The Breast 73 (2024) 103599

Contents lists available at ScienceDirect

The Breast

journal homepage: www.journals.elsevier.com/the-breast

Assessment of deep learning-based auto-contouring on interobserver consistency in target volume and organs-at-risk delineation for breast cancer: Implications for RTQA program in a multi-institutional study

Min Seo Choi [a,1], Jee Suk Chang [a,1], Kyubo Kim [b,v], Jin Hee Kim [c], Tae Hyung Kim [d], Sungmin Kim [e], Hyejung Cha [f], Oyeon Cho [g], Jin Hwa Choi [h], Myungsoo Kim [i], Juree Kim [j], Tae Gyu Kim [k], Seung-Gu Yeo [l], Ah Ram Chang [m], Sung-Ja Ahn [n], Jinhyun Choi [o], Ki Mun Kang [p], Jeanny Kwon [q], Taeryool Koo [r], Mi Young Kim [s], Seo Hee Choi [t], Bae Kwon Jeong [u], Bum-Sup Jang [ae], In Young Jo [w], Hyebin Lee [x], Nalee Kim [y], Hae Jin Park [z], Jung Ho Im [aa], Sea-Won Lee [ab], Yeona Cho [ac], Sun Young Lee [ad], Ji Hyun Chang [ae], Jaehee Chun [a], Eung Man Lee [b], Jin Sung Kim [a,*], Kyung Hwan Shin [ae,**], Yong Bae Kim [a]

[a] Department of Radiation Oncology, Yonsei University College of Medicine, Seoul, Republic of Korea
[b] Department of Radiation Oncology, Ewha Womans University College of Medicine, Seoul, Republic of Korea
[c] Department of Radiation Oncology, Dongsan Medical Center, Keimyung University School of Medicine, Daegu, Republic of Korea
[d] Department of Radiation Oncology, Nowon Eulji Medical Center, Eulji University School of Medicine, Seoul, Republic of Korea
[e] Department of Radiation Oncology, Dong-A University Hospital, Dong-A University College of Medicine, Busan, Republic of Korea
[f] Department of Radiation Oncology, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea
[g] Department of Radiation Oncology, Ajou University School of Medicine, Suwon, Republic of Korea
[h] Department of Radiation Oncology, Chung-Ang University Hospital, Seoul, Republic of Korea
[i] Department of Radiation Oncology, Incheon St Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea
[j] Department of Radiation Oncology, Ilsan CHA Medical Center, CHA University School of Medicine, Goyang, Republic of Korea
[k] Department of Radiation Oncology, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine, Changwon, Republic of Korea
[l] Department of Radiation Oncology, Soonchunhyang University College of Medicine, Soonchunhyang University Hospital, Bucheon, Republic of Korea
[m] Department of Radiation Oncology, Soonchunhyang University College of Medicine, Seoul, Republic of Korea
[n] Department of Radiation Oncology, Chonnam National University Medical School, Gwangju, Republic of Korea
[o] Department of Radiation Oncology, Jeju National University Hospital, Jeju University College of Medicine, Republic of Korea
[p] Gyeongsang National University Changwon Hospital, Gyeongsang National University College of Medicine, Jinju, Republic of Korea
[q] Department of Radiation Oncology, Chungnam National University School of Medicine, Daejeon, Republic of Korea
[r] Department of Radiation Oncology, Hallym University Sacred Heart Hospital, Hallym University College of Medicine, Anyang, Republic of Korea
[s] Department of Radiation Oncology, Kyungpook National University Chilgok Hospital, Daegu, Republic of Korea
[t] Department of Radiation Oncology, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Republic of Korea
[u] Department of Radiation Oncology, Gyeongsang National University Hospital, Gyeongsang National University College of Medicine, Jinju, Republic of Korea
[v] Department of Radiation Oncology, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam, Republic of Korea
[w] Department of Radiation Oncology, Soonchunhyang University Hospital, Cheonan, Republic of Korea
[x] Department of Radiation Oncology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[y] Department of Radiation Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[z] Department of Radiation Oncology, Hanyang University College of Medicine, Seoul, Republic of Korea
[aa] Department of Radiation Oncology, CHA Bundang Medical Center, CHA University School of Medicine, Seongnam, Republic of Korea
[ab] Department of Radiation Oncology, Eunpyeong St. Mary's Hospital, Catholic University of Korea College of Medicine, Seoul, Republic of Korea
[ac] Department of Radiation Oncology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea
[ad] Department of Radiation Oncology, Chonbuk National University Hospital, Jeonju, Republic of Korea
[ae] Department of Radiation Oncology, Seoul National University College of Medicine, Seoul, Republic of Korea

* Corresponding author. Yonsei University College of Medicine, 50Yonsei-ro, Shinchon-dong, Seodaemun-gu, Seoul, Republic of Korea.
** Corresponding author. Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul, Republic of Korea.
E-mail addresses: jinsung@yuhs.ac (J.S. Kim), radiat@snu.ac.kr (K.H. Shin).
[1] Min Seo Choi and Jee Suk Chang equally contributed to this work as first author.

https://doi.org/10.1016/j.breast.2023.103599
Received 18 September 2023; Received in revised form 30 October 2023; Accepted 9 November 2023
Available online 15 November 2023

ARTICLE INFO

Keywords:
RTQA
Inter-observer variation
Auto-contouring
Breast cancer
Deep learning

ABSTRACT

Purpose: To quantify interobserver variation (IOV) in target volume and organs-at-risk (OAR) contouring across 31 institutions in breast cancer cases and to explore the clinical utility of deep learning (DL)-based auto-contouring in reducing potential IOV.

Methods and materials: In phase 1, two breast cancer cases were randomly selected and distributed to multiple institutions for contouring six clinical target volumes (CTVs) and eight OAR. In Phase 2, auto-contour sets were generated using a previously published DL Breast segmentation model and were made available for all participants. The difference in IOV of submitted contours in phases 1 and 2 was investigated quantitatively using the Dice similarity coefficient (DSC) and Hausdorff distance (HD). The qualitative analysis involved using contour heat maps to visualize the extent and location of these variations and the required modification.

