# Title Page

**Title:** Development of a High-Performance Multiparametric MRI Oropharyngeal Primary Tumor Auto-Segmentation Deep Learning Model and Investigation of Input Channel Effects: Results from a Prospective Imaging Registry

**Author names and affiliations:** Kareem A. Wahid[1], Sara Ahmed[1], Renjie He[1], Lisanne V. van Dijk[1], Jonas Teuwen[2], Brigid A. McDonald[1], Vivian Salama[1], Abdallah S.R. Mohamed[1], Travis Salzillo[1], Cem Dede[1], Nicolette Taku[1], Stephen Y. Lai[3], Clifton D. Fuller[1]*, Mohamed A. Naser[1]*

[1]Department of Radiation Oncology, University of Texas MD Anderson Cancer Center, Houston, TX
[2]Department of Medical Imaging, Radboud University Medical Centre, Nijmegen, The Netherlands
[3]Department of Head and Neck Surgery, University of Texas MD Anderson Cancer Center, Houston, TX
* co-corresponding authors.

**Contact information:** Kareem A. Wahid, kawahid@mdanderson.org.

**Present/permanent address:** MD Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77030.

**Funding Statement:** This work was supported by the National Institutes of Health (NIH) through a Cancer Center Support Grant (P30-CA016672-44). K.A. Wahid and T. Salzillo are supported by training fellowships from The University of Texas Health Science Center at Houston Center for Clinical and Translational Sciences TL1 Program (TL1TR003169). S. Ahmed and M.A. Naser are supported by an NIH National Institute of Dental and Craniofacial Research (NIDCR) Award (R01 DE028290-01). R. He, A.S.R. Mohamed, and S.Y. Lai are supported by a NIH NIDCR Award (R01 DE025248). L.V. van Dijk receives funding and salary support from the Dutch organization NWO ZonMw during the period of study execution via the Rubicon Individual career development grant. B.A. McDonald receives research support from an

NIH NIDCR Award (F31DE029093) and the Dr. John J. Kopchick Fellowship through The University of Texas MD Anderson UTHealth Graduate School of Biomedical Sciences. C.D. Fuller received funding from an NIH NIDCR Award (1R01 DE025248-01/R56 DE025248) and Academic-Industrial Partnership Award (R01 DE028290); the National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679); the NIH Big Data to Knowledge (BD2K) Program of the National Cancer Institute (NCI) Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01 CA214825); the NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01 CA218148); the NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30 CA016672); the NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50 CA097007); and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Program (R25 EB025787). He has received direct industry grant support, speaking honoraria, and travel funding from Elekta AB.

# Main Text

## Abstract

*Background and Purpose:* Oropharyngeal cancer (OPC) primary gross tumor volume (GTVp) segmentation is crucial for radiotherapy. Multiparametric MRI (mpMRI) is increasingly used for OPC adaptive radiotherapy but relies on manual segmentation. Therefore, we constructed mpMRI deep learning (DL) OPC GTVp auto-segmentation models and determined the impact of input channels on segmentation performance.

*Materials and Methods:* GTVp ground truth segmentations were manually generated for 30 OPC patients from a clinical trial. We evaluated five mpMRI input channels (T2, T1, ADC, Ktrans, Ve). 3D Residual U-net models were developed and assessed using leave-one-out cross-validation. A baseline T2 model was compared to mpMRI models (T2+T1, T2+ADC, T2+Ktrans, T2+Ve, all 5 channels [ALL]) primarily using the Dice similarity coefficient (DSC). Sensitivity, positive predictive value, Hausdorff distance (HD), false-negative DSC (FND), false-positive DSC, surface DSC, 95% HD, and mean surface distance were also assessed. For the best model, ground truth and DL-generated segmentations were compared through a Turing test using physician observers.

*Results:* Models yielded mean DSCs from 0.71 (ALL) to 0.73 (T2+T1). Compared to the T2 model, performance was significantly improved for HD, FND, sensitivity, surface DSC, and 95% HD for the T2+T1 model ($p<0.05$) and for FND for the T2+Ve and ALL models ($p<0.05$). There were no differences between ground truth and DL-generated segmentations for all observers ($p>0.05$).

*Conclusion:* DL using mpMRI provides high-quality segmentations of OPC GTVp. Incorporating additional mpMRI channels may increase the performance of certain evaluation metrics. This pilot study is a promising step towards fully automated MR-guided OPC radiotherapy.

## 1. Introduction

Oropharyngeal cancer (OPC), a type of head and neck squamous cell carcinoma (HNSCC), is among the most common malignancies globally [1]. Treatment for OPC often includes radiotherapy because of its high cure rate [2]. Segmentation (also termed contouring) of the primary gross tumor volume (GTVp) on radiologic imaging is necessary for the OPC radiotherapy workflow. The GTVp, with a clinical and planning safety margin, acts as a target volume to deliver the radiotherapy dose. Therefore, inadequate GTVp definition may cause under-dosage of the tumor or over-dosage of surrounding normal tissues [3,4]. However, the current clinical standard is manual segmentation by physician experts, which is labor-intensive and subject to high inter-observer variation [5–7]. Therefore, an auto-segmentation tool would be a promising alternative to the current manual standard in OPC radiotherapy workflows.

Deep learning (DL) has found wide success in auto-segmentation [8,9], with many HNSCC auto-segmentation studies applying DL to CT imaging [10–12]. Although CT is the most commonly used imaging modality in OPC radiotherapy planning, MRI has been increasingly recognized as essential for tumor segmentation because of its exceptional soft-tissue contrast [13,14]. Additionally, the emergence of MR-Linac technology, an image-guided adaptive radiotherapy approach, has further incentivized the incorporation of MRI in OPC radiotherapy planning. Importantly, we recently demonstrated the utility of DL for HNSCC organ-at-risk auto-segmentation using MRI, with improvements in performance, execution time, and dosimetric differences compared to other auto-segmentation methods [15]. While several DL tumor auto-segmentation studies for nasopharyngeal cancer using MRI have been published [16–25], to our knowledge, only one study has been published for OPC [26]. Since HNSCC tumors at different anatomical sites have distinct anatomic boundaries and characteristics [27,28], it is

crucial that tumor segmentation models are developed for each site accordingly. Consequently, there exists an unmet need for OPC DL tumor segmentation tools using MRI.

Multiparametric MRI (mpMRI) incorporates multiple sequence acquisitions that highlight anatomical and functional information in tumors. For example, dynamic contrast-enhanced (DCE) MRI and diffusion-weighted imaging (DWI) can quantify tumor perfusion and diffusion patterns, respectively, and may affect OPC treatment guidance [29,30]. Recent studies of PET/CT OPC DL auto-segmentation [11,20,31–35] have demonstrated increased segmentation performance when combining functional and anatomical modalities. However, investigations that combine anatomical with functional MRI in HNSCC to achieve acceptable DL auto-segmentation performance are lacking [36,37].

