2-21-2015

A proposed framework for consensus-based lung tumour volume auto-segmentation in 4D computed tomography imaging.

Spencer Martin

Mark Brophy

David Palma

Alexander V Louie

Edward Yu

See next page for additional authors

Authors
Spencer Martin, Mark Brophy, David Palma, Alexander V Louie, Edward Yu, Brian Yaremko, Belal Ahmad, John L Barron, Steven S Beauchemin, George Rodrigues, and Stewart Gaede

A proposed framework for consensus-based lung tumour volume auto-segmentation in 4D computed tomography imaging

2015 Phys. Med. Biol. 60 1497

(http://iopscience.iop.org/0031-9155/60/4/1497)

IOP Publishing | Institute of Physics and Engineering in Medicine Physics in Medicine & Biology

Phys. Med. Biol. **60** (2015) 1497–1518 doi:10.1088/0031-9155/60/4/1497

A proposed framework for consensus-based lung tumour volume auto-segmentation in 4D computed tomography imaging

Spencer Martin[1], Mark Brophy[2], David Palma[3], Alexander V Louie[3], Edward Yu[3], Brian Yaremko[3], Belal Ahmad[3], John L Barron[2], Steven S Beauchemin[2], George Rodrigues[1,3,4,5] and Stewart Gaede[1,4,5,6]

[1] Department of Medical Biophysics, University of Western Ontario, London, Ontario, Canada
[2] Department of Computer Science, University of Western Ontario, London, Ontario, Canada
[3] Department of Oncology, University of Western Ontario, London, Ontario, Canada
[4] Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada
[5] Departments of Biomedical Engineering, University of Western Ontario, London, Ontario, Canada

E-mail: stewart.gaede@lhsc.on.ca

Received 14 July 2014, revised 28 November 2014
Accepted for publication 11 December 2014
Published 22 January 2015

Abstract

This work aims to propose and validate a framework for tumour volume auto-segmentation based on ground-truth estimates derived from multi-physician input contours to expedite 4D-CT based lung tumour volume delineation. 4D-CT datasets of ten non-small cell lung cancer (NSCLC) patients were manually segmented by 6 physicians. Multi-expert ground truth (GT) estimates were constructed using the STAPLE algorithm for the gross tumour volume (GTV) on all respiratory phases. Next, using a deformable model-based method, multi-expert GT on each individual phase of the 4D-CT dataset was propagated to all other phases providing auto-segmented GTVs and motion encompassing internal gross target volumes (IGTVs) based on GT estimates (STAPLE) from each respiratory phase of the 4D-CT dataset. Accuracy assessment of auto-segmentation employed graph cuts for 3D-shape reconstruction and point-set registration-based analysis yielding volumetric and distance-based measures. STAPLE-based auto-segmented

[6] Author to whom any correspondence should be addressed. 790 Commissioners Rd E, N6A 4L6 London, Ontario, Canada

GTV accuracy ranged from (81.51 ± 1.92) to (97.27 ± 0.28)% volumetric overlap of the estimated ground truth. IGTV auto-segmentation showed significantly improved accuracies with reduced variance for all patients ranging from 90.87 to 98.57% volumetric overlap of the ground truth volume. Additional metrics supported these observations with statistical significance. Accuracy of auto-segmentation was shown to be largely independent of selection of the initial propagation phase. IGTV construction based on auto-segmented GTVs within the 4D-CT dataset provided accurate and reliable target volumes compared to manual segmentation-based GT estimates. While inter-/intra-observer effects were largely mitigated, the proposed segmentation workflow is more complex than that of current clinical practice and requires further development.

Keywords: 4D-CT, STAPLE, lung cancer

(Some figures may appear in colour only in the online journal)

1. Introduction

High-dose image-guided radiotherapy (IGRT) and intensity modulated radiation therapy (IMRT) with radical intent for non-small cell lung cancer (NSCLC) treatment requires the delineation or segmentation of the primary tumour volume (the gross tumour volume or GTV) within 3D or 4D computed tomography (CT) images by the physician for treatment plan optimization purposes. Subsequent margins are added for clinical target volumes and planning target volumes (CTV and PTV) accounting for uncertainties such as microscopic disease and setup error, respectively. Highly accurate segmentation of the GTV is of the utmost importance in order to maximize the therapeutic ratio. More advanced treatment planning algorithms and radiation delivery techniques further work to enhance radiotherapy precision and patient outcomes. However, as precision improves and prescription dose increases, segmentation error and variability in target volume definition can have greater impact on treatment efficacy. Segmentation-related errors including inter- and intra-observer variability has been demonstrated with respect to lung cancer in 3D image acquisition (Giraud *et al* 2002, Van de Steene *et al* 2002). Clinical margins exist to account for internal motion of the tumour volume, setup error, geographic miss of the target volume and irradiation of healthy tissue, however, if variability and segmentation-related error exceeds the clinical margins, patient outcomes and treatment efficacy can be compromised. Jameson *et al* stated that variability in anatomical segmentation is the largest contributor to uncertainty in the radiotherapy process (Jameson *et al* 2010). This uncertainty can magnify other sources of error such as setup, inter- and intra-fractional motion. Techniques incorporating functional and/or metabolic imaging (Caldwell *et al* 2001, Steenbakkers *et al* 2006) and automatic segmentation strategies (Gaede *et al* 2011) have been reported to reduce segmentation variability due to inter- and intra-observer variance and expedite the process of image segmentation while maintaining accuracy.

The principle issue in segmentation of lung tumours in 3D-CT is the inability of the image acquisition and reconstruction process to account for tumour motion during respiration. Although breath-hold techniques can mitigate the effects of respiratory motion during treatment (Josipovic *et al* 2013), patients with compromised lung function may not be able to perform voluntary or involuntary breath-hold during treatment. Therefore, assessing the motion of lung tumours is essential for accurate radiotherapy treatment planning and delivery

purposes that still allow patients to breathe freely during treatment. Movement of tumours during respiration results in different types of image artifacts, which have been documented in past studies (Ekberg *et al* 1998, Barnes *et al* 2001, Erridge *et al* 2003, Plathow *et al* 2004, Persson *et al* 2010, 2011). These artifacts influence manual, automatic, and/or assisted segmentation of the tumour volume. The addition of arbitrary margins to the segmented GTV defined on the artifact inclusive CT scan can subsequently lead to geographic miss of the target volume and irradiation of normal, healthy tissue during radiotherapy treatment delivery.

Four-dimensional computed tomography (4D-CT) has become the optimal strategy for acquiring artifact-free image data and assessing respiratory-induced lung tumour motion. Numerous studies have led to improvements in 4D-CT acquisition/reconstruction techniques (Low *et al* 2003, Vedam *et al* 2003, Pan *et al* 2004). 4D-CT image reconstruction allows for free breathing patient CT images to be retrospectively sorted into 8–12 respiratory phase-based bins. These phase-binned image datasets are then used for target volume segmentation during different phases of the breathing cycle that can provide for phase-specific GTV segmentations and an envelope of the GTV throughout a respiratory cycle called the internal gross tumour volume (IGTV). While 4D-CT provides an improvement over conventional CT in terms of image integrity, defining the IGTV may result in increased segmentation error and geometric uncertainty as multiple 3D image volumes comprise the 4D dataset. A straightforward approach to IGTV segmentation is to have experts segment the GTV on each individual respiratory phase and then integrate those segmentations.

Very little consensus exists on the extent of tumour movement due to respiratory motion as it can be quite complex and highly variable between patients while showing dependence on the location of the tumour itself (Yan *et al* 2008). This makes motion-encompassing methods for target segmentation somewhat unreliable compared to manual segmentation on multiple respiratory phases. However, time requirements and considerable amounts of inter- and intra-observer variability limit the clinical feasibility of this practice. Many studies have reported on IGTV segmentation in 4D-CT datasets (Ezhil *et al* 2009, van Dam *et al* 2010, Speight *et al* 2011). Reports on the effectiveness of maximum- and average-intensity projection CT images offer marginal improvement and mixed results in the context of plan dosimetry (Ehler and Tomé 2008, Huang *et al* 2010). End exhalation and end inhalation phases of the 4D-CT dataset have also been used for target volume segmentation but these strategies do not account for hysteresis during intra-fraction motion and may miss motion outside of the boundary defined by the two phases. Auto-segmentation strategies have the potential to reduce segmentation time and mitigate segmentation-related geometric uncertainties in 4D-CT. Strategies utilizing deformable surface models and/or deformable image registration techniques for this purpose have been reported with varying levels of success (Kaus *et al* 2001, 2004, Ragan *et al* 2005, Pevsner *et al* 2006). It is vital that auto-segmentation strategies in 4D-CT are both efficient and highly accurate to limit the need for physician editing. While auto-segmentation reduces the clinical workload and the potential for intra-physician variance, inter-physician variance remains a problem. Additionally, when quantitatively evaluating image segmentation, numerous questions arise in the context of multi-expert studies and assessing automatic segmentation performance, many of which are summarized concisely by Gordon *et al* (Gordon *et al* 2009). Ground-truth (GT) estimation algorithms aim to address those questions by incorporating multiple expert markings and providing for a reference standard segmentation in quantitative evaluation and validation of medical image segmentation analysis (Chalana and Kim 1997, Gordon *et al* 2009). Utilizing multiple expert image marking allows for generation of estimates of lesion locations that take into account multiple physicians' input with an expectation that a result is a closer representation to the tumour's true location due to the assumed minimization of each expert's subjectivity in segmentation (Gordon *et al* 2009, Biancardi *et al* 2010).

