BJR © 2021 The Authors. Published by the British Institute of Radiology

Received: 29 September 2020 | Revised: 23 December 2020 | Accepted: 25 January 2021

https://doi.org/10.1259/bjr.20201177

Cite this article as:
Leonardi MC, Pepa M, Gugliandolo SG, Luraschi R, Vigorito S, Rojas DP,  et al. Geometric contour variation in clinical target volume of axillary lymph nodes in breast cancer radiotherapy: an AIRO multi-institutional study. *Br J Radiol* 2021; **94**: 20201177.

## FULL PAPER

# Geometric contour variation in clinical target volume of axillary lymph nodes in breast cancer radiotherapy: an AIRO multi-institutional study

**[1]MARIA CRISTINA LEONARDI, MD, [1]MATTEO PEPA, MSc, [1]SIMONE GIOVANNI GUGLIANDOLO, MSc, [2]ROSA LURASCHI, MSc, [2]SABRINA VIGORITO, MSc, [1]DAMARIS PATRICIA ROJAS, MD, [3]MARIA ROSA LA PORTA, MD, [3]DOMENICO CANTE, MD, [4]EDOARDO PETRUCCI, MSc, [5]LORENZA MARINO, MD, [6]GIUSEPPINA BORZÌ, MSc, [7]EDY IPPOLITO, MD, [8]MARISTELLA MARROCCO, MSc, [9]ALESSANDRA HUSCHER, MD, [10]MATTEO CHIEREGATO, MSc, [11]ANGELA ARGENONE, MD, [12]LUCIANO IADANZA, MSc, [13]FIORENZA DE ROSE, MD, [13]FRANCESCA LOBEFALO, MSc, [14]FRANCESCA CUCCIARELLI, MD, [15]MARCO VALENTI, MSc, [16]MARIA CARMEN DE SANTIS, MD, [17]ANNA CAVALLO, MSc, [18]FRANCESCA ROSSI, MD, [19]SERENELLA RUSSO, MSc, [20]AGNESE PRISCO, MD, [21]MARIKA GUERNIERI, MSc, [22]ROBERTA GUARNACCIA, MD, [23]TIZIANA MALATESTA, MSc, [24]ILARIA MEAGLIA, MD, [25]MARCO LIOTTA, MSc, [25]PAOLA TABARELLI DE FATIS, MSc, [26]ISABELLA PALUMBO, MD, [27]MARTA MARCANTONINI, MSc, [28]SARAH PIA COLANGIONE, MD, [29]EMILIO MEZZENGA, MSc, [30]SARA FALIVENE, MD, [31]MARIA MORMILE, MSc, [32]VINCENZO RAVO, MD, [32]CECILIA ARRICHIELLO, MSc, [33]ALESSANDRA FOZZA, MD, [34]MARIA PAOLA BARBERO, MSc, [24]GIOVANNI BATTISTA IVALDI, MD, [35]GIANPIERO CATALANO, MD, [36]CRISTIANA VIDALI, MD, [26]CYNTHIA ARISTEI, MD, [37]CATERINA GIANNITTO, MD, [1]ELEONORA MIGLIETTA, MSc, [38]ANTONELLA CIABATTONI, MD, [39,40]ICRO MEATTINI, MD, [41]ROBERTO ORECCHIA, MD, [2]FEDERICA CATTANI, MSc, [1,42]BARBARA ALICJA JERECZEK-FOSSA, MD PhD and on behalf of the Breast Study Group (BSG) of the Italian Association of Radiotherapy and Clinical Oncology (AIRO)**

