Elsevier Editorial System(tm) for Radiotherapy and Oncology

Manuscript Draft

Manuscript Number:

Title: Benefits of deep learning for delineation of organs at risk in head and neck cancer

Article Type: Full Length Article

Section/Category: Head & Neck Cancers

Keywords: Head and Neck neoplasms, Radiotherapy, Organs at Risk, Observer Variation, Neural Networks (computer), Delineation

Corresponding Author: Dr. Julie van der Veen, M.D.

Corresponding Author's Institution:

First Author: Julie van der Veen, M.D.

Order of Authors: Julie van der Veen, M.D.; Siri Willems; Sarah Deschuymer; David Robben; Wouter Crijns; Frederik Maes; Sandra Nuyts

Abstract: Purpose/objective: Precise delineation of organs at risk (OARs) in head and neck cancer (HNC) is necessary for accurate radiotherapy. Although guidelines exist, significant interobserver variability (IOV) remains. The aim was to validate a 3D convolutional neural network (CNN) for semi-automated delineation of OARs with respect to delineation accuracy, efficiency and consistency compared to manual delineation.

Material/Methods: 16 OARs were manually delineated in 15 new HNC patients by two trained radiation oncologists (RO) independently, using international consensus guidelines. OARs were also automatically delineated by applying the CNN and corrected as needed by both ROs separately. Both delineations were performed two weeks apart and blinded to each other. IOV between both ROs was quantified using Dice similarity coefficient (DSC) and average symmetric surface distance (ASSD). To objectify network accuracy, differences between automated and corrected delineations were calculated using the same similarity measures.

Results: Average correction time of the automated delineation was 33% shorter than manual delineation (23 vs 34 minutes)($p<10-6$). IOV improved significantly with network initialisation for nearly all OARs ($p<0.05$), resulting in decreased ASSD averaged over all OARs from 1.9 to 1.2 mm. The network achieved an accuracy of 90% and 84% DSC averaged over all OARs for RO1 and RO2 respectively, with an ASSD of 0.7 and 1.5 mm, which was in 93% and 73% of the cases lower than the IOV.

Conclusion: The CNN developed for automated OAR delineation in HNC was shown to be more efficient and consistent compared to manual delineation, which justify its implementation in clinical practice.

Cover Letter

Prof dr. Michael Baumann
Editor-in-Chief
Radiotherapy and Oncology Editorial Office
Deutsches Krebsforschungszentrum (DKFZ)
Im Neuenheimer Feld 280, 69120, Heidelberg, Germany
(E-mail: RadioTherOncol@dkfz.de)

February 14, 2019

Dear Editor-in-Chief,

We are pleased to submit our original research paper entitled "Benefits of deep learning for delineation of organs at risk in head and neck cancer" by Julie van der Veen, Siri Willems, Sarah Deschuymer, David Robben, Wouter Crijns, Frederik Maes, and Sandra Nuyts for consideration for publication in *Radiotherapy and Oncology.*

In this manuscript, we validated the use of a convolutional neural network for delineation of organs at risk in head and neck cancer. This network was based on manual delineation of 16 organs at risk in 70 patients, using international consensus guidelines of Brouwer et al. (2015) and Christianen et al. (2011). We showed that this network performs well, creating delineations that only need small corrections before they can be used for treatment planning. In addition to this, delineation of organs at risk is more time efficient and results in less interobserver variability between experiences radiation oncologists.

We believe that this manuscript is appropriate for publication by *Radiotherapy and Oncology* because the guidelines used for creating the network were also published in the Green Journal. Artificial intelligence is a hot topic and here we show its clinical implications. We anticipate that this manuscript is not only timely, but will also appeal to a wide readership.

This manuscript has not been published and is not under consideration for publication elsewhere and we declare that it comprises original unpublished work. All authors have agreed with submission in its present form. All authors declare that they have no competing interests.

If you feel that the manuscript is appropriate for your journal, we suggest professor dr. Jesper Eriksen (Aarhus University Hospital) because of his expertise in radiotherapy and head and neck cancer and his interest in automated delineation.

Thank you for your consideration. We look forward to your response. Please do not hesitate to contact us with any further questions.

On behalf of all authors,

Prof.dr. Nuyts Sandra
Department of Radiation Oncology
University Hospitals Leveun

sandra.nuyts@uzleuven.be
Tel secr. +32 16 34 76 00
UZ Leuven | campus Gasthuisberg | Herestraat 49 | B - 3000 Leuven | www.uzleuven.be

Benefits of deep learning for delineation of organs at risk in head and neck cancer

Van der Veen J*[1], Willems S.*[2], Deschuymer S.[1], Robben D.[2, 3], Crijns W.[1], Maes F.[2], Nuyts S.[1α]

[1]KU Leuven, Dept. Oncology, Laboratory of Experimental Radiotherapy, & UZ Leuven, Radiation Oncology, B-3000 Leuven, Belgium
[2] KU Leuven, Dept. ESAT, Processing Speech and Images (PSI), & UZ Leuven, Medical Imaging Research Center, B-3000 Leuven, Belgium
[3] Icometrix, B-3000 Leuven, Belgium

** These authors contributed equally to this work*
[α] Corresponding author at: KU Leuven, Dept. Oncology, Laboratory of Experimental Radiotherapy, & UZ Leuven, Radiation Oncology, B-3000 Leuven, Belgium. E-mail address: sandra.nuyts@uzleuven.be

Abstract

Purpose/objective: Precise delineation of organs at risk (OARs) in head and neck cancer (HNC) is necessary for accurate radiotherapy. Although guidelines exist, significant interobserver variability (IOV) remains. The aim was to validate a 3D convolutional neural network (CNN) for semi-automated delineation of OARs with respect to delineation accuracy, efficiency and consistency compared to manual delineation.

Material/Methods: 16 OARs were manually delineated in 15 new HNC patients by two trained radiation oncologists (RO) independently, using international consensus guidelines. OARs were also automatically delineated by applying the CNN and corrected as needed by both ROs separately. Both delineations were performed two weeks apart and blinded to each other. IOV between both ROs was quantified using Dice similarity coefficient (DSC) and average symmetric surface distance (ASSD). To objectify network accuracy, differences between automated and corrected delineations were calculated using the same similarity measures.

Results: Average correction time of the automated delineation was 33% shorter than manual delineation (23 vs 34 minutes)($p<10-6$). IOV improved significantly with network initialisation for nearly all OARs ($p<0.05$), resulting in decreased ASSD averaged over all OARs from 1.9 to 1.2 mm. The network achieved an accuracy of 90% and 84% DSC averaged over all OARs for RO1 and RO2 respectively, with an ASSD of 0.7 and 1.5 mm, which was in 93% and 73% of the cases lower than the IOV.

Conclusion: The CNN developed for automated OAR delineation in HNC was shown to be more efficient and consistent compared to manual delineation, which justify its implementation in clinical practice.

Introduction

Ranked as the seventh most common cancer and cause of cancer death worldwide, the burden of head and neck cancer (HNC) on society and the health sector should not be underestimated [1]. Radiotherapy (RT) plays an important role in the curative treatment of HNC and allows organ preservation and improved function preservation in selected cases, compared to surgery. Intensification of RT regimens by means of altered fractionation and concomitant chemotherapy have been beneficial for overall survival and loco-regional control [2,3], although disease recurrence remains an issue [4,5]. At the same time, intensification of RT regimens has induced an increase in acute and late toxicity, limiting further treatment intensification [6]. To compensate for this, implementation of intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) have allowed the dose delivered to the tumour to be shaped, resulting in better sparing of normal and critical tissue, decreasing toxicity [7]. Proton therapy has the potential to spare organs at risk (OARs) even more due to its more favourable dose-depth characteristics, including a sharp localized high dose delivery at the Bragg peak, with a low exit dose [8]. Adaptive radiotherapy, with the intention of sparing normal tissue even better and still provide sufficient coverage of target volumes (TVs) is also finding its way to RT centres worldwide.

The drawback of these newer techniques however, is that delineation is time consuming due to the complex head and neck anatomy. Delineation of OARs alone can take up to one hour, and of TVs up to two hours [9], which will be more for unexperienced radiation oncologists (ROs). On top of that, correct delineation is essential for optimal treatment planning and is one of the weakest links in the chain of actions needed to treat a patient with RT. Firstly, it is mainly performed manually and therefore susceptible to intra- and inter-observer variability (IOV) [10]. Secondly, variations in OAR delineation may influence the treatment plan, including the dose to OARs, which can also impact results of multicentre trials [11–13]. Thirdly, delineation errors remain present during the entire RT course so their impact can be larger than expected. IOV in TV and OAR delineation affect quality of RT, treatment outcomes and evaluation of clinical research [14,15]. The introduction of delineation guidelines for OARs has improved IOV [16], although Brouwer et al. [10] showed that there was still room for improvement.

In previous research, automated segmentation using machine learning approaches has been widely investigated to overcome drawbacks of manual segmentation procedures in medical imaging [17]. Available algorithms in current RT software are mainly atlas-based methods, which incorporate prior knowledge in the form of atlases and are registered to the daily images using deformable image registration (DIR). In particular for HNC patients, atlas-based models achieved acceptable results for segmentation of OARs [18,19]. Recently, deep learning approaches based on convolutional neural networks are gaining popularity thanks to their successes in many segmentation tasks in medical imaging [20,21], including RT [22–25].

Precise delineation of OARs in HNC is necessary for accurate radiotherapy treatment planning, correct interpretation of dose volume histograms (DVH) and reduction of therapeutic variability. The aim of this study was to evaluate the potential of a 3D convolutional neural network (CNN) for automated delineation of OARs most commonly delineated in HNC patients, which could significantly reduce delineation time and the burden of human intervention and IOV. The clinical implementation of the validated automated delineation tools could eventually result in a shorter interval between simulation and start of RT, affect treatment capacity and facilitate paradigm shifts such as online adaptive planning.

