Published in final edited form as:
Int J Radiat Oncol Biol Phys. 2014 May 1; 89(1): 214–221. doi:10.1016/j.ijrobp.2014.01.010.

Statistical modeling approach to quantitative analysis of inter-observer variability in breast contouring

Jinzhong Yang, Ph.D.[1], **Wendy A. Woodward, M.D.**[2], **Valerie K. Reed, M.D.**[2], **Eric A. Strom, M.D.**[2], **George H. Perkins, M.D.**[2], **Welela Tereffe, M.D.**[2], **Thomas A. Buchholz, M.D.**[2], **Lifei Zhang, Ph.D.**[1], **Peter Balter, Ph.D.**[1], **Laurence E. Court, Ph.D.**[1], **X. Allen Li, Ph.D.**[3], and **Lei Dong, Ph.D.**[1,4]

[1]Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX

[2]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX

[3]Department of Radiation Oncology, Medical College of Wisconsin, Milwaukee, WI

[4]Scripps Proton Therapy Center, San Diego, CA

Abstract

Purpose—To develop a new approach for inter-observer variability analysis.

Methods and Materials—Eight radiation oncologists specializing in breast cancer radiotherapy delineated a patient's left breast from scratch and from a template that was generated using deformable image registration. Three of the radiation oncologists had previously received training in RTOG consensus contouring for breast cancer atlas. The simultaneous truth and performance level estimation algorithm was applied to the eight contours delineated from scratch to produce a group-consensus contour. Individual Jaccard scores were fitted to a beta distribution model. We also applied this analysis to two more patients which were contoured by nine breast radiation oncologists from eight institutions.

Results—The beta distribution model had a mean of 86.2%, standard deviation of 5.9%, skewness of −0.7, and excess kurtosis of 0.55, exemplifying broad inter-observer variability. The three RTOG-trained physicians had higher agreement scores than average, indicating that their contours were close to the group-consensus contour. One physician had high sensitivity but lower specificity than the others, which implies that he/she tended to contour a structure larger than others. Two other physicians had low sensitivity but similar specificity as others, which implies

Corresponding author: Jinzhong Yang, Ph.D., Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Unit 94, Houston, TX 77030; (713) 792-2814; jyang4@mdanderson.org.

Meeting presentations: This work was partially presented at the 2010 RSNA meeting.

Conflict of interest: none.

that they tended to contour a structure smaller than others. With this information, they could adjust their contouring practice to be more consistent with others if desired. When contouring from the template, the beta distribution model had a mean of 92.3%, standard deviation of 3.4%, skewness of −0.79, and excess kurtosis of 0.83, which indicated a much better consistency among individual contours. Similar results were obtained for the analysis of two additional patients.

Conclusions—The proposed statistical approach was able to measure inter-observer variability quantitatively and to identify individuals who tend to contour differently from the others. The information could be useful as feedback to improve contouring consistency.

Introduction

In breast cancer radiotherapy, intensity-modulated radiation therapy (IMRT) is able to deliver a more conformal radiation dose than the traditional opposed tangential field technique (1, 2). Delineation of the whole breast is an important step for the breast IMRT (3). However, several studies have shown large intra- and inter-observer variability in contouring the whole breast, even for experienced radiation oncologists (4–6). This variability was taken into account in the National Surgical Adjuvant Breast and Bowel Project (NSABP) protocol B-39/Radiation Therapy Oncology Group (RTOG) protocol 0413 randomized study. Although RTOG has established consensus contouring for breast cancer atlas (7), intra- or inter-observer variability still exists because of differences in training, clinical experience, and quality of computed tomography (CT) images on which the contours are drawn.

Quantitative approaches have been proposed to demonstrate the existence of inter-observer variability in previous studies (3–6); however, these studies have not been able to gain insight into the causes of this variability or to suggest a feasible solution to improve consistency in contouring. Moreover, previous studies measured inter-observer variability by computing the distance/difference between the individual contour and the average contour of all observers (4, 6). This measurement is prone to the impact of outliers. If one observer draws the contour significantly different from the others, whose contours are similar, the one outlier will exaggerate the inter-observer variability. Therefore, a new quantitative approach is needed to overcome this deficiency.

The goals of this study were to develop a new quantitative approach to analyze inter-observer variations and recommend a learning experience from the result of the quantitative analysis. We applied this method to two independent studies on contouring the whole breast. We generated a group-consensus contour from a set of individual contours delineated by different physicians. The agreement between individual contour and the group-consensus contour was measured and used as the score of agreement. These scores were then fitted into a mathematical model for quantitative analysis. With this model, we were able to come up with possible solutions to improve the contouring consistency. We also compared two contouring methods, contouring from scratch and contouring from a template, to explore the potential benefits of using a template to improve contouring consistency.

