Evaluating clinical acceptability of organ-at-risk segmentation In head & neck cancer using a compendium of open-source 3D convolutional neural networks

Joseph Marsilla[1,2], Jun Won Kim[1,3], Sejin Kim[1,2], Denis Tkachuck[1], Katrina Rey-McIntyre[4,5], Tirth Patel[1,4,5], Tony Tadic[1,4,5], Fei-Fei Liu[1,2,4,5], Scott Bratman[1,2,4,5] , Andrew Hope[1,2,4,5], Benjamin Haibe-Kains[1,2,6,7,8]

[1]Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada
[2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[3]Department of Radiation Oncology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Korea
[4]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada
[5]Radiation Medicine Program, Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada
[6]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[7]Vector Institute, Toronto, Ontario, Canada
[8]Ontario Institute for Cancer Research, Toronto, Ontario, Canada

ABSTRACT

Deep learning-based auto-segmentation of organs at risk (OAR) holds the potential to improve efficacy and reduce inter-observer variability in radiotherapy planning; yet training robust auto-segmentation models and evaluating their performance is crucial for clinical implementation. Clinically acceptable auto-segmentation systems will transform radiation therapy planning procedures by reducing the amount of time required to generate the plan and therefore shortening the time between diagnosis and treatment. While studies have shown that auto-segmentation models can reach high accuracy, they often fail to reach the level of transparency and reproducibility required to assess the models' generalizability and clinical acceptability. This dissuades the adoption of auto-segmentation systems in clinical environments. In this study, we leverage the recent advances in deep learning and open science platforms to reimplement and compare the performance of eleven published OAR auto-segmentation models on the largest compendium of head-and-neck cancer imaging datasets to date. To create a benchmark for current and future studies, we made the full data compendium and computer code publicly available to allow the scientific community to

scrutinize, improve and build upon. We have developed a new paradigm for performance assessment of auto-segmentation systems by giving weight to metrics more closely correlated with clinical acceptability. To accelerate the rate of clinical acceptability analysis in medically oriented auto-segmentation studies, we extend the open-source quality assurance platform, QUANNOTATE, to enable clinical assessment of auto segmented regions of interest at scale. We further provide examples as to how clinical acceptability assessment could accelerate the adoption of auto-segmentation systems in the clinic by establishing 'baseline' clinical acceptability threshold(s) for multiple organs-at-risk in the head and neck region. All centers deploying auto-segmentation systems can employ a similar architecture designed to simultaneously assess performance and clinical acceptability so as to benchmark novel segmentation tools and determine if these tools meet their internal clinical goals.

Keywords: deep learning, semantic segmentation, auto-segmentation, convolutional neural networks, turing test, organs at risk, head and neck cancer, clinical acceptability

INTRODUCTION

In radiotherapy (RT) planning, organs at risk (OARs) are delineated to limit the radiation dose to these organs and minimize organ-specific toxicities. Automated OAR delineation has been a long-standing goal to increase efficiency and decrease manual effort required in RT planning. Early deformable atlas-based registration approaches [1–6] have been replaced by deep learning (DL)-based approaches using convolutional neural networks (CNN) as the state-of-the-art platform for automated segmentation of OARs in the head neck region since 2016 [7–19]. In a recent study, Van Dijk *et al*. have shown that radiation oncologists preferred DL-based auto-contours compared with Atlas Based auto-segmentation contours [18]. In addition, a clinical assessment of a commercially available OAR segmentation algorithm, showed the potential in using these systems to optimize clinical contouring workflows [14]. These studies confirm the translation potential of integrating DL-based auto-contouring models into the RT workflow for the treatment of head and neck cancer (HNC) as well as cancers of other primary sites. Performance metrics like the Volumetric Sørensen–Dice Coefficient (DICE) or the 95th Percentile Hausdorff Distance (95HD) are commonly used in comparing auto-contours and ground truth contours; yet, quantitative evaluation of clinical acceptability of an auto-segmentation model is hindered by inter-observer variability and previous studies have analyzed clinical acceptability mostly in qualitative settings. Correlating qualitative performance in terms of acceptability testing with quantitative metrics outlining a network's performance could prove invaluable to the future optimization and development of clinically acceptable auto segmentation systems.

Recent publications have shown the utility of properly trained and optimized simple CNN architectures when applied to complex segmentation tasks [20]. This paradigm shift suggests that model complexity does not necessarily correlate with increased model performance and concludes that details in method configuration affect performance at a higher level than structural architectural variations. In the context of OAR segmentation for HNC most recent publications have used a structural variant of original UNET architectures as their model of choice [21,22], indicating greater dependence on this backbone for semantic segmentation of OARs in the HN region [8,10–19,23–27]. Unfortunately, the published DL methods often suffer from lack of transparency regarding releasing code and underlying data [28]. Open-source release of data and code is essential to allow external model validation and to test the robustness of the findings in question and assess their potential clinical acceptability.

To address the lack of open source resources to validate findings of previously published studies, we performed a large comparative analysis of open source DL networks applied to

OAR segmentation for radiation therapy of HNC. We hypothesized that simple 3D UNET architectures with little to no architectural modifications may outperform more complex networks when segmenting OARs in HNC. We conducted this study by evaluating the performance of eleven open-source 3D segmentation models originally engineered for medical image segmentation in a radiological context to segment 19 OAR classes commonly used in RT planning. We found that the simplest 3D-UNET architecture proved to produce superior segmentations across all 19 OARs analyzed. Blinded clinical acceptability testing was conducted for each OAR in the study using a novel extension to a previously published web-based RT quality assurance tool. The results were then compared with six quantitative segmentation metrics for each OAR to determine those most closely correlated with clinical acceptability.

METHODS

Dataset Curation

We have built the largest dataset containing imaging, RT structures (OARs, gross tumor volume and neck lymph node levels), treatment, demographic and clinical information for HNC patients treated at the University Health Network (UHN) from 2005-2017. This original dataset of 3,211 patients is referred to as RADCURE and is under submission at The Cancer Imaging Archive [29,30]. A UHN institutional review board approved our study and waived the requirement for informed consent (REB 17-5871); we performed all experiments in accordance with relevant guidelines and ethical regulations of Princess Margaret Cancer Center. Only PM data used for network training was involved in the REB. As the associated clinical data has been collected prospectively as part of the PM Anthology of Outcomes [31], the clinical endpoints associated with this imaging resource are of high quality and have been used to drive significant changes in the management of HNC [32]. This dataset has been previously used for radiomic prediction of survival in HNC [33] and in studies addressing dental artifact reduction [34]. At PM, all patient contours are required to conform to standard nomenclature for both OARs and targets. To search for all the OARs available in the dataset, we implemented a series of regularized expression strings that were used to identify and standardize the names of the OARs which were reviewed by a radiation oncologist specializing in HNC (AJH). A list of the regular expressions used for OAR naming standardization can be found in the utilities folder on the study's github page. The RTSTRUCT files revealed a total of 34 OARs in the head and neck region had been contoured over the study period. For this analysis, we selected 19 OARs that

were consistently delineated in 582 patients (Supplementary Figure 1). The study cohort consisted of 378 oropharyngeal, 123 nasopharyngeal, 10 hypopharyngeal, 10 oral cavity, and 7 laryngeal cancer patients as well as 55 patients of unknown or other primary-site cancers. The 19 OARs included acoustics (L/R), brachial plexuses (L/R), brainstem, chiasm, esophagus, eyeballs (L/R), larynx, lenses (L/R), lips, mandible, optic nerves (L/R), parotid glands (L/R), and spinal cord. (Figure 1)

Selection of Published Segmentation Models

A literature search was conducted from January 2016 to April 2020 to find medical image segmentation studies using DL-based modeling approaches. Original segmentation studies that did not release code on publication or did not actively maintain the repository were excluded from the analysis (Table 1). We also excluded 2D architectures whose convolutional schemes could not be directly updated to 3D without disrupting architectural integrity. Primary selection criteria included models coded originally in PyTorch [35], that were 3D or could be minimally modified to accept 3D inputs. Popularity of repositories on github quantified by it's public github star rating was also a factor used when selecting architectures with similar architectural layouts. Selected networks were to be trained end-to-end directly, with no complex architectural or computational restrictions and with minimal modifications made to original source code provided by the original authors to keep architectural integrity. We refactored all tested models into a Pytorch Lightning framework [36] to increase readability, reusability and re-implementation of the code.

Training of Open Source Segmentation Models

Selected initial models were trained using a combined loss scheme to address the heavy pixel-wise class imbalance present within our dataset between the individual classes of OARs. We combined a modified and weighted TopK Cross entropy loss [11], which selects the top 10% of most difficult pixels in the cross entropy loss calculation and only added the contribution of these pixels to the loss with the Focal Tversky Loss previously defined to excel at semantic segmentation tasks where pixel-wise class imbalance between classes is high [37,38]. The RADAM optimizer was used to minimize the loss during training [39,40]. The initial learning rate of the optimizer was set to .001. The validation loss was monitored and the learning rate decreased by a factor of 0.5 if there was no significant change in validation loss after 12 epochs. The voxel spacing of each patient scan was standardized to 1mm x 1mm x 3mm using SimpleITK [41]. An augmentation scheme was implemented during training. Translation, mirroring, zooming, and

rotation were applied in-plane (x-y plane only) and at random during training. We used random translations between -32 and 32 pixels in each plane; uniform scaling factors between 0.9 and 1.1; and mirrored images with a probability of 0.5. Patient scans were cropped to a size of (192x192x64) pixels for use during training. The ground truth mask was first used to identify the coordinates of the patient's center of mass on which cropping was based. All CT volumes used in analysis were clipped to a HU range of -200 to 300 (a common soft tissue window applied by radiation oncologists) before augmentations were applied. Z-score normalization was applied after all augmentations were completed. For each experiment the same 80/10/10 split, corresponding to 479, 44, and 59 scans for training, validating, and testing, respectively, was used (Figure 1). Each model was trained on 4 NVIDIA Tesla P100 GPUs for 3 days or until convergence. Early stopping was implemented if no significant change (0.1 decrease in loss magnitude) was made in validation loss minimization after 50 epochs. After testing each model on the hold out set of 59 patients, the best model ranked by average classical segmentation metrics across all OARs (DICE, 95HD) were chosen for a 2nd re-training phase.

The best performing model (highest DICE and lowest 95HD) was selected and re-trained on full-resolution patient scans for an additional three days on 4 NVIDIA Tesla V100 GPUs and tested using MONAI based sliding-window inference [42]. The following additional changes were made to the final training protocol: initial learning rate was reduced to 4e-4, validation loss was monitored and learning rate was decayed by a factor of 0.96 after no change in loss for 1 epoch, all CT volumes used in the final analysis were clipped to a HU range of -500 to 1000. Cropping regimen was modified to the size of (192x192x128) pixels. Images were not resampled for re-training. We used random translations between -64 and 64 pixels while the other randomized augmentations and Z-score normalization scheme were kept unchanged. The initial convolutional setting of the best network was set to 48 feature maps instead of 32 feature maps used during the validation phase of our study. Although large parameter dense CNNs could lead to network overfitting (and poor generalizability) when data used for training is sparse [43], increasing the parameters by increasing feature maps at each network layer, increases the model's capacity to learn complex relationships like learning to segment complex anatomical features from a medical image.

Integrating Blinded Clinical Acceptability Testing Into Open-Source Quality Assessment Tool

Multiple observers were invited to review and evaluate clinical acceptability of OAR contours generated by the best performing model. Observers were able to access the contours online

from different locations and were blinded to the assessments by other observers. In order to conduct this analysis, we used QUANNOTATE (www.quannotate.com), a web-based quality assurance labeling platform which recently allowed researchers to conduct a wide range of anatomy labeling tasks primarily focused on post-therapy assessment of clinical target volume coverage around primary tumor and neck level regions used in RT for HNC [44]. In this study, we re-engineered QUANNOTATE using MERN (MongoDB, React JS, Express, Node JS) as the primary development stack [44] due to its speed and adaptability. The interface is open-source and can be accessed at (https://github.com/bhklab/quannotate). The application was modified so that each user is able to assess specific contours by individual OAR category. Inside the test dashboard, observers are able to slide through an entire patient scan, in an OAR dependent manner, using the integrated in-browser window slider. Users can select between different clinical CT windowing levels of the patient's CT image where the contours have been superimposed. There are currently three windowing level settings that can be used during the assessment, these include bone, soft tissue and lung views. A clinical assessed contour evaluation protocol was defined to determine clinical validity of the contours produced by the best method after retraining the best performing model (Figure 1). Users were asked a series of questions about each contour set. The first question, "Who delineated this contour?" assesses the hypothetical author of the contour under analysis. The options that can be selected include "I don't know", "Human" or "Computer". The next question assesses the clinical acceptability by getting the user to 'rate' the contour being assessed. We adopted a previously developed 5 point rating clinical acceptability scale for our study [17,45]. The contour acceptability scale ranged from 1 (Poor, large areas need minor or major edits, is unusable for planning purposes), 2 (Fair, needs significant edits to be used for planning purposes), 3 (Good, needs minor edits to be used for planning purposes), 4 (Within acceptable inter physician variation for planning purposes) and 5 (Perfect, indistinguishable from physician drawn contours for planning purposes). The users were blinded to the origin of the contour at the time of the questionnaire, and answers were exported to be analyzed after all users completely assessed all contours. Individual ratings of 4 or higher will be considered acceptable, the condition being no additional edits are required to transition set contour into a RT plan. When analyzing ratings averaged across all observers, a contour will be considered clinically acceptable when mean acceptability rating (MAR) for that contour across all observers is above 3.5, and not-clinically acceptable when MAR is less than 3.5 **(**Supplementary Figure 2).

Assessing Clinical Acceptability of the Best Performing Model

Due to time constraints, ten randomized patients were randomly selected from the hold out test set per OAR category and uploaded to QUANNOTATE for further scrutiny. Each observer used the QUANNOTATE interface to complete the blinded questionnaire defined above for two sets of contours for each OAR (Ground Truth or AI Generated). In total, answers were recorded for 380 scans per observer, where nscans = (10 Ground Truth Scans + 10 Paired Deep Learning Contours) x 19 OAR Categories. Results were extracted and Mean Acceptability Rating (MAR) for each OAR was calculated by averaging the ratings obtained for each contour across all observers. Corresponding heatmap(s) showing the distribution of MAR(s) for manual and DL generated contours were extracted. After inference, segmentation network performance was assessed by calculating three different types of performance metrics. Classical segmentation metrics including DICE and 95HD, boundary metrics (Surface Distance [SD], Added path length [APL]) , and false negative metrics (False Negative Volume and False Negative Length) [46]. These metrics were extracted for each OAR contour. To determine the weight each metric should have when assessing acceptability, a correlation heatmap was extracted and used to compare metrics at the OAR level against its corresponding MAR recorded using QUANNOTATE. Metrics were considered 'more clinically acceptable' if they showed significantly greater correlation with MAR. Suggestions will be made as to how these metrics can be used to optimize a network's segmentations for clinical acceptability. **(**Supplementary Figure 2)

Ensemble Modeling for Improved Performance and Generalizability

Ensembling by averaging the predictions of similar but distinct CNN models has proven to boost network performance while increasing a model's generalizability potential. To minimize variance and spurious model predictions, we decided to use five random K-Fold training/validation splits to build a WOLNET segmentation ensemble. The training scheme was modified as follows: random crops of size (112px 176px x 176px) were used to train the network with a batch size of 2. Each fold was trained on 4 NVIDIA Tesla V100 GPUs with 32G RAM, on the original multi-class segmentation task for 3 days or until convergence. To mimic a live clinical environment, the cropping regiment was modified as follows. Otsu thresholding was used to create a binary mask of the entire body section in the scan [47]. This mask was then used to find the center of mass of the patient, from which the original patient image was cropped to a size of (224px x 224px) in the x-y plane. The depth of the scan (z-plane) was left un-cropped. The following (224px x224px x Zpx) image was then used for inference. In order to calculate predictions over the entire volume, we employed an in-house sliding window function. Using the

ROI size defined above, multiple sub-crops of size (112px 176px x 176px) were generated and passed through each model in the ensemble. A total of 224 sub-crops were generated for each scan. The relative positions of each subcrop were tracked in relation to the original scan and used to average predictions from overlapping sections upon completion of inference. The final predictions from each scan were calculated by averaging the predictions from each model in the ensemble.

