Taylor & Francis
Taylor & Francis Group

Validated guidelines for tumor delineation on magnetic resonance imaging for laryngeal and hypopharyngeal cancer

Elise Anne Jager, Hans Ligtenberg, Joana Caldas-Magalhaes, Tim Schakel, Marielle E. Philippens, Frank A. Pameijer, Nicolien Kasperts, Stefan M. Willems, Christiaan H. Terhaard & Cornelis P. Raaijmakers

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Validated guidelines for tumor delineation on magnetic resonance imaging for laryngeal and hypopharyngeal cancer

Elise Anne Jager[a], Hans Ligtenberg[a], Joana Caldas-Magalhaes[a], Tim Schakel[a], Marielle E. Philippens[a], Frank A. Pameijer[b], Nicolien Kasperts[a], Stefan M. Willems[c], Christiaan H. Terhaard[a] and Cornelis P. Raaijmakers[a]

[a]Department of Radiation Oncology, University Medical Center Utrecht, Utrecht, The Netherlands; [b]Department of Radiology, University Medical Center Utrecht, Utrecht, The Netherlands; [c]Department of Pathology, University Medical Center Utrecht, Utrecht, The Netherlands;

ABSTRACT

Background: Validation of magnetic resonance imaging (MRI) and development of guidelines for the delineation of the gross tumor volume (GTV) is of utmost importance to benefit from the visibility of anatomical details on MR images and to achieve an accurate GTV delineation. In the ideal situation, the GTV delineation corresponds to the histopathologically determined 'true tumor volume'. Consequently, we developed guidelines for GTV delineation of laryngeal and hypopharyngeal tumors on MRI and determined the accuracy of the resulting delineation of the tumor outline on histopathology as gold standard.

Material and methods: Twenty-seven patients with T3 or T4 laryngeal/hypopharyngeal cancer underwent a MRI scan before laryngectomy. Hematoxylin and eosin sections were obtained from surgical specimens and tumor was delineated by one pathologist. GTV was delineated on MR images by three independent observers in two sessions. The first session (del1) was performed according to clinical practice. In the second session (del2) guidelines were used. The reconstructed specimen was registered to the MR images for comparison of the delineated GTVs to the tumor on histopathology. Volumes and overlap parameters were analyzed. A target margin needed to assure tumor coverage was determined.

Results: The median GTVs (del1: 19.4 cm^3, del2: 15.8 cm^3) were larger than the tumor volume on pathology (10.5 cm^3). Comparable target margins were needed for both delineation sessions to assure tumor coverage. By adding these margins to the GTVs, the target volumes for del1 (median: 81.3 cm^3) were significantly larger than for del2 (median: 64.2 cm^3) ($p \leq 0.0001$) with similar tumor coverage.

Conclusions: In clinical radiotherapy practice, the delineated GTV on MRI is twice as large as the tumor volume. Validated delineation guidelines lead to a significant decrease in the overestimation of the tumor volume.

Radiotherapy is developing towards a precision technique, delivering a high radiation dose to the tumor with tight treatment margins. Therefore, accurate three-dimensional (3D) target volume definition has become a crucial step. Target delineation, however, is one of the largest sources of uncertainty in head-and-neck cancer radiotherapy [1]. Even with the introduction of new imaging techniques the inter-observer variability remains relatively large [2–5]. For an accurate definition of the target volume, validation of gross tumor volume (GTV) delineation is fundamental. For this validation, histopathology is the gold standard. However, this validation is a complex procedure and few studies have been performed for head-and-neck cancer [5–7]. For radiotherapy purposes, a detailed comparison of histopathology and imaging is required to validate the actual size, shape and location of the tumor. This involves 3D reconstruction of the pathology specimen after slicing and matching of this specimen to the in vivo images.

Currently, computed tomography (CT) is the standard imaging modality for staging and delineation of laryngeal carcinoma. Magnetic resonance imaging (MRI) is gaining ground in radiation oncology as availability and image quality have been improved dramatically [7,8]. Several studies [2,3,5,9–11] have been performed, using various imaging modalities, to determine the agreement among observers when delineating the GTV in head-and-neck cancer. Studies in which CT was compared to MRI demonstrated contradictory results concerning the added value of MRI based on the inter-observer variation [3,4,11]. However, a considerable number of studies reported an increased visualization of boundaries between different tissues and increased soft tissue contrast with MRI compared to CT [3,4,9–12]. Thus, in head-and-neck cancer, especially in the laryngeal region, there is no consensus on the added value of MRI for improving agreement between observers. However, there is agreement that visualization of soft tissue structures is better

on MRI. A possible explanation for this contradiction is the increased visibility of anatomical details on MR images in combination with the lack of clear interpretation and delineation guidelines, what might result in an increased variability [4]. According to a study performed by Rasch et al. [3], the use of guidelines for delineation might be of value.

The aim of this study was to determine the accuracy of GTV delineation on MR images using delineation guidelines.

Material and methods

Patient selection

Thirty-six patients, treated with a total laryngectomy (TLE) for primary T3 or T4 laryngeal or hypopharyngeal cancer, were included in this study according to the inclusion criteria as further described in this section.

The first six patients were excluded for optimization of the pathology imaging registration procedure. Patient 12 was excluded because of a biopsy between preoperative imaging and surgery. The tumor of Patient 21 was too large for our whole mount standard analysis. The exclusion of Patient 30 was due to an incohesive tumor.

