FDG VALIDATION

Automated Lesion Detection in FDG PET/CT Images

Methods

1928 FDG PET/CT images from 1092 patients from 15 cancer types were included for training and validation of an AI model for lesion detection. A target of 27 patients per disease type were chosen as an external hold-out set to quantify the performance of lesion detection algorithms (see Appendix A: Power Calculations for more information). Thus, the following disease types were determined to have a sufficient number of patients: lung cancer, head and neck cancer, lymphoma, soft tissue sarcoma, and breast cancer. All detected lesions were manually contoured by an expert reader.
Training and validation were completed with 957 patients (1655 scans) total, with 95% for the training dataset and 5% for a validation dataset, which was used to monitor training performance during training. After training was completed, inference was performed on the baseline images of 135 held out patients.

The performance was quantified using the lesion detection sensitivity and number of false positives per patient (FPs/patient). FPs/patient was used to assess performance instead of voxel-level specificity, as is standard in convolutional neural network (CNN) lesion detection algorithms[1,2]. Unlike specificity, FPs/patient is not constrained between 0 and 1, and a lower FP/patient value is better. The milestone of sensitivity and specificity ≥ 0.8 was thus modified to a targeted sensitivity ≥ 0.8 and FPs/patient ≤ 1.5, which would equal state-of-the-art performance for detection algorithms in patients with lesions present throughout the body[1-3].

Results

Across all 135 patients, a total of 771 lesions were manually segmented. Of these lesions, 682 (88%) had SUV_{max} greater than 2.5 g/ml, and 531 of them (69%) had SUV_{max} greater than 4 g/ml. The patients with lymphoma had the largest number of lesions, with over half of the total number of lesions (447/771). The spread of lesions across all cancer sites is shown in Table 1.

The target of identifying 80% of lesions with fewer than 1.5 FPs/patient was achieved in all cancer types except lymphoma, as shown in Table 1. However, this target was achieved in lymphoma when considering only lesions with SUV_{max} greater than 4 g/ml. Lesion detection performance in patients with lung cancer, head/neck cancer, breast cancer, and soft tissue sarcoma was excellent when considering only lesions with SUV_{max} greater than 4, with over 75% of patients showing 100% lesion detection sensitivity with 0 false positives.

Table 1: Lesion detection performance across all cancer types, as a function of lesion SUV_{max} cut-offs. For all performance metrics, the median and interquartile range are shown.

Cancer Type	Lesions	Sensitivity	FPs/Patient	Sensitivity	FPs/Patient	Sensitivity	FPs/Patient
	Median [range]	All Lesions		Lesions with SUV_{max}>2.5		Lesions with SUV_{max}> 4	
Lung	4 [1, 12]	0.80 [0.60, 1.0]	1.0 [0.0, 1.7]	0.95 [0.70, 1.0]	1.0 [0.0, 1.75]	1.0 [1.0, 1.0]	0.0 [0.0, 1.0]
H/N	2 [0, 8]	1.0 [0.54, 1.0]	0.0 [0.0, 1.0]	1.0 [0.70, 1.0]	0.0 [0.0, 0.0]	1.0 [1.0, 1.0]	0.0 [0.0, 0.0]
Lymphoma	12 [1, 83]	0.75 [0.52, 1.0]	1.0 [0.0, 1.5]	0.75 [0.60, 1.0]	1.0 [0.0, 1.5]	0.94 [0.75, 1.0]	0.75 [0, 1.5]
Breast	3 [0, 15]	1.0 [0.50, 1.0]	0.0 [0.0, 1.0]	1.0 [0.81, 1.0]	0.0 [0.0, 0.0]	1.0 [1.0, 1.0]	0.0 [0.0, 0.0]
STS	1 [1, 2]	1.0 [1.0, 1.0]	0.0 [0.0, 1.0]	1.0 [1.0, 1.0]	0.0 [0.0, 0.5]	1.0 [1.0, 1.0]	0.0 [0.0, 0.0]