Methods

Data acquisition

Patients were recruited between August and November 2018 and included consecutive patients with a newly diagnosed HNC, scheduled for RT and without total laryngectomy. In total, 15 patients were included in the study (see Table 1 for patient characteristics). Each patient underwent a contrast-enhanced planning CT scan in the supine position with custom thermoplastic mask for immobilization, according to the conventional clinical protocol. The CT images were made on a multidetector-row spiral CT scanner (Somatom Sensation Open, 40 slice configuration; Siemens Medical Solutions, Erlangen, Germany). The acquisition parameters were: 120 kvP / 230 mAs (quality reference mAs with CARE Dose4D), no gantry tilt, spiral mode, rotation time 1s, 40 detector rows at 0.6mm intervals, table speed 21.6mm/rotation (pitch=0.9), reconstruction interval 3mm using Kernel B30s medium smooth, matrix size 512×512, pixel spacing 0.97×0.97mm.

3D convolutional neural network

A 3D CNN was previously developed and trained for automated delineation of 16 OARs in planning CT images of HNC patients based on a training set of 70 cases delineated according to international consensus guidelines [26,27] (see appendix A for more details). This CNN was applied to all 15 images in this study. The computation time needed by the CNN to automatically delineate all 16 OARs was about 3 minutes per image [28]. We refer to the original, unmodified delineations generated by the CNN as "automated delineations" further on.

Study design

Two ROs (JV and SD), well trained in delineation of OARs for HNC RT and familiar with the delineation guidelines [26,27], each delineated the 16 OARs on all 15 CT scans twice using Eclipse (Varian Medical Systems, Palo Alto, CA) in 2 separate, uninterrupted sessions for each patient: once manually ("manual delineations") and once by modifying and correcting the presented automated delineations generated by the CNN ("corrected delineations") (Figure 1). The 2 delineation sessions by the same RO for the same patient were performed with an average interval of 15.5 days, with manual delineations being performed in the 1st session for about half of the cases and in the 2nd session for the other half, and blinded for any other delineation result to avoid observer bias. All delineations were verified and approved without modification by a third expert in HNC RT (SN) to ensure their clinical validity.

Validation

The benefits of the use of a CNN based automated delineation tool in clinical practice were assessed in terms of its accuracy, impact on IOV and time efficiency.

Accuracy

The accuracy of the automated delineation tool was assessed for each 3D OAR separately by comparing it to the corrected delineations using the Dice Similarity Coefficient (DSC) and the Average Symmetric Surface Distance (ASSD). DSC is a measure for the overlap between two delineations *A* and *B*, yielding a value of 1 in case of perfect overlap and a value of 0 if no overlap:

$$DSC = 2 * \frac{|A \cap B|}{|A| + |B|}$$

with $|A|$ and $|B|$ the volumes of each delineation and $|A \cap B|$ the volume of their intersection. ASSD represents the mean distance between two delineations *A* and *B* in mm:

$$ASSD\ (A,B) = \frac{h(A,B) + h(B,A)}{2}$$

$$h(A,B) = \underset{a \in A}{\text{mean}} \{\min_{b \in B}\{d(a,b)\}\}$$

with $d(a,b)$ the 3D distance between point a on delineation A and point b on delineation B. Both DSC and ASSD provide an indication for the amount of corrections necessary for clinical approval.

Variability

The impact of the use of the automated delineation tool on IOV between different observers was assessed for each 3D OAR separately by computing DSC and ASSD between the manual delineations and between the corrected delineations made by both ROs (with larger DSC and smaller ASSD indicating less IOV). In addition, IOV for the same observer was assessed by comparing manual delineations and corrected delineations made by the same RO.

Efficiency

The efficiency of the automated delineation tool was quantified by comparing the time needed for manual delineation to the time needed for correction of the automated delineations. Both ROs recorded the total delineation time per patient for each of the 2 delineation sessions. This included the time for adjusting window settings, navigating between slices and creating or correcting all delineations for all 16 OARs.

Statistical Analysis

Statistically significant differences for DSC and ASSD were assessed with a two-sided, paired Wilcoxon signed rank test, using significance level α=0.05. To asses reduction in delineation time, a one-sided, Wilcoxon signed rank test was used, using significance level α=0.05.

Results

Accuracy

Examples of manual and automated delineations are shown in Figure 2. Mean DSC and ASSD for automated versus corrected delineations for each OAR are summarised in Table 2. Based on DSC, the network performed best for brainstem, left cochlea, mandible, parotid glands, submandibular glands and spinal cord (DSC>90%). The average corrections necessary for clinical acceptance were below 1 mm for cochleae, glottic larynx, mandible, right parotid gland, middle pharyngeal constrictor muscle (PCM) and right submandibular gland (ASSD<1mm). Average corrections for all other structures were below 2 mm.

Figure 3 plots for each 3D OAR separately the extent of the corrections made to the automated delineations versus IOV of the manual delineations, both quantified by ASSD, which shows that the accuracy of the automated delineation tool was better than the manual delineation variability for most OAR (grey shaded area; RO1: 93% of OARs; RO2: 72%).

Observer variability

Table 3 represents IOV (intra and inter) as assessed by DSC and ASSD. IOV for the corrected delineations is significantly lower than IOV for the manual delineations for almost all OARs. The left cochlea did not show a significant reduction in IOV with DSC and both cochleae showed no improvement with ASSD. The oral cavity and brainstem showed no significant improvement in DSC but ASSD reduced significantly. For most OARs, IOV for the corrected delineations is smaller than intra-observer variability.

Time

The time needed to correct the automated delineations was for both ROs significantly shorter than the time needed for manual delineation (RO1: 17 vs 30 minutes; RO2: 27 vs 38 minutes; mean: 23 vs 34 minutes, $p<10^{-6}$). Manual delineation time per patient ranged from 22 to 44 minutes, correction time ranged from 13 to 33 minutes, and time gain ranged from 6 to 19 minutes (33% on average).

Discussion

The aim of this study was to evaluate the benefits for clinical practice of the use of an automated delineation tool with respect to delineation accuracy, efficiency and reduction of IOV. Automated delineation was performed by a 3D CNN, which was trained using manual expert delineations of 16 OARS of HNC patients in agreement with international consensus guidelines of Brouwer et al. [26] and Christianen et al. [27].

Increased delineation efficiency by the use of the tool was demonstrated by the reduction in the time needed to correct the automated delineations versus the time needed for manual delineation. Even though one of the ROs was faster in both manual delineation and correction of automated delineations, both ROs delineated faster using the automated delineations. The average time needed per patient to verify and correct the complete 3D delineations for all 16 OARs together was 23 minutes, but could be as short as 13 minutes, i.e. less than 1 minute per 3D structure.

The accuracy of the CNN was examined by comparing the automated delineations to the corrected delineations. Table 2 shows that overall only small corrections needed to be made for the delineations to be clinically acceptable (ASSD<1.1 mm on average). Figure 3 shows that for most OARs, the differences between the automated and the corrected delineations were smaller than interobserver variability obtained with manual delineation, i.e. the delineation variability as typically observed in clinical practice. Two sets of outliers can be observed in Figure 3. A first set of outliers, in the top left corner, involves few automated delineations that needed relatively large corrections (between 4 and 6 mm on average) for some individual OARs, namely one left-sided submandibular gland (1), one left parotid gland (2), one inferior PCM (3) and one supraglottic larynx (4). For outliers (1) and (2), this was due to nearby enlarged lymph nodes, which were mistakenly delineated in the same volume as the OARs. For outlier (3), this patient had a low supraglottic larynx tumour originating in the left ary-epiglottic fold. Outlier (4) was a patient with an oropharyngeal tumour, invading the base of tongue, which made delineation of the cranial part of the supraglottic larynx challenging. The CNN thus sporadically showed some difficulty with adenopathies or primary tumours near OARs. Scatter due to dental fillings did not impact the accuracy as illustrated in Figure 2, likely because such artefacts were also present in some of the images in the training set.

A second set of outliers, in the bottom right corner of Figure 3, involves the spinal cord, which shows a large IOV between the manual delineations and large differences in the corrections made by RO2. Upon inspection of the contours, it was found that different cranial and caudal borders were used by the ROs. According to the guidelines, the cranial border of the spinal cord should be the tip of the dens of C2, but the ROs differed on the tip location from 1-3 CT slices (3-6 mm difference). The caudal border of the spinal cord showed even more variation between ROs, from the bottom of the CT scan to the cranial border of T3 (30-93 mm difference). According to the guidelines, the caudal border of the spinal cord should reach to the upper edge of T3, although for caudal tumours it should reach 5 cm under the planning target volume (PTV). To accommodate for these different cases, the CNN was trained to delineate the spinal cord to the most caudal slice of the planning CT scan. One of the ROs systematically corrected the automated delineations with respect to the caudal border of the spinal cord, while the other did not. On axial planes, however, no differences in spinal cord delineations were observed, as illustrated in Figure 2.

As shown in Table 3, IOV as quantified by DSC and ASSD reduced significantly for most OARs by using automated delineations. ASSD decreased significantly for all OARs, except for both cochlea, which are very small structures (~ 1 ml) for which DSC and ASSD are very susceptible to small delineation differences. DSC did not improve significantly for brainstem and oral cavity, while ASSD did. This is because DSC is volume dependent and for these large OARs only large delineation differences will impact DSC. Moreover, IOV between manual delineations was already small for oral cavity when measured with DSC (94% on average) and an improvement in DSC to 96% using the automated delineations was therefore not significant. The spinal cord unexpectedly showed one of the worst DSC and ASSD results compared to the other OARs. This is mainly due to the difference in caudal border chosen by one of the ROs as already explained above, resulting in an underestimation of the benefit of automated delineation on IOV for this structure. However, this difference is not clinically relevant: even though it has an influence on DVH, it has no impact on plan creation, evaluation and acceptance because the maximum (D_{max}) and not the average dose (D_{mean}) to the spinal cord is taken into account. In a serial OAR, like the spinal cord, loss of function in one part will cause the entire organ to stop functioning. A high dose to a small volume can cause serious toxicity and therefore the risk of damage is dominated by D_{max}. In a parallel OAR like the salivary glands, loss of function in one part of the OAR can be compensated by an unaffected part. Therefore, there is a threshold volume effect and the risk of injury, in this case resulting in xerostomia, is dominated by the D_{mean} over the whole OAR. Mucosa, like that of the oral cavity, is neither serial nor parallel, but behaves clinically like a parallel OAR, as desquamation of a large area of mucosa is more problematic than a small area [29]. The requirements and importance of correct OAR delineation and its impact on treatment planning thus depends on the type of OAR.

