Technical Innovations & Patient Support in Radiation Oncology 10 (2019) 8–12

Contents lists available at ScienceDirect

Technical Innovations & Patient Support in Radiation Oncology

journal homepage: www.elsevier.com/locate/tipsro

Research article

Assessment of contour variability in target volumes and organs at risk in lung cancer radiotherapy

Yatman Tsang [a,*], Peter Hoskin [a,f], Emiliano Spezi [b,g], David Landau [c], Jason Lester [d], Elizabeth Miles [a], John Conibear [e]

[a] NIHR Radiotherapy Trials Quality Assurance Group, Mount Vernon Cancer Centre, Rickmansworth Rd, Northwood HA6 2RN, UK
[b] Dept. of Medical Physics, Velindre Cancer Centre, Cardiff, UK
[c] Dept. of Clinical Oncology, Guy's and St. Thomas' Hospital, London SE1 7EH, UK
[d] Dept. of Clinical Oncology, Velindre Cancer Centre, Velindre Road, Cardiff CF14 2TL, UK
[e] Dept. of Clinical Oncology, St. Bartholomew's Hospital, West Smithfield, London EC1A 7BE, UK
[f] Division of Cancer Sciences, Faculty of Biology Medicine and Health, University of Manchester, Manchester, UK
[g] School of Engineering, Cardiff University, UK

ARTICLE INFO

Article history:
Received 5 April 2019
Received in revised form 19 May 2019
Accepted 21 May 2019

Keywords:
Contouring variability
Lung radiotherapy
Trials quality assurance

ABSTRACT

Aims: This study aimed to examine whether any significant differences existed in trial protocol compliance in target volumes (TV) and organs at risk (OARs) contouring amongst clinical oncologists specialised in lung cancer radiotherapy.

Materials/methods: Two lung radiotherapy trials that require all prospective investigators to submit pre-trial outlining quality assurance (QA) benchmark cases were selected. The contours from the benchmark cases were compared against a set of reference contours which were defined by the trial management group (TMG). In order to quantify the degree of variation in TV and OARs contouring, the matching index (MI), Dice coefficient (DICE), Jaccard index (JI), Van't Riet Index and geographical miss index (GMI) were calculated.

Results: A total of 198 structures contoured by 21 clinicians were collected from the outlining benchmark cases. There were 40 clinical target volumes (CTV), 32 spinal cord, 36 oesophagus, 36 heart and 54 lungs volumes included in the study. Analysis of the pre-trial benchmark cases revealed statistically significant differences ($p \leq 0.05$) in trial protocol compliances between clinical oncologists' target volume and organs at risk contours. Our results demonstrated that the lung contours had the highest level of conformity, followed by heart, CTV, spinal cord and oesophagus respectively.

Conclusions: This study showed that there was a statistically significant difference in trial protocol compliance for lung clinical oncologists' TV and OARs contouring within the pre-trial QA benchmark cases. Trial protocol compliances of TV and OARs delineation can be identified through assessing outlining QA benchmark cases.

Background

In the last decade, there have been progressive developments in the technology used for planning and delivery of lung radiotherapy (RT). Advanced RT techniques such as intensity modulated RT (IMRT), volumetric arc therapy (VMAT) and image guided radiotherapy (IGRT) facilitate a change from a 'one-size-fits-all' approach to more personalised radiation treatments [1,2]. The concept of isotoxic lung RT which allows the radiation dose prescription to the tumour to be tailored based upon normal tissue constraints was recently explored in the Isotoxic Dose Escalation and Acceleration in Lung Cancer ChemoRadiotherapy (IDEAL-CRT) and ISoToxic Accelerated RadioTherapy in locally advanced non-small cell lung cancer (I-START) trials [3,4].

