Multiple training interventions significantly improve reproducibility of PET/CT-based lung cancer radiotherapy target volume delineation using an IAEA study protocol

Konert, T., Vogel, W., Everitt, S., MacManus, M. P., Thorwarth, D., Fidarova, E., ... Hanna, G. G. (2016). Multiple training interventions significantly improve reproducibility of PET/CT-based lung cancer radiotherapy target volume delineation using an IAEA study protocol. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology, 121(1), 39-45. DOI: 10.1016/j.radonc.2016.09.002

Published in:
Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology

Document Version:
Publisher's PDF, also known as Version of record

Radiotherapy and Oncology 121 (2016) 39–45

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

PET/CT target delineation

Multiple training interventions significantly improve reproducibility of PET/CT-based lung cancer radiotherapy target volume delineation using an IAEA study protocol

Tom Konert [a,b], Wouter V. Vogel [a,b], Sarah Everitt [c], Michael P. MacManus [c], Daniela Thorwarth [d], Elena Fidarova [e], Diana Paez [e], Jan-Jakob Sonke [b], Gerard G. Hanna [f,*]

[a] Nuclear Medicine Department; [b] Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands; [c] Division of Radiation Oncology, Peter MacCallum Cancer Centre, East Melbourne, Australia; [d] Department of Radiation Oncology, University Hospital Tübingen, Germany; [e] Department of Nuclear Sciences and Application, International Atomic Energy Agency, Vienna, Austria; and [f] Centre for Cancer Research and Cell Biology, Queen's University of Belfast, United Kingdom

ARTICLE INFO

Article history:
Received 6 May 2016
Received in revised form 1 September 2016
Accepted 4 September 2016
Available online 20 September 2016

Keywords:
Lung cancer
PET/CT
Radiation treatment planning
Radiotherapy
Target volume delineation
Training interventions

ABSTRACT

Background and purpose: To assess the impact of a standardized delineation protocol and training interventions on PET/CT-based target volume delineation (TVD) in NSCLC in a multicenter setting.
Material and methods: Over a one-year period, 11 pairs, comprised each of a radiation oncologist and nuclear medicine physician with limited experience in PET/CT-based TVD for NSCLC from nine different countries took part in a training program through an International Atomic Energy Agency (IAEA) study (NCT02247713). Teams delineated gross tumor volume of the primary tumor, during and after training interventions, according to a provided delineation protocol. In-house developed software recorded the performed delineations, to allow visual inspection of strategies and to assess delineation accuracy.
Results: Following the first training, overall concordance indices for 3 repetitive cases increased from 0.57 ± 0.07 to 0.66 ± 0.07. The overall mean surface distance between observer and expert contours decreased from -0.40 ± 0.03 cm to -0.01 ± 0.33 cm. After further training overall concordance indices for another 3 repetitive cases further increased from 0.64 ± 0.06 to 0.80 ± 0.05 ($p = 0.01$). Mean surface distances decreased from -0.34 ± 0.16 cm to -0.05 ± 0.20 cm ($p = 0.01$).
Conclusion: Multiple training interventions improve PET/CT-based TVD delineation accuracy in NSCLC and reduce interobserver variation.

Lung cancer is the most common cause of death from cancer worldwide, estimated to be responsible for nearly 17% of the total [1] and it is estimated that more than 80% of patients in low and middle income countries are diagnosed with lung cancer in an advanced stage (III and IV) [2,3]. The use of fused ^{18}F-fluorodeoxyglucose Positron Emission Tomography/Computed Tomography (FDG-PET/CT) imaging is now the standard method of acquiring FDG-PET images for the purpose of baseline staging and RT treatment preparation [4], since it has been shown to be superior to either PET or CT alone [5,6]. The number of PET/CT scanners has increased in low and middle income countries in the last decade [7] and additional training in the use of PET/CT in radiotherapy planning (RTP) is vital to ensure appropriate interpretation of PET/CT with the hope, that the use of PET/CT will improve outcomes for patients treated with radiotherapy.

Due to advancements in radiotherapy techniques, accuracy in treatment delivery is improving and precise target volume definition has become more important, particularly in the era of dose escalation [8,9]. However, gross tumor volume (GTV) delineation is very sensitive to observer variation [10] and hence there is a potential risk of geographic miss of tumor [11]. PET has been shown to have a significant impact when used in the radiation treatment planning process and in particular when used for target volume delineation (TVD), where a significant reduction in interobserver variability (IOV) has been noted [11–15]. It is recommended that a radiation oncologist (RO) and a nuclear medicine physician (NMP)/PET radiologist should be both involved where PET is used for TVD [16,17]. Complex cases of GTV delineation in lung cancer patients should always be discussed in a multi-disciplinary quality control meeting. Most clinical studies have used a visual interpretation technique, while others have reported the use of a range of

* Corresponding author at: Centre for Cancer Research and Cell Biology, Queen's University of Belfast, 97 Lisburn Road, Belfast BT9 7AE, Northern Ireland, United Kingdom.
E-mail address: g.hanna@qub.ac.uk (G.G. Hanna).

http://dx.doi.org/10.1016/j.radonc.2016.09.002
0167-8140/

automated segmentation techniques to either guide or generate the relevant target volume [18–21]. There is no clear consensus on which method most closely approximates to the tumor position and tumor edge, and pathological correlation has proven difficult [22]. Preoperative PET imaging shows a remarkably good correlation with resected pathological specimens [23], although it is acknowledged that those specimens are affected by processing artifacts. Most recent guidance advises the use of visual interpretation of the PET signal when drawing the final contour, even in cases where auto-contouring is used to generate an initial draft for editing, if PET is to be used to inform the target volume [17].

