The effect of reporting environment factors and double reading on the intra-observer concordance of major diagnostic plain radiograph evaluation features

Poster No.: C-1893
Congress: ECR 2014
Type: Scientific Exhibit
Authors: T. Fitzgerald, E. Barnhill, S. Mirsadraee; Edinburgh/UK
Keywords: Professional issues, Plain radiographic studies, Digital radiography, PACS, Observer performance, Perception image, Medico-legal issues, Education and training
DOI: 10.1594/ecr2014/C-1893

Any information contained in this pdf file is automatically generated from digital material submitted to EPOS by third parties in the form of scientific presentations. References to any names, marks, products, or services of third parties or hypertext links to third-party sites or information are provided solely as a convenience to you and do not in any way constitute or imply ECR's endorsement, sponsorship or recommendation of the third party, information, product or service. ECR is not responsible for the content of these pages and does not make any representations regarding the content or accuracy of material in this file.

As per copyright regulations, any unauthorised use of the material or parts thereof as well as commercial reproduction or multiple distribution by any traditional or electronically based reproduction/publication method is strictly prohibited.

You agree to defend, indemnify, and hold ECR harmless from and against any and all claims, damages, costs, and expenses, including attorneys' fees, arising from or related to your use of these pages.

Please note: Links to movies, ppt slideshows and any other multimedia files are not available in the pdf version of presentations.

www.myESR.org
Aims and objectives

Identification and auditing of error and variation in plain radiograph reporting is important because of the volume of work and paucity of feedback\(^1\).

It has not been established how much scope for improvement in plain radiograph reporting accuracy exists at an experienced radiologist level, nor by which method(s) this might be effectively procured.

The primary aim of this study was to evaluate the effect of short-interval double-reading and simultaneously optimising the working environment for the second read, on one individual's plain radiograph interpretation accuracy, and internal validity.

Methods and materials

A single radiologist with 23 years consultant experience double read 488 patient's radiographs from general departmental reporting work lists within a four month period. Examinations were performed for all body parts although this was heavily weighted to chest radiographs comprising 65% of the total. The only exclusions from the analysis were all examinations comprising a repeat evaluation for essentially the same pathology within the same disease interval admission or episode.

Both readings were made on identical dedicated departmental picture archiving communication system reporting (PACS) workstations. The presentation of technically identical exams herein ensured any differences between readings were observer-derived, whether due to environmental or personal factors, or a combination.

An initial read followed by an interval (at least 24 hours later as a practical period in a clinical setting) reading was carried out. Comparison to the previous interpretation and the subsequent reference standard was made.

(see Figure 1 Flow diagram demonstrating the study design).

The first reading took place in shared reporting station areas. First reports were verbally dictated and transcribed onto a provisional typed report platform initially, and any exigent or unexpected results were communicated orally at that stage as is usual practice.

The radiographs were then interval re-read by the same individual at verification of the initial transcribed report at least 24 hours later. This second read occurred outside
standard working hours to optimise quiet, environmentally standardised, conditions without interruption nor time pressure. Second reading emphasis was on de-focusing from previously dictated features towards complete review of the image(s).

The presenting clinical details were also available at the second reading. The first read transcribed report was revised after the documentation of findings from the second read when the clinical report was verified for access by referring clinicians. Any diagnostic feature changes inserted into the original text were noted separately on a proforma.

Follow-up at one year principally using electronic patient record (EPR) clinical notes and imaging data was the reference standard. Any radiograph feature even visible only in retrospect was documented and enumerated.

Demographic and technical information recorded for data analysis also included: age, gender, a combined subject and imaging process 'contrast achievement score' and also a 'radiograph complexity score'.

The 'contrast achievement score' of the examination- was subjectively and pragmatically ordinally scored 1-3 through determination of a combination of patient contrast and technical quality achievement. The patient contrast was assessed relative to bone detail/mineralisation initially, but the ability to co-operate, the standard achieved on previous examinations, and/or any deleterious features due to pathology or patient non-co-operation which may have impacted adversely on technical quality within were also included in the scoring. Inadequate scored 1, adequate for primary clinical purpose scored 2 and optimal for all-purposes scored 3.