Results: Over 800 pairwise comparisons were analysed for each structure in each case. Quantitative phase 2 metrics showed significant improvement in the mean DSC (from 0.69 to 0.77) and HD (from 34.9 to 17.9 mm). Quantitative analysis showed increased interobserver agreement in phase 2, specifically for CTV structures (5–19 %), leading to fewer manual adjustments. Underlying IOV differences causes were reported using a questionnaire and hierarchical clustering analysis based on the volume of CTVs.

Conclusion: DL-based auto-contours improved the contour agreement for OARs and CTVs significantly, both qualitatively and quantitatively, suggesting its potential role in minimizing radiation therapy protocol deviation.

1. Introduction

Radiation therapy (RT) protocol deviations and non-compliance can be associated with a higher risk of treatment failure and mortality in clinical trials [1,2]. Among them, interobserver variation (IOV) in RT contouring is one of the biggest concerns [3,4]. As modern RT techniques allow for more accurate beam delivery with high conformality and precision, contouring has become increasingly important. However, data show that even radiation oncologist experts have difficulty contouring [5]. Furthermore, multiple studies have demonstrated significant IOVs in delineating target volumes, including clinical target volumes (CTVs) and organs-at-risk (OARs), in various types of cancers, both within and outside clinical trials [6–10]. The magnitude of IOV is often larger than that of variations related to organ movement and set-up uncertainties [11].

In recent decades, emphasis has been placed on improving IOV in contouring. These strategies include the widespread publication of site-specific atlases and consensus guidelines [12], often trial-specific guidance documents, education, audits, and peer-review as RT quality assurance (RTQA) tools. Benchmark studies, often referred to as dummy runs, are routinely conducted at the outset of clinical trials or individual case reviews (ICR) [13,14]. These studies, whether conducted prospectively or retrospectively, serve as established and essential components of the RTQA process in clinical trials. While previous methods have demonstrated their ability to enhance Interobserver Variability (IOV), they each come with their own limitations. This underscores the necessity of exploring alternative approaches to address IOV in contouring [15].

Deep learning (DL) is based on artificial neural networks and is being progressively introduced to aid radiation oncology. DL-based auto-contouring has been actively studied inhead and neck, prostate, and breast cancers, demonstrating significant benefits in terms of time-saving and improved contouring IOV [16]. Additionally, auto-contour is more interactive than static guidelines and atlases. This enhanced interactivity simplifies the process of tailoring anatomical contours to individual patients, which is particularly valuable in cases involving varying body shapes and treatment positioning [6]. To the best of our knowledge, no studies have investigated the clinical utility of DL-based auto-contouring within RTQA programs. In this context, the Korean Radiation Oncology Group (KROG) conducted a study on IOV in contouring the CTV and OARs in breast RT after breast conservation surgery or mastectomy with immediate breast reconstruction (KROG 21–01). Members of the KROG group delineated contours on simulated computed tomography (CT) images. Here we report the results of contouring IOV and assess the impact of DL-based auto-contouring on reducing IOV.

2. Materials and methods

2.1. Study

The contouring study consisted of two phases: 1) Phase 1 (Baseline), which aimed to investigate manual contouring variations, and 2) Phase 2 (AI Intervention), which aimed to investigate whether there were any changes in contouring variations when auto-contouring was provided (Supplementary Fig. 1). Two left-sided breast cancer cases were randomly chosen by an independent institution to avoid bias, and then sent to the members of the KROG for a preclinical dry-run contouring study. Case 1 was a T1cN1M0 (tumour size 1.5 cm, three positive axillary lymph nodes, triple negative subtype, and histologic grade 3) who had previously undergone breast conservation surgery. Case 2 was T3N1M0 (tumour size 9.5 cm, one positive axillary lymph node, luminal A type, and histologic grade 2), who underwent mastectomy with immediate subpectoral implant-based breast reconstruction. These two cases were used to evaluate contouring IOV at multiple institutions. For each case, several key images from the same CT and magnetic resonance imaging (MRI) scans were distributed to the participating investigators, along with other clinical information for radiation therapy treatment. This study was approved by the institutional review board of Seoul National University Hospital (H-2102-029-1193) and ethical review board of Korean Radiation Oncology Group (KROG 21–01).

In phase 1, each participant was instructed to contour the target volumes and OARs. The European Society for Radiation and Oncology consensus guideline [12,17] was suggested to aid the contouring of CTVs (CTV axillary levels 1, 2, and 3 [CTVn_L1, 2, 3], intramammary node [CTVn_IMN], supraclavicular node [CTVn_SCL or CTVn_L4], and CTV [CTVp_breast]); however, clinical discretion was allowed based on their experience and knowledge. The planning target volume (PTV) was generated using a non-isotropic geometrical expansion based on the participants' institutional policy. OARs included the heart, contralateral breast (CLB), thyroid, esophagus, spinal cord, left and right lungs (Lung R, L), and left anterior descending artery (LAD). Thirty-one institutions participated in the first phase of this study. From each participating institution, one board-certified radiation oncologist who specializes in breast cancer radiation therapy and actively performs target volume delineation for treatment planning was selected. The median years of experience of the observers is 10.5 years, with a standard deviation of 7.4 years.

Auto-contour sets containing target CTVs and OARs were generated on the test cases (i.e., Cases 1 and 2) using a previously published in-house DL model that has been used in clinics since 2020 [18]. The DL model used in this study had previously been tested on both internal and external cohorts of breast cancer patients, demonstrating robust

performance in left-sided, right-sided, and bilateral breast cancer. The model was chosen because of the limited availability of contour tools that encompass all the CTVs used in this study. Six months after phase 1, the same participants took part in phase 2. In Phase 2, participants were instructed to use auto-contour sets, but were given flexibility to deviate from them if they disagreed with their quality or found them uncomfortable to use.

2.2. IOV analysis

The quantitative metrics for performance evaluation included the Dice similarity coefficient (DSC), 95 % Hausdorff distance (HD), and added path length (APL). DSC is the most widely used metric in medical image segmentation, which defines the overlap between two volumes of interest [19]. Surface DSC is a variation of the DSC that quantifies the deviation between OAR surface contours [20]. HD is a measure of the surface distance between two point sets, A and B, defined as Equation (1).

$$H(A,B) = H(A,B) = max\{h(A,B), h(B,A)\} \quad (1)$$

HD95 denotes the maximum surface-to-surface separation among the 95th percentile of the ground-truth and segmentation surface points. Lastly, APL is a more recently introduced metric that represents all manual adjustments made in terms of the number of pixels added or removed and was previously reported to be more closely related to actual editing time [21].