Automated delineation of OARs on HNC planning CT scans using a CNN was previously investigated by Ibragimov et al. [21]. They concluded that their method performed better or comparable to state-of-the-art algorithms and commercial software for spinal cord, mandible, larynx and pharynx, and inferior for parotid- and submandibular glands. The average DSCs they found ranged between 69% and 90%, for pharynx and mandible respectively. Our CNN outperformed their results for all OARs, except for spinal cord (DSC 87% vs 76%). However, Ibragimov et al. did not delineate the spinal cord itself but the spinal canal, which is the bony structure through which the spinal cord runs and which is easier to detect and delineate than the spinal cord. Moreover, their DSC values were only calculated on the length of the spinal cord that could be affected during treatment, whereas in our study the entire delineated structure was considered.

The main strength of our study is the use of international consensus guidelines [26,27] to train the CNN to delineate a large number of OARs (16), including the different PCMs and laryngeal sub-volumes. In case delineation guidelines were to be modified in the future, the CNN could be easily retrained to adapt to these changes. A possible limitation is that the contours used to train the CNN in this study were delineated by only one RO, possibly introducing some observer bias despite following consensus guidelines. Nevertheless, the use of the automated delineation tool was shown to result in a shorter delineation time for both ROs in this study and in less IOV between them.

The implementation of this tool in clinical practice could be especially beneficial to ROs in training to reduce delineation time, and facilitate the recognition of OARs, thus benefitting training and resulting in a steeper learning curve. Moreover, with more efficient OAR delineation, more time could be spent on other aspects of RT, such as delineation of TVs and clinical follow-up of patients. Automated delineation is also very relevant for adaptive RT regimens. When patient anatomy or tumour volume change during treatment, re-contouring is very labour-intensive and automated delineation could make this process significantly more efficient. Although with photon therapy not

all HNC patients are eligible for adaptive RT, the need for adaptive RT will presumably be higher for proton therapy [30,31].

Conclusion

Validation in clinical setting of a CNN trained for automated delineation of OARs in HNC patients based on international consensus guidelines showed that automated delineation is not only significantly more efficient than manual delineation, but also reduces IOV. The automated delineations mainly require only minor corrections to be approved for treatment planning.

Conflicts of interest

None

Acknowledgments

Siri Willems is supported by a Ph.D. fellowship of the research foundation – Flanders (FWO). David Robben is supported by an innovation mandate of Flanders Innovation & Entrepreneurship (VLAIO).

[1] Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015. JAMA Oncol 2017;3:524–48. doi:10.1001/jamaoncol.2016.5688.

[2] Pignon J-P, Maître A le, Maillard E, Bourhis J. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): An update on 93 randomised trials and 17,346 patients. Radiother Oncol 2009;92:4–14. doi:10.1016/j.radonc.2009.04.014.

[3] Bourhis J, Overgaard J, Audry H, Ang KK, Saunders M, Bernier J, et al. Hyperfractionated or accelerated radiotherapy in head and neck cancer: a meta-analysis. Lancet 2006;368:843–54. doi:10.1016/S0140-6736(06)69121-6.

[4] Due AK, Vogelius IR, Aznar MC, Bentzen SM, Berthelsen AK, Korreman SS, et al. Recurrences after intensity modulated radiotherapy for head and neck squamous cell carcinoma more likely to originate from regions with high baseline [18F]-FDG uptake. Radiother Oncol 2014;111:360–5. doi:10.1016/j.radonc.2014.06.001.

[5] Bayman E, Prestwich RJD, Speight R, Aspin L, Garratt L, Wilson S, et al. Patterns of Failure after Intensity-modulated Radiotherapy in Head and Neck Squamous Cell Carcinoma using Compartmental Clinical Target Volume Delineation. Clin Oncol 2014;26:636–42. doi:10.1016/j.clon.2014.05.001.

[6] Nuyts S, Dirix P, Clement PMJ, Poorten V Vander, Delaere P, Schoenaers J, et al. Impact of Adding Concomitant Chemotherapy to Hyperfractionated Accelerated Radiotherapy for Advanced Head-and-Neck Squamous Cell Carcinoma. Int J Radiat Oncol Biol Phys 2009;73:1088–95. doi:10.1016/j.ijrobp.2008.05.042.

[7] Nutting CM, Morden JP, Harrington KJ, Urbano TG, Bhide SA, Clark C, et al. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial. Lancet Oncol 2011;12:127–36. doi:10.1016/S1470-2045(10)70290-4.

[8] Holliday EB, Frank SJ. Proton Radiation Therapy for Head and Neck Cancer: A Review of the Clinical Experience to Date. Int J Radiat Oncol 2014;89:292–302. doi:10.1016/J.IJROBP.2014.02.029.

[9] Harari PM, Song S, Tomé WA. Emphasizing Conformal Avoidance Versus Target Definition for IMRT Planning in Head-and-Neck Cancer. Int J Radiat Oncol Biol Phys 2010;77:950–8. doi:10.1016/j.ijrobp.2009.09.062.

[10] Brouwer CL, Steenbakkers RJ, van den Heuvel E, Duppen JC, Navran A, Bijl HP, et al. 3D

Variation in delineation of head and neck organs at risk. Radiat Oncol 2012;7:32. doi:10.1186/1748-717X-7-32.

[11] Nelms BE, Tomé WA, Robinson G, Wheeler J. Variations in the Contouring of Organs at Risk: Test Case From a Patient With Oropharyngeal Cancer. Int J Radiat Oncol 2012;82:368–78. doi:10.1016/j.ijrobp.2010.10.019.

[12] Dirix P, Nuyts S. Evidence-based organ-sparing radiotherapy in head and neck cancer. Lancet Oncol 2010;11:85–91. doi:10.1016/S1470-2045(09)70231-1.

[13] Peng Y lin, Chen L, Shen G zhu, Li Y ning, Yao J jin, Xiao W wei, et al. Interobserver variations in the delineation of target volumes and organs at risk and their impact on dose distribution in intensity-modulated radiation therapy for nasopharyngeal carcinoma. Oral Oncol 2018;82:1–7. doi:10.1016/j.oraloncology.2018.04.025.

[14] Gwynne S, Spezi E, Sebag-Montefiore D, Mukherjee S, Miles E, Conibear J, et al. Improving radiotherapy quality assurance in clinical trials: Assessment of target volume delineation of the pre-accrual benchmark case. Br J Radiol 2013;86:1–11. doi:10.1259/bjr.20120398.

[15] Piotrowski T, Gintowt K, Jodda A, Ryczkowski A, Bandyk W, Ba¸k B, et al. Impact of the Intra- and Inter-observer Variability in the Delineation of Parotid Glands on the Dose Calculation During Head and Neck Helical Tomotherapy. Technol Cancer Res Treat 2014:tcrtexpress.201. doi:10.7785/tcrtexpress.2013.600278.

[16] Yi SK, Hall WH, Mathai M, Dublin AB, Gupta V, Purdy JA, et al. Validating the RTOG-endorsed brachial plexus contouring atlas: An evaluation of reproducibility among patients treated by intensity-modulated radiotherapy for head-and-neck cancer. Int J Radiat Oncol Biol Phys 2012;82:1060–4. doi:10.1016/j.ijrobp.2010.10.035.

[17] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. doi:10.1016/j.media.2017.07.005.

[18] Fortunati V, Verhaart RF, Van Der Lijn F, Niessen WJ, Veenland JF, Paulides MM, et al. Tissue segmentation of head and neck CT images for treatment planning: A multiatlas approach combined with intensity modeling. Med Phys 2013;40. doi:10.1118/1.4810971.

[19] Tao C, Yi J, Chen N, Ren W, Cheng J, Tung S, et al. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma : A multi-institution clinical study. Radiother Oncol 2015;115:407–11. doi:10.1016/j.radonc.2015.05.012.

[20] Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 2017;36:61–78. doi:10.1016/j.media.2016.10.004.

[21] Olaf Ronneberger, Philipp Fischer and TB, Computer. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2015;9351:234–41. doi:10.1007/978-3-319-24574-4.

[22] Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J, et al. Deep Deconvolutional Neural Network for Target Segmentation of Nasopharyngeal Cancer in Planning Computed Tomography Images. Front Oncol 2017;7:1–9. doi:10.3389/fonc.2017.00315.

[23] Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks: Med Phys 2017;44:6377–89. doi:10.1002/mp.12602.

[24] Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys 2017;44:547–57.

[25] Nikolov S, Blackwell S, Mendes R, De Fauw J, Meyer C, Hughes C, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy 2018:1–31. doi:arXiv:1809.04430v1.

[26] Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC,

HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiother Oncol 2015;117:83–90. doi:10.1016/j.radonc.2015.07.041.

[27] Christianen MEMC, Langendijk JA, Westerlaan HE, Water TA Van De, Bijl HP. Delineation of organs at risk involved in swallowing for radiotherapy treatment planning. Radiother Oncol 2011;101:394–402. doi:10.1016/j.radonc.2011.05.015.

