Published in final edited form as:

J Radiat Oncol. 2016 March ; 5(1): 55–61. doi:10.1007/s13566-015-0230-7.

Quantitative assessment of target delineation variability for thymic cancers: Agreement evaluation of a prospective segmentation challenge

Emma Holliday, MD[1], **Clifton D. Fuller, MD, PhD**[1,*], **Jayashree Kalpathy-Cramer, PhD**[2], **Daniel Gomez, MD**[1], **Andreas Rimner, MD, PhD**[3], **Ying Li, MD, PhD**[4], **Suresh Senan, MD, PhD**[5], **Lynn D. Wilson, MD, MPH**[6], **Jehee Choi, MD**[7], **Ritsuko Komaki, MD**[1], and **Charles R. Thomas Jr., MD**[8]

[1]Division of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas [2]Department of Radiology, Massachusetts General Hospital, Boston, MA [3]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY [4]Department of Radiation Oncology, The University of Texas Health Science Center San Antonio, San Antonio, TX [5]Department of Radiation Oncology, University Medical Center, Amsterdam, The Netherlands [6]Department of Therapeutic Radiology, Yale Cancer Center, New Haven, CT [7]Department of Radiation Oncology, Kaiser Permanente Southern California, Los Angeles, CA [8]Department of Radiation Medicine, Oregon Health and Science University Knight Cancer Center, Portland, OR

Abstract

Objectives—We sought to quantitatively determine the inter-observer variability of expert radiotherapy target-volume delineation for thymic cancers, as part of a larger effort to develop an expert-consensus contouring atlas.

Methods—A pilot dataset was created consisting of a standardized case presentation with pre- and post-operative DICOM CT image sets from a single patient with Masaoka-Koga Stage III thymoma. Expert thoracic radiation oncologists delineated tumor targets on the pre- and post-operative scans as they would for a definitive and adjuvant case, respectively. Respondents completed a survey including recommended dose prescription and target volume margins for definitive and post-operative scenarios. Inter-observer variability was analyzed quantitatively with Warfield's simultaneous truth, performance level estimation (STAPLE) algorithm and Dice similarity coefficient (DSC).

Results—Seven users completed contouring for definitive and adjuvant cases; of these, 5 completed online surveys. Segmentation performance was assessed, with high mean±SD STAPLE-estimated segmentation sensitivity for definitive case GTV and CTV at 0.77 and 0.80, respectively, and post-operative CTV sensitivity of 0.55; all volumes had specificity of ≥0.99. Inter-observer agreement was markedly higher for the definitive target volumes, with mean±SD

[*]For correspondence and reprint requests contact: Clifton David Fuller, MD, PhD, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Unit 97, Houston, TX, 77030, Phone: 713-563-2334 Fax: 713-563-2366 cdfuller@mdanderson.org.

Presented, in part, at the 14th World Conference on Lung Cancer, July 3–7, 2011, Amsterdam, The Netherlands.

DSC of 0.88±0.03 and 0.89±0.04 for GTV and CTV respectively, compared to post-op CTV DSC of 0.69±0.06 (Kruskal-Wallis $p<0.01$.

Conclusion—Expert agreement for definitive case volumes was exceptionally high, though significantly lower agreement was noted post-operatively. Technique and dose prescription between experts was substantively consistent, and these preliminary results will be utilized to create an expert-consensus contouring atlas to aid the non-expert radiation oncologist in the planning of these challenging, rare tumors.

Keywords

Radiotherapy; thymoma; thymic carcinoma; treatment planning; target delineation; consensus guidelines

Introduction

Radiation therapy (RT) is recommended in the post-operative treatment of thymoma and thymic carcinoma in the cases of microscopic or macroscopic positive margins and can be considered for patients with more advanced disease who have had a complete resection (1). RT is also given for patients who are inoperable either for tumor-related or medical comorbidity reasons (2–3). The studies initially describing RT for thymic malignancies were performed in the 1980s and 1990s and utilized cobalt-60 units to deliver radiation to the entire mediastinum and the lung and pleura adjacent to the resected tumor via an anterior and posterior field arrangement until cord tolerance was reached, then angled off-cord oblique fields were commonly used (4). Reported grade 3 to 5 toxicities from therapy in this era ranged from 10–15% (4) and included tracheo-arterial fistula (5), pericarditis (6), pneumonitis (7), cardiac conduction abnormalities (8) and esophageal stricture (9).

