An automated workflow for patient-specific quality control of contour propagation

DOI:
10.1088/1361-6560/61/24/8577

Document Version
Accepted author manuscript

An automated workflow for patient-specific quality control of contour propagation

William J Beasley[1,2], **Alan McWilliam**[1], **Nicholas J Slevin**[1,3], **Ranald I Mackay**[1,2], **Marcel van Herk**[1]

[1] Division of Molecular and Clinical Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK

[2] Christie Medical Physics and Engineering (CMPE), The Christie NHS Foundation Trust, Manchester, UK

[3] Department of Clinical Oncology, The Christie NHS Foundation trust, Manchester, UK

E-mail: william.beasley@christie.nhs.uk

Abstract. Contour propagation is an essential component of adaptive radiotherapy, but current contour propagation algorithms are not yet sufficiently accurate to be used without manual supervision. Manual review of propagated contours is time-consuming, making routine implementation of real-time adaptive radiotherapy unrealistic. Automated methods of monitoring the performance of contour propagation algorithms are therefore required. We have developed an automated workflow for patient-specific quality control of contour propagation and validated it on a cohort of head and neck patients, on which parotids were outlined by two observers. Two types of error were simulated – mislabelling of contours and introducing noise in the scans before propagation. The ability of the workflow to correctly predict the occurrence of errors was tested, taking both sets of observer contours as ground truth, using receiver operator characteristic analysis. The area under the curve was 0.90 and 0.85 for the observers, indicating good ability to predict the occurrence of errors. This tool could potentially be used to identify propagated contours that are likely to be incorrect, acting as a flag for manual review of these contours. This would make contour propagation more efficient, facilitating the routine implementation of adaptive radiotherapy.

1. Introduction

Intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) have become an integral part of modern radiotherapy (Cozzi et al. 2004; Palma et al. 2008; Vanetti et al. 2009). Their ability to create highly conformal dose distributions has enabled the dose to organs at risk (OARs), and therefore radiation-induced toxicities, to be reduced (Nutting et al. 2011; Gulliford et al. 2012). Accurate delineation of OARs is therefore essential to fully realise the benefits afforded by IMRT and VMAT.

However, significant anatomic changes often occur during radiotherapy (Muren et al. 2003; Fiorino et al. 2005; Hong et al. 2005), and the steep dose gradients present in IMRT and VMAT mean that treatment plans can be particularly sensitive to such changes. As a result, a treatment plan based on anatomy imaged before treatment is often sub-optimal (Hansen et al. 2006; Bhide et al. 2010; Nishi et al. 2013). Adaptive radiotherapy (ART), in which a treatment plan is modified in response to anatomic changes, can be used to mitigate resulting dosimetric consequences (Yan et al. 1997).

On-treatment imaging is an integral component of ART, and accurate delineation of structures on these images is important. Manual delineation is impractical, and so automatic segmentation is essential to enable ART to be introduced into routine clinical use (Zhang et al. 2007; Muren & Thwaites 2013). The most common form of on-treatment automatic segmentation is contour propagation (Thor et al. 2011; Hardcastle et al. 2013; Kumarasiri et al. 2014). For this, an on-treatment image is non-rigidly registered with the planning computed tomography (pCT) image, and the resulting deformation vector field is used to deform the contours from the pCT to the new anatomy of the on-treatment image.

Initial validation of a contour propagation algorithm is essential before clinical use, and several algorithms have been evaluated (Thor et al. 2011; Hou et al. 2011; Hardcastle et al. 2013). However, it is generally considered that they cannot yet replace manual contour review. Furthermore, initial validation cannot guarantee the absence of contour propagation failures in routine clinical use. For contour propagation to be implemented into routine use, patient-specific quality control (QC) of contour propagation is essential. This would facilitate routine implementation of ART.

We present here an automated workflow for real-time patient-specific QC of contour propagation. This workflow monitors contour propagation quality, identifying situations in which the propagated contours are more likely to be subject to large uncertainties. It is validated on a cohort of head and neck cancer patients with two different sources of propagation error.

