The role of the free-response receiver operating characteristic method for dose and image quality optimisation

An evaluation of low-resolution computed tomography

John D. Thompson

Directorate of Radiography

University of Salford

2014

Thesis for the Degree of Doctor of Philosophy in Radiography

PhD by Published Works

Directorate of Radiography

School of Health Sciences

University of Salford

Greater Manchester

M6 6PU

For Kirsty, Harry and Olivia

Without the unwavering support, patience and understanding of my loving family this PhD would not have been possible. You will get me back soon!

Student Declaration

I hereby declare that **John David Thompson** has completed this thesis under the support and guidance of a dedicated team of supervisors at the University of Salford for the award of Doctor of Philosophy through the route of published works.

I have maintained confidentiality of participants throughout the research and I have correctly and adequately acknowledged significant contribution and made reference to supporting texts.

Signature _____ Date / /

Contents

List of Tables

List of Figures

Acknowledgements

Professor Peter Hogg

Lead PhD Supervisor; Professor of Radiography, University of Salford

Peter has provided unparalleled support, guidance and mentorship throughout my PhD by Published Works. His dedication to my learning has had an immeasurable impact on the development of my research skills.

Professor David J. Manning

PhD Co-Supervisor; Professor of Radiography, Lancaster University, University of Salford

David has been a mentor to me as I have been trying to establish myself in observer performance. His expert advice and guidance in all aspects of observer performance and visual perception has been invaluable to my development as an academic.

Katy Szczepura

PhD Co-Supervisor; Medical Physics Lecturer, University of Salford

Katy's knowledge of imaging system characteristics and her technical understanding of imaging systems has been a valuable influence. Katy has provided me with the necessary support and encouragement required to complete the research projects for my PhD.

Stephen Thompson

Independent Software Developer; Bury St. Edmunds

Stephen has invested a vast amount of time in the development and programming of the ROCView software that has been integral to this body of work; without these skills and dedication this PhD would not have been possible.

Professor Dev Chakraborty

Professor of Radiology; University of Pittsburgh, PA, USA

Dev has offered me expert advice and technical assistance with many aspects of performing free-response studies. Dev has been integral in the development of my analytical skills while also helping develop my academic writing style. Two periods of work with Professor Chakraborty (in-person) have been invaluable to my development and consolidation of my understanding of ROC/FROC methodology.

Professor Richard Lawson

Honorary Consultant Medical Physicist; Central Manchester Nuclear Medicine Centre. Honorary Professor; University of Salford. Honorary Senior Lecturer; University of Manchester.

Richard has been outstanding in providing me technical advice about the functionality of hybrid imaging systems. Richard has also given me regular advice on statistics to aid my understanding of data analysis and improve my reporting of results.

Dr Julie Nightingale

PhD Supervisor (Registration Period) Director of Radiography and Occupational Therapy, University of Salford

Julie has provided me with valuable guidance in the final stages of preparing this thesis. Her insight into citation analysis has been particularly valuable.

Abstract

This thesis describes the value of the free-response receiver operating characteristic (FROC) paradigm for dose and image quality optimisation in a niche area of imaging. The empirical works discussed in this thesis focus on the diagnostic value of the low-resolution computed tomography (CT) images acquired for attenuation correction (AC) – a process primarily used to correct for photon attenuation with images produced merely as a consequence of the exposure. The potential discovery of incidental findings on these images was investigated.

The observers taking part in the empirical studies were generally lacking in significant experience of interpreting CT images. As a consequence it was also deemed valuable to investigate the value of the novice observer in free-response studies. A further methodological consideration for studies of this kind is consistent and reliable image display and FROC data collection. Prototype software, ROCView, was designed and developed to make this an easy process and the key functionality and impact is analysed here.

In addition to the empirical works, two review papers, aimed at the technologists and radiographers performing low-resolution CT for AC, are summarised. They explain the value of the FROC paradigm and the jackkinfe alternative FROC (JAFROC) analysis method to a wide audience in nuclear medicine.

Abbreviations and Definitions

AC	Attenuation Correction *A method to correct for photon attenuation induced artefacts.*
AFROC	Alternative Free-response Receiver Operating Characteristic *A hybrid curve where the trapezoidal area beneath the curve is equivalent to the JAFROC figure-of-merit.*
ALARP	As Low As Reasonably Practicable *A principle often used to describe dose optimisation.*
AR	Acceptance Radius *A tool used to classify observer localisations as correct/incorrect.*
AUC	Area Under the ROC Curve *A summary index by which different diagnostic tests can be compared.*
*p*AUC	Partial Area Under the ROC Curve *A summary index useful for evaluations of high sensitivity or specificity.*
CD	Contrast Detail *(Phantom) Used to provide a simplified observer test using a signal known exactly/background known exactly method.*
CT	Computed Tomography *Imaging modality evaluated in this thesis.*
CTDI	Computed Tomography Dose Index *An estimate of the dose delivered in CT.*
DICOM	Digital Imaging and Communications in Medicine *Standard image format used in radiology.*
FPF	False Positive Fraction *1-specificity.*
FROC	Free-response Receiver Operating Characteristic *The method used to capture free-response data.*
HU	Hounsfield Unit *(Scale) A transformation of linear attenuation coefficient measurements where water is defined as zero (0) and air is defined as -1000.*
i, j, k	Modality, Reader, Case *Common denotation in ROC/FROC equations.*
JAFROC	Jackknife Alternative Free-response Receiver Operating Characteristic *The figure of merit most frequently associated with free-response methodology. It is the probability that a lesion rating exceeds any rating on normal cases.*
LD-LCS	Low-dose Lung Cancer Screening *Using low-dose CT acquisitions in a screening programme.*
LL	Lesion Localisation *A correct localisation of disease on an abnormal case.*

LLF	Lesion Localisation Fraction
	The number of lesions correctly localised divided by the total number of lesions, at each threshold.
MPI	Myocardial Perfusion Imaging
	A tomographic imaging procedure to reveal coronary artery disease. Images are frequently corrected with CT-based attenuation correction to account for photon attenuation artefacts.
MRMC	Multiple-reader Multiple-case
	Defined as common set of observers interpreting a set of images common to all modalities being evaluated.
NL	Non-lesion Localisation
	An incorrect localisation on normal or abnormal cases.
NLF	Non-lesion Localisation Fraction
	The number of non-lesion localisations divided by the number of cases.
ROC	Receiver Operating Characteristic
	A method for measuring the performance of observers and imaging modalities.
SDT	Signal Detection Theory
	Used to analyse the detection and discrimination process of an observer seeing a signal.
SPECT/CT	Single Photon Emission Computed Tomography/Computed Tomography
	A hybrid imaging modality combining functional and anatomical imaging.
TPF	True Positive Fraction
	Sensitivity.

Publications

This thesis is based on the following published papers. Papers are listed in order of their publication date:

1. Thompson JD, Hogg P, Thompson SM, Manning DJ, Szczepura K. ROCView: prototype software for data collection in jackknife alternative free-response receiver operating characteristic analysis. *The British Journal of Radiology* 2012;85:1320-1326.
2. Thompson J, Hogg P, Szczepura K, Manning D. Analysis of CT acquisition parameters suitable for use in SPECT/CT: A free-response receiver operating characteristic study. *Radiography* 2012;18:238-243.
3. Thompson J, Hogg P, Higham S, Manning D. Accurate localisation of incidental findings on the computed tomography attenuation correction image: the influence of tube current variation. *Nuclear Medicine Communications* 2013;34:180-184.
4. Thompson JD, Manning DJ, Hogg P. The value of observer performance studies in dose optimisation: A focus on free-response receiver operating characteristic methods. *Journal of Nuclear Medicine Technology* 2013;41:57-64.
5. Thompson JD, Hogg P, Manning DJ, Szczepura K, Chakraborty DP. A Free-response Evaluation Determining Value in the Computed Tomography Attenuation Correction Image for Revealing Pulmonary Incidental Findings: A Phantom Study. *Academic Radiology* 2014;21:538-545.
6. Thompson JD, Manning DJ, Hogg P. Analysing data from observer studies in medical imaging research: an introductory guide to free-response techniques. *Radiography* 2014;20:295-299.
7. Buissink C, Thompson JD, Voet M, Sanderud A, Kamping LV, Savary L, Mughal M, Rocha CS, Hart GE, Parreiral R, Martin G, Hogg P. The influence of observer training in a group of novice observers: a jackknife alternative free-response receiver operating characteristic analysis. *Radiography* 2014;20:300-305.

The papers will be referred to in the text using the numeric value (1-7). The full papers are presented in Appendices A-G.

Preliminary Reports

In addition to the published articles listed above, the research undertaken as part of this PhD by Published Works has been disseminated both orally and as poster presentations within conference proceedings and research seminars.

Conference Proceedings

Szczepura, K. Thompson, J. Tootell, A. Driver, J. Manning, D. Hogg, P. An analysis of phantom image data to determine optimal CT exposure factors for use in SPECT/CT. British Nuclear Medicine Society: Harrogate. Nuclear Medicine Communications 2010;31:463.

Thompson, J. Szczepura, K. Tootell, A. Sil, J. Manning, D. Hogg, P. An analysis of phantom image data to determine optimal CT exposure factors for use in SPECT/CT. World Federation of Nuclear Medicine and Biology: Cape Town. World Journal of Nuclear Medicine 2010;9;1:S158.

Thompson, J. Hogg, P. Szczepura, K. Tootell, A. Sil, J. Manning, D. Determination of optimal CT exposure factors for lung lesions using an anthropomorphic chest phantom for SPECT-CT. European Association of Nuclear Medicine: Vienna. European Journal of Nuclear Medicine and Molecular Imaging, 2010;37;Suppl 2:S494.

Thompson, J. Hogg, P. Thompson, S. ROCView: prototype software moving toward easier data collection in JAFROC analysis. Proceedings of the UK Radiological Congress (UKRC): Manchester. 2011:58 http://bjr.birjournals.org/site/misc/Proceed_2011.pdf

Thompson, J. Szczepura, K. Manning, D. Hogg, P. Lesion detection in the CT attenuation correction image of 5 different low resolution SPECT/CT systems: a multi-centre study. British Nuclear Medicine Society: Harrogate. Nuclear Medicine Communications 2012;33(5):548.

Thompson, J. Higham, S. Hogg, P. Manning, D. Szczepura, K. The impact of tube current variation on lesion detection in the attenuation correction image co-incidentally acquired for myocardial perfusion imaging in SPECT/CT: a phantom based study. Proceedings of the UK Radiological Congress (UKRC) 2012, Manchester

Thompson, J. Szczepura, K. Manning, D. Hogg, P. Lesion detection in the CT attenuation correction image of 5 different low resolution SPECT/CT systems: a multi-centre study. Proceedings of the UK Radiological Congress (UKRC) 2012, Manchester

Thompson, J. Hogg, P. Manning, D. Szczepura, K. Chakraborty, D. Lesion detection in the CT attenuation correction image of 5 SPECT/CT systems: a multi-centre study. European Association of Nuclear Medicine: Milan. European Journal of Nuclear Medicine and Molecular Imaging 2012;39(Suppl 2):625.

Thompson, J. Hogg, P. Higham, S. Manning, D. The impact of tube current variation on lesion detection in the attenuation correction image acquired for myocardial perfusion SPECT/CT: a phantom based study. European Association of Nuclear Medicine: Milan. European Journal of Nuclear Medicine and Molecular Imaging 2012;39(Suppl 2);625.

Buissink C, Thompson J, Voet M, Sanderud A, Kamping LV. Savary L, Mughal M, Rocha CS, Hart GE, Parreiral R, Martin G, Hogg P. The influence of observer training in the detection of pulmonary lesions in chest phantom single CT images: a JAFROC analysis. Portuguese Association of Radiotherapy, Radiology and Nuclear Medicine (ATARP) 2013, Lisbon.

Buissink C, Thompson JD, Voet M, Sanderud A, Kamping LV, Mughal M, Rocha CS, Hart GE, Parreiral R, Martin G, Hogg P. The influence of experience and observer training for the detection of simulated pulmonary lesions: a jackknife alternative free-response receiver operating characteristic analysis: Proceedings of the UK Radiological Congress (UKRC) 2014: Manchester.

Intellectual Ownership and Contribution

The intellectual ownership and type and percentage contribution of all co-authors for each paper (1-7) included in this PhD thesis are displayed in Table 1. The type of contribution is summarised as:

a) Concept and Design
b) Data Collection
c) Data Analysis
d) Drafting and Revision
e) Final Approval

This has been based on a subset of categories for authorship as recommended by the International Committee of Medical Journal Editors (International Committee of Medical Journal Editors, n.d.). However, contributors were not denied the opportunity to become authors if they did not fulfil each of the criteria. All co-authors listed had the opportunity to review the final submission.

Authors	Papers / Contribution (%) and Type						
	1	2	3	4	5	6	7
JDT	47.5 a d e	62.5 a b c d e	55.0 a b c d e	60.0 a d e	45.0 a b c d e	65.0 a d e	60.0 a b c d e
PH	12.5 a d e	17.5 a c d e	15.0 a b c d e	15.0 a d e	15.0 a b c d e	10.0 a d e	10.0 a b c d e
SMT	17.5 a d e						
DJM	12.5 a d e	10.0 a c d e	15.0 a c d e	25.0 a d e	12.5 c d e	25.0 a d e	
KS	10.0 a d e	10.0 a c d e			12.5 c d e		
SH			15.0 a b c d				
DPC					15.0 c d e		
CB							10.0 a b d e
MV							5.0 b c
AS							5.0 b c
et al							10.0 c d

Table 1: The contribution of each author to the papers included in the PhD by Published Works Thesis.

Substantiated Author Contributions

The contribution of all those eligible for authorship has been recognised by John David Thompson in all the papers included in this PhD by Published Works Thesis.

- No eligible author has been denied authorship of the opportunity to contribute
- No ineligible author has been included on any publication
- Where appropriate, acknowledgements have been made to participants who do not satisfy enough criteria to be considered a co-author
- The contribution and ownership displayed in Table 1 is accurate

I hereby declare that the above statements have been satisfied. I sign to acknowledge the contribution of all authors[1] in accordance with the University of Salford Code of Conduct:

Signed:

Professor Peter Hogg, Associate Head of Research, University of Salford

Lead PhD Supervisor and Co-author

[1] It was no longer possible to make contact with all co-authors. Author contribution substantiated by the lead PhD supervisor who was a co-author on all papers.

Citation Analysis

Citation analysis is a relevant activity for a thesis drawing on published work. The number of citations per article can be a good indication of the impact of the work to other researchers but it does not necessarily correlate with the quality of the paper (Nightingale & Marshall, 2012). Furthermore, the work discusses a niche area of imaging and prospective nature of the work means that the work has not been in press long enough to generate a large number of citations. Consequently it is also important to consider other relevant metrics, such as abstract and full-text views and downloads. Citations are usually considered to originate from journal articles but can appear in a wide range of sources (Nightingale & Marshall, 2012). The impact of the published work is demonstrated in Table 2.

Paper	1	2	3	4	5	6	7
Impact Factor	1.217	-	1.379	-	1.914	-	-
Google Scholar Citations	6 (5)	3 (2)	3 (2)	2 (0)	2 (0)	1 (0)	1 (0)
Scopus Citations	6 (5)	3 (2)	3 (2)	0	2 (0)	1 (0)	1 (0)
Tweets	0	0	0	1	0	0	0
Mendeley Saves	9	1	0	7	3	0	0
Research Gate Citations	3 (3)	0	2 (1)	0	0	0	0
Views	128	29	68	49	60	8	18
Downloads	0	0	0	0	27[2]	0	6

Table 2: Citation analysis collated from Google Scholar, Scopus and Research Gate. Note the variation in the number of citations from different sources. Scopus also includes a social media output; in each case the publisher made the Tweet. Date of analysis: 16/12/2014. Self-citations in parentheses.

[2] Paper 5 is freely available to download due to NIH Public Access.

Ethics in Observer Performance Research

Ethical Principles in Medical Informatics

For medical research the role of human participants in research is summarised by the Declaration of Helsinki (World Medical Association, 2013). The interests and wellbeing of the research participants must be the primary concern during the research, and while this document refers to participants as patients, the same standards should be applied to those who participate as observers.

In terms of medical informatics the systems in place should support research, protect from harm and maintain confidentiality (Duquenoy, George, & Solomonides, 2008). This has some application to the current body of work. Broadly speaking, ethical issues serve to protect the interests and integrity of individuals and uphold standards. In the observer studies completed for this PhD by PW it was essential to maintain the confidentiality of all the research participants, making sure that any recorded data was not identifiable to them. Ethical standards in relation to electronic data and records, based on the code of ethics, have been proposed (Duquenoy et al., 2008):

- Security
- Integrity
- Material Quality
- Usability
- Accessibility

Data Protection Act 1998

The Data Protection Act (DPA) 1998 (Legislation.gov.uk, 1998) is applicable to all observer studies. All data processing should be ‘adequate, relevant and not excessive’. It should therefore not exceed what is required for a reliable result to be formed, unless there are other components to the research. Participants must be given appropriate information to enable them to make a reasoned decision about taking part in a study. Consent may be required, but explicit (written) consent is only required for sensitive data.

