The suitability of common metrics for assessing parotid and larynx autosegmentation accuracy

William J. Beasley,[1,2a] Alan McWilliam,[1,2] Adam Aitkenhead,[1,2] Ranald I. Mackay,[1,2] Carl G. Rowbottom[1,3]

Institute of Cancer Sciences,[1] The University of Manchester, Manchester, UK; Christie Medical Physics and Engineering,[2] The Christie NHS Foundation Trust, Manchester, UK; Physics Department,[3] The Clatterbridge Cancer Centre NHS Foundation Trust, Bebington, UK
william.beasley@christie.nhs.uk

Received 12 June, 2015; accepted 9 December, 2015

Contouring structures in the head and neck is time-consuming, and automatic segmentation is an important part of an adaptive radiotherapy workflow. Geometric accuracy of automatic segmentation algorithms has been widely reported, but there is no consensus as to which metrics provide clinically meaningful results. This study investigated whether geometric accuracy (as quantified by several commonly used metrics) was associated with dosimetric differences for the parotid and larynx, comparing automatically generated contours against manually drawn ground truth contours. This enabled the suitability of different commonly used metrics to be assessed for measuring automatic segmentation accuracy of the parotid and larynx. Parotid and larynx structures for 10 head and neck patients were outlined by five clinicians to create ground truth structures. An automatic segmentation algorithm was used to create automatically generated normal structures, which were then used to create volumetric-modulated arc therapy plans. The mean doses to the automatically generated structures were compared with those of the corresponding ground truth structures, and the relative difference in mean dose was calculated for each structure. It was found that this difference did not correlate with the geometric accuracy provided by several metrics, notably the Dice similarity coefficient, which is a commonly used measure of spatial overlap. Surface-based metrics provided stronger correlation and are, therefore, more suitable for assessing automatic segmentation of the parotid and larynx.

PACS number(s): 87.57.nm, 87.55.D, 87.55.Qr

Key words: automatic segmentation, adaptive radiotherapy, treatment planning, head and neck

I. INTRODUCTION

Intensity-modulated radiotherapy (IMRT) and volumetric-modulated arc therapy (VMAT) are capable of creating highly conformal treatment plans, with steep dose gradients providing efficient organ at risk (OAR) sparing.[1,2] IMRT has been shown to benefit patients in the head and neck,[3–5] with the PARSPORT trial (Institute of Cancer Research, London, UK) demonstrating reduced incidence of xerostomia in patients treated with parotid-sparing IMRT relative to those treated with conformal radiotherapy.[6] However, in order to realize the benefits afforded by IMRT, accurate delineation of targets and normal structures is essential.[7]

[a] Corresponding author: William J. Beasley, Christie Medical Physics and Engineering, The Christie NHS Foundation Trust, Wilmslow Rd., Manchester M20 4BX, UK; phone: (+44) 161 9187768; fax: (+44) 161 4463545; email: william.beasley@christie.nhs.uk

Contouring in the head and neck is time-consuming and labor-intensive,[8,9] but automatic segmentation has shown potential to reduce interobserver variation and improve efficiency by reducing the time required for outlining.[7,10,11] This is of particular benefit to adaptive radiotherapy (ART), and there has therefore been much interest in automatic segmentation, with several algorithms having been assessed for accuracy.[7,11–14]

In such studies, the accuracy of automatic segmentation algorithms has been assessed by measuring geometric agreement between automatically generated structures and 'ground truth' structures provided by manual delineation. A wide variety of metrics have been reported in the literature, and can be broadly separated into volume-based and surface-based metrics. Volume-based metrics, such as the Dice similarity coefficient (DSC), which measures the spatial overlap of two volumes (see Fig. 1), and the conformity index (CI), which measures the relative difference in volumes, are commonly used.[15–17] Whilst these metrics are relatively simple to understand, they are difficult to interpret and are sensitive to the volumes of the structures being assessed.[18,19] Surface-based metrics provide a quantitative measure of the concordance of two surfaces and are typically based on distance-to-agreement (DTA). DTA is calculated by computing the minimum distance from a point on a reference surface to any point on a target surface (see Fig. 1), which is repeated for all points on the reference surface. From this, a DTA histogram can be produced. Several different metrics can be derived from this DTA histogram and some of the most commonly reported include the mean- and maximum-DTA[14,20] and the 95%-Hausdorff distance (95%-HD),[21,22] which is defined as the 95th percentile of the DTA histogram. Although these metrics provide a measure of the distance between two structures, they too can be difficult to translate into clinical relevance.[19]

