Received: 26 April 2023 | Revised: 31 October 2023 | Accepted: 13 December 2023

DOI: 10.1002/mp.16924

RESEARCH ARTICLE

MEDICAL PHYSICS

vOARiability: Interobserver and intermodality variability analysis in OAR contouring from head and neck CT and MR images

Gašper Podobnik[1] | Bulat Ibragimov[1,2] | Primož Peterlin[3] | Primož Strojan[3] | Tomaž Vrtovec[1]

[1] Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

[2] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[3] Institute of Oncology Ljubljana, Ljubljana, Slovenia

Correspondence
Tomaž Vrtovec, University of Ljubljana, Faculty of Electrical Engineering, Tržaška cesta 25, SI-1000 Ljubljana, Slovenia.
Email: tomaz.vrtovec@fe.uni-lj.si

Funding information
Slovenian Research and Innovation Agency (ARIS), Grant/Award Numbers: J2-1732, P2-0232, P3-0307; Novo Nordisk Fonden, Grant/Award Number: NFF20OC0062056; Ethics Committee of the Institute of Oncology Ljubljana, Slovenia, Grant/Award Number: ERID-EK/139

Abstract

Background: Accurate and consistent contouring of organs-at-risk (OARs) from medical images is a key step of radiotherapy (RT) cancer treatment planning. Most contouring approaches rely on computed tomography (CT) images, but the integration of complementary magnetic resonance (MR) modality is highly recommended, especially from the perspective of OAR contouring, synthetic CT and MR image generation for MR-only RT, and MR-guided RT. Although MR has been recognized as valuable for contouring OARs in the head and neck (HaN) region, the accuracy and consistency of the resulting contours have not been yet objectively evaluated.

Purpose: To analyze the interobserver and intermodality variability in contouring OARs in the HaN region, performed by observers with different level of experience from CT and MR images of the same patients.

Methods: In the final cohort of 27 CT and MR images of the same patients, contours of up to 31 OARs were obtained by a radiation oncology resident (junior observer, JO) and a board-certified radiation oncologist (senior observer, SO). The resulting contours were then evaluated in terms of interobserver variability, characterized as the agreement among different observers (JO and SO) when contouring OARs in a selected modality (CT or MR), and intermodality variability, characterized as the agreement among different modalities (CT and MR) when OARs were contoured by a selected observer (JO or SO), both by the Dice coefficient (DC) and 95-percentile Hausdorff distance (HD_{95}).

Results: The mean (±standard deviation) interobserver variability was 69.0 ± 20.2% and 5.1 ± 4.1 mm, while the mean intermodality variability was 61.6 ± 19.0% and 6.1 ± 4.3 mm in terms of DC and HD_{95}, respectively, across all OARs. Statistically significant differences were only found for specific OARs. The performed MR to CT image registration resulted in a mean target registration error of 1.7 ± 0.5 mm, which was considered as valid for the analysis of intermodality variability.

Conclusions: The contouring variability was, in general, similar for both image modalities, and experience did not considerably affect the contouring performance. However, the results indicate that an OAR is difficult to contour regardless of whether it is contoured in the CT or MR image, and that observer experience may be an important factor for OARs that are deemed difficult to contour. Several of the differences in the resulting variability can be also attributed to adherence to guidelines, especially for OARs with poor visibility or without distinctive boundaries in either CT or MR images. Although considerable contouring differences were observed for specific OARs, it can be concluded that almost all OARs can be contoured with a similar degree of variability in either the CT or MR modality, which works in favor of MR images from the perspective of MR-only and MR-guided RT.

KEYWORDS
contouring, computed tomography, head and neck cancer, image segmentation, magnetic resonance, organs-at-risk, radiotherapy, variability analysis

1 INTRODUCTION

In radiotherapy (RT) cancer treatment planning, accurate and consistent contouring of organs-at-risk (OARs) from medical images represents one of the key steps in producing patient-specific dose plans, assuring to deliver high dose to the tumor and spare OARs from excessive irradiation so as to reduce the associated toxicity.[1] Contouring has been identified as a task with the highest risk priority when reviewing RT treatment plans,[2] and therefore requires additional attention, particularly for regions such as the head and neck (HaN),[3] where it is common to define contours for more than 25 OARs.[4]

In clinical practice, manual OAR contouring is usually performed by radiation oncologists who commonly identify this task as tedious, labor-intensive and time-consuming (e.g., 5–6 h per image[5]). Moreover, besides being affected by image variability in the form of various imaging artifacts (i.e., noise, intensity inhomogeneities, partial volume effects, etc.) and variable anatomy appearance (i.e., natural biological variability, pathological changes, etc.),[1] contouring is ultimately biased by the experience of observers as well as by their subjective interpretation of OAR image boundaries, reflected in intra- and interobserver variability.[6–9] To reduce observer variability for OARs in the HaN region, several contouring guidelines[4,10] have been published, as well as initiatives have been recently launched to quantify observer variability[10–12] and provide quality assurance.[13–16] On the other hand, automated contouring (i.e., automated segmentation, auto-segmentation) performed by computer-assisted algorithms[17] has witnessed a revival with the introduction and integration of artificial intelligence approaches, such as deep learning,[18–26] which has outperformed the previously established atlas-based auto-segmentation.[27] As a result, computational challenges were organized to evaluate the quality of auto-segmentation results,[28] and several datasets were made publicly available for benchmarking different auto-segmentation methodologies[20,28–31] and evaluating their clinical acceptability.[32] However, even with sophisticated auto-segmentation approaches, manual contouring is still the method of choice for evaluating and benchmarking the performance of the developed algorithms.

Most manual contouring as well as auto-segmentation approaches still rely on computed tomography (CT) images, which are required for RT planning as they contain electron density information used for the calculation of the radiation beam energy absorption. However, because of the often insufficient CT image contrast for soft tissues, several studies recommended the integration of complementary magnetic resonance (MR) modality.[4,33] This is important especially from the perspective of OAR contouring,[4,34,35] synthetic MR image generation for MR-aided RT,[36,37] synthetic CT image generation for MR-only RT[38–41] and MR-guided RT.[42,43] While some OARs can be accurately and reliably contoured in CT images (i.e., bone structures such as, e.g., the mandible), MR images are often used to better visualize soft tissues. A common clinical practice is to align both images in the same coordinate system via image registration,[44] which allows the observer to switch between the two modalities and therefore better characterize OAR boundaries.

The MR modality has been long recognized as valuable for contouring OARs in the HaN region,[45–48] however, to the best of our knowledge, the accuracy and consistency of the resulting OAR contours have not been yet objectively evaluated. In this study, we therefore analyze the interobserver and intermodality variability of manual contouring of up to 31 OARs in the HaN region, performed by observers with different level of experience from CT and MR images of the same patients. Besides providing valuable insights to the levels of both interobserver and intermodality variability from the perspective of manual OAR contouring, the obtained results can be also viewed as a baseline for an objective evaluation of methods for auto-segmentation of OARs in the HaN region,[31] which have been rapidly evolving during the past decade due to the integration of artificial intelligence, and received a considerable boost in performance due to the advances in deep learning.[18–26]

