Radiotherapy and Oncology 167 (2022) 57–64

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

ProCaLung – Peer review in stage III, mediastinal node-positive, non-small-cell lung cancer: How to benchmark clinical practice of nodal target volume definition and delineation in Belgium ☆

Florian Charlier [a], Thomas Descamps [a], Yolande Lievens [b], Xavier Geets [c], Vincent Remouchamps [d], Maarten Lambrecht [e], Luigi Moretti [a,*]

[a] Radiation Oncology Department, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium; [b] Radiation Oncology Department, Ghent University Hospital and Ghent University, Ghent, Belgium; [c] Radiation Oncology Department, Cliniques Universitaires Saint-Luc, Brussels; [d] Radiation Oncology Department, CHU UCL Namur - site Sainte Elisabeth, Namur; and [e] Department of Radiation Oncology, University Hospitals Leuven, Belgium

ARTICLE INFO

Article history:
Received 22 July 2021
Received in revised form 24 November 2021
Accepted 30 November 2021
Available online 7 December 2021

Keywords:
Lung neoplasms
Radiotherapy
Quality assurance
Health care
Peer review
Lymph nodes
Lymphatic irradiation
Mediastinum
Observer variation
Practice guideline
Qualitative research

ABSTRACT

Background and purpose: The Quality Assurance project for stage III non-small cell lung cancer radiotherapy ProCaLung performed a multicentric two-step exercise evaluating mediastinal nodal Target Volume Definition and Delineation (TVD) variability and the opportunity for standardization. The TVD variability before and after providing detailed guidelines and the value of qualitative contour reviewing before applying quantitative measures were investigated.

Materials and methods: The case of a patient with stage III NSCLC and involved mediastinal lymph nodes was used as a basis for this study. Twenty-two radiation oncologists from nineteen centers in Belgium and Luxembourg participated in at least one of two phases of the project (before and after introduction of ProCaLung contouring guidelines). The resulting thirty-three mediastinal nodal GTV and CTV contours were then evaluated using a qualitative-before-quantitative (QBQ) approach. First, a qualitative analysis was performed, evaluating adherence to most recent guidelines. From this, a list of observed deviations was created and these were used to evaluate contour conformity. The second analysis was quantitative, using overlap and surface distance measures to compare contours within qualitative groups and between phases. A 'most robust' reference volume for these analyses was created using the STAPLE-algorithm and an averaging method.

Results: Five GTV and seven CTV qualitative groups were identified. Second step contours were more often in higher-conformity groups (p = 0.012 for GTV and p = 0.024 for CTV). Median Residual Mean Square Distances improved from 2.34 mm to 1.36 mm for GTV (p = 0.01) and from 4.53 mm to 1.58 mm for CTV ($p < 0.0001$). Median Dice coefficients increased from 0.81 to 0.84 for GTV (p = 0.07) and from 0.82 to 0.89 for CTV ($p \leq 0.001$). Using HC-contours only to generate references translated in more robust quantitative evaluations.

Conclusion: Variability of mediastinal nodal TVD was reduced after providing the ProCaLung consensus guidelines. A qualitative review was essential for providing meaningful quantitative measures.

It is well established that toxicity and tumor response obey dose-dependent relationships in locally advanced non-small cell lung cancer (LA-NSCLC) radiation treatments [1]. To maximize the therapeutic window, continuous improvement of spatial accuracy is necessary. On one hand, assessments of nodal involvement and localization are now more accurate owing to advances in diagnostic and staging procedures integrating imaging with (minimally-)invasive nodal sampling [2]. On the other hand, new computing algorithms enable more accurate and precise dose calculations, whereas technological evolutions in linear accelerators and progress in image guidance have been continuously improving treatment delivery accuracy.

Among the tasks performed by radiation oncologists for each treatment, two are arguably complex to evaluate and improve: the target volume definition, the selection of the malignant nodes to treat, and its delineation or contouring, the delimitation of this target on the planning CT-scan. Optimal target volume definition and delineation (TVD) is a prerequisite to ensure the safety and

☆ This study was supported by the Belgian College for Physicians in Radiation Oncology.

* Corresponding author at: Département de Radiothérapie-Oncologie, Institut Jules Bordet, Rue Héger-Bordet 1, 1000 Bruxelles, Belgium.
E-mail address: luigi.moretti@bordet.be (L. Moretti).

https://doi.org/10.1016/j.radonc.2021.11.034
0167-8140/

efficacy of promising isotoxic or even dose-escalated treatments [3]. In recent clinical trials, detailed contouring guidelines and quality assurance processes are provided to radiation oncologists to ensure consistent target volumes across time and study sites [4–6]. In routine clinical practice, TVD may be less rigorously performed, resulting in interobserver variability (IOV) [7], which has been recognized as a major source of spatial imprecision in radiation oncology (RO) [8]. Even within big-data-oriented initiatives, IOV might hinder knowledge extraction such that defining the best TVD method from these comprehensive data analyses may turn out to be impossible. Therefore, continued efforts are needed from the RO community to reduce this variability.

A possible explanation for the persistence of IOV may be the lack of description of the anatomical differences between contours in many studies which mainly reported numerical expressions of IOV [6,9,10], instead of the information required for clinicians to adapt their methods. Furthermore, quantitative approaches used to analyze the quality of TVD [11] are generally applied to a reference volume generated from multiple observers' inputs [12,13] irrespective of their similarity. It can be hypothesized that performing a qualitative review to select the contours to feed the consensus algorithm's may improve its robustness.