All participants should also be made aware of the purpose of the data collection, how long the data will be kept and what will be done with the data when it has been collected. For all empirical works reported in this thesis, participants were told that their data would be used in a research project with the intention of publication; it was made clear to all participants that

their data would be destroyed (electronic record deleted) if they decided to withdraw from the study – which they could do at any time. Identifiable data was kept in a password-protected file, to ensure that there was no unauthorised access of data.

Ethical Considerations: Phantom versus Patient

Working with an anthropomorphic chest phantom has many obvious benefits as well as drawbacks in comparison to imaging patients. A recent editorial highlights the ethical issues of reviewing a dose optimisation study that has been conducted on patients (Achenbach, Chandrashekhar, & Narula, 2013). The main concern was the repeated exposure of patients to illustrate that dose and image quality could be optimised in computed tomography. It is clear in this example that the research team and institutional review board did not consider non-maleficence (do no harm), and are in breach of one of the core ethical principles. While this editorial's main concern was the ethics in research in publishing, it also adds support to the ethical justification of using an anthropomorphic chest phantom to investigate dose optimisation. Knowledge of true disease status is required for observer studies. Although this is not strictly an ethical issue, it is relevant to the argument between patient and phantom based studies. Ascertaining this 'truth' can be problematic in patients and recent work cites this as a problem in observer performance studies (Kundel, 2006). In phantom work, this is much easier to control as the investigators are often responsible for the placements of 'disease'.

Acquiring Observer Performance Data at the University of Salford

The overwhelming majority of the observer studies were conducted within the University of Salford medical imaging facility. All articles within this PhD by PW were subject to the same high standards of ethical requirements, as they were constituted within the University of Salford's ethical framework. All participants were consented prior to participating; this included giving a detailed explanation of what was required through use of an information sheet and/or an explanatory PowerPoint presentation. Data were stored in an anonymous fashion on a password-protected computer. Participants had the option to receive feedback on their performance and most took up this opportunity. Observers received a data-sheet that could be used in a CPD portfolio; identifying their individual and observer averaged performance. Most volunteers benefitted from 2-4 hours education on FROC methodologies.

Introduction

Observer performance has been applied in radiology to assess the diagnostic performance of imaging systems and techniques. This type of assessment method is useful for measuring diagnostic performance when the observer is considered to be an integral component of the imaging system. These methods have become popular and have particular value in comparing the performance of existing and new imaging techniques, where the aim is to establish if the new technique offers any statistically significant advantage over the current gold-standard or best current alternative.

The receiver operating characteristic (ROC) method has been described extensively (Chesters, 1992; Hanley & McNeil, 1983; Hanley, 1989; Kundel, 2006; Metz, 1978; Swets, 1996) and for many years was the method of choice for observer studies assessing a single pathology. The ROC method was a natural successor to signal detection theory (SDT) and the original application was in radio detection and ranging (radar) (Manning, 1998). The first use of ROC in radiology occurred in the 1960's (Lusted, 1960) and has since been used in a large number of laboratory studies. However, the ROC method is limited by not taking into account all the available information that an image contains – specifically it does not deal well with multiple pathologies and does not discriminate on the basis of location, where only a single rating of confidence can be obtained for each image.

The free-response ROC (FROC) paradigm was developed to overcome these problems. This method is location sensitive and requires observers to localise all suspicious areas of an image with precision, and provide a confidence rating for each localisation; the earliest implementation of this paradigm was reported in the 1990's (Chakraborty & Winter, 1990). The key methodological aspects of FROC have double effect; not only is this method more reflective of the clinical task but it also holds a statistical advantage over traditional ROC methods (Krupinski & Jiang, 2008).

The development of the FROC paradigm and the suitability of it to this research theme have been explored in the published works and will be critically examined in this thesis.

Rationale

Advancing technology and new techniques ensure the important role of radiology in the diagnosis and management of disease. The non-invasive nature of many imaging tests can be preferential for patients and referrers in comparison to interventional or surgical procedures, despite the accompanying radiation risk.

The demand for radiological procedures is ever increasing, with the collective dose of computed tomography (CT) increasing by nearly 30% in the UK over a 10-year period (Hart, Wall, Hillier, & Shrimpton, 2010). There are persisting concerns about the dose associated with CT and the contribution of CT to the total dose received by a single patient (Hara et al., 2009). This is evident in the United Kingdom, where CT examinations are responsible for the highest dose in terms of the total radiation dose delivered (Hart et al., 2010). Additionally, CT is the only diagnostic modality where the dose has increased in recent years (Dawson, 2004). In the United States of America the increasing use of CT has been attributed to an increased role in paediatric diagnosis and adult screening, where reduced scanning time has been cited as major contributing factors (Brenner & Hall, 2007).

The other side to the increasing use of CT is the potential clinical benefit. It has been reported that a significant increase in the use of CT in emergency care has improved clinical decision making. In particular, patients with abdominal pain are more likely to receive a CT scan, with an overall reduction in the number of hospitalizations and transfers (Kocher et al., 2011).

The use of CT has also developed rapidly in nuclear medicine. The initial use of CT in nuclear medicine was to provide attenuation correction (AC) during myocardial perfusion imaging (MPI) (Bateman & Cullom, 2005; Malkerneker et al., 2007; Pazhenkottil et al., 2011). The CT images are acquired as a high photon flux (good statistically in comparison to radioactive line sources) transmission map, giving a good representation of tissue attenuation. Large volumes of individual projections are acquired to provide a transmission measurement of each slice of tissue in the desired volume of the body (Patton & Turkington, 2008). This can then be used to correct for photon attenuation of radioactive isotopes by body structures that are apparent in single photon emission computed tomography (SPECT), which is the tomographic imaging mode used in nuclear medicine where planar images are acquired at multiple angles around the body area of interest. This is typically 60 planar images at 3° increments over 180° for cardiac imaging (Patton & Turkington, 2008).

The transmission imaging provided by AC had previously been conducted using a moving radioactive line source such as Gadolinium, ^{153}Gd, or Germanium, ^{68}Ge; and while this improved the specificity of the technique (Gallowitsch et al., 1998) it is associated with several drawbacks in comparison to CT-based transmission imaging for AC. These include reduced photon flux, long acquisition times, the need to replace the radioactive line source and cross-talk of radiopharmaceutical energies (Zaidi & Hasegawa, 2003). CT-based AC is also higher resolution and can provide anatomical detail, but is also associated with a higher dose than ^{68}Ge in PET imaging (Wu et al., 2004).

The use of CT for AC in conjunction with SPECT and positron emission tomography (PET) has led to significant improvements in MPI for the diagnosis of coronary artery disease (CAD) (Gaemperli, Bengel, & Kaufmann, 2011; Kaufmann, 2009). AC is typically performed when images are degraded as a result of patient dependant, non-uniform and occasionally focal, photon attenuation. In males this occurs because of the diaphragm and in females because of breast tissue (Burrell & MacDonald, 2006). The CT data provide a high quality map of attenuation coefficients that are used to correct for radionuclide photon attenuation in SPECT and PET (Koepfli et al., 2004; Pazhenkottil et al., 2011).

Beyond AC, CT is now used routinely for localisation of radioactive foci, with particular success in cases of specific tracer uptake and also for diagnosis (Bockisch, Freudenberg, Schmidt, & Kuwert, 2009), where PET/CT has embraced this sooner and more universally than SPECT/CT.

Despite the existence of three clear roles for CT in nuclear medicine, the ‘quality’ of CT required to accomplish each task is less well defined and there is likely to be overlap between the quality/dose of CT required for each role. To amplify this overlap, there is also great variation in the capability of the CT systems used to perform AC. The primary focus of the empirical work reported in this thesis was to evaluate the potential clinical value of the low-resolution CT image that is produced during AC in a range of systems.

The clinical value of these images is an evolving problem. Early SPECT/CT systems used basic CT units designed to provide AC. The quality of the CT images produced was poor, and in comparison to a conventional CT image they are likely to have been considered non-diagnostic. Despite this, some previous work has been completed to assess the diagnostic value of these low-resolution CT images. Using an early SPECT/CT system, the low-resolution CT images of 200 consecutive patients undergoing MPI with AC were evaluated for incidental findings. Large numbers (234 abnormalities) were found, with some considered of major clinical significance that were previously unknown to the patient (Goetze, Pannu, & Wahl, 2006). A large number of abnormalities were reported in and around the thorax.

CT technology in nuclear medicine has improved since the work of Goetze et al (2006) was completed and the images produced during CT-based AC are improving in quality. This may have prompted a shift in the perception of the value of these images. The Ionising Radiation (Medical Exposure) Regulations, 2000, Regulation 7 require that a clinical interpretation is completed for each exposure (Department of Health, 2000) – this implies that the CT images produced during AC should be reported. This has recently gained some support, with recommendations from The British Nuclear Medicine Society now recommending that CT images should be reported (Arumugam, Harbinson, Reyes, Sabharwal, & Tonge, 2012). However, they limit this to images considered 'diagnostic quality' and appear to exclude 'low-dose' or 'low-resolution' images from this recommendation. The work presented in this thesis tests the strength of this guidance by assessing a range of SPECT/CT systems, with different CT unit capability, using the latest observer performance methods.

In the context of dose optimisation, this guidance, with the knowledge of the potential discovery of incidental findings, has led to some difficult and controversial decisions regarding the quality of the CT exposure used for AC. The trade-off concerns the extra information that may be gained by performing a higher dose scan against the original clinical question and the primary reason for the exposure – AC. This is further complicated by the fact that an adequate attenuation map (mu; μ) can be created, irrespective of the exposure parameters used (Kamel et al., 2002; Preuss et al., 2008; Wells, Soueidan, Vanderwerf, & Ruddya, 2012; Xia, Alessio, & Kinahan, 2009).

The advantage of using a higher-dose (if not diagnostic) CT acquisition for AC with the premise of detecting incidental findings in MPI is unclear. Dose optimisation is a key aspect of radiology and all examinations should be dose efficient, conforming to the as low as reasonably practicable (ALARP) principle.

Objectives

The empirical work evaluated in this thesis was completed using the free-response receiver operating characteristic paradigm and the latest analysis methods that are associated with these observer performance methods. This thesis will discuss the suitability of these methods for the research reported and evaluated in this thesis. The research was facilitated by prototype software that allowed multiple concurrent users to complete image evaluations in any location since it is run from a web-based platform. The key functionality and suitability of the software will be critically evaluated.

The focus of the empirical works was to evaluate the diagnostic value and potential optimisation of radiation dose of the low-resolution CT images acquired during AC. The observer task in this research was often completed by participants that would be considered novice, in part because of the difficulties in accessing large numbers of expert observers. The publications reported in this thesis are critically evaluated for their potential impact and develop the following objectives for critical evaluation:

1. Examine the fundamental principles of measuring observer performance with a focus on the free-response paradigm.

2. Develop a consistent and reliable method for image display and response capture in free-response studies.

3. Assess the potential for lesion detection and dose and image quality optimisation in a range of CT and SPECT/CT systems.

4. Assess the role of the novice observer for suitability in observer performance research.

Objective 1

Examine the fundamental principles of measuring observer performance with a focus on the free-response paradigm.

In order to proceed with the research it was necessary to develop a strong underpinning knowledge of observer performance, with a particular emphasis on the methods and analysis of data. This review begins with a brief overview of signal detection theory before making the natural step on to the ROC and FROC paradigms.

The following topics will be explored:

- Signal Detection Theory (SDT)
- Receiver Operating Characteristic (ROC) Analysis
- ROC Rating (Confidence) Scales
- ROC Curves
- Multi-Reader Multi-Case Analysis

To proceed with the investigation of low-resolution CT using the proposed method it was first important to develop a broad underpinning knowledge of observer performance. This begins by reviewing the development of receiver operating characteristic methods, understanding the limitations and embracing the recent developments in the methodology that have led to a statistically powerful and clinically relevant method of analysis.

The text will look at the key methodological aspects of the free-response paradigm and explore the empirical construction of the curves associated with free-response methods and the figures of merit that are used in statistical testing. Two key papers (4 and 6) have demonstrated and supported this process. The choice of journals for publication was vital, since the aim was to raise awareness of, and explain, the important role that observer performance plays in the evaluation of imaging techniques to two key groups: Radiographers and Nuclear Medicine Technologists.

Signal Detection Theory

For many years signal detection theory (SDT) has been used as a theoretical framework to understand decision making under uncertainty (Jang, Wixted, & Huber, 2009). SDT is used

to analyse the detection and discrimination process of an observer seeing a signal (i.e. in a background of increasing noise), while attempting to reveal the decision criterion (threshold) adopted by the observer during the detection process (Jäkel & Wichmann, 2006).

SDT research began with an application of the Rose model using a contrast detail (CD) phantom in fluoroscopy (Burgess, 2010). This was the first attempt at dose and image quality optimisation in radiology, where the CD phantom was used to assess observer performance. Albert Rose developed the absolute scale of quantum efficiency to evaluate system performance and what is now known as the Rose model to assess signal detectability in human observers (Burgess, 2010). These methods have maintained their popularity in medical imaging. Applications of detectability of sharp-edged objects in a uniform background of changing noise are still being used in quality control for fluoroscopy, radiography, mammography (Kotre, 2006) and nuclear medicine (Rzeszotarski, 1999), Figure 1. The type of test described is commonly referred to as a signal known exactly / background known exactly (SKE/BKE) test (Burgess, 2010; Kotre, 2006).

In the following example, the observer is tasked with deciding which size and density of sharp-edged object can be detected. Once this has been completed for a range of object sizes and levels of image noise, contrast detail curves can be plotted to show detection rates at different levels of contrast and object sizes, Figure 2. This example reveals that reduced counting statistics in nuclear medicine have a negative impact on object detectability due to a loss of spatial frequency. Theory supports the findings that noise impairs image interpretation (Manning, 1998).

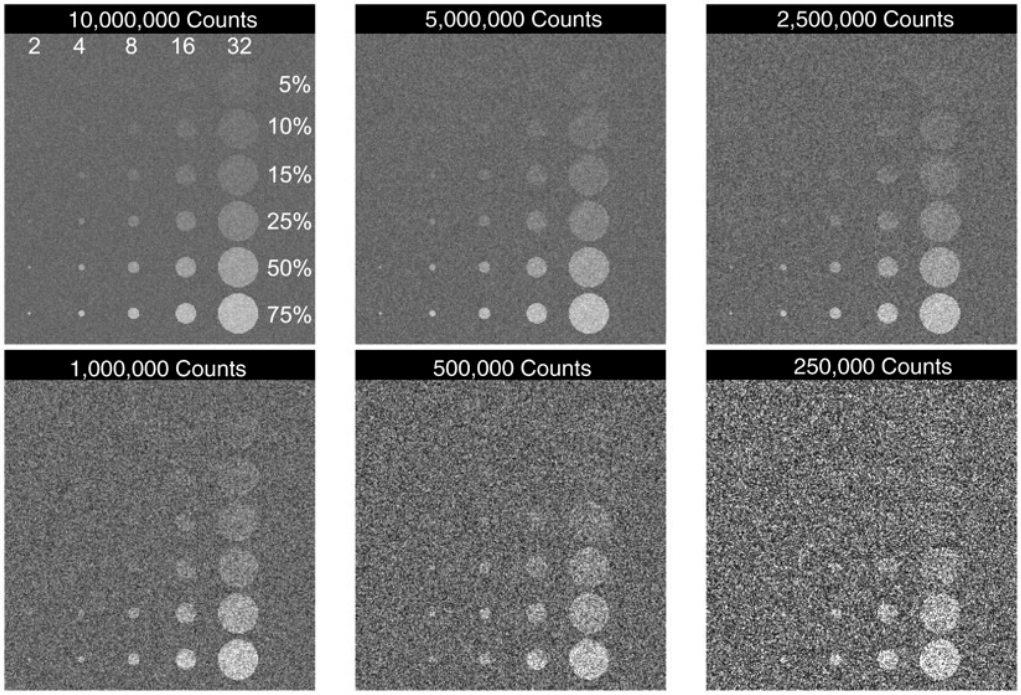

Figure 1: Contrast detail images produced by a planar image acquisition on a gamma camera in nuclear medicine. The Rose model predicts the line of demarcation for object detection; a diagonal line from the lower left to upper right of each image – objects above this line are barely detectable The level of image noise dictates the position of the diagonal line. Image reproduced from Rzeszotarski (1999).

