Clinical and Translational Radiation Oncology 15 (2019) 83–92

Contents lists available at ScienceDirect

Clinical and Translational Radiation Oncology

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

Original Research Article

An *in-silico* quality assurance study of contouring target volumes in thoracic tumors within a cooperative group setting ☆

Hesham Elhalawani [a,*], Baher Elgohari [a], Timothy A. Lin [a,b], Abdallah S.R. Mohamed [a,c], Thomas J. Fitzgerald [d], Fran Laurie [d], Kenneth Ulin [d], Jayashree Kalpathy-Cramer [e], Thomas Guerrero [f], Emma B. Holliday [a], Gregory Russo [g,1], Abhilasha Patel [h], William Jones [h], Gary V. Walker [a,i], Musaddiq Awan [j,2], Mehee Choi [k,3], Roi Dagan [l], Omar Mahmoud [m,4], Anna Shapiro [n], Feng-Ming (Spring) Kong [o], Daniel Gomez [a], Jing Zeng [p], Roy Decker [q], Femke O.B. Spoelstra [r], Laurie E. Gaspar [s], Lisa A. Kachnic [t], Charles R. Thomas Jr. [u,*], Paul Okunieff [v], Clifton D. Fuller [a,*]

[a] Department of Radiation Oncology, University of Texas M.D. Anderson Cancer Center, TX 77030, USA
[b] Baylor College of Medicine, TX 77030, USA
[c] Department of Clinical Oncology and Nuclear Medicine, Alexandria University, Alexandria, Egypt
[d] Imaging and Radiation Oncology Core QA Center Rhode Island, University of Massachusetts Medical School, Worcester, Massachusetts, USA
[e] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Massachusetts, USA
[f] Department of Radiation Oncology, Beaumont Health System, Royal Oak, MI, USA
[g] Department of Radiation Oncology, Boston Medical Center, Massachusetts, USA
[h] Department of Radiation Oncology, University of Texas Health Sciences Center at San Antonio, TX, USA
[i] Department of Radiation Oncology, Banner MD Anderson Cancer Center, Gilbert, Arizona, USA
[j] Department of Radiation Oncology, Case Western Reserve University, OH, USA
[k] Department of Radiation Oncology, Northwestern University, IL, USA
[l] University of Florida Health Proton Therapy Institute, FL, USA
[m] Department of Radiation Oncology, University of Miami, FL, USA
[n] Department of Radiation Oncology, Upstate Cancer Center, SUNY Upstate Medical University, NY, USA
[o] Department of Radiation Oncology, University Hospitals Cleveland Medical Center, OH, USA
[p] Department of Radiation Oncology, University of Washington Medical Center, WA, USA
[q] Department of Therapeutic Radiology, Yale University School of Medicine, Connecticut, USA
[r] Department of Radiation Oncology, Amsterdam University Medical Centers, Vrije Universiteit, Amsterdam, The Netherlands
[s] Department of Radiation Oncology, Vanderbilt University, TN, USA
[t] Department of Radiation Oncology, Vanderbilt University Medical Center, Tennessee, USA
[u] Department of Radiation Medicine, Oregon Health & Science University, Oregon, USA
[v] SWOG, Department of Radiation Oncology, University of Florida College of Medicine, Florida, USA

ARTICLE INFO

Article history:
Received 6 December 2018
Revised 3 January 2019
Accepted 4 January 2019
Available online 6 January 2019

Keywords:
Target volumes
Contouring
Quality assurance

ABSTRACT

Introduction: Target delineation variability is a significant technical impediment in multi-institutional trials which employ intensity modulated radiotherapy (IMRT), as there is a real potential for clinically meaningful variances that can impact the outcomes in clinical trials. The goal of this study is to determine the variability of target delineation among participants from different institutions as part of Southwest Oncology Group (SWOG) Radiotherapy Committee's multi-institutional *in-silico* quality assurance study in patients with Pancoast tumors as a "dry run" for trial implementation.

Methods: CT simulation scans were acquired from four patients with Pancoast tumor. Two patients had simulation 4D-CT and FDG-FDG PET-CT while two patients had 3D-CT and FDG-FDG PET-CT. Seventeen SWOG-affiliated physicians independently delineated target volumes defined as gross primary and nodal

☆ A portion of the data presented in this manuscript was presented at the SWOG Radiation Oncology Committee Spring 2018 Meeting, San Francisco, CA, USA April 13, 2018 American Radium Society 100th Annual Meeting, Orlando, FL, USA May 8, 2018.

* Corresponding authors.
E-mail addresses: hmelhalawani@mdanderson.org (H. Elhalawani), roy.decker@yale.edu (R. Decker), thomasch@ohsu.edu (C.R. Thomas Jr.), cdfuller@mdanderson.org (C.D. Fuller).
[1] Present address: Section of Radiation Oncology, Dartmouth Hitchcock Medical Center and Norris Cotton Cancer Center, Lebanon, NH, USA.
[2] Present affiliation: Department of Radiation Oncology, Medical College of Wisconsin, Wisconsin, USA.
[3] Present affiliation: Loyola University Medical Center, IL, USA, 60153.
[4] Present affiliation: Radiation Oncology Department, Rutgers, the State University of New Jersey, New Brunswick, New Jersey, USA.

https://doi.org/10.1016/j.ctro.2019.01.001

QA
Pancoast tumor
Thoracic

tumor volumes (GTV_P & GTV_N), clinical target volume (CTV), and planning target volume (PTV).

Six board-certified thoracic radiation oncologists were designated as the 'Experts' for this study. Their delineations were used to create a simultaneous truth and performance level estimation (STAPLE) contours using ADMIRE software (Elekta AB, Sweden 2017). Individual participants' contours were then compared with Experts' STAPLE contours.

Results: When compared to the Experts' STAPLE, GTV_P had the best agreement among all participants, while GTV_N showed the lowest agreement among all participants. There were no statistically significant differences in all studied parameters for all TVs for cases with 4D-CT versus cases with 3D-CT simulation scans.

Conclusions: High degree of inter-observer variation was noted for all target volume except for GTV_P, unveiling potentials for protocol modification for subsequent clinically meaningful improvement in target definition. Various similarity indices exist that can be used to guide multi-institutional radiotherapy delineation QA credentialing.

1. Introduction

The advent of conformal radiotherapy, as well as the developing paradigm of image-guided radiotherapy, affords delivery of tumoricidal radiation doses to user-defined target volumes while minimizing dose to proximal organs at risk (OARs). Nonetheless, the steep dosimetric gradients involved mean that even minor geometric uncertainties may result in substantial dose variation which may in turn reduce delivered dose to the tumor and/or increase exposure to radiosensitive tissues [1,2].

In an effort to ameliorate toxicity, improve clinical outcomes and increase the overall therapeutic ratio, incorporation of multimodality imaging inputs in the target volume delineation process has become normative for specific organ sites. The addition of multimodality imaging parameters, such as fluorodeoxyglucose positron emission tomography fused with computed tomography (FDG PET-CT) [3], and magnetic resonance imaging (MRI), has been demonstrated in several series to improve reproducibility of target volume definition among users in selected anatomic applications [4,5]. Additionally, temporal imaging techniques, like 4D-CT or cine-MR have increasingly been explored as a mechanism to accurately localize the target and OARs, and/or reduction of geometric margins required to ensure adequate dose prescription (i.e. planning target volume 'PTV' reduction) [6].

