Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/rpor

Review

A review of automatic lung tumour segmentation in the era of 4DCT

Nadine Wong Yuzhen, Sarah Barrett[*]

Applied Radiation Therapy Trinity, Discipline of Radiation Therapy, Trinity College Dublin, Ireland

ARTICLE INFO

Article history:
Received 10 August 2018
Received in revised form
24 November 2018
Accepted 21 January 2019
Available online 22 February 2019

Keywords:
Lung cancer
Delineation
4DCT
Automatic segmentation

ABSTRACT

Aim: To review the literature on auto-contouring methods of lung tumour volumes on four-dimensional computed tomography (4DCT).

Background: Manual delineation of lung tumour on 4DCT has been the gold standard in clinical practice. However, it is resource intensive due to the high volume of data which results in longer contouring duration and uncertainties in defining target. Auto-contouring may present as an attractive alternative by decreasing manual inputs required, thus improving the contouring process. This review aims to assess the accuracy, variability and contouring duration of automatic contouring compared with manual contouring in lung cancer on 4DCT datasets.

Materials and methods: A search and review of literature were conducted to identify studies regarding lung tumour contouring on 4DCT. Manual and auto-contours were assessed and compared based on accuracy, variability and contouring duration.

Results: Thirteen studies were included in this review and their results were compared. Accuracy of auto-contours was found to be comparable to manual contours. Auto-contouring resulted in lesser inter-observer variation when compared to manual contouring, however there was no significant reduction in intra-observer variability. Additionally, contouring duration was reduced with auto-contouring although long computation time could present as a bottleneck.

Conclusion: Auto-contouring is reliable and efficient, producing accurate contours with better consistency compared to manual contours. However, manual inputs would still be required both before and after auto-propagation.

1. Background

Worldwide incidence and mortality of lung cancer are high.[1] Radiation therapy is commonly used to treat from early- to advanced-staged lung cancer either as an adjuvant treatment after surgery, concurrently with chemotherapy or for palliation of symptoms.[2] An important goal of radiotherapy planning phase is to correctly localise and contour the tumour. This ensures that the radiation is delivered to the correct target volume while sparing surrounding normal tissues.[3] Inaccuracies in the contouring stage will produce systematic errors which will result in a geographical miss of the target.[4] The impact of this error may be amplified with the use of conformal treatment modalities such as intensity modulated radiation therapy (IMRT). Hence, it is crucial to define tumour accurately to ensure optimal tumour control and to minimise toxicities.[5]

[*] *Corresponding author.*
E-mail address: barrets7@tcd.ie (S. Barrett).

https://doi.org/10.1016/j.rpor.2019.01.003

Three-dimensional computed tomography (3DCT) has been the main imaging modality in radiotherapy treatment planning. However, the presence of significant respiratory motion, and therefore tumour motion, in lung cases is not accounted for in 3DCT. Alternatively, four-dimensional computed tomography (4DCT) may be used for lung cancer treatment planning which has been shown to increase the accuracy of target delineation.[6] Radiation therapy centres using 4DCT for lung cancer patients have reported a reduction in image artefacts present in 3DCT.[3] In 4DCT, each breathing cycle is divided into ten phases and subsequently the processed images are binned into datasets for each phase. Gross tumour volume (GTV) in each dataset are contoured and combined to form the internal gross tumour volume (IGTV).[7] The use of 4DCT, however, has greatly decreased the efficiency of the work process due to the additional workload of delineation on 10 different phases.[8,9] Manual contouring on 4DCT is time-consuming and may be subjected to high levels of variability and inconsistency.[10] Maximum intensity projection (MIP) reconstruction is used clinically to reduce workload from manually contouring in all phases. The MIP dataset represents the full area of target motion and is used to derive the internal target volume (ITV).[11] However, the reliability of contouring on MIP is questionable, with studies suggesting that risks of over or underestimating of the lung tumour volume exist.[12,13] Other challenges with manual contouring on MIP includes motion artefacts on reconstructed 4DCT images, irregular tumour volumes and indistinguishable surrounding organs-at-risks (OARs) due to similar densities.[14] Ideally, the GTV should be delineated and optimised with reference to each individual phase of the breathing cycle.[15]

In order to save time and minimise errors while avoiding the disadvantages of delineating tumour from MIP, the proposed use of auto-contouring has been an area of active research. Automatic interventions have shown to mitigate problems with manual contouring. Auto-contouring ranges from being a simple to a multi-step process with the use of a number of complex algorithms. Deformable registration is frequently used in the auto-contouring process which involves the propagation of manually delineated contours on a single phase to the remaining phases. A wide range of techniques have been employed in studies, such as deformable template-based segmentation, automatic thresholding, iterative thresholding,[16] model-based deformable image registration algorithm,[17] multi-seed point segmentation,[18] principal surfaces with propagation of contours[19] and competitive region-growing based algorithm[20] among many others. A fully automated contouring process is seldom available, therefore semi-auto-contouring is more applicable in clinical settings as it minimises user input and contouring duration, while achieving acceptable contours. The process of semi-auto-contouring involves automatic segmentation combined with manual inputs in the form of contouring on a reference dataset which would then be used to propagate contours to the remaining images.[15,17] This is followed by visual assessment and manual editing as a post-auto-contouring step.[20]

When evaluating contours, the challenge to objectively assess contours exists due to the lack of a "gold standard". Ideally, surgical specimens should be used to reflect the actual tumour size.[18,20] However, this information is rarely available and itself may be subjected to inaccuracies due to distortion in specimens.[20] Since manual contouring has been the main method of defining tumour volume for radiotherapy planning, it is widely accepted as the reference contour of which auto-contours are compared with. In an attempt to reduce bias from individual contours, manually derived contours obtained from multiple observers were used to assess accuracy instead.[6,13,17,21–23] For instance, Martin, S., et al.,[10] compared individual manual contours to the ground truth estimate segmentations derived from simultaneous truth and performance level estimation (STAPLE) algorithm of 6 physicians as this provide the average contours for comparison. Hence, it was not possible to prove that auto-contour is more accurate than manual contour since all studies assessed auto-contours used manual contours as reference. By showing that auto-contours are as accurate or similar to manual contours, it can be deemed effective and thus suitable to be used clinically.