In this pilot study, we evaluated the effects of anatomical and functional mpMRI inputs on OPC GTVp segmentation performance. Using open-source DL frameworks with standardized clinical trial data, we trained and evaluated DL models based on variable mpMRI input channels. We then compared the models qualitatively and quantitatively to determine which channel combinations led to the best segmentation results. Finally, we characterized the clinical acceptability of the best-performing model using physician experts.

## 2. Methods

*2.1. Imaging Data:* We acquired pre-radiotherapy T2-weighted (T2), contrast-enhanced T1-weighted Dixon fat-suppressed (T1), DCE, and DWI MRI sequences in Digital Imaging and Communications in Medicine (DICOM) format for 124 HNSCC patients from a prospective clinical trial investigating longitudinal mpMRI (NCT03145077). Images were collected from August 2018-August 2019 under a HIPAA-compliant protocol approved by The University of

Texas MD Anderson Cancer Center's IRB (RCR03-0800). The protocol included a waiver of informed consent. We curated 30 OPC patients with a visible GTVp based on the complete availability of T2, T1, DCE, and DWI image sets (**Fig. S1**). Demographic characteristics of the patients are shown in **Table S1**. Imaging was performed on a Siemens Aera scanner with a magnetic field strength of 1.5 T and standardized acquisition parameters (**Table S2**). All patients were immobilized with a thermoplastic mask. Apparent diffusion coefficient (ADC) parametric maps were derived from DWI sequences through a proprietary Siemens algorithm (Munich, Germany) using a monoexponential model. The Tofts model was used to generate parametric maps from DCE sequences for the volume transfer constant (Ktrans) and the extravascular extracellular volume fraction (Ve) [38]. Additional details regarding DCE parametric map generation can be found in our previous publication [39]. GTVp structures were manually segmented in the DICOM-RT Structure format by a physician (radiologist with >5 years of expertise in HNSCC) in Velocity AI v.3.0.1 (Atlanta, GA, USA). GTVp structures were segmented on the T2 MRI, but the physician could consult the other images. An example of the mpMRI images used in this study and overlying GTVp segmentation for one patient is shown in **Figure 1A**.

*2.2. Image Processing:* To ensure adequate MRI comparability between patients [40], we performed intensity standardization for all images. Anatomical sequences (T2, T1) were standardized using a Z-score (mean=0, standard deviation=1), while functional parametric maps (ADC, Ktrans, Ve) were truncated to the 10th and 90th percentile for all patients and rescaled to [-1, 1] as per a previous study [36]. All images were cropped to the smallest field of view (Ktrans, Ve) and resampled to the T2 resolution. An example of the image processing workflow is shown in **Figure 1B**.

*2.3. Segmentation Model Architecture and Implementation:* A DL convolutional neural network based on the 3D Residual U-net architecture [41,42] was implemented in the Medical Open Network for Artificial Intelligence (MONAI) software package [43] (**Fig. 1C**). The GTVp mask was used as the ground truth target to train the segmentation model. The MRI images acted as variable-channel inputs to the models. We investigated the following channel combinations as separate models: T2, T2+T1, T2+ADC, T2+Ktrans, T2+Ve, and all five input channels (ALL). The T2 model acted as a baseline of comparison for all other models. We implemented an Adam optimizer with a Sørensen-Dice similarity coefficient (DSC) loss function. The models were trained for 700 iterations with a learning rate of $2 \times 10^{-4}$ for the first 550 iterations and $1 \times 10^{-4}$ for the remaining 150 iterations. Data augmentation was used to mitigate overfitting. Additional details on the DL architecture and implementation are found in **Supplementary Methods**.

*2.4. Model Evaluation:* Model performance was primarily assessed using DSC. We also implemented additional spatial similarity metrics, including Hausdorff distance (HD), false-negative DSC (FND), false-positive DSC (FPD), sensitivity, positive predictive value (PPV), surface DSC, 95% HD, and mean surface distance (MSD). For surface DSC, a tolerance of 3.0 mm was selected as suitable from previous inter-observer variability studies on T2 MRI of OPC GTVp [44]. Surface distance metrics were calculated using the surface-distance Python package [45], while all other metrics were calculated in Elekta ADMIRE v.2.9 (Stockholm, Sweden). Each model was trained and evaluated using leave-one-out cross-validation (LOOCV) (**Fig. 1D**).

*2.5. Clinical Evaluation*: For our best-performing model, we assigned three physician expert observers (radiologist from *2.1* >1-year post-segmenting, two radiation oncologists) to evaluate the ground truth and corresponding DL-generated segmentations using subjective scoring

criteria based on a 4-point Likert scale. The score categories were: 1 = requires corrections, large errors; 2 = requires corrections, minor errors; 3 = clinically acceptable, errors not clinically significant; 4 = clinically acceptable, highly accurate. Additionally, we asked observers to predict the source of the segmentations as either human (ground truth) or DL-generated through a modified Turing test [46]. Ground truth and DL-generated segmentations for all 30 patients were anonymized and randomly presented to experts for clinical evaluation. Experts were blinded to the segmentation source.

*2.6. Statistical Analysis:* After performing a Shapiro-Wilk test, we found that our data were not normally distributed ($p<0.05$); therefore, we utilized nonparametric statistical tests. We used one-sided Wilcoxon signed-rank tests (alternative hypothesis of greater than for DSC, sensitivity, surface DSC, and PPV; alternative hypothesis of less than for HD, FND, FPD, 95% HD, and MSD) to evaluate differences between our baseline T2 model and models with additional channels. We used Mann-Whitney U tests to detect differences in model performance based on tumor subsite (base of tongue vs. tonsil). Additionally, to assess correlations of tumor size with model performance, we calculated Pearson correlation coefficients with corresponding p-values of ground truth volume against DSC, HD, and surface DSC for every model. Finally, to assess the clinical evaluation of ground truth against DL-generated segmentations, for each observer we implemented a two-sided Wilcoxon signed-rank test for scores and a McNemar test for source predictions. For all statistical analyses, p-values less than 0.05 were considered significant. Analyses were performed in Python v.3.7.9. Code notebooks can be found at GitHub (https://github.com/kwahid/mpMRI_OPC_GTVp_segmentation).

## 3. Results

*3.1. Model Performance:* **Figure 2** shows boxplots of model performance for all tested input channel combinations with respect to different evaluation metrics. T2+T1 was the best performing model overall with the best mean scores in DSC, HD, sensitivity, surface DSC, and 95% HD. ALL was the worst performing model overall with the worst mean scores in in DSC, FPD, and PPV. T2 performed best in FPD and PPV, but worst in sensitivity and 95% HD. T2+Ve performed best in FND and MSD but worst in HD. T2+Ktrans performed worst in MSD. T2+T1 had significantly better performance ($p<0.05$) than the baseline T2 model for HD, FND, sensitivity, surface DSC, and 95% HD. T2+Ve and ALL had significantly better performance ($p<0.05$) than the baseline T2 model for FND. **Figure S2** shows a heatmap of p-values comparing channel combinations to the baseline T2 model. A subgroup analysis revealed no significant differences in model performance for any combination of models and metrics based on OPC subsite, as all p-values were > 0.05 (**Table S3**).

