(2):205 214.
14. Swensson RG, Hessel SJ, Herman PG. The value of searching films with out specific preconceptions. Invest Radiol 1985; 20(1):100 107.
15. Eldevik OP, Dugstad G, Orrison WW. The effect of clinical bias on the interpretation of myelography and spinal computed tomography. Radiol ogy 1982; 145(1):85 89.
16. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clini cal information a systematic review. JAMA 2004; 292(13):1602 1609.
17. Bunch P, Hamilton J, Sanderson G, et al. A Free Response Approach to The Measurement And Characterization Of Radiographic Observer Per formance, 0127. SPIE, 1977.
18. Starr SJ, Metz CE, Lusted LB, et al. Visual detection and localization of radiographic images. Radiology 1975; 116(3):533 538.
19. Chakraborty DP. A status report on free response analysis. Radiat Prot Dosimetry 2010; 139(1 3):20 25.
20. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis, and validation. Med Phys 2004; 31 (8):2313 2330.
21. Berbaum KS, El Khoury GY, Franken Jr EA, et al. Impact of clinical his tory on fracture detection with radiography. Radiology 1988; 168(2): 507 511.
22. Aideyan UO, Berbaum K, Smith WL. Influence of prior radiologic informa tion on the interpretation of radiographic examinations. Acad Radiol 1995; 2(3):205 208.

23. Robinson JW, Brennan PC, Mello Thoms C, et al. Reporting instructions significantly impact false positive rates when reading chest radiographs. Eur Radiol 2016; 26(10):3654 3659.
24. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta analysis of diagnostic test accuracy studies: the PRISMA DTA statement. JAMA 2018; 319(4):388 396.
25. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS 2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155(8):529 536.
26. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005; 37(5):360 363.
27. Eldevik OP, Dugstad G, Orrison WW, et al. The effect of clinical bias on the interpretation of myelography and spinal computed tomography. Radiology 1982; 145(1):85 89.
28. Berbaum KS, Franken EA, Elkhoury GY. Impact of clinical history on radiographic detection of fractures: a comparison of radiologists and orthopedists. Am J Roentgenol 1989; 153(6):1221 1224.
29. Rickett AB, Finlay DB, Jagger C. The importance of clinical details when reporting accident and emergency radiographs. Injury 1992; 23(7):458 460.
30. Berbaum KS, Franken Jr. EA, Dorfman DD, et al. Influence of clinical history on perception of abnormalities in pediatric radiographs. Acad Radiol 1994; 1(3):217 223.
31. Aideyan UO, Berbaum K, Smith WL. Influence of prior radiologic information on the interpretation of radiographic examinations. Acad Radiol 1995; 2(3):205 208.
32. Knight JA, Libstug AR, Moravan V, et al. An assessment of the influence of clinical breast examination reports on the interpretation of mammograms in a breast screening program. Breast Cancer Res Treat 1998; 48 (1):65 71.
33. Leslie A, Jones AJ, Goddard PR. The influence of clinical information on the reporting of CT by radiologists. Br J Radiol 2000; 73(874):1052 1055.
34. Dijkman BG, Busse JW, Walter SD, et al. The impact of clinical data on the evaluation of tibial fracture healing. Trials 2011; 12:237.
35. Mouden M, Ottervanger JP, Timmer JR, et al. The influence of coronary calcium score on the interpretation of myocardial perfusion imaging. J Nucl Cardiol 2014; 21(2):368 374.
36. Babcook CJ, Norman GR, Coblentz CL. Effect of clinical history on the interpretation of chest radiographs in childhood bronchiolitis. Invest Radiol 1993; 28(3):214 217.
37. Berbaum KS, Franken EA, Anderson KL, et al. The influence of clinical history on visual search with single and multiple abnormalities. Invest Radiol 1993; 28(3):191 201.
38. Ciray I, Astrom G, Sundstrom C, et al. Assessment of suspected bone metastases CT with and without clinical information compared to CT guided bone biopsy. Acta Radiol 1997; 38(5):890 895.
39. Mullins ME, Lev MH, Schellingerhout D, et al. Influence of availability of clinical history on detection of early stroke using unenhanced CT and diffusion weighted MR imaging. Am J Roentgenol 2002; 179(1):223 228.
40. Irwig L, Macaskill P, Walter S, et al. New methods give better estimates of changes in diagnostic accuracy when prior information is provided. J Clin Epidemiol 2006; 59(3):299 307.
41. Littlefair S, Brennan P, Mello Thoms C, et al. Outcomes knowledge may bias radiological decision making. Acad Radiol 2016; 23(6):760 767.
42. Littlefair S, Brennan P, Reed W, et al. Does expectation of abnormality affect the search pattern of radiologists when looking for pulmonary nodules? J Digit Imaging 2017; 30(1):55 62.
43. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016; 6(11):e012799. e012799.
44. Kinnunen J, Göthlin J, Tötterman S, et al. The effects of clinical data on radiological diagnostic performance in middle face trauma. Rofo 1987; 147(2):143 146.
45. Elmore JG, Wells CK, Howard DH, et al. The impact of clinical history on mammographic interpretations. JAMA 1997; 277(1):49 52.
46. Quekel LGBA, Goei R, Kessels AGH, et al. Detection of lung cancer on the chest radiograph: impact of previous films, clinical information, double reading, and dual reading. J Clin Epidemiol 2001; 54(11):1146 1150.
47. Cereser L, Zuiani C, Graziani G, et al. Impact of clinical data on chest radiography sensitivity in detecting pulmonary abnormalities in immunocompromised patients with suspected pneumonia. Radiol Med 2010; 115(2):205 214.
48. Millet I, Alili C, Bouic Pages E, et al. Acute abdominal pain in elderly patients: effect of radiologist awareness of clinicobiologic information on CT accuracy. Am J Roentgenol 2013; 201(6):1171 1179.
49. Tudor GR, Finlay D, Taub N. An assessment of inter observer agreement and accuracy when reporting plain radiographs. Clin Radiol 1997; 52 (3):235 238.
50. McNeil BJ, Hanley JA, Funkenstein HH, et al. Paired receiver operating characteristic curves and the effect of history on radiographic interpretation CT of the head as a case study. Radiology 1983; 149(1):75 77.
51. Berbaum KS, Franken EA, Dorfman DD, et al. Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs. Invest Radiol 1986; 21(7):532 539.
52. Berbaum KS, Franken EA, Dorfman DD, et al. Influence of clinical history upon detection of nodules and other lesions. Invest Radiol 1988; 23(1):48 55.
53. Cooperstein LA, Good BC, Eelkema EA, et al. The effect of clinical history on chest radiograph interpretations in a PACS environment. Invest Radiol 1990; 25(6):670 674.
54. Song KS, Song HH, Park SH, et al. Impact of clinical history on film interpretation. Yonsei Med J 1992; 33(2):168 172.
55. Ehara S, Katsuragawa S. Influence of clinical information on the detection of wrist fractures in children. Tohoku J Exp Med 1999; 189(2):147 153.
56. Houssami N, Irwig L, Simpson JM, et al. The influence of clinical information on the accuracy of diagnostic mammography. Breast Cancer Res Treat 2004; 85(3):223 228.
57. Houssami N, Irwig L, Simpson JM, et al. The influence of knowledge of mammography findings on the accuracy of breast ultrasound in symptomatic women. Breast J 2005; 11(3):167 172.
58. Baek SE, Kim MJ, Kim EK, et al. Effect of clinical information on diagnostic performance in breast sonography. J Ultrasound Med 2009; 28 (10):1349 1356.
59. Filippone A, Cianci R, Iezzi R, et al. Effect of clinical history in focal liver lesion detection and classification on 4 detector row computed tomography and gadoxetic acid enhanced mr imaging in oncologic patients. J Comput Assist Tomogr 2009; 33(6):851 857.
60. Saba L, Ajossa S, Ledda G, et al. Does the clinical information play a role in the magnetic resonance diagnostic confidence analysis of ovarian and deep endometriosis? Br J Radiol 2019; 92(1096).
61. Roelofs AJ, Karssemeijer N, Wedekind N, et al. Importance of comparison of current and prior mammograms in breast cancer screening. Radiology 2007; 242(1):70 77.
62. Soh BP, Lee WB, McEntee MF, et al. Mammography test sets: Reading location and prior images do not affect group performance. Clin Radiol 2014; 69(4):397 402.
63. Dhingsa R, Qayyum A, Coakley FV, et al. Prostate cancer localization with endorectal MR imaging and MR spectroscopic imaging: effect of clinical data on reader accuracy. Radiology 2004; 230(1):215 220.
64. Littlefair S, Mello Thoms C, Reed W, et al. Increasing prevalence expectation in thoracic radiology leads to overcall. Acad Radiol 2016; 23 (3):284 289.
65. Shikino K, Ikusaka M, Ohira Y, et al. Influence of predicting the diagnosis from history on the accuracy of physical examination. Adv Med Educ Pract 2015; 6:143 148.
66. Schechter MT, Sheps SB. Diagnostic testing revisited: pathways through uncertainty. Can Med Assoc J 1985; 132(7):755 760.
67. Chang PJ. Bayesian analysis revisited: a radiologist's survival guide. AJR Am J Roentgenol 1989; 152(4):721 727.
68. Obuchowski NA, Lieber ML, Powell KA. Data analysis for detection and localization of multiple abnormalities with application to mammography. Acad Radiol 2000; 7(7):516 525.
69. Chakraborty DP. Data analysis for detection and localization of multiple abnormalities with application to mammography (comment). Acad Radiol 2000; 7(7):553 554.
70. Chakraborty DP. New developments in observer performance methodology in medical imaging. Semin Nucl Med 2011; 41(6):401 418.
71. Chakraborty D. Recent developments in FROC methodology. In: Krupinski ESaE, ed. The Handbook of Medical Image Perception and Techniques, Cambridge, UK: Cambridge University Press; 2010:216 239.
72. Croskerry P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. Acad Emerg Med 2002; 9(11):1184 1204.
73. Ryan JT, Haygood TM, Yamal JM, et al. The ''Memory Effect'' for repeated radiologic observations. Am J Roentgenol 2011; 197(6):W985 W991.
74. Evans KK, Marom EM, Godoy MC, et al. Radiologists remember mountains better than radiographs, or do they? J Med Imaging (Bellingham) 2016; 3(1):011005.
75. Orne MT. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. Am Psychol Assoc 1962; 17(11):776 783.

CHAPTER 4

4.1 Bridging section to Chapter 4

Diagnostic performance studies can be subject to bias due to shortcomings in design and conduct, where the results may not apply to other patient groups or settings. To improve the completeness and transparency of these studies, the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement was developed. By following the STARD guidelines, relevant design issues and sources of bias can be addressed, thereby ensuring that good quality studies on diagnostic performance can be undertaken.

The systematic review in chapter two found a lack of high quality evidence comparing the diagnostic performance of CBCT and digital PA radiography in endodontic disease detection. It identified methodological issues which led to limitations of the current literature. These include the lack of a valid reference standard, the use of a PA radiolucency as an index test metric, skewed population sampling using a very high disease prevalence, undisclosed severity of disease and non-disease in the included cases, low reader numbers and minimal data on reader characteristics, measurement of image interpretation at the case level, and a reading task that does not replicate the clinical task. These limitations affect the external validity of the results and the generalisability of the findings due to low statistical power. In addition, the observer performance methodologies employed in published studies do not account for images with multiple abnormalities nor provide abnormality location information, which may incorrectly reward incorrect decisions on image interpretation. Consequently, the review in chapter two provided recommendations to overcome the limitations in previous studies and provide valid and relevant data on diagnostic performance in the endodontic domain.

Building on the findings and recommendations in chapter two, the work presented in chapter four compares the diagnostic performance of CBCT and digital PA radiography in endodontic disease detection. It addresses the aforementioned limitations by using several study design elements never used in published studies on these imaging modalities. A factorial study in the free-response paradigm was undertaken, accounting for changes in the following independent variables: case type (diseased or non-diseased), case severity (subtle, moderate or obvious), reader type (low or high level of experience) and modality (CBCT or digital PA radiography). This allowed readers to mark images with abnormalities as many times as they wish, awarded correct abnormality location and penalised incorrect location. Data analysis included the weighted Alternative Free Response Operating Characteristic 1 (wAFROC1) metric, a measure of the probability on marked images that lesion localisations are rated higher than non-localisations. It corresponds to the area under the wAFROC1 curve. For non-diseased cases, specificity was the main metric.

This study is able to provide data on diagnostic performance of CBCT and digital PA radiography in endodontic disease detection in different case and reader types. The results of the study show that CBCT performed better than PA imaging in non-diseased cases, with CBCT having greater specificity in obvious non-diseased cases. Conversely, PA outperformed CBCT imaging in diseased cases, with higher wAFROC1 values in subtle and moderate diseased cases. Readers with more clinical experience had fewer false positive marks in obvious diseased CBCT cases than those with fewer years of experience. Generally, reader performance was suboptimal across PA radiography and CBCT and neither modality showed a performance advantage over the other, suggesting a need for improvement in interpretive skills of readers for both PA and CBCT imaging.

CBCT had some tendency for fewer lesion localisations (true positives – TPs) in diseased cases and greater false positives (FPs) in both diseased and non-diseased cases than in PA radiography. These findings could be considered a criterion shift for each case category in the domain of signal detection theory, representing to the willingness of the reader to make a positive call for disease (changing β in the ROC paradigm). However, this cannot be measured in the same manner in the free response paradigm, because readers are not limited to one rating per image and there is no limit nor relationship between the number of TPs and FPs being made.

This paper is published in Journal of Endodontics as:

Periapical radiography versus cone beam computed tomography in endodontic disease detection: a free-response, factorial study. Journal of Endodontics 2023, 49(4):419-429.
https://doi.org/10.1016/j.joen.2023.02.001

Periapical Radiography versus Cone Beam Computed Tomography in Endodontic Disease Detection: A Free-response, Factorial Study

Kehn E. Yapp, BDSc, DCD (Endo), Mo'ayyad Suleiman, PhD, Patrick Brennan, PhD, and Ernest Ekpo, PhD

ABSTRACT

Aim: To assess and compare reader performance in interpreting digital periapical (PA) radiography and cone beam computed tomography (CBCT) in endodontic disease detection, using a free-response, factorial model. **Materials and Methods:** A reader performance study of 2 image test sets was undertaken using a factorial, free-response design, accounting for the independent variables: case type, case severity, reader type, and imaging modality. Twenty-two readers interpreted 60 PA and 60 CBCT images divided into 5 categories: diseased–subtle, diseased–moderate, diseased–obvious, nondiseased–subtle, and non-diseased–obvious. Lesion localization fraction, specificity, false positive (FP) marks, and the weighted alternative free-response receiver operating characteristic figure of merit were calculated. **Results:** CBCT had greater specificity than PA in the obvious nondiseased cases ($P = .01$) and no significant difference in the subtle nondiseased category. Weighted alternative free-response receiver operating characteristic values were higher for PA than CBCT in the subtle diseased ($P = .02$) and moderate diseased ($P = .01$) groups with no significant difference between in the obvious diseased groups. CBCT had higher mean FPs than PA ($P < .05$) in subtle diseased cases. Mean lesion localization fraction in the moderate diseased group was higher in PA than CBCT ($P = .003$). No relationships were found between clinical experience and all diagnostic performance measures, except for in the obvious diseased CBCT group, where increasing experience was associated mean FP marks ($P = .04$). **Conclusions:** Reader performance in the detection of endodontic disease is better with PA radiography than CBCT. Clinical experience does not impact upon the accuracy of interpretation of both PA radiography and CBCT. *(J Endod 2023;49:419–429.)*

KEY WORDS

CBCT; diagnostic performance; free-response; factorial study; PA

SIGNIFICANCE

PA radiography had better reader performance than CBCT in endodontic disease detection. In non-diseased images, CBCT had greater specificity than PA when the non-diseased cases were obvious.

Endodontic disease is estimated to affect between 7% and 86% of the population[1] and 22 million endodontic procedures are estimated to be performed in the United States of America every year[2]. Endodontically related toothache is estimated to affect up to 88% of people, making it a very common type of pain[3]. Thus, endodontic disease is a significant problem that compromises patients' quality of life[4]. However, endodontic misdiagnosis has been identified in the literature with details of several cases being described[5]. Proper diagnosis is essential in ensuring patient centerd outcomes are met, by correctly treating diseased teeth and not treating healthy teeth.

Periapical (PA) radiography is the main imaging modality and an integral part of endodontic diagnosis and treatment planning[6]. Whilst commonly used, some studies have evaluated its diagnostic performance. An early cadaver study[7] showed simulated pathologic changes in cancellous bone which were seen as radiolucencies if there was cortical bone perforation. An early study showed that PA radiography has good diagnostic performance in detecting endodontic disease[8]. Other cadaver studies detecting apical periodontitis, using histopathology findings as the reference standard reported wide variation in sensitivity ranging from 27% to 60%[9,10] and specificity of 99%[9]. The limitation in using cadaver

From the Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging Science, Faculty of Medicine and Health, School of Health Sciences, The University of Sydney, Camperdown, New South Wales, Australia

The study was approved by the institutional review board of the University of Sydney (Approval number: 2020/477).

Address requests for reprints to Kehn E. Yapp, Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging Science, Faculty of Medicine and Health, School of Health Sciences, The University of Sydney, Camperdown, NSW, Australia.
E mail address: keyapptdo@gmail.com
0099 2399/$ see front matter

https://doi.org/10.1016/j.joen.2023.02.001

studies is that histopathological presence of PA inflammation does not necessarily indicate a disease entity affecting the patient[11]. A recent systematic review revealed a lack of robust evidence on the diagnostic performance of PA radiography and cone beam computed tomography (CBCT) and no studies using clinical data to assess the performance of PA radiography[12].

The use of CBCT in dentistry has increased over time[13], with Australian data showing the national presence of 232 machines in 2014, increasing to 707 machines in 2020[14]. There are now more studies evaluating the diagnostic performance of CBCT than PA radiography, however, these studies have several limitations. No reference standard has been used to validate the index test in CBCT studies[15], published studies utilized sample populations with a disease prevalence of 100%, no data exist on CBCT performance in a nondiseased population[16] and no study has accounted for the severity of disease or nondisease in the population studied[17]. A recent review identified many methodological flaws in the literature on the diagnostic efficacy of PA radiography and CBCT in endodontic disease detection and provided recommendations for addressing these limitations[12]. These limitations include study design issues and high risk of sampling bias. First, the presence of PA radiolucency was often used as the index test to indicate endodontic disease; however, radiolucency is not always pathognomonic of disease[18]. Endodontic disease can also occur in the absence of PA radiolucency[19]; this radiographic change has not been shown to be a valid surrogate outcome measure for endodontic disease. Secondly, study samples are skewed with high disease prevalence. When samples are skewed to have a high prevalence of diseased patients, diagnostic performance metrics from 1 patient group cannot be generalized to another if their clinical features differ[20]. Ideally, the main purpose of a diagnostic performance study is "to provide an estimate of how it would perform on the average in those similar cases and observers and readings that were not studied[21]." This would include patients with endodontic disease that clinically present in an identical manner to nondiseased patients and vice versa, in addition to symptomatic patients with and without clear imaging features of endodontic disease. Thirdly, many studies assessing the diagnostic performance of PA and CBCT imaging used a small sample of readers and were underpowered[12]. Lastly, eader factors such as training[22] and years of reading experience[23] affect observer performance, but no PA radiography or CBCT observer evaluation study has examined the impact of these factors on endodontic disease detection using these imaging modalities. These limitations emphasise the need for well designed studies that consider a wide range of patients with and without the disease who have a broad distribution of clinical features[24], a larger sample of readers with varying experience, a reference standard established by a Delphi panel[25], and abnormality level analysis accounting for readers' confidence in the presence of an abnormality[26].

This study aims to assess and compare reader performance in interpreting digital PA radiography and CBCT in endodontic disease detection, using a free response, factorial study design. The null hypothesis tested is that there is no difference in diagnostic performance between digital PA radiography and CBCT in endodontic disease detection. Understanding the diagnostic performance of each imaging modality in the range of diseased and nondiseased cases will provide evidence to inform imaging pathways for patients with endodontic disease.

METHODS

Study Design

The Standards for Reporting of Diagnostic Accuracy Studies[27] guidelines were used to design a factorial, free response study (Fig. 1). Changes in the independent variables were controlled: case type (diseased and nondiseased), case severity (subtle, moderate, and obvious), reader type (low, medium, and high level of clinical experience) and reading modality (digital PA radiography or CBCT). The study was approved by the institutional review board of the University of Sydney (Approval number: 2020/477).

Image Test Set

Purposeful sampling of 60 cases for each modality (PA radiography and CBCT imaging) was performed and cases divided into 5 categories: diseased subtle (n 14), diseased moderate (n 12), diseased obvious (n 8), nondiseased subtle (n 14), and nondiseased obvious (n 12). Greater weight was allocated to the categories with more diagnostic uncertainty (ie the subtle cases). The moderate diseased category was added because it was deemed a clinically relevant group (Fig. 2). This broad spectrum of cases used deidentified patient data from private endodontic referral practices in Australia, Canada, and the United States of America. These cases were diagnostic in nature not screening and had a presenting concern as determined by the patient and/or referring dentist. Some cases included abnormalities additional to a tooth with endodontic disease. Examples of abnormalities included were dental caries, endodontic disease, periodontal bone loss, tooth resorption, and fracture. Image test sets were unique, for each modality 120 cases were different (60 for PA radiography, 60 for CBCT). This helps reduces demand characteristics regarding observer behaviour[28]. The main tooth distribution was identical for both imaging modalities. Additional abnormalities other than endodontic disease were found in both modalities. For the PA radiography test set, 4 cases had 5 additional abnormalities. The CBCT test set had 26 cases with 34 additional abnormalities.

The consensus of a Delphi panel consisting of 3 endodontists with a total of 79 years' specialist clinical experience was used as the reference standard. In each case, preoperative clinical and radiographic data and follow up data of minimum 6 months' duration were provided to the Delphi panel. The real presenting clinical history and clinical findings taken at each appointment were used. Details of the Delphi consensus procedure establishing the reference standard are described elsewhere[25]. Digital PA radiographs were acquired using the size 1, 1200 × 1600 pixel RVG 6100 sensor (Carestream, GA). CBCT images were acquired using the CS9000 3D unit (Carestream, GA) which produced a 38 × 50 mm field of view volume with 76μ voxel resolution. Exposure settings (mA and kVp) were identical for all cases.

Reader Recruitment

Participant recruitment utilized direct contact, print, and online advertisements targeted at dentists in Victoria, Australia. Dentists with varying amounts of experience were invited to participate. Medical radiologists who reported CBCT images were encouraged to participate. Continuing education credits were given to participants. No monetary reimbursement was given.

