International Journal of
Radiation Oncology
biology ● physics

www.redjournal.org

Physics Contribution

Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function

Carlos E. Cardenas, MS,[*] Rachel E. McCarroll, BS,[*] Laurence E. Court, PhD,[*] Baher A. Elgohari, MD,[†,‡] Hesham Elhalawani, MD,[†] Clifton D. Fuller, MD, PhD,[†] Mona J. Kamal, MD, PhD,[†,§] Mohamed A.M. Meheissen, MD, MSc,[†,||] Abdallah S.R. Mohamed, MD, MSc,[†,||] Arvind Rao, PhD,[¶] Bowman Williams, BS,[†] Andrew Wong, MD,[†] Jinzhong Yang, PhD,[*] and Michalis Aristophanous, PhD[#]

Departments of [*]Radiation Physics, [†]Radiation Oncology, and [¶]Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas; [‡]Department of Clinical Oncology and Nuclear Medicine, Faculty of Medicine, Mansoura University, Mansoura, Egypt; [§]Department of Clinical Oncology and Nuclear Medicine, Faculty of Medicine, Ain Shams University, Cairo, Egypt; [||]Department of Clinical Oncology and Nuclear Medicine, Faculty of Medicine, University of Alexandria, Alexandria, Egypt; and [#]Department of Radiation and Cellular Oncology, The University of Chicago, Chicago, Illinois

Received Sep 25, 2017, and in revised form Jan 2, 2018. Accepted for publication Jan 31, 2018.

Reprint requests to: Carlos E. Cardenas, MS, Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Pickens Academic Tower, 1400 Pressler St, Unit 1420, Houston, TX 77030-3722. Tel: (713) 653-6393; E-mail: cecardenas@mdanderson.org

A preliminary analysis of a portion of our data set was presented as an oral presentation at the 59th American Association of Physicists in Medicine Annual Meeting, August 1, 2017, Denver, CO.

Dr Fuller is supported by the Andrew Sabin Family Foundation; is a Sabin Family Foundation Fellow; receives funding and salary support from the National Institutes of Health (NIH), including the National Institute for Dental and Craniofacial Research Award (grant 1R01DE025248-01/R56DE025248-01), a National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (grant 1R01CA225190-01 and NSF grant 1557679); the NIH Big Data to Knowledge Program of the National Cancer Institute (NCI) Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (grant 1R01CA214825-01), NCI Early Phase Clinical Trials and Imaging and Image-Guided Interventions Program (grant 1R01CA218148-01), an NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from the University of Texas MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (grant P30CA016672), and an NIH/NCI Head and Neck Specialized Programs of Research Excellence Developmental Research Program Award (grant P50 CA097007-10); and has received direct industry grant support and travel funding from Elekta AB. Drs Meheissen and Elgohari receive funding from the Egyptian Ministry of Higher Education.

Conflict of interest: none.

Supplementary material for this article can be found at www.redjournal.org.

Acknowledgments—We gratefully acknowledge the support of the Nvidia Corporation with the donation of the Tesla K40 GPU used for the present research. In addition, we acknowledge the Texas Advanced Computing Center at The University of Texas at Austin for providing high-power computing resources that have contributed to the research results we report. Finally, we acknowledge Varian Medical Systems for their financial support.

https://doi.org/10.1016/j.ijrobp.2018.01.114

Summary

Clinical target volume (CTV) delineation can vary among physicians and practices. In an effort to reduce this variability, we developed a deep learning auto-delineation algorithm for high-risk CTVs with a Dice similarity parameter optimization function. Most of our predicted volumes had high agreement with the physician ground-truth volumes. The predicted contours could be implemented clinically with only minor or no changes.

Purpose: Automating and standardizing the contouring of clinical target volumes (CTVs) can reduce interphysician variability, which is one of the largest sources of uncertainty in head and neck radiation therapy. In addition to using uniform margin expansions to auto-delineate high-risk CTVs, very little work has been performed to provide patient- and disease-specific high-risk CTVs. The aim of the present study was to develop a deep neural network for the auto-delineation of high-risk CTVs.

Methods and Materials: Fifty-two oropharyngeal cancer patients were selected for the present study. All patients were treated at The University of Texas MD Anderson Cancer Center from January 2006 to August 2010 and had previously contoured gross tumor volumes and CTVs. We developed a deep learning algorithm using deep auto-encoders to identify physician contouring patterns at our institution. These models use distance map information from surrounding anatomic structures and the gross tumor volume as input parameters and conduct voxel-based classification to identify voxels that are part of the high-risk CTV. In addition, we developed a novel probability threshold selection function, based on the Dice similarity coefficient (DSC), to improve the generalization of the predicted volumes. The DSC-based function is implemented during an inner cross-validation loop, and probability thresholds are selected a priori during model parameter optimization. We performed a volumetric comparison between the predicted and manually contoured volumes to assess our model.

Results: The predicted volumes had a median DSC value of 0.81 (range 0.62-0.90), median mean surface distance of 2.8 mm (range 1.6-5.5), and median 95th Hausdorff distance of 7.5 mm (range 4.7-17.9) when comparing our predicted high-risk CTVs with the physician manual contours.

Conclusions: These predicted high-risk CTVs provided close agreement to the ground-truth compared with current interobserver variability. The predicted contours could be implemented clinically, with only minor or no changes.

Introduction

Manual delineation of clinical target volumes (CTVs) remains a time-consuming task in radiation oncology. CTVs are tissue volumes that contain the demonstrable gross tumor volume (GTV) and provide coverage for any suspected microscopic disease and pathways of tumor spread such as regional lymph nodes (1). Because the radiation dose is prescribed to these volumes and adequate coverage is required to achieve cure, accurate CTV delineation is essential in radiation therapy. Although established guidelines are available to delineate site-specific CTVs, these volumes are still subject to high intra- and interobserver variability for most treatment sites (2-6). This variability in delineation and the heterogeneity in clinical practice have hindered our ability to systematically assess the quality of the radiation therapy plans and are considered major sources of uncertainty (7).

When treating head and neck (H&N) cancer, radiation therapy prevails as the principal nonsurgical treatment option. For this site in particular, the complexity of radiation treatment planning and the time required to delineate the target and normal tissue volumes are significantly increased (8) owing to the large number of organs at risk located near H&N tumors. To add to this complexity, H&N treatment plans typically require several CTVs, which are used to deliver different radiation dose levels, depending on the risk of recurrence for that region (ie, high-, intermediate-, and low-risk volumes). In particular, accurate delineation of the high-risk CTV is imperative, and failure to provide adequate coverage has the potential to reduce tumor control and increase the risk of locoregional recurrence (9, 10).