**Figure 3** shows examples of model segmentations compared to ground truth segmentations for high-, medium-, and low-performance cases, based on DSC scores across all models. For the high-performance case, the T2 model demonstrates a DSC of 0.88, with the incorporation of additional channels leading to DSC scores of 0.87-0.90. For the medium-performance case, the T2 model demonstrates a DSC of 0.71, with the incorporation of additional channels leading to DSC scores of 0.72-0.78. For the low-performance case, the T2 model demonstrated a DSC of 0.37 and many spuriously predicted voxels in the posterior region of the head, with the incorporation of additional channels reducing the number of spurious voxels and leading to DSC scores of 0.52-0.61.

*3.2. Size Dependence of Models:* **Figure 4** shows correlation graphs of the various models comparing tumor size to DSC, HD, and surface DSC. The range of values for tumor size were 1.74-45.19cc. Every model showed non-significant positive correlations for DSC ($p>0.05$) and

significant positive correlations for HD ($p<0.005$), except for T2+Ve ($r=0.33$, $p=0.079$) and ALL ($r=0.06$, $p=0.76$). Every model also showed significant negative correlations for surface DSC ($p<0.05$), except for T2+ADC ($r=-0.34$, $p=0.07$), T2+Ve ($r=-0.30$, $p=0.11$), and ALL ($r=-0.30$, $p=0.1$).

*3.3. Clinical Evaluation:* **Table 1** shows the categorical scores and predicted sources for ground truth and DL-generated (T2+T1 model) segmentations for each observer. The mean scores for ground truth vs. DL-generated segmentations were 3.0 vs. 2.5, 2.5 vs. 2.7, and 3.0 vs. 3.0 for observers 1, 2, and 3, respectively. Significance testing revealed no observer could differentiate between the scores ($p>0.05$) or source ($p>0.05$) of the ground truth segmentations compared to the DL-generated segmentations.

## 4. Discussion

In this pilot study, we determined the impact of mpMRI input channel combinations (T2, T2+T1, T2+ADC, T2+Ktrans, T2+Ve, ALL) on DL model segmentation performance. Recent work has suggested that the average agreement between physicians measured in DSC for OPC tumor segmentation is exceptionally low [44]. Notably, compared to previous fully-automated primary tumor segmentation studies of HNSCC patients, we achieved promising average DSC performance (**Table 2**). While it is difficult to directly compare DSCs between studies due to different datasets and model training, our models seemingly improve upon the only other fully-automated OPC tumor segmentation study to our knowledge (DSC=0.55), which exclusively investigated anatomical MRI [26].

The best average DSC performance was achieved by the T2+T1 model (DSC=0.73), which was higher than the baseline T2 model (DSC=0.72) but not statistically significant. Moreover,

average DSC decreased when combining all input channels (DSC=0.71), though non-significantly. However, a previous similar study by Bielak et al. investigating HNSCC tumors with segmentations derived from T2 MRI demonstrated an increased DSC after the inclusion of all available mpMRI channels [36], which is in direct opposition to our results. Importantly, the authors used a smaller number of patients (n=18) than our study and implemented repeat imaging at different time-points, which could confound their results. Additionally, their results may be more relevant for a specific HNSCC tumor site, but no analysis was performed to verify this. Furthermore, it should be noted that the average DSC for their best model was ~0.30, which was substantially lower than all our models. Notably, auto-segmentation studies in prostate cancer have also reported conflicting results on the additive effects of additional mpMRI input channels for DSC when using ground truth annotations derived from T2 MRI [47–49]. Therefore, further investigations are likely needed to verify if a significant positive DSC effect exists for mpMRI input channel combinations in OPC tumor auto-segmentation.

While most auto-segmentation studies have focused on DSC as an evaluation metric, it has been argued that other metrics should also be taken into consideration, depending on the use-case of the auto-segmentation tool [50,51]. Therefore, to increase the robustness of our analysis, we have included complimentary metrics (HD, FND, FPD, sensitivity, PPV, surface DSC, 95% HD, and MSD) to evaluate our models. Like DSC, most metrics show high performance across various models, with some models demonstrating significantly better values than the baseline T2 model. Interestingly, we demonstrated that in certain edge cases (low-performance example), the inclusion of additional channels could circumvent spurious voxel predictions derived from the baseline T2 model (a possible byproduct of model overfitting), which may increase model robustness. These results indicate that the additional channels may contain underlying additive information to improve performance for aspects other than traditional DSC-based evaluation. Notably, the specific anatomic subsite of the tumor (base of tongue or

tonsil) had no significant effect on performance for any models for any evaluation metric, indicating that the models were robust to the spatial location of the OPC.

Previous studies [16,36] have suggested small tumors may be more difficult for DL models to segment, which would hinder the incorporation of models into radiotherapy workflows. Importantly, there were no significant correlations between tumor size and DSC for any of our models. However, it should be noted that surface distance metrics, such as the HD and surface DSC, demonstrate some size dependence, with larger and smaller tumors being easier for our models to segment, respectively. Interestingly, the surface distance metrics do not demonstrate a significant size dependence for some models that utilize additional channels, particularly those that correspond to functional parametric maps. Therefore, the inclusion of additional channels may strengthen the robustness of models to tumor size for surface distance metric performance, but further confirmatory work is needed.

The acceptability of segmentations used in a radiotherapy workflow is ultimately determined by physician judgment, with physician rating scales considered the gold standard for clinically relevant segmentation quality [51]. While subjective evaluation through rating scales is common in auto-segmentation studies, the established variability of OPC tumor segmentation between observers [44] highlights the difficulty in the interpretation of multi-observer segmentation quality analysis. Therefore, we implemented a comparative approach for each observer to determine if significant clinical differences were present between the ground truth segmentations and the corresponding segmentations of the best DL model (T2+T1). We demonstrated that experts were unable to determine differences between the ground truth and the DL-generated segmentations or identify the source of the segmentations. Therefore, our model “passed” the Turing test, which highlights its potential clinical utility. Of note, the radiologist who provided the original ground truth segmentations was the closest among the observers to correctly

discriminating the segmentation sources but was still unable to achieve statistical significance. Moreover, for the radiation oncologist observers the mean clinical acceptability score of the DL-generated segmentations was equal to or higher than the ground truth segmentations, which may indicate a slight preference towards DL-generated OPC tumor segmentations for radiotherapy end users.

One limitation of our study is the use of a small cohort with standardized acquisition parameters. However, we have taken steps to optimally utilize our data by implementing a LOOCV approach and investigating various evaluation metrics. Moreover, we plan to include additional prospectively acquired data for model training and use external heterogenous validation sets in future studies to increase model generalizability. Another limitation of our study is that we have constrained our analysis of input image channels based on those that were investigated in previous literature [36]. However, mpMRI input channels can be further investigated through additional quantitative parametric maps (e.g., extended Tofts model [52], advanced DWI fitting models [53], etc.). Therefore, we plan to include additional input channels in future analyses. A final limitation of our study is the lack of overt image registration. Our images were acquired from a standardized clinical trial with patient immobilization; therefore, implicit co-registration was deemed adequate for tumor overlap. However, small amounts of motion artifacts may cause the segmentation mask to overlap improperly on mpMRI image channels, impacting auto-segmentation quality. Furthermore, though no geometric distortion was observed on any parametric maps, distortions were not explicitly quantified. Future studies should investigate the role of additional OPC-specific registration algorithms and geometric distortion correction in combination with mpMRI DL auto-segmentation algorithms.