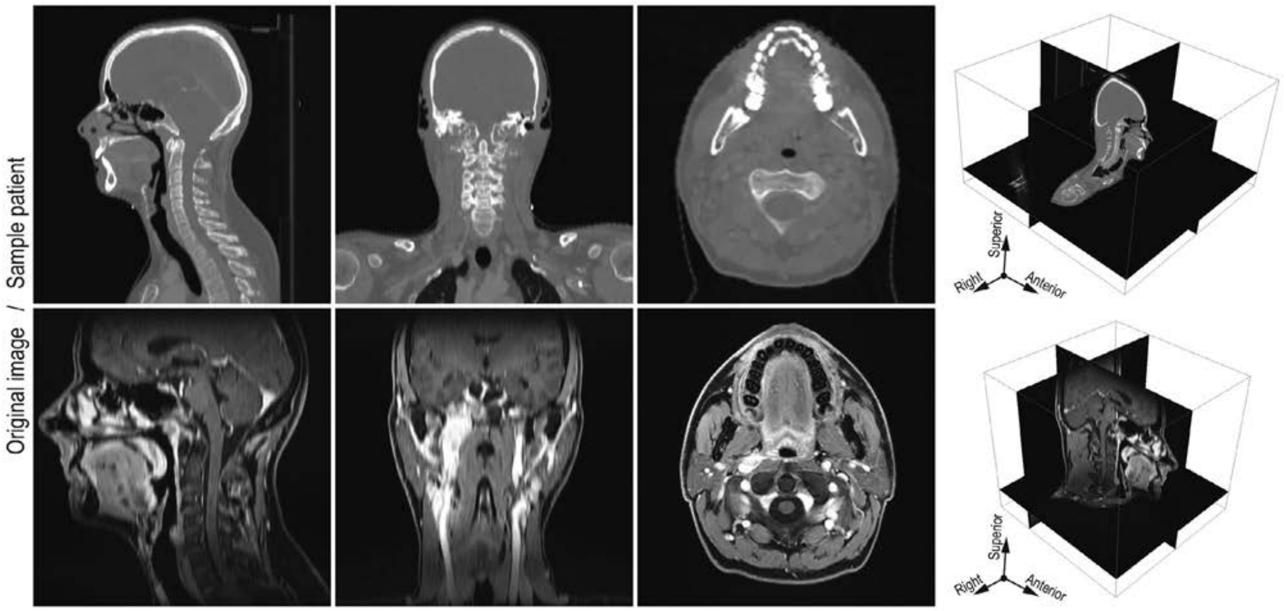

FIGURE 1 Example of a CT (*top row*) and MR (*bottom row*) image of the same patient from the devised cohort, displayed in mid-axial (*left*), mid-coronal (*middle*), and mid-sagittal (*right*) cross-sections. (For visualization purposes, the original cross-sections were zoomed to the region of interest.) CT, computed tomography; MR, magnetic resonance.

2 | METHODS

2.1 | Images

We devised an initial cohort of 30 retrospectively collected anonymized patients (24 males, 6 females) that underwent both CT and MR image acquisition for the purpose of image-guided RT in the HaN region at the Institute of Oncology Ljubljana, Slovenia, in 2019, with a mean age (± standard deviation, SD) of 60.4 ± 10.2 years (range: 36–78 years). The images are part of the HaN-Seg dataset[31] (refer to for image acquisition details), and an example of a CT and MR image pair is shown in Figure 1.

2.2 | Manual contouring

Manual contouring was performed by radiation oncologists from the Institute of Oncology Ljubljana, Slovenia, who used the *ARIA Oncology Information System* software (v15.6, Varian Medical Systems, Palo Alto, CA, USA) for image manipulation and manual pixel-wise delineation of up to 31 OARs (or 23, considering the left and right instances of paired OARs as one instance) according to the standard RT planning practice and by following the established OAR contouring guidelines for the HaN region.[4] Each CT or MR image was independently (without using the other modality) contoured twice, once by a junior observer (JO), that is, a radiation oncology resident, from a group of 10 JOs, and once by a senior observer (SO), that is, a board-certified radiation oncologist, from a group of eight SOs, in order to reduce the bias of individual contouring styles. Moreover, each JO or SO was assigned CT images that were not paired with MR images, meaning that also each modality was independently contoured (i.e., without help from the other modality). In the Supplementary Material, Table S1 and Table S2 contain the observer-to-patient/image modality and patient/image modality-to-observer allocation tables, respectively, while Figure S1 and Figure S2 show detailed statistical distributions of the estimated over-contouring and under-contouring that resulted from individual contouring styles of different observers.

2.3 | Image registration

To compare among contour sets from different modalities and analyze the corresponding intermodality variability, CT and MR images of each patient were registered by using *SimpleElastix* (v0.10.0, https://simpleelastix.github.io), an extension of the open-source image registration toolbox *elastix*.[49] By finding the optimal rigid (i.e., translation and rotation) and non-rigid (i.e., B-splines) geometrical alignment between each pair of CT and MR images, the OAR contours were mapped into the same coordinate system. For the purpose of evaluating registration results, six control points were manually placed by an experienced medical imaging researcher at anatomical locations that can be reliably identified in both CT and MR images, that is, at the nasal tip (#1), the posterior edge of the left/right angle of the mandible (#2/#3), the posterior edge of the skull at the height of the nasal tip (#4), and at the approximate junction between the vertebral lamina and left/right transverse process of a C3 – C6 vertebral level (#5/#6). The registration performance was then quantitatively evaluated by computing the Euclidean distance between each control point and its corresponding registered pair, which was averaged across all six control points for each patient to obtain the target registration error (TRE).

2.4 | Variability analysis

Manual contouring was evaluated in terms of interobserver variability, characterized as the agreement among different observers (JO and SO) when contouring OARs in a selected modality (CT or MR), and intermodality variability, characterized as the agreement among different modalities (CT and MR) when OARs were contoured by a selected observer (JO or SO). Both were measured by the Dice coefficient (DC), a standard metrics for volume overlap, and the 95-percentile Hausdorff distance (HD_{95}), a standard metrics for surface mutual proximity (https://github.com/deepmind/surface-distance)[27]:

$$DC = \frac{2|X \cap Y|}{|X| + |Y|},$$

$$HD_{95} = \max\{d_{k95\%}(X, Y), d_{k95\%}(Y, X)\}, \qquad (1)$$

where X and Y are the two volumetric contours under comparison, $|X|$, $|Y|$ and $|X \cap Y|$ represent the number of voxels in X, Y and their volumetric overlap, respectively, and $d_{k95\%}(X, Y)$ and $d_{k95\%}(Y, X)$ are the 95-percentile of the Euclidean distances from voxels in X to the surface of Y, and vice-versa, respectively. Statistically significant differences were observed by applying paired t-tests at $p = 0.05$ significance level. The results are presented with mean values and SD, and in the form of box plots indicating the median value and quartile range.

3 | RESULTS

3.1 | Images

From the initial cohort, two patients were excluded because of an insufficient MR image field of view (FoV), and one because of poor MR image quality. The final cohort consisted of 27 patients (21 males, 6 females), with a mean age of 60.8 ± 10.4 years (range: 36–78 years), where for each patient one CT and one T1-weighted MR image of the HaN region was available. The images used in this study are part of the publicly available HaN-Seg dataset[31] (i.e., 20 images; https://doi.org/10.5281/zenodo.7442914) as well as of the privately withheld dataset used for the HaN-Seg challenge (i.e., 7 images; https://hanseg2023.grand-challenge.org).

3.2 | Manual contouring

The appointed JO and SO independently contoured up to 31 OARs in each of the 27 CT and 27 MR images. In specific cases, some OARs were left out due to poor visibility. In general, the mandible and cochleae were not contoured in MR images, and the optic chiasm was not contoured in CT images (the few attempts of such contouring were omitted from analysis). Because of the commonly smaller FoV, several OARs were not contoured in MR images. Each patient was therefore assigned four contour sets, that is, one set obtained by JO in the CT image (JO/CT), one set obtained by JO in the MR image (JO/MR), one set obtained by SO in the CT image (SO/CT) and one set obtained by SO in the MR image (SO/MR). By following the American Association of Physicists in Medicine (AAPM) Task Group 263 nomenclature[50] and contouring guidelines,[4] the OAR contours were named `A_Carotid_L/R` (carotid artery), `Arytenoid` (arytenoids), `Bone_Mandible` (mandible), `Brainstem` (brainstem), `BuccalMucosa` (buccal mucosa), `Cavity_Oral` (oral cavity), `Cochlea_L/R` (cochlea), `Cricopharyngeus` (cricopharyngeal inlet), `Esophagus_S` (cervical esophagus), `Eye_AL/R` (anterior segment of the eyeball), `Eye_PL/R` (posterior segment of the eyeball), `Glnd_Lacrimal_L/R` (lacrimal gland), `Glnd_Submand_L/R` (submandibular gland), `Glnd_Thyroid` (thyroid gland), `Glottis` (glottic larynx), `Larynx_SG` (supraglottic larynx), `Lips` (lips), `Musc_Constrict` (pharyngeal constrictor muscles, PCMs), `OpticChiasm` (optic chiasm), `OpticNrv_L/R` (optic nerve), `Parotid_L/R` (parotid gland), `Pituitary` (pituitary gland), and `SpinalCord` (spinal cord), where `L/R` denotes the left/right OAR instances. An example is shown in Figure 2.