Cooperative group studies that plan to implement these novel forms of image-guided radiotherapy (IGRT) in protocols necessitate consideration of target volume variability as a function of quality control [7]. Even minor geometric variation in target volumes may impact local control outcomes. Additionally, it is known that clinical trial radiation deviation may result in measurable decrement in clinical outcomes [8]. Furthermore, while medical physics quality assurance parameters face strict scrutiny, multicenter clinical trials may enroll patients from diverse radiation treatment centers. Radiation oncologists may vary regarding degree of expertise in target delineation generally, unknown acumen in target delineation within the anatomic region of interest, and/or unknown standard practices of incorporating either functional imaging (FDG PET, FDG PET-CT) or temporally-indexed imaging (4D-CT) into target definition[9]. Even when the aforementioned clinical practices are known, the practical consistency by a given center may still vary.

Efforts are increasingly made to formalize target volume strategies (e.g. the recent advent of standardized nodal atlases) [10,11], and educate radiation oncologists regarding specific expert-derived skillsets (e.g. ASTRO IMRT guidelines and symposia) [12]. Nonetheless, there remains scant data regarding how to optimize clinical trial protocol implementation of multimodality functional imaging combined with temporal information (e.g. 4D-CT) [13]. This represents an unmet need given the fact that 4D-CT is currently the standard-of-care for RT planning of potentially curative lung cancers [14]. Furthermore, a recent systematic review by Frood et al. [15] shed some light on the discrepancies in target volume definition using 4D FDG PET-CT versus 3D FDG PET-CT.

Moreover, quality assurance studies using computer-estimated consensus contours have been tested in multiple cancer sites [16–18], however -to our knowledge- not specifically in Pancoast tumors. Thus, in addition to testing the feasibility of computer-generated consensus contouring with multimodality functional and/or temporal imaging, we sought to assess inter-observer variability in the setting of Pancoast tumor contouring. Guideline adherence may theoretically reduce the risk of radiation-induced toxicities such as brachial plexopathy [19] or pneumonitis [20].

Consequently, the Southwest Oncology Group (SWOG) Radiation Treatment Committee designed this prospective *in-silico* quality assurance pilot study to determine how diverse institutions participating in clinical trials are incorporating FDG PET-CT, 3D-, and 4D-CT data in RT planning. Additionally, we sought to provide insight into how such integration might be optimized, standardized, and scrutinized by central review boards, like the IROC Rhode Island (formerly QARC).

The specific aims of this study are to:

1. Determine the feasibility of multi-site electronic data collection methods for evaluation of target delineation and user survey.
2. Summarize the effect of 4D-CT incorporation upon target volume delineation as assessed by custom target delineation evaluation software.
3. Determine specific criteria for target volume "credentialing" in trials where IMRT is to be implemented.
4. Generate pilot data and testable hypotheses for future research efforts.

2. Material and methods

This study was approved by the SWOG Radiation Therapy Committee for execution [21], and started accruing radiation oncologists in spring 2012. The study was designed in a multi-user fashion combining objective evaluation of specific target delineation and dose parameters with user reported data. Pancoast tumors were selected as test cases for this concept study; given the relatively strong agreement on target and OARs delineation among radiation oncologists [22,23]. Radiotherapy guidelines were drafted and multimodality simulation images were acquired as a part of a then-existing SWOG Pancoast tumor clinical trial protocol.

2.1. Data

2.1.1. Scans and simulation guidelines

Non-contrast-enhanced CT simulation scans–both bone and soft tissue windows- were acquired from four patients with Pancoast tumors from a single center. These DICOM (Digital Imaging and Communications in Medicine) files belonged to actual patients treated with radiotherapy (RT), albeit without any Health Insurance Portability and Accountability Act of 1996 (HIPPA)-defined patient-specific information [24]. This was performed by anonymizing patient DICOM datasets using 'DCMAnonymize' DICOM Validation Toolkit (https://www.dvtk. org). This software tool removes all patient specific data (e.g. name, ID codes, date of scan, gender) from DICOM file headers.

Two patients had simulation 4D-CT and 18-FDG-FDG PET-CT scans while two patients had simulation 3D-CT and FDG-FDG PET-CT scans. The 4D-CT scans were contoured using an average CT that was generated by averaging 4D CT acquisitions over ten phases of the full breathing cycle. These scans were made available on the IROC Rhode Island (formerly QARC), website at www.QARC.org for participants to download scans, instructions and participation survey. IROC Rhode Island provides radiation therapy, diagnostic imaging and data management services for SWOG among other National Clinical Trials Network (NCTN) institutes [25] (Fig. 1).

2.1.2. Summary of thoracic radiotherapy

All contouring efforts for this study were undertaken on participants' treatment planning system (TPS) of choice, provided the capacity for DICOM-RT export was present.

Participants were supplied with relevant imaging and clinical evaluation information to help them define target volumes precisely. Additionally, a priori definitions for target volumes, organs at risk, dose prescription and constraints were specifically provided as a part of broader consensus guidelines. For example, clinical target volume (CTV) for radiation fields was defined as the primary tumor, plus *involved* mediastinal, ipsilateral paratracheal lymph nodes (levels 2 and 4), and ipsilateral supraclavicular nodes. The total planned dose was set to be 54 Gy prescribed to an isodose line that encompassed the planning target volume (PTV) and that satisfies the dose uniformity guidelines. Participants were tasked to create an intensity-modulated RT (IMRT) plan for a treatment course that was to be delivered in 30, once daily fractions of 1.8 Gy each, to a total dose of 54 Gy to the PTV. The percent of normal lung volume receiving 20 Gy or more (V20) must be kept at less than 37%. 54 Gy was chosen in anticipation of testing this dose in a future prospective, cooperative group combined-modality clinical trial for Pancoast tumors.

2.1.3. Target volume definition

To that end, the definitions of tumor and target volumes as well as guidelines for delineation were clearly provided for participants. These were in accordance with International Commission on Radiation Units and Measurements (ICRU) Reports 50 and 62 in order to mitigate sources of delineation congruency among participants [26,27]. The instructions included standardized nomenclature of target volumes, hence facilitating subsequent analysis in a time-efficient manner. Tumor volumes encompassed gross tumor volume (GTV) either of the primary tumor (GTV_P) or the clinically positive lymph nodes (GTV_N). Other target volumes included: clinical target volume (CTV), planning target volume (PTV), and internal target volume (ITV). Detailed descriptions are depicted in Table 1. Of note, the fourth set of scans belonged to a patient with no nodal disease, hence no GTV_N delineation was required.

2.1.4. Prescribed doses and fractionation guidelines

The total planned dose was set to be 54 Gy prescribed to an isodose line that encompassed the PTV and that satisfied the dose uniformity guidelines. The 54 Gy dose was based upon SWOG recommendations to incorporate higher doses than the classic 45 Gy (per SWOG-9416 and SWOG-0220) for Pancoast trials [28–30]. This was to be planned as thirty, once daily fractions of 1.8 Gy. The entire PTV was to receive at least 93% of the protocol dose and a contiguous volume of no>2 cc within the PTV was not to exceed 120% of the protocol dose [31].