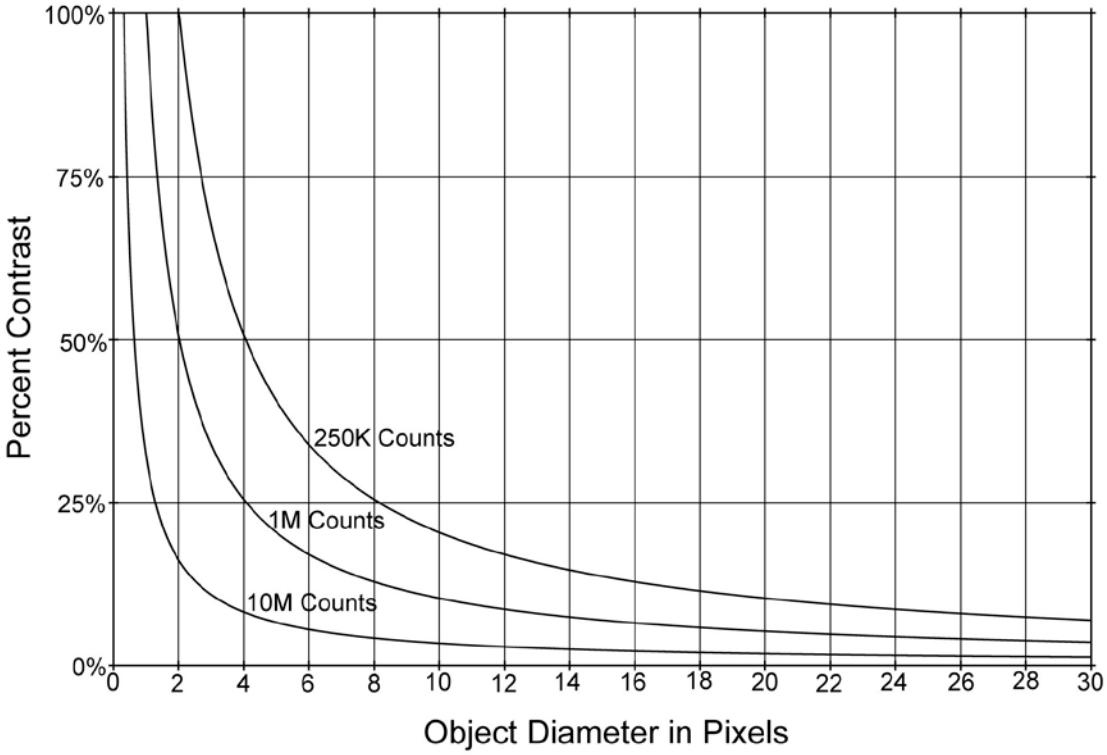

Figure 2: The contrast detail curves for different counting statistics, as illustrated in Figure 1. The sizes of object above each line indicates those that can be detected. In this example, smaller and lower contrast objects can be seen with increasing image counts. Image reproduced from Rzeszotarski (1999).

Despite the value of the SKE/BKE test in radiology and the obvious benefit of a low cost and simple solution to comparing system performance, it is questionable whether the results can be extrapolated into the clinical setting as a prediction of detectability (Kotre, 2006). Furthermore, no visual search is required. Where there is a low threshold for detection the known location of the signal may contribute to detection rates for objects that otherwise would have been inconspicuous. SDT was a natural stepping-stone toward the development of receiver operating characteristic (ROC) analysis for the observer performance in radiology (Burgess, 2011).

The Limitations of a Binary Decision-Making Process

When an observer interprets a series of images with signal detection as the primary objective, there are four possible outcomes of each interpretation: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). In this basic example the observer is making a simple decision about the presence of a signal – either it is present or absent. When the observer completes this task they use a decision threshold to determine whether the signal is present of not. For a single decision threshold, a basic binary (yes/no) response is made. This yields a single measure of sensitivity and specificity. However, this decision threshold is variable for different observers and tasks. It is this variation that can lead to different estimations of sensitivity and specificity for the same test (Obuchowski, 2005). A further drawback of this is that it is difficult to compare two different summary figures (i.e sensitivity and specificity).

In radiology a single decision threshold may be acceptable for cases that are easily classified as normal or diseased. However, the simple cases do not generally cause major problems in radiology; it is the difficult cases that require greater attention, where the boundaries between error and acceptable variation are less clear (Robinson, 1997). Furthermore, a single value of sensitivity and specificity does not account for an imaging system's ability to distinguish between actually negative and actually positive patients (Metz, 2006).

The receiver operating characteristic (ROC) method builds on the limitations of single decision threshold, using a confidence scale to rate the perceived likelihood of disease presence (Metz, 2006; Obuchowski, 2005). This provides an analysis through the entire range of sensitivity and specificity values, over which the ROC curve is plotted (Zou, O'Malley, & Mauri, 2007). The multiple thresholds used in ROC provide a single value (the figure of merit) which can be more easily compared.

Receiver Operating Characteristic Analysis

The receiver operating characteristic (ROC) method was originally used for RADAR detection in the 1950's prior to its first application in radiology in the 1960's (van Erkel & Pattynama, 1998; Vining & Gladish, 1992). ROC methods focus on the correct classification of disease and are frequently used to assess the diagnostic accuracy of imaging techniques in which the observer (frequently a radiologist) is considered an integral part of the system (Chakraborty, 2002). ROC methods are particularly useful for comparing a new imaging technique with an existing 'gold-standard', when the true disease status of the patient is known (Zou, Liu, Bandos, Ohno-Machedo, & Rockette, 2012). The popularity of the ROC method arises from the ability of the method to provide a description of diagnostic accuracy for a full range of sensitivity and specificity of a diagnostic test (Gur et al., 2010; Metz, 2006). When comparing the diagnostic accuracy of two tests a detectability index, such as the area under the ROC curve (AUC), is used to show the level at which the two diagnostic tests are different.

Rating scales are used in ROC analysis to allow an observer to provide a numeric rating to an image based on the perceived likelihood that disease is present (Chakraborty, 2002). These are frequently on a scale of 1 to 5, described as ordinal or discrete, where an increasingly large value signifies an increased level of confidence in the observer's decision. Rating scales can also be quasi-continuous, which can provide a more precise measure of diagnostic accuracy if used well (Hadjiiski, Chan, Sahiner, Helvie, & Roubidoux, 2007). Neither ordinal nor quasi-continuous scales are suitable for all tasks (Rockette & Gur, 2008) and the type of scale used must compliment the diagnostic task. Each value on a rating scale denotes a different threshold value of sensitivity and specificity (Metz, 2006). The use of a rating scale allows a measure of diagnostic performance that is independent of both disease prevalence and decision threshold.

In an ROC study each image is rated and a score of '1' is often reserved for those images that the observer believes to be disease free; ratings 2-5 indicate an increasing suspicion of disease for each scored image. For future reference, the ratings associated with ROC analysis will be classed as *case-based* decisions.

The ROC Curve and Area Under the Curve

The ROC curve is a simple graphical method of displaying the diagnostic accuracy of a test, as discovered during ROC analysis (Zou et al., 2012). It provides a useful visual comparison of the performance of different tests or observers. An ROC curve is produced for each imaging modality or diagnostic test, and each observer being evaluated. The ROC curve is a plot of true positive fraction (TPF, the proportion of correct decisions on actually abnormal

cases) against false positive fraction (FPF, the proportion of incorrect decisions on actually normal cases) (Chesters, 1992; Vining & Gladish, 1992; Zanca et al., 2012). This is a display of the relationship between sensitivity and specificity for a full range of decision thresholds – as determined by the confidence scale used.

For statistical analysis it is common to summarise the ROC curve using the AUC index; it is also frequently referred to as A_z (Jiang, Metz, & Nishikawa, 1996), but this is an estimate of the binormal model estimation only. It is a standardised measure of accuracy of combined observer and modality performance (Chakraborty, 2002). The AUC is defined as the probability that a randomly selected abnormal case has a test result more indicative of abnormality than that of a randomly chosen normal case (Chesters, 1992; Hanley & McNeil, 1982). AUC estimated by the trapezoidal area is equivalent to the Mann-Whitney statistic (Zou, Tempany, Fielding, & Silverman, 1998).

All ROC curves include the trivial operating points of 0,0 and 1,1 that could be calculated by using the strictest and most lax decision thresholds. The number of points on the confidence scale determines the number of non-trivial operating points. Typical ROC curve appearances are shown in Figure 3. A perfect test runs from 0,0 to 0,1 and to 1,1. This covers the entire plot area and the AUC is equal to 1. The chance diagonal represents the curve of a test that is no better than random guessing, which means the test is as likely to be incorrect as it is correct. Tests with high diagnostic accuracy are expected to have an AUC of around 0.9; those with moderate accuracy have an AUC of around 0.75 (Obuchowski, 2000).

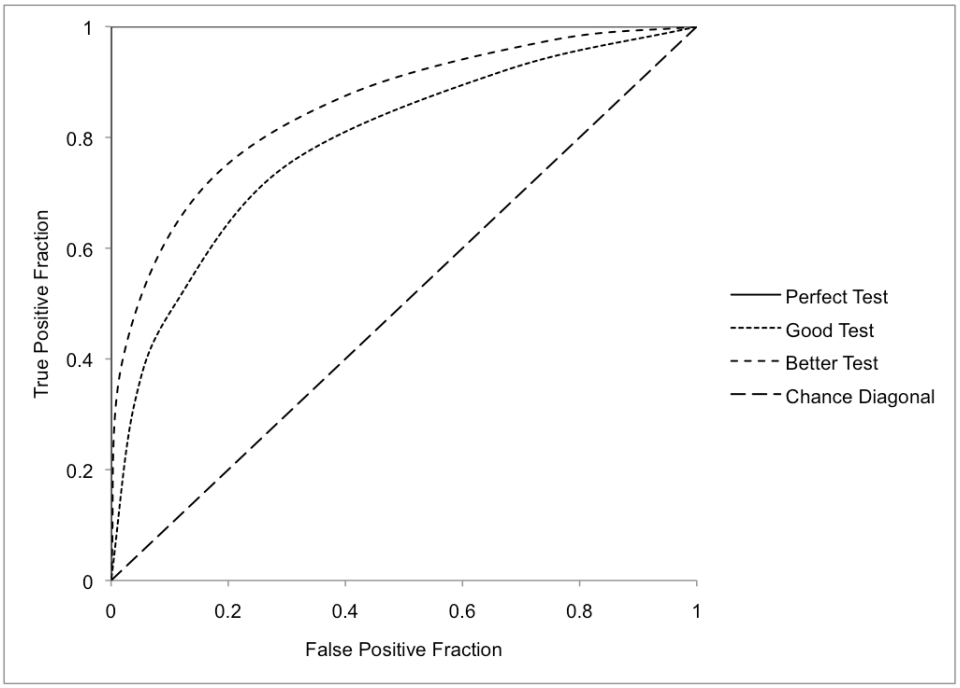

Figure 3: Typical ROC Curve appearances.

Measurements of full AUC give an overall probability that diseased and non-diseased cases are being correctly classified. However, it can be valuable to look at only a certain portion of the curve, known as the partial area.

Partial Area

The partial area (*p*AUC) is frequently defined as the area between two false positive points on the ROC curve (x-axis) (Obuchowski, 2000); this looks at specificity. To look at the high sensitivity portion of the curve, it is the area between two true positive points (y-axis) that should be measured. The case has been made for a partial area measure for tests that require high sensitivity, such as mammography (Jiang et al., 1996). Typically a measurement of partial area is defined as the area to the left of a specified FPF value to measure specificity, or above a specified TPF to measure sensitivity (Zanca et al., 2012), Figure 4.

A full AUC summary implies that all decision thresholds are equally important. Additionally, it can be distracting to rely on the full AUC measurement when comparing two tests with a similar figure-of-merit (FOM), since each test may reveal better performance at different portions of the ROC curve (Zou et al., 2007). The AUC is a FOM to quantify observer performance for detection tasks and is directly related to the two-alternative forced choice (2AFC) test (Clarkson & Shen, 2010).

The chest X-ray (CXR) can be used to describe an example where a full area measure may not be optimal. When looking for solitary pulmonary nodules you want to ensure that the test has high sensitivity, since overlooking a lesion may have serious consequences for the patient. In this situation it may be desirable to look at the high sensitivity portion of the curve. Conversely, it may be of interest to ensure that the test has high specificity; since it would be undesirable to have patients undergoing high dose follow-up examinations unnecessarily.

Partial area can be very useful for comparing two tests that have intersecting ROC curves and a full AUC that is statistically similar. In this situation it can be problematic to determine which is the better test. An example of intersecting ROC curves and the value of partial area is presented in Figure 4. Despite the advantages of the partial area measurement, in the clinical environment one must be mindful of the real clinical impact of one test over another when interpreting ROC curves. It may be that availability, cost and dose become considerations that are of equal importance.

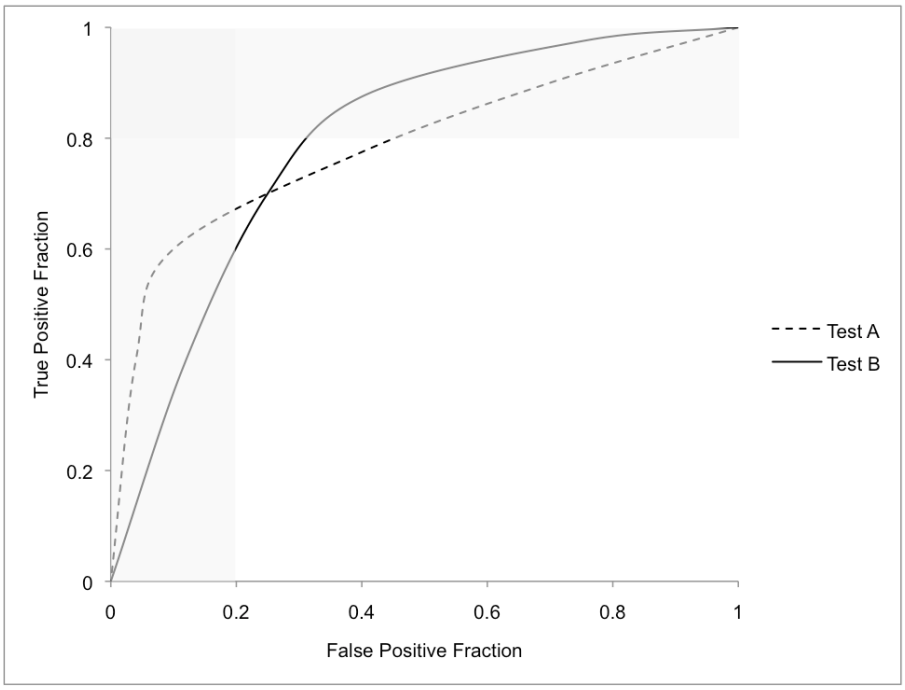

Figure 4: Two intersecting ROC curves. Despite having statistically similar summary indices (AUC) the two curves have areas that outperform each other. The vertical grey band below the curves shows that Test A has better specificity. The horizontal grey band to the right of the curves shows that Test B has better sensitivity.

ROC Plot Example

Consider the example of a series of chest X-ray (CXR) images, 100 normal and 100 abnormal (containing a solitary pulmonary nodule). The observer task is to classify the cases using a rating scale of 1-5, where an increasing number represents a decision of increasing confidence (Table 3). The sensitivity and specificity is calculated for each decision threshold using the standard formulae. This is then converted to TPF and FPF so that the ROC curve can be plotted.

The plotted example shown in Figure 5 is described as the empirical ROC curve, since it is connected with straight lines. The AUC of the empirical curve can be calculated using the trapezoidal rule, where empirical estimations have shown that the AUC is equivalent to the Mann-Whitney U-statistic (Zou et al., 1998). Note the difference to the curves displayed in Figure 3 and Figure 4. These curves are smooth and continuous, described as fitted, representing the correctly modelled ROC curve. However, curve fitting should only be used if there is independence between cases (Chakraborty, 2011). In the empirical work described in my PhD by PW no curve fitting has been applied for two reasons; (i), the cases are not independent as they arise from a single phantom, and (ii), the analysis software does not currently support curve fitting.

The operating points in Table 3 produce an empirical curve with an AUC of 0.794, which can be calculated using the trapezoidal rule. If independence of cases is assumed and the curve is fitted using maximum likelihood estimation by a web-based calculator (Eng, 2014) then the AUC is 0.804, Figure 6. Typically, a computer program will estimate empirical or parametric area by numerical integration, which in the case of empirical estimations by the trapezoidal rule, tends to underestimate the true value (Krzanowski & Hand, 2009).

Rating Labels	1	2	3	4	5	TOTALS
TP	8	12	22	26	32	100
FP	33	37	15	11	4	100
	>1	>2	>3	>4		
Sensitivity (%)	92	80	58	32		
Specificity (%)	33	70	85	96		
TPF (Sens.)	0.92	0.80	0.58	0.32		
FPF (1-Spec.)	0.67	0.30	0.15	0.04		

Table 3: Example data of 200 CXR images, with 50% containing a solitary pulmonary nodule. Observer decisions are distributed across the scale in the expected manner, with the majority of diseased cases scored with high confidence and the majority of normal cases scored with low confidence.