[28] Willems S, Crijns W, La Greca Saint-Esteven A, Van Der Veen J, Robben D, Depuydt T, et al. Clinical Implementation of DeepVoxNet for Auto-Delineation of Organs at Risk in Head and Neck Cancer Patients in Radiotherapy, 2018, p. 223–32. doi:10.1007/978-3-030-01201-4_24.

[29] Chang DS, Lasley FD, Das IJ, Mendonca MS, Dynlacht JR. Normal Tissue Radiation Responses. Basic Radiother. Phys. Biol., Cham: Springer International Publishing; 2014, p. 265–75. doi:10.1007/978-3-319-06841-1_26.

[30] Brouwer CL, Steenbakkers RJHM, Schaaf A Van Der, Sopacua CTC, Dijk LV Van, Kierkels RGJ, et al. Selection of head and neck cancer patients for adaptive radiotherapy to decrease xerostomia. Radiother Oncol 2016;120:36–40. doi:10.1016/j.radonc.2016.05.025.

[31] Brown E, Owen R, Harden F, Mengersen K, Oestreich K, Houghton W, et al. Predicting the need for adaptive radiotherapy in head and neck cancer. Radiother Oncol 2015;116:57–63. doi:10.1016/j.radonc.2015.06.025.

Conflicts of interest statement

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Table 1

Table 1: Tumour site and TNM staging for the 15 HNC patients in the study.

Patient	Site	T	N	p16
1	oropharynx	4a	3b	NS
2	parotid left (postoperative)	4a	2b	
3	supraglottic	2	3b	
4	oropharynx	1	1	+
5	oropharynx + hypopharynx	2 / 4	0 / 2b	-
6	oral cavity	1	2b	
7	oropharynx	2	1	-
8	oropharynx	3	2	+
9	oral cavity (postoperative)	2	N2c	
10	oropharynx (postoperative)	2	1	+
11	hypopharynx	2	1	
12	oropharynx	1	1	+
13	oropharynx	4	2	+
14	larynx	4a	0	
15	larynx	3	0	

Abbreviations: TNM: tumour staging according to the TNM-8 staging system (2017); T: clinical tumour stage; N: clinical nodal stage; p16: p16 protein expression, correlated with human papilloma virus status; NS: not specified

Table 2

Table 2: Evaluation of accuracy of CNN based automated OAR delineation as perceived by each observer (Acc1, Acc2). The automated delineation is compared to the corrected delineation made by each observer (RO1, RO2) for each 3D OAR separately by computing their DSC and ASSD. Volumes for manual, automated and corrected OAR delineations are reported as well. All values are reported as mean ± STD for all patients (n=15) and for both observers for manual and corrected volumes.

	Acc1 (RO1)		Acc2 (RO2)		Volume (ml)		
OARs	**DSC (%)**	**ASSD (mm)**	**DSC (%)**	**ASSD (mm)**	**Manual**	**Automated**	**Corrected**
Brainstem	94.9±2.1	0.7±0.2	98.1±1.6	1.7±0.6	228.8±44.4	263.1±32.4	231.1±41.7
Cochlea left	98.7±1.1	0.1±0.1	95.1±7.2	1.1±0.5	0.8±0.2	1.6±0.6	1.2±0.4
Cochlea right	96.4±8.4	0.2±0.2	80.0±25.8	1.1±0.4	1.0±0.3	1.4±0.7	1.2±0.4
U Esophagus	83.1±17.2	1.3±0.9	79.3±17.9	2.0±1.3	49.4±17.8	49.8±17.5	49.4±14.7
Glottic larynx	76.8±16.1	0.7±0.6	64.4±14.7	1.3±0.6	38.0±15.9	26.6±11.4	35.8±15.7
Mandible	98.8±0.6	0.2±0.1	91.2±2.2	0.7±0.2	620.2±106.2	596.5±116.7	619.3±117.2
Oral cavity	93.1±7.3	1.3±0.8	85.4±10.1	2.1±1.0	1040.0±196.3	1025.3±234.0	1087.8±223.6
S Glottic larynx	74.2±23.7	1.5±1.1	67.1±23.5	2.1±1.3	160.6±49.5	124.2±47.9	156.8±45.0
PG left	96.1±2.9	0.9±1.1	93.4±5.3	1.3±1.3	276.9±86.1	275.5±81.1	276.5±86.5
PG right	96.1±3.3	0.5±0.3	92.9±5.2	0.9±0.5	303.1±104.0	281.7±88.4	290.3±95.7
PCM inferior	80.1±23.7	1.3±1.3	71.3±21.0	1.7±1.1	43.8±9.7	41.7±15.9	44.4±13.3
PCM middle	84.1±1.0	0.6±0.3	76.4±10.0	1.1±0.4	50.3±17.9	43.2±13.8	48.2±16.7
PCM superior	82.9±15.4	0.8±0.5	75.5±13.0	1.2±0.4	84.4±25.2	73.5±26.4	80.3±21.0
SG left	96.5±4.9	1.0±1.2	91.4±8.3	1.4±1.2	91.1±28.5	90.3±41.0	87.7±27.9
SG right	98.2±2.7	0.3±0.2	95.1±7.0	0.6±0.6	92.7±31.5	77.2±31.3	85.4±28.4
Spinal cord	97.8±3.1	0.2±0.2	92.8±3.6	3.6±2.3	136.3±23.7	166.8±32.8	152.2±28.5

Abbreviations: OARs: organs at risk; DSC: Dice similarity coefficient; ASSD: average symmetric surface distance; STD: standard deviation; RO: radiation oncologist; PCM: pharyngeal constrictor muscles; PG: parotid gland; SG: submandibular gland; U: upper; S: supra.

Table 3

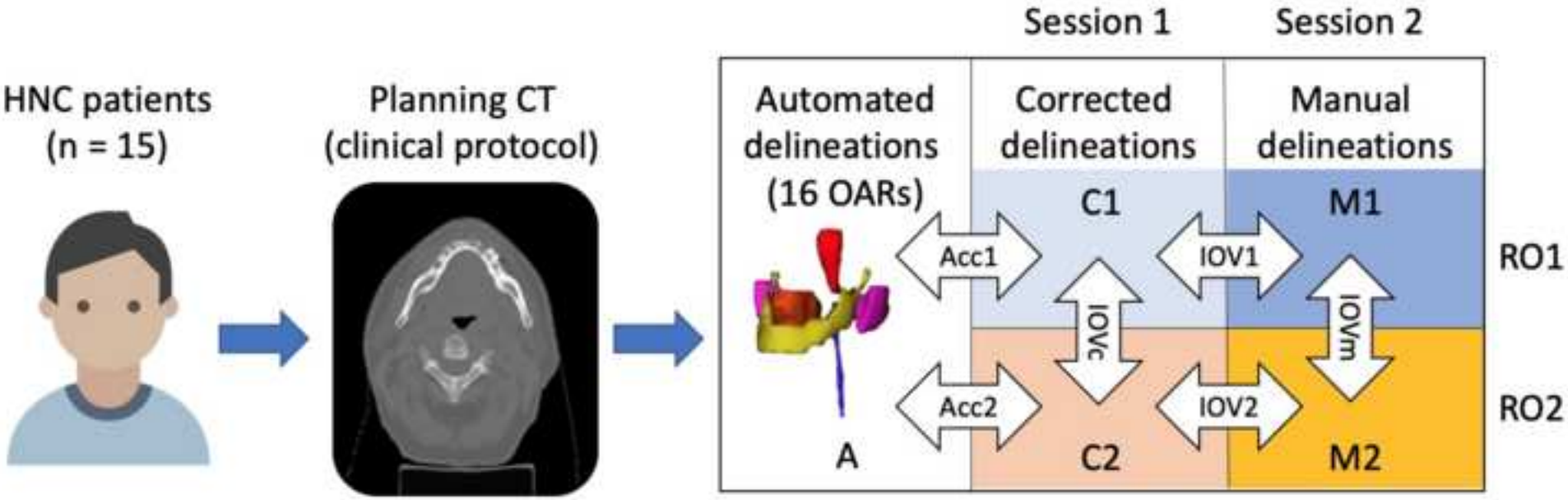
Table 3: Evaluation of intra- and inter-observer variability between the manual and the corrected automated OAR delineations. Intra-observer variability (IOV1, IOV2) is assessed by comparing the manual to the corrected delineations for each 3D OAR for either observer separately (RO1, RO2) using DSC and ASSD. Inter-observer variability (IOVm, IOVc) is assessed by comparing the delineations of RO1 to those of RO2 for each 3D OAR for the manual and corrected delineations separately. All values are reported as mean ± STD for all patients (n=15). Statistically significant differences ($p < 0.05$) in inter-observer variability for the corrected versus the manual delineations (IOVc vs IOVm) are indicated in bold.