[1]Division of Radiation Oncology, IEO Istituto Europeo di Oncologia IRCCS, Milano, Italy
[2]Unit of Medical Physics, IEO Istituto Europeo di Oncologia IRCCS, Milano, Italy
[3]Radiotherapy Department, ASL TO4 Ivrea Community Hospital, Ivrea, Italy
[4]Unit of Medical Physics, ASL TO4 Ivrea Community Hospital, Ivrea, Italy
[5]Radiotherapy Unit, REM Radioterapia, Viagrande (CT), Italy
[6]Unit of Medical Physics, REM Radioterapia, Viagrande (CT), Italy
[7]Department of Radiotherapy, Campus Bio-Medico University, Roma, Italy
[8]Unit of Medical Physics, Campus Bio-Medico University, Roma, Italy
[9]Radiotherapy Unit, Fondazione Poliambulanza, Brescia, Italy
[10]Unit of Medical Physics, Fondazione Poliambulanza, Brescia, Italy
[11]Division of Radiation Oncology, Azienda Ospedaliera di Rilievo Nazionale San Pio, Benevento, Italy
[12]Unit of Medical Physics, Azienda Ospedaliera di Rilievo Nazionale San Pio, Benevento, italy
[13]Radiotherapy and Radiosurgery Department, Humanitas Clinical and Research Centre IRCCS, Milano, Italy
[14]Department of Internal Medicine, Radiotherapy Institute, Ospedali Riuniti Umberto I, G.M. Lancisi, G. Salesi, Ancona, Italy
[15]Unit of Medical Physics, Ospedali Riuniti Umberto I, G.M. Lancisi, G. Salesi, Ancona, Italy
[16]Radiotherapy Unit 1, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy
[17]Unit of Medical Physics, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy
[18]Radiotherapy Unit, Usl Toscana Centro, Ospedale Santa Maria Annunziata, Firenze, Italy
[19]Unit of Medical Physics, Usl Toscana Centro, Ospedale Santa Maria Annunziata, Firenze, Italy
[20]Department of Radiotherapy, ASUFC - P.O. “ Santa Maria della Misericordia” di Udine, Udine, Italy
[21]Unit of Medical Physics, ASUFC - P.O. “ Santa Maria della Misericordia” di Udine, Udine, Italy
[22]Radiotherapy Unit, Ospedale Fatebenefratelli Isola Tiberina, Roma, Italy
[23]Unit of Medical Physics, Ospedale Fatebenefratelli Isola Tiberina, Roma, Italy
[24]Radiation Oncology Unit, Istituti Clinici Scientifici Maugeri IRCCS, Pavia, Italy
[25]Medical Physics Unit, Istituti Clinici Scientifici Maugeri IRCCS, Pavia, Italy
[26]Radiation Oncology Section, University of Perugia and Perugia General Hospital, Perugia, Italy
[27]Medical Physics Unit, Perugia General Hospital, Perugia, Italy
[28]Radiotherapy Unit, Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) IRCCS, Meldola, Italy
[29]Medical Physics Unit, IRCCS Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) "Dino Amadori", Meldola (FC), Italy
[30]Department of Radiotherapy, ASL Napoli 1 Centro - Ospedale del Mare, Napoli, Italy
[31]Unit of Medical Physics, ASL Napoli 1 Centro - Ospedale del Mare, Napoli, Italy
[32]Unit of Radiotherapy, Istituto Nazionale Tumori – IRCCS - Fondazione G. Pascale, Napoli, Italy
[33]Division of Radiation Oncology, Azienda Ospedaliera Nazionale SS. Antonio e Biagio e Cesare Arrigo, Alessandria, Italy
[34]Unit of Medical Physics, Azienda Ospedaliera Nazionale SS. Antonio e Biagio e Cesare Arrigo, Alessandria, Italy
[35]Department of Radiotherapy, IRCCS MultiMedica, Sesto San Giovanni (MI), Italy
[36]Department of Radiation Oncology, Azienda Sanitaria Universitaria Integrata di Trieste (ASUI-TS), Trieste, Italy
[37]Division of Radiology, IEO Istituto Europeo di Oncologia IRCCS, Milano, Italy
[38]Department of Radiotherapy, San Filippo Neri Hospital, ASL Roma 1, Roma, Italy