Resultantly, 27 patients (median age 62 years, range 49–79 years, two female and 25 male) with primary T3 (N = 4) or T4 (N = 23) laryngeal (supraglottic including the glottic: N = 2, supraglottic: N = 7, transglottic: N = 4, glottic: N = 2) or hypopharyngeal (N = 12) carcinoma were used in this study for analysis. The patients underwent a TLE as a primary treatment in our institution between March 2009 and August 2014. The patients are numbered according to their study numbers.

This study is part of an extensive image validation project using positron emission tomography (PET), CT and MR images. Criteria of exclusion were: contraindications for MRI at 1.5 Tesla, contraindication for CT contrast administration as defined in the protocols of the radiology department and insulin dependent diabetes mellitus. The study was approved by our ethical review board.

Preoperative image acquisition

Before surgery, all patients underwent an MRI scan with T1-weighted (T1w) images in transverse, sagittal and coronal directions as well as transverse T2-weighted (T2w) and T1-weighted Gadolinium-enhanced images (T1w-Gd) while immobilized in a head and shoulder radiotherapy mask and using two small flexible surface coils (Intera, Philips Medical Systems) [12]. The MRI scan was made on a 1.5 Tesla spectrometer producing high resolution images. Imaging parameters can be found in Supplementary (available online at http://www.informahealthcare.com) Table 1. The median time interval between MRI and surgery was 14 days (range 1–35 days).

Processing of the surgical specimen

Macroscopy

The pathology procedure was described in detail previously [6]. Briefly, the fresh larynx specimen was fixated in 10% formaldehyde directly after surgery. After fixation, the specimen was embedded in an agarose block and transversely sliced in approximately 3-mm thick slices, which were subsequently photographed and digitized.

Microscopy

From each 3-mm thick slice, a 4-μm section was obtained and stained with hematoxylin and eosin (H&E). Histopathological analysis was performed by a dedicated head-and-neck pathologist who delineated all tumor tissue on the H&E sections using a microscope. The delineated tumor is referred to as tumor$_{H\&E}$. The H&E sections were subsequently digitized. These delineations were used as gold standard to validate the delineations on MRI.

Image registration and 3D reconstruction

First, the H&E sections were registered to the corresponding thick-slice photos using cartilage landmarks to perform a point-based rigid registration with scaling. For 3D reconstruction of the pathology specimen and registration to the MR images we used a previous published method [6]. To optimize the registration, the rigid registration between the pathology and MRI was visually verified and manually adjusted to correct for deformations of the pathology specimen. This was done by rigidly registering the tumor$_{H\&E}$ contour to the MRI and then manually adjusting the tumor$_{H\&E}$ contour to anatomical structures visible both on histopathology and MRI. The manual adjustments were done in consensus between two experts.

Delineation of GTV

Three dedicated and MRI-trained head-and-neck specialists independently delineated the GTV on MRI in Volumetool [13] an in-house developed clinical and research software application.

After a pilot study of five patients (not included in this study) followed by a meeting where the individual delineations were discussed, the approach was fine-tuned according to clinical practice and the experience of the observers. Consensus was reached on the following approach for the first delineation session (del1) using the various MRI sequences:

1. T1w with respect to the anatomy: delineation of all abnormal anatomy (i.e. presumed tumor).
2. T2w with respect to signal intensity: delineation of all hyper intense areas, with the exception of evident stasis of saliva. No differentiation between the primary tumor and surrounding edema and/or soft tissue swelling adjacent to the presumed tumor bulk.
3. T1w-Gd with respect to signal intensity: delineation of all enhancing areas that were suspect for tumor/tumor growth based on clinical experience of the observer.

T1w, T2w and T1w-Gd transverse and T1w sagittal and coronal images were available for image analysis.

The sagittal and coronal views permit an overview of the consistency of the delineations in non-axial planes and

Table 1. Volumes and overlap analysis.

Patient study number	Tumor_H&E (cm³)	del1				del2			
		GTV (cm³)	Sens	PPV	CI	GTV (cm³)	Sens	PPV	CI
15	3.4	12.8	0.99	0.26	0.26	7.3	0.94	0.44	0.43
9	3.4	9.2	0.98	0.36	0.36	5.8	0.89	0.52	0.49
17	4.7	8.2	0.93	0.53	0.51	6.2	0.87	0.66	0.60
14	5.2	13.4	0.89	0.34	0.33	9.5	0.87	0.49	0.45
19	5.5	18.3	0.97	0.29	0.29	16.4	0.94	0.34	0.33
13	5.9	10.5	0.91	0.52	0.49	7.5	0.81	0.63	0.55
8	6.7	19.4	0.97	0.34	0.33	10.1	0.93	0.62	0.59
16	7.3	16.7	0.87	0.38	0.36	13.9	0.90	0.47	0.45
23	7.3	17.2	0.98	0.42	0.42	18.2	0.98	0.40	0.40
27	7.7	25.5	0.96	0.30	0.29	18.9	0.95	0.39	0.38
35	7.9	15.1	0.93	0.49	0.47	13.3	0.92	0.57	0.54
29	8.5	22.4	0.94	0.37	0.36	13.3	0.82	0.53	0.47
10	8.7	26.4	0.99	0.33	0.33	15.5	0.96	0.54	0.52
22	10.5	17.5	0.91	0.55	0.52	13.0	0.86	0.69	0.62
34	10.6	15.2	0.95	0.66	0.64	12.9	0.89	0.74	0.68
20	12.9	13.3	0.70	0.68	0.52	10.9	0.65	0.77	0.54
32	13.6	18.9	0.80	0.61	0.52	15.8	0.83	0.72	0.62
7	14.2	31.6	0.98	0.44	0.44	23.1	0.94	0.58	0.56
31	15.2	34.6	0.98	0.44	0.43	27.3	0.95	0.53	0.52
26	15.8	26.9	0.87	0.51	0.48	20.7	0.84	0.66	0.58
25	16.1	33.3	0.98	0.48	0.47	27.5	0.92	0.54	0.52
11	17.7	36.0	0.99	0.49	0.49	24.7	0.95	0.69	0.67
28	29.0	41.8	0.97	0.68	0.66	35.9	0.94	0.76	0.73
18	30.0	46.3	0.96	0.63	0.61	40.5	0.95	0.70	0.67
24	39.6	71.5	0.94	0.52	0.51	59.0	0.91	0.61	0.58
33	47.6	58.1	0.90	0.74	0.68	50.9	0.88	0.82	0.74
36	68.6	86.4	0.88	0.70	0.64	67.2	0.78	0.80	0.66
median	10.5	19.4	0.95	0.49	0.47	15.8	0.91	0.61	0.55