1.1. Aim

The purpose of this review is to present evidence and evaluate auto-contouring in relation to manual contouring of lung tumour on 4DCT datasets by comparing accuracy, variability and duration between manual and auto-contouring.

2. Materials and methods

2.1. *Search strategy for identification of studies*

A systematic approach to identify relevant publications was conducted on three databases; "Science Direct", "Pubmed" and "Embase" using the keywords "Lung cancer" OR "Lung tumour" AND "automatic" AND "delineation" OR "contouring" OR "segmentation".

Randomised controlled trials, prospective and retrospective studies were included for this review. Studies without full text or non-English studies were excluded. No limit was placed on the number of participants in studies.

Patients with different lung cancer staging, tumour size and location were included. For patients with lymph nodes involvement, only the delineation of their primary tumour mass was assessed. No age, gender or performance status restrictions were applied. Patients with lung co-morbidities such as severe fibrosis, chronic obstructive pulmonary disease (COPD), pneumothorax or pleural effusion were excluded from this study.

Automatic and semi-automatic contouring methods conducted on 4DCT were assessed against manual contours which were deemed as the "gold standard" contour. Contouring on other imaging modalities such as 3DCT, positron emission tomography (PET) or magnetic resonance imaging (MRI) were excluded for this review. No exclusion criteria were placed on the type of automatic or semi-automatic segmentation approaches used.

After duplicated studies were removed, titles and abstracts of remaining studies were screened for relevance based on the inclusion and exclusion criteria. Full text from the identified studies was assessed and a secondary manual search on the reference list was conducted to identify additional relevant studies. The Downs and Black checklist was used to assess the quality of selected studies, providing a score based on the methodology of the studies. A total score between 0 and 7 were deemed as poor quality studies, 8 to 14 were deemed as fair quality studies, 15 to 21 were good quality studies and 22 to 28 were excellent quality studies.[24]

2.2. *Outcome measures*

The effectiveness of contouring methods was assessed and compared based on accuracy, variability and duration.

2.3. *Accuracy*

A range of metrics were reported in the studies to evaluate the accuracy of auto-contouring. The Dice Similarity Coefficient (DSC) measures the spatial overlap between two contours and was the most frequently reported quantitative metric used to assess accuracy of contours. A high degree of overlap would yield a value of 1 which represent complete spatial overlap, while a value of 0 represents no spatial overlap.[13] A DSC value of more than 0.7 is considered to have achieved good agreement.[17] Qualitative analysis included in studies such as acceptance rate and magnitude of corrections required were also considered. The metrics were used to analyse and compare auto-contours with manual contours.

2.4. *Variability*

Similar to accuracy outcomes, DSC was the most frequently reported quantitative metric for variability. The metrics are used to calculate inter- and intra-observer variations both within auto-contours and manual contours which can then be compared.

Mean values of spatial overlap metrics provided between manual and auto-contours were used to determine accuracy while standard deviation (SD) and coefficient of variance (COV) were used to assess the degree of variability between contours. Comparison of spatial overlap metrics derived between two contours created with the same method to demonstrate inter- and intra-observer variability.

2.5. *Contouring duration*

Mean, SD and range of contouring time for each contouring method were measured in minutes. For auto-contouring, contouring on reference phase time and computation time make up the total contouring duration. For manual contouring observers' level of experience was also considered in the analysis.

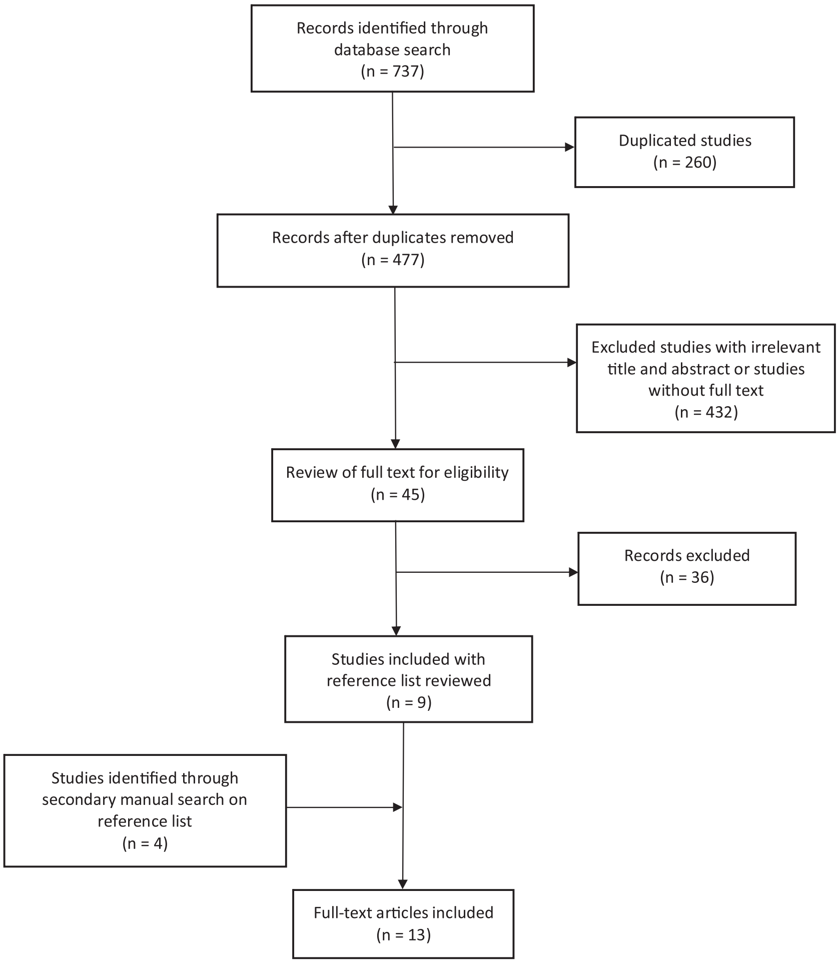
Records identified through database search (n = 737)
Duplicated studies (n = 260)
Records after duplicates removed (n = 477)
Excluded studies with irrelevant title and abstract or studies without full text (n = 432)
Review of full text for eligibility (n = 45)
Records excluded (n = 36)
Studies included with reference list reviewed (n = 9)
Studies identified through secondary manual search on reference list (n = 4)
Full-text articles included (n = 13)

Fig. 1 – Flow diagram showing search results.