Methods and Materials

Patient data

This retrospective study has been approved by the institutional review board of MD Anderson Cancer Center (Approval number: RCR03-0400). A template patient who underwent consecutive treatment to the left breast was identified from our institutional database. A team consisting of radiation oncologists, a breast surgical oncologist, and a radiologist collectively defined the whole breast clinical target volume (CTVwb) on the CT image of this template patient on a Pinnacle Treatment Planning System (Philips Medical Systems, Fitchburg, WI). A test patient with left breast treatment was also identified who underwent CT simulation in the same position as the template patient. Eight radiation oncologists, including seven breast cancer radiation oncologists and one senior medical resident, delineated the CTVwb on the CT image of the test patient using the following two methods: contouring from scratch and contouring from the template, as described below in details. Three of these eight radiation oncologists had previously received training in RTOG consensus contouring for breast cancer atlas prior to this study.

Contouring from scratch was performed on the Pinnacle Treatment Planning System too. The external skin contour was first automatically generated, and then each physician used editing tools to further contour the CTVwb individually. When contouring from the template, we first performed deformable image registration between the template patient and the test patient using an accelerated “demons” algorithm (8) and then mapped the CTVwb contour from the template patient to the test patient using the deformation field that resulted from the deformable image registration (9). Each physician individually modified this autosegmented contour as necessary.

Group-consensus contours

We introduced group-consensus contours to quantify the inter-observer variability. The group-consensus contour is defined as the best agreement of a group of individual contours. The difference between each individual contour and the group-consensus contour was measured and used as the agreement score for the individual contour. Let G denote the volume enclosed by the group-consensus contour and D denote the volume enclosed by one individual contour. We computed the sensitivity and specificity as the score of agreement for the individual contour D (10). Sensitivity and specificity were defined as follows:

$$Sensitivity=\frac{D\cap G}{G},\ \ Specificity=\frac{\overline{D}\cap\overline{G}}{\overline{G}}, \quad (1)$$

where $\bar{D}$ and $\bar{G}$ denote the space outside the volumes D and G, respectively. Sensitivity was the true-positive rate when comparing the individual contour against the group-consensus contour, while specificity described the true-negative rate. The combination of sensitivity and specificity quantifies the inter-observer variability.

The group-consensus contour representing the best agreement of a group of contours can be obtained by maximizing the sensitivity and specificity between each individual contour and the group-consensus contour. The simultaneous truth and performance level estimation

(STAPLE) algorithm (11) was used to achieve this goal (see Appendix eI for details). The STAPLE algorithm takes into account the spaces both inside (sensitivity) and outside (specificity) the region of interest for a maximum likelihood estimation. In order to balance the weights of sensitivity and specificity in the estimation, we selected a small region enclosing all observer contours and forced the space outside the region of interest to be roughly equivalent to the volume inside the region of interest, i.e., $G \cong \bar{G}$. With this constraint, the sensitivity and specificity were in the same scale and the estimated group-consensus contour best approximated the underlying true segmentation (9).

Statistical modeling to quantitative analysis

The agreement score composed of the sensitivity and the specificity may also be described using the union overlap, i.e., the Jaccard coefficient (12), between an individual contour and the group-consensus contour. The Jaccard coefficient is defined as:

$$Jaccard = \frac{D \cap G}{D \cup G}. \quad (2)$$

It has a value between 0 and 1, with 1 indicating perfect agreement and 0 indicating no overlap. The Jaccard coefficients for a group of contours have been shown to follow a beta distribution (13). We fitted the Jaccard coefficients to a beta distribution model with a probability density function as follows:

$$f(y|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad (3)$$

where y is a variable of the Jaccard coefficient, $\alpha > 0$ and $\beta > 0$ are unknown parameters characterizing the model, and $B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$ is the beta function used as the normalization constant. The parameters α and β were the maximum likelihood estimates from the measured Jaccard coefficients (14). This model can be graphically displayed with a curve, which intuitively illustrates the inter-observer variability (13).

The mean (M), standard deviation (SD), skewness (SK), and excess kurtosis (EK) of the beta distribution model were used to quantitatively evaluate the inter-observer variability (15):

$$M = \frac{\alpha}{\alpha+\beta}, \quad SD = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}, \quad SK = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{\alpha+\beta+2\sqrt{\alpha\beta}}, \quad EK = \frac{6\left[(\alpha-\beta)^2(\alpha+\beta+1)-\alpha\beta(\alpha+\beta+2)\right]}{\alpha\beta(\alpha+\beta+2)\cdot(\alpha+\beta+3)}. \quad (4)$$

The mean value represents the average agreement of all individual contours from the group-consensus contour; the larger the mean value, the smaller the inter-observer variability. Standard deviation denotes the distribution of agreement scores; the smaller the standard deviation, the smaller the inter-observer variability. Skewness represents the symmetry of the beta distribution. Skewness less than 0 means the distribution curve spread toward the left side, with higher possibility of agreement scores less than the mean value, thus tending

to a larger inter-observer variability. Kurtosis is a measure of the peakedness of the beta distribution model. The larger the kurtosis, the less distribution of the scores to the two ends, thus tending to a smaller inter-observer variability. The tools that generated the group-consensus contours and the beta distribution models for quantitative analysis were developed in-house.

