Challenges in the target volume definition of lung cancer radiotherapy

Susan Mercieca1,2, José S. A. Belderbos3, Marcel van Herk4

1Faculty of Health Science, University of Malta, Msida, Malta; 2The University of Amsterdam, Amsterdam, The Netherlands; 3Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands; 4University of Manchester, Manchester Academic Health Centre, The Christie NHS Foundation Trust, Manchester, UK

Contributions: (I) Conception and design: All authors; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: S Mercieca; (V) Data analysis and interpretation: S Mercieca; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Susan Mercieca. Faculty of Health Science, University of Malta, Msida, Malta; The University of Amsterdam, P.O. Box 19268 1000 GG, Amsterdam, The Netherlands. Email: susan.mercieca@um.edu.mt.

Abstract: Radiotherapy, with or without systemic treatment has an important role in the management of lung cancer. In order to deliver the treatment accurately, the clinician must precisely outline the gross tumour volume (GTV), mostly on computed tomography (CT) images. However, due to the limited contrast between tumour and non-malignant changes in the lung tissue, it can be difficult to distinguish the tumour boundaries on CT images leading to large interobserver variation and differences in interpretation. Therefore the definition of the GTV has often been described as the weakest link in radiotherapy with its inaccuracy potentially leading to missing the tumour or unnecessarily irradiating normal tissue. In this article, we review the various techniques that can be used to reduce delineation uncertainties in lung cancer. The findings of this review indicate that to date, it is still not possible to eliminate interobserver variation in the definition of GTV. Positron Emission Tomography (PET-CT) has an important role in improving the staging accuracy and the definition of the tumour. Various autosegmentation tools have also been proposed to fully or partially automate the delineation process. However, their development is currently hindered by the unavailability of absolute gold standards that can be used to train and validate these algorithms. Hence, manual delineation is still considered to be the gold standard. Nevertheless, auto-segmented contours can provide a good starting point, eventually reducing the delineation time and interobserver variation. Improvements in image quality can also reduce the delineation uncertainty in some cases. The main factor leading to interobserver variation is image interpretation differences between clinicians. Therefore, protocols, training and peer review checks of delineated contours are essential to address this challenge. The development of the MR-linac will also present new challenges and opportunities in optimising the definition of the target volume as well as in the development of adaptive radiotherapy strategies.

Keywords: Interobserver variation; lung cancer; radiotherapy; gross tumour volume (GTV)

Submitted May 01, 2020. Accepted for publication Jun 15, 2020.
doi: 10.21037/tlcr-20-627
View this article at: http://dx.doi.org/10.21037/tlcr-20-627

Introduction

Radiotherapy with or without systemic treatment has an important role in the management of lung cancer. This treatment involves the precise delivery of ionising radiation to the tumour, with the aim to minimise the dose to normal tissue and hence reduce treatment side effects. Accurate definition of the treatment area is one of the most important steps in high-precision radiotherapy. This process involves defining the gross visible tumour volume (GTV) on
computed tomography images. Margins are added around the GTV to account for microscopic disease, as well as random and systematic set-up errors to form the planning target volume (PTV) (1).

Failure to define the GTV accurately will, therefore, result in a systematic error and lower the precision of the overall radiotherapy workflow. Ironically, the definition of the GTV has also been described as the ‘weakest link’ in the radiotherapy treatment chain (1). Numerous studies have shown that this process is prone to interobserver variation and human errors, particularly for lung cancer (2-7). Lung tumours are often surrounded by interstitial lung tissue changes or atelectasis that look similar to the tumour making it difficult to distinguish tumour boundaries (Figure 1). Furthermore, the definition of the GTV requires the clinician to make complex judgements based on the patient’s clinical history, diagnostic images, and anatomical knowledge to identify the target and potential routes of spread.

Another important limitation is that the final GTV delineation represents a snapshot of the tumour shape and position in time. The tumour can change during the course of treatment as a result of changes in respiratory motion, tumour baseline shifts, regression and progression and anatomical changes caused by pleural effusion and infiltrative changes (9). Large safety margins are required to account for this uncertainty, potentially limiting dose escalation. Image-guided radiotherapy has, therefore, a crucial role in identifying these changes during treatment and various techniques have been proposed to adapt the treatment accordingly.

In this article, the extent of this problem will be discussed together with techniques that could be used to reduce uncertainties in target volume delineation.

Quantifying interobserver variation

Interobserver variation in the definition of the GTV can be classified as minor or major (10). Minor interobserver variation includes small deviations caused by the difficulty to outline “fuzzy” tumour boundaries on the images using the contouring tools available. Major variations are clinically significant changes that may lead to a geographical tumour miss or unnecessary dose to healthy tissue. These are generally caused by differences in image interpretation and human errors, for example, by failing to contour involved lymph nodes or tumour extensions (4,10).