Although an abundance of work auto-delineating normal structures using atlas-based registration techniques is available (11-13), little work has been performed to auto-delineate H&N CTVs, especially to auto-delineate high-risk target volumes. Machine learning and deep learning normal tissue auto-segmentation approaches have increased in popularity during the past few years. Some improvements in normal tissue segmentation have been observed using these novel techniques; however, a need remains to investigate these approaches for auto-delineation of CTVs. To the best of our knowledge, no registration-based approaches are available to auto-delineate high-risk CTVs. This is not surprising owing to the lack of significant features on computed tomography (CT) images (limited by coverage of possible microscopic disease) and the high variability in GTV geometric shape, location, and subsite involvement. Although definition of the high-risk CTV is guided by the anatomic structures, the high-risk CTV is neither a distinct structure, such as the GTV, nor a specific anatomic structure, such as elective nodal chains. These limitations have hindered the development of auto-delineation algorithms for these volumes.

Our previous work (14) has shown that distance metrics can provide sufficient information to automate the delineation of high-risk CTVs and that deep auto-encoders (15, 16) provide a venue for good generalization even when few patients are used for training. This is primarily because these models were trained on a voxel by voxel basis providing hundreds of thousands of inputs per patient for training. In addition, a preliminary study (16) from our group showed that clustering patients per site and nodal status provided improvement in prediction performance for oropharyngeal patients.

Automating the CTV delineation process for H&N tumors would offer many clinical advantages. First, it has the potential to reduce the variability in target design and clinical practice among radiation oncologists. This reduction in variability would provide better data for multi-institutional studies in which clinical practices can vary greatly (2). Second, it would aid in reducing the physician contouring time. This would allow physicians to spend more time with patients to provide better quality of care.

In the present report, we propose a novel method to auto-delineate high-risk CTVs that overcomes several of the current limitations. Our approach requires only a limited amount of training data and performed well compared with manual contours. More specifically,

- We propose a deep learning approach in which the model is trained on anatomic structure distance map information to produce patient-specific high-risk CTVs.
- We have addressed, to the best of our knowledge for the first time, a nonuniform margin approach to the auto-delineation of high-risk CTVs for H&N patients.
- We introduce a novel threshold selection function to convert probability maps into binary volumes.
- Finally, we present an evaluation of our method and show that our predicted volumes are in close agreement with manually drawn contours.

Methods and Materials

Patient and image characteristics

A total of 52 oropharyngeal cancer patients (11 base of tongue node-negative, 15 base of tongue node-positive, 15 tonsil node-negative, and 11 tonsil node-positive) who had undergone curative-intent intensity modulated radiation therapy for H&N squamous cell carcinoma from January 2006 to August 2010 at The University of Texas MD Anderson Cancer Center were selected from an institutional review board—approved protocol. All patients had available simulation CT scans with previously manually contoured GTVs (primary and nodal, as applicable) and high-risk CTVs used for treatment planning. Each CT image included the H&N region and had matrix sizes of 512 × 512 × number of slices (median 152, range: 47-348). The voxel size was 0.976 mm × 0.976 mm × 2.5 to 3.0 mm. The contours delineated on these images were used in the present study.

Stacked auto-encoders

We chose to use stacked auto-encoders owing to their ability to speed up training and provide improvement in predictions by initializing weights through unsupervised learning (17). During unsupervised learning, only the input data are provided, and the auto-encoders learn a general representation of the data set. Hidden layer neurons were activated using the logistic function. After this unsupervised learning step, we trained the output layer through supervised learning and used cross-validation to fine-tune the network architecture. During the supervised learning step, our algorithm fine-tuned the architecture by updating the network's weights to match the training set's inputs to the training set's known output. Our deep auto-encoders are composed of 2 hidden layers, followed by a soft-max layer for binary classification. An illustration of the network's architecture is provided Fig. E1 (available online at www.redjournal.org). To provide an improvement in generalization, we implemented L_2-norm and sparsity (Kullback-Leibler divergence) regularization (18) into the mean squared error cost function (Eq. 1) used during unsupervised training; the cross-entropy cost function (Eq. 2) was used during supervised training and fine-tuning:

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(x_{kn} - \widehat{x}_{kn}\right)^2 + \lambda * \Omega_{\text{weights}} + \beta * \Omega_{\text{sparsity}} \quad (1)$$

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk} + (1 - t_{nk})\ln(1 - y_{nk}) \quad (2)$$

where λ is the coefficient for the L_2-norm regularization term, β is the coefficient for the sparsity regularization term, t_{nk} is the nkth entry of the target matrix, and y_{nk} is the nkth output from the auto-encoder when the input vector is x_n. Finally, scaled conjugate gradient optimization (19) was selected for training owing to the greater convergence speed and classification performance (20, 21).

Model features and outcome

The patients' GTV (primary and nodal for node-positive patients), high-risk CTV, and anatomic structures (eg, mandible, skull, vertebral body, pharyngeal and nasopharyngeal air cavities, left and right parotid glands, maxillary arch, hyoid, thyroid cartilage, and skin) were manually contoured in the Pinnacle treatment planning system (Phillips Medical Systems, Amsterdam, Netherlands). The high-risk CTVs and GTVs had been previously contoured for the purpose of planning the patients' radiation therapy and underwent rigorous peer-review by a group of subspecialized H&N radiation oncologists (22); the selected anatomic structures were contoured specifically for the present study. All the volumes were converted to 3-dimensional (3D) binary masks using MATLAB R2016b (MathWorks, Natick, MA).

Three-dimensional distance maps were calculated from the GTV and anatomic structure's binary masks as follows: for each voxel, $v(x,y,z)$, in the CT image space, minimum Euclidian distance vectors $(r,\ \theta,\ \varphi)$ were calculated for each structure (GTVs and anatomic structures) such that for each $v(x,y,z)$ we would have 12 distance vectors, $V(r,\ \theta,\ \varphi)$, 13 if node-positive, and 36 (or 39) distance input features. Signed distances were used to differentiate voxels that were located inside the contoured volume. In addition to these features, we extracted each voxel's corresponding class (0 or 1) based on the high-risk CTV mask providing the following relationship:

$$CTV(x,y,z) \sim v_{GTV}(x,y,z) + v_{\mathrm{Mandible}}(x,y,z) + \cdots + v_{\mathrm{Skin}}(x,y,z) \tag{3}$$

To reduce the computational time, we only included voxels within 5 cm of the GTV for training and predicting for new patients. This was a conservative value because all high-risk CTVs used in the present study were within 2 cm of the GTV. Once the models were trained and used to predict with a test patient, the output from the test patient was a patient-specific probability map of the high-risk CTV. Our preliminary work (16) showed that training models by grouping patients per site and nodal status improved the overall prediction performance; thus, this approach was implemented in the present study.