To parallel the term 'geometric accuracy', which quantifies the spatial agreement of two different structures, we introduce the term 'dosimetric accuracy' to quantify the difference in dose between two structures within a given dose distribution. In the case of automatic segmentation, the goal is to create automatically generated structures with high geometric accuracy relative to the ground truth. Similarly, within the context of treatment planning and evaluation, in which an automatically generated contour might be used for treatment planning, it is also important that the dose reported to an automatically generated contour agrees with the dose reported to the corresponding ground truth structure.[23] Erroneous dose reporting may ultimately lead to a suboptimal plan.

With such a variety of spatial metrics available, there is no consensus as to the most suitable metric for assessing geometric accuracy.[18,24] As both the geometric and dosimetric accuracy are important for treatment planning and evaluation, it can be argued that suitable spatial metrics are those that provide results related to dosimetric accuracy.[25] A geometrically accurate contour, as measured with a suitable spatial metric, should therefore be reflected in a small dosimetric difference, and vice versa.

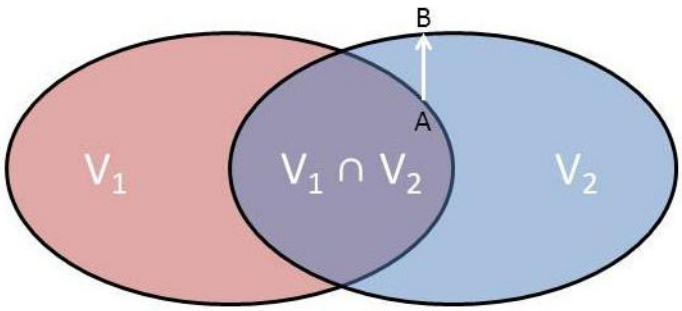

FIG. 1. DSC and DTA. DSC measures the spatial overlap between two volumes, and DTA describes the shortest distance between two surfaces for a specific point.

In the present study, the geometric and dosimetric accuracy are measured for the parotid and larynx in head and neck VMAT treatment planning. The relationship between the geometric and dosimetric accuracy is measured, thus identifying suitable spatial metrics.

II. MATERIALS AND METHODS

Five clinicians outlined the parotids and larynx for 10 head and neck cancer patients. These contours were created as part of a recent study at our institution assessing the geometric accuracy of a commercial automatic segmentation algorithm. The contouring has previously been described,[14] but is briefly outlined here. Contouring was performed according to locally agreed protocols, and all observers contoured the structures independently, with access to the same clinical information. The observers were free to adjust the windowing and level according to personal preference. For each structure, the five clinician contours were combined into a single ground truth contour using the simultaneous truth and performance level estimation (STAPLE) algorithm, which computes a probabilistic estimate of the ground truth from multiple segmentations of the same structure.[26] The resulting STAPLE contours were used as the reference standard (i.e., the ground truth) against which automatic contours were compared.

A. Dosimetric and geometric accuracy

For each patient, dual-arc 6 MV Elekta VMAT plans (Elekta, Stockholm, Sweden) were retrospectively created using the Philips Pinnacle^3 v9.6 treatment planning system (Philips Radiation Oncology Systems, Andover, MA), according to standard departmental protocols (see Table 1). Planning target volumes (PTVs) were created from a uniform 4 mm expansion of relevant clinical target volumes (CTVs), which had been drawn at the point of initial treatment; automatic segmentation of target volumes was not investigated. Automatically generated normal structures were created using the Philips Smart Probabilistic Contouring Engine (SPICE) software. These automatically generated contours were then used directly in the plan optimization, with a 5 mm uniform margin applied to the spinal cord and brainstem to create planning organ at risk volumes (PRVs). The STAPLE contours for the parotids and larynx were imported into the treatment plan and the mean doses to these structures (the 'true' doses) were compared to those of the corresponding automatically generated contours. The mean dose was used as this is the dosimetric parameter of interest when assessing a treatment plan for these structures. The dosimetric accuracy was then defined as the percentage difference between the mean dose to the automatically generated and ground truth structures, relative to the dose to the ground truth structure.