Contouring from scratch versus contouring from the template

We analyzed and quantitatively compared the inter-observer variability for two contouring methods: contouring from scratch and contouring from the template. The group-consensus contours were generated first for these two methods. The Jaccard coefficients between individual contours and the group-consensus contour were computed, and then the beta distribution models were built from the Jaccard coefficients. The beta distribution models for both methods were plotted in one coordinate system for comparison and their mean, standard deviation, skewness, and excess kurtosis values were computed for quantitative comparison.

More examples

Additional analysis was performed on two breast cancer patients from an independent source with their contours delineated by nine radiation oncologists specializing in breast radiotherapy from eight institutions (5). Patient A had Stage I (T1c, N0, M0) left breast cancer and Patient B had Stage IIIA (T2, N2, M0) right breast cancer. CT images of both patients were acquired with various external metallic markers. Nine oncologists delineated the left breast volume for Patient A and the right breast volume for Patient B. They were instructed to delineate these structures using their own segmentation tools with a window/level setting of 600/40 for soft tissue. No specific guideline was provided as how to delineate the breast volume. We applied our method to quantitatively analyze the inter-observer variability in contouring these two patients. The sensitivity, specificity, and Jaccard coefficient were recorded for analysis. The beta distribution models were built for both patients to graphically demonstrate the inter-observer variability.

Results

Group-consensus contours

Group-consensus contours were generated for both contouring from scratch and contouring from the template. The physicians' individual contours and the group-consensus contour were plotted overlaid on the CT image, as shown in Figure 1. The inter-observer variability can be visualized from this figure, mostly existing in the medial and lateral borders of the breast. Figure 1 suggests that contouring from the template may have had less inter-observer variability than contouring from scratch. When generating the group-consensus contour, the sensitivity and specificity between individual contour and the group-consensus contour were also measured at the same time for each observer, as listed in Table 1. These values are the maximum likelihood estimates from the STAPLE algorithm and best evaluate the inter-observer variability among the group of eight observers who drew the contours.

Statistical modeling to quantitative analysis

The Jaccard coefficients shown in Table 1 were calculated using Equation (2) for every individual contour. For each contouring method, the Jaccard coefficients were fitted into the beta distribution model to obtain the distribution of observer agreement scores. The distribution models were plotted with each individual's agreement score indicated on the curve (Figure 1). The distribution model showed that the possible scores ranged between 60% and 100% for contouring from scratch, and between 80% and 100% for contouring from the template. Observers 1, 2, and 3 had received training in RTOG consensus contouring for breast cancer atlas. Their scores distributed to the right side of the curve toward a perfect agreement, representing relatively high agreement scores, which means that their contours were close to the group-consensus contour. This implies that the RTOG training may improve consensus contouring. In addition, Table 1 shows that observer 5 had consistent high sensitivity but lower specificity than others, which implies that this observer tended to contour this structure larger than others. On the other hand, observers 4 and 6 had low sensitivity but specificity similar to the others, which implies that these observers tended to contour this structure smaller than others. This was validated by comparing the individual contours with the group-consensus contour. The contour drawn by observer 5 was mostly larger than the group-consensus contour [Figure 2(a)] and the contours drawn by observers 4 and 6 were mostly smaller than the group-consensus contour [Figure 2(b)]. Therefore, our quantitative analysis could be used to instruct a particular physician to adjust his/her contouring practice to be more consistent with others. If an observer has consistently lower specificity than others, he/she may need to contour small when in doubt; on the other hand, if an observer has consistently lower sensitivity than others, he/she may need to contour large when in doubt.

Contouring from scratch versus contouring from the template

We compared the inter-observer variability quantitatively for contouring from scratch and contouring from the template. The beta distribution models were plotted in one coordinate system for comparison (Figure 3). Furthermore, we computed the mean, standard deviation, skewness, and excess kurtosis for these two distribution models. Contouring from scratch had a mean score of 86.2%, standard deviation of 5.9%, skewness of −0.70, and excess kurtosis of 0.55, with the possible scores distributing from 60% to 100%, representing a broad inter-observer variability. However, contouring from the template had a mean score of 92.3%, standard deviation of 3.4%, skewness of −0.79, and excess kurtosis of 0.83, with the possible scores distributing from 80% to 100%, which shows a much-improved consistency among individual contours. This improvement can also be observed from the individual evaluations in Table 1. Observer 6 contoured much more consistently with others when using the template than from scratch, with a change in agreement score from 76.8% to 94.2%. This demonstrates that contouring from a template improved the contouring consistency among the individual observers.