Postprocessing and probability threshold selection

Once the probability map for a patient was created, we used a 3D Gaussian filter ($\sigma = 1$) to obtain a smooth probability map. Although most machine learning algorithms use a threshold of 0.5 to convert probabilities into binary classes, we chose to optimize this probability threshold selection by including a Dice similarity coefficient (DSC) (23) probability threshold selection function during cross-validation in model training. This provided a more useful metric than the area under the curve and classification error owing to the imbalance in classes. In addition, we evaluated the performance of the DSC loss function compared with that of distance metrics such as the mean surface distance (MSD) and the 95th percentile Hausdorff distance (95HD) between the predicted and ground-truth volumes. These metrics were evaluated by converting the probability maps into binary volumes by increasing the probabilities from 0 to 1 in 0.005 steps. Three-dimensional and 2-dimensional closing and opening algorithms were used on the binary images for additional postprocessing before evaluation. To prevent overfitting our models set, we evaluated the models' predictive accuracy, based on DSC, for different training epochs. An epoch is a complete pass through a given data set, meaning that at the end of each epoch, all patient data in the trained model were seen at least once by the neural network.

Evaluation

Three-dimensional volume metrics were used to assess the performance of the predicted volumes. In addition to calculating the DSC, MSD, and 95HD between the predicted volumes and the manually contoured high-risk CTVs, we calculated the difference between the volumes, false-negative Dice (FND), false-positive Dice (FPD) (24), and the normalized volumetric difference (VD). The FND and FPD can be used as surrogates for potential near misses and overtreatment, respectively.

$$DSC = \frac{2*TP}{2*TP+FN+FP} \tag{4}$$

$$MSD = \frac{1}{2}\left(\bar{d}_{DNN,G} + \bar{d}_{G,DNN}\right) \tag{5}$$

$$95HD = \mathrm{percentile}\left(d_{DNN,G} \cup d_{G,DNN}, 95^{th}\right) \tag{6}$$

$$FND = \frac{2*FN}{2*TP+FN+FP} \tag{7}$$

$$FPD = \frac{2*FP}{2*TP+FN+FP} \tag{8}$$

$$VD = \frac{V_{DNN} - V_G}{V_G} \tag{9}$$

where TP, FN, and FP indicate true positive, false negative, and false positive, respectively; DNN and G, the auto-delineated and ground-truth (manual) contours, respectively; and $d_{\mathrm{DNN,G}}$ is a vector containing all minimum Euclidean distances from each surface voxel on volume DNN to volume G. In addition, we compared the CTVs generated using uniform margin expansions to our ground-truth CTVs using these same metrics. A uniform expansion of 0.5 cm from the GTV was selected, because it was systematically used by the Danish Head and Neck Cancer group (25). Finally, we evaluated the differences in planning target volumes (PTVs) when adding a 0.3-cm margin to the ground-truth (PTV), DNN, and uniform margin CTVs.

Cross-validation

During model training, we used nested leave-one-out cross-validation (LOOCV) for parameter-tuning using a grid-search approach. The parameters optimized during the grid search were the number of layers, number of nodes per layer, L2 weight regularization value at each layer, sparsity regularization value at each layer, and sparsity proportion at each layer. In our nested LOOCV method (Fig. 1). All voxels from the test patient were excluded from training and were not used to predict a volume until the parameters had been optimized through an internal cross-validation loop. In this internal LOOCV loop, models were trained, leaving out all voxels for the cross-validation patient. Every time a model was trained in the internal LOOCV loop, the model was used to predict the high-risk CTV of the cross-

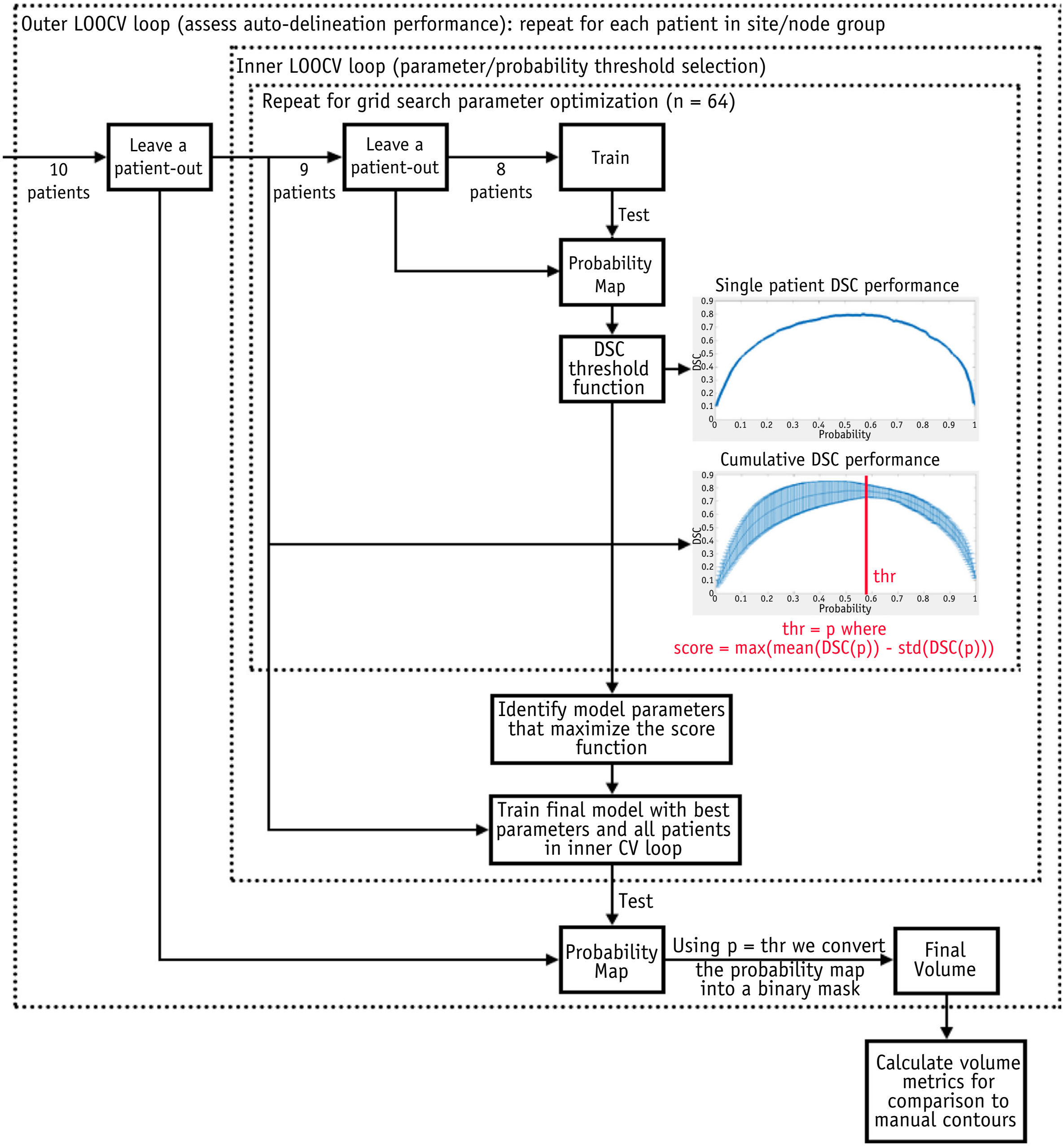
Outer LOOCV loop (assess auto-delineation performance): repeat for each patient in site/node group
Inner LOOCV loop (parameter/probability threshold selection)
Repeat for grid search parameter optimization (n = 64)
10 patients
Leave a patient-out
9 patients
Leave a patient-out
8 patients
Train
Test
Probability Map
DSC threshold function
Single patient DSC performance
DSC
Probability
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Cumulative DSC performance
DSC
Probability
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
thr
thr = p where
score = max(mean(DSC(p)) - std(DSC(p)))
Identify model parameters that maximize the score function
Train final model with best parameters and all patients in inner CV loop
Test
Probability Map
Using p = thr we convert the probability map into a binary mask
Final Volume
Calculate volume metrics for comparison to manual contours

