Acta Oncologica, 2016; 55: 391–395

Taylor & Francis
Taylor & Francis Group

LETTER TO THE EDITOR

Interobserver variation among pathologists for delineation of tumor on H&E-sections of laryngeal and hypopharyngeal carcinoma. How good is the gold standard?

ELISE ANNE JAGER[1], STEFAN M. WILLEMS[2], TIM SCHAKEL[1], NINA KOOIJ[3], PIETER J. SLOOTWEG[4], MARIELLE E. P. PHILIPPENS[1], JOANA CALDAS-MAGALHAES[1], CHRIS H. J. TERHAARD[1] & CORNELIS P. J. RAAIJMAKERS[1]

[1]*Department of Radiation Oncology, University Medical Center Utrecht, Utrecht, The Netherlands,* [2]*Department of Pathology, University Medical Center Utrecht, Utrecht, The Netherlands,* [3]*Department of Pathology, Laboratory Oost-Nederland, Hengelo, The Netherlands and* [4]*Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands*

To the Editor,

In tumor delineation studies for radiotherapy, histopathology is used for validation purposes [1–7]. Validation of tumor delineation is a complex procedure and relatively few studies have been performed in the research field of head-and-neck cancer [3,5,8]. In these studies, whole mount sections of laryngectomy specimens were obtained. The tumor was delineated by a pathologist and used to validate various imaging modalities, e.g. computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) for their ability to distinguish tumor tissue. For the interpretation of validation studies, the variation of tumor outline on histopathology is crucial, as it is used as gold standard for tumor delineation in clinical imaging studies as currently performed by our institute [3,5,9]. However, a study on the reproducibility of tumor outline is missing. The aim of this study is to determine the variation of tumor delineation among pathologists on H&E-sections for laryngeal and hypopharyngeal carcinoma to quantify the uncertainties in the gold standard in the context of imaging validation studies for laryngeal and hypopharyngeal carcinoma.

Material and methods

Ten patients were randomly selected from a database of 22 patients enrolled in the imaging-validation study performed at our institute [3]. This study was approved by the Ethics committee of the University Medical Center Utrecht, the Netherlands and Informed Consent was given by the patients included in this study.

All tumors from patients selected for this study were T3 or T4 squamous cell carcinoma of the larynx or hypopharynx, eligible for surgical resection (Table I).

Three dedicated head-and-neck pathologists manually delineated carcinomatous tissue on the whole mount hematoxylin and eosin (H&E) stained sections. Overlap and distance analyses were performed.

Delineation on H&E-sections

H&E-sections were obtained from whole mount sections of the laryngectomy-specimens according to the procedure used in our imaging-validation study. The sets of H&E-sections of the whole specimens of 10 patients consisted of in total 279 H&E-sections. The interval between the H&E-sections obtained from the specimens was 3 mm. Three head-and-neck pathologists from different institutions independently delineated tumor tissue on the H&E-sections using a permanent marker pen. The pathologists were blinded to each others' results. One pathologist used magnifying glasses and a light microscope for areas of doubt. The other two pathologists used a light microscope. No guidelines for specific magnifications

Correspondence: E. A. Jager, Department of Radiation Oncology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands. Tel: + 31 88 7550591. Fax: + 31 88 7555560. E-mail: E.A.Jager-3@umcutrecht.nl

(Received 28 April 2015; accepted 29 April 2015)

ISSN 0284-186X print/ISSN 1651-226X online © 2015 Informa Healthcare
DOI: 10.3109/0284186X.2015.1049661

Table I. Tumor characteristics per patient.

Patient	Primary site	T-stage
1	hypopharyngeal	T4
2	hypopharyngeal	T3
3	supraglottic	T4
4	supraglottic	T3
5	supraglottic&glottic	T4
6	supraglottic	T4
7	transglottic	T4
8	glottic	T4
9	hypopharyngeal	T4
10	supraglottic	T4

Tumor site and tumor stage for each patient.

and settings were given. The observers were instructed to manually draw a line around the tumor tissue on the H&E-sections including cartilage invasion but excluding any positive lymph nodes. After delineation, the sections, including the tumor outlines, were scanned at 300 dpi resolution. For separate digitization of the lines drawn by the pathologist an in-house developed software package [10], used in clinical radiotherapy practice, was applied to manually trace the tumor outline by a researcher. The digitized lines were then projected on scanned H&E-sections without delineations. Re-evaluation of H&E-sections with remarkable discrepancies between delineations of the pathologists was performed. The delineations used in the analysis were not adjusted after re-evaluation.

Observation parameters

Volumetric analysis. The volumes of the delineated tumors were determined by multiplying the number of voxels contained within a contour by the size of the voxel. This value was multiplied by 3 mm which is the interval between the sections. Mean volumes and standard deviations were calculated.