The advent of conformal, image-guided RT allows for the delivery of tumoricidal radiation doses to physician-delineated target volumes while minimizing dose to nearby organs-at-risk and potentially reduces significant pulmonary toxicity rates down to 5–10% (10). Intensity modulated RT (IMRT) further allows for an improved therapeutic ratio through the use of “inverse planning” to optimize dose to the tumor while preferentially sparing organs at risk (11). However, target volumes and organs-at-risk for treatment planning are necessarily defined by human users, introducing possible geometric variability due to variation in target delineation (12). *Under*-contouring can lead to local relapse, and *over*-contouring can lead to unnecessary normal tissue toxicity. Accurate target delineation by the treating physician is not only the most critical step of the process, but also the most susceptible to error (13). Several groups have sought to account for systematic variability introduced in target delineation in order to optimize volume definition and treatment planning margins (14–16). The location of the target volume appears to influence the frequency and magnitude of contouring error. Fritton et al reported 8mm interobserver variability with CT-based planning for mediastinal non-small-cell lung cancer (NSCLC) (17).

Thymic neoplasms present a unique set of challenges in terms of target delineation for conformal radiotherapy. The rarity of thymic tumors means that heterogenous radiotherapy approaches are often reported in retrospective series. An expert-consensus atlas would not only be invaluable as a resource for physicians who infrequently treat thymic tumors but could also serve a validated benchmark for future clinical trials using conformal radiotherapy. Through this project, we sought to quantitatively determine the relative inter-observer variability of expert target volume delineation, which we hypothesized would be similar to that noted in NSCLC contouring atlases: mean Dice similarity coefficient (DSC) between 0.6 and 0.8 (17).

Methods

Case selection

IRB approval for anonymized data utilization for this study has been approved as “exempt” under University of Texas Health Science Center IRB # HSC200B0166E. A pilot dataset was created consisting of a standardized case presentation with anonymized pre- and post-operative DICOM CT image sets from a single patient with Masaoka-Koga Stage III thymoma. Datasets consisted of a treatment planning CT, and fused MRI and FDG-PET. Each case session consisted of specific axial slices through derived from the DICOM file of an actual patient without any HiPPA-defined patient-specific information. This was performed by anonymizing the patient DICOM data using DCMAnonymize DICOM Validation Toolkit (DVTk.org). Associated diagnostic imaging were also available for reference.

Data collection

The anonymized DICOM file was made available along with a PDF containing an anonymized clinical vignette. Participating thoracic radiation oncologists contoured, using a treatment planning system of their choice, the gross tumor volume (GTV), and clinical target volumes (CTV) as they would in their standard practice, in both a definitive (pre-op scan) and adjuvant (post-op) setting. After submission of contours as DICOM RTSTRUCT files for central analysis users completed a survey detailing the dose prescription and PTV margins they would recommend in definitive and post-operative (i.e. R1 vs R2) scenarios. Data submission and analysis were semi-automated using TaCTICS (“Target Contour Testing/Instructional Computer Software” http://skynet.ohsu.edu/tactics/), a target delineation statistical software with a graphical user interface (GUI), developed at Oregon Health & Science University (18–19). TaCTICS allows for near real-time data analysis and reporting of quantitative scoring metrics that compare user-derived structures with reference sets derived from other expert users. Additionally, TaCTICS serves as a cost-effective central repository of contouring data for planned central analysis.

Data analysis

Inter-observer variability was analyzed quantitatively with Warfield's simultaneous truth and performance level estimation (STAPLE) algorithm to generate a composite segmentation estimate of a “ground truth” target volume, which was then compared to each individual experts contours as a Dice similarity coefficient (DSC) [where DSC=1 indicates total agreement, while DSC=0 indicates no overlapping TV voxels] using nonparametric analysis (20–22).

Results

A total of 7 expert thoracic radiation oncologists who are members of the International Thymic Malignancy Interest Group (ITMIG) completed the contouring tasks for definitive and post-operative cases; of these 5 completed online surveys. Details are listed in Table 1.

Hausdorff distances were calculated for the GTV, pre-op CTV and post-op CTV and are shown in Figure 1a. The Hausdorff distance measures how far two subsets are from each other and is defined as the greatest of all the distances from a point in one set to the closest point in the other set. Segmentation performance was assessed, with high mean±SD STAPLE-estimated segmentation sensitivity for definitive case GTV and CTV at 0.77 and 0.80, respectively, and post-operative CTV sensitivity of 0.55; all volumes had specificity of ≥0.99 (Figure 1b).