Experimental Setting and Reading Strategy

Readers were instructed to view the entire PA radiograph and CBCT imaging volume and report the location of any abnormality. This was designed to replicate the clinical task. They were instructed to provide a brief description for every abnormality with a confidence rating from 1 to 4. A rating of 1 was assigned if no abnormality in the entire case was detected. Ratings of 2, 3, and 4 denoted lowest, middle, and highest levels of confidence for each assigned abnormality. Readers were asked to mark all abnormalities seen in each case. These mark

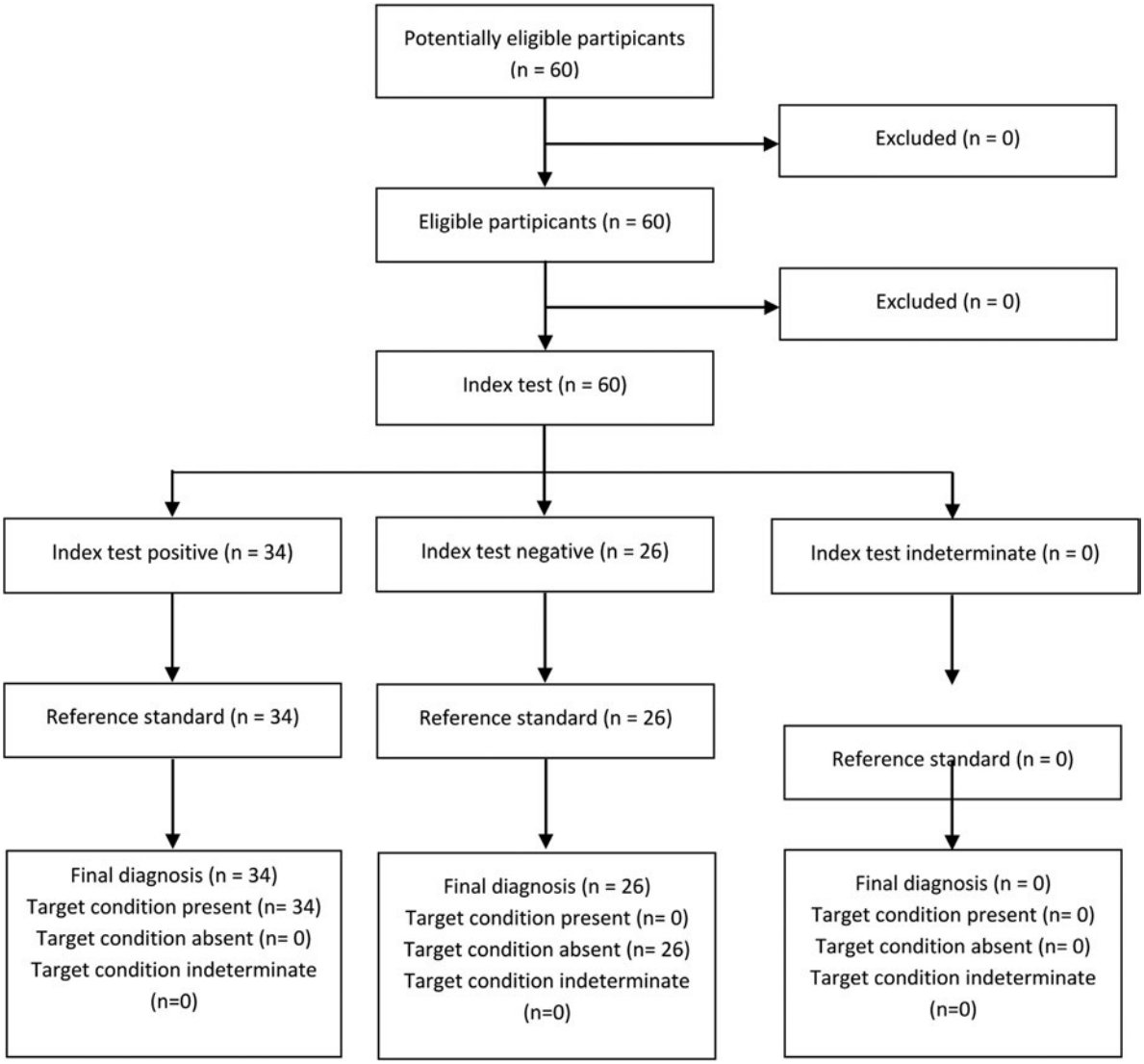


FIGURE 1 – STARD flow chart for both PA radiography and CBCT imaging. STARD, Standards for Reporting of Diagnostic Accuracy Studies; CBCT, Cone beam computed tomography.

rating pairs are measurement characteristics of the free response paradigm where readers are free to identify as many abnormalities as they wish[29]. Participants were not given information about disease or abnormality prevalence only that these cases either had zero abnormalities or at least one abnormality. No information was provided about the expected number of abnormalities per case. No clinical history was provided, and participants had no knowledge of the reference standard. No time limit was imposed; readers could take as much time as required and could take breaks when necessary.

Reading Environment

Images were displayed on an LED monitor with 4096 × 2304 resolution (Apple, CA). The PA radiographs were viewed on a web based platform (DetectED X). Participants were shown how to use the viewing software (Carestream, GA) to view the CBCT images. Both platforms allowed readers to change window/level settings as required, including brightness, contrast, and zoom options. Ambient light measurements were recorded (Lux light meter free Android smartphone application). Screen recording software (Camtasia, TechSmith, MI) was used for verification purposes, if a tooth was mislabeled, or another administrative issue had to be clarified.

Data Analysis

Data were analyzed in the free response paradigm[30]. A correctly localized abnormality, as determined by the Delphi panel, was deemed a lesion localiZation (LL). Every mark not corresponding to an abnormality was classed a non localisation (NL). A non diseased case with at least one NL (rating of 2, 3 or 4) was deemed 0 for specificity. Diagnostic performance metrics were true positive marks or LLs, false positive (FP) marks or non localizations (NLs), specificity, and abnormality sensitivity or LLfraction (LLF). The alternative free response receiver operating characteristic (AFROC1) curve, a plot of the LLF versus the FP fraction in all cases, was used. The figure of merit was a weighted AFROC1 (wAFROC1) statistic, corresponding to the area under the wAFROC1 curve. This gives equal importance to all cases, regardless of the number of abnormalities per case[31]. The wAFROC1 figure of merit is a summary measure of diagnostic performance, which is the probability that a LL will be correctly chosen above a NL, when a choice of 2 marks

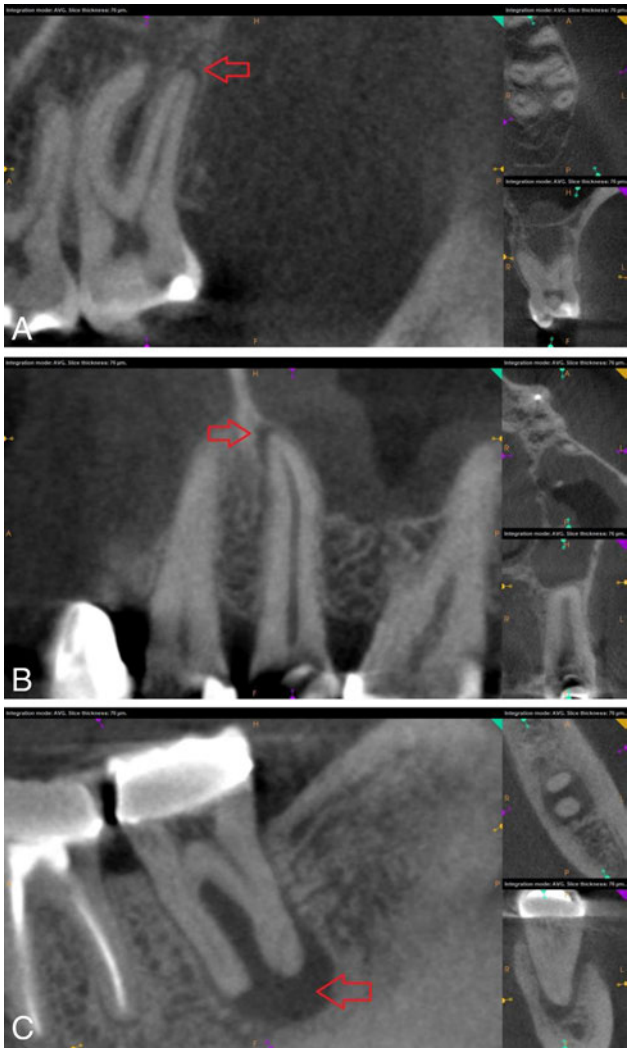


FIGURE 2 – Case examples (red arrows) of (*A*) diseased subtle (*B*) diseased moderate (*C*) diseased obvious.

is given. Different diagnostic performance metrics were used to test diseased and nondiseased test sets. In the diseased groups (subtle, moderate, and obvious) a wAFROC1 statistic, NLs(or FPs), and LLF were used to assess performance. In the nondiseased groups, specificity and FPs were used to assess reader performance. The PA radiography nondiseased cases did not have any additional abnormalities; therefore, comparative wAFROC1 analyses could not be performed for both modalities in the nondiseased groups.

Power calculation aimed to have a minimum 80% power with α 0.05, based on a sample size of 60 cases and an effect size of a wAFROC1 figure of merit of 0.075. Recruiting 20 readers exceeded the minimum to have 91.56% power (The RJafroc Quick Start Book, https://dpc10ster.github.io/RJafrocQuickStart/). Pilot data were obtained from the CBCT interpretations of like cases by 3 experienced endodontists not involved in the study. This provided a mean wAFROC1 difference of 0.075 between readers as the effect size. Data distribution was established using a Shapiro Wilk test, which showed that the data were non normal (W 0.863, $P < .01$). A Kruskal Wallis test was used to analyze subgroups. The mean diagnostic performances of all readers were compared using a Wilcoxon signed rank test. *P* values of <.05 were deemed significant. All analyses were conducted using RJafroc 2.0.1 in RStudio 4.0.3 (RStudio, PBC, MA).

TABLE 1 - Reader Characteristics

	Number	Mean	Range
Clinical experience			
0 9 y	4	3.8	0 8
10 19 y	12	14.4	12 19
20 or more y	6	38.8	23 48

RESULTS

Data from the 22 readers who consented to the study were used, 20 dentists (14 general and 6 specialists) and 2 medical radiologists. All data were included from the participants, with no exclusions due to missing or incomplete responses. Reading conditions had a mean ambient light value of 34.5 Lux (range 16 62 Lux). Clinical experience ranged from 0 to 48 years with a mean of 19.1 years. Reader characteristics are shown in Table 1.

Nondiseased Group

Reader performance analysis in the nondiseased subtle category showed no difference between PA and CBCT in specificity (mean: PA 0.48 [range: 0.30 0.52] versus CBCT 0.46 [range: 0.43 0.46]; *P* .67). Specificity was significantly higher for CBCT than PA (mean: CBCT 0.55 [range: 0.53 0.60] versus PA 0.38 [range: 0.25 0.43]; *P* .01) in the obvious nondiseased group. Years of clinical experience had no effect on specificity for any of the nondiseased subgroups in both PA and CBCT. There were also no significant differences in specificity between PA and CBCT for each experience groupings (Table 2).

FP marks were not significantly different between PA radiography and CBCT ($P > .05$ for both subtle and obvious cases). For the nondiseased subtle cases, the mean FP mark was 11.8 (range: 10.1 16.8) for PA and 13.6 (range: 11.2 16.8) for CBCT. For obvious cases, the mean FP marks were 12.5 (range: 11.1 16.8) and 10.1 (range: 9.2 10.9) for PA and CBCT respectively. Clinical experience did not significantly affect the number of FPs, neither was the type of modality in each experience category.

Diseased Group

Reader performance as assessed using wAFROC1 metrics showed that for both modalities, diagnostic performance improved with increasing disease severity (Table 2). PA radiography performed significantly better in the subtle disease (PA 0.54 vs CBCT 0.48; *P* .02) and moderate diseased (PA 0.60 vs CBCT 0.50; *P* .01) groups (Table 3). No significant

TABLE 2 - Non-diseased Case Types Mean Values 95% Confidence Intervals in Brackets

	Nondiseased - subtle						Nondiseased - obvious					
	Specificity			False positives			Specificity			False positives		
	PA	CBCT	*P* value	PA	CBCT	*P* value	PA	CBCT	*P* value	PA	CBCT	*P* value
Overall												
All readers	0 48 (0 40–0 56)	0 46 (0 38–0 53)	67	11 8 (9 2–14 6)	13 6 (10 3–16 8)	52	0 38 (0 31–0 45)	0 55 (0 47–0 64)	05*	12 5 (10 3–14 7)	10 1 (7 4–12 9)	09
Clinical experience												
0–9 y	0 30 (0 44–0 56)	0 43 (0 39–0 61)	47	16 8 (9 7–14 3)	16 8 (4 2–15 8)	56	0 25 (0 09–0 24)	0 60 (0 72–0 95)	08	16 8 (15 6–20 4)	9 3 (–0 1–6 1)	11
10–19 y	0 52 (0 43–0 57)	0 46 (0 21–0 36)	44	10 1 (11 9–16 1)	13 8 (22 0–27 9)	17	0 43 (0 44–0 56)	0 53 (0 07–0 26)	17	11 1 (12 2–15 8)	10 9 (21 9–28 1)	45
20 or more y	0 52 (0 55–0 74)	0 46 (0 59–0 70)	52	12 2 (2 3–9 7)	11 2 (4 3–7 7)	94	0 36 (0 58–0 75)	0 56 (0 53–0 63)	09	12 5 (2 6–7 4)	9 2 (3 9–8 1)	23
P value	15	99		16	92		22	81		18	77	

CBCT Cone beam computed tomography PA periapical
*Significant at $P < .05$ from Kruskal-Wallis tests

difference was found between PA and CBCT in the obvious diseased group (PA 0.68 vs CBCT 0.70; *P* .44). When these modalities were compared according to the years of experience of the readers, PA demonstrated higher wAFROC1 values than CBCT for all categories except for 2 (obvious cases, 10 19 and 20+ years of experience with CBCT having higher values of 0.01 and 0.09 respectively). However, the differences in wAFROC1 between modalities across years of experience were not significant, neither were the wAFROC1 values within each imaging modality different across readers with different years of experience.

Table 4 shows data on FP marks; CBCT imaging had significantly higher mean FPs than PA radiography (9.8 vs 7.1 respectively; $P < .05$) in subtle cases. No significant differences were found in diseased moderate cases (PA 8.7 for vs CBCT 8.6; *P* .74) and in obvious cases (PA 4.9 vs CBCT 5.6; *P* .63). Years of clinical experience also did not significantly impact upon the differences in the number of FP marks between PA and CBCT ($P > .05$ for all). However, in CBCT, readers with more years of experience demonstrated fewer FP marks only in the obvious group (*P* .04).

LL fraction differed across diseased groups (Table 5). Mean LLF was slightly higher for CBCT than PA in subtle (CBCT 0.44 vs PA 0.43; *P* .42) and obvious (CBCT 0.72 vs PA 0.69; *P* .37) groups but these were not significantly different. In the moderate disease category, PA demonstrated a significantly higher LLF than CBCT (0.63 vs 0.45 respectively; *P* .003). Years of clinical experience had no significant effect on LLF within modalities ($P > .24$ for all). When comparing modalities at each level of readers' years of clinical experience, significant differences were only observed for moderate disease cases, where PA had a higher LLF than CBCT (0.73 vs 0.51 respectively; *P* .02).

DISCUSSION

Our study showed that when a significant difference in diagnostic performance was found, PA radiography outperformed CBCT imaging in all but one category, the obvious nondiseased group where CBCT demonstrated better performance. Overall, the differences in specificity for both modalities in both subtle and obvious nondiseased images were small. A previous study using clinical records on CBCT performance in detecting a PA radiolucency with low disease prevalence reported wide variations (0.26 0.80) in reader specificity, depending on criteria for a

TABLE 3 - Diseased Case Types–Mean wAFROC1 Statistic Values 95% Confidence Intervals in Backets

	Diseased subtle			Diseased moderate			Diseased obvious		
	PA	CBCT	*P* value	PA	CBCT	*P* value	PA	CBCT	*P* value
Overall									
All readers	0 54 (0 50–0 58)	0 48 (0 43–0 54)	< 05*	0 60 (0 56–0 65)	0 50 (0 44–0 56)	< 05*	0 68 (0 63–0 72)	0 70 (0 66–0 75)	44
Clinical experience									
0–9 y	0 49 (0 39–0 47)	0 38 (0 26–0 38)	15	0 67 (0 64–0 73)	0 50 (0 44–0 54)	08	0 70 (0 60-0 67)	0 68 (0 71–0 77)	56
10–19 y	0 54 (0 36–0 46)	0 49 (0 29–0 40)	34	0 59 (0 43–0 51)	0 48 (0 31–0 46)	16	0 68 (0 62-0 69)	0 69 (0 58–0 67)	84
20 or more y	0 57 (0 66–0 75)	0 54 (0 44–0 52)	63	0 60 (0 74–0 85)	0 54 (0 48–0 57)	52	0 65 (0 73-0 85)	0 74 (0 57–0 65)	34
P value	43	18		52	72		83	68	

CBCT Cone beam computed tomography PA periapical
*Significant at $P < 05$

nondiseased image[32]. This wide variation may have been due to the different characteristics of the index test of a CBCT PA index (CBCT PAI) scale. Instead of a clinically relevant dichotomous call about the presence or absence of disease, it used the measurement of the size of a PA radiolucency, without reference to the state of disease or level of abnormality[33]. The CBCT PAI may also have a broad range of results because a PA radiolucency is not pathognomonic for disease[18]. The absence of a PA radiolucency can also occur when endodontic disease is present, such as in irreversible pulpitis[19]. Cadaver studies using histopathology as the reference standard for endodontic disease had much higher specificity values of 0.99 for PA and 1.00 for CBCT[9] and specificity of 0.79 for CBCT[10]. These differences are likely due to methodological issues of using a binary classifier[9], which may bias the reader into declaring an absence of a "PA lesion" if the option of selecting a subtle radiolucency is unavailable. Another issue is measurement at the root level[10], which biases reader behavior by directing attention to areas of the image that may not have been visualized, and rewards readers for determining if some roots are normal and others are abnormal within the same tooth. These cadaver studies also have unknown disease prevalence and severity in their samples, which may skew the images to be easier to correctly identify the absence (or presence) of endodontic disease. Our reader assignment using the free response paradigm is designed to replicate the clinical task, which both rewards correct localizations and penalizes NLs at the same time[29]. It allows for and counts both true positive (LL) and FP (NL) marks to occur in the same image, a clinically relevant feature that has not been used in previous studies evaluating diagnostic image performance in the endodontic domain[12]. This overcomes errors previously overlooked in other studies and has been shown to yield higher power than Receiver Operating Characteristic (ROC) methodology that uses measurement at the case level[34].

In diseased images, PA radiography demonstrated a greater performance trend with higher wAFROC1 values and fewer FPs than CBCT in many categories. Diagnostic performance improved with increasing disease severity across both PA radiography and CBCT, as well as fewer CBCT FPs as disease became more severe. The decrease in FP marks may be similar to a satisfaction of search phenomenon, where readers satisfy a "quest for meaning"[35] once a more obvious abnormality has been found and are less likely to make further marks on the image. An increasing trend in diagnostic performance

TABLE 4 - Diseased Case Types Mean False Positives, 95% Confidence Intervals in Brackets

	Diseased subtle			Diseased moderate			Diseased obvious		
	PA	CBCT	*P* value	PA	CBCT	*P* value	PA	CBCT	*P* value
Overall									
Overall	7.1 (5.3 8.9)	9.8 (6.9 12.7)	<.05*	8.7 (6.8 10.6)	8.6 (5.9 11.4)	.74	4.9 (3.6 6.2)	5.6 (4.1 7.1)	.63
Clinical experience									
0 9 y	8 (8.7 11.4)	17.5 (11.5 18.5)	.14	11.5 (4.4 9.6)	9.8 (4.4 9.6)	.89	6.5 (3.9 8.2)	7 (3.8 6.2)	.67
10 19 y	7.3 (13.0 17.0)	8.5 (18.3 23.7)	.77	7.8 (7.3 10.7)	9.8 (10.7 17.3)	.58	4.3 (3.9 6.1)	6.6 (4.4 7.6)	.12
20 or more y	6.2 (2.3 5.7)	7.2 (4.7 7.3)	.52	8.7 (0.3 3.7)	5.7 (1.8 4.2)	.23	5.2 (−0.1 2.1)	2.7 (2.2 3.8)	.11
P value	.72	.12		.41	.51		.73	<.05*	

CBCT, Cone beam computed tomography; PA, periapical.
*Significant at $P < .05$.

with disease severity has been previously reported in a study comparing film and digitized film PA radiography in detecting endodontic and periodontal disease using the ROC paradigm[36]. However, eye tracking studies are needed to establish the phenomenon responsible for lower FP rates in obvious diseased cases and if this finding affects performance in terms of LLs and NLs. Previous studies have reported that CBCT outperformed PA in terms of sensitivity[10] and area under the ROC curve[9]. An animal study has also reported higher sensitivity for CBCT compared to PA imaging in detecting apical periodontitis[37]. These histopathological studies used a reader task of determining the presence of apical periodontitis or PA lesions, which is not representative of a clinical task that is describing the presence of any abnormality in the entire image as used in our study. A localizing prompt embedded in the CBCT search task has been shown to direct the reader's attention to a region with an abnormality, which is more readily identified[38]. These methodological differences may explain the difference between our results and that of published studies.

Whilst diagnostic performance can be influenced by technological and reader factors, it should be noted that the results represent the image interpretation by the same cohort of readers.

The years of clinical experience of readers did not affect performance except in the diseased obvious disease group, where readers with more years of experience demonstrated significantly fewer FPs. Whilst the same trend was found in the subtle disease group for both PA and CBCT, these were not significantly greater than zero. The lower FP rates for readers with more years of experience may be due to older adults having increased effectiveness of emotion regulation and a greater reluctance to make decisions[39]. Years of experience reading normal PA images may increase the threshold for calling an abnormality. A similar finding has been reported in mammography, where older radiologists demonstrated higher specificity[40] and lower FP rate[41]. More years of experience interpreting mammograms had decreased FP rates[42]. Our study accounted for reader experience; however, as PA radiography has been the long standing modality for endodontic disease detection and treatment planning, it is possible that readers' familiarity with PA radiography and low experience with CBCT may have influenced the results.

One important point worth mentioning is the overall suboptimal performance, with average specificity, sensitivity, and LL all below 75%. These findings highlight that specialist training and development of expertise in radiological image interpretation requires a substantial amount of effort and time opportunities not available for dentists and specialists. The results indicate that specialized training programs to improve interpretation skills of dentists are needed, particularly for the newly introduced CBCT. Studies have shown that strategies that support readers to interact with image test sets and receive immediate feedback on their performance and mitigate diagnostic errors improved performance by 34%, with trainees benefitting from immediate feedback[43]. Such strategies can be explored to improve the interpretive skills of dentists and specialists so that the benefits of imaging patients with endodontic issues can be fully accrued.