In addition to measuring the dosimetric accuracy for the automatically generated structures, the difference in mean dose to the individual clinician contours relative to the true dose (dose to the STAPLE contour) was also measured for each patient. This provided a measurement of the dosimetric interobserver variation in mean dose for each patient, ultimately defining the range within which the dose to the automatically generated structure is acceptable.

Table 1. OAR dose constraints used for creating the VMAT plans.

OAR	*Dose Constraint/cGy*
Spinal cord PRV	Max < 4800 Max 1 cm^3 < 4500
Brainstem PRV	Max < 5400 Max 1 cm^3 < 5000
Contralateral parotid	Mean < 2600
Larynx	Mean < 4500
Oral cavity	Mean < 4500

A number of commonly used metrics were used to measure the geometric accuracy of the automatically generated contours relative to the ground truth structures, using an in-house MATLAB script (MathWorks, Natick, MA). Two volume-based metrics were investigated: the conformity index (CI), which is the ratio of the volumes of the two structures; and DSC, which is a measure of the spatial overlap of two structures, defined as $DSC = 2(V_1 \cap V_2)/(|V_1| + |V_2|)$ (see Fig. 1). The centroid separation, which is the magnitude of the distance between the centers of mass of two structures, was also measured.

The other metrics were based on the surface agreement of two structures, and an in-house MATLAB script was used to calculate a DTA histogram for each structure pair. DTA is defined for a particular point on a reference surface, A, as the shortest distance to any point on surface B (see Fig. 1). This is performed for each point on surface A, and a cumulative DTA histogram is created. From this DTA histogram, the mean and maximum DTA were measured, along with the 95%-Hausdorff distance (95%-HD), measuring the 95th percentile of the cumulative DTA histogram.

B. Relationship between geometric and dosimetric accuracy

The correlation between the dosimetric accuracy and the different metrics was measured using the Pearson product-moment correlation coefficient; the strength of the correlation indicated the strength of the relationship between the geometric and dosimetric accuracy.

III. RESULTS

A. Dosimetric and geometric accuracy

The mean dosimetric accuracy was measured to be -4.8 ± 3.4% and -8.4 ± 2.3% (dose to the automatically generated contours lower than that to the STAPLE contours) for the parotids and larynx, respectively. The uncertainties were estimated from the mean standard deviations in the interobserver variation in mean dose, which provides an estimate of the uncertainty in the dose delivered to the ground truth contours. Figures 2 and 3 show box plots of the dosimetric interobserver variation for the individual parotid and larynx structures, respectively. The boxes indicate the interquartile range and the whiskers indicate the maximum and minimum range of variation in mean dose to the five clinician-drawn structures relative to the STAPLE contour (equal to the interobserver variation). The dosimetric accuracy of the automatically generated contours is also indicated for each structure by the black circles; it can be seen that the dose to

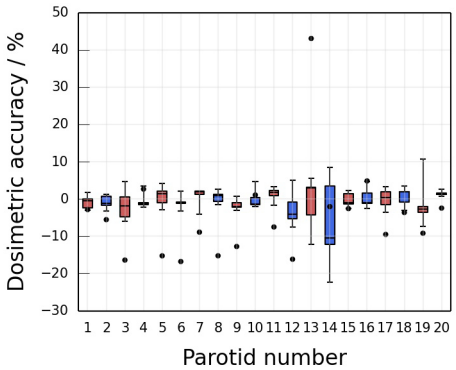

FIG. 2. Dosimetric interobserver variation for the parotids. Box plot showing the interobserver variation in dosimetric accuracy relative to the STAPLE contours for the parotid glands. Red boxes indicate right hand parotid glands and blue boxes indicate left hand glands. The boxes indicate the interquartile range, the whiskers indicate the minimum and maximum variation, and the horizontal lines indicate the median accuracy of the five clinician contours. The mean dosimetric accuracy of the automatically generated contours is indicated by the circles.