	Intra (manual vs corrected)				**Inter (RO1 vs RO2)**			
	IOV1 (RO1)		**IOV2 (RO2)**		**IOVm (manual)**		**IOVc (corrected)**	
	DSC (%)	**ASSD (mm)**	**DSC (%)**	**ASSD (mm)**	**DSC (%)**	**ASSD (mm)**	**DSC (%)**	**ASSD (mm)**
Brainstem	90.0±4.1	1.6±0.2	90.4±3.8	1.4±0.4	68.1±11.7	2.2±0.5	70.8±11.9	**1.6±0.6**
Cochlea left	57.5±20.0	1.2±0.4	71.1±17.9	0.7±0.3	48.7±13.4	1.1±0.3	53.9±17.7	1.1±0.5
Cochlea right	72.2±19.3	1.1±0.3	69.0±17.8	0.8±0.3	51.4±12.0	1.1±0.4	**63.3±19.8**	1.0±0.4
U Esophagus	81.1±11.3	1.6±0.3	77.4±18.1	1.5±0.6	64.3±15.3	2.0±0.6	**79.3±16.5**	**1.1±0.7**
Glottic larynx	88.2±4.8	1.1±0.2	81.8±15.6	1.0±0.3	73.1±17.8	1.4±0.4	**89.8±9.0**	**0.9±0.3**
Mandible	91.9±4.3	1.2±0.2	95.4±3.1	0.7±0.2	93.7±2.9	1.2±0.2	**98.8±1.4**	**0.7±0.2**
Oral cavity	84.7±7.5	2.5±0.4	91.5±6.0	2.0±0.7	94.0±4.7	2.9±0.6	96.2±3.2	**1.6±0.7**
S Glottic larynx	87.7±5.4	1.5±0.2	87.1±10.0	1.4±0.5	86.3±10.9	1.8±0.4	**93.9±3.2**	**1.3±0.4**
PG left	81.3±22.2	1.8±0.5	83.8±22.8	1.5±0.4	88.8±3.4	1.7±0.1	**93.7±5.2**	**0.8±0.6**
PG right	88.0±5.9	1.9±0.4	91.2±3.9	1.5±0.5	89.1±3.3	1.8±0.3	**95.1±3.8**	**0.7±0.4**
PCM inferior	79.2±7.3	1.8±0.3	82.0±7.9	1.5±0.4	76.0±8.2	1.7±0.3	**87.7±7.8**	**1.2±0.5**
PCM middle	75.8±8.8	1.5±0.2	78.3±7.9	1.2±0.4	71.5±7.2	1.5±0.2	**86.4±8.2**	**0.8±0.3**
PCM superior	70.0±10.4	1.8±0.3	69.8±13.0	1.5±0.5	53.5±8.0	2.1±0.3	**77.6±12.7**	**1.2±0.5**
SG left	77.5±30.5	1.6±0.3	78.9±31.2	1.3±0.4	86.5±6.5	1.5±0.2	**92.7±7.9**	**0.8±0.6**
SG right	79.0±31.3	1.5±0.2	80.2±32.0	1.3±0.4	88.4±3.9	1.4±0.2	**96.0±5.3**	**0.4±0.6**
Spinal cord	77.6±6.7	1.2±0.2	84.9±6.5	1.1±0.4	70.8±6.8	4.4±1.9	**75.7±6.5**	**3.7±2.3**

Abbreviations: OARs: organs at risk; DSC: dice similarity coefficient; ASSD: average symmetric surface distance; STD: standard deviation; RO: radiation oncologist; PCM: pharyngeal constrictor muscles; PG: parotid gland; SG: submandibular gland; U: upper; S: supra.

Figure 1

Click here to download high resolution image



Caption figure 1

Figure 1: Overview of study design. Automated delineations (A) of 16 OARs in conventional planning CT images of 15 HNC patients were corrected by 2 different ROs (RO1: C1, RO2: C2) and were also manually delineated by the same ROs (RO1: M1, RO2: M2) in different delineation sessions. Accuracy of automated delineation was assessed by comparing automated and corrected contours for each RO (Acc1: C1 vs A, Acc2: C2 vs A). Intra-observer variability was assessed by comparing corrected and manual delineations by the same RO (IOV1: C1 vs M1, IOV2: C2 vs M2). Inter-observer variability was assessed by comparing corrected and manual delineations by different ROs (IOVc: C1 vs C2, IOVm: M1 vs M2).