Inter-observer agreement was markedly higher for the definitive case target volumes, with mean±SD DSC of 0.88±0.03 and 0.89±0.04 for GTV and CTV respectively, compared to post-op CTV DSC of 0.69±0.06 (Kruskal-Wallis $p<0.01$; Figure 1c) for all experts vs. the STAPLE composite(s). Figure 2 shows a representative screenshot of two expert-segmented target volumes simultaneously displayed on a single axial CT slice.

Survey results indicated 60% of experts would routinely add margins of 0.5 cm to CTV volumes for PTV generation, with 40% advocating 1 cm expansions. All respondents suggested IMRT should be used, 80% suggested 4DCT should be utilized for simulation, and 40% suggested PETCT fusion should be implemented, if available, for contour delineation. For definitive cases, a PTV prescription of 50–54 Gy was suggested. Post-operatively, for R1 resections, a PTV prescription of 60–66 Gy was suggested, with 60–70 Gy for an R2 resection.

Discussion

This prospective, pilot study sought to quantify inter-observer variability between expert-delineated GTV and CTV volumes for a standardized case of stage III thymoma in both the pre-operative and post-operative setting. There was excellent agreement for pre-operative GTV (DSC of 0.88±0.03) and CTV (DSC of 0.89±0.04), but the agreement for post-operative CTV was significantly lower (DSC of 0.69±0.06). The majority of experts surveyed stated they would use 0.5 cm PTV margins, and all stated they would use IMRT. Recommended doses mirrored those published in international guidelines.

Currently, radiation oncologists gain skill and expertise in target and organ-at-risk delineation, also known as contouring, through clinical practice during their residency training. Differences in case volume and diversity as well as clinical teaching and supervision lead to an unavoidably heterogeneous skill-set in residency graduates. More recently, the development of educational tools and activities to assist clinicians in developing standardized target delineation skill-sets has been an active area of research.

Contouring atlases are an efficient way for experts in the field to communicate guidelines for delineating both target volumes and organs at risk to large numbers of trainees and practicing clinicians. These have been developed and published for several common cancers such as lung (23–24), prostate (25–26), breast (27), rectum (28), pancreas (29) anus (30), cervical and endometrial cancer (31–32) and head and neck cancer (33). These atlases are typically assembled via consensus after target volumes and organs at risk are delineated by several established specialists for standardized test cases. Although some atlases were developed through discussion and verbal consensus among experts, DSC is often used to quantify level of agreement, as was performed in our study. The DSCs in our study ranged from 0.69±0.06 for the post-operative CTV volume to 0.89±0.04 for the pre-operative CTV volume and 0.88±0.03 for the pre-operative GTV volume. This compares favorably to other published target volume delineation projects. For groups publishing consensus prostate cancer target volumes, post-operative CTV volumes also had lower agreement between expert contours with a mean specificity of 0.5± 0.25 to 0.62± 0.22 (26). For groups publishing breast cancer target volumes, mean DSCs ranged from 0.56–0.95 in one study (27) with the most variability in the nodal target volumes and the least variability in the tumor bed (34).

There has been recent interest in quantifying the effects of contouring atlases on target volumes delineated in clinical practice. Studies have been recently published that demonstrate substantial improvement in target volume delineation through the use of such consensus atlases in gastrointestinal, breast and head and neck tumor sites (35–37); however, limited improvement was reported after a similar education intervention in lung tumor volume delineation (38). It can be argued that expert-consensus atlases may be even more valuable for rare tumors with which many radiation oncologists are less familiar. There is no current atlas for thymic malignancies, although recent guidelines for simulation, target volume definition and dose-volume constraints have recently been published (39).

Further, consistent target volumes have yet to be established across series; for instance, some series report routine supraclavicular nodal coverage (40), while others do not (41). Thoracic anatomy is complex, and pre-operative evaluation of primary tumor extension is often more subtle than in other anatomic regions. The high inter-observer agreement seen when delineating pre-operative GTV in this study suggests, at least in the hands of experts in the fields, the intact primary tumor volume can be reliably and reproducibly delineated. In practice, surgical clips are often used to help guide CTV target delineation in the post-operative setting, but number and location of clip placement often vary, and discussion of post-operative imaging with the surgeon can also help to define the area at risk.

