Published in final edited form as:
Radiother Oncol. 2010 December ; 97(3): 572–578. doi:10.1016/j.radonc.2010.06.009.

Tools for Consensus Analysis of Experts' Contours for Radiotherapy Structure Definitions

Rawan Allozi, B.A.[1], **X. Allen Li, Ph.D.**[2], **Julia White, M.D.**[2], **Aditya Apte, M.S.**[1], **An Tai, Ph.D.**[2], **Jeff M. Michalski, M.D.**[1], **Walter R. Bosch, Ph.D.**[1], and **Issam El Naqa, Ph.D.**[1]

[1] Washington University, Saint Louis, MO, USA

[2] Medical College Wisconsin, Milwaukee, WI, USA

Abstract

Background and Purpose—To demonstrate and examine the ability of a newly developed software tool to estimate and analyze consensus contours from manually created contours by expert radiation oncologists.

Material and Methods—Several statistical methods and a graphical user interface were developed. For evaluation purposes, we used three breast cancer CT scans from the RTOG Breast Cancer Atlas Project. Specific structures were contoured before and after the experts' consensus panel meeting. Differences in the contours were evaluated qualitatively and quantitatively by the consensus software tool. Estimates of consensus contours were analyzed for the different structures and Dice similarity and Dice-Jaccard coefficients were used for comparative evaluation.

Results—Based on kappa statistics, highest levels of agreement were seen in the left breast, lumpectomy, and heart. Significant improvements between pre- and post-consensus contours were seen in delineation of the chestwall and breasts while significant variations were noticed in the supraclavicular and internal mammary nodes. Dice calculations for all pre-consensus STAPLE estimations and final consensus panel structures reached 0.80 or greater for the heart, left/right breast, case-A lumpectomy, and chestwall.

Conclusions—Using the consensus software tool incorporating STAPLE estimates provided the ability to create contours similar to the ones generated by expert physicians.

Keywords

Structure definition; radiotherapy treatment planning; experts' consensus; statistical modeling; software tools

Corresponding Author: Issam El Naqa, Ph.D. Department of Radiation Oncology Washington University School of Medicine, Campus-Box 8224 St. Louis, MO 63110 elnaqa@wustl.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest Statement
None.

INTRODUCTION

The ultimate goal of radiotherapy treatment is to achieve high rates of local control through tumoricidal doses, which cover all the gross and sub-clinical disease, while limiting toxicity to surrounding normal tissues. However, uncertainties associated with target volumes and organs-at-risk (OAR) definitions, organ motion, and patient setup errors stand as obstacles to achieving better outcomes in radiotherapy planning and treatment delivery [1-4]. In the era of 3D conformal radiotherapy (3D-CRT) and Intensity Modulated Radiation Therapy (IMRT), accurate delineation of tumor volumes and surrounding normal structures becomes the physicians' bottleneck task for subsequent treatment planning optimization and delivery. When designing a 3D-CRT treatment plan, the treating physician draws contours of the gross target volume (GTV) and its surrounding OARs following ICRU guidelines [5]. These guidelines state that knowledge of uncertainties such as patient positioning and organ motion are required to accurately define the PTV [5]. However, this would presume that the previous steps in treatment planning, namely, the delineation of the GTV and the clinical target volume (CTV), are accurate. For some tumor locations, inconsistencies in GTV and CTV definitions maybe the most important error in radiation therapy planning and delivery [6]. Several studies have noted significant uncertainties that arise during the delineation process of the GTV, CTV, and OARs in different cancer sites [2-4,7]. For instance, Mourik *et al.* found that inter-observer variation in breast tumor target volume delineation was larger than setup accuracies [3] while Vorwerk *et al.* found a much larger inter-observer variation in the delineation of the GTV than in the delineation of the PTV in lung cancer [4].

Variability in contouring structures may lead to under-dosage, causing a decrease in tumor control probability (TCP), or over-dosage, resulting in an increase in normal tissue complication probability (NTCP). Thus, uncertainty analysis of GTV/CTV and OARs is critical for adequate coverage of the tumor and sparing of surrounding normal organs [6]. Towards this goal, we have developed a new tool for uncertainty analysis of contoured structures to guide dosimetrists and oncologists in defining different structures for treatment planning purposes. The tool utilizes statistical methods and graphical means to quantify experts' levels of agreement of contours at selected confidence levels, and estimate a consensus structure contour derived from multiple experts' contours. Specifically, three statistical methods are used within the software tool and discussed below in details: apparent agreement, kappa-corrected algorithm [8], and expectation-maximization (EM) algorithm for simultaneous truth and performance level estimation (STAPLE) [9]. Figure 1 illustrates the general workflow of the tool. For convenience, the tool is integrated with a publicly available in-house research treatment planning system called CERR [10].