Interobserver variation. In order to quantify the variation between the observers, the generalized conformity index (CIgen) was calculated [11]. CIgen is defined as the sum of the common volumes of the various observer pairs divided by the sum of the encompassing volumes of these pairs (observer A&B, A&C, B&C) and is defined for three observers as:

$$CI_{gen} = \frac{(A \cap B) + (A \cap C) + (B \cap C)}{(A \cup B) + (A \cup C) + (B \cup C)}$$

The common volume (CV) is the volume that is part of all individual delineations for one patient. The encompassing volume (EV) is the volume encompassing all individual delineations for one patient. A CIgen of 1.00 indicates perfect overlap (identical delineations, 100% agreement), whereas a CI of 0.00 indicates no overlap at all. This index is independent of the number of observers or delineated volumes [11].

Analysis of variation in distance

From the three contours a common and an encompassing contour were derived. For each H&E-section the distance for each point on the common contour to the closest point on the encompassing contour was calculated. The root mean squares (RMS) of these distances were calculated.

Statistical analysis

The correlation between the mean size of the tumor and variation between the delineated volumes for that tumor was analyzed with the Spearman's rank correlation test (two-tailed) which was also used to analyze the correlation between delineated tumor volumes and the CIgen. The comparison between the overlap of the various observer pairs was analyzed with related samples Friedman's two-way analyses of variance ranks. The Wilcoxon-signed ranks test was applied for analyzing the distribution of the delineated tumor volumes between observers.

Results

In total 124 of the 279 H&E-sections were delineated by the pathologists resulting in 372 delineations.

In general the agreement between observers appeared high. However, for several H&E-sections considerable variation between delineations was observed (Figure 1).

Volumetric analysis and interobserver variation

The mean delineated volume by the three observers was 12.95 cm^3 (SD 0.3, range 3.0–39.8). The variation between the delineated volumes per patient was approximately 2% and was not related to the size of the tumor (rho = -0.32, p = 0.37) (Table II).

One pathologist (observer A) delineated a larger volume in eight of the 10 cases compared to the other observers (Table II). The distribution of the delineated tumor volumes was significantly different between observer A and B (p = 0.022) and observer A and C (p = 0.007). No difference between observer B and C was observed (p = 0.74).

The mean interobserver agreement, expressed as the generalized conformity index, was 0.87 (SD 0.04, range 0.82–0.95) (Table II).

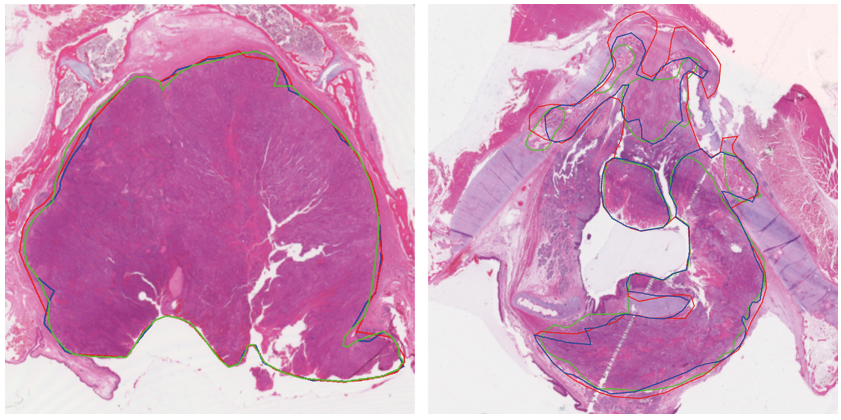

Figure 1. H&E-stained sections obtained from a laryngectomy-specimen with tumor delineations of the three pathologists. Good agreement between observers is perceived on the left section. The right H&E-stained section of a tumor with irregular shaped areas showed larger interobserver variation.

There was no systematic difference in overlap (CV/EV) between the observer pairs ($p = 0.42$) and no correlation between delineated tumor volumes and CIgen was observed (rho = 0.37, $p = 0.30$).

Re-evaluation H&E-sections with discrepant tumor delineations

Ten sections with remarkable discrepancies between pathologists were re-evaluated by all three observers in order to determine if the variation was due to interpretation, if tumor was overlooked and/or if the observers would come to the same conclusion after re-evaluation. The criteria for re-evaluation were: discrepant areas larger than 0.8 cm^2 and if one or more isolated delineated areas which were not delineated by the other observers. From this re-evaluation it was concluded that most discrepancies were based on the interpretation of the extension of cartilage invasion and the in- or exclusion of parts of the necrotic cartilage (Figure 2). After re-evaluation all the observers agreed that these parts of the cartilage were necrotic as a consequence of tumor invasion. Difficulties to distinguish a lymph node with metastatic disease from the main tumor mass caused larger variation on two H&E-sections. For two cranial sections, small fragments of the tumor were clearly overlooked by two pathologists.