Fig. 1. Block diagram of nested leave-one-out cross-validation (LOOCV). In the inner loop, model parameters were selected by maximizing the score function using Dice similarity coefficient (DSC) curves for all patients in the inner loop. The probability threshold value identified for the corresponding model parameters was used after training the final model to convert the predicted probability map into a binary structure on a test patient (outer loop). This final volume was then evaluated using overlap and distance metrics to compare it to the physician manually delineated high-dose clinical target volumes. *Abbreviations:* max = maximum; std = standard deviation.

validation patient, and their prediction performance was used to determine the optimal parameter selection.

Results

Using an Intel Xeon central processing unit (2.8 GHz × 10 cores) and a Tesla K40 graphic processing unit, training required on average 2.51 ± 0.85 hours per patient, and the predicted high-risk CTVs were created within a mean time of 2.75 ± 0.62 seconds. Although the predictions were almost instantaneous, calculating the distance maps for each patient before predicting the new volumes required on average 9.0 ± 3.3 minutes. The volume statistics for the manually contoured GTVs and high-risk CTVs, DNN CTV, uniform CTV, and their respective PTVs are listed in Table 1, and their respective distributions are shown in Figure 2. The

Table 1 Volume statistics (in cm^3) for manually contoured GTVs and high-risk CTVs and predicted CTV1s

Variable	Minimum	Median	Maximum	Mean	SD
GTV	1.0	22.4	103.9	26.8	20.7
CTV1	14.3	101.2	255.2	102.9	58.7
CTV1-DNN	16.1	88.6	273.8	101.1	55.6
CTV1-Uniform	5.3	56.7	195.9	62.4	39.2
PTV1	24.2	147.0	389.5	151.1	80.5
PTV1-DNN	29.4	127.1	423.3	145.3	77.7
PTV1-Uniform	11.3	87.3	267.4	93.9	53.4

Abbreviations: CTV = clinical target volume; CTV1 = ground-truth CTV; DNN = auto-delineated contours; GTV = gross tumor volume; PTV = planning target volume; PTV1 = ground-truth PTV; Uniform = uniform margin expansion; SD = standard deviation.

mean volume difference between the DNN and ground-truth CTV was 1.0 ± 29.5 cm^3 (range −73.3 to 63.9). The corresponding difference between the uniform and ground-truth CTV was 47.7 ± 30.5 cm^3 (range 2.3-126.6). All CTVs generated with uniform margins were smaller than the ground-truth CTVs, and 50% of DNN-predicted volumes were smaller than their corresponding ground-truth volumes. When comparing volume overlap between the ground-truth PTV and PTV DNN, we found a mean DSC of 0.81 ± 0.05 (range 0.67-0.90). The DSC values between the ground-truth PTV and PTV-uniform margin were significantly reduced ($P < .0001$, Wilcoxon rank sum test), with a mean of 0.73 ± 0.10 (range 0.35-0.87).

When comparing the DSC threshold selection function to the distance metric performance (Fig. 3), both functions produced similar results for the training, cross-validation, and test sets when choosing the maximum DSC and minimum MSD and 95HD scores for probability threshold selection. Because calculation of the DSC requires minimum computational resources, we opted to use this metric for probability threshold selection moving forward.

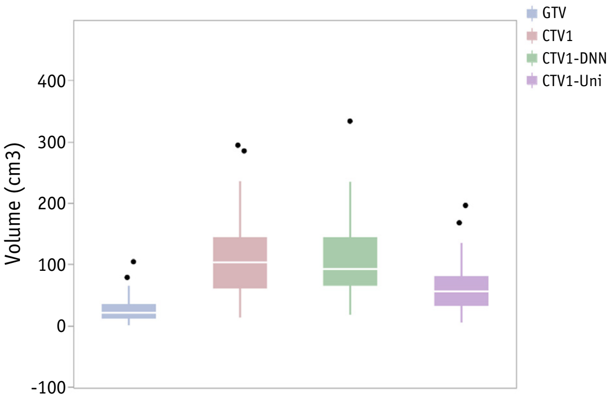

Fig. 2. Volume distributions of gross tumor volume (GTV), ground-truth clinical target volume (CTV1), DNN predicted CTV (CTV-DNN), and uniform margin expansion CTV (CTV-Uni). The volume distributions of the CTV1 and CTV1-DNN were similar.

Evaluation of the epochs used for training showed an initial increase in performance that was followed by a decrease in performance on the cross-validation and test sets when using 500 epochs. This decrease in performance was not observed in the training set, hinting that the models began overfitting approximately between 250 and 500 epochs (Fig. 4).

A comprehensive evaluation between the DNN auto-delineated volumes and physician manual contours is provided Table E1 (available online at www.redjournal.org) and Figure 5. The predicted volumes for 4 patients are illustrated in Figure 6. These volumes showed good agreement between these volumes and the physician

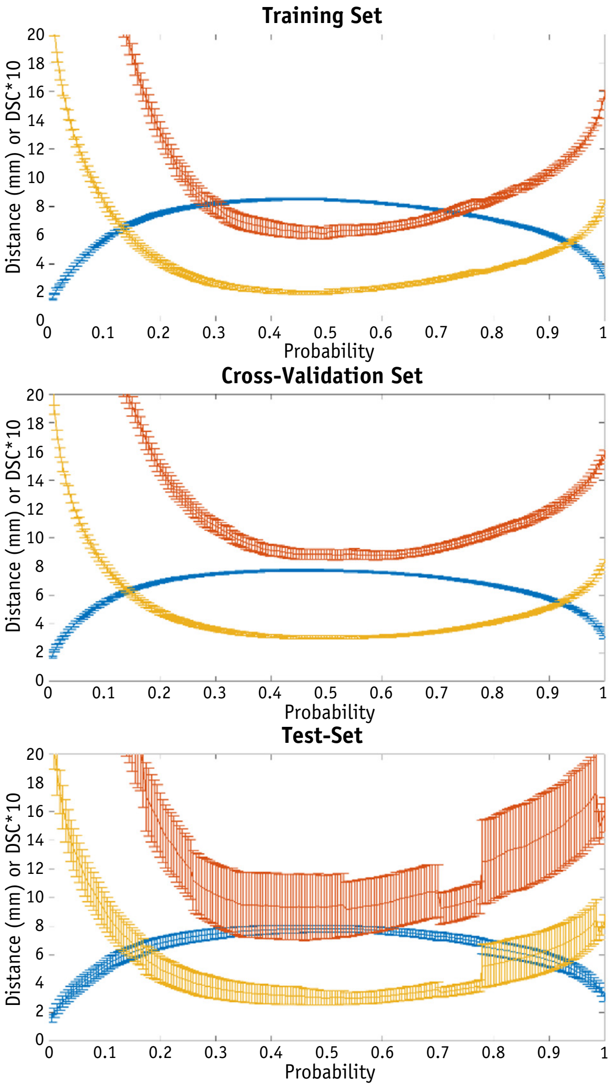

Fig. 3. Comparison of Dice similarity coefficient (DSC) and distance metrics for probability threshold selection. Mean DSC (plus standard error) depicted in blue, with mean 95th percentile Hausdorff distance and mean surface distance in yellow and red, respectively. DSC displayed as DSC × 10 for visual comparison.