For qualitative evaluation, a questionnaire containing the following questions was sent to all observers who participated in this study:

- Question 1: "How much time did it take to complete the contours for each phase?" Answers were "<30 min," "30–60 min," and >1 h"
- Question 2: "How would you rate the auto-contour quality?" where the answers were given on a 5-point scale ranging from 1 (not useable) to 5 (no edits needed).
- Question 3: 'Do you think auto-contouring will help reduce IOV in the future?' on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree).
- Question 4: "How much auto-contour did you utilize in Phase 2?" on a 5-point scale from 1 (not at all) to 5 (very much).

We included both two-dimensional (2D) and three-dimensional (3D) heat maps to visualize interobserver agreement and areas of manual edits with respect to the edited auto-contour (reference contour). A radiation oncologist with 9 years of experience (author J.S·C) edited the auto-contour sets, and an independent panel of three radiation oncologists (authors K·K, K·H·S, and Y·B·K) finalized it as a reference. The 2D heatmap shows variations among observers, with values ranging from 0 to 31 (Supplementary Fig. 3). The areas with the greatest and least overlap are indicated in red and blue, respectively. A three-dimensional heatmap was created to show the average adjustment of the participants projected on the reference shape of each OAR. The nearest point on the participant's 3D surface was determined using reference 3D surface. Subsequently, we determined whether the point was outside or inside the closed reference surface. Depending on the degree of adjustment, each point was represented by a colour map ranging from red (i.e., maximum outward expansion of 10 mm) to blue (i.e., maximum inward shrinkage of 10 mm).

Hierarchical clustering was conducted in R 1.0.12 using the pheatmap package, only including 26 observers who contoured all structures of interest. The reference volume was the ground-truth reference contour. Euclidean distance and complete linkage were used as the distance metric and linkage algorithm, respectively. Values were standardized in the column direction.

3. Results

In total, 31 and 30 institutions participated in phases 1 and 2, respectively. To assess IOV, participants contoured 15 structures in two cases, resulting in 465 and 435 paired comparisons per case, and 930 and 870 paired comparisons per structure. Significantly improved IOV and stronger alignment with the consensus contour were observed in Phase 2 (Table 1). Surface DSC results were lower than DSC in CTVs, although higher values were observed for smaller structures like the thyroid and LAD. HD also decreased across all structures. For example, the DSC for the LAD increased from 0.44 to 0.61. CTVs had a greater IOV (i.e., less DSC) than OARs, except for the LAD. Additionally, the phase 1 contours were evaluated against an unedited DL model (Supplementary Table 1C). The DL model showed similar resemblance to the consensus contour, but lower similarity in structures like the LAD (DSC of 0.19 vs. consensus of 0.50) and the spinal cord (DSC of 0.66 vs. consensus of 0.76).

Phase 2 had larger areas of strong interobserver agreement than Phase 1, as shown by the smaller blue regions in the CTVs (Supplementary Fig. 3). For case 1, the percentage of CTV breast with high agreement indicated by red was 16.2 % in Phase 2 versus 8.7 % in Phase 1. For case 2, the difference was even higher with 25.0 % in Phase 2 versus 9.9 % in Phase 1. This trend was consistent in other CTVs such as CTVn_L2 and CTVn_SCL, as shown in Supplementary Figs. 4 and 5. Interestingly, there was a mixed response in the area of the reconstructed breast in case 2. In Phase 1, a larger number of observers (approximately 20) opted to encompass the central portion of the breast implant, whereas in Phase 2, a smaller contingent (approximately 10) made this choice. Additionally, there were notable enhancements in OARs from Phase 1 to Phase 2, although to a lesser extent (Supplementary Fig. 6).

Three-dimensional heatmaps were used to confirm the areas of each structure where most edits were made. It was discovered that observers were less likely to edit in phase 2, as implied by fewer red and blue regions that indicate outward or inward contour modification, respectively (Fig. 1). The severity of contour adjustment was generally much lower in Phase 2. Phase 2 contours had a higher proportion of green regions (i.e., areas similar to the reference), as best shown by the CTVn_L1 and CTVn_IMN contours. In phase 1, physicians were more likely to draw contours towards the skin for CTVn_L1, whereas, in phase 2, this was significantly reduced. Furthermore, based on the heatmaps, the starting points for CTVn_L2 and CTVp_breast varied more in phase 1 than in phase 2.

For phase 1, the percentage of time taken to complete contours was greatest at 30–60 min time intervals (Fig. 2A). Conversely, the greatest percentage of time taken was in the less than 30 min interval in phase 2, with more than half of the observers responding to this category. As shown in Fig. 2B, for the target structures, the percentage of minor edits was the greatest, followed by major edits, mostly acceptable and not useable, at 53.2 %, 36.2 %, 6.4 %, and 4.3 %, respectively. For OARs, observers were generally satisfied with the auto contour, with contour quality requiring only minor edits or less, although no one answered that they were perfect. The extent to which users used the AI-produced baseline was positively correlated with both the R2 contour assessment score (Question 2) ($r = 0.88$) and future score (Question 3) ($r = 0.82$). The degree of utilization was positively correlated with both the contour evaluation score (Question 2) and future score (Question 3), with R^2 values of 0.88 and 0.82, respectively (Fig. 2C).

Using a hierarchical clustering heatmap, we discovered that radiation oncologists could be classified based on their contouring style, whether they drew smaller or larger than the auto-contours (Fig. 3). Observers who failed to include the required structures fully were excluded from this part of the analysis, leaving 26 observers. The outermost clustering indicates that observers are divided into two major groups. Out of 26 observers, approximately 20 % tended to contour larger than the reference. The rest did not have distinct contouring

Table 1
Quantitative evaluation through interobserver comparison (a) and with reference to consensus contour (b). Abbreviations: DSC = dice similarity coefficient; SD = standard deviation; HD = Hausdorff distance; CTVn_L1 = CTV axillary level 1; CTVn_L2 = CTV axillary level 2; CTVn_L3 = CTV axillary level 3; CTVn_IMN = intramammary node; CTVn_SCL = supraclavicular node; CTVp_breast = clinical target volume; CLB = contralateral breast; Lung R = right lung; Lung L = left lung; LAD = left anterior descending artery.