Analysis of variation in distance

Ninety-five percent of the measured distances between the encompassing and the common lines were smaller than 2.0 mm (SD 0.7) and in 90% smaller than 1.4 mm (SD 0.4) (Table III).

The 5% of the calculated distances larger than 2.0 mm were mostly found in irregularly shaped tumor areas and in cartilage with tumor invasion (Figure 1). The RMS of the distances between the common and the encompassing contours amounted to 1.0 mm.

Table II. Delineated tumor volumes per pathologist and interobserver variation according to the generalized conformity index.

Patient	Observer A (cm^3)	Observer B (cm^3)	Observer C (cm^3)	Mean (cm^3)	SD (cm^3)	RSD(%)	CIgen
2	3.10	2.93	2.91	2.98	0.10	3.50	0.88
4	3.37	3.47	3.07	3.30	0.21	6.30	0.85
6	5.15	4.92	4.79	4.95	0.18	3.68	0.82
5	5.36	5.13	4.97	5.15	0.20	3.80	0.86
3	6.30	6.25	5.65	6.07	0.36	5.96	0.85
8	9.78	9.87	9.82	9.82	0.05	0.46	0.85
1	13.17	12.17	11.71	12.35	0.75	6.04	0.85
7	14.30	13.95	13.22	13.82	0.55	3.99	0.85
10	31.57	31.37	30.81	31.25	0.39	1.26	0.95
9	40.52	38.48	40.45	39.82	1.16	2.91	0.91
mean	13.26	12.85	12.74	12.95	0.27	2.12	0.87

CIgen, generalized conformity index; RSD, relative standard deviation as percentage SD of mean; SD, standard deviation.

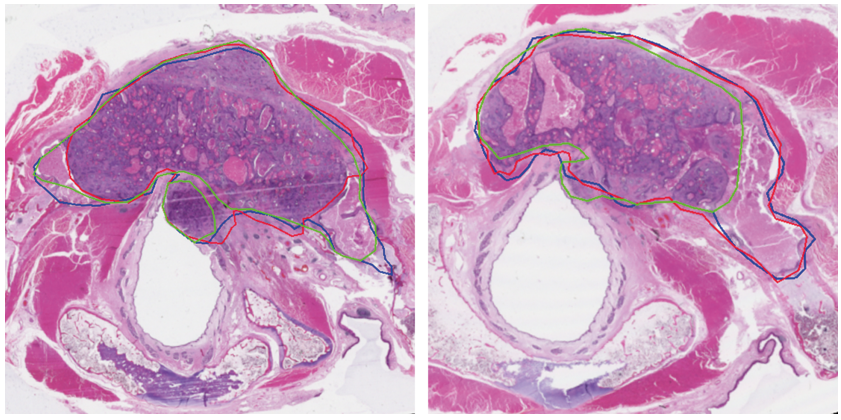

Figure 2. H&E-stained sections obtained from a laryngectomy-specimen with tumor delineations of the three pathologists. For this specimen the largest distances between delineations were observed due to exclusion of thyroid cartilage by one of the observers. Distances up to 9.9 mm were measured.

Discussion

To our knowledge this is the first study in which the variance of tumor delineation on H&E-sections by pathologists was investigated. The variation between tumor delineation on H&E-sections by the pathologists was relatively low and histopathology as the gold standard for imaging validation studies was highly reliable. The mean overlap between the delineations (expressed as CIgen) amounted to 0.87. This implies that on average the observers agreed on 87% of the total delineated volume. The distances between the delineations were in 95% of the measured distances smaller than 2 mm. Larger distances were found in irregularly shaped tumor areas and in the presence of cartilage invasion. The inclusion or exclusion of cartilage increased variation although after re-evaluation there was consensus about whether or not the cartilage was affected. The variation between the pathologists measured in this study was also caused by several other factors. The thickness of the used pencil (0.7 mm), the delineation style, e.g. the decision to delineate along the outside of the tumor border or along the inside; how precise the observer decided to delineate along the tumor border, and the in- or exclusion of necrotic tissue. Increasing accordance in delineation style by clear delineation guidelines would further decrease the variation between pathologists.

Table III. Distance analysis.

Patient	p90(mm)	p95(mm)
1	1.3	1.9
2	0.9	1.0
3	1.3	1.8
4	1.3	1.7
5	1.6	2.2
6	1.6	2.2
7	1.8	2.5
8	2.0	3.6
9	1.3	1.6
10	1.0	1.3
mean (SD)	1.4 (0.4)	2.0 (0.7)

Distances in mm measured between each point on the common and the encompassing delineation per patient. p90, p95: 90% respectively 95% of the measured distances is smaller than the values shown in the table.