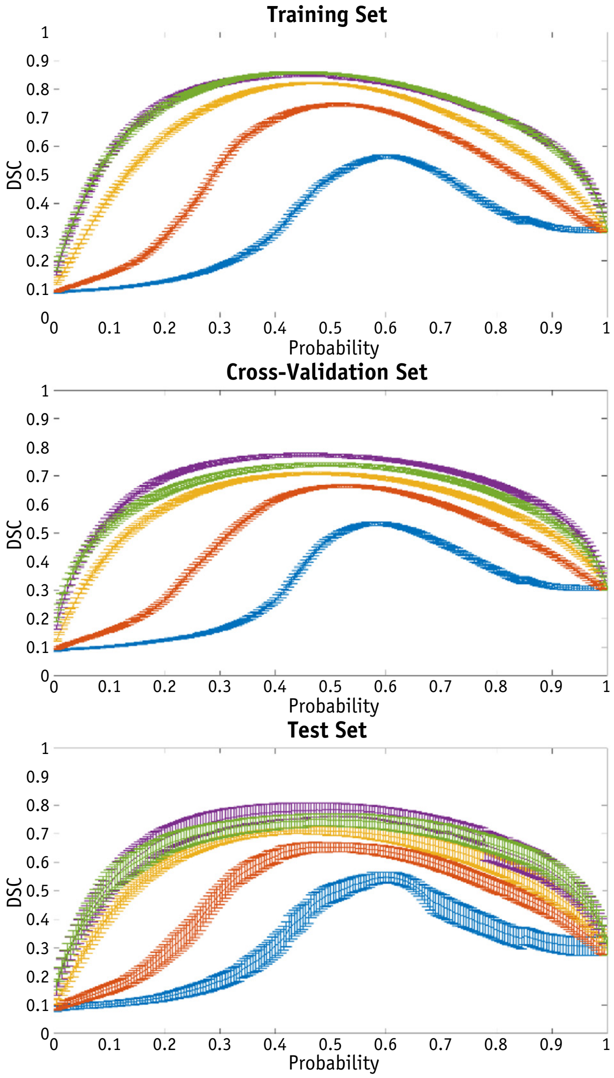

Fig. 4. Epoch analysis results for training, cross-validation, and test sets. Epochs used were 15 (blue), 50 (orange), 150 (yellow), 250 (purple), and 500 (green). Error bars provide standard error from the mean Dice similarity coefficient (DSC) value at each probability threshold.

manual contours. For the 52 patients, the median DSC was 0.814 (range 0.622-0.904), median MSD was 2.75 mm (range 1.57-5.47), and median 95HD was 7.49 mm (range 4.74-17.85). Overall, the auto-delineated volumes were slightly larger on average (0.034 ± 0.265), with a median FPD of 0.199 (range 0.004-0.652) and median FND of 0.151 (range 0.010-0.623). Evaluation of the ground-truth volumes showed large variability in the GTV-to-CTV expansions in the craniocaudal direction, with a mean expansion of 10.7 ± 5.1 mm (range 3.0-26.6) in the cranial direction and 9.7 ± 6.2 mm (range 0.0-30.0) in the caudal direction. The variability measured in the craniocaudal expansion of the ground-truth CTVs affected the accuracy of the DNN predicted volumes in these directions showing high FND (undertreatment) and FPD (overtreatment) values for some patients.

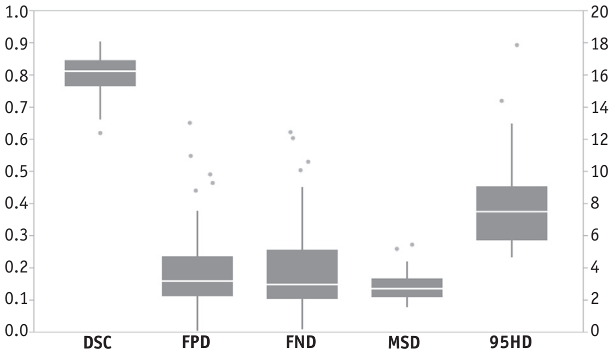

Fig. 5. Volumetric comparison between the auto-delineated and manually contoured volumes. Dice similarity coefficient (DSC), false-negative Dice (FND), and false-positive Dice (FPD) values reflected by the left vertical axis and mean surface distance (MSD) and 95th percentile Hausdorff distance (95HD) given in millimeters and correspond to the right vertical axis.

The interdisease site and nodal group comparison is presented in Figure 7. The predicted volumes from patients with nodal disease showed slightly greater overlap agreement, in terms of DSC, to the ground truth than did those without nodal disease (median DSC 0.83 vs 0.77). However, this difference was not statistically significant ($P = .25$, Wilcoxon rank sum test). In addition, the DSC values for the nodal volumes showed less variability (Table 2).

Discussion

The use of deep learning in medical image segmentation has become more popular during the past few years. Most efforts have focused on auto-segmenting normal tissues, with very little work performed to automate the delineation of CTVs. In our approach, we used distance maps from normal structures and GTVs to learn the physician patterns in auto-delineating high-risk CTVs. This approach was chosen owing to the lack of visible anatomic edges on CT imaging and the high variability in GTV location and size. In addition, because our algorithm uses the binary contours to compute the inputs for our model, the normal tissue and GTV contours created using any modality (ie, magnetic resonance imaging) could be used to generate automated high-risk CTVs. Furthermore, this approach could be used to train physician- or institution-specific models to automate this process and retaining the patterns used for the desired clinical practice. This remains to be evaluated because it was outside the scope of the present study.