Factors causing IOV in TVD are variable interpretation of guidelines, lack of differentiation between normal structures and tumor, incorrect interpretation of radiological images, lack of knowledge in cross sectional radiological anatomy, and suboptimal imaging techniques e.g. lack of IV contrast [24–26]. The use of a rigorous contouring protocol in which clinicians follow a detailed set of instructions and the use of a teaching intervention may help in minimizing IOV [21,31,32,34,35]. To ensure adequate and reproducible visual interpretation and application of PET images for RTP, this procedure should be standardized. A recent publication provided guidance on the use and role of PET/CT imaging for RTP in NSCLC patient [17]. This study evaluates the impact on the use of these practical guidelines through active teaching using multiple training interventions involving multiple centers with minimal experience in PET/CT-based TVD.

Methods

Target volume delineation assignments

PET/CT-based TVD was assessed through the use of repeated delineation assignments. In all contouring assignments a team consisting of a RO and a NMP were asked to delineate tumor volumes of primary tumor (GTV). Before the training, participants were asked to delineate as per their local delineation protocol and then again after the first training intervention according to a standardized delineation protocol [17]. Fully anonymized patient cases were used, including three dimensional FDG-PET and CT image data sets acquired for the purpose of radiotherapy planning. No intravenous contrast agent was used. Comprehensive case specific medical reports were included in all assignments to avoid bias due to incorrect diagnosis. An overview of the patient cases used during this training program is given in Table 1. In each case two senior ROs and a senior NMP delineated one reference 'expert' contour (GTV_{exp}) in agreement in the absence of a histopathologically proven gold standard.

Participants

The participants in this study were from eleven medical centers from nine different countries (Brazil, Estonia, India, Jordan, Pakistan, Poland, Turkey, Uruguay, and Vietnam). Each center was represented by a RO and a NMP. Before the training program centers were asked if they already performed PET/CT-based RTP. Five out of eleven centers already had limited experience in TVD with PET/CT. Other centers used PET/CT imaging for diagnostic and staging purposes only. Participants were not able to see delineations of other centers.

Big Brother target volume delineation software

Software developed in the Netherlands Cancer Institute, called Big Brother, was used throughout this multicenter study as platform for image viewing and analysis, and TVD in FDG PET/CT imaging [10]. As soon as the observer starts the Big Brother software and initiates TVD, any interaction with the software is recorded such as mouse motion, window/level and use of delineation tools. This feature allows visual inspection of strategies and comparison with expert contours to assess delineation accuracy.

Target volume delineation training program

The training program consisted of four contouring assignments, two training events and three additional clinical cases for practice (see Fig. 1). Contouring assignments 1 and 2 were performed before the first training event without the use of a standardized delineation protocol and were used as a baseline measurement. The first training event was face-to-face over a three-day period and included various lectures about relevant topics in nuclear medicine and radiation oncology and a delineation workshop on the use of PET/CT for RTP in NSCLC. The delineation workshop contained three more clinical cases which were again performed without the IAEA delineation protocol. The delineation protocol as described in the IAEA consensus document was introduced during the workshop [17]. The differences between the results and the IAEA protocol constituted the basis for a teaching discussion, consequentially clarifying protocol ambiguities. More contouring assignments followed after this training to evaluate its impact on delineation accuracy and IOV. Contouring assignment 3 was performed three months after contouring assignment 2 and contained the same clinical cases. To allow the participants to practice more with the delineation protocol three additional clinical cases were added.

After results were obtained from the above described assignments, an interim analysis was performed with the aim of identifying difficult areas in TVD and to ensure delineation occurred

Table 1
Sequence of events and characteristics of the included patients. Abbreviations: T = Primary tumor, N = Regional Lymph Nodes according to the 7th edition TNM classification.

	Case number	T	N	M	Stage	Lymph nodes
Contouring Assignment 1	1	2	1	0	IIB	11L
	2	2	2	0	IIIA	10R, 7, 8, 4R
Contouring Assignment 2	3	3	0	0	IIB	
	4	2	0	0	IIA	
	5	4	2	0	IIIB	7, 4R, 2R
Training 1	6	3	2	0	IIIA	4R, 2R
	7	3	0	0	IIB	
	8	1	2	0	IIIA	10R, 4R
Contouring Assignment 3	Consisting of cases 3, 4 and 5 (repeat assignment)					
Practice	9	2	2	0	IIIA	10R, 7, 4R
	10	2	2	0	IIIA	10R, 7
	11	2	2	0	IIIA	7, 4R
Training 2	Webinar/feedback reports					
Contouring Assignment 4	Consisting of cases 1, 6 and 7 (repeat assignment)					

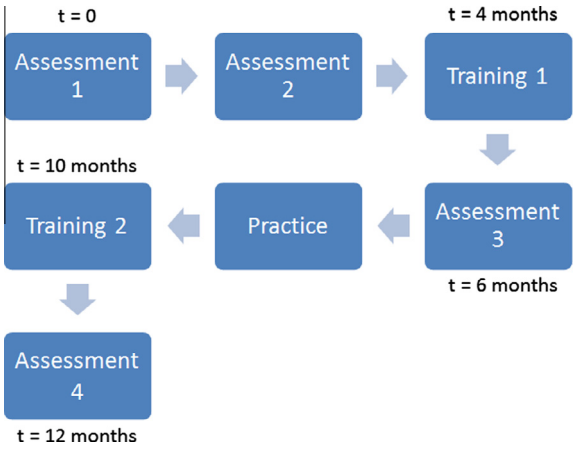

Fig. 1. Schematic view of the training program. Over a one year period, 11 pairs of a radiation oncologist and a nuclear medicine physician performed four contouring assignments and three more cases for practice, and also attended two training events. Assessment 1, 2 and cases in Training 1 functioned as baseline measurements. Assessment 2 and 3 contained the same cases and results were compared to assess the impact of the first training event. Assessment 4 contained one case from Assessment 1 and two from Training 1 and the results were compared to assess the impact of the complete training.