2.6. Comparative analysis

To evaluate the significance of outcomes, the range of statistical test used by the studies were collected and reviewed. These include paired student *T*-test, paired two-tailed *T*-test, Wilcoxon signed-rank test, *F*-test, ANOVA analysis and post hoc analysis. Due to heterogeneity within the data collected it was not possible to quantitatively compare the findings in a meta-analysis.

3. Results

The search strategy identified 737 studies of which 477 studies remained after removal of duplicates. Following review of titles and abstracts, Assessment of title and abstracts identified 45 studies based on inclusion and exclusion criteria. Following a full text review and a secondary manual search on the reference list, 13 studies[6,10,13,15,17,21–23,25–29] were deemed relevant for this analysis (Fig. 1). No prospective studies were identified. A summary of selected studies relevant to this review and their corresponding Downs and Black scores are stated in Table 1.

3.1. Accuracy

10 of 13 studies reported on the accuracy of contours.[15,17,21–23,25–29] The main details of studies are presented in Table 2 with all detail available as supplementary material. All studies reported comparable accuracy of auto-contours with manual contours. Auto-contours were propagated either from the mid-ventilation,[15] end-exhalation[17,22,25–28] or end-inhalation[25,29] phase. Auto-contours were assessed against expert-drawn contours and showed good agreement.[15,17,21–23,25–29] Three studies[15,17,26] calculated DSC between auto- and manual contours which range from 0.77 to 0.95. Two studies[15,26] calculated normalised DSC (nDSC), a DSC with the inclusion of an uncertainty index, which range from 0.89 to 1.04. Two studies[27,28] used other metrics such as volume overlap index (VOI), mean overlap index, root-mean-square (RMS) distance and Hausdorff distance and showed that auto-contours were highly similar to manual contours. Three studies[21,22,29] found that auto-contouring achieved accuracy comparable with inter-observer variability in manual contouring. One study by Weiss, E., et al.[23] demonstrated that auto-contouring produced contours that were 9% larger than manual contours.

All three studies[15,26,29] that conducted qualitative analysis of auto-contour reported that auto-contouring produced accurate contours. Speight, R., et al.[15] compared two different auto-contouring algorithms with manual contouring and qualitative assessment was conducted based on clinical acceptability and need for manual editing. Wijesooriya, K., et al.[29] recorded the number of patients requiring manual correction on the propagated contours. Both studies showed that minimal editing was required on auto-contours. Ezhil, M., et al.[26] conducted visual inspection of auto-contours and concluded that propagated contours were mostly within manually drawn contours.

3.2. Variability

Seven studies assessed the variability of contouring methods.[6,10,13,15,17,22,26] Details of studies are presented in Table 3. In the two studies[6,10] that assessed manual contouring only, there were high variabilities in both inter- and intra-observer variations. Louie, A.V., et al.[6] demonstrated that inter-observer variation was more significant than intra-observer in manual contours.

Two studies[13,15] reported lesser inter-observer variation in auto-contours than in manual contours, with mean DSC between observers in auto-contours and manual contours ranging from 0.97 to 0.99 and 0.78 to 0.93 respectively. One study[26] assessed different auto-contouring methods and showed low levels of variations between contours. One study[22] stated that the inter-observer mean distance difference over all angles for manual contours as 2.1 mm.

Three studies[13,15,17] assessed intra-observer variability in auto-contours as compared to manual contours. Speight, R., et al.[15] showed that intra-observer variability was reduced, with mean DSC improving from 0.78 to 0.99. Gaede, et al.[17] and van Dam, I.E., et al.[13] showed that intra-observer variability was not significantly reduced with auto-contouring with mean DSC ranging from 0.82 to 0.95.

3.3. Contouring duration

Eight studies recorded the contouring duration for each patient.[6,13,15,17,22,26–28] Details on contouring duration and physician experience are listed in Table 4. In the three studies that compared the duration of manual with auto-contouring methods,[13,15,17] two studies reported that contouring duration was reduced with auto-contouring while one study showed otherwise. One study reported on the duration of manual contouring only[6] and three studies reported on the duration for auto-contouring only.[22,26–28] Overall, manual contouring duration per patient range from 15 to 90 min while total auto-contouring duration per patient, including manual contouring on reference phase and automatic propagation to other phases, ranging from 18 to 76 min. Computation time range from 10 to 51 min.

4. Discussion

Manual lung tumour delineation on 4DCT datasets are subjected to inaccuracies, variabilities and the process is highly labour intensive.[15] The use of automatic segmentation techniques has been investigated and evaluated for its suitability as an alternative to manual delineation in these datasets.

Accuracy in tumour contouring is a key step in treatment planning, affecting the geometric accuracy of plans and hence tumour control probability. The reference 4DCT dataset used for automatic propagation varies among the studies as seen in Table 2. Most commonly, the end-exhale phase was used to manually construct the reference tumour volume before auto-propagation as it is highly reproducible in predicting respiratory motion.[30] However, the end-inhale phase was reported in one study to be more suitable for contour propagation as it consists of more known than unknown data.[30]

Table 1 – Summary of selected studies.

Author (year), [reference number]	Study type	Auto-contouring algorithm	Volume	Participants (*n*)	Observers (*n*)	Tumour location	Comparison	Types of data	Time assessed?	Downs and Black score
Martin, S., et al. (2015)[10]	R	NA	GTV	10	6	NA	Manual vs Ground truth estimation	VOE, RMS symmetric surface distance	No	13
Speight, R., et al. (2011)[15]	R	Atlas-Based DIR: B-Spline algorithm Demons algorithm	ITV	25	1	All	Two auto vs manual	DSC, MDA, nDSC, QA	Yes	15
Gaede, S., et al. (2011)[17]	R	Model-based DIR	GTV	10	6	RLL, LLL, RUL and LUL	Inter and intra-observer variability, Auto vs Manual	DSC Centroid motion	Yes	15
van Dam, I.E., et al. (2010)[13]	R	B-spline algorithm	GTV, ITV	6	Manual: 2 Auto: 3	Peripheral lungs	Inter and intra-observer variability, Auto vs Manual	DSC	Yes	14
Louie, A.V., et al. (2010)[6]	R	NA	GTV	10	6	RLL, LLL, RUL and LUL	Interobserver and intra-observer variability	COV, VOI	Yes	18
Ehler, E.D., et al. (2009)[25]	R	Model based segmentation, Rigid translational image registration, Deformable surface mesh	GTV, ITV	6	NA	Peripheral and central of RL and LL	Accuracy of automatic contouring, Manual vs auto	ICI, ECI, Centroid position and volume	No	12
Wijesooriya, K., et al. (2008)[29]	R	Large deformable diffeomorphic image registration	GTV	13	1	NA	Auto vs manual	COM comparison, Surface congruence analysis, QA	No	12
Weiss, E., et al. (2008)[23]	R	Large deformation diffeomorphic image registration algorithm	GTV	12	4	Peripheral and central of RL and LL	Auto vs manual, Dosimetric evaluation		No	15