Image Registration, Lesion Matching, and Response Assessment

Methods

A multi-step process was developed to automatically register baseline and follow-up PET/CT images, determine which lesions on each scan are matched over time, and calculate quantitative response parameters as published in Santoro-Fernandes *et al.*[4]. For registration, CT images were registered using 3D deformable registration, which is performed using free-form deformation with 3rd order B-splines interpolation and hierarchical control grids. After registration, it is assumed that structures (e.g., lesions) on follow-up scans overlap structures on the baseline scan that are similar or identical. However, to account for minor errors that can occur during registration, lesion contours are dilated to ensure overlap. Following registration and dilation, lesions are clustered to account for the possibility of multiple lesions on baseline merging into a single lesion on follow-up, or vice versa. Finally, lesions are numbered and matched via the Munkres assignment algorithm[5], which optimizes lesion matching based on the lesion intersection volume between the scan pairs. Following registration and lesion tracking, quantitative metrics can be tracked across time for individual lesions.

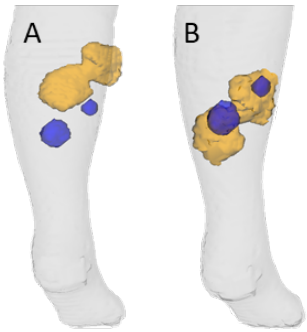

Figure 1: Illustration of two melanoma lesions on baseline (blue) that merged into a sir lesion on follow-up (yellow) before registration (A) and after registration (B)

The above algorithm was validated in a dataset of 172 FDG or ^{18}F-FLT PET/CT images of 32 patients with metastatic cancers, including melanoma, breast, lung, and prostate cancers. Lesions in all scans were manually identified, segmented, and matched across scans by an experienced imaging scientist. Algorithm performance was assessed using the overall accuracy of matching. Response assessment was evaluated using changes in SUV_{total}.

Results

In the 172 PET/CT images of the 32 patients, 140 scan pairs (two consecutive scans) were available for analysis. The imaging scientist identified a total of 736 matching decisions, which included 298 matching lesions and 438 cases of new or disappearing lesions. Without dilating lesion boundaries, the overall accuracy of the automated lesion matching method was 0.90. Accuracy was higher in matching decisions involving a new or disappearing lesions (accuracy of 1.0) compared to lesions that were present on both scans (accuracy of 0.75). When a dilation of 25 mm was implemented on lesion boundaries, the overall accuracy of the approach increased to 0.98. This approach was used to perform lesion matching in the NSCLC dataset for the remainder of Aim 1. An example of a scan pair with a complicated lesion matching decision is shown in **Figure 1.**

Automated Organ Segmentation

Methods

Aa total of 439 CT images that were acquired as part of PET/CT studies were used for training and testing organ segmentation models. Together, this dataset consisted of many disease types (prostate cancer, neuro-endocrine tumors, lung cancer, breast cancer, and systemic amyloidosis) and scanner types

(>20 different scanner models). Organ contours of the liver, spleen, lungs, thyroid, kidneys, pancreas, bladder, aorta, bowel, stomach, and heart were manually generated by a radiographer or experienced imaging researcher. 34 images were held out as a testing data.

To assess the segmentation performance, the DSC was calculated for each organ for each patient in the testing set. Median and interquartile range are reported. The target performance was set individually for each organ, based on DSC values measured across multiple physicians (interphysician performance). Interphysician performance was gathered from literature[6]. In the case of the thyroid, interphysician DSC values were not available from literature. Thus, values were obtained internally by having a second reader manually contour the thyroid in a subset of 49 patients.

Results

The results for segmentation DSC in all organs are reported in Table 2. For all 11 organs, performance values of the algorithm exceeded those of interphysician performance, indicating all targets for performance were achieved.

Table 2: Performance of the automated organ segmentation algorithm across the 34 CT images in the test dataset. Median and interquartile range are reported, as well as reference values from interphysician variability values.

Organ	DSC, Median	DSC, Interquartile range
Liver	0.96	[0.95, 0.96]
Spleen	0.93	[0.92, 0.95]
Lung	0.97	[0.96, 0.97]
Thyroid	0.73	[0.68, 0.80]
Kidney	0.93	[0.91, 0.94]
Pancreas	0.81	[0.70, 0.85]
Bladder	0.86	[0.76, 0.91]
Aorta	0.92	[0.90, 0.93]
Bowel	0.90	[0.89, 0.93]
Stomach	0.91	[0.88, 0.93]
Heart	0.94	[0.92, 0.95]

Automated Skeletal Segmentation

Methods

A total of 148 CT images that were acquired as part of PET/CT studies were collected. Together, this dataset consisted of many disease types (prostate cancer, neuro-endocrine tumors, melanoma, breast cancer, and soft-tissue sarcoma) and scanner types (>15 different scanner models). The entire skeleton was segmented and split manually into 27 segments by an experienced radiographer or imaging expert. The full list of skeletal segments is included in Table 3.