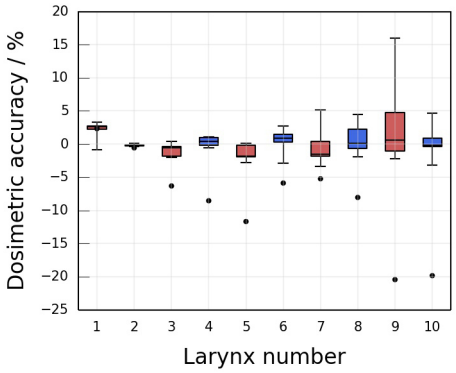

FIG. 3. Dosimetric interobserver variation for the larynx. Box plot showing the interobserver variation in dosimetric accuracy relative to the STAPLE contours for the larynx. The boxes indicate the interquartile range, the whiskers indicate the minimum and maximum variation, and the horizontal lines indicate the median accuracy of the five clinician contours. The mean dosimetric accuracy of the automatically generated contours is indicated by the circles.

the automatically generated contour was outside the dosimetric interobserver variation for 16 out of the 20 parotid glands and nine of the ten larynx contours. Note that parotid 13 (Fig. 2) has a dosimetric accuracy of + 43%. This was caused by the gland's being in a region of low dose (mean doses of 376 cGy and 514 cGy to the STAPLE and automatically generated contour, respectively), resulting in a large relative difference in mean dose between the SPICE and STAPLE contours. Similarly, parotid 14 also received a low mean dose (approximately 500 cGy), so the interobserver variation was relatively large for this gland.

B. Relationship between geometric and dosimetric accuracy

Table 2 shows the correlation coefficients between the dosimetric accuracy and the various metrics. There was no correlation between the volume-based metrics (DSC and CI) and dosimetric accuracy for the parotids. Metrics based on surface agreement (DTA) showed statistically significant ($p < 0.05$) correlations with dosimetric accuracy, with meanDTA and 95%-HD showing strong correlation. The strongest correlate for the parotid was found to be the centroid separation. This can be seen in Fig. 4, which shows scatter plots of the dosimetric accuracy as a function of centroid separation (left hand plot), for which correlation was strong and statistically significant, and DSC (right hand plot), for which correlation was weak and not statistically significant ($p > 0.05$).

Centroid separation did not correlate with dosimetric accuracy for the larynx, and weak correlation was observed for the volume-based metrics. Strong correlation was observed for the surface-based metrics.

TABLE 2. Correlation coefficients between the different metrics and the dosimetric accuracy.

Metric	*Parotid*	*Larynx*
DSC	-0.35	-0.59[a]
CI	-0.33	0.58[a]
Centroid separation	0.82[a]	0.50
maxDTA	0.55[a]	0.60[a]
meanDTA	0.69[a]	0.64[a]
95%-HD	0.61[a]	0.63[a]

[a] Statistical significance at $p < 0.05$.

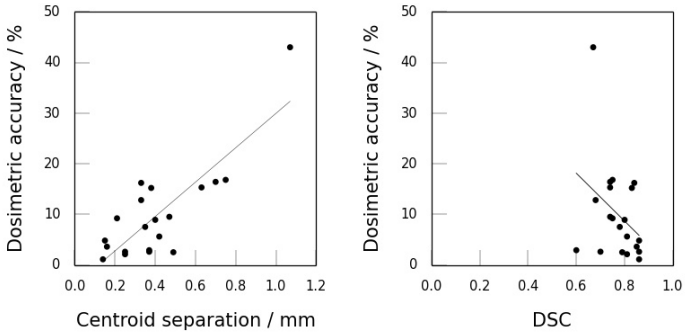

FIG. 4. Scatter plots showing the relationship between dosimetric and geometric accuracy for the parotid. The left hand plot shows the relationship for the centroid separation (R = 0.82), and the right hand plot shows the relationship for DSC (R = -0.35). Lines of best fit are also shown.