The 'radiograph complexity grade' of 1-3 was similarly denoted to account for the examination type related to the usual degree of diagnostic difficulty encountered within that image dataset. A chest x-ray was scored 3 due to the 'multi-task' of viewing for perception three quite different densities namely air, bone and soft tissue. Intra-thoracic structures visible within each of these density-differentials require sequentially separate focusing periods as part of the hard-working and cognitively-laden second analytical phase of image detail perception and interpretation beyond any initial (rapidly-formed) impression. This second phase is required to satisfy diagnostic expectations; for example a chest film always remains a valid, if recalcitrantly vulnerable, test for neoplasia beyond the initial diagnostic query.

An abdomen or axial skeletal examination scored as 2 reflects less complexity from the number of features and densities usually considered.

All simpler or single issue views scored 1 on the scale (for example Musculo-skeletal radiographs).

Clinical outcome for each case was determined at one-year follow-up. This was the reference standard, determined by the individual case circumstances. The standard
comprised a judgment based on disease clinical features and course usually through the EPR; positively by electronic system entry of a clinical record or subsequent definitive imaging, or by negative inference where appropriate. Less frequently an experienced specialist peer reviewer re-read the examination (3.5%) In one case (Figs 2 and 3) an outpatient visit occurred just beyond the one year mark which replaced one of the peer review values.

Data collection also included a score for the veracity/relevance of available clinical information data which had been supplied at referral, determined at the one year follow up. A score of 1 was entered for inadequate or misleading information, 2 for adequate information for the purpose and 3 for adequate to excellent information for all purposes normally subsumed in the broadest utility for that examination.

Statistical analysis:

Descriptive analysis included calculation of the frequencies, means and standard deviations (SD). The Kappa statistic was used as a measure of agreement between different rounds of image interpretation and the reference standard. The K value was interpreted as follows

<table>
<thead>
<tr>
<th>Value of K</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.20</td>
<td>Poor</td>
</tr>
<tr>
<td>0.21 - 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 - 0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61 - 0.80</td>
<td>Good</td>
</tr>
<tr>
<td>0.81 - 1.00</td>
<td>Very good</td>
</tr>
</tbody>
</table>

All statistical analysis was performed SPSS Statistics for Windows v10.

The impact of the second viewing on accuracy of diagnosis as compared to the reference standard was measured using McNemar’s Chi-squared test with continuity correction.

In addition, each scan was rated with the following factors:

# Age (mu=65, sigma=21)

# Sex (50% male)

# Mineralisation, degree of bone detail (rated 1-3)

# Complexity of the examination (rated 1-3)

# Technical quality of the examination (rated 1-3)
# Quality of patient history information (rated 1-3)

The relationships of these variables to the diagnostic outcomes were analysed using ordinal logistic regression. The predictive power of the model was measured using McFadden's pseudo $R^2$. Non-significant variables were eliminated through backward stepwise regression to a minimum adequate model after it was confirmed that the stepwise regression had minimal effect on p-values of significant variables or the coefficient of determination. Regression analysis was performed for three different dependent variables:

# Errors in first reading

# Errors in second reading

# Improvement from first to second reading

A p-value of $< 0.05$ was accepted as significant. All statistical analysis was performed using R Studio for Linux.\(^5\)

**Images for this section:**

**Fig. 1:** Flow diagram
Results

- The percentage discordant feature rate from the reference method calculated on the first read of 84/489 (17.2%) was approximately halved on the second read to 39/489 (8.0%).
- There was very good (k>0.8) agreement in the number of abnormalities seen on each reading compared to the reference standard.
- Results from McNemar's test are shown in Table 1. The second viewing improved diagnostic accuracy from 84% to 92% and the test was significant with p < 10^{-6}.
- Minimal adequate models from the ordinal linear regression are shown in Table 2. Sex showed significant correlation for two of the dependent variables, and quality of patient history showed significant correlation for two of the dependent variables. However, as the great majority of the diagnoses were accurate across all factors, the explanatory power of the models was extremely limited. In each model the coefficient of determination was less than 0.04.
- Table 3 Rated values-numbers within categories 1-3 for quality of clinical history, contrast of subject and image viewed (percentages).
- Figures 2-12 illustrate five case examples of True and False imaging results at various stages of the study readings.

Table 1. McNemar's test result.