This study has some limitations. First, readers interpreted the images in an experimental setting, which may introduce demand characteristics, experimental artifacts that alters behavior to suit the perceived goals of the study[28]. It is not practical for dentists to only read the images in a clinical setting without seeing the patient, as performed in a medical radiology study[44], because the dentist is the ordering and interpreting clinician for any radiographs, and often the intervening clinician for patient treatment. However, a mammography study has shown a good level of concordance between reader performance in clinical and laboratory settings[45]. The reader population included mostly dentists with only 2 radiologists, who may not be representative of the population of radiologists who do interpret dental radiographs; however, the prevalence of radiologists reporting dental imaging is unknown. Secondly, readers were mostly general dentists who constitute the majority of the dental practitioner population ordering and interpreting CBCT and PA radiography images in Australia. In addition, the radiologists who participated in the study were not CBCT specialists although they interpret CBCT images in their practices. Therefore, the performance values recorded in our study may reflect the abilities of the dentists rather than the diagnostic performance of CBCT and PA radiography. Thus, our findings should be interpreted with caution since most of the readers may have limited experience with CBCT interpretation.

Thirdly, it would have been better to assess the association between years of experience reading CBCT images and diagnostic performance. However, since CBCT is a relatively new imaging tool, there were not many readers with significant or varying years of experience reading CBCT images to allow for a robust comparison. In addition, a previous study[46] showed no association between annual volume of CBCT images read per year and reader performance. Therefore, it was logical to assess the association between clinical experience and diagnostic performance. Our findings suggest that until general dentists who routinely interpret dental CBCT images improve their CBCT interpretation skills, the benefits of CBCT in endodontic practice can be fully achieved. Thus, programs to educate and improve the performance of general dentists in the interpretation of CBCT images are urgently needed.

Another potential limitation of the study is that our analysis considers CBCT as a primary imaging tool, albeit most clinical settings use CBCT as an adjunct to PA

TABLE 5 - Diseased Case Types–Mean Lesion Localisation Fraction 95% Confidence Intervals in Brackets

	Diseased subtle			Diseased moderate			Diseased obvious		
	PA	CBCT	*P* value	PA	CBCT	*P* value	PA	CBCT	*P* value
Overall									
Overall	0 43 (0 36–0 49)	0 44 (0 38–0 50)	42	0 63 (0 55–0 70)	0 45 (0 38–0 52)	< 05*	0 69 (0 63–0 75)	0 72 (0 67–0 77)	37
Clinical experience									
0–9 y	0 42 (0 28–0 34)	0 51 (0 51–0 55)	08	0 73 (0 67–0 71)	0 51 (0 62–0 74)	< 05*	0 75 (0 67–0 72)	0 69 (0 71–0 83)	56
10–19 y	0 41 (0 17–0 33)	0 40 (0 29–0 44)	60	0 58 (0 22–0 39)	0 42 (0 38–0 53)	08	0 68 (0 54–0 66)	0 70 (0 65–0 74)	82
20 or more y	0 47 (0 46–0 54)	0 47 (0 21–0 31)	75	0 64 (0 55–0 68)	0 47 (0 20–0 34)	15	0 68 (0 63–0 77)	0 77 (0 56–0 67)	42
P value	72	24		33	68		52	45	

CBCT Cone beam computed tomography PA periapical
*Significant at $P < 05$

radiography. Other disciplines have recommended to "Avoid taking conventional 2D radiographs if the clinical examination indicates that a CBCT study is indicated for proper diagnosis and/or treatment planning or if a recent CBCT study is available[47]." Whilst the use of CBCT as a primary imaging tool may challenge legacy concepts of "As Low As Reasonably Attainable" and guidelines of the clinician gathering all clinical history and 2D imaging prior to ordering CBCT[48], this study provides evidence for the use of CBCT as a primary imaging modality.

The strengths of our study include the factorial design which allows analysis of the relevant variables, namely case type and disease severity, free response analysis which measures diagnostic performance at the abnormality level, and the use of a valid and independent reference standard. These address limitations of previous studies on diagnostic performance of PA and CBCT identified in a recent systematic review[12]. The purposeful sampling of the dataset ensured that the main abnormalities were proportionately equally distributed across the 2 imaging modalities. In published image interpretation studies, only disease status and severity have been identified to have an impact on diagnostic performance. Other characteristics such as patient age, gender, and regional location have not been shown to affect diagnostic performance. Whilst it may be possible that there are confounding factors considered to be "unknown unknowns," our study has controlled for established factors that influence performance. Another strength is that readers were not given clinical history, prompts or other patient information that is an interpretive bias and has been shown to affect CBCT image interpretation[46]. It may be argued that this could be a limitation and favor PA radiography interpretation, given the potentially smaller field of view. However, CBCT, like any other volumetric imaging tool, was developed to help overcome the limitations of 2D imaging. This provides opportunities to maximize diagnosis. In addition, the use of the CBCT software used in our study can create a pseudopanoramic, pseudoPA view, where the reconstruction shows the entire effective volume in a single view. Therefore, the reader does not necessarily have to always scroll through all volumetric data or be distracted or inconvenienced by the volumetric nature of CBCT.

CONCLUSIONS

Reader performance in the detection of endodontic disease is better with PA radiography than CBCT, but reading CBCT is

associated with fewer FP errors as disease becomes more severe. Reader performance with PA radiography and CBCT improve as disease features become more prominent. Clinical experience does not impact upon the accuracy of interpretation of both PA radiography and CBCT; however, experienced readers make fewer FPs errors when interpreting CBCT images with obvious disease features.

ACKNOWLEDGMENTS

This is a nonfunded project.

The authors deny any conflicts of interest.

REFERENCES

1. Persoon IF, Özok AR. Definitions and epidemiology of endodontic infections. Curr Oral Health Rep 2017;4:278–85.
2. 2005–2006 Survey of Dental Services Rendered. Am Dent Assoc Chicago 2007:5.
3. Kakoei S, Parirokh M, Nakhaee N, et al. Prevalence of toothache and associated factors: a population-based study in southeast Iran. Iran Endod J 2013;8:123–8.
4. Locker D, Quiñonez C. To what extent do oral disorders compromise the quality of life? Community Dent Oral Epidemiol 2011;39:3–11.
5. Sirotheau Corrêa Pontes F, Paiva Fonseca F, Souza de Jesus A, et al. Nonendodontic lesions misdiagnosed as apical periodontitis lesions: series of case reports and review of literature. J Endod 2014;40:16–27.
6. Berman L, Rotstein I. Diagnosis. In: Hargreaves K, Berman L, editors. Pathways of the Pulp. St Louis, MO: Elsevier; 2016. p. 2–32.
7. Bender IB, Seltzer S. Roentgenographic and direct observation of experimental lesions in bone: I. J Am Dent Assoc 1961;62:152–60.
8. Reit C, Gröndahl HG. Application of statistical decision theory to radiographic diagnosis of endodontically treated teeth. Scand J Dent Res 1983;91:213–8.
9. Kanagasingam S, Lim CX, Yong CP, et al. Diagnostic accuracy of periapical radiography and cone beam computed tomography in detecting apical periodontitis using histopathological findings as a reference standard. Int Endod J 2017;50:417–26.
10. Kruse C, Spin-Neto R, Kraft DCE, et al. Diagnostic accuracy of cone beam computed tomography used for assessment of apical periodontitis: an *ex vivo* histopathological study on human cadavers. Int Endod J 2019;52:439–50.
11. Villa-Machado PA, Restrepo-Patiño DM, Calvo-Trejos JP, et al. Cone beam computed tomographic and micro computed tomographic evaluations of the root apexes of teeth with posttreatment apical periodontitis. J Endod 2020;46:1695–701.
12. Yapp KE, Brennan P, Ekpo E. Endodontic disease detection: digital periapical radiography versus cone beam computed tomography – a systematic review. J Med Imaging (Bellingham) 2021;8:041205.
13. Brown LF, Monsour P. The growth of medicare rebatable cone beam computed tomography and panoramic radiography in Australia. Aust Dent J 2015;60:511–9.
14. Lam M, Critchley S, Zhang A, Monsour P. Current trends in the adoption and education of cone beam computed tomography and panoramic radiography machines across Australia. Dentomaxillofac Radiol 2021;50:20200380.
15. Estrela C, Bueno M, Leles C, et al. Accuracy of cone beam computed tomography and panoramic and periapical radiography for detection of apical periodontitis. J Endod 2008;34:273–9.
16. Patel S, Wilson R, Dawood A, Mannocci F. The detection of periapical pathosis using periapical radiography and cone beam computed tomography – Part 1: pre-operative status. Int Endod J 2012;45:702–10.
17. Abella F, Patel S, Duran-Sindreu F, et al. An evaluation of the periapical status of teeth with necrotic pulps using periapical radiography and cone-beam computed tomography. Int Endod J 2014;47:387–96.
18. Molven O, Halse A, Fristad I, MacDonald-Jankowski D. Periapical changes following root canal treatment observed 20–27 years postoperatively. Int Endod J 2002;35:784–90.

19. Abella F, Patel S, Duran Sindreu F, et al. Evaluating the periapical status of teeth with irreversible pulpitis by using cone beam computed tomography scanning and periapical radiographs. J Endod 2012;38:1588 91.

20. Scott IA, Greenberg PB, Poole PJ. Cautionary tales in the clinical interpretation of studies of diagnostic tests. Intern Med J 2008;38:120 9.

21. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging 1989;29:307 35.

22. Buissink C, Thompson JD, Voet M, et al. The influence of experience and training in a group of novice observers: a jackknife alternative free response receiver operating characteristic analysis. Radiography 2014;20:300 5.

23. Rawashdeh M, Lee W, Bourne R, et al. Markers of good performance in mammography depend on number of annual readings. Radiology 2013;269:61 7.

24. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design related bias in studies of diagnostic tests. JAMA 1999;282:1061 6.

25. Stheeman SE, van't Hof MA, Mileman PA, van der PF. Stelt Use of the Delphi technique to develop standards for quality assessment in diagnostic radiology. Community Dent Health 1995;12:194 9.

26. Chakraborty DP. A brief history of free response receiver operating characteristic paradigm data analysis. Acad Radiol 2013;20:915 9.

27. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6:e012799.

28. Orne MT. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. Washington, DC: Am Psychologist; 1962. p. 776 83.

29. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free response paradigm. Phys Med Biol 2006;51:3449 62.

30. Bunch P, Hamilton J, Sanderson G. A free response approach to the measurement and characterization of radiographic observer performance. Appl Opt Instrum Med VI 1977;0127:124 35.

31. Chakraborty DP, Zhai XT. On the meaning of the weighted alternative free response operating characteristic figure of merit. Med Phys 2016;43:2548.

32. Pope O, Sathorn C, Parashos P. A comparative investigation of cone beam computed tomography and periapical radiography in the diagnosis of a healthy periapex. J Endod 2014;40:360 5.

33. Estrela C, Bueno MR, Azevedo BC, et al. A new periapical index based on cone beam computed tomography. J Endod 2008;34:1325 31.

34. Chakraborty D. Statistical power in observer performance studies: Comparison of the receiver operating characteristic and free response methods in tasks involving localization. Acad Radiol 2002;9:147 56.

35. Tuddenham WJ. Visual search, image organization, and reader error in roentgen diagnosis. Studies of the psycho physiology of roentgen image perception. Radiology 1962;78:694 704.

36. Khademi J. The Effects of Digitization on Diagnostic Performance in Endodontics and Periodontics: an ROC study. Iowa: The University of Iowa; 1994.

37. de Paula Silva FWG, Wu MK, Leonardo MR, et al. Accuracy of periapical radiography and cone beam computed tomography scans in diagnosing apical periodontitis using histopathological findings as a gold standard. J Endod 2009;35:1009 12.

38. Nguyen A. The effect of a localizing prompt on cone beam computed tomography interpretation. St Louis, MO: Saint Louis University; 2017.

39. Mather M. A Review of decision making processes: weighing the risks and benefits of aging. In: Carstensen L, Hartel C, editors. National Research Council (US) Committee on Aging Frontiers in Social Psychology, Personality, and Adult Developmental Psychology. Washington, DC: National Academies Press (US); 2006.

40. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004;96:1840 50.

41. Tan A, Freeman DH, Goodwin JS, Freeman JL. Variation in false positive rates of mammography reading among 1067 radiologists: a population based assessment. Breast Cancer Res Treat 2006;100:309 18.

42. Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology 2009;253:641–51.

43. Suleiman WI, Rawashdeh MA, Lewis SJ, et al. Impact of breast reader assessment strategy on mammographic radiologists' test reading performance. J Med Imaging Radiat Oncol 2016;60:352–8.

44. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. Am J Roentgenol 1981;137:1055–8.

45. Soh BP, Lee WB, McEntee MF, et al. Mammography test sets: reading location and prior images do not affect group performance. Clin Radiol 2014;69:397–402.

46. Yapp KE, Brennan P, Ekpo E. The effect of clinical history on diagnostic performance of endodontic cone beam CT interpretation. Clin Radiol 2023. https://doi.org/10.1016/j.crad.2022.12.005. in press.

47. Clinical recommendations regarding use of cone beam computed tomography in orthodontics. Position statement by the American Academy of Oral and Maxillofacial Radiology. Oral Surg Oral Med Oral Pathol Oral Radiol 2013;116:238–57.

48. Use of Cone Beam Computed Tomography in Endodontics 2015/2016 Update. Chicago, IL: American Association of Endodontists and the American Academy of Oral and Maxillofacial Radiology; 2016. p. 1–6.

CHAPTER 5

5.1 Bridging section to Chapter 5

The systematic review in chapter three provided updated information on the effect of clinical history on diagnostic image interpretation. It found that there was a lack of data using more advanced methodologies for assessing diagnostic performance. Limitations in the published literature include a lack of data on location information of abnormalities and adjustment for cases with multiple abnormalities (McNeil et al. 1983, Good et al. 1990, Quekel et al. 2001). Another finding was that disease prevalence in the studied samples had large variation and readers had a limited range of experience. The type of history provided varied from limited information to a sham clinical history. The methodology bias regarding balancing reading sessions and accounting for the time between sessions was also poorly controlled.

Chapter five describes a study undertaken to overcome the limitations identified in studies by using a factorial, free-response, crossover study. Readers interpreted the same 60 CBCT images twice – with and without history – over two reading sessions, with an adequate interval between reading sessions that were each divided into cases provided with and without real clinical history. This occurred without knowledge of the "double reading" nature of the study. The following independent variables were accounted for: case type (diseased or non-diseased), case severity (subtle, moderate or obvious), reader type (low or high level of experience) and reading modality (CBCT with or without clinical history). Data analysis used the weighted Alternative Free Response Operating Characteristic 1 (wAFROC1) metric, lesion localisation fraction, false positive fraction, and specificity.

Clinical history was found to improve performance in diseased cases and slightly reduce performance in non-diseased cases. Readers with less experience and low monthly reading volume had greater performance in subtle diseased cases with clinical history. These findings suggest that there is a mixed effect of having clinical history prior to interpreting CBCT images, with no clear evidence supporting the interpretation of CBCT imaging with history. This could also be considered a criterion shift, if this study were conducted in the ROC paradigm. However, the free response paradigm does not limit readers in marking image locations for suspected disease. Because there is no relationship between true and false positive calls, the criterion shift cannot be measured in the free response paradigm. The findings of the study provide new insight into the effect of clinical history on diagnostic performance of CBCT in endodontic disease detection in different case and reader types.

This paper is published in Clinical Radiology as:

Yapp KE, Brennan P, Ekpo E. The effect of clinical history on diagnostic performance of endodontic cone beam computed tomography interpretation. Clinical Radiology 2023, 78(5):e433-e441. https://doi.org/10.1016/j.crad.2022.12.005

References

Good, B. C., et al. (1990). "Does knowledge of the clinical history affect the accuracy of chest radiograph interpretation?" American Journal of Roentgenology **154**(4): 709-712.
McNeil, B. J., et al. (1983). "Paired receiver operating characteristic curves and the effect of history on radiographic interpretation - CT of the head as a case study." Radiology **149**(1): 75-77.
Quekel, L. G. B. A., et al. (2001). "Detection of lung cancer on the chest radiograph: impact of previous films, clinical information, double reading, and dual reading." Journal of Clinical Epidemiology **54**(11): 1146-1150.

Clinical Radiology 78 (2023) e433 e441

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

The effect of clinical history on diagnostic performance of endodontic cone-beam CT interpretation

K.E. Yapp*, P. Brennan, E. Ekpo

Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging Science, School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Australia

ARTICLE INFORMATION

Article history:
Received 25 July 2022
Received in revised form
21 November 2022
Accepted 9 December 2022

AIM: To assess the effect of clinical history on the interpretation of endodontic disease in dental cone beam computed tomography (CBCT).

MATERIALS AND METHODS: A reader performance study of an image test set was undertaken using a factorial, free response, crossover design, accounting for the independent variables: case type, case severity, reader type, and reading modality. Twenty three readers interpreted 60 CBCT images twice over two reading sessions using a balanced design, once with access to clinical history and once without, where 30 in each session included history. Lesion localisations, specificity, false positive marks and the weighted alternative free response receiver operating characteristic (wAFROC1) figure of merit were calculated.

RESULTS: Clinical history had no significant effect on specificity and false positive rates in non diseased cases ($p>0.05$), but improved lesion localisation in subtle and obvious diseased cases ($p<0.01$). wAFROC1 values were higher with clinical history for subtle (0.58 versus 0.48; $p<0.001$) and obvious (0.77 versus 0.71; p 0.006) diseased categories. No associations were observed between clinical history and both readers' years of experience and reading volume in the non diseased categories. Readers with fewer (p 0.03) and moderate (p 0.008) years of experience and low (p 0.002) CBCT reading volume demonstrated better lesion localisation in subtle diseased cases when clinical history was available.

CONCLUSIONS: Clinical history improved the interpretation of CBCT images with disease without affecting the interpretation of images without disease. Less and moderately experienced readers and low volume readers benefitted more from availability of clinical history.

Introduction

Medical image interpretation is subject to cognitive errors and biases.[1] Cognitive biases can lead to diagnostic errors, including misdiagnosis, oversight of abnormalities, or image misinterpretation.[2] Access to a patient's clinical history is a cognitive bias (attribution bias), particularly if the presentation of clinical information influences the

* Guarantor and correspondent: Kehn E Yapp, Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging Science, School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, NSW, Australia.
E mail address: keyapptdo@gmail.com (K.E. Yapp).

https://doi.org/10.1016/j.crad.2022.12.005
0009-9260/

diagnostic decision.[3] An early review on clinical history showed diagnostic accuracy improved at the case level when clinical information was available.[4]

Dentists begin with[5] and are trained to obtain[6] a clinical history before ordering and interpreting radiographic images. The American Academy of Pediatric Dentistry recommends, "the need for dental radiographs can be determined only after consideration of the patient's medical and dental histories".[7] These recommendations are consistent with the practice of dentists collecting clinical history and examination findings before ordering and interpreting radiographic imaging. Interestingly, the inherent bias in these recommendations can lead to an increased risk of interpretive errors by directing (or misdirecting) the image interpretation based on the clinical history already gathered.[8] A recent systematic review focusing on diagnostic performance measurement at the abnormality level found that clinical history did not improve performance, but instead had a tendency to increase false positives.[9] Because many studies have measured diagnostic performance and image interpretation at the case level, they have been unable to account for abnormality location nor multiple abnormalities per case.[10] Other interpretive errors can go unnoticed when measuring image interpretation at the case level, such as a location-level false positive and false negative occurring on the same image when a location of an abnormality is misidentified. An image-level positive occurs, which strictly is a correct finding; two errors occur: an incorrect abnormality location reported and an abnormality missed.[11] This is clinically significant as treatment can occur at the wrong location and/or the missed abnormality may grow and increase in severity.[12] These limitations recommend for studies to examine the effect of clinical history on diagnostic image interpretation and allow for measuring multiple abnormalities per image, and reward correct and penalise incorrect abnormality locations.[9]

Cone-beam computed tomography (CBCT) has been used in dentistry for two decades and has gained popularity[13] as an adjunct to intra-oral radiography and orthopantomography. The three-dimensional imaging of CBCT has supported its use in many disciplines such as endodontics, implant placement, and oral surgery.[14] Guidelines on CBCT indications have not been shown to be based on experimental evidence evaluating diagnostic performance. The SEDENTEXCT European Commission has "evidence-based guidelines" on CBCT for dental and maxillofacial radiology stating, "CBCT examinations must not be carried out unless a history and clinical examination have been performed".[15] A joint position statement from the American Association of Endodontists and American Academy of Oral and Maxillofacial Radiology similarly states, "CBCT should be used only when the patient's history and a clinical examination demonstrate that the benefits to the patient outweigh the potential risks".[16] The consistent culture of dentistry is that the clinical history must be acquired and processed before determining if diagnostic imaging should be ordered and interpreted.

The effect of the ordering recommendation on the diagnostic performance of CBCT interpretation is unclear. Having the clinical history before reading CBCT images may be subject to the same cognitive bias as those for other diagnostic tests.[9] Although there is a tendency in dentistry to introduce the clinical history bias prior to image interpretation, its effect has never been evaluated in dental CBCT interpretation. Therefore, the aim of this study was to assess the effect of clinical history on diagnostic performance on endodontic (pulpal and/or periradicular) disease detection in dental CBCT interpretation. Understanding the effect of clinical history on diagnostic efficacy may provide evidence to validate the recommendations around CBCT imaging and data to inform training programmes in CBCT interpretation.

Materials and methods

Study design

This study followed the Standards for Reporting of Diagnostic Accuracy Studies (STARD)[17] guidelines and used a factorial, free-response, crossover study design. This controlled for changes in the independent variables: case type (diseased and non-diseased), case severity (subtle, moderate, and obvious), reader type (low, medium, and high level of clinical experience), and reading modality (with and without access to clinical history prior to image interpretation). The study was approved by the institutional review board of The University of Sydney.

Image dataset and grouping

Purposeful retrospective sampling of 60 cases was performed and cases divided into five categories: diseased—subtle (n=14), diseased—moderate (n=12), diseased—obvious (n=8), non-diseased—subtle (n=14), and non-diseased—obvious (n=12). Obvious was defined when there were clear and compelling findings indicating that endodontic disease was present or absent. Subtle was deemed to be difficult and mild findings that endodontic disease was present or absent. Moderate was defined as an intermediate disease category. Examples are in Fig 1. This broad spectrum of cases used de-identified patient data from private endodontic referral practices in Australia, Canada, and the USA. These cases were diagnostic in nature (not screening) with a presenting concern as determined by the patient and/or referring dentist. Some cases included abnormalities additional to a tooth with endodontic disease, for example, dental caries, endodontic disease, periodontal bone loss, and tooth resorption and fracture. In total, there were 34 additional abnormalities in 26 cases across all case categories.

