AFROC analysis of reporting radiographer’s performance in CT head interpretation

P. Lockwood⁹, K. Piperᵇ

Allied Health Department, Canterbury Christ Church University, Medway Campus, Rowan Williams Building, 30 Pembroke Court, Chatham Maritime, Kent, ME4 4UF, UK

Keywords: Reporting Radiographer, Consultant Radiologist, Computed Tomography, Image Interpretation, Diagnostic Accuracy.
Abstract

Aim: A preliminary small scale study to assess the diagnostic performance of a limited group of reporting radiographers and consultant radiologists in clinical practice undertaking computer tomography (CT) head interpretation.

Method: A multiple reader multiple case (MRMC) alternative free response receiver operating characteristic (AFROC) methodology was applied for this study. Utilising an image bank of 30 CT head examinations, with a 1:1 ratio of normal to abnormal cases. A reference standard was established by double reporting the original reports using two additional independent consultant radiologists with arbitration of discordance by the researcher. Twelve observers from six southern National Health Service (NHS) trusts were invited to participate. The results were compared for accuracy, agreement, sensitivity, specificity. Data analysis used AFROC and area under the curve (AUC) with standard error against the ground truth.

Results: The reporting radiographers results demonstrated a mean sensitivity rate of 88.7% (95% CI 82.3 to 95.1%), specificity 95.6% (96% CI 90.1 to 100%) and accuracy of 92.2% (95% CI 89.3 to 95%). The consultant radiologists mean sensitivity rate was 83.35% (95% CI 80 to 86.7%), specificity 90% (95% CI 86.7 to 93.3%) and accuracy of 86.65% (95% CI 83.3 to 90%). Observer performance between the two groups was compared with AFROC, AUC, and standard error analysis (p=0.94, SE 0.202).

Conclusion: The findings of this research indicate that within a limited study, a small group of reporting radiographers demonstrated high levels of diagnostic accuracy in the interpretation of CT head examinations that was equivalent to a small selection of consultant radiologists.
Highlights

- We assessed reporting radiographers and consultant radiologists in a clinical setting.
- This was a small scale retrospective multi-reader multi-case multi-site study.
- AFROC used lesion-based decisions rather than case-based decisions.
- Within a limited study the observer performance was high in CT head interpretation.
- Parallels were drawn with published results from other CT head interpretation studies.
Introduction

The number of head injured patients attending district general hospitals has been estimated by the United Kingdom (UK) Acquired Brain Injury Forum\(^1\) during 2011-2012 to be around 353,059 UK patients. These figures estimate around 558 per 100,000 of the population experience head injuries each year. This represents a 33.5% increase in the last ten years (10,000-20,000 per year in the UK) of admissions for severe traumatic brain injuries.

Both the National Healthcare Service (NHS) and the Department of Health (DoH)\(^2, 3, 4, 5, 6\) have a strong ethos of developing and improving patient outcomes and service delivery. With the NHS currently undertaking the ‘Nicholson Challenge’ (2006-2015)\(^7\) to generate extra productivity and service quality improvement, set by Sir David Nicholson. Within radiology additional NHS drivers for change include pressures from DoH targets of the acute 4 hour waiting time\(^8\), cancer ‘referral to treatment’ 18 week target waits\(^6\), and the National Diagnostics Imaging Board\(^9\) policies on reporting targets. Specifically within computed tomography (CT) as a modality, National Institute for Health and Clinical Excellence (NICE) guidelines\(^10,11\) of reporting turnaround timeframes for stroke and head injury examinations have changed historic working practices with the need for urgent 30 minute to 1 hour verbal and written CT head reports. This coupled with an increase in the amount of CT examinations that have increased by 33.5% a year since 2008\(^12\) have emphasised the need to re-evaluate how the service delivery can accommodate future pressure.