Collection and Preprocessing of External Datasets

We have also gathered a set of 5 publicly available datasets for external validation of our segmentation models that span a total of 335 patients [11,48–51] (Supplementary Table 1). These datasets contain varying distributions of OARs that overlap with those included in our study. Two out of the five datasets were extracted and curated using an internal preprocessing methodology. These scripts were used in tandem with Imgtools (github.com/bhklab/imgtools), another in-house image processing package, to make these imaging data directly amenable to DL. Processed datasets including other in-house pre-processing scripts used during curation are made publicly available and can be accessed at (github.com/bhklab/ptl-oar-segmentation/utils). To provide an accurate “clinic-like” environment for external validation, no ground-truth information will be used to identify the center of mass of any scan during inference. During inference, otsu thresholding was used to first create a ‘mask’ of the body, from which the patient’s center of mass was calculated and used to crop the image in the x-y plane with dimensions of 292x292px. Sliding window inference was applied to the cropped version of the patient scan. Final predictions were averaged across each fold and performance was assessed using the classical segmentation metrics described above. The weights for each fold, and processed versions of each external dataset, were saved and made available on (github.com/bhklab/ptl-oar-segmentation/inference**).** External publicly available datasets were governed by individual REBs by it’s institution of origin.

RESULTS

A complete workflow summarizing the study overview was generated to highlight the most important parts of this analysis (Figure 1).

Model selection

Out of the 15 OAR segmentation papers published from 2016 to 2020 only three provided open-source code to allow the community to validate their findings. Out of the three studies that

did release their code, only one (Anatomy-Net [13]) could be re-trained in an end-to-end fashion without major architectural modifications. Additionally, only one open-source repository from these studies was effectively maintained or updated their code[12] and therefore the published findings from the majority of OAR segmentation studies in the HN region could not be directly reproduced without major modifications to their pipelines (i.e. re-implementation). In total, 60 additional studies with a medical image segmentation theme were reviewed in the conducted literary search to augment networks we could test during our analysis. Out of the 60 studies examined, only 31 studies made a part of their code available to the open-source community. Out of the 31 open source architectures analyzed, eleven open source networks passed the defined selection criteria [13,22,52–60]. Studies were excluded from this analysis based on code availability, computational capacity, model integrability and repository maintenance. A total of eleven segmentation models were selected to be trained **(**Supplementary Table 2**)**.

Performance of the best Segmentation model

We compared the segmentation models using the DICE and 95HD metrics for the combined set of OARs, however, we would note that quantitative performance of any given contour may not correlate well when observed through the qualitative scope of clinical acceptability testing. This is to say that not all top OARs with high-volumetric dice or low hausdorff distance between the corresponding ground truth contour could be considered acceptable when assessed by field experts. This puts an emphasis to produce networks that are optimized for metrics most closely associated with acceptability. When analyzing the mean performance metrics of all OARs for each open source model trained, the top three segmentation models were UNET variants. WOLNET [61] performed the best among the 11 models and achieved the highest average DICE (0.765±0.10) (Supplementary Figure 3a) and lowest average HD (2.63±2.61**)** (Supplementary Figure 3b). Altered 3D versions of UNET3+DEEPSUP (DICE 0.74±0.10; HD 2.98±2.63) and UNET++ (DICE 0.73±0.10; HD 3.10±2.83) were the second and third top-performing segmentation models. The metrics for individual OARs were compared for the top-three networks. WOLNET consistently outperforms other methods for all OARs in our analysis**.** These results show a significant performance difference between WOLNET and all other segmentation models when comparing average performance metrics for each method across all OARs (Supplementary Figure 3c). WOLNET was re-trained according to procedures outlined in (Methods: Training of Open Source Segmentation Models**)**, resulting in a final average DICE of (0.77± 0.09) and a 95HD of (3.42± 4.05). Four other common segmentation metrics were also extracted for each OAR category. (Supplementary Table 3). Mandible (0.92±0.03), Eyes (0.88 ±

0.04), and Spinal Cord (0.85± 0.05) received the highest mean DICE, while Chiasm (0.41 ± 0.18), Brachial plexus (0.70 ± 0.12), and Optic Nerves (0.71 ± 0.10) received the lowest mean DICE respectively. Chiasm and the Acoustics (L/R) had significantly higher variance in DICE than other OARs (Figure 2a). Lenses (1.44 ± 0.53), Eyes (1.81 ± 0.75) and Spinal Cord (1.98±0.73) had the lowest HD, whereas Acoustics (3.66 ± 9.93), Brachial Plexus (4.98 ± 7.35) and Chiasm (6.59 ± 4.46) had the highest recorded 95HD respectively. Brachial Plexus (L/R) and Acoustics (R/L) had significantly higher variance in 95HD than other OARs (Figure 2b, Supplementary Table 3). Exported contours from a random test set patient were compared against the ground truth human generated contours for that patient, showing high correlation between contours (Figure 2c).

Clinical Acceptability

The complete set of contours were analyzed by four practicing radiation oncologists with 10+ years of experience using the QUANNOTATE platform. A total of 20 samples, 10 manual (human) contours and 10 paired computer (deep learning) contours) for each of the 19 OARs (a total of 380 contours) were assessed by each observer. In total, 40 ratings for each OAR in each category were recorded (10 ratings per oncologist). According to the un-averaged clinical acceptability ratings, 73% of manual contours analyzed were considered acceptable, compared with 48% of deep learning contours **(**Figure 2d). Deviance from perfect acceptability for the manual contour assessment could be concerning. Certain factors could affect the outcome of the acceptability analysis. Firstly, as more rigorous contouring standards have developed at our center over time, contours delineated prior to those ‘standard upgrades’ would be considered ‘unacceptable’ by current delineation requirements. Secondarily, external observers participated in the acceptability test and contouring practices for a given OAR analyzed could vary at their center which may have more or less stringent contouring requirements than at Princess Margaret. Taking this into account, MAR analysis still provides valuable information for clinical acceptability of any ROI used in radiation therapy planning. A MAR threshold of 3.5 or more was considered to be ‘acceptable’ for this analysis meaning that these contours require no edits to be confidently adopted into RT planning. The manual contours received a global MAR of 3.81±0.88 and compared with 3.37±0.97 for DL contours. When comparing MAR for all OARs, 78% of manual contours are considered acceptable (Supplementary Figure 4a) compared with 52% of DL contours (Supplementary Figure 4b). Histogram of mean acceptability rating (MAR) was plotted for manual ground truth contours (Supplementary Figure 4c) and deep learning generated contours (Supplementary Figure 4d) at the OAR level. When analyzing individual

OAR categories, manual contours were considered more acceptable than deep learning contours for 15 out of the 19 OARs assessed. When assessing whether certain OAR categories passed the mean acceptability cutoff of 3.5, 15 manually delineated OARs on average were considered clinically acceptable, requiring no edits for planning purposes, compared with 9 OARs generated by DL. When analyzing categories of OARs requiring minor edits for their contours to be accepted into RT plans (3.0 < MAR < 3.5), 7 DL generated OARs compared with 4 manually contoured OARs met this criteria (Supplementary Table 4). The least clinically acceptable manually contoured OAR was the chiasm (lightest pink shade on histogram) with an average acceptability rating of 3.03±1.27, while the most acceptable manually contoured OAR was the mandible (reddish shade on histogram) with an average acceptability rating of 4.15±0.89. The least clinically acceptable deep learning generated OAR was the Larynx (bronze shade on histogram) with an average acceptability rating of 2.38±0.90, while the most acceptable deep learning generated OAR was the lenses (turquoise shade on histogram) with an average acceptability rating of 3.90±0.85.

Defining Clinically Acceptable Performance Metrics

There exists a divide between quantitative performance of segmentation networks with the qualitative evaluation of the contours that are produced by them. In other words, just because a network can segment a contour with a higher than average DICE score, does not necessarily mean that contour is good enough to be integrated within a clinical RT plan. Same logic applies to low 95HD scores. Acceptability testing is therefore an essential component in any RT focused auto-segmentation study to give clinical relevance to deep learning contours generated. The results of the acceptability test and MAR correlation are plotted in a heatmap for each individual OAR (Figure 3a-b). When correlating DL contour MAR with performance metrics, boundary distance metrics (HD, SD) were found to be most significantly correlated indicating optimizing networks for these metrics could prove beneficial in producing more clinically relevant contours (Figure 3c). When analyzing global correlation with MAR across all OARs, there were no significant correlations between acceptability ratings and performance metrics (Supplementary Figure 5). However, when looking at individual OARs, HD was significantly correlated with MAR for 6 out of the 19 OARs (These include left eye, brain stem, right parotid gland, Left Lens and Chiasm (Supplementary Figure 6).

External Validation

A model's application potential is directly related to how well that model can perform on external datasets with characteristics outside the cohort of the one used to train the predictive model [72,73].

Depending on the model's robustness when applied to external data, additional rounds of finetuning may be required to optimize the model's performance.We therefore developed a WOLNET ensemble, with 5 models each trained on different training set divisions. In order to test the full extent of the ensemble model's generalizability potential, 5 datasets were collected and curated. The ensemble was applied to the processed images of each datasets and classical metrics were extracted for each OAR category that had contours delineated in the dataset under scrutiny **(**Supplementary Table 1**;** Supplementary Figure 7).

The WOLNET ensemble outperformed the singular WOLNET model that was used for acceptability testing both in terms of DICE (Supplementary Table 5) and 95HD (Supplementary Table 6) for most OARs. When looking at metric variation by OAR volume, we see a positive trend correlating high volume OARs with higher VolDice and lower 95HD ratings (Supplementary Figure 8). These findings are consistent for all OARs segmented except low-volume chiasm (which is poorly defined on a CT scan and therefore contours are more variable) and high-volume mandible (which is precisely defined on a CT scan and relatively easy to segment). We found that DICE was lowly and more positively correlated with OAR volume than 95HD (Supplementary Figure 8e). We found extensive variability of ground truth information extracted from each external dataset and only a subset of OARs segmented in this study overlapped (Supplementary Table 5). Median classical performance metrics extracted for each OAR present in each dataset (Observer) were compared in separate scatterplots for each dataset (Figure 4). Barplots of 3D volumetric Dice and 95HD were also extracted for each dataset (Supplementary Figure 8). In comparison to the performance of the singular WOLNET model used for QUANNOTATE testing, improvements in terms of DICE and 95HD were made for each OAR category. We can expect to see similar improvements in OAR acceptability when clinical acceptability testing is conducted. Two datasets had the most overlapping OAR categories with our Radcure dataset. All results for external validation of the WOLNET ensemble on each external dataset for each OAR category that overlapped with RADCURE were extracted (Average DICE - Supplementary Table 6; Average 95HD - Supplementary Table 7). Dataset 1 (HNSCC-3DCT-RT) [74] and Dataset 5 (StructSeg19) [49] had 14 out of the 19 RADCURE OARs delineated in their dataset. The top performing OAR categories when compared against GT contours in dataset 1 were the eyeballs (L/R) both achieving DICE scores of 0.83±0.05 and 95HD scores of 2.83±0.87 and 2.45±0.64 respectively. The lowest performing OAR in Dataset 1 was the chiasm achieving an average DICE of 0.27±0.21 The top performing OAR categories when compared against GT contours in dataset 5 were the parotids (L/R) both

achieving DICE scores of 0.86±0.04 and 95HD scores of 3.0±1.73 and 3.0±3.0 respectively. The lowest performing OAR in Dataset 5 was also the Chiasm with an average DICE of 0.33±0.17 and average 95HD of 5.14±3.53. While performance on external datasets doesn't eclipse that of the internal radcure test set for most OARs, findings show that the WOLNET ensemble without finetuning is generalizable on external data. Major causes of performance variations include: inter-observer variations between delineation protocols during ground truth contouring; variability in patient cohorts and scanners used for imaging. Ensemble finetuning can further enhance performance for certain smaller OARs that may show extensive delineation variability at the dataset level.

DISCUSSION

In the current study, we demonstrated that a simple 3D Unet (named WOLNET) provides the best objective performance of all published models but complete acceptability (where generated contours require no edits to be inducted into RT plans) of the all OAR contours generated by this approach has yet to be achieved. We are confident that 16 out of 19 OARs generated by this simple open-source network can be inducted into RT plans with at least minor edits before direct inclusion into RT plans. Critically, final weights and code for all 11 open-source 3D segmentation models used in this study, including the quannotate web-platform used for acceptability testing and preprocessed datasets used for the external validation have been publicly shared with the scientific community. This is the first deep learning study addressing automated OAR segmentation for RT planning conducted with reproducibility and data sharing as a foundational motivation for this analysis. Additionally, we provide a foundation on top of which future automated segmentation systems can be benchmarked and assessed in terms of performance and clinical acceptability for multiple OARs in the HN region. Adopting similar protocols in future studies and clinical trials assessing performance of segmentation systems in clinical settings will accelerate adoption of these methods and their ability to dramatically improve standard of care. We selected open-source segmentation architectures constructed for medical image segmentation, assessed the segmentation quality of each network, and provided the community with pre-trained weights of the top-performing models. These results provide a set of 'open-source' controls on top of which future segmentation studies can improve upon. For a study to be fully open source, code, data, coding environment and model weights used in the analysis have to be released with the publication. Of those who released their code, lack of compliance to current technological reproducibility standards prevents future validation [28]. We encourage authors whose studies follow ours to adopt an

open-source friendly, reproducibility centric policy when releasing their work including any significant improvements to our methods.

The current study showed the relative effectiveness of simple UNET based networks on complex multi-class segmentation tasks. For this analysis, we choose a loss based approach to combat the large class imbalance problem that exists within the OAR segmentation problem of the HNC region. However, other studies have shown that loss based approaches combined with more complex architectural designs may improve segmentation results [16,19,27] It was not a surprise, however, that a simple UNET variant beat out more complex networks tested in this study. State of the art segmentation is dominated by the UNET and its variants, proving that the simplest of networks can compete with the most advanced networks of our day to achieve state of the art performances. Our results agree with other analyses which have concluded that the issue affecting poor auto-segmentation model performance does not lie at the level of model complexity, but rather the steps surrounding model training. They have found that optimizing data preprocessing and postprocessing at the target level can dramatically improve model performance [75].

There can be differences by which experts passively identify errors in contours before assessing contour acceptability. Experts may focus on detecting errors at the boundaries of OARs. Additionally, experts may focus on determining whether a contour displays adequate volumetric coverage over the target of interest. Keeping this in mind, it is intriguing to note, that when Mean Acceptability Rating (MAR) was compared against all the performance metrics used to adjudicate model accuracy across all OARs, the most significant correlation occurred with respect to boundary distance metrics (Figure 3c 95HD : -.26 , SD : -.30). In other words, our findings suggest that clinicians are more likely to focus on detecting errors made around the boundaries of OARs when assessing acceptability, rather than focusing on volumetric errors in the complete 3D coverage of a contour (which can be modeled by a volumetric based statistic like DICE). Large errors made around contour boundaries imply high boundary distance metrics, which in turn, results in low acceptability ratings.