	(a) Interobserver Comparison						(b) Comparison to consensus contour					
	DSC (±SD)		Surface DSC (±SD)		HD (±SD)		DSC (±SD)		Surface DSC (±SD)		HD (±SD)	
	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2
CTVn_L1	0.58 ± 0.13	0.67 ± 0.22	0.43 ± 0.14	0.55 ± 0.25	42.96 ± 29.29	31.65 ± 33.63	0.64 ± 0.12	0.73 ± 0.17	0.44 ± 0.14	0.58 ± 0.18	27.93 ± 18.27	19.89 ± 23.64
CTVn_L2	0.51 ± 0.17	0.70 ± 0.25	0.44 ± 0.16	0.64 ± 0.25	46.22 ± 35.59	24.83 ± 24.20	0.52 ± 0.20	0.65 ± 0.21	0.46 ± 0.19	0.60 ± 0.20	56.12 ± 32.07	36.45 ± 27.58
CTVn_L3	0.47 ± 0.14	0.55 ± 0.23	0.45 ± 0.14	0.54 ± 0.23	41.47 ± 38.70	28.72 ± 28.12	0.49 ± 0.18	0.61 ± 0.20	0.47 ± 0.18	0.60 ± 0.21	35.56 ± 30.95	21.31 ± 23.80
CTVn_IMN	0.52 ± 0.13	0.61 ± 0.16	0.65 ± 0.16	0.72 ± 0.16	35.46 ± 29.96	18.78 ± 19.67	0.49 ± 0.15	0.64 ± 0.15	0.59 ± 0.19	0.74 ± 0.19	31.74 ± 27.27	13.84 ± 15.66
CTVn_SCL	0.51 ± 0.14	0.62 ± 0.20	0.40 ± 0.14	0.52 ± 0.21	48.26 ± 32.26	38.47 ± 34.94	0.30 ± 0.14	0.32 ± 0.13	0.29 ± 0.11	0.34 ± 0.10	96.30 ± 38.43	105.48 ± 42.84
CTVp_breast	0.75 ± 0.12	0.80 ± 0.13	0.59 ± 0.16	0.72 ± 0.20	22.52 ± 16.91	16.01 ± 20.22	0.73 ± 0.13	0.81 ± 0.14	0.64 ± 0.16	0.79 ± 0.18	18.02 ± 11.95	11.99 ± 17.85
Heart	0.90 ± 0.05	0.95 ± 0.03	0.68 ± 0.14	0.82 ± 0.12	16.36 ± 12.21	8.38 ± 6.06	0.92 ± 0.05	0.95 ± 0.02	0.73 ± 0.16	0.84 ± 0.11	12.36 ± 8.84	7.21 ± 4.54
Contralateral breast	0.81 ± 0.06	0.89 ± 0.10	0.61 ± 0.17	0.79 ± 0.21	21.70 ± 15.67	15.22 ± 25.78	0.84 ± 0.06	0.92 ± 0.08	0.67 ± 0.19	0.87 ± 0.18	15.57 ± 11.16	9.17 ± 20.11
Thyroid	0.75 ± 0.12	0.79 ± 0.12	0.86 ± 0.11	0.89 ± 0.11	9.72 ± 15.83	7.03 ± 15.28	0.79 ± 0.10	0.82 ± 0.08	0.90 ± 0.09	0.92 ± 0.07	8.10 ± 12.29	5.47 ± 10.98
Esophagus	0.77 ± 0.06	0.81 ± 0.07	0.89 ± 0.07	0.91 ± 0.06	31.62 ± 59.29	8.52 ± 17.51	0.81 ± 0.05	0.83 ± 0.04	0.92 ± 0.05	0.94 ± 0.04	15.70 ± 39.07	5.18 ± 10.86
Spinal cord	0.68 ± 0.12	0.79 ± 0.14	0.80 ± 0.13	0.89 ± 0.11	112.88 ± 111.66	33.93 ± 64.35	0.76 ± 0.09	0.82 ± 0.07	0.87 ± 0.09	0.94 ± 0.10	65.09 ± 86.21	18.27 ± 44.42
Lung R	0.97 ± 0.01	0.98 ± 0.01	0.95 ± 0.03	0.97 ± 0.03	3.86 ± 2.50	2.56 ± 1.80	0.97 ± 0.01	0.98 ± 0.01	0.96 ± 0.02	0.97 ± 0.02	2.76 ± 1.18	1.81 ± 0.86
Lung L	0.87 ± 0.01	0.98 ± 0.02	0.95 ± 0.03	0.97 ± 0.03	3.63 ± 1.98	2.35 ± 1.50	0.97 ± 0.01	0.98 ± 0.01	0.96 ± 0.02	0.97 ± 0.01	2.86 ± 0.84	2.15 ± 0.80
LAD	0.44 ± 0.21	0.61 ± 0.18	0.71 ± 0.21	0.80 ± 0.15	52.14 ± 59.74	14.08 ± 15.30	0.50 ± 0.16	0.59 ± 0.01	0.77 ± 0.14	0.81 ± 0.10	39.03 ± 51.87	12.96 ± 13.48

patterns, although further grouping was performed.

4. Discussion

This study first assessed the magnitude of IOV in CTV and OAR contouring in left-sided breast RT with regional nodal irradiation. Similar to a previous study (KROG 19–01) [22], there was substantial IOV among observers in contouring some OARs and the majority of CTVs in Phase 1. The widespread use of intensity-modulated radiotherapy in breast RT, as supported by recent phase 3 randomized trials [23–25], suggests that contouring deviations may have a more significant impact on RT-related toxicity and local control than previously thought. We investigated the role of DL-based auto-outlines as novel interventions which can be incorporated into RTQA programs in multicentre clinical trials. Our study demonstrated a significant reduction in IOV in OARs and various CTVs through interventions. Our findings included a local assessment of contour editing and a classification of the contouring preferences and styles of the observers, which we consider a key feature that distinguishes the current study from others. More specifically, the participants' contours were successfully aligned to auto-contour reference shapes in 2D and 3D representations, allowing us to pinpoint the local anatomical regions of common adjustments among the participants. Furthermore, cluster analysis based on the volume of various CTVs could classify the participants based on the observer contouring style. Such information can be useful in providing feedback on unacceptable protocol deviations and may allow local investigators to improve the contouring of enrolled patients.