The goal of this study is to be one of the first to couple consensus-based tumour volume segmentation by way of the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm (Warfield *et al* 2004) in 4D-CT datasets based on *N* physicians input segmentations with deformable model-based automatic segmentation technique capable of integration into existing treatment planning systems (TPS) for multi-expert 4D-CT tumour volume auto-segmentation. Little to no literature exists on this topic and it has the potential to improve target volume segmentation by incorporating multiple expert markings with the capability of eliminating inter-/intra-observer variability for IGRT purposes. As depicted in the workflow presented in figure 1, each 4D image sequence is manually segmented allowing for each 3D image volume within the 4D dataset forming a reference, STAPLE derived segmentation (GTV_i). Those segmentations are subsequently used as the basis for auto-segmentation to the other respiratory phase specific 3D image volumes ($GTV_{i,k}$), which are then analyzed via comparison to the manually derived STAPLE segmentations. Reference segmentations (GTV_i) are enveloped to form a motion encompassing STAPLE segmentation for each patient (IGTV). Respiratory phase-based target segmentations ($GTV_{i,k}$) are also enveloped to create phase-based motion encompassing volumes ($IGTV_i$) that are compared to the reference IGTVs. Through this process, depicted in figure 1, we can incorporate the STAPLE algorithm into existing auto-segmentation strategies and validate both the efficacy and accuracy of auto-segmentated multi-expert consensus volumes. Additionally, by selecting each respiratory-phase within the 4D-CT dataset as the basis of propagation within the proposed workflow, we are able to determine approximately which state of the breathing cycle (e.g. maximum exhalation, mid-inhalation, maximum inhalation, mid-exhalation, etc) provides the optimal basis for tumour volume auto-segmentation in the context of 4D imaging.

2. Methods and materials

2.1. Image acquisition and reconstruction

4D-CT imaging was performed on ten patients with NSCLC. Patient demographics and disease information are presented in table 1. Local REB approval was obtained and all data was anonymized prior to any segmentation. A Philips 16-slice Brilliance Big Bore CT scanner (Philips Medical Systems, Cleveland, USA) was used with the pulmonary gating application to image patients. The Real-Time Position Management (RPM) respiratory gating system (Varian Medical Systems, Palo Alto, USA) was used as a respiratory surrogate. The RPM system uses an infrared camera that follows reflective markers placed on the patient's chest or abdomen. For all ten patients, a long spiral CT scan with pitch < 0.1 was performed to encompass the entire thorax. Pulmonary signal data was collected from the RPM system simultaneously with CT data. CT data was then reconstructed at ten different respiratory phases (*i*). Respiratory phases were tagged according to temporal location along the respiratory cycle, indicating temporal steps from one full inspiration phase to another ($i = 0, 10, \ldots, 90\%$). This form of image reconstruction allows for visualization of tumour volume displacement at ten equally spaced points in time throughout the respiratory cycle. Due to this 4D reconstruction method, for all patients the phases (*i*) correspond to each other.

2.2. Manual segmentation

The GTV was segmented on each of the ten respiratory phases for ten patients by six radiation oncologists with clinical experience ranging from 1 to 25 years experience.

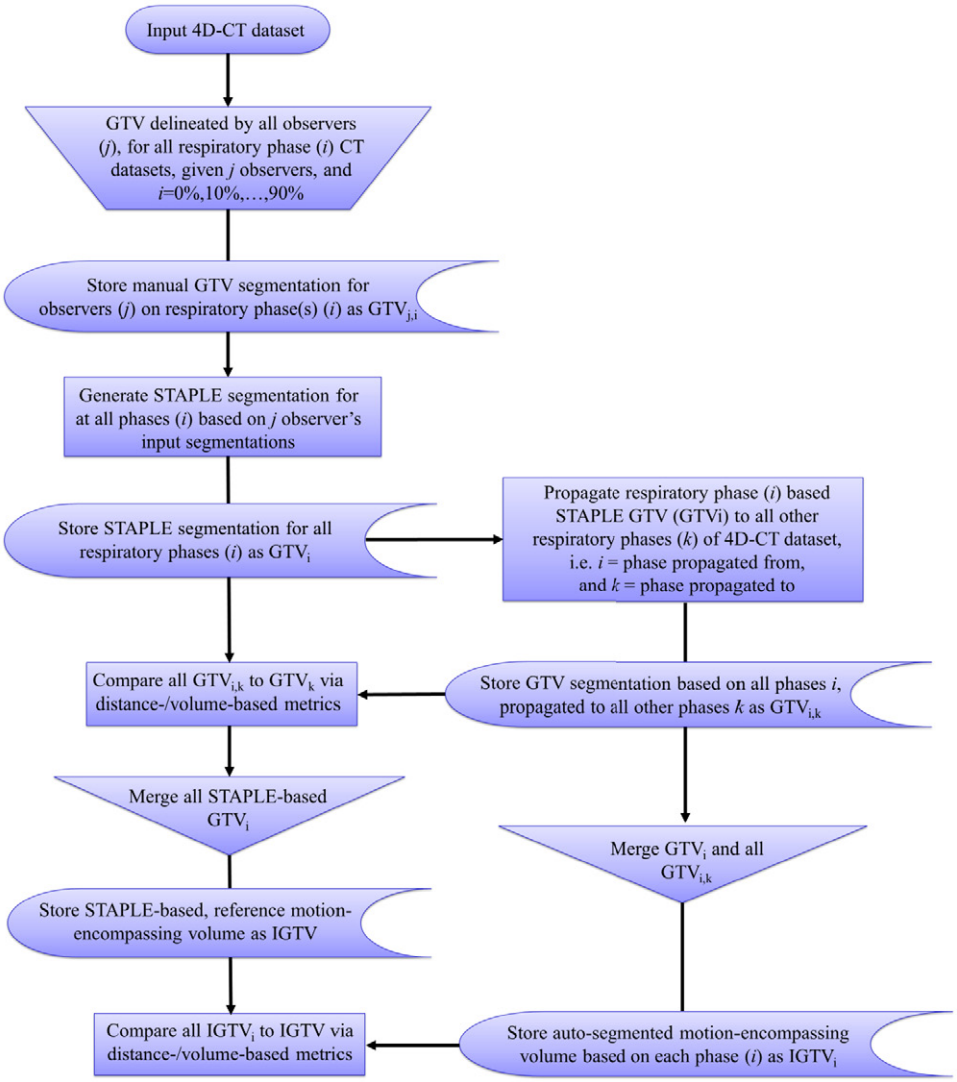

Figure 1. Methodology flow chart. Each 4D image sequence is manually segmented on all phases by j expert observers to allow for phase-specific STAPLE derived segmentations (GTV_i). Those segmentations are subsequently used as the basis for auto-segmentation to k remaining respiratory phase specific 3D image volumes ($GTV_{i,k}$), which are then assessed via comparison to the manually derived STAPLE segmentations (GTV_i). Reference segmentations (GTV_i) are also enveloped to form a motion encompassing STAPLE segmentation for each patient (IGTV). Respiratory phase-based target segmentations ($GTV_{i,k}$) are subsequently enveloped to create phase-based motion encompassing volumes ($IGTV_i$) that are compared to the reference IGTVs.

Target volume segmentation was done within the Pinnacle Treatment Planning System 9.1y (Philips Medical Systems, Fitchburg, WI). Default visualization parameters were used for lung tumour segmentation by all physicians in accordance with recommendations by Giraud *et al* (Giraud *et al* 2002) (−600/1600 HU for lung window, +20/400 HU

Table 1. Patient demographic, disease information, and manual segmentation difficulty.

Patient	Staging	MSD[a]	Location	Volume (cm^3)
A	IIIA	3.00 (Difficult)	RLL	272.63
B	IIIA	4.50 (Difficult)	RLL	67.70
C	IIB	3.00 (Difficult)	RLL	29.86
D	IIIA	1.83 (Easy)	RLL	46.99
E	IIIA	1.33 (Easy)	LLL	10.42
F	IIB	3.50 (Difficult)	RUL	17.95
G	IIIA	2.17 (Easy)	LUL	2.24
H	IIIA	3.33 (Difficult)	LUL	20.80
I	IIB	1.83 (Easy)	RLL	28.18
J	IIIA	2.50 (Easy)	RUL	90.10

[a] MSD: Mean Segmentation Difficulty.

for the mediastinal window). The IGTV envelope, defined as the union of GTV segmentation from all respiratory phases, was created using MiMVista v.5.2 (MiM Software, Cleveland, USA). Experts were blinded to one another's segmentations and were provided with a representative axial 2D CT image indicating location of the primary GTV. Upon completion of manual segmentation, experts were asked to assign a difficulty level on a scale of 1 (least difficult) to 5 (most difficult) for each case. The five cases with average difficulty above 2.5 were classified as difficult while the remaining cases were classified as easy. Patient characteristics and manual segmentation difficulty scores can be seen below in table 1.

2.3. Consensus segmentation

The STAPLE algorithm has been shown to be both highly robust and adept at incorporating and fusing multiple expert segmentations into one GT estimate segmentation or multi-expert GT segmentation (Biancardi *et al* 2010). The STAPLE technique, proposed by Warfield *et al* (Warfield *et al* 2004), takes a collection of *J* binary image segmentations as inputs and simultaneously computes a probabilistic estimate of the true tumour segmentation and a measure of the performance level of each observer's input segmentation. STAPLE utilizes an expectation-maximization (EM) approach to improve the initial GT estimate in an iterative fashion. In the E-step, the unobserved true segmentation is computed as a probability map where each voxel is assigned a probability of being a part of the segmented object (i.e. tumour volume). In the M-step, observer's performance level parameters are estimated by maximizing the complete data log-likelihood using the current reference segmentation generated at the E-step. The final step of the STAPLE algorithm is the generation of the final estimate of GT based on the construction of a hidden Markov Random Field (MRF). Further details of the algorithm are reported elsewhere (Boykov and Kolmogorov 2004, Commowick and Warfield 2010).

In using the STAPLE algorithm for GT estimation, it is important to remember that the algorithm itself benefits from prior information made available by way of input segmentation reliability. This means that the performance of the GT estimate relies on the level of expertise of the observers input segmentation that are used to formulate the GT estimate. As such, radiation oncologists from the London Regional Cancer Program (LRCP) were selected to provide input manual segmentations for this study.

