Impact of combining training data from multiple disease types on lesion detection performance in two CNN architectures

Amy J Weisman[1], Matthew D La Fontaine[1], Ojaswita Lokre[1], Rajkumar Munian-Govindan[1], Timothy G Perk[1]

[1]AIQ Solutions

AAPM 2022
JULY 10–14 | WASHINGTON, DC
64TH ANNUAL MEETING & EXHIBITION

INTRODUCTION

It is often the case that the performance of a convolutional neural network (CNN) trained for automated lesion detection improves as the dataset size increases. However, it is difficult and time-consuming to collect large amounts of data of a single disease type for automated lesion detection methods.

Here, we investigate how lesion detection performance of convolutional neural networks is impacted when dataset size is increased through combining data from multiple disease types.

MATERIALS AND METHODS

- Lesions were manually contoured on baseline and follow-up FDG PET/CT images of patients with:
  - Diffuse large B-cell lymphoma ($N_{patients}$=133, N_{scans}=415)
  - Head/neck cancer ($N_{patients}$=594, N_{scans}=898),
  - Non-small cell lung cancer ($N_{patients}$=225, N_{scans}=339)
- Two CNN architectures were implemented (**Figure 1**) to produce binary lesion masks with PET/CT images as inputs

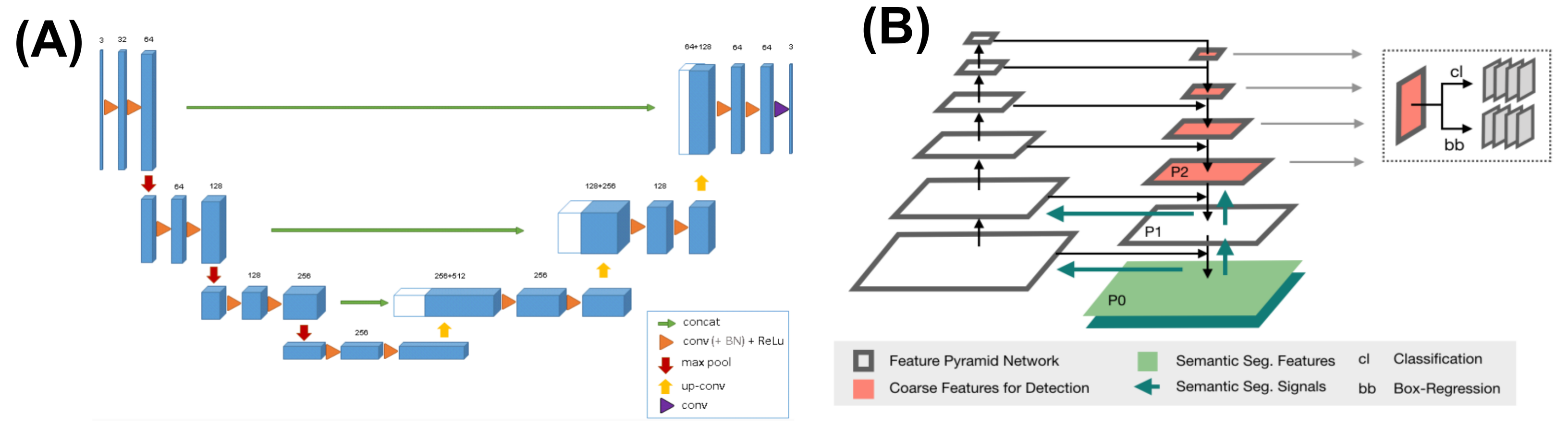

Image from Cicek et al., 2016 *Image from Jaeger et al., 2018*

Figure 1: Architecture for the (A) U-net [1] and (B) Retina U-net [2] model in this study.