## 5. Conclusions

In summary, using mpMRI inputs, we built OPC primary tumor DL auto-segmentation models that demonstrated excellent performance across multiple evaluation metrics, with average DSC scores as high as 0.73. Compared to our baseline model trained on T2 MRI only, we find that adding T1 MRI significantly improved HD, FND, sensitivity, surface DSC, and 95% HD. Moreover, adding Ve or using all input channels simultaneously significantly improved FND. Additionally, certain favorable aspects of model construction, including decreased spurious voxel predictions and robustness to tumor size when considering surface distance metric performance, are apparent for models that leverage additional input channels. Finally, physician experts could not differentiate ground truth from DL-generated segmentations, demonstrating our model “passed” the Turing test. These promising results should be further verified in large independent datasets. Overall, our pilot study is an important step towards fully automated MR-guided OPC radiotherapy workflows.

**Acknowledgements:** We thank Ms. Ann Sutton from the Editing Services Group at The University of Texas MD Anderson Cancer Center Research Medical Library for editing this article.

## References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.
[2] De Felice F, Tombolini V, Valentini V, de Vincentiis M, Mezi S, Brugnoletti O, et al. Advances in the Management of HPV-Related Oropharyngeal Cancer. J Oncol 2019;2019:9173729.
[3] Njeh CF. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. J Med Phys 2008;33:136–40.
[4] Njeh CF, Dong L, Orton CG. Point/Counterpoint. IGRT has limited clinical value due to lack of accurate tumor delineation. Med Phys 2013;40:040601.
[5] Vorwerk H, Zink K, Schiller R, Budach V, Böhmer D, Kampfer S, et al. Protection of quality and innovation in radiation oncology: The prospective multicenter trial the German Society of Radiation Oncology (DEGRO-QUIRO study). Strahlentherapie Und Onkologie 2014;190:433–43. https://doi.org/10.1007/s00066-014-0634-0.
[6] Segedin B, Petric P. Uncertainties in target volume delineation in radiotherapy--are they relevant and what can we do about them? Radiol Oncol 2016;50:254–62.
[7] Rasch C, Steenbakkers R, van Herk M. Target definition in prostate, head, and neck. Semin Radiat Oncol 2005;15:136–45.
[8] Guo Y, Liu Y, Georgiou T, Lew MS. A review of semantic segmentation using deep neural networks. International Journal of Multimedia Information Retrieval 2018;7:87–93.
[9] Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A Review on Deep Learning Techniques Applied to Semantic Segmentation. ArXiv [CsCV] 2017.
[10] Maleki F, Le WT, Sananmuang T, Kadoury S, Forghani R. Machine Learning Applications for Head and Neck Imaging. Neuroimaging Clin N Am 2020;30:517–29.
[11] Lo Faso EA, Gambino O, Pirrone R. Head–Neck Cancer Delineation. NATO Adv Sci Inst Ser E Appl Sci 2021;11:2721.
[12] Kosmin M, Ledsam J, Romera-Paredes B, Mendes R, Moinuddin S, de Souza D, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiother Oncol 2019;135:130–40.
[13] Zima AJ, Wesolowski JR, Ibrahim M, Lassig AAD, Lassig J, Mukherji SK. Magnetic resonance imaging of oropharyngeal cancer. Top Magn Reson Imaging 2007;18:237–42.
[14] Lewis-Jones H, Colley S, Gibson D. Imaging in head and neck cancer: United Kingdom National Multidisciplinary Guidelines. J Laryngol Otol 2016;130:S28–31.
[15] McDonald B, Cardenas C, O'Connell N, Ahmed S, Naser M, Xu J, et al. Autosegmentation On Low-Resolution T2-Weighted MRI of Head and Neck Cancers for Off-Line Dose Reconstruction in MR-Linac Adapt-To-Position Workflow, 2021.
[16] Lin L, Dou Q, Jin Y-M, Zhou G-Q, Tang Y-Q, Chen W-L, et al. Deep Learning for Automated Contouring of Primary Tumor Volumes by MRI for Nasopharyngeal Carcinoma. Radiology 2019;291:677–86.
[17] Ye Y, Cai Z, Huang B, He Y, Zeng P, Zou G, et al. Fully-Automated Segmentation of Nasopharyngeal Carcinoma on Dual-Sequence MRI Using Convolutional Neural Networks. Front Oncol 2020;10:166.
[18] Ma Z, Wu X, Song Q, Luo Y, Wang Y, Zhou J. Automated nasopharyngeal carcinoma segmentation in magnetic resonance images by combination of convolutional neural networks and graph cut. Exp Ther Med 2018;16:2511–21.
[19] Li Q, Xu Y, Chen Z, Liu D, Feng S-T, Law M, et al. Tumor Segmentation in Contrast-Enhanced Magnetic Resonance Imaging for Nasopharyngeal Carcinoma: Deep Learning with Convolutional Neural Network. Biomed Res Int 2018;2018:9128527.