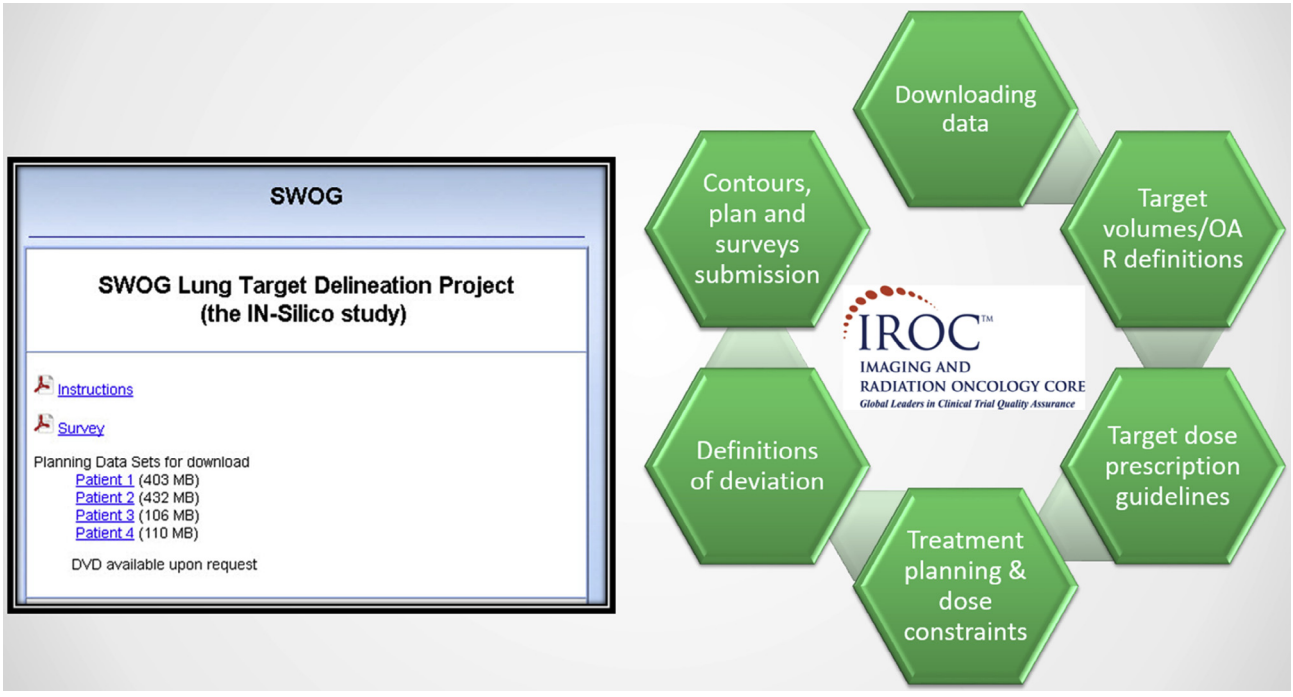

Fig. 1. SWOG *in-silico* study tasks completed via the IROC Rhode Island platform.

Table 1
Definitions and specified designations for target volumes of interest: instructions for participants.

Target volume	Definition	Specified designation
Gross tumor volume (GTV)	The primary tumor (**GTV_P**) and clinically positive lymph nodes (**GTV_N**) seen on the pretreatment PET scan (standardized uptake value; SUV > 3), diagnostic CT scan, and/or treatment planning CT (>1 cm short axis diameter) will comprise the GTV. The GTV will always be located in the apex of the ipsilateral lung in this particular study. This volume(s) may be disjointed.	GTV_P GTV_N
Internal target volume (if used)	The **ITV** includes the envelope that encompasses the tumor motion for a complete respiratory cycle.	ITV
Clinical target volume	The **CTV** is defined to be the GTV plus a 0.5 cm to 1 cm margin as appropriate to account for microscopic tumor extension. The ipsilateral paratracheal lymph nodes (levels 2 and 4) and supraclavicular fossa lymph nodes will be defined as comprising part of the CTV for this protocol. If an ITV approach is used then the ITV plus 0.5 cm to 1 cm is added to the ITV to form the CTV. Elective treatment of the entire mediastinum will not be done. If a uniform margin is used for CTV expansion, please specify **CTV_xxMM**, where xx = the margin expansion in millimeters, as structure name (e.g. **CTV_05MM** denotes a 5 mm expansion margin); if a non-uniform margin is used, please designate the volume **CTV.**	CTV
Planning target volume	The PTV margin should account for setup uncertainties and may be individualized. The PTV will comprise the CTV with a minimum 0.5 cm (if daily imaging correction will be used), or a minimum 1.0 cm if daily imaging will not be performed. If a uniform margin is used for PTV expansion, please specify **PTV_xxMM**, where xx = the margin expansion in millimeters, as structure name (e.g. **PTV_05MM** denotes a 5 mm expansion margin); if a non-uniform margin is used, please designate the volume **PTV.**	PTV_xxMM

2.2. Participants

Seventeen SWOG-affiliated radiation oncology participants (physicians), with median career experience of 11 years (IQR: 2.5–18.75), independently delineated treatment target volumes as well as OARs. Each participant used his/her preferred TPS for target delineation. Moreover, six board-certified thoracic radiation oncologists were designated as the 'Experts' for this study. Selection criteria for 'Experts' included all the following: (1) board certification in radiation oncology; (2) a minimum of 10 years of practice; Thoracic radiation oncology subspecialty. After plan completion, each participant, identified by a specific ID and login, submitted the plan via the internet using FTP transfer (www.qarc.org). Recorded information extracted from the contouring session were then available for central analysis.

2.3. Data submission and curation

Target, plan, and dose data were collected centrally via the QARC platform. Analysis was performed using quantitative software-assisted analysis of contour, dose and imaging data. Each participant was asked to fill out an on-line survey assessing subjective responses to contouring tasks. This survey was made available via QARC website. It included categorical Likert scale evaluation of institutional practices regarding FDG PET/CT and 4DCT incorporation into target delineation, as well as subjective responses regarding the implemented target delineation instruction materials, TPS used, applied software tool for FDG PET-CT registration and 4D-CT analysis, among others (**Supplementary Material 1**).

After data collection was completed, plan suitability analysis was performed. Designated Experts evaluated each contouring session in order to determine if the plan was clinically acceptable without modification, or whether it represented a major or minor deviation from expert-determined acceptability, using criteria as specified in the initial protocol.

2.4. Imaging analytics and evaluation metrics

The six designated Experts' contours were incorporated into the final definition of target volumes to serve as comparators for subsequent analysis. This expert composite, i.e. ground truth, was constructed using Warfield's Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm, a feature embedded in Elekta ABAS (Atlas-based auto-segmentation software) (http://www.elekta.com/ABAS). Segmentation simply implies "classifying" whether each image voxel belongs to the volume of interest or the background. The STAPLE algorithm allows for multiple segmentations fusion by automatically estimating the segmentation