This study is the first of its kind to establish robust baselines of clinical acceptability for both manual contours and DLC. Previous studies only reported showing observers one slice of a 3D contour when asking them to assess acceptability[18]. This is a poor representation of acceptability as a large majority of the OAR contour using previous methods fails to be

assessed. We engineered a more robust version of clinical acceptability testing in QUANNOTATE that gave expert observers the ability to analyze the entire extent contour in question on all CT imaging slices when performing the assessment. In order for DLC to be superior and pass the test, the contour must be generated accurately across the entire span of the 3D OAR in question, not just a single 2D slice. In theory, OARs with high-ranked performance metrics could fail the test if any part of the 3D contour under examination looks 'questionable'. Clinical Acceptability of entire 3D volumes allows us to paint a more accurate picture of deep learning contour fidelity by finding errors present at the edges of complex ROI(s). When analyzing results we noticed this occurred during the assessment of deep learning generated larynx contours. When analyzing the contours produced, the failure in performance was caused by broken segmentations produced at the upper and lower bounds of the larynx. The poor performance could be explained by the complexity of the larynx itself but could also represent a novel limitation of current segmentation models when presented with poorly defined 'edges' of therapy structures (Supplementary Figure 9). As an OAR the larynx is composed of 3 CT densities (cartilage/bone, soft tissue, and air) while the other OARs have soft tissue or bone. The complexity of the larynx contour could have caused inconsistencies when the ground truth contours were originally produced. We found that some ground truth contours included the cartilage, while others included only the soft tissue. These inconsistencies within our training data caused poor convergence at the boundaries of the larynx. A robust assessment of clinical acceptability like the one performed in this study is required before applying any deep learning based auto contouring system in a clinical setting.

Conducting external validation experiments on multiple datasets without fine tuning to external data was essential to assess the direct generalizability of the WOLNET implementation. Regardless of the dataset, the WOLNET ensemble performed well at segmenting large OARs like the Mandible and Brain Stem and OARs with homogeneous structures like the eyeballs and parotids. Despite adequate performance across most OARs in external datasets tested, we are confident that fine tuning these networks to external datasets will improve performance at the dataset level. In addition to network finetuning, improvement of protocols related to the following can also boost future network generalizability. These include: improved inference using adaptable sliding windows; increasing receptive field of the network during training; ensuring equal distribution of images take from different scanners and centers is used during training; leveraging full use of the datasets at hand by conditioning networks to learn from all data we have at our disposal and optimizing networks using custom loss functions engineered from

metrics most correlated with clinical acceptability. Future work seeks to establish acceptability standards for simple segmentation networks applied to multiple sites outside the head and neck region.

CONCLUSION

These results show that simple open-source 3D architectures consistently outcompete more complex networks by quantitative measures. Qualitative assessment for clinical acceptability may not agree with quantitative performance, especially when the entire range of OARs is evaluated. Greater weight should be placed to optimize auto-segmentation systems for boundary distance metrics to produce more 'clinically acceptable' contours. Clinical acceptability testing and the open-source frameworks that provide networks and interfaces on which these tests can be conducted, will prove invaluable to the future adoption of deep learning based auto segmentation systems in clinical RT practices.

REFERENCES

1. Sims, R. *et al.* A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. *Radiotherapy and Oncology* vol. 93 474–478 (2009).
2. Stapleford, L. J. *et al.* Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **77**, 959–966 (2010).
3. Qazi, A. A. *et al.* Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach. *Medical Physics* vol. 38 6160–6170 (2011).
4. Fortunati, V. *et al.* Tissue segmentation of head and neck CT images for treatment planning: a multiatlas approach combined with intensity modeling. *Med. Phys.* **40**, 071905 (2013).
5. Thomson, D. *et al.* Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiat. Oncol.* **9**, 173 (2014).
6. Fritscher, K. D. *et al.* Automatic segmentation of head and neck CT images for radiotherapy

treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. *Medical Physics* vol. 41 051910 (2014).

7. Fritscher, K. *et al.* Deep Neural Networks for Fast Segmentation of 3D Medical Images. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* 158–165 (Springer International Publishing, 2016).
8. Ibragimov, B. & Xing, L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Medical Physics* vol. 44 547–557 (2017).
9. Močnik, D. *et al.* Segmentation of parotid glands from registered CT and MR images. *Phys. Med.* **52**, 33–41 (2018).
10. Ren, X. *et al.* Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images. *Medical Physics* vol. 45 2063–2075 (2018).
11. Nikolov, S. *et al.* Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv [cs.CV]* (2018).
12. Tappeiner, E. *et al.* Multi-organ segmentation of the head and neck area: an efficient hierarchical neural networks approach. *Int. J. Comput. Assist. Radiol. Surg.* **14**, 745–754 (2019).
13. Zhu, W. *et al.* AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med. Phys.* **46**, 576–589 (2019).
14. van Rooij, W. *et al.* Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation. *Int. J. Radiat. Oncol. Biol. Phys.* **104**, 677–684 (2019).
15. Zhong, T. *et al.* Boosting-based Cascaded Convolutional Neural Networks for the Segmentation of CT Organs-at-risk in Nasopharyngeal Carcinoma. *Med. Phys.* **46**, 5602–5611 (2019).
16. Tang, H. *et al.* Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nature Machine Intelligence* **1**, 480–491 (2019).

17. Rhee, D. J. *et al.* Automatic detection of contouring errors using convolutional neural networks. *Med. Phys.* **46**, 5086–5097 (2019).

18. van Dijk, L. V. *et al.* Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiother. Oncol.* **142**, 115–123 (2020).

19. Guo, D. *et al.* Organ at Risk Segmentation for Head and Neck Cancer using Stratified Learning and Neural Architecture Search. *arXiv [cs.CV]* (2020).

20. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).

21. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer International Publishing, 2015).

22. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* 424–432 (Springer International Publishing, 2016).

23. Tong, N., Gou, S., Yang, S., Ruan, D. & Sheng, K. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Med. Phys.* **45**, 4558–4567 (2018).

24. Gao, Y. *et al.* FocusNet: Imbalanced Large and Small Organ Segmentation with an End-to-End Deep Neural Network for Head and Neck CT Images. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* 829–838 (Springer International Publishing, 2019).

25. Wang, Y., Zhao, L., Wang, M. & Song, Z. Organ at Risk Segmentation in Head and Neck CT Images Using a Two-Stage Segmentation Framework Based on 3D U-Net. *IEEE Access* **7**, 144591–144602 (2019).

26. Zhang, S. *et al.* A slice classification model-facilitated 3D encoder–decoder network for segmenting organs at risk in head and neck cancer. *J. Radiat. Res.* **62**, 94–103 (2020).
27. Gao, Y. *et al.* FocusNetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images. *Med. Image Anal.* **67**, 101831 (2021).
28. Haibe-Kains, B. *et al.* Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16 (2020).
29. Prior, F. W. *et al.* TCIA: An information resource to enable open science. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2013**, 1282–1285 (2013).
30. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
31. Wong, K. *et al.* Point-of-care outcome assessment in the cancer clinic: audit of data quality. *Radiother. Oncol.* **95**, 339–343 (2010).
32. O'Sullivan, B. *et al.* Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. *Lancet Oncol.* **17**, 440–451 (2016).
33. Haibe-Kains, B. *et al.* A Machine Learning Challenge for Prognostic Modelling in Head and Neck Cancer Using Multi-modal Data. doi:10.21203/rs.3.rs-185311/v1.
34. Arrowsmith, C. *et al.* Automated detection of dental artifacts for large-scale radiomic analysis in radiation oncology. *Phys Imaging Radiat Oncol* **18**, 41–47 (2021).
35. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv [cs.LG]* (2019).
36. Falcon, W. *et al. PyTorch Lightning: The lightweight PyTorch wrapper for high-performance AI research. 1.3.6 release*. (2021). doi:10.5281/zenodo.3828935.
37. Abraham, N. & Khan, N. M. A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation. in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* 683–687 (2019).

38. Jun Ma. *SegLoss: A collection of loss functions for medical image segmentation*. (Github).

39. Liu, L. *RAdam*. (Github).

40. Liu, L. *et al.* On the Variance of the Adaptive Learning Rate and Beyond. *arXiv [cs.LG]* (2019).

41. Beare, R., Lowekamp, B. & Yaniv, Z. Image segmentation, registration and characterization in R with SimpleITK. *J. Stat. Softw.* **86**, (2018).

42. Eric Kerfoot, Raghav Mi, Tom Vercauteren, Wenqi Li. *MONAI: AI Toolkit for Healthcare Imaging Release 0.5.3*. (Github).

43. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv [cs.LG]* (2016).

44. Kim, J. W. *et al.* Development of web-based quality-assurance tool for radiotherapy target delineation for head and neck cancer: quality evaluation of nasopharyngeal carcinoma. doi:10.1101/2021.02.24.21252123.

45. Cardenas, C. E. *et al.* Generating High-Quality Lymph Node Clinical Target Volumes for Head and Neck Cancer Radiation Therapy Using a Fully Automated Deep Learning-Based Approach. *International Journal of Radiation Oncology*Biology*Physics* vol. 109 801–812 (2021).

46. Vaassen, F. *et al.* Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol* **13**, 1–6 (2020).

47. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).

48. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).

49. Li, H. & Chen, M. Automatic Structure Segmentation for Radio Therapy Planning Challenge 2020. in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru* 4–8 (2020).

50. Raudaschl, P. F. *et al.* Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. *Med. Phys.* **44**, 2020–2036 (2017).
51. Bejarano, T., De Ornelas-Couto, M. & Mihaylov, I. B. Longitudinal fan-beam computed tomography dataset for head-and-neck squamous cell carcinoma patients. *Med. Phys.* **46**, 2526–2537 (2019).
52. Lee, K., Zung, J., Li, P., Jain, V. & Sebastian Seung, H. Superhuman Accuracy on the SNEMI3D Connectomics Challenge. *arXiv [cs.CV]* (2017).
53. Li, W. *et al.* On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task. *arXiv [cs.CV]* (2017).
54. Fang, X. & Yan, P. Multi-Organ Segmentation Over Partially Labeled Datasets With Multi-Scale Feature Abstraction. *IEEE Trans. Med. Imaging* **39**, 3619–3629 (2020).
55. Huang, H. *et al.* UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1055–1059 (2020).
56. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018)* **11045**, 3–11 (2018).
57. Yu, L. *et al.* Automatic 3D Cardiovascular MR Segmentation with Densely-Connected Volumetric ConvNets. *arXiv [cs.CV]* (2017).
58. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A. & Bengio, Y. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. *arXiv [cs.CV]* (2016).
59. Zhang, H. *et al.* RSANet: Recurrent Slice-Wise Attention Network for Multiple Sclerosis Lesion Segmentation. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* 411–419 (Springer International Publishing, 2019).
60. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *arXiv [cs.CV]* (2016).

61. Wolny, A. *et al.* Accurate and versatile 3D segmentation of plant tissues at cellular resolution. *Elife* **9**, (2020).

62. Wolny, A. *pytorch-3dunet: 3D U-Net model for volumetric semantic segmentation written in pytorch*. (Github).

63. Nikolas, A. *HighResNet3D.py at master · black0017/MedicalZooPytorch*. (Github).

64. *PIPO-FAN: PIPO-FAN for multi organ segmentation over partial labeled datasets using pytorch*. (Github).

65. *UNet-Version*. (Github).

66. *UNet-Version*. (Github).

67. Zhu, W. *AnatomyNet-for-anatomical-segmentation: AnatomyNet: Deep 3D Squeeze-and-excitation U-Nets for fast and fully automated whole-volume anatomical segmentation*. (Github).

68. Nikolas, A. *DenseVoxelNet.py at master · black0017/MedicalZooPytorch*. (Github).

69. Fortuner, B. *pytorch_tiramisu: FC-DenseNet in PyTorch for Semantic Segmentation*. (Github).

70. tinymilky. *RSANet: RSANet: Recurrent Slice-wise Attention Network for Multiple Sclerosis Lesion Segmentation (MICCAI 2019)*. (Github).

71. Macy, M. *vnet.pytorch: A PyTorch implementation for V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*. (Github).

72. Chollet, F. *Deep Learning with Python, Second Edition*. (Simon and Schuster, 2021).

73. Ho, S. Y., Phua, K., Wong, L. & Bin Goh, W. W. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. *Patterns (N Y)* **1**, 100129 (2020).

74. Bejarano, T., De Ornelas-Couto, M. & Mihaylov, I. B. Longitudinal fan-beam computed tomography dataset for head-and-neck squamous cell carcinoma patients. *Med. Phys.* **46**, 2526–2537 (2019).

75. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M. & Maier-Hein, K. H. No New-Net. in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* 234–244 (Springer International Publishing, 2019).

TABLES & FIGURES

Table 1. Summary of published OAR segmentation studies and their open-source status. Only AnatomyNet [13] (in bold) could be included in our study as it.

Study	Code	Name (if Applicable)	Model Type	Training Data	Model Weights	# Patients	# OARs
Ibragimov, 2017 [8]	No		3D CNN	No	No	50	14
Močnik, 2018 [9]	No		2D CNN	No	No	85	1
Ren, 2018 [10]	No		3D VGG	No	No	98	3
Tappeiner , 2019 [12]	Yes		3D RESNET	NA	Yes	40	9
Nikolov, 2018 [11]	No		2.5D UNET	No	No	663	21
Tong, 2018 (Tong et al. 2018)	No	FCN+SRM	3D FCN	No	No	32	8
Zhu, 2019 [13]	**Yes**	**AnatomyNet**	**3D UNET**	**No**	**No**	**271**	**9**
van Dijk, 2019 [18]	No		2D UNET	No	No	693	22
Rooij, 2019 [14]	No		3D UNET	No	No	157	11
Gao, 2019 (Gao et al. 2019)	No	FocusNetv1	3D UNET	No	No	50	18
Zhong 2019 [15]	No		2D RESNET	No	No	140	5
Tang, 2019 [16]	Yes	UaNet	3D UNET	No	Yes	215	28
Wang, 2019 (Wang et al. 2019)	No		3D UNET	No	No	50	14
Rhee, 2020 [17]	No		3D UNET	No	No	3693	16
Guo, 2020 [19]	No	SOARS	3D P-HNN	No	No	142	42
Zhang, 2021 (Zhang et al. 2020)	No		3D FCN	No	No	170	12
Gao, 2021[24]	Yes	FocusNetv2	3D UNET	No	No	1164	22