[20] Chen H, Qi Y, Yin Y, Li T, Liu X, Li X, et al. MMFNet: A multi-modality MRI fusion network for segmentation of nasopharyngeal carcinoma. Neurocomputing 2020;394:27–40.
[21] He Y, Yu X, Liu C, Zhang J, Hu K, Zhu HC. A 3D Dual Path U-Net of Cancer Segmentation Based on MRI. 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), 2018, p. 268–72.
[22] Ke L, Deng Y, Xia W, Qiang M, Chen X, Liu K, et al. Development of a self-constrained 3D DenseNet model in automatic detection and segmentation of nasopharyngeal carcinoma using magnetic resonance images. Oral Oncology 2020;110:104862. https://doi.org/10.1016/j.oraloncology.2020.104862.
[23] Wang Y, Zu C, Hu G, Luo Y, Ma Z, He K, et al. Automatic tumor segmentation with deep convolutional neural networks for radiotherapy applications. Neural Process Letters 2018;48:1323–34.
[24] Ma Z, Zhou S, Wu X, Zhang H, Yan W, Sun S, et al. Nasopharyngeal carcinoma segmentation based on enhanced convolutional neural networks using multi-modal metric learning. Phys Med Biol 2019;64:025005.
[25] Huang J-B, Zhuo E, Li H, Liu L, Cai H, Ou Y. Achieving Accurate Segmentation of Nasopharyngeal Carcinoma in MR Images Through Recurrent Attention. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing; 2019, p. 494–502.
[26] Rodríguez Outeiral R, Bos P, Al-Mamgani A, Jasperse B, Simões R, van der Heide UA. Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning. Phys Imaging Radiat Oncol 2021;19:39–44.
[27] Shiga K, Ogawa T, Katagiri K, Yoshida F, Tateda M, Matsuura K, et al. Differences between oral cancer and cancers of the pharynx and larynx on a molecular level. Oncol Lett 2012;3:238–43.
[28] Rothenberg SM, Ellisen LW. The molecular pathogenesis of head and neck squamous cell carcinoma. J Clin Invest 2012;122:1951–7.
[29] van der Heide UA, Houweling AC, Groenendaal G, Beets-Tan RGH, Lambin P. Functional MRI for radiotherapy dose painting. Magn Reson Imaging 2012;30:1216–23.
[30] Salzillo T, Taku N, Wahid K, McDonald B, Wang J, van Dijk L, et al. Advances in Imaging for HPV-Related Oropharyngeal Cancer: Applications to Radiation Oncology. Semin Radiat Oncol 2021.
[31] Andrearczyk V, Oreiller V, Vallières M, Castelli J, Elhalawani H, Jreige M, et al. Automatic Segmentation of Head and Neck Tumors and Nodal Metastases in PET-CT scans. In: Arbel T, Ayed IB, de Bruijne M, Descoteaux M, Lombaert H, Pal C, editors. Proceedings of the Third Conference on Medical Imaging with Deep Learning, vol. 121, Montreal, QC, Canada: PMLR; 2020, p. 33–43.
[32] Moe YM, Groendahl AR, Mulstad M, Tomic O, Indahl U, Dale E, et al. Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers. ArXiv [EessIV] 2019.
[33] Huang B, Chen Z, Wu P-M, Ye Y, Feng S-T, Wong C-YO, et al. Fully Automated Delineation of Gross Tumor Volume for Head and Neck Cancer on PET-CT Using Deep Learning: A Dual-Center Study. Contrast Media Mol Imaging 2018;2018:8923028.
[34] Naser MA, van Dijk LV, He R, Wahid KA, Fuller CD. Tumor Segmentation in Patients with Head and Neck Cancers Using Deep Learning Based-on Multi-modality PET/CT Images. Head and Neck Tumor Segmentation, Springer International Publishing; 2021, p. 85–98.
[35] Iantsen A, Visvikis D, Hatt M. Squeeze-and-Excitation Normalization for Automated Delineation of Head and Neck Primary Tumors in Combined PET and CT Images. Head and Neck Tumor Segmentation, Springer International Publishing; 2021, p. 37–43.
[36] Bielak L, Wiedenmann N, Berlin A, Nicolay NH, Gunashekar DD, Hägele L, et al. Convolutional neural networks for head and neck tumor segmentation on 7-channel

multiparametric MRI: a leave-one-out analysis. Radiat Oncol 2020;15:181.
[37] Bielak L, Wiedenmann N, Nicolay NH, Lottner T, Fischer J, Bunea H, et al. Automatic Tumor Segmentation With a Convolutional Neural Network in Multiparametric MRI: Influence of Distortion Correction. Tomography 2019;5:292–9.
[38] Gaddikeri S, Gaddikeri RS, Tailor T, Anzai Y. Dynamic Contrast-Enhanced MR Imaging in Head and Neck Cancer: Techniques and Clinical Applications. AJNR Am J Neuroradiol 2016;37:588–95.
[39] Mohamed ASR, He R, Ding Y, Wang J, Fahim J, Elgohari B, et al. Quantitative Dynamic Contrast-Enhanced MRI Identifies Radiation-Induced Vascular Damage in Patients With Advanced Osteoradionecrosis: Results of a Prospective Study. International Journal of Radiation Oncology*Biology*Physics 2020;108:1319–28. https://doi.org/10.1016/j.ijrobp.2020.07.029.
[40] Wahid KA, He R, McDonald BA, Anderson BM, Salzillo T, Mulder S, et al. MRI Intensity Standardization Evaluation Design for Head and Neck Quantitative Imaging Applications. MedRxiv 2021.
[41] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing; 2015, p. 234–41.
[42] Naser MA, Deen MJ. Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. Comput Biol Med 2020;121:103758.
[43] Ma N, Li W, Brown R, Wang Y, Gorman B, Behrooz, et al. Project-MONAI/MONAI: 0.5.0. 2021. https://doi.org/10.5281/zenodo.4679866.
[44] Blinde S, Mohamed ASR, Al-Mamgani A, Newbold K, Karam I, Robbins JR, et al. Large interobserver variation in the international MR-LINAC oropharyngeal carcinoma delineation study. Int J Radiat Oncol Biol Phys 2017;99:E639–40.
[45] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. ArXiv [CsCV] 2018.
[46] Gooding MJ, Smith AJ, Tariq M, Aljabar P, Peressutti D, van der Stoep J, et al. Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test. Med Phys 2018;45:5105–15.
[47] Nai Y-H, Teo BW, Tan NL, Chua KYW, Wong CK, O'Doherty S, et al. Evaluation of Multimodal Algorithms for the Segmentation of Multiparametric MRI Prostate Images. Comput Math Methods Med 2020;2020:8861035.
[48] Zhao X, Xie P, Wang M, Li W, Pickhardt PJ, Xia W, et al. Deep learning–based fully automated detection and segmentation of lymph nodes on multiparametric-mri for rectal cancer: A multicentre study. EBioMedicine 2020;56:102780.
[49] Pellicer-Valero OJ, Marenco Jiménez JL, Gonzalez-Perez V, Ramón-Borja JLC, García IM, Benito MB, et al. Deep Learning for fully automatic detection, segmentation, and Gleason Grade estimation of prostate cancer in multiparametric Magnetic Resonance Images. ArXiv [PhysicsMed-Ph] 2021.
[50] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29.
[51] Sherer MV, Lin D, Elguindi S, Duke S, Tan L-T, Cacicedo J, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother Oncol 2021;160:185–91.
[52] Sourbron SP, Buckley DL. On the scope and interpretation of the Tofts models for DCE-MRI. Magn Reson Med 2011;66:735–45.
[53] Fujima N, Sakashita T, Homma A, Shimizu Y, Yoshida A, Harada T, et al. Advanced diffusion models in head and neck squamous cell carcinoma patients: Goodness of fit, relationships among diffusion parameters and comparison with dynamic contrast-enhanced

perfusion. Magn Reson Imaging 2017;36:16–23.

## Figure Captions

**Figure 1.** Annotation, processing, and analysis of data used in this study. **(A)** Multiparametric MRI input channels for oropharyngeal tumor segmentation. The white dotted line depicts the primary gross tumor volume segmentation. Anatomical sequence images are outlined in grey boxes, while functional sequence parametric map images are outlined in red boxes. **(B)** Image processing steps which included image cropping, resampling, and rescaling. **(C)** An illustration of the 3D Residual U-net model architecture. For illustrative purposes, only one input channel (T2-weighted image) is shown, but multiple input channel combinations were used throughout the analysis as separate models. **(D)** Overall study design which incorporated multi-channel input combinations coupled to a leave-one-out cross-validation (LOOCV) evaluation approach. T2=T2-weighted MRI, T1=T1-weighted MRI, ADC=apparent diffusion coefficient, Ktrans=volume transfer constant, Ve=extravascular extracellular volume fraction, ALL=all 5 input channels. BN=Batch normalization, PReLU=parametric rectified linear unit activation function.