More examples

The sensitivity, specificity, and Jaccard coefficient between individual contour and the group-consensus contour were shown in Table 2 for both patients. We noticed that, when

contouring the left breast of Patient A, Observer 2 had a much lower specificity than others, indicating he contoured much larger than others, which was verified by his contour shown in Figure 4. On the other hand, Observers 1, 4, and 9 had a lower sensitivity when contouring the right breast of Patient B, indicating they contoured smaller than others, which was also validated by their contours in Figure 4. In addition, Figure 4 showed that Observers 1 and 2 had their scores distributing to the left of the curve in contouring one patient, but to the right of the curve in contouring the other patient. This demonstrated the intra-observer variability for them in contouring these two patients. On the other hand, for Observers 3, 4, and 9 who had their scores distributing to the left of the curve in both patients, they had consistent low sensitivity or specificity. With this information, they might adjust their practice to be more consistent with others if necessary.

Discussion

We proposed a quantitative approach to analyzing inter-observer variability in contouring and demonstrated this method in breast cancer radiotherapy. To our knowledge, it is the first time to integrate the STAPLE algorithm and a beta distribution model for such analysis. Through statistical modeling, we were able to graphically visualize the quantified inter-observer variability. This method also facilitated the comparison of inter-observer variability for different contouring methods. In addition, our approach suggested ways to potentially improve an individual's contouring practice when a consensus contour is agreed upon. Although this method was demonstrated in the scope of breast contouring, it is also applicable to other delineation tasks.

Our approach was based on the volume overlap indices and did not contain spatial or shape information of contouring structures. Although our approach can instruct a particular physician when her contouring practice varies from that of colleagues, it cannot tell specifically where to adjust the contour. On the other hand, it worth mentioning that each observer was a radiation oncology breast specialist diligently contouring using their standard clinical approach, and the test patients are common cases in clinic; therefore, each observer's practice is very likely to reproduce when contouring other similar patients, although intra-observer variability may exist. Feedback from our approach may serve as a guideline for individuals to improve contouring consistency if desirable.

Our approach generated a group-consensus contour as the basis for quantitative analysis. Unlike the average contour used in many previous studies (4, 6), the group-consensus contour is not subject to the impact of outliers when it is used for inter-observer variability analysis. The group-consensus contour is generated by the maximum likelihood estimate from the STAPLE algorithm. Because an outlier contour does not agree with others, the STAPLE algorithm, to achieve the maximum likelihood estimate, will assign a small weight to the outlier contour automatically when estimating the group-consensus contour, thus minimizing the impact from the outlier contour. Furthermore, the group-consensus contour is the best estimation of the underlying ground truth of structure segmentation. This originates from the nature of the STAPLE algorithm and has been validated by several previous studies in multi-atlas auto-segmentation (9, 16–18).

Our study compared contouring from scratch and contouring from a template and demonstrated that contouring from the template improved contouring consistency. This conclusion is consistent with those of previous studies (6, 19, 20). Furthermore, contouring from a template required substantially less time and effort than contouring from scratch, especially for structures with low contrast and unclear anatomic boundaries on the CT images (21–23). In addition, we should also note that there are other approaches to facilitating contouring the breast, such as determining the breast boundaries using wires at the time of CT imaging (2), which we have not included in our analysis.

Contouring consistency continues to be an important aspect of radiotherapy. Variation in contouring of course contributes to variation in dose-volume histogram (DVH) reporting and evaluation. Significantly reducing it will improve the accuracy of DVH toxicity and local control correlations in volume based planning. Concerted efforts have been made to improve contouring consistency, such as the RTOG consensus contouring guideline (7) and ASTRO eContouring Learning Lab (24). Our findings verify that RTOG guideline improved contouring consistency and suggest an effective way to improve contouring consistency through consensus contouring training. Our findings also suggest using a well-defined template as the starting point for contouring to improve consistency. Furthermore, our approach is able to identify individuals whose practice tends to contouring larger or smaller than others. Bearing this information in mind, those individuals are able to adjust their practice to be more consistent with others, if consistency is desired.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Kathryn Hale from the Department of Scientific Publication for reviewing the manuscript. This research was supported in part by the National Institutes of Health through MD Anderson's Cancer Center Support Grant CA016672.