Our deep learning approach was able to auto-delineate high-risk CTVs with DSC values (mean DSC 0.81) comparable to those observed for normal tissue auto-segmentation techniques (26). McCarroll et al showed

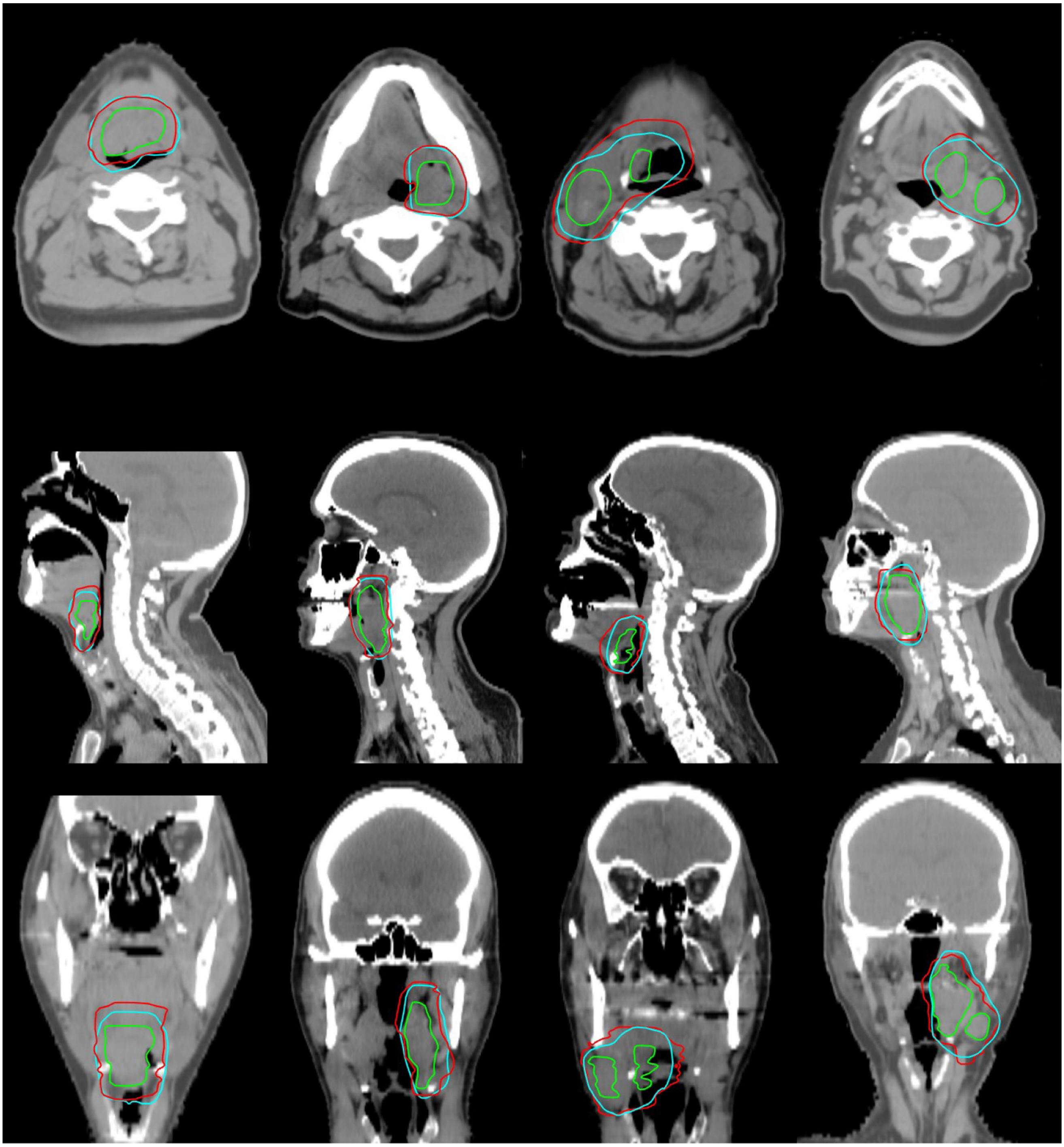

Fig. 6. Comparison between predicted ground-truth clinical target volume (CTV1) (blue) and physician manual contours (red) for 4 oropharyngeal cancer patients. The primary and nodal gross tumor volume is included (green). From left to right, we illustrate a case from each site and nodal status (base of tongue node-negative, tonsil node-negative, base of tongue node-positive, and tonsil node-positive).

that after clinically implementing a normal tissue auto-segmentation tool, the average DSC between the auto-contours and physician-edited volumes was 0.78 for 8 H&N normal structures. Overall, our DSC values ranged from 0.62 to 0.90, and the median MSD was 2.75 mm. These results are comparable to those reported by inter- and intraobservational studies for manual delineation of these volumes (27, 28). When using uniform margin expansions, we found that all auto-generated CTVs were smaller in volume than the ground-truth volumes and that mean difference in volumes between the DNN auto-delineated and ground-truth CTVs was 1.0 ± 29.5 cm^3. The variability in the craniocaudal extent of the GTV-to-CTV margin expansion made it difficult to assess the occurrence of

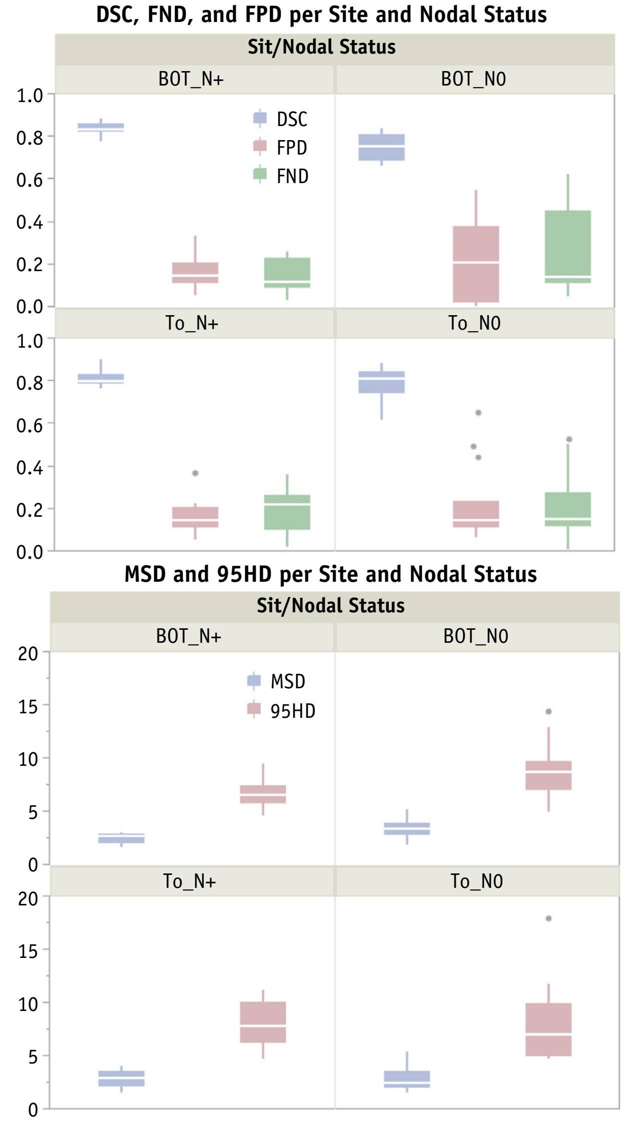

Fig. 7. Volumetric comparison between predicted and manual volumes per disease site and nodal status. (Top) Overlap metrics (Dice similarity coefficient [DSC], false-negative Dice [FND], and false-positive Dice [FPD]) between the 4 disease site and nodal status groups. (Bottom) Comparison between the 4 disease site and nodal status groups according to distance metrics. *Abbreviations:* BOT_N+ = base of tongue node-positive; BOT_N0 = base of tongue node-negative; 95HD = 95th percentile Hausdorff distance; MSD = mean surface distance; To_N+ = tonsil node-positive; To_N0 = tonsil node-negative.