Figure 2Am

Click here to download high resolution image

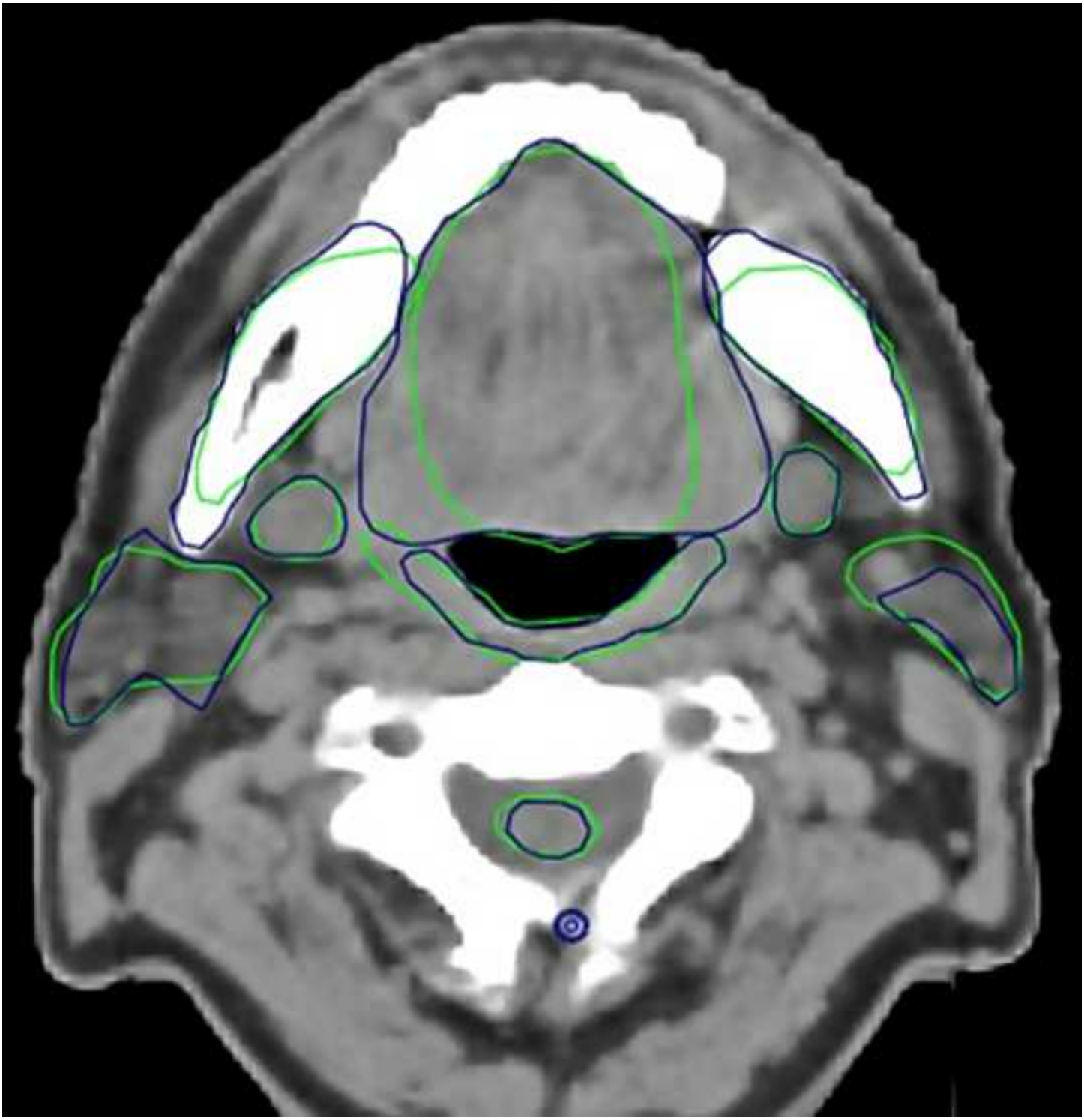


Figure 2Ac

Click here to download high resolution image

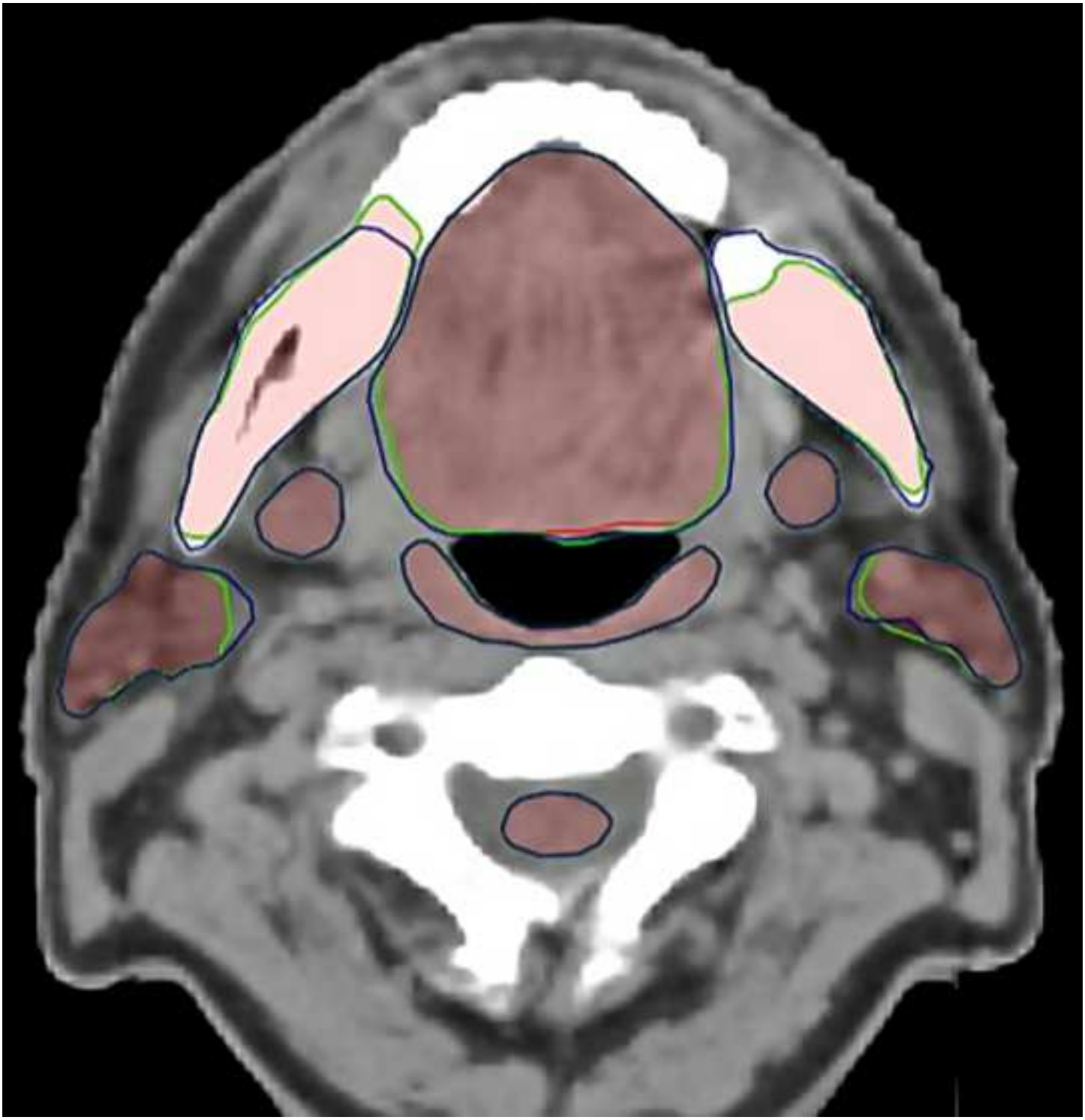


Figure 2Bm

Click here to download high resolution image

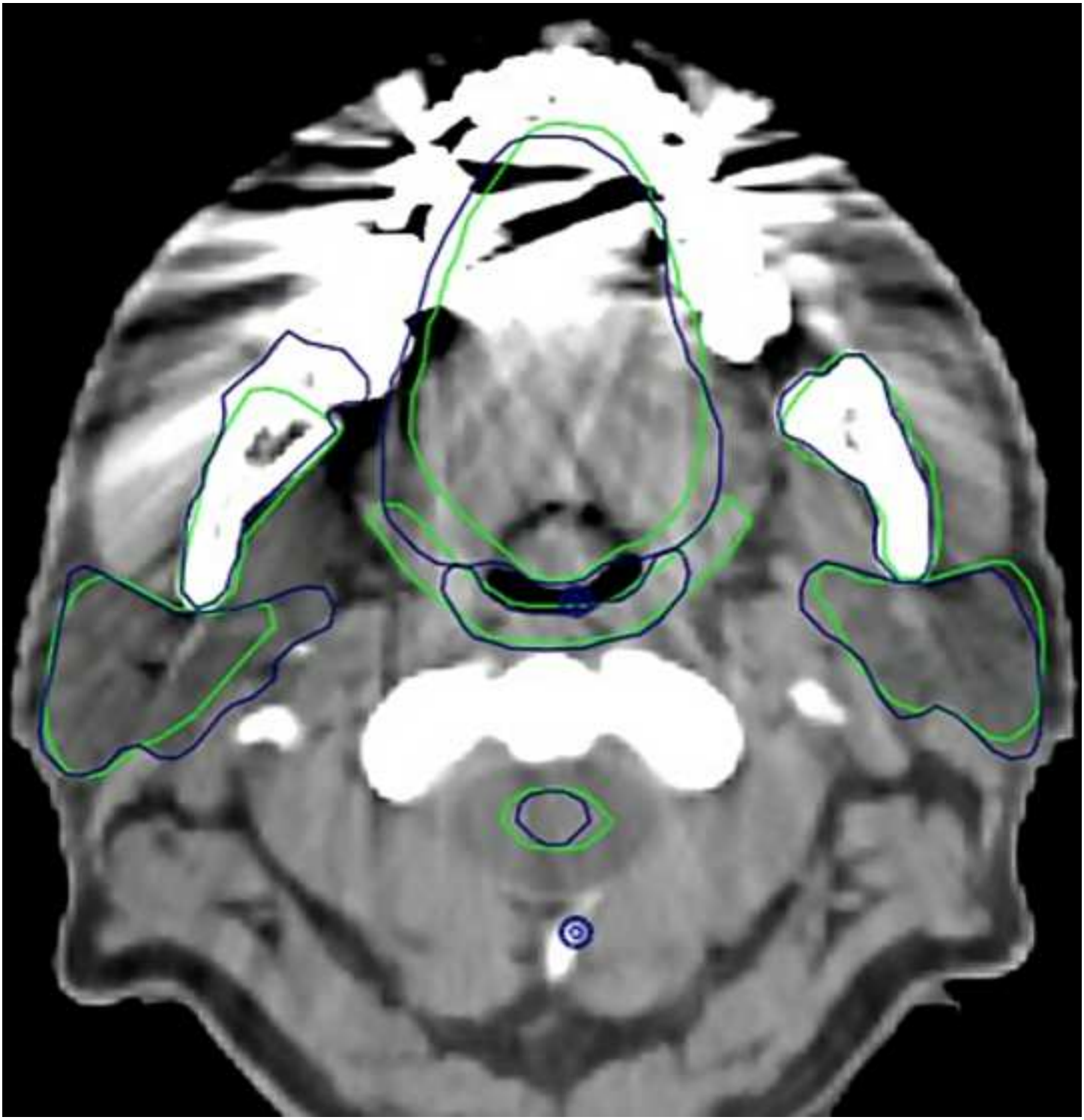


Figure 2Bc

Click here to download high resolution image

Figure 2Cm

Click here to download high resolution image

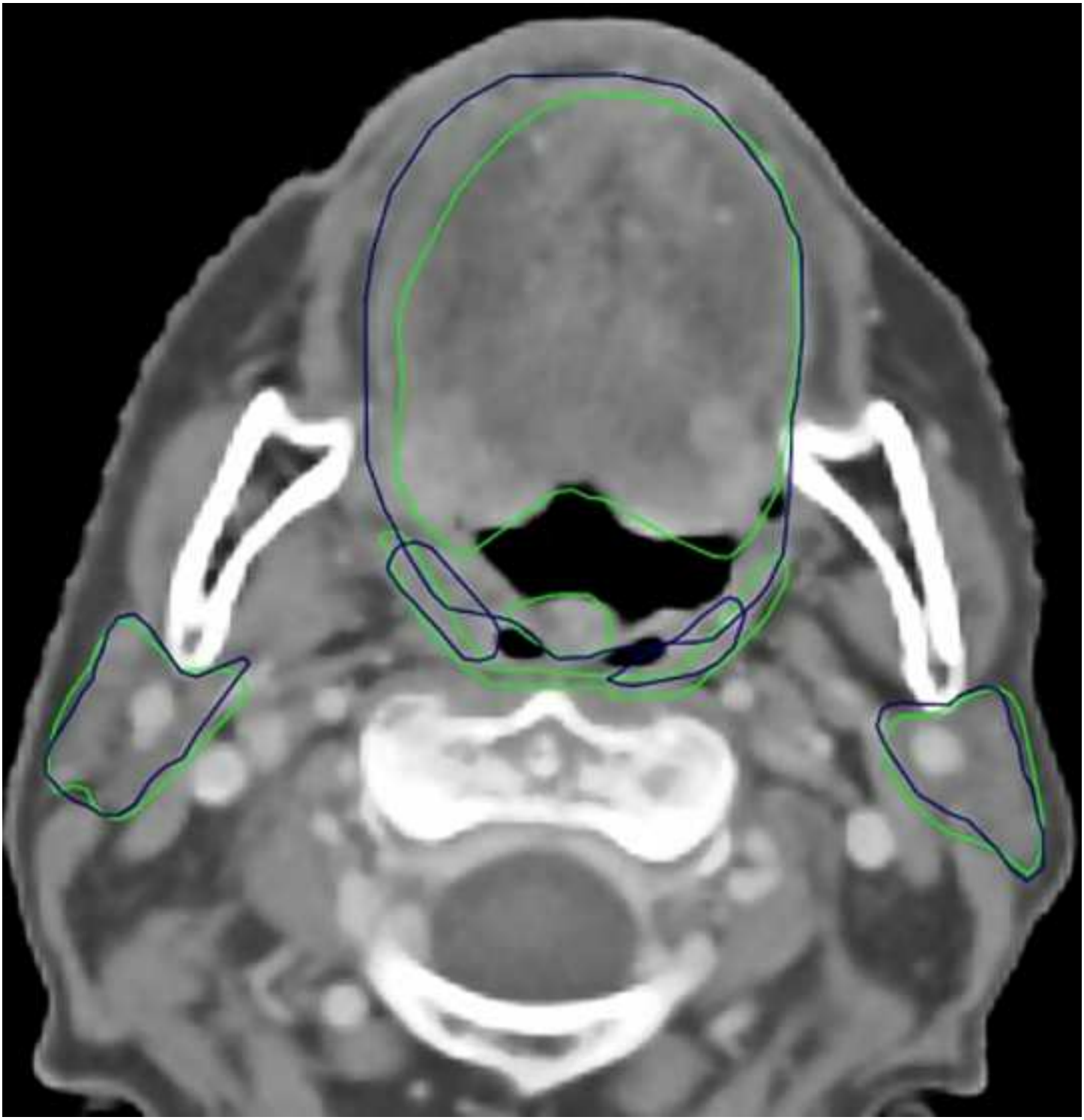


Figure 2Cc

Click here to download high resolution image

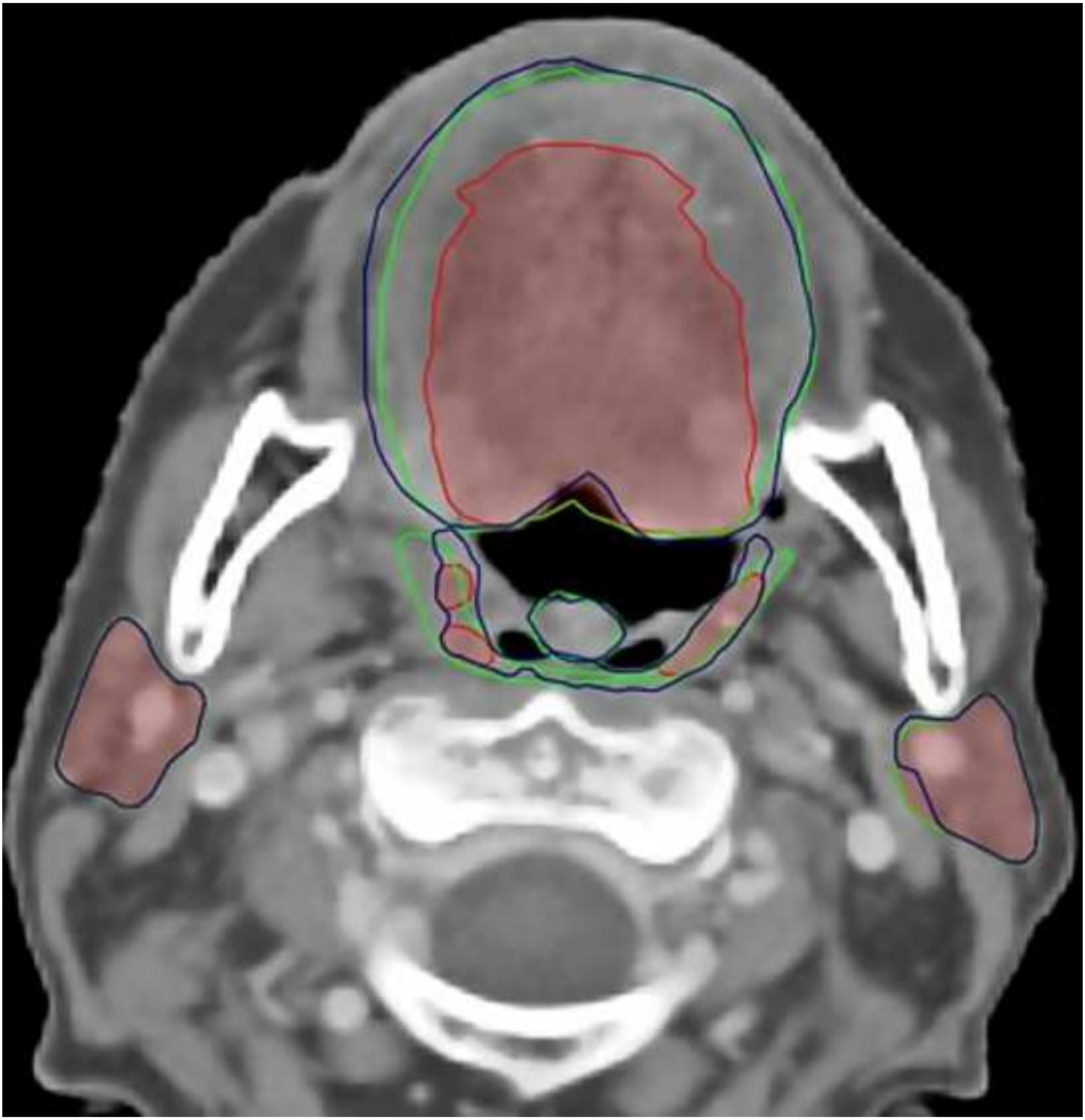


Figure 2D1

Click here to download high resolution image