2.4. Auto-segmentation

Deformable surface models have been used previously in the setting of respiratory motion and 4D-CT imaging, demonstrating robustness with respect to image artifacts and consistent capturing of deformation at all respiratory phases. Target volume auto-segmentation employs a deformable model-based method, where a triangular surface mesh is adapted to an image through iterative process of surface detection and reconfiguration of triangle vertices through minimization of the energy function ($E = E_{\text{ext.}} + \alpha E_{\text{int.}}$) . The internal shape energy term ($E_{\text{int.}}$) maintains the vertex configuration of an intial mesh while the α parameter regulates the influence of the external feature energy term ($E_{\text{ext.}}$) . This drives the mesh toward detected surface points. Surface detection is performed for each triangle center and achieved by maximizing a cost function that evaluates displacements along the triangle normal to deform to a feature based on grey-scale value transitions while simultaneously restricting large displacement vectors (Kaus *et al* 2004). This deformable model-based method was chosen as it was integrated into the Pinnacle Treatment Planning System 9.1y (Philips Medical Systems, Fitchburg, WI) analogous to the work by Gaede *et al* (Gaede *et al* 2011). No auto-segmented volumes in this study were subject to any form of observer review to allow for unbiased assessment of the auto-segmentation accuracy.

2.5. Accuracy validation

To assess the accuracy of the proposed consensus-based segmentation strategy in this study, reference respiratory phase specific STAPLE GTV segmentations were propagated to all other respiratory phases within the 4D-CT dataset for all patients. Subsequently, respiratory-phase specific, auto-segmented IGTV segmentations were also generated. These auto-segmented volumes were then compared to their respective respiratory phase specific, manually derived, true STAPLE segmentations that were available for all respiratory-phase GTV and IGTV segmentations. Via this proposed methodology, auto-segmented STAPLE segmentations for all possible tumour volume segmentations can be compared back to a manually derived reference STAPLE volume to assess accuracy. A workflow of this methodology can be seen in figure 1.

Following the work by Heimann *et al*, metrics were chosen to assess segmentation accuracy in terms of both average and maximum surface distance errors as well as volumetric error to better convey information regarding segmentation quality and estimate overall segmentation accuracy (Heimann *et al* 2009). To analyze the segmentation volumes in this study, a global optimization framework (e.g. graph-cut) for 3D shape reconstruction was implemented to allow for 3D analysis of segmented volumes outside of the treatment planning system (TPS) in which the volumes were originally segmented. The method proposed by Lempitsky and Boykov was used for this as it provided for robust, high-resolution surface reconstruction with global optimality that ensured a set of closed, minimal surfaces were generated for comparison (Lempitsky and Boykov 2007). This technique was used in conjunction with a surface-extraction technique with higher-order smoothness (Lempitsky 2010) for improved visualization and qualitative analysis of the reconstructed segmentation volumes as well.

2.5.1. Volume-based analysis. Volume-based metrics are some of the most commonly used metrics in analysis of tumour volume segmentation accuracy and quality. Numerous studies have implemented volume metrics yielding different results and different opinions regarding the effectiveness of said metrics and are summarized in studies by Fotina *et al* and Jameson *et al* (Jameson *et al* 2010, Fotina *et al* 2012). Two of the most common metrics are the Dice Similarity Coefficient (DSC) and the Jaccard Index (JI). These metrics have both been used

extensively in segmentation analysis and are closely associated mathematically, with little preference shown for one or the other. However, a recent study by Fotina *et al* demonstrated that the DSC allots double value to the overlap area and tends to over-estimate the amount of agreement between two segmentations (Fotina *et al* 2012). As such, the JI was chosen as the volumetric overlap metric in this study. The volumetric overlap (VO) is then given as a percent of agreement between two volumes and is defined as $100\% \times (V_A \cup V_B / V_A \cup V_B)$. A value of 0 represents complete disagreement between volumes, while a value of 100% represent perfect volumetric overlap. The VO calculations were made possible by way of the globally optimal surface reconstruction technique utilized in this study.

2.5.2. Distance-based analysis. For this study, the root mean square (RMS) symmetric surface distance and the maximum symmetric surface distance (or Hausdorff Distance) (HD) measures were calculated as they represent the average and maximum error measures between segmentations, respectively. The RMS symmetric surface distance is measured in millimeters (mm) and is based on surface voxels of two separate segmentations, whose surfaces are given by $S(A)$ and $S(B)$, and point clouds representing segmentations given by A and B, respectively. As each segmentation is a 3D point cloud, the coherent point drift (CPD) algorithm (Myronenko and Song 2010) was used to calculate distance-based metrics in this study. This registration technique was chosen to provide for symmetric correspondence in measurement between point clouds. Typically, in calculating the RMS distance between two point clouds, say the distance from each point a in cloud A to its corresponding point b in B, the metric will be weighted largely by non-overlapping parts and registration and measurement of distance between the two will be asymmetric. To overcome this and establish symmetric correspondence, a methodology was adopted for calculating the RMS symmetric surface distance by first implementing CPD then utilizing an approximate nearest neighbour technique. Using CPD, we register A to B, yielding A' and finding the nearest neighbour correspondence of each point in A' in B to yield B-neighbors, a new unique set. We then register B-neighbors to A, yielding B-neighbors′ and calculate the approximate nearest neighbors of each point of B-neighbors′ in A. This yields the unique set A-neighbors. Using the two point clouds, A-neighbors and B-neighbors, we can calculate the RMS distance between the two with symmetric correspondence. This is done by calculating and storing the squared Euclidean distance between each set of corresponding points. The average RMS symmetric distance is then defined as the average of all stored distances, where 0 represents a perfect segmentation. Formally, the shortest distances between arbitrary corresponding voxels in $S(A_{\text{neighbors}})$ and $S(B_{\text{neighbors}})$ is given by:

$$d\left(s_{A_n}, S\left(B_n\right)\right) = \min_{s_{B_n} \in S(B)} \left\| s_{A_n} - s_{B_n} \right\| \tag{1}$$

$$d\left(s_{B_n}, S\left(A_n\right)\right) = \min_{s_{A_n} \in S(A)} \left\| s_{B_n} - s_{A_n} \right\| \tag{2}$$

Where s_X represents an arbitrary surface voxel in surface $S(X)$. The symmetric RMS distance between corresponding points in $S(A_{\text{neighbors}})$ and $S(B_{\text{neighbors}})$ is then given by:

$$\begin{aligned} \text{RMSD}\left(A, B\right) = &\left(\sqrt{\frac{1}{|S\left(A_n\right)| + |S\left(B_n\right)|}} \right) \\ &\times \sqrt{\sum_{s_{A_n} \in S(A)} d^2\left(s_{A_n}, S\left(B_n\right)\right) + \sum_{s_{B_n} \in S(B)} d^2\left(s_{B_n}, S\left(A_n\right)\right)} \end{aligned} \tag{3}$$

As stated by Heimann *et al*, this metric is one of the most important in evaluating segmentation accuracy (Heimann *et al* 2009). The maximum symmetric surface distance (MSD) is also measured in millimeters and is determined implicitly with the RMS symmetric distance. This measure is better known as the Hausdorff distance (HD) (Huttenlocher *et al* 1993) and as is determined as the maximum Euclidean distance or maximum symmetric surface distance. A perfect match between two segmentations yields a value of zero, and this measure is formally given as:

$$\mathrm{MSD}\,(A,B) = \mathrm{HD}\,(A,B) = \max\left\{ \max_{s_{A_n}\in S(A)} d\left(s_{A_n}, S\left(B_n\right)\right), \max_{s_{B_n}\in S(B)} d\left(s_{B_n}, S\left(A_n\right)\right) \right\} \tag{4}$$

This metric was included as it gives the maximum error in segmentation analysis with sensitivity to outliers. To determine the statistical significance of the measures used in this study, two analysis methods were used. To determine statistical significance between GTV and IGTV measurements (between-groups), post-hoc, unequal variance *t*-tests were used (Welch's *t*-test). This allowed for comparison of unequal populations (GTV and IGTV measurements) for each patient. In determining the statistical differences between respiratory phase-specific measurements (within groups), a one-way ANOVA test was used with a Bonferroni correction of $\alpha / 10$. This test allowed us to determine statistical differences between respiratory phase-based GTV measurements on an intra-patient basis with correction for multiple comparisons. For within groups analysis, p-values < 0.005 were considered statistically significant and p-values < 0.05 were considered statistically significant in between groups analysis.

3. Results

Figure 2 shows representative slices in the axial, sagittal, and coronal planes and 3D view of 6 physicians manual IGTV segmentation for four different patients. Figure 3 shows an example of the STAPLE-based probabilistic estimate of the true segmentation from multiple expert inputs in all three planes. A summary of volumetric overlap for respiratory phase-based GTV and IGTV auto-segmentation for all ten patients is shown in table 2. GTV auto-segmentation was subject to varying accuracy across the patient group with respect to volumetric overlap, ranging from (81.51 ± 1.92) to (97.27 ± 0.28)% agreement, and a median value of 92.16% across all phases and patients. IGTV auto-segmentation accuracy was significantly improved with reduced variance for all ten patients ranging from 90.87 to 98.57% volumetric overlap and a median value of 95.68% across all phases and patients. Histograms of all GTV ($N = 900$) and IGTV ($N = 100$) overlap measurements are shown in figure 4.