under- and overtreatment. On physician review of all cases with FND and FPD values >0.450, the craniocaudal extent of the DNN-predicted volumes was considered acceptable. The dosimetric effects of using auto-delineated CTVs is difficult to assess without clinical outcomes data. Owing to the high overlap between the predicted and ground-truth volumes, we would expect minimal changes to the normal tissue doses. In a preliminary study (16), 5 radiation oncologists visually inspected a subset of the DNN-predicted volumes as a part of a blinded study. They found that 85% of the auto-delineated and 93% of the ground-truth volumes would be acceptable for clinical use with only minor changes.

Table 2 Volumetric comparison between DNN and ground-truth CTV per tumor site

Variable	DSC	FPD	FND	MSD (mm)	95HD (mm)
ToN0					
Minimum	0.622	0.068	0.010	1.6	4.8
Median	0.814	0.147	0.151	2.5	7.0
Maximum	0.887	0.652	0.531	5.5	17.9
Mean	0.788	0.218	0.206	2.8	7.8
SD	0.076	0.172	0.153	1.1	3.6
ToN+					
Minimum	0.769	0.056	0.025	1.6	4.8
Median	0.803	0.145	0.219	3.0	7.8
Maximum	0.904	0.370	0.364	4.1	11.3
Mean	0.816	0.174	0.194	3.0	8.1
SD	0.037	0.084	0.105	0.8	2.2
BOTN0					
Minimum	0.664	0.004	0.050	1.9	5.0
Median	0.756	0.213	0.142	3.4	8.8
Maximum	0.843	0.549	0.623	5.2	14.4
Mean	0.755	0.234	0.257	3.5	8.8
SD	0.059	0.195	0.210	1.0	2.8
BOTN+					
Minimum	0.781	0.055	0.033	1.7	4.7
Median	0.838	0.149	0.122	2.7	6.6
Maximum	0.887	0.338	0.262	3.1	9.5
Mean	0.840	0.172	0.147	2.5	6.7
SD	0.031	0.080	0.076	0.5	1.3

Abbreviations: 95HD = 95th percentile Hausdorff distance; BOTN+ = base of tongue node-positive; BOTN0 = base of tongue node-negative; CTV = clinical target volume; DNN = auto-delineated contours; DSC = Dice similarity coefficient; FND = false-negative Dice; FPD = false-positive Dice; MSD = mean surface distance; SD = standard deviation; ToN+ = tonsil node-positive; ToN0 = tonsil node-negative.

Very little work has been performed to auto-delineate high-dose CTVs. Belshi et al (29) proposed automating these volumes using a 3D uniform margin expansion from manually contoured GTVs for conformal radiation therapy. However, the introduction of intensity modulated radiation therapy allowed for more complex, conformal, and patient-specific radiation plans, which required more accurate target definition to ensure the tumor is not undertreated and to limit the dose to the surrounding normal tissues. Chao et al (30) showed that using a 1-cm uniform margin expansion from the GTV reduced observer variability when auto-delineating high-risk CTVs for 2 H&N patients. However, their study did not provide an overlap comparison of these volumes to the ground-truth volume (volume used for treatment); thus, the ability of a 1-cm uniform margin to reproduce the physicians' goal was unclear.

Hong et al (2) conducted a survey to investigate differences in CTV delineation among experienced H&N

radiation oncologists and found significant heterogeneity between physician contours and clinical practice. They found that high-risk CTVs had a large standard deviation (43 cm^3). In our analysis, we found that the mean volume difference between the predicted and ground-truth volumes was 1.0 cm^3, just a small fraction of the variability found in the study by Hong et al (2), showing that volume variability can be reduced through auto-delineation. Finally, their study showed that although published guidelines are available to standardize H&N target volume delineation, significant variations between experienced physician contours still exist and standardization of this process is urgently needed. A more recent study by Blinde et al (28) showed in abstract form that they observed a large variability when delineating high-risk CTVs in a group of >20 radiation oncologists. In their preliminary results, they observed volume differences of up to a factor of 8. The lack of standardization in CTV delineation can be problematic for many reasons. The heterogeneity in target design and clinical practice increases the variation in clinical information. Reducing this through standardized target volumes could help produce better quality clinical data. Our results are promising because our approach can be implemented in multiple institutions to improve standardization of radiation therapy, which could, in turn, reduce uncertainties in radiation therapy clinical trials.

An inherent product of the auto-delineation of CTVs is the reduction of physician contouring time. This benefit is increased when planning treatment of H&N tumors because data have suggested that target delineation in this region is comparatively difficult to contour and results in greater interobserver variability than other anatomic sites (8). Hong et al (2) reported that the H&N CTV average contouring time was 102.5 minutes (range 60-210). Although it is unknown how much time was required to delineate the ground-truth CTV alone, it is clear that any reduction in contouring time could benefit the treatment planning workflow. The model we have presented produces high-risk CTVs with a mean time requirement of 10 minutes, with almost 99% of computational time devoted to preparing the inputs before volume prediction. This computationally expensive process could be improved by optimizing the currently used algorithm to compute the inputs using graphic processing units in which voxel-based distance measurements could be calculated in a parallel fashion.

It has been shown that the quality of the radiation therapy plan greatly depends on delineation accuracy and physician experience (31-33). Our institution's H&N service treats ~400 patients annually, and every patient treated undergoes our head and neck planning and development clinic in which the attending physician's CTV contours are peer-reviewed by the H&N group (22). It is our belief that this peer-review process aids in the reduction of interobserver variability and provides high-quality contours for deep learning approaches. Thus, the use of automatically delineated CTVs could help physicians bridge this gap in quality assurance when peer review is not available. A CTV auto-delineation tool could be used to provide physicians with contours before the peer-review sessions at which the radiation oncologists would assess the contours' coverage and make any edits, if necessary, before approval of the target volumes.