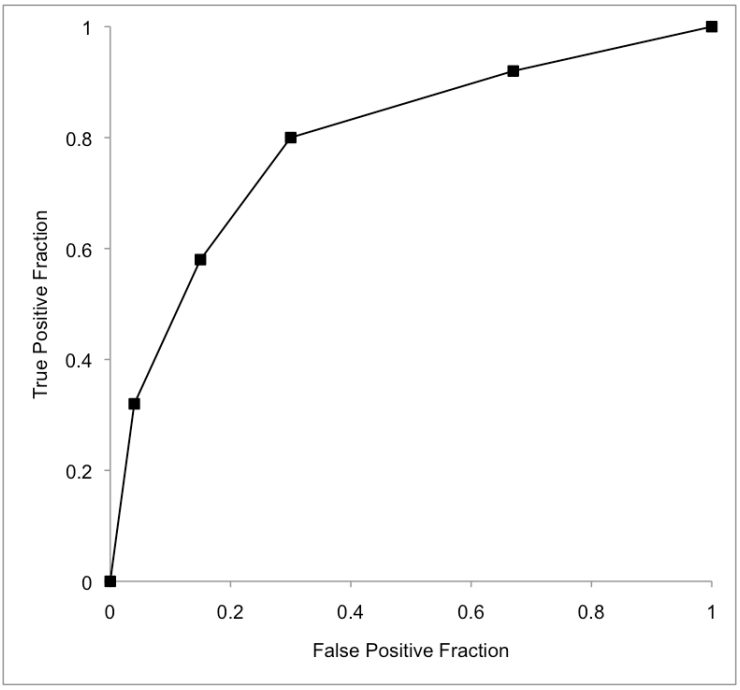

Figure 5: The ROC curve plotted from the operating points calculated in Table 3.

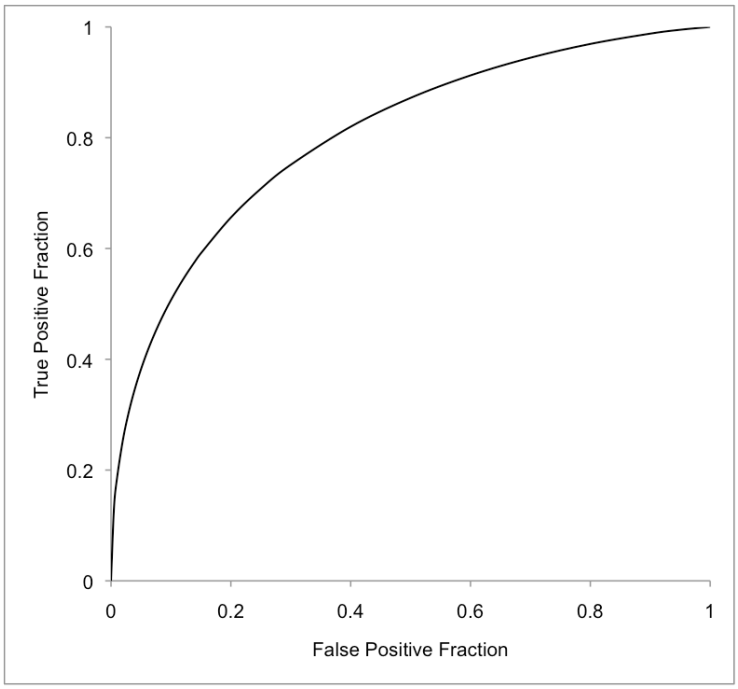

Figure 6: The fitted ROC curve from the rating data in Table 3.

The data in Table 3 have been used to perform curve fitting. If the data are not normally distributed a parametric calculation (fitted) of area may be misleading and the empirical method is preferable (Krzanowski & Hand, 2009).

Free-response Receiver Operating Characteristic Analysis

The free-response receiver operating characteristic (FROC) paradigm was first introduced to radiology in the late 1970's (Bunch, Hamilton, Sanderson, & Simmons, 1978). The method then became further developed by the 1990's (Chakraborty & Winter, 1990). It seeks to overcome some of the limitations of traditional ROC analysis by incorporating location sensitive data into the analysis. ROC methods are limited due to an inability to deal with multiple disease sites effectively; in addition it is also not concerned with localisation and as such it is possible, indeed likely, that false identifications are treated as true identifications on diseased cases.

The FROC method shows greater consideration of the radiological search model, where a global opinion of an image is formed by the observer's peripheral vision. Areas of the images deemed suspicious are then given greater attention, where the observer must apply their decision threshold for disease. Visual search has previously focussed on detection but this is

not as straightforward in the clinical setting due to an unknown priori assumption of disease (Chakraborty, 2006).

The FROC method gives due consideration to this and allows observers to make multiple localisations of disease in each case (Gur & Rockette, 2008); this means that a single case can contain both correct and incorrect localisations (Chakraborty, 2002). The key aspect of the FROC method is the ability to reward correct localisations and penalise incorrect localisations, making efficient use of location information (Chakraborty, 2011b). This method can have more clinical relevance than the ROC paradigm for focal disease but is a demanding perceptual task for the observer (Chesters, 1992).

It is relevant to acknowledge some important differences to the ROC paradigm, specifically:

- Cases can contain multiple ratings in FROC studies; only a single rating for each case in ROC studies
- Cases do not have to be 'rated' if the observer believes the case is disease free; a rating is always applied in ROC
- Individual localisations are classified as correct or incorrect by a proximity criterion and the risk of false localisations being incorrectly classified is minimal

'Mark-Rating' Pairs

Under the FROC paradigm all suspicious areas of an image that exceed the observer's reporting criteria threshold are localised; this is termed a 'mark'. A confidence score accompanies each mark, in the same way that an entire case is scored in the ROC method; this is termed the rating. It can be normal for a series of 'mark-rating' pairs to be produced for each case (Chakraborty & Yoon, 2008). ROC studies result in case-based decisions, whereas FROC studies produce perceived lesion-based decisions.

Proximity Criterion

For a FROC study to be reliable, all localisations should be within a clinically acceptable distance from the true site of disease (Chakraborty, 2002). Localisations that are close enough to the true site of disease are classified as lesion localisation (LL) and those that are too far from the true site are classified as non-lesion localisation (NL) (Chakraborty, 2008; Zanca et al., 2012). While it is easy to understand LL marks, NL marks can occur for two reasons, (i) inadequate proximity to the true disease site; and (ii) a failure to identify a lesion. The clinically acceptable distance, which will vary on the type of pathology present in the

study, is known as the proximity criterion (Chakraborty, 2006). The choice of criterion is important; while it is useful to allow some tolerance for localisations due to hand jitter (observer error in localisation), it must be remembered that the size of the criterion will have an effect on the figure of merit – less strict criteria lead to an inflated figure of merit (Gur et al., 2008).

One of the more common criteria used is the acceptance radius. This is applicable to studies of small spherical lesions, as described in Papers 2, 3, 5 & 7, where the radii length emanates from the centre of the lesion. For this type of criterion it has been recently recommended that the size of acceptance radius should be based on the largest lesion in the study (Haygood et al., 2012). It is also noted that the proximity criterion should be similar for the modalities being compared as a larger value for one modality may cause a bias (Chakraborty, 2011b). This can be a complicated issue and it may require that data be analysed with different sizes of acceptance radii if, for example, the resolution of the imaging modalities being compared is different. The efficacy of this has been explored in Paper 5.

The FROC Curve

The FROC curve is a plot of non-lesion localisation fraction (NLF, the total number of NL ratings above a threshold divided by the total number of images), against the lesion localisation fraction (LLF, which is the total number of LL marks above a threshold divided by the total number of lesions) (Chakraborty, 2010; Zanca et al., 2012). To clarify, the NLF (x-axis) is normalised to the number of cases and the LLF (y-axis) is normalised to the number of lesions (Chakraborty, 2011a). The FROC curve is therefore defined as the plot of probability of LL against the mean number of NL marks per image as the threshold is adjusted (Chakraborty & Yoon, 2008).

The FROC curve (Figure 7) always starts at the origin (0,0) and ends at coordinates of (λ,v), where λ is the mean number of noise sites (no lesion) per image and v is the probability that a lesion is considered for marking (Chakraborty & Yoon, 2008). The FROC curve can have value for determining whether the observer is making full use of the rating scale; a steep start to the curve indicates high confidence and the curve should approach a plateau towards the end of the curve (Chakraborty, 2011a). The FROC curve can extend continuously and as a result the area below the curve cannot be effectively measured to provide a summary index. The FROC curve of a perfect observer has area of zero, and according to the model ends at a finite point and does not extend to an abscissa of unity.

The Alternative FROC (AFROC) Curve

The AFROC curve is a hybrid plot of the ROC and FROC curves. It takes the x-axis from the ROC curve, FPF, and the y-axis from the FROC curve, LLF. On the AFROC curve the FPF is the fraction of diseased cases with NL marks. The AFROC curve has much more value than the FROC curve since it is contained to a plot area of one (1). The area under the AFROC curve defines the reward for LL marks and penalises for NL marks on the basis of the confidence score (Haygood et al., 2012). Additionally, the trapezoidal area of the empirical AFROC curve is equivalent to the jackknife AFROC (JAFROC) figure of merit – frequently used in the analysis of FROC data.

Raw FROC Data for Curve Construction						
Case	Lesions	LL Ratings			NL Ratings	Highest Rating
1	1	4			-	4
2	1	-			-	0
3	1	3			-	3
4	1	9			-	9
5	1	10			-	10
6	2	3	-		-	3
7	0	-			-	0
8	1	-			3	3
9	1	-			5	5
10	3	4	-	-	3	4
11	3	6	2	-	-	6
12	3	10	-	-	-	10
13	1	10			3, 7	10
14	1	8			8	8
15	2	-			-	0
16	2	-			-	0
17	3	3	5	-	-	5
18	2	10	8		3	10
19	0	-			-	0
20	0	-			-	0
21	0	-			-	0
22	0	-			-	0
23	0	-			6	6
24	0	-			-	0
25	0	-			-	0
26	0	-			-	0

Table 4: Raw FROC data is presented, listing the LL and NL ratings and also the highest rating, regardless of classification. These values can now be binned to allow curve construction.

Constructing Curves from FROC data

Three types of curve can be constructed from FROC data. This will be illustrated using an example from Paper 5. The raw FROC data for a single observer is presented in Table 4. In this study there were 26 cases, 17 abnormal cases containing 29 lesions and 9 normal cases. Data are then binned (placed into the interval confidence rating and summed). This is useful if there is a requirement to plot an ROC curve alongside the FROC and AFROC curves. This may be advantageous if wishing to compare performance on the basis of lesion-based and case-based decisions. In order to do this from FROC data, one must determine what the highest rating is on each image – it is the highest rating from which the ROC curve can be inferred. For TP results the highest rating is the highest rating of all LL and NL marks on a case; for FP results it is the highest NL rating, since there are no LL ratings on normal images. This process is also required for AFROC curve construction since this is a plot of LLF against FPF. The operating points of a highest rating inferred ROC curve are displayed in Table 5.

Value	Confidence Score / Number of Observer Decisions										
	0	1	2	3	4	5	6	7	8	9	10
Lesion True Positive (TP)	3	0	0	3	2	2	1	0	1	1	4
No Lesion False Positive (FP)	8	0	0	0	0	0	1	0	0	0	0
	Calculation of Operating Points										
Threshold		≥1	≥2	≥3	≥4	≥5	≥6	≥7	≥8	≥9	≥10
Sensitivity (%)	100	82.4	82.4	82.4	64.7	52.9	41.2	41.2	35.3	29.4	23.5
Specificity (%)	0	88.9	88.9	100	100	100	100	100	100	100	100
True Positive Fraction (TPF)	1	0.824	0.824	0.824	0.647	0.529	0.412	0.412	0.353	0.294	0.235
False Positive Fraction (FPF)	1	0.111	0.111	0.111	0.111	0.111	0	0	0	0	0

Table 5: The raw FROC data from Table 4 has been binned and operating points calculated for each decision threshold.

The operating points for a FROC curve are calculated in a similar way to those for a ROC curve, but they are normalised to a different denominator. The LLF is normalised to the number of lesions while the NLF remains normalised to the number of cases. The important difference between the calculations of the operating points for the FROC curve (Table 6) is that it accounts for all the ratings made in each image, whereas the operating points for the ROC curve account for the highest rating made only.

Value	Bin Number										
	0	1	2	3	4	5	6	7	8	9	10
Lesion Localisation (LL)	14	0	1	3	2	1	1	0	2	1	4
Non-Lesion Localisation (NL)	-	0	0	4	1	1	1	1	1	0	0
	Calculation of Operating Points										
Threshold	-	≥1	≥2	≥3	≥4	≥5	≥6	≥7	≥8	≥9	≥10
Lesion Localisation Fraction (LLF)	-	0.517	0.517	0.483	0.379	0.310	0.276	0.241	0.241	0.172	0.138
Non-Lesion Localisation Fraction (NLF)	-	0.308	0.308	0.308	0.154	0.154	0.115	0.077	0.038	0.000	0.000

Table 6: The FROC curve operating points calculated from the raw FROC data. Note that this includes all ratings (LL and NL) and not just the highest rating. Therefore there are 35 ratings used in these calculations compared to 26 for the highest rating inferred ROC calculations.

Value	Operating Points									
Lesion Localisation Fraction (LLF)	1.000	0.517	0.483	0.379	0.310	0.276	0.241	0.172	0.138	0.000
False Positive Fraction (FPF)	1.000	0.111	0.111	0.111	0.111	0.111	0.000	0.000	0.000	0.000

Table 7: The unique operating points for the AFROC curve for the FROC data in Table 4.

The operating points of the AFROC curve (Table 7) for this single observer example can be extracted from Table 5 and Table 6. Data were crosschecked against JAFROC software version 4.2 (www.devchakraborty.com) for confirmation of accuracy.

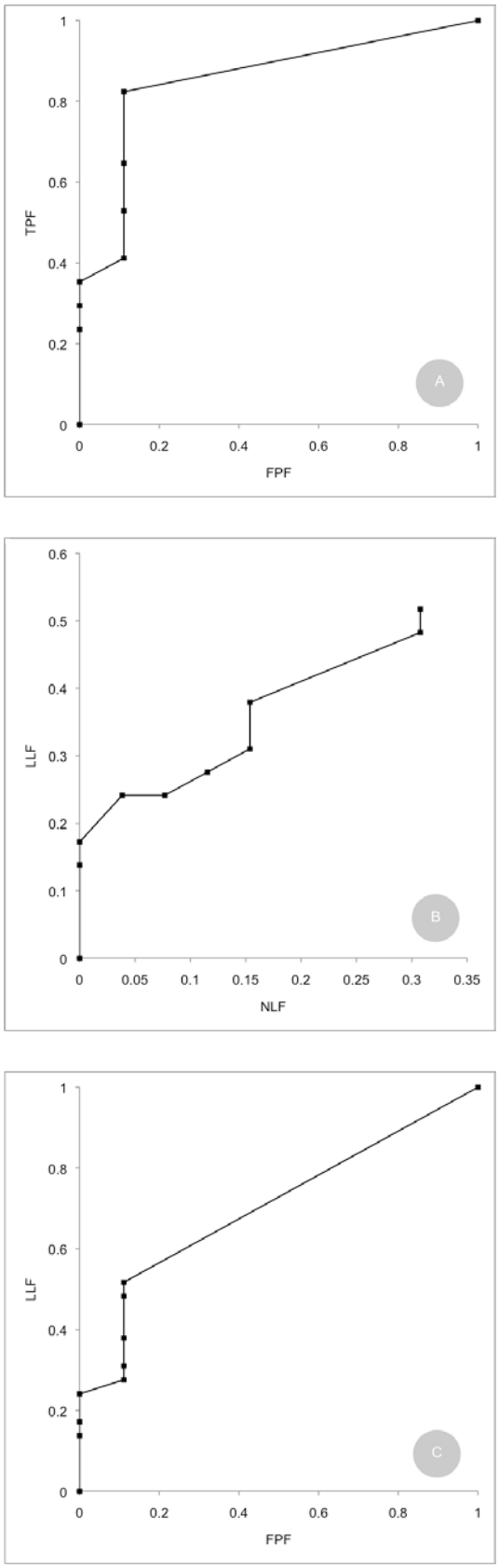

Figure 7: The three curves produced from the raw FROC data in Table 4; A, the highest rating inferred ROC curve; B, the FROC curve; and C, the AFROC curve. Operating points are marked on all curves. The area under the ROC curve is 0.853 and the area under the AFROC curve is 0.703.

Data Analysis

The previous examples have evaluated the data and curves produced by a single observer and a single imaging modality. When conducting an observer performance study it is normal to use a group of observers who interpret all images from all the imaging modalities being compared. This type of design is referred to as multi-reader multi-case (MRMC) (Zanca et al., 2012). Different types of analysis are presented for ROC and FROC data.

Multi-Reader Multi-Case (MRMC) Analysis

The aim of conducting an MRMC study is to reduce the influence of observer experience and variability in the difficulty of the images being interpreted such that any observed difference between the modalities being compared is not masked (Chakraborty, 2011b). It is optimal to generate a result that is applicable to the populations of observers and cases; however, it is important to note that this was not possible in the work reported in this thesis, since it is not possible to report results to a population of phantoms. Despite this, the statistical power is good since the use of a phantom removes the influence of case variability. There are two well-known methods for MRMC analysis, (i) the Dorfman-Berbaum-Metz (DBM) method, and (ii) the Obuchowski-Rockette (OR) method. All the analyses in this work have employed the DBM method and this will be the focus of the discussion. However the DBM and OR methods are equivalent and are exactly equal when the jackknife method is used (Hillis, Berbaum, & Metz, 2008).