The data was split into a training dataset (N=117 scans), a validation dataset (N=6 scans) and a testing dataset (N=25 scans). To assess the segmentation performance, the DSC was calculated for each organ for each patient in the testing set. Median and interquartile range are reported. A target DSC of 0.8 was set for all bone segments.

Results

The results for segmentation DSC in all skeletal segments is reported in Table 3. For all 27 segments, performance values of the algorithm exceeded the target values of DSC=0.8.

Table 3: Performance of the automated skeletal segmentation algorithm across the 25 CT images in the test dataset. Median and interquartile range are reported.

Bone	DSC, Median	DSC, Interquartile range	Bone	DSC, Median	DSC, Interquartile range
Skull	0.89	[0.88, 0.9]	R Hand	0.81	[0.76, 0.88]
Mandible	0.91	[0.88, 0.93]	L Hand	0.85	[0.83, 0.89]
Ribs	0.86	[0.83, 0.88]	Illium	0.95	[0.93, 0.96]
Cervical Spine	0.92	[0.9, 0.94]	Pubis	0.93	[0.91, 0.94]
Thoracic Spine	0.92	[0.91, 0.93]	Ischium	0.94	[0.93, 0.95]
Lumbar Spine	0.93	[0.91, 0.94]	R Femur	0.96	[0.95, 0.97]
Sacrum	0.93	[0.92, 0.94]	L Femur	0.96	[0.95, 0.97]
R Shoulder	0.91	[0.88, 0.92]	R Tibia/Fibula	0.96	[0.95, 0.97]
L Shoulder	0.90	[0.86, 0.92]	L Tibia/Fibula	0.95	[0.94, 0.97]
Sternum	0.91	[0.89, 0.92]	R Foot	0.95	[0.93, 0.96]
R Humerus	0.94	[0.92, 0.95]	L Foot	0.94	[0.92, 0.95]
L Humerus	0.94	[0.91, 0.95]	R Patella	0.95	[0.94, 0.96]
R Radius/Ulna	0.84	[0.69, 0.89]	L Patella	0.93	[0.92, 0.96]
L Radius/Ulna	0.85	[0.76, 0.91]			

FDG Presentations

Impact of combining training data from multiple disease types

This study investigated how lesion detection performance of convolutional neural networks is impacted when dataset size is increased through combining data from multiple disease types. 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 on scans from 120 patients with 2,058 lesions (lung, lymphoma, head/neck cancer) were analyzed. Results indicated that 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.

Weisman A, la Fontaine M, Lokre O, Munian-Govindan R, Perk T. Impact of Training with Data From Multiple Disease Types On Lesion Detection Performance in Two CNN Architectures. In: The American Association of Physicists in Medicine Annual Meeting; 2022, Appendix B

Generalizability of segmentation models across scanner manufacturers

This study assesses whether convolutional neural networks (CNNs) trained for organ segmentation can generalize well across scanner manufacturers by comparing scanner-specific and scanner-mixed training approaches. CT images of 405 patients were acquired using 20 scanner models (GE, Siemens). 16 structures were manually contoured on each CT image and 3D U-nets were trained using across-scanner and scanner-mixed approaches. Results indicated that manufacturer impact on segmentation of organs was minimal.