Our approach had a few limitations. First, it relied on manual delineation of normal structures and the GTV for which some interobserver variability has been reported. Furthermore, image quality and dental artifacts could affect the accuracy of these segmentations; however, because the manual segmentations used in the present study were reviewed by ≥2 physicians, we believe this peer-review process will provide better quality normal tissue and target volume segmentations. The time-consuming task of manually delineating the structures used in the present study could be overcome by implementing auto-segmentation of these volumes by way of atlas-based segmentation and positron emission tomography-based segmentation of normal structures and the GTV, respectively. This could aid in the reduction of interobserver variability. Second, our patient set had high variability in disease presentation. This variability was reduced by training auto-delineation models using patient data according to disease site and nodal status. However, even for patients within each disease site and nodal status group, secondary sites of disease (eg, tonsil tumor invading the soft palate) could translate into poor predictive performance. Using a larger number of patients and clustering these according to disease extent could improve pattern recognition in physician delineation patterns; however, this remains to be investigated. Finally, all volumes used for training our models were collected from a single institution, and these might not represent the clinical practice at other institutions.

Conclusions

By implementing a DSC-based threshold selection function, our DNN auto-delineation algorithm accurately identified physician patterns to predict clinically acceptable high-risk CTV contours. Our models allowed for the prediction of new volumes within a few minutes and have the potential to greatly reduce physician contouring time. Most of the predicted high-risk CTVs were in close agreement with the physician manual contours and could be implemented clinically with only minor or no changes.

References

1. *International Commission on Radiation Units and Measurements. Prescribing, Recording and Reporting Photon Beam Therapy (Supplement to ICRU Report 50).* Bethesda: International Commission on Radiation Units and Measurements; 1999.
2. Hong TS, Tome WA, Harari PM. Heterogeneity in head and neck IMRT target design and clinical practice. *Radiother Oncol* 2012;103: 92-98.

3. Eminowicz G, Mccormack M. Variability of clinical target volume delineation for definitive radiotherapy in cervix cancer. *Radiother Oncol* 2015;117:542-547.
4. Li XA, Tai A, Arthur DW, et al. Variability of target and normal structure delineation for breast cancer radiotherapy: A RTOG multi-institutional and multiobserver study. *Int J Radiat Oncol Biol Phys* 2009;73:944-951.
5. Lütgendorf-caucig C, Fotina I, Stock M, et al. Feasibility of CBCT-based target and normal structure delineation in prostate cancer radiotherapy: Multiobserver and image multimodality study. *Radiother Oncol* 2011;98:154-161.
6. Lütgendorf-caucig C, Fotina I, Gallops-Evans E, et al. Multicenter evaluation of different target volume delineation concepts in pediatric Hodgkin's lymphoma: A case study. *Strahlenther Onkol* 2012;188: 1025-1030.
7. van Herk M. Errors and margins in radiotherapy. *Semin Radiat Oncol* 2004;14:52-64.
8. Multi-Institutional Target Delineation in Oncology Group. Human-computer interaction in radiotherapy target volume delineation: A prospective, multi-institutional comparison of user input devices. *J Digit Imaging* 2011;24:794-803.
9. Lee N, Xia P, Fischbein NJ, et al. Intensity-modulated radiation therapy for head-and-neck cancer: The UCSF experience focusing on target volume delineation. *Int J Radiat Oncol Biol Phys* 2003;57:49-60.
10. Eisbruch A, Foote RL, Sullivan BO, et al. Intensity-modulated radiation therapy for head and neck cancer: Emphasis on the selection and delineation of the targets. *Semin Radiat Oncol* 2002;12:238-249.
11. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med Phys* 2014;41:050902.
12. Yang J, Zhang Y, Zhang L, et al. Automatic segmentation of parotids from CT scans using multiple atlases. *Med Image Anal Clin A Gd Chall* 2010;2010:323-330.
13. Han X, Hoogeman MS, Levendag PC, et al. Atlas-based auto-segmentation of head and neck CT images. *Med Image Comput Comput Assist Interv* 2008;11(Pt 2):434-441.
14. Cardenas C, Wong A, Mohamed A, et al. Delineating high-dose clinical target volumes for head and neck tumors using machine learning algorithms. *Med Phys* 2016.
15. Cardenas CE, McCarroll R, Court LE, et al. Deep learning on clinically-clustered patients improves auto-delineation of oropharyngeal high-risk clinical target volumes. *Med Phys* 2017;44:3052.
16. Cardenas CE, McCarroll R, Court LE, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. *Med Phys* 2017;44:3160-3161.
17. Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks. *Adv Neural Inf Process Syst* 2007;19:153-160.
18. Olshausen BA, Fieldt DJ. Sparse coding with an overcomplete basis set: Strategy employed by V1? *Vision Res* 1997;37:3311-3325.
19. Meiller MF. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks* 1993;6:525-533.
20. Orozco J, García CAR. Detecting pathologies from infant cry: Applying scaled conjugate gradient neural networks. In: Proceedings of the European Symposium on Artificial Neural Networks. 2003;60: 349−354.
21. Sharma B, Venugopalan PK. Comparison of neural network training functions for hematoma classification in brain CT images. *J Comput Eng* 2014;16:31-35.
22. Cardenas CE, Mohamed ASR, Tao R, et al. Prospective qualitative and quantitative analysis of real-time peer review quality assurance rounds incorporating direct physical examination for head and neck cancer radiation therapy. *Int J Radiat Oncol Biol Phys* 2016;98:532-540.
23. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945;26:297-302.
24. Babalola KO, Patenaude B, Aljabar P, et al. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. *Neuroimage* 2009;47:1435-1447.
25. Hansen CR, Johansen J, Samsøe E, et al. Consequences of introducing geometric GTV to CTV margin expansion in DAHANCA contouring guidelines for head and neck radiotherapy. *Radiother Oncol* 2018;126: 43-47.
26. McCarroll R, Yang J, Cardenas CE, et al. Machine learning for the prediction of physician edits to clinical auto-contours in the head-and-neck. *Med Phys* 2017;44:3160.
27. Awan M, Zafereo M, Lewis CM, et al. Interdisciplinary variation in segmentation of high-risk postoperative tumor volumes in the head and neck. *Int J Radiat Oncol Biol Phys* 2013;87:S584-S585.
28. Blinde S, Mohamed ASR, Newbold K, et al. Large interobserver variation in the international MR-LINAC oropharyngeal carcinoma delineation study. *Int J Radiat Oncol Biol Phys* 2017;99:E639-E640.
29. Belshi R, Pontvert D, Rosenwald J-C, et al. Automatic three-dimensional expansion of structures applied to determination of the clinical target volume in conformal radiotherapy. *Radiat Oncol* 1997; 37:731-736.
30. Chao KS, Hide S, Hen H, et al. Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach. *Int J Radiat Oncol Biol Phys* 2007;68:1512-1521.
31. Boero IJ, Paravati AJ, Xu B, et al. Importance of radiation oncologist experience among patients with head-and-neck cancer treated with intensity-modulated radiation therapy. *J Clin Oncol* 2016;34:684-690.
32. Peters LJ, O'Sullivan B, Giralt J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: Results from TROG 02.02. *J Clin Oncol* 2010;28:2996-3001.
33. Wuthrick EJ, Zhang Q, Machtay M, et al. Institutional clinical trial accrual volume and survival of patients with head and neck cancer. *J Clin Oncol* 2015;33:156-164.