[39]Radiation Oncology Unit - Oncology Department, Azienda Ospedaliero-Universitaria Careggi, Firenze, Italy
[40]Department of Experimental and Clinical Biomedical Sciences "M. Serio", University of Florence, Firenze, Italy
[41]Scientific Direction, IEO Istituto Europeo di Oncologia IRCCS, Milano, Italy
[42]Department of Oncology and Hemato-Oncology, University of Milan, Milano, Italy

Address correspondence to: Mrs Damaris Patricia Rojas
E-mail: *damarojas@gmail.com*

**Objectives:** To determine interobserver variability in axillary nodal contouring in breast cancer (BC) radiotherapy (RT) by comparing the clinical target volume of participating single centres (SC-CTV) with a gold-standard CTV (GS-CTV).
**Methods:** The GS-CTV of three patients (P1, P2, P3) with increasing complexity was created in DICOM format from the median contour of axillary CTVs drawn by BC experts, validated using the simultaneous truth and performance-level estimation and peer-reviewed. GS-CTVs were compared with the correspondent SC-CTVs drawn by radiation oncologists, using validated metrics and a total score (TS) integrating all of them.
**Results:** Eighteen RT centres participated in the study. Comparative analyses revealed that, on average, the SC-CTVs were smaller than GS-CTV for P1 and P2 (by −29.25% and −27.83%, respectively) and larger for P3 (by +12.53%). The mean Jaccard index was greater for P1 and P2 compared to P3, but the overlap extent value was around 0.50 or less. Regarding nodal levels, L4 showed the highest concordance with the GS. In the intra-patient comparison, L2 and L3 achieved lower TS than L4. Nodal levels showed discrepancy with GS, which was not statistically significant for P1, and negligible for P2, while P3 had the worst agreement. DICE similarity coefficient did not exceed the minimum threshold for agreement of 0.70 in all the measurements.
**Conclusions:** Substantial differences were observed between SC- and GS-CTV, especially for P3 with altered arm setup. L2 and L3 were the most critical levels. The study highlighted these key points to address.
**Advances in knowledge** The present study compares, by means of validated geometric indexes, manual segmentations of axillary lymph nodes in breast cancer from different observers and different institutions made on radiotherapy planning CT images. Assessing such variability is of paramount importance, as geometric uncertainties might lead to incorrect dosimetry and compromise oncological outcome.

## INTRODUCTION

Throughout the radiotherapy (RT) workflow, from simulation to treatment planning, the clinical target volume (CTV) delineation is one of the most crucial steps,[1,2] since inaccuracy can lead to a systematic error downstream.[3] Technological advances allowing the delivery of more and more conformal RT are bound to enhance the impact of uncertainties in contouring.[4,5] Significant inter-observer variability in CTV delineation of many tumours, including breast cancer (BC), has been previously described.[6–9] Against the backdrop of lack of standardization of methodology for contouring comparison, different metrics and contouring procedures have been used. Vinod and coll.[10] reviewed the variety of tools used for analysis of contours, encompassing group consensus, reference model, average, median, randomly selected contour, Gold Standard (GS), Simultaneous Truth and Performance Level Estimation (STAPLE).[11] Even if the large number of available metrics makes the results from different studies difficult to be compared, their great value is to create awareness about the most common sources of sub-optimal contouring.[12–14]

The axillary nodal contouring variability at multi- and intra-institutional level was investigated in a previous study[15] endorsed by the Breast Study Group (BSG) of the Italian Association of Radiotherapy and Clinical Oncology (AIRO), where three radiation oncologists (ROs) with different expertise (the Expert, the Senior, the Junior) worked on three representative patients with different complexity (P1 "the simple anatomy", P2 "the obese", P3 "the altered arm set-up"). The group consensus was generated using the STAPLE algorithm and acted as reference mean contours for comparison calculations.