Various metrics have been proposed to quantify interobserver variation in relation to a gold standard including; simple volume and volume overlap measurements, the centre of mass, measures of surface shape variations and dosimetric analysis (11-13). A summary of these metrics, together with their advantages and limitations, is provided in Table S1. The accuracy of these metrics is case dependent and may not always reveal the impact of interobserver variation on the dose to the tumour, organs at risk (OARs) and ultimately, clinical outcomes (14). Furthermore, the lack of an absolute gold standard makes it difficult to accurately validate the accuracy of a delineated contour (10,15,16). It is, therefore recommended to use more than one metric to quantify interobserver variation (17). A qualitative assessment can also be performed whereby an expert or expert panel visually evaluates the contours and classify these as acceptable or unacceptable according to a consensus delineation protocol (9,16,18). The limitation of the latter approach is that it is subjective and time consuming (19). However, when used alongside other quantitative metrics, a qualitative assessment can provide a better understanding of the factors leading to interobserver variation.

Factors contributing to interobserver variation in lung cancer

Numerous studies have been conducted to assess the interobserver variation in lung cancer (3-7,20,21). These
Optimising the definition of the GTV

Although interobserver variation in the definition of the GTV can be classified as a systematic error it is difficult to account for this variation through the use of margins since this variation is often not uniform, case depended and way too large in particularly for interpretational differences leading to an unacceptably large margin.

In view of this, various methods have been proposed in the literature to reduce the interobserver variation in target volume definition including; use of clearer protocols (4,20,31), inclusion of multimodality images (6,28), autosegmentation (32-34), respiratory motion management (35), training (20), and the introduction of peer review checks (18,19,21,25,36) (Table 1).

Multimodality images for target definition

CT is still considered to be the gold standard imaging modality in lung radiotherapy as it provides both 3D anatomical information and tissue densities, necessary for dose calculation. However, the contrast between tumour surrounding soft tissue and malignant changes is often limited.

When using 3DCT, a margin is added around the CTV to account for respiratory tumour motion to form the internal target volume. This margin is based on population respiratory motion data. It does not account for the patient's individual respiratory motion, potentially leading to either an overestimation or an underestimation of the margin required to account for this uncertainty (37). These limitations can be overcome by improving the contrast and spatial resolution on CT. Additional imaging modalities including; positron emission computed tomography (PET-CT), magnetic resonance imaging (MRI), and/or respiratory correlated computed tomography (4DCT) also have an important role.

With the introduction of multimodality imaging, however, there is a need for improved protocols and collaboration between oncologists, radiologists, and nuclear medicine physicians (20,22). Most radiotherapy centres do not have “dedicated” PET-CT and MRI scanners that allow scanning of the patient in the treatment position for radiotherapy planning and therefore, a planning CT is required. When these images are not acquired with the patient in the treatment position mis-registration between the diagnostic images and planning CT is likely making it difficult to identify corresponding structures on the planning CT leading to misinterpretation. Hence, maintaining clear patient set-up and imaging protocols is essential to facilitate the use of multiple images. Furthermore, since the tumour can change over the course of treatment, image-guided radiotherapy can be used to identify the changes and adapt the treatment accordingly.

Improving the CT spatial and contrast resolution

Intravenous iodine contrast can be used to improve the contrast between the tumour tissue and blood vessels. However, due to underlying co-morbidities, not all patients can tolerate intravenous contrast (38). Diagnostic high-resolution CT scan can be used alongside treatment planning CT scans to improve the assessment of interstitial lung disease and lymph node involvement (39).

Role of FDG PET-CT

The tumour activity can be quantified by measuring the standard uptake value (SUV) of radioactive tracer on the PET-CT within a predefined region of interest (40). The PET image provides biological information but very limited anatomical detail. To overcome this problem, a CT is also acquired that is inherently registered (spatially aligned)
<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>No of cases and observers</th>
<th>Intervention</th>
<th>Assessment metrics</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steenbakkers et al. (4)</td>
<td>Evaluated impact of using FDG PET-CT and delineation protocol on interobserver variation using the big brother software</td>
<td>22 NSCLC cases, 11 consultant radiation oncologists</td>
<td>FDG PET-CT protocol</td>
<td>Mean local SD, delineation time</td>
<td>The introduction of FDG PET-CT and delineation protocol reduced the mean local SD from 1.0 to 0.4 cm. The largest reduction in the observer variation was seen in the atelectasis region (local SD 1.9 cm reduced to 0.5 cm). The mean delineation time was reduced from 16 to 12 minutes (P&lt;0.001)</td>
</tr>
<tr>
<td>Fitton et al. (5)</td>
<td>Evaluated the impact of using FDG PET-CT on interobserver variation based on tumour stage and location</td>
<td>22 NSCLC cases, 11 consultant radiation oncologist</td>
<td>FDG PET-CT</td>
<td>Mean local SD</td>
<td>Mean local SD for tumours surrounded by lung tissue was 0.4 cm on CT and reduced to 0.3 cm when using FDG PET-CT (P=0.162). The mean local SD for tumours invading the mediastinum, vessels or pericardium was significantly higher on CT (1.3 cm) as opposed to 0.4 cm when using FDG PET-CT (P&lt;0.001) highlighting the need to use FDG PET-CT for these cases</td>
</tr>
<tr>
<td>Persson et al. (3)</td>
<td>Quantify the interobserver delineation variation for peripheral SBRT lung tumours on 3DCT</td>
<td>22 NSCLC cases, 3 radiologists and 3 radiation oncologists</td>
<td>N/A</td>
<td>Local SD, CI</td>
<td>The mean local SD was 0.15 and 0.26 cm in the transverse and craniocaudal plane, respectively. Tumours with pleural contact had a significantly larger local SD than tumours surrounded by lung tissue. A larger margin in the craniocaudal direction is recommended</td>
</tr>
<tr>
<td>Peulen et al. (7)</td>
<td>Evaluated interobserver variation of early stage NSCLC cases using mid-V planning technique</td>
<td>11 Radiation oncologists, 16 early stage NSCLC cases</td>
<td>N/A</td>
<td>Local SD and PTV margins</td>
<td>A relatively small target delineation uncertainty of 1.2–1.8 mm was observed for early stage NSCLC. A 3.4–5.9 mm GTV-to-PTV margin was required to account for this uncertainty alone</td>
</tr>
<tr>
<td>Dewas et al. (22)</td>
<td>Comparative study of a NSCLC case delineated by 120 residents before and after a radioanatomy lecture</td>
<td>120 trainee and 9 senior radiation oncologists. Single case</td>
<td>Training</td>
<td>Volume, degree of overlap, Kappa indices and dosimetry</td>
<td>The delineated volume of the trainees was larger but not significantly different from the expert consensus before and after the course. There was no difference in the overlap and kappa indices before and after course as the pre-course contours were already good. V20 for lung was higher in the residents’ group compared to the experts’ group (23.2% versus 36.5%)</td>
</tr>
<tr>
<td>Jameson et al. (12)</td>
<td>Evaluated the relationship between contouring variation, TCP and equivalent uniform dose (EUD) for 3D conformal NSCLC radiotherapy</td>
<td>7 NSCLC cases, 3 radiation oncologists</td>
<td>COM, volume and maximum mediolateral volume variation</td>
<td>All contouring metrics showed a correlation with TCP and EUD for NSCLC with the mediolateral volume dimension showing the highest correlation followed by the anteroposterior dimension, volume, CI, COM and superior-inferior dimension</td>
<td></td>
</tr>
</tbody>
</table>
Table 1 (continued)