following the standardized approach. Detailed personal feedback reports were written with the aim of correcting misinterpretations of the delineation protocol and to advise on specific areas prone to deviation from the IAEA expert contour. This served as a preliminary to the webinar which was held as a second training event. An update on PET/CT-based TVD in RTP and general feedback was given in the webinar. Afterward the content was discussed with the group. As a final step participants performed contouring assignment 4 with three clinical cases, which they had performed earlier during the training program, eight months after the first training event.

Data analysis

To examine the impact of the training interventions the contours from the participants were analyzed. Various parameters such as the GTV size, miss of GTV_{exp}, Concordance Index (CI) and surface mean distance were calculated, and also volumetric and 3-dimensional analysis was performed as described by Deurloo et al. [27]. The CI is defined as the intersection of two delineated volumes divided by their union:

$$CI = \frac{(A \cap B)}{A \cup B}$$

The CI can vary between 0 and 1 where 0 means there is a complete disagreement between the observers and a CI of 1 indicates a perfect agreement [28]. It was calculated for measuring the delineation accuracy relative to the expert contour (CI_{expert}). Intragroup agreement (CI_{group}) was also calculated using the surface median as a reference. The surface median was obtained as described by Rasch et al. [29]. The mean (absolute) surface distance between the observers' GTV and expert GTV and the distance between each observers' GTV and surface median were both calculated. For all parameters, the mean ± SD is reported unless stated otherwise. Wilcoxon's signed rank tests were used to estimate the significance of any differences after the training events and p-values of 0.05 or less were considered significant.

Results

In all contouring assignments, teams were asked to delineate GTV of the primary tumor as per study protocol. One of the pairs with a RO, who was not board certified at the time of the training, was excluded from the analysis. After the first training event the overall CI_{expert} slightly increased from 0.57 ± 0.07 to 0.66 ± 0.07. The mean CI_{expert} and CI_{group} per case are given in Fig. 2. Observer volumes were larger after the training and miss of GTV_{exp} was significantly reduced from 127.32 ± 42.43 cc to 59.94 ± 48.94 cc. A detailed summary with p-values is given in Table 2. The overall mean surface distance and mean absolute surface distance compared to the reference contour decreased from −0.40 ± 0.03 cm to −0.01 ± 0.33 cm and from 0.47 ± 0.08 cm to 0.45 ± 0.17 cm respectively. The overall CI_{group} decreased from 0.81 ± 0.07 to 0.75 ± 0.10.

After the second teaching event overall CI_{expert} for another 3 repetitive cases increased from 0.64 ± 0.06 to 0.80 ± 0.05 (see Fig. 3 for more details). A reduction of GTV_{exp} miss from

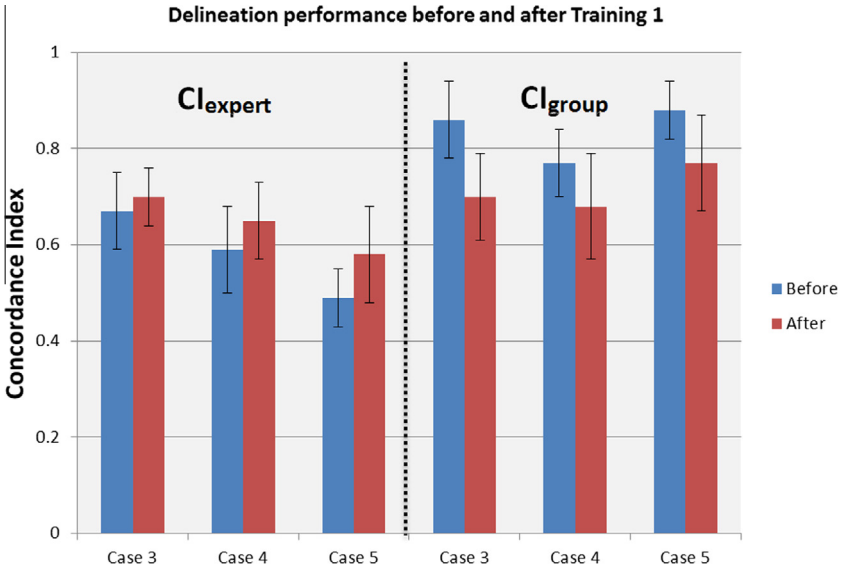

Fig. 2. Delineation accuracy relative to expert versus intragroup with 95%-CI before and after Training 1 using a standardized delineation protocol. CI_{group} = median concordance index between the observers' GTV and the median surface contour. CI_{expert} = median concordance index between the observers' GTV and expert GTV.

Table 2
Comparison of results from contouring the GTV before and after the first training event, and before and after a complete training in the use of a standardized delineation protocol. CI_{expert} = median concordance index between the observers' GTV and expert GTV. Mean distance = mean surface distance between the observers' GTV and expert GTV. Mean |distance| = mean absolute surface distance.