Diagnostic radiologists often use multimodality imaging for enhanced clinical evaluation. In the setting of NSCLC, it has been suggested that reviewing PET imaging with a radiologist led to significant changes to GTV volumes for stage IA-IIIB patients (42). However, optimal multimodality techniques for tumor definition such as 18FDG-PET (43–44), MRI (45–46) have yet to be determined for thymic malignancies. Additionally, physiologic motion of the thorax provides an impediment to accurate imaging due to motion artifact on PET, CT and MRI imaging as well as introduces uncertainty during therapy due to intra- and inter-fraction motion of the tumor and surrounding normal tissues. Additionally, while 4DCT for motion correction is recognized as useful for thoracic tumors generally, implementation remains unstandardized with regard to how to optimally implement motion-correction for target volume delineation/treatment of mediastinal tumors. In our survey, 80% of expert respondents suggested implementation of 4DCT for simulation, and 40% suggested implementation of PETCT. These suggestions, as well as the ranges of doses and PTV margin, are in line with a recently published consensus statement (39).

The initial results of this pilot study are encouraging, and the applications of these data are many. First, further case studies will be released for expert contouring in order to create an online atlas. In addition to use in clinical practice, our hope is that this atlas can also serve as a tool to help standardize radiation therapy planning for future multi-institutional clinical trials. Many ongoing and recently closed trials have provided participating institutions and physicians with contouring atlases for assistance with consistent and accurate target delineation. Furthermore, prior pilot studies in rectal cancer have demonstrated that consensus guideline atlas implementation improves inter-observer agreement and a greater approximation of expert target volumes in the cooperative group setting (36, 47).

Obviously, our use of a single case for this analysis presents a significant limitation on the generalizability of this dataset. Nonetheless, this effort represents, to our knowledge, the first prospectively collected quantitative assessment of target volume delineation for thymic malignancies, and is a landmark effort for ITMIG. Our data suggest that experts can, with little oversight, achieve statistically comparable target volumes for gross tumor and pre-therapy clinical target volumes, but that agreement post-operatively is much lower. This suggests that, either a consensus atlas intervention as demonstrated in rectal (36) or head and neck (37) could be used to standardize volumes for prospective radiotherapy trials. Alternatively, contouring of pre-surgical volumes (which exhibit high agreement) with subsequent registration to post-therapy scans could be utilized, provided sufficient quality assurance of the registration method is performed (48–49).

Conclusions

Expert agreement for definitive case GTV and CTV volumes was exceptionally high, though significantly lower agreement was noted for post-operative CTV. Technique and dose prescription between experts was substantively consistent. Data from this preliminary analysis will be used to generate educational materials to assist non-expert target delineation through guideline and atlas implementation. Future efforts will seek to assess the impact of 4DCT and PET imaging on target delineation variability.

Acknowledgements

Dr. Fuller received/receives grant support from: the SWOG Hope Foundation Dr. Charles A. Coltman, Jr. Fellowship in Clinical Trials; the National Institutes of Health Paul Calabresi Clinical Oncology Award Program (K12 CA088084) and Clinician Scientist Loan Repayment Program (L30 CA136381-02); Elekta AB/MD Anderson Consortium; GE Medical Systems/MD Anderson Center for Advanced Biomedical Imaging In-Kind Award; the MD Anderson Center for Radiation Oncology Research, and an MD Anderson Institutional Research Grant Program Award. These listed funders/supporters played no role in the study design, collection, analysis, interpretation of data, manuscript writing, or decision to submit the report for publication. The rest of the authors have nothing to disclose.