Several methods have been used to reduce IOV. As reviewed by Vinod et al. [14] and recommended by the UK Royal College of Radiologists [15] and in a recent systematic literature review [16], guidelines and peer-review are valuable approaches. The Belgian College for Physicians in Radiation Oncology (College) is a federal body tasked with the mission to evaluate and improve the quality of RO practice across the country, relying on centers' voluntary participation. After the rectal and breast cancer projects (PROCARE [17] and PROCAB[18,19], resp.), ProCaLung was introduced to monitor the quality of LA-NSCLC radiotherapy through peer-review of mediastinal nodal TVD. The project started with a benchmark-case delineation exercise in two steps which explored the magnitude of real-life TVD variability across centers before and after the publication of ESTRO-ACROP guidelines [20]. These recommendations provide guidance for defining the nodes to irradiate, but still allow for very different Clinical Target Volumes (CTV) depending on the physician's choice. Hence, the College organized an expert consensus meeting to establish less permissive recommendations, based on these ESTRO-ACROP guidelines, below referred to as ProCaLung guidelines, where target definition is also based on the algorithm described by Peeters et al. [2], but delineation recommendations are more specific (described in Methods). These guidelines were provided to the participants for the second step of the exercise to evaluate their impact on the standardization of the mediastinal nodal TVD in Belgium.

This paper describes our qualitative and quantitative analyses of this nodal target definition and delineation exercise with an in-depth evaluation of IOV.

Materials and methods

In January 2017, radiation oncologists with expertise in thoracic oncology from the 25 radiotherapy centers in Belgium and Luxembourg were invited to participate in a delineation exercise of a

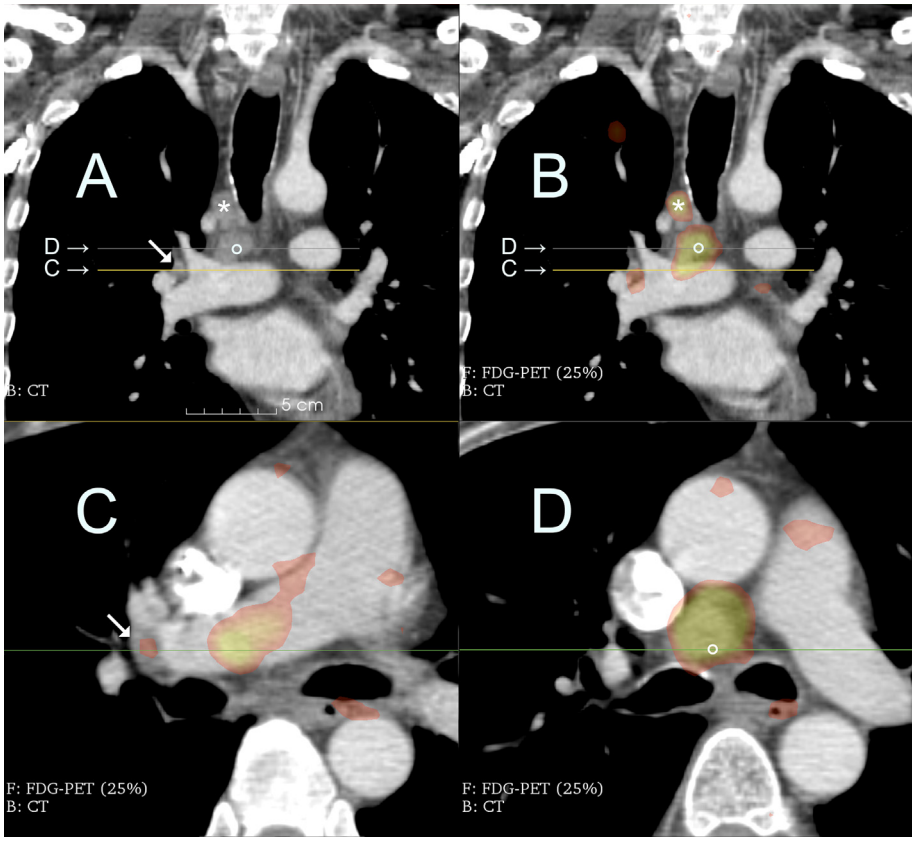

Fig. 1. Delineation exercise case. The exercise case was a patient with a squamous cell carcinoma of the right upper lobe. Regarding the regional workup, the chest CT-scan described a 1.1 × 1.7 cm node in upper 4R nodal station (* in frames A and B) and contiguous nodes up to 2 × 2.4 cm in lower 4R station (pre-carinal) (° in frames A–C) and a node in station 10R with a 9 mm short axis (white arrow in frames A and C). The PET/CT report described lymph nodes with SUVmax of 12.2 and 13 in upper (*) and lower 4R (°), respectively. Note a longitudinal stretch of the PET signal possibly due to breathing motion, best seen on the pulmonary artery on frame B and C. The results of the EBUS procedure confirmed the infiltration of the upper 4R node, while the analysis of the lower 4R punction revealed fatty tissue without neoplastic cell. Abbreviations: EBUS = Endobronchial Ultrasound, CT = Computed Tomography, FDG = Fluorodeoxyglucose, PET = Positron Emission Tomography, SUVmax = Maximum Standard Uptake Value.

stage III NSCLC patient with a squamous cell carcinoma of the right upper lobe and positive mediastinal nodes (Fig. 1). It is important to note that this patient is just one test case. The prospective segment of the ProCaLung study will include all stage III non-small-cell lung cancer patients with mediastinal node involvement (thus excluding those with T4 N0 and T3/4 N1 disease). Inclusion will require histological confirmation, and all non-small-cell lung cancer subtypes will be included. Workup details were sent to the participants alongside the contrast-enhanced Positron Emission Tomography/Computed Tomography (PET/CT) images. In June 2018, a new invitation to delineate the same patient was sent to all Belgian radiation oncologists with expertise in lung cancer, regardless of their participation in step 1, with the ProCaLung guidelines attached (Supplementary Materials). These included the recommendations of Peeters et al. [2] for target definition and those of ESTRO-ACROP [20] for delineation. However, it was decided to add two additional restrictions in order to standardize contouring: the use of a 5 mm CTV margin around the GTV instead of whole nodal stations and the exclusion of elective irradiation of adjacent nodal stations.