Recently, the European Society of Therapeutic Radiology and Oncology (ESTRO) and the Radiation Therapy Oncology Group (RTOG) initiated several independent studies for creating consensus guidelines for delineation of tumor volumes and surrounding organs-at-risk. Initial versions of the consensus tool have already been effectively utilized by multiple RTOG consensus panels in performing statistical analyses and generating atlases for different cancer sites [11-13]. In Michalski *et al.*, the consensus tool was utilized to measure the level of agreement for the CTV of postoperative prostate cancer between participating physicians. Generated STAPLE contours were used as the starting point for the final CTV atlas [12]. Myerson *et al.* also utilized the consensus tool in order to combine individual contours and gather consensus on the elective CTVs to be used in IMRT planning for anal and rectal cancers [13]. Additionally, Lawton et al. made use of the consensus tool to find the levels of discrepancy in the CTV definition of pelvic lymph nodes by 14 genitourinary radiation oncologists [11].

The primary objective of the current study is to evaluate the ability of the consensus software tool to analyze and generate structure estimates similar to the ones generated independently by a consensus of expert physicians. Other applications of the tool would include training resident/medical students on contouring oncology structures and generating atlases for auto-segmentation algorithms.

METHODS AND MATERIALS

Dataset

Materials from the RTOG Breast Cancer Atlas Project were utilized in order to evaluate the ability of the software tool to mimic experts' estimations of consensus manually. The data included pre- and post- consensus contours generated by expert radiation oncologists independent of the consensus tool. Treatment planning CT scans from three representative cases of breast cancer patients were used. The scans were 3D free-breathing scans with 120 kV and 250mA source and slice thickness of 2.5 mm. Nine physicians from 8 institutions participated in the study, more details could be found in [14].

Manual consensus generation

Participating physicians independently contoured target volumes and OARs on the same three CT scans. Physicians were instructed to contour specified structures using their own segmentation tools with a window/level setting of 600/400[14]. For the first run of segmentations, instructions as to how to delineate the volumes were not provided to the physicians in order to quantify multi-institutional and multi-observer variability [14]. For the second run of segmentations, detailed instructions to delineate the volumes based on a consensus document from the participating radiation oncologists were provided. Then, the consensus contour set was generated by manually averaging the contours from the second run in a follow-up face-to-face meeting. The final averaged contours can be found at: http://www.rtog.org/atlases/breastCancer/main.html.

Consensus software tool description

The tool can function as a plug-in within CERR and its user interface consists of multiple metric panels. In addition, an interactive figure is used for selecting the operating confidence agreement level, which is estimated from the agreement probability maps, as illustrated in figure 1. Examples of physicians' and software generated contours are provided in figure 2.

In this study, the contoured structures from all physicians were imported from DICOM files and were merged onto a single scan for each case. The following three statistical metrics are presented in the different panels:

1. Apparent agreement—this is the apparent agreement between the experts, where the apparent agreement probability of the i^{th} voxel is calculated as:

$$p_i=\frac{\sum_{j=1}^{m} r_j}{m}, i=1, \ldots, n \tag{1}$$

r_j = rate by which the j^{th} expert selects the current voxel.

In this case of inter-observer analysis, it is 0 or 1.

m=number of experts

n=number of voxels selected by any of the experts

2. Kappa-corrected agreement—In order to account for agreement among participating physician experts beyond what could be expected by chance, the consensus tool calculates generalized kappa statistics [8]. Kappa is a commonly used measure of agreement in imaging studies. Generalized kappa is recommended for evaluating inter-rater agreement when there are more than two raters [15]. The kappa coefficient was calculated using the following formula:

$$Kappa \quad (\kappa)=\frac{(Apparent\_Agreement-Chance\_Agreement)}{(1-Chance\_Agreement)} \tag{2}$$

Chance_Agreement = the expected agreement by chance alone and is based on marginal totals: $\prod_{j=1}^{m} p\left(r_j=1\right)+\prod_{j=1}^{m} p\left(r_j=0\right)$

The metric yields a value ranging between -1 and 1, with a value of -1 representing complete disagreement, 0 representing no agreement above chance, and 1 representing perfect agreement. Interested reader is referred to [16] for more details. Following Landis and Koch's benchmarks for the interpretation of strength of agreement, kappa <0.00 is poor, 0.00-0.20 is slight, 0.21-0.40 is fair, 0.41-0.60 is moderate, 0.61-0.80 is substantial, and 0.81-1.00 is almost perfect agreement [17].