The reference standard was established by a Delphi panel consisting of three endodontists with a total of 79 years' specialist clinical experience. The preoperative clinical and radiographic data and follow-up data of a minimum of 6-months duration were included for each member of the Delphi panel. Clinical history provided to the readers included the real presenting history and clinical findings taken at each appointment. This included a detailed pain history, clinical observations, and clinical test findings.

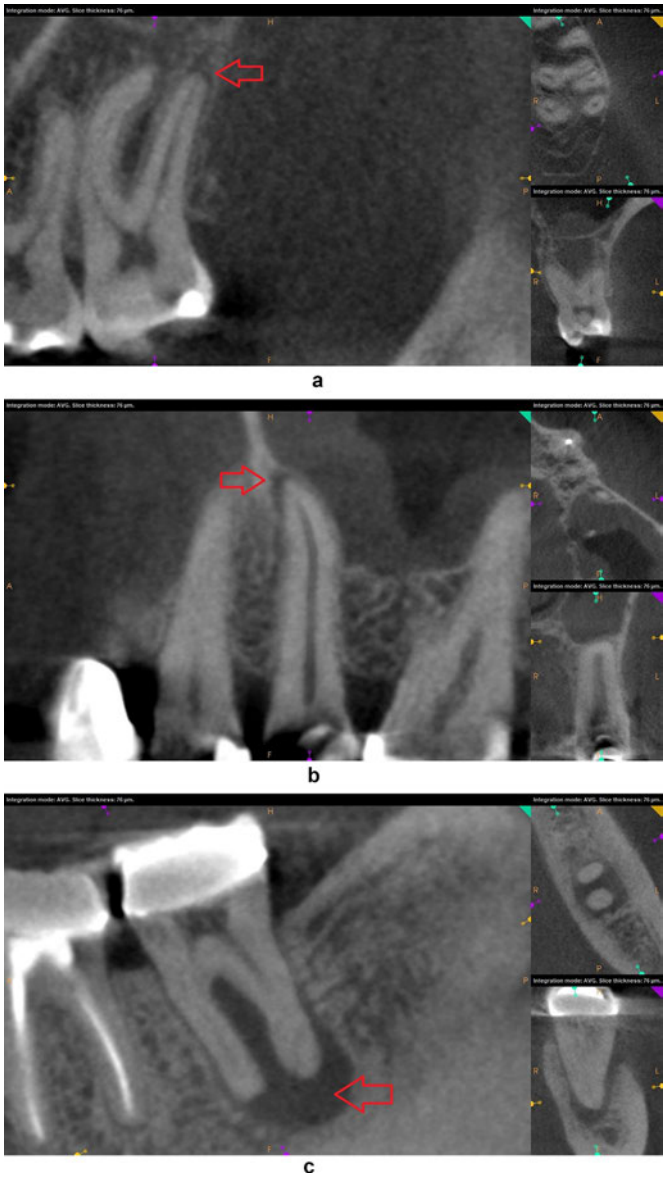


Figure 1 Case examples (red arrows) of (a) diseased subtle, (b) diseased moderate, and (c) diseased obvious.

Participants had no access to the reference standard. CBCT imaging was acquired using the CS9000 3D unit (Carestream, GA, USA) with a 38 × 50 mm volume and 76 μm voxel resolution.

Reader participants

Readers were recruited by direct invitation and from print and online advertisements targeted at dentists and radiologists with a wide range of experience working in Victoria, Australia. All participants received continuing education credits for participating.

Experimental setting and reading strategy

Readers were instructed to view the entire imaging volume, report the location and description of any abnormality, and provide a confidence rating from 1 to 4. Rating 1 was assigned if no abnormality was detected. Ratings of 2, 3, and 4 denoted lowest, middle, and highest levels of confidence for each abnormality. Readers were asked to mark as many abnormalities as they wished in each case. These mark–rating pairs are measurement characteristics of the free-response paradigm where readers are free to identify as many abnormalities as they wish.[18] Participants were informed that cases either had zero, or at least one, abnormality. No data on abnormality prevalence or expected abnormalities per case were provided.

A crossover design was used, with participants reading the entire test set of 60 cases twice over two reading sessions, once with the accompanying clinical history and once without. A balanced design was used for consistency in each session. For each of the five diseased and non-diseased case categories, half had the clinical history provided and the other half had no history for each session. Each reader had a randomised reading order, and each session was divided into two halves of 30 cases with a short break between.

To limit clustering and to ensure that all case types were distributed evenly throughout the reading sessions, each half of every session had an approximately equal number of diseased and non-diseased case types. The number of cases with history provided for each case type was divided into equal numbers per half. Throughout the sessions, there was an equal distribution of both case types (diseased and non-diseased) with and without history. Cases in the first reading session were labelled 1–60. At least 10 weeks elapsed between reading sessions to limit memory bias. The second reading session had the accompanying history reversed for all 60 cases. The same balance and consistency were used as in the first session; however, before randomisation, 15 cases in the first half were swapped with 15 cases in the second half to limit order bias. Cases in the second reading session were labelled 61–120 to avoid reader suspicion of repeat cases. Participants were informed that all 120 cases were different.

Reading environment

Images were displayed on a monitor with 4,096 × 2,304 resolution (Apple, CA, USA). Participants were shown how to use the viewing software (Carestream) to view the CBCT images, in particular window/level settings and navigation using curved, oblique, and orthogonal slicing. Ambient light measurements were recorded (lux light meter Android application). Screen recording software (Camtasia, TechSmith, MI, USA) was used for verification, in the event that a tooth was mislabelled, or another administrative issue needed clarification.

Data analysis

Data were analysed using the free-response paradigm.[11] When an abnormality determined by the Delphi panel was localised correctly, a lesion localisation (LL) was awarded. Every mark not corresponding to an abnormality was deemed a non-localisation (NL). Diagnostic performance metrics were true-positive marks or lesion localisations

(LL), false-positive marks or non-localisations (NL), specificity, and abnormality sensitivity or lesion localisation fraction (LLF: LLs divided by the number of abnormalities). The alternative free-response receiver operating characteristic (AFROC) curve was evaluated, a plot of the LLF versus the false-positive fraction (FPF: NLs divided by the number of non-diseased cases), as was the AFROC1 curve: LLF versus the FPF using all cases (NLs divided by the number of all diseased and non-diseased cases). The figure of merit used was the weighted AFROC1 statistic (wAFROC1), corresponding to the area under the wAFROC1 curve. This gives equal importance to all cases, regardless of the number of abnormalities per case.[19]

A power calculation based on a sample size of 60 cases and 20 readers and an effect size of a wAFROC figure of merit of 0.075 and $\alpha = 0.05$ demonstrated 91.56% power. The effect size of 0.075 was based on pilot data, from the mean wAFROC difference between three experienced endodontists not involved in the study. Data were grouped into three categories of clinical experience and the number of CBCT readings per month to assess the association between reader factors and performance. A Shapiro–Wilk test for sample population normality ($W = 0.63029$, $p<0.01$) showed the data were non-parametric. Therefore, a Kruskal–Wallis test was used to analyse subgroups. A Wilcoxon signed-rank test was used to compare mean performance with and without clinical history for the entire reader sample. The significance level was set at $p<0.05$. All analyses were conducted using RJafroc 2.0.1 in RStudio 4.0.3 (RStudio, PBC, MA, USA).

Results

Twenty-one dentists and two medical radiologists completed the test-set. No data were excluded from the participants. Mean clinical experience for all readers was 19 years (Table 1). Reading conditions had a mean ambient light of 34.1 lx (range 16–62 lx). A mean of 22 weeks (range 10–41 weeks) elapsed between reading sessions.

Non-diseased groupings

Tables 2 and 3 show reader performance in the non-diseased (subtle and obvious) categories. No differences were observed in reader performance with and without clinical history. Mean specificity was 0.46 without and 0.44 with history for the subtle non-diseased group ($p=0.78$), and 0.55 and 0.46, respectively, in the obvious non-diseased group ($p=0.07$). There were slightly more false-positive marks when clinical history was available compared to reading without clinical history (564 and 544, respectively; $p=0.64$). In both non-diseased categories, fewer non-diseased cases were identified correctly as non-diseased with clinical history than without (269 and 300, respectively). When the analyses were nested within readers' years of experience and CBCT reading volume, there was no significant trend in specificity and false-positive rates ($p>0.05$).

Table 1
Reader characteristics.

Clinical experience	Number	Mean	Range
0–9 years	4	3.8	0–8
10–19 years	13	14.5	12–17
20 or more years	6	38.8	23–48
CBCT studies read per month			
0–8	15	2.8	0–8
9–40	4	10	9–12
41 or more	4	66.3	60–80

CBCT, cone beam computed tomography.

Diseased groupings

Tables 4–6 show reader performance in the diseased (subtle, moderate, and obvious) groups. wAFROC1 values were significantly higher with clinical history in the diseased subtle ($p<0.01$) and obvious ($p=0.01$) groups than without history. Reader averaged wAFROC1 curves are in Fig 2. Lesion localisation was better with clinical history across all groups: subtle ($p<0.01$), moderate ($p<0.01$), and obvious ($p=0.03$). More false-positive marks occurred when history was available than without history (565 versus 547; $p=0.55$). Mean false-positive rates were not significantly different between readings made with and without clinical history ($p>0.05$). wAFROC1 values increased over the three categories as disease became more obvious. When the readings without clinical history or those with clinical history were considered independently, no interaction was found between reading volume and performance ($p>0.05$ for all); however, readers with low CBCT reading volume demonstrated better wAFROC1 values when clinical history was available ($p=0.002$) compared to when history was not provided. In the subtle disease group, readers with moderate reading volume demonstrated better lesion localisation when clinical history was available ($p=0.008$).

When the analyses were nested within readers' years of experience (Tables 4–6), and readings without clinical history and readings with clinical history were considered independently, no interaction was found between reader experience and performance ($p>0.05$ for all). Although false-positive calls decreased with more clinical experience in all diseased categories both with and without history, they were not significantly different. When readings with clinical history were compared to readings without history, no significant differences were observed in wAFROC1 ($p>0.05$ for all experience groupings); however, readers with fewer ($p=0.03$) and moderate ($p=0.04$) years of experience demonstrated better lesion localisation in subtle diseased cases when clinical history was available.

Discussion

The present findings show that clinical history improves the detection of endodontic disease in CBCT images without significantly negatively affecting the interpretation of non-diseased images. Access to clinical history decreased specificity for non-diseased images and increased false positives in all groups but was not significantly different to having no

Table 2
Non diseased subtle cases.

	Specificity			False positives		
	No history	History	p Value	No history	History	p Value
All readers	0.46 (0.39 0.54)	0.44 (0.37 0.51)	p=0.78	13.4 (10.2 16.6)	12.4 (10 14.8)	p=0.71
Years of experience						
0 9	0.43 (0.17 0.69)	0.41 (0.26 0.57)	p=0.56	16.8 (2.9 30.6)	13.3 (7 19.5)	p=0.89
10 19	0.47 (0.38 0.57)	0.43 (0.34 0.51)	p=0.64	13.5 (9.7 17.3)	13.1 (10 16.1)	p=0.89
20+	0.46 (0.36 0.57)	0.5 (0.32 0.67)	p=0.87	11.2 (7.9 14.5)	10.3 (5 15.7)	p=0.69
	p=0.99	p=0.76		p=0.94	p=0.58	
CBCT examinations read per month						
0 8	0.44 (0.35 0.54)	0.46 (0.35 0.56)	p=0.82	15 (10.4 19.6)	12.9 (9.3 16.5)	p=0.60
9 40	0.59 (0.43 0.74)	0.46 (0.37 0.56)	p=0.25	8.3 (5.6 10.9)	10.8 (8.7 12.8)	p=0.25
41+	0.41 (0.26 0.57)	0.38 (0.31 0.44)	p=0.67	12.8 (9.6 15.9)	12 (8.9 15)	p=0.67
	p=0.31	p=0.59		p=0.30	p=0.87	

Data are mean values, 95% confidence intervals in parentheses.
CBCT, cone beam computed tomography.

Table 3
Non diseased obvious cases.

	Specificity			False positives		
	No history	History	p Value	No history	History	p Value
All readers	0.55 (0.46 0.63)	0.46 (0.37 0.55)	p=0.07	10.2 (7.6 12.9)	12.1 (9.5 14.8)	p=0.18
Years of experience						
0 9	0.60 (0.33 0.88)	0.39 (0.03 0.76)	p=0.39	9.3 (1.9 16.6)	14.3 (5.5 23)	p=0.39
10 19	0.53 (0.41 0.65)	0.47 (0.38 0.57)	p=0.47	11 (7.1 14.9)	11.9 (8.5 15.4)	p=0.52
20+	0.56 (0.46 0.66)	0.46 (0.28 0.64)	p=0.47	9.2 (5.1 13.2)	11.2 (6.3 16)	p=0.81
	p=0.79	p=0.69		p=0.776	p=0.615	
CBCT examinations read per month						
0 8	0.49 (0.38 0.59)	0.46 (0.34 0.58)	p=0.69	11.9 (8.3 15.5)	12.6 (8.8 16.4)	p=0.82
9 40	0.73 (0.61 0.85)	0.42 (0.18 0.66)	p=0.08	4.8 (3.1 6.4)	11.3 (4.5 18)	p=0.11
41+	0.58 (0.47 0.70)	0.48 (0.34 0.62)	p=0.31	9.5 (5.4 13.6)	11.3 (9.8 12.7)	p=0.67
	p=0.12	p=0.93		p=0.09	p=0.79	

Data are mean values, 95% confidence intervals in parentheses.
CBCT, cone beam computed tomography.

Table 4
Diseased subtle cases.

	wAFROC1			Lesion localisations			False positives		
	No history	History	p Value	No history	History	p Value	No history	History	p Value
All readers	0.48 (0.43 0.53)	0.58 (0.52 0.63)	p<0.01[a]	8.4 (7.2 9.6)	12 (10.7 13.3)	p<0.01[a]	9.8 (7 12.6)	10.8 (7.8 13.7)	p=0.40
Years of experience									
0 9	0.38 (0.24 0.52)	0.50 (0.32 0.69)	p=0.15	9.8 (8.8 10.7)	12.8 (11.3 14.2)	p=0.03[a]	17.5 (9.1 25.9)	14.3 (4.4 24.1)	p=0.66
10 19	0.48 (0.41 0.56)	0.60 (0.52 0.67)	p=0.04[a]	7.7 (5.9 9.5)	11.9 (9.9 13.9)	p=0.008[a]	8.6 (5.2 12)	10.5 (6.6 14.4)	p=0.44
20+	0.54 (0.47 0.62)	0.59 (0.49 0.68)	p=0.52	9 (7.1 10.9)	11.7 (9.3 14.1)	p=0.07	7.2 (4.6 9.7)	9.2 (4.8 13.6)	p=0.58
	p=0.16	p=0.45		p=0.24	p=0.94		p=0.12	p=0.73	
CBCT examinations read per month									
0 8	0.47 (0.40 0.54)	0.58 (0.51 0.65)	p=0.06	8.1 (6.6 9.5)	12.1 (10.6 13.6)	p=0.002[a]	11.1 (7.3 15)	11.8 (7.6 16)	p=0.87
9 40	0.46 (0.37 0.55)	0.54 (0.44 0.64)	p=0.56	7.8 (5.2 10.3)	9.8 (6.3 13.2)	p=0.31	7.3 (1.7 12.8)	7.8 (3.6 11.9)	p=0.88
41+	0.55 (0.39 0.70)	0.61 (0.44 0.77)	p=0.56	10.3 (7.7 12.8)	14 (11.3 16.7)	p=0.15	7.3 (2.8 11.7)	10 (4.6 15.4)	p=0.56
	p=0.77	p=0.82		p=0.48	p=0.21		p=0.56	p=0.67	

Data are mean values, 95% confidence intervals in parentheses.
CBCT, cone beam computed tomography.
[a] Significant at p<0.05.

history. Many factors linked to the effect of clinical history on search, perception, and decision-making, and the diagnostic nature of our dataset may have contributed to these findings. Distinguishing information in the clinical history may direct the attention of the reader to a particular image region containing an abnormality, which is more readily identified.[20] When disease has subtle and ambiguous features, a clinical history slightly suggestive of disease may influence the reader's judgement of the presence of an abnormality that would not be identified without the history. Therefore, the description of the clinical scenario is enough to elevate the prior probability where an abnormality

Table 5
Diseased moderate cases.

	wAFROC1			Lesion localisations			False positives		
	No history	History	p Value	No history	History	p Value	No history	History	p Value
All readers	0.51 (0.44 0.57)	0.54 (0.48 0.60)	$p=0.29$	10.2 (8.7 11.8)	12.2 (10.8 13.6)	$p<0.01$[a]	8.6 (5.9 11.3)	8.7 (6.7 10.7)	$p=0.43$
Years of experience									
0 9	0.50 (0.39 0.61)	0.50 (0.37 0.63)	$p=0.77$	11.3 (8 14.5)	12.8 (11.8 13.7)	$p=0.66$	9.8 (3.5 16)	12.3 (5.8 18.7)	$p=0.39$
10 19	0.49 (0.39 0.59)	0.53 (0.46 0.61)	$p=0.59$	9.8 (7.6 12.1)	12 (9.9 14.1)	$p=0.18$	9.6 (5.4 13.8)	8.8 (6.3 11.4)	$p=0.89$
20+	0.54 (0.46 0.63)	0.59 (0.46 0.71)	$p=0.63$	10.3 (7.3 13.4)	12.2 (9.2 15.1)	$p=0.38$	5.7 (3.4 8)	6 (4.7 7.3)	$p=0.75$
	$p=0.80$	$p=0.61$		$p=0.86$	$p=0.97$		$p=0.48$	$p=0.25$	
CBCT examinations read per month									
0 8	0.49 (0.40 0.58)	0.52 (0.46 0.59)	$p=0.62$	9.5 (7.6 11.3)	11.9 (10.2 13.7)	$p=0.07$	9.7 (5.8 13.7)	9.8 (7.1 12.5)	$p=0.59$
9 40	0.48 (0.43 0.54)	0.47 (0.34 0.60)	$p=0.39$	9.3 (5.4 13.1)	10.3 (7.6 12.9)	$p=0.56$	5.8 (3.6 7.9)	7.8 (4.5 11)	$p=0.25$
41+	0.59 (0.49 0.68)	0.68 (0.53 0.82)	$p=0.39$	14 (11.3 16.7)	15 (12.6 17.4)	$p=0.56$	7.3 (5.1 9.4)	5.5 (4.5 6.5)	$p=0.31$
	$p=0.41$	$p=0.12$		$p=0.08$	$p=0.11$		$p=0.71$	$p=0.28$	

Data are mean values, 95% confidence intervals in parentheses.
CBCT, cone beam computed tomography.
[a] Significant at $p<0.05$.

Table 6
Diseased obvious cases.

	wAFROC1			Lesion localisations			False positives		
	No history	History	p Value	No history	History	p Value	No history	History	p Value
All readers	0.71 (0.67 0.75)	0.77 (0.73 0.80)	$p=0.01$[a]	9.3 (8.7 10)	10 (9.7 10.5)	$p=0.03$[a]	5.4 (3.9 6.9)	5.1 (3.4 6.8)	$p=0.48$
Years of experience									
0 9	0.68 (0.61 0.75)	0.77 (0.70 0.83)	$p=0.08$	9 (7 11)	9.8 (8.5 11)	$p=0.56$	7 (4.1 9.9)	6.8 (1.9 11.6)	$p=1$
10 19	0.71 (0.64 0.77)	0.75 (0.70 0.80)	$p=0.34$	9.2 (8.4 9.9)	10.1 (9.6 10.5)	$p=0.09$	6.2 (4 8.3)	5.8 (3.4 8.2)	$p=0.82$
20+	0.74 (0.66 0.82)	0.81 (0.73 0.89)	$p=0.17$	10 (8.7 11.3)	10.3 (9.4 11.3)	$p=0.75$	2.7 (1.2 4.2)	2.3 (0.6 4)	$p=0.58$
	$p=0.76$	$p=0.41$		$p=0.47$	$p=0.79$		$p=0.07$	$p=0.15$	
CBCT examinations read per month									
0 8	0.69 (0.64 0.74)	0.76 (0.71 0.81)	$p=0.06$	9.1 (8.3 9.8)	10.1 (9.6 10.6)	$p=0.05$	6.5 (4.6 8.5)	6.1 (3.7 8.6)	$p=0.71$
9 40	0.71 (0.58 0.83)	0.76 (0.69 0.84)	$p=0.47$	8.8 (7.3 10.2)	9.8 (8.8 10.7)	$p=0.39$	3.5 (1.6 5.4)	3.5 (2.2 4.8)	$p=0.88$
41+	0.78 (0.69 0.87)	0.78 (0.68 0.89)	$p=0.66$	11 (10.2 11.8)	10.5 (9.2 11.8)	$p=0.56$	3 (0.9 5.1)	2.8 (1 4.4)	$p=0.88$
	$p=0.29$	$p=0.95$		$p=0.04$[a]	$p=0.63$		$p=0.16$	$p=0.47$	

Data are mean values, 95% confidence intervals in parentheses.
CBCT, cone beam computed tomography.
[a] Significant at $p<0.05$.

is deemed present, regardless of the image features.[21] Conversely, when a non-diseased case shows ambiguous image perturbations, a clinical history slightly suggestive of disease may bias the decision of the reader and lead to the reporting of an abnormality that is not actually present.[22] The elevated prior probability for these diagnostic images may have been responsible for the slightly higher false-positive marks.

This effect of clinical history has not been reported previously in the dental literature because previous studies assessed reader performance at case level using the receiver operating characteristic (ROC) paradigm, which limits performance measurement at abnormality level.[10] To the authors' knowledge, the present study is the first CBCT perception study to separate diseased and non-diseased cases, grade their severity, and perform a factorial study design in the free-response paradigm following the STARD guidelines. The methodology also overcomes previous study design limitations by analysing non-diseased cases separately and enabling performance measurement of the effect of clinical history at the abnormality level. This factorial study design is valuable in isolating factors that may influence diagnostic performance.