Volumes and overlap analysis for delineation on MRI without (del1) and with (del2) guidelines. Patients are listed by tumor size. Tumor_H&E: tumor volume delineated by a pathologist on H&E sections. CI: mean conformity index determined for the overlap between the GTV contour of the observers and the pathology contour; GTV: mean GTV volumes for three observers delineated on MRI. Overlap analysis; PPV: mean positive predictive value of the three observers; Sens: mean sensitivity of the three observers.

determination of the cranio-caudal tumor extension (Figure 1). The GTV was delineated on the T1w-Gd image and was modified according to the other MRI sequences. For all study patients, the observers were aware of the findings during endoscopy, which was performed before the MRI.

After an interval of one year, the tumors were re-delineated (del2) this time by using guidelines derived from criteria for diagnosis of neoplastic cartilage invasion on MR images [14] applied on soft tissue structures. The following delineation guidelines were used to analyze cartilage and soft tissue structures. The delineation guidelines were:

1. T2w or T1w-Gd with signal intensity higher than that of the adjacent tumor bulk was considered to indicate inflammation and was not included in the GTV (Figure 2).
2. Areas of strongly increased T2w signal intensity in the immediate surroundings of the tumor were considered to be stasis of saliva or trapped secretions and these were not included in the GTV (Figure 3).

Patients 19 and 20 were re-delineated because of flaws in one of the observers' delineations.

Analysis

Volumetric analysis and overlap analysis

The volume of the tumor_H&E was compared to the various GTVs. Three parameters were calculated to quantify the overlap. The sensitivity (1), which reflects the part of the tumor_H&E volume that was included in the GTV; the positive predictive value (PPV) (2), which reflects the part of the GTV that actually was tumor; and the conformity index (CI) (3) that quantifies the similarity between the two volumes (Figure 4).

$$Sensitivity = \frac{|tumorH\&E \cap GTV|}{|tumorH\&E|} \quad (1)$$

$$Positive\ predictive\ value = \frac{|tumorH\&E \cap GTV|}{|GTV|} \quad (2)$$

$$Conformity\ index = \frac{|tumorH\&E \cap GTV|}{|tumorH\&E \cup GTV|} \quad (3)$$

Distance analysis

From the GTV and the tumor_H&E contour, a common contour was derived for each patient per observer.

Two types of distances were calculated. Type I was the distance from each point of the contour of the tumor_H&E to the closest point on the common contour in 3D. This measure quantifies the distances resulting from missing tumor (underestimation). Type II was the distance from each point on the contour of the GTV to the closest point of the common contour and quantifies the overestimation (Figure 4).