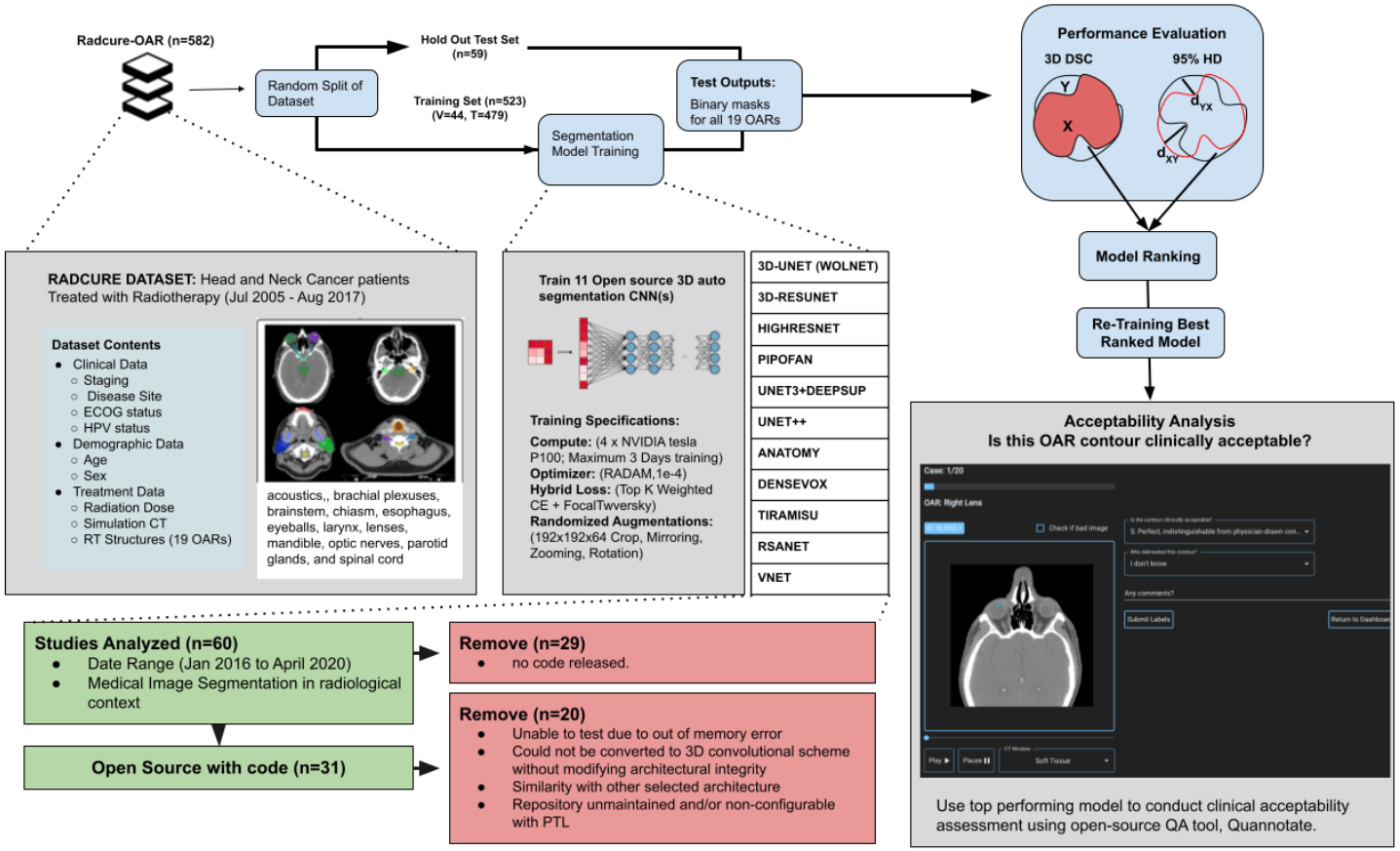


Figure 1. Study Overview: A general overview of the study design. A total of 582 patients met selection criteria having the same 19 OAR(s) delineated in their RTSTRUCT files. Data was extracted and used to train 11 different open source 3D segmentation networks. The 11 networks were selected as a subset of 60 studies proposing implementing image segmentation networks in a medical context. A subset of 29 studies did not release code along with their publication, and therefore, these models could not be directly assessed. Another 20 studies were removed because of one or more of the following: networks could not be tested with original configuration due to limitation of computational power, 2D networks that could not be converted to 3D convolutional scheme without modifying it's architectural integrity, close similarity or overlap between other architectures, code released with the study was unmaintained and/or could not be integrated into PytorchLightning. The networks were ranked based on overall performance on a hold out test set of 59 patients across all 19 OARs. The top model was then chosen to be fine-tuned and used in a blinded clinical acceptability assessment conducted by 4 expert radiation oncologists on the open source Quannotate platform.

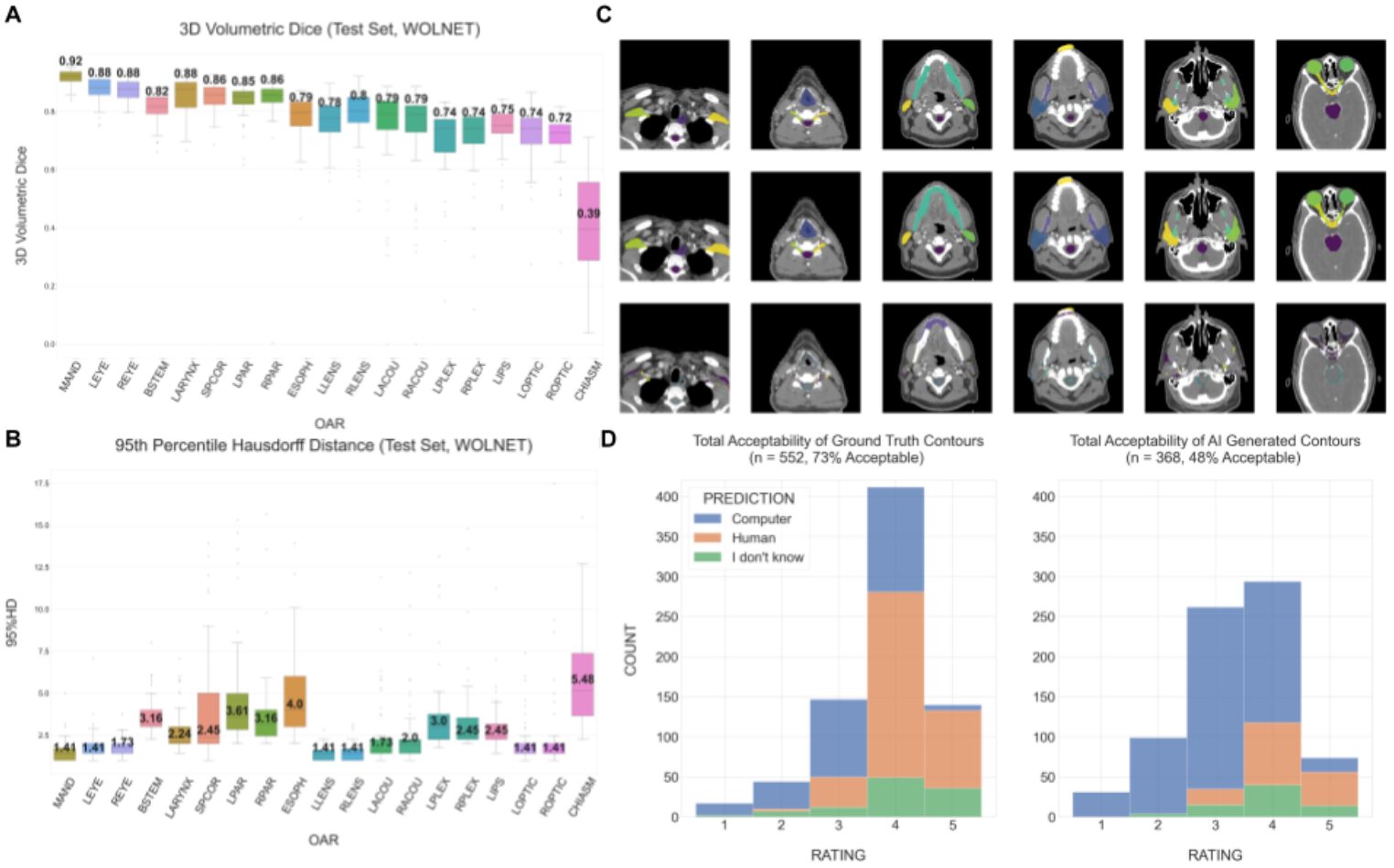


Figure 2. Preliminary results for organs-at-risk (OARs) segmentation for the best performing network of the 11 reimplemented models (WOLNET) on test set samples in RADCURE. A. 3D Volumetric Dice for each OAR B. Contours from random test set patient C. 95th Percentile Hausdorff Distance for each OAR D. Preliminary results for the clinical acceptability test of the WOLNET predictions using our open-source where 73% of manual contours analyzed were considered acceptable, compared with 48% of deep learning contours.

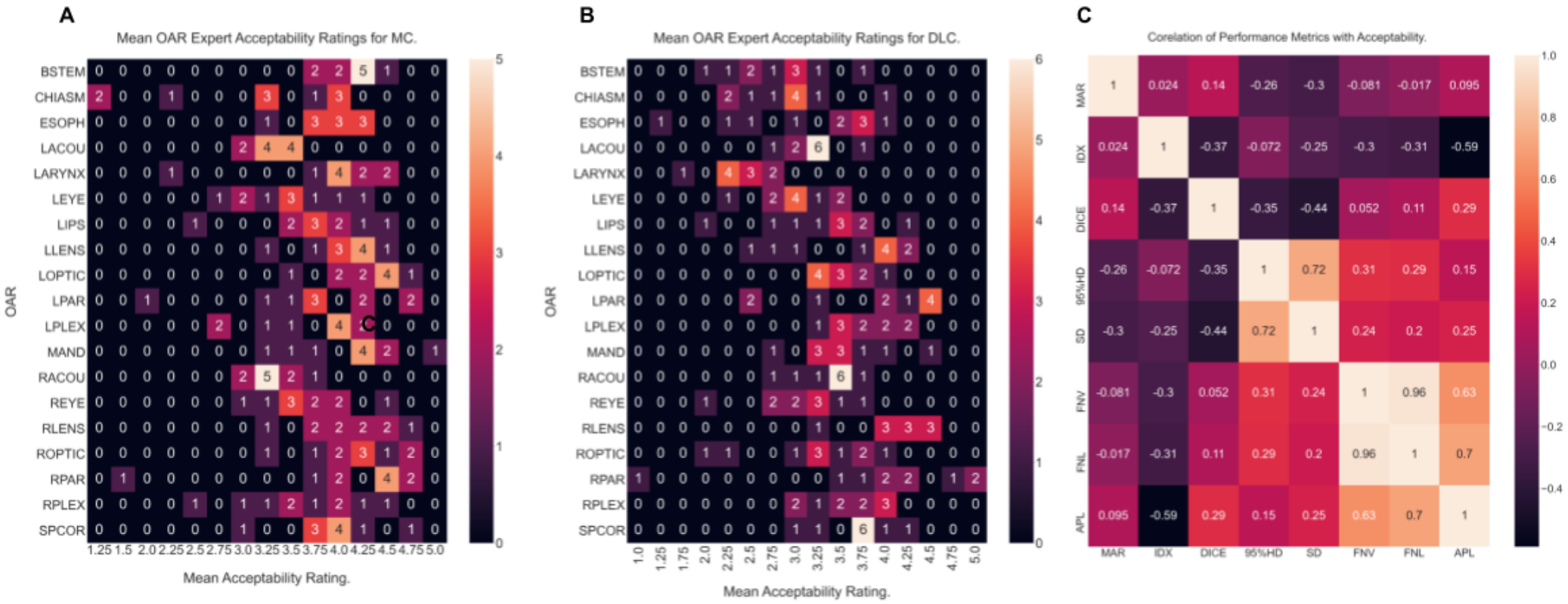
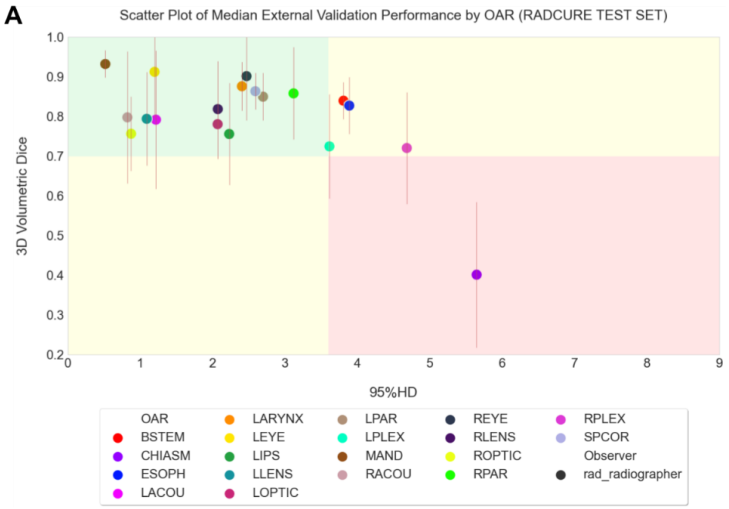
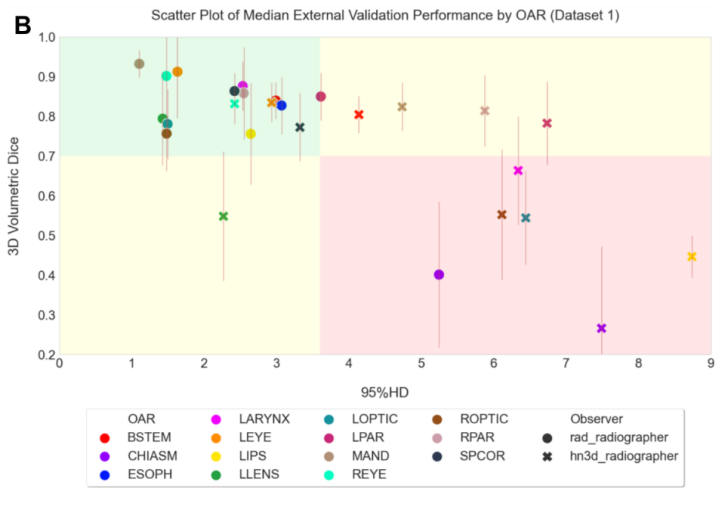


Figure 3. Panel A shows the results of the acceptability test by representing mean acceptability rating (MAR) counts for each OAR in a heatmap for manual ground truth contours. The higher the value of a box the more contours of that given OAR (row) had any given MAR value (column) and the lighter that box will be. Notice a shift to the left when examining the heatmap of mean acceptability ratings for deep learning contours examined for each OAR indicating a greater degree of clinical acceptance for manual contours as depicted by figure 4D. Manual contours received a significantly higher mean rating of 3.75 than deep learning contours which were rated 3.23 when all OARs were considered (3.75 ± 0.77 vs. 3.23 ± 0.86, p <0.01). Panel C plots mean acceptability rating correlation with 6 common segmentation metrics. Mean acceptability rating showed significant negative correlation with boundary distance metrics like 95% Hausdorf and Surface distances. (~-0.26 for 95% Hausdorff Distance and ~-0.30 for Surface Distance).