When diseased cases were interpreted with clinical history, performance improved, with higher wAFROC1 values and more LLs compared to without history mainly in the subtle disease group. Performance with and without history improved with increased severity. This observation may be due to more salient features of disease that capture attention and aid in disease identification as severity increased. This is consistent with earlier findings,[23] which follows a logical explanation that more obvious abnormal features are easier to detect and correctly identify as an abnormality. The present findings suggest that although clinicians have no knowledge of the disease status prior to image interpretation, clinical history can increase their prior probability before reviewing the imaging features of disease. This increased probability may have contributed to better performance in diseased images[24] and higher false-positive errors in both disease and non-diseased cases. The present study was able to assess the effect of disease and non-disease severity on performance, particularly in more difficult cases with diagnostic uncertainty, as was recommended previously.[25]

To establish if the present findings were influenced by intrinsic reader factors, the impact of years of experience

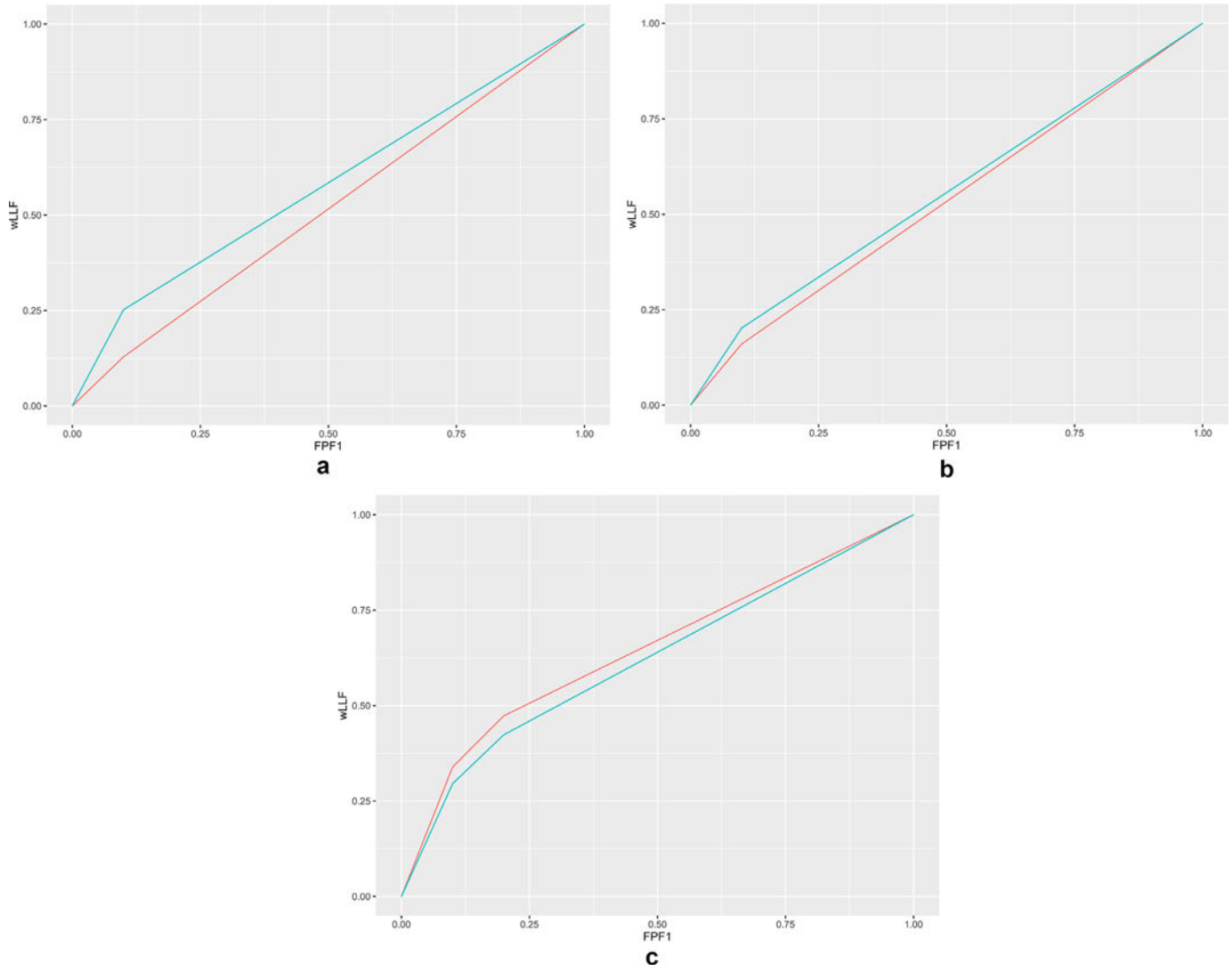


Figure 2 Reader averaged wAFROC1 curves for diseased cases. (a) Diseased subtle cases, (b) diseased moderate cases, and (c) diseased obvious cases. Red, no history; green, history.

and monthly CBCT images read on reader performance was examined. Clinical experience and reading volume did not have a clear relationship with performance in non-diseased cases, with and without history. These findings may be due to the recent introduction of CBCT into dentistry and scarcity of specific CBCT training programmes. There were readers with decades of clinical experience who had very little experience in reading CBCT images. When performance with and without clinical history were compared across different levels of clinical experience, readers with fewer (0–9) and moderate (10–19) years of experience benefitted from having clinical information when reading cases with subtle features of disease. This is demonstrated by higher wAFROC1 values in readers with moderate experience and more mean LLs in those with fewer and moderate years of experience when history was available. These findings suggest that clinical history is relevant to less and moderately experienced practitioners for accurate detection and interpretation of subtle abnormalities. The similarity in wAFROC1 values, LLs, and false-positive rates between readers with different clinical experience across moderate and obvious diseased categories support the decreased utility of history as imaging features of disease become more obvious. False-positive calls in all diseased categories except one group (diseased moderate without history) decreased with increased years of experience and monthly CBCT reading volume, although not significantly greater than chance.

The present study is the first to study the effect of reader factors on the interpretation of CBCT with and without history in endodontic disease detection. A recent review found no dental imaging studies examining the effect of history on diagnostic imaging performance or the influence of clinical experience or reading volume on diagnostic performance.[9] In other imaging domains such as mammography, radiologists with more experience outperformed less experienced residents,[26] made fewer false-positive errors,[27] performed better in localising lesions and simultaneously

rating their level of malignancy,[28] and demonstrate higher sensitivity.[29] Annual reading volume was also shown to improve diagnostic performance,[28] location sensitivity, and ROC area under the curve values[29]; however, it is difficult to compare findings between different study types. First, the mammography studies did not compare reader performance with and without history. Secondly, they did not consider diseased and non-diseased images independently. Lastly, mammography studies mostly used screening datasets and analysed all cases together without assessing images with subtle (difficult cases) features separately from prominent (less difficult) imaging features. The present study was based on a diagnostic dataset and considered the imaging presentations of abnormalities. Thus, the present results highlight the need for further evaluation of the impact of reader factors in CBCT interpretation.

This study has some limitations. First, readers interpreted all images knowing it was an experiment and not in a clinical setting, which introduces demand characteristics,[30] an experimental artefact due to altered behaviour to suit the perceived goals of the study. An early study used cases in an clinical setting for a radiologist[31]; this method would be impractical because dentists order and interpret radiographic images before any patient intervention. It would be unlikely for a dentist to be a radiologist without any patient interaction, except for a specialist oral and maxillofacial radiologist; however, previous mammography-based studies have reported that reader performance in real-life clinical settings correlates positively within a laboratory setting.[32] Secondly, it can be argued that reading the same images twice may introduce memory bias; however, a chest radiograph study showed that memory erodes to chance level within 7 weeks,[33] with another study showing the memory effect to be weak after 1–3 days, and "more striking" abnormalities being recognised more accurately.[34] The required duration for reducing memory bias has not been studied in dental CBCT imaging. Recollection of the same case in the present study is unlikely (the mean interval was 22 weeks) and if so, it would not necessarily affect diagnostic performance without access to the reference standard. The real clinical history collected by skilled clinicians was used, which may be of greater diagnostic value than a vague history missing relevant details and increase the study's external validity. A qualitative analysis comparing relevant and general clinical history may be difficult to perform; however, it may be why improved diagnostic performance in diseased cases with history was observed.

In conclusion, clinical history biases the interpretation and the diagnostic performance of CBCT imaging in endodontic disease detection. Although the availability of clinical history had no significant impact on the interpretation of CBCT images without disease, it improved diagnostic performance in abnormal images. Inconsistent results on the effect of clinical history on interpretation across difference disease types warrant further studies to provide more evidence to support guidelines around taking a patient's history prior to the interpretation of dental CBCT images.

Declaration of competing interest

The authors declare no conflict of interest.

References

1. Samei E, Krupinski E. *The handbook of medical image perception and techniques*. Cambridge: Cambridge University Press; 2019.
2. Lee CS, Nagy PG, Weaver SJ, *et al*. Cognitive and system factors contributing to diagnostic errors in radiology. *AJR Am J Roentgenol* 2013;**201**:611 7.
3. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. *RadioGraphics* 2018;**38**:236 47.
4. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information a systematic review. *JAMA* 2004;**292**:1602 9.
5. Mitchell D, Mitchell L, McCaul L. *Oxford handbook of clinical dentistry*. 6th edn. Oxford: Oxford University Press; 2014.
6. Stefanac S, Nesbit S. *Diagnosis and treatment planning in dentistry*. 3rd edn. St Louis: Elsevier; 2016.
7. American Academy of Pediatric Dentistry. Prescribing dental radio graphs for infants, children, adolescents, and individuals with special health care needs. In: *The reference manual of pediatric dentistry*. Chicago: American Academy of Pediatric Dentistry; 2021. p. 273 6.
8. Croskerry P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Acad Emerg Med* 2002;**9**:1184 204.
9. Yapp KE, Brennan P, Ekpo E. The effect of clinical history on diagnostic im aging interpretation a systematic review. *Acad Radiol* 2022;**29**:255 66.
10. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys* 2004;**31**:2313 30.
11. Bunch P, Hamilton J, Sanderson G, *et al*. A free response approach to the measurement and characterization of radiographic observer perfor mance. In: *Proc. SPIE 0127, application of optical instrumentation in medi cine VI*. Bellingham, WA: SPIE; 1977, https://doi.org/10.1117/12.955926.
12. Chakraborty D. Recent developments in FROC methodology. In: Samei E, Krupinski E, editors. *The handbook of medical image perception and techniques*. Cambridge: Cambridge University Press; 2010. p. 216 39.
13. Brown LF, Monsour P. The growth of Medicare rebatable cone beam computed tomography and panoramic radiography in Australia. *Aust Dent J* 2015;**60**:511 9.
14. Alamri HM, Sadrameli M, Alshalhoob MA, *et al*. Applications of CBCT in dental practice: a review of the literature. *Gen Dent* 2012;**60**:390 400. ; quiz 401 2.
15. European Commission. Energy Protection Radiation No. 172. Cone beam CT for dental and maxillofacial radiology (evidence based guidelines). Available at: https://www.sedentexct.eu/files/radiation_protection_172.pdf. [Accessed 21 November 2022].
16. Special Committee to Revise the Joint AAE/AAOMR Position Statement on use of CBCT in Endodontics. AAE and AAOMR joint position state ment: use of cone beam computed tomography in endodontics 2015 update. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2015;**120**:508 12.
17. Cohen JF, Korevaar DA, Altman DG, *et al*. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;**6**:e012799. e012799.
18. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free response paradigm. *Phys Med Biol* 2006;**51**:3449 62.
19. Chakraborty DP, Zhai XT. On the meaning of the weighted alternative free response operating characteristic figure of merit. *Med Phys* 2016;**43**:2548.
20. Nguyen A. *The effect of a localizing prompt on cone beam computed to mography interpretation*. St. Louis, MO, USA: Saint Louis University; 2017.
21. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980;**302**:1109 17.
22. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926 30.
23. Khademi J. *The effects of digitization on diagnostic performance in end odontics and periodontics: an ROC study*. Iowa City, IA, USA: The Uni versity of Iowa; 1994.
24. Schechter MT, Sheps SB. Diagnostic testing revisited: pathways through uncertainty. *Can Med Assoc J* 1985;**132**:755 60.

25. Scott IA, Greenberg PB, Poole PJ. Cautionary tales in the clinical interpretation of studies of diagnostic tests. *Intern Med J* 2008;**38**:120 9.
26. Nodine CF, Kundel HL, Mello Thoms C, *et al*. How experience and training influence mammography expertise. *Acad Radiol* 1999;**6**:575 85.
27. Elmore JG, Jackson SL, Abraham L, *et al*. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology* 2009;**253**:641 51.
28. Rawashdeh MA, Lee WB, Bourne RM, *et al*. Markers of good performance in mammography depend on number of annual readings. *Radiology* 2013;**269**:61 7.
29. Suleiman WI, Lewis SJ, Georgian Smith D, *et al*. Number of mammography cases read per year is a strong predictor of sensitivity. *J Med Imag (Bellingham)* 2014;**1**:015503.
30. Orne MT. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *Am Psychol* 1962;**17**:776 83.
31. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. *AJR Am J Roentgenol* 1981;**137**:1055 8.
32. Soh BP, Lee WB, McEntee MF, *et al*. Mammography test sets: reading location and prior images do not affect group performance. *Clin Radiol* 2014;**69**:397 402.
33. Evans KK, Marom EM, Godoy MC, *et al*. Radiologists remember mountains better than radiographs, or do they? *J Med Imag (Bellingham)* 2016;**3**:011005.
34. Ryan JT, Haygood TM, Yamal JM, *et al*. The "memory effect" for repeated radiologic observations. *AJR Am J Roentgenol* 2011;**197**:W985 91.

CHAPTER 6

6.1 Bridging section to Chapter 6

When diagnostic images are interpreted, there may be abnormalities additional to the primary search task. Detection of these incidental findings is important because any disease entity that remains unknown and unmanaged can progress and worsen outcomes for patients. Undocumented incidental findings have been described in many settings. A study on abdomen and pelvic computed tomography (CT) scans showed that 62% of patients had undocumented incidental imaging findings (Devine et al. 2010). One third of emergency patients with CT scans had incidental findings, but only 9.8% were reported to patients (Thompson et al. 2011). A recent CT study on low-energy falls in elderly patients had incidental findings in 74% of patients (Niedermeier et al. 2022). The evidence shows significant frequency of incidental imaging findings, which if undiagnosed can lead to an increase in disease severity and negatively impact upon patient management and outcomes.

Classification of these missed (false negative) findings includes search, recognition and decision errors (Kundel et al. 1978), as well as "satisfaction of search" errors (Berbaum et al. 1990) or "subsequent search misses" (Adamo et al. 2013). These occur when the abnormality is not fixated (search), is briefly fixated without the observer realising (recognition) or when the abnormality is fixated, but the observer decides that no abnormality is present, or the stimuli is not strong enough to be reported (decision). The satisfaction of search error occurs when after detection of an abnormality, an additional abnormality remains undetected (Berbaum et al. 1990) when the reader has satisfied their "quest for meaning" (Tuddenham 1962). Subsequent search misses are thought to occur due to perceptual set (observers, upon finding a target, are biased to search for similar targets and more likely to miss dissimilar targets) and resource depletion (after expending attention and working memory resources to detect a first target, fewer resources are available to detect an additional target). Another reason is attentional template (where detecting a first target creates an attentional template, incurs a search cost, reduces the cognitive resources required to recognise a new target, thereby increasing the likelihood of missing a subsequent search target) (Adamo et al. 2021).

There are several studies on missed incidental findings in medical radiology (Devine et al. 2010, Thompson et al. 2011, Niedermeier et al. 2022), but limited data is available on the detection of incidental abnormalities on dental cone beam computed tomography (CBCT). Furthermore, the effect of clinical history on detection of these additional abnormalities has not been widely studied. When CBCT cases with two abnormalities were analysed both with and without history, 80% of findings were reported with history and 98% were reported without history (Hiserote 2015). A similar parallel group study, also with CBCT cases with only two abnormalities, found that a localising prompt increased abnormality detection (Nguyen 2017). These methodological limitations of sample power, 100% disease prevalence, sham history and a non-replicative clinical reading task emphasise the need

for well-powered and quality studies to provide a detailed understanding of the effect of clinical history on the detection of incidental abnormalities.

Chapter six is a study evaluating the effect of clinical history on incidental abnormality detection using a factorial, free-response study design. It also analyses the effect of history on reader confidence in incidental abnormality detection. The methodology used in the study presented in chapter five accounts for changes in the independent variables: case type (diseased or non-diseased), case severity (subtle, moderate or obvious), reader type (low or high level of experience) and reading modality (CBCT with or without clinical history). Data analysis used the metrics of lesion localisation fraction, the incidental abnormalities correctly localised, and reader confidence ratings (rating 2, 3 or 4) of both reading modalities. This study provides data on the effect of clinical history on incidental abnormality detection and reader confidence in endodontic CBCT imaging. The results show that clinical history improved detection of incidental abnormalities in subtle non-diseased cases. When history was available, reader confidence was highest, but not associated with an overall improvement in diagnostic performance.

This paper has been submitted for publication as:

Yapp KE, Ekpo E. Clinical history and incidental abnormality detection in endodontic cone beam computed tomography (Journal of Medical Imaging submitted 2023).

References

Adamo, S. H., et al. (2013). "Self-induced attentional blink: a cause of errors in multiple-target search." Psychol Sci **24**(12): 2569-2574.
Adamo, S. H., et al. (2021). "From "satisfaction of search" to "subsequent search misses": a review of multiple-target search errors across radiology and cognitive science." Cogn Res Princ Implic **6**(1): 59.
Berbaum, K. S., et al. (1990). "Satisfaction of Search in Diagnostic Radiology." Investigative Radiology **25**(2): 133-140.
Devine, A. S., et al. (2010). "Frequency of incidental findings on computed tomography of trauma patients." West J Emerg Med **11**(1): 24-27.
Hiserote, D. (2015). The Effect of Patient Dental History on Cone Beam Computed Tomography Interpretation of Periapical Findings in Endodontics, Saint Louis University.
Kundel, H. L., et al. (1978). "Visual scanning, pattern recognition and decision-making in pulmonary nodule detection." Invest Radiol **13**(3): 175-181.
Nguyen, A. (2017). The effect of a localizing prompt on cone beam computed tomography interpretation, Saint Louis University.
Niedermeier, S., et al. (2022). "Incidental CT Findings in the Elderly with Low-Energy Falls: Prevalence and Implications." Diagnostics (Basel, Switzerland) **12**(2): 354.
Thompson, R. J., et al. (2011). "Incidental Findings on CT Scans in the Emergency Department." Emergency medicine international **2011**: 624847-624847.
Tuddenham, W. J. (1962). "Visual search, image organization, and reader error in roentgen diagnosis. Studies of the psycho-physiology of roentgen image perception." Radiology **78**: 694-704.

Title: Clinical history and incidental abnormality detection in endodontic cone beam computed tomography

Authors

Kehn E Yapp, BDSc DCD (Endo)*

Ernest Ekpo, PhD*

*Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging Science, School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Australia.

Corresponding author: Kehn E Yapp, Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging Science, School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, NSW, Australia.

Email: keyapptdo@gmail.com

Disclosures: none

Acknowledgements: The authors deny any conflicts of interest.

ABSTRACT

Introduction: To assess the effect of clinical history on incidental abnormality detection, false positive marks and diagnostic confidence in endodontic cone beam computed tomography (CBCT) imaging.

Methods: A reader performance study using a free-response, factorial study design was undertaken, which accounted for changes in the independent variables: native case type, native case severity, reader type and reading modality. Twenty-three readers interpreted 26 cases (18 diseased, 8 non-diseased) twice, once with and once without access to clinical history. Each case had at least one incidental abnormality that was not a native endodontic finding. Lesion localisation, non-localisations (false positives) and diagnostic confidence (rating 2, 3 or 4 – lowest, middle, and highest respectively) of incidental abnormalities were analysed.

Results: Clinical history increased the detection of incidental abnormalities in non-diseased subtle cases (76 vs 59, p=0.04). Reader experience and monthly CBCT reading volume did not affect incidental abnormality detection. False positives were not affected by clinical history nor reader characteristics. The highest confidence rating was most often used in each case type when clinical history was available. For this rating, history had significantly greater lesion localisations in subtle diseased (53 vs 41, p=0.03) and non-diseased images (53 vs 33, p=0.02).

Conclusions: Clinical history improved the detection of incidental endodontic abnormalities in non-diseased subtle CBCT images and did not affect the number of false positive marks. Reader confidence in correctly identified abnormalities was higher with clinical history when disease and non-disease was subtle but was not associated with an improvement in diagnostic performance.

Keywords: history, incidental, endodontic, cone beam computed tomography

Clinical history and incidental abnormality detection in endodontic cone beam computed tomography

Introduction

Detection of incidental abnormalities, additional to the primary image search task and interpretation, can be clinically relevant. Several reports have highlighted the nature of undocumented incidental findings. In trauma patients undergoing abdomen and pelvic computed tomography (CT) scans, 62% of patients had incidental imaging findings with no documentation in the electronic medical record, with the presumption that many of these patients were not informed of these findings[1]. Emergency patients receiving CT scans were reported to have incidental findings in one third of cases but only 9.8% were reported to patients in discharge paperwork[2]. Another study on elderly patients with low-energy falls showed that in CT examinations, 74% of patients had incidental findings[3]. The significance of these missed incidental findings is that, if a disease entity remains undetected and unmanaged, disease can progress, and patient outcomes can worsen.

Factors related to missed radiological findings have been described as cognitive and system errors[4]. Some cognitive biases related to image interpretation error include: anchoring bias – failure to adjust an initial impression, despite receiving additional information; confirmation bias – searching for data to reaffirm an existing hypothesis; framing bias – drawing different conclusions from the same information depending on how the information is presented; satisfaction of search – decreasing vigilance and/or awareness for additional abnormalities after differentiating the first abnormality; and inattentional bias – missing findings hiding in plain sight due to unexpected location or nature[5]. System factors relating to diagnostic error focus on visual and mental fatigue[4]. The reasons for missed findings are likely multifactorial, however errors in perception have also been identified as a relevant contributor.

Perceptual false negative errors have been classified into search, recognition and decision error[6]. Search errors occur when the abnormality is not fixated by the observer. Recognition errors occur when the abnormality is briefly fixated without being distinguished from the background parenchyma. Decision errors occur when the observer fixates the abnormality and decides no

abnormality is present or it is benign and chooses not to report it. Another perceptual error, satisfaction of search[7] was later called the "subsequent search miss" effect[8]. Perception has been described to be dependent on an innate "search for meaning", with the suggestion that once the "quest for meaning" is "satisfied" with a positive finding, the observer is likely to stop the imaging search task "whether or not the entire film has been scrutinized"[9]. Other explanations for this observation include perceptual set (observer bias to search for similar targets and miss dissimilar targets) and resource depletion (fewer available resources to detect additional targets after expending attention and working memory resources to detect a first target)[10]. These factors that affect identification of primary imaging findings may also affect detection of incidental findings. The relationship between misses or detections of primary abnormalities and incidental findings, and the mechanism linking the two, is not well understood.