References

1. Gomez, DR.; Fuller, CD.; Chennupati, S., et al. Mediastinal and Tracheal Cancer. In: Halperin, EC.; Wazer, DE.; Perez, CA.; Brady, LW., editors. Perez and Brady's Principles and Practice of Radiation Oncology. 6th edition. Williams & Wilkins; Philadelphia, PA: © 2013 by Lippincott
2. National Comprehensive Cancer Network. [Accessed 09/29/2014] Thymomas and Thymic Carcinomas. Version 12014. http://www.nccn.org/professionals/physician_gls/pdf/thymic.pdf
3. Falkson CB, Bezjak A, Darling G, Gregg R, Malthaner R, Maziak DE, et al. Lung cancer disease site group of cancer care Ontario's Program in evidence-based care. The managmenet of thymoma: a systematic review and practice guideline. J Thorac Onc. 2009; 4:911–919.
4. Hsu HC, Huang EY, Wang CJ, et al. Postoperative radiotherapy in thymic carcinoma: treatment results and prognostic factors. Int J Radiat Oncol Biol Phys. 2002; 52:801–805. [PubMed: 11849804]
5. Curran WJ Jr, Kornstein MJ, Brooks JJ, Turisi A 3rd. Invasive thymoma: the role of mediastinal irradiation following complete or incomplete surgical resection. J Clin Oncol. 1988; 6:1722–1727. [PubMed: 3183702]
6. Ogawa K, Uno T, Toita T, Onishi H, Yoshida H, Kakinohana Y, et al. Postoperative radiotherapy for patients with completely resected thymoma: a multi-institutional retrospective review of 103 patients. Cancer. 2002; 94:1405–1413. [PubMed: 11920495]
7. Mornex F, Resbeut M, Richaud P, et al. Radiotherapy and chemotherapy for invasive thymomas: a multicentric retrospective review of 90 cases. The FNCLCC trialists. Federation Nationale des Centres de Lutte Contre le Cancer. Int J Radiat Oncol Biol Phys. 1995; 32:651–659. [PubMed: 7790251]
8. Nako T, Kanaya H, Namura M, et al. Complete atrioventricular block following radiation therapy for malignant thymoma. Jpn J Med. 1990; 29:104–110. [PubMed: 2214340]
9. Hirota S, Tsujino K, Hishikawa Y, Watanabe H, Kono K, Soejima T, et al. Endoscopic findings of radiation esophagitis in concurrent chemoradiotherapy for intrathoracic malignancies. Radiother Oncol. 2001; 58:273–278. [PubMed: 11230888]
10. Lucchi M, Mussi A, Basolo F, et al. The multimodality treatment of thymic carcinoma. Eur J Cardiothorac Surg. 2001; 19:566–569. [PubMed: 11343932]
11. Gomez DR, Komaki R. Technical advances of radiation therapy for thymic malignancies. J Thorac Oncol. 2010; 5:S336–343. [PubMed: 20859129]
12. Jeanneret-Sozzi W, Moeckli R, Valley JF, et al. The reasons for discrepancies in target volume delineation: a SASRO study on head-and-neck and prostate cancers. Strahlenther Onkol. 2006; 182:450–457. [PubMed: 16896591]
13. Njeh CF. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. J Med Phys. 2008; 33:136–140. [PubMed: 19893706]
14. Leunens G, Menten J, Weltens C, et al. Quality assessment of medical decision making in radiation oncology: variability in target volume delineation for brain tumours. Radiother Oncol. 1993; 29:169–175. [PubMed: 8310142]
15. Muijs CT, Schreurs LM, Busz DM, et al. Consequences of additional use of PET information for target volume delineation and radiotherapy dose distribution for esophageal cancer. Radiother Oncol. 2009; 93:447–453. [PubMed: 19765847]
16. Rasch C, Barillot I, Remeijer P, et al. Definition of the prostate in CT and MRI: a multi-observer study. Int J Radiat Oncol Biol Phys. 1999; 43:57–66. [PubMed: 9989514]
17. Fitton I, Steenbakkers RJ, Gihuijs K, et al. Impact of anatomic location on value CT-PET co-registration for delineation of lung tumors. Int J Radiat Oncol Biol Phys. 2008; 70:1403–1407. [PubMed: 17980511]
18. Kalpathy-Cramer J, Fuller CD. Target contour testing/instructional computer software (TaCTICS): A novel training and evaluation platform for radiotherapy target delineation. AMIA Annu Symp Proc. Nov 13.2010 2010:361–365. [PubMed: 21347001]
19. Kalpathy-Cramer J, Bedrick SD, Boccia K, et al. A pilot prospective feasibility study of organ-at-risk definition using Target Contour Testing/Instrutional Computuer Software (TaCTICS), a

training and evaluation platform for radiotherapy target delineation. AMIA Annu Symp Proc. 2011; 2011:654–663. [PubMed: 22195121]