The current investigation drew on the same datasets and represents an expansion of the above-mentioned study.[15] Instead of using group consensus, a GS CTV (GS-CTV) was generated to assess geometric variation. Comparative analyses between GS-CTVs and the nodal CTVs delineated by the participating centres, named single-centre CTVs (SC-CTV), quantified the interobserver variability from a different perspective and laid the ground for the subsequent study evaluating dosimetric variation.

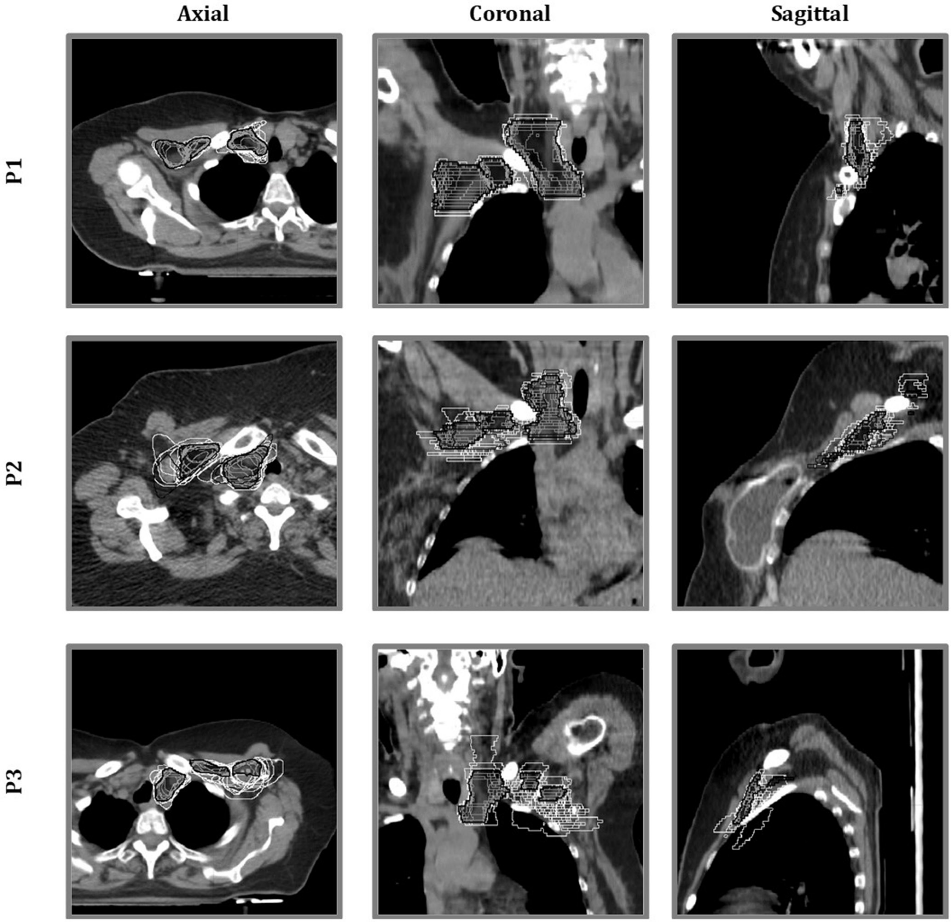
## METHODS AND MATERIALS

The present study is part of an investigation on nodal contouring variability in BC RT initiated and developed within the Breast Study Group (BSG) of the Italian Association of Radiotherapy and Clinical Oncology (AIRO). Of the 15 initial participating RT institutions, 14 accepted to continue the collaboration. In addition, four new centres, where junior ROs familiar with the project have moved to, joined the investigation. The patients were treated at the IEO, European Institute of Oncology, IRCCS, Milan, Italy. They gave written informed consent for the use of their anonymized clinical and image data for research and training purposes. The study was conducted within the research project on intensity-modulated RT (IMRT) and hypofractionation on BC notified to the Ethical Committee of the European Institute of Oncology (IEO) (26 May 2016, Milan, Italy) and approved by the review board.