This individualised isotoxic treatment approach depends heavily on a clinician's own interpretation of radiological cross-sectional anatomy and requires clinicians to differentiate between 'normal' and 'abnormal' body tissues. This makes the process of both target volume (TV) and organs at risk (OAR) delineations highly observer dependent and at significant risk of

* Corresponding author at: Radiotherapy Department, Mount Vernon Cancer Centre, Northwood, Middlesex HA6 2RN, UK.
E-mail address: yatmantsang@nhs.net (Y. Tsang).

https://doi.org/10.1016/j.tipsro.2019.05.001

inter-observer variation [5]. In lung external beam RT, most published research has focused on the extent of inter-observer variation concerning the TV delineation, and much less work has been conducted on inter-observer variability involving OAR [6–9]. Against this background, this study was carried out to quantify the degree of variation in TV and OARs amongst clinical oncologists participating in the IDEAR-CRT and I-START trials.

Materials and methods

IDEAL-CRT and I-START trials quality assurance (QA)

The IDEAL-CRT trial was a phase I/II multi-centre trial evaluating the toxicity, feasibility and potential clinical effectiveness of isotoxic, dose-escalated RT with concurrent chemoradiotherapy in patients with stage II or stage III non-small cell lung cancer (NSCLC) [3]. The I-START trial was a phase II multi-centre trial evaluating the toxicity, feasibility and potential clinical effectiveness of isotoxic, accelerated and dose-escalated RT sequential to chemotherapy in patients with more locally advanced stage II to stage IIIB NSCLC [4].

In the UK, the National Cancer Research Institute (NCRI) has established the National Radiotherapy Trials QA group (RTTQA) to co-ordinate clinical trial QA work. As part of the trial QA programme for both the IDEAL-CRT and I-START trials, all participating centres were asked to complete two outlining QA benchmark exercise prior to site activation. The same outlining QA benchmark exercises were utilised by both trials. The QA exercises were locally advanced, stage III lung cancers with tumours located centrally within the chest. Background case histories, diagnostic imaging scans (PET-CT) with the reports were provided along with the planning CT scan. The QA instructions requested clinicians to create structures including the clinical target volume (CTV), spinal cord, oesophagus, heart and lungs.

Contouring variations in TV and OARs

All completed outlining QA benchmark cases were submitted to the RTTQA group for review. These were first reviewed by the trial QA team for trial protocol outlining compliance. The TV and OARs contours from each clinician's QA exercise submission were compared against a set of reference contours which was defined by the trial management group (TMG) (Fig. 1). This QA approach, whereby a TMG reference contour set is utilised, has already been validated by the UK's RTTQA group [10]. Once the trial QA team had completed their review, a detailed feedback report was created and sent back to the submitting clinician.

For this analysis, and to allow the quantification of the degree of variation that exists amongst clinicians participating in the IDEAL-CRT and iSTART trials, the TV and OAR contours from clinician's first submissions were collected from both QA benchmark cases and analysed against the TMG reference contour set. In total, one simple volume measurement index: matching index (MI) and four common conformity indices (CI) including the Dice coefficient (DICE), Jaccard index (JCI), Van't Riet Index (RIET) and geographical miss index (GMI) were calculated to allow assessment of under and over outlining. These indices were selected on the basis that they offer useful variations in the assessment of clinician contours [11–16]. MI compares two simple contours incorporating the differences in volumes: 100% implies a perfect match between two contours. In this analysis, 1-GMI was calculated and used instead of GMI to permit the use of a single scale of conformity for all four CI whereby 1.0 represented perfect contour conformity and 0 represented no conformity.

Analysis

For each CI, Kruskai-Wallis ANOVA was performed to detect whether any statistically significant differences in trial protocol compliances between TV and OAR, volumes existed. This was then followed by a Bonferroni-type multiple comparison to establish the hierarchy. A Bonferroni correction was utilised to counteract the problem of multiple comparisons.