IV. DISCUSSION

Automatic segmentation will be an essential component of treatment planning and ART, and it is important to assess the geometric accuracy of automatic segmentation algorithms before clinical implementation. There are many widely accepted spatial metrics for assessing geometric accuracy, but there is no consensus as to the most appropriate metrics to use. For treatment planning and evaluation, it is also important that the dose to an automatically generated structure agrees with the dose to its corresponding ground truth structure. This work proposes that an appropriate spatial metric is one that correlates with dosimetric accuracy, and aims to identify spatial metrics suitable for assessing automatic segmentation accuracy of the parotid and larynx in the head and neck.

The results have indicated that several commonly used geometric metrics do not correspond to dosimetric accuracy and are not suitable for assessing automatic segmentation performance for certain OARs in the head and neck. Notably, it has been shown that DSC is a poor surrogate for dosimetric accuracy for the parotids. This is highlighted by the fact that the mean parotid DSC in this study was 0.77, which is generally considered to be clinically acceptable,[14,27] but the mean dose to the automatically generated structures was outside the range of dosimetric interobserver variation for 16 out of the 20 parotid glands investigated.

Although DSC and CI did not correlate with dosimetric accuracy for the parotids, the remaining metrics provided statistically significant correlations. The centroid separation, which is the magnitude of the distance between the centers of mass of the automatically generated and STAPLE contours, provided the strongest correlation. This is likely explained by the fact that the parotids are often in close proximity to the target volume and in the region of a unidirectional steep dose gradient, such that global differences in organ position have a large effect on dosimetric accuracy. In contrast, the centroid separation did not correlate with dosimetric accuracy for the larynx. This was likely caused by the fact that the larynx is often in the region of several dose gradients, and so a global shift of position does not necessarily change the mean dose.

The surface-based metrics correlated with dosimetric accuracy for both the parotids and larynx, although the correlation for maxDTA was weaker than for meanDTA and 95%-HD, probably due to the fact that a discrepancy in a single point on a surface does not necessarily have a large effect on the mean dose to that structure. Nevertheless, the correlation of the surface-based metrics with dosimetric accuracy suggests that these metrics are suitable for assessing automatic segmentation accuracy.

Whilst there have been many studies reporting the geometric accuracy of various automatic segmentation algorithms,[7,13,14,28] there have been relatively few that have investigated the dosimetric effect of automatic segmentation uncertainties in head and neck IMRT.[25,29] Tsuji

et al.[25] investigated the dosimetric accuracy in 16 patients treated with head and neck IMRT. They compared the doses delivered to automatic and manual contours, and reported that the dosimetric differences were significant for the targets, but minor for the OARs. In contrast, Eiland et al.[29] reported significant dosimetric differences between automatic and manual contours in seven head and neck IMRT plans, and concluded that automatic segmentation cannot yet replace manual delineation for treatment planning. This is in agreement with our findings, which show that the dose delivered to automatic contours is generally outside the range of interobserver variation.

The above studies measured the difference between dosimetric parameters for different structures in the head and neck for automatic and manual contours, using a single observer to define the ground truth. However, in our study, five clinicians outlined each structure, providing the interobserver variation in dose for individual patients. This enabled a more realistic assessment of the acceptability of automatic contours, as the dose to an automatically generated structure could be compared to the interobserver variation for the specific patient in question. Additionally, the use of a STAPLE volume provided a better estimate of the ground truth than for a single observer.[26]

Metrics suitable for assessing the geometric accuracy of automatic segmentation algorithms should be related to dosimetric accuracy, and there is no consensus as to which metrics are suitable for use in head and neck VMAT. Tsuji et al.[25] investigated the relationship between geometric and dosimetric accuracy for two metrics: DSC and the overlap index (OI), which measures the proportion of the manual contour within the automatic contour. Although they reported correlation between GTV dosimetric agreement and the OI, there was no correlation for the OARs. This supports the results obtained in the present study, where the volume-based metrics did not correlate with differences in mean dose; however, the authors did not investigate surface-based metrics. Nelms et al.[23] did investigate a surface-based metric, the 'linear penalty', which is a modified DTA giving more weight to larger contour discrepancies. Although they reported that this metric was related to dosimetric accuracy, the linear penalty is not commonly used; nevertheless, this supports our findings that surface-based metrics are suitable for measuring automatic segmentation accuracy.