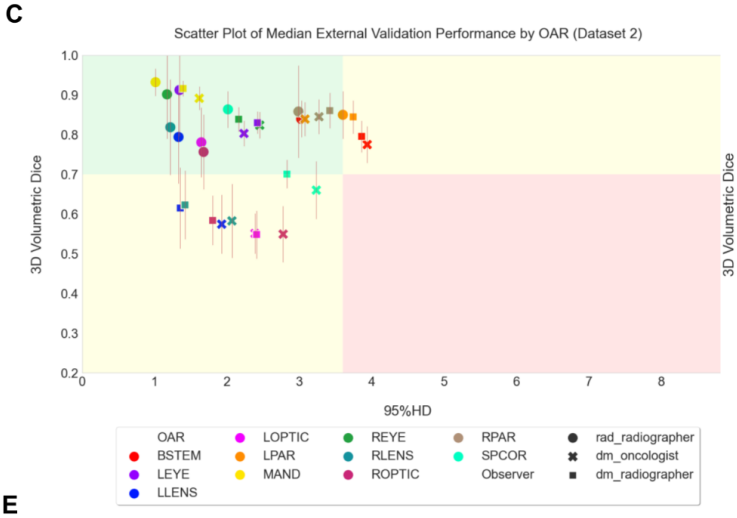
A



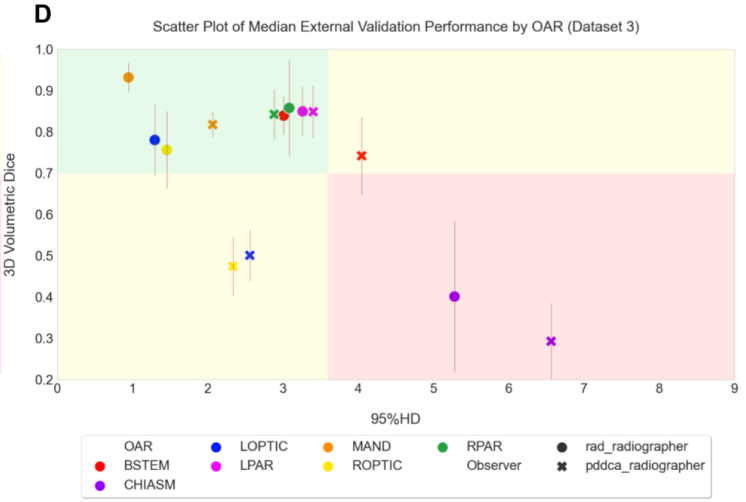
B



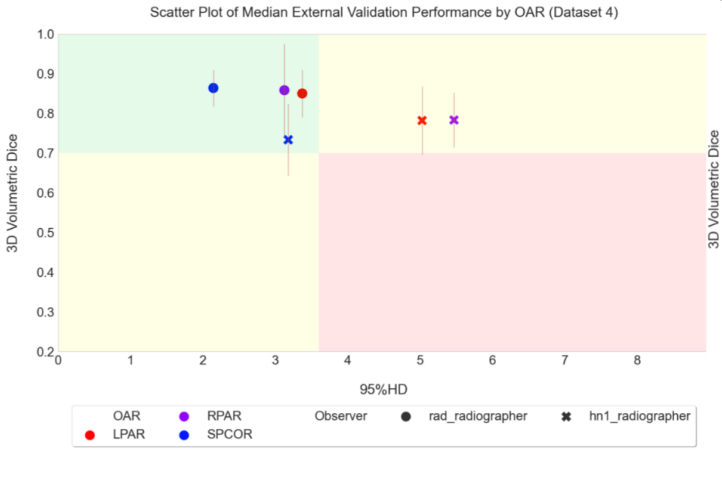
C



D



E



F

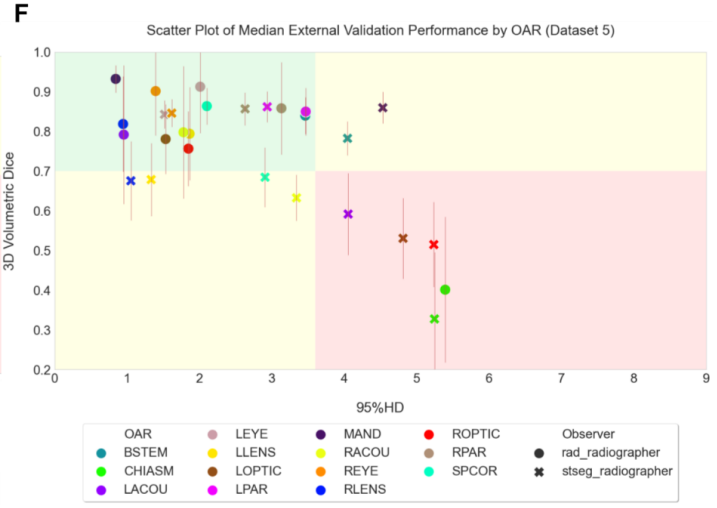
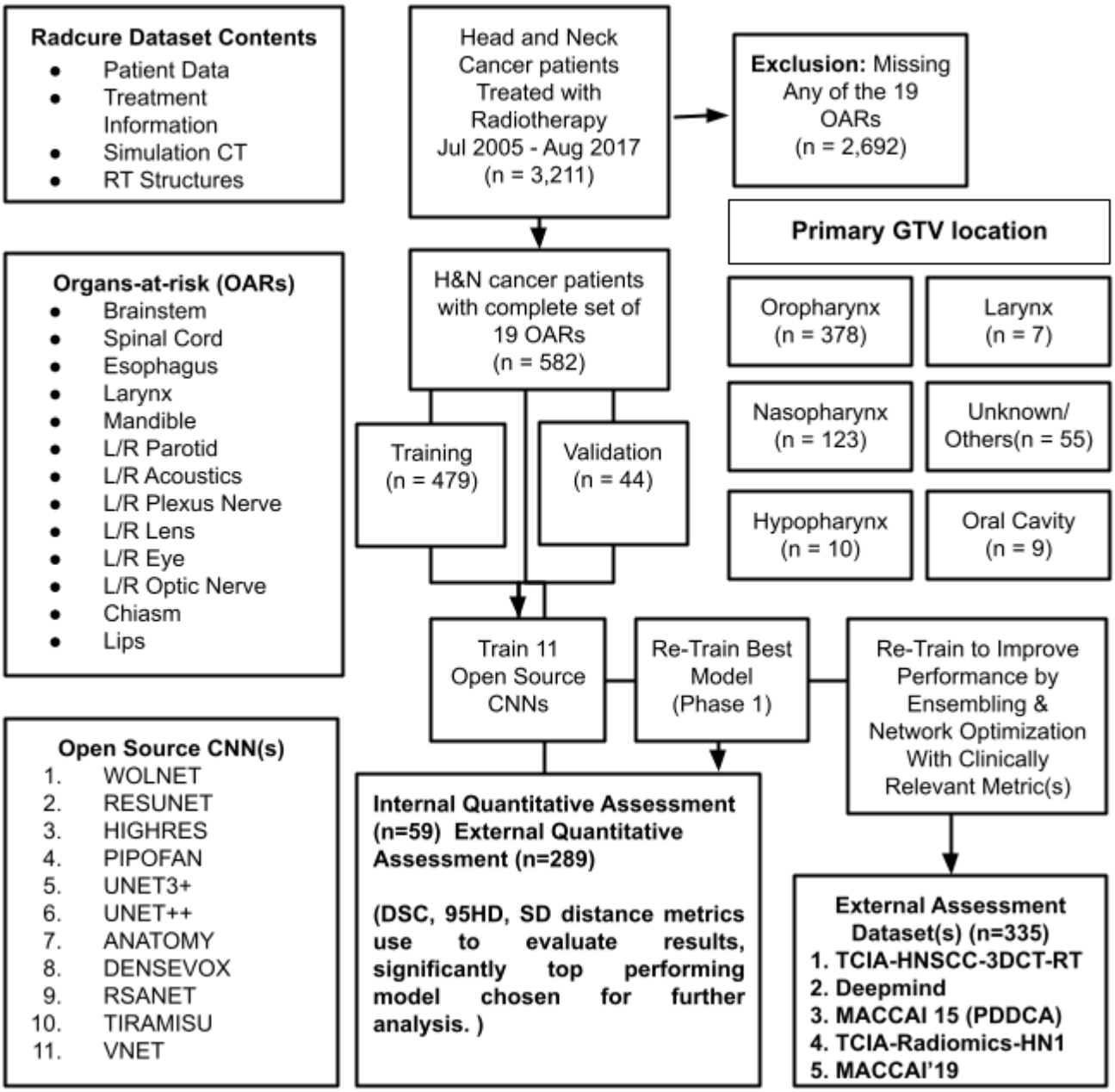
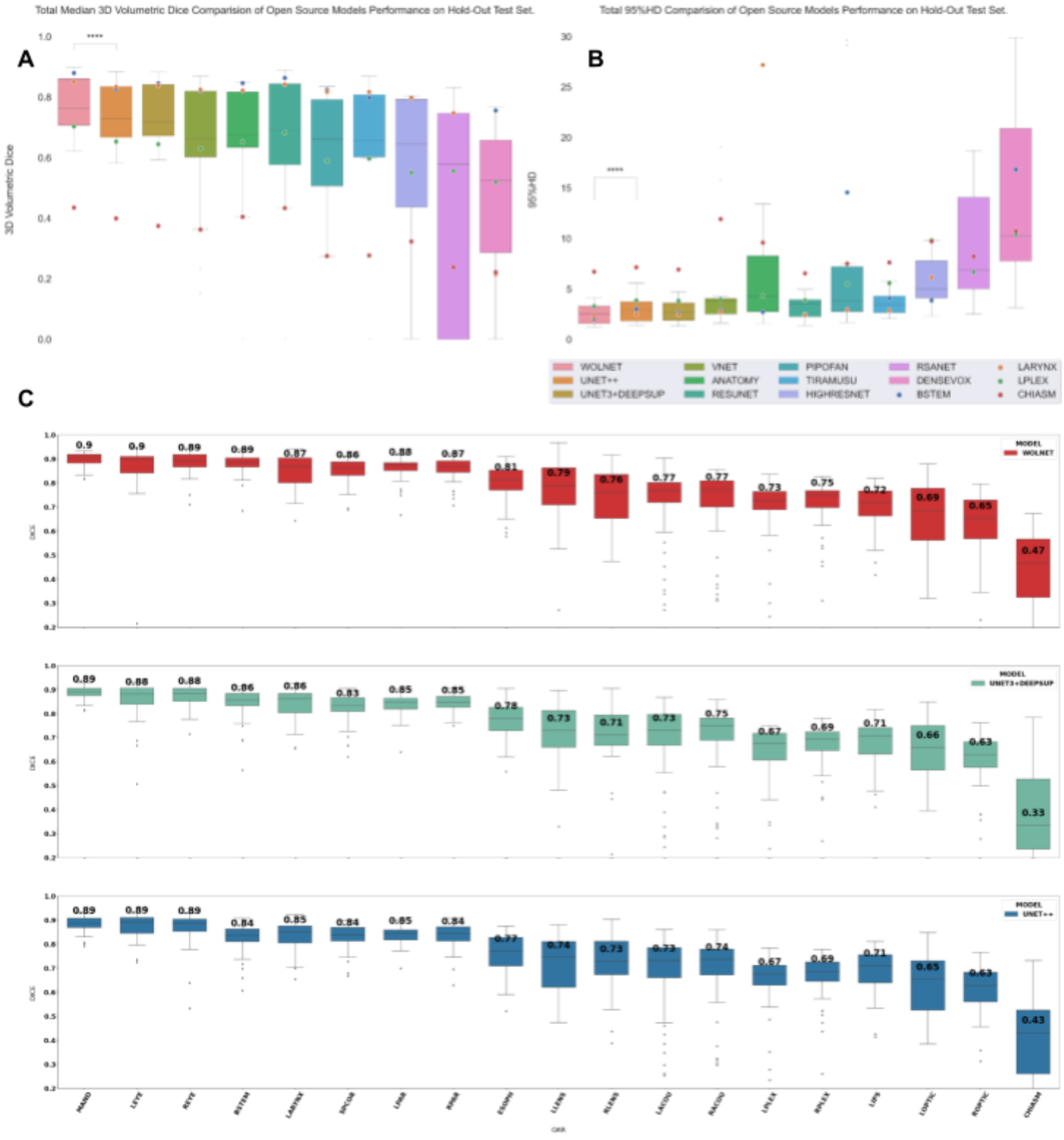


Figure 4: External Validation Performance of Optimized WOLNET Ensemble: Scatterplots plot median classical performance metrics (DICE, 95HD) for each OAR category. A: Radcure Test Set (n=59), B: Dataset 1: TCIA-HNSCC-3DCT-RT (n=83) C: Dataset 2: Deepmind (n=35), D: Dataset 3: MACCAI'15 - PDDCA (n=48), E: Dataset 4: TCIA-HN1-RADIOMICS (n=119) F: Dataset 5: MACCAI'19 - StructSeg19 (n=50)