Barriers to improving current working practices include staff shortages to implement new guidelines, and the current dilemma of implementing a full 7 day service delivery with restricted service capacity. Within diagnostic imaging, the Royal College of Radiologists (RCR) Clinical Radiology Workforce Report\(^12\) recommended a level of 47 consultant radiologists per million of the population for the UK. The reported RCR\(^12\) level for the south of England was 30 per million, the lowest of all regional variations. With a deficit of 210 unfilled NHS consultant radiologist posts in the UK, the RCR\(^12\) report advised that the current consultant radiologist workforce does not meet the required needs of the radiology service demand. The report indicated 85% of UK radiology departments reporting workload was not being adequately completed by the consultant radiologist workforce. The RCR\(^12\) estimated the shortfall in reporting to be 47% of all examinations were left unreported in 2011, whilst recommending the best approach to tackle the deficit has been the adoption of reporting radiographers.

In identifying potential ways to reduce reporting delays and increase service provision, a skills mix of reporting has been promoted and endorsed jointly by the RCR and the Society and College of Radiographers (SCoR)\(^13, 14, 15\). Examples of such an approach have been demonstrated in surveys by
the SCoR\textsuperscript{16, 17} which showed at least 17 NHS trusts in the UK had adopted and supported role
extension of reporting radiographers to supplement their service provision by 2012\textsuperscript{16}. This has
helped to improve service delivery of reporting traumatic injuries and assisted in the early detection
of pathological conditions and cancers\textsuperscript{2, 4, 5}.

**Aims and Objectives**

The study hypothesis predicted reporting radiographers would have a diagnostic accuracy
comparable or equal to consultant radiologists in CT head interpretation in a clinical setting. To
answer the hypothesis, the research study set inter and intra-participant objectives within the study:
Identifying statistical interpretation results for variation or equivalence rates between two groups of
participants (consultant radiologists and reporting radiographers) undertaking the same image bank
analysis.

**Methodology**

The design followed a multiple reader multiple case (MRMC) retrospective study of CT head
interpretation by reporting radiographers ($n=6/6$) and consultant radiologists ($n=2/6$) at 6 NHS
hospitals within the southern region of the UK.

Chang\textsuperscript{18} suggests that any experimental study which evaluates the efficiency of reporting standards
by Bayesian analysis must use an explicitly defined reference standard. The study adopted a
retrospective method using patient cases with known true disease status from a collection of 125
cases previously obtained by the University for teaching and research. This had been additionally
double reported by two independent consultant radiologists. Brealey\textsuperscript{19} and Robinson\textsuperscript{20} advise that
employing a triple approach to obtaining a retrospective reference standard enforces validity of the
reference standard.

Brealey\textsuperscript{19} discusses issues of internal validity of research as the amount and range of presenting
conditions used in the control group (image case bank) in diagnostic performance studies. The CT
head examinations reflected a suitable range of subtle and textbook examples to determine high
levels of accuracy to remove internal validity concerns. Displaying a fair representation of
pathologies as recommended by Robinson et al\textsuperscript{20} and Brealey\textsuperscript{19}, and similar to methods used in
studies by Briggs et al\textsuperscript{21}, McCarron et al\textsuperscript{22}, Erly et al\textsuperscript{23}, Strub et al\textsuperscript{24}, and Gallagher et al\textsuperscript{25}. Concerning
the relative frequency of cases with and without disease in the study sample, Brealey\textsuperscript{19}, Metz\textsuperscript{26},
Brealey and Scally\textsuperscript{27}, Thompson et al\textsuperscript{28}, and Piper, Paterson and Ryan\textsuperscript{29} endorsed a balanced
approach to the ratio of normal to abnormal conditions (1:1).
The test bank was reviewed within the participant’s clinical departments under ambient lighting settings for radiological reporting environments. The images were displayed on a Toshiba Windows Notebook Laptop with a Liquid Crystal Display (LCD) monitor with resolution of 1280x1024. The laptop had been calibrated to the Digital Imaging and Communications in Medicine (DICOM) part 14 Greyscale Standard Display Function (GSDF) with the VeriLUM software programme. Quality checks were performed on the Laptop LCD monitor prior to each test with a standard diagnostic imaging Society of Motion Picture and Television Engineers (SMPTE) reference pattern for spatial uniformity of luminance and temporal luminance stability as recommended by the RCR. An independent PACS system of iQ-View software programme was used to display the cases in a sequential order.