### Gold standard creation and study procedure

The axillary nodal GS-CTV for each of the three case patients was created in DICOM format by the median of all CTVs drawn by the BC experts of the RT centres originally involved, validated

Figure 1. Qualitative assessment of areas of variability: for all the three representative patients (P1, P2 and P3), all clinical target volumes (CTVs) (single-centre CTVs in white and gold standard CTV in black and grey-filled) were overlaid in the axial, coronal and sagittal CT planes, to visually quantify the interobserver variability.



using the STAPLE algorithm and peer-reviewed by three independent BC experts (CG, GBI, CV) and one radiologist (CG) to reach a consensus outline (Figure 1). Subsequently, the nodal CTVs independently drawn by the 14 expert/senior and the four junior ROs (SC-CTV) on the same CT images of the three patients were retrieved from the previous work and compared to the newly formed GS-CTVs, whose volumes are reported in Table 1. Such comparisons were performed both considering the GS- and the SC-CTVs as a whole and by breaking down the nodal level (L2, L3, L4) CTVs using the MIM software v. 6.1.7 (MIM Software, Cleveland, OH). L1 was not investigated in this phase. The participants were asked to follow the AIRO contouring guidelines.[15–17]

Table 1. Volumetric characteristics of gold standard-clinical target volumes (GS-CTVs) of the nodal levels (L2, L3 and L4), individually and collectively, by representative patient (P1, P2 and P3).

| | P1 | P2 | P3 |
|---|---|---|---|
| **L2 GS-CTV** | 29.00 | 45.29 | 17.96 |
| **L3 GS-CTV** | 13.72 | 24.22 | 11.76 |
| **L4 GS-CTV** | 43.07 | 59.28 | 28.49 |
| **Whole GS-CTV** (L2 + L3+L4) | 85.78 | 128.80 | 58.19 |

GS-CTV, gold standard clinicaltarget volume; L2, L3, L4, lymph node level 2, 3 and 4; P1, P2, P3, patient 1, 2 and 3.
The table reports volumes expressed in cm$^3$.

## Indexes for geometric comparisons

Geometric differences between the GS-CTV and the SC-CTV for each patient were measured according to some of the metrics proposed by the American Association of Physicists in Medicine Task Group 132[18] for evaluating differences in image registration. Specifically, differences in terms of shape, size and position between the volume encased by the GS contour ($V_{GS}$) and by the SC contour ($V_{SC}$), as well as between single nodal levels were assessed using the following geometric metrics:

- **Jaccard Index (JI)**
  This index is defined as the volume intersected by $V_{GS}$ and $V_{SC}$, divided by the union of the two volumes[19] :

$$JI = \frac{V_{GS} \cap V_{SC}}{V_{GS} \cup V_{SC}} = \frac{V_{GS} \cap V_{SC}}{V_{GS} + V_{SC} - V_{GS} \cap V_{SC}}$$

  This index ranges from 0 (no overlap) to 1 (perfect overlap).

- **Mean Distance to Conformity (MDC)**
  This index is defined as the mean distance, expressed in mm, between each point in $V_{GS}$ ($P_{GS,i}$), and the nearest point in $V_{SC}$ ($P_{SC,i}$). One mathematical translation of that is, approximately:

$$MDC = \frac{\sum_{i=1}^{N} d\left(P_{GS_i}, P_{SC_i}\right)}{N} = \frac{\sum_{i=1}^{N} \sqrt{\left(x_{GS_i}-x_{SC_i}\right)^2+\left(y_{GS_i}-y_{SC_i}\right)^2+\left(z_{GS_i}-z_{SC_i}\right)^2}}{N}$$

  where $N$ is the total number of considered points. If contours perfectly overlap, this index is equal to 0, otherwise it increases as agreement decreases.
- **Volume Difference (VD)**
  It is the relative percentage difference between $V_{GS}$ and $V_{SC}$.