Qualitative analysis

GTV delineations from step 1 (baseline) and step 2 (with ProCaLung guidelines) were visually inspected for compliance with ProCaLung TVD recommendations. Upon inspection, multiple deviations from recommendations were discovered (Figs. S1 and S2), and a list was created containing all types of deviations that were observed (Table 1, left column). The disparities were then categorized into groups and ranked based on their estimated impact on the contours' similarity to the GTV expected per protocol. Each GTV was then re-examined using this list to ensure that no deviation was missed, and then allocated to a group based on its largest deviation observed. The first two groups included delineations with target definition deviations and were collectively labeled "non-conform", while the last group included those with the highest conformity to guidelines (labeled HC for Highly Conformal).

The same procedure (Fig. S3) was also applied to categorize CTV delineations. However, the first CTV group included all contours with "non-conform" GTV (Fig. S4). Contours with the fewest and smallest deviations and whose respective GTV had been assigned to HC were included in the last CTV group (HC).

This categorization was used to compare the exercise's steps qualitatively and to guide the reference volume determination for quantitative evaluations.

Quantitative analyses

Independently for GTV and CTV, we determined the reference volume for quantitative comparisons by identifying the one best translating the qualitative review, i.e., that best discriminated HC vs all other groups (non-HC). To assess the importance of selecting contours to use when generating that reference, three contour sets were considered: delineations from all participants (ALL), those from the second step only, as guidelines were provided (STEP2), and only the contours most conforming to guidelines after the qualitative review (HC). To reduce the risk of findings being associated with the consensus algorithm used, two methods were applied on these three contour sets, generating a total of six different volumes: the Simultaneous Truth and Performance Level Esti-

Table 1
Qualitative evaluation of contours.

GTV deviations from guidelines	Step 1 (n = 18) Nb of RO | %		Step 2 (n = 15) Nb of RO | %		GTV qualitative group (if 1st deviation)	Step 1 (n = 18) Nb of RO | %		Step 2 (n = 15) Nb of RO | %	
Inclusion of hilum node	3	17%	1	7%	#1 – GTV inclusion criteria unmet	6	33%	1	7%
Inclusion of 2R node	1	6%	0	0%					
No inclusion of lower 4R node	1	6%	0	0%					
Inclusion of part of right brachiocephalic artery (other nodal station)	1	6%	0	0%					
Inclusion of part of the pericardium (posterior aspect of the superior aortic recess)	7	39%	4	27%	#2 – Non nodal tissue included in GTV	6	33%	4	27%
Imprecise delineation: Missing slices or part(s) of a node or inclusion of a significant amount of non-nodal tissue, defined as "clearly visible when using a mediastinum viewing window".	15	83%	9	60%	#3 – Other gross imprecision in GTV delineation	4	22%	3	20%
SNN included in GTV *	15	83%	6	40%	#4 – SNN included in GTV *	1	6%	2	13%
No deviation from guidelines	1	6%	5	33%	#5 – GTV High conformity (HC) (n = 6)	1	6%	5	33%
CTV deviations from guidelines	**Step 1 (n = 18)**		**Step 2 (n = 15)**		**CTV qualitative groups (if 1st deviation)**	**Step 1 (n = 18)**		**Step 2 (n = 15)**	
GTV in groups #1 or #2	12	67%	5	33%	#1 – GTV not conform	12	67%	5	33%
CTV delineated in stations without GTV	4	22%	1	7%	#2 – Elective CTV	1	6%	1	7%
Delineated more than 5-mm margin around GTV (mostly whole nodal stations)	8	44%	3	20%	#3 – Selective CTV but not a 5 mm margin	1	6%	0	0%
No or partial inclusion of SNN	5	28%	1	7%	#4 – CTV missing SNN	1	6%	1	7%
No or partial crop to lungs or air in trachea	7	39%	7	47%	#5 – Missing crop to lung parenchyma and/or air in trachea	2	11%	2	13%
No crop to pericardium recess and no or partial crop to blood vessels. Most often missed: at the bottom of the contour (into the main pulmonary artery), or at the anterior part of the azygos vein	18	100%	15	100%	#6 – All criteria met except crop to vessels and pericardium and GTV not HC *	1	6%	2	13%
					#7 – CTV High conformity (HC) (n = 4)	0	0%	4	27%

Counts of radiation oncologists' contours for each deviation and group allocation. Percentages shown are calculated with respect to total number of contours sent at each step (n). * Groups where duplicated contours were excluded from group comparisons but kept included in steps comparisons.
Abbreviations: CTV = Clinical Target Volume; GTV = Gross Tumor Volume; RO = Radiation Oncologist; SNN = Small Neighboring Nodes, infra-centimetric PET-negative lymph nodes within a 5 mm distance from the positive nodes.

mation algorithm, applied with default parameters (STAPLE) [21], and an averaging method, selecting voxels included by a majority of the physicians (VOTE) [22]. Computations were performed in 3D on rasterized versions of the contours, i.e., a CT-sized matrix (512 × 512×181) where voxels included in the volume equal 1 and the others 0. The analyses consisted of the Dice Similarity Coefficient (DSC) and Surface Distance metrics (SD): the Residual Mean Square Distance (RMSD) [23], Hausdorff Distance (HDmax), and the 95th and 99th percentile partial distances (HD95 and HD99) [24]. Specifically, the maximum value of directed measure was considered for SDs (RMSD excepted). The best values for these metrics are 1 for DSC and 0 for SDs if absolutely identical volumes are compared.

The best volume identified was used as a reference to compare the exercise steps quantitatively and to assess the performance of an arbitrarily chosen 3 mm HD95 as a threshold for a classification method of HC vs. non-HC contours.

Statistical methods and software

Cochran-Armitage trend tests were used to compare the distribution of contours in qualitative groups between the two exercise steps. Repeated measures analyses of variance on ranks [25] were applied for DSC and RMSD for each method to detect if the contour set used to generate references yields different results, correcting the *p*-value for the lack of sphericity [26] with Greenhouse-Geisser epsilon. These were performed separately for HC and non-HC contours, as the metrics would vary in opposite directions if the robustness of the generated volumes is improved by the selection process, and included GTVs and CTVs to increase statistical power. Post-hoc comparisons were performed with Wilcoxon's signed rank tests. Comparisons of contour metrics between exercise steps were performed with the Brunner-Munzel test for independent data [27].