The 95th percentiles of the distances per patient for each observer were determined. For each patient an average 95th percentile distance was calculated. The median value (p50) of the values (the average 95th percentile) of all patients was determined as the final type I and type II distance (p95 type I/II distances).
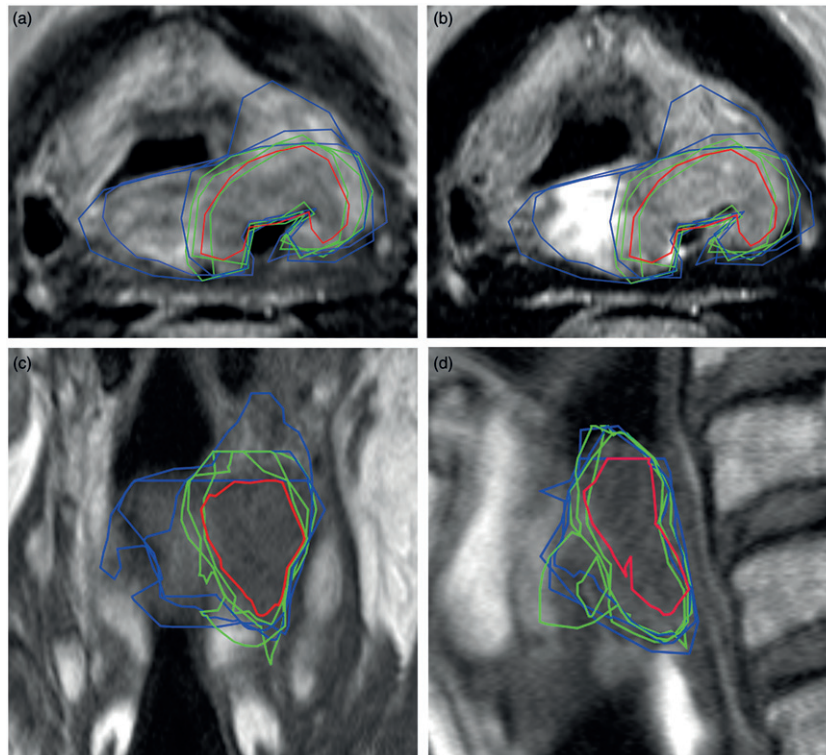
Figure 1. Axial MR image at the level of the supraglottic larynx showing a T3 hypopharyngeal tumor centered at the left piriform sinus. GTV delineations of three observers according to clinical practice (del1: blue contours) and with guidelines (del2: green contours) and tumor delineations made by the pathologist (tumor$_{H\&E}$: red contour). T1w Gd-enhanced (a) and T2w (b) MR images showing a large asymmetric area of soft tissue on the right site of the tumor. This area was included in the first round (del1) of GTV delineations by two of the three observers. However, based on the increased signal intensity compared to the tumor, this area was excluded in the second round (del2) by all the three observers. (c,d) corresponding coronal and sagittal T1w reformations with contours permitting an overview of the delineations in non-axial planes. The delineated volumes of the second session (del2), where guidelines were used, show a large reduction in delineated volumes and reduction in overestimation of the tumor (red contour) compared to the delineated volumes in the first round (del1).

The 95th percentile of the distances indicates that on average 95% of the outer contour (i.e. the tumor$_{H\&E}$ contour for type I distance, and the GTV contour for type II distance) lies within these final type I and type II distances.

The 95th percentile as a cutoff point for the distance analysis was chosen to limit the influence of residual registration errors and deformations.

Derivation of target margin

From the type I distance, a margin that accounts for the underestimation was derived.

First, the 95th percentiles of the type I distances per patient for each observer were determined. Subsequently, a gamma distribution over these values was assumed per observer. The 95th percentile of this gamma distribution was determined per observer, resulting in three observer depended margins. Averaging these margins resulted in the margin accounting for the underestimation, i.e. the target margin.

For determination of the target margin, the 95th percentile of the gamma distribution was chosen to prevent extensive margins caused by the worst case and, consequently, overestimation of nearly all tumors.

Target volume determination

The target margin was added to the GTV contour of observer with the most average type I distances for delineation

without and with guidelines. By adding the target margin around the GTV contour while correcting for anatomical boundaries such as air, pharyngeal constrictor muscles and vertebrae, target volumes were calculated.

Statistical analysis

Pairwise comparison of the volumes, overlap and distance parameters between the tumor$_{H\&E}$ and GTV were performed with a two-tailed Wilcoxon signed rank test.

Results

Volumetric analysis and overlap analysis

A large variation in tumor volume was observed (Table 1). In general a significant overestimation of the tumor was observed, which was reduced when guidelines were used for delineation (p < .0001) (Table 1, Figure 1–3). In some cases very good agreement between pathology and MRI delineations was observed and the delineations nearly completely covered the tumor (maximum sensitivity of 0.99). In other cases larger deviations were found and in the worst case 35% of the tumor was not included in the GTV (sensitivity of 0.65).

The median sensitivity of the three observers was larger for del1 (0.95) compared to del2 (0.91) (p < .0001). The amount of overestimation (PPV) varied largely although generally the overestimation for del2 was reduced compared to
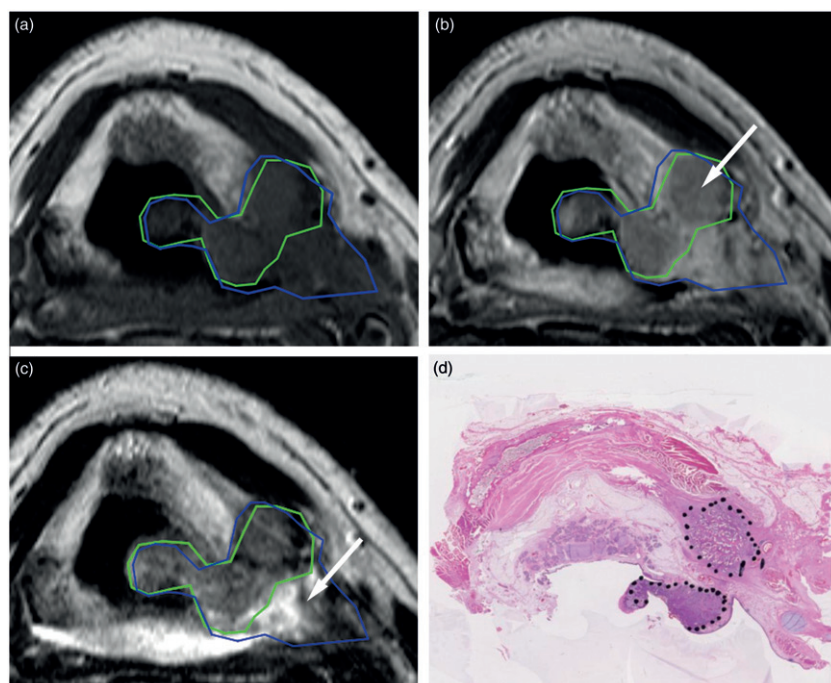
Figure 2. Patient with T4 hypopharyngeal tumor centered at the left aryepiglottic fold. GTV delineations of one observer according to clinical practice (del1: dark grey contour) and with guidelines (del2: light grey contour) and tumor delineations made by the pathologist (tumor$_{H\&E}$: black dotted contour). (a) axial T1w MR image at the level of the supraglottic larynx showing a tumor centered in the left aryepiglottic fold, (b) axial T1w Gd-enhanced MR image at the same level as (a), showing extensive involvement of the left paralaryngeal space (arrow). (c) T2w axial MR image at the same level as (a,b). In the first session of GTV delineation (del1) the observer included a large area on the left (arrow). In the second round (del2), this lateral extension was excluded because of high T2 signal intensity (considered as stasis of saliva or trapped secretions according to the guidelines). This separated the tumor from the left extralaryngeal soft tissues, which were not involved. (d) corresponding H&E section with the tumor$_{H\&E}$ contour.