<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>No of cases and observers</th>
<th>Intervention</th>
<th>Assessment metrics</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giraud et al.</td>
<td>Compare the delineation of the GTV of by radiologists and radiation</td>
<td>10 NSCLC cases, 9</td>
<td>Experience and radiologists input</td>
<td>Volume and CI</td>
<td>Radiologists tended to delineate smaller volumes than radiation oncologists and encountered fewer difficulties to delineate ‘difficult’ cases. Junior physicians, regardless of their speciality, also tended to delineate smaller and more homogeneous volumes than senior physicians, especially for ‘difficult’ cases.</td>
</tr>
<tr>
<td>(23)</td>
<td>oncologists with experience in the field in various centres</td>
<td>8 radiation oncologists</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Konert et al.</td>
<td>Assessed the impact of a standardized delineation protocol and training</td>
<td>11 radiation oncologists</td>
<td>Protocol and 2 training interventions</td>
<td>CI, local SD</td>
<td>Following the first training, overall conformity indices for 3 repetitive cases increased from 0.57 to 0.66. The local SD between observer and expert contours decreased from −0.40±0.03 to −0.01±0.33 cm. After further training, overall CIs for another 3 repetitive cases further increased from 0.64 to 0.80 (P=0.01). Mean local SD decreased from −0.34 to −0.05 cm (P=0.01). Findings suggest that multiple training interventions are required to reduce interobserver variation in NSCLC</td>
</tr>
<tr>
<td>(20)</td>
<td>in NSCLC in a multicentre setting</td>
<td>and 11 nuclear medicine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>physicians from different countries, 6 NSCLC cases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cui et al. (24)</td>
<td>Assessed the impact of a contouring atlas in reducing observer variation</td>
<td>12 institutes, 3 NSCLC</td>
<td>Protocol</td>
<td>Cl mean distance to reference contour and dosimetry</td>
<td>The PTV contouring consistency did not show improvement with an atlas, but considerable improvement was noted on OARs. Variations in PTV volume also affected dose distribution in surrounding tissues significantly</td>
</tr>
<tr>
<td></td>
<td>cases</td>
<td>cases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tsang et al.</td>
<td>Assessment of contour variability in target volumes and OARs in</td>
<td>2 benchmark stage 3 NSCLC</td>
<td>Peer review</td>
<td>Various conformity indexes</td>
<td>A statistically significant difference in trial protocol compliance for both GTV and OARs</td>
</tr>
<tr>
<td>(25)</td>
<td>lung cancer radiotherapy. Data from 2 UK lung cancer clinical trials</td>
<td>cases, 21 clinical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>oncologists</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groom et al.</td>
<td>Impact of protocol deviations in the CONVERT lung cancer trial on</td>
<td>94 SCLC cases, no of</td>
<td>Peer review</td>
<td>Survival</td>
<td>19.1% of the reviewed cases had unacceptable variation. PTV coverage was the most common violation. Patients with increasing number of protocol deviations had worse survival. High recruiting centres had the least deviations</td>
</tr>
<tr>
<td>(28)</td>
<td>survival</td>
<td>centres or reviewers not</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>specified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rooney et al.</td>
<td>Impact of peer review on lung cancer plans</td>
<td>121 lung cases, reviewed</td>
<td>Peer review</td>
<td>Qualitative</td>
<td>Twenty-one (17%) had a change in the GTV</td>
</tr>
<tr>
<td>(21)</td>
<td></td>
<td>by at least 2 oncologists</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lo et al. (19)</td>
<td>Impact of peer review on SBRT NSCLC plans</td>
<td>40 NSCLC PTVs, 2/3</td>
<td>Peer review</td>
<td>Qualitative dosimetry</td>
<td>43% of PTVs required minor changes, while 18% required major changes to avoid a violation of dose limits. A smaller proportion of changes recommended on peer review in the later versus earlier plans suggested an institutional learning curve. Peer review is recommended as a starting point to improve the consistency of SBRT PTVs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>radiation oncologists</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>reviewed each case</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PET, positron emission tomography; CT, computed tomography; GTV, gross tumour volume; PTV, planning target volume.
with the PET to obtain anatomical information. The use of FDG PET-CT in radiotherapy has been shown to reduce interobserver variation, especially when defining tumours surrounded by atelectasis (4,5). Furthermore, it facilitates the detection of both metastatic lymph nodes and distant metastasis hence improve staging accuracy (Figure 2) (41,42). However, FDG PET-CT also has a number of limitations. PET has a low spatial resolution and cannot detect very small nodules (<1 cm). False negatives and positives may occur in diabetic patients with high blood glucose levels at the time of scanning. Increased FDG uptake is observed in many non-neoplastic lesions, granulation tissue (e.g., wound healing), infections and other inflammatory processes, eventually resulting in false negatives and false positives (43).