– Table 1 (*Continued*)

Author (year), [reference number]	Study type	Auto-contouring algorithm	Volume	Participants (*n*)	Observers (*n*)	Tumour location	Comparison	Types of data	Time assessed?	Downs and Black score
Wang, H., et al. (2008)[28]	R	Intensity-based (volume) DIR algorithm	GTV	9	NA	NA	Auto vs manual or physician modified deformed contours	VOI	Yes*	12
Ezhil, M., et al. (2008)[26]	R	Rigid image registration in house vs vendor vs addition of deformable adaptation	GTV	20	1	RLL, LLL, RUL, LUL and central of RL and LL	Rigid vs adaptive vs manual	DSC	Yes*	14
Shekhar, R., et al. (2007)[27]	R	Free-form deformation based non-rigid image registration algorithm	GTV	5	1	NA	Auto vs manual	VOI, RMS distance, Hausdorff distance	Yes*	10
Orban de Xivry, J., et al. (2007)[21]	R	‘B-spline’ registration	GTV, ITV	13	2	NA	Auto vs manual	CI	No	11
Pevsner, A., et al. (2006)[22]	R	Deformable-object matching model	GTV	6	4	RLL, LLL, RUL, LUL and middle lobe	Auto vs manual	Contour comparison algorithm, Vector displacement	Yes*	13

R: retrospective; GTV: gross tumour volume; ITV: internal tumour volume; DIR: deformable image registration; VOE: volume overlap error; RMS: root mean square; DSC: dice similarity coefficient; nDSC: normalised dice similarity coefficient; MDA: mean distance to agreement; RLL = right lower lobe; LLL = left lower lobe; RUL = right upper lobe; LUL = left upper lobe; RL: right lung; LL: left lung; COV: coefficient of variation; VOI: volume overlap index; COM: centre of mass; ICI: intrinsic conformation index; ECI: extrinsic conformation index; CI: concordance index; QA: qualitative analysis; NA: not applicable.

* Computation time only.

Table 2 – Studies evaluating accuracy of contours. Main reported metric displayed, all reported metrics are available in supplementary materials.

Author (year) [Reference]	Dataset	Metric reported	Statistical test and significance	Qualitative analysis	Conclusion
Speight, R., et al. (2011)[15]	Manual: MV, MI and ME (union) Auto: MV phase	Mean DSC (SD) B-Spline: 0.859 (± 0.044) Demons: 0.856 (± 0.045)	Paired student t-test ($p > 0.05$): Statistically insignificant between the two auto-algorithm.	No. of patients with nDSC < 1 (1 mm uncertainty): 7 No significant diff between b-spline and demons.	Auto-contouring produce contours with good agreement.
Gaede, S., et al. (2011)[17]	Manual: all phases Auto: ME phase	Mean DSC > 0.8	Paired two-tailed t-test ($p < 0.05$ was considered to be significant): No significant difference Mean DSC in 7 out of 10 cases.	NA	Good agreement between auto-contours with manual contours.
Ehler, E.D., et al. (2009)[25]	Manual: all phases Auto: either ME or MI based on amount of artefacts	Mean ICI GTV Propagated: 0.905 Adapted: 1.008	NA	NA	Good agreement between ITV and GTV with manual contours in primary breathing phase.
Wijesooriya, K., et al. (2008)[29]	Manual: all phases Auto: propagated from MI phase	Fractional Volume agrees within 0.2 ± 0.1 Mean COM Agrees within 0.5 ± 1.5 mm	NA	Number of patients requiring manual correction based on expiration phase is minimal.	Good agreement between auto and manual contours.
Weiss, et al. (2008)[23]	Manual: all phases Auto: propagate from MI phase	Volume Mean difference: T2 = 2.14 cm^3 (SD 4.04 cm^3) T5 = 3.00 cm^3 (SD 5.08 cm^3)	NA	NA	Auto-contouring is accurate, but resulted in larger contours (up to 9%.)
Wang, H., et al. (2008)[28]	Manual: all phases Auto: propagated from ME phase	Mean VOI: 98.3% (range 96.1–99.3%)	NA	NA	Deformed contours agree well with physician drawn contours.
Ezhil, M., et al. (2008)[26]	Manual: all phases Auto: propagated from ME phase	Mean DSC (SD) Rigid in-house = 0.88 (±0.04) Rigid commercial = 0.88 (±0.06) Adaptive = 0.77 (±0.10)	NA	The propagated IGTVs were mostly within the mIGTVs.	Rigid-body propagation method is accurate in generating ITV within a 1 mm margin of uncertainty.
Shekhar, R., et al. (2007)[27]	Manual: ME (expert) and MI phase Auto: ME phase	Mean COM distance 1.96 mm (range 0.2–4.1)	NA	NA	Algorithm used produced more accurate segmentation results than those in previously published reports.
Orban de Xivry, J., et al. (2007)[21]	Manual: all phases Auto: One reference phase[*]	CI showed auto comparable accuracy with manual inter-observer variability.	The Wilcoxon signed-rank test ($p < 0.01$): No significant difference in inter-observer agreement and the method–physician 1 agreement in most patients.	NA	Accuracy was comparable to interobserver variability.

– Table 2 (*Continued*)

Author (year) [Reference]	Dataset	Metric reported	Statistical test and significance	Qualitative analysis	Conclusion
Pevsner, A., et al. (2006)[22]	Manual: MI phase (guided by the same ME contours) Auto: ME phase	Mean distance differences over all angles Interobserver = 2.1 mm (range: 1.1–3.7 mm) Model-observer = 2.6 mm (range: 1.1–5.1 mm)	Paired two-sided *t* test: No significance between interobserver and model–observer GTV differences ($p = 0.39$)	NA	Discrepancy in magnitude between observer variations versus model-observer variations was not statistically significant.