2. Method

2.1. Automated workflow for contour propagation QC

Figure 1 illustrates the automated workflow for contour propagation QC. The workflow requires two sets of ground truth contours: one on the pCT, and one on an image taken on the first treatment fraction; in this case a cone beam CT (CBCT) (A). Note that the second ground truth can originate from contours propagated from the pCT. At treatment fraction *n*, contours are propagated from pCT onto CBCTn; it is these structures on which QC is performed (B). For this, these structures are propagated back onto CBCT1 (C), such that there are now two sets of structures on CBCT1. The concordance of these structures on CBCT1 is measured using the Dice similarity coefficient (DSC) and the mean distance-to-agreement (DTA). These "consistency metrics" are then used to infer the accuracy of the contours propagated onto CBCTn.

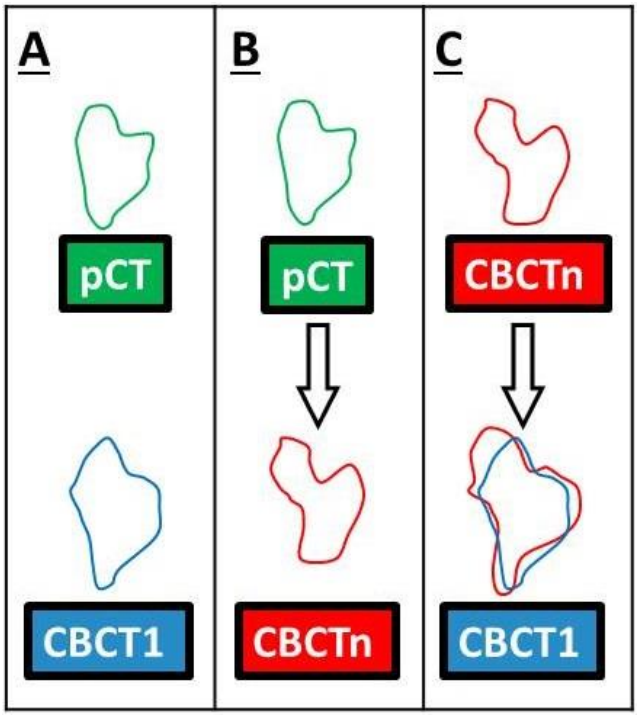

Figure 1: Illustration of the contour QC workflow. Consistency metrics are calculated from the concordance of structures on CBCT1. These consistency metrics are used to infer the quality of contour propagation onto CBCTn.

This workflow was implemented using an in-house Python (v 2.7) script and ADMIRE (ADMIRE v 1.11, Elekta AB, Stockholm, Sweden), an automatic segmentation algorithm. The Python script was used to run the propagations in ADMIRE via a batch file, and to calculate the consistency metrics.

2.2. Image and contour data

Ten head and neck cancer patients who received weekly CBCT imaging as part of a previous study at our institution (Ho et al. 2012) were included in the study. Two observers (GT1 and GT2)

independently contoured the parotids on the pCT and each weekly CBCT for each patient; these structures were taken as the ground truth. ADMIRE was used to propagate these ground truth parotids from the pCT onto each CBCT, and the accuracy of the propagations was measured with DSC and mean DTA. In addition to the accuracy of the propagated contours, the inter-observer variation was estimated from the concordance of the two sets of ground truth structures.

2.3. Workflow validation: error scenario I

The ability of the automated workflow to detect gross propagation errors was tested by copying contours to incorrect images for a subset of patients for a single observer (GT1). Propagated contours on CBCTs 3-6 were copied onto CBCT2, such that the contours on CBCT2 originated from a different image set. The automated workflow was performed on these structures and the consistency metrics were measured. The ability of the uncertainty metrics to identify these errors was investigated.