**Role of MRI**

MRI in lung cancer radiotherapy is mainly used to delineate Pancoast tumours a particular type of lung tumour located in the upper lobes of the lungs that tend to spread into the chest wall and nerves. The use of MRI in lung cancer radiotherapy is currently limited by the lack of tissue density information required for dose calculations, the low proton density of the lung tissue and motion artefacts introduced by the long duration of the scan. New imaging sequences are currently being developed to facilitate the introduction of MRI in lung cancer radiotherapy triggered by the development of the MRI guided adaptive radiotherapy (44,45).

**Role of 4DCT**

4DCT can be used to account for the patient’s individual tumour motion. With this technique, the respiratory cycle is measured using devices such as an abdominal belt or infrared marker. A large number of CT images are then acquired and correlated with the breathing cycle. These images are then sorted during reconstruction into 8 to 10 equal respiratory bins with each bin representing either a specific phase (from 0 to 100%) or amplitude position of the respiratory cycle (37).

While 4DCT can be used to account for the patient’s individual respiratory motion, it also introduces new challenges. The 4DCT is typically not used to calculate the dose distribution, and therefore a 3DCT is reconstructed from this data. Furthermore, 4DCT imaging is prone to motion artefacts, particularly in patients with irregular breathing patterns (46). This occurs due to a mismatch between the data acquisition and respiratory phases. Several methods have been proposed to reduce these artefacts, including improvements in signal acquisition, gating, sorting and post-processing techniques. Visual and audio respiratory coaching can be used to regularise the breathing pattern and hence reduce these artefacts (47,48). However, the reported effectiveness of these techniques varies among patients. They are also time consuming and complex to implement clinically (47). Alternatively, the CT images can be acquired only at specific phases or amplitudes of the respiratory cycle (gating) and therefore, data from irregular breathing patterns is excluded. While gating reduces the number of artefacts, it comes at the cost of prolonging the scanning time.

Image sorting can be performed based on either the respiratory phase or amplitude. Amplitude sorting is less affected by outliers in the breathing cycle unless there are
gaps in the respiratory signal (49). Furthermore, sorting based on the movement of internal anatomy such as the diaphragm rather than external surrogates was found to reduce artefacts as it is more likely to represent the true internal anatomical movement (50). Alternatively, image post-processing techniques can be used to reduce artefacts (51).

Respiratory motion management

Several methods can be used to account for respiratory motion including; gating, tracking, internal target volume (ITV), mid-ventilation (Mid-V) and mid-position (Mid-P) (38).

Gating involves delivering the treatment only during specific amplitudes or phases within the respiratory cycle. Tracking involves continuously aligning and reshaping the radiation beam in real-time to account for variations in tumour position (37). However, while these techniques result in a very small PTV, they are complex and time consuming to implement clinically and therefore not widely used (37).

The ITV technique involves defining the CTV on either all or a selection of the 4DCT breathing phases. The ITV is then determined to be the envelope of motion of the CTV. When using this technique, the CTV has to be defined multiple times, making the delineation process time consuming (Figure 3). An alternative approach is to reconstruct the 4DCT image into a 3DCT that represents the full tumour motion (52). The GTV is delineated, and a margin is added to account for microscopic spread to form the ITV. Since the delineated GTV on the reconstructed 4DCT includes the tumour motion, it is referred to as the internal gross target volume (IGTV).