Overall, these measures showed consistency in the relative location accuracy of auto-segmented GTV and IGTV structures. Symmetric RMS surface distances are summarized in table 3. These distance-based measurements displayed a similar trend to volumetric overlap measurements with reduced error for IGTV segmentations. GTV surface-to-surface deviations ranged from (3.18 ± 0.05) to (4.73 ± 0.05) mm and a median value of 3.64 mm in surface-to-surface distance for respiratory phase-specific GTV segmentations across all phases and patients. IGTV measurements showed a reduced range of value with smaller variance, ranging from 2.68 to 4.21 mm, with a median value of 3.10 mm across all phases and patients. Histograms of GTV ($N = 900$) and IGTV ($N = 100$) surface distances are shown in figure 5. Statistically significant reduction in both surface-to-surface distances and volumetric overlap were observed for all patients, suggesting that the GTV envelope or IGTV segmentation reduces the uncertainty in 4D-CT based target volume segmentation. For GTV segmentation, 1-way ANOVA demonstrated that the choice of respiratory phase for the basis

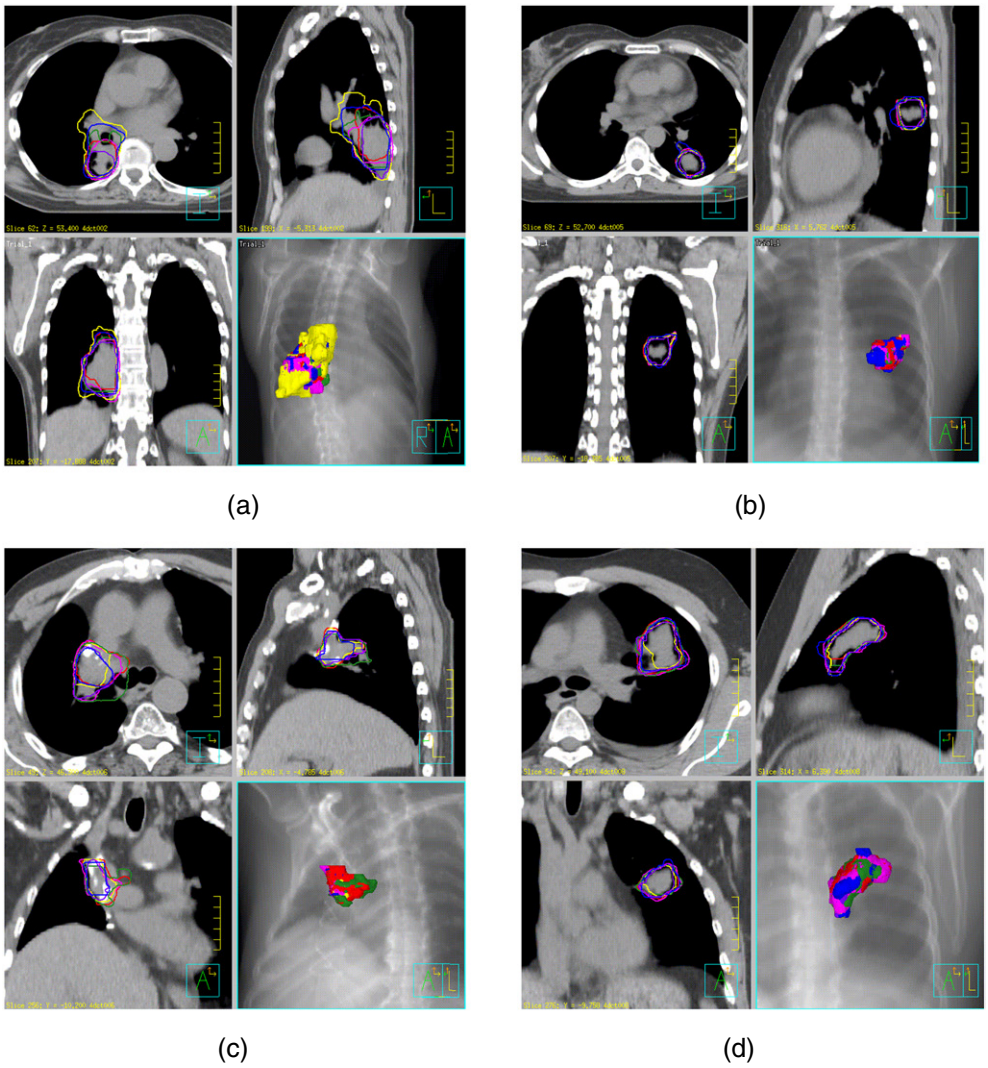

Figure 2. Representative slices in the axial, sagittal, and coronal planes and 3D view of 6 physicians manual IGTV segmentation for patients B, E, F, and H, in panels (*a*)–(*d*), respectively (from Pinnacle TPS 8.1y Beta).

of auto-segmentation provided for statistically significant differences in accuracy with respect to both volumetric overlap and RMS distance measures. Currently, no clinical interpretation of volumetric overlap indices exists making the metric difficult to analyze in the context of potential treatment efficacy and patient outcomes. However, table 2 shows that the difference in average volumetric overlap (Δ_{VO}) between phases was <5% for all patients except patient C. Therefore, while statistically different, these differences are quite small and are not indicative of optimality for any specific respiratory phase as a basis for auto-segmentation. While symmetric RMS distances also showed significant differences between respiratory phases, they also did not indicate optimality for any specific respiratory phase as a basis for auto-segmentation. This was due to differences in symmetric RMS measures being sub-millimeter and clinical margins in the context of radiotherapy are considered on the order of

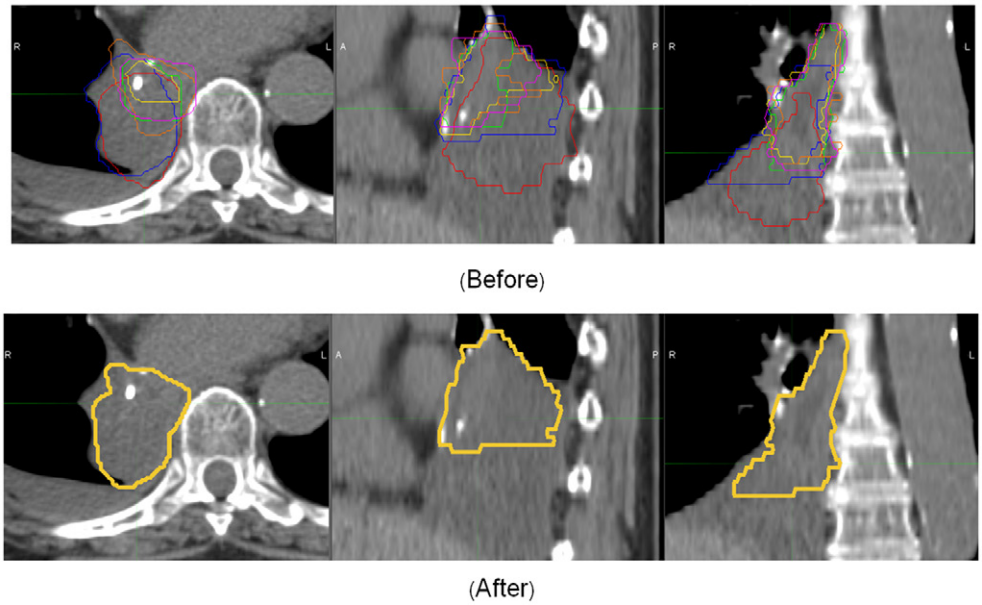

Figure 3. Example result of STAPLE-based probabilistic estimate of the true segmentation from multiple expert input segmentations in the axial, sagittal, and coronal image planes. Top panels represent multiple physicians' manual tumour volume segmentations with each colour representing a different observer. Bottom panels represent STAPLE algorithm GT estimate segmentation (from Pinnacle TPS 8.1y Beta).

millimeters. Therefore, while differences due to the choice of propagation phase were statistically significant, they were not clinically relevant. These two observations indicate that the choice of respiratory phase for the basis of auto-segmentation may be arbitrary, and certain phases such as end-exhale (50%) or end-inhale (0%) do not provide for better auto-segmentation results than any other individual phase when propagating tumour volume segmentation through the entire 4D-CT dataset. Table 4 shows Hausdorff distances for GTV and IGTV segmentations. Hausdorff distances were quite variable across the patient groups and showed no specific trends. GTV and IGTV segmentations showed no statistically different results, and increased Hausdorff distances were observed in cases with increased manual segmentation difficulty as graded by physicians. Based upon 3D surface reconstruction, the largest Hausdorff distances were typically due to segmentation variability confined to the base and/or apex of the target volume (see figure 6). ANOVA between respiratory phases again showed significant differences in segmentation. However, for this metric, deviations were on the order of millimeters meaning they were of clinical relevance. No particular respiratory phase showed a propensity for lower Hausdorff distance calculations, and its value in this study is somewhat limited by the lack of statistical significance or visible trend in the data. However, this metric could be of value in a clinical setting by allowing physicians to view regions of largest disagreement between two segmentations. In the context of high dose fractionation for radiotherapy techniques such as stereotactic ablative radiotherapy (SABR) (McGarry *et al* 2005) and radio-surgery (Whyte *et al* 2003), error on the scale of millimeters and centimetres must be avoided to ensure healthy tissue and organs at risk are spared adequately. Figures 6 and 7 show surfaced rendered manually derived STAPLE IGTV segmentations with surface meshes of end-exhalation/inhalation based auto-segmented IGTV overlaid for difficult and easy cases,

Table 2. Volumetric overlap measurements from each respiratory phase derived auto-segmentation. Average values are shown for GTV structures, single values shown for IGTV structures.