The recruitment criteria of participants required completion of SCoR accredited training and qualification in CT Head reporting, with completion of a period of post training experience of independent reporting within an NHS hospital trust. Obuchowski proposes designs of an MRMC Phase 1 pilot study only requires a small selection of 10-50 cases, of which we choose 30 cases from the bank of 125 cases to be double reported. Obuchowski also suggests in MRMC studies of difficult cases in terms of disease prevalence and appearances should include between 5-10 observers to compare groups of observer’s performance.

Six reporting radiographers were invited to participate (n=6/6 completed the study), and six consultant radiologists were invited to participate in the study (n=2/6 completed the study). Each participant was provided with a copy of instructions detailing the patient history, presenting symptoms, age, gender and referral source, for each case. The participants received these instructions in person by the researcher and were collected after each participant session for compiling of the raw data.

The study required participants to record their findings as either normal or abnormal. If the case was normal they marked the case 0, and moved on to the next case. If the participant deemed the case to be abnormal, they recorded a score of 1-4 (very low to very high confidence of an abnormality) and recorded the name of the pathological condition seen, the anatomical location of the condition/disease and their confidence score of the interpreted pathology. The confidence classification score and free response text allowed the results to be analysed by true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Allowing calculations of accuracy, agreement, sensitivity and specificity using a method adopted by Piper, Ryan and Paterson and Piper, Buscall and Thomas. When considering the accuracy of interpreting radiographic examinations, Obuchowski suggests high accuracy to be 90% (specificity / sensitivity 80%).
Statistical evaluation employed alternative free response receiver operating characteristic (AFROC) curve analysis and area under the curve (AUC) comparison to measure performance. In MRMC studies the use of the AFROC method is ideal when the amount of abnormalities and locations are required to be identified, and ranked each against values according to the confidence levels. Particular attention to the location of the lesion identified to within an acceptance radius (proximity criterion emanating for the centre of the suspected lesion-location (LL) Thompson et al) allowed the researcher to class the participant’s responses as LL (true location of abnormality = TP) or non-lesion (NL) location (wrong location of abnormality = FP or FN).

Chakraborty cautions that the conventional receiver operating characteristic (ROC) paradigm does not distinguish statistical differences for incorrect location (FP), if multiple lesions are present the ROC would classify a TP result even if all the abnormalities were not identified or anatomical location described correctly. Significant clinical implications which may impact on treatment cannot be accounted for in this scenario. Chakraborty advocates AFROC curves over conventional ROC curves, as they provide an increased power due to lesion localization.

Jackknife free-response ROC (AJFROC) calculations were considered for the data analysis but were rejected on the grounds that the output and statistical tests assume paired analysis of two modalities not readers. The use of single modalities violates the assumption of the calculations. Additionally a test run produced a zero score for the incorrect localisation fraction (ILF), thus it had in this instance no power advantage over AFROC analysis.

Conventional ROC plotting generates a curve using the axis of true positive fraction (TPF) in this case sensitivity, versus false positive fraction (FPF) which is calculated as 1-specificity (Thompson et al). AFROC plotting uses a mixture of conventional ROC methodology and free response ROC (FROC) calculations. FROC is a variant of ROC which was designed to reduce the ROC limitations of a binary yes/no answer and instead determine scoring of multiple lesions per case with unlimited location identification (Thompson et al). FROC calculations replace the FPF with non-lesion fractions (NLF) on the x-axis, and number of lesions (lesion location fraction (LLF) on the y-axis. AFROC is a combination of both paradigms and uses LLF on the y-axis (the same as FROC) and FPF on the x-axis (the same as conventional ROC calculations) Thompson et al.