References

1. Chang SX, Deschesne KM, Cullip TJ, et al. A comparison of different intensity modulation treatment techniques for tangential breast irradiation. International Journal of Radiation Oncology*Biology*Physics. 1999; 45:1305–1314.
2. Ahunbay EE, Chen G-P, Thatcher S, et al. Direct aperture optimization–based intensity-modulated radiotherapy for whole breast irradiation. International Journal of Radiation Oncology*Biology*Physics. 2007; 67:1248–1258.
3. Landis DM, Luo WX, Song J, et al. Variability among breast radiation oncologists in delineation of the postsurgical lumpectomy cavity. International Journal of Radiation Oncology Biology Physics. 2007; 67:1299–1308.
4. Hurkmans CW, Borger JH, Pieters BR, et al. Variability in target volume delineation on CT scans of the breast. International Journal of Radiation Oncology*Biology*Physics. 2001; 50:1366–1372.
5. 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. International Journal of Radiation Oncology*Biology*Physics. 2009; 73:944–951.

6. Reed VK, Woodward WA, Zhang L, et al. Automatic Segmentation of Whole Breast Using Atlas Approach and Deformable Image Registration. International Journal of Radiation Oncology*Biology*Physics. 2009; 73:1493–1500.
7. RTOG. Breast Cancer Atlas for Radiation Therapy Planning: Consensus Definitions. Cited June 18, 2013; Available from: http://www.rtog.org/CoreLab/ContouringAtlases/BreastCancerAtlas.aspx
8. Wang H, Dong L, Lii MF, et al. Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy. International Journal of Radiation Oncology*Biology*Physics. 2005; 61:725–735.
9. Yang, J.; Zhang, Y.; Zhang, L., et al. Automatic Segmentation of Parotids from CT Scans Using Multiple Atlases. In: vanGinneken, B.; Murphy, K.; Heimann, T., et al., editors. Medical Image Analysis for the Clinic: A Grand Challenge. Beijing, China: MICCAI Society; 2010. p. 323-330.
10. Altman DG, Bland JM. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. BMJ. 1994:308. [PubMed: 8124120]
11. 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]
12. Jaccard P. The Distribution of the Flora in the Alpine Zone. New Phytologist. 1912; 11:37–50.
13. Yang J, Wei C, Zhang L, et al. A statistical modeling approach for evaluating auto-segmentation methods for image-guided radiotherapy. Computerized Medical Imaging and Graphics. 2012; 36:492–500. [PubMed: 22673541]
14. Hahn, GJ.; Shapiro, SS. Statistical Models in Engineering. New York, NY: John Wiley & Sons, Inc; 1994.
15. Gupta, AK.; Nadarajah, S. Mathematical properties of the Beta distribution. In: Gupta, AK.; Nadarajah, S., editors. Handbook of Beta Distribution and Its Applications. New York, NY: CRC Press; 2004. p. 33-54.
16. Yang J, Amini A, Williamson R, et al. Automatic contouring of brachial plexus using a multi-atlas approach for lung cancer radiotherapy. Practical Radiation Therapy. 2013 In press.
17. Stapleford LJ, Lawson JD, Perkins C, et al. Evaluation of Automatic Atlas-Based Lymph Node Segmentation for Head-and-Neck Cancer. International Journal of Radiation Oncology * Biology * Physics. 2010; 77:959–966.
18. Han, X.; Hibbard, LS.; O'Connell, NP., et al. Automatic Segmentation of Parotids in Head and Neck CT Images using Multi-atlas Fusion. In: vanGinneken, B.; Murphy, K.; Heimann, T., et al., editors. Medical Image Analysis for the Clinic: A Grand Challenge. 2010. p. 297-304.
19. Chao KSC, Bhide S, Chen H, et al. Reduce in Variation and Improve Efficiency of Target Volume Delineation by a Computer-Assisted System Using a Deformable Image Registration Approach. International Journal of Radiation Oncology*Biology*Physics. 2007; 68:1512–1521.
20. Wang H, Garden AS, Zhang L, et al. Performance Evaluation of Automatic Anatomy Segmentation Algorithm on Repeat or Four-Dimensional Computed Tomography Images Using Deformable Image Registration Method. International Journal of Radiation Oncology*Biology*Physics. 2008; 72:210–219.
21. Amini A, Yang J, Williamson R, et al. Dose Constraints to Prevent Radiation-Induced Brachial Plexopathy in Patients Treated for Lung Cancer. International Journal of Radiation Oncology*Biology*Physics. 2012; 82:e391–e398.
22. Teguh DN, Levendag PC, Voet PWJ, et al. Clinical Validation of Atlas-Based Auto-Segmentation of Multiple Target Volumes and Normal Tissue (Swallowing/Mastication) Structures in the Head and Neck. International Journal of Radiation Oncology * Biology * Physics. 2010; 81:950–957.
23. Voet PWJ, Dirkx MLP, Teguh DN, et al. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. Radiotherapy and Oncology. 2011; 98:373–377. [PubMed: 21269714]
24. ASTRO. eContouring Learning Lab. Cited August 27, 2013; Available from: https://www.astro.org/Meetings-and-Events/2013-Annual-Meeting/Registration-Information/eContouring-Learning-Lab.aspx

Summary

We developed a statistical modeling approach to evaluate inter-observer variability in contouring the whole breast for radiotherapy treatment planning. This approach verified that training based on RTOG consensus contouring guideline improved contouring consistency. In addition, the method could recommend ways to improve an individual's contouring practice. The study also demonstrated that the use of a well-defined template as the starting point for contouring will improve contouring consistency.