Propagation phase (%)	Patient A		Patient B		Patient C		Patient D		Patient E	
	GTV ± SD (%)	IGTV (%)	GTV ± SD (%)	IGTV (%)	GTV ± SD (%)	IGTV (%)	GTV ± SD (%)	IGTV (%)	GTV ± SD (%)	IGTV (%)
0.0	96.58 ± 0.61	98.10	92.49 ± 0.87	95.40	87.57 ± 1.48	91.45	94.20 ± 0.43	96.43	92.65 ± 1.12	94.80
10.0	96.50 ± 0.42	97.96	92.34 ± 0.50	95.91	86.90 ± 0.97	92.59	94.12 ± 0.42	96.99	90.19 ± 0.97	94.24
20.0	97.19 ± 0.30	98.14	93.57 ± 0.78	96.09	86.24 ± 0.71	91.18	93.19 ± 0.35	96.17	91.79 ± 1.28	94.51
30.0	97.04 ± 0.14	98.10	93.94 ± 0.50	95.89	87.95 ± 0.88	92.78	93.92 ± 0.30	96.75	92.48 ± 1.07	95.49
40.0	97.07 ± 0.42	97.99	93.17 ± 0.82	95.73	89.19 ± 2.68	91.91	95.16 ± 0.39	97.00	90.23 ± 1.83	95.04
50.0	95.07 ± 0.67	98.57	92.71 ± 1.20	96.23	89.54 ± 0.86	92.42	93.01 ± 0.52	97.23	90.90 ± 1.38	95.14
60.0	96.25 ± 0.15	97.88	93.81 ± 0.89	96.07	87.98 ± 1.92	90.87	95.17 ± 0.25	96.45	91.32 ± 1.26	94.32
70.0	96.90 ± 0.24	97.82	93.06 ± 0.71	96.18	81.51 ± 1.93	92.14	94.83 ± 0.66	96.37	92.48 ± 1.01	94.07
80.0	97.21 ± 0.28	98.03	92.54 ± 0.82	96.54	86.58 ± 3.03	91.60	93.88 ± 0.68	96.46	90.80 ± 1.72	94.30
90.0	97.27 ± 0.28	98.26	93.59 ± 0.60	96.05	90.10 ± 1.47	92.01	94.95 ± 0.29	96.96	91.93 ± 1.01	94.46
p(within group)	$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$	
p(between groups)		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$
Propagation phase (%)	Patient F		Patient G		Patient H		Patient I		Patient J	
	GTV ± SD (%)	IGTV (%)	GTV ± SD (%)	IGTV (%)	GTV ± SD (%)	IGTV (%)	GTV ± SD (%)	IGTV (%)	GTV ± SD (%)	IGTV (%)
0.0	89.21 ± 1.15	92.26	88.30 ± 2.06	96.03	89.93 ± 0.62	95.85	92.11 ± 0.75	95.86	96.31 ± 0.40	96.42
10.0	89.43 ± 1.05	91.41	88.27 ± 2.75	95.22	87.85 ± 0.48	94.94	93.28 ± 1.32	95.94	95.12 ± 0.31	97.07
20.0	88.71 ± 0.96	93.50	90.10 ± 1.24	93.69	87.89 ± 1.00	94.89	93.57 ± 1.18	96.21	96.46 ± 0.37	96.80
30.0	87.97 ± 1.10	91.40	89.18 ± 1.55	92.21	89.05 ± 0.68	95.41	92.84 ± 0.71	95.56	95.94 ± 0.27	97.25
40.0	88.44 ± 1.25	93.29	89.96 ± 1.01	93.58	88.97 ± 0.80	95.34	94.08 ± 0.77	96.28	96.07 ± 0.51	96.58
50.0	87.25 ± 1.28	91.23	88.99 ± 2.18	95.05	89.57 ± 0.75	95.63	95.36 ± 0.58	95.59	95.27 ± 0.40	96.80
60.0	88.45 ± 1.24	91.58	89.30 ± 1.37	92.68	89.12 ± 0.67	95.00	91.92 ± 1.08	96.17	95.89 ± 0.34	96.40
70.0	89.49 ± 2.06	92.84	90.74 ± 0.93	94.96	89.18 ± 0.80	95.76	93.63 ± 0.98	96.33	96.52 ± 0.58	96.92
80.0	90.74 ± 1.57	92.49	89.87 ± 1.42	94.52	90.58 ± 0.79	94.85	93.91 ± 1.00	95.91	95.52 ± 0.19	96.59
90.0	88.20 ± 0.98	91.53	89.02 ± 2.96	94.83	89.80 ± 0.71	95.57	91.91 ± 1.50	96.05	96.02 ± 0.66	96.88
p(within group)	$p < 0.0001$		0.1237		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$	
p(between groups)		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$

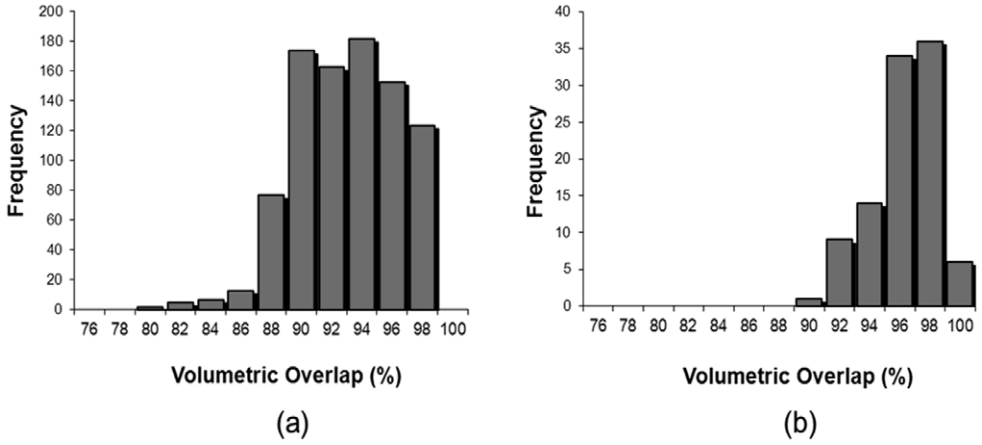

Figure 4. Histograms of volumetric overlap measurements after auto-segmentation. GTV and IGTV volumetric measurements are shown in (*a*) and (*b*), respectively.

respectively. Qualitative analysis shows that the auto-segmentation has a propensity to over-estimate target volumes. However, the geometry of auto-segmented volumes remains highly consistent with the reference STAPLE segmentations demonstrating that the deformable model-based technique is capable of robust segmentation in the presence of both simple and complex tumour volume geometries under the influence of respiratory motion.

4. Discussion

In this study, we have included consensus-based, multi-expert GT estimate segmentation with a deformable model-based automatic segmentation technique to establish a novel framework for anatomical delineation of lung tumours in 4D-CT image datasets for radiotherapy. The need for more advanced segmentation strategies increases in using 4D-CT, as respiratory motion influences on thoracic target volumes must be compensated for across larger amounts of 3D image information. This entails multiple segmentations of the GTV across anywhere from 8 to 12 respiratory phase specific CT volumes and generation of the enveloping IGTV. However, while necessary for treatment efficacy, this is an impractical task for any clinical workflow. The framework established in this study incorporates the information of the entire 4D-CT dataset while largely mitigating inter- and intra-observer variability, one of the largest contributors to segmentation related geometric uncertainty currently experienced in image-guided radiotherapy, by probabilistically estimating the GT via multiple expert segmentations. The proposed methodology highlights the strengths and possibilities for auto-segmentation strategies in 4D-CT incorporating GT estimate algorithms while systematically assessing the accuracy and mitigation of inter- and intra-observer variability. Areas for improvement still exist. The implementation of automated or assisted segmentation techniques within individual respiratory phase datasets of the 4D-CT may further reduce the time needed to segment the target volume. For example, multi-expert manual segmentation of base/apex slices and subsequent intra-respiratory phase contour propagation may provide increased timesaving and more accurate STAPLE segmentation calculation. Zhang *et al* have applied this type of proposed methodology previously in kidney segmentation (Zhang *et al* 2013).

A limitation in this study is the small number of patients and observers. While only ten patients and six observers were utilized from a single institution, this provided for approximately

Table 3. Surface distance measurements from each respiratory phase derived auto-segmentation. Average values are shown for GTV structures, single values shown for IGTV structures.

Propagation phase	Patient A		Patient B		Patient C		Patient D		Patient E	
	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)
0.0	3.67 ± 0.12	2.91	4.17 ± 0.17	3.57	4.22 ± 0.26	3.33	3.50 ± 0.10	2.96	3.23 ± 0.17	2.77
10.0	3.80 ± 0.11	2.86	4.08 ± 0.22	3.49	4.17 ± 0.32	4.21	3.54 ± 0.19	2.94	3.58 ± 0.17	2.85
20.0	3.75 ± 0.10	3.05	3.98 ± 0.26	3.72	4.73 ± 0.28	4.18	3.62 ± 0.17	2.91	3.31 ± 0.13	2.68
30.0	3.52 ± 0.08	2.87	3.94 ± 0.26	4.49	4.05 ± 0.27	3.69	3.46 ± 0.07	2.86	3.43 ± 0.23	2.86
40.0	3.67 ± 0.16	3.11	3.96 ± 0.22	3.95	4.26 ± 0.37	3.06	3.46 ± 0.14	2.82	3.41 ± 0.18	3.00
50.0	3.65 ± 0.12	2.88	4.17 ± 0.25	3.66	4.32 ± 0.14	3.52	3.71 ± 0.13	3.04	3.38 ± 0.27	2.95
60.0	3.85 ± 0.17	3.07	4.21 ± 0.24	3.88	4.28 ± 0.40	3.54	3.57 ± 0.21	3.02	3.55 ± 0.34	3.18
70.0	3.72 ± 0.16	3.05	4.16 ± 0.26	3.71	3.91 ± 0.30	3.61	3.56 ± 0.21	3.29	3.49 ± 0.28	3.19
80.0	3.87 ± 0.15	3.00	4.08 ± 0.20	3.39	3.95 ± 0.29	3.95	3.33 ± 0.08	2.91	3.39 ± 0.13	2.86
90.0	3.65 ± 0.14	3.04	4.31 ± 0.23	3.56	3.84 ± 0.35	3.54	3.34 ± 0.14	2.90	3.43 ± 0.26	2.92
p(within group)	$p < 0.0001$		0.0159		$p < 0.0001$		$p < 0.0001$		0.0552	
p(between groups)		$p < 0.0001$		$p < 0.0001$		0.0013		$p < 0.0001$		$p < 0.0001$
Propagation phase	Patient F		Patient G		Patient H		Patient I		Patient J	
	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)
0.0	3.62 ± 0.13	3.04	3.46 ± 0.57	3.18	3.42 ± 0.08	3.29	3.65 ± 0.23	3.06	3.18 ± 0.05	2.78
10.0	3.60 ± 0.09	3.52	3.46 ± 0.49	3.09	3.68 ± 0.16	3.07	3.71 ± 0.25	3.45	3.46 ± 0.14	2.80
20.0	3.92 ± 0.15	3.48	3.65 ± 0.15	3.48	3.74 ± 0.17	3.14	3.45 ± 0.21	3.10	3.52 ± 0.10	2.88
30.0	3.77 ± 0.12	3.60	3.32 ± 0.47	3.51	3.57 ± 0.10	3.10	3.51 ± 0.21	3.07	3.29 ± 0.08	2.83
40.0	3.68 ± 0.09	3.21	3.51 ± 0.58	3.07	3.63 ± 0.15	3.16	3.62 ± 0.26	3.16	3.39 ± 0.08	2.71
50.0	3.64 ± 0.22	3.42	3.63 ± 0.46	3.33	3.49 ± 0.15	3.28	3.38 ± 0.30	3.19	3.35 ± 0.11	2.87
60.0	3.71 ± 0.13	3.73	3.76 ± 0.49	3.33	3.87 ± 0.22	3.02	3.65 ± 0.31	2.89	3.50 ± 0.15	2.70
70.0	3.62 ± 0.22	3.58	3.61 ± 0.46	2.97	3.70 ± 0.22	3.02	3.62 ± 0.20	3.09	3.50 ± 0.14	2.85
80.0	3.60 ± 0.21	3.42	3.61 ± 0.30	3.32	3.61 ± 0.17	3.21	3.45 ± 0.15	3.08	3.40 ± 0.09	2.79
90.0	3.84 ± 0.17	3.32	3.40 ± 0.45	3.57	3.39 ± 0.09	3.18	3.67 ± 0.22	3.29	3.27 ± 0.10	2.70
p(within group)	0.0070		0.76704		$p < 0.0001$		0.0605		$p < 0.0001$	
p(between groups)		0.002 112		0.0037		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$

