Reporter confidence of an AI-assisted PET/CT reading workflow in pre-treatment assessment of high-grade lymphoma : multi-centre reader study

ECR 2023 - electronic poster presentation

R. Frood[1,2], J. M. Y. Willaime[3], B. Miles[4], N. Brooks[4], P. Strouhal[4], C. Patel[1], A. Scarsbrook[1,2]

1. Department of Radiology, Leeds Teaching Hospitals NHS Trust, Leeds, UK
2. Leeds Institute of Health Research, University of Leeds, Leeds, UK
3. Mirada Medical Ltd., Oxford, UK
4. Alliance Medical Ltd., Warwick, UK

Purpose or Learning Objective

The incidence of lymphoma is increasing worldwide [1] with Fluorine-18 fluorodeoxyglucose (FDG)-positron emission tomography-computed tomography (PET-CT) being the gold standard imaging technique for staging and response assessment of high-grade lymphoma [2]. The time taken to report each case varies on the complexity of the case, the quality of the images and the reporter's experience. Given the increasing workload and the global radiology workforce crisis, artificial intelligence (AI) assisted reading may offer the ability to improve efficiency and accuracy of reporting [3, 4]. However, what is not clear is how AI may affect the reporter's decision making particularly when being presented with flawed data.

The aim of the study was to assess the effect on reporter confidence and ability to identify mistakes when using prototype PET-CT reading software (Mirada Medical Ltd., UK) incorporating AI-assistance.

Methods or Background

Fifteen adult high-grade lymphoma cases were retrospectively selected from PET/CT studies performed between January 2008 and January 2020 for staging purposes at a large single tertiary centre. The cases consisted of stage II, stage IIE/stage IE and stage IV disease as well as cases with prominent physiology uptake due to metformin-related bowel activity or brown fat activity. These were chosen to represent a mix of routinely encountered scenarios within real-world clinical practice as well as providing situations were under or over-estimation of disease could affect patient management.

Nine blinded reporters (three trainees, three junior (<5 years' experience) and three senior consultants (>5 years' experience) from three centres within the UK participated in the reader study. Each reporter read the fifteen cases using a standard workflow, which closely mimicked their standard clinical environment using the Alliance Medical PET/CT reporting platform and a research prototype built on Mirada XD configured with their usual user preferences. Then after a 6-week washout period the same cases, presented in a different order and with different case numbers, were re-read with the same Alliance Medical

PET/CT reporting platform and Mirada research prototype. However, this time the AI-assisted module was enabled to display pre-segmented sites of disease (**Fig 1**).