The Dorfman-Berbaum-Metz Method

The DBM-MRMC method is the most commonly applied method in ROC analysis, where the AUC is used as the FOM and data are tested using analysis of variance (ANOVA) (Chakraborty, 2011b; Zanca et al., 2012). The DBM method separates the variables of case variance and the variance within and between observers to assess whether the observed difference is caused by the observers or the cases of the study (Kundel & Nodine, 2010). This is known as the jackknife procedure, where pseudovalues are calculated by leaving out one case at a time and then looking at the difference in accuracy estimates between the estimate of all data and with that case removed (Hillis, 2010). For ROC analysis the AUC pseudovalue for case *k*, denoted by Y_{ijk} is:

$$Y_{ijk} = cA\hat{U}C_{ij} - (c-1)A\hat{U}C_{ij(K)}$$

Where *c* is the number of cases, $\hat{AUC}_{ij}$ is the AUC estimate for modality *i* and reader *j* computed for *ALL* data. $\hat{AUC}_{ij(k)}$ is the estimate with case *k* removed (Hillis, 2010).

The number of pseudovalues is determined by the multiplication of *ijk*; for example, an ROC study of two imaging modalities (*i*), five observer (*j*) and fifty cases (*k*) would generate 500 psuedovalues. A random effect ANOVA is then applied to the series of dependant variables (pseudovalues). Modalities are treated as a fixed factor and observers and cases are treated as random such that the results can be applied to the populations.

Once the model has been estimated the next natural step is to use data to determine whether the modalities being compared have statistically different diagnostic performance. Variance components typically of σ_R^2, observer factor, σ_C^2 case factor, and both $\sigma_{\mu R}^2$ and $\sigma_{\mu C}^2$ as modality-observer and modality-case factors respectively, show the variability in the samples. As an example, the variability within the case sample (σ_C^2) arises from the fact that some cases are easier to interpret than others. For $\sigma_{\mu C}^2$ it defines whether a particular case causes more variability in one modality than another. Finally, an error term is represents the residual variability of ijk contributions to the pseudovalue, σ_ε^2 (Chakraborty, 2011a).

The variance components are estimated via ANOVA, which determines whether the relative difference between modalities is due to chance. The final outcome of the analysis is still a single figure of merit for each modality (and observer). The benefit of this method is revealing the individual contribution of each case; removing a high scored diseased case or a low scored normal case will cause the pseudovalue to increase – the converse is true for a low scored diseased case or high scored normal case (Chakraborty, 2011a).

After the jackknife procedure is complete, and the overall figure of merit calculated it is necessary to calculate the 95% confidence intervals. The Fisher F-statistic is calculated with a p-value, where the p-value is the probability that the F-statistic is larger than the observed value (Chakraborty, 2011a). The 95% confidence interval tells us that with 95% certainty the figure of merit is between an upper and lower boundary, for example a θ of 0.602 with limits of 0.549 and 0.655 reveals that the value lies between these limits with 95% certainty. However, it does not mean that 95% of the sample data is between these limits. The calculation of confidence intervals is performed with the following equation:

$$\bar{x} \pm z\frac{s}{\sqrt{n}}$$

where $\bar{x}$ is the mean of the sample, Z is a constant derived from the t-distribution with n-1 degrees of freedom (t-distribution 20DF, Z=2.086), S is the best estimate of population standard deviation and n is the number of values in the sample. For the example above, 0.6022409 (0.54959674,0.65488505) the equation is as follows:

$$0.602 \pm 2.086 \frac{0.116}{\sqrt{21}}$$

for a mean θ of 0.602, a standard deviation of 0.116 in a sample of 21 (n) observers. This calculation derives lower and upper 95% confidence intervals of 0.54959582 and 0.65488597, correct to five decimal places for random readers and fixed cases.

Jackknife Alternative Free-Response Receiver Operating Characteristic (JAFROC) Method

The JAFROC method is the equivalent method for analysing free-response rating data. The JAFROC figure of merit is the trapezoidal are under the AFROC curve, θ. The JAFROC figure of merit is defined by:

$$\theta = \frac{1}{N_N N_L} \sum_{i=1}^{N_N} \sum_{j=1}^{N_L} \psi\left(X_i, Y_j\right)$$

Where X_i is the highest noise rating (mark on a normal case) for a case i, Y_j is the rating for the jth lesion, N_N is the number of normal cases, and N_L is the number of lesions. The psi (ψ) function is the comparison of the highest rated noise rating and the lesion rating (Chakraborty, 2010c). For the ψ function, if X_i (abnormal) is greater than Y_j (normal) then ψ=1; if X_i is less than Y_j then ψ=0; if they are equal ψ=0.5. It therefore follows that the JAFROC figure of merit is defined as the probability that lesions ratings (LL marks) are rated higher than all noise sites (NL marks) on normal images (Chakraborty & Berbaum, 2004). Since the free-response method collects location information for multiple mark-rating pairs, it offers an advantage over ROC analysis on normal cases since the highest rated NL is considered a more stable value of multiple localisations used on normal cases; additionally, NL marks on abnormal images are not used. Remember that the AFROC curve is a plot of LLF (LL marks on abnormal cases) against FPF (normalised to cases). However, multiple lesions on abnormal cases each carry their own rating; missed lesions and low confidence LL marks reduce the LLF and thus the area under the AFROC curve (Chakraborty, 2010b).

The significance testing procedure is identical to the DBM method previously described. The only difference to the DBM procedure for ROC analysis is the figure of merit used; ROC figure of merit is the AUC whereas the JAFROC figure of merit is θ (theta) (Chakraborty, 2010a) and each case is represented by a single pseudovalue, despite individual cases containing multiple mark rating pairs. The similarities are recognised by the pseudovalue denotation in JAFROC analysis:

$$PV_{ijk} = N_T\theta_{ij} - (N_T - 1)\theta_{ij(k)}$$

Where the pseudovalue for each modality *i*, observer *j* and case *k* is defined by PV_{ijk}, N_T is the total number of cases, θ_{ij} is the figure of merit for modality i and observer j for all data, and $\theta_{ij(k)}$ is the figure of merit with case k removed – identifying the dependence of each case (Chakraborty, 2010c).

Comparing Figures of Merit

Visual inspection of ROC/AFROC curves can be valuable for comparing the diagnostic performance of different tests if they are obviously different. As previously mentioned this is easy when the sensitivity and specificity of one test exceeds the other and when curves do not intersect. However, for one to be certain of the incremental value of one test over another it important to compare figures of merit statistically.

The null hypothesis, H_0, is that there is no difference in the figure of merit. An appropriate statistical test must be used to determine whether there is enough evidence to reject the null hypothesis. Focussing on the index, either AUC, pAUC or θ, it is always the figure of merit with the higher value that is superior (Tourassi, 2010). It is conventional to look for statistical significance at α=0.05; this controls the probability of Type I error, where modalities are not found to be the same if they are indeed different; and the Type II error, that they are not found to be different if they are indeed statistically similar.

All of the empirical work evaluated in this thesis considers paired data; i.e. the same cases were exposed to the same set of modalities and observers.

Objective 2

Develop a consistent and reliable method for image display and response capture in free-response studies.

Paper 1 describes the development of a reliable tool for the collection of free-response data. This was integral to the completion of the empirical works (Papers 2, 3, 5 & 7) completed for this thesis. ROCView was used as the sole data collection tool for all free-response studies.

Identifying the Task

The empirical work used CT images of varying image quality, either within a single CT system, or from a range of SPECT/CT systems. In each study the images from each system depicted the same set of lesions in the same anatomical position. Additionally, all studies used a conventional MRMC design, so all images from all the imaging modalities being compared were viewed by the same group of observers. Consequently it was necessary to randomise the order in which the images were displayed to the groups of observers; this would reduce the dependence of the order the images were displayed on the figure of merit achieved.

It is important to qualify that in all the empirical works described the CT images were displayed singularly. This deviates from the clinical routine of reporting images but the method was consistent throughout and substantial effort was made to ensure that lesions were shown at maximum visibility and at the same anatomical position. This was occasionally complicated by variation in the reconstructed slice, as per Paper 5. In this example the localisations were classified as LL and NL marks by two different acceptance radii, 20 and 40 pixels. In this case there was minor inflation of the FOM with the larger radii, suggesting that some marks were incorrectly being classified as LL with a lax proximity criterion.

Access to suitable participants is a barrier to conducting an observer study with optimal statistical power. Low numbers of observers dictate that the caseload requirements on each individual observer can be high. Improving the researchers access to participants, and

improving the participants' access to images was thought to be a suitable step to performing studies with good statistical power.

Development of ROCView

In 2009 the author and Stephen Thompson developed ROCView, with intellectual input from the supervisory team at the University of Salford, as a prototype software solution to data collection in free-response studies. The software was initially developed to run on a single PC that had to be transported to the location where the image evaluations would be completed. The software was adapted in 2010 to run as a web-service, hosted by eUKHost. This had a significant improvement in access to participants. Initially, only a single observer could complete an image evaluation at one time, and with image evaluations typically lasting 45minutes to 1 hour, this would incur a lot of supervision time for the researcher. Developing ROCView as a web-service allowed large numbers of observers to complete the same study, using the same set of images, concurrently. The imaging laboratory at the University of Salford allowed 14 observers to use ROCView concurrently, on the same specification of monitor.

The web-based nature of ROCView was particularly valuable to Paper 7 where the research was completed in five different countries without the need for specialist software to be installed or for a PC to be transported around Europe. The data collection period was completed in a very short period of time and many evaluations were completed concurrently. This type of data collection process can be limited by the variation in specification of monitor. For the research conducted in Paper 7, each centre was required to perform a monitor calibration to ensure that the display quality was adequate for the image evaluation task. There were only minor variations between each centre. However, it is more difficult to control ambient lighting, but previous research on the impact on diagnostic performance is not conclusive on the effect on soft-copy reading (Goo et al., 2004; Park et al., 2008; Pollard, Chawla, Delong, Hashimoto, & Samei, 2008; Uffmann et al., 2005).

Key Functionality of ROCView

To satisfy the requirements of the free-response paradigm the following key aspects of functionality were required:

- Selection of an appropriate rating scale
- Simple lesion localisation method

- Simple confidence scoring method
- Setting the 'Truth'
- Variable proximity criterion
- Reliable storage of data
- Easy extraction and manipulation of data for JAFROC analysis

Ordinal (discrete category) and quasi-continuous rating scales were supported to allow observers to rate confidence[3]. Rating scales are used to allow the observer to provide a numeric rating of confidence on the perceived likelihood of disease presence (Chakraborty, 2002) and it is important to provide both categories of scale since there is no optimal scale that suits all studies (Rockette & Gur, 2008). Despite the value of rating scales for helping to define the full range of sensitivity and specificity on the basis of threshold, a scale is rarely truly representative of the clinical task (Zou et al., 2007).

Lesion localisation was a simplistic task in ROCView, where all localisations were made with individual mouse clicks. Hovering over the desired localisation and selecting 'remove click' could remove unwanted localisations. Once this localisation had been made a pop-up box appeared with the chosen rating scale. The ordinal scale was presented to the observer as a series of selectable buttons marked 1-5 and with statements of increasing confidence (Table 8). These rating scales were used in Papers 2 and 3. The feedback from observers was mixed with this design of scale and although it was deemed easy to use from a practical (point and click) viewpoint, some observers did not feel comfortable assigning a statement to each localisation. Novice observers were used in both of these studies and it is likely that they do not regularly provide a verbal/written statement in clinical tasks, while they may also be unfamiliar with research tasks of this nature. However, an ROC/FROC response can also be unfamiliar to expert observers so the significance of the type of rating scale used may be in question and therefore requires further investigation. Consequently a decision was made to use quasi-continuous scales for future work. This category of rating scale is enabled by a slider-bar (Figure 8).

[3] The choice of rating scale was determined by the researcher at the beginning of each study.

Rating	Example 1	Example 2
1	Very unlikely a lesion is present	Not at all confident
2	Unlikely a lesion is present	Not very confident
3	A lesion may be present	Fairly confident
4	Likely a lesion is present	Confident
5	Very likely a lesion is present	Very Confident

Table 8: Two examples of discrete rating scale employed by ROCView. Statements were accompanied by a rating of 1-5 once the localisation and rating had been completed and listed next to the image being evaluated.

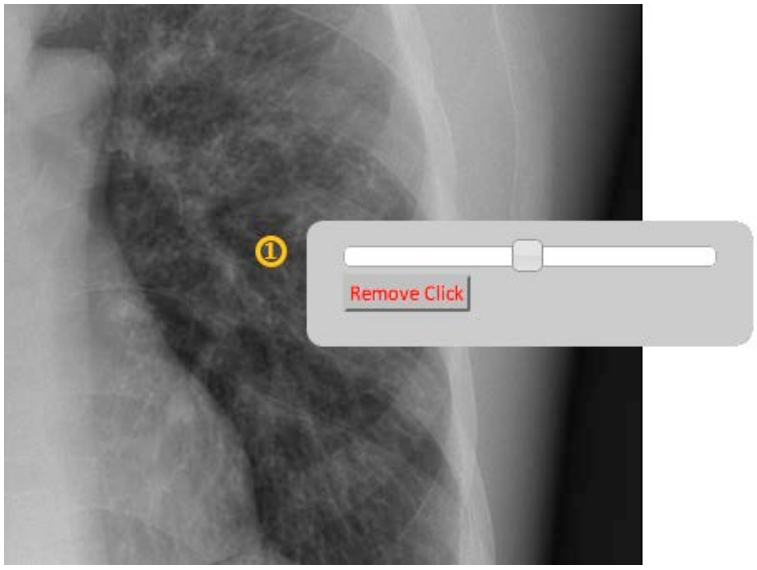

Figure 8: The slider-bar quasi-continuous rating scale that appeared as a pop-up box following a localisation. Moving the slider further to the right indicated a higher level of confidence. The marker containing 1 indicates that this is the first localisation on this case.

The slider-bar received positive feedback from the observers completing the studies reported in Papers 5 and 7 due to ease of use and a lack of ambiguity in ratings. As with the ordinal scales a numerical rating appeared on-screen with each localisation as a list. However, the 100-point data were not analysed as being quasi-continuous and were re-sampled to 10-point rating scale. It can be seen from Figure 8 that the slider-bar scale is unmarked (i.e. no graduations) and it has been previously suggested that observers are poor at effectively utilising scales that have a large range of rating points (Rockette & Gur, 2008). The efficacy of re-binning continuous data to a discrete scale of 1-5, 1-10 and 1-20 bins of equal width has been assessed; it was found not to cause a significant difference in the result generated while producing data that is suitable for reliable curve fitting (Hadjiiski et al., 2007). For Papers 3, 5 & 7, the data were re-binned to 10 equal width bins of 10 ratings.

Once the mark-rating pairs have been created it is important to be able to classify the marks as either LL or NL. This requires two separate functions; (i) setting the 'truth' status for each study, and (ii) using a proximity criterion to classify marks as LL or NL. For all the empirical works in this thesis the researcher established the truth status. While this is acceptable for phantom images where the location of lesions has been tightly controlled and recorded, for clinical images it may require a radiologist to perform this task. However, it must be re-iterated that any images used must have had the true disease status established and the 'truth' must not be subject to variable interpretation.

In ROCView, completing each study and marking, with mouse clicks, the centre of each lesion with high precision, created the truth. The process was similar to that experienced by the observers completing the image evaluations, but no confidence score was given. ROCView records the x-y coordinate of all localisations and it is the coordinates generated under the administrator account that are used as the reference marks. In each study, the mark-rating pairs made by each observer (participant) are compared to the truth on the basis of the x-y coordinates.

The proximity criterion employed was an acceptance radius (AR). The AR emanated from the pixel of the mouse click for each localisation, classifying localisations as LL if they fell within the radius, and as NL otherwise. The classification of mark-rating pairs is not performed until the researcher initiates the download of the stored FROC data. When this is done the researcher has the opportunity to select the AR size. This can be useful in the assessment of observer accuracy, or when the pixel size (spatial resolution) of the modalities or tests being compared is not equal and one wishes to assess the impact of a tight/relaxed proximity criterion. This was the case in Paper 5, where a 20-pixel and 40-pixel AR were used due to the variation in pixel size between the SPECT/CT systems being compared.

Data download is a simplistic procedure where the researcher must select the required study, the observers who have completed the evaluations, the AR size and whether to download LL or NL marks. Following this, only minor adjustment is required for the file to be read by the JAFROC software (current version 4.2). A truth sheet, stating the number of lesions in each case, must also be supplied.

Future Developments to ROCView

ROCView will continue to be used in medical imaging research. A move into radiographic imaging, still using a phantom has recently begun. A phantom study of lesion detection on

postero-anterior chest X-rays has required to additional functions to be added to the software; (i) a blanking image (noise/snow) to prevent the observer from seeing changes in lesions positions as the image changes and (ii) a grey-scale inversion function, similar to that found in a picture archiving and communications system (PACS).

It may be necessary to adapt the software to use digital imaging and communications in medicine (DICOM) standard images. If this was the case it may also be necessary to implement an 'off-line' version of the software due to the large file size associated with series of CT images and singular radiographic images.