The maximum intensity projection (MIP) is one of the most commonly used reconstruction techniques (37). The MIP displays the highest density value encountered along the viewing ray for each pixel of volumetric data throughout the respiratory cycle (53,54). As such, these projections overlay all the CT phases and eventually represent the tumour position throughout the whole respiratory cycle. Delineations on the MIP generally show a good agreement with ITV generated from the 4DCT (54). However, the MIP reconstructed image is blurry making it difficult to distinguish the boundaries between tumour and tissue of equal tissue density such as blood vessels, diaphragm or mediastinum (53) potentially increasing the delineation uncertainty. Moreover, the ITV technique tends to overestimate the size of the PTV (35).

To overcome these issues, The Netherlands Cancer Institute (NKI), developed two new 4DCT image reconstruction techniques; Mid-Ventilation (Mid-V), and Mid-position (35,55). The Mid-V technique selects the frame of the 4D acquisition, where the tumour is closest to its mean time-weighted position. This frame can be selected visually or using rigid registration algorithms. The mid-position (Mid-P) technique uses deformable image registration to reconstruct every part of the anatomy in every frame to its average time-weighted mean position and...
then combines all frames.

The advantage of using the Mid-V and Mid-P over the MIP technique is that respiratory motion is decoupled from the GTV definition and is taken into account as a random error to be combined quadratically with other error sources and not linearly (55). These methods result in generally a smaller PTV (about 33% smaller), eventually sparing normal tissue (35,56). Peulen et al. (56), reported that the Mid-V technique was safely and easily implemented clinically at NKI with a 2-year local control rate of 98% for patients treated with SBRT (n=297). However, more clinical trials are required to assess the impact of using different motion management techniques on clinical outcomes.

Moreover, since the Mid-P reconstruction does not depend on a single frame, the motion artefacts are reduced, potentially facilitating the delineation process (51,55). However, this improvement comes at the cost of a somewhat reduced spatial resolution (Figure 3).

Mercieca et al. (2), compared the impact of using these three image reconstructions on interobserver variation in lung cancer. The overall difference in interobserver variation between the MIP, Mid-V and Mid-P was small. The benefit of using the Mid-V and Mid-P was more prominent in some specific tumour interfaces including the lung, chest wall and regions with a large tumour motion. An advantage of using the Mid-V and Mid-P technique is that it does not require the observer to review the delineations on the 4DCT making it easier to define tumour boundaries resulting in reduced interobserver variation in regions with large tumour motion. There was no benefit in using the Mid-P for lymph node delineation due to interpretational differences when incorporating diagnostic data in the delineation.

Role of 4D FDG PET-CT

A limitation of 3D FDG PET-CT is that respiratory motion can degrade the quality of the images in particular for small tumours located close to the diaphragm that tends to be more mobile. This can eventually result in mis-registration between the PET and the CT leading to interpretational difference when defining the GTV and inaccurate attenuation correction. Furthermore, the SUV measurements are blurred, eventually leading to an inaccurate segmentation of the GTV. An alternative approach is to acquire the images using deep inspiration breath-hold. However, a study by Nygård et al. (57) found that the deep inspiration breath-hold scans did not have a clinically relevant impact on the uptake metrics and did not improve the test-retest repeatability of FDG uptake metrics in lung cancer patients when compared with free-breathing scans.

To overcome these issues, the use of 4D FDG PET-CT has been proposed. This technique improves the diagnostic accuracy in particular for the detection of lymph nodes and small lung tumours (58), eventually reducing the interobserver variation in the definition of central lung target tumours (59). The benefit of using 4D FDG PET-CT for radiotherapy planning is also hampered by the long acquisition time eventually increase the chances of patient movement during the scan while also lowering the machine throughput. As a result, 4D FDG PET-CT is not commonly used in clinical practice. An alternative approach to 4D FDG PET-CT is the use of a motion-compensated Mid-P PET-CT scan as proposed by Kruis et al. (60). This technique could be used to reduce the blurring of the SUV signal improving the appearance of both tumour and boost volumes. However, this improvement was mainly noted for tumours with respiratory motion amplitude larger than 10 mm. Compared to a 3D PET scan, the lesions in the motion-compensated scans had higher SUV values and a smaller 50% SUVmax volumes, eventually altering the volume used in PET boost studies. Kruis et al. (60), also noted that an irregular breathing cycle could increase the number of artefacts.

Image-guided adaptive radiotherapy (ART)

With the integration of cone-beam computed tomography (CBCT) and MRI on the linear accelerator, it is now possible to identify intrathoracic anatomical changes prior to treatment and adapt the treatment accordingly if necessary. During adaptive radiotherapy, the planning CT is first registered with the localisation image, and any variations in the tumour and OAR shape and position are assessed. This is then followed by the application of an adaptive strategy. These strategies can be divided into two categories, ‘adapt-to-position’ (ATP) and ‘adapt-to-shape’ (ATS) (61). For ATP, rigid image registration is used to assess and account for variations in the isocentre position only (for e.g., by adjusting the couch position). On the other hand, for ATS strategies, deformable image registration is used to transfer anatomic contours and dose between the CBCT and planning CT images. This is used to assess dose deviations caused by the intrathoracic tumour and anatomical changes, providing guidance to when the
The MR linac allows for the acquisition of high-quality soft-tissue contrast images with functional information without using ionising radiation, allowing the oncologist to make daily treatment adaptation. Furthermore, MR-linacs now allow for cross-sectional beam-on imaging, making it possible to monitor tumour and organ at risk motion during treatment delivery without the need to use external surrogates or statistical respiratory models. This, together with the ability to acquire images in the sagittal and coronal plane results in higher image quality with less binning artefact, and more realistic motion estimation as the uncertainty from an imperfect external-internal surrogate is eliminated. Moreover, it also facilitates the use of gating and tumour tracking techniques.