All reported *p*-values are two-sided, with 0.05 as the chosen statistical significance threshold. The Benjamini-Hochberg [28] multiple comparisons correction was also applied to the abovementioned tests, accepting a 5% false discovery rate.

The proportion of contours with HD95 $\leq$3 mm was evaluated between steps with Fischer's exact test.

Software used for this work is detailed in the supplementary material.

Results

Eighteen physicians (15 centers) participated in the first step, fifteen physicians (14 centers) participated in the second step, of which eleven participated in both. This means that in total, 22 physicians from 19 centers (out of a total of 25 centers in Belgium and Luxemburg) participated in the delineation study. Additionally, there was a good geographic spread, representative of each region in Belgium, as well as a good balance between academic and non-academic centers. One physician sent identical contours for both steps. While only one set of these contours was included in the group categorization analyses to prevent artificial imbalance, both were considered in between-step comparisons. A diagram of the analyses process is presented on Fig. S5.

Qualitative analysis

The 32 different GTV and CTV contours were visually inspected and qualitatively categorized. Table 1 shows the distribution of observed deviations from guidelines and assignments in qualitative groups. For GTV, 5 groups were identified. 7 contours defining other nodes than those expected in station 4R as target were assigned to the first group, while 6 GTV delineations without any deviation constituted the HC-group. CTVs were categorized into 7 groups. Seventeen CTV contours whose GTV belonged to the two least-conforming groups were assigned to the first CTV group. Four CTV contours with the expected lowest-impact deviations (missing or partial crop to blood vessels and pericardium fold) had their respective GTV included in GTV HC and were used to constitute the CTV HC group. Table 1 also presents the different categorization of contours in qualitative groups between the exercise steps (p = 0.012 for GTV and p = 0.024 for CTV).

Quantitative analyses

Fig. 2 presents the DSC and RMSD between groups for the different reference volumes. Both for STAPLE and VOTE methods, the correlation of quantitative metrics with qualitative evaluations improved when discarding non-compliant contours from reference generation (Table 2). Visual and quantitative comparisons of the $VOTE_{HC}$ and $STAPLE_{HC}$ volumes showed remarkable agreement for GTV with a DSC of 0.98 and RMSD of 0.34 mm, while the CTV volumes were slightly more different (DSC 0.94, RMSD 0.87 mm). CTV $STAPLE_{HC}$ had converged on a single observer contour with imperfections that were not observed on the $VOTE_{HC}$ contour. Hence, further analyses used $VOTE_{HC}$ contours as reference. All quantitative results computed are found on Table S1. The prespecified maximum HD95 value of 3 mm had 100% sensitivity to identify HC delineations for both structures, with 85% and 90% specificity for GTV and CTV, respectively. Furthermore, using any HC contour as a reference to calculate HD95 also yielded 100% sensitivity to identify the other HC contours for both structures using the same 3 mm threshold, with 86% (95% CI 0.83–0.89) and 90% (95% CI: 0.84–0.93) specificity for GTV and CTV, respectively.

Quantitative comparisons confirmed statistically significant improvements of almost all metrics in step 2 (Table 3). The proportion of contours with HD95 under 3 mm went for GTV from 17% (3/18) in step 1 to 47% (7/15) in step 2 (p = 0.13) and for CTV from 6% (1/18) to 40% (6/15) (p = 0.03).

Discussion

This report describes a stepwise approach to assess IOV, i.e., an in-depth qualitative review followed by quantitative analyses of contours (Qualitative Before Quantitative approach, QBQ), applied to both steps of the ProCaLung delineation exercise. It builds on and improves previous initiatives on IOV. Van de Steene et al. proposed four groups of factors explaining GTV variability [29]. While differentiating involved nodes from other structures, pathological or not, is less common, likely owing to the improvement of imaging techniques, the remaining two groups explained almost all reasons for the IOV observed in this work: methodology and lack of knowledge. Our analyses additionally included the CTV, suggesting that similar factors are also the primary drivers of CTV IOV, also after providing TVD guidelines. Still, contours from the second step were significantly closer to what ProCaLung guidelines advise, confirming that detailed recommendations enable decreasing IOV, as previously reported [14]. However, benchmark case studies performed prior to clinical trials showed that providing instructions is not sufficient to ensure per-protocol TVD. In the PET-plan trial's dummy run, GTV volume and overlap metrics showed significant IOV before and after an additional teaching session [30]. The Southwest Oncology Group (SWOG), observed a median DSC for nodal GTV of only 0.35 as the target definition was highly uneven among participants [10]. We confirmed that agreement on target definition is a fundamental condition to guarantee similar contours. Data on nodal target volumes IOV outside the context of clinical trials are scarce. When analyzing the ESTRO delineation