- Four CNNs were trained for each architecture: one per disease type and one with all train images combined (**Figure 2**)

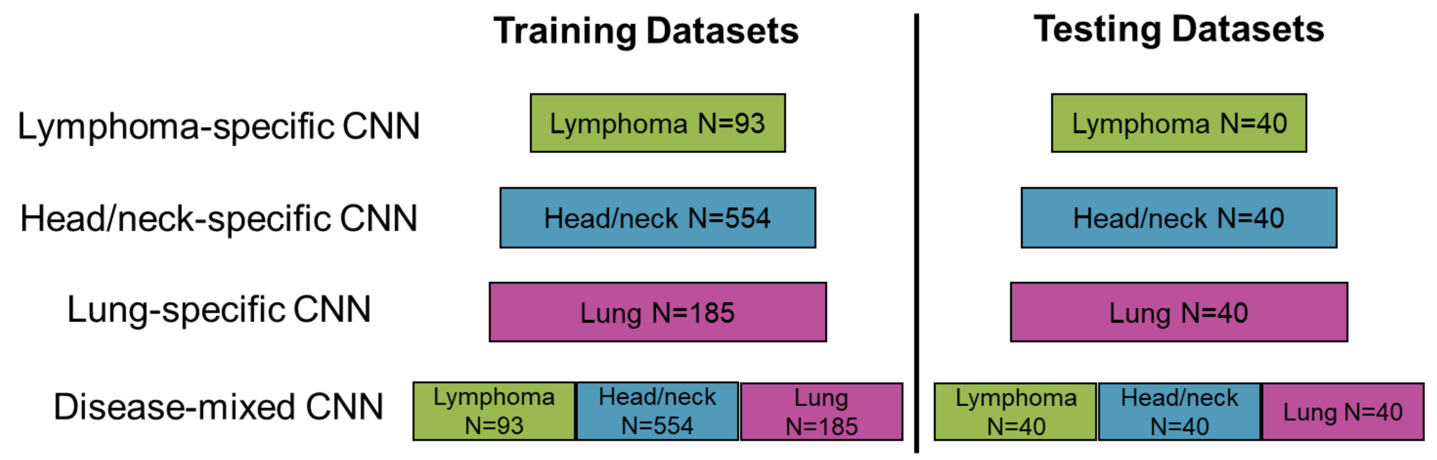

Figure 2: The four CNN approaches assessed on both U-net and retina U-net architectures. Testing data was kept identical across disease-mixed and disease-specific models.

- Performance differences of disease-mixed vs disease-specific training on the lesion detection sensitivity and number of false positives per patient (FPs/patient) was assessed using Wilcoxon signed-rank tests (paired).

RESULTS

- The 40 test patients per disease type had the following number of FDG PET/CT images: 127 lymphoma scans (1,452 lesions), 55 head/neck scans (190 lesions), and 65 lung scans (416 lesions).
- Overall differences in sensitivity and FPs/patient of disease-mixed compared to disease-specific training is shown in **Table 1**. Results for all patients and Wilcoxon p-values comparing the methods are shown in **Figure 3**.
- Performance changes were mixed across disease types and performance metrics, but consistent across network architectures

Table 1: Overall difference in performance metrics, differences are calculated as disease-mixed performance minus disease-specific performance. Overall impressions of results are shown in far right column

	Change in Sensitivity		Change in FPs/patient		Overall Assessment
Disease	U-net	Retina U-net	U-net	Retina U-net	
Lymphoma	-20	-13	-1.3	-2.2	Worse sensitivity, but fewer FPs
Head/neck	13	5	0.1	0.4	Better sensitivity, but more FPs
Lung	-1	-7	0.3	2.1	Worse sensitivity, more FPs