**Figure 2.** Boxplots comparing evaluation metrics of models built with different input channels. Evaluation metrics correspond to Dice similarity coefficient (DSC) (**A**), Hausdorff distance (HD) (**B**), false-negative DSC (FND) (**C**), false-positive DSC (FPD) (**D**), sensitivity (**E**), positive predictive value (PPV) (**F**), surface DSC (**G**), 95% HD (**H**), and mean surface distance (MSD) (**I**). Boxes show quartiles and median lines, while whiskers extend to the remaining distribution. Mean ± standard deviation is shown inside or adjacent to the corresponding box. The single and double stars above the boxplots correspond to significantly lower or higher values, respectively, compared to the baseline model for that metric. T2=T2-weighted MRI, T1=T1-weighted MRI, ADC=apparent diffusion coefficient, Ktrans=volume transfer constant, Ve= extravascular extracellular volume fraction, ALL=all 5 input channels.

**Figure 3.** 2D axial slice representations of ground truth segmentations (red dotted outline) and predicted segmentations (yellow dotted outline) for high- (green), medium- (blue), and low-

(orange) performance cases. Slices for each case are shown in rows superiorly to inferiorly (top, middle, and bottom). Models are shown in columns. The DSC scores for corresponding models are shown in the top left corners. The high-performance case corresponds to a left tonsillar T4 tumor. The medium-performance case corresponds to a left base of tongue T4 tumor. The low-performance case corresponds to a right base of tongue T4 tumor. T2=T2-weighted MRI, T1=T1-weighted MRI, ADC=apparent diffusion coefficient, Ktrans=volume transfer constant, Ve=extravascular extracellular volume fraction, ALL=all 5 input channels.

**Figure 4.** Dependence of tumor size on the Dice Similarity Coefficient (DSC) (**A**), Hausdorff Distance (HD) (**B**), and surface DSC (**C**), for various input channel models. T2=T2-weighted MRI, T1=T1-weighted MRI, ADC=apparent diffusion coefficient, Ktrans=volume transfer constant, Ve=extravascular extracellular volume fraction, ALL=all 5 input channels.