MV: mid ventilation; MI: max inhale; ME: max exhale; DSC: dice similarity coefficient; nDSC: normalised dice similarity coefficient; ICI: intrinsic conformation index; COM: centre of mass; VOI: volume overlap index; ITV: internal tumour volume; GTV: gross tumour volume; IGTV: internal GTV; mIGTV; manual IGTV; SD: standard deviation; CI: Concordance index; NA: not applicable.

* No mention as to which specific phase.

Furthermore, motion or binning artefacts in 4DCT images can result in poor image quality.[8] Therefore, the propagation phase used and the amount of artefacts present contribute to inaccuracies in contouring.

Generally, a good agreement was found between auto- and manual contours in all studies. Auto-contours was also able to produce contours within inter-observer variability in manual contours,[21,22,29] thus showing that auto-contouring can localise lung tumours with similar accuracy to manual observers. The studies included in this review have mainly used manual contours as the reference contour for auto-propagation and in the calculation of quantitative measurements. As auto-contouring often requires a set of manual contours as reference for propagation, the accuracy of auto-contours will be affected by inconsistency and inaccuracies in the manually contoured reference phase.[15] Shekhar, R., et al.[27] observed poor VOI, RMS and Hausdorff distance between auto- and manual contours in one lung cancer case due to poorly contoured reference phase caused by the characteristic of the tumour. Tumours that are small, diffused or have spiky appearance are harder to contour as their boundaries are ambiguous to observers. Auto-contouring propagates these errors to other phases, resulting in an inaccurate volume being delineated. Hence, to ensure that manual contouring of tumour is done accurately the provision of comprehensive guidelines, use of multiple imaging modalities for target definition, regular learning or training workshops and a robust quality assurance process before contour propagation should be implemented.[5]

A wide range of tumour locations, as listed in Table 1, were included in the studies to allow for a more holistic review of auto-contouring. Assessment of the impact of tumour location on the accuracy of auto-contouring was not conducted by the studies. However, the location of the tumour may affect the accuracy of auto-contouring. Previous studies have shown that tumours located in close proximity to blood vessels or chest wall can pose a challenge for the auto-contouring tool to differentiate between structures with similar densities.[20,31] Uninvolved healthy structures might be delineated as part of the tumour which is an inaccurate representation of the actual tumour volume. Future studies should evaluate the impact tumour location have on the accuracy of auto-contouring. It may be suggested that auto-contouring should only be used for well-circumscribed tumours, while more advanced segmentation algorithms would be needed for diffused or tumour located near other structures.[32]

Although studies[15,26,29] have concluded that auto-contours were accurate based on qualitative analysis, a degree of manual correction was still required for contours to be clinically acceptable. This is in line with clinical practice as contours should be visually inspected and edited accordingly as part of the quality assurance process.[20]

The level of observers' experience was not widely addressed by the studies. Some studies merely mentioned that contours were completed by or under the supervision of an experienced physician.[13,23,29] Moreover, the definition of an 'experienced' physician was not provided in the studies. As seen in some studies, observers' experience is usually defined by the number of years in practice or their area of expertise.[33–35] The accounting for any correlation between observers' experience with accuracy of contours was also lacking. This may be due to the limited reporting of such information. This is not in keeping with published data which suggested that observers' years of experience may affect the accuracy of tumour delineation as inexperience or inadequate training can lead to imprecise localisation of tumour such as including adjacent normal structures into tumour contour[36] or inability in identifying microscopic diseases.[37]

The subjective nature of tumour contouring can lead to high rates of inter- and intra- observer variability in manual contouring. The window settings of which the observer used for contouring affects the visibility of the lesion possibly causing variation in contours. In addition it was shown that auto-contouring of GTV can vary when applied on 4DCT datasets of different breathing phases, between maximum inhalation and exhalation.[8] These variabilities exist between and within observers, leading to inter- and intra-observer variabilities in manual contouring respectively.

Inter- and intra-observer variation was observed in manual contours.[6,10] Two studies demonstrated that auto-contouring reduces inter-observer variability while intra-observer variability was not significantly reduced.[13,17] On the contrary, Speight, R., et al.[15] demonstrated that intra-observer variation was significantly reduced. However, this outcome was based on the contours of a single patient by a single observer. Since intra-observer variability is lesser than inter-observer

Table 3 – Studies evaluating variability of contours.

Author (year) [Reference]	Dataset manually contoured	Mean DSC (SD)	Volume	Distances (mm)	Conclusion
Martin, S., et al. (2015)[10]	Manual: All phases	NA	Mean observer VOE range: SD = 0.6–13.2% COV = 6.30–93.25%	Symmetric RMS range: SD = 0.2–1.2 COV = 4.02–28.56%	High variability among observer with presence of outliers.
Speight, R., et al. (2011)[15]	Manual: MV, MI and ME phases, Auto: MV phase	**Intra-observer** B-spline: 0.99 cm Manual: 0.78	NA	Mean MDA = 0.05 cm	Auto-contouring has high reproducibility and is more consistent than current manual method.
Gaede, S., et al. (2011)[17]	Manual: all phases Auto: ME phase	**Inter-observer** Range: Manual = 0.55–0.91 Auto = 0.50–0.92 **Intra-observer** >0.8* Range 0.82–0.95	NA	NA	Significant decrease in inter-observer variations with auto-contouring. Low intra-observer variation in both methods.
van Dam, I.E., et al. (2010)[13]	Manual: all phases and MIP Auto: MI phase	**Inter-observer** Manual all phases: 0.93 (1SD 0.02) Auto: 0.89 (1SD 7.2%) Exclude outlier = 0.97 (1SD = 0.01) **Intra-observer** Auto: 0.93 (*T*-test: $p < 0.02$)	**Inter-observer** Manual all phases: $R > 0.87$ Auto: Mean overlap = 86% (10.6% 1SD) (*T*-test: $p < 0.02$) **Intra-observer** Auto: Mean overlap = 92.9% (4.2% 1SD)	NA	Lesser inter-observer variations. Similar intra-observer variation.
Louie, A.V., et al. (2010)[6]	Manual: all phases	NA	**Inter-observer** Mean VOI = 0.802 (range 0.556–0.915, mean SD 0.064) **Intra-observer** Mean VOI = 0.802 (range 0.770–0.825, mean SD 0.059)	NA	Inter-observer variation is a more significant source of error than intra-observer variability. Case difficulty is significant for inter-observer variability.
Ezhil, M., et al. (2008)[26]	Manual: all phases Auto: ME phase	Rigid in-house = 0.88 (± 0.04) Rigid commercial = 0.88 (±0.06) Adaptive = 0.77 (±0.10)	NA	NA	Automatic segmentation is consistent.
Pevsner, A., et al. (2006)[22]	Manual: MI phase (guided by the same ME contours) Auto: ME phase	NA	NA	**Inter-observer** Differences over all angles: Mean = 2.1 (range 1.1–3.7)	Discrepancy in magnitude between observer variations versus model-observer variations was not statistically significant.