Nevertheless, there are a number of challenges that need to be addressed for the clinical implementation of the MR-linac (67). Patient movement can increase as a result of the prolonged treatment time and the claustrophobic environment of the MRI. Workflows and imaging sequences need to be developed for radiotherapy purposes. Software also needs to be developed to account for the lack of tissue density information required for dose calculations and the time consuming step of contour propagation, editing and QA should be optimised, for instance by introducing simultaneous remote review. Ultimately, the clinical and cost-effectiveness of this technique must be proven with well designed clinical trials.

Auto-segmentation

Auto-segmentation involves converting an image into a collection of pixels that share the same characteristics such as intensity, shape or texture, thus facilitating the distinction between tumour and normal tissue. The advantage of incorporating FDG PET-CT into radiotherapy planning is that FDG tends to accumulate in cancer cells, thus facilitating tumour localisation and the development of auto-segmentation tools. On the other hand, CT based auto-segmentation tools are more complex to develop due to the poor contrast between tumour and adjacent soft tissue. Numerous semi-automatic and fully automated segmentation algorithms have been developed to facilitate this process, including algorithms based on thresholding, region growing, edge detection, statistical and machine learning algorithms (33,68). These algorithms tend to vary significantly in complexity, accuracy, degree of user intervention and availability.

Threshold algorithms are the simplest and most widely used (68). This technique defines the tumour by selecting all
the image voxels above a certain SUV intensity threshold, usually the SUVmax (hottest pixel) within a pre-defined region of interest.

The development of auto-segmentation algorithms for FDG PET-CT remains challenging as the SUV measurements are affected by physiological factors such as body mass and plasma glucose, biological characteristics of the tumour, the low resolution of the PET images and variations in scan parameters (69). Shepherd et al. (33), reviewed 30 different segmentation algorithms used in 13 different institutions. The findings of this study indicate that manual contouring is still the most accurate. However, simple threshold segmentation algorithms performed well compared to more complex algorithms. Mercieca et al. (70) compared such segmentations with pathology data and concluded that the threshold algorithm on the maximum SUV or SUVpeak performed equally well. The provision of auto-segmentation tools followed by manual editing has been found to reduce contouring time and, interobserver variation, and correlated well with pathology data (32,33,71).

Machine and deep learning methods are also showing promising results for OARs delineations as the shape of these organs are similar for most patients (68). However, these algorithms are highly dependent on the accuracy of predefined contours. Since the shape and texture of lung tumours can vary significantly between patients, it is more difficult to develop these algorithms to delineate lung tumours. Moreover, the lack of reliable gold standards makes it difficult to validate the accuracy of these algorithms.

The main barrier for clinical implementation of machine and deep learning algorithms is the availability of high-quality clinical contouring data for training. This data is often stored in secure servers across a number of hospitals that are not linked. Improvements in workflows and logistics would be required in order to securely link all the patient data required to develop contouring databases (72). An important question is whether to include all the oncologists’ contours in the training database. Training of algorithms can be supervised whereby the algorithm learns from labelled datasets (i.e., good contours) or unsupervised whereby the algorithms tries to make sense of unlabelled data (i.e., not providing contours that have been peer-reviewed) by independently extracting features and patterns from the images.

Training of algorithms using non-reviewed physician contours can introduce a bias by the particular physician’s medical training, experience, goals, or misconceptions, eventually leading to an inaccurate segmentation (23,73). An extensive database is required to reduce the effects of major outliers, but this will also increase the computation time. This problem can be resolved through the use of supervised training data whereby only the contours that have been delineated using a specific protocol and peer-reviewed by experts are included in the database (72). Alternatively, only the contours from patients that had acceptable local control rates and toxicity could be used to develop the training database. The latter would automatically exclude cases whereby the tumour recurred as a result of a geographical miss or cases that had unacceptable toxicities due to an excessive inclusion of normal tissue. The limitation of this approach is that it still requires a manual intervention to label the data making it time consuming to develop the algorithm. Also, cases where the PTV coverage is compromised due to proximity or OARs may need to be excluded.

**Delineation protocols**

Numerous consensus delineation guidelines (38,74,75) have been published by professional bodies providing detailed information to facilitate the interpretation of clinical information, diagnostic images and biopsies necessary to define the GTVp and GTVln. These protocols also provide information on the process that should be followed to define the GTV such as setting the optimal window/level on CT based on the tumour location and provide guidelines on how to incorporate diagnostic images to facilitate the definition of the GTV.

For the definition of the lymph node GTV, the European Society of Radiotherapy and Oncology together with the Advisory Committee in Radiation Oncology Practice (ESTRO-ACROP) (38,75) proposed an algorithm that could be used to identify the lymph nodes that should be included in the GTV based on the diagnostic CT, FDG PET-CT and biopsy information.