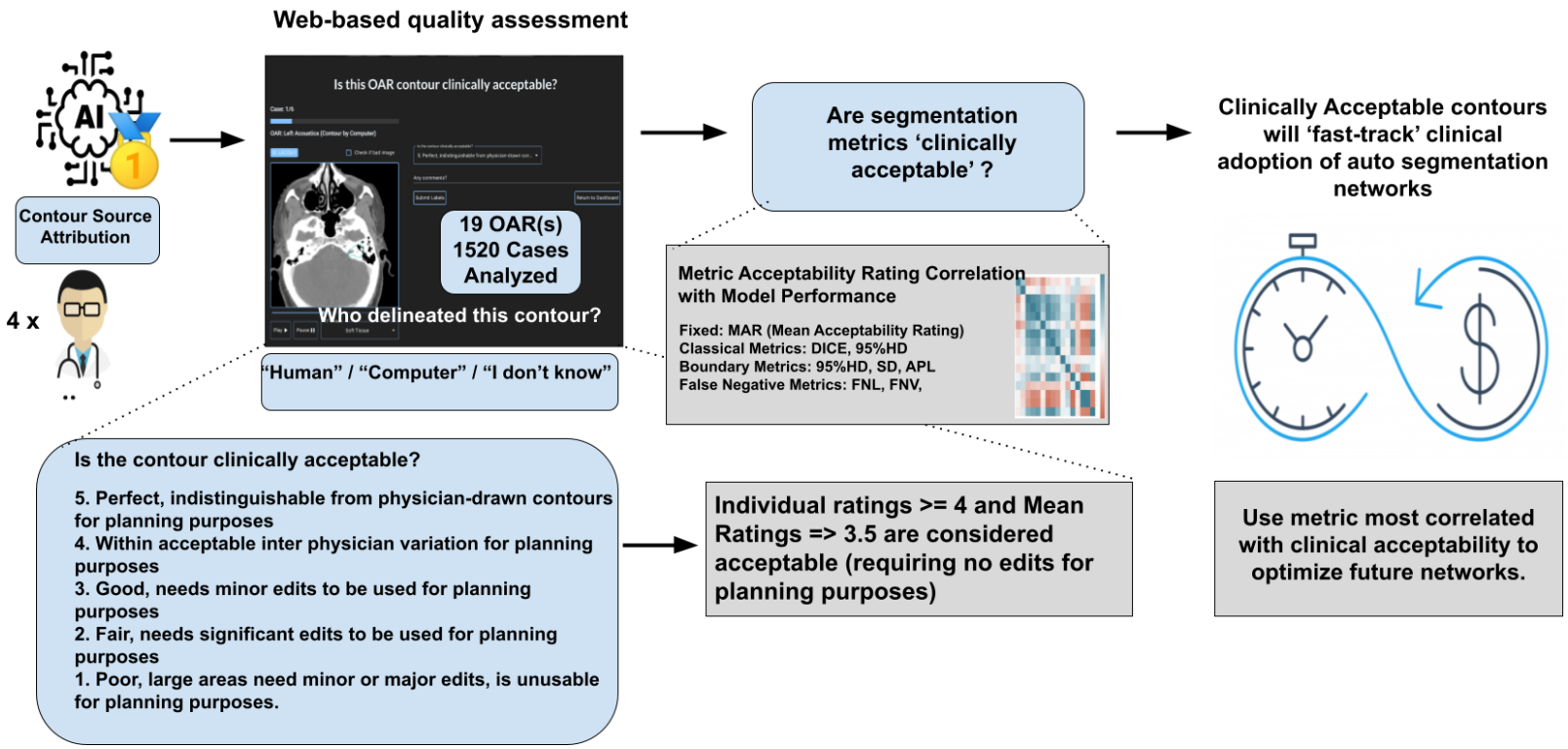
SUPPLEMENTARY MATERIALS



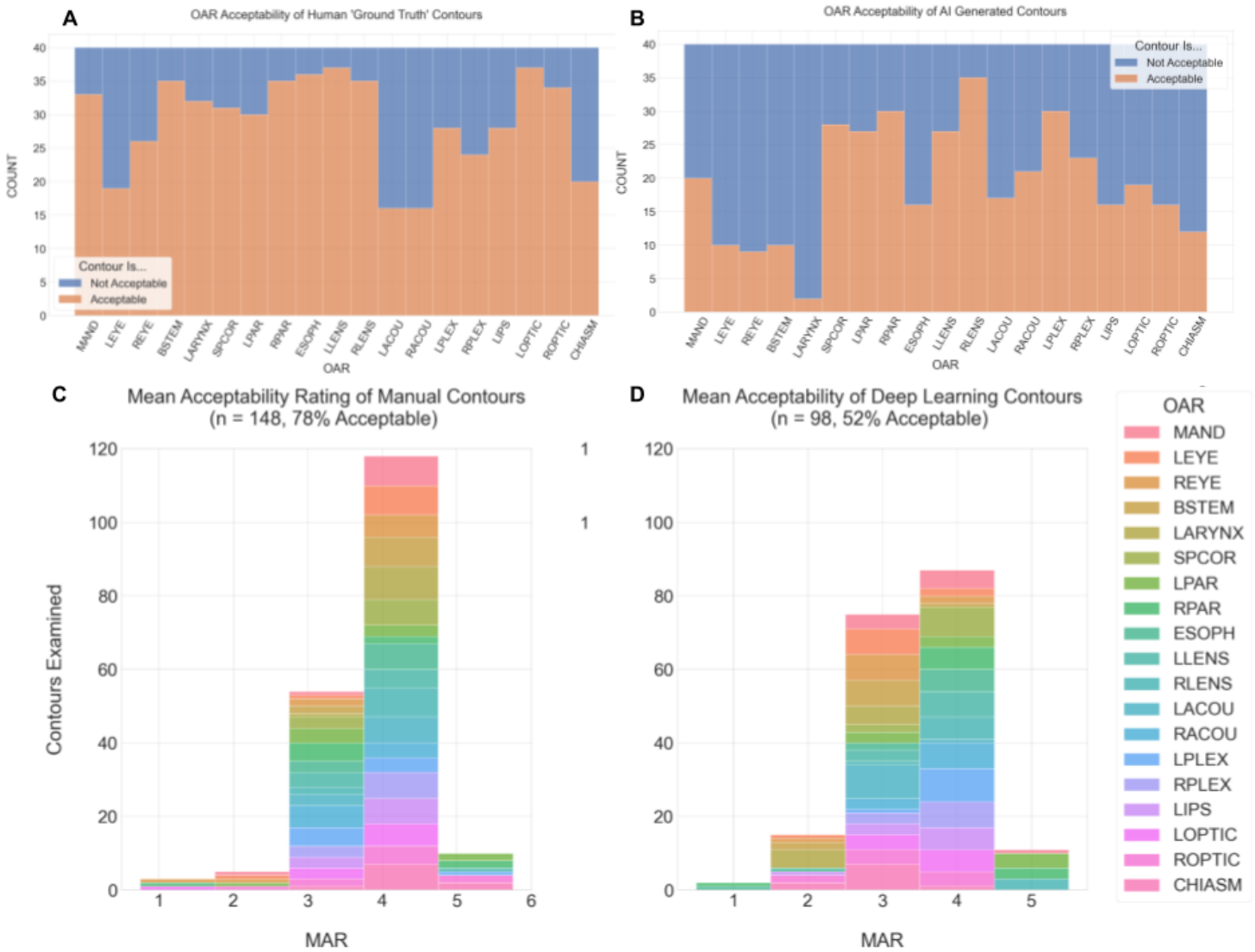
Supplementary Figure 1: Cohort Diagram



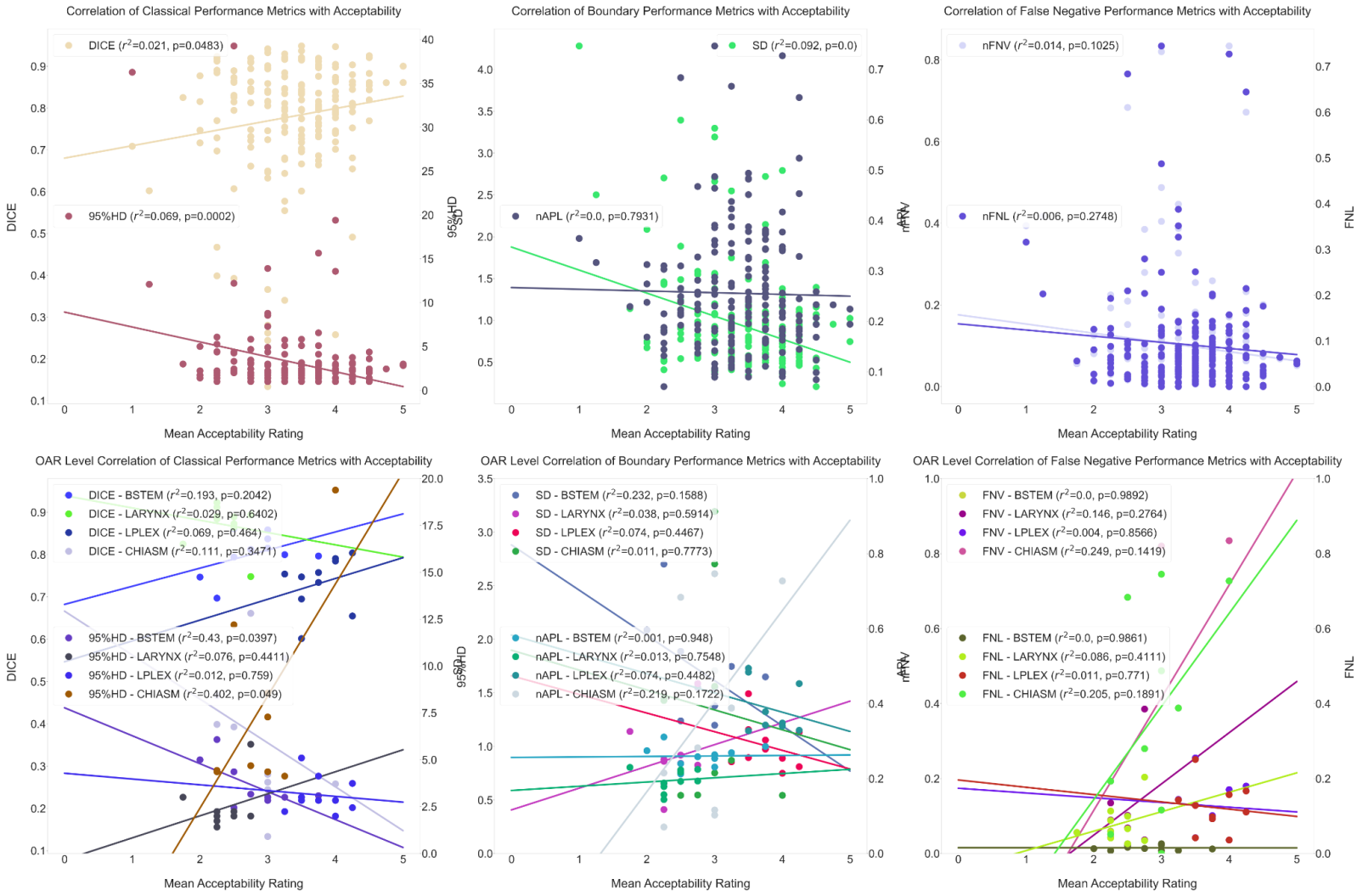
Supplementary Figure 3: a-b) Comparison of quantitative performance of all models using classical metrics (Hausdorff Distance and Volumetric Dice Coefficient) **c) Quantitative performance (DSC) of 3 best performing models for 19 OARs** Plotted are the distribution of volumetric dice values for the top performing models when applied to our test set, notice WOLNET is the single 3D CNN that produced superior segmentations for every OAR in our analysis.



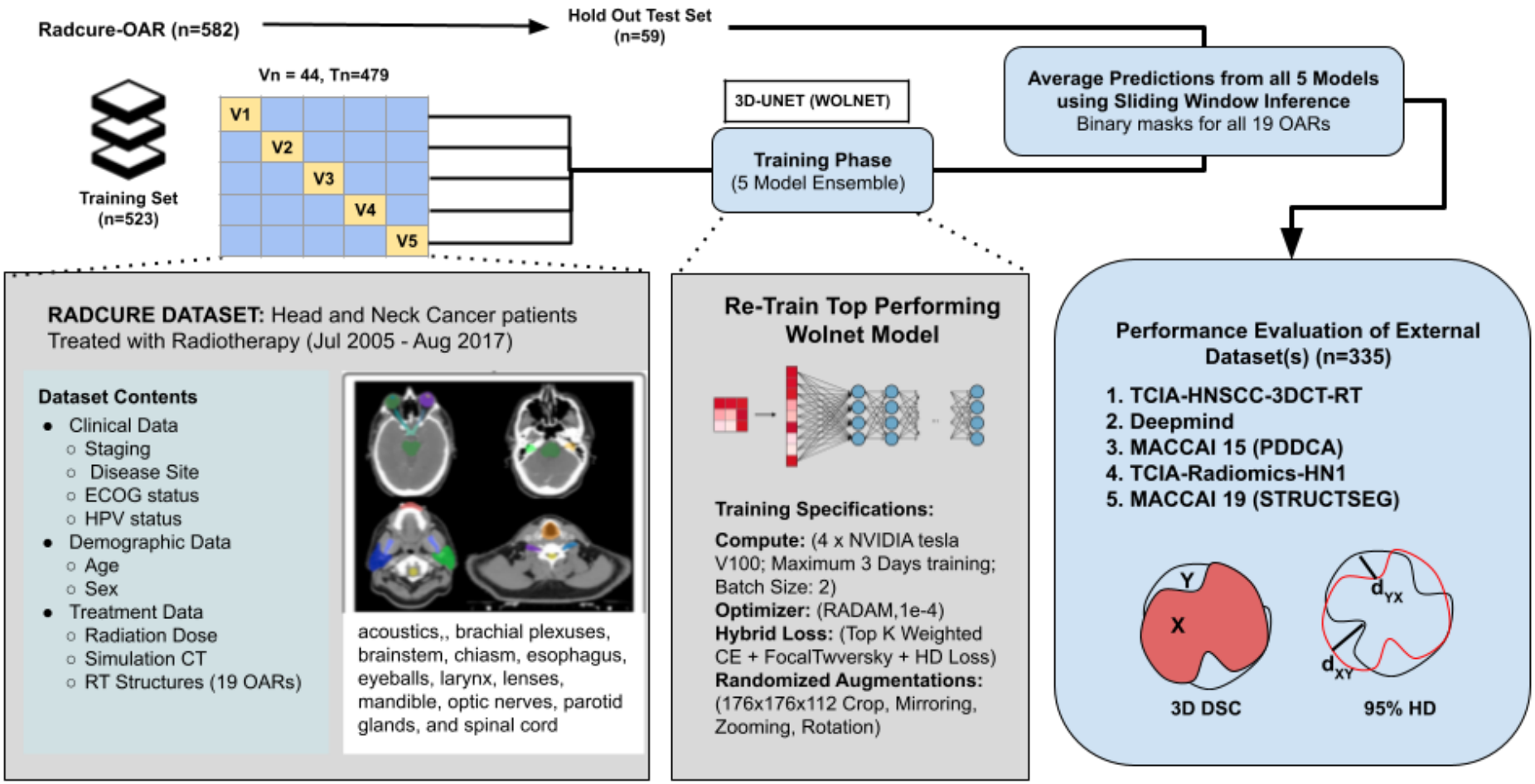
Supplementary Figure 2. Overview of Clinical Acceptability Protocol, an open source web-based quality assurance tool from our previous study was modified for clinical acceptability testing of radiation therapy contours. For this analysis, 4 expert radiation oncologists were each given the opportunity to assess the acceptability of deep learning or manual ground truth contours in a blinded fashion. 10 ground truth contours paired with 10 deep learning generated contours for the same patient were extracted for each of the 19 OARs. A total of 380 3D contours were assessed by each observer. They were asked to rate acceptability on a 5 point scale taking the complete volume of the entire OAR contour into context. A rating of 4.0 and higher can be considered 'acceptable' in that no edits are required by the examining physician for planning purposes. They were also asked to 'guess' the observer that generated the contour ("Human"/ "Computer" / "I don't Know") before submitting their rating. Mean Acceptability Ratings were then calculated for each OAR, and analysis assessing correlation of acceptability with 6 different segmentation performance metrics was conducted. These metrics include 3D Volumetric Dice Overlap Coefficient, 95% Hausdorff Distance, Applied Path Length, False Negative Length, and False Negative Volume



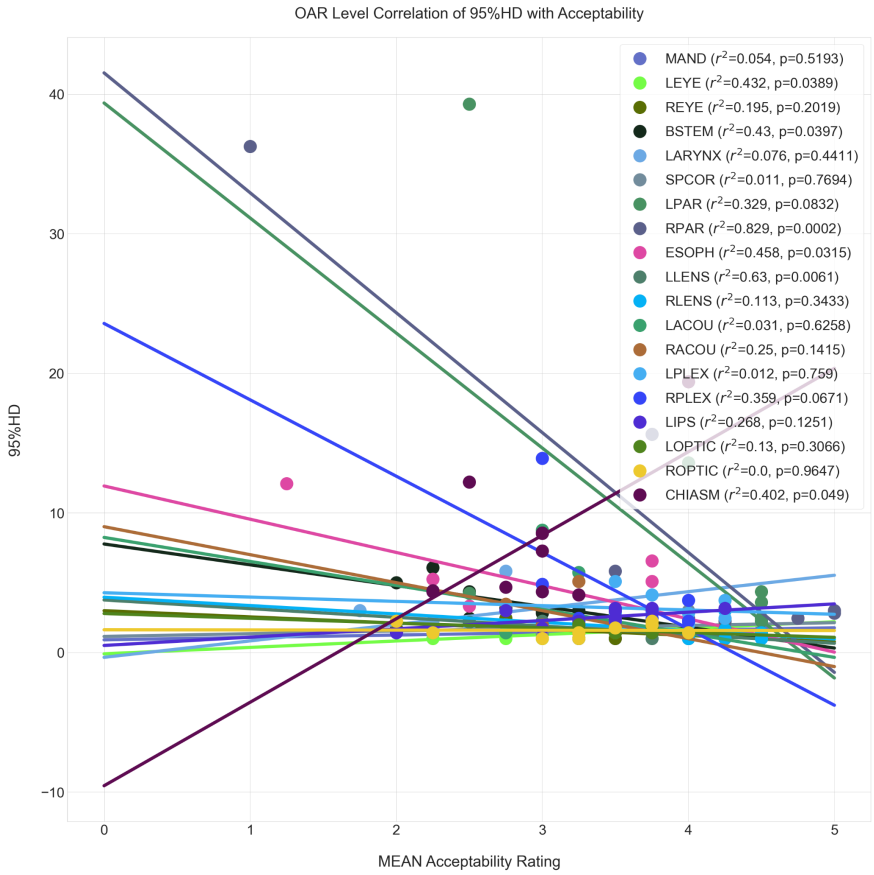
Supplementary Figure 4: (a) Bar plots of the unaveraged acceptability ratings for ground truth contours across all OARs **(b)** Bar plots of the unaveraged acceptability ratings for deep learning generated contours across all OARs. A single acceptability rating of 4 or more was considered 'acceptable' for this analysis (acceptable - orange; unacceptable - blue) **(c)** Histogram of mean acceptability rating (MAR) for manual ground truth contours **(d)** Histogram of mean acceptability rating (MAR) for deep learning generated contours. A MAR threshold of 3.5 or more was considered 'acceptable' for this analysis meaning that contour required no edits to be included into a radiation therapy plan. When analyzing individual OAR categories, Manual contours were considered more acceptable than deep learning contours for 15 out of the 19 OARs assessed. When assessing whether certain OAR categories passed the mean acceptability cutoff of 3.5, 15 manually delineated OARs on average were considered clinically acceptable, requiring no edits for planning purposes, compared with 9 OARs generated by deep learning.