MV: mid ventilation; MI: max inhale; ME: max exhale; DSC: dice similarity coefficient; SD: standard deviation; VOE: volume overlap error; COV: coefficient of variance; RMS: root-mean-square; MDA: mean distance to agreement; *R*: correlation test; VOI: volume overlap index; NA: not applicable.

* Between auto and manual edited by the same physician.

Table 4 – Studies evaluating contouring time.

Author (year) [Reference]	Auto-contouring algorithm	Manual time (minutes per case)	Auto time (minutes per case)	Observer experience

Speight, R., et al. (2011)[15]	B-spline
	Demons
		Mean = 53
		Range = 15–74
			Mean clinician = 25
			Mean computation:
			B-spline (51 min 3 s)
			Demons (34 min 22 s)
				Not reported

Gaede, S., et al. (2011)[17]	Model based DIR	Mean = 42.7
		S.D = 18.6
		Range = 15–90
			Mean = 17.7
			S.D = 5.4
			Range = 9.5–33 (includes time to delineate reference phase, auto-propagation, and review time)
				2 radiation oncology residents and 4 radiation oncologists.

van Dam, I.E., et al. (2010)[13]	B-spline	GTV range = 24–89 ITV range = 4–12	Total range = 40–58 Mean computation = 36	Not reported
Louie, A.V., et al. (2010)[6]	NA	Mean: All cases = 42.7 (range 15–90) Difficult cases = 49.5 Easy cases = 35.9	NA	6 radiation oncologists with 1–20 years of experience.
Wang, H., et al. (2008)[28]	Intensity-based DIR	NA	Mean computation: 27	NA
Ezhil, M., et al. (2008)[26]	Rigid image registration Deformable adaptation	Not reported	Propagation time: Rigid (5 min) Adaptive (20 min)	A radiation oncologist
Shekhar, R., et al. (2007)[27]	Free-form deformable non-rigid image registration	NA	Mean computation: Standard PC = 10–12 h Advance microprocessor = reduced to minutes (not specified)	NA
Pevsner, A., et al. (2006)[22]	Deformable object matching model	NA	Computation range: 10–15	NA

NA: not applicable.

variability,[13,15] it was expected that an improvement in intra-observer variability will not be as significant as inter-observer variation even with the use of auto-contouring.

As mentioned, observers' experience was not widely evaluated in the studies. One study by Louie, A.V., et al.,[6] found that the amount of variability in contours was not significantly affected by observer's experience. However, previous study has shown that variations exist between junior and senior physicians.[38] The lack of consistent training and learning of contouring skills may lead to increased variability especially in inexperienced physicians.[39] Interestingly, Louie, A.V., et al.[6] assessed the effect of case difficulty on inter-observer variations in manual contours which was found to be significant. This was caused by increase amount of image artefacts in difficult cases, making the visualisation of tumour vague. To mitigate these existing problems, the implementation of guidelines, consistent training and use of auto-contouring can help to further reduce inter-observer variation in challenging cases.[5]

It is important to note the limitations of overlap indices used in these studies. Overlap metrics such as DSC and VOI are frequently used in studies that assess contours[40,41] to determine the similarity between two sets of contours.[27] However, they fail to account for spatial information and are more sensitive in smaller tumours compared to large tumours.[27] These metrics only accounts for the size of contours relative to each other but not their location. Furthermore, the same amount of disagreement between contours of a small tumour will result in poorer overlap indexes than larger tumours as it will account for a larger proportion of the area or volume of the contour. This necessitates the need for additional metrics to be used such as RMS distance, Hausdorff distance, mean distance to agreement (MDA) and centre of mass (COM) displacements to provide a more comprehensive and accurate method of quantitative analysis.

Manual contouring in all phases of 4DCT is practiced clinically, however this process is resource intensive. Auto-contouring may present as a time-saving solution that reduces the amount of manual input needed for contouring. From the studies included, both Gaede, S., et al.[17] and van Dam, I.E., et al.[13] compared duration of manual contouring with the total duration of auto-contouring which includes the time taken for manual delineation on a reference phase, followed by auto-propagation and a final review of auto-contours by physicians. Evaluation of total contouring time is a fair and appropriate way to compare the efficiency of contouring methods as it accounts for the total time taken for the entire process to construct clinically acceptable contours. Both studies reported shorter total contouring duration with auto-contouring. In the study conducted by Speight et al.,[15] although clinician's involvement in the contouring process was reduced by a mean of 28 min, the total time required for auto-contouring was still greater than manual contouring with the largest difference in mean time of 23 min between the two methods. This contrasting result can be attributed to the long computation time in the study which took up majority of the total contouring time.

Computation time varied between studies due to differences in algorithms and computer processors used. Several different auto-contouring algorithms were discussed in the studies. They are mainly categorised into rigid and non-rigid deformable image registrations (DIR). DIR is a process of elastic deformation to create a layout of corresponding features.[15] Its ability to accommodate respiratory motion is especially useful in lung tumour segmentation.[29] Basis-spline (B-spline) algorithm was the most commonly investigated mathematical algorithm followed by demon and model-based algorithms. Intensity-based DIR is a non-rigid DIR method and was used by some studies.[27,28] It is a suitable method for contouring on CT images due to intensity consistency in CT.[42] However, intensity-based DIR is extremely slow when used on standard PCs,[27] with computation duration of up to 10 to 12 h. From the studies, mean computation time for B-spline and demon algorithms was 51 min 3 s and 34 min 22 s respectively.[15] Computation time was exceptionally long in the study conducted by Shekhar, R., et al.,[27] where mean computation time took 10 to 12 h on a standard computer processor which is impractical. Other computation time reported range from five to 51 min.[13,15,22,26–28] The availability of fast computer processor in the department is thus an important factor in determining the time-saving advantage of auto-contouring. It is useful to note that although overall duration for contouring did not decrease with auto-contouring and in some cases longer than manual contouring, clinicians' involvement in the contouring process was greatly reduced with auto-contouring. The time required for clinicians to contour the reference phase and to review the final contours was lesser than the total time needed for manual contouring. Therefore, this reduces clinicians' workload, allowing time to be devoted to other duties. Departments planning to adopt auto-contouring into their workflow should utilise fast and efficient computer processors to gain maximum time-saving with auto-contouring.