the del1 (p < .0001). The median CI increased by the use of guidelines (del1: 0.47, range 0.26–0.68, del2: 0.55, range 0.33–0.74) (p < .0001) (Table 1).

For Patients 20, 24, 26, 32, 33 and 36 larger tumor volumes were missed compared to other patients. The mean missed tumor volume of these patients was 4.0 cm^3 (SD 2.2) for del1 and 5.6 cm^3 (SD 4.7) for del2. The other patients, for whom less tumor volume was missed, the missed tumor volume was 0.45 cm^3 (SD 0.35) (del1) and 0.55 cm^3 (SD 0.48) (del2). Visual inspection of the tumors with larger missed tumor volumes showed that Patient 20, the irregular shape of the tumor for might have hampered an accurate delineation resulting in the lowest sensitivity of 0.70 (del1) and 0.65 (del2). For Patient 24 [sensitivity: 0.94 (del1), 0.91(del2)] and 36 [sensitivity: 0.88 (del1), 0.78 (del2)], the latter being an outlier with respect to tumor size (68.6 cm^3) and with on average the largest missed volume [7.9 (del1), 14.8 (del2) cm^3], the tumors encountered more deformations and registration problems. Furthermore, parts of tumor-invaded thyroid cartilage were not included in the GTV. For Patient 26 [sensitivity: 0.87 (del1), 0.84 (del2)] parts of the tumor were missed in retropharyngeal direction and for Patient 32 [sensitivity: 0.80 (del1), 0.83 (del2)] the tumor extended into the pharynx and the constrictor muscles crossing the midline. This extension was not clearly visible on the MR images and was missed by the observers. The tumor of Patient 33 [sensitivity: 0.90 (del1), 0.88 (del2)] showed extra laryngeal growth, which was missed.

Distance analysis

For type I distances, quantifying the distances resulting from missing tumor, the median (p50) of the 95% of the measured distances of the tumor extensions outside the GTV (distance type I, p95), were smaller than 1.5 mm (del1) and 2.0 mm (del2) (Table 2). The maximal distances from the GTV border where tumor extensions were found were 14.7 mm (del1) and 13.0 mm (del2). The median (p50) of the 95th percentile for type II distances, indicating the overestimation, was significantly smaller for delineation with guidelines (6.5 mm) compared to delineation without guidelines (10.3 mm) (p < .0001) (Table 2).