The study was approved by the university research ethics and governance committee and conformed to Section 33 of the UK Data Protection. All the cases had been obtained from a pre-existing DICOM digital teaching library (DTL). The radiology source data (identifying narrative elements including staff names, hospital name, and identifying patient data) had been manually removed to anonymise the images. This practice follows Cosson and Willis guidance from the

Results

The results for the reporting radiographers (n=6, Ranked RR1-RR6) from six NHS hospitals judged against the reference standard are shown in Table 1. The conjectured accuracy predictor by Obuchowski\textsuperscript{37} for intra-observer variability listed high accuracy to be 90% (specificity / sensitivity 80%). For the reporting radiographers, 4 out of 6 scored higher than 90% in accuracy (the lowest score was 88.3%, mean 92.2%), for sensitivity 5 out of 6 scored over 80% (lowest score 78%, mean 88.7%), and for specificity all scored over 80% (lowest score 86.7%, mean 95.6%). Comparison of the AUC was calculated using MedCalc\textsuperscript{41} to obtain individual AFROC plotting (Graph 1 and 2, and Table 2), and a mean AUC value of 0.903 (95% CI 0.835 to 0.948). MedCalc\textsuperscript{41} calculations to produce the AUC used methodology by Metz\textsuperscript{42}, Griner et al\textsuperscript{43} and Zweig and Campbell\textsuperscript{44} which advised would give increased power and sensitivity to the results from this method than from using traditional t-test comparison calculations.

Further calculations using MedCalc\textsuperscript{41} which applied DeLong, DeLong and Clarke-Pearson\textsuperscript{45}, Hanley and Haijian-Tilaki\textsuperscript{46} and Hanley and McNeil\textsuperscript{47, 48} sampling comparison methodology produced a reporting radiographer mean standard error (SE) analysis of 0.020033.

The results for the consultant radiologists (n=2, Ranked CR1-CR2) judged against the reference standard are shown in Table 3 and 4. The consultant radiologists for sensitivity scored 80% and 86.7% respectively, for specificity all scored over 80% (86.7%, and 93.3%), accuracy was judged to be 83.3% and 90%. Comparison of the AUC was calculated using MedCalc\textsuperscript{41} to obtain individual AFROC plotting (Graphs 3 and 4, and Table 3 and 4), and a mean AUC value of 0.888 (95% CI 0.817 to 0.936) and a SE of 0.026. A test of the comparison between the RR and CR AUC and SE, resulted in p=0.9408 and SE 0.202, inferring that the AUC was not statistically different between the cohorts.

Discussion

A common issue with conventional ROC scoring of participants raw data is the potential for degenerative data. Metz\textsuperscript{26} discussed controversies of converting raw data into ROC curve plotting; where the data scale is too discrete and implies it contains degenerate data to produce inappropriate ROC curve shapes and AUC calculations. In response to these practical issues Metz\textsuperscript{26} advised to use AFROC plotting to obtain AUC scores for valid statistical significance in MRMC studies of reader variation. This decision process requires the participant’s results to be scored against the amount of lesions and locations present in the images banks. In this study the participant’s case
bank of 30 CT head examples contained a possible 115 scores of LL or NL with associated location and confidence scores to give an accurate description of the participant’s diagnostic threshold.

Obuchowski\textsuperscript{33} and Chakraborty\textsuperscript{37} recommend ROC curves and AUC as a global measure of accuracy and performance. In pathology interpretation where false negative scores could have significant complications, a high sensitivity (TP rate) and specificity (TN rate) is recommended. Obuchowski\textsuperscript{34} advises the use of sensitivity at a FP score equal to or less than 0.10 (specificity >0.90). This high level of sensitivity and specificity in ROC studies has been set to a standard that reflects the seriousness of the interpretation of pathology on patient outcomes and treatments (avoidance of surgery or other diagnostic tests, hospital stay, or abandonment of clinical treatment). Fineburg et al\textsuperscript{49}, Fryback and Thornbury\textsuperscript{50} and Brealey\textsuperscript{19} emphasize the interpretation of imaging in the chain of clinical efficacy must set high standards to reduce the risk of error and harmful patient outcomes.

Six electronic databases (Cochrane, Medline, Europe Pubmed Central, CINAHL, ScienceDirect and Google Scholar) were searched to find comparative CT head interpretation studies. The literature search located 45 papers; only one non-peer review journal paper displayed the results of a reporting radiographer’s CT head interpretation study\textsuperscript{51}. The paper did not provide sufficient details as to the methodology, data, sample size or statistical analysis used, although the limited results displayed a high sensitivity and specificity.