Weisman A, la Fontaine M, Lokre O, Munian-Govindan R, Perk T. Assessment of the Generalizability of Organ Segmentation CNNs Across CT Scanner Manufacturers. In: The American Association of Physicists in Medicine Annual Meeting; 2022. Appendix C

References

1. Weisman AJ, Kieler MW, Perlman SB, et al. Convolutional Neural Networks for Automated PET/CT Detection of Diseased Lymph Node Burden in Patients with Lymphoma. *Radiology: Artificial Intelligence*. 2020;2(5):e200016. doi:10.1148/ryai.2020200016
2. Roth HR, Lu L, Seff A, et al. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. *Med Image Comput Comput Assist Interv*. 2014;17(Pt 1):520-527.
3. Sibille L, Seifert R, Avramovic N, et al. 18F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks. *Radiology*. 2019;294(2):445-452. doi:10.1148/radiol.2019191114
4. Santoro-Fernandes V, Huff DT, Scarpelli ML, et al. Development and validation of a longitudinal soft-tissue metastatic lesion matching algorithm. *Physics in Medicine & Biology*. 2021;(111):0-13. doi:10.1088/1361-6560/ac1457
5. Munkres J. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*. 1957;5(1). doi:10.1137/0105003
6. Trägårdh E, Borrelli P, Kaboteh R, et al. RECOMIA—a cloud-based platform for artificial intelligence research in nuclear medicine and radiology. *EJNMMI Physics*. 2020;7(1). doi:10.1186/s40658-020-00316-9

Appendix A: Power Calculations

An experienced biostatistician performed the following power calculations to determine the target number of patients in each validation dataset. The primary prediction endpoint was lesion detection sensitivity: the proportions of lesions identified as positive using the AI model out of the total number of positive lesions. The predictive power of the AI model in identifying true positives can be quantified by calculating the area under the curve (AUC) of the corresponding receiver operating characteristic (ROC) curve. An AUC of at least 80% is generally considered as acceptable and can be used as a benchmark.

In order to determine the minimum rate for validation samples, a formal sample size calculation can be conducted, based on the test H0: AUC ≤0.5 (no predictive power) vs. AUC ≥ 0.8 (acceptable predictive power for the validation analysis). Table 4 shows the minimum rates for the validation samples to detect an AUC of 80% with 90% power at the one-sided 0.05 significance level, under various assumptions regarding the true lesion rates and lesion sensitivity. As 100% of baseline images for the patients contained at least one lesion, a target of no fewer than 25-27 scans per disease type was set for the validation data.

Table 4: Required number of samples (scans) for validation assuming for detecting an acceptable AUC of at least 80% with 90% power at the one-sided 0.05 significance level, assuming a true lesion sensitivity level of 80% and true lesion rates ranging from 20-100%.

Primary cancer type	Number of Pts/Scans	Percentage of patients with at least one lesion					
		20%	30%	40%	50%	75%	100%
Head/neck	428/574	138	92	69	52	29	29
Lung	203/339	136	88	68	51	38	27
Lymphoma	133/422	131	84	63	51	34	25

Appendix B. American Association of Physicists in Medicine Meeting; 2022

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

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

AIQSolutions

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

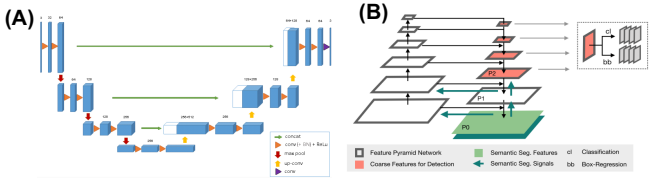

Image from Cicek et al., 2016 *Image from Jaeger et al., 2018*

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**)

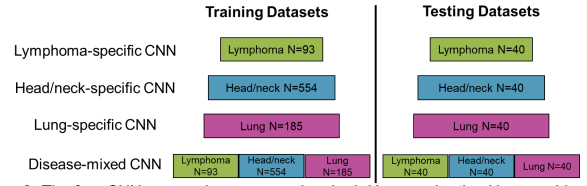

Figure 2: The four CNN approaches assessed on both U-net and retina U-net architectures. Testing data was kept identical across disease-mixed and disease-specific models.