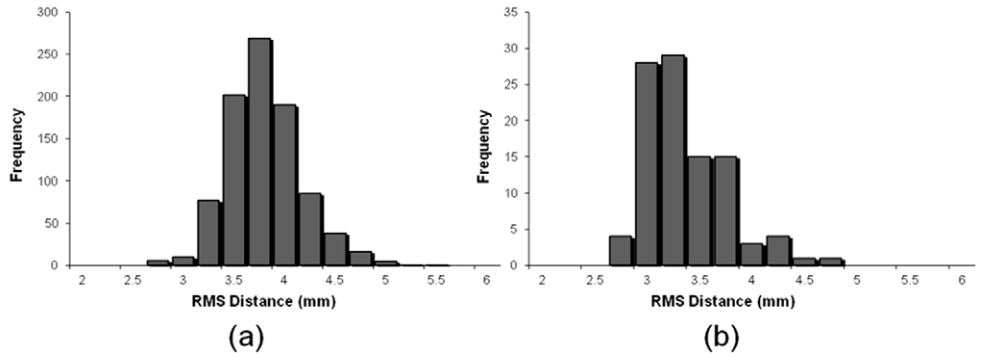

Figure 5. Histograms of symmetric RMS surface distance measurements after auto-segmentation. GTV and IGTV surface measurements are shown in (*a*) and (*b*), respectively.

six hundred manually defined primary and nodal GTV segmentations. While this yielded a considerable amount of data, it isn't enough to support clinical adoption of the technique at this time. Potential future pilot studies involving a larger number of patients would be required to outline the logistics of the proposed segmentation workflow in clinic while determining the exact amount of manual input that is required to implement the proposed methodology and precisely quantifying the time-savings per patient compared to a fully manual workflow. Future work focusing on clinical implementation of the technique proposed in this study and its impact on radiotherapy treatment plan dosimetry would require a significantly larger number of patients and observers from multiple institutions to ensure a sound clinical workflow is developed to account for any and all forms of variability. Subsequent clinical use of the methodology proposed in this study would require further streamlining of manual segmentation as well as collaboration amongst disease-site physician groups within the clinic in addition to subsequent physician review and necessary correction of automated segmentations, in particular for high risk ROIs, areas of increased uncertainly and geometrically complex tumour volumes that may be subject to irregular respiratory motion. Supplementary work analyzing the dosimetric impact of this segmentation technique is also required in the context of target volume segmentation to assess local control, and critical structures to assess dose to healthy tissue and OARs.

The deformable model-based technique proved capable of providing robust segmentation irrespective of 4D-CT image quality, tumour volume geometry, and the choice of respiratory phase CT as the basis for auto-segmentation. The results of this study seem to indicate that the choice of reference respiratory phase has little to no effect on the accuracy of auto-segmentation although it may be possible that the accuracy of the STAPLE algorithm's calculation of the consensus segmentation itself may depend on the respiratory phase of the 4D-CT dataset, and this could be an area for future study. Accuracy was observed to be more so dependent on physical characteristics of the tumour volume, as complex volumes (i.e. patient B and C) were subject to larger error in deformation and auto-segmentation due to factors such as size and proximity to the lung periphery. These observations correlate with those made by Erhardt *et al* who noted that larger, adherent volumes were also subject to increased predicted tumour position error (Ehrhardt *et al* 2011). Wu *et al* commented on the impact of these errors in the clinical setting when treating lung cancer patients with IGRT based on 4D-CT images. They also noted that the clinical margin to compensate for tumour motion in 4D-CT is typically ~5 mm

Table 4. Hausdorff distance measurements from each respiratory phase derived auto-segmentation. Average values are shown for GTV structures, single values shown for IGTV structures.

Propagation phase	Patient A		Patient B		Patient C		Patient D		Patient E	
	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)
0.0	5.99 ± 0.50	7.11	9.07 ± 2.48	10.47	7.52 ± 1.39	15.86	6.42 ± 1.28	4.23	3.68 ± 0.23	6.31
10.0	6.71 ± 0.60	6.84	9.38 ± 0.84	8.97	7.13 ± 1.50	15.08	5.80 ± 1.46	4.69	6.94 ± 1.87	4.88
20.0	7.21 ± 2.13	5.81	7.55 ± 1.72	6.60	13.87 ± 0.99	8.43	6.27 ± 1.00	3.52	4.70 ± 1.31	5.02
30.0	6.03 ± 0.87	6.55	7.32 ± 1.84	7.60	7.58 ± 1.94	11.15	5.63 ± 1.03	4.62	5.16 ± 0.92	4.14
40.0	9.51 ± 3.09	5.73	8.19 ± 1.56	13.97	13.03 ± 3.00	7.74	6.25 ± 1.00	4.47	6.56 ± 1.01	4.08
50.0	5.53 ± 0.58	5.26	10.01 ± 2.47	6.52	8.98 ± 1.66	14.52	5.27 ± 0.75	4.23	4.62 ± 1.15	3.58
60.0	7.38 ± 2.89	7.04	8.72 ± 1.59	9.18	8.16 ± 2.75	7.81	5.95 ± 1.02	5.86	4.81 ± 1.29	3.71
70.0	7.08 ± 2.14	7.44	8.81 ± 1.98	11.07	6.32 ± 1.56	10.20	5.16 ± 0.77	7.65	4.63 ± 1.64	4.19
80.0	7.81 ± 1.30	7.44	8.99 ± 2.69	10.26	7.00 ± 1.30	13.98	5.30 ± 1.41	5.18	5.82 ± 0.88	6.58
90.0	5.83 ± 0.59	6.55	11.54 ± 1.05	7.83	6.40 ± 1.85	11.53	5.21 ± 0.94	4.83	5.91 ± 1.31	4.93
p(within group)	0.0002		0.0008		$p < 0.0001$		0.0883		$p < 0.0001$	
p(between groups)		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$
Propagation phase	Patient F		Patient G		Patient H		Patient I		Patient J	
	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)	GTV ± SD (mm)	IGTV (mm)
0.0	5.07 ± 0.47	9.56	3.79 ± 0.67	3.52	4.97 ± 0.84	6.36	6.10 ± 1.04	6.26	4.96 ± 1.25	3.71
10.0	5.55 ± 0.92	5.26	3.30 ± 0.77	4.46	4.61 ± 0.50	5.03	6.08 ± 1.76	7.05	6.47 ± 1.3	5.73
20.0	6.68 ± 1.11	5.81	3.83 ± 0.39	3.96	4.79 ± 0.56	5.61	5.85 ± 1.23	4.10	5.70 ± 0.76	5.86
30.0	5.74 ± 1.36	5.30	4.22 ± 1.32	4.07	5.18 ± 0.69	5.40	5.29 ± 1.13	6.94	4.80 ± 0.54	5.26
40.0	5.67 ± 1.17	6.08	4.07 ± 1.12	4.17	7.32 ± 1.21	5.84	6.94 ± 1.31	5.84	4.88 ± 0.55	6.05
50.0	6.05 ± 1.40	8.39	3.86 ± 0.75	4.10	4.91 ± 0.77	6.74	5.84 ± 1.52	6.35	5.58 ± 0.85	6.29
60.0	5.77 ± 1.49	6.68	4.33 ± 1.01	3.27	5.27 ± 0.66	5.51	6.67 ± 1.07	4.01	5.55 ± 0.83	5.73
70.0	5.59 ± 1.30	6.55	3.48 ± 0.72	4.82	4.93 ± 0.71	5.84	5.40 ± 1.39	6.97	5.06 ± 0.98	6.44
80.0	5.84 ± 1.15	5.94	4.40 ± 0.73	4.82	5.42 ± 0.50	5.03	5.33 ± 1.09	7.53	5.25 ± 1.14	6.18
90.0	6.58 ± 1.08	4.88	3.55 ± 0.67	4.74	4.30 ± 0.46	6.36	6.08 ± 1.68	4.72	4.82 ± 1.14	4.08
p(within group)	0.1644		0.0965		$p < 0.0001$		0.1695		0.0095	
p(between groups)		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$		$p < 0.0001$