	Case No.	Expert volume (cc)	Observer volume (cc ± SD)	Miss (cc ± SD)	CI_{expert} (±SD)	Mean distance (cm ± SD)	Mean |distance| (cm ± SD)
Results per case before and after Training 1 (contouring assignment 2 versus 3)							
Before	3	388.38	282.75 ± 46.28	127.32 ± 42.43	0.67 ± 0.10	−0.41 ± 0.14	0.44 ± 0.13
After			409.48 ± 115.12	59.94 ± 48.94	0.70 ± 0.07	0.05 ± 0.40	0.37 ± 0.23
			(p = 0.03)	(p = 0.03)	(p = 0.34)	(p = 0.03)	(p = 0.92)
Before	4	50.86	30.58 ± 6.99	20.43 ± 6.09	0.59 ± 0.11	−0.27 ± 0.18	0.33 ± 0.10
After			51.35 ± 20.94	8.88 ± 9.34	0.65 ± 0.10	−0.11 ± 0.29	0.30 ± 0.10
			(p = 0.03)	(p = 0.03)	(p = 0.06)	(p = 0.17)	(p = 0.14)
Before	5	164.46	84.93 ± 11.67	84.26 ± 11.66	0.49 ± 0.07	−0.64 ± 0.11	0.66 ± 0.12
After			108.49 ± 41.23	62.18 ± 23.14	0.58 ± 0.12	−0.43 ± 0.64	0.50 ± 0.45
			(p = 0.05)	(p = 0.05)	(p = 0.08)	(p = 0.05)	(p = 0.46)
Overall results for 3 repeated cases (contouring assignment 2 versus 3)							
Before			123.96 ± 18.35	79.01 ± 17.04	0.57 ± 0.07	−0.40 ± 0.03	0.47 ± 0.08
After			191.38 ± 57.29	42.86 ± 25.02	0.66 ± 0.07	−0.01 ± 0.33	0.45 ± 0.17
			(p = 0.03)	(p = 0.05)	(p = 0.12)	(p = 0.03)	(p = 0.75)
Results per case before and after the complete training program (contouring assignment 1 versus 4)							
Before	6	377.99	280.39 ± 92.41	115.45 ± 52.16	0.68 ± 0.11	−0.41 ± 0.35	0.43 ± 0.19
After			370.78 ± 37.06	26.73 ± 18.34	0.85 ± 0.03	−0.01 ± 0.16	0.20 ± 0.09
			(p = 0.07)	(p = 0.01)	(p = 0.01)	(p = 0.07)	(p = 0.02)
Before	7	254.68	157.99 ± 80.15	103.70 ± 43.86	0.59 ± 0.10	−0.55 ± 0.40	0.60 ± 0.16
After			220.24 ± 81.53	39.72 ± 25.87	0.82 ± 0.10	−0.18 ± 0.34	0.24 ± 0.18
			(p = 0.04)	(p = 0.04)	(p = 0. 05)	(p = 0.07)	(p = 0.07)
Before*	1	134.07	78.08 ± 17.60	55.66 ± 9.58	0.56 ± 0.05	−0.44 ± 0.17	0.50 ± 0.06
After*			110.20 ± 14.51	26.23 ± 9.68	0.80 ± 0.04	−0.19 ± 0.11	0.21 ± 0.04
			(p = 0.08)	(p = 0.03)	(p = 0.05)	(p = 0.08)	(p = 0.05)
Overall results for 3 repeated cases (contouring assignment 1 versus 4)							
Before			190.82 ± 38.07	78.89 ± 22.95	0.64 ± 0.06	−0.34 ± 0.16	0.49 ± 0.10
After			246.32 ± 43.11	42.86 ± 25.01	0.80 ± 0.05	−0.05 ± 0.20	0.27 ± 0.09
			(p = 0.01)	(p = 0.01)	(p = 0.01)	(p = 0.01)	(p = 0.01)

* For this case (before and after) two observers were excluded from analysis, because PET positive lymph nodes adjacent to the tumor were included in the GTV of the primary tumor.

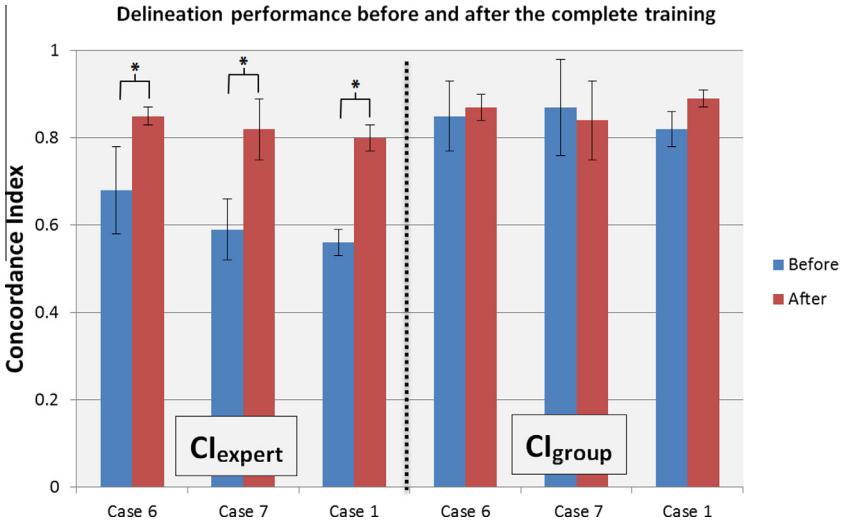

Fig. 3. Delineation accuracy relative to expert versus intragroup with 95%-CI before (blue bar) and after (red bar) a complete training in the use of a standardized delineation protocol. CI_{group} = median concordance index between the observers' GTV and the median surface contour. CI_{expert} = median concordance index between the observers' GTV and expert GTV. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