Elective lymph node irradiation is no longer recommended as this procedure leads to increased toxicity while it also limits dose escalation (38,76,77). The ESTRO-ACROP guidelines identified two acceptable methods that can be used to determine the boundary for the GTVln (38). The GTVln can be defined by either defining the positive lymph node with an 8mm expansion to account for microscopic spread or by defining the entire lymph node station.

Both lymph node delineation methods have been used in large multicentre clinical trials without unacceptable out-field mediastinal recurrence rates (38). However, the
definition of the lymph node station results in a larger GTV when compared with defining only the involved node, potentially increasing the toxicity for the patient. Anatomical atlases illustrating the location of specific lymph node stations and how to define the GTV for specific cases have also been developed (74,78,79).

Studies have shown that guidelines can reduce the interobserver variability in the definition of the GTV and OARs in lung cancer (53,61). However, significant interobserver variation remains even amongst experts (24), and therefore training is essential to ensure the correct interpretation and application of these guidelines in routine clinical practice. It is essential to acknowledge that the use of different protocols between different centres may also result in variations when defining the GTV, thus highlighting the need to harmonise protocols. Moreover, protocols may not always provide guidance for all clinical scenarios, and hence discussion of difficult cases in a multidisciplinary team is recommended.

**Training**

Vinod et al. (80) evaluated the impact of several training programmes on reducing interobserver variation. The impact of training varied across studies as the delivery method as well as the target audience varied. Larger group didactic lectures did not have a significant impact on interobserver variation while courses that had a practical component and provided individual feedback were reported to be more effective in reducing interobserver variation (80). An international delineation study conducted by Konert et al. (20), showed that more than one training intervention might be required in order to have a significant impact in reducing interobserver variation when delineating the GTV in lung cancer and eventually lead to a change to clinical practice. Mercieca et al. (81) compared individually made delineations to delineation made by group consensus, and showed that the latter had more improvement than training, illustrating the need for collaboration and peer review.

**Peer review**

Training and clear protocols are important to improve consistency in contouring. However, these may not necessarily lead to a change in routine clinical practice or eliminate human errors (20).

Furthermore, the task of defining the GTV requires a range of expertise from radiologists, physicist, radiographers and oncologists. Numerous studies have shown that peer review by a second oncologist or within a multidisciplinary team can reduce the interobserver variation in target volume definition and facilitate the identification of unacceptable gross errors (19,82,83). Studies have shown that when peer review is introduced, unacceptable errors are identified in about 17% of target volumes (19,21,84). These errors have been linked with worse survival in clinical trials (18,26).

As a result, several professional bodies have now issued guidelines to establish minimum standards for peer review as part of the Radiotherapy department’s quality assurance processes (10,38,85,86). Although these guidelines indicate that peer review is essential, it is not a common practice in many radiotherapy centres (10,38). Various barriers exist for the routine implementation of peer review in clinical practice including allocated time to review contours, shortage of staff, availability of radiology services, delays to start treatment, availability of workstations and appropriate software (10,87,88). Workflows and cases reviewed also varied widely across centres (87). Outcomes from peer-review should be clearly documented, and the data generated used to improve delineation protocols, training and the accuracy of autosegmentation tools.

Artificial intelligence could also be used to develop computer-assisted peer review software. Hui et al. (89) developed an algorithm that could be used to evaluate OARs in the thoracic region. In this study, the researchers simulated common delineation errors, including boundary deviations, missing slices, incorrect labelling, and craniocaudal over-extension for OARs in the thoracic region. The algorithm was able to detect 37% of the minor and 85% of the major errors. The reason for lack of precision in detecting minor errors was attributed to the fact that these errors were inconsistently judged by the reviewers. The use of this tool also improved the reviewers’ error detection sensitivity from 61% to 68% for minor errors and from 78% to 87% for major error. The findings of these studies suggest that such tools could be used to assist the oncologists in reviewing contours, but they should not be used to replace human judgement. Over-reliance on the system might end up becoming counterproductive and actually reduce the ability of the reviewer to identify errors. Further research is required to develop similar algorithms for lung tumours.

**Conclusions**

The findings of this review indicate that to date, it is
still not possible to eliminate interobserver variation in the definition of GTV. Positron Emission Tomography (PET-CT) has an important role in improving the staging accuracy and the definition of the tumour. Various autosegmentation tools have also been proposed to fully or partially automate the delineation process. However, their development is currently hindered by the unavailability of absolute gold standards that can be used to validate these algorithms as well as the wide morphological and shape variations of lung tumours. Hence, manual delineation is still considered to be the gold standard. Nevertheless, autosegmented contours can provide a good starting point, eventually reducing the delineation time and interobserver variation. Improvements in image quality can also reduce the delineation uncertainty in some cases. However, the main factor leading to interobserver variation is image interpretation differences between clinicians. Therefore, protocols, training and peer review of contours are essential to address this challenge.

Acknowledgments

Professor Marcel van Herk is supported by NIHR Manchester Biomedical Research Centre.

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editors (Jacek Jassem and Rafal Dziadziuszko) for the series “Radiotherapy in thoracic malignancies” published in Translational Lung Cancer Research. The article was sent for external peer review organized by the Guest Editors and the editorial office.