In the present study, a single commercial automatic segmentation algorithm was used to generate automatic contours. It should be emphasized that the relationship between the geometric and dosimetric accuracy would be independent of the specific automatic segmentation algorithm used.

The present study used a small dataset of 10 head and neck patients to assess the relationship between dosimetric and geometric accuracy for the parotid and larynx. These structures are considered parallel organs, and the results cannot be extrapolated to serial organs in the head and neck, such as the spinal cord and brainstem. The location of these structures relative to typical dose gradients, as well as the fact that the dosimetric parameter of interest is the maximum dose, means that further work is required to determine which metrics are suitable for assessing such serial organs.

Similarly, extrapolation of the results presented here to other treatment sites should be performed with caution. By using the mean dose to quantify dosimetric accuracy, a complex three-dimensional dose distribution has been collapsed into a single dosimetric parameter, disregarding any positional information about the dose distribution. However, this was mitigated in our study by planning all patients with the same head and neck VMAT class solution, such that the dose distributions of all 10 patients were similar. This means that the results presented here apply to the parotid and larynx when used for head and neck VMAT treatment planning. For example, although it might be expected that metrics useful for the parotid might also be useful for the rectum in prostate radiotherapy, as they are both in close proximity to a target volume and are in a region of a single steep dose gradient, further work would be required to verify this.

This study has assessed the relationship between geometric and dosimetric accuracy for several spatial metrics commonly used for assessing automatic segmentation accuracy.

Specifically, the suitability of these metrics has been assessed for the parotid and larynx in head and neck VMAT treatment planning. The results have indicated that the suitability of a spatial metric is dependent on the structure. In particular, the common volume-based metrics, such as DSC and OI, are not related to dosimetric accuracy for the parotid, and only weakly related to dosimetric accuracy for the larynx. The surface-based metrics were related to dosimetric accuracy and, therefore, suitable for assessing automatic segmentation accuracy for both the parotids and larynx.

V. CONCLUSIONS

There are several spatial metrics available for assessing automatic segmentation accuracy, and there is no consensus on which metrics should be used. For treatment planning and evaluation, both geometric and dosimetric accuracy are important, as inaccurate contours can result in a suboptimal treatment plan. A suitable spatial metric should therefore be related to dosimetric accuracy. This study has measured the relationship between geometric and dosimetric accuracy for several commonly used metrics in head and neck VMAT treatment planning. We found that this relationship is structure-dependent, and that there was no statistically significant relationship between volume-based metrics and dosimetric accuracy for the parotids, with only a weak correlation for the larynx. The surface-based metrics correlated with dosimetric accuracy for both structures, indicating that these metrics are more suitable measures of automatic segmentation accuracy of the parotid and larynx in the head and neck.