Results

A total 198 structures were contoured in 40 case submissions and they were used in the analysis. There were 40 CTV, 32 spinal cord, 36 oesophagus, 36 heart and 54 lungs (27 left and 27 right lungs) volumes included in the analysis. As suggested in Fig. 2,

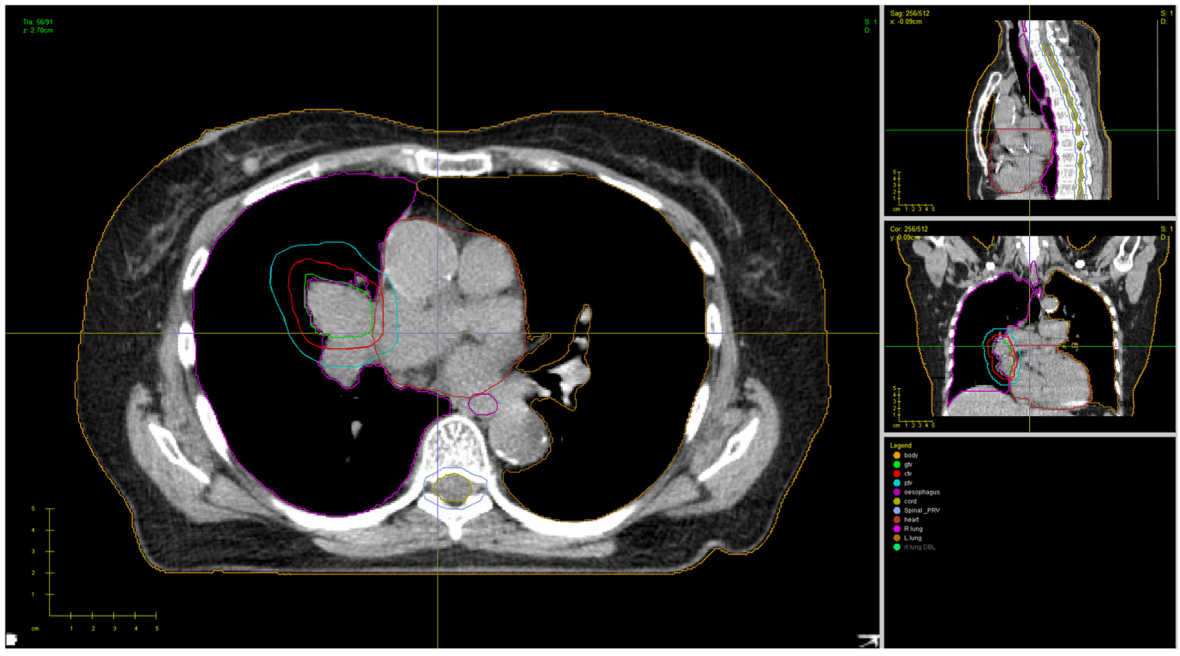

Fig. 1. Displaying an example CT Slice from the QA case with TMG Reference Contours.

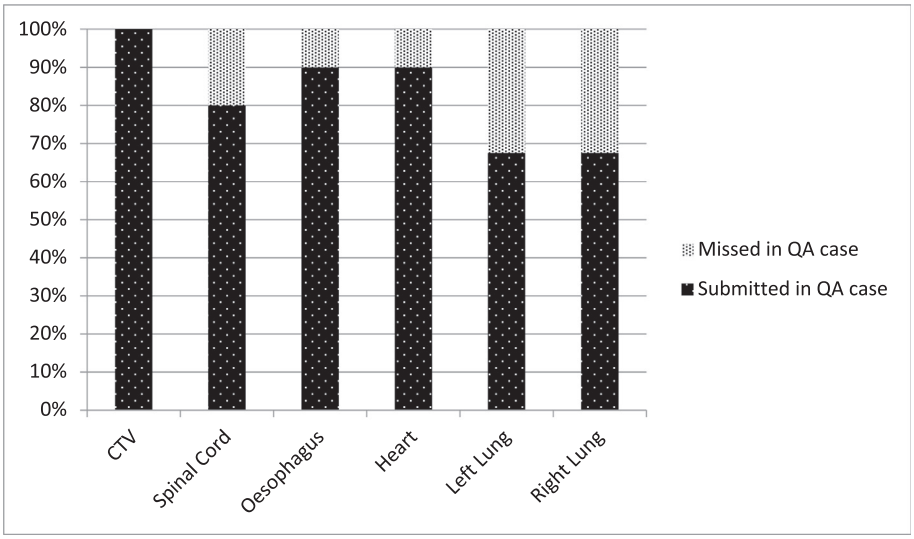

Fig. 2. Displays the number of submissions per structure in the QA cases.

there was a 100% submission rate for the CTV contours while there were 32% of the QA cases without the lung volumes delineated.