2.4. Workflow validation: error scenario II

For the second error scenario, Gaussian noise was added to the CBCT images (CBCT2-6) using an in-house Python script. Gaussian kernels with standard deviations of 20 HU, 100 HU, 300 HU, 500 HU and 1000 HU were used. Structures from pCT were propagated onto these noisy images, and the propagation accuracy was measured by comparison with the manual delineations as before. The automated workflow was performed and the consistency metrics were evaluated for each propagation.

For this error scenario, the definition of an error was based on the accuracy of the propagation, using a threshold to determine whether a propagation error had occurred. This threshold was calculated for DSC and mean DTA relative to the ground truth CBCT contour, flagging anything further than three standard deviations from the mean inter-observer variation on pCT. A propagated contour with an accuracy that exceeded this error threshold for either DSC or mean DTA was therefore classified as an error. Note that the threshold was one-sided, such that only propagated contours with discrepancies larger than the mean inter-observer variation were classified as errors; any propagated contours with deviations smaller than the mean inter-observer variation were not classified as errors.

A logistic regression model was trained to detect these errors using the consistency metrics alone. This model was implemented using the Python library ‘scikit-learn’, an open-source machine learning library (Pedregosa et al. 2011). The data were pre-processed to ensure that the consistency metrics had a mean centred on zero and unity standard deviation. Stratified three-fold cross validation was used to split the data into training and test datasets; the model was trained on the training set and was then used to predict whether an error had occurred in the test dataset. The error predictions were compared with the known errors, and receiver operating characteristic (ROC) analysis was performed. The area under the curve (AUC), which provides a measure of the model performance, was calculated. This was performed for each iteration of the stratified three-fold cross validation, and the mean ROC curve

and AUC were used to summarise the model performance. This was performed for both sets of ground truth contours independently.

3. Results

3.1. Propagation performance

The mean accuracy and standard deviations of the propagated contours, as well as the inter-observer variation, are shown in Table 1. Note that there were consistent discrepancies between the parotid contours drawn by the two observers for three patients (on pCT and all CBCTs); the inter-observer variations after excluding these patients are denoted with * in the table. It can be seen from the table that the accuracy and standard deviations of the propagated contours is better than those of the inter-observer variation, indicating good performance of ADMIRE for contour propagation.

Table 1: Mean accuracy and standard deviations of the propagated structures relative to the ground truth structures, and inter-observer variation for the CBCT images and pCT. Note that inter-observer variations calculated excluding the three patients with large discrepancies between observers are denoted with *.

Propagation	DSC	Mean DTA / mm
Propagation accuracy (GT1)	0.82 ± 0.02	1.64 ± 0.26
Propagation accuracy (GT2)	0.79 ± 0.06	1.96 ± 0.43
Inter-observer variation (CBCT)	0.74 ± 0.05	3.52 ± 1.49
Inter-observer variation (CBCT)*	0.75 ± 0.06	3.05 ± 1.13
Inter-observer variation (pCT)	0.84 ± 0.03	2.20 ± 1.18
Inter-observer variation (pCT)*	0.86 ± 0.02	1.57 ± 0.21

3.2. Workflow validation: error scenario I

A plot of the consistency metrics from error scenario I is shown in Figure 2. The errors, illustrated as red stars, are clearly separate from the standard propagations with no errors (shown as green circles). This implies that the consistency metrics can identify these types of gross error using a simple threshold.