3. Consensus generation by maximum likelihood estimation—The STAPLE algorithm is utilized by the software to generate consensus contours. In this approach, the true contouring decisions at each image voxel are formulated as maximum-likelihood estimates from the observed contours by optimizing sensitivity and specificity parameters of each expert's performance using the EM algorithm assuming a binomial distribution. Using the collection of manually drawn contours provided by raters, STAPLE computes a probabilistic estimate of the ‘true contour’ that represents the desired anatomy or tumor and measures the performance of each individual segmentation. The probabilistic estimate of the true segmentation is created utilizing three factors: an estimate of the optimal combination of the segmentations, the weight of each segmentation depending on performance, and the incorporation of an *a priori* model for the spatial distribution of structures being contoured [9].

Statistical evaluation

In order to test the ability of the tool to create a ‘true contour’ similar to that made by human consensus, Dice metric was calculated for each pre-consensus STAPLE estimate and the final consensus panel structures. Dice similarity coefficient (DSC) measures the similarity between two sets and was calculated using the following formula:

$$\mathrm{DSC}=\frac{2\,(\mathrm{A}\cap\mathrm{B})}{(\mathrm{A}+\mathrm{B})}, \tag{3}$$

where A is the first set, B is the second set, and is the intersection of the two sets. A closely related similarity metric called Dice-Jaccard is also calculated [18]:

$$\mathrm{DJC}=\frac{(\mathrm{A}\cap\mathrm{B})}{(\mathrm{A}\cup\mathrm{B})}. \tag{4}$$

EXPERIMENTAL RESULTS

The variability of each set of contours for all 3 cases is summarized in tables 1-3. Each structure had 7, 8, or 9 expert contours used for analysis. The highest levels of agreement were seen in the left-breast, case-A lumpectomy, and heart. The highest agreements for pre-consensus contours were in case-A lumpectomy (κ = .82) and left-breast (κ =.81). The highest agreement for post-consensus contours was in case-A left-breast contours (κ=.88). The most significant improvements in agreement were seen in chestwalls and breasts. For chestwall, kappa increased from .66 to .77; for left-breast, from .81 to .88; and for right-breast, from .71 to .80. The higher levels of agreement seen in the left-breast compared to right-breast may reflect the general clinical consensus that the target volume in case-A was the left breast following lumpectomy. In contrast, the higher variation in the right-breast is likely due to clinical disagreement on the extent of the chestwall to include within the targeted breast contour as illustrated in figure 2. As expected, significant variation (agreement levels of moderate or below) was seen in the pre- and post-consensus delineations of the supraclavicular nodes (SCV), internal mammary nodes (IMN), and the axillary apex. This is likely due to vague imaging details on CT scans. The most significant improvement was seen in the axillary apex, with a decrease in union volume from 40.97 to 28.02 and an increase in kappa from .14 to .39. However, the intersection volume remained at 0. Furthermore, the union volumes for SCV and IMN increased from pre- to post-consensus.

Dice calculations for pre-consensus STAPLE estimates and final consensus panel structures reached .80 or greater for heart, left-breast, right-breast, case-A lumpectomy, case-C lumpectomy, and chestwall. The highest DSCs were between STAPLE estimated left-breast contour and the consensus panel's final contour (DSC=0.955 at 100% confidence) as illustrated in figure 3a. The heart, case-A lumpectomy, left-breast, and right-breast all exhibited DSC greater than 0.9, which demonstrate the high level of similarity between the final manual consensus structures and estimated contours by the tool. Although the DSCs were not as high for SCV, IMN and axillary apex they still reached a range of .70-.79 as seen in figures 4 and 5. In figure 5, it is noted that DSC of the axillary apex decreased rapidly at higher confidence levels, which is likely due to the high variability in these contours (pre-consensus κ = .14). On the other hand, the DJC values followed a similar trend to DSC but provided more conservative estimates. This is expected since DSC and DJC metrics are related via $DSC = 2DJC/(DJC+1)$ [19] as can be seen in figures 3, 4, and 5.