- 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

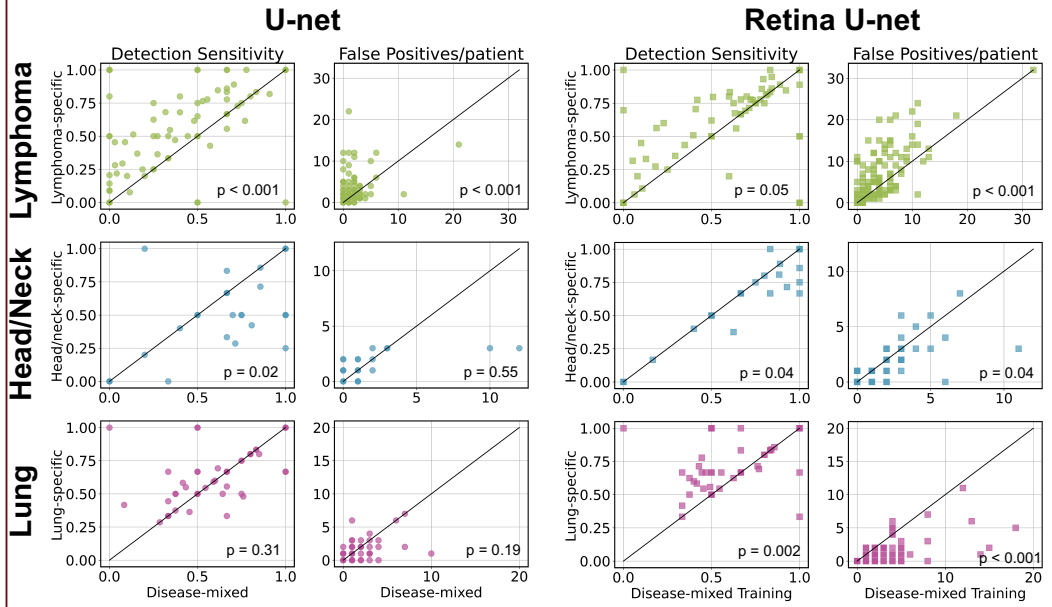

Figure 3: Comparison of performance for disease-specific & disease-mixed training for all disease types (rows) for U-net (left) and retina U-net (right). P-values are calculated using Wilcoxon paired tests comparing the two methods.

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

Appendix C. American Association of Physicists in Medicine Meeting; 2022

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

Determination of generalizability for CT organ segmentation CNNs across scanner manufacturer

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

AIQSolutions

INTRODUCTION

There are many useful applications for automated CT organ segmentation methods using convolutional neural networks (CNNs). Gathering data for training CNNs is a burdensome task, especially when datasets must be well balanced across potential biases such as patient sex, disease, or scanner models and manufacturers.

Here, we assess whether CNNs trained for organ segmentation can generalize well across scanner manufacturers by comparing scanner-specific and scanner-mixed training approaches.

MATERIALS AND METHODS

- CT images of 405 patients were retrospectively acquired from:
  - Siemens Healthineers scanners (N=186, 12 scanner models)
  - GE Medical Systems scanners (N=219, 8 scanner models)
- 16 structures were manually contoured on each CT image (see x-axis of Figure 3 for full list)
- 3D U-nets (Figure 1) were trained using across-scanner and scanner-mixed approaches, outlined in Figure 2
- **Performance Assessment:**
  - Test performance was quantified for all approaches using Dice Similarity Coefficient (DSC) for each structure
  - Differences in across-scanner and scanner-mixed approaches was quantified using median DSC value and Wilcoxon signed rank tests.

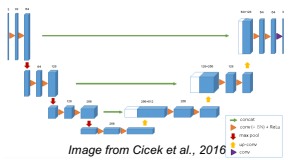

Figure 1: Architecture for the 3D U-net [1] used in this study. CT images were used as inputs to contour the 16 structures

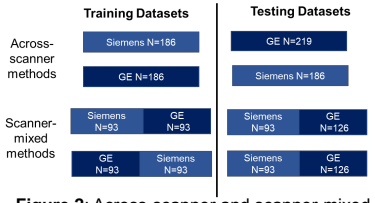

Figure 2: Across-scanner and scanner-mixed training/testing datasets. All training regimens had 186 train images. An approach similar to 2-fold cross validation was taken in scanner-mixed approaches to shuffle train data.