Whilst there are studies on missed incidental findings in medical radiology, little is known regarding dental imaging, in particular cone beam computed tomography (CBCT). One CBCT pilot study using readers that read each case twice – with and without a fabricated patient history – and cases with two periapical radiolucencies where one was more obvious, found that 80% of findings were reported with history and 98% were reported without history[11]. Another study using two parallel reader groups and two endodontic CBCT cases, each with two abnormalities and a contrived referral letter with a localising prompt referring to one of the two teeth with abnormalities, found that the localising prompt led to 100% detection of the prompted abnormality and reduced detection of the unprompted abnormality by half[12]. The localising prompt per case was alternated between groups and this finding was observed in all cases. The suggestion from this study was that a localising prompt reduced overall abnormality detection likely by inducing visual neglect. These pilot studies have limitations, including small sample size, 100% disease prevalence with no data on non-diseased images, sham history and referral letter details, and a reading task not replicative of the clinical task. Further studies are therefore required to overcome these limitations and to provide adequate data on the effect of biases such as clinical history on incidental findings in dental CBCT imaging. Reader factors, such as years of experience and reading volume have been shown to affect diagnostic

performance in medical image perception, but little is known regarding the effect on performance regarding incidental findings in dental CBCT imaging.

Incidental abnormalities often detected in endodontic images include dental caries, periodontal bone loss, tooth resorption and fracture. The detection and management of these abnormalities is relevant because undetected caries and other dental diseases will inevitably lead to worse health outcomes for patients. Therefore, this study aims to assess the effect of clinical history and reader factors on incidental abnormality detection and false positive marks in endodontic CBCT imaging by using a free-response, factorial study design. It also aims to analyse the effect of history on reader confidence in incidental abnormality detection.

Materials and methods

Study design: This experiment followed the Standards for Reporting of Diagnostic Accuracy Studies (STARD)[13] guidelines and a factorial, free-response, crossover study design. Changes in the independent variables were analysed: case type (diseased and non-diseased), case severity (subtle, moderate and obvious), reader type (low, medium and high level of clinical experience) and reading modality (with and without access to clinical history prior to image interpretation). Institutional review board approval at The University of Sydney was granted (Approval number: 2020/477).

Image dataset and grouping: The 26 cases included were 18 diseased and 8 non-diseased, part of a larger study[14]. There were 20 additional abnormalities in the diseased group and 14 in the non-diseased group, with a total of 52 abnormalities – 18 native and 34 additional. The case types based on the native abnormality were divided into five categories: diseased – subtle (5), diseased – moderate (8), diseased – obvious (5), non-diseased – subtle (3) and non-diseased – obvious (5) (Figure). Obvious was defined as a case with clear and compelling findings indicating that endodontic disease was present or absent. Subtle was deemed to be difficult and mild findings that endodontic disease was present or absent. Moderate was defined as an intermediate disease category. The moderate diseased category was added because it was deemed a clinically relevant group, with a moderate non-diseased group excluded because it was deemed to not be clinically relevant. These cases were

diagnostic in nature – not screening – from private endodontic referral practices in Australia, Canada and the United States of America and had a presenting concern as determined by the patient and/or referring dentist. The primary abnormality was a tooth with endodontic disease and the incidental abnormalities included dental caries, periodontal bone loss, tooth resorption and fracture.

The reference standard was established by a Delphi panel consisting of three endodontists with collectively 79 years of specialist clinical experience. The pre-operative clinical and radiographic data and follow up data of a minimum six months' duration were included for each case. Clinical history included the real presenting history and clinical findings taken at each appointment, related to the reason for presentation. Abnormalities were classified descriptively (eg. dental caries on tooth 16 mesial), rather than using a digitally assigned location. Cone beam computed tomography imaging was acquired using the CS9000 3D unit (Carestream, Georgia, USA) which produced a 38 x 50mm field of view volume with 76µ voxel resolution.

Participants: Participants were recruited from print and online advertisements targeted at dentists in Victoria, Australia as well as being contacted directly to participate. A broad range of dentists with varying amounts of experience were encouraged to participate. Medical radiologists with some experience in reporting cone beam computed tomography were directly invited to participate. No data were collected on the participants who declined or did not respond to an invitation. All participants received continuing education credits for participating and no monetary incentives were offered.

Experimental setting and reading strategy: Readers were instructed to view the entire imaging volume and report the location of any abnormality. They were instructed to provide a brief description for every abnormality with a confidence rating of abnormality presence from 1 to 4. A rating of 1 was assigned if no abnormality was detected. Ratings of 2, 3 and 4 denoted lowest, middle, and highest levels of confidence in the presence of each assigned abnormality respectively. Readers were asked to mark as many abnormalities as they wish for each case. These mark-rating pairs are measurement characteristics of the free-response paradigm[15]. Participants were informed that some cases had no

abnormalities and others had at least one abnormality. No information was provided about the abnormality prevalence in the test set nor expected number of abnormalities per case.

A crossover design was used, where every participant read the image data set twice over two reading sessions – once with the accompanying clinical history and once without. A balanced design was used for consistency in each session, having an even number of each of the case type categories with and without clinical history provided for each reading session. To limit clustering and to ensure that all case types were evenly distributed throughout the reading sessions, each half of every session had an approximately equal number of diseased and non-diseased case types. Cases with history provided for each case category had equal numbers per half of the reading session. Throughout the sessions, there was an equal distribution of both case types (diseased and non-diseased) with and without history. At least 10 weeks elapsed between reading sessions to limit memory bias. The second reading session had the accompanying history reversed for all cases. The same balance and consistency were used as in the first session. Participants were not informed about the double reading nature of the study and were told that all cases amongst reading modalities were different.

Reading environment: Images were displayed on a monitor with 4096x2304 resolution (Apple, California, USA). Participants were shown how to use the viewing software (Carestream, Georgia, USA) to view the CBCT images, in particular window/level settings and navigation using curved, oblique and orthogonal slicing. Ambient light measurements were recorded (Lux light meter free smartphone application). Screen recording software (Camtasia, TechSmith, Michigan, USA) was used for verification purposes, in the event that a tooth was mislabelled or another administrative issue had to be clarified.

Data Analysis: Data were analysed using the free-response paradigm[16]. When an abnormality, as determined by the Delphi panel, was correctly localised, a lesion localisation (LL) was awarded. Every mark that did not correspond to an abnormality was deemed a non-lesion localisation (NL) or false positive (FP). Diagnostic performance metrics were true positive marks or lesion localisations (LL). Reader confidence in the classification of abnormalities was also examined by comparing their

confidence ratings (rating 2, 3 or 4) between reading modalities. To analyse subgroups, a Kruskal-Wallis test was used and a Wilcoxon signed-rank test compared the mean diagnostic performance for all readers. Significance was deemed to be $p<0.05$.

Results

All 23 readers provided usable datasets. No data were excluded from the participants. Reader characteristics are shown in Table 1. In non-diseased subtle cases, images read with clinical history had a greater number of identified incidental abnormalities, a statistically significant ($p=0.04$) finding for all readers. The non-diseased obvious cases had a similar trend, but this was not significant. In the diseased group, as disease features became more obvious, clinical history was associated with an opposite finding – fewer reported incidental abnormalities in moderate and obvious cases. Subtle diseased cases had more incidental findings with history, however this finding did not reach statistical significance. Clinical history did not affect the detection of incidental abnormalities when reader groups were analysed by years of experience and monthly CBCT reading volume. Details of the detected incidental abnormalities by case type are shown in Tables 2 and 3. False positive marks were not affected by clinical history in both diseased and non-diseased case types, for all readers and when years of experience and reading volume were analysed (Tables 4 and 5).

When analysing confidence for all readers, for the lowest and medium confidence ratings (2 and 3), incidental abnormality LLs were not significantly different between reading with and without clinical history (Table 6). The reverse was found with the highest confidence rating (4), where readings with history demonstrated a significantly greater number of incidental abnormality LLs than cases read without history ($p=0.02$). When the numbers of LLs in each rating category were grouped by reader experience and number of monthly CBCTs read, no significant differences were found in incidental abnormality detection with and without clinical history. When reader confidence in rating abnormalities was grouped by case type, readers demonstrated higher confidence in rating (rating 4) incidental abnormalities when clinical history was provided in only the non-diseased subtle and diseased subtle categories (Table 7). No significant differences were noted in any other case or rating group.

Discussion

Our findings show that history improves the detection of incidental abnormalities in non-diseased images with subtle features but is associated with a decreasing trend in the detection of incidental abnormalities as native disease becomes more obvious. However, this decrease in detection of incidental abnormalities was not significantly greater than zero.

False negative perceptual errors have been described as scanning, recognition and decision errors[6]. It is unclear which type of false negative errors occurred in these more obvious diseased cases, however there may be a relationship between history improving diagnostic performance in more obvious diseased cases – as shown in the results of the larger study[14] – and reducing incidental abnormality detection at the same time. One reason for this observation could be the satisfaction of search effect, later renamed the "subsequent search miss" effect[8]. Theories regarding why they occur include satisfaction (the reader has satisfied a "quest for meaning" once the first target is localised), perceptual set (observers, upon finding a target, are biased to search for similar targets and more likely to miss dissimilar targets) and resource depletion (after expending attention and working memory resources to detect a first target, fewer resources are available to detect an additional target)[10]. These authors also proposed a new "Attentional Template" theory, where detecting a first target creates an attentional template, incurs a search cost, reduces the cognitive resources required to recognise a new target, thereby increasing the likelihood of missing a subsequent search target[10]. It is likely that a combination of the above features is contributing to the reduced detection of incidental abnormalities in more obvious diseased cases read with history.

Decision errors relating to missed abnormalities occur when image perturbations are fixated for longer periods, but not reported as abnormal. An eye tracking study on lung cancer detection on chest radiographs found that of all false negative calls, 65% were fixated for over one second, with the mean dwell time being 3.1 seconds[17]. Although our study did not use eye tracking, future studies are required to analyse fixation and detection of primary and incidental abnormalities in CBCT imaging.

The increased detection of incidental abnormalities when history was available in non-disease subtle cases may be due to the history being suggestive of endodontic disease, the case being non-suggestive of endodontic disease, which then led the reader to be more vigilant in detecting abnormalities.

Strategies to mitigate errors in image interpretation have been described[18]. These include peer review, departmental meetings ("quality improvement"), structured reporting and checklists and perceptual training. Whilst training strategies can help in native abnormality detection, further studies are required to determine if these strategies can help mitigate errors in the detection of incidental abnormalities. This is because unreported incidental findings could lead to disease entities progressing and providing worse outcomes to patients.

False positive errors were not affected by clinical history, consistent with findings from a larger study[14]. This suggests that in cases with incidental abnormalities, access to clinical history does not impact the number of non-localisations, in cases with and without endodontic disease.

An interesting finding was that the highest confidence rating was often used when clinical history was available, particularly in the diseased subtle and non-diseased subtle categories. These findings suggest that in cases with greater difficulty in interpretation, history is more likely to increase a reader's confidence in abnormality detection. It should be noted that of all ratings in each subcategory, most of the correctly localised abnormalities for both history and no history had the highest confidence rating (4) in all but one group – 0-9 years of experience. This may reflect the reader cohort who was more likely to make correct high confidence calls, or the types of native abnormalities – where more subtle findings are more likely to have lower confidence calls. However, higher reader confidence with history was not associated with increased incidental abnormality detection, indicating that while history increased reader confidence, it did not affect diagnostic performance. This observation was also found in a study comparing mammography and breast tomosynthesis, where one modality improved diagnostic confidence but did not affect diagnostic performance[19]. This suggests that there is no clear relationship between diagnostic confidence and diagnostic performance regarding correctly localised abnormalities.

As far as we are aware, this is the first study analysing the effect of history on incidental abnormality detection in any diagnostic imaging modality. The use of the free response paradigm allows for

measurement and correct localisation at the abnormality level, a feature not available when evaluating images through the Receiver Operating Characteristic paradigm at the case level[20].

This study has limitations. The type of incidental abnormality (subtle or obvious) was not measured and may have been a factor in detection. However, the scope of this study was to determine if the native abnormality and case type would affect incidental abnormality detection. Further study with more subcategorisation would be required to determine if features of the incidental abnormality and clinical history would impact on detection. Another limitation is the lack of eye tracking in determining reasoning behind the false negative errors. A focused perception study with eye tracking would be required to deeply analyse perceptual issues in CBCT interpretation and incidental abnormality detection.

Conclusions

Clinical history improves the detection of incidental endodontic abnormalities in non-diseased subtle CBCT images. Reader confidence in correctly identified abnormalities was higher with clinical history when disease and non-disease was subtle, but this was not associated with improvement in diagnostic performance.

Disclosures

The authors have no conflicts of interest, financial or otherwise, to disclose.

Table 1. Reader characteristics

	Number of readers	Mean	Range
Clinical experience			
0-9 years	4	3.8	0-8
10-19 years	13	14.5	12-17
20 or more years	6	38.8	23-48
CBCT studies read per month			
0-8	15	2.8	0-8
9-40	4	10	9-12
41 or more	4	66.3	60-80

Table 2. Incidental abnormalities detected in non-diseased cases, mean per reader in parentheses

	Non-diseased subtle			Non-diseased obvious		
	No history	History		No history	History	
All readers	59 (2.6)	76 (3.3)	p=0.04*, T=36	60 (2.6)	68 (3)	p=0.38, T=27.5
Years of experience						
0-9	7 (1.8)	13 (3.3)	p=0.11, H=2.52	13 (3.3)	12 (3)	p=0.67, H=0.19
10-19	39 (3)	44 (3.4)	p=0.80, H=0.07	32 (2.5)	38 (2.9)	p=0.56, H=0.35
20+	13 (2.2)	19 (3.2)	p=0.13, H=2.31	15 (2.5)	18 (3)	p=0.38, H=0.78
CBCTs read per month						
0-8	40 (2.7)	54 (3.6)	p=0.20, H=1.65	38 (2.5)	44 (2.9)	p=0.55, H=0.36
9-40	9 (2.3)	10 (2.5)	p=0.56, H=0.33	8 (2)	12 (3)	p=0.31, H=1.02
41+	10 (2.5)	12 (3)	p=0.19, H=1.69	14 (3.5)	12 (3)	p=0.67, H=0.19

*significant at $p<0.05$

Table 3. Incidental abnormalities detected in diseased cases, mean per reader in parentheses

	Diseased subtle			Diseased moderate			Diseased obvious		
	No history	History		No history	History		No history	History	
All readers	57 (2.5)	67 (2.9)	p=0.08, T=23.5	86 (3.7)	85 (3.7)	p=0.82, T=71.5	58 (2.5)	53 (2.3)	p=0.26, T=20
Years of experience									
0-9	8 (2)	10 (2.5)	p=0.39, H=0.75	17 (4.3)	12 (3)	p=0.19, H=1.69	8 (2)	7 (1.8)	p=0.89, H=0.02
10-19	30 (2.3)	37 (2.8)	p=0.29, H=1.11	43 (3.3)	47 (3.6)	p=0.88, H=0.02	30 (2.3)	30 (2.3)	p=0.94, H=0.01
20+	19 (3.2)	20 (3.3)	p=0.47, H=0.52	26 (4.3)	26 (4.3)	p=0.81, H=0.06	20 (3.3)	16 (2.7)	p=0.26, H=1.26
CBCTs read per month									
0-8	35 (2.3)	42 (2.8)	p=0.31, H=1.03	52 (3.5)	51 (3.4)	p=0.69, H=0.16	37 (2.5)	36 (2.4)	p=0.92, H=0.01
9-40	8 (2)	10 (2.5)	p=0.56, H=0.33	14 (3.5)	13 (3.3)	p=0.77, H=0.08	8 (2)	9 (2.3)	p=0.67, H=0.19
41+	14 (3.5)	15 (3.8)	p=0.56, H=0.33	20 (5)	21 (5.3)	p=0.56, H=0.33	13 (3.3)	8 (2)	p=0.15, H=2.08

313 Table 4. False positive marks in non-diseased cases, mean per reader in parentheses

	Non-diseased subtle			Non-diseased obvious		
	No history	History		No history	History	
All readers	65 (2.8)	54 (2.3)	p=0.22, T=44	88 (3.8)	94 (4.1)	p=0.82, T=109
Years of experience						
0-9	14 (3.5)	8 (2)	p=0.11, H=2.52	12 (3)	20 (5)	p=0.53, H=0.33
10-19	39 (3)	35 (2.7)	p=0.66, H=0.19	58 (4.5)	55 (4.2)	p=0.78, H=0.08
20+	12 (2)	11 (1.8)	p=0.63, H=0.23	18 (3)	19 (3.2)	p=1, H=0
CBCTs read per month						
0-8	47 (3.1)	38 (2.5)	p=0.41, H=0.69	69 (4.6)	65 (4.3)	p=0.68, H=0.17
9-40	9 (2.3)	6 (1.5)	p=0.31, H=1.02	9 (2.3)	18 (4.5)	p=0.47, H=0.52
41+	9 (2.3)	10 (2.5)	p=0.77, H=0.08	10 (2.5)	11 (2.8)	p=0.77, H=0.08

314

315 Table 5. False positive marks in diseased cases, mean per reader in parentheses

	Diseased subtle			Diseased moderate			Diseased obvious		
	No history	History		No history	History		No history	History	
All readers	80 (3.5)	76 (3.3)	p=0.70, T=60.5	124 (5.4)	123 (5.3)	p=0.98, T=94.5	65 (2.8)	70 (3)	p=0.69, T=68
Years of experience									
0-9	24 (6)	22 (5.5)	p=0.89, H=0.02	22 (5.5)	36 (9)	p=0.31, H=1.02	13 (3.3)	13 (3.3)	p=1, H=0
10-19	43 (3.3)	35 (2.7)	p=0.38, H=0.76	84 (6.5)	70 (5.4)	p=0.52, H=0.41	47 (3.6)	47 (3.6)	p=0.88, H=0.02
20+	13 (2.2)	19 (3.2)	p=0.58, H=0.31	18 (3)	17 (2.8)	p=0.81, H=0.06	6 (1)	10 (1.7)	p=0.42, H=0.64
CBCTs read per month									
0-8	58 (3.9)	50 (3.3)	p=0.56, H=0.34	96 (6.4)	92 (6.2)	p=0.97, H=0	56 (3.7)	59 (3.9)	p=0.95, H=0
9-40	12 (3)	10 (2.5)	p=0.67, H=0.19	13 (3.3)	18 (4.5)	p=0.31, H=1.02	5 (1.3)	7 (1.8)	p=0.47, H=0.52
41+	10 (2.5)	16 (4)	p=0.31, H=1.02	15 (3.8)	13 (3.3)	p=0.77, H=0.08	4 (1)	4 (1)	p=1, H=0

316

317

318

319

320

321

322

323

324

Table 6. Correctly localised incidental abnormalities by confidence ratings and reader type, percentage of all ratings in parentheses

	Rating 2 (lowest)			Rating 3 (medium)			Rating 4 (highest)		
	No history	History		No history	History		No history	History	
All readers	32 (10)	28 (8)	p=0.70, T=40	80 (25)	72 (21)	p=0.50, T=78	209 (65)	248 (71)	p=0.02*, T=35
Years of experience									
0-9	9 (17)	11 (20.4)	p=0.89, H=0.02	22 (41.5)	25 (46.3)	p=0.67, H=0.18	22 (41.5)	18 (33.3)	p=0.39, H=0.75
10-19	19 (12)	15 (9)	p=0.54, H=0.38	40 (26)	32 (18)	p=0.43, H=0.63	98 (62)	129 (73)	p=0.12, H=2.45
20+	2 (2)	2 (2)	p=1, H=0	17 (18)	14 (14)	p=0.58, H=0.31	74 (80)	83 (84)	p=0.42, H=0.64
CBCTs read per month									
0-8	24 (12)	23 (10)	p=0.92, H=0.01	66 (33)	57 (25)	p=0.63, H=0.23	112 (55)	147 (65)	p=0.27, H=1.21
9-40	3 (6)	1 (2)	p=0.47, H=0.52	5 (10)	7 (13)	p=1, H=0	40 (84)	45 (85)	p=0.39, H=0.75
41+	5 (7)	4 (6)	p=0.56, H=0.33	9 (13)	8 (12)	p=0.77, H=0.08	57 (80)	56 (82)	p=0.39, H=0.75

*significant at $p<0.05$

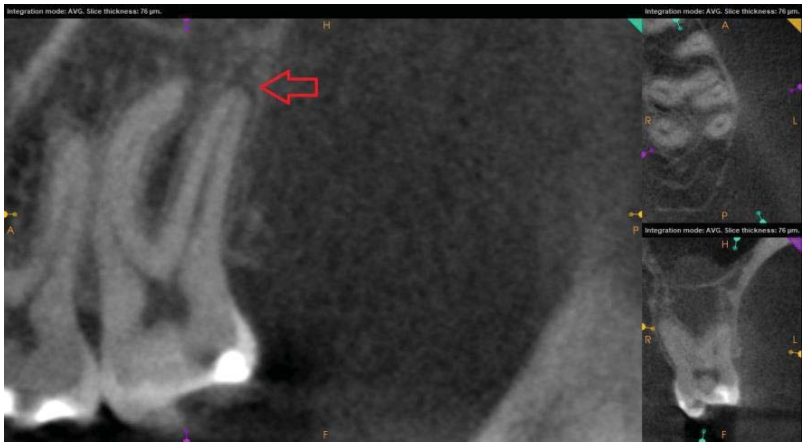
Table 7. Correctly localised incidental abnormalities, by confidence ratings and case type, percentage of all ratings in parentheses

	Rating 2 (lowest)			Rating 3 (medium)			Rating 4 (highest)		
	No history	History		No history	History		No history	History	
Non-diseased group									
Non-diseased subtle	11 (18)	7 (9)	p=0.37, T=2.5	16 (27)	17 (22)	p=0.79, T=35.5	33 (55)	53 (69)	p=0.02*, T=26
Non-diseased obvious	9 (15)	9 (13)	p=1, T=27.5	16 (26)	21 (29)	p=0.40, T=28	36 (59)	41 (58)	p=0.44, T=60
Diseased group									
Diseased subtle	3 (6)	5 (8)	p=0.47, T=7	12 (21)	7 (11)	p=0.15, T=10	41 (73)	53 (81)	p=0.03*, T=22.5
Diseased moderate	9 (11)	5 (6)	p=0.16, T=3.5	28 (33)	22 (27)	p=0.24, T=28	48 (57)	56 (67)	p=0.20, T=31.5
Diseased obvious	0 (0)	3 (6)	p=0.19, T=0	7 (12)	4 (7)	p=0.32, T=8	51 (88)	45 (87)	p=0.25, T=28.5

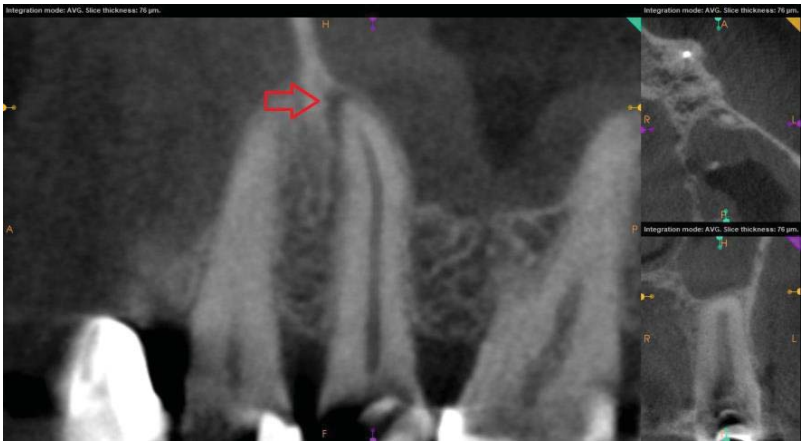
*significant at $p<0.05$

Figure. Case examples (red arrows) of (a) diseased subtle (b) diseased moderate (c) diseased obvious

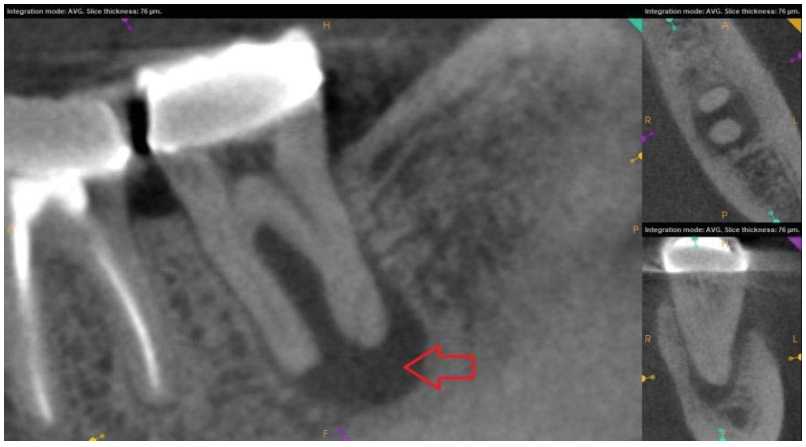
(a)



(b)



(c)



References

1. Devine, A.S., Jackson, C.S., Lyons, L., and Mason, J.D., *Frequency of incidental findings on computed tomography of trauma patients.* West J Emerg Med, 2010. **11**(1): p. 24-7.
2. Thompson, R.J., Wojcik, S.M., Grant, W.D., and Ko, P.Y., *Incidental Findings on CT Scans in the Emergency Department.* Emergency medicine international, 2011. **2011**: p. 624847-624847.