High levels of similarity were also seen between the final consensus panel's contours and the post-consensus STAPLE estimations for the heart, left-breast, case-A lumpectomy, chestwall, case-B SCV, right-breast, and case C-lumpectomy: each had a DSC/DJC metrics of 0.95/0.90, 0.96/0.92, 0.94/0.88, 0.92/0.85, 0.85/0.74, 0.93/0.87, 0.90/0.82, respectively. This indicates that these final structures generated by the consensus panel were similar to the final contours delineated by the physicians. Lower levels of similarity were seen for the case-C SCV (0.64/0.48), case-C IMN (0.60/0.43), case-B IMN (0.70/0.54) and axillary apex (0.81/0.69), indicating the inherent difficulty of generating a final structure based on contours of high variability.

DISCUSSION

Several studies have reported on delineation variability of the GTV and CTV for radiotherapy treatment planning, particularly in the case of prostate cancer, however, a wider range of variations is noticed in head and neck, esophageal, and lung tumors [6]. In order to facilitate this process and provide atlases based on experts’ agreement studies, we have created an interactive, user-friendly software tool that allows for analysis and estimation of consensus contour from experts' data. In this study, we validated the ability of this tool to generate a contour estimate similar to manual expert's consensus contours and demonstrated its potential role in radiation oncology research. This tool has already been utilized for creating consensus structures in a variety of ways as described earlier [11-13]. Ultimately, the tool may aid reader agreement studies with a variety of applications, including developing reliable diagnostic rules, understanding variability in treatment recommendations, evaluating effects of training on interpretation consistency, determining the reliability of classification systems (lexicon development), and comparing consistency of different sources of medical information [15]. The consensus tool may also be used as a learning tool for residents. For instance, residents could draw contours and gauge their progress compared to experts' consensus.

In this validation study, contours from three representative breast cancer cases were utilized to evaluate the consensus tool. These cases provided an independent dataset of pre- and post- consensus contours generated by expert radiation oncologists. However, in future studies we expect to apply the tool to more complex sites and consider possible overlap between adjacent structures using multi-category labeling STAPLE [9].

Although the consensus software tool can facilitate reader agreement studies, there might be a potential pitfall in relying totally on the objectivity of the software tool and not taking advantage of its interactive nature, especially in cases of outliers. As illustrated by Michalski *et al.* and Myerson *et al.*, the tool can be used in a way that does not eliminate the input of expert physicians. Furthermore, it should be noted that STAPLE estimates generated by the consensus tool relies on the contours delineated by physicians; the consensus tool's validity is limited by the validity of the physicians' contours. While the physicians participating in a reader agreement study could agree on a region being the correct area of interest, it cannot be unequivocally concluded that the physicians are correct. In efforts to minimize this error, it is important to ensure that the physicians participating in atlas generation are experts in their respective fields.

One of the main challenges facing clinicians when delineating structures is the lack of contrast on CT images to define boundaries. Van de Water *et al.* found that providing guidelines for OARs in the head and neck area and CT-based illustrations could not resolve the difficulties and uncertainties in defining the parts of the tongue with minor salivary glands [2]. Our results also illustrated the difficulties in delineating SVC, IMN, and the axillary apex despite a consensus panel meeting. However, CT remains the gold standard for treatment planning today, and providing better atlases for practicing clinicians is therefore a necessity. An alternative is to provide computer-assisted segmentation tools for defining the physical [1,20] or even biological target volumes [21].