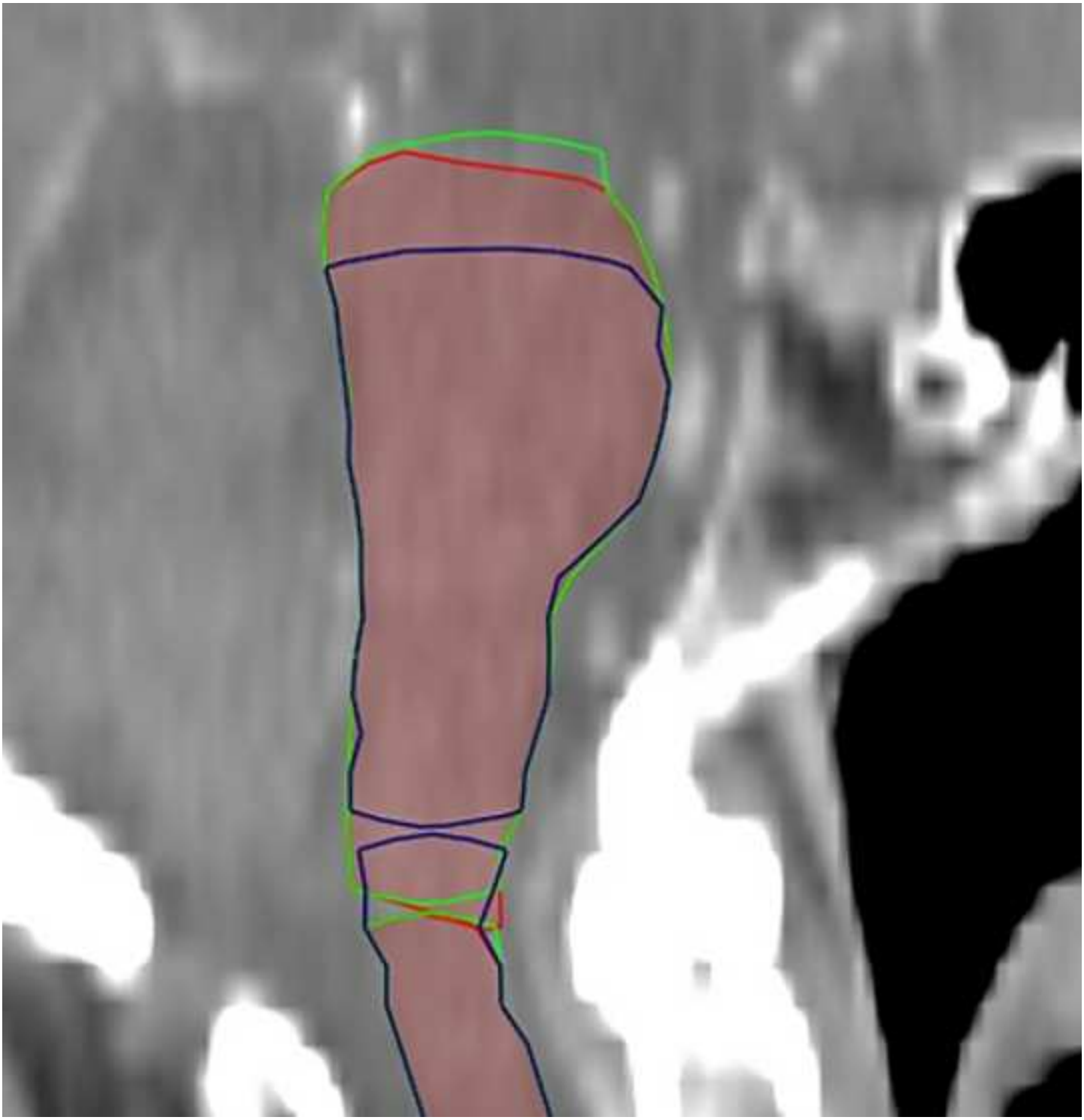


Figure 2D2

Click here to download high resolution image

Caption figure 2

Figure 1: Illustration of intra- and inter-observer variability between manual delineations (A_m - C_m) and corrected delineations (A_c - C_c), for both observers (RO1, RO2). Notice a decrease in IOV in A_c compared to A_m, and B_c compared to B_m, even with scatter artefacts. The decrease in IOV observed in C_c compared to C_m is due to a difference in delineation by RO2, independent of the network. Figures D_1 and D_2 show the difference in cranial and caudal border selection by the two ROs for brainstem and spinal cord.