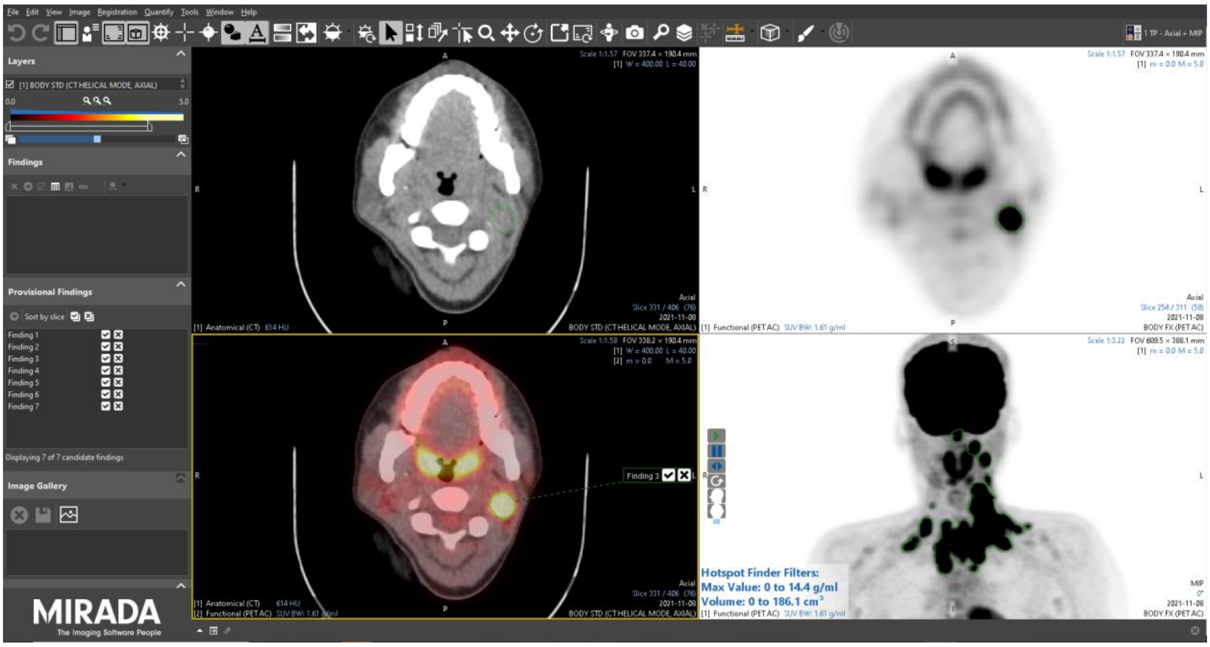

File Edit View Image Registration Quantify Tools Window Help
Layers
[1] BODY STD (CT HELICAL MODE, AXIAL)
0.0
5.0
Findings
Provisional Findings
Sort by slice
Finding 1
Finding 2
Finding 3
Finding 4
Finding 5
Finding 6
Finding 7
Displaying 7 of 7 candidate findings
Image Gallery
MIRADA
The Imaging Software People
[1] Anatomical (CT) 614 HU
[2] Functional (PET AC) SUV BW: 1.61 g/ml
BODY STD (CT HELICAL MODE, AXIAL)
[1] Functional (PET AC) SUV BW: 1.61 g/ml
BODY FX (PET AC)
Axial
Slice 331 / 406 (76)
2021-11-08
Finding 3
A
P
R
L
Hotspot Finder Filters:
Max Value: 0 to 14.4 g/ml
Volume: 0 to 186.1 cm³
MIP
0°

Fig 1: Screenshot of the Mirada Medical research prototype with the AI-assisted module enabled.

An even split of PET/CTs with gold standard (GS) 5/15, false positive (FP) 5/15, and false negative (FN) 5/15 segmentations were provided as part of the AI assisted read. Examples of the FN and FP cases are provided in **Fig 2**. Following evaluation of each case, the reporters completed an online questionnaire (https://www.sogolytics.com) documenting their confidence in identifying all disease sites and, for the AI-assisted read, how confident they felt the AI-assistance was in identifying the full extent of disease and if there were any FP or FNs. They were also asked if they felt biased by the segmentations presented to them. Significance was calculated using Wilcoxon ranked-signed and Mann-Whitney U tests with a p-value of less than 0.05 regarded as being significant.