Descriptive statistics were summarised for all indices of each structure in Table 1. In terms of the CTV contouring, it showed an excellent mean MI to the TMG reference contour set of 99.34%. The mean DICE and JCI values of CTV were 0.84 (95%CI 0.81 – 0.87) and 0.73 (95%CI 0.69–0.77) respectively. Reviewing the standard deviations stated in Table 1, the highest variations were found in spinal cord and oesophagus in all five indices. Statistically significant differences in trial protocol compliances were detected for all five indices when comparing between the structures. This warranted for the pairwise comparisons to establish the hierarchy of trial protocol compliance of each structure.

As Table 2 demonstrates, significant differences were found in all Bonferroni-type multiple comparisons, apart from Heart vs Lungs in DICE, and CTV vs Spinal Cord, CTV vs Heart and Heart vs Lungs in 1-GMI.

The lung contours had the highest level of conformity to the TMG reference contour set, followed by heart, CTV, spinal cord and oesophagus in all indices.

Discussion

This study has examined whether any significant differences exist in trial protocol compliance in specialist lung cancer clinical oncologist's TV and OARs contouring through the analysis of two

Table 1
Summarises the descriptive statistics of MI, DICE, JACCARD, RIET and 1-GMI for CTV, spinal cord, oesophagus, heart and lungs contours.

		Mean index value	Standard deviation	95% Confidence interval for mean		Kruskai-Wallis ANOVA
				Lower bound	Upper bound	
MI	CTV	99.34%	0.4%	99.22%	99.45%	P < 0.05
	Spinal Cord	97.16%	1.1%	96.76%	97.57%	
	Oesophagus	96.03%	0.6%	95.84%	96.22%	
	Heart	99.87%	0.1%	99.86%	99.88%	
	Lungs	99.95%	0.1%	99.94%	99.96%	
DICE	CTV	0.84	0.10	0.81	0.87	P < 0.05
	Spinal Cord	0.74	0.12	0.69	0.78	
	Oesophagus	0.64	0.13	0.60	0.69	
	Heart	0.92	0.04	0.91	0.93	
	Lungs	0.97	0.01	0.97	0.98	
JCI	CTV	0.73	0.13	0.69	0.77	P < 0.05
	Spinal Cord	0.60	0.14	0.54	0.65	
	Oesophagus	0.48	0.14	0.44	0.53	
	Heart	0.86	0.06	0.84	0.88	
	Lungs	0.95	0.02	0.94	0.95	
RIET	CTV	0.72	0.14	0.68	0.77	P < 0.05
	Spinal Cord	0.58	0.15	0.52	0.63	
	Oesophagus	0.44	0.15	0.39	0.49	
	Heart	0.85	0.06	0.83	0.87	
	Lungs	0.95	0.02	0.94	0.95	
1-GMI	CTV	0.86	0.12	0.82	0.90	P < 0.05
	Spinal Cord	0.82	0.17	0.76	0.88	
	Oesophagus	0.67	0.18	0.61	0.73	
	Heart	0.91	0.06	0.89	0.93	
	Lungs	0.96	0.03	0.95	0.97	

Table 2
Summaries the results of the Bonferroni-type multiple comparisons between contours for the MI, DICE, JACCARD, RIET and 1-GMI. The cells with non-significant P values (>0.05) have been highlighted.