78.89 ± 22.95 cc to 42.86 ± 25.01 cc was observed, next to an increase in observer volume. Overall mean surface distances between observers and the expert contour decreased from −0.34 ± 0.16 cm to −0.05 ± 0.20 cm. A decrease from 0.49 ± 0.10 cm to 0.27 ± 0.09 cm in overall mean absolute surface distance was observed. The overall CI_{group} increased from 0.80 ± 0.08 to 0.85 ± 0.08. Examples of improvement in IOV and delineation accuracy before and after the training program are given in Fig. 4 and in supplementary Fig. 1.

Discussion

Many studies have reported on the effect of guidelines or protocols and testing of the effect of specific teaching and measured contouring variability before and after an intervention [30]. This study is the first to report on the impact of a training program about the use of the recently published IAEA guidelines for PET/CT based radiotherapy planning in lung cancer patients in an international multicenter trial. To the best of our knowledge this

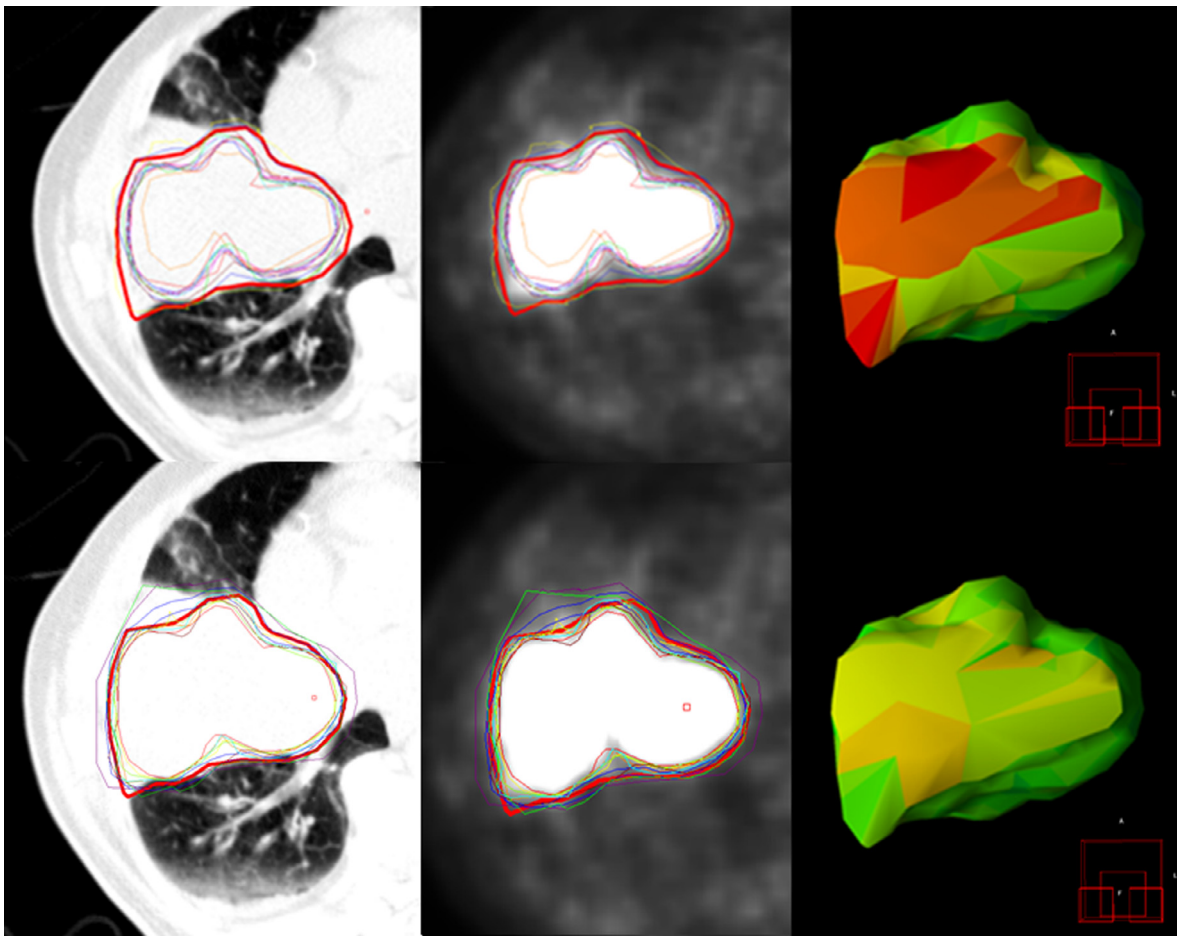

Fig. 4. Results from contouring the GTV in case 6 before (top images) and after (bottom images) the complete training. From left to right an image slice in the axial plane from CT, PET, and a 3D model of the expert contour with mean surface distance errors projected (from blue to red, corresponding with a mean absolute surface distance of 0 to > 1 cm). The bold red line represents the GTVexp.

is the first study about the use of multiple teaching interventions using not only face-to-face training, but also providing an online learning platform in the form of a webinar which showed to play an effective role in harmonizing the delineation process globally. In terms of practicalities of delivering this type of training, the use of innovative technology such as the delivery of live webinars may significantly reduce cost without significantly reducing educational quality. Any economic assessment regarding the cost benefit ratio undertaking repeated tests of training exceeded the scope of this study.