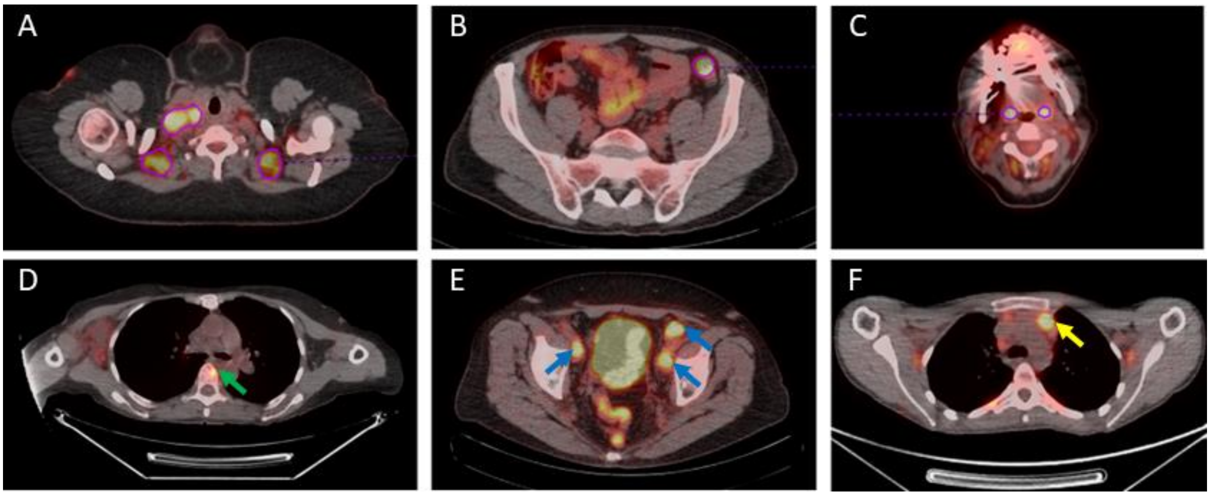


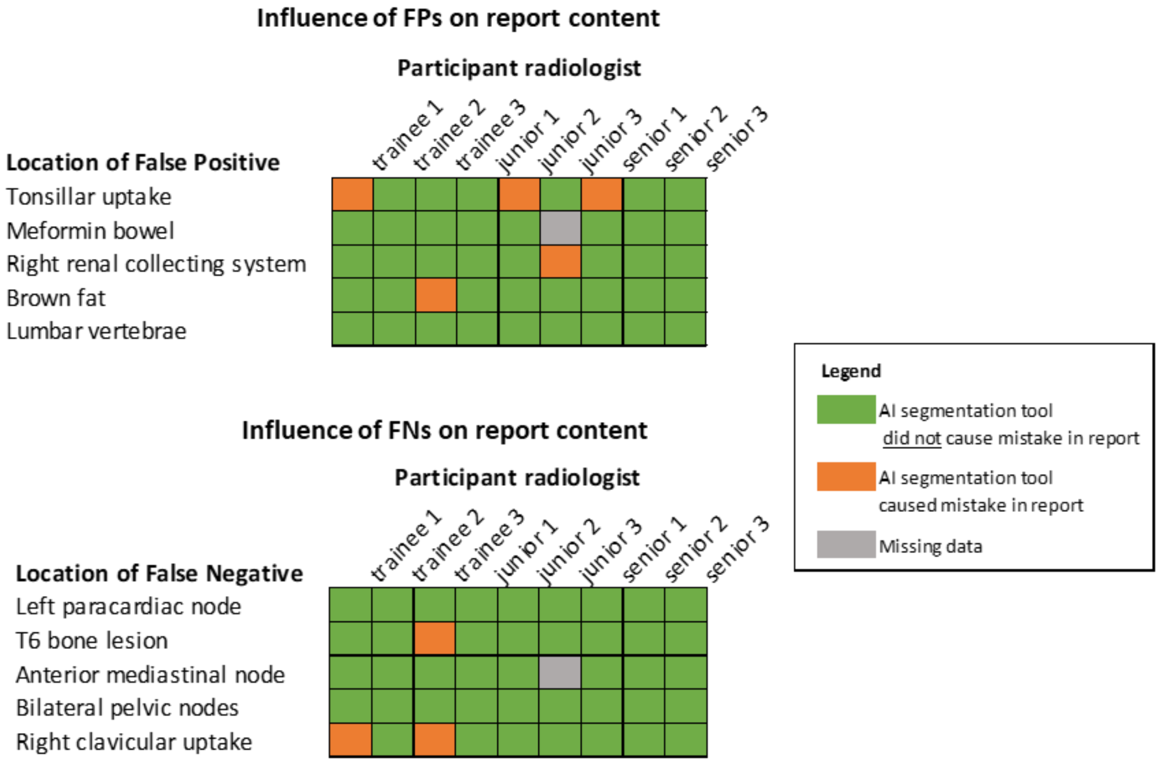
Fig 2: Examples of false positive (FP) (A-C) and false negative (FN) (D-F) segmentations. D demonstrates the missing inclusion of osseous disease within the vertebral column (green arrow), E demonstrates missing pelvic side wall lymph node segmentation (blue arrow) and F demonstrates missing anterior mediastinal lymphadenopathy (yellow arrow)

Results or Findings

All but one reporter completed a questionnaire for each case, one junior reporter completed 13/15 of the AI-assisted read questionnaires. Only cases where both the non-AI and AI-assisted read questionnaires were included in the analysis (trainees = 45 cases, junior consultants = 43 cases, senior consultants = 45 cases).

There was a significant increase in self-reported confidence in disease identification with AI-assistance compared to the non-AI assisted read (median 8/10 vs 7/10, $p<0.001$). Although participants' confidence in the AI-assistance tool significantly decreased when comparing the GS and FN segmentations (median 8 vs 6, $p<0.001$). There was no significant difference between GS and FP cohorts.

6/9 participants reported that they had been biased by the AI-assisted segmentations. However, in 91% cases (80/88), the FP/FN contours did not influence report content compared to baseline. When specifically asked about FN findings, 3/15 (20%) segmentations were not identified as being erroneous by the trainee reporters, whereas all the senior and junior reporters identified all the disease missed by the provided segmentations (**Fig 3**). Regarding the FP findings, trainees misinterpreted 2/15 (13%) cases, junior reporters misinterpreted 2/14 (14%) cases and senior reporters misinterpreted 1/15 (7%) of the erroneous segmentations as being correct (**Fig 3**).



Influence of FPs on report content

Participant radiologist

Location of False Positive	trainee 1	trainee 2	trainee 3	junior 1	junior 2	junior 3	senior 1	senior 2	senior 3
Tonsillar uptake	🟧	🟩	🟩	🟩	🟧	🟩	🟧	🟩	🟩
Meformin bowel	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩	🟩
Right renal collecting system	🟩	🟩	🟩	🟩	🟩	🟧	🟩	🟩	🟩
Brown fat	🟩	🟩	🟧	🟩	🟩	🟩	🟩	🟩	🟩
Lumbar vertebrae	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Influence of FNs on report content

