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

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

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.

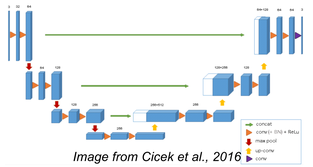

Image from Cicek et al., 2016

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

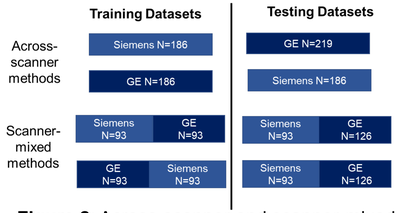

Figure 2: Across-scanner and scanner-mixed training/testing datasets. All training regimens had 186 train images. An approach similar to 2-fold cross validation was taken in scanner-mixed approaches to shuffle train data.

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.

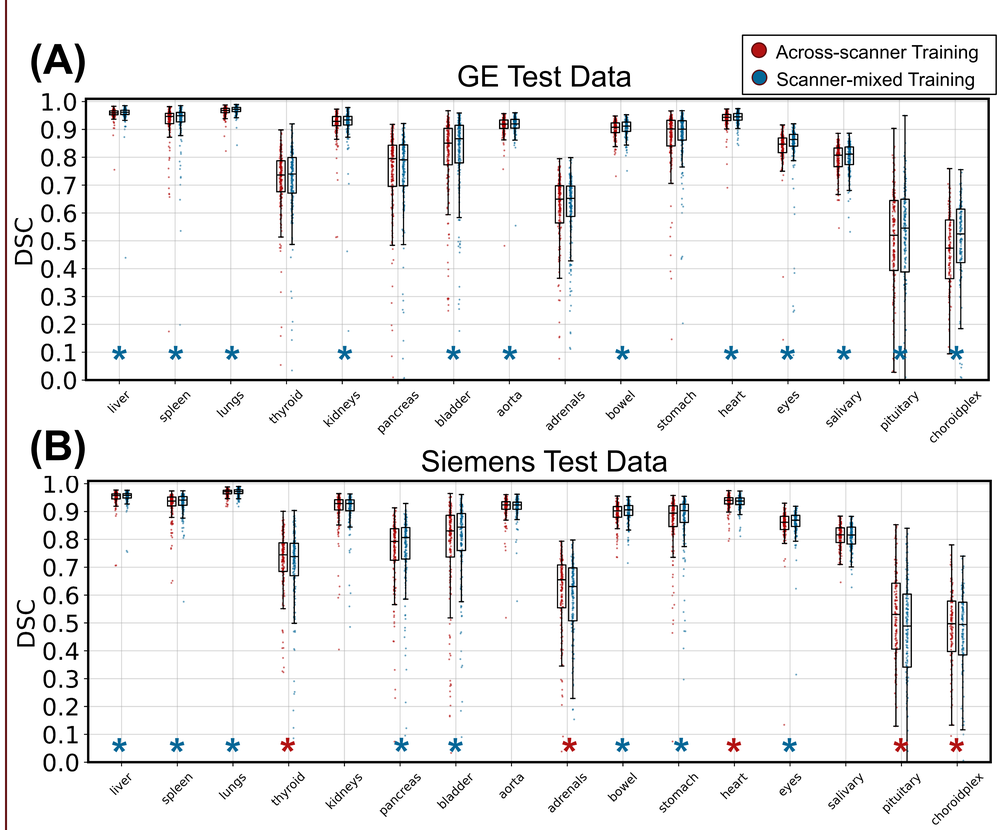

Figure 3: Dice similarity coefficient (DSC) for all organs based on across-scanner training (red) and scanner-mixed training (blue). Asterisks below box plots show significant Wilcoxon paired test p-values across methods, colored based on which method did significantly better (e.g., blue asterisk indicates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

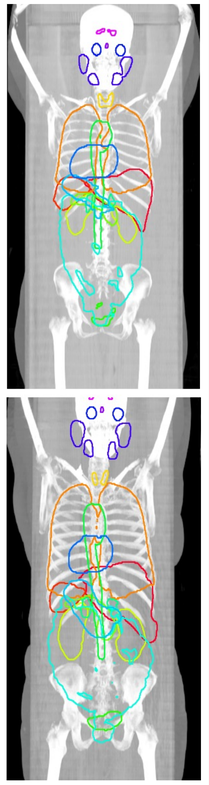

Figure 4: Example maximum intensity projections (MIPs) of CTs and organ contours for two patients.

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