COPYRIGHT

REFERENCES

1. Cozzi L, Fogliata A, Bolsi A, Nicolini G, Bernier J. Three-dimensional conformal vs. intensity-modulated radiotherapy in head-and-neck cancer patients: Comparative analysis of dosimetric and technical parameters. Int J Radiat Oncol Biol Phys. 2004;58(2):617–24.
2. Vanetti E, Clivio A, Nicolini G, et al. Volumetric modulated arc radiotherapy for carcinomas of the oro-pharynx, hypo-pharynx and larynx: a treatment planning comparison with fixed field IMRT. Radiother Oncol. 2009;92(1):111–17.
3. Chao LS, Deasy JO, Markman J, et al. A prospective study of salivary function sparing in patients with head-and-neck cancers receiving intensity-modulated or three-dimensional radiation therapy: initial results. Int J Radiat Oncol Biol Phys. 2001;49(4):907–16.
4. Saarilahti K, Kouri M, Collan J, et al. Intensity modulated radiotherapy for head and neck cancer: evidence for preserved salivary gland function. Radiother Oncol. 2005;74(3):251–58.
5. Saarilahti K, Kouri M, Collan J, et al. Sparing of the submandibular glands by intensity modulated radiotherapy in the treatment of head and neck cancer Radiother Oncol. 2006;78(3):270–75.
6. Nutting CM, Morden JP, Harrington KJ, et al. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial. Lancet Oncol. 2011;12(2):127–36.
7. Stapleford LJ, Lawson JD, Perkins C, et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2010;77(3):959–66.
8. Miles EA, Clark CH, Urbano MT, et al. The impact of introducing intensity modulated radiotherapy into routine clinical practice. Radiother Oncol. 2005;77(3):241–46.
9. Harari PM, Song S, Tomé WA. Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2010;77(3):950–58.
10. Chao KS, Bhide S, Chen H, et al. Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach. Int J Radiat Oncol Biol Phys. 2007;68(5):1512–21.
11. Walker GV, Awan M, Tao R, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. Radiother Oncol. 2014;112(3):321–25.

12. Zhang T, Chi Y, Meldolesi E, Yan D. Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy. Int J Radiat Oncol Biol Phys. 2007;68(2):522–30.
13. Al-Mayah A, Moseley J, Hunter S, et al. Biomechanical-based image registration for head and neck radiation treatment. Phys Med Biol. 2010;55(21):6491–500.
14. Thomson D, Boylan C, Liptrot T, et al. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. Radiother Oncol. 2014;9(1):173.
15. Dice L. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297–302.
16. Simmat I, Georg P, Georg D, Birkfellner W, Goldner G, Stock M. Assessment of accuracy and efficiency of atlas-based autosegmentation for prostate radiotherapy in a variety of clinical conditions. Strahlenther Onkol. 2012;188(9):807–15.
17. Veiga C, McClelland J, Moinuddin S, et al. Toward adaptive radiotherapy for head and neck patients: feasibility study on using CT-to-CBCT deformable registration for 'dose of the day' calculations. Med Phys. 2014;41(3):031703.
18. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. Med Phys. 2014;41(5):050902.
19. Sykes J. Reflections on the current status of commercial automated segmentation systems in clinical practice. J Med Radiat Sci. 2014;61(3):131–34.
20. Hoffmann C, Krause S, Stoiber EM, et al. Accuracy quantification of a deformable image registration tool applied in a clinical setting. J Appl Clin Med Phys. 2014;15(1):4564.
21. Hou J, Guerrero M, Chen W, D'Souza WD. Deformable planning CT to cone-beam CT image registration in head-and-neck cancer. Med Phys. 2011;38(4):2088–94.
22. Huger S, Graff P, Harter V, et al. Evaluation of the Block Matching deformable registration algorithm in the field of head-and-neck adaptive radiotherapy. Phys Med. 2014;30(3):301–08.
23. Nelms BE, Tomé WA, Robinson G, Wheeler J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. Int J Radiat Oncol Biol Phys. 2012;82(1):368–78.
24. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2012;30(9):1234–48.
25. Tsuji SY, Hwang A, Weinberg V, Yom SS, Quivey JM, Xia P. Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2010;7(3):707–14.
26. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004;23(7):903–21.
27. Mattiucci GC, Boldrini L, Chiloiro G, et al. Automatic delineation for replanning in nasopharynx radiotherapy: what is the agreement among experts to be considered as benchmark? Acta Oncol. 2013;52(7):1417–22.
28. Qazi AA, Pekar V, Kim J, Xie J, Breen SL, Jaffray DA. Auto-segmentation of normal and target structures in head and neck CT images: a feature-driven model-based approach. Med Phys. 2011;38(11):6160–70.
29. Eiland RB, Maare C, Sjöström D, Samsøe E, Behrens CF. Dosimetric and geometric evaluation of the use of deformable image registration in adaptive intensity-modulated radiotherapy for head-and-neck cancer. J Radiat Res. 2014;55(5):1002–08.