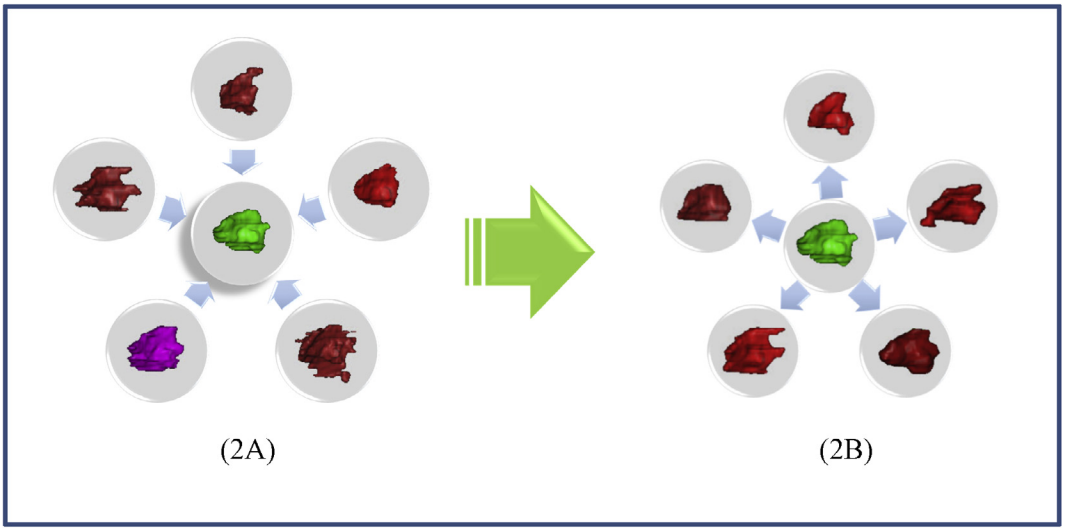

Fig. 2. (A) Defining 'ground truth' and (B) comparing individual contours against 'ground truth' (The green volume in the center of each figure represents the Experts' composite). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

quality (as characterized by the sensitivity and specificity parameters) of each classifier while simultaneously deriving a weighted combination of multiple classifiers [32]. This approach provides a probabilistic estimate of the true delineation and a measure of the performance level represented by each expert user (Fig. 2A).

Contours by each of 17 participants were then compared with the experts' STAPLE composite contour, i.e. ground truth (Fig. 2B). The following evaluation metrics were used to assess delineations similarity:

A. *Dice Similarity Coefficient* [33]: a spatial overlap index and a reproducibility validation metric. The value of a DSC ranges from 0, indicating no spatial overlap to 1, indicating complete overlap between two sets of binary segmentation results [34] (Fig. 3**A**),
B. *Hausdorff metric*: gives the mean (Mean Surface Distance) or the largest length (Maximum Hausdorff Distance) out of the set of all distances between each point of a set (individual contour) to the closest point of a second set (expert composite) [35] (Fig. 3**B**),
C. *Volume Overlap Ratio* (VOR, or the Jaccard similarity coefficient) was computed to represent the ratio of the volume of intersection to the volume of union per target volume (TV) per case [36] (Fig. 3**C**). All three aforementioned metrics were calculated using a commercially available image registration software (VelocityAI™ 3.0.1)
D. *Intraclass correlation coefficient* (ICC) was calculated as a measure of concordance of clustered segmentations between participants using **IBM SPSS Statistics 22.0.** The ICC is basically a signal-to-noise ratio, where higher values close to 1 indicate higher concordance between volumes of the same group [37].

2.5. Statistical analysis

Based on estimates from our previous pilot series [7], power and sample size analysis (G*Power 3 statistical software) [38] was performed, assuming a minimum possible asymptotic relative efficiency of $\geq$0.86423, using an a priori power goal 1-β of 0.8, non-Bonferroni corrected one-tailed $\alpha = 0.5$, for detection of an effect size of 0.8 (large effect). This resulted in a minimum requisite sample size of at least 16 participants, hence the inclusion of 17 radiation oncologists.

Statistical assessment was performed using JMP v 11Pro (SAS institute, Cary, NC). One-way analysis of variance (ANOVA) was applied to assess whether there were any statistically significant differences among the mean values of evaluation metrics per target volume. Additionally, the differential impact of incorporating 3D-CT versus 4D-CT on delineation accuracy was assessed using the ANOVA test.

3. Results

Seventeen SWOG-affiliated physicians, with median career experience of 11 years (IQR: 2.5-18.75) participated in this study. Of note, four out of 17 participants didn't submit the questionnaires.

In this paper, we're reporting target volume variation, namely: GTV_P, GTV_N, CTV, and PTV in comparison to reference expert composites. We excluded 'ITV' from our analysis as the vast majority of participants didn't assign them. More importantly, there were not enough expert 'ITV' contours per case to feed into the STAPLE algorithm to define the 'ground truth'. Almost all other target volumes in all 4 cases were not unanimously retrievable from QARC database or contoured by 17 participants (range: 9–17 contours/target volume/case).

Moreover, GTV_N contours showed some inter-observer variability of included thoracic nodal stations. One example was GTV_N in case #1. Out of 12 available delineations of GTV_N: 2 participants delineated (1L) nodal station, 5 participants included stations (3A) and (6), while the rest included all 3 stations. All contours were included in the analysis as long as there were sufficient expert contours to serve as an input to the STAPLE algorithm to define the ground truth.

For each case, all 4 evaluation metrics mentioned in the 'Methods' section were derived by comparing individual participants' target volumes to the corresponding Experts' STAPLE composite. Median values and corresponding interquartile ranges are also provided in Table 2. When compared to Experts' STAPLE, GTV_P had the best agreement among all participants with median DSC of

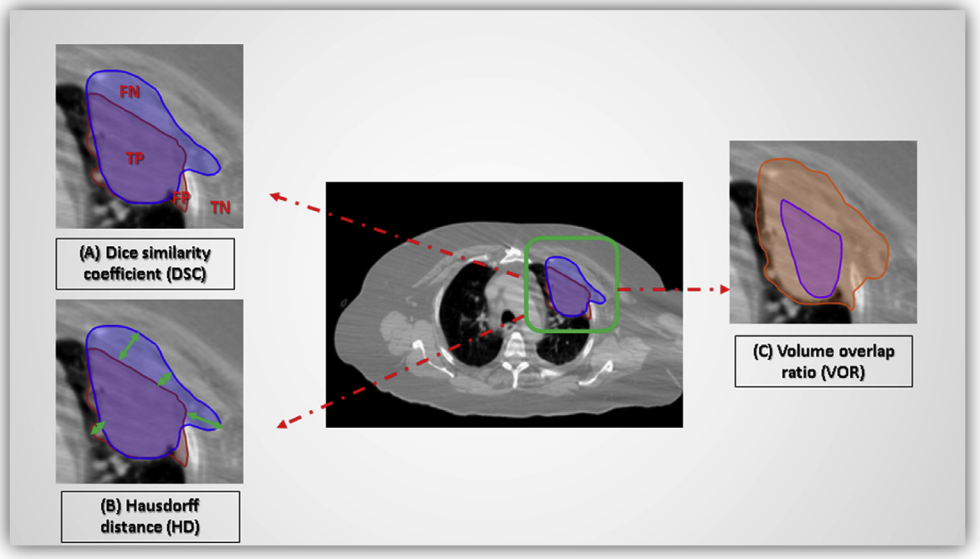

Fig. 3. Evaluation metrics for assessment of delineations similarity; (A) Dice similarity coefficient, (B) Hausdorff distance (HD), and (C) Volume overlap ratio. (FN: false negative; TP: true positive; FP: false positive; TN: true negative).

Table 2
Median values of individual metrics of all-participants' delineations across all cases; including interquartile ranges (DSC = Dice coefficient; HD = maximum Hausdorff distance; MSD = Mean Surface Distance; GTV_P = gross primary tumor volume; GTV_N = gross nodal tumor volume; CTV = clinical target volume; PTV = planning target volume; mm = millimeters; IQR = Inter-Quartile Range).