For the best case the overlap (CIgen) between the observers was 0.95. The variation of 0.05 was merely caused by the thickness of the pencil and the delineation style. This value may consequently be considered as the maximum overlap value. As this study is unique with regard to its purpose and method we were not able to adequately compare our findings with results from other studies. However, the results of this study can be compared to delineation studies performed on various imaging modalities. In an imaging-validation study [3] performed at our institution, the registration errors between various imaging modalities and histopathology using a three-dimensional (3D) registration method, were determined. The calculated registration errors (RMSE CT; 1.5 mm PET; 3.3 mm MRI; 3.0 mm) exceeded the variation between the pathologists determined in this study. Therefore, in a study-setting in which pathology imaging registration is performed, the registration inaccuracy is larger than the variation of the gold standard. Much larger delineation inaccuracies varying from 3.1 to 16.1 mm were reported for delineation of the GTV by various observers delineating on CT images [5]. Rasch et al. [12] calculated the RMS of the standard deviation of distances between delineations of radiation-oncologists and the median surface for tumor delineation for nasopharyngeal carcinoma. This resulted in 4.4 mm on CT and 3.3 mm on CT combined with MRI with an overlap agreement of 36% (CT) and 64% (CT+MRI). These values are considerably larger than the values reported in our study. Earlier work performed by our

research group showed a CIgen of 0.61 for delineation of supraglottic laryngeal carcinoma by three radiation-oncologists [13]. Therefore, it can be concluded that delineation inaccuracies on images are much larger than on histopathology.

Declaration of interest: This work was supported by the Dutch Cancer Society, Grant No. 2011-5152. The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

[1] Zhang Y, Hu J, Li J, Wang N, Li W, Zhou Y, et al. Comparison of imaging-based gross tumor volume and pathological volume determined by whole-mount serial sections in primary cervical cancer. Oncol Targets Ther 2013;6:917–23.

[2] Stroom J, Schlief A, Alderliesten T, Peterse H, Bartelink H, Gilhuijs K. Using histopathology breast cancer data to reduce clinical target volume margins at radiotherapy. Int J Radiat Oncol Biol Phys 2009;74:898–905.

[3] Caldas-Magalhaes J, Kasperts N, Kooij N, van den Berg CA, Terhaard CH, Raaijmakers CP, 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.

[4] Borren A, Moman MR, Groenendaal G, Boeken Kruger AE, van Diest PJ, van der Groep P, et al. Why prostate tumour delineation based on apparent diffusion coefficient is challenging: An exploration of the tissue microanatomy. Acta Oncol 2013;52:1629–36.

[5] Caldas-Magalhaes J, Kooij N, Ligtenberg H, Jager EA, Schakel T, Kasperts N, et al. The accuracy of target delineation in laryngeal and hypopharyngeal cancer. Acta Oncol Epub 2015 Mar 3:1–7.

[6] Campbell S, Poon I, Markel D, Vena D, Higgins K, Enepekides 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.

[7] Groenendaal G, Moman MR, Korporaal JG, van Diest PJ, van Vulpen M, Philippens ME, et al. Validation of functional imaging with pathology for tumor delineation in the prostate. Radiother Oncol 2010;94:145–50.

[8] Daisne JF, Duprez T, Weynand B, Lonneux M, Hamoir M, Reychler H, 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.

[9] Driessen JP, Caldas-Magalhaes J, Janssen LM, Pameijer FA, Kooij N, Terhaard CH, et al. Diffusion-weighted MR imaging in laryngeal and hypopharyngeal carcinoma: Association between apparent diffusion coefficient and histologic findings. Radiology 2014;272:456–63.

[10] Bol GH, Kotte AN, van der Heide UA, Lagendijk JJ. Simultaneous multi-modality ROI delineation in clinical practice. Comput Methods Programs Biomed 2009;96: 133–40.

[11] Kouwenhoven E, Giezen M, Struikmans H. Measuring the similarity of target volume delineations independent of the number of observers. Phys Med Biol 2009;54:2863–73.

[12] Rasch CR, Steenbakkers RJ, Fitton I, Duppen JC, Nowak PJ, Pameijer FA, et al. Decreased 3D observer variation with matched CT-MRI, for target delineation in Nasopharynx cancer. Radiat Oncol 2010;5:21–717X–5–21.

[13] Jager E, Kasperts N, Caldas-Magalhaes J, Philippens M, Pameijer FA, Terhaard C, et al. GTV delineation in supraglottic laryngeal carcinoma: Interobserver agreement of CT versus CT-MR delineation. Radiat Oncol 2015;10:26.