<table>
<thead>
<tr>
<th></th>
<th>First Viewing</th>
<th>Second Viewing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matches Reference</td>
<td>410</td>
<td>451</td>
</tr>
<tr>
<td>Standard Does not match gold standard</td>
<td>78</td>
<td>37</td>
</tr>
</tbody>
</table>

Table 2. Minimum adequate models for diagnostic results

**Error - First Reading**

<table>
<thead>
<tr>
<th>Factor</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>-2.17</td>
<td>0.032*</td>
</tr>
<tr>
<td>Quality of patient history</td>
<td>-2.26</td>
<td>0.012*</td>
</tr>
</tbody>
</table>

**McFadden Pseudo-$R^2$**: 0.021

**Error - Second Reading**
Factor  |  t-value  |  p-value
---|---|---
Sex  | -1.16  | 0.032* 

_McFadden Pseudo-$R^2$ : 0.0067

**Improvement from First to Second Reading**

Factor  |  t-value  |  p-value
---|---|---
Sex  | -1.49  | 0.13 
Quality of patient history  | -3.02  | 0.0025** 

_McFadden Pseudo-$R^2$ : 0.031

Table 3 Ratings table (n=488)

<table>
<thead>
<tr>
<th>Rating score</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient contrast score (%)</td>
<td>45 (9.2)</td>
<td>271 (55.5)</td>
<td>172 (35.3)</td>
</tr>
<tr>
<td>Radiograph quality achieved (%)</td>
<td>17 (3.5)</td>
<td>205 (42)</td>
<td>266 (54.5)</td>
</tr>
<tr>
<td>Patient information quality (%)</td>
<td>178 (36.5)</td>
<td>168 (34.4)</td>
<td>142 (29.1)</td>
</tr>
</tbody>
</table>

_Images for this section:_
Fig. 2: Case 1 Image 1 True Positive calcaneal fracture noted on second read-after first read False Negative. Later confirmed true positive at 1 year image on Figure 3. Inferior arrow at perceived fracture line, superior arrow at effusion (not unexpected as active rheumatoid arthritis) and some talar dome collapse, measured against calcaneal height posteriorly.
**Fig. 3:** Case 1 Image 2 Later interval radiological confirmation-just after one year follow-up. Clinically severe loss of functionality for one year as a result of calcaneal compression injury. Note inferior arrows at fractured cortex at inferior aspect of fracture line, further calcaneal collapse measured in side bar and smaller effusion outlined by superior arrow.
Fig. 4: Case 2 Image 1 Chest x-ray. First and second reads designated False Negative.
Fig. 5: Case 2 Image 2 Subsequent coronal CT - confirming paraspinal mass secondarily from infective discitis. 6 days delay
**Fig. 6:** Case 2 Image 3 Sagittal MR-confirming paraspinal mass was secondarily from infective discitis. 6 days delay
**Fig. 7:** Case 3 Image 1 Initial True Positive for possible destructive bone lesion, altered to False Negative on second read.

**Fig. 8:** Case 3 Image 2 Axial MR-Large destructive neoplasm predominantly of right side of sacrum.
Fig. 9: Case 4 Image 1 False Negative first read, altered to True Positive for dense ribs on second read
**Fig. 10:** Case 4 Image 2 Pelvic radiograph. Subsequent confirmation of prostatic carcinoma secondaries.
Fig. 11: Case 4 Image 3 Isotope bone 'superscan'. Subsequent confirmation of prostatic carcinoma secondaries to bone—note no renal uptake visible.
**Fig. 12:** Case 5 False Negative on first read changed to True Positive—bone destruction at superior sacral margin of right sacro-iliac joint discerned on second read.
Conclusion

The task in plain radiograph reporting involves scrutiny for perception and interpretation within large sets of non-unambiguous data within the attenuation image of the patient. The high volume of plain radiograph work and very low levels of feedback (to calibrate rules and cues) are important factorials; hence it is unsurprising that human fallibility inherent in visual/radiological tasks is universal and sometimes profound.\textsuperscript{1, 6, 7}

Wide intra-observer kappa value ranges for separate individuals indicate that any strategic improvement applicability may also be individual-specific, related to personality type and simultaneously suggests an opportunity to tap this potential towards improving accuracy.\textsuperscript{1}

The constituent components of the reality of reporting imperfection, namely variation and error\textsuperscript{7, 8} are not distinguished herein and are merely combined within a 'discordance rate' (or a reciprocal, the concordance rate), resolving any inference of 'blame'.

Interval re-reading of radiographs which were technically identically presented ensured that any variation between readings would only be observer-derived, whether that was due to environmental and/or personal factors. A delayed second read was both an enhancement to the normal process of service provision and provided opportunity for collection of this natural experimental dataset via applying the same experienced consultant radiologist eye twice.