Multiple studies have evaluated the effectiveness of interventions, including guidelines and atlases [26] and teaching [27], and these interventions were associated with improved IOV in contouring. This study found that using DL-based auto-contours increased the average DSC from 0.69 to 0.77. These figures are slightly higher, yet consistent with previous studies [27,28]. An Italian study reported an overall median DSC of 0.66 in nodal delineation in the presence of guidelines [29]. Another aspect that distinguishes our study from others was the visual demonstration of improvement in contouring IOV in 2D and 3D heatmaps, identifying specific regions of IOV improvements. Compared with guidelines and atlases, our findings indicated that providing auto-contours to edit saves time and can be utilized to measure one's performance against a benchmark reference. Although there are well-known benefits to using DL-based auto-segmentation, there are also some criticisms that should be considered, such as limited generalizability, the need for manual correction, and a lack of transparency. While external validation and model updates are crucial in the aspects of algorithm development and model management, human bias, manifested as diverse individual expectations, is one of the most significant factors reducing clinical acceptance on the user's end. Unsurprisingly, the extent of the disagreement with provided auto-contours was inversely correlated with the degree of auto-contour use (Fig. 2). Ciardo et al. [29] found that the degree of expertise significantly impacted the volume size of nodal targets and IOV, with junior oncologists having lower IOV than seniors or experts. This implies that similar to guideline intervention, human factors such as the observers' own experience and knowledge accounted for residual contouring IOV after interventions. As reported by McIntosh et al. [30], these human factors were also the primary cause for the decreased likelihood of selecting a DL-based RT plan during the prospective deployment phase, highlighting the importance of addressing them.

Vaassen et al. [31] recently analysed user adjustments after DL auto-contouring in the thorax region of nearly 700 cases, including breast cancer, and found large adjustments in some specific regions for most OARs. In our study, 2D and 3D heatmaps were used to visualize the areas of IOV and the degree to which each user adjusted the structures to suit their preferences in six directions relative to reference contours. We found that most auto-contours were under-segmented to some extent and most CTVs had an asymmetric distribution of IOV regions. In

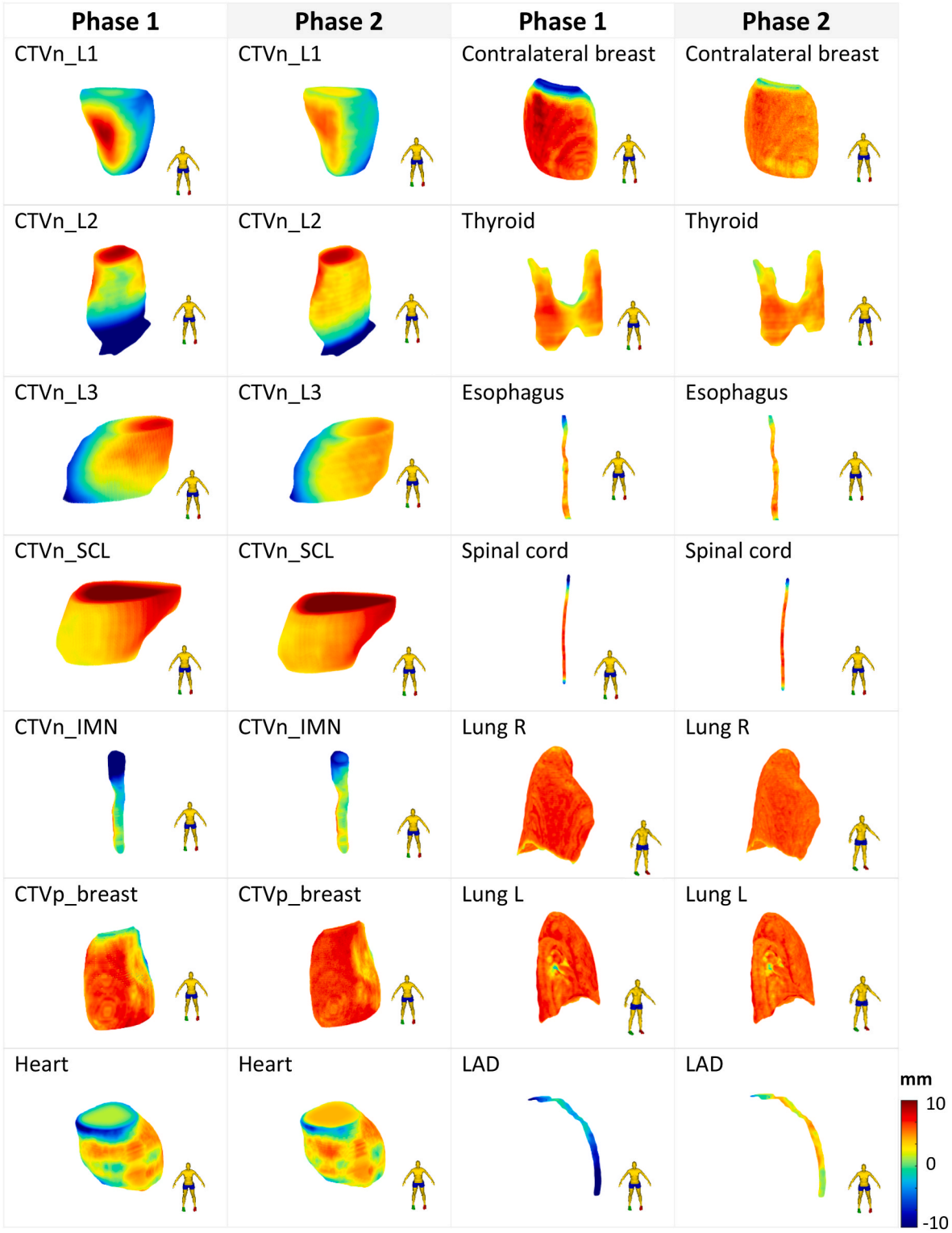

Fig. 1. A three-dimensional projection of the observers' average adjustments for Phases 1 and 2 onto the reference shape of each structure. Higher adjustments indicate outward adjustments on a scale of −10 mm–10 mm. Abbreviations: CTVn_L1 = CTV axillary level 1; CTVn_L2 = CTV axillary level 2; CTVn_L3 = CTV axillary level 3; CTVn_IMN = intramammary node; CTVn_SCL = supraclavicular node; CTVp_breast = clinical target volume; CLB = contralateral breast; Lung R = right lung; Lung L = left lung; LAD = left anterior descending artery.