The review of literature evaluating consultant radiologist’s interpretation of CT head scans allowed analysis of the summary estimates to calculate a broad estimation of the combined results. The most statistically detailed study found was Erly et al\textsuperscript{52} who studied 15 consultant radiologists reviewing 716 CT head scans (649 were normal). The results produced an agreement level of 95%, sensitivity 85.7%, specificity 99.7% and accuracy of 99.4%.

Further published studies found limited statistical details on diagnostic thresholds for consultant radiologist’s interpretation CT head examinations. Nagaraja et al\textsuperscript{53}, studied 6 consultant radiologists reviewing 270 paediatric CT head examinations of subtle fractures and congenital abnormalities, found 84.1% agreement and 15.9% disagreement. Le et al\textsuperscript{54} on the findings of 10 consultant radiologists reviewing 1,736 cases of which 48 were reported as discordant, gave a concordance rate of 97.2%. A similar study by Briggs et al\textsuperscript{21} produced a 66% agreement and 44% discordance rate. McCarron et al\textsuperscript{22} studied 9 consultant radiologists reviewing 77 CT head examinations, obtained an agreement of 86.6%. Schriger et al\textsuperscript{55} used a multiple site, multiple case methodology of 36 consultant radiologists reviewing 56 CT scans established an accuracy of 83%.
When considering the accuracy of interpretation, Obuchowski\textsuperscript{34} suggests high accuracy to be 90% (specificity / sensitivity 80%). The literature search and analysis provided a reasonable estimation of consultant radiologists from the published literature reviewed studies. The averaged estimated consultant radiologist reference standard was 83% accuracy, and 85.5% agreement (95% CI 73.0 to 97.0%, p<0.271) from results by Schriger et al\textsuperscript{55}; Erly et al\textsuperscript{52}; Le et al\textsuperscript{54}, Briggs et al\textsuperscript{21}; Nagaraja et al\textsuperscript{53} and McCarron et al\textsuperscript{22}. The literature showed that the majority of consultant radiologist study results had not supplied sufficient data to accurately calculate a pooled sensitivity or specificity for consultant radiologists. In comparison from the limited small sample of observers which is not generalizable to the greater population, our preliminary study found reporting radiographers mean accuracy to be 92.2%, and from the small sample of consultant radiologists 86.6%, which is above the mean of the published literature.

**Conclusion**

The overall aim of this limited scale preliminary research was to achieve an understanding of the degree of image interpretation accuracy of a small sample of CT head reporting radiographers and consultant radiologists in a clinical environment. Particularly, the relationship between the calibre of results (intra observer analysis) and in comparison (inter observer analysis) to each other and the published consultant radiologists diagnostic threshold.

The study findings suggested that a small sample of reporting radiographers displayed a high level of accuracy in the interpretation of CT head examinations, which was equivalent to a small sample of consultant radiologists, and were consistent with the published findings of other studies in this field. It is recommended further funded research needs to be undertaken to establish the degree of accuracy of a larger sample of participants. Further research would also encourage debate on role extension of radiographers reporting, and foster discussion on improving and modernising the workforce roles for future service delivery within this modality.
References


30. IMAGE Smith, Inc. (VeriLUM (Version 5.2.1); 2008. Available at: http://verilum.software.informer.com/


32. IMAGE Information Systems. iQ-View (Version 2.8.0); 2013. Available at: http://www.k-pacs.net/10555.html


Acknowledgements

The authors would like to thank all the radiographers and radiologists who took part in interpreting the image banks in this study, and for their time and dedication to the study.

Funding

There are no financial conflicts of interest.