3.3 | Image registration

Image registration was successfully obtained for each of the 27 CT and MR image pairs, with a mean TRE of 1.7 ± 0.5 mm (median: 1.6 mm; maximum: 3.3 mm) across the whole cohort. According to the resulting geometrical transformation, the MR contours were mapped to the coordinate system of the corresponding CT image, therefore enabling a one-to-one comparison of OAR contours and analysis of intermodality variability. An example of the resulting image registration along with detailed TRE results is shown in Figure 3.

3.4 | Variability analysis

Besides the personal choice of the observer whether to contour an OAR or not, occurring especially when the OAR was only partially visible in the image (i.e., because

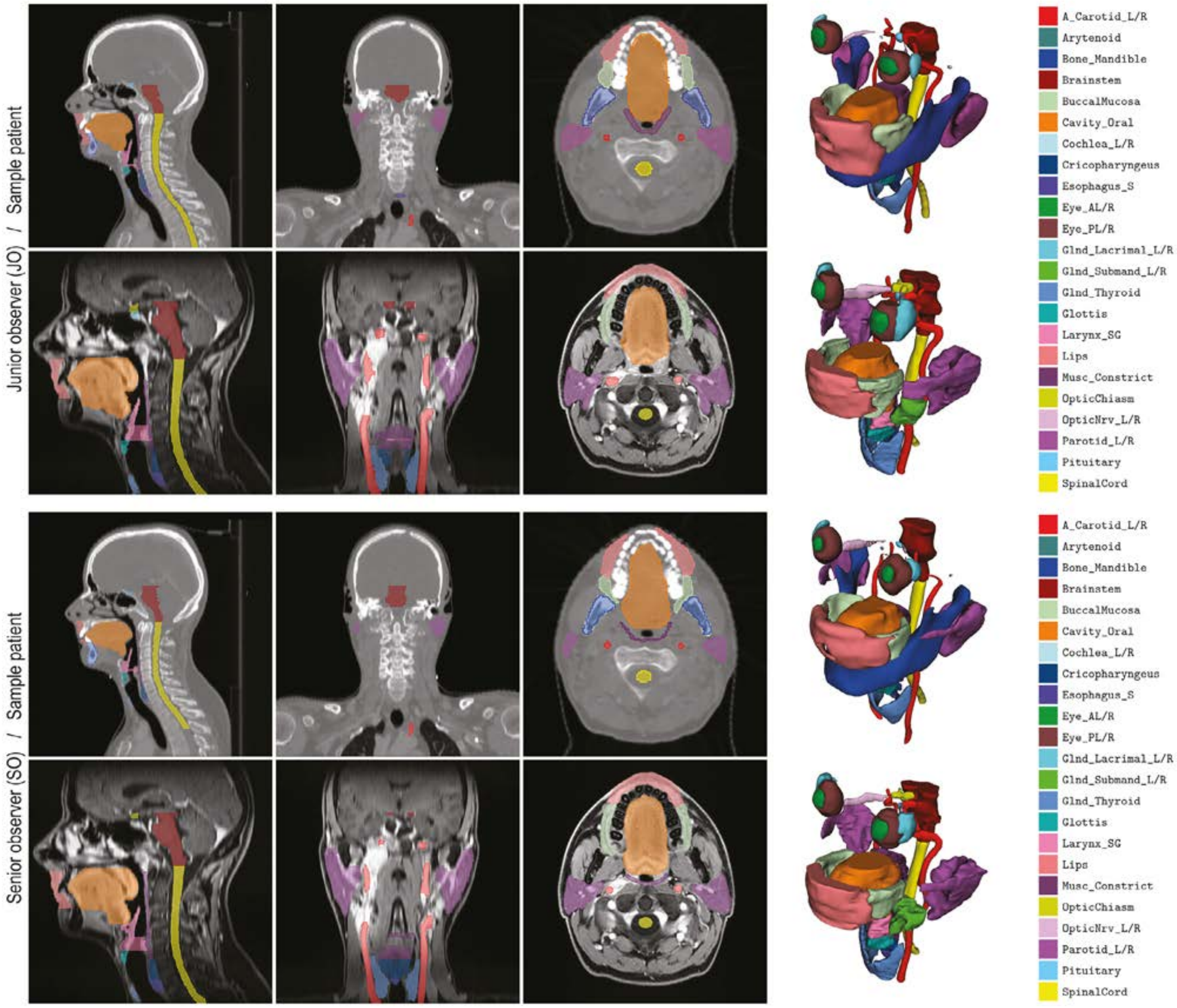

FIGURE 2 Example of manual contouring of head and neck OARs in a CT (*left, top row*) and MR (*left, bottom row*) image of the same patient from Figure 1, as performed by different observers, along with three-dimensional reconstructions of the corresponding OAR contours (*right*). CT, computed tomography; MR, magnetic resonance; OAR, organ-at-risk.

of a limited FoV), the following observer inconsistencies against contouring guidelines[4] were identified. First, the cranial contours of carotid arteries were often not extended to the optic chiasm but only to the pituitary gland, while their caudal ends often did not reach the brachiocephalic trunk or the aortic arch. As a result, when comparing two A_Carotid_L/R contours, they were cut-off at their cranial and caudal ends at the axial cross-sections that contained both contours. Next, the caudal end of the spinal cord was often not extended to the superior edge of the T3 vertebra, and similarly the caudal end of the cervical esophagus was often not extended to the inferior edge of the C7 vertebra. As a result, when comparing two SpinalCord or two Esophagus_S contours, they were cut-off at their caudal ends at the axial cross-sections containing both contours. On the other hand, the cranial end of PCMs was often not extended to the inferior tip of the pterygoid plates, therefore, when comparing two Musc_Constrict contours, they were cut-off at their cranial ends at the axial cross-section that contained both contours. Finally, because of a smaller FoV of MR images, some OARs were cropped by image registration, and therefore only those contour pairs that were in volume larger than 25% of the original contour volume before registration were retained. By performing the described operations, our variability analysis becomes more focused on actual contouring differences rather than guideline interpretation and adherence.