One of the major limitations of the work completed for this PhD by PW is the method of image display. Single CT image slices were used for evaluation in all of the empirical works. This is an obvious deviation from a typical evaluation of a CT examination, where all images would be evaluated sequentially. If ROCView were to be updated to allow this it would also be necessary to design a volumetric proximity criterion that applied to a series of slices (3D) rather than a single slice (2D).

Other Studies and Future Work using ROCView

Preliminary reports have been made for a further three papers using ROCView (Jessop et al., 2014; Vamvakas, Hogg, Thompson, Manning, & Szczepura, 2013; Wareing, Hogg, Thompson, Manning, & Szczepura, 2013). The author is currently completing a study of radiographic imaging of the chest. Applications are also expected in mammography to coincide with the Diagnostic Imaging Research Programme at the University of Salford.

Objective 3

Assess the potential for lesion detection and dose and image quality optimisation in a range of CT and SPECT/CT systems.

An Appropriate Test Tool

The ethical considerations of this work have been previously discussed; however, it is worth re-affirming that work of this nature (repeatedly exposing the same subject for the purpose of dose optimisation) is not suitable for a patient population. The phantom used in the empirical works permitted repeated exposures without any concern of the harmful effect of radiation.

The phantom ("*Lungman" N1 Multi-purpose chest phantom, Kyoto Kagaku Company Limited, Japan*) used has an intricate internal structure that seeks to simulate the vascular network within the lungs. This can be removed from the main body of the phantom so that simulated lesions can be placed at any position within the vascular structure.

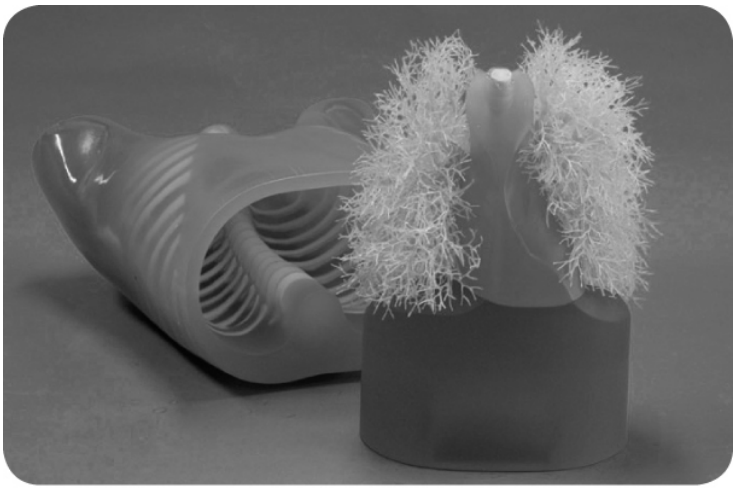

Figure 9: The anthropomorphic chest phantom used in the empirical works (Lungman N1 Multi-Purpose Chest Phantom; Kyoto Kagaku Company Limited Online, www.kyotokagaku.com/products).

The images provided by a CT acquisition of this phantom made for an observer study with a suitable level of difficulty. An example of the images produced can be found in Appendix C – Paper 3 and Appendix E – Paper 5. The observers were generally impressed with the representation provided by the phantom. However, a frequent comment and accepted limitation of the phantom is a lack of respiratory motion.

Method Outline

The method used in all the empirical works was based on the same core steps. Literature searching completed for Paper 2 revealed that there is an approximately even distribution of lesions throughout the thorax (nodules in human subjects) with a higher proportion (70%) found peripherally (within 2cm of the pleura) (Pauls et al., 2008; Rubin et al., 2005; Wright et al., 1996; Xu, Lou, Zhang, Pan, & Zhang, 2007). However, literature review of lesion size and density was less conclusive (Kang et al., 2008; Pauls et al., 2008; Rubin et al., 2005; Wormanns et al., 2004; Xu et al., 2007) (Table 9) and therefore a range of lesions was used in the research.

Study	Mean Size Measurement (mm)	Locations	Hounsfield Unit (HU) Density
(Kang et al., 2008)	8.1±8.8		
(Marten & Grabbe, 2003)			20-60
(Pauls et al., 2008)	7.3	53% Apical 47% Lower	-121.6
(Rubin et al., 2005)	5.1±2.3	71% Peripheral 29% Central	
(Wormanns et al., 2004)	7.4±4.5		35±16.3
(Wright et al., 1996)	Majority <5	28% Upper 44% Middle 28% Lower	
(Xu et al., 2007)	67% <5	Peripheral if within 2cm of pleura	

Table 9: Characteristics of lesion size, location and density.

Testing the Method

Access to clinical SPECT/CT systems can be limited due to the availability of staff and scanner time during normal working hours. It was therefore imperative that the method was tested thoroughly in the research environment at the University of Salford prior to beginning the empirical work in a clinical environment.

A modern CT scanner (Toshiba Aquilion 16-slice MDCT) was used to assess the potential for dose optimisation in CT by a variation of acquisition parameters. This would also provide an opportunity to test ROCView for display and response capture and also to perform an analysis of data.

Scanning in a single CT unit allows an opportunity for lesion configurations to be represented in the same anatomical position and in the same slice number since no physical movement of the phantom is required between acquisitions; the table of the CT scanner performs the only movement that occurs. Between each CT acquisition the table position was set to zero to ensure the same start point for each scan. This would ensure that lesion appearances within the resultant images were 'matched'.

In this first empirical work (Paper 2) there were eight variations of acquisition protocol; two variants of helical pitch and four variants of tube current while all other acquisition and image reconstruction parameters remained constant. A decision was also made to look at *case-based* rather than *lesion-based* decisions in the analysis, using the DBM-MRMC method. Unfortunately, this led to a reduction in statistical power. In addition, using eight different 'modalities' for comparison in either JAFROC or DMB-MRMC analysis exceeds what is statistically satisfying and the large number of inter-comparisons of statistical similarity can be difficult to interpret.[4] In this study the case-based analysis showed no statistical difference in lesion detection on the basis of changing the acquisition parameters. The overall outcome was that dose optimisation was possible. However, it is accepted that a failure to reject the null hypothesis does not mean that the alternative hypothesis is true.

Dose Optimisation in a Single Hybrid System

In nuclear medicine the CT component of a hybrid system is used for attenuation correction (AC), lesion localisation or diagnosis. When performing a scan with the primary objective of AC the acquisition parameters should be set to incur the lowest dose to the patient, as several previous works have shown that the acquisition parameters used have no measurable impact on the quality of AC provided (Kamel et al., 2002; Preuss et al., 2008; Wells et al., 2012). However, low-dose data should not be overlooked and there is compelling evidence that incidental findings can be detected on these low-resolution images. With this in mind, the focus of Paper 3 was to evaluate the influence of tube current on the ability of observers to accurately localise simulated lesions on the resultant CT images of a scan performed for AC. In this work, twenty observers evaluated images from four different tube current settings, searching for lesions of different spherical sizes and density. Observers were made aware of the case mix and general lesion characteristics.

[4] Since the completion of the empirical works there has been discussion with Professor Dev Chakraborty of Pittsburgh University, Pennsylvania, to discuss what to do with this type of analysis. Although this new development is beyond the scope of this PhD thesis, a new JAFROC analysis, termed crossed-modality, has been developed to deal with the analysis of two co-existing factors.

The SPECT/CT system evaluated in Paper 3 is commonly used in practice. Many of the image acquisition and reconstruction parameters are fixed with very little operator input. However, the tube current and rotation speed are selectable. The tube current range of 1, 1.5, 2 and 2.5mA sounds small and the difference in dose received by the patient may seem insignificant given the small range. However, in hybrid imaging the tube rotation time can be long 23.1 or 30 seconds, so the effective mAs, accounting for pitch, can be considerably reduced if using the lowest tube current.

In this study the data collection method was identical to that in Paper 2. However, on this occasion the data were analysed using the JAFROC method, taking *lesion-based* decisions into account. Despite the tube current range, lesion detection was statistically equal at each dose level, suggesting that dose could be optimised at the lowest tube current setting. This was an encouraging result when the primary purpose of the acquisition is AC. Even if lesion detection had been statistically superior at a higher tube current, the justification of using a higher dose acquisition for AC would have been questionable if the same primary objective (reliable AC) could be achieved with a lower dose.

Comparing Multiple SPECT/CT Systems

Following a successful evaluation of a single SPECT/CT system in Paper 3, the method was applied to a series of SPECT/CT systems, representing the current range of SPECT/CT systems available in clinical practice. Paper 5 was the most challenging empirical work to complete, but also the most rewarding and the paper that has the most valuable impact to the nuclear medicine community. Using the described method, the phantom was imaged on five SPECT/CT systems in different geographic locations. The logistics of this study dictated that the phantom was only loaded with one configuration of lesions, yielding a low number of cases for the free-response study. It would have been impractical to load the phantom with lesions and move it between sites on more than one occasion.

The outcome of this work suggested that the detection of incidental findings as a result of the CT acquisition parameters used was subject to significant variation, with some SPECT/CT systems offering good lesion detectability and others offering poor lesion detectability. Although the review of the CT data is not the primary purpose of the investigation it does suggest that a patient population may receive a different quality of examination, depending on the SPECT/CT system and the acquisition parameters in operation. However, it is important to look beyond this examination in isolation and consider what may happen to a patient if images are reviewed and incidental findings are revealed. The images being

reviewed are generally low-resolution and not of the diagnostic standard associated with a conventional CT scan. Therefore it is almost certain that patients with incidental findings will be sent for further imaging, thus increasing the dose. If the incidental finding turns out to be of clinical significance, then it is a worthwhile exposure. However, if the incidental finding is a false positive identification, as a result of an interpretation of the low-resolution images then the efficacy of reviewing the images acquired for attenuation correction is called into question since this end result could be wasted resources, increased radiation dose to the patient and potential risks of unnecessary treatment. This is a niche area of practice but some parallels can be drawn with low-dose lung cancer screening; this will be discussed later on.

A low number (n=27) of cases were used in this study, but since it is a phantom study the low number of cases is not a problem since there is no case variability. Furthermore, a large number of observers were used (n=21) to reduce variability. The observers used were all nuclear medicine practitioners studying a hybrid imaging module as part of the Postgraduate Diploma in Nuclear Medicine, but in terms of image interpretation they would not be classed as expert observers. The issue of expertise and access to suitable observers are evaluated further in Paper 7.

Summary

A significant amount of work has looked at the technical performance and capability of both CT and SPECT/CT systems, focussing on physical measures of image quality and the quality of AC provided. However, the author believes that the work collated for this PhD by published works represents the first observer studies of the low-resolution CT images that are acquired in SPECT/CT imaging. The method used is reliable, despite the use of single CT images rather than a contiguous set of image data.

Objective 4

Assess the role of the novice observer for suitability in observer performance research.

The search experience and expectations of radiologists are suppressed in phantom studies since the environment is different to that within which they normally work. Consequently it is possible that suitably trained novices could perform to an equal standard if appropriate training is given. In all the empirical papers completed for this thesis the level of CT experience (years working in CT, hybrid or conventional) was generally very low; this is summarised in Table 10.

Study	Observers	Experience (years)	Mean
Paper 2	9	0, 1, 0, 2, 4, 7, 0, 1, 16	3.44±5.25
Paper 3	20	18, 2, 0, 9, 0, 0, 0, 24, 8, 6, 0, 0, 7, 0, 0, 5, 0, 0, 0, 20	4.45±6.74
Paper 5	21	1, 0, 0, 1, 0, 0, 4, 2.5, 2, 1.5, 0, 2.5, 0.5, 2, 0.5, 0, 0, 1.5, 1.5, 1, 3	1.17±1.17
Paper 7	34	ALL 0 (*undergraduate students*)	-

Table 10: Summary of observer experience in the empirical works completed for this PhD by PW.

The key aspect of the empirical works reported here is the lack of case variability. This was enforced by the use of the anthropomorphic chest phantom, and while this has many obvious benefits, particularly in relation to the level of control in the work, it is limited in its representation of the clinical task. Radiologists, typically considered expert observers, have a well-developed mental picture of normality as a result of caseload experience. This does not apply in phantom work since it is far enough removed from real patient images. This immediately removes a distinct advantage that an expert observer typically holds over a novice observer. In Paper 7 the ability of the novice observer was put to the test. Although the method was limited (i.e. knowledge of prevalence in the second evaluation), it showed that observer performance could improve at a second read, without being able to place an exclusive authority on a reason for this.

It can be seen from Table 10 that Paper 7 was the first empirical work to use novice observers exclusively. As in the other empirical works the novices were evaluating low-resolution CT images acquired for AC – of which they had no previous experience. Prior to the first image evaluation the observers were trained in the use of ROCView and shown lesion and phantom appearances. In the second evaluation the number of NL localisation was reduced by more than 50%. As previously mentioned, it was not possible to determine a singular cause for this since there were several contributory factors in the second evaluation (experience gained, knowledge of prevalence, training in ROC/FROC methods).

More work is required to investigate the value of the novice observer in observer performance. A suitable strategy to assess the novice observer requires careful planning such that they could be compared to expert performance. A test and re-test strategy, exposing novices and experts to the same series of images on two occasions, with no other confounding variables, may allow a better assessment of the novice group. A remaining issue would be the experienced gained from a first evaluation, but at least it would then be possible to state whether or not the experience gained was significant on observer performance.

Wider Applications 1

Low-dose Lung Cancer Screening and Incidental Findings

The empirical work described in this thesis focuses on a niche area of imaging that has previously had minimal investigation. The acquisition parameters used for AC imaging are vastly different to those used for diagnostic CT. However, low-dose imaging is also performed in CT for lung cancer screening.

Study	Acquisition Protocol
Paper 5 (AC) SPECT/CT System 5	120kVp, Pitch 0.94, 50mAs, 16x0.75mm detector configuration, 5mm reconstructed slice, 1.5 s rotation time
ELCAP (LCS) (Henschke, 2011)	140kVp, Pitch 2, 40mAs, <0.8 s rotation time, 10mm slice reconstruction
ELCAP (LCS) (Cagnon et al., 2006)	120kVp, Pitch 0.98-2, 50-160mA or 26.7-59.2mAs, <1.0s rotation time, 64x0.6mm-4x2.5mm configuration, 2.0-3.2mm slice reconstruction at 1.8-2.0mm intervals
Mayo Clinic (LCS) (Cagnon et al., 2006)	120kVp, Pitch 1.5, 40mAs, 5mm slice acquisition, 3.75mm slice reconstruction
NLST (LCS) (Aberle, Berg, et al., 2011)	120-140kVp, 20-30mAs, 1.0-3.2mm slice reconstruction at 1.0-2.5mm intervals
PLuSS (LCS) (Wilson et al., 2008)	140kVp, 40-60mAs, 2.5mm slice reconstruction
NELSON (LCS) (Ru Zhao et al., 2011)	80-140kVp, Pitch 1.3, variable mAs, 16x0.75mm detector configuration
Depiscan (LCS) (Blanchon et al., 2007)	100-140kVp, Pitch 1-1.5, 20-100mA (variable mAs), 1-1.5mm detector configuration, 1.25-3mm slice reconstruction

Table 11: A comparison of CT acquisition parameters used in lung cancer screening (LCS) and for attenuation correction (AC).

Some of the reported CT acquisition parameters used for lung cancer screening (LCS) are comparable with the low-resolution imaging used for attenuation correction (Table 11); however, spatial resolution is not compromised to the same extent and a larger matrix size

and smaller field of view are used in LCS since there is no requirement to match transmission and emission data.

A comparison of LCS and AC imaging protocols can be found in Table 11. It is clear that acquisition protocols for LCS show the same level of variation as those for AC; this is more than likely representative of the time period over which these screening trials have been conducted and also the ever changing CT technology. The Early Lung Cancer Action Project (ELCAP) commenced in 1992 and CT has seen dramatic change in this time. Table 11 reflects this, where two different ELCAP protocols are presented (Cagnon et al., 2006; Henschke, 2011). It is interesting to note there is still scope for variation in the CT acquisition protocols, despite calls for standardisation (Cagnon et al., 2006).

The controversy of low-dose CT for LCS (radiation dose balanced against clinical benefit) is similar to that of searching AC images for incidental findings in terms of the efficacy of the procedure. On the negative side, there are reports of misdiagnosis and over-diagnosis (Patz et al., 2014; Veronesi et al., 2012) whereas other work highlights the reduced morbidity and mortality rates associated with LCS (Humphrey et al., 2013).

Of further interest, a review paper of LCS revealed that in most cases, but not all, the radiologist would search for incidental findings when reviewing the datasets (Humphrey et al., 2013). The key finding was that 7.5% of all LCS CT scans contained an abnormality of clinical significance that was not suspicious for lung cancer. It is now interesting to compare this finding to those of a clinical study of incidental findings for patients undergoing CT based attenuation correction during a myocardial perfusion scan. In this study a large proportion of patients (119/200; 59.5%) demonstrated incidental findings, however the number considered of potentially major clinical significance was low, 21 in 200 (10.5%) (Goetze et al., 2006). On initial inspection these figures sound similar and it may be expected that similar numbers of major incidental findings may be discovered. However, while the LCS patients receive a full thoracic CT, the patients reported in the study of Goetze et al., (2006) only received a CT scan of a small area (typically 13cm) that covered the heart and thus provided a suitable region for AC. So, it could be hypothesised that if the patients attending for MPI had received a full thoracic CT the number of significant findings could have been larger. This leads to a pertinent question; why? The LCS patients are classed as asymptomatic but are considered to be a sample of the population that are ‘at risk’ (i.e. exposure to tobacco or asbestos) (Nanavaty, Alvarez, & Alberts, 2014). Patients attending for MPI will be there on the basis of clinical suspicion of coronary artery disease (CAD) or ischemic heart disease (IHD) and could be considered symptomatic. Larger studies of incidental findings revealed on AC images would be required to confirm if the trend observed by Goetze et al., (2006) is

representative; a sample of 200 patients is small in comparison to the larger LCS trials; for example the National Lung Cancer Screening Trial (NLST) conducted in the US used 53,454 patients to compare the value of CT and CXR in the detection of lung cancer (Aberle, Adams, et al., 2011).