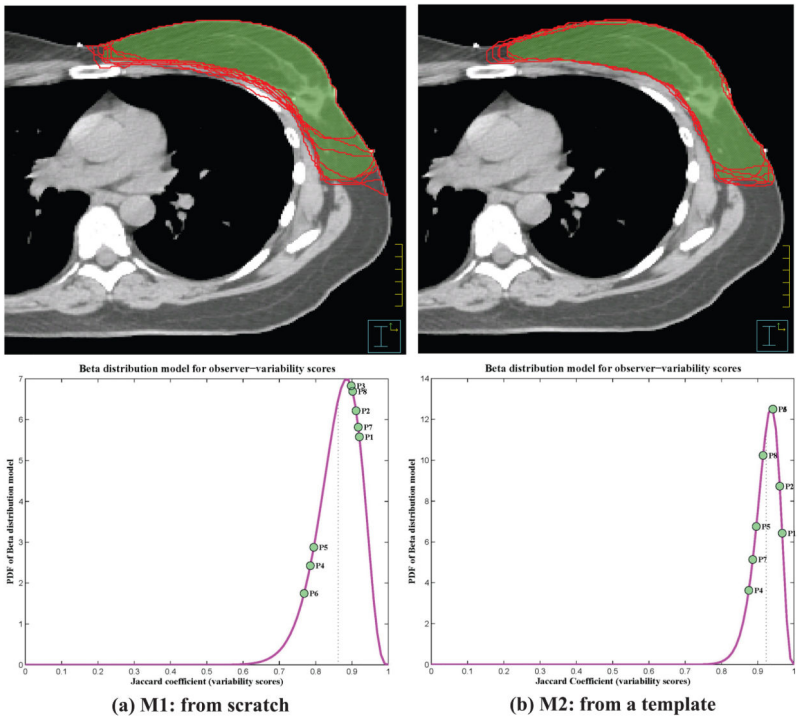


Figure 1.

Top row: The individual contours by eight physicians (red solid lines) and the group-consensus contour (green colorwash) for the left whole breast plotted overlaid on the CT image. Bottom row: Visualization of beta distribution models (as probability density function [PDF]) with individual agreement score on the curve. (For interpretation of the references to color, the reader is referred to the web version of this article.)

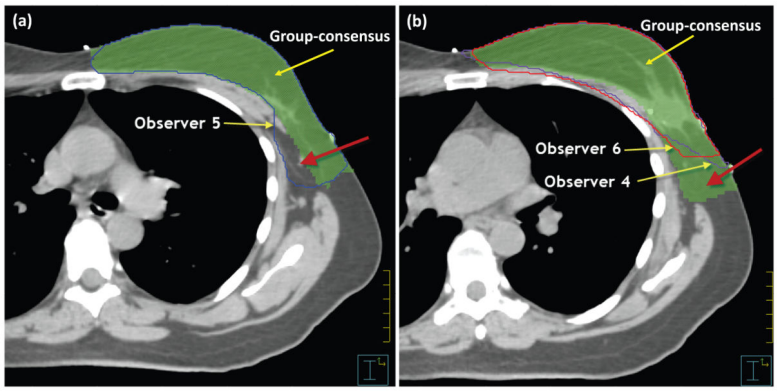


Figure 2.

(a) Contour drawn by Observer 5 showed larger than the group-consensus contour; while (b) contours drawn by Observers 4 and 6 showed smaller than the group-consensus contour.