Target margin and target volume analysis

Required target margins were comparable for delineation according to clinical practice (del1: 7.0 mm) and with guidelines (del2: 6.9 mm) (Table 2: p95 of the distance type I, p95). When applying these margins, significantly larger target volumes were calculated for del1 (median: 64.4 cm^3, range 33.5–194.7 cm^3) compared to del2 (median: 50.3 cm^3, range 26.2–167.5 cm^3) (p < .0001), with a decrease in overestimation (median del1: 53.9 cm^3, median del2: 39.8 cm^3) of more than 25% (p < .0001). By using this target margin, the maximal missed tumor volumes was 0.28 cm^3 (del1) and 0.32 cm^3 (del2) for Patient 20 and Patient 26 respectively. Only a small fraction of the total tumor volume (del1: 0.17%, del2: 0.23%) was missed after applying a target margin. These missed
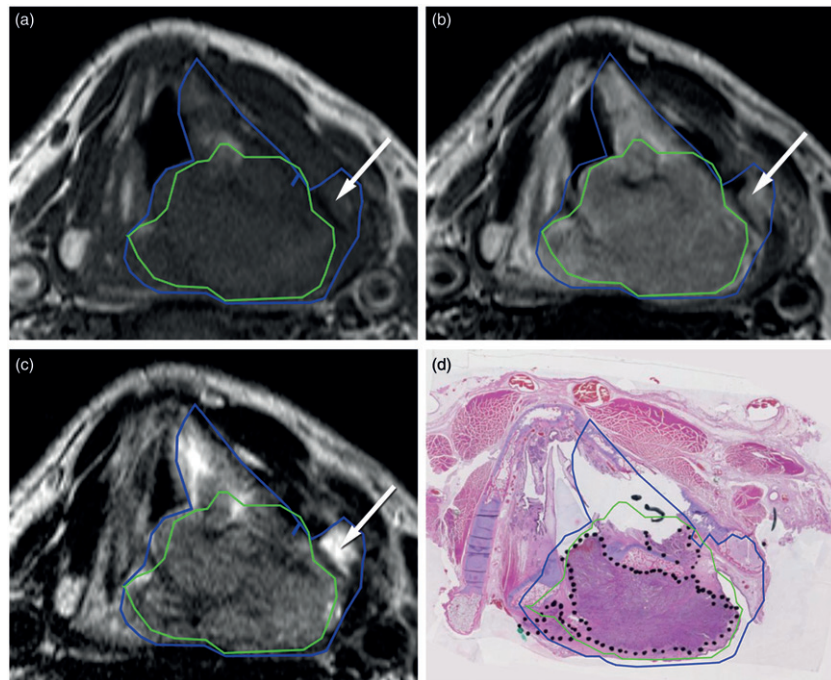
Figure 3. Patient with a T4 hypopharyngeal tumor centered in the left piriform sinus. GTV delineations of one observers according to clinical practice (del1: dark grey contour) and with guidelines (del2: light grey contour) and tumor delineations made by the pathologist (tumor$_{H\&E}$: black dotted contour). (a,b) T1w and T1w Gdenhanced axial MR image of the larynx at the level of the false vocal cord showing a tumor centered in the left piriform sinus. Note: enhancing area in the left thyroid laminae (arrow) adjacent to the tumor on the T1w Gd image. This area was included in the first round (del1) of GTV delineation by two of the three observers. The signal intensity of this area is higher compared to the tumor. Therefore in the second round (del2) considered as inflammation and excluded from the GTV. (c) T2w axial MR image of the larynx at the same level as (a,b). The same area (arrow) displays a higher T2 signal intensity compared to the tumor. Based on the guidelines this area was excluded in the second round (del2). (d) corresponding H&E section with the tumor$_{H\&E}$ contour.
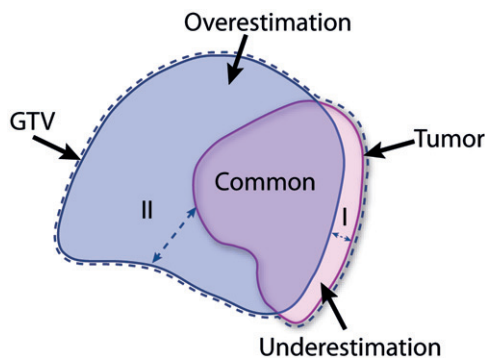

Figure 4. Schematic diagram of overlap and distance parameters. GTV: gross tumor volume (dark grey) and tumor$_{H\&E}$ (light grey). Blue dotted line reflects the encompassing contour. Arrow I: type I distance, arrow II: type II distance. The common volume is enclosed by a contour indicating the overlap between the GTV and tumor$_{H\&E}$ contours. Overestimation is delineation of tissue by the observers that is not tumor while underestimation is the part of the tumor that is not included in the GTV delineation, i.e. missed volume.

volumes were not significantly different (p = .53) between del1 and del2 and were found in five of the 27 patients [number 20 (del1), 26 (del2), 32, 36 and 24].

Discussion

In clinical practice, the delineated GTV on MRI is twice as large as the tumor volume determined on pathology (Figure 1, Table 1). Clear delineation guidelines lead to a decrease of tumor volume overestimation and a smaller target volume.

For GTV delineation on MRI according to current clinical practice, the overestimation results in relatively large radiation fields. This overestimation might be explained by the inclusion of tissue with abnormal appearance other than tumor tissue, e.g. edema or inflamed tissue, in the GTV. Subsequently, delineation guidelines were used with the aim to differentiate between tumorous and non-tumorous tissue and, consequently, the exclusion of non-tumorous tissue from the GTV. The PPV implied that, on average, 49% of the delineated GTV on MRI consisted of tumor tissue while 5% of the tumor was missed (sensitivity). The delineation guidelines increased the percentage of tumor in the GTV with 12% at the expense of a slight decrease in sensitivity of 4% (0.95–0.91).

An explanation for this decrease in sensitivity is that by reduction of the GTV volume, the chance of missing tumor was increased for del2 compared to del1. By decreasing the delineated volumes, the effect of the remaining registration errors and deformations might become more apparent. Furthermore, by trying to delineate the tumor as accurately as possible, small delineation inconsistencies become apparent even though the correct tumor border has been delineated. The CI, however, increased from 0.47 to 0.55 indicating more resemblance of the GTV with the true tumor volume by the use of guidelines. This parameter takes into account the overestimation as well as the underestimation.

For some patients it was more difficult to include all tumor tissue in the GTV. In a previous study, in which the inter-observer variation between pathologists for delineation of the tumor on H&E sections was investigated, the largest

Table 2. Distance analysis.