CONCLUSIONS

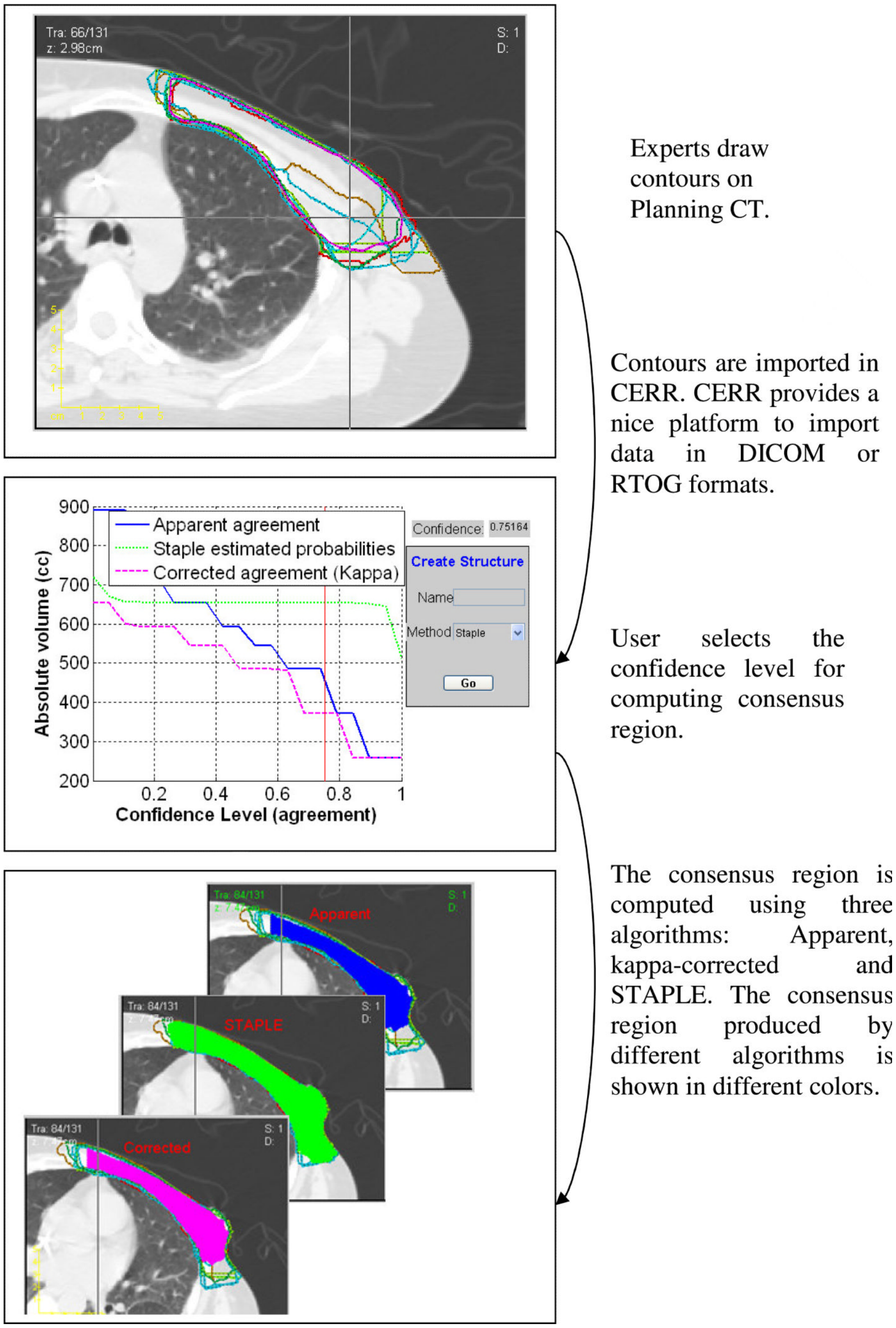
We have demonstrated that the developed consensus software tool could be a useful aid in radiation oncology research. Using different statistical methods and STAPLE estimations, the tool can provide means to generate and analyze consensus estimates similar to those generated manually by expert physicians and therefore expedite and facilitate the process of creating a final contour.

Acknowledgments

This work was supported in part by Barnes-Jewish Hospital Foundation (6661-01) and by NIH U24 grant CA81647.