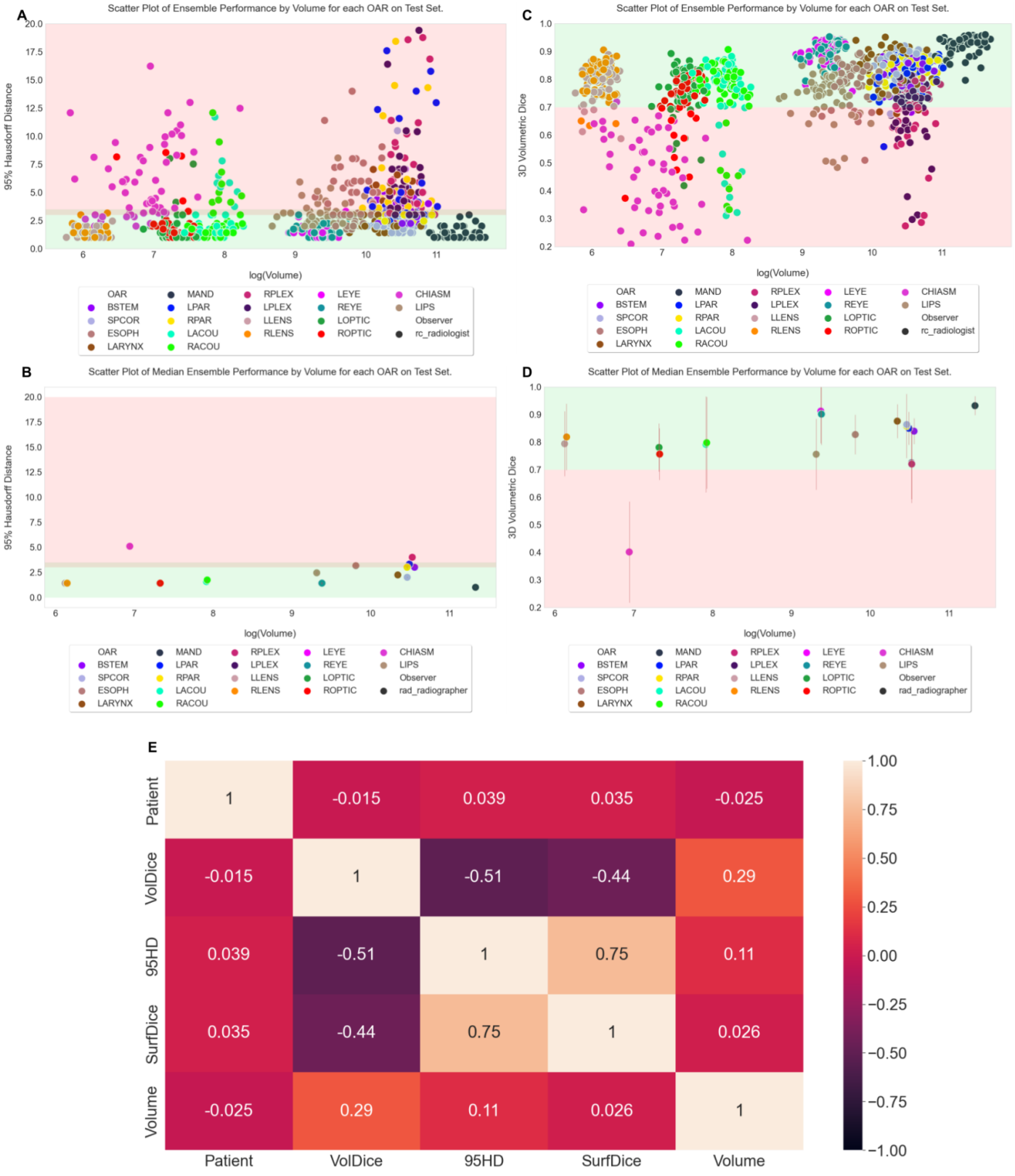
Supplementary Figure 5: Individual Performance Metric Correlation with Mean Acceptability Rating (MAR) Correlation of Mean acceptability with a) classical segmentation metrics (DICE, 95th percentile Hausdorff) b) advanced boundary metrics (surface distance (SD), added path length (APL)) c) false negative metrics (false negative volume (FNV), false negative length (FNL)) across all OARs. d) - f) Correlation of mean acceptability rating with various performance metrics for brain stem (BSTEM), Larynx, left brachial plexus (LPLEX), and Chiasm.



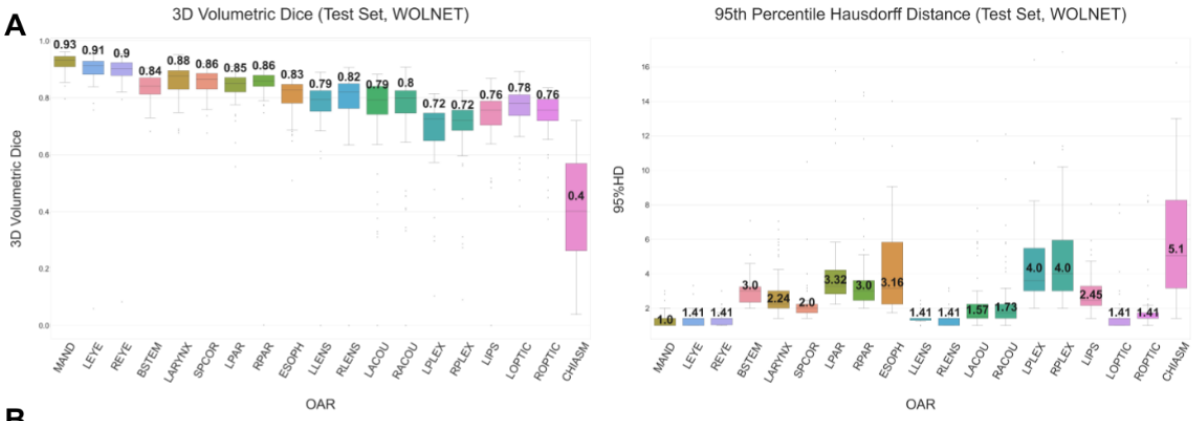
Supplementary Figure 7: Overview of Ensemble & WOLNET External Validation

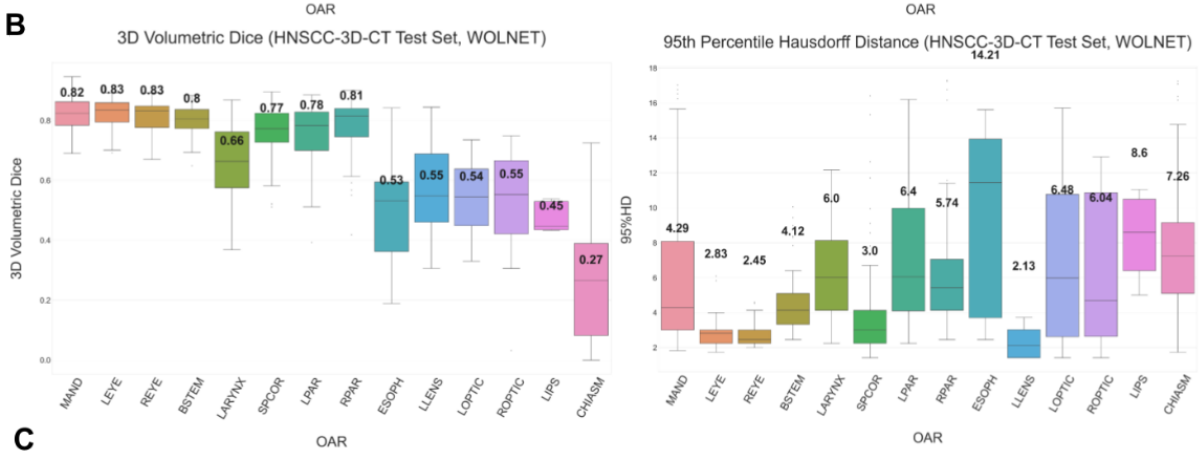


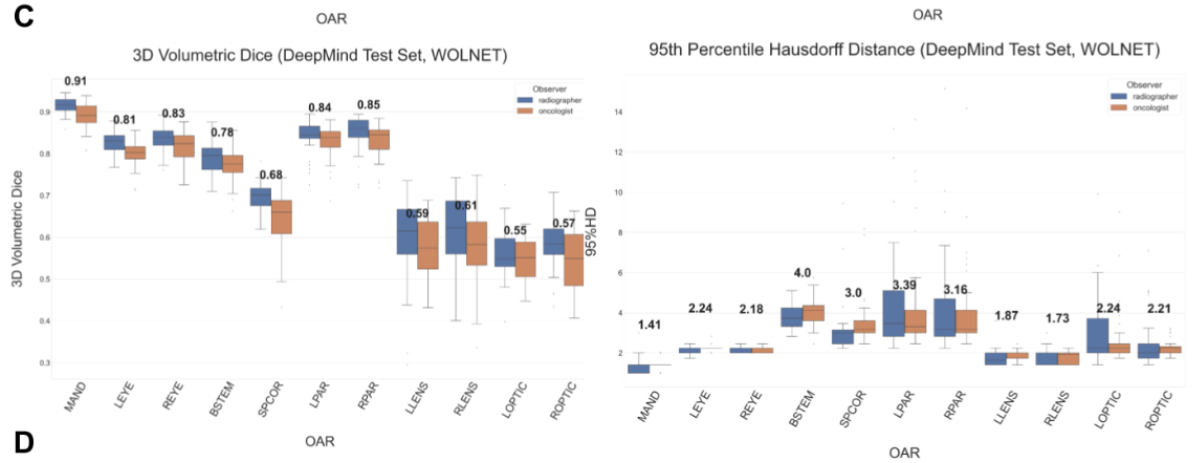
Supplementary Figure 6: Correlation of 95th Percentile Hausdorff Distance with Mean Acceptability Rating. Mean Acceptability Rating was significantly correlated with HD for 7 out of the 19 OARs analyze

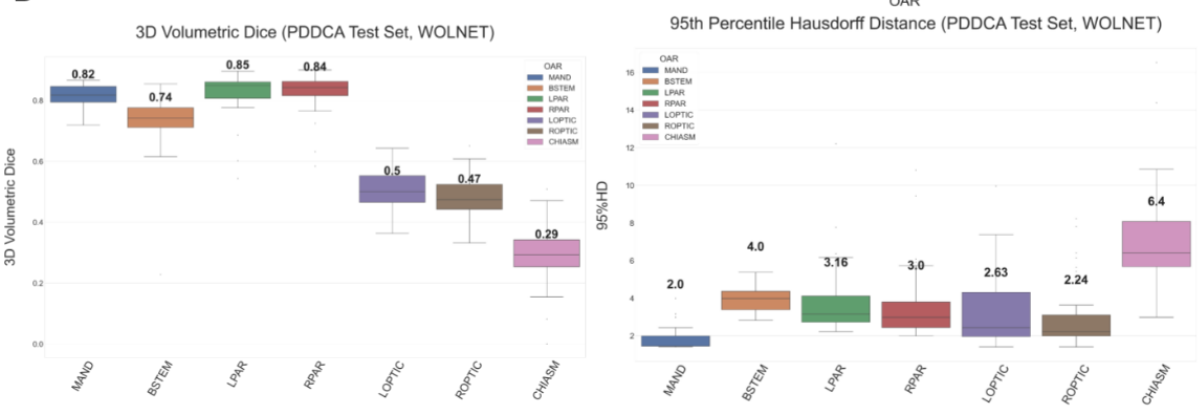


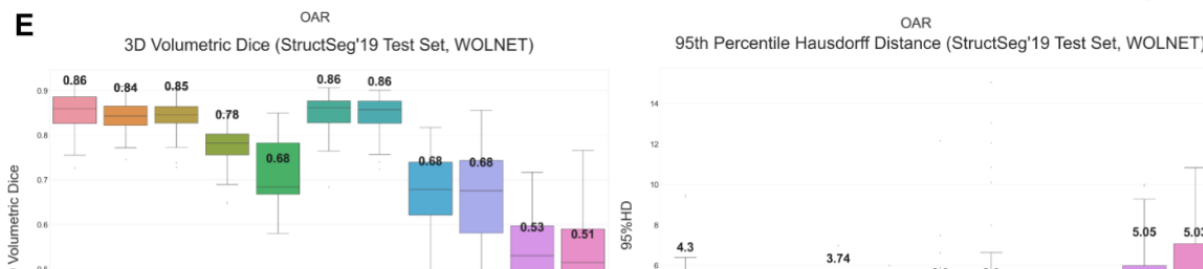
Supplementary Figure 8: Scatterplot of Ensemble Performance against OAR volume for DICE (A-B) 95HD (C-D); (E) Heatmap of correlation between classical metrics and OAR volume.



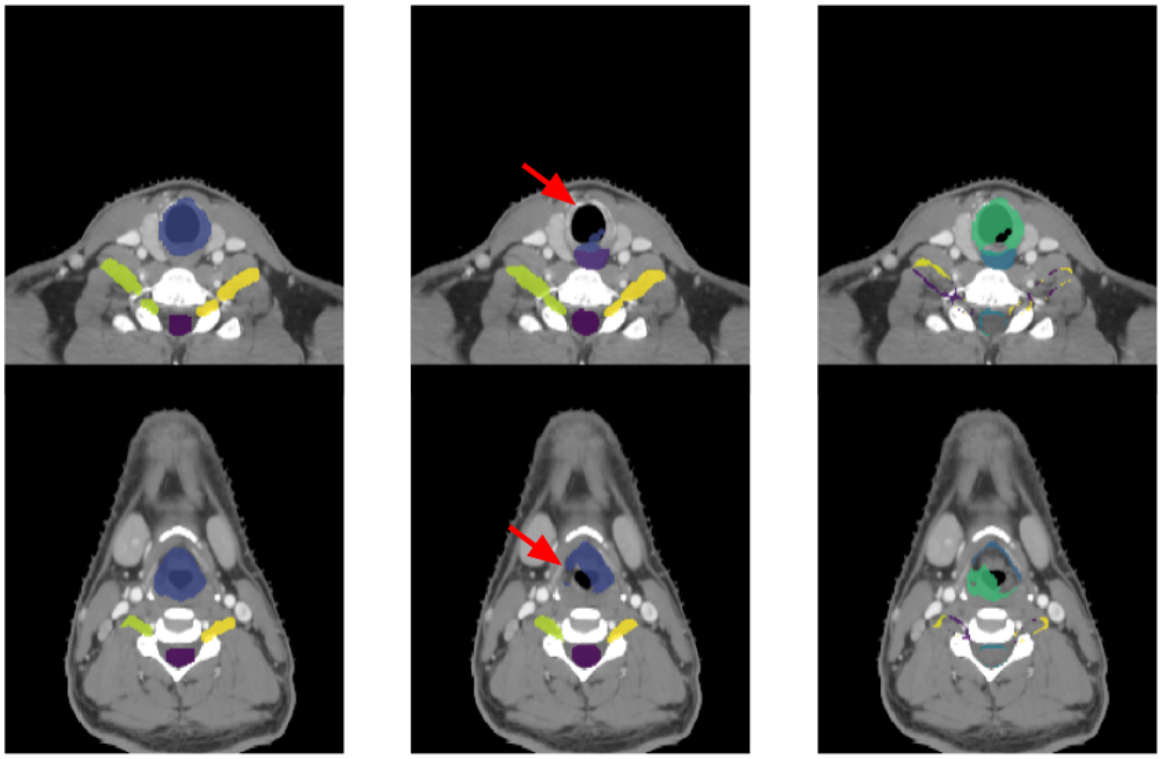








Supplementary Figure 9: Barplots of 3D volumetric Dice (left) and 95%HD (right) for each external dataset to test WOLNET ensemble generalizability.