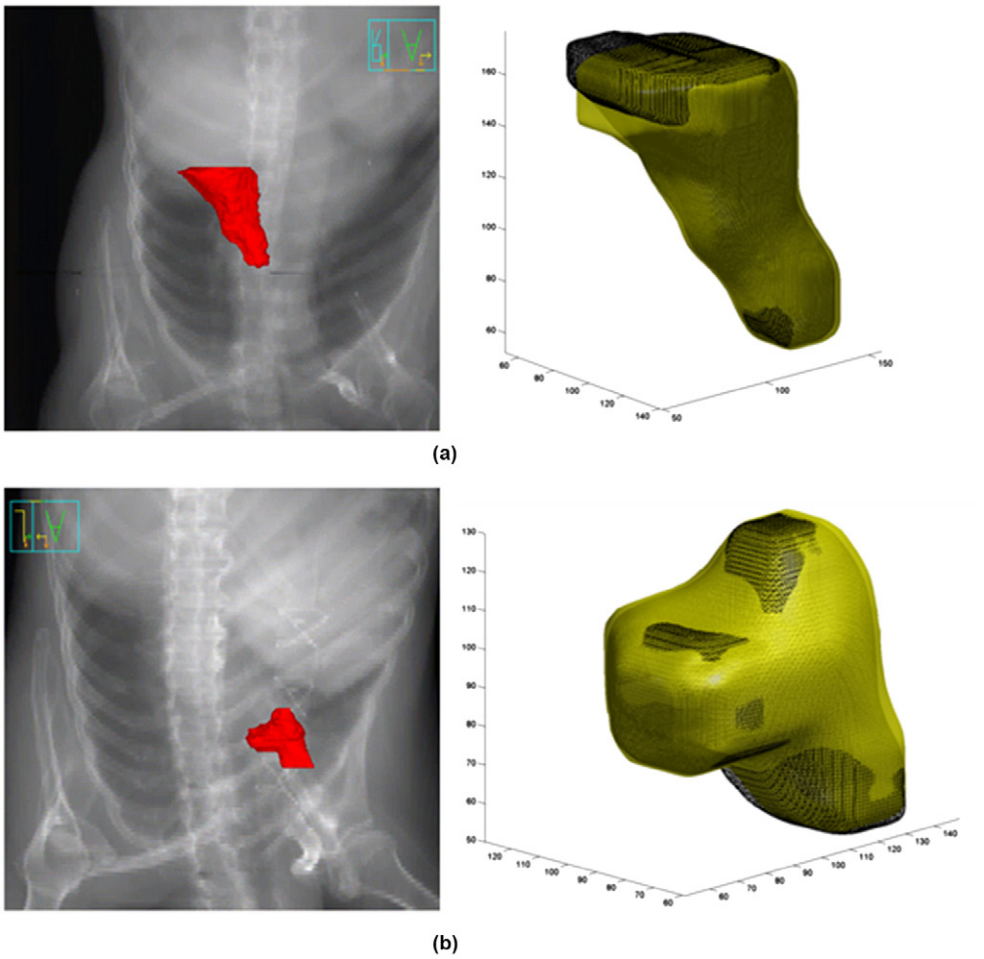

Figure 6. CT images depicting STAPLE-based IGTVs (red) (left panels), and surface renderings of IGTV auto-segmentation based on end-inhalation phase (0%) (black mesh) and end-exhalation phase (50%) (transparent yellow) (right panels) for difficult cases such as patient C (*a*) and patient F (*b*).

added to the CTV with an additional 7 mm margin added for the ITV for a total margin size of 12 mm (Wu *et al* 2013). As distance-based errors were comparable between the two studies, the results of this study seem to support the observation made by Wu *et al* that dose margins could be reduced in order to accommodate more precise treatment of lung tumours using 4D-CT. Although the deformable model-based method used in this study was strictly implemented for the purpose of auto-segmentation, future work could implement this algorithm to analyze both the changes in amplitude and deformation of the tumour caused by respiratory motion.

Image registration and deformable model-based techniques utilizing grey-scale transitions has been shown to be of limited use in the presence of contrast enhancement as images of dissimilar intensities and/or variable contrast enhancement provide for improper displacement estimates (Varadhan *et al* 2013). At the same time, the literature on functional imaging, contrast enhancement, and/or hybrid imaging in segmentation of lung tumours has increased

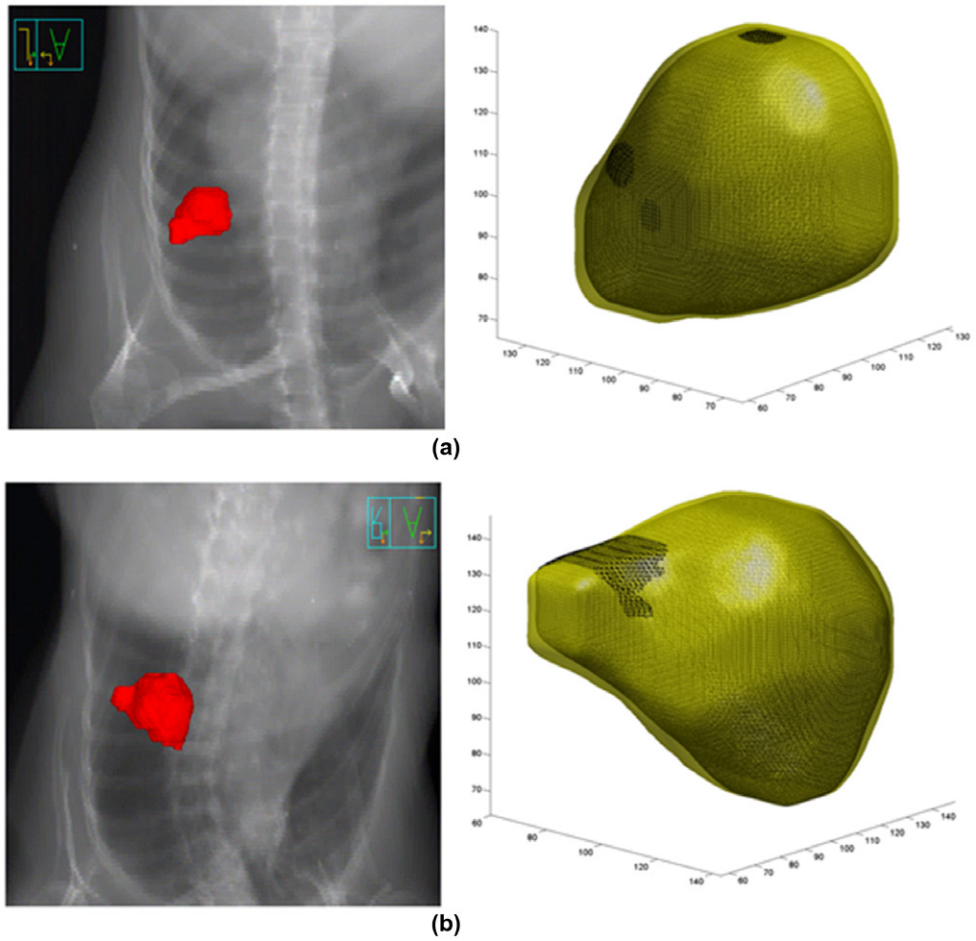

Figure 7. CT images depicting STAPLE-based IGTVs (red) (left panels), and surface renderings of IGTV auto-segmentation based on end-inhalation phase (0%)(black mesh) and end-exhalation phase (50%) (transparent yellow) (right panels) for easy cases such as patient E (*a*) and patient I (*b*).

in recent years. The use of contrast and/or additional image modalities to help differentiate tumours from normal structures and healthy tissue is well documented in efforts to improve both segmentation accuracy and subsequent radiotherapy treatment efficacy (Caldwell *et al* 2001, Steenbakkers *et al* 2006, van Baardwijk *et al* 2007, Apostolova *et al* 2010). The use of such imaging techniques has shown to decrease inter- and intra-observer variability while also mitigating potential treatment dosimetry errors, contrast enhancement and/or multimodal imaging is not standard for clinical treatment of lung cancer and no current clinical guidelines exist concerning the absolute/gradient thresholds in manual segmentation. Additionally, difficulties in defining edges and borders in PET/CT image volumes is also an issue in manual segmentation, as observed by MacManus *et al* (MacManus *et al* 2001). While the Radiation Therapy Oncology Group (RTOG) is currently conducting a clinical trial (0515) focused on determining the impact of PET/CT fusion in GTV segmentation for NSCLC carcinoma patients, these issues currently remain unresolved and lacking in general consensus (Bradley

et al 2012). Thus, even as functional becomes more prevalent as an ancillary tool for segmentation of lung tumours and target volume definition, the need for segmentation techniques operating solely in the context of 4D-CT imaging remains high.

While manual segmentation provides arguably the most accurate basis for any auto-segmentation method, the ability to incorporate multiple expert segmentations is highly desirable to maximize the utility of available resources within the clinic. Further improvement is still possible, and studies must be performed to assess the accuracy of any proposed technique as any time saved in segmenting tumour volumes for IGRT is largely negated if accuracy is compromised in doing so. While some general consensus exists on what metrics should be calculated, very little agreement exists on the best way to calculate them. Although the CPD algorithm and a novel surface reconstruction technique based on global optimization were utilized in this study, the authors concede that different datasets, anatomies, and imaging acquisition techniques may require different methodologies and tools for measuring segmentation accuracy and quality. As such, it is important to review the literature to ensure the most effective and accurate methodology for future studies. The development of auto-segmentation techniques in the context of contrast enhanced, standalone 4D-CT, and 3D-CT image datasets is of great interest for future work in segmentation of not only lung tumours, but also other disease sites and different anatomies.

5. Conclusions

Our findings suggest that the proposed framework for STAPLE-based lung tumour segmentation in 4D-CT images is accurate when compared to manually derived STAPLE segmentations, especially for IGTV segmentation. Additionally, lung tumour volume segmentation is subject to significantly less inter-/intra-observer variability compared to the manual method. Thus, the reduced variability of our proposed segmentation strategy compared to manual techniques suggests that it can be used to infer accurate lung tumour locations within 4D-CT datasets across a diverse patient group indicating sensitivity to small or large changes in lung tumour dimensions and volume. Further improvements have been presented to reduce manual segmentation workload which could make our suggested framework suitable for translation into the clinical independent of ancillary imaging information.

Acknowledgments

The authors would like to thank Jeff Kempe for his help in software analysis and implementation. This work was funded by the Ontario Institute for Cancer Research (OICR) through funding provided by the government of Ontario and by the Canadian Institute of Health Research (CIHR), CIHR Strategic Training Initiative in Cancer Research.