REFERENCES

1. Price GJ, Moore CJ. A method to calculate coverage probability from uncertainties in radiotherapy via a statistical shape model. Phys Med Biol. 2007; 52:1947–1965. [PubMed: 17374921]
2. van de Water TA, Bijl HP, Westerlaan HE, Langendijk JA. Delineation guidelines for organs at risk involved in radiation-induced salivary dysfunction and xerostomia. Radiother Oncol. 2009; 93:545–552. [PubMed: 19853316]
3. van Mourik AM, Elkhuizen PH, Minkema D, Duppen JC, van Vliet-Vroegindeweij C. Multiinstitutional study on target volume delineation variation in breast radiotherapy in the presence of guidelines. Radiother Oncol. 2010; 94:286–291. [PubMed: 20199818]
4. 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]
5. International Commission on Radiation Units and Measurements (ICRU). Prescribing, Recording, and Reporting Photon Beam Therapy (Supplement to ICRU Report 50). Bethesda, MD: 1999.
6. Weiss E, Hess CF. The Impact of Gross Tumor Volume (GTV) and Clinical Target Volume (CTV) Definition on the Total Accuracy in Radiotherapy. Strahlentherapie und Onkologie. 2003; 179:21–30. [PubMed: 12540981]
7. Rasch; Steenbakkers; Herk, v. Target Definition in Prostate, Head, and Neck. Seminars in Radiation Oncology. 2005; 15:136–145. [PubMed: 15983939]
8. Fleiss, JL.; Levin, BA.; Paik, MC. Statistical methods for rates and proportions. J. Wiley; Hoboken, N.J.: 2003.
9. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging. 2004; 23:903–921. [PubMed: 15250643]
10. Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys. 2003; 30:979–985. [PubMed: 12773007]
11. Lawton CA, Michalski J, El-Naqa I, et al. Variation in the definition of clinical target volumes for pelvic nodal conformal radiation therapy for prostate cancer. Int J Radiat Oncol Biol Phys. 2009; 74:377–382. [PubMed: 18947941]
12. Michalski JM, Lawton C, El Naqa I, et al. Development of RTOG Consensus Guidelines for the Definition of the Clinical Target Volume for Postoperative Conformal Radiation Therapy for Prostate Cancer. Int J Radiat Oncol Biol Phys. 2009
13. Myerson RJ, Garofalo MC, Naqa IE, et al. Elective Clinical Target Volumes for Conformal Therapy in Anorectal Cancer: An Radiation Therapy Oncology Group Consensus Panel Contouring Atlas. Int J Radiat Oncol Biol Phys. 2008
14. 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]
15. Crewson PE, Applegate KE. Data Collection in Radiology Research. Am. J. Roentgenol. 2001; 177:755–761. [PubMed: 11566667]
16. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005; 37:360–363. [PubMed: 15883903]
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]
18. Wondergem BCM. Bommel, Pv, Weide, TPvd. Matching Index Expressions for Information Retrieval. Information Retrieval. 2000; 2:24.
19. Shattuck D, Sandor-Leahy S, Schaper K, Rottenberg D, Leahy R. Magnetic Resonance Image Tissue Classification Using a Partial Volume Model. NeuroImage. 2001; 13:21.
20. Sims R, Isambert A, Gregoire V, et al. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. Radiother Oncol. 2009; 93:474–478. [PubMed: 19758720]
21. Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging. 2010



Experts draw contours on Planning CT.

Contours are imported in CERR. CERR provides a nice platform to import data in DICOM or RTOG formats.

User selects the confidence level for computing consensus region.

The consensus region is computed using three algorithms: Apparent, kappa-corrected and STAPLE. The consensus region produced by different algorithms is shown in different colors.

Figure 1.

Workflow of the consensus tool.

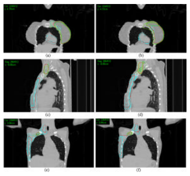


Figure 2.

(a) Case-A pre-consensus contours of the heart, left-breast, and lumpectomy with 95% STAPLE estimations (red) (b) Case-A post-consensus contours of the heart, left-breast, and lumpectomy with final consensus panel contours (red) (c) Case-B pre-consensus contours of the chestwall and supraclavicular nodes with 95% STAPLE estimations (red) (d) Case-B post-consensus contours of the chestwall and SVC nodes with final consensus panel contours (red) (e) Case-C pre-consensus contours of the right-breast, axillary apex, and SVC nodes with 95% STAPLE estimations (red) (f) Case-C post-consensus contours of the right breast, axillary apex, and SVC nodes with final consensus panel contours (red).

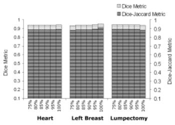


Figure 3.

Comparison analysis of case-A using DSC and DJC with pre-consensus at different confidence levels and final consensus panel contours.