RESULTS

- For the 219 GE images, the scanner-mixed approach had overall better performance (Figure 3A):
  - The scanner-mixed models had significantly higher DSC for 12 structures compared to the across-scanner approach with median improvements in DSC ranging from +0.001 to +0.03
  - The remaining 4 structures did not show significant differences in across-scanner vs scanner-mixed training
- In the 186 Siemens images, mixed results were found (Figure 3B):
  - The scanner-mixed showed a significantly higher DSC in 8 structures and significantly lower DSC in 5 structures
  - In all structures, median DSC differences of the Siemens data ranged from -0.03 to +0.02.

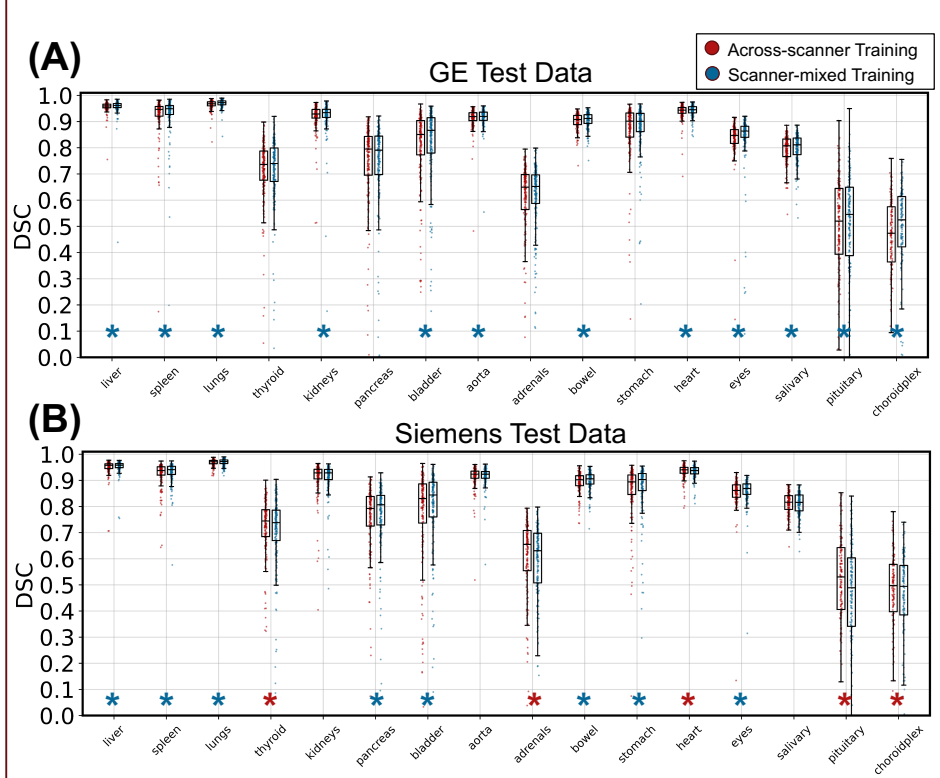

Figure 3: Dice similarity coefficient (DSC) for all organs based on across-scanner training (red) and scanner-mixed training (blue). Asterisks below box plots show significant Wilcoxon paired test p-values across methods, colored based on which method did significantly better (e.g., blue asterisk indicates scanner-mixed training achieved significantly higher DSC, red indicates across-scanner was better).

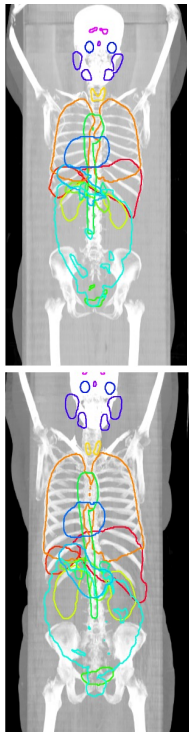

Figure 4: Example maximum intensity projections (MIPs) of CTs and organ contours for two patients.

KEY FINDINGS

When testing the generalizability of CNNs trained on multiple scanners from a single manufacturer, results varied by manufacturer and anatomic structure.

Our results indicate that manufacturer impact on segmentation of organs was minimal, even when the DSC changes were significant..

LIMITATIONS

Further research is needed to investigate whether these trends are maintained when training on a single scanner model and applying to other scanner models or manufacturers

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.

DICSLOSURES

All authors are employed by AIQ Solutions.

CONTACT INFORMATION

Amy.Weisman@aiq-solutions.com
@aj_weisman