Patient study number	Distance type I, p95 (mm)		Distance type II, p95 (mm)		Distance type I, max (mm)	
	del1	del2	del1	del2	del1	del2
15	0.1	0.8	10.6	6.8	0.9	2.4
9	0.4	1.5	8.9	4.9	1.8	4.1
17	1.5	1.9	6.2	4.2	3.8	3.3
14	2.0	2.7	8.9	6.0	4.9	5.7
19	1.0	2.0	11.8	12.9	4.3	4.3
13	2.0	3.1	9.6	6.6	5.4	5.3
8	0.6	1.2	20.5	5.1	2.8	3.4
16	2.9	2.1	12.9	8.7	4.8	4.3
23	0.6	0.3	11.8	12.5	3.3	2.8
27	0.6	1.1	12.0	9.4	2.4	3.1
35	1.6	1.4	7.5	6.1	3.8	3.5
29	1.5	4.7	11.4	9.1	3.9	8.5
10	0.1	1.0	12.6	6.2	3.3	4.4
22	2.7	3.9	9.4	5.3	7.2	7.9
34	1.3	2.2	6.9	6.1	3.3	4.8
20	6.5	6.1	6.9	3.4	12.0	10.3
32	7.7	6.4	8.1	4.8	14.7	12.3
7	0.3	1.7	9.6	6.7	4.1	5.9
31	0.4	1.2	13.3	8.9	3.7	4.8
26	8.0	7.2	11.9	7.0	13.2	12.2
25	0.7	2.0	7.9	6.5	3.7	5.4
11	0.4	1.0	10.6	5.5	1.9	3.3
28	1.0	1.3	6.1	4.5	3.2	4.3
18	1.9	1.8	14.1	11.1	5.0	5.3
24	2.3	3.7	13.8	9.2	9.1	11.4
33	3.3	3.6	6.3	5.1	8.7	9.3
36	3.5	6.3	10.3	8.1	10.0	13.0
median (p50)	1.5	2.0	10.3	6.5	3.9	4.8
p95	7.0	6.9			10.2	8.3

From three individual GTV contours and the tumor$_{H\&E}$ contour a common and an encompassing contour were derived. Type I distance: distance from each point of the common contour to the closest point on the contour of the tumor$_{H\&E}$. Type II distance: distance from each point of the common contour to the closest point on the contour of the GTV. Max: maximum distance measured for this patient, mean of the three observers. Distance type I and distance type II, p95 (columns): 95% of the mean distances of the three observers are smaller than the presented value. P95 (value in de last row): the 95th percentile of the 95th percentile of the mean distances of the three observers (distance type I, p95 and distance type II, p95). The target margins were based on the p95 of the distance type I, p95. P95 (last row) of distance type I, max: 95th percentile of mean of the maximum type I distances of the three observers. P50: the median values. The patients are listed by tumor size.

variation was observed for Patient 20. In general, the inter-observer variation between pathologists was small [15]. However, the irregular shape of this tumor, with a large amount of swelling around the tumor, hampered an accurate delineation on both the H&E sections and the MR images. Another difficulty was tumor invading the cartilage structures leading to misses (Patients 24 and 36). In clinical radiotherapy practice, radiation oncologists sporadically delineate tumors with cartilage invasion because these tumors are seldom primary treated with radiation. Lack of experience in evaluating and delineating cartilage invasion on MRI might have caused this finding.

Important work on validation of imaging techniques for head-and-neck tumors has been performed by Daisne et al. [7]. A comparison between photographs of nine TLE specimens and delineated GTVs on MRI resulted in a lower sensitivity (0.83) and PPV (0.44) than what was found in our study. A shortcoming of this study was that microscopic evaluation was not performed. Furthermore, in our work a dedicated MRI protocol was used for GTV delineation in radiotherapy

[12]. Overestimation of the tumor on MRI has also been reported for oropharyngeal and oral cavity cancer by Seitz et al., although this overestimation was smaller than observed in our study (tumor volume: $16.6 \pm 18.6 \, cm^3$, GTV MRI: $17.6 \pm 19.1 \, cm^3$) [16].

In our study, patients with surgery as primary treatment were included. For lower tumor stages (T1 and T2T1b-T3) radiotherapy usually is the treatment of choice and the lower stage tumors mostly concern smaller tumors. In this study the use of guidelines results in a reduction of overestimation for smaller tumors as well as for larger tumors, indicating that these guidelines could be of added value for reduction of overestimation for laryngeal and hypopharyngeal tumors that are treated with radiation therapy alone.

The median time interval between MRI and surgery was 14 days (range 1–35 days). For 24 patients surgery was performed within 17 days after MRI acquisition, whereas for three patients this time interval exceeded one month. This increased time interval might have enabled significant tumor growth between MRI and surgery, although no clear data on tumor growth rate are available. If the tumor on MRI was smaller than the tumor volume identified by histopathology, this might have caused a decrease in overestimation or an increase in tumor volume that was missed (decrease in sensitivity) by MRI. However, no significant correlation between the sensitivity and time interval between surgery and MRI (del1: Spearman's rho -0.080, p = .69 del2: -.34 p = .87) was found.

Due to flaws in one of the observers' delineation, two patients were re-delineated. These flaws were caused by an unfinished delineation and incorrect use of guidelines and were considered not to occur in clinical practice. These re-delineations resulted in more reliable results and were used for the analysis.

Although we optimized the registration between pathology and the MR images and manually corrected deformations of the specimen according to the MR images, we could not fully prevent some registration mismatch. Furthermore, a variable amount of shrinkage of tumor specimens after fixation has been found for head-and-neck cancer depending on the fixation method, tumor site, and tissue of origin [17–20]. For oral tongue cancer, the tumors shrank on average 20.2% [17]. This is much more than what has been found by Caldas-Magalhaes et al. [6]. The shrinkage of larynx specimens due to fixation with formaldehyde was on average 3% within the cartilage skeleton, which included the tumor bulk. The larynx is a privileged site with a strong and rigid skeleton, which helps to maintain the shape and the size of the tissues inside it resulting in less shrinkage.