The resulting overall interobserver variability across all OARs was, respectively in terms of DC and HD_{95}, equal to 71.2 ± 18.2% and 4.9 ± 4.1 mm for JO/CT

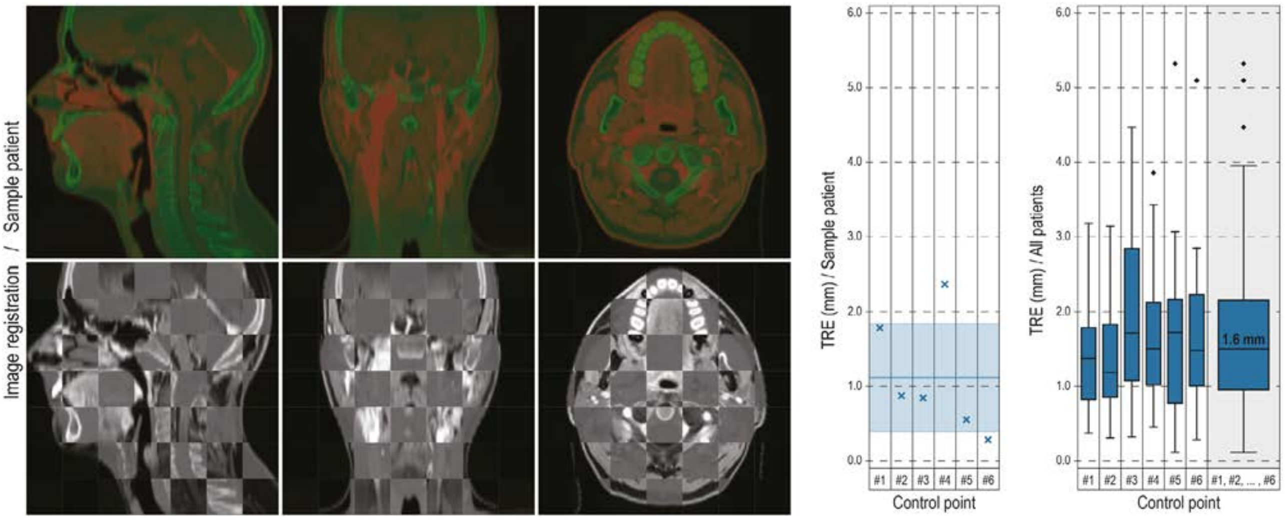

FIGURE 3 Example of registration results for a selected pair of CT and MR images of the same patient from Figure 1, shown as a semi-transparent color-coded overlay (*left, top row*) and checkerboard (*left, bottom row*) of CT/MR cross-sections. The registration results were quantitatively evaluated by the TRE of six corresponding control points, as shown for the patient from Figure 1 (*right, first plot*) as well as an aggregate box plot for all patients (*right, second plot*). CT, computed tomography; MR, magnetic resonance; TRE, target registration error.

versus SO/CT, 66.6 ± 22.1% and 5.4 ± 4.0 mm for JO/MR versus SO/MR, and 69.0 ± 20.2% and 5.1 ± 4.1 mm for JO versus SO. The overall intermodality variability across all OARs was equal to 61.9 ± 19.6% and 6.0 ± 4.3 mm for CT/JO versus MR/JO, 61.3 ± 18.4% and 6.2 ± 4.4 mm for CT/SO versus MR/SO, and 61.6 ± 19.0% and 6.1 ± 4.3 mm for CT versus MR. The results for individual OARs are shown in Table 1. Statistically significant differences were only found for specific OARs and can be observed from the results for DC in Figure 4 and HD_{95} in Figure 5. As the contours of the mandible and cochleae were available only for CT images, no interobserver variability for `Bone_Mandible` and `Cochlea_L/R` is reported for MR images. Similarly, the contours of the optic chiasm were available only for MR images, therefore the interobserver variability for `OpticChiasm` is not reported for CT images. As a result, the intermodality variability is also not reported for these three OARs. For the remaining OARs, the number of contoured instances was always 27 or less, depending on the FoV of corresponding images. Some major inconsistencies in manual contouring that appear as outliers in the analysis of the interobserver and intermodality variability are presented in Figure S3 to Figure S8 of the Supplementary material.

4 | DISCUSSION

To mitigate the contouring variability for OARs in the HaN region, several well-defined guidelines have been published, as reported on the *eContour* (https://econtour.org) web portal.[10] The most established consensus[4] encompasses a complete set of OARs, while other guidelines are focused on OARs relevant to the case of nasopharyngeal carcinoma, swallowing, salivary functioning, hearing and balance, brachial plexopathy, and optic neuropathy.[10] However, even if guidelines are followed, manual contouring is still biased by the subjective interpretation of the observer, and therefore it is strongly recommended to perform basic observer training with joint delineation review sessions,[7] and to include additional modalities to improve the visibility of structure boundaries.[6] By considering the emerging role of the MR modality in RT[33] and its value for OAR contouring,[48] in this study we performed a variability analysis of OAR contouring in the HaN region from CT and MR images of the same patients.

The interobserver variability of HaN OAR contouring has been so far evaluated in several studies.[6–9,27] Brouwer et al.[6] obtained contours of seven OARs from CT images of six patients by five different observers, and reported large contouring variations[a] for the glottic larynx and spinal cord, and moderate for the thyroid, submandibular and parotid glands. While large variations were mostly attributed to poor compliance with the guidelines, they noted that the addition of MR images may improve the visibility of boundaries between tissues. Nelms et al.[7] evaluated contours of five OARs from a CT image of a single patient that were provided by up to 32 different clinical institutions. They identified the brainstem contours as the most variable, followed by contours of the parotid glands, spinal cord and mandible. Van der Veen et al.[9] obtained contours of 13 OARs from CT images of five patients by up to 14 observers, and concluded that small variations

[a] The originally reported concordance index (CI), also known as the Jaccard index, can be calculated as CI=100%·DC/(200%−DC), therefore DC can be calculated as DC=200%·CI/(100%+CI).

TABLE 1 Interobserver and intermodality variability in contouring OARs in CT and MR images of the same patients, performed by a JO and a SO, and reported as mean ± standard deviation of the DC and HD_{95}, computed from *n* instances of individual OARs.