# Figures

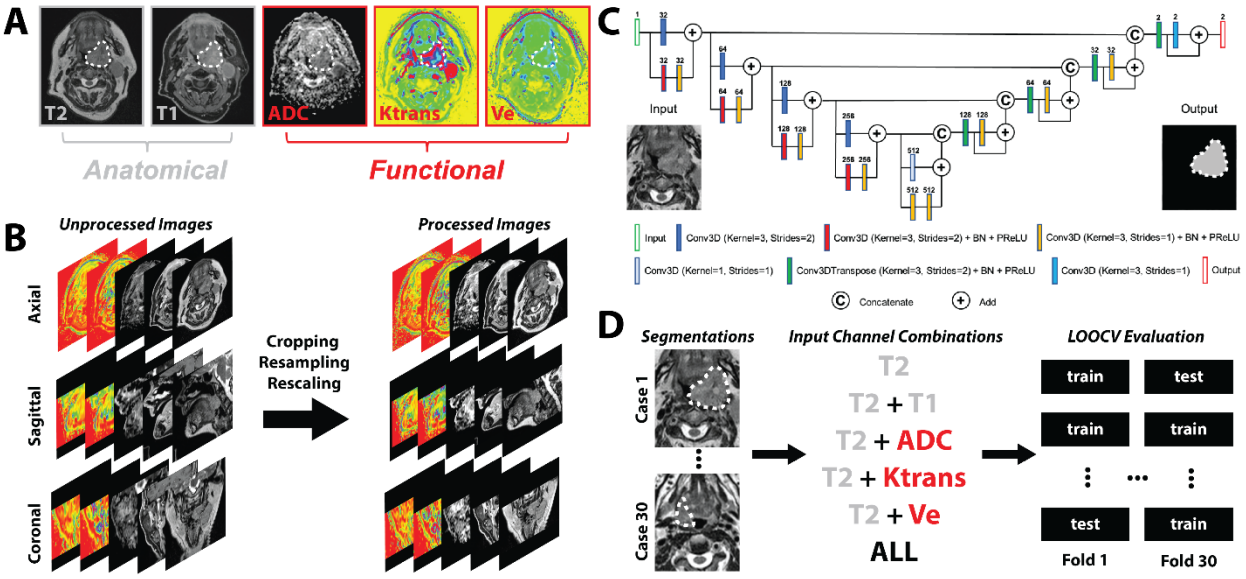


**Figure 1.**

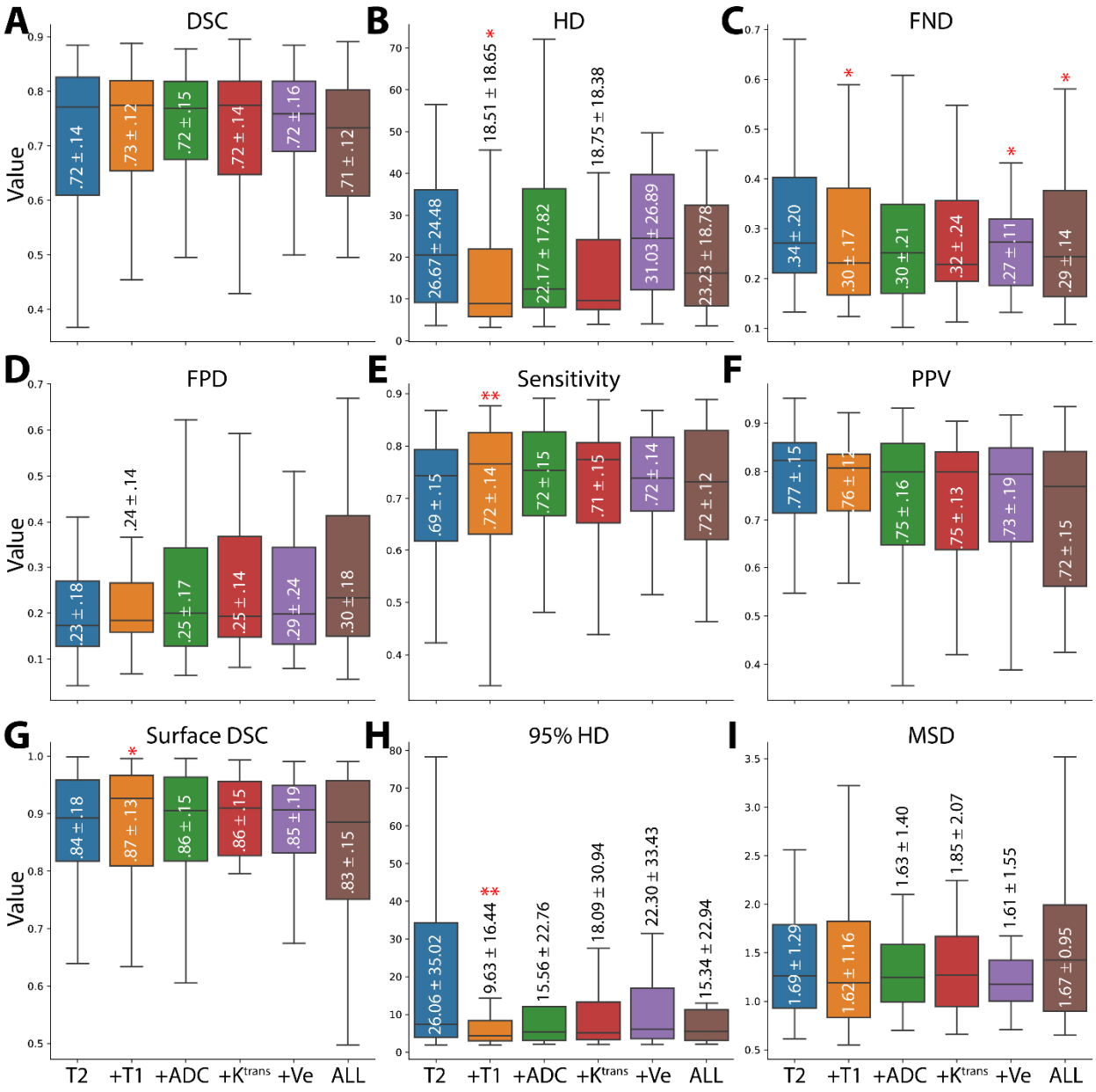


**Figure 2.**

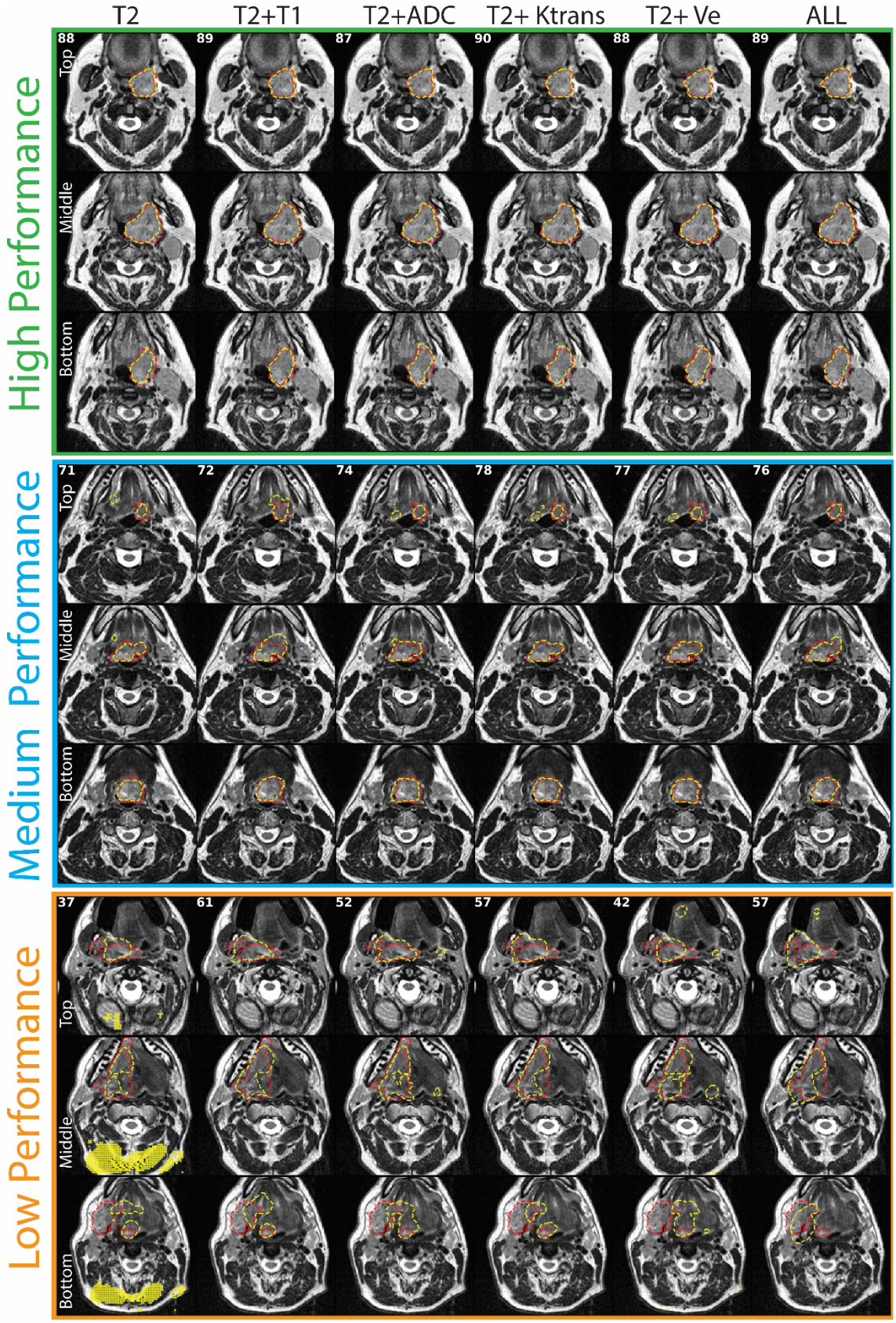


**Figure 3.**

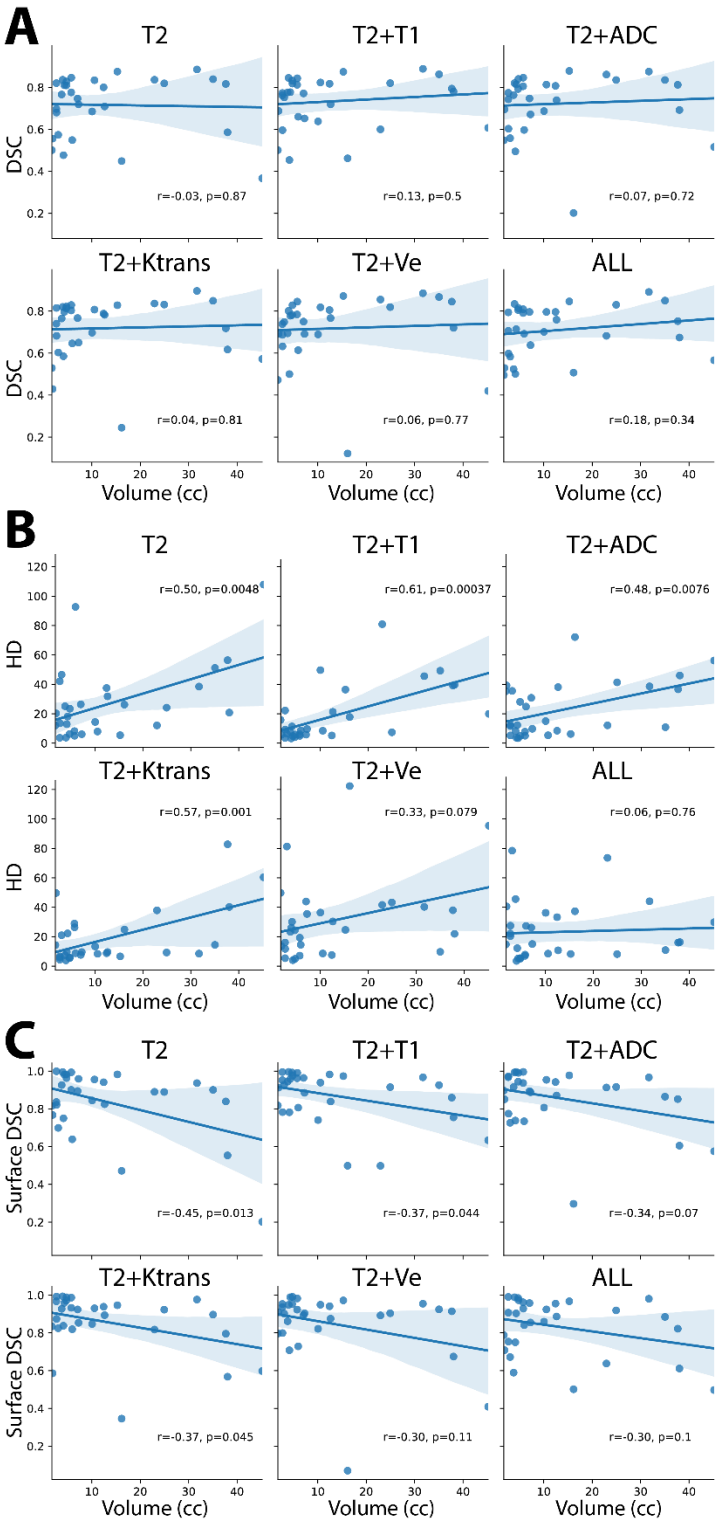


**Figure 4.**

# Tables

**Table 1.** Clinical evaluation and Turing test results for three physician expert observers. Each observer was asked to score blinded ground truth (GT) or deep learning (DL)-generated segmentations on a 4-point Likert scale (1 = requires corrections, large errors; 2 = requires corrections, minor errors; 3 = clinically acceptable, errors not clinically significant; 4 = clinically acceptable, highly accurate) and asked to identify the source of the segmentation (GT or DL). DL-generated segmentations corresponded to the best DL model tested (T2-weighted + T1-weighted).

| Observer | Score | GT (#) | DL (#) | p-value[1] | Source | GT (#) | DL (#) | p-value[2] |
|---|---|---|---|---|---|---|---|---|
| **Observer 1 (Radiologist)** | 1 | 3 | 6 | 0.13 | GT | 16 | 10 | 0.18 |
| | 2 | 7 | 11 | | DL | 14 | 20 | |
| | 3 | 7 | 4 | | | | | |
| | 4 | 13 | 9 | | | | | |
| **Observer 2 (Radiation-Oncologist)** | 1 | 3 | 6 | 0.44 | GT | 14 | 14 | 1.00 |
| | 2 | 10 | 4 | | DL | 16 | 16 | |
| | 3 | 16 | 13 | | | | | |
| | 4 | 1 | 7 | | | | | |
| **Observer 3 (Radiation-Oncologist)** | 1 | 1 | 3 | 0.98 | GT | 9 | 12 | 0.61 |
| | 2 | 8 | 6 | | DL | 21 | 18 | |
| | 3 | 11 | 10 | | | | | |
| | 4 | 10 | 11 | | | | | |

[1] Two-sided Wilcoxon signed rank tests were used for score comparisons. [2] McNemar tests were used for source prediction comparisons.

**Table 2.