	GTV_P	GTV_N	CTV	PTV
Number of participants				
Case 1	13	12	16	17
Case 2	14	9	16	16
Case 3	15	13	15	15
Case 4	16	0	16	16
DSC median (IQR)	0.87 (0.78–0.94)	0.35 (0.10–0.46)	0.80 (0.72–0.90)	0.81 (0.70–0.90)
HD median in mm (IQR)	12.6 (10.09–17.75)	68.27 (41.20–80.31)	43.73 (23.48–64.90)	34.22 (24.80–59.18)
MSD median in mm (IQR)	0.37 (0.15–0.81)	17.5 (8.5–25.45)	2.03 (0.40–4.31)	1.56 (0.64–3.98)
VOR (%)				
Case 1	21.9	7.1*	14	13
Case 2	67.1	0	46.4	48.1
Case 3	41	0	23.8	28
Case 4	7.2	N/A**	4.4	9.9

* Given the inter-observer variability in selection of included lymph node stations, this VOR was calculated based on the volumes segmented by group of participants with most harmonious approach (6 out of 12).
** N/A: Not Applicable because this scan belongs to a patient with no nodal disease.

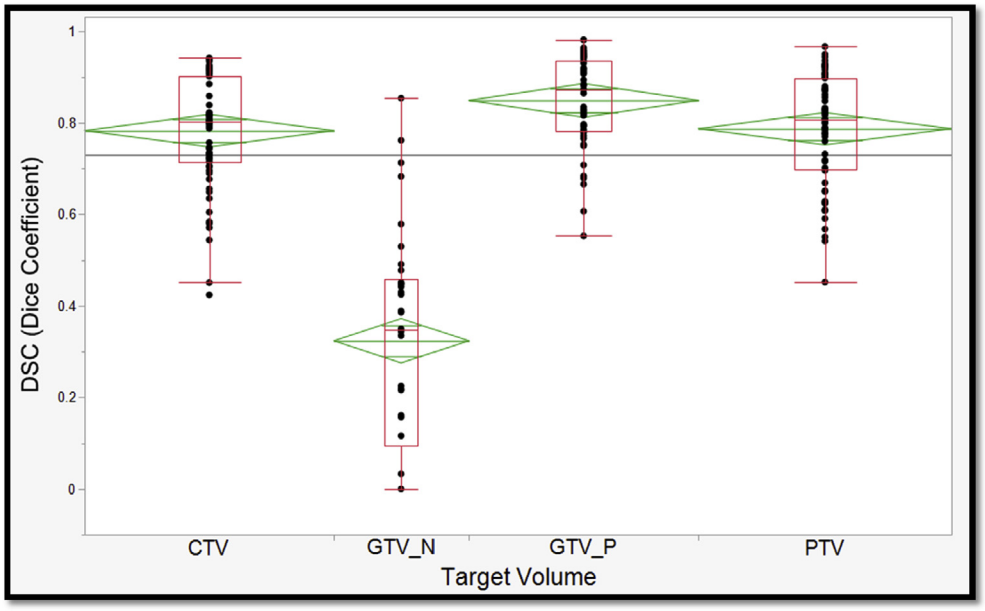

Fig. 4. One-way analysis of Variance of Dice similarity coefficient across target volumes. (Green diamonds represent mean and standard deviation; and red boxes encompass interquartile ranges with the transverse line representing median value). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

0.87. On the contrary, GTV_N demonstrated the lowest agreement with median DSC of 0.35 as shown in Fig. 4. Likewise, GTV_P demonstrated the lowest median HD & MSD values; 12.6 mm and 0.37 mm, respectively. Whereas, GTV_N showed the largest median HD and MSD values; 68.27 and 17.5, respectively (Figs. 5 and 6).

In between lies PTV and CTV which were associated with the second and third best median DSC, respectively (Fig. 4). Similarly, the second and third highest HD and MSD were reported with CTV and PTV as tabulated in Table 2 and shown in Figs. 5 and 6. The median VOR for all TVs in the included 4 cases was 0.14 (range 0–0.67). Along the same lines, the VOR was consistently highest for GTV_P and lowest for GTV_N across all 4 cases. Given the inter-observer variability in selection of included lymph node stations, the GTV_N VOR was calculated based on the volumes segmented by group of participants with most harmonious approach (6 out of 12). Intraclass correlation coefficients (ICC) for DSC, HD, MSD metrics are also reported in Table 2.

Interestingly, there were no statistically significant differences in all studied metrics for all TVs for cases with 4D-CT (i.e. contoured on average CT) versus cases with 3D-CT simulation scans (Fig. 7). Table 2 depicts the number of participants' delineations for each of the studied TVs as well as medians and interquartile ranges (IQRs) of all individual metrics derived from all cases.

4. Discussion

Target delineation is a key element in modern radiotherapy treatment planning, particularly following the introduction of intensity modulated radiotherapy (IMRT)[39]. Accurate target identification has become as important as beam delivery to ensure tumor coverage, and reduce unnecessary normal tissue exposure [40]. Despite efforts to provide guidelines for more reproducible target delineation across different institutes [14,16,41], inter-observer variations are still a barrier that faces multi-

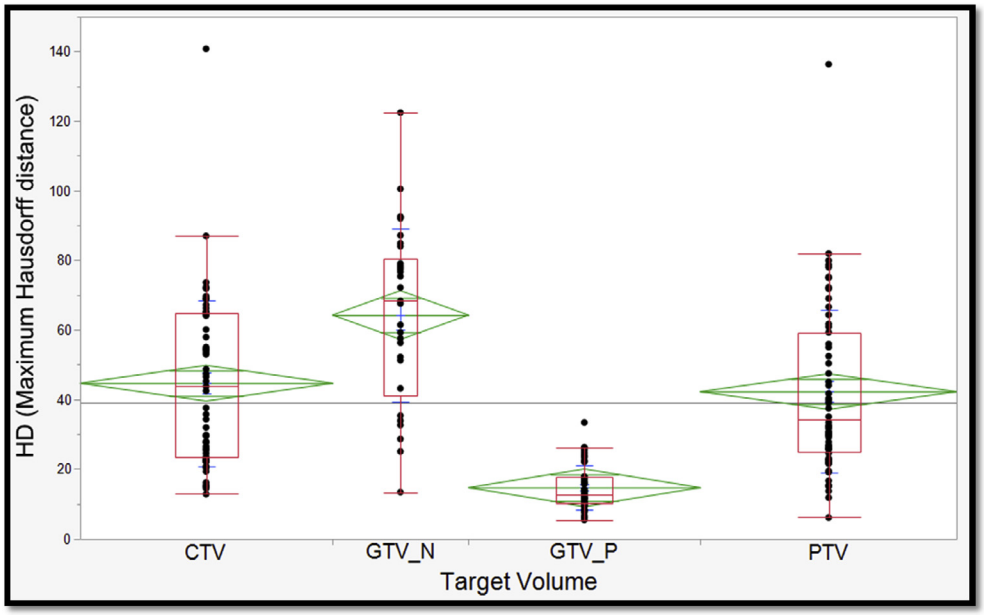

Fig. 5. One-way analysis of Variance of maximum Hausdorff distance across target volumes. (Green diamonds represent mean and standard deviation; and red boxes encompass interquartile ranges with the transverse line representing median value). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