A recent clinical study of incidental findings has built on the early work of Goetze et al. (2006). In the study of (Coward et al., 2014) the AC images of patients undergoing MPI were assessed at a number of different imaging centres, using a variety of acquisition parameters, reflective of those discussed in Paper 5. In their study there were incidental findings in 27% of patients (423/1819) compared to 59.5% (119/200) of (Goetze et al., 2006). However, the number of findings to be of clinical significance was much lower (0.2%) in the more recent study. The authors ultimately question the overall benefit of reviewing the low-resolution CT images when the nature of the images (low quality) means that disease cannot necessarily be ruled out even if it is not identified on the images.

The work of Coward et al. (2014) is consistent with Paper 5, where the SPECT/CT systems used in the study did not appear to perform equally due to the quality of the images produced. Although the authors did not have matched cases in all centres, since this was a patient based study, they found that the positive predictive value (PPV) was significantly better for the SPECT/CT system with diagnostic capability compared to those that are considered low-resolution.

Incidental findings are described as an ‘inevitable consequence’ of screening (Edey & Hansell, 2009). While it is accepted that life-threatening disease must be acted upon, they can obscure the findings of a trial. The authors then continue to explain that investigators completing The Dutch Belgian Randomised Lung Cancer Screening Trial (NELSON) found a much lower prevalence of incidental findings (1%) and considered that the systematic review of extra-thoracic pathology should not be performed on low-dose screening of the chest (Edey & Hansell, 2009). This may be a difficult conclusion to accept when there is a chance of detecting a clinically significant incidental finding, but this must be balanced against the costs of searching for and detecting incidental findings (time, money, radiation dose, stress to the patient).

Wider Applications 2

FROC and Radiographic Trauma Imaging

It is the aim of the researcher to develop ROCView to allow the FROC method to be applied to the reporting of radiographic trauma images by radiographers. With the continuing developments in radiology, where the radiologist is trying to relinquish responsibility of tasks that could be completed by others, it is important that the performance of those taking on these reporting roles is monitored. Radiographers are currently involved in the reporting of CT head and fluoroscopic examinations in addition to radiographic image appearances. However, the scope of pathology available to report upon in CT and fluoroscopy is likely to be too broad for a tightly controlled FROC evaluation and it is the authors opinion that it is better suited to a single pathology type in a focussed group of patients (i.e. fracture detection). That is not to say that FROC methodology cannot be used effectively in other areas of imaging, but the methods currently being employed by the researcher are better suited to fracture detection on trauma imaging.

If the development of ROCView for monitoring fracture detection by radiographers is successful, it is anticipated that it should be possible to provide a similar self-assessment as mammography PERFORMS (PERsonal performance in Mammographic Screening) (Gale, 2010) to help reporting radiographers monitor their trauma image interpretation skills.

Conclusions

The free-response method has been successfully applied to dose and image optimisation in low-resolution CT. Scope for dose reduction is evident in the empirical work and it should now be the priority to extend the method to the clinical setting. Lesion detection can be highly variable in low-resolution CT images as a result of the CT acquisition parameters selected by the operator, but controversy still surrounds the use of low-resolution images when interpretation of CT data for incidental findings is not the primary reason for the exposure. Future work will look to assess new CT technology, such as iterative reconstruction, focussing on the impact this has on lesion detection.

Reference List

Aberle, D. R., Adams, A. M., Berg, C. D., Black, W. C., Clapp, J. D., Fagerstrom, R. M., Gareen, I. F., Gatsonis, C., Marcus, P. M., & Sicks, J. D. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine*, *365*, 395–409. doi:10.1016/j.yrad.2012.03.010

Aberle, D. R., Berg, C. D., Black, W. C., Church, T. R., Fagerstrom, R. M., Galen, B., Gareen, I. F., Gatsonis, C., Goldin, J., Gohagan, J. K., Hillman, B., Jaffe, C., Kramer, B. S., Lynch, D., Marcus, P. M., Schnall, M., Sullivan, D. C., Sullivan, D., & Zylak, C. J. (2011). The National Lung Screening Trial: overview and study design. *Radiology*, *258*, 243–253. doi:10.1148/radiol.10091808

Achenbach, S., Chandrashekhar, Y., & Narula, J. (2013). The ethics of publishing medical imaging research. *JACC. Cardiovascular Imaging*, *6*(12), 1351–3. doi:10.1016/j.jcmg.2013.10.003

Arumugam, P., Harbinson, M., Reyes, E., Sabharwal, N., & Tonge, C. (2012). *Procedure Guidelines for Radionuclide Myocardial Perfusion Imaging with Single-Photon Emission Computed Tomography (SPECT)*. Retrieved from http://www.bnms.org.uk/images/stories/Procedures_and_Guidelines/MPS_procedure_guidelines_Final_12.pdf

Bateman, T. M., & Cullom, S. J. (2005). Attenuation correction single-photon emission computed tomography myocardial perfusion imaging. *Seminars in Nuclear Medicine*, *35*, 37–51. doi:10.1053/j.semnuclmed.2004.09.003

Blanchon, T., Bréchot, J. M., Grenier, P. A., Ferretti, G. R., Lemarié, E., Milleron, B., Chagué, D., Laurent, F., Martinet, Y., Beigelman-Aubry, C., Blanchon, F., Revel, M. P., Friard, S., Rémy-Jardin, M., Vasile, M., Santelmo, N., Lecalier, A., Lefébure, P., Moro-Sibilot, D., Breton, J. L., Carette, M. F., Brambilla, C., Fournel, F., Kieffer, A., Frija, G., & Flahault, A. (2007). Baseline results of the Depiscan study: A French randomized pilot trial of lung cancer screening comparing low dose CT scan (LDCT) and chest X-ray (CXR). *Lung Cancer*, *58*, 50–58. doi:10.1016/j.lungcan.2007.05.009

Bockisch, A., Freudenberg, L. S., Schmidt, D., & Kuwert, T. (2009). Hybrid Imaging by SPECT/CT and PET/CT: Proven Outcomes in Cancer Imaging. *Seminars in Nuclear Medicine*. doi:10.1053/j.semnuclmed.2009.03.003

Brenner, D. J., & Hall, E. J. (2007). Computed tomography--an increasing source of radiation exposure. *The New England Journal of Medicine*, *357*, 2277–2284. doi:10.1056/NEJMra072149

Bunch, P., Hamilton, J., Sanderson, G., & Simmons, A. (1978). A free-response approach to the measurement and characteriztion of radiographic-observer performance. *Journal of Applied Photographic Engineering*, *4*, 166–171.

Burgess, A. E. (2010a). Signal detection in radiology. In E. Samei & E. A. Krupinski (Eds.), *The Handbook of Medical Image Perception and Techniques* (pp. 47–72). New York: Cambridge University Press.

Burgess, A. E. (2010b). Signal detection theory - a brief history. In E. Samei & E. A. Krupinski (Eds.), *The Handbook of Medical Image Perception and Techniques* (pp. 26–71). New York: Cambridge University Press.

Burgess, A. E. (2011). Visual perception studies and observer models in medical imaging. *Seminars in Nuclear Medicine*, *41*, 419–36. doi:10.1053/j.semnuclmed.2011.06.005

Burrell, S., & MacDonald, A. (2006). Artifacts and pitfalls in myocardial perfusion imaging. *Journal of Nuclear Medicine Technology*, *34*, 193–211; quiz 212–214.

Cagnon, C. H., Cody, D. D., McNitt-Gray, M. F., Seibert, J. A., Judy, P. F., & Aberle, D. R. (2006). Description and implementation of a quality control program in an imaging-based clinical trial. *Academic Radiology*, *13*, 1431–1441. doi:10.1016/j.acra.2006.08.015

Chakraborty, D. P. (2002). Statistical power in observer-performance studies: Comparison of the receiver operating characteristic and free-response methods in tasks involving localization. *Academic Radiology*, *9*, 147–156. doi:10.1016/S1076-6332(03)80164-2

Chakraborty, D. P. (2006). Analysis of Location Specific Observer Performance Data: Validated Extensions of the Jackknife Free-Response (JAFROC) Method. *Academic Radiology*, *13*, 1187–1193. doi:10.1016/j.acra.2006.06.016

Chakraborty, D. P. (2008). Validation and Statistical Power Comparison of Methods for Analyzing Free-response Observer Performance Studies. *Academic Radiology*, *15*, 1554–1566. doi:10.1016/j.acra.2008.07.018

Chakraborty, D. P. (2010a). A Status report on free-response analysis. *Radiation Protection Dosimetry*, *139*, 20–25. doi:10.1093/rpd/ncp305

Chakraborty, D. P. (2010b). Clinical relevance of the ROC and free-response paradigms for comparing imaging system efficacies. *Radiation Protection Dosimetry*, *139*, 37–41. doi:10.1093/rpd/ncq017

Chakraborty, D. P. (2010c). Recent developments in FROC methodology. In E. Samei & E. A. Krupinski (Eds.), *The Handbook of Medical Image Perception and Techniques* (pp. 216–239). New York: Cambridge University Press.

Chakraborty, D. P. (2011a). New developments in observer performance methodology in medical imaging. *Seminars in Nuclear Medicine*, *41*, 401–418. doi:10.1053/j.semnuclmed.2011.07.001

Chakraborty, D. P. (2011b). Recent developments in imaging system assessment methodology, FROC analysis and the search model. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *648*. doi:10.1016/j.nima.2010.11.042

Chakraborty, D. P., & Berbaum, K. S. (2004). Observer studies involving detection and localization: Modeling, analysis, and validation. *Medical Physics*, *31*(8), 2313–2330. doi:10.1118/1.1769352

Chakraborty, D. P., & Winter, L. H. (1990). Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*, *174*, 873–881.

Chakraborty, D. P., & Yoon, H.-J. (2008). Operating characteristics predicted by models for diagnostic tasks involving lesion localization. *Medical Physics*, *35*, 435–445. doi:10.1118/1.2820902

Chesters, M. S. (1992). Human visual perception and ROC methodology in medical imaging. *Physics in Medicine and Biology*, *37*, 1433–1476. doi:10.1088/0031-9155/37/7/001

Clarkson, E., & Shen, F. (2010). Fisher information and surrogate figures of merit for the task-based assessment of image quality. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *27*, 2313–2326. doi:10.1364/JOSAA.27.002313

Coward, J., Lawson, R., Kane, T., Elias, M., Howes, A., Birchall, J., & Hogg, P. (2014). Multi-centre analysis of incidental findings on low-resolution CT attenuation correction images. *The British Journal of …*, *87*(1042). doi:10.1259/bjr.20130701

Dawson, P. (2004). Patient dose in multislice CT: why is it increasing and does it matter? *The British Journal of Radiology*, *77 Spec No*, S10–S13. doi:10.1259/bjr/23162044

Department of Health. The Ionising Radiation (Medical Exposure) Regulations 2000, Health (San Francisco) 1–12 (2000). Retrieved from http://www.legislation.gov.uk/uksi/2000/1059/pdfs/uksi_20001059_en.pdf

Duquenoy, P., George, C., & Solomonides, A. (2008). Considering something “ELSE”: ethical, legal and socio-economic factors in medical imaging and medical informatics. *Computer Methods and Programs in Biomedicine*, *92*(3), 227–37. doi:10.1016/j.cmpb.2008.06.001

Edey, A. J., & Hansell, D. M. (2009). CT lung cancer screening in the UK. *The British Journal of Radiology*, *82*(979), 529–31. doi:10.1259/bjr/17503608

Eng, J. (2014). ROC Analysis: A web based calculator for ROC curves. Retrieved August 19, 2014, from http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html

Gaemperli, O., Bengel, F., & Kaufmann, P. A. (2011). Cardiac hybrid imaging. *European Heart Journal*, *32*, 2100–2108.

Gale, A. (2010). Maintaining quality in the UK breast screening program. In D. J. Manning & C. K. Abbey (Eds.), *Proc SPIE 7627, Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*. doi:10.1117/12.846036

Gallowitsch, H. J., Sykora, J., Mikosch, P., Kresnik, E., Unterweger, O., Molnar, M., Grimm, G., & Lind, P. (1998). Attenuation-corrected thallium-201 single-photon emission tomography using a gadolinium-153 moving line source: Clinical value and the impact of attenuation correction on the extent and severity of perfusion abnormalities. *European Journal of Nuclear Medicine*, *25*, 220–228. doi:10.1007/s002590050220

Goetze, S., Pannu, H. K., & Wahl, R. L. (2006). Clinically significant abnormal findings on the "nondiagnostic" CT portion of low-amperage-CT attenuation-corrected myocardial perfusion SPECT/CT studies. *Journal of Nuclear Medicine*, *47*, 1312–1318.

Goo, J. M., Choi, J.-Y., Im, J.-G., Lee, H. J., Chung, M. J., Han, D., Park, S. H., Kim, J. H., & Nam, S.-H. (2004). Effect of monitor luminance and ambient light on observer performance in soft-copy reading of digital chest radiographs. *Radiology*, *232*, 762–766. doi:10.1148/radiol.2323030628

Gur, D., Bandos, A. I., Klym, A. H., Cohen, C. S., Hakim, C. M., Hardesty, L. A., Ganott, M. A., Perrin, R. L., Poller, W. R., Shah, R., Sumkin, J. H., Wallace, L. P., & Rockette, H. E. (2008). Agreement of the Order of Overall Performance Levels Under Different Reading Paradigms. *Academic Radiology*, *15*, 1567–1573. doi:10.1016/j.acra.2008.07.011

Gur, D., Bandos, A. I., Rockette, H. E., Zuley, M. L., Hakim, C. M., Chough, D. M., Ganott, M. A., & Sumkin, J. H. (2010). Is an ROC-type Response Truly Always Better Than a Binary Response in Observer Performance Studies? *Academic Radiology*, *17*, 639–645. doi:10.1016/j.acra.2009.12.012

Gur, D., & Rockette, H. E. (2008). Performance Assessments of Diagnostic Systems Under the FROC Paradigm. Experimental, Analytical, and Results Interpretation Issues. *Academic Radiology*, *15*, 1312–1315. doi:10.1016/j.acra.2008.05.006

Hadjiiski, L., Chan, H. P., Sahiner, B., Helvie, M. A., & Roubidoux, M. A. (2007). Quasi-Continuous and Discrete Confidence Rating Scales for Observer Performance Studies. Effects on ROC Analysis1. *Academic Radiology*, *14*, 38–48. doi:10.1016/j.acra.2006.09.048

Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging*, *29*, 307–335.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36. doi:10.1148/radiology.143.1.7063747

Hanley, J. A., & McNeil, B. J. (1983). A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases. *Radiology*, *148*(3), 839–843. Retrieved from http://pubs.rsna.org/doi/abs/10.1148/radiology.148.3.6878708

Hara, A. K., Paden, R. G., Silva, A. C., Kujak, J. L., Lawder, H. J., & Pavlicek, W. (2009). Iterative reconstruction technique for reducing body radiation dose at CT: feasibility study. *AJR. American Journal of Roentgenology*, *193*, 764–771. doi:10.2214/AJR.09.2397

Hart, D., Wall, B., Hillier, M., & Shrimpton, P. (2010). *Frequency and collective dose for medical and dental X-ray examinations in the UK, 2008*. *Health Protection Agency HPA-CRCE-012* (pp. 1–52). Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Frequency+and+Collective+Dose+for+Medical+and+Dental+X-ray+Examinations+in+the+UK+,+2008#0

Haygood, T. M., Ryan, J. T., Brennan, P. C., Li, S., Marom, E. M., McEntee, M., Itani, M., Evanoff, M. G., & Chakraborty, D. P. (2012). On the choice of acceptance radius in free-response observer performance studies. *British Journal of Radiology*. doi:10.1259/bjr/42313554

Henschke, C. I. (2011). *International Early Lung Cancer Action Program: Enrollment and Screening Protocol*. Retrieved from http://www.ielcap.org/sites/default/files/ielcap.pdf

Hillis, S. L. (2010). Multireader ROC analysis. In E. Samei & E. A. Krupinski (Eds.), *The Handbook of Medical Image Perception and Techniques* (pp. 204–215). New York: Cambridge University Press.