	Interobserver variability						Intermodality variability					
	JO/CT versus SO/CT			JO/MR versus SO/MR			CT/JO versus MR/JO			CT/SO versus MR/SO		
OAR	*n*	DC (%)	HD_{95} (mm)	*n*	DC (%)	HD_{95} (mm)	*n*	DC (%)	HD_{95} (mm)	*n*	DC (%)	HD_{95} (mm)
A_Carotid_L	26	76.2 ± 14.8	3.1 ± 4.1	27	76.8 ± 9.9	2.7 ± 2.5	26	65.5 ± 10.1	3.0 ± 1.4	26	61.2 ± 15.2	4.2 ± 3.3
A_Carotid_R	26	75.8 ± 14.7	2.7 ± 3.4	26	78.6 ± 7.1	2.0 ± 1.2	26	65.0 ± 8.4	2.9 ± 1.2	25	60.6 ± 13.7	3.8 ± 2.3
Arytenoid	26	59.5 ± 20.8	4.2 ± 5.7	25	25.7 ± 24.4	6.0 ± 3.2	24	27.0 ± 21.2	6.1 ± 3.3	24	32.1 ± 21.4	6.0 ± 4.9
Bone_Mandible	26	91.6 ± 3.8	2.1 ± 2.0									
Brainstem	26	76.6 ± 8.6	7.0 ± 4.0	26	79.2 ± 14.9	9.0 ± 7.2	26	71.4 ± 11.7	9.4 ± 5.7	25	75.9 ± 7.7	7.4 ± 4.3
BuccalMucosa	25	56.0 ± 15.2	8.9 ± 4.8	22	54.2 ± 16.7	8.8 ± 3.9	21	56.5 ± 10.8	8.9 ± 4.7	24	47.8 ± 18.2	9.6 ± 5.7
Cavity_Oral	25	84.9 ± 6.1	8.1 ± 4.1	26	83.2 ± 5.6	8.4 ± 3.2	26	79.1 ± 8.4	10.1 ± 6.3	24	78.5 ± 8.8	10.3 ± 4.8
Cochlea_L	26	46.0 ± 25.5	3.2 ± 2.0									
Cochlea_R	26	44.8 ± 24.3	3.6 ± 2.2									
Cricopharyngeus	26	58.6 ± 13.2	7.9 ± 5.1	24	49.1 ± 16.8	7.8 ± 3.8	21	52.4 ± 16.2	6.9 ± 4.7	20	51.6 ± 16.0	6.1 ± 3.6
Esophagus_S	23	83.0 ± 11.3	1.9 ± 1.4	15	80.3 ± 9.3	2.5 ± 1.4	11	73.1 ± 12.2	3.5 ± 2.5	10	67.9 ± 10.3	3.4 ± 0.8
Eye_AL	27	69.2 ± 12.8	3.5 ± 2.0	25	69.1 ± 14.4	3.2 ± 2.2	20	64.7 ± 9.3	3.8 ± 1.3	20	56.4 ± 11.5	4.7 ± 1.8
Eye_AR	27	72.1 ± 13.4	3.3 ± 2.1	25	66.2 ± 14.1	3.5 ± 2.4	20	63.9 ± 10.0	3.5 ± 1.5	20	59.1 ± 15.6	4.6 ± 2.1
Eye_PL	27	88.6 ± 3.9	2.4 ± 0.8	26	84.2 ± 5.4	3.3 ± 1.4	22	82.1 ± 5.1	3.8 ± 1.7	22	79.4 ± 5.8	3.5 ± 1.1
Eye_PR	26	88.1 ± 4.5	2.4 ± 0.9	26	85.2 ± 5.2	2.9 ± 1.0	21	81.4 ± 4.9	3.6 ± 1.2	20	81.5 ± 4.7	3.6 ± 1.3
Glnd_Lacrimal_L	26	59.7 ± 12.9	4.9 ± 3.3	26	54.1 ± 25.4	4.9 ± 3.6	19	39.8 ± 18.1	6.4 ± 3.9	19	45.1 ± 17.1	5.3 ± 2.7
Glnd_Lacrimal_R	26	60.4 ± 15.9	4.6 ± 2.8	26	53.7 ± 26.3	5.5 ± 4.5	20	38.4 ± 17.8	6.0 ± 2.7	20	43.1 ± 16.3	6.6 ± 3.5
Glnd_Submand_L	26	85.7 ± 5.2	3.7 ± 2.7	26	85.8 ± 5.8	3.6 ± 2.8	26	78.7 ± 5.4	4.4 ± 2.2	25	77.3 ± 6.1	5.3 ± 3.6
Glnd_Submand_R	26	83.7 ± 5.8	4.7 ± 3.5	24	85.2 ± 5.8	4.0 ± 3.6	24	79.0 ± 7.4	4.5 ± 2.3	25	76.2 ± 7.3	6.0 ± 4.5
Glnd_Thyroid	26	85.9 ± 4.9	3.3 ± 3.7	20	78.5 ± 9.0	5.5 ± 2.6	15	71.0 ± 9.0	6.6 ± 3.4	15	69.7 ± 15.5	5.8 ± 5.6
Glottis	26	66.9 ± 14.1	4.7 ± 4.0	24	58.4 ± 19.8	6.2 ± 3.7	25	49.6 ± 15.7	5.6 ± 2.7	24	48.9 ± 15.5	7.4 ± 4.5
Larynx_SG	26	72.3 ± 16.0	6.0 ± 3.7	25	58.8 ± 9.9	8.7 ± 3.3	26	58.0 ± 12.1	7.4 ± 3.1	24	58.0 ± 11.3	7.4 ± 2.7
Lips	27	66.1 ± 9.5	9.4 ± 4.4	24	63.5 ± 13.3	9.6 ± 4.6	24	59.7 ± 10.0	10.5 ± 4.8	24	56.5 ± 12.5	9.8 ± 4.5
Musc_Constrict	25	61.6 ± 9.5	6.7 ± 2.8	25	57.5 ± 7.2	6.7 ± 2.9	26	51.7 ± 12.1	6.5 ± 3.2	23	50.5 ± 7.1	7.8 ± 3.7
OpticChiasm				26	39.5 ± 19.0	5.7 ± 3.1						
OpticNrv_L	26	60.8 ± 13.1	4.7 ± 2.8	26	55.9 ± 13.4	4.5 ± 2.1	22	48.7 ± 13.7	4.9 ± 3.9	22	45.0 ± 11.7	5.4 ± 2.6
OpticNrv_R	27	64.6 ± 13.5	5.3 ± 3.7	26	58.0 ± 16.1	4.3 ± 3.4	22	48.6 ± 11.2	4.7 ± 2.7	22	50.6 ± 8.0	5.8 ± 3.7
Parotid_L	27	82.0 ± 5.5	8.2 ± 5.0	26	85.6 ± 3.9	4.8 ± 2.5	26	79.6 ± 5.5	7.0 ± 4.9	26	78.6 ± 4.8	8.7 ± 7.0
Parotid_R	26	79.4 ± 9.3	8.8 ± 6.4	26	84.2 ± 5.0	5.9 ± 3.0	25	77.7 ± 8.8	9.3 ± 5.9	25	77.0 ± 6.8	7.8 ± 5.3
Pituitary	25	54.0 ± 19.4	3.4 ± 1.4	25	35.1 ± 25.3	5.5 ± 4.8	23	32.3 ± 23.2	6.1 ± 4.6	25	48.1 ± 14.5	4.0 ± 1.9
SpinalCord	27	81.0 ± 4.8	2.8 ± 2.6	26	81.3 ± 9.1	4.9 ± 5.8	25	71.2 ± 8.8	4.1 ± 3.9	26	74.3 ± 5.6	5.2 ± 5.1

Abbreviations: CT, computed tomography; DC, Dice coefficient; HD_{95}, 95-percentile Hausdorff distance; JO, junior observer; MR, magnetic resonance; OARs, organs-at-risk; SO, senior observer.

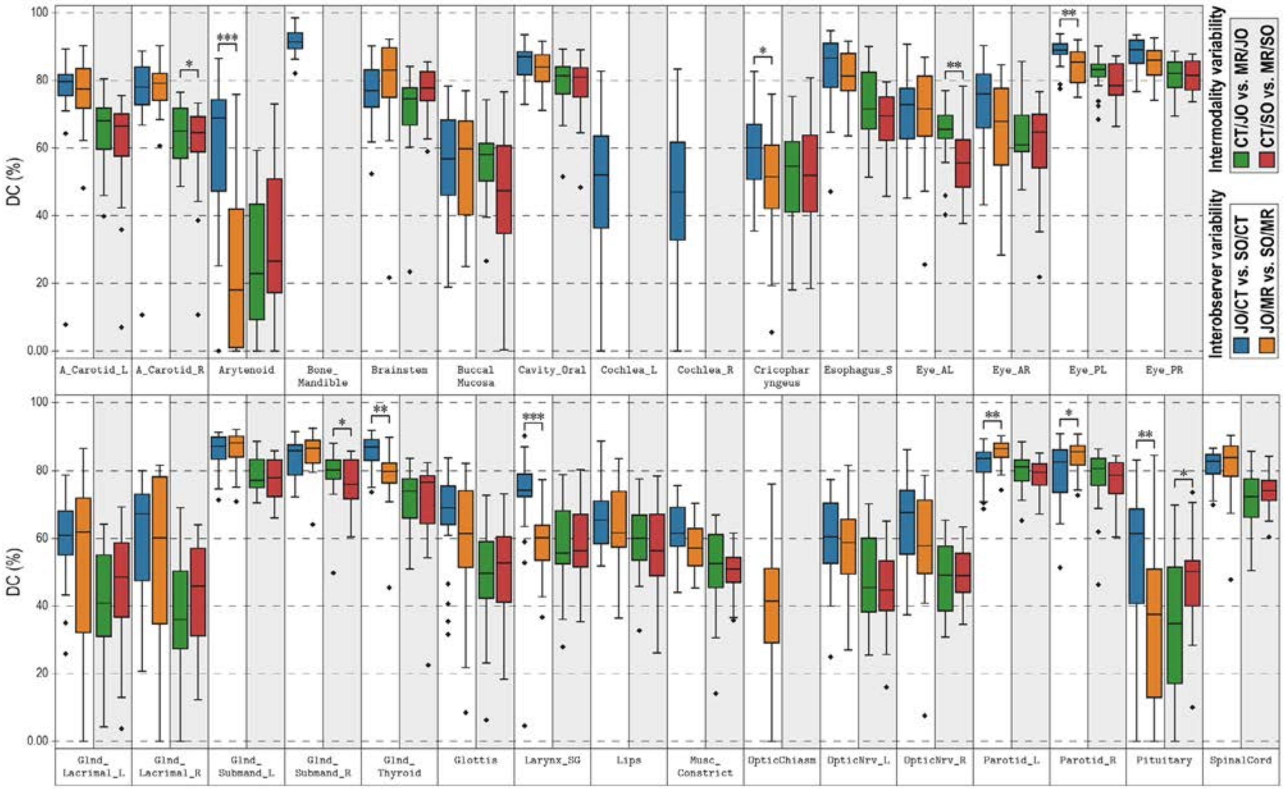

FIGURE 4 Box plots of the interobserver and intermodality variability in contouring organs-at-risk in CT and MR images of the same patients, performed by a JO and a SO, and reported in terms of the DC. The diamond symbols denote outliers (⧫), while asterisk symbols denote statistically significant differences ($* \rightarrow 0.05 > p > 0.01$; $** \rightarrow 0.01 > p > 0.001$; $*** \rightarrow 0.001 > p$). CT, computed tomography; DC, Dice coefficient; JO, junior observer; MR, magnetic resonance; SO, senior observer.