$$VD = \left(\frac{V_{SC}- V_{GS}}{V_{GS}} \cdot 100\right) \%$$

  The absolute value of VD and level of agreement are inversely correlated.
- **DICE Similarity Coefficient (DSC)**
  It is defined as twice the volume encompassed by both contours divided by the sum of the two volumes. In symbols:

$$DSC = 2\,\frac{V_{GS} \cap V_{SC}}{V_{GS}+ V_{SC}}$$

  The higher the DSC, the higher the agreement.
- **Sensitivity Index (SI)**
  It quantifies the intersection volume between $V_{SC}$ and $V_{GS}$ compared to the volume of the latter. It measures the probability that the SC contour matches the GS one.[16]
  In symbols:

$$SI = \frac{V_{GS} \cap V_{SC}}{V_{GS}}$$

  The higher the SI, the higher the agreement.
- **Inclusion Index (II)**

  It quantifies the intersection volume between $V_{SC}$ and $V_{GS}$ compared to the volume of the former. It measures the probability that a voxel of the SC contour is actually a voxel of the GS one.[16] In symbols:

$$II = \frac{V_{GS} \cap V_{SC}}{V_{SC}}$$

  The higher the II, the higher the agreement.

For a better interpretation of the results and to give an immediate indication of the degree of concordance with the GS of patients and nodal levels, an additional parameter, namely Total Score (TS), was created by assigning to each index a point value from 1 to 3 (from worst to best). Being JI, DSC, SI and II directly correlated to the level of agreement, higher scores were given to higher indexes. On the other hand, since MDC and the absolute value of VD increase as the agreement decreases, higher scores were assigned to lower indexes. In this way, a higher TS corresponded to a higher level of agreement with the GS. All the metrics derived for contour analysis were computed using the ImSimQA software (v4.2, Oncology Systems Limited, Shrewsbury, UK).[20]

### Statistical analysis

For every type of score, any significant difference between the distributions of segmentation results across patients and across lymph nodes was checked with the Kruskal-Wallis test. Differences among patients were assessed both on individual nodal level and the entire CTV. If any statistical significance was found ($p$-values lower than 0.05), the deviating patient/nodal level was identified as "odd-one-out" (OOO).

## RESULTS

A total of 18 RT centres participated in the study. All SC-CTVs for P1 were analysed, whereas 1 SC-CTV for P2 and 2 SC-CTVs for P3 were excluded due to technical issues in the uploading process of DICOM images. For each representative case, the SC-CTV was compared to the GS-CTV (Table 2). Overall, for P1 and P2, the SC-CTV was significantly underestimated by the ROs (by −29.25% and −27.83%, respectively), whereas for P3 it was slightly overestimated (by +12.53%). Given the mathematical definition of SI and II, which are inversely correlated with $V_{GS}$ and $V_{SC}$, respectively, this observation was further supported by the average higher value of SI and the lower average value of II (0.63 and 0.59, respectively). The GS volumes, expressed in $cm^3$, are reported in Table 1. DSC values for both individual nodal levels and the whole SC-CTV did not exceed 0.70, which is considered the minimum threshold for agreement. Average JI was lower than 0.50 in almost all the comparisons. The mean distance between each point of GS and the nearest point of SC volumes, that is MDC, was greater than 5 mm. Many indexes presented high standard deviation (SD), which confirmed the high contouring variability.

### Inter-patient and intra-lymph node level analysis

Table 2 reported the agreement between the SC and GS contour volumes, considering both the whole CTV and the single nodal levels, for the three case patients. When the whole CTV was considered, the lowest TS was assigned to P3 and was mainly imputable to the lowest scores achieved in L2 and L3. In particular, the difference between P3 and the other patients was statistically significant when considering JI, VD and II. L4 showed the greatest concordance, even in P3. No difference between the SC and GS contour volumes of L2, L3 and the entire CTV was observed for P1 and P2. The JI, which expresses the extent of overlap, was quite low (less than 0.50 for L2 and L3 and slightly higher or close to 0.50 for L4). Figure 1 provides a graphical visualization of the interobserver variability, showing SC-CTVs (white-coloured) overlap and discrepancy with the GS-CTV (black-coloured and grey-filled).