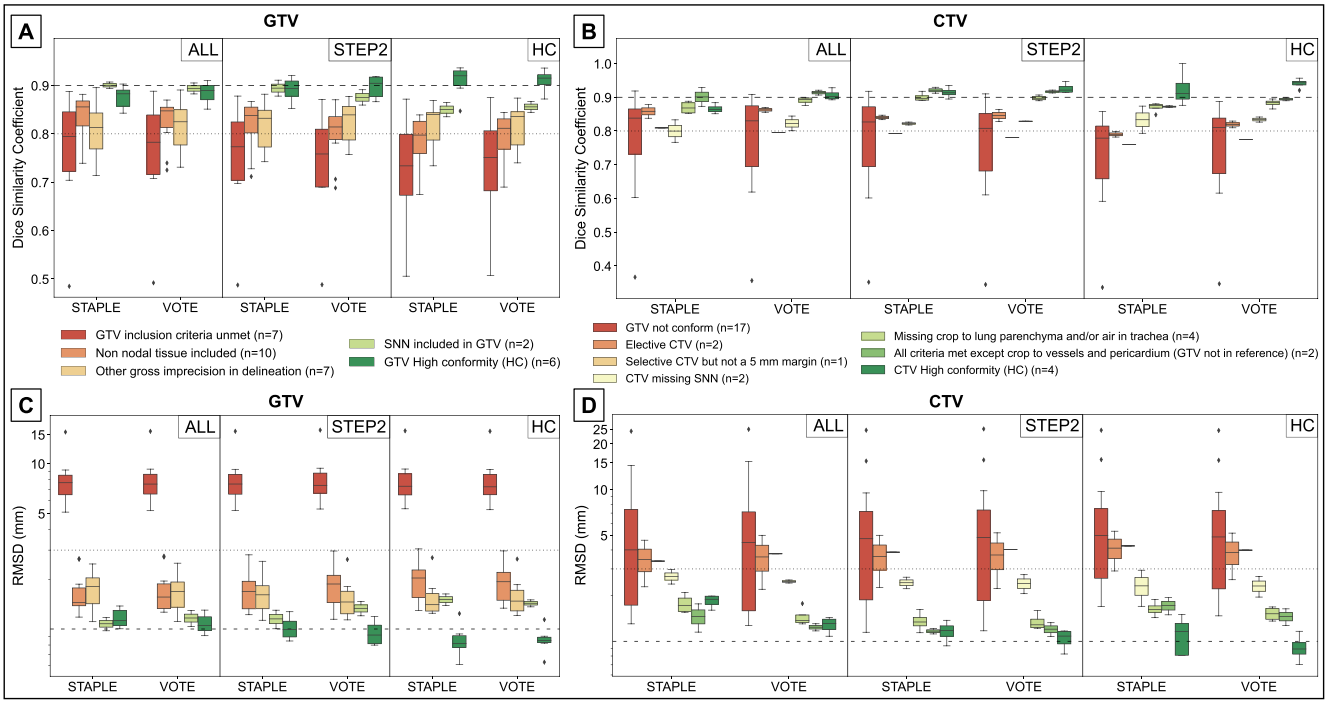

Fig. 2. Quantitative comparisons of qualitative groups per reference volume used. A and B: Dice Similarity Coefficient (DSC, lines at 0.8 and 0.9) – C and D: Residual Mean Square Distance (RMSD, lines at 3 and 1 mm). Selecting HC contours to generate the reference volume enables a better quantitative discrimination of qualitative groups. Abbreviations: ALL = All delineation exercise contours, CTV = Clinical Target Volume, GTV = Gross Tumor Volume, HC = contours meeting the largest number of qualitative criteria of conformity with ProCaLung guidelines, STEP 2 = contours received during the second step only.

Table 2
Discriminative power of the quantitative evaluation depending on the contour set used to generate the reference volume (GTV and CTV considered together).

Staple											
metric	Contourset	ALL (1)		STEP 2 (2)		HC (3)		RM ANOVAon ranks		Wilcoxonsigned-rank test	
		Median (95% CI IQR)		Median (95% CI IQR)		Median (95% CI IQR)		F	Adj p	(1) vs. (2)	(2) vs. (3)
DSC	HC (n = 10)	0.87	(0.85–0.89)	0.91	(0.88–0.92)	0.92	(0.89–0.93)	13.0	0.0007 *	0.007 *	0.21
	Others (n = 54)	0.85	(0.81–0.86)	0.84	(0.81–0.86)	0.80	(0.78–0.82)	27.9	<0.0001 *	0.031	<0.0001 *
RMSD	HC (n = 10)	1.36	(1.08–1.89)	1.10	(0.91–1.23)	0.89	(0.77–1.24)	13.9	0.0011 *	0.0051 *	0.11
	Others (n = 54)	2.17	(1.73–2.97)	2.12	(1.67–2.81)	2.54	(1.89–3.06)	9.26	0.0007 *	0.08	<0.0001 *
VOTE											
metric	**Contourset**	**ALL (1)**		**STEP 2 (2)**		**HC (3)**		**RM ANOVAon ranks**		**Wilcoxonsigned-rank test**	
		Median (95% CI IQR)		**Median (95% CI IQR)**		**Median (95% CI IQR)**		**F**	**Adj p**	**(1) vs. (2)**	**(2) vs. (3)**
DSC	HC (n = 10)	0.90	(0.88–0.91)	0.91	(0.89–0.92)	0.92	(0.91–0.94)	27.6	<0.0001 *	0.009 *	0.007 *
	Others (n = 54)	0.85	(0.80–0.87)	0.83	(0.81–0.85)	0.82	(0.80–0.84)	35.7	<0.0001 *	0.0001 *	<0.0001 *
RMSD	HC (n = 10)	1.16	(1.00–1.31)	1.00	(0.82–1.16)	0.86	(0.76–1.02)	20.1	0.0002*	0.0051 *	0.07
	Others (n = 54)	2.15	(1.61–3.01)	2.12	(1.60–3.50)	2.24	(1.86–3.37)	11.8	<0.0001 *	0.0015 *	0.21

Abbreviations: ALL = All delineation exercise contours, CI = Confidence Interval, CTV = Clinical Target Volume, DSC = Dice Similarity Coefficient, GTV = Gross Tumor Volume, HC = High-conforming contours, IQR = Interquartile range, RM ANOVA = Repeated Measures Analysis Of Variance, RMSD = Residual Mean Square Distance, STEP 2 = contours received during the second step only.
RM ANOVA on ranks results are reported with raw F-statistics and adjusted p-values. * = Significant after Benjamini Hochberg multiple testing correction for all RM ANOVA on ranks and all Wilcoxon tests, (3) vs (1) included and all significant, not shown.

courses, substantial variability was observed for the delineation of nodal GTV, as even the experts did not always agree on target definition [31].