20. Koch GG, Landis JR, Freeman JL, et al. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics. 1977; 33:133–158. [PubMed: 843570]
21. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977; 33:363–374. [PubMed: 884196]
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]
23. Kong FM, Ritter T, Quint DJ, et al. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs and brachial plexus. Int J Radiat Oncol Biol Phys. 2011; 81:1442–1457. [PubMed: 20934273]
24. Chapet O, Kong FM, Quint LE, et al. CT-based definition of thoraric lymph node stations: an atlas from the University of Michigan. Int J Radiat Oncol Biol Phys. 2005; 63:170–178. [PubMed: 16111586]
25. Taylor A, Rockall AG, Powell ME. An atlas of the pelvic lymph node regions to aid radiotherapy target volume definition. Clin Oncol (R Coll Radiol). 2007; 19:542–550. [PubMed: 17624745]
26. Michalski M, Lawton C, El Naqa I, et al. Development of RTOG consensus guidelines for the definition of clinical target volume for postoperative conformal radiation therapy for prostate cancer. Int J Radiat Oncol Biol Phys. 2010; 76:361–368. [PubMed: 19394158]
27. Nielsen MH, Berg M, Pedersen AN, et al. Delineation of target volumes and organs at risk in adjuvant radiotherapy of early breast cancer: national guidelines and contouring atlas by the Danish Breast Cancer Cooperative Group. Acta Oncol. 2013; 52:703–710. [PubMed: 23421926]
28. Myerson RJ, Garofalo MC, El Naqa I, 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. 2008; 74:824–830. [PubMed: 19117696]
29. Goodman KA, Regine WF, Dawson LA, et al. Radiation Therapy Oncology Group consensus panel guidelines for the delineation of the clinical target volume in the postoperative treatment of pancreatic head cancer. Int J Radiat Oncol Biol Phys. 2012; 83:901–908. [PubMed: 22483737]
30. Ng M, Leong T, Chander S, 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; 83:1455–1462. [PubMed: 22401917]
31. Small W Jr, Mell LK, Anderson P, et al. Consensus guidelines for delineation of clinical target volume for intensity modulated pelvic radiotherapy in postoperative treatment of endometrial and cervical cancer. Int J Radiat Oncol Biol Phys. 2008; 71:428–434. [PubMed: 18037584]
32. Toita T, Ohno T, Kaneyasu Y, et al. A consensus-based guideline defining the clinical target volume for pelvic lymph nodes in external beam radiotherapy for uterine cervical cancer. Jpn J Clin Oncol. 2010; 40:456–463. [PubMed: 20133334]
33. Gregoire V, Ang K, Budach W, 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 guideline. Radiother Oncol. 2014; 110:172–81. [PubMed: 24183870]
34. Li XA, Tai A, Arthur DW, et al. Variability of target and normal structure delineation for breast cancer radiotherapy: an RTOG multi-institutional and multiobserver study. Int J Radiat Oncol Biol Phys. 2009; 73:944–951. [PubMed: 19215827]
35. Berkelman S, Wolden S, Lee N. Head-and-neck target delineation among radiation oncology residents after a teaching intervention: a prospective, blinded pilot study. Int J Radiat Oncol Biol Phys. 2009; 73:416–423. [PubMed: 18538494]
36. Fuller CD, Nijkamp J, Duppen J, 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. 2010; 72:481–489. [PubMed: 20400244]
37. Awan M, Kalpathy-Cramer J, Gunn GB, 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; 3:186–193. [PubMed: 24674363]

38. Vorwerk H, Beckmann G, Bremer M, et al. The delineation of target volumes for radiotherapy of lung cancer patients. Radiother Oncol. 2009; 91:455–460. [PubMed: 19339069]

39. Gomez D, Komaki R, Yu J, et al. Radiation therapy definitions and reporting guidelines for thymic malignancies. Zhongguo Fei Ai Za Zhi. 2014; 17:110–115. [PubMed: 24581161]

40. Kim ES, Putnam JB, Komaki R, et al. Phase II study of a multidisciplinary approach with induction chemotherapy, followed by surgical resection, radiation therapy, and consolidation chemotherapy for unresectable malignant thymomas: final report. Lung Cancer. 2004; 44:369–379. [PubMed: 15140551]

41. Cardillo G, Carleo F, Giunti R, et al. Predictors of survival in patients with locally advanced thymoma and thymic carcinoma (Masaoka stages III and IVa). Eur J Cardiothorac Surg. 37:819–823. [PubMed: 19948412]

42. Hanna GG, Carson KJ, Lynch T, et al. 18F-fluorodeoxyglucose positron emission tomography/computed tomography-based radiotherapy target volume definition in non-small-cell lung cancer: delineation by radiation oncologists vs. joint outlining with a PET radiologist? Int J Radiat Oncol Biol Phys. 2010; 78:1040–1051. [PubMed: 20350798]