Hillis, S. L., Berbaum, K. S., & Metz, C. E. (2008). Recent Developments in the Dorfman-Berbaum-Metz Procedure for Multireader ROC Study Analysis. *Academic Radiology*, *15*, 647–661. doi:10.1016/j.acra.2007.12.015

Humphrey, L., Deffebach, M., Pappas, M., Baumann, C., Artis, K., Priest Mitchell, J., Zakher, B., Fu, R., & Slatore, C. (2013). Screening for Lung Cancer With Low-Dose Computed Tomography: A systematic review to update the U.S. Preventative Services Task Force Recommendation. *Annals of Internal …*, *159*(6), 411–420. Retrieved from http://www.atsjournals.org/doi/abs/10.1164/ajrccm.165.4.2107006

International Committee of Medical Journal Editors. (n.d.). Defining the Role of Authors and Contributors. Retrieved July 04, 2014, from http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, *6*, 1307–1322. doi:10.1167/6.11.13

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology. General*, *138*, 291–306. doi:10.1037/a0015525

Jessop, M., Thompson, J. D., Sil, J., Sanderud, A., Jorge, J., Groot, M. de, Lanca, L., & Hogg, P. (2014). Lesion detection performance in an anthropomorphic chest phantom: comparative analysis of low-dose hybrid CT systems. In *International Society of Radiographers and Radiological Technologists*. Helsinki.

Jiang, Y., Metz, C. E., & Nishikawa, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, *201*, 745–750.

Kamel, E., Hany, T. F., Burger, C., Treyer, V., Lonn, A. H., Von Schulthess, G. K., & Buck, A. (2002). CT vs 68Ge attenuation correction in a combined PET/CT system: Evaluation of the effect of lowering the CT tube current. *European Journal of Nuclear Medicine*, *29*, 346–350. doi:10.1007/s00259-001-0698-9

Kang, M. C., Kang, C. H., Lee, H. J., Goo, J. M., Kim, Y. T., & Kim, J. H. (2008). Accuracy of 16-channel multi-detector row chest computed tomography with thin sections in the detection of metastatic pulmonary nodules. *European Journal of Cardio-Thoracic Surgery*, *33*, 473–479. doi:10.1016/j.ejcts.2007.12.011

Kaufmann, P. A. (2009). Cardiac Hybrid Imaging: state of the art. *Annals of Nuclear Medicine*, *23*, 325–331.

Kocher, K. E., Meurer, W. J., Fazel, R., Scott, P. A., Krumholz, H. M., & Nallamothu, B. K. (2011). National trends in use of computed tomography in the emergency department. *Annals of Emergency Medicine*. doi:10.1016/j.annemergmed.2011.05.020

Koepfli, P., Hany, T. F., Wyss, C. A., Namdar, M., Burger, C., Konstantinidis, A. V, Berthold, T., Von Schulthess, G. K., & Kaufmann, P. A. (2004). *CT attenuation correction for myocardial perfusion quantification using a PET/CT hybrid scanner. Journal of nuclear medicine : official publication, Society of Nuclear Medicine* (Vol. 45, pp. 537–542).

Kotre, C. J. (2006). Short communication: an investigation of search pattern extent in the threshold contrast detection task. *The British Journal of Radiology*, *79*, 437–440. doi:10.1259/bjr/13489819

Krupinski, E. A., & Jiang, Y. (2008). Anniversary Paper: Evaluation of medical imaging systems. *Medical Physics*, *35*(2), 645. doi:10.1118/1.2830376

Krzanowski, W., & Hand, D. (2009). *ROC Curves for Continuous Data*. *Hand The* (Vol. 111, p. 241). doi:10.1201/9781439800225

Kundel, H. (2006). History of research in medical image perception. *Journal of the American College of Radiology*, *3*(6), 402–8. doi:10.1016/j.jacr.2006.02.023

Kundel, H., & Nodine, C. (2010). A short history of image perception in medical radiology. In E. Samei & E. A. Krupinski (Eds.), *The Handbook of Medical Image Perception and Techniques* (pp. 9–20). New York: Cambridge University Press.

Legislation.gov.uk. (1998). Data Protection Act 1998. Retrieved from http://www.legislation.gov.uk/ukpga/1998/29/contents

Lusted, L. B. (1960). Logical Analysis in Roentgen Diagnosis. *Radiology*, *74*(2), 178–193. doi:http://dx.doi.org/10.1148/74.2.178

Malkerneker, D., Brenner, R., Martin, W. H., Sampson, U. K. A., Feurer, I. D., Kronenberg, M. W., & Delbeke, D. (2007). CT-based attenuation correction versus prone imaging to decrease equivocal interpretations of rest/stress Tc-99m tetrofosmin SPECT MPI. *Journal of Nuclear Cardiology*, *14*, 314–323. doi:10.1016/j.nuclcard.2007.02.005

Manning, D. J. (1998). Evaluation of diagnostic performance in radiography. *Radiography*, *4*(1), 49–60. doi:10.1016/S1078-8174(98)80030-8

Marten, K., & Grabbe, E. (2003). The challenge of the solitary pulmonary nodule: Diagnostic assessment with multislice spiral CT. *Clinical Imaging*, *27*, 156–161. doi:10.1016/S0899-7071(02)00541-7

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283–298. doi:http://dx.doi.org/10.1016/S0001-2998(78)80014-2

Metz, C. E. (2006). Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *Journal of the American College of Radiology*, *3*, 413–422. doi:10.1016/j.jacr.2006.02.021

Nanavaty, P., Alvarez, M. S., & Alberts, W. M. (2014). Lung cancer screening: advantages, controversies, and applications. *Cancer Control : Journal of the Moffitt Cancer Center*, *21*, 9–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24357736

Nightingale, J. M., & Marshall, G. (2012). Citation analysis as a measure of article quality, journal influence and individual researcher performance. *Radiography*, *18*(2), 60–67. doi:10.1016/j.radi.2011.10.044

Obuchowski, N. A. (2000). Sample size tables for receiver operating characteristic studies. *AJR. American Journal of Roentgenology*, *175*, 603–608. doi:10.2214/ajr.175.3.1750603

Obuchowski, N. A. (2005). Fundamentals of clinical research for radiologists: ROC Analysis. *American Journal of Roentgenology*, *184*, 364–372. doi:10.2214/ajr.184.2.01840364

Park, C. M., Lee, H. J., Goo, J. M., Han, D. H., Kim, J. H., Lim, K. Y., Kim, S. H., Kang, J. J., Kim, K. G., Lee, C. H., Chun, E. J., & Im, J. G. (2008). Comparison of observer performance on soft-copy reading of digital chest radiographs: High resolution liquid-crystal display monitors versus cathode-ray tube monitors. *European Journal of Radiology*, *66*, 13–18. doi:10.1016/j.ejrad.2007.05.023

Patton, J. A., & Turkington, T. G. (2008). SPECT/CT physical principles and attenuation correction. *Journal of Nuclear Medicine Technology*, *36*, 1–10. doi:10.2967/jnmt.107.046839

Patz, E. F., Pinsky, P., Gatsonis, C., Sicks, J. D., Kramer, B. S., Tammemägi, M. C., Chiles, C., Black, W. C., & Aberle, D. R. (2014). Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Internal Medicine*, *174*, 269–74. doi:10.1001/jamainternmed.2013.12738

Pauls, S., Kürschner, C., Dharaiya, E., Muche, R., Schmidt, S. A., Krüger, S., Brambs, H. J., & Aschoff, A. J. (2008). Comparison of manual and automated size measurements of lung metastases on MDCT images: Potential influence on therapeutic decisions. *European Journal of Radiology*, *66*, 19–26. doi:10.1016/j.ejrad.2007.05.022

Pazhenkottil, A. P., Ghadri, J.-R., Nkoulou, R. N., Wolfrum, M., Buechel, R. R., Küest, S. M., Husmann, L., Herzog, B. A., Gaemperli, O., & Kaufmann, P. A. (2011). Improved outcome prediction by SPECT myocardial perfusion imaging after CT attenuation correction. *Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine*, *52*, 196–200. doi:10.2967/jnumed.110.080580

Pollard, B. J., Chawla, A. S., Delong, D. M., Hashimoto, N., & Samei, E. (2008). Object detectability at increased ambient lighting conditions. *Medical Physics*, *35*, 2204–2213. doi:10.1118/1.2907566

Preuss, R., Weise, R., Lindner, O., Fricke, E., Fricke, H., & Burchert, W. (2008). Optimisation of protocol for low dose CT-derived attenuation correction in myocardial perfusion SPECT imaging. *European Journal of Nuclear Medicine and Molecular Imaging*, *35*, 1133–1141. doi:10.1007/s00259-007-0680-2

Robinson, P. J. (1997). Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image. *The British Journal of Radiology*, *70*, 1085–1098. doi:10.1259/bjr.70.839.9536897

Rockette, H. E., & Gur, D. (2008). Selection of a Rating Scale in Receiver Operating Characteristic Studies: Some Remaining Issues. *Academic Radiology*, *15*, 245–248. doi:10.1016/j.acra.2007.10.011

Ru Zhao, Y., Xie, X., de Koning, H. J., Mali, W. P., Vliegenthart, R., & Oudkerk, M. (2011). NELSON lung cancer screening study. *Cancer Imaging : The Official Publication of the International Cancer Imaging Society*, *11 Spec No*, S79–84. doi:10.1102/1470-7330.2011.9020

Rubin, G. D., Lyo, J. K., Paik, D. S., Sherbondy, A. J., Chow, L. C., Leung, A. N., Mindelzun, R., Schraedley-Desmond, P. K., Zinck, S. E., Naidich, D. P., & Napel, S. (2005). Pulmonary nodules on multi-detector row CT scans: performance comparison of

radiologists and computer-aided detection. *Radiology*, *234*, 274–283. doi:10.1148/radiol.2341040589

Rzeszotarski, M. S. (1999). Counting statistics. *Radiographics*, *19*, 765–82. doi:10.1148/radiographics.19.3.g99ma33765

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics : collected papers*. *Scientific psychology series* (p. Chp 11).

Tourassi, G. (2010). Receiver operating characteristic analysis: basic concepts and practical applications. In E. Samei & E. A. Krupinski (Eds.), *The Handbook of Medical Image Perception and Techniques* (pp. 187–203). New York: Cambridge University Press.

Uffmann, M., Prokop, M., Kupper, W., Mang, T., Fiedler, V., & Schaefer-Prokop, C. (2005). Soft-copy reading of digital chest radiographs: effect of ambient light and automatic optimization of monitor luminance. *Investigative Radiology*, *40*, 180–185. doi:10.1097/01.rli.0000153658.15272.91

Vamvakas, I., Hogg, P., Thompson, J. D., Manning, D. J., & Szczepura, K. (2013). The influence of adaptive iterative dose reduction on lesion detection performance within an anthropomorphic chest phantom: a free-response receiver operating characteristic study. In *European Congress of Radiology*. Vienna. doi:10.1594/ecr2013/C-0601

Van Erkel, A. R., & Pattynama, P. M. (1998). Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *European Journal of Radiology*, *27*, 88–94. doi:10.1016/S0720-048X(97)00157-5

Veronesi, G., Maisonneuve, P., Bellomi, M., Rampinelli, C., Durli, I., Bertolotti, R., & Spaggiari, L. (2012). Estimating overdiagnosis in low-dose computed tomography screening for lung cancer: a cohort study. *Annals of Internal Medicine*, *157*, 776–84. doi:10.7326/0003-4819-157-11-201212040-00005

Vining, D. J., & Gladish, G. W. (1992). Receiver operating characteristic curves: a basic understanding. *Radiographics*, *12*, 1147–1154. Retrieved from http://radiographics.rsna.org/content/12/6/1147.full.pdf

Wareing, A., Hogg, P., Thompson, J. D., Manning, D. J., & Szczepura, K. (2013). Lesion detection performance of high-dose and low-dose CT acquisitions on a PET/CT system: a free-response receiver operating characteristic study. In *European Congress of Radiology*. Vienna. doi:10.1594/ecr2013/C-1526

Wells, R. G., Soueidan, K., Vanderwerf, K., & Ruddya, T. D. (2012). Comparing slow-versus high-speed CT for attenuation correction of cardiac SPECT perfusion studies. *Journal of Nuclear Cardiology*, *19*, 719–726. doi:10.1007/s12350-012-9555-4

Wilson, D. O., Weissfeld, J. L., Fuhrman, C. R., Fisher, S. N., Balogh, P., Landreneau, R. J., Luketich, J. D., & Siegfried, J. M. (2008). The Pittsburgh Lung Screening Study (PLuSS): outcomes within 3 years of a first computed tomography scan. *American*

Journal of Respiratory and Critical Care Medicine, *178*, 956–961. doi:10.1164/rccm.200802-336OC

World Medical Association. (2013). World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Journal of the American Medical Association*, *310*(20), 2191–4. doi:10.1001/jama.2013.281053

Wormanns, D., Kohl, G., Klotz, E., Marheine, A., Beyer, F., Heindel, W., & Diederich, S. (2004). Volumetric measurements of pulmonary nodules at multi-row detector CT: In vivo reproducibility. *European Radiology*, *14*, 86–92. doi:10.1007/s00330-003-2132-0

Wright, A. R., Collie, D. A., Williams, J. R., Hashemi-Malayeri, B., Stevenson, A. J., & Turnbull, C. M. (1996). *Pulmonary nodules: effect on detection of spiral CT pitch. Radiology* (Vol. 199, pp. 837–841). doi:10.1148/radiology.199.3.8638014

Wu, T. H., Huang, Y. H., Lee, J. J. S., Wang, S. Y., Wang, S. C., Su, C. T., Chen, L. K., & Chu, T. C. (2004). Radiation exposure during transmission measurements: Comparison between CT- and germanium-based techniques with a current PET scanner. *European Journal of Nuclear Medicine and Molecular Imaging*, *31*, 38–43. doi:10.1007/s00259-003-1327-6

Xia, T., Alessio, A. M., & Kinahan, P. E. (2009). Limits of ultra-low dose CT attenuation correction for PET/CT. In *IEEE Nuclear Science Symposium Conference Record* (pp. 3074–3079). doi:10.1109/NSSMIC.2009.5401665

Xu, X. J., Lou, F. L., Zhang, M. M., Pan, Z. M., & Zhang, L. (2007). Usefulness of low-dose CT in the detection of pulmonary metastasis of gestational trophoblastic tumours. *Clinical Radiology*, *62*, 998–1003. doi:10.1016/j.crad.2007.03.009

Zaidi, H., & Hasegawa, B. H. (2003). Determination of the attenuation map in emission tomography. *Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine*, *44*, 291–315.

Zanca, F., Hillis, S. L., Claus, F., Van Ongeval, C., Celis, V., Provoost, V., Yoon, H.-J., & Bosmans, H. (2012). Correlation of free-response and receiver-operating-characteristic area-under-the-curve estimates: Results from independently conducted FROC/ROC studies in mammography. *Medical Physics*. doi:10.1118/1.4747262

Zou, K. H., Liu, A., Bandos, A. I., Ohno-Machedo, L., & Rockette, H. E. (2012). *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. Boca Raton, USA: Chapman and Hall.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*, 654–657. doi:10.1161/CIRCULATIONAHA.105.594929

Zou, K. H., Tempany, C. M., Fielding, J. R., & Silverman, S. G. (1998). Original smooth receiver operating characteristic curve estimation from continuous data: Statistical

methods for analyzing the predictive value of spiral CT of ureteral stones. *Academic Radiology*, *5*, 680–687. doi:10.1016/S1076-6332(98)80562-X

Appendix A – Paper 1

Thompson JD, Hogg P, Thompson SM, Manning DJ, Szczepura K. ROCView: prototype software for data collection in jackknife alternative free-response receiver operating characteristic analysis. *The British Journal of Radiology* 2012;85:1320-1326.

Appendix B – Paper 2

Thompson J, Hogg P, Szczepura K, Manning D. Analysis of CT acquisition parameters suitable for use in SPECT/CT: A free-response receiver operating characteristic study. *Radiography* 2012;18:238-243.

Appendix C – Paper 3

Thompson J, Hogg P, Higham S, Manning D. Accurate localisation of incidental findings on the computed tomography attenuation correction image: the influence of tube current variation. *Nuclear Medicine Communications* 2013;34:180-184.

Appendix D – Paper 4

Thompson JD, Manning DJ, Hogg P. The value of observer performance studies in dose optimisation: A focus on free-response receiver operating characteristic methods. *Journal of Nuclear Medicine Technology* 2013;41:57-64.

Appendix E – Paper 5

Thompson JD, Hogg P, Manning DJ, Szczepura K, Chakraborty DP. A Free-response Evaluation Determining Value in the Computed Tomography Attenuation Correction Image for Revealing Pulmonary Incidental Findings: A Phantom Study. *Academic Radiology* 2014;21:538-545.

Appendix F – Paper 6

Thompson JD, Manning DJ, Hogg P. Analysing data from observer studies in medical imaging research: an introductory guide to free-response techniques. *Radiography* 2014;20:295-299.

Appendix G – Paper 7

Buissink C, Thompson JD, Voet M, Sanderud A, Kamping LV, Savary L, Mughal M, Rocha CS, Hart GE, Parreiral R, Martin G, Hogg P. The influence of observer training in a group of novice observers: a jackknife alternative free-response receiver operating characteristic analysis. *Radiography* 2014;20:300-305.