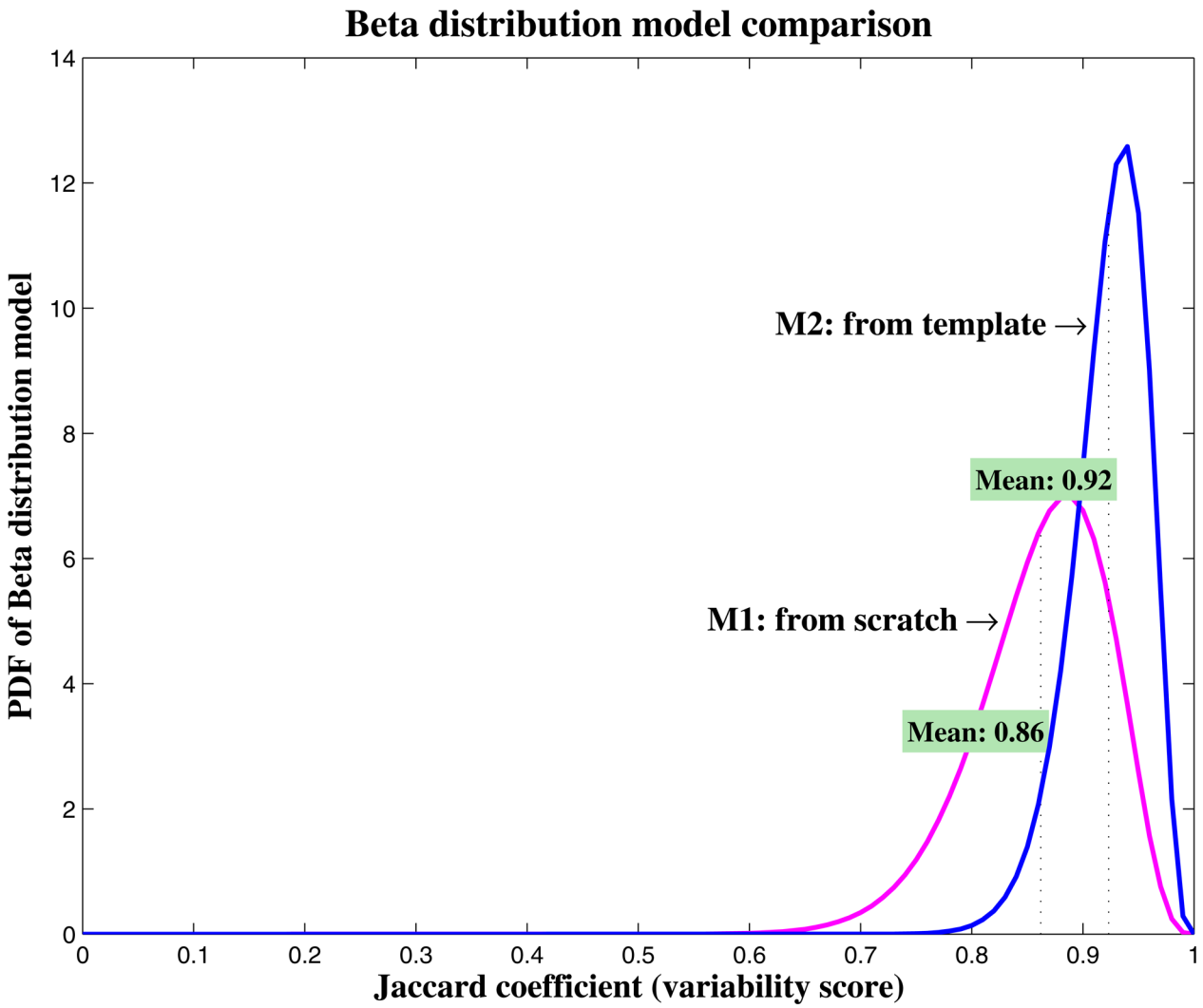
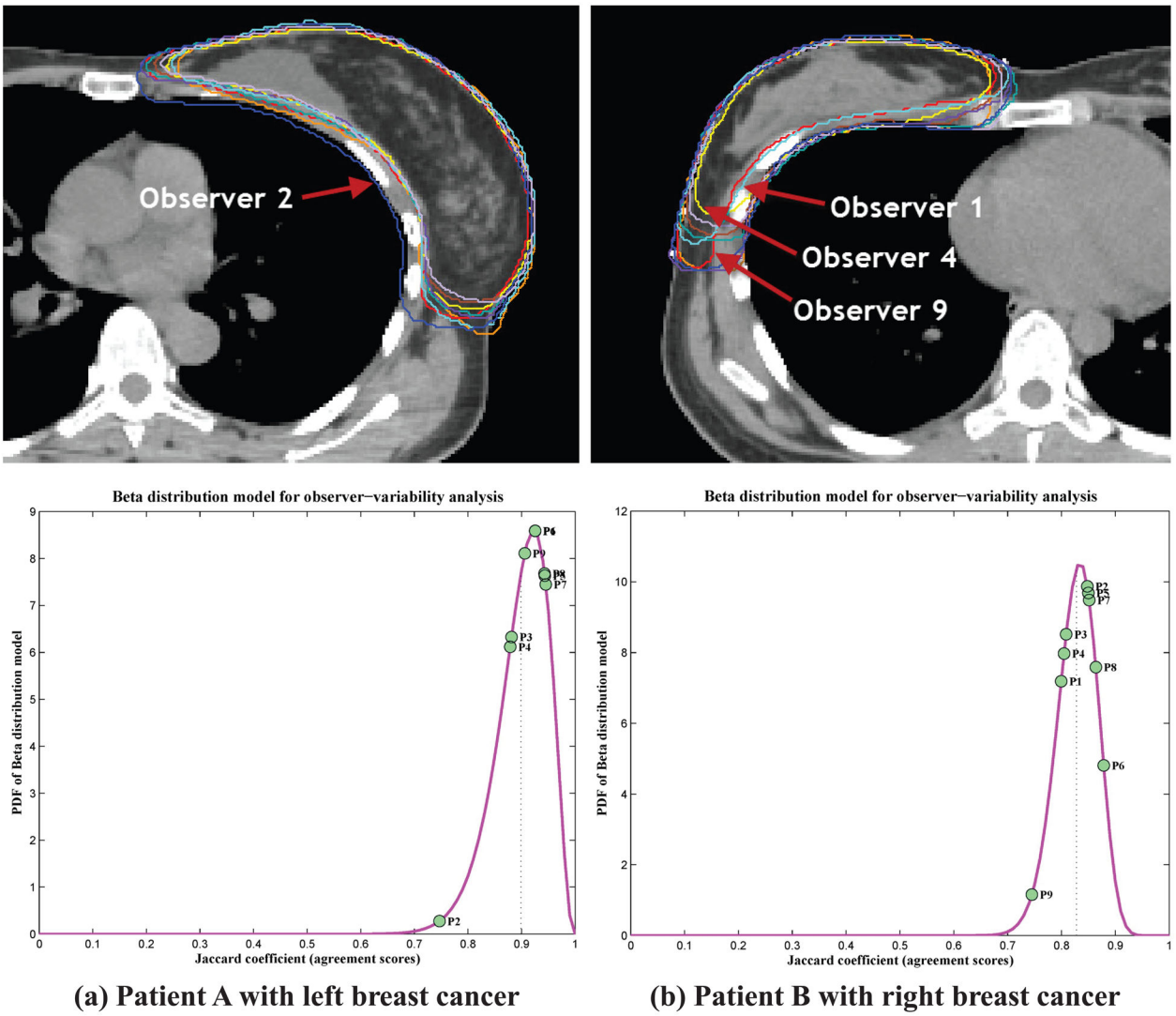