Before any training was given, centers delineated their GTV based on local delineation protocols and thus a variety of approaches were observed. Results after the first training suggest that the use of a delineation protocol increased delineation accuracy, however a significant reduction in IOV and better adherence of the outlining protocol were only achieved through additional extended training.

A major contribution in reducing IOV may have been feedback participants received from the interim analysis that was performed before the second training intervention. While it is acknowledged that the cases selected for the practice case (see Table 1) may have been more challenging to delineate than the test cases, the use of these helped identify several areas that caused difficulty in TVD for the participants. These areas of variation included atelectasis, PET window level settings, nearby normal structures (e.g. pulmonary veins or arteries) that were seen as tumor, and/or suspicious areas on CT showing low FDG uptake. These areas of variation were documented in the individual feedback reports, and were also included in the general feedback and discussed during the webinar. This may have contributed to the significant impact on delineation accuracy seen in the last assignment. This emphasizes the importance of additional training events to correct errors in delineation still occurring after the first training. Schimek-Jasch et al. found out that the use of a dummy-run and study group meetings as part of Quality Assurance (QA) in multicenter clinical trials helps to identify misinterpretations of a standardized delineation method which helped in reducing IOV [32]. However, the outcome of that study did not show a significant effect which may underline the take-home message of our study that changing behavior requires multiple and multi-faceted interventions. Lack of other studies investigating the impact of multiple training interventions in TVD makes comparison of outcomes difficult. Further studies with more observers would be needed to validate the results in this study. Currently the IAEA conducts a multicenter international study investigating the impact of blended distance learning (with additional training interventions) in the field of RT contouring on quality of delineation (CRP E33040).

Several studies concluded that the use of a standardized delineation protocol helps in decreasing IOV [31,32,34] and our results concur with that. There was a non-significant increase in IOV seen after the first training event, however a decrease was noted after the complete training program. In spite of the first training, visual inspection with Big Brother showed that some observers misinter-

preted or did not comply with the guidelines thus their contours did not more closely resemble the CI_{expert} contour. Due to the fact that some observers drew contours more similar to that of the expert in contrary to others in the group, the CI_{expert} still increased slightly. This increase in variation between observers reduced the CI_{group} (see Fig. 2). The second training event was necessary to correct any misinterpretations of practical guidelines and this approach of having training interventions may seem more effective than a single event training program. This again highlights the difficulty of changing behavior in order to obtain reproducibility and underlines the importance of teaching through multiple interventions to improve adherence to contouring guidelines. However, participants still had difficulties in determining the tumor boundary in cases with suspicious areas showing FDG uptake comparable to background activity. Decisions, as to whether or not to include these suspect areas within the GTV, are namely based on experience and expertise and this contributes to a degree of IOV. This emphasizes the importance of multidisciplinary meetings in case the RO experiences difficulties in contouring.

Contrary to the delineation accuracy, GTV size did increase significantly among the observers after the first training. This was mainly due to specific training in standardization of PET window level settings which is applicable in most commercially available radiotherapy planning systems [17]. The difficulty with manual delineation in PET/CT imaging is that the apparent boundaries of the FDG avid tumor are highly dependent on the chosen PET window level settings. Tumors will appear larger when delineation occurs using a high window setting and smaller with a low window setting on the PET display. Observers were trained in using standardized PET level window settings consequently showing an apparently larger tumor than before and this contributed to an increase in GTV size. Before the 1st training intervention most delineations were drawn too tightly around the tumor, which may possibly lead to geographical tumor miss. Caution has to be taken in circumstances where, in the absence of respiration correlated CT, PET/CT may be substantially less accurate in defining the motion pathway of highly mobile lung tumors and in tumors with low FDG uptake [33].

When observers delineate the same tumor repeatedly this creates systematic errors and contributes to intra-observer variability, since it is unlikely that any manual contour would be reproduced identically at different time points. In this study we did not examine what contribution this effect had on intra-observer variability when repeating the same case in a short timeframe. However it is hypothesized this effect is negligible compared to learning effects over a longer timeframe.

There have been a number of studies which have examined IOV in TVD using PET based delineation with a number of these studies focusing on the comparison of automatic delineation methods with manual delineation [17–21]. Doll et al. used one patient case and found overall concordance indices between experts, interdisciplinary pairs and single field specialists ranging from 0.49 to 0.67 which is similar to our results after the first training [16]. The experts showed the highest intragroup concordance of 0.67 and if that is representative then the outcome of our training program could be seen as successful. However, since the expert group in Doll et al. only performed one case and the study did not use a standardized delineation protocol this comparison is not valid. This study used only one expert contour per case as reference, which limited the conclusion whether an observer met a certain minimum level of quality in TVD. An intragroup expert concordance value could help in determining such a minimum required level of quality. Further research has to determine which deviation from the intragroup concordance value would be acceptable.

Another limitation was the amount of cases available for the repeated assignments. Not all cases in Training 1 were suitable for inclusion in Assessment 4 due to the small tumor size in case 8 and therefore one case of the first assignment (case 1) had to be included. It is acknowledged that observers were less familiar with the software in the beginning than later in the training program, however the delineation software is similar to any other delineation tool used in clinic and it is hypothesized that if there is any learning effect present, this only has an impact on the delineation speed. Above that, a similar increase is seen in all cases after the complete training suggesting that the learning effect of the delineation software was negligible (see Fig. 3).