Legend figure 2
Click here to download high resolution image

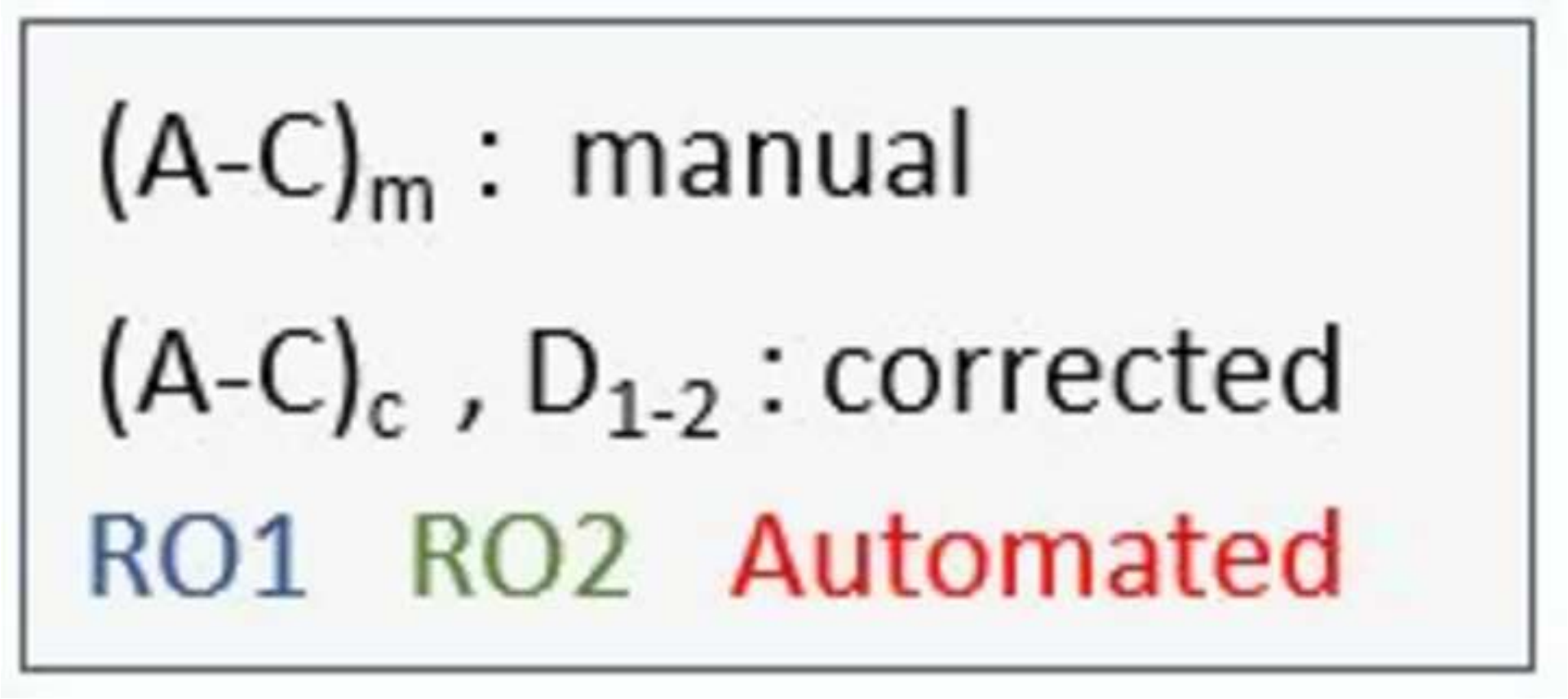


Figure 3A

Click here to download high resolution image

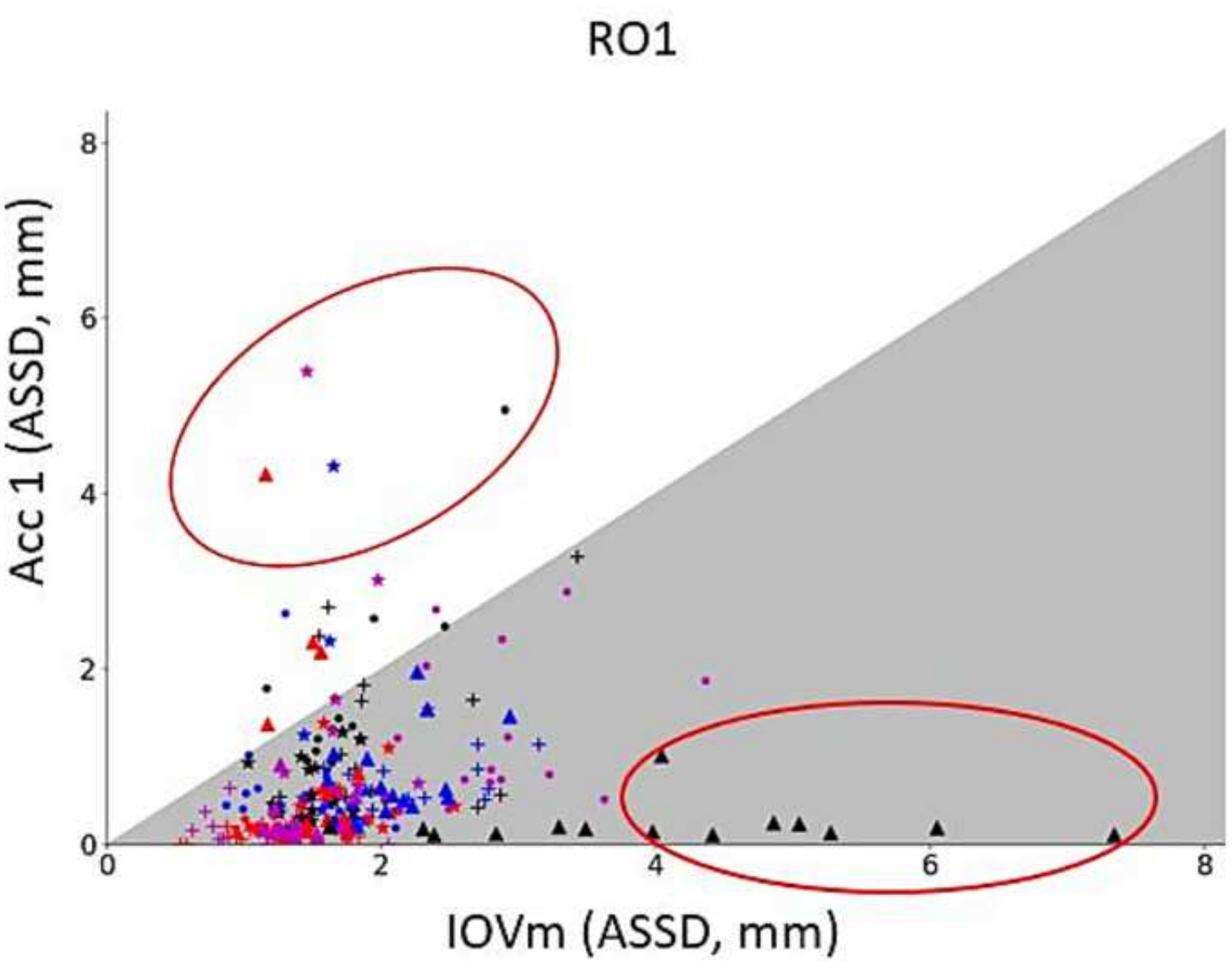
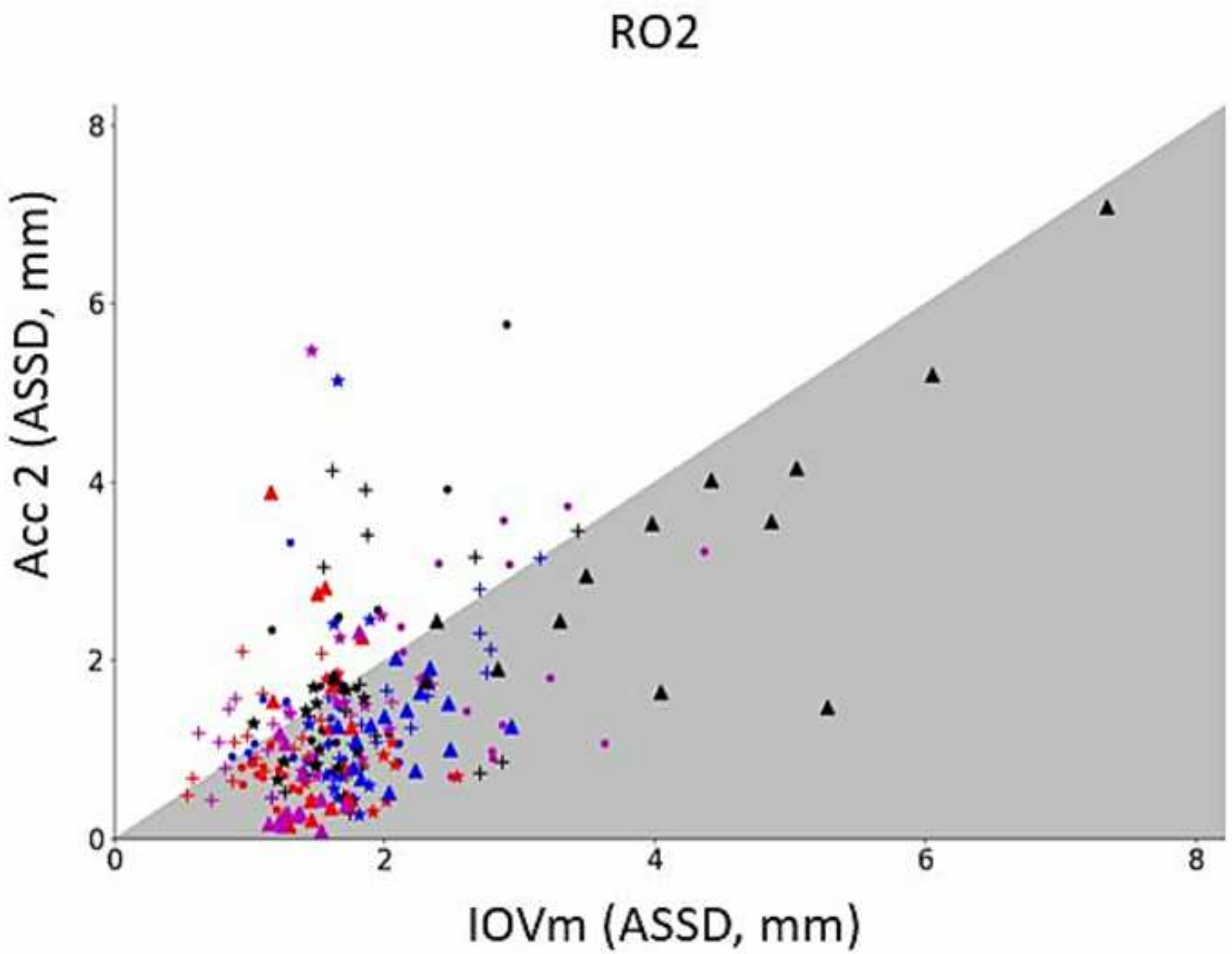


Figure 3B

Click here to download high resolution image



Caption figure 3

Figure 1: Network accuracy, corrections needed by radiation oncologist 1 (A) and 2 (B) vs manual interobserver variability. Network accuracy quantified by average corrections needed before the automated delineations were clinically acceptable. This was compared to interobserver variability between manual delineations. Each data point represents an organ at risk from one patient. For all structures in the grey zone, the corrections are smaller than variability in clinical practice. Abbreviations: ASSD: average symmetric surface distance; mm: millimetres; Acc: accuracy of network; RO: radiation oncologist; PCM: pharyngeal constrictor muscles; PG: parotid gland; SG: submandibular gland; U: upper; S: supra; IOVm: manual interobserver variability.

Legend figure 3

Click here to download high resolution image

Supplementary Files

Click here to download Supplementary Files: Appendix A.docx

Highlights

- Validation of automated delineation of organs at risk in head and neck cancer
- Automated delineation is more time efficient
- Automated delineation reduces interobserver variability