Conflicts of interest

None to declare

References

Apostolova I *et al* 2010 Combined correction of recovery effect and motion blur for SUV quantification of solitary pulmonary nodules in FDG PET/CT *Eur. Radiol.* **20** 1868–77

Barnes E A, Murray B R, Robinson D M, Underwood L J, Hanson J and Roa W H 2001 Dosimetric evaluation of lung tumour immobilization using breath hold at deep inspiration *Int. J. Radiat. Oncol. Biol. Phys.* **50** 1091–8

Biancardi A M, Jirapatnakul A C and Reeves A P 2010 A comparison of ground truth estimation methods *Int. J. Comput. Assist. Radiol. Surg.* **5** 295–305

Boykov Y and Kolmogorov V 2004 An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 1124–37

Bradley J *et al* 2012 A phase II comparative study of gross tumour volume definition with or without PET/CT fusion in dosimetric planning for non-small-cell lung cancer (NSCLC): Primary Analysis of Radiation Therapy Oncology Group (RTOG) 0515 *Int. J. Radiat. Oncol. Biol. Phys.* **82** 435–41

Caldwell C B, Mah K, Ung Y C, Danjoux C E, Balogh J M, Ganguli S N and Ehrlich L E 2001 Observer variation in contouring gross tumour volume in patients with poorly defined non-small-cell lung tumours on CT: the impact of 18FDG-hybrid PET fusion *Int. J. Radiat. Oncol. Biol. Phys.* **51** 923–31

Chalana V and Kim Y 1997 A methodology for evaluation of boundary detection algorithms on medical images *IEEE Trans. Med. Imaging* **16** 642–52

Commowick O and Warfield S K 2010 Estimation of inferential uncertainty in assessing expert segmentation performance from STAPLE *IEEE Trans. Med. Imaging* **29** 771–80

Ehler E D and Tomé W A 2008 Lung 4D-IMRT treatment planning: an evaluation of three methods applied to 4D data sets *Radiother. Oncol.* **88** 319–25

Ehrhardt J, Werner R, Schmidt-Richberg A and Handels H 2011 Statistical modeling of 4D respiratory lung motion using diffeomorphic image registration *IEEE Trans. Med. Imaging* **30** 251–65

Ekberg L, Holmberg O, Wittgren L, Bjelkengren G and Landberg T 1998 What margins should be added to the clinical target volume in radiotherapy treatment planning for lung cancer? *Radiother. Oncol.* **48** 71–7

Erridge S C, Seppenwoolde Y, Muller S H, Van Herk M and De Jaeger K 2003 Portal imaging to assess set-up errors, tumour motion and tumour shrinkage during conformal radiotherapy of non-small cell lung cancer *Radiother. Oncol.* **66** 75–85

Ezhil M, Vedam S, Balter P, Choi B, Mirkovic D, Starkschall G and Chang J Y 2009 Determination of patient-specific internal gross tumour volumes for lung cancer using 4D computed tomography *Radiat. Oncol.* **4** 4

Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R and Georg D 2012 Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy *Strahlenther. Onkol.* **188** 160–7

Gaede S, Olsthoorn J, Louie A V, Palma D, Yu E, Yaremko B, Ahmad B, Chen J, Bzdusek K and Rodrigues G 2011 An evaluation of an automated 4D-CT contour propagation tool to define an internal gross tumour volume for lung cancer radiotherapy *Radiother. Oncol.* **101** 322–8

Giraud P *et al* 2002 Conformal radiotherapy for lung cancer: different delineation of the gross tumour volume (GTV) by radiologists and radiation oncologists *Radiother. Oncol.* **62** 27–36

Gordon S, Lotenberg S, Long R, Antani S, Jeronimo J and Greenspan H 2009 Evaluation of uterine cervix segmentations using ground truth from multiple experts *Comput. Med. Imaging Graph.* **33** 205–16

Heimann T *et al* 2009 Comparison and evaluation of methods for liver segmentation from CT datasets *IEEE Trans. Med. Imaging* **28** 1251–65

Huang L, Park K, Boike T, Lee P, Papiez L, Solberg T, Ding C and Timmerman R D 2010 A study on the dosimetric accuracy of treatment planning for stereotactic body radiation therapy of lung cancer using average and maximum intensity projection images *Radiother. Oncol.* **96** 48–54

Huttenlocher D P, Klanderman G A and Rucklidge W J 1993 Comparing images using the hausdorff distance *IEEE Trans. Pattern Anal. Mach. Intell.* **15** 850–63

Jameson M G, Holloway L C, Vial P J, Vinod S K and Metcalfe P E 2010 A review of methods of analysis in contouring studies for radiation oncology *J. Med. Imaging Radiat. Oncol.* **54** 401–10

Josipovic M, Persson G F, Håkansson K, Damkjær S M S, Bangsgaard J P, Westman G, Riisgaard S, Specht L and Aznar M C 2013 Deep inspiration breath hold radiotherapy for locally advanced lung cancer: comparison of different treatment techniques on target coverage, lung dose and treatment delivery time *Acta Oncol.* **52** 1582–86

Kaus M R, Pekar V, Lorenz C, Truyen R, Lobregt S, Richolt J and Weese J 2001 Automated 3D PDM construction using deformable models *Computer Vision, 2001. ICCV 2001. Proc. 8th IEEE Int. Conf.* (*Vancouver, BC*) vol 1 pp 566–72

Kaus M R, Netsch T, Kabus S, Pekar V and Mcnutt T 2004 Estimation of organ motion from 4D CT for 4D radiation therapy planning of lung cancer *Med. Image Comput. Comput.-Assist. Intervent.–MICCAI 2004* vol 3217 (Berlin: Springer) pp 1017–24

Lempitsky V 2010 Surface extraction from binary volumes with higher-order smoothness *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (*San Francisco, CA, 13–18 June 2010*) pp 1197–04

Lempitsky V and Boykov Y 2007 Global optimization for shape fitting *2007 IEEE Conf. on Computer Vision and Pattern Recognition* (*Minneapolis, June 2007*) pp 1–8

Low D A *et al* 2003 A method for the reconstruction of 4D synchronized CT scans acquired during free breathing *Med. Phys.* **30** 1254

MacManus M P, Hicks R J, Matthews J P, Hogg A, McKenzie A F, Wirth A, Ware R E and Ball D L 2001 High rate of detection of unsuspected distant metastases by pet in apparent stage III non-small-cell lung cancer: implications for radical radiation therapy *Int. J. Radiat. Oncol. Biol. Phys.* **50** 287–93

McGarry R C, Papiez L, Williams M, Whitford T and Timmerman R D 2005 Stereotactic body radiation therapy of early-stage non-small-cell lung carcinoma: phase I study *Int. J. Radiat. Oncol. Biol. Phys.* **63** 1010–5

Myronenko A and Song X 2010 Point-set registration: coherent point drift *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 2262–75

Pan T, Lee T-Y, Rietzel E and Chen G T Y 2004 4D-CT Imaging of a volume influenced by respiratory motion on multi-slice CT *Med. Phys.* **31** 333

Persson G F, Nygaard D E, Brink C, Jahn J W, Munck Af Rosenschöld P, Specht L and Korreman S S 2010 Deviations in delineated GTV caused by artefacts in 4D-CT *Radiother. Oncol.* **96** 61–6

Persson G F, Nygaard D E, Munck Af Rosenschöld P, Vogelius I R, Josipovic M, Specht L and Korreman S S 2011 Artifacts in conventional computed tomography (CT) and free breathing 4D ct induce uncertainty in gross tumour volume determination *Int. J. Radiat. Oncol. Biol. Phys.* **80** 1573–80

Pevsner A *et al* 2006 Evaluation of an automated deformable image matching method for quantifying lung motion in respiration-correlated CT images *Med. Phys.* **33** 369

Plathow C, Ley S, Fink C, Puderbach M, Hosch W, Schmähl A, Debus J and Kauczor H-U 2004 Analysis of intrathoracic tumour mobility during whole breathing cycle by dynamic MRI *Int. J. Radiat. Oncol. Biol. Phys.* **59** 952–9

Ragan D, Starkschall G, McNutt T, Kaus M, Guerrero T and Stevens C W 2005 Semiautomated 4D computed tomography segmentation using deformable models *Med. Phys.* **32** 2254

Speight R, Sykes J, Lindsay R, Franks K and Thwaites D 2011 The evaluation of a deformable image registration segmentation technique for semi-automating internal target volume (ITV) production from 4D-CT images of lung stereotactic body radiotherapy (SBRT) patients *Radiother. Oncol.* **98** 277–83

Steenbakkers R J H M *et al* 2006 Reduction of observer variation using matched CT-PET for lung cancer delineation: a 3D analysis *Int. J. Radiat. Oncol. Biol. Phys.* **64** 435–48

van Baardwijk A *et al* 2007 PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumour and involved nodal volumes *Int. J. Radiat. Oncol. Biol. Phys.* **68** 771–8

van Dam I E, van Sörnsen de Koste J R, Hanna G G, Muirhead R, Slotman B J and Senan S 2010 Improving target delineation on 4D-CT scans in stage I NSCLC using a deformable registration tool *Radiother. Oncol.* **96** 67–72

Van de Steene J, Linthout N, de Mey J, Vinh-Hung V, Claassens C, Noppen M, Bel A and Storme G 2002 Definition of gross tumour volume in lung cancer: inter-observer variability *Radiother. Oncol.* **62** 37–49

Varadhan R, Karangelis G, Krishnan K and Hui S 2013 A framework for deformable image registration validation in radiotherapy clinical applications *J. Appl. Clin. Med. Phys.* **14** 4066

Vedam S S, Keall P J, Kini V R, Mostafavi H, Shukla H P and Mohan R 2003 Acquiring a 4D computed tomography dataset using an external respiratory signal *Phys. Med. Biol.* **48** 45–62

Warfield S K, Zou K H and Wells W M 2004 Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation *IEEE Trans. Med. Imaging* **23** 903–21

Whyte R I, Crownover R, Murphy M J, Martin D P, Rice T W, DeCamp M M, Rodebaugh R, Weinhous M S and Le Q-T 2003 Stereotactic radiosurgery for lung tumours: preliminary report of a phase I trial *Ann. Thorac. Surg.* **75** 1097–101

Wu G, Wang Q, Lian J and Shen D 2013 Estimating the 4D respiratory lung motion by spatiotemporal registration and super-resolution image reconstruction *Med. Phys.* **40** 031710

Zhang P, Liang Y, Chang S and Fan H 2013 Kidney segmentation in CT sequences using graph cuts based active contours model and contextual continuity *Med. Phys.* **40** 081905

Zhu G, Yang J, Lu M, Ajlouni M, Kim J H and Yin F-F 2008 The investigation on the location effect of external markers in respiratory-gated radiotherapy *J. Appl. Clin. Med. Phys.* **9** 2758