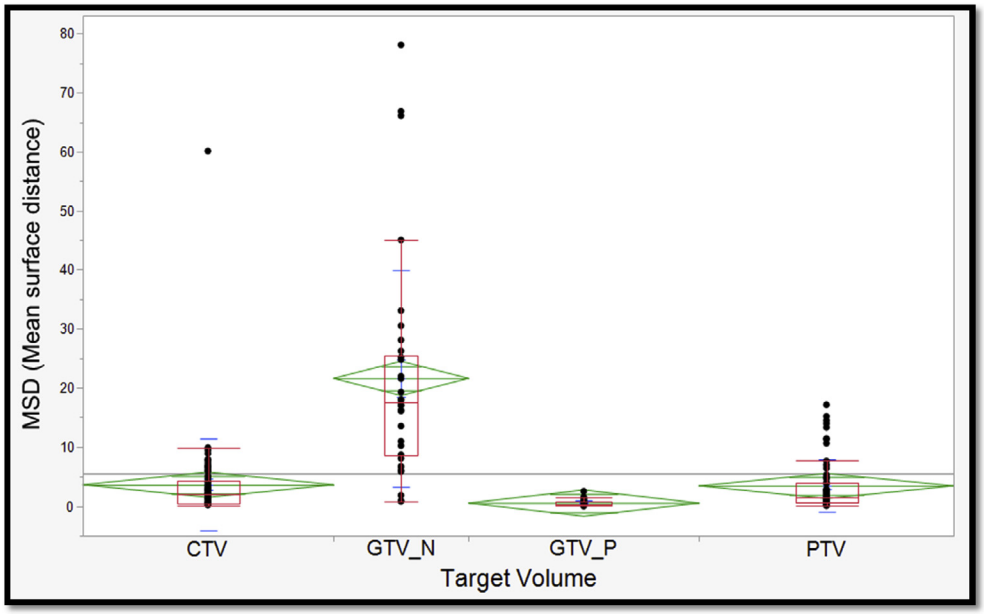

Fig. 6. One-way analysis of Variance of mean surface distance (MSD) across target volumes. (Green diamonds represent mean and standard deviation; and red boxes encompass interquartile ranges with the transverse line representing median value). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

institutional trials [2]. Another barrier is the absence of an optimized protocol to account for different imaging modalities when it comes to RT planning for Pancoast tumors [22]. Consequently, our group sought to quantitatively determine the inter-observer variability of expert radiotherapy target-volume delineation for Pancoast tumors, towards developing an expert-consensus contouring atlas.

Per our study, we have shown that the best agreement was in GTV_P across all participants when compared to experts' STAPLE with a median Dice coefficient of 0.873 (range 0.781–0.935). Also, we showed that the lowest agreement was in GTV_N with a median DSC of 0.347 (range 0.095–0.458). No statistically significant difference was shown between different imaging simulation scans (i.e. 4D-CT and 3D-CT).

Some limitations faced our study including long time taken by experts and other participants to complete the needed contours (an average of 18 months). The 4D-CT scans were contoured using only a single static time point (i.e. different phases of respiration were not accounted for), which may explain why a statistically significant difference was not detected between 3D-CT and 4D-CT simulation studies. Of note, participants used FDG PET images to define tumor volumes for all four cases. Consequently, investigat-

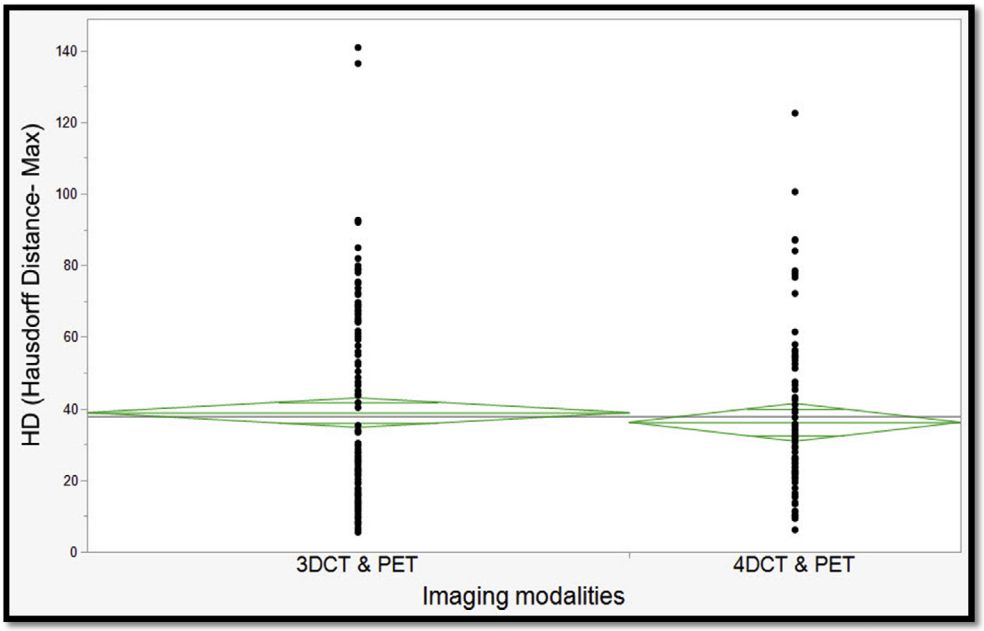

Fig. 7. One-way analysis of Variance of maximum Hausdorff distance by simulation imaging modality. (Green diamonds represent mean and standard deviation). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ing the value of FDG PET fusion in improving or standardizing target volumes delineation was not feasible. The selection of participants was limited to SWOG institutions which might represent a selection bias. Some target volumes contours were not included in the analysis of all 4 cases. This was attributed to either available contours being not retrievable from IROC Rhode Island database or contours not originally made by all 17 participants (range: 9–17 contours available for analysis per target volume for each case).

The instruction set assigned GTV_N to the clinically positive lymph nodes seen on pre-treatment PET scan (SUV > 3), diagnostic CT scan, and/or treatment planning CT (>1 cm short axis diameter). However, wide inter-observer variability in GTV_N contours was observed. This is chiefly attributed to neither enumerating affected thoracic nodal stations in the instruction set nor suggesting standardized reference lymph node mapping atlas; another two areas for improvement for future studies [11,42]. A similar study by Mercieca et al also named interpretational differences among observers as the leading cause for large variations in GTV_N delineation [43]. For example, in case #2, among the 9 participants who delineated GTV_N, only 3 participants included the same nodal stations. On the other hand, the remaining 6 participants included partly or completely different nodal stations in 5 different permutations. This significantly affected the interpretability of the overlapping/similarity metrics we selected for evaluating inter-observer delineation agreement. Similarly, GTV_N VOR was calculated as zero for cases #2 and #3.

To avoid this pitfall in case #1, GTV_N VOR was calculated based on volumes segmented by group of participants with most harmonious choice of nodal stations (6 out of 12). Also, it was difficult to interpret DSC for GTV_N given the small size of lymph nodes and the multiple non-contiguous objects within one VOI, i.e. one GTV_N volume may encompass multiple affected lymph nodes. Hausdorff distance may represent an appropriate alternative in such a situation where the probability of overlap is small [44].

Nonetheless, this study represents one of the earliest multi-institutional efforts to automate target volume delineation quality assurance for Pancoast tumors. To that end, fully anonymized inter-institutional data sharing was optimized via the IROC Rhode Island platform. Additionally, we provided all the participating radiation oncologists with contouring guidelines, target volumes definitions and standardized nomenclatures.

Our work follows previous SWOG efforts made to address inter-observer variability across different organs sites [9]. In previous work, we showed that educational intervention with a SWOG-approved consensus atlas reduced inter-observer variability in rectal cancer target volume delineation [7]. However, that study was performed using a CT-only dataset. The goal of the present study was quantitative assessment of variation in target volumes delineation for Pancoast tumors in a clinical trial group setting.

Previous work by other groups suggested that atlas and real time feedback hugely improve OARs delineation in head and neck cancers [45]. Similar results were reported by the Australasian Gastrointestinal Trials Group (AGITG), showing that planning guidelines could be significantly improved and help radiation oncologists optimize IMRT delivery in anal cancer [41]. Another study was also conducted to assess atlas implementation effect on anorectal target delineation, and also showed that use of atlas and contouring guidelines do reduce inter-observer variability [46].