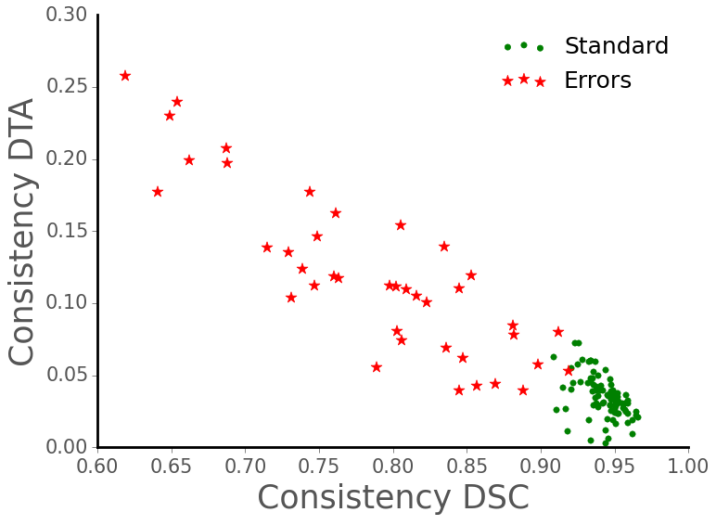

Figure 2: Plot of the consistency metrics for error scenario I. The errors are shown as red stars, and the standard propagations as green circles. The clear separation between the standard propagations and errors means that these errors could be identified from the consistency metrics alone.

3.3. Workflow validation: error scenario II

For error scenario II, noise was added to CBCTs 2-6 for each patient to reduce propagation performance and induce propagation errors. As expected, the accuracy of the contour propagation algorithm reduced with increasing levels of noise, as shown in Figure 3, albeit surprisingly slowly.

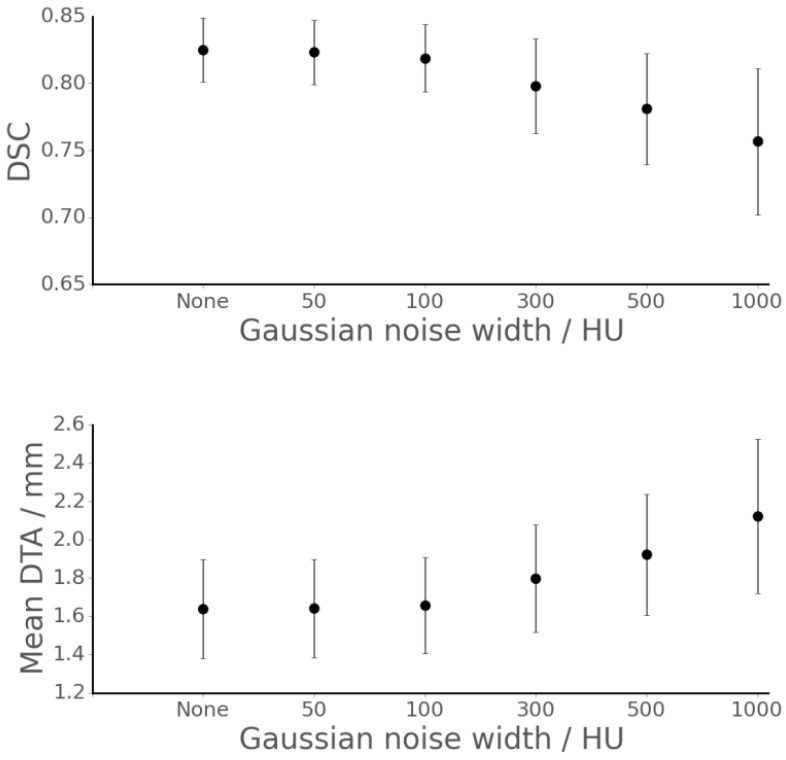

Figure 3: Propagation accuracy reduced with increasing noise.

The automated workflow was applied to these noisy images, and a logistic regression model was trained to predict the occurrence of contour propagation errors. A contour propagation error was defined as any propagated contour with an accuracy (as quantified by DSC **or** mean DTA) more than three standard deviations from the mean inter-observer variation on the pCT (Table 1). Due to large differences between observers on three patients, these were excluded when calculating the error threshold. An error was therefore defined as any propagated contour with a mean DTA greater than 2.20 mm, or with a DSC less than 0.80.