occurred for the contours of the mandible, brainstem, and submandibular and parotid glands, while larger variations were observed for the contours of the oral cavity, cochleae, PCMs, and glottic and supraglottic larynx. The only known study of interobserver contouring variability in MR images is the pilot study of Yuan et al.,[8] who evaluated contours of five OARs from T1-weighted MR images of eight healthy subjects by two observers. The extremely small reported variability[a] for the submandibular glands, parotid glands and spinal cord was probably due to the fact that these OARs can be distinctively observed in MR images. Although not set as the primary focus, several other studies investigated the interobserver variability for the purpose of validating the proposed novel OAR auto-segmentation methods.[27] On the other hand, the intermodality variability in contouring OARs from CT and MR images has, to the best of our knowledge, not been investigated yet.

4.1 Interobserver variability

In our study, we evaluated the interobserver contouring variability for a complete set of 31 OARs in the HaN region by basing our analysis on the comparison of two contours, obtained for the same OAR in the same CT or MR image by two observers with different experience, that is, JO and SO. After a detailed inspection, we can attribute several of the major resulting contouring differences to different guideline interpretations. Such examples include the selection of the axial cross-sections where one OAR ends and the other begins, for example, supraglottic larynx - glottic larynx, brainstem - spinal cord, PCMs - cricopharyngeal inlet or cricopharyngeal inlet - cervical esophagus, and detection of the correct OAR anatomy, for example, anterior segment of the eyeball - lens. Other reasons for contouring differences can be attributed to specific OARs without distinctive boundaries, for example, lips, buccal mucosa and oral cavity, where it was a personal choice of the observer to include their superior- or inferior-most parts. Finally, some OARs are often poorly visible, such as the optic chiasm, lacrimal glands or pituitary glands.

For individual OARs, the results can be best assessed by observing the reported DC and HD_{95} in Table 1, Figures 4 and 5. When observing CT contouring, it can be concluded that several OARs are subjected to a large interobserver variability, such as the arytenoids, buccal mucosa, cochleae, cricopharyngeal inlet, anterior segment of the eyeball, lacrimal glands, glottic and supraglottic larynx, optic nerves, and pituitary gland. On the other hand, moderate variability can be attributed

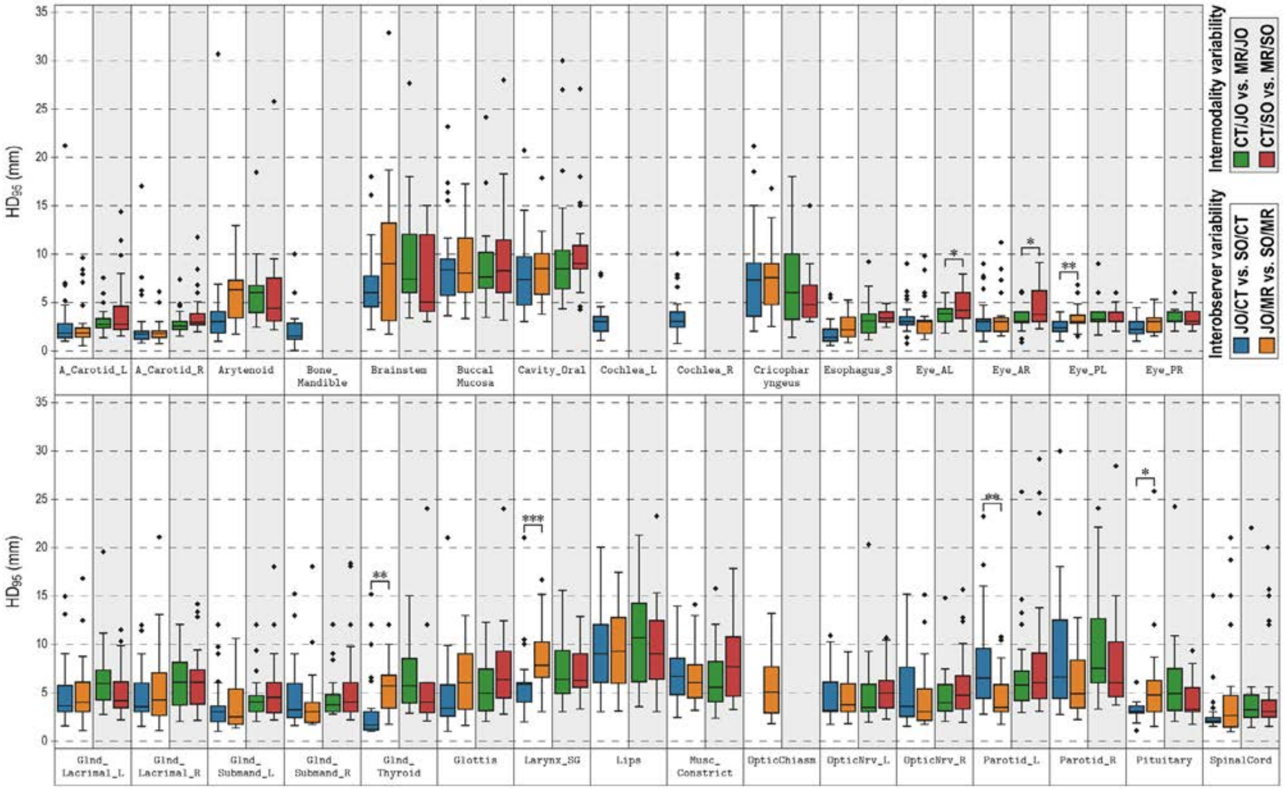

FIGURE 5 Box plots of the interobserver and intermodality variability in contouring organs-at-risk in CT and MR images of the same patients, performed by a JO and a SO, and reported in terms of the HD$_{95}$. The diamond symbols denote outliers (◆), while asterisk symbols denote statistically significant differences ($* \rightarrow 0.05 > p > 0.01$; $** \rightarrow 0.01 > p > 0.001$; $*** \rightarrow 0.001 > p$). CT, computed tomography; HD$_{95}$, 95-percentile Hausdorff distance; JO, junior observer; MR, magnetic resonance; SO, senior observer.

to the carotid arteries, brainstem, cervical esophagus, lips and PCMs, while a small variability was observed for the mandible, oral cavity, posterior segment of the eyeball, submandibular, thyroid and parotid glands, and spinal cord. Interestingly, almost the same conclusions can be drawn for the interobserver variability in MR contouring, excluding of course the mandible and cochleae that were contoured only in CT images, and additionally observing a large contouring variability of the optic chiasm, which was contoured only in MR images. The only major difference is that a moderate instead of small variability was observed for the thyroid gland, which however exhibits a considerable contrast in CT intensities against surrounding tissues.[4] The obtained results are not only in accordance with existing studies,[6–9] but also indicate that the level of interobserver variability does not considerably depend on the modality. It can be therefore concluded that an OAR is difficult to contour regardless of whether it is contoured in the CT or MR image (e.g., due to its small size, poor visibility, indistinctive boundaries).