The non-blinded nature of the studies may introduce bias as clinicians were aware of being timed. This might have affected the accuracy of duration measurements as timing individuals in an experimental setting might not be reflective of actual clinical situation.

4.1. *Future work*

Due to the retrospective nature of the studies, direct comparison between different algorithms was not possible. Factors such as the difficulty of lung cancer cases used, type of algorithms, speed of computer processor, observers' training and experience and image window settings were not consistent between the studies. Moreover, the studies were limited by the number of patients and observers. For future considerations, further prospective controlled studies conducted in the future would provide a more reliable assessment of auto-contouring. Furthermore, future progression should focus on the development of fully automated contouring algorithms which eliminates any human intervention. This would remove potential human errors from the contouring process thus increasing accurate and decreasing variability while reducing contouring duration.

5. Conclusion

Overall, auto-contouring was found to be accurate, reproducible and efficient when compared to manual contouring alone. Accuracy outcomes were limited by the lack of 'gold standard' contours to be compared with, hence auto-contours were not proven to be more accurate than manual contours. However, contours from both methods were comparable making auto-contouring a suitable substitute for manual contouring. Automation decreases inter-observer variability due to lesser human interventions while intra-observer variations remain relatively consistent for both methods. Auto-contouring is a time-saving solution for contouring on 4DCT images as clinicians' contouring time was reduced. Long overall time with auto-contouring was due to extended computation time which can be further reduced by developing an efficient contouring algorithm and using faster computer processors. Prospective blinded trials conducted in the future would therefore provide more reliable results and thus aid in clinical implementation of auto-contouring.

Conflict of interest

None declared.

Financial disclosure

None declared.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.rpor.2019.01.003.

REFERENCES

1. Parashar B, Arora S, Wernicke AG. Radiation therapy for early stage lung cancer. *Sem Interv Radiol* 2013;**30**(2):185–90.
2. Maciejczyk A, Skrzypczyńska I, Janiszewska M. Lung cancer. Radiotherapy in lung cancer: actual methods and future trends. *Rep Pract Oncol Radiother* 2014;**19**(6):353–60.