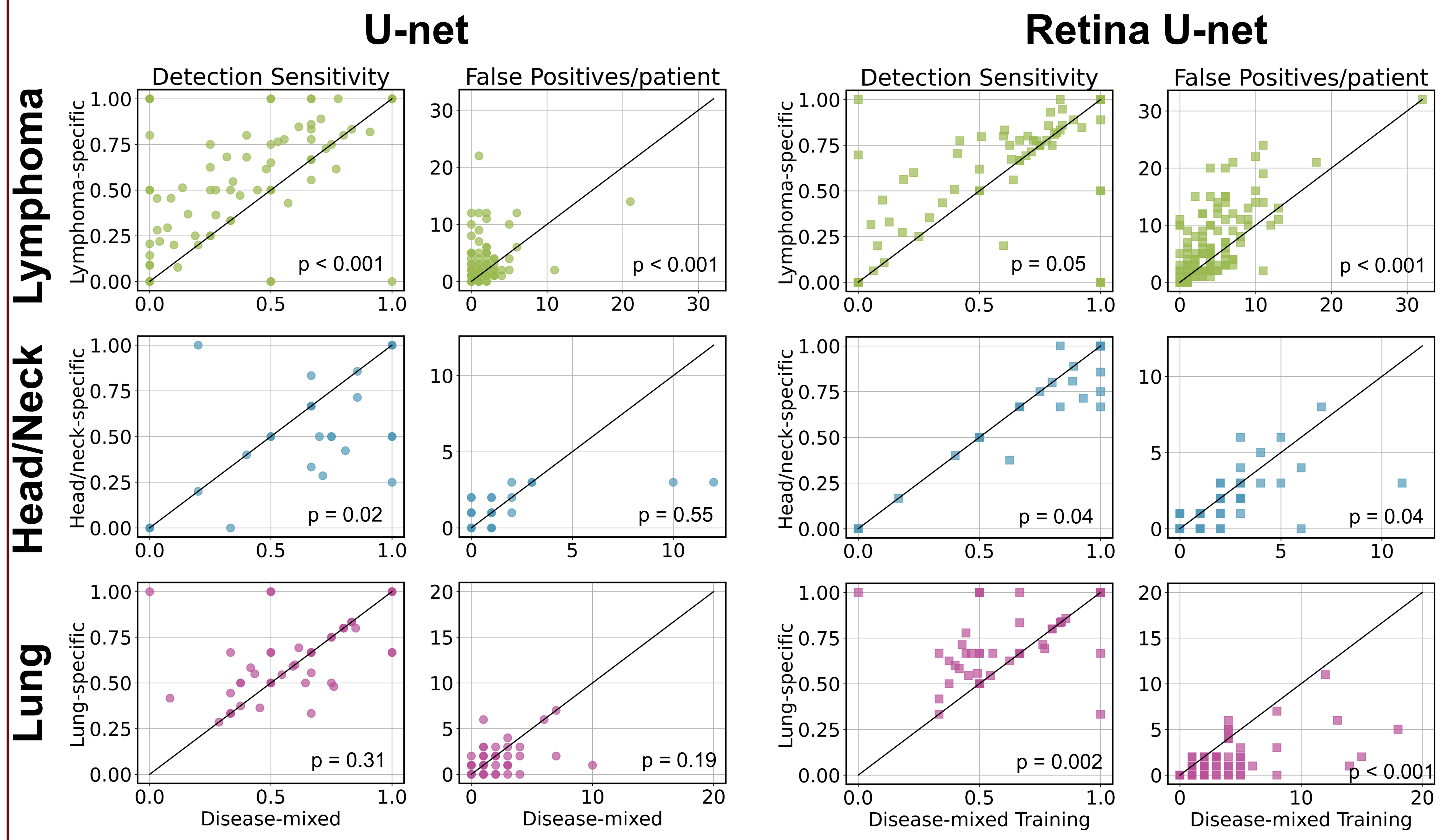

Figure 3: Comparison of performance for disease-specific & disease-mixed training for all disease types (rows) for U-net (left) and retina U-net (right). P-values are calculated using Wilcoxon paired tests comparing the two methods.

KEY FINDINGS

This study shows that for some disease types, performance may be significantly impacted with the inclusion of other disease types in the training dataset, while others may show unchanged performance.

While it may be advantageous in some scenarios to have a single model for the detection of multiple diseases, disease-mixed models should always be compared to disease-specific models to ensure performance is optimized.

LIMITATIONS

Note the purpose of this study was only to assess change due to training approach, not to achieve optimal performance of each individual training approach as hyperparameters of the CNNs were not tuned and absolute performance was not assessed.

REFERENCES

[1] Çiçek, Özgün, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2016.

[2] Jaeger, Paul F., et al. "Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection." Machine Learning for Health Workshop. PMLR, 2020.

DICSLOSURES

All authors are employed by AIQ Solutions.

CONTACT INFORMATION

Amy.Weisman@aiq-solutions.com

@aj_weisman