Figure 3.

Comparison of beta distribution models (as probability density function [PDF]) for contouring from scratch and contouring from the template. The higher, narrower, and sharper the curve, the smaller the inter-observer variability. Contouring from the template showed a smaller inter-observer variability than contouring from scratch.



(a) Patient A with left breast cancer **(b) Patient B with right breast cancer**

Figure 4.

Top row: Variability in contouring the breast volumes. Observer 2 contoured much larger than others on Patient A with left breast cancer, while Observers 1, 4, and 9 contoured much smaller than others on Patient B with right breast cancer. Bottom row: Visualization of beta distribution models (as probability density function [PDF]) with individual agreement score on the curve.

Table 1

The agreement scores between individual contours and group-consensus contours for eight observers in both contouring from scratch and contouring from the template. The boldface numbers indicate the observers have lower sensitivity or specificity than others.

Observer	Contouring from scratch			Contouring from template		
	Sensitivity	Specificity	Jaccard	Sensitivity	Specificity	Jaccard
1	96.6%	95.8%	92.1%	97.2%	99.2%	96.7%
2	96.8%	94.7%	91.2%	98.5%	97.4%	96.1%
3	98.3%	92.0%	89.8%	99.0%	95.3%	94.2%
4	**86.9%**	92.0%	78.5%	**90.7%**	96.6%	87.6%
5	95.4%	**83.6%**	79.5%	95.9%	**94.0%**	89.6%
6	**81.8%**	95.6%	76.8%	97.4%	96.7%	94.2%
7	95.2%	96.5%	91.8%	93.7%	95.2%	88.6%
8	93.3%	97.4%	90.2%	94.2%	97.5%	91.5%

Table 2

The agreement scores between individual contours and group-consensus contours for nine observers in contouring the left breast for Patient A and the right breast for Patient B.

Observer	Patient A with left breast cancer			Patient B with right breast cancer		
	Sensitivity	Specificity	Jaccard	Sensitivity	Specificity	Jaccard
1	97.6%	95.9%	92.6%	80.7%	98.8%	80.0%
2	97.8%	76.4%	74.7%	93.6%	90.1%	84.8%
3	95.5%	93.5%	88.2%	97.4%	80.9%	80.9%
4	91.8%	96.7%	87.9%	81.9%	98.0%	80.5%
5	99.0%	96.3%	94.4%	96.0%	87.7%	85.0%
6	99.0%	94.7%	92.5%	97.3%	90.1%	87.8%
7	96.0%	98.9%	94.5%	97.6%	86.2%	85.2%
8	95.3%	99.1%	94.3%	88.9%	97.2%	86.4%
9	92.0%	98.8%	90.6%	75.7%	98.0%	74.5%