Impact of training dataset size on technical performance of a deep learning model for detection and quantification of lymphomatous disease on ^{18}F-FDG PET/CT

G.V. Ionescu[1], R. Frood[2,3], A.F. Scarsbrook[2,3] and J.M.Y. Willaime[1]

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

Objective

FDG PET/CT is widely used for staging high-grade lymphoma. Artificial intelligence has the potential to improve efficiency and enable use of advanced quantification methods in a clinical setting. Here we investigate the impact of the amount of data used to train a deep learning (DL) model on detection and segmentation performance.

Materials & Methods

Pre-treatment FDG PET/CT scans of 420 patients with a total of 6150 lymphoma lesions segmented as ground truth by experienced PET- reporters were randomly split into training (300) and test sets (120).

A DL model, consisting of an ensemble of patch-based 3D DenseNet, was trained using various dataset sizes: N = 50, 100, 150, 200, 250 and 300, randomly sampled from a total of 300 cases.

Lesion detection performance was assessed using sensitivity and false positives (FPs) per patient, and true positives to false positives ratio (TPs/FPs) across the test set.

Segmentation and quantification performance were evaluated using sensitivity, positive predictive value (PPV), Dice score and non-parametric Bland Altman analysis for SUV_{max} and SUV_{mean} per lesion, and total metabolic volume (TMV) and total lesion glycolysis (TLG) per patient.

Results

Lesion detection sensitivity varied between 82% to 88%, whilst FPs per patient decreased with more training data (see **Table 1**). TPs/FPs improved as the training dataset size increased.

Table 2 shows the segmentation performance for the six models: voxel-wise sensitivity, PPV and Dice score.

Bland Altman analysis showed improvement in Limits of Agreement (LoA) for lesion volume, TMV and TLG with more training data (See **Figure 2**).

Dataset size	50	100	150	200	250	300
Sensitivity (%)	82	83	88	83	83	86
FPs	9	4	4	4	3	3
TPs/FPs	0.73	1.42	1.40	1.43	1.67	1.69

Table 1: Lesion detection performance obtained for different dataset sizes. ***Median sensitivity was similar across models. FPs decreased with increasing dataset size. TPs/FPs ratio improved with more training data.***

Dataset size	50	100	150	200	250	300
Sensitivity (%)	91	93	92	91	89	93
PPV (%)	75	82	83	86	88	88
Dice (%)	78	83	84	85	85	86

Table 2: Segmentation performance. Median sensitivity, PPV and Dice obtained for different dataset sizes.

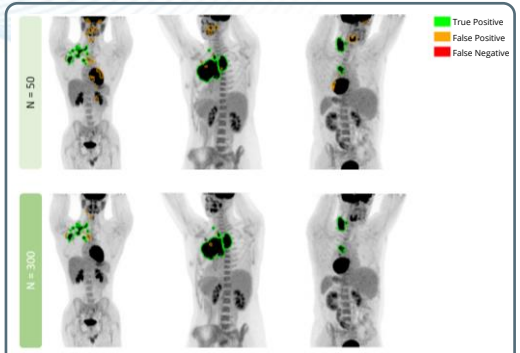

Figure 1: Examples of predictions for N=50 and N=300. Maximum intensity projections (MIP) of PET images with overlaying contours, showing true positives (green), false positives (orange) and false negatives (red).

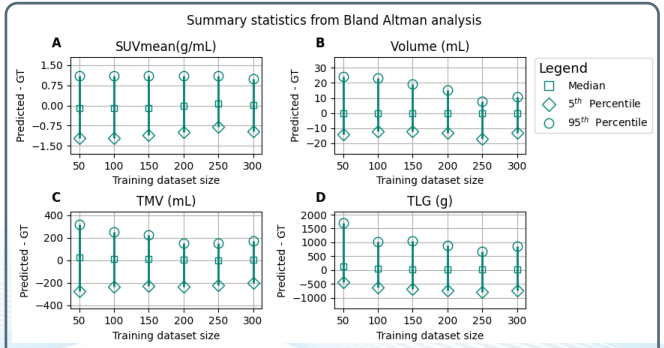

Figure 2: Statistics from Bland Altman analysis for SUVmean (A), Volume (B), TMV (C) and TLG (D). Median difference, lower and upper limits of agreement (LoA) calculated as 5th and 95th percentile are reported for the six models trained with different dataset sizes.

Conclusion

A deep learning model was relatively unaffected by the size of the training dataset in its ability to detect lymphoma lesions on PET/CT scans. However, more training data reduced FP rate, and improved agreement between prediction and ground truth segmentations for lesion volume SUV_{mean}, TMV and TLG.

Acknowledgements

R Frood and A Scarsbrook received funding from Innovate UK via a National Consortium of Intelligent Medical Imaging grant (Ref 104688). The funders had no role in study design, data collection and analysis or preparation of this presentation. We thank Matt Clark and Chirag Patel who assisted with the ground truth segmentation and verification process.