Supplementary Table 2, a comparison of the findings with existing literature demonstrates that disagreements in most CTVs were most noticeable, which implies the limited generalizability of the DL-based auto-segmentation model, particularly in target volumes.

In the RTQA process, various means of feedback can be provided to the recruiting centre for any unacceptable deviation. A review of deviations in target volume delineation indicated that concerns about bias in the feedback process persisted, and processes for feedback on the quality of contouring deviations were not standardized in previous trials [32]. In this context, identifying specific physician styles in CTV

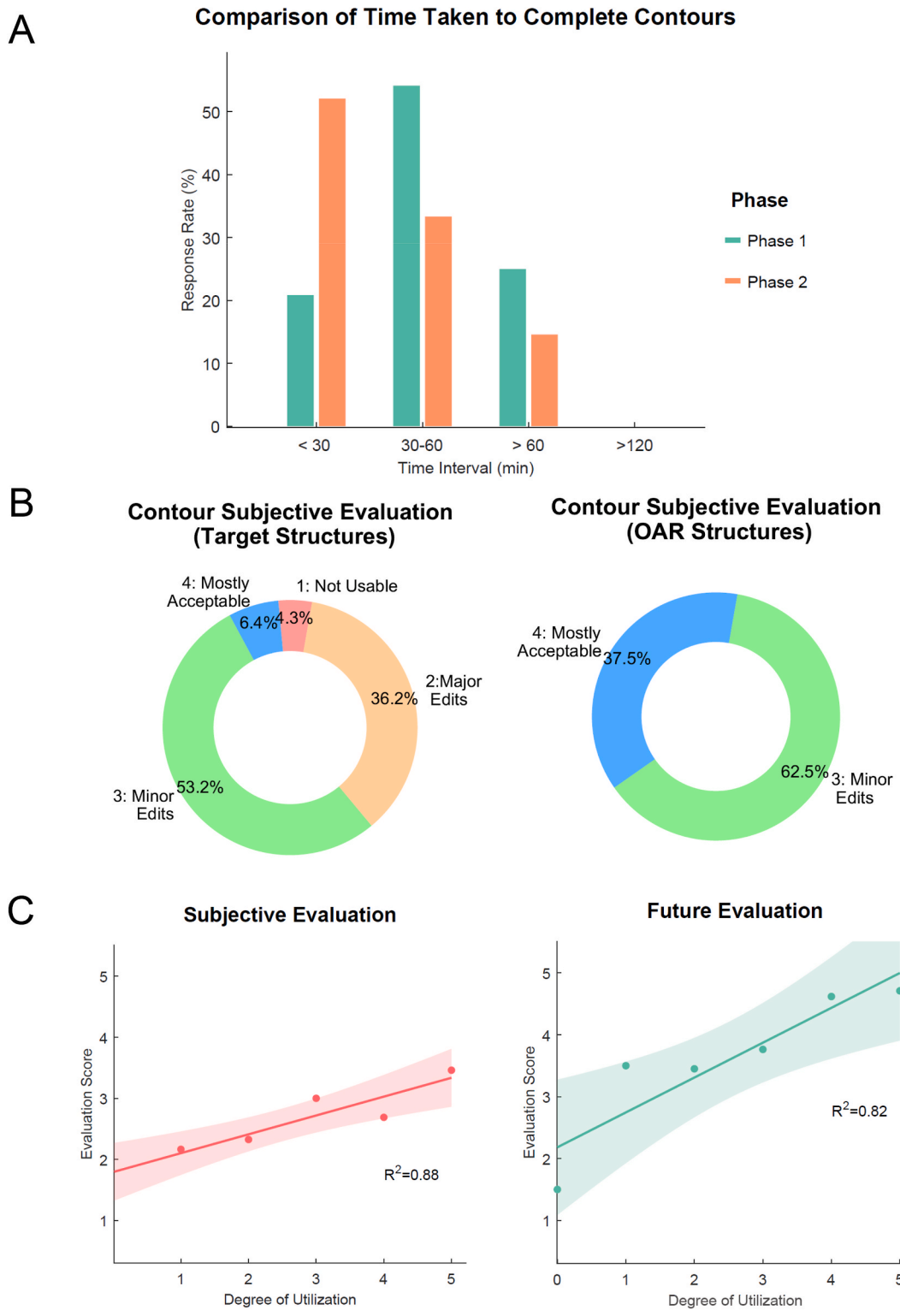

Fig. 2. (A) Time comparison, (B) Subjective evaluation comparison, (C) Left: Relationship between the degree of auto-contour utilization and the evaluation score for Cases 1 and 2, Right: Relationship between the degree of auto-contour utilization and the future perspective evaluation score.

contouring based on volume size (Fig. 3) can produce unbiased and individualized forms of feedback; however, it depends on the nature of the referenced contours. Ownership of the reference contours, and knowledge of the input to the DL algorithm, can increase the likelihood of the provided feedback being accepted. The classification in this study was performed using only one criterion (the size of the contouring). For example, our findings regarding the classification of physicians' styles and identification of groups with a propensity to contour larger areas can be utilized later to either provide individual instructions for using auto-contour in a broader manner or to personally recommend contouring in a more restrained manner. However, future research will need to evaluate the factors that affect a physician's contouring approach and determine which factors lead to a more individualized model before implementing the benchmark credentialing exercise. Additionally, we