Comparisons	MI	DICE	JCI	RIET	1-GMI
CTV vs Spinal Cord	P<0.05	P<0.05	P<0.05	P<0.05	P=1.00
CTV vs Oesophagus	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05
CTV vs Heart	P<0.05	P<0.05	P<0.05	P<0.05	P=0.61
CTV vs Lungs	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05
Spinal Cord vs Oesophagus	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05
Spinal Cord vs Heart	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05
Spinal Cord vs Lungs	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05
Oesophagus vs Heart	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05
Oesophagus vs Lungs	P<0.05	P<0.05	P<0.05	P<0.05	P<0.05
Heart vs Lungs	P=1.00	P=0.08	P<0.05	P<0.05	P=0.57

pre-trial benchmark cases used as part of the radiotherapy QA for two UK national lung cancer trials, IDEAL-CRT and I-START. Our study found that high concordances of TV and OARs to TMG reference contour set in these two lung cancer trials. There were statistically significant differences in trial protocol compliance in TV and each OAR contouring for both volume measurement index and CIs analysed. It suggested that lung contours had the highest level of conformity to the TMG reference contour set, followed by heart, CTV, spinal cord and oesophagus for all four indices analysed.

As demonstrated in Table 1, the mean value of JCI for the CTV was 0.73 (CI 0.69–0.77). This was comparable to the findings reported by Grills et al. and Konert et al. [17,18]. The reason of selecting CTV instead of gross tumour volume (GTV) for QA analysis in this study is to take account of the delineation accuracy of how well the clinicians revised the CTV according to surrounding anatomical boundaries after growing it from the GTV.

When defining the target volume, clinicians often try to avoid missing out macroscopic tumour or areas considered to be at risk of harbouring microscopic disease from their volume. Hence, it would seem more likely that clinicians tend to over outline rather than under outline their target volumes [6]. This concept might explain why the CTV had such low levels of under outlining when assessed using the DICE (mean value 0.84) and 1-GMI (mean value 0.86). Assessment of the RIET for the target group (mean value 0.72) was lower than that of the DICE supporting this theory.

Several studies have shown the benefits of the addition of complimentary imaging modalities in improving TV delineation accuracy [7,9,17,18]. The outlining QA benchmark exercise included background case histories and diagnostic PET-CT imaging scans with reports to aid the clinicians with their TV delineation. However, such detailed guidance was not implemented to aid clinicians in their OAR contouring. This could be explained on the basis that the target volume contains the tumour and is therefore deemed by the clinician to be the most important volume to be defined. This is supported by Fig. 1 indicating that there was a 100% submission rate of CTV volumes and only 68% of the QA cases were with lung volumes submitted.

Sub-optimal TV definition negatively impacts on the chances of cancer cure. Hence it is therefore likely for clinicians to put more focus on implementing different contouring strategies to improve the accuracy of TV delineation.

As illustrated in Table 1, the lungs had the highest levels of conformity whilst the oesophagus had the lowest in the context of the IDEAL and I-START trials QA. The excellent conformity in the lungs' contour could be explained by the fact that these were largely auto-contoured as per the gold standard contours using the auto-segmentation function of the radiotherapy treatment planning system (TPS) with minimal clinician input. Therefore, the degree of contouring variation witnessed would seem to be dramatically minimised when the human interaction is largely removed from the delineation process. In line with the published data by La Macchia et al., our findings (as shown in the standard deviations of Table 1) validate the ability of the auto-segmentation algorithm to accurately outline normal lung contours [19]. Besides, it is noted that the lung volumes are generally much larger in comparison to OARs such as spinal cord and oesophagus. This may cause a higher chance of geometrically overlap in lungs contour, resulting in high conformity to TMG reference set and low variations between cases.

Regarding the oesophagus, the analysis on the indices suggested that both under-outlining (not including the actual organ) and over-outlining (including structures which are not the organ) seemed to be present over the anatomical course of the organ. Based on the oesophagus' anatomical location and its proximity to other central mediastinal structures, it can be difficult for clinicians to interpret the organ's precise boundaries and anatomical course. This is particularly problematic in studies of isotoxic dose escalation where the dose to the oesophagus is a critical part of the dose escalation algorithm. It is unclear what complementary imaging techniques could address this problem.