43. Endo M, Nakagawa K, Ohde Y, et al. Utility of 18FDG-PET for differentiating the grade of malignancy in thymic epithelial tumors. Lung Cancer. 2008; 61:350–355. [PubMed: 18304691]

44. Kumar A, Regmi SK, Dutta R, et al. Characterization of thymic masses using (18)F-FDG PET-CT. Ann Nucl Med. 2009; 23:569–577. [PubMed: 19585212]

45. Inoue A, Tomiyama N, Fujimoto K, et al. MR imaging of thymic epithelial tumors: correlation with World Health Organization classification. Radiat Med. 2006; 24:171–181. [PubMed: 16875304]

46. Tomiyama N, Honda O, Tsubamoto M, et al. Anterior mediastinal tumors: diagnostic accuracy on CT and MRI. Eur J Radiol. 2009; 69:280–288. [PubMed: 18023547]

47. Fuller CD, Duppen J, Rasch CR, et al. A Prospective Randomized Pilot Study of Site-specific Atlas Incorporation into Target Volume Delineation Instructions in the Cooperative Group Setting: Preliminary Results from a Southwest Oncology Group Pilot using Big Brother. Int J Radiat Oncol Biol Phys. 2009; 75:S136–S137.

48. Castillo R, Castillo E, Guerra R, et al. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Phys Med Biol. 2009; 54:1849–1870. [PubMed: 19265208]

49. Castillo R, Castillo E, Fuentes D, et al. A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive. Phys Med Biol. 2013; 58:2861–2877. [PubMed: 23571679]

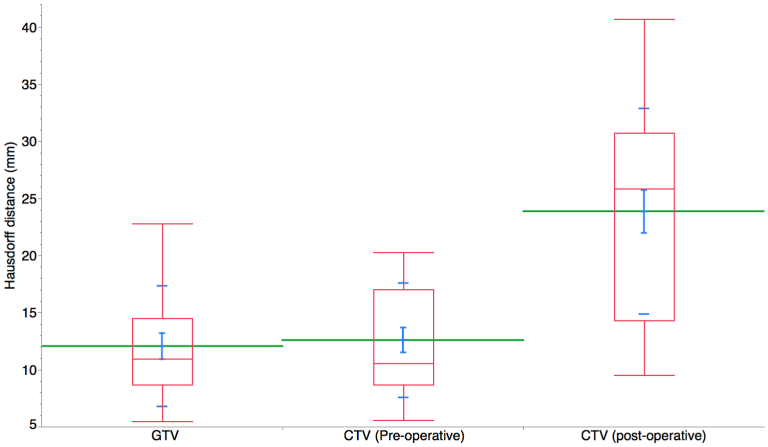

Figure 1A.

Hausdorff distances were calculated for the preGTV, preCTV and postCTV. The Hausdorff distance measures how far two subsets are from each other and is defined as the greatest of all the distances from a point in one set to the closest point in the other set. (postCTV=post-operative CTV; preGTV=definitive case GTV; preCTV=definitive case CTV).

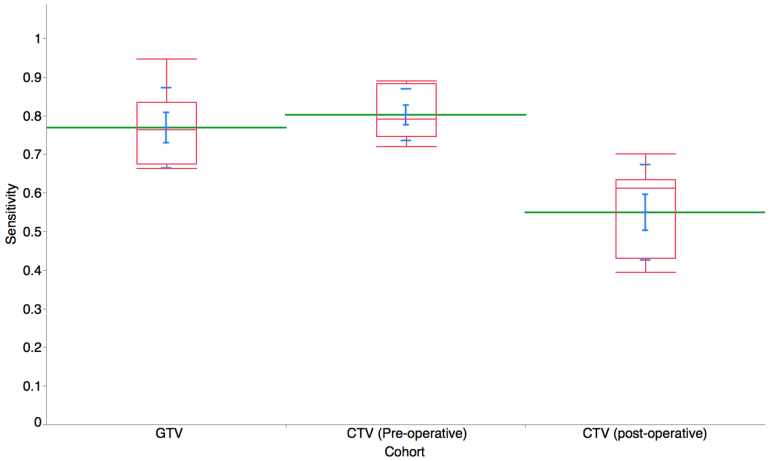

Figure 1B.