Using these thresholds to define an error, stratified three-fold cross validated ROC analysis was performed using a logistic regression model, the results of which are illustrated in Figure 4. Shown are the resulting ROC curves for both sets of ground truth contours. The red dotted line represents the random guess line. The AUC for the two observers, GT1 and GT2, were 0.90 and 0.85, respectively. It is clear that the model works well for detecting these errors, although the performance was better for GT1 than for GT2. This is likely caused by the larger intra-observer variation in GT2, which can be inferred from the larger standard deviation in the propagation accuracy for GT2 in Table 1.

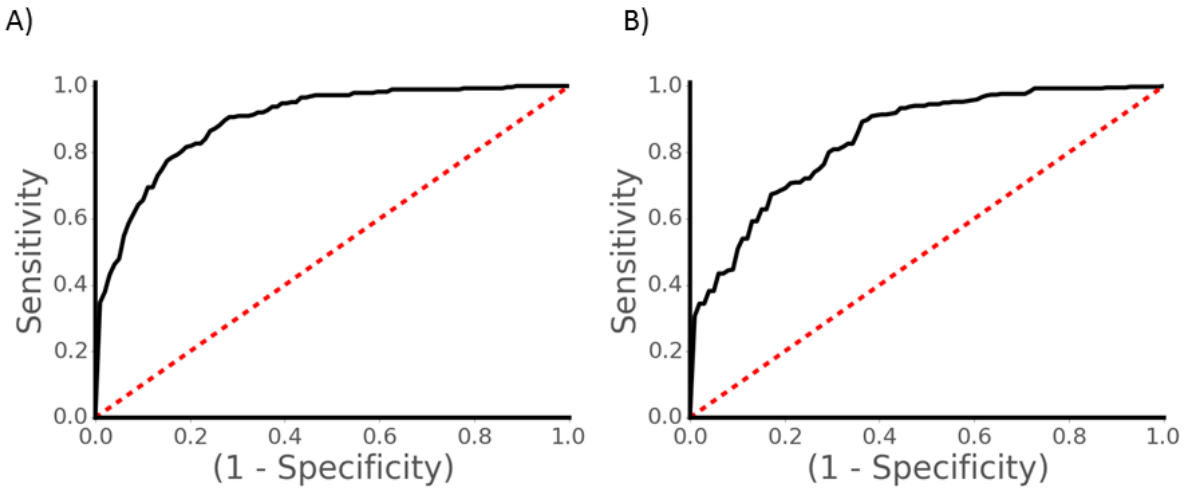

Figure 4: ROC curves for the three-fold validated logistic regression model for GT1 (A) and GT2 (B). The solid black line shows the average curve, and the red dotted line shows the random guess line. The AUC was 0.90 for GT1 and 0.85 for GT2.

4. Discussion

Accurate automatic contour propagation is essential for clinical implementation of ART (Zhang et al. 2007). Although pre-clinical evaluation of an algorithm is important, successful contour propagation cannot be guaranteed in routine clinical use. It is therefore important to perform patient-specific QC on propagated contours, and we have introduced here an automated workflow for QC of contour propagation. This workflow uses consistency metrics, which quantify the consistency of contour propagation over multiple registrations, to monitor propagation performance.

We have shown that this workflow can be used for patient-specific QC of contour propagation, identifying simulated propagation errors. Automatic segmentation is not yet able to fully replace manual delineation, and to facilitate its introduction into clinical use, tools for monitoring propagation performance are important. To the best of the authors' knowledge, no such tools have been reported before.

The concept of automated contour review is not in itself a new concept, however. Machine learning has been suggested as a technique for detecting contouring errors (Kohlberger et al. 2012; McIntosh et al. 2013; Chen et al. 2015). Kohlberger et al. used linear and non-linear regression models to predict the underlying ground truth accuracy of manually-drawn contours using 42 intensity-based and geometric features. McIntosh et al. extended this concept with a random forest model that also included features describing the relative position of structures. They reported that their model could predict manual contouring errors with an AUC of 0.75. Chen et al. 2015 reported a similar method, which was implemented in a graphical user interface to highlight potentially erroneous contours.