The complex nature of advanced radiotherapy can introduce uncertainty in the reproducibility and accuracy of treatment. This may be amplified in a multi-centre clinical trial setting without a comprehensive QA programme, especially when the advanced radiotherapy techniques required are unfamiliar to participating centres [20–22]. Our study has demonstrated the statistically significant variation that exists in trial protocol compliances of TV and OAR contouring during the pre-trial QA period for two UK lung cancer radiotherapy trials and therefore highlights the importance of performing outlining QA for the purposed of clinician feedback to help minimise contouring variation. Misinterpretations of per-protocol TV and OARs delineation can be identified through assessing outlining QA benchmark cases and potential solutions developed through discussion between the RTTQA team and TMG. Such collaborations between the RTTQA team and TMG has already been demonstrated to be an invaluable resource for providing UK centres with a strong, cooperative network and safe environment for implementing new advanced radiotherapy techniques [10,20,21].

Recently it has been recommended by the Royal College of Radiologists in the UK, that radiotherapy departments should have

processes in place to facilitate regular systematic peer reviews of TV and OAR [23]. The QA benchmark exercises in this study can be utilised as a common test case for departments to include in their own annual audit programme. Our analysis on TV and OAR contouring in two UK lung radiotherapy trials provide valuable insight on the current standard for clinical oncologists contouring and permits the ability for clinicians and radiotherapy departments to be benchmarked against each other.

It is acknowledged that the main limitation of our study is that the results only represent a snapshot of UK lung cancer focused clinical oncologists' contouring conformity during the pre-accrual trial QA benchmark period. In order to ensure that the level of consistency in TV and OARs contouring is maintained, it is suggested that all participating clinicians should repeat the benchmark QA case exercise at pre-defined time points whilst the trial remains open to recruitment i.e. every 12 months. Analysis of these subsequent benchmark cases could be used to detect whether inter-observer variation in TV and OAR contouring deteriorates beyond the completion of the initial pre-trial benchmark period.

Conclusion

Our findings suggest that high concordances of TV and OARs to TMG reference contour set were found in the pre-trial QA of IDEAL-CRT and I-START. There are statistically significant differences in trial protocol compliances of TV and OAR contours produced by specialists within the pre-trial QA benchmark cases. With the introduction of an individualised isotoxic RT approach in a clinical trial, it is important to ensure that both TV and OARs are contoured consistently and accurately according to a comprehensive and clearly defined trial protocol because inter-observer variation in OAR delineation could have a significant impact on both the final prescription dose and the dose received by the contoured OAR within the isotoxic radiotherapy setting.

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

Acknowledgements

Both IDEAL-CRT trial (C13530/A10424) and the I-START trial (CRUK/10/005) were funded by Cancer Research UK. The RTTQA group is funded by the National Institute for Health Research.

The authors would like to thank all centres particpating in the two trials and all clinical oncologists that undertook the QA cases.

References

[1] Van Baardwijk A, Bosmans G, Boersma L, Wanders S, Dekker A, Dingemans A, et al. Individualized radical radiotherapy of non–small-cell lung cancer based on normal tissue dose constraints: a feasibility study. Int J Radiat Oncol Biol Phys 2008;71(5):1394–401.

[2] Van Baardwijk A, Wanders S, Boersma L, Borger J, Öllers M, Dingemans A, et al. Mature results of an individualized radiation dose prescription study based on normal tissue constraints in stages I to III non–small-cell lung cancer. J Clin Oncol 2010;28(8):1380–6.

[3] Landau D, Hughes L, Baker A, Bates A, Bayne M, Counsell N, et al. IDEAL-CRT: a phase 1/2 trial of isotoxic dose-escalated radiation therapy and concurrent chemotherapy in patients with stage II/III non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2016;95(5):1367–77.

[4] Lester J, Nixon L, Mayles P, Mayles H, Tsang Y, Ionescu A, et al. 156 The I-START trial: ISoToxic Accelerated RadioTherapy in locally advanced non-small cell lung cancer. Lung Cancer 2012;75:S51.

[5] Roques TW. Patient selection and radiotherapy volume definition – can we improve the weakest links in the treatment chain? Clin Oncol (R Coll Radiol) 2014;26(6):353–5.

[6] 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.

[7] Steenbakkers R, Duppen J, Fitton I, Deurloo K, Zijp L, Comans E, et al. Reduction of observer variation using matched CT-PET for lung cancer delineation: a three-dimensional analysis. Int J Radiat Oncol Biol Phys 2006;64(2):435–48.