3. Nielsen TB, Hansen CR, Westberg J, Hansen O, Brink C. Impact of 4D image quality on the accuracy of target definition. *Australas Phys Eng Sci Med* 2016;**39**(1):103–12.
4. Njeh CF. Tumor delineation: the weakest link in the search for accuracy in radiotherapy. *J Med Phys* 2008;**33**(4):136–40.
5. Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. *J Med Imaging Radiat Oncol* 2016;**60**(3):393–406.
6. Louie AV, Rodrigues G, Olsthoorn J, Palma D, Yu E, Yaremko B, et al. Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era. *Radiother Oncol* 2010;**95**(2):166–71.
7. Li F, Li J, Zhang Y, Xu M, Shang D, Fan T, et al. Geometrical differences in gross target volumes between 3DCT and 4DCT imaging in radiotherapy for non-small-cell lung cancer. *J Radiat Res (Tokyo)* 2013;**54**(5):950–6.
8. Wei J, Li G. Automated lung segmentation and image quality assessment for clinical 3D/4D computed tomography. *IEEE J Transl Eng Health Med* 2014;**2**.
9. Adamczyk M, AUTP. Respiratory motion and its compensation possibilities in the modern external beam radiotherapy of lung cancer. *Nowotwory* 2017;**67**(5):292–6.
10. Martin S, Johnson C, Brophy M, Palma DA, Barron JL, Beauchemin SS, et al. Impact of target volume segmentation accuracy and variability on treatment planning for 4D-CT-based non-small cell lung cancer radiotherapy. *Acta Oncol* 2015;**54**(3):322–32.
11. Wang L, Fan J, Lin T, Jin L, Ma C. SU-E-T-521: is cone beam CT (CBCT) equivalent to 4-dimensinoal (4D) MIP images to account for target motion in SBRT lung treatment? *Med Phys* 2011;**38**(6 Part 18):3608–9.
12. Muirhead R, McNee SG, Featherstone C, Moore K, Muscat S. Use of Maximum Intensity Projections (MIPs) for target outlining in 4DCT radiotherapy planning. *J Thorac Oncol* 2008;**3**(12):1433–8.
13. van Dam IE, de Koste JRvS, Hanna GG, Muirhead R, Slotman BJ, Senan S. Improving target delineation on 4-dimensional CT scans in stage I NSCLC using a deformable registration tool. *Radiother Oncol* 2010;**96**(1):67–72.
14. Van Elmpt W, Van Der Stoep J, Van Soest J, Lustberg T, Gooding M, Dekker A. Atlas-based segmentation reduces interobserver variation and delineation time for OAR in time OAR NSCLC. *Radiother Oncol* 2017;**123**:S661.
15. Speight R, Sykes J, Lindsay R, Franks K, Thwaites D. The evaluation of a deformable image registration segmentation technique for semi-automating internal target volume (ITV) production from 4DCT images of lung stereotactic body radiotherapy (SBRT) patients. *Radiother Oncol* 2011;**98**(2):277–83.
16. Jin Rim Y, Shouliang Q, van Triest HJW, Yan K, Wei Q. Automatic segmentation of juxta-pleural tumors from CT images based on morphological feature analysis. *Bio-Medical Materials & Engineering* 2014;**24**(6):3137–44.
17. Gaede S, Olsthoorn J, Louie AV, Palma D, Yu E, Yaremko B, et al. An evaluation of an automated 4D-CT contour propagation tool to define an internal gross tumour volume for lung cancer radiotherapy. *Radiotherapy and Oncology* 2011;**101**(2):322–8.
18. Rios Velazquez E, Aerts HJ, Gu Y, Goldgof DB, De Ruysscher D, Dekker A, et al. A semiautomatic CT-based ensemble segmentation of lung tumors: comparison with oncologists' delineations and with the surgical specimen. *Radiother Oncol* 2012;**105**(2):167–73.
19. *A Novel Application of Principal Surfaces to Segmentation in 4D-CT for Radiation Treatment Planning*; 2010. p. 758.
20. Velazquez ER, Parmar C, Jermoumi M, Mak RH, Van Baardwijk A, Fennessy FM, et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Scientific reports* 2013;**3**:3529.
21. Orban de Xivry J, Janssens G, Bosmans G, De Craene M, Dekker A, Buijsen J, et al. Tumour delineation and cumulative dose computation in radiotherapy based on deformable registration of respiratory correlated CT images of lung cancer patients. *Radiother Oncol* 2007;**85**(2):232–8.
22. Pevsner A, Davis B, Joshi S, Hertanto A, Mechalakos J, Yorke E, et al. Evaluation of an automated deformable image matching method for quantifying lung motion in respiration-correlated CT images. *Medical Physics* 2006;**33**(2):369–76.
23. Weiss E, Wijesooriya K, Ramakrishnan V, Keall PJ. Comparison of intensity-modulated radiotherapy planning based on manual and automatically generated contours using deformable image registration in four-dimensional computed tomography of lung cancer patients. *International Journal of Radiation Oncology Biology Physics* 2008;**70**(2):572–81, e2.
24. Sara HD, Nick B. The Feasibility of Creating a Checklist for the Assessment of the Methodological Quality Both of Randomised and Non-Randomised Studies of Health Care Interventions. *Journal of Epidemiology and Community Health (1979-)* 1998;**6**:377.
25. Ehler ED, Bzdusek K, Tom, eacute WA. A method to automate the segmentation of the GTV and ITV for lung tumors. *Medical dosimetry* 2009;**34**(2):145–53.
26. Ezhil M, Choi B, Starkschall G, Bucci MK, Vedam S, Balter P. Comparison of Rigid and Adaptive Methods of Propagating Gross Tumor Volume Through Respiratory Phases of Four-Dimensional Computed Tomography Image Data Set. *International Journal of Radiation Oncology Biology Physics* 2008;**71**(1):290–6.
27. Shekhar R, Lei P, Castro-Pareja CR, Plishker WL, D'Souza WD. Automatic segmentation of phase-correlated CT scans through nonrigid image registration using geometrically regularized free-form deformation. *Medical Physics* 2007;**34**(7):3054–66.
28. Wang H, Garden AS, Zhang L, Wei X, Ahamad A, Kuban DA, et al. Performance Evaluation of Automatic Anatomy Segmentation Algorithm on Repeat or Four-Dimensional Computed Tomography Images Using Deformable Image Registration Method. *International Journal of Radiation Oncology Biology Physics* 2008;**72**(1):210–9.
29. Wijesooriya K, Weiss E, Dill V, Dong L, Mohan R, Joshi S, et al. Quantifying the accuracy of automated structure segmentation in 4D CT images using a deformable image registration algorithm. *Medical Physics* 2008;**35**(4):1251–60.
30. Renchao J, Yongchuan L, Mi C, Sheng Z, Enmin S. Contour propagation for lung tumor delineation in 4D-CT using tensor-product surface of uniform and non-uniform closed cubic B-splines. *Phys Med Biol* 2018;**63**(1):1.
31. Zhao B, Yankelevitz D, Reeves A, Henschke C. Two-dimensional multi-criterion segmentation of pulmonary nodules on helical CT images. *Medical Physics* 1999;**26**(6):889–95.
32. Mukhopadhyay S. A Segmentation Framework of Pulmonary Nodules in Lung CT Images. *Journal of Digital Imaging* 2016;**29**(1):86–103.
33. de Hoop B, Gietema H, van Ginneken B, Zanen P, Groenewegen G, Prokop M. A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: what is the minimum increase in size to detect growth in repeated CT examinations. *Eur Radiol* 2009;**19**(4):800–8.
34. Dinkel J, Khalilzadeh O, Hintze C, Fabel M, Puderbach M, Eichinger M, et al. Inter-observer reproducibility of semi-automatic tumor diameter measurement and volumetric analysis in patients with lung cancer. *Lung Cancer* 2013;**82**(1):76–82.

35. Echegaray S, Nair V, Kadoch M, Leung A, Rubin D, Gevaert O, et al. A Rapid Segmentation-Insensitive "Digital Biopsy" Method for Radiomic Feature Extraction: Method and Pilot Study Using CT Images of Non-Small Cell Lung Cancer. *Tomography* 2016;**2**(4):283–94.
36. Van de Steene J, Linthout N, de Mey J, Vinh-Hung V, Claassens C, Noppen M, et al. Definition of gross tumor volume in lung cancer: inter-observer variability. *Radiother Oncol* 2002;**62**(1):37–49.
37. Lin R. Target volume delineation and margins in the management of lung cancers in the era of image guided radiation therapy. *J Med Radiat Sci* 2014;**61**(1):1–3.
38. Giraud P, Elles S, Helfre S, De Rycke Y, Servois V, Carette M-F, et al. Conformal radiotherapy for lung cancer: different delineation of the gross tumor volume (GTV) by radiologists and radiation oncologists. *otherapy and Oncology* 2002;**62**(1):27–36.
39. Weiss E, Hess CF. The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy theoretical aspects and practical experiences. *Strahlenther Onkol* 2003;**179**(1):21–30.
40. Wang S, Zhou M, Liu Z, Liu Z, Gu D, Zang Y, et al. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical Image Analysis* 2017;**40**:172–83.
41. Yip SSF, Parmar C, Blezek D, Estepar RSJ, Pieper S, Kim J, et al. Application of the 3D slicer chest imaging platform segmentation algorithm for large lung nodule delineation. *PLoS ONE* 2017;**12**(6.).
42. Wu Q, Cao R, Pei X, Jia J, Hu L. Deformable image registration of CT images for automatic contour propagation in radiation therapy. *Bio-Medical Materials and Engineering* 2015;**26**:S1037–44.