The dosimetric impact of TVD IOV, evaluated in several studies and reported in a systematic literature review, revealed 'unacceptable' or 'major' deviations in up to 13.4% of plans, while the specific differences were usually not reported [32]. In our exercise, major deviations were considerably less common in the second step. However, deviations persisted both for GTV and CTV, mainly in relation to three structures: infra-centimetric PET-negative nodes within 5 mm distance from positive nodes (small neighboring nodes, SNN, Fig. 3-A/B), a pericardial recess (Fig. 3-C/D), and blood vessels. To the best of our knowledge, no published contouring recommendations for LA-NSCLC outside clinical trials addressed the matter of SNNs, and they were not specifically mentioned in ProCaLung guidelines either. While the nodal selection algorithm [2]

Table 3
Quantitative comparison of ProCaLung delineation exercise steps.

GTV	Step 1			Step 2			BM
metric	Median	IQR	Range	Median	IQR	Range	*p*-value
DSC	0.81	(0.71, 0.85)	(0.51, 0.91)	0.84	(0.81, 0.89)	(0.75, 0.94)	0.07
RMSD	2.34	(1.42, 5.63)	(0.85, 15.71)	1.36	(1.02, 1.72)	(0.63, 9.2)	0.01 *
HD95	5.79	(3.44, 21.23)	(2.0, 37.63)	3.07	(2.11, 4.3)	(1.38, 27.46)	0.01*
HD99	7.89	(4.84, 23.62)	(2.43, 39.65)	4.37	(3.51, 5.84)	(2.0, 29.37)	<0.01 *
HDmax	9.71	(6.43, 24.52)	(2.96, 40.96)	5.28	(4.66, 7.46)	(2.8, 31.27)	0.01*
CTV	**Step 1**			**Step 2**			**BM**
metric	**Median**	**IQR**	**Range**	**Median**	**IQR**	**Range**	***p*-value**
DSC	0.82	(0.69, 0.84)	(0.35, 0.89)	0.89	(0.84, 0.91)	(0.64, 0.96)	0.0008 *
RMSD	4.53	(2.28, 6.72)	(1.36, 24.55)	1.58	(1.22, 2.09)	(0.7, 9.59)	<0.0001*
HD95	12.7	(5.59, 21.05)	(2.23, 49.88)	3.41	(2.18, 4.8)	(1.38, 25.27)	<0.0001 *
HD99	15.43	(7.45, 23.88)	(3.09, 55.81)	4.12	(3.4, 6.52)	(2.0, 29.19)	<0.0001*
HDmax	17.25	(9.39, 25.39)	(4.8, 59.81)	5.88	(4.12, 7.93)	(2.8, 33.56)	<0.0001*

GTV and CTV contours of the exercise's second step are also quantitatively closer to the reference volume.
*Significant after Benjamini Hochberg multiple testing correction.
Abbreviations: BM = Brunner-Munzel two-sided test, DSC = Dice Similarity Coefficient, IQR = Interquartile range, RMSD = Residual Mean Square Distance.

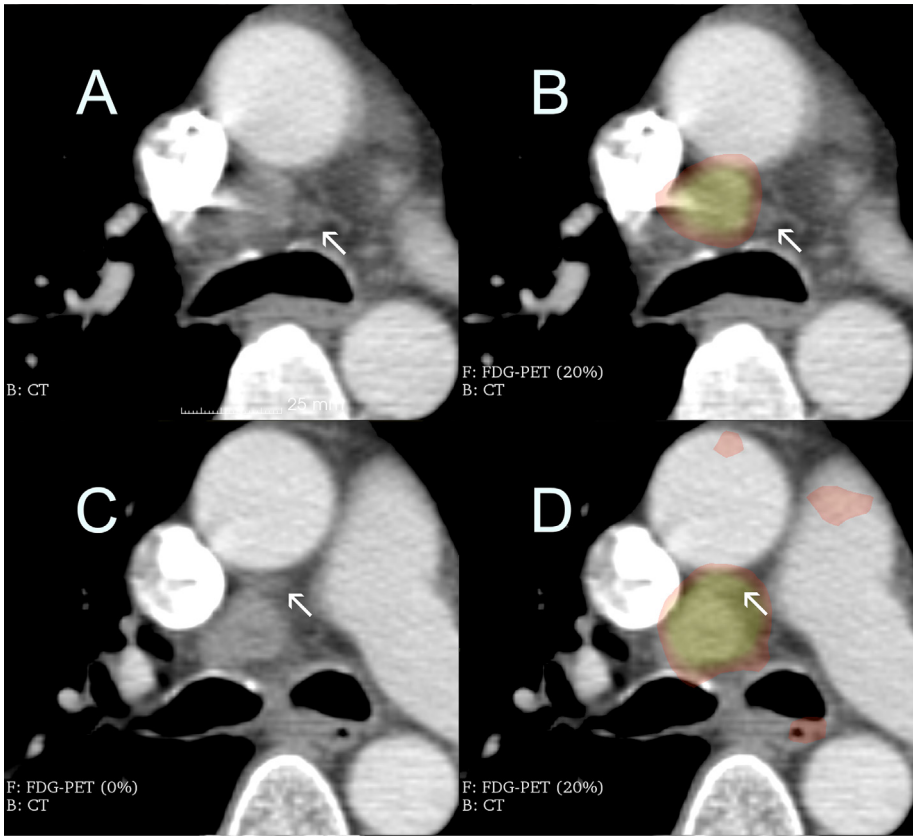

Fig. 3. Small Neighboring Nodes and pericardial fold. Frames A and B (with PET) show the same slice at the carina level. The white arrows indicate the Small Neighboring Nodes (SNN) which should not be included in GTV as per ProCaLung guidelines. On frames C and D, a few slices below A/B, the arrows point to the posterior aspect of the pericardial superior aortic fold, which should also not be included as target volume. Note that the longitudinal stretch of the PET signal possibly due to breathing motion (see Fig. 1) gives the appearance of high metabolic activity in this structure. Abbreviations: GTV = Gross Tumor Volume, PET = Positron Emission Tomography.

does not consider them as GTV, 40% of the second step contours did. Per definition, SNNs are within 5 mm of positive nodes and should therefore be included within the CTV margin and treated at the highest dose level. However, inclusion in the GTV means that the CTV contour would be extended by an additional unwarranted 5 mm CTV margin around them. Part of the pericardial superior aortic recess was also often included in GTV by physicians at both steps. Better knowledge of the mediastinal anatomy could therefore help reducing IOV. Moreover, although the pericardium is considered an organ-at-risk in RO, it is never specifically mentioned as a 'crop-to' structure, and no contour reviewed here excluded it completely. Also, cropping to blood vessels was always partial even though none of them showed signs of involvement. In the RETRACE study [7] of IOV in the adaptive setting for lung and head-and-neck patients, no delineation instructions were given. In contrast with head-and-neck cases, experts rarely edited the isotropic margins for lung cases, contrary to published guideline recommendations [20]. This may indicate an unusual approach in lung cancer TVD

practice, which could benefit from specific consideration within teaching initiatives.