Peer Review File: Available at http://dx.doi.org/10.21037/tlcr-20-627

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/tlcr-20-627). The series “Radiotherapy in thoracic malignancies” was commissioned by the editorial office without any funding or sponsorship. MvH reports being supported by the NIHR Manchester Biomedical Research Centre, outside the submitted work. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

anatomical changes in lung cancer patients during the course of radiotherapy. Radiother Oncol 2014;113:392-7.
32. van Baardwijk A, Bosmans G, Boersma L, et al. PET-CT-Based Auto-Contouring in Non-Small-Cell Lung Cancer Correlates With Pathology and Reduces Interobserver Variability in the Delineation of the Primary Tumor and Involved Nodal Volumes. Int J Radiat Oncol Biol Phys


### Table S1 Summary of metrics used to assess interobserver variation. The accuracy of the metrics depends on their ability to assess variations in volume, shape, location, margins required to account for interobserver variations and ultimately, treatment outcomes. Some metrics are easily exported from the treatment planning software and are more widely used (12,16,19,90-94).

<table>
<thead>
<tr>
<th>Metric type</th>
<th>Method</th>
<th>Perfect value</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple volume measurements</td>
<td>Compares delineated volume with a reference contour; e.g., volume A, volume B</td>
<td>1</td>
<td>Easily exported from planning software; correlates well with NTCP; provides information on over or under outlining</td>
<td>Contours can have the same volume but different shape and location; cannot be used to calculate margins</td>
</tr>
<tr>
<td>Centre of mass (COM)</td>
<td>Calculates the difference in the centre coordinates (x,y,z) of different contours</td>
<td>0</td>
<td>Easily exported from planning software</td>
<td>Contours can have the same COM but different shape and volume; cannot be used to calculate margins</td>
</tr>
<tr>
<td>Overlap metrics</td>
<td>The overlap between observer contour (A) and reference contour (B) can be calculated using: Jaccard = ( \frac{A \cap B}{A \cup B} ) or Dice = ( \frac{2(A \cap B)}{A + B} ). The general conformity index (GCI) can be used to calculate the overlap between many pairs of observers. No reference contour required. GCI = ( \sum_{\text{pairs}}</td>
<td>A \cap B</td>
<td>)</td>
<td>1</td>
</tr>
<tr>
<td>Over or under outlining</td>
<td>General miss index (GMI): calculates the amount of under outlining. GMI = ( \frac{B - (A \cap B)}{B} ). Discordance index (DI) calculates the amount of over outlining DI = ( \frac{A - (A \cap B)}{A} )</td>
<td>0</td>
<td>Easily exported from planning software; provides information about over and under outlining</td>
<td>Provides no information on shape and volume variations; site and case dependent; cannot be used to calculate margins</td>
</tr>
<tr>
<td>Shape surface metrics</td>
<td>Local SD measures the perpendicular distance (d') from the reference contour (B) to the observers’ contours (A). The standard deviation in the distance between all observers is calculated at each point, and then the average is calculated using the root mean square. Other similar algorithms include Mean distance to agreement, ComGrad distance and bidirectional local distance [96]. These vary in the method used to measure the distance from reference contour</td>
<td>0</td>
<td>Widely used in the literature; provides information about shape and location; can be used to estimate margins</td>
<td>Requires specialised software; no information about volume; overall score influenced by outliers; accuracy for irregularly shaped contours depends on the algorithm</td>
</tr>
<tr>
<td>3D shape surface method</td>
<td>Distribution of local SD over tumour surface area plotted as a histogram or 3D surface map</td>
<td>0</td>
<td>Provides information about the percentage tumour surface area affected by large interobserver variation</td>
<td>Requires specialised software; no information about volume; accuracy for irregularly shaped contours depends on the algorithm</td>
</tr>
<tr>
<td>Dosimetric assessment</td>
<td>Involves applying the dose distribution from a reference expert plan to each of the observers’ contours with or without plan optimisation according to the observers’ contours. Or plan on each observer and evaluate coverage of reference contours. Dosimetric deviations from the reference plan are evaluated.</td>
<td>Within acceptable oars and PTV tolerances according to clinical guidelines</td>
<td>Provides a correlation with clinical outcomes</td>
<td>Time consuming; not widely used in interobserver studies; depends on the ability of the planner to optimise the plan</td>
</tr>
<tr>
<td>Semi-quantitative analysis</td>
<td>An algorithm creates a percentage score by identifying the voxels falling outside or missing from the reference contour. A penalty can be applied by the teacher based on the distance of the voxels from the reference contour and severity of error [94,95]</td>
<td>100%</td>
<td>Provides a correlation with clinical outcomes</td>
<td>Limited research on the penalties that should be applied; specialised software required</td>
</tr>
<tr>
<td>Expert qualitative visual analysis</td>
<td>Expert’s visually classify contours as acceptable or unacceptable based on the clinical impact of the error [18] there is little research documenting its impact in the setting of stereotactic body radiation therapy (SBRT)</td>
<td>Accept</td>
<td>Provides a correlation with clinical outcomes; no specialised software required; facilitates the identification of factors leading to interobserver variation</td>
<td>Subjective; time consuming; no quantitative measurement provided</td>
</tr>
</tbody>
</table>

Local SD, local standard deviation; TCP, tumour control probability; NTCP, normal tissue complication probability.
References