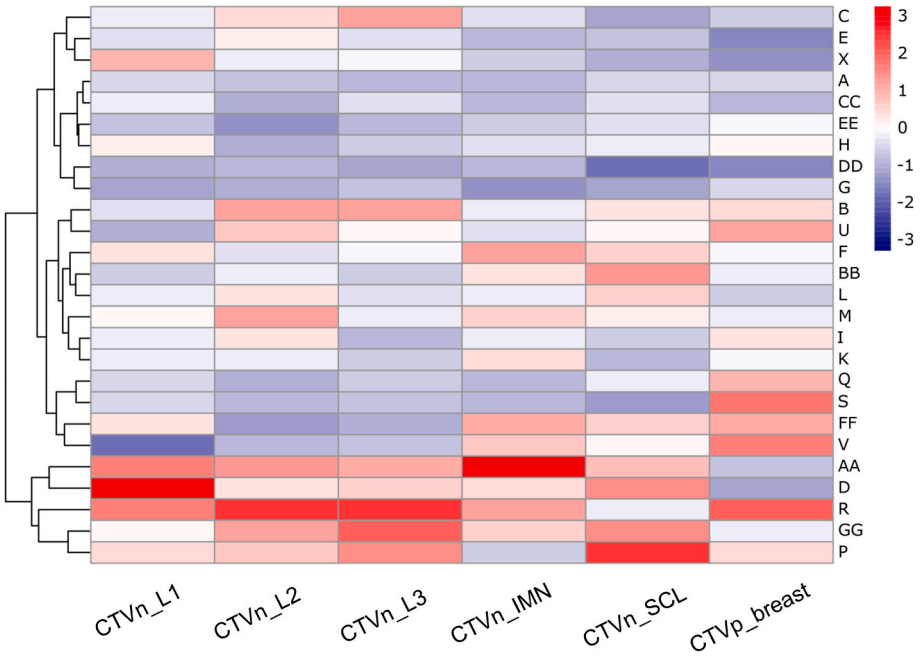

Fig. 3. Cluster heatmap of observers based on contour volume. Positive values indicate that expert drawn contour has a higher volume than the ground-truth reference contour. Abbreviations: CTVn_L1 = CTV axillary level 1; CTVn_L2 = CTV axillary level 2; CTVn_L3 = CTV axillary level 3; CTVn_IMN = intramammary node; CTVn_SCL = supraclavicular node; CTVp_breast = clinical target volume.

believe that identifying the extents and areas of disagreement, along with user education, can increase the likelihood of capitalizing on auto-contours.

This study has some limitations that should be acknowledged. First, our scope was limited to Korean institutions in terms of both participating observers and the dataset used, where relatively smaller breasts and body sizes are common. Also, evaluating the IOV in only two cases would be limiting. The degree and direction of IOV can vary among multinational observers and patient cohorts. Nonetheless, this study serves as the foundation for future collaborative research. Secondly, this study primarily examined the contouring component of RTQA and did not explore other factors such as their impact on planning, delivery, and infrastructure. In the next phase, we plan to expand our research to include the planning and dosimetric aspects, which can offer insights into potential clinical significance. Another limitation of this pseudo-trial is its retrospective nature, as multicentre clinical trials may not have been completely captured in our study. Finally, due to its retrospective nature, our study did not assess other important metrics reflecting the quality of auto-contours, such as dosimetric impact, time savings, qualitative scoring of each case, and the clinical acceptance rate.

5. Conclusion

In summary, we evaluated the impact of DL-based auto-contouring on interobserver variability in a multicenter setting. Our findings indicate that the adoption of DL-based auto-contouring technology leads to a significant improvement in contour agreement, both qualitatively and quantitatively, for OARs and CTVs. Incorporating DL-based auto-contouring into RT trials and RTQA programs, and including it in educational materials for RTQA feedback, could be a novel and promising approach to IOV assessment that should be evaluated in future trials. A significant cause of the remaining disagreement in contouring appears to be the human element, including human knowledge and experience, which may result in the rejection of the offered DL-based auto-contours. To reduce the risk of human bias in future clinical studies that use automated tools, researchers should continuously work to mitigate human bias and consider how receptive users are to these tools.

Author responsible for statistical analysis

Min Seo Choi.
Email: minseochoi135@gmail.com.

Conflict of interest statement for all authors

JSK is co-founder of Oncosoft and serves as an advisor to Rayence. JSC holds stock in Oncosoft.

Funding statement

This study was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1C1C1009359).

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C2008623).

Data availability statement for this work

Research data are stored in an institutional repository and will be shared upon request to the corresponding author.

Acknowledgements

Presented at the 64th Annual Meeting of the American Society for Radiation Oncology, San Antonio, TX, October 24, 2022.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.breast.2023.103599.

References

[1] Peters LJ, O'Sullivan B, Giralt J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. J Clin Oncol Jun 20 2010;28(18):2996–3001. https://doi.org/10.1200/jco.2009.27.4498.

[2] Kearvell R, Haworth A, Ebert MA, et al. Quality improvements in prostate radiotherapy: outcomes and impact of comprehensive quality assurance during the TROG 03.04 'RADAR' trial. J Med Imaging Radiat Oncol Apr 2013;57(2):247–57. https://doi.org/10.1111/1754-9485.12025.

[3] van Mourik AM, Elkhuizen PH, Minkema D, Duppen JC, Dutch Young Boost Study G, van Vliet-Vroegindeweij C. Multiinstitutional study on target volume delineation variation in breast radiotherapy in the presence of guidelines. Radiother Oncol Mar 2010;94(3):286–91. https://doi.org/10.1016/j.radonc.2010.01.009.

[4] 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 Mar 1 2009;73(3):944–51. https://doi.org/10.1016/j.ijrobp.2008.10.034.

[5] Byun HK, Chang JS, Choi MS, et al. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. Radiat Oncol Oct 14 2021;16(1):203. https://doi.org/10.1186/s13014-021-01923-1.

[6] Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. Radiother Oncol Nov 2016;121(2):169–79. https://doi.org/10.1016/j.radonc.2016.09.009.

[7] Caravatta L, Macchia G, Mattiucci GC, et al. Inter-observer variability of clinical target volume delineation in radiotherapy treatment of pancreatic cancer: a multi-institutional contouring experience. Radiat Oncol 2014;9:1–9.