The aforementioned issues (no information on how to handle SNNs and the absence of crop-to structures) may have influenced the IOV in this exercise. For this reason, we updated the protocol for the prospective ProCaLung study to include an algorithm on how to handle SNNs in the vicinity of the nodal CTV and a crop-to structure atlas. These adaptations should aid in reducing IOV by delivering a very precise protocol with little room for interpretation.

In studies where a reference volume is determined based on concordance of multiple observers, all available contours were usually used, even with reportedly significant differences between them [12,13,33], possibly limiting the robustness of subsequent calculations. As seen in RETRACE, the TVD variability between experts can be significant [7]. In the SWOG study, STAPLE was applied to six expert contours to analyze IOV between 17 other participants. However, the similarity of the experts' delineations was not reported. Our QBQ approach enabled the generation of volumes well-suited for a quantified expression of IOV when closeness to a reference is meaningful without assumptions on the observers' expert competence. This methodology selected the best contours based on actual performance rather than expectation. Without a detailed qualitative review, the most logical delineation set to use would have been STEP2, as instructions were available. Our reference volumes based on the HC sets (GTV and CTV) were, naturally, closer to each HC contour, but also further from the others, highlighting their validity using some of the most discriminating 3D metrics available. For CTV, the $VOTE_{HC}$ method was more robust to any one contour's imperfections than $STAPLE_{HC}$ was, probably due to the limited number of HC contours (four). In summary, we recommend performing a qualitative review before generating a reference volume for quantitative analyses and advise caution with small sample sizes.

This QBQ approach, which allowed to confirm the quantitatively higher conformity in the exercise's second step, also enabled to quantify the impact of qualitative criteria. Selecting a different set of nodes to be treated or including whole nodal stations as CTV explained the largest quantitative differences, but most deviations evaluated are noticeable. Unnecessarily larger volumes would result in higher doses to organs at risk, even from a small systematic unnecessary margin [34] such as the lack of CTV crop. We do believe that respecting a few well-defined qualitative criteria are sufficient to significantly limit the interobserver variability and its dosimetric impact.

QBQ requires comprehensive and stringent guidelines, which may limit the transferability to other clinical settings. Had the same approach been applied based on ESTRO-ACROP instead of ProCaLung guidelines, results would have been less informative. Nonetheless, the qualitative examination allowed the identification of a few criteria, and the constituted list will likely be further expanded when applied to larger sets of patients before it can be validated. Even though it initially is a time-consuming approach, it is rewarding, as it should thereafter be applicable to contour reviewing in any clinical, research, or educational setting.

A strength of this study is the prospective snapshot of target definition and delineation for LA-NSCLC across a country within numerous centers of varying sizes, academic or not, and physicians using their usual software, in a way that could be as close as possible to the real-world practice of a large panel of radiation oncologists. The similarity of high-conformity contours is encouraging for any TVD quality assurance program such as the planned prospective phase of ProCaLung (NCT04726358). Indeed, the metrics calculated when comparing submitted delineations to the reviewer's HC ones should be meaningful and trustworthy. ProCaLung will offer peer-review for actual patients in a real-life setting and is expected to further reduce the residual variability persisting after providing TVD guidelines, thus positively impacting the quality of NSCLC radiotherapy for the benefit of patients [16,31,35].

Conclusion

Variability of mediastinal nodal TVD was reduced in the second step of this nation-wide exercise. A qualitative review was essential to achieving meaningful quantitative comparisons of contours. Residual IOV supports the relevance of ProCaLung's upcoming peer-review initiative in Belgium. The aim is to implement a two-year peer-review project at national level to further reduce IOV, and improve quality of nodal mediastinal delineation in Stage III N2 NSCLC patients.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Y.L.: Personal fees from AstraZeneca (advisory board) and RaySearch (speaker fee), not related to this project. F.C., T.D., L.M., V.R., X.G., M.L.: None.

Acknowledgements

The authors would like to thank Eszter Hortobágyi for handling the practicalities of contour transfers with radiation oncology departments, as well as all healthcare professionals involved in their department's participation. This work was supported in part by a grant from "Les Amis de l'Institut Jules Bordet" (Belgium). Finally, we would like to thank Marianne Paesmans, MS, who has conducted a review of the implementation and results of the statistical analyses performed in this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2021.11.034.

References

[1] Perez CA, Pajak TF, Rubin P, Simpson JR, Mohiuddin M, Brady LW, et al. Long-term observations of the patterns of failure in patients with unresectable non-oat cell carcinoma of the lung treated with definitive radiotherapy report by the radiation therapy oncology group. Cancer 1987;59:1874–81.

[2] Peeters ST, Dooms C, Van Baardwijk A, Dingemans A-M, Martinussen H, Vansteenkiste J, et al. Selective mediastinal node irradiation in non-small cell lung cancer in the IMRT/VMAT era: How to use E(B)US-NA information in addition to PET–CT for delineation? Radiother Oncol 2016;120:273–8.

[3] Nestle U, Schimek-Jasch T, Kremp S, Schaefer-Schuler A, Mix M, Küsters A, et al. Imaging-based target volume reduction in chemoradiotherapy for locally advanced non-small-cell lung cancer (PET-Plan): a multicentre, open-label, randomised, controlled trial. Lancet Oncol 2020;21:581–92.