### Inter-lymph node level and intra-patient analysis

Table 3 reported the agreement between the SC and the GS contour volumes of single nodal levels within the same patient. For all patients, L4 exhibited the greatest concordance with GS and therefore achieved the highest TS. Especially for P3, L4 differed from the other nodal levels to such an extent to be considered the "odd-one-out" (OOO) for almost all indexes. P1 achieved the best agreement for any axillary level ($p$-values > 0.05). Conversely, virtually all the indexes for P3 revealed

Table 2. Inter patient comparison: degree of agreement between single centre- and gold standard- clinical target volumes by axillary nodal level (L2, L3 and L4). Mean values are reported

| Contour | Patient | JI | MDC (mm) | VD (%) | DSC | SI | II | TS |
|---|---|---|---|---|---|---|---|---|
| L2 CTV | P1 | 0.44 (0.16) | 8.70 (4.18) | −39.64 (16.46) | 0.61 (0.17) | 0.49 (0.17) | 0.83 (0.19) | 2.3 |
| | P2 | 0.40 (0.12) | 9.05 (2.70) | −32.33 (20.18) | 0.60 (0.12) | 0.51 (0.14) | 0.76 (0.12) | 2.2 |
| | P3 | 0.28 (0.13) | 12.22 (5.80) | 38.21 (53.41) | 0.45 (0.18) | 0.54 (0.24) | 0.40 (0.16) | **1.5** |
| | *p*-value | **0.017** | 0.111 | **<0.001** | **0.024** | 0.537 | **<0.001** | |
| | OOO | P3 | | P3 | P3 | | P3 | |
| | AVE | 0.37 | 9.99 | 36.73 | 0.55 | 0.51 | 0.66 | |
| L3 CTV | P1 | 0.42 (0.16) | 10.03 (10.43) | −23.23 (44.11) | 0.59 (0.22) | 0.54 (0.21) | 0.69 (0.26) | 1.8 |
| | P2 | 0.43 (0.13) | 8.64 (6.09) | −13.33 (41.50) | 0.64 (0.14) | 0.60 (0.17) | 0.74 (0.16) | 3.0 |
| | P3 | 0.35 (0.19) | 11.80 (12.08) | 44.03 (95.12) | 0.51 (0.25) | 0.57 (0.26) | 0.51 (0.29) | **1.2** |
| | *p*-value | 0.296 | 0.676 | **0.012** | 0.230 | 0.681 | **0.021** | |
| | OOO | | | P3 | | | P3 | |
| | AVE | 0.40 | 10.16 | 26.86 | 0.58 | 0.57 | 0.65 | |
| L4 CTV | P1 | 0.52 (0.12) | 7.39 (1.97) | −25.63 (15.63) | 0.70 (0.10) | 0.61 (0.13) | 0.83 (0.10) | 2.3 |
| | P2 | 0.49 (0.08) | 7.78 (1.76) | −32.66 (15.40) | 0.67 (0.07) | 0.57 (0.11) | 0.85 (0.07) | **1.5** |
| | P3 | 0.53 (0.13) | 8.28 (4.23) | −18.04 (22.90) | 0.70 (0.12) | 0.63 (0.13) | 0.80 (0.16) | 2.2 |
| | *p*-value | 0.429 | 0.625 | 0.192 | 0.471 | 0.681 | 0.886 | |
| | OOO | | | | | | | |
| | AVE | 0.51 | 7.82 | 25.44 | 0.69 | 0.60 | 0.83 | |
| Whole CTV (L2 +L3+L4) | P1 | 0.48 (0.10) | 7.79 (1.91) | −29.25 (13.44) | 0.68 (0.08) | 0.58 (0.09) | 0.83 (0.09) | 2.2 |
| | P2 | 0.48 (0.07) | 8.28 (1.67) | −27.83 (14.21) | 0.67 (0.06) | 0.58 (0.09) | 0.81 (0.06) | 2.2 |
| | P3 | 0.39 (0.11) | 9.77 (2.90) | 12.53 (32.13) | 0.60 (0.12) | 0.63 (0.14) | 0.59 (0.15) | **1.7** |
| | *p*-value | **0.033** | 0.119 | **<0.001** | 0.112 | 0.225 | **<0.001** | |
| | OOO | P3 | | P3 | | | P3 | |
| | AVE | 0.45 | 8.61 | 23.20 | 0.65 | 0.60 | 0.74 | |