** Survey of relevant DL auto-segmentation literature for comparison with our study. Only studies ≤3 years old and with sample sizes ≥30 were selected for comparison. DL=deep learning, N=number of images sets used in study, GTVp=primary gross tumor volume, DSC=Dice similarity coefficient, OPC=oropharyngeal cancer, NPC=nasopharyngeal cancer, HNSCC=head and neck squamous cell carcinoma, T2=T2-weighted MRI, T1=T1-weighted MRI, DCE=dynamic contrast enhanced MRI, DWI=diffusion weighted imaging MRI, LOOCV=leave-one-out cross-validation, CV=cross-validation, CNN=convolutional neural network.

| Author, Year | Site | Modality | DL Architecture | N (Train, Test) | GTVp DSC (average) |
|---|---|---|---|---|---|
| This study, 2021 | OPC | MRI (T1, T2, DCE, DWI) | 3D Residual Unet | 30 (LOOCV) | 0.73 (best model, T2+T1) |
| Outeiral et al., 2021 [26] | OPC | MRI (T1, T2) | 3D Unet | 171 (151, 20) | 0.55 |
| Andrearczyk et al., 2020 [31] | OPC | CT, PET | 2D Unet | 202 (LOOCV) | 0.48 (CT), 0.58 (PET), 0.6 (PET/CT) |
| Moe et al., 2019 [32] | OPC | CT, PET | 2D Unet | 197 (157, 40) | 0.65 (CT), 0.71 (PET), 0.75 (PET/CT) |
| Naser et al., 2020 [34] | OPC | CT, PET | 3D Unet | 201 (5-fold CV) | 0.69 |
| Iantsen et al., 2020 [35] | OPC | CT, PET | 3D Unet | 201 (4-fold CV) | 0.745 |
| Ma et al., 2018 [18] | NPC | MRI (T1) | 3D CNN + graph-cut | 30 (LOOCV) | 0.85 |
| Ye et al., 2020 [17] | NPC | MRI (T1, T2) | 3D Unet | 44 (10-fold CV) | 0.62 (T1), 0.64 (T2), 0.72 (T1+T2) |
| Chen et al., 2020 [20] | NPC | MRI (T1, T2) | 3D Encoder-decoder network | 149 (5-fold CV) | 0.72 |
| Huang et al., 2019 [25] | NPC | MRI (T1, T2) | 2D CNN + recurrent attention | 596 (430, 166) | 0.78 |
| Lin et al., 2019 [16] | NPC | MRI (T1, T2) | 3D CNN | 1021 (818, 203) | 0.79 |
| Ke et al., 2020 [22] | NPC | MRI (T1, T2) | 3D DenseNet + multi-task learning | 3142 (2792, 350) | 0.77 |
| Li et al., 2018 [19] | NPC | MRI (T1) | 2D CNN | 87 (LOOCV) | 0.89 |
| Ma et al., 2019 [24] | NPC | CT, MRI (T1) | 2D CNN | 90 (5-fold CV) | 0.752 |
| Bielak et al., 2020 [36] | HNSCC | MRI (T1, T2, DCE, DWI) | 3D CNN (DeepMedic) | 36 (LOOCV) | ~0.30* |

* Average DSC interpreted from manuscript figure.