Performance in the detection of lung cancer is degraded as viewing time decreases illustrating the 'hard-work' phase of reporting and overcoming recalcitrant 'satisfaction of search' issues.\textsuperscript{7} Heterogeneity is noted amongst observer-participants who are time-constrained; half deteriorated with time constrained while others improved markedly.\textsuperscript{9}

Accuracy rates within performance studies are necessarily constructed artificially with the laudable intent of scientific-modelling validity.\textsuperscript{10} Over-emphasis on accuracy may not be entirely appropriate given the prevalent in-exactitude when dealing with real-life patients whose problems are a complex 'mix'; often multiple features of one disease may combine with equally important co-morbidities.\textsuperscript{11} The mortality rate in this current study was a considerable 25% at one-year follow-up confirming a high prevalence of disease in the study population. Thus being...
"right" is not the whole story. Safeguards both in the clinical areas and within discussion of the diagnostic differential and case-management recommendation content of radiology reports, assist safe handling even if this seems contrary to the crux of the Radiology report. Reflecting the protean nature of imaging and of the varying course of the pathologies sought or fallen upon, radiologists are trained and expected to 'call' a small percentage of false positives as intelligent, good practice within 'performance'.

The genesis and maintenance of expertise remains elusive to define let alone measure. Pragmatism and variation pervade even its definition. Recognition as an expert may comprise 'individuals with [the] commitment to [being able to reliably independently report a chest or other plain radiographic study] the task. Others may require more specifically expert 'attributes' of discipline in strategy, experience, the pre-emptive acquisition of new knowledge, and context awareness. Nonetheless from the literature on behavioural decision theory for expert performance (often disappointing) or by cognitive scientific study of the enhanced processes of experts, the 'street-level' of radiological diagnosis, visual tracking and on signal detection theory, the gestaltic mystery of the human radiological eye and brain's unsurpassed ability to operate as well as it does within such a complex domain remains unsolved.

Misdiagnosis might not be seen as an opportunity to learn if accuracy is considered too important and an intolerance of consideration of error has developed. At the other extreme overexposure or over-reaction to what's unexpectedly discovered at follow up can imbue a 'creeping determinism' which may diminish or trivialise the investigation of past and present events.

A better clinical history rating assisted increasing accuracy on second reading herein. Real-life factors encountered, as here, of insufficient, irrelevant or misleading clinical input information seem deleterious to radiological processing even beyond priming perception for relevant abnormality; examinations have to be appropriate to, and of sufficiently high quality for, the 'real' clinical situation. Indeed, as one of the expert attributes is contextual awareness, this is an unsurprising finding whether considered theoretically or via some artificial test studies.

Beyond human factors, for improvement in healthcare the task complexity, team deficiencies, environmental and organisational problems are omni-present barriers. From these items, failures emerge unpredictably in time, frequency and severity. While it is useful to compare top-down safety management with
bottom-up error elimination strategies\textsuperscript{19,20}, the double read/report is a useful and proven management strategy for preventing errors.\textsuperscript{7,8}

In summary, this study demonstrates variations in the accuracy of plain radiograph readings by an experienced radiologist in two different episodes, one performed in a non-optimal condition, the other performed in optimal reading conditions. Whilst the results confirmed very good agreement in the number of abnormalities seen on each reading compared to the reference standard, there was a significantly improved diagnostic accuracy from 84\% to 92\% in the second episode. This demonstrates the importance of work environment in the accuracy of reports generated by an experienced radiologist. Personality type, experience, double reading intervention and the quality of reporting facilities/conditions will comprise important constituent factors in this individual radiologist’s diagnostic accuracy performance in plain radiograph interpretation.

Ethical considerations-this was a second reporting enhancement to usual practice. All cases illustrated here were presented and discussed at our discrepancy meeting.\textsuperscript{19} Double reading has been adopted for all subsequent interpretive reporting by the index radiologist.

Personal information

References

1 Tudor GR, Finlay DB. Is there an improvement in performance on radiographs that are re-reported at 24 hours Br. J. Radiol. 1999; 72: 465-468.


11 Gunderman RB, Nyce JM. Tyranny of Accuracy in Radiologic Education. Radiology 2002; 222:297-300


13 Norman GR, Coblenz CL Brooks LR et al Expertise in Visual Diagnosis-a review of the Literature, Academic Medicine 1992; 67: 10; S78-83


20 Flach J M For those condemned to live in the future, Commentary Qual Saf Health Care 2003;12:311-312 doi:10.1136/qhc.12.4.311