[8] Wills L, Hudson E, Hanna L, Williams L, Macbeth F, Lester J. Variability in lung cancer target volume definition between clinicians: a comparative study. Lung Cancer 2008;60:S30.

[9] Hanna G, McAleese J, Carson K, Stewart D, Cosgrove V, Eakin R, et al. 18F-FDG PET-CT based target volume definition in non-small cell lung cancer reduces inter-observer variation in already PET-CT staged patients. Lung Cancer 2009;63:S34–5.

[10] Gwynne S, Spezi E, Sebag-Montefiore D, Mukherjee S, Miles E, Conibear J, et al. Improving radiotherapy quality assurance in clinical trials: assessment of target volume delineation of the pre-accrual benchmark case. Brit J Radiol 2013;86(1024):20120398.

[11] Gwynne S, Spezi E, Wills L, Nixon L, Hurt C, Joseph G, et al. Toward semi-automated assessment of target volume delineation in radiotherapy trials: the SCOPE 1 pretrial test case. Int J Radiat Oncol Biol Phys 2012;84(4):1037–42. https://doi.org/10.1016/j.ijrobp.2012.01.094.

[12] Hanna GG, Hounsell AR, O'Sullivan JM. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. Clin Oncol 2010;22(7):515–25.

[13] Riet A, Mak A, Moerland M, Elders L, van der Zee W. conformation number to quantify the degree of conformality in brachytherapy and external beam irradiation: application to the prostate. Int J Radiat Oncol Biol Phys 1997;37(3):731–6.

[14] Kouwenhoven E, Giezen M, Struikmans H. Measuring the similarity of target volume delineations independent of the number of observers. Phys Med Biol 2009;54(9):2863–73.

[15] Kepka L, Bujko K, Garmol D, Palucki J, Zolciak-Siwinska A, Guzel-Szczepiorkowska Z, et al. Delineation variation of lymph node stations for treatment planning in lung cancer radiotherapy. Radiother Oncol 2007;85(3):450–5.

[16] Muijs C, Schreurs L, Busz D, Beukema J, van der Borden A, Pruim J, et al. Consequences of additional use of PET information for target volume delineation and radiotherapy dose distribution for esophageal cancer. Radiother Oncol 2009;93(3):447–53.

[17] Grills I, Yan D, Black Q, Wong C, Martinez A. Kestin L. Clinical implications of defining the gross tumor volume with combination of CT and 18FDG-positron emission tomography in non–small-cell lung cancer. Int J Radiat Oncol Biol Phys 2007;67(3):709–19.

[18] Konert T, Vogel W, Everitt S, MacManus M, Thorwarth D, Fidarova E, et al. Multiple training interventions significantly improve reproducibility of PET/CT-based lung cancer radiotherapy target volume delineation using an IAEA study protocol. Radiother Oncol 2016;121(1):39–45.

[19] La Macchia M, Fellin F, Amichetti M, Cianchetti M, Gianolini S, Paola V, et al. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. Radiat Oncol 2012;7(1):160.

[20] Venables K, Tsang Y, Ciurlionis L, Coles CE, Yarnold JR. Does participation in clinical trials influence the implementation of new techniques? A look at changing techniques in breast radiotherapy in the UK. Clin Oncol (R Coll Radiol) 2012;24:e100–5. https://doi.org/10.1016/j.clon.2012.06.010.

[21] Tsang Y, Ciurlionis L, Kirby A, Locke I, Venables K, Yarnold J, et al. Clinical impact of IMPORT HIGH trial (CRUK/06/003) on breast radiotherapy practices in the United Kingdom. Brit J Radiol 2015;88(1056):20150453. https://doi.org/10.1259/bjr.20150453.

[22] Weber DC, Tomsej M, Melidis C, Hurkmans CW. QA makes a clinical trial stronger: evidence-based medicine inradiation therapy. Radiother Oncol 2012;105:4–8.

[23] The Royal College of Radiologists. Radiotherapy target volume definition and peer review – RCR guidance. London: The Royal College of Radiologists. Ref No. BFCO(17)2; 2017.