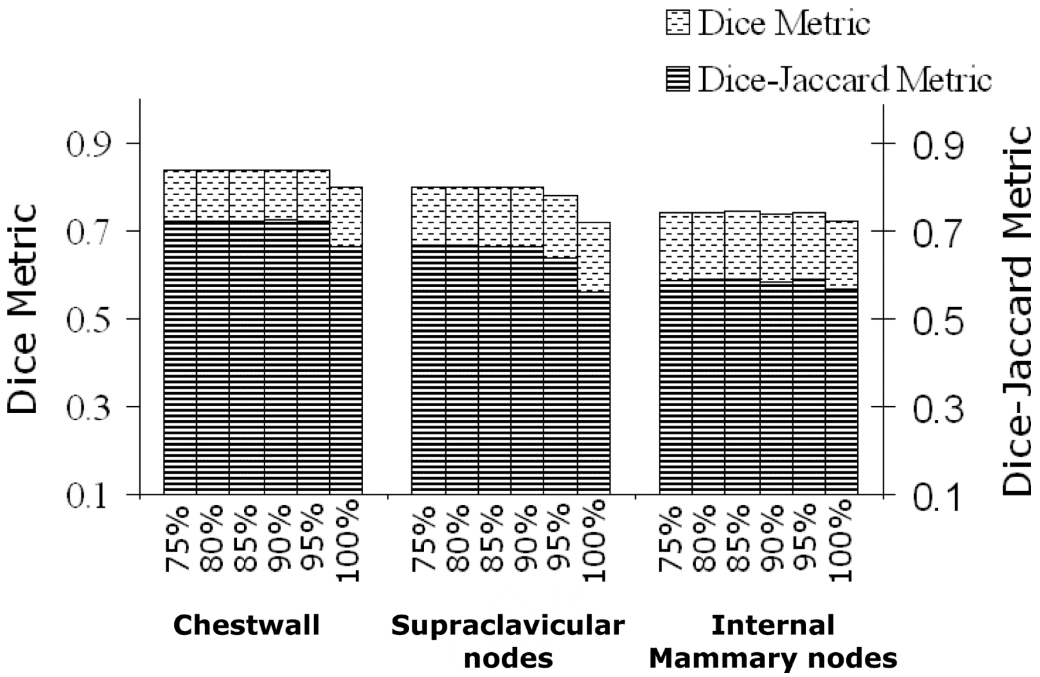


Figure 4.

Comparison analysis of case-B using DSC and DJC with pre-consensus at different confidence levels and final consensus panel contours.

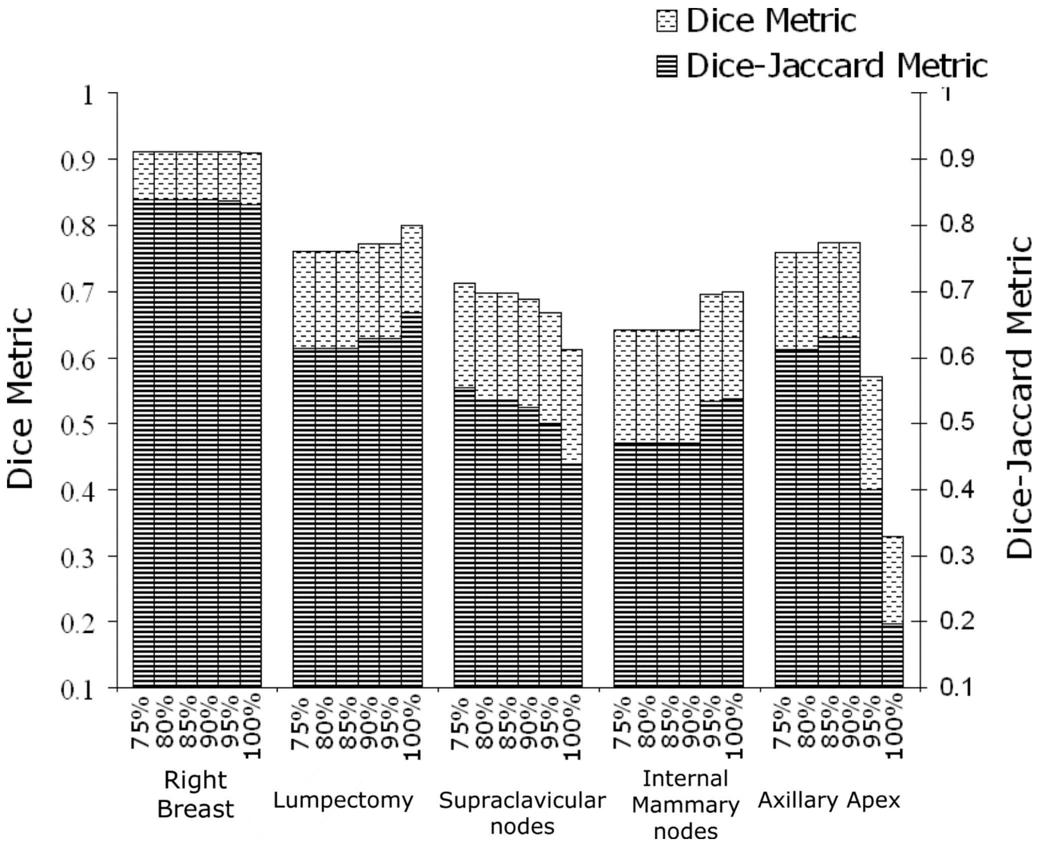


Figure 5.

Comparison analysis of case-C using DSC and DJC with pre-consensus at different confidence levels and final consensus panel contours.

Table 1

Summary analysis results for Case-A.