4.2 | Intermodality variability

One of the challenges for the analysis of intermodality variability is the fact that it is based on registering the CT and MR images of the same patient laying in the same position during both imaging sessions, so that the corresponding OAR contours can be compared in the same coordinate system. We evaluated the quality of our registration by means of TRE, which is, in general, consistent with the AAPM Task Group 132 recommendations[44] stating that a mean TRE below 2 mm and a maximal TRE below 5 mm are desired for the majority of clinical applications. For only seven out of 27 patients, the mean TRE exceeded 2 mm, but the corresponding maximal TRE was always below 5 mm except for one patient (i.e., 5.3 mm). However, for this specific patient, two control points could not be defined due to a relatively small FoV of the MR image, which was probably also the reason for a less accurate registration. Nevertheless, the results indicate that the obtained registration enabled valid comparison and further variability analysis, and justify the choice of *elastix*[49] over more sophisticated approaches.[51,52]

For individual OARs, the results can be best assessed by observing the reported DC and HD$_{95}$ in Table 1, Figures 4 and 5. Large intermodality contouring variability can be attributed to the arytenoids, buccal mucosa, cricopharyngeal inlet, lacrimal glands, glottic larynx and pituitary gland, and small variability to the posterior segment of the eyeball, submandibular and parotid glands, and spinal cord. The majority of OARs were subjected to moderate contouring variability, that is, the carotid arteries, brainstem, oral cavity, cervical esophagus, anterior segment of the eyeball, thyroid gland, supraglottic larynx, lips, PCMs and optic nerves. As discussed for the interobserver variability, several of the differences in the intermodality variability can be also attributed to adherence to guidelines, especially for OARs with poor visibility or without distinctive boundaries in either CT or MR images. Although the results were, in general, within the same range, the overall statistics shows that the intermodality variability was slightly larger for SO in comparison to JO. Especially for OARs like the buccal mucosa, anterior segment of the eyeball and lips, the differences are more noticeable, which may reflect a lower level of SO attention when contouring OARs with indistinctive boundaries. However, for OARs that are deemed difficult to contour, such as the lacrimal and pituitary glands, the variability was smaller for SO, indicating that observer experience is in such cases important. Finally, although the intermodality variability is in general larger, it follows to a certain degree the interobserver variability, meaning that the agreement of contours between modalities is approximately the same as the agreement of contours between observers for a single modality.

4.3 | Implications for auto-segmentation

The variability analysis as reported in our study offers valuable insights for the development and validation of auto-segmentation methods, which is always a topic of great interest in the field of medical imaging.[27] The reported interobserver variability serves as a baseline performance that any auto-segmentation tool should aim to surpass. An efficient tool should, in practice, outperform the interobserver variability by leveraging the ability of deep learning models[18–26] to learn from multiple contouring styles, and generate segmentation masks that are more representative of both observers. Furthermore, the investigation of the intermodality variability provides important information regarding the visibility and distinguishability of OARs as represented by different imaging modalities. This knowledge can be leveraged to inform the design and optimization of auto-segmentation methods. By understanding such a perspective, researchers can tailor their approaches to exploit the strengths of each modality, and improve the accuracy, consistency and quality of deep learning auto-segmentation.[15,16] The baseline auto-segmentation experiments and results, performed and obtained for the images used in this study,[53] indicate that there is still room for improvements that can be leveraged by applying custom solutions, for example, tailored CT and MR modality feature fusion module techniques.[54]

Our study is not without limitations. First, although observers were asked to mimic clinical practice, contouring was performed retrospectively and the observers were aware that their results would not be used for RT planning. Second, there were only two contour sets available for each CT and MR image, and normally more contours would be required, preferably from multiple institutions, for a more reliable variability analysis. Finally, the variability analysis was performed by comparing the obtained contours, but preferably a consensus in the form of ground truth contours would represent a better comparison reference. Nevertheless, with the increasing use of MR in RT planning, our study indicates that OARs in the HaN can be contoured with a similar level of variability in either the CT or MR modality by either a JO or SO. The next steps are therefore to analyze the variability against a ground truth,[31] perform a multi-institutional variability study,[7] and evaluate how the interobserver and intermodality variability affect the RT dose calculation.[55]

5 | CONCLUSION

We evaluated the interobserver and intermodality variability in HaN OAR contouring from CT and MR images of the same patients. The major conclusion is that the contouring variability is, in general, similar for both image modalities (i.e., CT vs. MR), and that observer experience (i.e., JO vs. SO) does not considerably affect the contouring performance. Although we have identified considerable contouring differences for specific OARs, we can conclude that almost all OARs can be contoured with a similar degree of variability in either the CT or MR modality, which provides favorable support for MR images from the perspective of MR-only[38–40] and MR-guided RT.[42,43]

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research and Innovation Agency (ARIS) under grants J2-1732, P2-0232 and P3-0307, and partially by the Novo Nordisk Foundation under grant NFF20OC0062056. The study was approved by the Ethics Committee of the Institute of Oncology Ljubljana, Slovenia, under No. ERID-EK/139.

CONFLICT OF INTEREST STATEMENT

The authors have no relevant conflicts of interest to disclose.

REFERENCES

1. Vinod S, Jameson M, Min M, Holloway L. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiother Oncol*. 2016;121:169-179.
2. Ford E, Conroy L, Dong L, et al. Strategies for effective physics plan and chart review in radiation therapy: report of AAPM Task Group 275. *Med Phys*. 2020;47:e236-e272.

3. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71:209-249.
4. Brouwer C, Steenbakkers RJ, Bourhis J, 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.
5. Verhaart R, Fortunati V, Verduijn G, van Walsum T, Veenland J, Paulides M. CT-based patient modeling for head and neck hyperthermia treatment planning: manual versus automatic normal-tissue-segmentation. *Radiother Oncol*. 2014;111:158-163.
6. Brouwer C, Steenbakkers RJ, van den Heuvel E, et al. 3D variation in delineation of head and neck organs at risk. *Radiat Oncol*. 2012;7:32.
7. Nelms B, Tomé W, Robinson G, Wheeler J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys*. 2012;82:368-378.
8. Yuan J, Wong O, Law M, Ding Y, Cheung K, Yu, S,. Delineation variability of head and neck organs at risk on T1-weighted isotropic magnetic resonance images: a pilot study on healthy volunteers. *Int J Radiat Oncol Biol Phys*. 2016;96:E609.
9. van der Veen J, Gulyban A, Willems S, Maes F, Nuyts S. Interobserver variability in organ at risk delineation in head and neck cancer. *Radiat Oncol*. 2021;16:120.
10. Lin D, Lapen K, Sherer MV, et al. A systematic review of contouring guidelines in radiation oncology: analysis of frequency, methodology, and delivery of consensus recommendations. *Int J Radiat Oncol Biol Phys*. 2020;107:827-835.
11. Wong J, Fong A, McVicar N, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother Oncol*. 2020;144:152-158.
12. Lin D, Wahid KA, Nelms BE, et al. *E pluribus unum*: prospective acceptability benchmarking from the contouring collaborative for consensus in radiation oncology crowdsourced initiative for multiobserver segmentation. *J Med Imaging*. 2023;10:S11903.
13. Vandewinckele L, Claessens M, Dinkla A, et al. Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. *Radiother Oncol*. 2020;153:55-66.
14. Boukerroui D, Vasquez Osorio E, Brunenberg E, Gooding, M,. Analytic calculations and synthetic shapes for validation of quantitative contour comparison software. *Phys Imaging Radiat Oncol*. 2023;26:100436.
15. Duan J, Bernard ME, Castle JR, et al. Contouring quality assurance methodology based on multiple geometric features against deep learning auto-segmentation. *Med Phys*. 2023;50:2715-2732.
16. Duan J, Bernard ME, Rong Y, et al. Contour subregion error detection methodology using deep learning auto-segmentation. *Med Phys*. 2023;50:6673-6683.
17. Wu X, Udupa JK, Tong Y, et al. AAR-RT - a system for auto-contouring organs at risk on CT images for radiation therapy planning: principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. *Med Image Anal*. 2019;54:45-62.
18. Gao Y, Huang R, Yang Y, et al. FocusNetv2: imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images. *Med Image Anal*. 2021;67:101831.
19. Korte J, Hardcastle N, Ng S, Clark B, Kron T, Jackson P. Cascaded deep learning-based auto-segmentation for head and neck cancer patients: organs at risk on T2-weighted magnetic resonance imaging. *Med Phys*. 2021;48:7757-7772.
20. Nikolov S, Blackwell S, Zverovitch A, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. *J Med Internet Res*. 2021;23:e26151.
21. Jiang J, Elguindi S, Berry SL, et al. Nested block self-attention multiple resolution residual network for multiorgan segmentation from CT. *Med Phys*. 2022;49:5244-5257.
22. Koo J, Caudell JJ, Latifi K, et al. Comparative evaluation of a prototype deep learning algorithm for autosegmentation of normal tissues in head and neck radiotherapy. *Radiother Oncol*. 2022;174:52-58.
23. Udupa J, Liu T, Jin C, et al. Combining natural and artificial intelligence for robust automatic anatomy segmentation: application in neck and thorax auto-contouring. *Med Phys*. 2022;49:7118-7149.
24. Ye X, Guo D, Ge J, et al. Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. *Nat Commun*. 2022;13:6137.
25. Gardner M, Bouchta YB, Mylonas A, et al. Realistic CT data augmentation for accurate deep-learning based segmentation of head and neck tumors in kV images acquired during radiation therapy. *Med Phys*. 2023;50:4206-4219.
26. Zhao Q, Wang G, Lei W, et al. Segmentation of multiple organs-at-risk associated with brain tumors based on coarse-to-fine stratified networks. *Med Phys*. 2023;50:4430-4442.
27. Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Autosegmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. *Med Phys*. 2020;47:e929-e950.
28. Raudaschl P, Zaffino P, Sharp GC, et al. Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. *Med Phys*. 2017;44:2020-2036.
29. Joint Head and Neck MRI-Radiotherapy Development Cooperative, Kiser K, Meheissen MAM, Mohamed ASR, et al. Prospective quantitative quality assurance and deformation estimation of MRI-CT image registration in simulation of head and neck radiotherapy patients. *Clin Transl Radiat Oncol*. 2019;18:120-127.
30. Tang H, Chen X, Liu Y, et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nat Mach Intell*. 2019;1:480-491.
31. Podobnik G, Strojan P, Peterlin P, Ibragimov B, Vrtovec T. HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset. *Med Phys*. 2023;50:1917-1927.
32. Baroudi H, Brock KK, Cao W, et al. Automated contouring and planning in radiation therapy: what is 'clinically acceptable'? *Diagnostics*. 2023;13:667.
33. McGee K, Tyagi N, Bayouth JE, et al. Findings of the AAPM Ad Hoc committee on magnetic resonance imaging in radiation therapy: unmet needs, opportunities, and recommendations. *Med Phys*. 2021;48:4523-4531.
34. Hague C, McPartlin A, Lee LW, et al. An evaluation of MR based deep learning auto-contouring for planning head and neck radiotherapy. *Radiother Oncol*. 2021;158:112-117.
35. Zhong Z, He L, Chen C, et al. Full-scale attention network for automated organ segmentation on head and neck CT and MR images. *IET Image Process*. 2023;17:660-673.
36. Liu Y, Lei Y, Fu Y, et al. Head and neck multi-organ auto-segmentation on CT images aided by synthetic MRI. *Med Phys*. 2020;47:4294-4302.
37. Dai X, Lei Y, Wang T, et al. Automated delineation of head and neck organs at risk using synthetic MRI-aided mask scoring regional convolutional neural network *Med Phys*. 2021;48:5862-5873.
38. Lei Y, Harms J, Wang T, et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med Phys*. 2019;46:3565-3581.
39. Qi M, Li Y, Wu A, Lu X, Zhou L, Song T. Multi-sequence MR generated sCT is promising for HNC MR-only RT: a comprehensive evaluation of previously developed sCT generation networks. *Med Phys*. 2022;49:2150-2158.
40. Zhao Y, Wang H, Yu C, et al. Compensation cycle consistent generative adversarial networks (Comp-GAN) for synthetic CT

generation from MR scans with truncated anatomy. *Med Phys*. 2023;50:4399-4414.

41. Yuan S, Liu Y, Wei R, Zhu J, Men K, Dai J. A novel loss function to reproduce texture features for deep learning-based MRI-to-CT synthesis. *Med Phys*. 2024. https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.16850
42. Boeke S, Mönnich D, van Timmeren J, Balermpas P. MR-guided radiotherapy for head and neck cancer: current developments, perspectives, and challenges. *Front Oncol*. 2021;11:616156.
43. Huynh E, Boyle S, Campbell J, et al. Toward implementation of MR-guided radiation therapy for laryngeal cancer with healthy volunteer imaging and a custom MR-CT larynx phantom. *Med Phys*. 2022;49:1814-1821.
44. Brock K, Mutic S, McNutt T, Li H, Kessler M. Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM Radiation Therapy Committee Task Group No. 132. *Med Phys*. 2017;44:e43-e76.
45. Eekers D, In 't Ven L, Roelofs E, et al. The EPTN consensus-based atlas for CT- and MR-based contouring in neuro-oncology. *Radiother Oncol*. 2018;128:37-43.
46. Cardenas C, Mohamed ASR, Yang J, et al. Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. *Med Phys*. 2020;47:2317-2322.
47. Kieselmann J, Fuller C, Gurney-Champion O, Oelfke U. Cross-modality deep learning: contouring of MRI data from annotated CT data only. *Med Phys*. 2021;48:1673-1684.
48. Paczona V, Capala ME, Deák-Karancsi B, et al. Magnetic resonance imaging-based delineation of organs at risk in the head and neck region. *Adv Radiat Oncol*. 2022;8:101042.
49. Klein S, Staring M, Murphy K, Viergever M, Pluim J. Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging*. 2010;29:196-205.
50. Mayo C, Moran JM, Bosch W, et al. American Association of Physicists in Medicine Task Group 263: standardizing nomenclatures in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2018;100:1057-1066.
51. Lee D, Alam S, Jiang J, Cervino L, Hu Y-C, Zhang P. Seq2Morph: a deep learning deformable image registration algorithm for longitudinal imaging studies and adaptive radiotherapy. *Med Phys*. 2023;50:970-979.
52. Yang S, Li H, Chen S, et al. Multiscale feature fusion network for 3D head MRI image registration. *Med Phys*. 2023;50:5609-5620.
53. Podobnik G, Ibragimov B, Strojan P, Peterlin P, Vrtovec T. Segmentation of organs-at-risk from CT and MR images of the head and neck: baseline results. In: *19th IEEE Symposium on Biomedical Imaging – ISBI 2022*. IEEE; 2022.
54. Podobnik G, Strojan P, Peterlin P, Ibragimov B, Vrtovec T. Multimodal CT and MR segmentation of head and neck organs-at-risk. In: *26th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Vol 14223. LNCS, Springer; 2023:745-755.
55. Babier A, Zhang B, Mahmood R, et al. OpenKBP: the open-access knowledge-based planning grand challenge and dataset. *Med Phys*. 2021;48:5549-5561.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Podobnik G, Ibragimov B, Peterlin P, Strojan P, Vrtovec T. vOARiability: Interobserver and intermodality variability analysis in OAR contouring from head and neck CT and MR images. *Med Phys*. 2024;51:2175–2186. https://doi.org/10.1002/mp.16924