One of the results of this study is that overestimation of the tumors on MRI was found. This overestimation is probably caused by the inclusion edema and/or inflamed tissue in the GTV. Histological investigation of these volumes is needed to confirm the presence of edema and inflamed tissue.

The target margin was added uniformly around the GTV and therefore, the direction of the extensions were not taken into account. It should be noted that the target margins are determined for research purposes. Detailed analysis of margins, which can be used to incorporate delineation

uncertainty in clinical practice for various imaging modalities is subject of future research.

Conclusions

In clinical practice, the delineated GTV on MRI is twice as large as the tumor volume determined on pathology. Delineation guidelines can be used as a starting point for MRI delineation for laryngeal and hypopharyngeal carcinoma, leading to a decrease in tumor volume overestimation. After the addition of a target margin around the GTVs to include all tumor tissue, the use of guidelines resulted in a smaller target volume with similar tumor coverage.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

1. Njeh CF. Tumor delineation: the weakest link in the search for accuracy in radiotherapy. J Med Phys 2008;33:136–40.
2. Geets X, Daisne JF, Arcangeli S, et al. Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: comparison between CT-scan and MRI. Radiother Oncol 2005;77:25–31.
3. Rasch CR, Steenbakkers RJ, Fitton I, et al. Decreased 3D observer variation with matched CT-MRI, for target delineation in Nasopharynx cancer. Radiat Oncol 2010;5:21-717X-5-21.
4. Jager E, Kasperts N, Caldas-Magalhaes J, et al. GTV delineation in supraglottic laryngeal carcinoma: interobserver agreement of CT versus CT-MR delineation. Radiat Oncol 2015;10:26.
5. Caldas-Magalhaes J, Kooij N, Ligtenberg H, et al. The accuracy of target delineation in laryngeal and hypopharyngeal cancer. Acta Oncol 2015;54:1181–1187.
6. Caldas-Magalhaes J, Kasperts N, Kooij N, et al. Validation of imaging with pathology in laryngeal cancer: accuracy of the registration methodology. Int J Radiat Oncol Biol Phys 2012;82:e289–98.
7. Daisne JF, Duprez T, Weynand B, et al. Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen. Radiology 2004;233:93–100.
8. Khoo VS, Dearnaley DP, Finnigan DJ, et al. Magnetic resonance imaging (MRI): considerations and applications in radiotherapy treatment planning. Radiother Oncol 1997;42:1–15.
9. Ahmed M, Schmidt M, Sohaib A, et al. The value of magnetic resonance imaging in target volume delineation of base of tongue tumors–a study using flexible surface coils. Radiother Oncol 2010;94:161–7.
10. Chung NN, Ting LL, Hsu WC, et al. Impact of magnetic resonance imaging versus CT on nasopharyngeal carcinoma: primary tumor target delineation for radiotherapy. Head Neck 2004;26:241–6.
11. Rasch C, Keus R, Pameijer FA, et al. The potential impact of CT-MRI matching on tumor volume delineation in advanced head and neck cancer. Int J Radiat Oncol Biol Phys 1997;39:841–8.
12. Verduijn GM, Bartels LW, Raaijmakers CP, et al. Magnetic resonance imaging protocol optimization for delineation of gross tumor volume in hypopharyngeal and laryngeal tumors. Int J Radiat Oncol Biol Phys 2009;74:630–6.
13. Bol GH, Kotte AN, van der Heide UA, et al. Simultaneous multimodality ROI delineation in clinical practice. Comput Methods Programs Biomed 2009;96:133–40.
14. Becker M, Zbaren P, Casselman JW, et al. Neoplastic invasion of laryngeal cartilage: reassessment of criteria for diagnosis at MR imaging. Radiology 2008;249:551–9.
15. Jager EA, Willems SM, Schakel T, et al. Interobserver variation among pathologists for delineation of tumor on H&E-sections of laryngeal and hypopharyngeal carcinoma. How good is the gold standard? Acta Oncol 2016;55:391–395.
16. Seitz O, Chambron-Pinho N, Middendorp M, et al. 18F-Fluorodeoxyglucose-PET/CT to evaluate tumor, nodal disease, and gross tumor volume of oropharyngeal and oral cavity cancer: comparison with MR imaging and validation with surgical specimen. Neuroradiology 2009;51:677–86.
17. Brotherston D, Poon I, Peerani R, et al. Tumor shrinkage associated with whole-mount histopathologic techniques in oral tongue carcinoma. Pathol Res Pract 2015;211:398–403.
18. Campbell S, Poon I, Markel D, et al. Evaluation of microscopic disease in oral tongue cancer using whole-mount histopathologic techniques: implications for the management of head-and-neck cancers. Int J Radiat Oncol Biol Phys 2012;82:574–81.
19. Chen CH, Hsu MY, Jiang RS, et al. Shrinkage of head and neck cancer specimens after formalin fixation. J Chin Med Assoc 2012;75:109–13.
20. Cheng A, Cox D, Schmidt BL. Oral squamous cell carcinoma margin discrepancy after resection and pathologic processing. J Oral Maxillofac Surg 2008;66:523–9.