These techniques show promise in identifying contouring errors, either manually-defined or automatically-generated. However, they are designed for baseline contour assessment, and do not explicitly test contour propagation. Monitoring the quality of the initial contours is important, as errors at baseline would be propagated onto on-treatment images. Such tests could therefore complement a workflow for monitoring contour propagation quality.

Contour propagation accuracy is closely linked to deformable image registration (DIR) accuracy. There are many metrics for measuring DIR uncertainty, and the automated workflow described here is similar in concept to the distance-to-discordant metric, described by Saleh et al. 2014. This metric measures the geometric uncertainty in each voxel over multiple non-rigid registrations, and they reported that it was related to known registration errors. Such methods could be useful for validating dose warping techniques, but the registration accuracy inside or outside a contour is not necessarily important for monitoring contour propagation accuracy. So tests for monitoring DIR quality may be unnecessarily strict for testing contour propagation accuracy.

The automated error detection workflow described here has been validated on a set of ten head and neck cancer patients, for whom two independent observers outlined the parotids on weekly CBCT images. The parotids were considered in the present study as they are known to undergo anatomic changes during radiotherapy, and the resulting increase in cumulative mean dose is often the reason for ART (Wu et al. 2009). In addition, contour propagation accuracy of the parotids was reasonable, allowing us to gradually introduce errors and test their effect. However, anatomic changes for the parotid are gradual, and so further work is required to assess the workflow for structures that experience different types of motion, such as the rectum or bladder, in which anatomic changes between fractions are more unpredictable.

The workflow has been implemented here with ADMIRE. This algorithm was chosen as its command line interface meant that it could be easily integrated with an in-house Python program, enabling automation of the QC process. Implementation of the workflow with an alternative algorithm should be possible, but its suitability should be carefully evaluated. Both inter- and intra- modality registration is required for the workflow (Figure 1), and any algorithm should be capable of performing both accurately. For example, an algorithm with poor CBCT-CBCT contour propagation performance would create more false positives due to errors introduced when creating the consistency metrics (step C in Figure 1).

Contour propagation accuracy was quantified with DSC and mean DTA, although many other metrics exist for measuring propagation accuracy. However, there is no consensus on which are the most appropriate to use. DSC and mean DTA have been used here as they are commonly-reported metrics, and there is evidence that mean DTA is related to discrepancies in the mean dose for the parotids (Beasley et al. 2016). However, inclusion of additional geometric metrics, as well as intensity-based metrics, could potentially improve the ability of the workflow to detect propagation failures.

One of the difficulties in assessing an automatic contouring algorithm is the uncertainty in the ground truth contours. Manual segmentations are inherently subject to intra- and inter-observer variation (Nelms et al. 2012), which limits the perceived contour propagation accuracy. This uncertainty in the ground truth contours also limits the ability of the automated workflow described here to detect propagation failures. Uncertainty in the propagation accuracy translates into uncertainty in the designation of propagation failures as defined by a threshold on DSC and mean DTA. A larger uncertainty in the ground truth causes a larger uncertainty in the definition of propagation errors, limiting the ability to detect errors. This was apparent in our data when the logistic regression model was tested on the second observer (GT2). The propagation accuracy was lower and the variance higher for GT2 than for GT1, implying that there was a larger amount of uncertainty in GT2. Indeed, when the model was applied to GT2, the AUC was lower than for GT1. This uncertainty in the ground truth contours therefore limits our ability to test whether a model can correctly identify propagation errors.

Propagation errors were defined as any propagated contour with discrepancies in DSC or mean DTA more than three standard deviations from the mean inter-observer variation on pCT. The inter-observer variation on pCT was chosen because this represented the ‘true’ inter-observer variation; as the same observers outlined the parotids on both pCT and CBCT, the larger inter-observer variation on CBCT was likely caused by the poorer image quality of the CBCT images. However, the exact choice of error threshold does not affect the overall functionality of the logistic regression model, as training the model implicitly accounts for the choice of threshold used to classify an error.