The authors of this study are planning to expand these pilot data into a protocol designed to serve as a template for future SWOG studies involving radiotherapy for Pancoast tumors. By providing clear instructions, well-defined standardized criteria and subsequent calibration tool for target delineation, future multi-institutional studies could execute credentialing/QA efforts. That would invariably facilitate the task of central review boards for multi-institutional clinical trials in a timely- and cost-effective manner. These results can also be integrated into development of educational tools for residency training programs or continued medical education programs to help clinicians improve their target delineation skillset [47].

5. Conclusions

In conclusion, a relatively high degree of inter-observer variation was noted for all target volumes except for GTV_P, revealing

potentials for future protocol modifications to improve the accuracy, and reduce the variability of target volume definitions. This study helps explain the variance that can be seen between radiation oncologists treating the same patient and thus provides a tool to better understand and assess the acceptable range of treatment volumes satisfactory to meet QA standards.

6. Link to research Data

https://www.qarc.org/

7. Co-authors specific contributions

All listed co-authors performed the following:

1. Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work;
2. Drafting the work or revising it critically for important intellectual content
3. Final approval of the version to be published
4. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Specific additional individual cooperative effort contributions to study/manuscript design/execution/interpretation, in addition to all criteria above are listed as follows:

- HE – Manuscript writing, statistical analysis, imaging data curation, and segmentation analysis
- BE, TL – Manuscript writing
- ASRM – Statistical analysis, and segmentation analysis
- RD,TJF, FL, KU, JKC, TG – data management and curation, platform support through IROC, and oversight of segmentation analysis
- ASRM, EBH, GR, AP, WJ, GW, MA, MC, RD, OM, AS, SK, DG, JZ, FS, LG, CRT– Image segmentation, contributory analytic support, clinical review and review of manuscript.
- LK, PO: contributory analytic support, clinical review and review of manuscript, conceptual feedback and support
- CRT, CDF: Co-corresponding authors; conceived, coordinated, and directed all study activities, responsible for data collection, project integrity, manuscript content, and editorial oversight and correspondence; direct oversight of trainee personnel. Statistical analysis and guarantors of statistical quality.

Acknowledgements

Multiple funders/agencies contributed to personnel salaries or project support during the manuscript preparation interval. This work was funded by the SWOG/Hope Foundation Dr. Charles A. Coltman, Jr. Fellowship Program Grant. Dr. Elhalawani is supported in part by the philanthropic donations from the Family of Paul W. Beach to Dr. G. Brandon Gunn, MD. Drs. Elhalawani and Fuller received salary support from National Institutes of Health (NIH)/National Cancer Institute (NCI) Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50 CA097007-10). Drs. Fuller and Mohamed receive(d) salary support from NIH, including: National Institute for Dental and Craniofacial Research Award (1R01DE025248-01/R 56DE025248-01); NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01CA218148-01); NCI Big Data to Knowledge (BD2K) Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825-01); NCI Cancer Center Support Grant (CCSG) Radiation Oncology and Cancer Imaging Program Pilot Research Program Award (P30CA016672); NCI/National Science Foundation Joint NSF/NIH Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) award (5R01CA225190-0); and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Program (R25EB025787). This research is supported by the Andrew Sabin Family Foundation; Dr. Fuller is a Sabin Family Foundation Fellow. Dr. Fuller has received direct institutional academic/industry grant support from Elekta AB via a Elekta AB/MD Anderson Department of Radiation Oncology Seed Grant. Dr. Fuller has received speaker travel and honoraria from Elekta AB. Dr. Elgohari is funded through joint supervision program by the Egyptian Ministry of Cultural and Higher Education. Dr. Kalpathy-Cramer is supported by the National Cancer Institute (U24 CA180927-03, U01 CA154601-06. Dr. Dagan is/was a speaker for Ion Beam Applications, and receives/received research funding from Elekta. Dr. Kachnic reported that she gave expert testimony for Andrew Skinner Inc; received grants from the National Cancer Institute and lecture fees from the American Society for Radiation Oncology, American Society of Clinical Oncology, and Chartrounds TM; and earned royalties from Up-to-Date. Drs. FitzGerald and Ulin and Ms. Laurie's efforts were supported in part by the NIH/NCI grants to the Imaging and Radiation Oncology Core-IROC and The Quality Assurance Review Center-QARC: U24CA180803 (IROC) and U10CA29511 (QARC). Dr. Gaspar has a consulting agreement with AstraZeneca.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ctro.2019.01.001.

References

[1] Jeanneret-Sozzi W et al. The reasons for discrepancies in target volume delineation: a SASRO study on head-and-neck and prostate cancers. Strahlentherapie und Onkologie:Organ der Deutschen Rontgengesellschaft [et al] 2006;182:450–7. https://doi.org/10.1007/s00066-006-1463-6.
[2] Njeh CF. Tumor delineation: the weakest link in the search for accuracy in radiotherapy. J Med Phys 2008;33:136–40. https://doi.org/10.4103/0971-6203.44472.
[3] Steenbakkers RJ 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. https://doi.org/10.1016/j.ijrobp.2005.06.034.
[4] Rasch C et al. Definition of the prostate in CT and MRI: a multi-observer study. Int J Radiat Oncol Biol Phys 1999;43:57–66.
[5] Schmidt MA, Payne GS. Radiotherapy planning using MRI. Phys Med Biol 2015;60:R323–61. https://doi.org/10.1088/0031-9155/60/22/R323.
[6] Rasch C et al. The potential impact of CT-MRI matching on tumor volume delineation in advanced head and neck cancer. Int J Radiat Oncol Biol Phys 1997;39:841–8.
[7] Fuller CD et al. Prospective randomized double-blind pilot study of site-specific consensus atlas implementation for rectal cancer target volume delineation in the cooperative group setting. Int J Radiat Oncol Biol Phys 2011;79:481–9. https://doi.org/10.1016/j.ijrobp.2009.11.012.
[8] Ohri N et al. Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials. JNCI J Natl Cancer Instit 2013;105:387–93. https://doi.org/10.1093/jnci/djt001.
[9] Holliday E et al. Quantitative assessment of target delineation variability for thymic cancers: agreement evaluation of a prospective segmentation challenge. J Radiat Oncol 2016;5:55–61. https://doi.org/10.1007/s13566-015-0230-7.
[10] Gregoire V et al. Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. Radiother Oncol J Europ Soc Therap Radiol Oncol 2014;110:172–81. https://doi.org/10.1016/j.radonc.2013.10.0.
[11] Lynch R, Pitson G, Ball D, Claude L, Sarrut D. Computed tomographic atlas for the new international lymph node map for lung cancer: a radiation oncologist perspective. Pract Radiat Oncol 2013;3:54–66. https://doi.org/10.1016/j.prro.2012.01.007.
[12] Hartford AC et al. American College of Radiology (ACR) and American Society for Radiation Oncology (ASTRO) Practice Guideline for Intensity-modulated