3. Niedermeier, S., Wania, R., Lampart, A., et al., *Incidental CT Findings in the Elderly with Low-Energy Falls: Prevalence and Implications.* Diagnostics (Basel, Switzerland), 2022. **12**(2): p. 354.
4. Lee, C.S., Nagy, P.G., Weaver, S.J., and Newman-Toker, D.E., *Cognitive and system factors contributing to diagnostic errors in radiology.* AJR Am J Roentgenol, 2013. **201**(3): p. 611-7.
5. Busby, L.P., Courtier, J.L., and Glastonbury, C.M., *Bias in Radiology: The How and Why of Misses and Misinterpretations.* Radiographics, 2018. **38**(1): p. 236-247.
6. Kundel, H.L., Nodine, C.F., and Carmody, D., *Visual scanning, pattern recognition and decision-making in pulmonary nodule detection.* Invest Radiol, 1978. **13**(3): p. 175-81.
7. Berbaum, K.S., Franken, E.A., Dorfman, D.D., et al., *Satisfaction of Search in Diagnostic Radiology.* Investigative Radiology, 1990. **25**(2): p. 133-140.
8. Adamo, S.H., Cain, M.S., and Mitroff, S.R., *Self-induced attentional blink: a cause of errors in multiple-target search.* Psychol Sci, 2013. **24**(12): p. 2569-74.
9. Tuddenham, W.J., *Visual search, image organization, and reader error in roentgen diagnosis. Studies of the psycho-physiology of roentgen image perception.* Radiology, 1962. **78**: p. 694-704.
10. Adamo, S.H., Gereke, B.J., Shomstein, S., and Schmidt, J., *From "satisfaction of search" to "subsequent search misses": a review of multiple-target search errors across radiology and cognitive science.* Cogn Res Princ Implic, 2021. **6**(1): p. 59.
11. Hiserote, D., *The Effect of Patient Dental History on Cone Beam Computed Tomography Interpretation of Periapical Findings in Endodontics*. 2015, Saint Louis University.
12. Nguyen, A., *The effect of a localizing prompt on cone beam computed tomography interpretation*. 2017, Saint Louis University.
13. Cohen, J.F., Korevaar, D.A., Altman, D.G., et al., *STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration.* BMJ open, 2016. **6**(11): p. e012799-e012799.
14. Yapp, K.E., Brennan, P., and Ekpo, E., *The effect of clinical history on diagnostic performance of endodontic cone-beam CT interpretation.* Clinical Radiology, in press 2023.
15. Chakraborty, D.P., *A search model and figure of merit for observer data acquired according to the free-response paradigm.* Physics in Medicine & Biology, 2006. **51**(14): p. 3449-62.
16. Bunch, P., Hamilton, J., Sanderson, G., and Simmons, A., *A Free Response Approach To The Measurement And Characterization Of Radiographic Observer Performance*. Application of Optical Instrumentation in Medicine VI. Vol. 0127. 1977: SPIE.
17. Manning, D.J., Ethell, S.C., and Donovan, T., *Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph.* British Journal of Radiology, 2004. **77**(915): p. 231-235.
18. Degnan, A.J., Ghobadi, E.H., Hardy, P., et al., *Perceptual and Interpretive Error in Diagnostic Radiology-Causes and Potential Solutions.* Academic Radiology, 2019. **26**(6): p. 833-845.
19. Hadadi, I., Rae, W., Clarke, J., McEntee, M., and Ekpo, E., *Breast cancer detection: Comparison of digital mammography and digital breast tomosynthesis across non-dense and dense breasts.* Radiography (Lond), 2021. **27**(4): p. 1027-1032.
20. Chakraborty, D.P. and Berbaum, K.S., *Observer studies involving detection and localization: modeling, analysis, and validation.* Med Phys, 2004. **31**(8): p. 2313-30.

CHAPTER 7

7.1 Thesis findings

This thesis aims to provide new data on the diagnostic performance of CBCT and digital PA radiography in endodontic disease detection, and the effect of clinical history on interpretation of CBCT images in endodontic disease and incidental abnormality detection. This was achieved by using a factorial, free-response study to compare: diagnostic performance of CBCT and digital PA radiography in endodontic disease detection, the effect of clinical history on diagnostic performance of CBCT in endodontic disease detection, and to analyse the effect of clinical history on incidental abnormality detection and diagnostic confidence in CBCT.

When CBCT and digital PA radiography were compared, CBCT had greater specificity than PA in the obvious non-diseased cases and no significant difference in the subtle non-diseased category. Diagnostic performance was greater for PA radiography than CBCT where wAFROC1 values were higher in the subtle diseased and moderate diseased groups, with no significant difference between in the obvious diseased group. False positive errors were higher for CBCT than PA radiography in subtle diseased cases. PA radiography had a higher mean lesion localisation fraction than CBCT in the moderate diseased group. When clinical experience and all diagnostic performance measures were compared, no relationships were found, except the obvious diseased CBCT group where increasing experience was associated with fewer mean false positives than PA radiography.

In comparing CBCT interpretation with and without clinical history, there was no significant effect on specificity and false positive rates in non-diseased cases. History improved diagnostic performance in diseased cases; it improved lesion localisation in subtle and obvious diseased cases and had higher wAFROC1 values in subtle and obvious diseased categories. No associations were observed between clinical history and both readers' years of experience and reading volume in the non-diseased categories. Readers with fewer and moderate years of experience and low CBCT reading volume demonstrated better lesion localisation in subtle diseased cases when clinical history was available.

Incidental abnormalities had increased detection with clinical history in non-diseased subtle cases. Reader experience and monthly CBCT reading volume did not affect incidental abnormality detection. The highest confidence rating was most often used in each case type when clinical history was available. For this rating (4), history had significantly greater lesion localisations in subtle diseased and non-diseased images. Insights from these studies are discussed below.

7.1.1 Periapical radiography versus CBCT in endodontic disease detection

This study found that CBCT outperformed PA imaging in non-diseased cases and PA outperformed CBCT imaging in diseased cases. Whilst the methods of data analysis were more rigorous and different to those in the existing literature, there were differences when results of similar measurement were able to be compared. CBCT and PA radiography specificity was substantially lower than the data from other studies. This indicates that diagnostic performance of both modalities in non-diseased cases has been previously overestimated, suggesting that readers would have a tendency to overcall normal images. In practice, this should be considered so that clinicians can be more careful and ensure that false positive calls are minimised.

These findings in non-diseased cases can be explained by differences in study methodology and diagnostic performance measurement. In this thesis, readers were instructed to view the entire image or volume and mark as many abnormalities as they wished, without any localising prompts nor specific search or interpretive tasks. This method replicates image interpretation in a clinical setting. In contrast, other studies had a reader task of rating: the diameter of PA radiolucency (Pope et al. 2014), presence of a "PA lesion" (Kanagasingam et al. 2017) or presence of apical periodontitis (Kruse et al. 2019). By limiting the reader to only interpreting a specific area of the image, which is not representative of the clinical task, it limits the number of ratings per image and prevents the opportunity to interpret other relevant parts of the image. The method used in these other studies reduces the risk of false positive marks by reducing the opportunity to make the marks in the first place. By doing so, it increases the probability of producing higher specificity. In this thesis, by giving readers the opportunity of making an infinite number of marks per image, it increases the theoretical risk of making false positive marks and reducing specificity.

CBCT had greater diagnostic performance than PA imaging in non-diseased cases, with significantly greater specificity in the obvious cases, no significant difference in the subtle cases and fewer false positives in subtle and obvious cases. This may be due to lack of reader experience with the newer CBCT imaging modality. Because of the limited time in the market as an imaging option to dental patients, the technology may not be well understood, and knowledge of the image generation and reconstruction process may not be widely known. This may have introduced hesitancy in assigning abnormalities on CBCT images due to unfamiliarity. Reduced performance in PA imaging may have been a function of overcall using a more familiar modality. To account for reader experience, recruitment included a wide range of readers with varying experience; however, experience measured by the number of years of clinical experience was not found to influence performance. This may be due to years of clinical experience not necessarily being a reflection of expertise in image interpretation, nor of training in this particular domain. This tendency for a performance difference

between modalities suggests a change in reader willingness to make a call, which may be considered a criterion shift in the ROC paradigm. A criterion shift is a change in perceived reader parameters in a binary classifier task, where one of two choices must be made for the entire task. The work presented in this thesis was based on the free response paradigm. Because the free response paradigm does not employ a two-alternative forced choice (2AFC) method of measurement, reader willingness cannot be measured nor explained by a criterion shift in this thesis. Thus, the findings presented in this thesis may not necessarily be a function of reader willingness because of the unlimited number of marks readers are allowed to make and the lack of relationship between the total number of true and false positive marks in the free response paradigm. This is because a reader's willingness to make a call in one part of the image does not mean they have the same willingness to make a call in another part of the same image.

In diseased cases, wAFROC1, lesion localisation fraction and false positive mark values were unable to be compared to the existing body of literature because no other studies measured performance at abnormality level. When comparing modalities, PA had significantly greater performance values than CBCT in most case categories with all methods of performance analysis. This could be explained by domain expertise in PA imaging due to its long history in dentistry and widespread experience amongst readers. This level of expertise has not yet been shown in CBCT imaging because of its recent introduction to dental imaging. Therefore, less opportunity and experience for readers to not only interpret CBCT imaging and receive feedback, limits the ability for skillset development in this modality. Another finding not reported in previous studies, was that for both CBCT and PA imaging, in all metrics of reader performance analysis in diseased cases, diagnostic performance increased as disease severity increased. This is likely due to distinguishing imaging features becoming more obvious and a reduced threshold for abnormality detection, with increasing severity. These results are similar to an earlier study that measured diagnostic performance of film and digitised film PA radiography at the case level (Khademi 1994). These findings suggest that while these imaging modalities perform better with increased disease severity, diagnostic performance decreases as disease becomes more subtle – thereby reducing the utility of these diagnostic tests in more difficult situations. Lower performance can be explained by observer difficulty in detecting and categorising abnormalities on these imaging tools when disease presents with difficult imaging features. Subtle lesions detected "must have a sufficient number of visible features to hold the observer's attention and allow them to discriminate between lesion and background noise" (Krupinski 2005). This suggests that "very subtle lesions that are missed simply do not have enough visible features to maintain the observer's interest or attention after it has been fixated". According to Krupinski et al., obvious abnormalities are more easily identified because they "have enough visible features to hold the observer's attention longer, but the observer is unable to recognize enough features or discriminate

enough of them from the background to decide that a lesion is present". This could explain the findings reported for PA radiography and CBCT imaging in chapter four.

Results of the study presented in chapter four indicate that performance levels are low (specificity and wAFROC1 values) and the error rate is high (false positives) for both PA radiography and CBCT. These findings suggest that diagnostic performance of PA and CBCT imaging in the endodontic domain when a rigorous observer performance methodology is used is much lower than studies previously published. The wider variation in observer performance reported in chapter four also indicates that reader factors may have also contributed to low specificity and high false positives. In relation to clinical practice, these findings emphasise the need for further training in image perception and interpretation – particularly for the newer CBCT imaging modality – is required. A better understanding of the technological basis for CBCT image reconstruction would also help readers in discerning relevant details of image features and their representation.

Reader factors, in particular clinical experience, were not shown to influence performance for both PA and CBCT imaging. Mammography-based studies have shown that, radiologists with more experience outperformed residents demonstrating lower false positives (Nodine et al. 1999, Tan et al. 2006, Elmore et al. 2009) and higher specificity (Barlow et al. 2004). This may be due to older adults having increased effectiveness of emotion regulation and a greater reluctance to make decisions (Mather 2006). The work presented in chapter four is the first to assess reader factors influencing diagnostic image interpretation in the endodontic domain and the results presented may be due to multiple factors. First, clinical experience is not necessarily a representation of image interpretation frequency or readers' interaction with images produced by either of these imaging modalities. Secondly, years of clinical experience is not a surrogate for specific training in any domain, and factors other than years of experience may contribute to the development of expertise. The study presented in chapter 4 included participants with a wide range of experience from 0 to 54 years, with just over half the sample in the 10–19-year group. Due to the recent introduction of CBCT technology into clinical practice, readers with decades of clinical experience in dentistry have at best, limited experience interpreting CBCT images. Therefore, it is not surprising that no performance difference was found between readers of low and high levels of experience.

Chapter four was able to address the limitations identified in published studies as described in chapters one and two. In particular, the methodology employed accounted for factors that impact upon performance to increase external validity of the results. First, unlike previous studies that used only diseased cases, a sample population with the entire spectrum of disease and non-disease severity was used to assess the performance of PA radiography and CBCT in chapter four. Secondly, the reader task was a relevant clinical task allowing for the identification of abnormalities in the entire image

and the inclusion of multiple abnormalities per case. Thirdly, a valid, independent, and clearly defined reference standard established by a Delphi panel using pre-operative and follow up clinical and radiographic data was used. Fourthly, data analysis involved measurement at the abnormality, not the case level, and rewarded correct decisions and penalised incorrect decisions. By addressing the limitations of previous studies, data presented in this thesis provides a better understanding of the diagnostic performance of both PA and CBCT imaging.

7.1.2 Effect of Clinical history on diagnostic performance of endodontic CBCT interpretation

The findings in chapter five show that clinical history improved diagnostic performance in diseased cases but not in non-diseased cases, where there was a slight reduction in performance. These findings are likely due to the cognitive and perceptual biases due to having access to history before image interpretation. The first feature of clinical history is that there is likely to be distinguishing information related to the patient presentation, directing the attention to a particular image region containing an abnormality. When a reader is prompted to refer to a localised area, abnormalities are more likely to be detected (Nguyen 2017). If the history is suggestive of a disease and the accompanying imaging features are not readily detected, the clinical description would not only influence reader judgement on the presence of an abnormality that would not be identified without history, but also be enough to elevate the prior probability of an abnormality regardless of imaging features (Pauker and Kassirer 1980). In a situation of a non-diseased case with subtle or ambiguous image perturbations, clinical history suggestive of a disease may bias the reader decision to report an abnormality that is not present (Ransohoff and Feinstein 1978), which would explain the reduction in diagnostic performance in non-diseased cases in chapter five. When readers read the same cases under two reading modality conditions – with and without history – and performance shifts were seen between modalities, this could be seen as history being a criterion shift (changing a reader's willingness to make a call). This is noted in studies using signal detection theory, using the ROC paradigm, where measurement occurs at the case level and there is a relationship between true and false positive calls. In the free response paradigm, this does not occur because the reader is free to mark as many areas of the image they consider to be abnormal. There is also no relationship between the willingness of the reader to make a call in multiple parts of the same image. Even though only the reading conditions differed, and all other variables were the same, the criterion shift cannot be an explanation for the findings in chapter five. This is because the free response measurement method does not allow for the criterion shift to be analysed – because there is no relationship between true and false positive calls being made. This would indicate that clinically, the effect of history would not be a criterion shift, but instead another type of bias or set of biases.

When comparing the results to the published literature on diagnostic image interpretation, all available studies did not separate analysis of diseased and non-diseased cases – therefore, a direct comparison was not possible. Most studies identified in the systematic review presented in chapter three measured diagnostic performance at the case level and found that reading images with history had better performance than without history. This study used measurement at the abnormality level and the results were similar to two studies with a similar methodology – their findings being that history reduced (Dhingsa et al. 2004), or had no significant effect on diagnostic performance (Littlefair et al. 2016). Measurement at abnormality level includes marking the location of every reported abnormality, which rewards correct and penalises incorrect locations. The greater discriminating power is able to account for errors that measurement at case level cannot. This explains the difference in results using different measurement paradigms; if a case with an abnormality has an incorrect localisation, two errors occur – a false negative (miss) and a false positive (incorrect location). Case level measurement wrongly assigns this to be a correct call, where abnormality level measurement assigns this as two errors. In the ROC (case level) paradigm, these errors will give the appearance of increased sensitivity. Because case level measurement cannot assign false positives to diseased cases, the number of false positives will appear to be lower than reality, reducing the false positive fraction. On a plot of an ROC curve, the increased true positive fraction (sensitivity) and reduced false positive fraction results in a greater area under the ROC curve than what the true results would show. Therefore, the methodology employed in chapter five overcomes these limitations and provides a comprehensive understanding of the effect of clinical history on diagnostic performance.

Reader characteristics and their effect on diagnostic performance were mixed between diseased and non-diseased images. No associations were observed between clinical history and both reader experience and reading volume in the non-diseased categories. This may be due to a lack of experience in readers of all levels of experience or amount of reading volume in interpreting CBCT imaging without endodontic disease. Guidelines in endodontics have been published indicating that clinicians should gather all clinical and two-dimensional diagnostic imaging prior to ordering a CBCT imaging study (2016), which would suggest that these tend to be cases already suspicious of having an endodontic issue. The recent introduction of CBCT into dentistry is also a factor, in addition to the lack of specific CBCT training programmes designed to improve performance. Thus, it is logical that dentists and radiologists interpreting CBCT images should demonstrate low reading volume and years of experience in CBCT interpretation. Some readers with decades of clinical experience had a low CBCT monthly reading volume and vice versa. These factors would reduce the number of available non-diseased CBCT images ordered for endodontic purposes. Therefore, it is unsurprising that no clear relationship was found between the two factors and diagnostic performance when clinical history was available. Because the opportunity to interpret CBCT images without endodontic disease is limited, the development of expertise is also restricted and performance differences between

reading non-diseased images with and without history would be difficult to detect due to a low level of expertise.

Readers interpreting diseased cases showed that clinical history improved performance for those with fewer (0-9) and moderate (10-19) years of experience having greater lesion localisations, and moderate years of experience having higher wAFROC1 values, in subtle diseased cases. Those with a low (0-8) monthly CBCT reading volume had better lesion localisation with history in the same subtle diseased cases. This is suggestive of clinical history influencing reader performance with lower levels of experience and reading volume, by directing attention to the relevant imaging areas with abnormalities, and/or elevating the prior probability of a patient having endodontic disease. It may be that the reader cohort with low CBCT reading experience required highlighting of the salient features in the patient presentation, which would help their detection or interpretation of abnormal imaging findings and improve performance.

Compared to studies in other imaging domains, these results are different regarding the effect of reader factors on diagnostic performance. In mammography studies, annual reading volume improved diagnostic performance (Rawashdeh et al. 2013), location sensitivity and area under the ROC curve values (Suleiman et al. 2014). Although these studies did not have consistent findings regarding reader characteristics on diagnostic performance, comparisons between different study types can be difficult. These mammography studies did not compare reader performance with and without history. Their datasets did not separate diseased and non-diseased images. The severity of disease was not disclosed, although mainly screening images were used, which may have combined those with subtle and obvious features of disease. As discussed previously and demonstrated in chapter four, the severity of the cases included in the dataset influences reader performance. Compared to dental images, the subtlety of mammographic lesions and heterogeneity of breast parenchyma impact upon reader performance in mammography. Therefore, it is difficult to compare mammography-based results and the data presented in this thesis.

By using this updated methodology, data can be analysed more precisely and more accurate conclusions about the effect of clinical history on diagnostic image interpretation can be made. This informs clinical practice by reminding clinicians of the bias of clinical history has on CBCT image interpretation and how this bias may influence diagnostic performance in the detection of endodontic disease and the interpretation of normal (non-diseased) images. By having access to clinical history, interpretation can both improve and reduce performance, depending on the case type – which the reader does not know. There are guideline recommendations that patient history should be taken and a clinical examination performed before interpreting CBCT images. However, the results show that having a clinical history prior to CBCT image interpretation will bias the results, in a way unknown to

the clinician. It is important that the potential impact of clinical history on reader decision making is considered when undertaking image interpretation or developing future guidelines around the request and interpretation of dental images.

7.1.3 Clinical history and incidental abnormality detection in endodontic CBCT

The results show that clinical history improved detection of incidental abnormalities only in subtle non-diseased cases. Reader confidence regarding abnormality identification in the highest category was greater when history was available, but not associated with an overall improvement in diagnostic performance. These findings of clinical history showing no sign of any effect on detecting incidental abnormalities in all diseased cases may be due to the presence of the native abnormality and its effect on the perceptual process, rather than the bias of clinical history. No difference in false positives in cases with and without clinical history was found, indicating no sign of history bias on imaging overcalls.