Segmentation performance was assessed, with high mean±SD STAPLE-estimated segmentation sensitivity for definitive case preGTV and preCTV, respectively, and lower sensitivity for postCTV sensitivity. (postCTV=post-operative CTV; preGTV=definitive case GTV; preCTV=definitive case CTV),

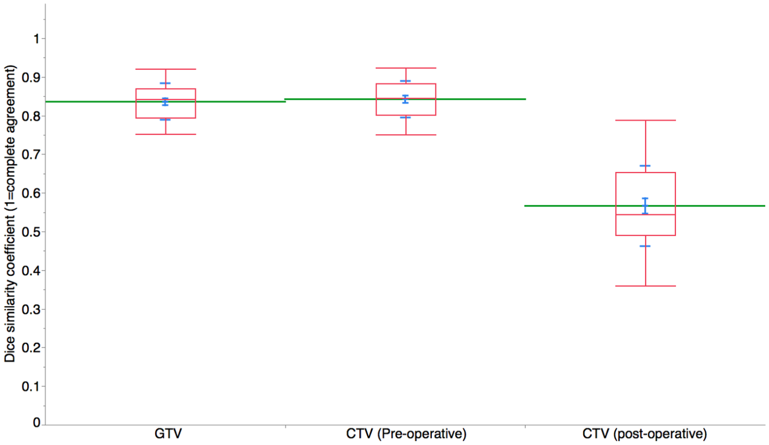

Figure 1C.

Dice similarity coefficients distribution of all expert participants, as compared to STAPLE-estimated ground truth (postCTV=post-operative CTV; preGTV=definitive case GTV; preCTV=definitive case CTV), where DSC=1 indicates total volumetric agreement.

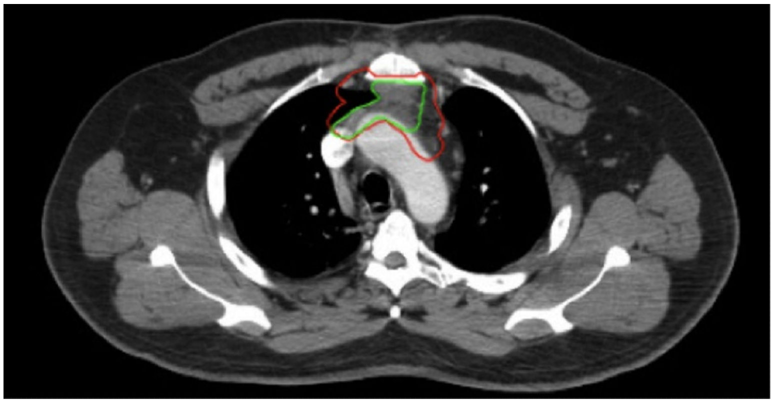

Figure 2.

Screenshot of TaCTICS (“Target Contour Testing/Instructional Computer Software”) software, simultaneously displaying two expert-segmented target volumes (green and red outlines).

Table 1

Each expert's contoured GTV, pre-operative CTV and post-operative CTV sensitivity and specificity compared to the “ground truth” target volume generated by the STAPLE algorithm

	Case 1- Stage III Thymoma preop/definitive				Case 1-Stage III Thymoma postop/adjuvant	
Expert	**GTV-sensitivity**	**GTV-specificity**	**CTV-sensitivity**	**CTV-specificity**	**CTV-sensitivity**	**CTV-specificity**
#1	0.84	>0.99	0.75	>0.99	0.70	>0.99
#2	0.95	>0.99	0.89	>0.99	0.40	>0.99
#3	0.81	>0.99	0.72	>0.99	0.43	>0.99
#4	0.68	>0.99	0.76	>0.99	0.61	>0.99
#5	0.68	>0.99	0.79	>0.99	0.44	>0.99
#6	0.76	>0.99	0.88	>0.99	0.63	>0.99
#7	0.66	>0.99	0.82	>0.99	0.63	>0.99
Mean±SD	0.77±0.10	>0.99	0.80±0.07	>0.99	0.55±0.12	>0.99

GTV = gross tumor volume, CTV = clinical target volume. Staging as per the Masaoka-Koga staging system. Sensitivity and specificity values for each expert contour Warfield's simultaneous truth and performance level estimation (STAPLE) algorithm to generate a composite segmentation estimate of a “ground truth” target volume, which was then compared to each individual experts contours.