[4] Lo AC, Liu M, Chan E, Lund C, Truong PT, Loewen S, et al. The impact of peer review of volume delineation in stereotactic body radiation therapy planning for primary lung cancer: a multicenter quality assurance study. J Thor Oncol 2014;9(4):527–33.

[5] Lambrecht M, Clementel E, Sonke J-J, Nestle U, Adebahr S, Guckenberger M, et al. Radiotherapy quality assurance of SBRT for patients with centrally located lung tumours within the multicentre phase II EORTC Lungtech trial: Benchmark case results. Radiother Oncol 2019;132:63–9.

[6] Spoelstra FOB, Senan S, Le Péchoux C, 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. Int J Radiat Oncol Biol Phys 2010;76:1106–13.

[7] Apolle R, Appold S, Bijl HP, Blanchard P, Bussink J, Faivre-Finn C, et al. Inter-observer variability in target delineation increases during adaptive treatment of head-and-neck and lung cancer. Acta Oncol 2019;58:1378–85.
[8] Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. Radiother Oncol 2016;121:169–79.
[9] Growcott S, Dembrey T, Patel R, Eaton D, Cameron A. Inter-Observer Variability in Target Volume Delineations of Benign and Metastatic Brain Tumours for Stereotactic Radiosurgery: Results of a National Quality Assurance Programme. Clin Oncol (R Coll Radiol) 2020;32:13–25.
[10] Elhalawani H, Elgohari B, Lin TA, Mohamed ASR, Fitzgerald TJ, Laurie F, et al. An in-silico quality assurance study of contouring target volumes in thoracic tumors within a cooperative group setting. Clin Transl Radiat Oncol 2019;15:83–92.
[11] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging 2015;15. https://doi.org/10.1186/s12880-015-0068-x.
[12] Martin S, Johnson C, Brophy M, Palma DA, Barron JL, Beauchemin SS, et al. Impact of target volume segmentation accuracy and variability on treatment planning for 4D-CT-based non-small cell lung cancer radiotherapy. Acta Oncol 2015;54:322–32.
[13] Steenbakkers RJHM, Duppen JC, Fitton I, Deurloo KEI, Zijp L, Uitterhoeve ALJ, 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.
[14] Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J. Med Imaging Radiat Oncol 2016;60(3):393–406.
[15] Radiologists TRCo. Radiotherapy target volume definition and peer review – RCR guidance. 2017.
[16] Lewis PJ, Court LE, Lievens Y, Aggarwal A. Structure and processes of existing practice in radiotherapy peer review: a systematic review of the literature. Clin Oncol 2021;33:248–60.
[17] Joye I, Lambrecht M, Jegou D, Hortobágyi E, Scalliet P, Haustermans K. Does a central review platform improve the quality of radiotherapy for rectal cancer? Results of a national quality assurance project. Radiother Oncol 2014;111:400–5.
[18] Verhoeven K, Weltens C, Remouchamps V, Mahjoubi K, Veldeman L, Lengele B, et al. Vessel based delineation guidelines for the elective lymph node regions in breast cancer radiation therapy – PROCAB guidelines. Radiother Oncol 2015;114:11–6.
[19] Kindts I, Vandermeulen A, Verhoeven K, Laenen A, Hortobágyi E, Weltens CG. A central review platform improves the quality of regional lymph node delineation for breast cancer radiation therapy. Int J Radiat Oncol Biol Phys 2016;96:E38–9.
[20] Nestle U, De Ruysscher D, Ricardi U, Geets X, Belderbos J, Pöttgen C, et al. ESTRO ACROP guidelines for target volume definition in the treatment of locally advanced non-small cell lung cancer. Radiother Oncol 2018;127:1–5.
[21] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23:903–21.
[22] Steenbakkers RJHM, Duppen JC, Fitton I, Deurloo KEI, Zijp LJ, Comans EFI, 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;64:435–48.
[23] Heimann T, van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, Bauer C, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging 2009;28:1251–65.
[24] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 1993;15:850–63.
[25] Conover WJ, Iman RL. Rank Transformations as a Bridge between Parametric and Nonparametric Statistics. The American Statistician. 1981;35:124-9.
[26] Greenhouse SW, Geisser S. On methods in the analysis of profile data. Psychometrika 1959;24:95–112.
[27] Brunner E, Munzel U. The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. Biomet J 2000;42:17–25.
[28] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodol) 1995;57:289–300.
[29] 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;62:37–49.
[30] Schimek-Jasch T, Troost EGC, 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 cancerVerbesserung der Zielvolumendefinition durch einen Dummy Run mit interventionellem Konturierungstraining beim lokal fortgeschrittenen nicht-kleinzelligen Lungenkarzinom: Reduktion der Interobserver-Variabilität in klinischen Multizenterstudien. Strahlenther Onkol 2015;191:525–33.
[31] Mercieca S, Belderbos JSA, van Baardwijk A, Delorme S, van Herk M. The impact of training and professional collaboration on the interobserver variation of lung cancer delineations: a multi-institutional study. Acta Oncol 2019;58:200–8.
[32] Cox S, Cleves A, Clementel E, Miles E, Staffurth J, Gwynne S. Impact of deviations in target volume delineation – Time for a new RTQA approach? Radiother Oncol 2019;137:1–8.
[33] Jameson MG, Kumar S, Vinod SK, Metcalfe PE, Holloway LC. Correlation of contouring variation with modeled outcome for conformal non-small cell lung cancer radiotherapy. Radiother Oncol 2014;112:332–6.
[34] Giraud P, Elles S, Helfre S, De Rycke Y, Servois V, Carette M-F, et al. Conformal radiotherapy for lung cancer: different delineation of the gross tumor volume (GTV) by radiologists and radiation oncologists. Radiother Oncol 2002;62:27–36.
[35] Riegel AC, Vaccarelli M, Cox BW, Chou H, Cao Y, Potters L. Impact of multi-institutional prospective peer review on target and organ-at-risk delineation in radiation therapy. Pract Radiat Oncol 2019;9:e228–35.