Spoelstra et al. had seen that a significant IOV in contouring confounded interpretation of post-operative radiotherapy and concluded that Quality Assurance (QA) procedures would need to be incorporated to tackle this problem [34]. A German multicenter PET study also covered a similar interesting topic i.e. harmonization of diagnostic viewing and reporting and also outlined the importance of QA [35]. They concluded that a structured interventional harmonization process significantly improved the IOV in their expert panel. However, no focus had been given on target volume delineation. In our study, additional training led to an increased delineation accuracy and decreased the IOV. In clinic, the IOV should also be assessed in order to see if it is necessary to provide more training in order to achieve reproducible results among ROs. Therefore it is recommended that assessment of IOV should be performed frequently next to having multi-disciplinary quality control meetings as part of the QA on TVD.

A study examining the influence of experience and qualification in PET based TVD concluded that IOV may be dependent on qualification, but not on years of experience [16]. It is known that some centers already had minimal experience in PET/CT-based TVD and that there were centers with no experience in this field. However, no significant difference in performance was seen after comparison of inexperienced versus minimal experienced participants.

Conventional 3-dimensional PET/CT imaging was chosen as the modality for TVD, since not all participants in the study had experience in PET/CT-based radiotherapy planning in NSCLC patients. The impact of 4-dimensional PET/CT has not been investigated and may be of interest to further increase accuracy. Furthermore, if a PET/CT acquired for diagnostic or staging purposes is used to inform the TVD process, care must be taken when registering a diagnostic PET/CT with a planning CT. Guidance regarding this has been described in the IAEA study protocol [17].

Conclusion

ROs and NMPs with limited experience in PET/CT-based TVD for lung cancer benefit significantly from receiving multiple training interventions with a standardized delineation protocol. Future research within a larger population should validate the results in this study to provide more evidence on the impact of multiple training interventions about PET/CT-based TVD for NSCLC.

Conflict of interest

None to declare.

Acknowledgements

The authors would like to thank Heloisa Carvalho and Paolo Duarte (Instituto do Câncer do Estado de São Paulo), Camilla Mosci and Carlos Zuliani (Hospital das Clínicas UNICAMP), Darja Altuhhova and Liina Karusoo (North Estonia Medical Center), Rakesh Kapoor and Ashwani Sood (Postgraduate Institute of Medical Education and Research), Jamal Khader and Akram Al-Ibraheem (King Hussein Cancer Center), Shahid Abubaker and Muhammad Numair

(Institute of Nuclear Medicine and Oncology), Bozena Birkenfeld and Bartlomiej Masojc (Zachodniopomorskie Centrum Onkologii), Cigdem Soydal and Tuğçe Kütük (Ankara University School of Medicine), Aldo Quarnetti and Omar Alonso (Centro Uruguayo de Imagenología Molecular), Quang Bieu Bui and Ngoc Ha Le (Tran Hung Dao General Hospital), Nguyen Xuan Canh and Tuan Anh Le (Cho Ray Hospital) for their active cooperation in the study. The authors also wish to thank Marcel van Herk for the support with the Big Brother software in this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.radonc.2016.09.002.