Participant radiologist

Location of False Negative	trainee 1	trainee 2	trainee 3	junior 1	junior 2	junior 3	senior 1	senior 2	senior 3
Left paracardiac node	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
T6 bone lesion	🟩	🟩	🟧	🟩	🟩	🟩	🟩	🟩	🟩
Anterior mediastinal node	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩	🟩
Bilateral pelvic nodes	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Right clavicular uptake	🟧	🟩	🟧	🟩	🟩	🟩	🟩	🟩	🟩

Legend

🟩 AI segmentation tool did not cause mistake in report

🟧 AI segmentation tool caused mistake in report

⬜ Missing data

Fig 3: The influence of false positive (FP) and false negative (FN) segmentations on individual reporters

Conclusion

Prior studies have explored the utility of AI-assistance in oncological CT, breast tomosynthesis, MRI spine and brain [5-8]. This study aimed to assess the influence of flawed segmentations on readers interpreting staging lymphoma PET/CT scans. It also assessed a reporter's ability to identify if AI was having a negative effect on their reports. Reporter confidence was significantly improved by AI-assistance, and in the vast majority (>90%) of FP/FN cases the reporter was able to correctly identify when there was a mistake with the presented segmentation. Trainees were less likely to identify if an area of disease had been missed from the presented segmentations when compared to junior and senior consultant colleagues. This may represent reader fatigue in completing questionnaires or reflect the potential for less experienced reporters to be more vulnerable to AI-derived mistakes. Trainees had the highest combined error rate for FP and FN (5/30) compared to 2/28 for junior and 1/30 for senior consultant reporters. Interestingly although neither junior nor senior consultants were influenced by FN segmentations both cohorts were influenced by FP (junior = 2/14, senior 1/15), albeit in small numbers. Although the limited study sample does not allow definitive conclusions to be drawn, it does highlight the idea that more experienced reporters have the potential to be influenced by these FPs. As with any double reporting, it is likely that over time, the strengths, and weakness of the second reporter (AI-assistance in this case) may become more apparent and can be adjusted for.

The main limitations of the study are the small numbers of readers and cases, this initial study was designed to inform a larger trial to develop an AI-assisted PET/CT tool which could be developed and validated in clinical practice. The AI-assisted read was always the second read, although there was a washout period, and whilst the order and details of cases were changed there is a possibility that some reader may have remembered individual cases. For the larger study the non-AI and AI-assisted readers will be mixed. Only the influence of flawed segmentation on the reporter was explored, and not the potential impact this could have had on patient management.

In conclusion, in this initial feasibility study less experienced reporters were more likely to be misled by inaccurate segmentations. The introduction of AI-assisted reading requires careful consideration regarding the sensitivity or specificity of the model created and the training and support provided to the reporters using the software.

References

1. Huang J, Pang WS, Lok V, Zhang L, Lucero-Prisno DE, Xu W, et al. Incidence, mortality, risk factors, and trends for Hodgkin lymphoma: a global data analysis. J Hematol Oncol [Internet]. BioMed Central; 2022;15:1–11. Available from: https://doi.org/10.1186/s13045-022-01281-9

2. El-Galaly TC, Villa D, Gormsen LC, Baech J, Lo A, Cheah CY. FDG-PET/CT in the management of lymphomas: current status and future directions. J Intern Med. 2018;284:358–76.

3. Royal College of Radiologists Clinical Radiology UK Census Report 2021. Available from https://www.rcr.ac.uk/clinical-radiology/rcr-clinical-radiology-census-report-2021#:~:text=Key%20findings&text=The%20total%20number%20of%20UK,2020%2C%20to%204%2C127%20in%202021 (Accessed 11/01/2023)

4. RSNA. Radiology Facing a Global Shortage [Internet]. RSNA NEWS. 2022. Available from: https://www.rsna.org/news/2022/may/Global-Radiologist-Shortage (Accessed 11/01/2023)

5. Mehralivand S, Harmon SA, Shih JH, Smith CP, Lay N, Argun B, et al. Multicenter Multireader Evaluation of an Artificial Intelligence–Based Attention Mapping System for the Detection of Prostate Cancer With Multiparametric MRI. Am J Roentgenol 2020; 215: 903–912

6. Guermazi A, Tannoury C, Kompel AJ, Murakami AM, Ducarouge A, Gillibert A, et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. Radiology 2022; 302: 627–636

7. Lee JH, Kim KH, Lee EH, Ahn JS, Ryu JK, Park YM, et al. Improving the Performance of Radiologists Using Artificial Intelligence-Based Detection Support Software for Mammography: A Multi-Reader Study. Korean J Radiol 2022; 23: 505–516

8. Lu SL, Xiao FR, Cheng JCH, Yang WC, Cheng YH, Chang YC, et al. Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. Neuro Oncol 2021; 23: 1560–1568