Conflict of interest statement

The author a is a programme leader and author b programme director for the postgraduate CT head reporting course at CCCU.
### Tables

<table>
<thead>
<tr>
<th>Participant</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>FP</th>
<th>FN</th>
<th>TP</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR1</td>
<td>96.6</td>
<td>91.8</td>
<td>94.1</td>
<td>1.25</td>
<td>1.25</td>
<td>14.25</td>
<td>14</td>
</tr>
<tr>
<td>RR2</td>
<td>88.1</td>
<td>98.4</td>
<td>93.3</td>
<td>0.25</td>
<td>1.75</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>RR3</td>
<td>78</td>
<td>98.4</td>
<td>88.3</td>
<td>0.25</td>
<td>3.25</td>
<td>11.5</td>
<td>15</td>
</tr>
<tr>
<td>RR4</td>
<td>91.7</td>
<td>86.7</td>
<td>89.2</td>
<td>2</td>
<td>1.25</td>
<td>13.75</td>
<td>13</td>
</tr>
<tr>
<td>RR5</td>
<td>90</td>
<td>100</td>
<td>95</td>
<td>0</td>
<td>1.5</td>
<td>13.5</td>
<td>15</td>
</tr>
<tr>
<td>RR6</td>
<td>88.1</td>
<td>98.4</td>
<td>93.3</td>
<td>0.25</td>
<td>1.75</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>88.75</strong></td>
<td><strong>95.61</strong></td>
<td><strong>92.2</strong></td>
<td><strong>0.66</strong></td>
<td><strong>1.79</strong></td>
<td><strong>13.66</strong></td>
<td><strong>14.5</strong></td>
</tr>
</tbody>
</table>

Table 1 Reporting radiographers results compared to the study reference standard

[Graph of Lesions Detected Fraction vs False Positive Fraction for RR1 (AUC 0.975)]

[Graph of Lesions Detected Fraction vs False Positive Fraction for RR2 (AUC 0.91)]
Lesions Detected Fraction

False Positive Fraction

RR3 (AUC 0.84)

RR4 (AUC 0.913)

RR5 (AUC 0.87)
Graph 1 Reporting radiographer AFROC and AUC results.

<table>
<thead>
<tr>
<th>Radiographer</th>
<th>AUROC</th>
<th>SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR1</td>
<td>0.975</td>
<td>0.0110</td>
<td>0.927 to 0.995</td>
</tr>
<tr>
<td>RR2</td>
<td>0.910</td>
<td>0.0193</td>
<td>0.842 to 0.955</td>
</tr>
<tr>
<td>RR3</td>
<td>0.840</td>
<td>0.0234</td>
<td>0.760 to 0.902</td>
</tr>
<tr>
<td>RR4</td>
<td>0.913</td>
<td>0.0252</td>
<td>0.845 to 0.902</td>
</tr>
<tr>
<td>RR5</td>
<td>0.870</td>
<td>0.0220</td>
<td>0.795 to 0.925</td>
</tr>
<tr>
<td>RR6</td>
<td>0.910</td>
<td>0.0193</td>
<td>0.842 to 0.955</td>
</tr>
<tr>
<td>Mean</td>
<td>0.975</td>
<td>0.0365</td>
<td>0.835 to 0.939</td>
</tr>
</tbody>
</table>

Table 2 Reporting radiographer AFROC results
Graph 2 Comparison of reporting radiographer AFROC results.

Table 3 Consultant radiologist results compared to the study reference standard

<table>
<thead>
<tr>
<th>Participant</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>FP</th>
<th>FN</th>
<th>TP</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR1</td>
<td>80</td>
<td>86.7</td>
<td>83.3</td>
<td>2</td>
<td>3</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>CR2</td>
<td>86.7</td>
<td>93.3</td>
<td>90</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Mean</td>
<td>83.35</td>
<td>90</td>
<td>86.65</td>
<td>1.5</td>
<td>2.5</td>
<td>12.5</td>
<td>13.5</td>
</tr>
</tbody>
</table>

Table 4 Consultant radiologist AFROC results

<table>
<thead>
<tr>
<th></th>
<th>AUROC</th>
<th>SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR1</td>
<td>0.831</td>
<td>0.0338</td>
<td>0.749 to 0.864</td>
</tr>
<tr>
<td>CR2</td>
<td>0.945</td>
<td>0.0182</td>
<td>0.886 to 0.979</td>
</tr>
<tr>
<td>Mean</td>
<td>0.888</td>
<td>0.026</td>
<td>0.817 to 0.921</td>
</tr>
</tbody>
</table>
Graph 3 Consultant radiologist AFROC results.
Graph 4: Comparison of consultant radiologist AFROC results.