Supplementary Figure 10: Clinical Acceptability of entire 3D volumes allows us to paint a more accurate picture of deep learning contour fidelity by finding errors present at the edges of complex ROI(s) like the Larynx. Note this was an error found after assessing contours produced by the network after the second training phase, these were the contours uploaded to QUANNOTATE for acceptability testing .

Supplementary Tables

Supplementary Table 1. Public HNC imaging and segmentation datasets curated by our team. Missing data are represented in gray. 'Available Data' types that are only available for a portion of patients are represented by a transparently coloured cell.

Dataset	Institution	Scans	Available Data				References
			Imaging	Clinical	OAR	GTV	
Radcure	UHN	2552					29
HNSCC-3DCT-RT	MIAMI	94					67
Deepmind	HMS+Multi-Site	35					11
PDDCA	HMS	48					45
Radiomics-HN1	MAASTRO	137					43
STRUCTSEG19	CAS	50					44
Head-Neck-CT-Atlas	MDACC	215					68

Supplementary Table 2: Architectures Selected for Comparative Analysis	
Name	**Description**
3D-UNET (WOLNET)	One popular pytorch implementations of 3D UNET paper [22] were taken and used as the baseline architecture. Named WOLNET after it's author. [61,62]
3D-RESUNET	We integrated a third party implementation [61,62] of a residual symmetric 3D UNET proposed by [52]. They introduced a residual skip connection to each valid convolutional module present in their network. To minimize information loss, this network does not down sample feature maps along the z dimension. To minimize the effects of anisotropy 2D convolutions are used in the modules at the lowest part of the network, which contain fine scale feature maps. Each regular residual module will apply in total 7x7x5 of nonlinear convolutions to the input. (To embed 2D features 3x3x1 convolutions are applied followed by subsequent 3x3x3 convolutions.).
HIGHRESNET	Li, W et al (2017) choose to integrate dilated convolutions and residual connections in their proposed 20 layer residual network. These residually connected dilated convolutions allowed for multi-scale feature preservation as training progressed. [53] A third party implementation of HighResNet3D was used for our analysis. [63]
PIPOFAN	We modified the original 2D Pyramid Input Pyramid Output Network proposed by Fang, et al (2020) to accept 3D volumes. This 3D pyramid abstraction network (PIPOFAN) processes the volumetric input by applying 3D Equal Depth Convolutions (EDC), after passing through the network the pyramid outputs are fused together which has been shown to improve subsequent organ segmentations. This network was originally used to segment multiple thoracic OARs on individual slices of a CT scan. [54,64]
UNET3+	We introduced a 3D version of a 2D UNET3+ proposed by Huang, H et al (2020) which was created as a modification of the UNET++ that incorporated full-scale skip-connections into the UNET++ network. This architecture was engineered to produce full-scale aggregated feature maps that allows deep supervision components to learn more comprehensive hierarchical feature maps with the hopes of producing more accurate contours. [55,65]
UNET++	We integrated a 3D version of the 2D Nested Unet Architecture (UNET++) proposed by Zhou et al (2018) using their code as the base network for the architecture used in the study. The authors redesigned sip connection pathways of the original UNET architecture with the intent to reduce the semantic gaps between the encoding and decoding feature maps. This is the first paper to propose and integrate deep supervision into their network where the outputs of each individual segmentation branch are averaged before the final softmax layer of the network. [56,66]

ANATOMY	AnatomyNet was the only segmentation model used in this analysis that was previously published on a HNC OAR segmentation task. This UNET variant incorporates squeeze and excitation residual building blocks in the downsampling/upsampling layers of the network. Code was refactored and updated to suit the newest versions of pytorch. [13,67]
DENSEVOX	A pytorch based third party implementation of DenseVoxNet 3D first proposed bt Yu, L et al 2017, for cardiac segmentation was integrated into our study. This network consists of two DenseBlocks in the downsampling part of the network which are densely connected. In total there are 24 transformation layers before upsampling. A long skip connection was used to stabilize the training process by connecting the transition layer to the output layer. [57,68]
TIRAMISU	We adapted a third party 2D implementation of a fully convolutional Densenet originally presented by Jegou et al, 2017 (named 100-layer tiramisu). This paper was the first to apply the DenseNet to the problem of 2D semantic segmentation. Densenets are constructed by concatenating each output of a subsequent densely connected convolutional block to the next block, therefore linearly augmenting the number of feature maps after each 'down transition'. This does not occur in the upsampling part of the network. The feature maps from the downsampling path are then concatenated with those of the upsampling path to produce a predicted segmentation mask at the resolution of the original input. [58,69]
RSANET	RSANet is a 3D recurrent slice-wise attention network proposed by Zhang, H et al (2019) could be directly integrated into our network. Originally constructed for Multiple Sclerosis lesion segmentation this network utilizes slice wise attention blocks to help capture long-range inter-slice dependencies along any direction of a 3D medical image. These blocks allow for the recurrent aggregation of information along multiple directions therefore providing a mechanism to help capture global contextual information, which can be used to produce more accurate segmentations. [59,70]
VNET	An updated third party implementation of VNET architecture proposed by Millerari et al (2016) was used in this study [60]. The VNET architecture is a fully convolutional network based on the original UNET. The authors choose to replace 3x3x3 convolutions present in UNET by 1 strided 5x5x5 convolutions. Additionally, in place of max-pooling, 2x2x2 convolutions with stride of 2 were used during down sampling. Finally, PReLU nonlinearities were chosen to replace original ReLUs throughout the network. [60,71]

Supplementary Table 3: Performance Metrics for Re-Trained Best Performing Model (WOLNET)

OAR	DICE	95HD	SD	FNV	APL	FNL
BSTEM	0.81±0.05	4.3±5.47	1.63±0.55	0.03±0.04	0.26±0.04	0.02±0.03
SPCOR	0.87±0.04	1.98±0.73	0.8±0.25	0.06±0.04	0.22±0.06	0.06±0.03
ESOPH	0.78±0.08	5.45±6.02	1.3±0.6	0.13±0.11	0.27±0.05	0.1±0.06
LARYNX	0.85±0.07	2.79±1.44	1.05±0.52	0.1±0.11	0.2±0.04	0.08±0.06
MAND	0.92±0.03	2.13±5.01	0.63±0.23	0.03±0.04	0.14±0.04	0.03±0.03
LPAR	0.84±0.06	5.31±5.5	1.52±0.83	0.08±0.06	0.23±0.04	0.07±0.05
RPAR	0.84±0.12	5.04±5.93	1.51±1.95	0.11±0.08	0.22±0.04	0.09±0.07
LACOU	0.74±0.17	3.69±9.92	1.98±8.53	0.1±0.15	0.31±0.13	0.09±0.11
RACOU	0.73±0.16	3.83±9.94	2.05±8.73	0.09±0.16	0.32±0.12	0.08±0.12
RPLEX	0.7±0.13	4.79±7.58	1.56±2.34	0.16±0.08	0.4±0.08	0.15±0.07
LPLEX	0.7±0.12	5.16±7.12	1.57±2.07	0.14±0.08	0.39±0.07	0.14±0.07
LLENS	0.77±0.07	1.440.44	0.57±0.18	0.08±0.11	0.21±0.1	0.08±0.11
RLENS	0.78±0.1	1.44±0.6	0.5±0.18	0.11±0.15	0.21±0.13	0.11±0.15
LEYE	0.88±0.04	1.83±0.92	0.76±0.28	0.04±0.03	0.18±0.05	0.04±0.03
REYE	0.87±0.04	1.79±0.54	0.76±0.2	0.03±0.02	0.18±0.05	0.03±0.02
LOPTIC	0.72±0.1	1.96±1.28	0.65±0.34	0.12±0.1	0.24±0.09	0.12±0.1
ROPTIC	0.7±0.1	2.34±2.56	0.73±0.48	0.11±0.11	0.23±0.09	0.11±0.11
CHIASM	0.41±0.18	6.59±4.46	1.73±1.71	0.31±0.28	0.32±0.26	0.3±0.27
LIPS	0.74±0.08	3.06±1.53	1.03±0.47	0.11±0.08	0.4±0.08	0.11±0.07
ALL	**0.77±0.09**	**3.42±4.05**	**1.18±1.6**	**0.1±0.1**	**0.26±0.08**	**0.09±0.08**

Supplementary Table 4: MC v. DLC Acceptability Ratings For Each OAR Category

ROI	MAR (MC)	MAR (DLC)	ROI	MAR (MC)	MAR (DLC)
MAND	**4.15±0.89**	3.53±1.13			
LEYE	**3.45±0.71**	3.00±0.78	RLENS	4.10±0.78	**4.15±0.70**
REYE	**3.68±0.62**	3.05±0.68	LACOU	**3.30±0.91**	3.20±0.97
BSTEM	**4.13±0.69**	2.8±0.91	RACOU	3.30±0.72	**3.38±0.74**
LARYNX	**3.95±0.85**	2.38±0.90	LPLEX	3.68±0.76	**3.78±0.58**
SPCOR	**3.93±0.76**	3.70±0.56	RPLEX	**3.63±0.84***	3.58±0.81
LPAR	3.80±0.99	**3.85±1.02**	LIPS	**3.75±0.71**	3.33±0.80
RPAR	**4.08±1.02***	3.95±1.23	LOPTIC	**4.28±0.68**	3.50±0.78
ESOPH	**3.93±0.80**	3.13±0.97	ROPTIC	**4.18±0.68**	3.20±0.79
LLENS	**4.05±0.68**	3.65±0.92	CHIASM	**3.03±1.27***	2.90±1.06

Supplementary Table 5: Distribution of Ground Truth OAR labels (masks) in external datasets

OAR	Internal Test - RAD (n=59)	Dataset 1 - Rad (n=83)	Dataset 2 - Onc (n=35)	Dataset 2 - Rad (n=35)	Dataset 3 - Rad (n=48)	Dataset 4 - Rad (n=119)	Dataset 5 - Rad (n=50)
BSTEM	59	82	33	33	48		50
CHIASM	59	61			48		50
ESOPH	59	12					
LACOU	59						50
LARYNX	59	47					
LEYE	59	46	33	33			50
LIPS	59	5					
LLENS	59	9	33	33			50
LOPTIC	59	27	33	33	48		50
LPAR	59	71	33	33	48	94	50
LPLEX	59						
MAND	59	60	33	33	48		50
RACOU	59						50
REYE	59	49	33	33			50
RLENS	59		33	33			50
ROPTIC	59	28	33	33	48		50
RPAR	59	71	33	33	48	93	50
RPLEX	59						
SPCOR	59	83	33	33		119	50

Supplementary Table 6: 3D Volumetric DICE Performance of Optimized WOLNET Ensemble on External Data

*** two contours (L/R) mandible were used as single GT mask**

OAR	Internal - Test (n=59)	Dataset 1 - Rad (n=83)	Dataset 2 - Onc (n=35)	Dataset 2 - Rad (n=35)	Dataset 3 - Rad (n=48)	Dataset 4 - Rad (n=119)	Dataset 5 - Rad (n=50)
BSTEM	0.84±0.05	0.8±0.05	0.77±0.05	0.8±0.04	0.74±0.09		0.78±0.04
CHIASM	0.4±0.18	0.27±0.21			0.29±0.09		0.33±0.17
ESOPH	0.83±0.07	0.53±0.21					
LACOU	0.79±0.17						0.59±0.1
LARYNX	0.88±0.06	0.66±0.14					
LEYE	0.91±0.12	0.83±0.05	0.8±0.03	0.83±0.03			0.84±0.04
LIPS	0.76±0.13	0.45±0.05					
LLENS	0.79±0.12	0.55±0.16	0.57±0.07	0.62±0.1			0.68±0.09
LOPTIC	0.78±0.09	0.54±0.12	0.55±0.05	0.55±0.06	0.5±0.06		0.53±0.1
LPAR	0.85±0.06	0.78±0.1	0.84±0.04	0.84±0.04	0.85±0.06	0.78±0.09	0.86±0.04
LPLEX	0.72±0.13						
MAND	0.93±0.03	0.82±0.06	0.89±0.03	0.92±0.02	0.82±0.03		0.86±0.04*
RACOU	0.8±0.17						0.63±0.06
REYE	0.9±0.11	0.83±0.05	0.82±0.03	0.84±0.03			0.85±0.03
RLENS	0.82±0.12		0.58±0.09	0.62±0.09			0.68±0.1
ROPTIC	0.76±0.09	0.55±0.16	0.55±0.07	0.58±0.06	0.47±0.07		0.51±0.11
RPAR	0.86±0.12	0.81±0.09	0.85±0.04	0.86±0.04	0.84±0.06	0.78±0.07	0.86±0.04
RPLEX	0.72±0.14						
SPCOR	0.86±0.05	0.77±0.09	0.66±0.07	0.7±0.04		0.73±0.09	0.68±0.07

Supplementary Table 7: 95%HD Performance of Optimized WOLNET Ensemble on External Data

*** two contours (L/R) mandible were used as single GT mask**

OAR	Internal - RAD (n=59)	Dataset 1 - Rad (n=83)	Dataset 2 - Onc (n=35)	Dataset 2 - Rad (n=35)	Dataset 3 - Rad (n=48)	Dataset 4 - Rad (n=119)	Dataset 5 - Rad (n=50)
BSTEM	3.0±0.88	4.12±1.77	4.12±0.73	3.74±0.68	4.0±2.16		3.74±0.85
CHIASM	5.1±3.94	7.26±3.75			6.4±2.62		5.14±3.53
ESOPH	3.16±5.48	14.21±9.13					
LACOU	1.57±9.79						3.74±1.14
LARYNX	2.24±1.38	6.0±2.71					
LEYE	1.41±2.98	2.83±0.87	2.24±0.16	2.24±0.15			1.41±0.39
LIPS	2.45±1.21	8.6±2.59					
LLENS	1.41±0.33	2.13±0.92	2.0±0.31	1.65±0.55			1.41±0.37
LOPTIC	1.41±1.29	6.48±5.28	2.24±1.53	2.24±1.88	2.63±4.44		5.05±2.19
LPAR	3.32±3.44	6.4±4.41	3.32±2.77	3.74±5.17	3.16±1.72	4.9±3.52	3.0±1.73
LPLEX	4.0±13.93						
MAND	1.0±12.19	4.29±4.47	1.41±0.32	1.0±0.42	2.0±23.97		4.3±14.11*
RACOU	1.73±9.95						3.73±0.97
REYE	1.41±2.97	2.45±0.64	2.24±0.14	2.0±0.13			1.41±0.3
RLENS	1.41±0.45		1.95±0.31	1.41±0.44			1.41±0.41
ROPTIC	1.41±1.6	6.04±5.14	2.24±0.42	2.0±1.34	2.24±1.65		5.03±2.68
RPAR	3.0±4.29	5.74±4.86	3.16±2.36	3.16±2.73	3.0±1.66	5.39±4.75	3.0±3.04
RPLEX	4.0±13.8						
SPCOR	2.0±1.3	3.0±4.35	3.16±1.27	3.0±1.41		3.0±12.3	2.83±0.83