References

[1] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:359–86.
[2] US National Institutes of Health. National Cancer Institute. SEER Cancer Statistics Review, 1975–2011, 2014. Accessed on 02-03-2016.
[3] National Cancer Intelligence Network. Stage Breakdown by CCG 2013. London: NCIN; 2015. Accessed on 02-03-2016.
[4] Ung JC, Bezjak A, Coakley N, Evans WK, Lung Cancer Disease Site Group. Positron emission tomography with 18Fluorodeoxyglucose in radiation treatment planning for non-small cell lung cancer. J Thorac Oncol 2011;6:86–97.
[5] Cerfolio RJ, Ojha B, Bryant AS, Raghuveer V, Mountz JM, Bartolucci AA. The accuracy of integrated PET-CT compared with dedicated PET alone for the staging of patients with non small cell lung cancer. Ann Thorac Surg 2004;78:1017–23.
[6] De Wever W, Ceyssens S, Mortelmans L, et al. Additional value of PET-CT in the staging of lung cancer: comparison with CT alone, PET alone and visual correlation of PET and CT. Eur Radiol 2007;17:23–32.
[7] Dondi M, Kashyap R, Paez D, Pascual T, Zaknun J, Bastos FM, et al. Trends in nuclear medicine in developing countries. J Nucl Med 2011;52:16–23.
[8] De Ruysscher D, Wanders S, Minken A, et al. Effects of radiotherapy planning with a dedicated combined PET-CT-simulator of patients with non-small cell lung cancer on dose limiting normal tissues and radiation dose-escalation: a planning study. Radiother Oncol 2005;77:5–10.
[9] van Elmpt W, De Ruysscher D, van der Salm A, Lakeman A, van der Stoep J, Emans D, et al. The PET-boost randomised phase II dose-escalation trial in non-small cell lung cancer. Radiother Oncol 2012;104:67–71.
[10] Steenbakkers RJ, Duppen JC, Fitton I, Deurloo KE, Zijp L, Uitterhoeve AL, et al. Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a 'Big Brother' evaluation. Radiother Oncol 2005;77:182–90.
[11] Rasch C, Belderbos J, van Giersbergen A, De Kok I, Laura T, Boer M, et al. The influence of a multi-disciplinary meeting for quality assurance on target delineation in radiotherapy treatment preparation international. J Radiat Oncol 2009;75:452–3.
[12] Hanna GG, McAleese J, Carson KJ, Stewart DP, Cosgrove VP, Eakin RL, et al. (18) F-FDG PET-CT simulation for non-small-cell lung cancer: effect in patients already staged by PET-CT. Int J Radiat Oncol Biol Phys 2010;77:24–30.
[13] Caldwell CB, Mah K, Ung YC, et al. Observer variation in contouring gross tumor volume in patients with poorly defined non small-cell lung tumors on CT: The impact of 18FDG-hybrid PET fusion. Int J Radiat Oncol Biol Phys 2001;51:923–31.
[14] Greco C, Rosenzweig K, Cascini GL, Tamburrini O. Current status of PET/CT for tumour volume definition in radiotherapy treatment planning for non-small cell lung cancer (NSCLC). Lung Cancer 2007;57:125–34.
[15] Fox JL, Rengan R, O'Meara W, Yorke E, Erdi Y, Nehmeh S, et al. Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non-small-cell lung cancer? Int J Radiat Oncol Biol Phys 2005;62:70–5.
[16] Doll C, Duncker-Rohr V, Rücker G, Mix M, MacManus M, De Ruysscher D, et al. Influence of experience and qualification on PET-based target volume delineation. When there is no expert–ask your colleague. Strahlenther Onkol 2014;190:555–62.
[17] Konert T, Vogel W, MacManus MP, Nestle U, Belderbos J, Grégoire V, et al. PET/CT imaging for target volume delineation in curative intent radiotherapy of non-small cell lung cancer: IAEA consensus report 2014. Radiother Oncol 2015;116:27–34.
[18] Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging 2010;37:2165–87.
[19] Werner-Wasik M, Nelson AD, Choi W, et al. What is the best way to contour lung tumors on PET scans? Multiobserver validation of a gradient-based method using a NSCLC digital PET phantom. Int J Radiat Oncol Biol Phys 2012;82:1164–71.
[20] Cui H, Wang X, Feng D. Automated localization and segmentation of lung tumor from PET-CT thorax volumes based on image feature analysis. In: Proceedings of the 34th annual international conference of the IEEE engineering in medicine and biology society, San Diego, Calif, USA; 2012. p. 5384–7.
[21] Bayne M, Hicks RJ, Everitt S, Fimmell N, Ball D, Reynolds J, et al. Reproducibility of "intelligent" contouring of gross tumor volume in non-small-cell lung cancer on PET/CT images using a standardized visual method. Int J Radiat Oncol Biol Phys 2010;77:1151–7.
[22] van Loon J, Siedschlag C, Stroom J, Blauwgeers H, van Suylen RJ, Knegjens J, et al. Microscopic disease extension in three dimensions for non-small-cell lung cancer: development of a prediction model using pathology-validated positron emission tomography and computed tomography features. Int J Radiat Oncol Biol Phys 2012;82:448–56.
[23] 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 2007;68:771–8.
[24] Van de Steene J, Linthout N, de Mey J, et al. Definition of gross tumor volume in lung cancer: interobserver variability. Radiother Oncol 2002;62:37–49.
[25] Giraud P, Elles S, Helfre S, De Rycke Y, Servois V, Carette MF, et al. Conformal radiotherapy for lung cancer: different delineation of the gross tumor volume (GTV) by radiologists and radiation oncologists. Radiother Oncol 2002;62:27–36.
[26] 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;85:450–5.
[27] Deurloo KE, Steenbakkers RJ, Zijp LJ. Quantification of shape variation of prostate and seminal vesicles during external beam radiotherapy. Int J Radiat Oncol Biol Phys 2005;61:228–38.
[28] Hanna GG, Hounsell AR, O'Sullivan JM. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. Clin Oncol (R Coll Radiol) 2010 Sep;22:515–25.
[29] Rasch CR, Steenbakkers RJ, Fitton I, Duppen JC, Nowak PJ, Pameijer FA, et al. Decreased 3D observer variation with matched CT-MRI, for target delineation in Nasopharynx cancer. Radiat Oncol 2010;5:21.
[30] 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;60:392–406.
[31] Bowden P, Fisher R, Mac Manus M, Wirth A, Duchesne G, Millward M, et al. Measurement of lung tumor volumes using three-dimensional computer planning software. Int J Radiat Oncol Biol Phys 2002;53:566–73.
[32] Schimek-Jasch T, Troost EG, Rücker G, Prokic V, Avlar M, Duncker-Rohr V, et al. A teaching intervention in a contouring dummy run improved target volume delineation in locally advanced non-small cell lung cancer: Reducing the interobserver variability in multicentre clinical studies. Strahlenther Onkol 2015;191:525–33.
[33] Hanna GG, van Sörnsen de, Koste JR, Dahele MR, et al. Defining target volumes for stereotactic ablative radiotherapy of early-stage lung tumours: a comparison of three-dimensional 18F-fluorodeoxyglucose positron emission tomography and four-dimensional computed tomography. Clin Oncol (R Coll Radiol) 2012;24:71–80.
[34] Spoelstra FO, Senan S, Le PéchouxC, Ishikura S, Casas F, Ball D, et al. Variations in target volume definition for postoperative radiotherapy in stage III non-small cell lung cancer: analysis of an international contouring study. Lung Adjuvant Radiotherapy Trial Investigators Group. Int J Radiat Oncol Biol Phys 2010;76:1106–13.
[35] Nestle U, Rischke HC, Eschmann SM, Holl G, Tosch M, Miederer M, et al. Improved inter-observer agreement of an expert review panel in an oncology treatment trial – Insights from a structured interventional process. Eur J Cancer 2015;51:2525–33.