Structure Measure	Left-Breast		Lumpectomy		Heart	
	Pre	Post	Pre	Post	Pre	Post
#experts	8	9	8	9	8	9
Vol.-Max	1011.84	1032.16	40.15	40.67	555.28	516.50
Vol.-Min	655.88	744.55	26.07	24.13	198.16	296.51
Vol.-Avg.	813.50	821.06	29.99	29.65	447.39	452.01
Vol.-Med.	826.12	809.34	29.33	27.09	474.00	462.74
Vol.- Std.	117.79	86.14	4.55	5.13	108.30	69.01
Vol.-Intersection	498.18	658.86	21.28	21.83	187.45	267.27
*Vol.-Union	1158.72	1137.39	43.52	42.53	580.30	573.05
**Agreement-Sensitivity (Avg±SD)	0.88±0.11	0.95±0.03	0.93±0.048	0.95±0.04	0.88±0.20	0.89±0.12
**Agreement- Specificity (Avg±SD)	0.98±0.02	0.99±0.02	0.97±0.04	0.96±0.05	0.98±0.030	0.98±0.02
***Kappa-statistics	0.81 Almost-Perfect	0.88 Almost-Perfect	0.82 Almost-Perfect	0.83 Almost-Perfect	0.77 Substantial	0.81 Almost-Perfect

All p-values <0.0001.

* The volume of the region containing all expert contours

** STAPLE estimates

*** Corrected for chance.

Table 2

Summary analysis results for Case-B.

Structure	Chestwall		SCV		IMN	
Measure						
	Pre	Post	Pre	Post	Pre	Post
#experts	8	9	8	9	7	9
Vol.-Max	958.12	688.11	51.91	62.77	4.43	16.31
Vol.-Min	362.46	447.27	7.90	11.91	1.05	2.86
Vol.-Avg.	545.73	559.42	22.38	36.77	3.10	6.22
Vol.-Med.	471.10	542.08	23.12	36.43	3.87	4.71
Vol.-Std.	213.08	91.36	14.47	16.17	1.49	4.34
Vol.-intersection	200.96	252.92	2.73	2.31	0.20	0.52
*Vol.-union	1051.03	891.07	68.96	113.71	9.25	25.45
**Agreement-sensitivity (Avg±SD)	0.72±0.17	0.79±0.10	0.57±0.26	0.61±0.22	0.51±0.26	0.54±0.14
**Agreement-specificity (Avg±SD)	0.99±0.02	0.99±0.01	0.98±0.03	0.97±0.04	0.99±0.01	0.99±0.02
***Kappa-statistics	0.66 Substantial	0.77 Substantial	0.43 Moderate	0.43 Moderate	0.40 Fair	0.38 Fair

All p-values <0.0001.

*The volume of the region containing all expert contours

**STAPLE estimates

***Corrected for chance.

Table 3

Summary analysis results for Case-C.

Structure	SCV		Axillary Apex		IMN		Lumpectomy		Right-Breast	
Measure										
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# experts	8	8	8	9	8	9	8	9	8	9
Vol.-Max	40.17	32.48	19.22	18.08	10.85	9.00	31.19	21.64	1087.12	737.44
Vol.-Min	2.78	9.41	2.86	2.40	0.15	0.26	8.38	3.66	400.44	481.82
Vol.-Avg.	16.86	23.24	8.07	8.99	3.30	3.96	13.22	8.87	588.91	615.91
Vol.-Med.	18.91	24.34	6.74	6.31	3.04	3.53	9.98	7.31	498.87	654.51
Vol.-Std.	12.68	7.07	5.64	6.24	3.29	3.15	7.67	5.53	221.21	94.25
Vol.-intersection	0	0.33	0	0	0.06	0.07	5.24	3.16	318.31	386.78
[*]Vol.-union	66.01	76.29	40.97	28.02	12.87	15.19	32.56	23.99	1124.96	833.87
[**]Agreement - sensitivity (Avg±SD)	0.46±0.28	0.45±0.16	0.33±0.32	0.50±0.32	0.40±0.28	0.38±0.28	0.76±0.15	0.74±0.22	0.84±0.10	0.87±0.10
[**]Agreement-specificity (Avg±SD)	0.98±0.03	0.98±0.01	0.98±0.01	0.99±0.01	0.99±0.02	0.99±0.01	0.97±0.07	0.97±0.06	0.98±0.03	0.98±0.02
[***]Kappa-statistics	0.27 Fair	0.33 Fair	0.14 Slight	0.39 Fair	0.27 Fair	0.30 Fair	0.59 Moderate	0.57 Moderate	0.71 Substantial	0.80 Substantial

All p-values <0.0001

[*] The volume of the region containing all expert contours

[**] STAPLE estimates

[***] Corrected for chance.