[8] Hong T, Tome W, Chappell R, Harari P. Variations in target delineation for head and neck IMRT: an international multi-institutional study. Int J Radiat Oncol Biol Phys 2004;60(1):S157–8.

[9] Jansen EP, Nijkamp J, Gubanski M, Lind PA, Verheij M. Interobserver variation of clinical target volume delineation in gastric cancer. Int J Radiat Oncol Biol Phys 2010;77(4):1166–70.

[10] Mukesh M, Benson R, Jena R, et al. Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help? Br J Radiol 2012;85(1016):e530–6.

[11] Segedin B, Petric P. Uncertainties in target volume delineation in radiotherapy - are they relevant and what can we do about them? Radiol Oncol Sep 1 2016;50(3):254–62. https://doi.org/10.1515/raon-2016-0023.

[12] Offersen BV, Boersma LJ, Kirkove C, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. Radiother Oncol Jan 2015;114(1):3–10. https://doi.org/10.1016/j.radonc.2014.11.030.

[13] Chung Y, Kim JW, Shin KH, et al. Dummy run of quality assurance program in a phase 3 randomized trial investigating the role of internal mammary lymph node irradiation in breast cancer patients: Korean Radiation Oncology Group 08-06 study. Int J Radiat Oncol Biol Phys. Feb 1 2015;91(2):419. https://doi.org/10.1016/j.ijrobp.2014.10.022. 26.

[14] Yoon HI, Yoon J, Chung Y, et al. Individual case review in a phase 3 randomized trial to investigate the role of internal mammary lymph node irradiation for breast cancer: Korean Radiation Oncology Group 08-06 study. Radiother Oncol Apr 2017;123(1):15–21. https://doi.org/10.1016/j.radonc.2017.01.017.

[15] van der Veen J, Gulyban A, Willems S, Maes F, Nuyts S. Interobserver variability in organ at risk delineation in head and neck cancer. Radiat Oncol 2021;16(1):120. https://doi.org/10.1186/s13014-020-01677-2. 2021/06/28.

[16] Poortmans PMP, Takanen S, Marta GN, Meattini I, Kaidar-Person O. Winter is over: the use of Artificial Intelligence to individualise radiation therapy for breast cancer. Breast Feb 2020;49:194–200. https://doi.org/10.1016/j.breast.2019.11.011.

[17] Kaidar-Person O, Vrou Offersen B, Hol S, et al. ESTRO ACROP consensus guideline for target volume delineation in the setting of postmastectomy radiation therapy after implant-based immediate reconstruction for early stage breast cancer. Radiother Oncol Aug 2019;137:159–66. https://doi.org/10.1016/j.radonc.2019.04.010.

[18] Choi MS, Choi BS, Chung SY, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. Radiother Oncol Dec 2020;153:139–45. https://doi.org/10.1016/j.radonc.2020.09.045.

[19] Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol Skr 1948;5:1–34.

[20] Nikolov S, Blackwell S, Zverovitch A, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. 2018. *arXiv preprint arXiv:180904430*.

[21] Vaassen F, Hazelaar C, Vaniqui A, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol Jan 2020;13:1–6. https://doi.org/10.1016/j.phro.2019.12.001.

[22] Kim K, Chun M, Jin H, et al. Inter-institutional variation in intensity-modulated radiotherapy for breast cancer in Korea (KROG 19-01). Anticancer Res Jun 2021;41(6):3145–52. https://doi.org/10.21873/anticanres.15100.

[23] Chang JS, Chang JH, Kim N, Kim YB, Shin KH, Kim K. Intensity modulated radiotherapy and volumetric modulated arc therapy in the treatment of breast cancer: an updated review. J Breast Cancer Oct 2022;25(5):349–65. https://doi.org/10.4048/jbc.2022.25.e37.

[24] Choi KH, Ahn SJ, Jeong JU, et al. Postoperative radiotherapy with intensity-modulated radiation therapy versus 3-dimensional conformal radiotherapy in early breast cancer: a randomized clinical trial of KROG 15-03. Radiother Oncol Jan 2021;154:179–86. https://doi.org/10.1016/j.radonc.2020.09.043.

[25] Horner-Rieber J, Forster T, Hommertgen A, et al. Intensity modulated radiation therapy (IMRT) with simultaneously integrated boost shortens treatment time and is noninferior to conventional radiation therapy followed by sequential boost in adjuvant breast cancer treatment: results of a large randomized phase III trial (IMRT-MC2 trial). Int J Radiat Oncol Biol Phys Apr 1 2021;109(5):1311–24. https://doi.org/10.1016/j.ijrobp.2020.12.005.

[26] Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J Med Imaging Radiat Oncol Jun 2016;60(3):393–406. https://doi.org/10.1111/1754-9485.12462.

[27] Eansor P, Norris ME, D'Souza LA, et al. Can we identify predictors of success in contouring education for radiation oncology trainees? An analysis of the anatomy and radiology contouring bootcamp. Pract Radiat Oncol Nov-Dec 2022;12(6):e486–92. https://doi.org/10.1016/j.prro.2022.05.016.

[28] Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J Med Imag Radiat On Jun 2016;60(3):393–406. https://doi.org/10.1111/1754-9485.12462.

[29] Ciardo D, Argenone A, Boboc GI, et al. Variability in axillary lymph node delineation for breast cancer radiotherapy in presence of guidelines on a multi-institutional platform. Acta Oncol Aug 2017;56(8):1081–8. https://doi.org/10.1080/0284186X.2017.1325004.

[30] McIntosh C, Conroy L, Tjong MC, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. Nat Med Jun 2021;27(6):999. https://doi.org/10.1038/s41591-021-01359-w.

[31] Vaassen F, Boukerroui D, Looney P, et al. Real-world analysis of manual editing of deep learning contouring in the thorax region. Phys Imaging Radiat Oncol Apr 2022;22:104–10. https://doi.org/10.1016/j.phro.2022.04.008.

[32] Cox S, Cleves A, Clementel E, Miles E, Staffurth J, Gwynne S. Impact of deviations in target volume delineation - time for a new RTQA approach? Radiother Oncol Aug 2019;137:1–8. https://doi.org/10.1016/j.radonc.2019.04.012.