The proposed workflow was validated using simulated errors, as the performance of the contour propagation software was acceptable for the available patients. Two types of error were simulated: gross errors, in which the contours were effectively not propagated according to the underlying deformation vector field (error scenario I); and errors resulting from uncertainty in the propagation due to noise in the images (error scenario II). Although these errors are not necessarily realistic, they tested different sources of potential failure. Error scenario I simulated a grossly incorrect contour, and it would be expected that any error detection workflow would be capable of detecting such errors. In error scenario II, noise was added to the CBCT images; although noise of this type is not necessarily realistic, it introduces uncertainty into the registration and so the ability of the workflow to detect propagation uncertainty was tested. Nevertheless, further work is required to verify the workflow on clinically-observed errors.

Automatic contour propagation is essential for ART, but manual review of propagated contours is often necessary. A recent study described a workflow for online plan adaptation using magnetic resonance image guided radiotherapy (Acharya et al. 2015). Although they reported a reasonable time for plan adaptation (median time of 26 minutes), it was necessary for a clinician to manually review the propagated contours at each treatment fraction. This would be a barrier for routine ART implementation. The automated workflow we have presented here is a potential solution to this problem, as it could be used to highlight contours that are likely to be incorrect, acting as a flag for manual review by a clinician. For this, the model parameters would be optimised to obtain an appropriate balance between false positives (incorrect prediction of an error), resulting in unnecessary contour review, and false negatives (incorrect prediction of the absence of an error), resulting in an incorrect contour going unnoticed. This would therefore ensure that only contours with potential errors would be manually reviewed, improving ART efficiency.

Patient-specific QC of contour propagation is important to facilitate the routine implementation of ART, and the automated workflow described here shows potential as a tool for patient-specific QC of contour propagation, enabling ART to become more feasible.

5. Conclusion

Contour propagation is an essential component of ART, but unreliable propagation limits its routine clinical implementation. There are currently no tools to aide patient-specific QC of contour propagation. An automated workflow for patient-specific QC of contour propagation, based on consistency metrics calculated from multiple registrations, has been presented and tested on a set of ten head and neck patients with simulated propagation errors. This workflow shows potential as a tool for quality control of contour propagation, and could help facilitate the clinical implementation of adaptive radiotherapy.

Acknowledgements

ADMIRE was supplied by Elekta as part of a research agreement.

References

Acharya, S. et al., 2015. Online magnetic resonance image guided adaptive radiation therapy: first clinical applications. *International Journal of Radiation Oncology, Biology, Physics*, 94(2), pp.394–403.

Beasley, W.J. et al., 2016. The suitability of common metrics for assessing parotid and larynx autosegmentation accuracy. *Journal of Applied Clinical Medical Physics*, 17(2), pp.41–49.

Bhide, S. a et al., 2010. Weekly volume and dosimetric changes during chemoradiotherapy with intensity-modulated radiation therapy for head and neck cancer: a prospective observational study. *International Journal of Radiation Oncology, Biology, Physics*, 76(5), pp.1360–8.

Chen, H.-C. et al., 2015. Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: A general strategy. *Medical Physics*, 42, pp.1048–1059.

Cozzi, L. et al., 2004. Three-dimensional conformal vs. intensity-modulated radiotherapy in head-and-neck cancer patients: Comparative analysis of dosimetric and technical parameters. *International Journal of Radiation Oncology, Biology, Physics*, 58(2), pp.617–624.

Fiorino, C. et al., 2005. Rectal and bladder motion during conformal radiotherapy after radical prostatectomy. *Radiotherapy and Oncology*, 74(2), pp.187–195.

Gulliford, S.L. et al., 2012. Dosimetric explanations of fatigue in head and neck radiotherapy: an analysis from the PARSPORT Phase III trial. *Radiotherapy and oncology*, 104(2), pp.205–12.

Hansen, E.K. et al., 2006. Repeat CT imaging and replanning during the course of IMRT for head-and-neck cancer. *International Journal of Radiation Oncology, Biology, Physics*, 64(2), pp.355–62.