In cases where there was endodontic disease present, failure to detect a subsequent abnormality is a false negative error. These have been historically described as scanning, recognition and decision errors (Kundel et al. 1978). These errors occur when the abnormality is not fixated by the observer (scanning), the abnormality is briefly fixated without being distinguished from the background parenchyma (recognition), or when the observer fixates the abnormality and decides no abnormality is present or it is benign and chooses not to report it (decision). To determine which type of false negative error has occurred, eye tracking is required to verify areas of fixation. Because this study did not perform eye tracking, the type/s of error in non-detection of these incidental abnormalities could not be established. Other reasons for the non-detection of incidental abnormalities in cases with native abnormalities include the satisfaction of search effect (Berbaum et al. 1990), later revised to be the "subsequent search miss" effect (Adamo et al. 2013), which is when subsequent targets are less likely to be detected after a first target is detected. Theories for why they occur could explain why incidental abnormalities were missed with and without clinical history prior to interpretation. These theories have been described as satisfaction (the reader has satisfied a "quest for meaning" once the first target is localised), perceptual set (observers, upon finding a target, are biased to search for similar targets and more likely to miss dissimilar targets) and resource depletion (after expending attention and working memory resources to detect a first target, fewer resources are available to detect an additional target) (Adamo et al. 2021). According to a new "Attentional Template" theory, detecting a first target creates an attentional template, incurs a search cost, reduces the cognitive resources required to recognise a new target, thereby increasing the likelihood of missing a subsequent search target. It is likely that a combination of the above factors is contributing to missed incidental abnormalities in cases with native endodontic disease. When there is a perception task that uses attentional resources, it

is possible that any bias from having clinical history prior to interpretation does not impact on a subsequent interpretive task – no effect was found for diseased images in chapter six. Although the relationship between detection and non-detection of the native and incidental abnormalities was beyond the scope of this study, it is an area that requires further assessment.

The increased detection of incidental abnormalities in non-disease subtle cases when clinical history was available may be due to conflicting probabilities of disease. The history would be suggestive of endodontic disease and the image features not suggestive of endodontic disease, which would then lead the reader to be more vigilant in detecting a subsequent abnormality. This would then lead to reader "satisfaction" or fulfilling a "quest for meaning" by detecting an incidental abnormality (Tuddenham 1962).

The effect of clinical history on reader confidence showed that in non-diseased subtle and diseased subtle cases using the highest confidence level, history had an increase in correctly localised incidental abnormalities. When each confidence rating was evaluated for all readers, history led to an increase in incidental lesion localisations for the highest level of confidence. This observation may be due to information in the clinical history providing an intermediate probability of endodontic disease, which then affects the vigilance and subsequent threshold of the reader detecting an additional abnormality. This would be a form of commission bias (Croskerry 2002), where there is an urge to do something – which in this setting is to make a confident call of an abnormality being present.

The published literature did not have any studies studying the effect of clinical history on incidental abnormality detection for any imaging modality; therefore, the literature does not provide any basis for comparison of the findings in chapter six with previous pieces of work. The only study that assessed diagnostic confidence focused on breast imaging and compared readers confidence in mammography and digital breast tomosynthesis and showed that diagnostic confidence increased without affecting diagnostic performance (Hadadi et al. 2021). This work has shown the limited benefit of clinical history prior to image interpretation regarding incidental abnormality detection. The relevance to clinical practice is that this data indicates clinical history is a bias prior to CBCT image interpretation. It is inconsistent with the guideline recommendations of having a patient history and clinical examination prior to CBCT image interpretation.

7.2 Thesis implications

There are several findings of this thesis.

- The first is that there is no clear overall advantage of CBCT imaging over PA radiography, with newer CBCT imaging having a limited benefit and no performance advantage over PA

radiography and vice versa. Therefore, any of these modalities can be used for examining patients presenting with a suspected endodontic issue.

- The second major finding is that clinical history has a benefit in endodontic disease detection with CBCT in diseased cases only with insignificant disadvantage in patients without endodontic disease. Because the reader has no knowledge of the patient's disease state, it is difficult to establish if clinical history will bias their diagnostic decision making and potentially lead to overdiagnosis and overtreatment.
- The third major finding is that clinical history offers limited improvement in detection of incidental abnormalities using CBCT imaging, with some advantage in subtle non-diseased cases only. Clinical history does not improve the detection of incidental abnormalities when native abnormalities are present. Therefore, the detection of incidental abnormalities should not be affected by the presence or absence of clinical information.
- The fourth major finding is that reading volume and years of clinical experience have no appreciable effect on detection of endodontic disease on CBCT images. This may be a function of the limited time that CBCT has been available in dentistry and the lack of long-term expertise development available in the dental imaging domain compared to other medical imaging domains.

In addition to the implications above, a prior probability factor was identified which is relevant but not emphasised in the existing endodontic or dental literature. Disease severity affects diagnostic performance and as disease became more obvious, performance increased, and as disease became more subtle and difficult to detect, performance decreased. This implies that the value of the diagnostic test is lower in the situations where the pre-test probability of disease is low where it requires further diagnostic information to determine if disease is present. It also indicates that disease severity is an important factor in addition to reader or modality factors regarding diagnostic interpretation.

This thesis can be used as a framework for future studies to be based upon. These findings show that there is clinical history bias prior to interpreting CBCT images. This is inconsistent with the recommendation of providing clinical history prior to CBCT interpretation. Future studies should focus on analysing the various settings where this bias can affect diagnostic performance and look for strategies that can improve performance.

7.3 Limitations

One of the limitations of this thesis is the experimental setting of the study, which can introduce demand characteristics, which are experimental artifacts that alter behaviour to suit the perceived

goals of the study (Orne 1962). To read these images in a clinical setting without seeing the patient is not practical for dentists because the dentist is the ordering and interpreting clinician for any radiographs, and usually the intervening clinician for patient treatment. However, a mammography-based study has shown that there is concordance between the performance of readers in a clinical and laboratory setting (Soh et al. 2014), which supports the external validity of the findings of this thesis.

Another limitation of the thesis is the reader sampling strategy. All the readers were sampled from Australia and included those who were available and willing to participate in the study. As the sample involved Australian clinicians, there may be cultural factors affecting behaviour that are not seen in readers from other countries. The incidental sampling strategy may have also influenced the results, as these readers may not completely capture the characteristics and experience of the population of dentists interpreting PA radiography and CBCT images. The minimal number of medical radiologists who participated in the study may not be representative of the medical radiologists who do interpret dental imaging; however, the prevalence of this occurring has not been studied.

The third limitation relates to the design of the reading strategy. Reading images twice, even without knowledge of the fact, may introduce memory bias. However, a chest radiography study found that when tested, memory erodes to chance level within seven weeks (Evans et al. 2016). Given the memory effect in dental CBCT imaging has not been studied and the mean interval between reading sessions in the history experiment was 22 weeks, the reading strategy is unlikely to influence the results of the studies presented in this thesis.

The fourth limitation is the actual clinical history was used in chapters five and six. Real clinical history, gathered by skilled clinicians, used in the study may be both a limitation and strength of the thesis. These details are likely to be of greater diagnostic value than a vague history missing relevant details, due to the ability to gather salient patient information and increase (or decrease) the prior probability of endodontic disease being present, thereby strengthening the effect of history. The history used in the study may not have identical qualitative value to a more general history, which may be missing relevant patient history and clinical examination information, making this a limitation due to the limited external validity.

7.4 Recommendations for future studies

Future studies should build upon the methods and findings of this thesis. In particular, studying the interpretive process in CBCT imaging, both with and without history in the same reading session, would be a clinically relevant experiment. This would involve a reading without history, followed by reading the history, then another look at the images. This method has been studied in chest

radiographs (Berbaum et al. 1994) and suggested in a later editorial (Griscom 2002). It would be able to provide data on interpretative changes both before and after having access to history. Another method of expanding upon this thesis is to measure the qualitative effect of clinical history and to assess which parts are relevant in increasing or decreasing the prior probability of endodontic disease.

7.5 Conclusion

Periapical radiography performs better than CBCT in the detection of endodontic disease, but worse in interpretation of non-diseased images. Access to clinical history improves the interpretation of CBCT images with disease without significantly affecting the interpretation of images without disease. Clinical history also improves the detection of incidental endodontic abnormalities in non-diseased subtle CBCT images only. Moderately experienced and low volume readers benefit more from having access to clinical history when interpreting CBCT images. Availability of clinical history also increases reader confidence in correctly identifying subtle abnormalities and images with no disease features. Insights from this thesis should be used to inform referral for endodontic imaging and choice of imaging modality to improve the diagnosis and treatment of endodontic disease.

References

American Association of Endodontists and the American Academy of Oral and Maxillofacial Radiology (2016). Use of Cone Beam Computed Tomography in Endodontics — 2015/2016 Update.

Adamo, S. H., et al. (2013). "Self-induced attentional blink: a cause of errors in multiple-target search." Psychol Sci **24**(12): 2569-2574.

Adamo, S. H., et al. (2021). "From "satisfaction of search" to "subsequent search misses": a review of multiple-target search errors across radiology and cognitive science." Cogn Res Princ Implic **6**(1): 59.

Barlow, W. E., et al. (2004). "Accuracy of screening mammography interpretation by characteristics of radiologists." J Natl Cancer Inst **96**(24): 1840-1850.

Berbaum, K. S., et al. (1990). "Satisfaction of Search in Diagnostic Radiology." Investigative Radiology **25**(2): 133-140.

Berbaum, K. S., et al. (1994). "Influence of clinical history on perception of abnormalities in pediatric radiographs." Acad Radiol **1**(3): 217-223.

Croskerry, P. (2002). "Achieving quality in clinical decision making: cognitive strategies and detection of bias." Acad Emerg Med **9**(11): 1184-1204.

Dhingsa, R., et al. (2004). "Prostate Cancer Localization with Endorectal MR Imaging and MR Spectroscopic Imaging: Effect of Clinical Data on Reader Accuracy." Radiology **230**(1): 215-220.

Elmore, J. G., et al. (2009). "Variability in Interpretive Performance at Screening Mammography and Radiologists' Characteristics Associated with Accuracy." Radiology **253**(3): 641-651.

Evans, K. K., et al. (2016). "Radiologists remember mountains better than radiographs, or do they?" J Med Imaging (Bellingham) **3**(1): 011005.

Griscom, N. T. (2002). "A Suggestion: Look at the Images First, Before You Read the History." Radiology **223**(1): 9-10.

Hadadi, I., et al. (2021). "Breast cancer detection: Comparison of digital mammography and digital breast tomosynthesis across non-dense and dense breasts." Radiography (Lond) **27**(4): 1027-1032.

Kanagasingam, S., et al. (2017). "Diagnostic accuracy of periapical radiography and cone beam computed tomography in detecting apical periodontitis using histopathological findings as a reference standard." Int Endod J **50**(5): 417-426.
Khademi, J. (1994). The Effects of Digitization on Diagnostic Performance in Endodontics and Periodontics: an ROC study Master Thesis, The University of Iowa.
Krupinski, E. A. (2005). "Visual search of mammographic images: influence of lesion subtlety." Acad Radiol **12**(8): 965-969.
Kruse, C., et al. (2019). "Diagnostic accuracy of cone beam computed tomography used for assessment of apical periodontitis: an ex vivo histopathological study on human cadavers." Int Endod J **52**(4): 439-450.
Kundel, H. L., et al. (1978). "Visual scanning, pattern recognition and decision-making in pulmonary nodule detection." Invest Radiol **13**(3): 175-181.
Littlefair, S., et al. (2016). "Increasing Prevalence Expectation in Thoracic Radiology Leads to Overcall." Academic Radiology **23**(3): 284-289.
Mather, M. (2006). A Review of Decision-Making Processes: Weighing the Risks and Benefits of Aging. National Research Council (US) Committee on Aging Frontiers in Social Psychology, Personality, and Adult Developmental Psychology. L. Carstensen and C. Hartel, National Academies Press (US).
Nguyen, A. (2017). The effect of a localizing prompt on cone beam computed tomography interpretation, Saint Louis University.
Nodine, C. F., et al. (1999). "How experience and training influence mammography expertise." Academic Radiology **6**(10): 575-585.
Orne, M. T. (1962). "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications." Am Psychologist: 776-783.
Pauker, S. G. and J. P. Kassirer (1980). "The threshold approach to clinical decision making." N Engl J Med **302**(20): 1109-1117.
Pope, O., et al. (2014). "A Comparative Investigation of Cone-beam Computed Tomography and Periapical Radiography in the Diagnosis of a Healthy Periapex." J Endod **40**(3): 360-365.
Ransohoff, D. F. and A. R. Feinstein (1978). "Problems of spectrum and bias in evaluating the efficacy of diagnostic tests." N Engl J Med **299**(17): 926-930.
Rawashdeh, M. A., et al. (2013). "Markers of Good Performance in Mammography Depend on Number of Annual Readings." Radiology **269**(1): 61-67.
Soh, B. P., et al. (2014). "Mammography test sets: Reading location and prior images do not affect group performance." Clinical Radiology **69**(4): 397-402.
Suleiman, W. I., et al. (2014). "Number of mammography cases read per year is a strong predictor of sensitivity." J Med Imaging (Bellingham) **1**(1): 015503.
Tan, A., et al. (2006). "Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment." Breast Cancer Research and Treatment **100**(3): 309-318.
Tuddenham, W. J. (1962). "Visual search, image organization, and reader error in roentgen diagnosis. Studies of the psycho-physiology of roentgen image perception." Radiology **78**: 694-704.

Appendix A: Participant information statement

THE UNIVERSITY OF SYDNEY

Discipline of Medical Imaging Science
School of Health Sciences
Faculty of Medicine and Health

ABN 15 211 513 464

Dr Ernest Ekpo
Academic Fellow

Room 215
Building M, C42 – Cumberland Campus
The University of Sydney
NSW 2006 AUSTRALIA
Telephone: +61 2 9351 9656
Facsimile: +61 2 9351 9146
Email: ernest.ekpo@sydney.edu.au
Web: http://www.sydney.edu.au/

Diagnostic efficacy of dental digital 2D periapical radiography and 3D cone beam computed tomography

PARTICIPANT INFORMATION STATEMENT

(1) What is this study about?

You are invited to take part in a research study about the detection of dental abnormalities in two-dimensional periapical radiography and three-dimensional cone beam computed tomography. The purpose of the study is to assess which of these imaging tools is better suited for diagnosing different dental abnormalities. The study aims to achieve three aims: 1) assess the diagnostic performance of digital 2D periapical radiography compared to 3D cone beam computed tomography in the detection of dental abnormalities; 2) assess the association between diagnostic performance and the characteristics of the person who interprets these images such as training (dental and radiology) and experience in diagnostic performance in periapical radiography and cone beam computed tomography interpretation; 3) investigate the effect of clinical and dental history on the accuracy of interpretation. Findings of the study will help to provide informed choices of technology and reader factors for improving diagnostic efficacy.

You have been invited to participate in this study because you are a radiologist or dentist interpreting images produced by digital 2D periapical radiography and 3D cone beam computed tomography. This Participant Information Statement tells you about the research study. Knowing what is involved will help you decide if you want to take part in the research. Please read this sheet carefully and ask questions about anything that you don't understand or want to know more about.

Participation in this research study is voluntary.

By giving your consent to take part in this study you are telling us that you:
- ✓ Understand what you have read.
- ✓ Agree to take part in the research study as outlined below.
- ✓ Agree to the use of your personal information as described.

You will be given a copy of this Participant Information Statement to keep.

(2) Who is running the study?

The study is being carried out by the following researchers:

- Dr Ernest Ekpo, Academic Fellow, The University of Sydney
- Professor Patrick Brennan, Professor of Diagnostic Imaging, The University of Sydney
- Dr Kehn Yapp, PhD Student, The University of Sydney.

STUDENT DECLARATION
Dr Kehn Yapp is conducting this study as the basis for the degree of Doctor of Philosophy at The University of Sydney. This will take place under the supervision of Dr Ernest Ekpo, Academic Fellow/Lecturer.

(3) What will the study involve for me?

As a participant in this study you will be asked to interpret 60 periapical and 120 cone beam computed tomography images, spread over two reading sessions. Therefore, you will perform a total of 180 image interpretations for this study. It should take two hours to complete each reading session. At the beginning of the study, we will ask you some demographic questions relating to your experience as a radiologist or dentist. You will have the opportunity to review your reports and receive feedback on your diagnostic performance.

(4) How much of my time will the study take?

We anticipate that it would take you approximately two hours to complete a session of reading. There are two reading sessions, which will be conducted approximately three months apart. Therefore, this study will take approximately four hours of your time.

(5) Who can take part in the study?

Participation in the study is restricted to dentists and radiologists who report digital 2D periapical radiography and 3D cone beam computed tomography images. This restriction is to avoid the inclusion of individuals who do not have expertise in these modalities, which could bias the results of the study.

(6) Do I have to be in the study? Can I withdraw from the study once I've started?

Being in this study is completely voluntary and you do not have to take part. Your decision whether to participate will not affect your current or future relationship with the researchers or anyone else at the University of Sydney.

If you decide to take part in the study and then change your mind later, you are free to withdraw at any time. You can do this by contacting the Chief Investigator, Dr Ernest Ekpo or PhD student, Dr Kehn Yapp and request the identification and deletion of your reports based on the email you have provided. Your reports will therefore be removed from any further recordkeeping and will not be included in the analysis of results. There are no consequences for withdrawing from the study.

(7) Are there any risks or costs associated with being in the study?

Aside from giving up your time, we do not expect that there will be any risks or costs associated with taking part in this study.

(8) Are there any benefits associated with being in the study?

We cannot guarantee that you will receive any direct benefits from being in the study.

(9) What will happen to information about me that is collected during the study?

For this study, we will collect demographic information such as age and gender and practice-related information such as training (dentist or radiologist), years since qualification, number of images read per week. We will also collect the scores of your image interpretation performance (sensitivity, specificity, accuracy, and false positive rates).

By providing your consent, you are agreeing to us collecting personal information about you for the purposes of this research study. Your information will only be used for the purposes outlined in this Participant Information Statement, unless you consent otherwise.
Your information will be stored securely, and your identity/information will be kept strictly confidential, except as required by law. Study findings may be published, but you will not be individually identifiable in these publications.

We will keep the information we collect for this study, and we may use it in future projects. By providing your consent you are allowing us to use your information in future projects. We don't know at this stage what these other projects will involve. We will seek ethical approval before using the information in these future projects.

We intend to submit the information from this project to a public database for research information, so that other researchers can access it and use it in their projects. Before we do so, we will take out all the identifying information so that the people we give it to won't know whose information it is. They won't know that you participated in the project and they won't be able to link you to any of the information you provided.

(10) Can I tell other people about the study?

Yes, you are welcome to tell other people about the study.

(11) What if I would like further information about the study?

When you have read this information, Dr Kehn Yapp who will be available at the time of consent will be available to discuss it with you further and answer any questions you may have. If you would like to know more at any stage during the study, please feel free to contact **Dr Ernest Ekpo (ernest.ekpo@sydney.edu.au)** or **Dr Kehn Yapp (keyapptdo@gmail.com)**

(12) Will I be told the results of the study?

You have a right to receive feedback about the overall results of this study. You can tell us that you wish to receive feedback by ticking the relevant box on the consent form. The study is designed such that when you complete the second reading session, you will be allowed to review your reports and receive feedback if you wish. This feedback will be in the form of a lay summary of the findings.

(13) What if I have a complaint or any concerns about the study?

Research involving humans in Australia is reviewed by an independent group of people called a Human Research Ethics Committee (HREC). The ethical aspects of this study have been approved by the HREC of the University of Sydney *[**protocol no – 2020/477**]*. As part of this process, we have agreed to carry out the study according to the *National Statement on Ethical Conduct in Human Research (2007).* This statement has been developed to protect people who agree to take part in research studies.

If you are concerned about the way this study is being conducted or you wish to make a complaint to someone independent from the study, please contact the university using the details outlined below. Please quote the study title and protocol number.

The Manager, Ethics Administration, University of Sydney:

- **Telephone:** +61 2 8627 8176
- **Email:** human.ethics@sydney.edu.au
- **Fax:** +61 2 8627 8177 (Facsimile)

This information sheet is for you to keep

Appendix B: Ethics approval

THE UNIVERSITY OF SYDNEY

Research Integrity & Ethics Administration
HUMAN RESEARCH ETHICS COMMITTEE

Tuesday, 18 August 2020

Dr Ernest Ekpo
Medical Radiation Sciences; Faculty of Medicine and Health
Email: ernest.ekpo@sydney.edu.au

Dear Ernest,

The University of Sydney Human Research Ethics Committee (HREC) has considered your application.

I am pleased to inform you that after consideration of your response, your project has been approved.

Details of the approval are as follows:

Project No.:	**2020/477**
Project Title:	**Diagnostic efficacy of dental digital 2D periapical radiography and 3D cone beam computed tomography**
Authorised Personnel:	**Ekpo Ernest; Brennan Patrick; Yapp Kehn;**
Approval Period:	**18 August 2020 to 18 August 2024**
First Annual Report Due:	**18 August 2021**

Documents Approved:

Date Uploaded	Version Number	Document Name
10/07/2020	Version 1	Recruitment flyer
21/05/2020	Version 1	Participant consent form
10/07/2020	Version 2	Participant information statement

Condition/s of Approval

- Research must be conducted according to the approved proposal.
- An annual progress report must be submitted to the Ethics Office on or before the anniversary of approval and on completion of the project.
- You must report as soon as practicable anything that might warrant review of ethical approval of the project including:
  - Serious or unexpected adverse events (which should be reported within 72 hours).
  - Unforeseen events that might affect continued ethical acceptability of the project.
- Any changes to the proposal must be approved prior to their implementation (except where an amendment is undertaken to eliminate *immediate* risk to participants).
- Personnel working on this project must be sufficiently qualified by education, training and experience for their role, or adequately supervised. Changes to personnel must be reported and approved.
- Personnel must disclose any actual or potential conflicts of interest, including any financial or other interest or affiliation, as relevant to this project.
- Data and primary materials must be retained and stored in accordance with the relevant legislation and University guidelines.

- Ethics approval is dependent upon ongoing compliance of the research with the *National Statement on Ethical Conduct in Human Research*, the *Australian Code for the Responsible Conduct of Research*, applicable legal requirements, and with University policies, procedures and governance requirements.
- The Ethics Office may conduct audits on approved projects.
- The Chief Investigator has ultimate responsibility for the conduct of the research and is responsible for ensuring all others involved will conduct the research in accordance with the above.

This letter constitutes ethical approval only.

Please contact the Ethics Office should you require further information or clarification.

Sincerely,

Dr Haryana Dillon
Chair
Human Research Ethics Committee (HREC 3)

The University of Sydney of Sydney HRECs are constituted and operate in accordance with the National Health and Medical Research Council's (NHMRC) National Statement on Ethical Conduct in Human Research (2018) and the NHMRC's Australian Code for the Responsible Conduct of Research (2018)