Radiation Therapy (IMRT). Am J Clin Oncol 2012;35:612–7. https://doi.org/10.1097/COC.0b013e31826e0515.
[13] Muirhead R, McNee SG, Featherstone C, Moore K, Muscat S. Use of Maximum Intensity Projections (MIPs) for target outlining in 4DCT radiotherapy planning. J Thoracic Oncol Off Publicat Int Associat Study Lung Cancer 2008;3:1433–8. https://doi.org/10.1097/JTO.0b013e31818e5db7.
[14] Ruysscher DD et al. European organisation for research and treatment of cancer recommendations for planning and delivery of high-dose, high-precision radiotherapy for lung cancer. J Clin Oncol 2010;28:5301–10. https://doi.org/10.1200/jco.2010.30.3271.
[15] Frood R et al. Effectiveness of respiratory-gated positron emission tomography/computed tomography for radiotherapy planning in patients with lung carcinoma – a systematic review. Clin Oncol (Royal College of Radiologists (Great Britain)) 2018;30:225–32. https://doi.org/10.1016/j.clon.2018.01.005.
[16] Wu AJ et al. Expert consensus contouring guidelines for intensity modulated radiation therapy in esophageal and gastroesophageal junction cancer. Int J Radiat Oncol Biol Phys 2015;92:911–20. https://doi.org/10.1016/j.ijrobp.2015.03.030.
[17] Myerson RJ et al. Elective clinical target volumes for conformal therapy in anorectal cancer: a radiation therapy oncology group consensus panel contouring atlas. Int J Radiat Oncol Biol Phys 2009;74:824–30. https://doi.org/10.1016/j.ijrobp.2008.08.070.
[18] Allozi R et al. Tools for consensus analysis of experts' contours for radiotherapy structure definitions. Radiother Oncol J Eur Soc Therap Radiol Oncol 2010;97:572–8. https://doi.org/10.1016/j.radonc.2010.06.009.
[19] Amini A et al. Dose constraints to prevent radiation-induced brachial plexopathy in patients treated for lung cancer. Int J Radiat Oncol Biol Phys 2012;82:e391–398. https://doi.org/10.1016/j.ijrobp.2011.06.1961.
[20] Kwa SL et al. Radiation pneumonitis as a function of mean lung dose: an analysis of pooled data of 540 patients. Int J Radiat Oncol Biol Phys 1998;42:1–9.
[21] Okunieff P et al. Report from the radiation therapy committee of the southwest oncology group (swog): research objectives workshop 2008. Clin Cancer Res Off J Am Associat Cancer Res 2009;15:5663–70. https://doi.org/10.1158/1078-0432.ccr-09-0357.
[22] Truntzer P et al. Superior sulcus non-small cell lung carcinoma: a comparison of IMRT and 3D-RT dosimetry. Rep Pract Oncol Radiother 2016;21:427–34. https://doi.org/10.1016/j.rpor.2016.03.006.
[23] Roy AEF, Wells P. Volume definition in radiotherapy planning for lung cancer: how the radiologist can help. Cancer Imaging 2006;6:116–23. https://doi.org/10.1102/1470-7330.2006.0019.
[24] From The Field. The politics of the health insurance portability and accountability act. Health Aff 1997;16:146–50. https://doi.org/10.1377/hlthaff.16.3.146.
[25] Okunieff P et al. Report from the SWOG radiation oncology committee: research objectives workshop 2017. Clin Cancer Res 2018. https://doi.org/10.1158/1078-0432.ccr-17-3202.
[26] ICRU. Report 50. Prescribing, recording, and reporting photon beam therapy ICRU 1993. Oxford University Press, Oxford, United Kingdom.
[27] ICRU. Report 62. Prescribing, recording, and reporting photon beam therapy (Supplement to ICRU Report 50)ICRU 1999 (Oxford University Press, Oxford, United Kingdom
[28] Rusch VW et al. Induction chemoradiation and surgical resection for non–small cell lung carcinomas of the superior sulcus: initial results of Southwest Oncology Group Trial 9416 (Intergroup Trial 0160). J Thorac Cardiovasc Surg 2001;121:472–83. https://doi.org/10.1067/mtc.2001.112465.
[29] Kernstine KH et al. Trimodality therapy for superior sulcus non-small cell lung cancer: Southwest Oncology Group-Intergroup Trial S0220. Ann Thorac Surg 2014;98:402–10. https://doi.org/10.1016/j.athoracsur.2014.04.129.
[30] Kappers I et al. Results of combined modality treatment in patients with non-small-cell lung cancer of the superior sulcus and the rationale for surgical resection. Eur J Cardiothorac Surg 2009;36:741–6. https://doi.org/10.1016/j.ejcts.2009.04.069.
[31] Lilenbaum R et al. Phase II trial of combined modality therapy with myeloid growth factor support in patients with locally advanced non-small cell lung cancer. J Thorac Oncol 2010;5:837–40. https://doi.org/10.1097/JTO.0b013e3181d6e141.
[32] 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. https://doi.org/10.1109/tmi.2004.828354.
[33] Lapa C et al. Prognostic value of positron emission tomography-assessed tumor heterogeneity in patients with thyroid cancer undergoing treatment with radiopeptide therapy. Nucl Med Biol 2015;42:349–54. https://doi.org/10.1016/j.nucmedbio.2014.12.006.
[34] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302. https://doi.org/10.2307/1932409.
[35] Rao M et al. Comparison of human and automatic segmentations of kidneys from CT images. Int J Radiat Oncol Biol Phys 2005;61:954–60. https://doi.org/10.1016/j.ijrobp.2004.11.014.
[36] Paul J. The distribution of the Flora in the alpine zone.1. New Phytologist 1912;11:37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.
[37] Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19:3–11. https://doi.org/10.2466/pr0.1966.19.1.3.
[38] Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39:175–91.
[39] Taylor A, Powell MEB. Intensity-modulated radiotherapy—what is it? Cancer Imaging 2004;4:68–73. https://doi.org/10.1102/1470-7330.2004.0003.
[40] Liauw SL, Connell PP, Weichselbaum RR. New paradigms and future challenges in radiation oncology: an update of biological targets and technology. Sci Transl Med 2013;5. https://doi.org/10.1126/scitranslmed.3005148.
[41] Ng M et al. Australasian Gastrointestinal Trials Group (AGITG) contouring atlas and planning guidelines for intensity-modulated radiotherapy in anal cancer. Int J Radiat Oncol Biol Phys 2012;83:1455–62. https://doi.org/10.1016/j.ijrobp.2011.12.058.
[42] Rusch VW et al. The IASLC lung cancer staging project: a proposal for a new international lymph node map in the forthcoming seventh edition of the TNM classification for lung cancer. J Thorac Oncol Off Publicat Int Assoc Study of Lung Cancer 2009;4:568–77. https://doi.org/10.1097/JTO.0b013e3181a0d82e.
[43] 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 2018;1–9. https://doi.org/10.1080/0284186X.2018.1529422.
[44] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29. https://doi.org/10.1186/s12880-015-0068-x.
[45] Awan M et al. Prospective assessment of an atlas-based intervention combined with real-time software feedback in contouring lymph node levels and organs-at-risk in the head and neck: quantitative assessment of conformance to expert delineation. Pract Radiat Oncol 2013;3:186–93. https://doi.org/10.1016/j.prro.2012.11.002.
[46] Mavroidis P et al. Consequences of anorectal cancer atlas implementation in the cooperative group setting: radiobiologic analysis of a prospective randomized in silico target delineation study. Radiotherap Oncol J Eur Soc Therap Radiol Oncol 2014;112:418–24. https://doi.org/10.1016/j.radonc.2014.05.011.
[47] Szumacher E et al. Effectiveness of educational intervention on the congruence of prostate and rectal contouring as compared with a gold standard in three-dimensional radiotherapy for prostate. Int J Radiat Oncol Biol Phys 2010;76:379–85. https://doi.org/10.1016/j.ijrobp.2009.02.008.