Quantifying Generalizability of Automated Prostate Segmentation

Fuyao Chen[1,5](email: fuyao.chen@yale.edu), Rajesh Venkataraman[6], Lawrence H. Staib[1,3,4], Jeffrey C. Weinreb[3], Amy C. Justice[7], John A. Onofrey[1,2,3]

Departments of [1] Biomedical Engineering, [2] Urology, [3] Radiology & Biomedical Imaging, [4] Electrical Engineering, [5] Medical Scientist Training Program, Yale University, New Haven, CT, USA, [6] Eigen Health, Grass Valley, CA, USA, [7] Department of Internal Medicine, VA Connecticut Healthcare, CT, USA

Introduction and overall goal: Deep learning (DL) techniques have shown promising results in automated prostate gland segmentation in magnetic resonance imaging (MRI) scans, which is essential for various prostate oncological applications, such as tumor detection, treatment planning, and disease prognosis prediction. However, it is vital to examine algorithm performances across different datasets for practical deployment in clinical settings. Thus, this study aims to investigate the performance of prostate gland segmentation DL algorithms on diverse internal and external datasets representative of different imaging sites.

Specific aims: To quantify the impact of dataset variations on algorithm prostate segmentation results through comparison to ground truth annotations.

Rationale and background: Previously proposed DL strategies for prostate gland segmentation achieve impressive results during training and testing on a specific dataset. However, their performance may deteriorate when applied to new, unseen prostate MRI data. This discrepancy in performance can arise from differences in MRI scanners, imaging protocols, patient demographics, pathological variations, etc. Exploration of these differences and their impact on algorithm results are crucial for the understanding of DL algorithm reliability.

Methods and materials: T2-weighted prostate MRI images from Stanford Hospital (226) (GE 3T MRI) and Yale New Haven Hospital (226) (Siemens 3T MRI) were employed for training a state-of-the-art 3D U-net model for the gland segmentation task. The resulting model was tested internally on unseen images from Stanford Hospital (73) and YNHH (65), and externally on public datasets of 39 images from Initiative for Collaborative Computer Vision Benchmarking (12CVB) (GE 1.5T and Siemens 3T MRI) and 1500 images from Prostate Imaging: Cancer AI (PI-CAI) challenge (Philips and Siemens 1.5/3T MRI). Dice similarity coefficients were computed between model results and ground truth segmentations by expert radiologists to assess model performance on different datasets. Expert radiologist ground truth is not available for PI-CAI data, so Dice was calculated based on results of external independently trained algorithm provided by PI-CAI. The results were further stratified by scanner variations, including vendor, model and field strength, and clinical heterogeneities, including patient age (terciles), prostate volume (PSA≥4), prostate specific antigen (PSA) level (terciles), and cancer histopathology categories (Gleason Grade Group) (GG≥2).

Results: Fig. 1 displays automated gland segmentation results. The DL model achieved high Dice scores range for all datasets, with the highest median Dice score observed in the YNHH dataset (Median [95% CI]: 0.94[0.89-0.96]) and lowest median Dice in I2CVB dataset (0.87[0.79-0.02]). The algorithm demonstrated stable performance across different MRI vendors and field strengths. The highest Dice variations are noted for Trio and Skyra models. For clinically relevant variables, the distributions of Dice values are consistent across age groups and Gleason Grade Groups. Categorization based on prostate volume indicates that low prostate volume results in larger Dice variation and lower Dice values.

Discussion and conclusion: Scanner model variation is observed to impact algorithm performance, while low prostate volume negatively affects both the precision and accuracy of automated segmentation. Overall, our algorithm performance is stable across diverse datasets, validating the reliability and generalizability of this model for the gland segmentation task in changeable clinical settings.

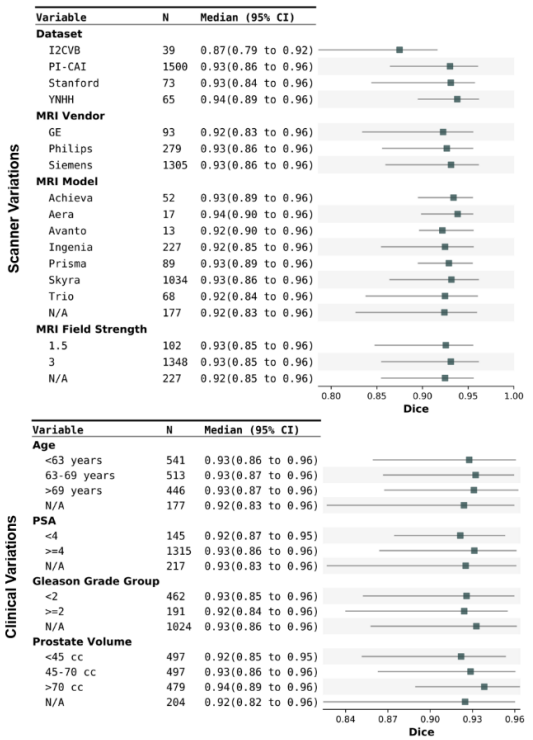

Fig 1. Median and 95% confidence interval of Dice similarity coefficients stratified based on scanner variations (MRI vendor, model and field strength) and clinical heterogeneity (age, PSA, prostate volume and Gleason grade group). Age and prostate volume categories were selected based on 33 and 66 percentiles of variable distributions. PSA categories were selected based on clinical trigger for cancer screening. Gleason grade categories were selected based on indication for clinically significant vs. non-significant cancer.