Hardcastle, N. et al., 2013. Accuracy of deformable image registration for contour propagation in adaptive lung radiotherapy. *Radiation Oncology*, 8(1), p.243.

Ho, K.F. et al., 2012. Monitoring dosimetric impact of weight loss with kilovoltage (kV) cone beam CT (CBCT) during parotid-sparing IMRT and concurrent chemotherapy. *International Journal of Radiation Oncology, Biology, Physics*, 82(3), pp.e375–82.

Hong, T.S. et al., 2005. The impact of daily setup variations on head-and-neck intensity-modulated radiation therapy. *International Journal of Radiation Oncology, Biology, Physics*, 61(3), pp.779–788.

Hou, J. et al., 2011. Deformable planning CT to cone-beam CT image registration in head-and-neck cancer. *Medical Physics*, 38(4), p.2088.

Kohlberger, T. et al., 2012. Evaluating segmentation error without ground truth. *Medical Image Computing and Computer-Assisted Intervention*, 15, pp. 528–36.

Kumarasiri, A. et al., 2014. Deformable image registration based automatic CT-to-CT contour propagation for head and neck adaptive radiotherapy in the routine clinical setting. *Medical Physics*, 41(12), p.121712.

McIntosh, C., Svistoun, I. & Purdie, T.G., 2013. Groupwise conditional random forests for automatic shape classification and contour quality assessment in radiotherapy planning. *IEEE Transactions on Medical Imaging*, 32(6), pp.1043–57.

Muren, L.P., Smaaland, R. & Dahl, O., 2003. Organ motion, set-up variation and treatment margins in radical radiotherapy of urinary bladder cancer. *Radiotherapy and Oncology*, 69(3), pp.291–304.

Muren, L.P. & Thwaites, D.I., 2013. The on-going quest for treatment precision and conformality in radiotherapy. *Radiotherapy and Oncology*, 109(3), pp.337–41.

Nelms, B.E. et al., 2012. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *International Journal of Radiation Oncology, Biology, Physics*, 82(1), pp.368–78.

Nishi, T. et al., 2013. Volume and dosimetric changes and initial clinical experience of a two-step adaptive intensity modulated radiation therapy (IMRT) scheme for head and neck cancer. *Radiotherapy and*

Oncology, 106(1), pp.85–9.

Nutting, C.M. et al., 2011. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial. *The Lancet Oncology*, 12(2), pp.127–36.

Palma, D. et al., 2008. Volumetric modulated arc therapy for delivery of prostate radiotherapy: comparison with intensity-modulated radiotherapy and three-dimensional conformal radiotherapy. *International Journal of Radiation Oncology, Biology, Physics*, 72(4), pp.996–1001.

Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

Saleh, Z.H. et al., 2014. The distance discordance metric-a novel approach to quantifying spatial uncertainties in intra- and inter-patient deformable image registration. *Physics in Medicine and Biology*, 59(3), pp.733–46.

Thor, M. et al., 2011. Deformable image registration for contour propagation from CT to cone-beam CT scans in radiotherapy of prostate cancer. *Acta Oncologica*, 50(6), pp.918–25.

Vanetti, E. et al., 2009. Volumetric modulated arc radiotherapy for carcinomas of the oro-pharynx, hypo-pharynx and larynx: A treatment planning comparison with fixed field IMRT. *Radiotherapy and Oncology*, 92(1), pp.111–117.

Wu, Q. et al., 2009. Adaptive replanning strategies accounting for shrinkage in head and neck IMRT. *International Journal of Radiation Oncology, Biology, Physics*, 75(3), pp.924–32.

Yan, D. et al., 1997. Adaptive radiation therapy. *Physics in Medicine and Biology*, 42(1), pp.123–132.

Zhang, T. et al., 2007. Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy. *International Journal of Radiation Oncology, Biology, Physics*, 68(2), pp.522–30.