AVE, average; CTV, clinical target volume; DSC, DICE similarity coefficient; II, inclusion index; JI, Jaccard index; L2,L3, L4, lymph node level 2, 3 and 4; MDC, mean distance to conformity; OOO, odd-one-out (the one differing from the others in the group); SI, sensitivity index; TS, total score; VD, volume difference.

Standard deviations (SDs) are put in parentheses. Statistically significant *p*-values (<0.05) are highlighted in bold.

statistically significant differences between the compared SC and GS volumes.

## DISCUSSION

Accurate target delineation represents one of the most critical steps of modern RT, which aims to deliver more and more conformal treatments.[21–24] Several studies have demonstrated substantial interobserver variations in the contouring of breast target volumes, especially with respect to tumour bed and nodal regions.[9,17,25,26] Reasons for clinical practice variability are multifactorial and partly due to the small target volumes, which enhance the relative differences in contours. In the study by Li and coll.,[26] the mean JI expressed as percent overlap was 39–51% for axillary nodes compared to 72–77% for the larger chest wall/ breast volumes.

The significantly higher variability found in the previous study[15] across patients (as the difficulty of the case increased, concordance decreased) and nodal levels (worse results with those centrally located) has been confirmed by the current one in presence of GS as a benchmark. These comparative analyses were restricted to L2-L4 because these nodal levels are generally included in the locoregional RT after nodal dissection.[27] L1 coverage will be the object of subsequent investigation. Overall, the mean DSC was very similar between the previous and the current study, being comprised between 0.60 and 0.70, confirming the moderate agreement. The mean JI of nodal levels remained quite low, around 0.50 or less, especially for P3. As a matter of fact, the SC-CTV for P3 was larger than that outlined for P1 and P2. This finding confirms the trend of drawing larger volumes to offset uncertainty linked to the different arm position. In the atlas published by Martinez-Monge and coll.[28] the breast nodal levels were delineated as large areas instead of separate entities to underline uncertainty about the exact locations. P3 with the altered arm set-up presented the greatest variability according to all the metrics, mainly affecting L2 and L3. This is

Table 3. Inter nodal level comparison: degree of agreement between single centre- and gold standard- clinical target volumes by patient (P1, P2 and P3). Mean values are reported

| Patient | Contour | JI | MDC (mm) | VD (%) | DSC | SI | II | TS |
|---|---|---|---|---|---|---|---|---|
| **P1** | L2 | 0.44 (0.16) | 8.70 (4.18) | −39.64 (16.46) | 0.61 (0.17) | 0.49 (0.17) | 0.83 (0.19) | 1.8 |
| | L3 | 0.42 (0.16) | 10.03 (10.43) | −23.23 (44.11) | 0.59 (0.22) | 0.54 (0.21) | 0.69 (0.26) | **1.5** |
| | L4 | 0.52 (0.12) | 7.39 (1.97) | −25.63 (15.63) | 0.70 (0.10) | 0.61 (0.13) | 0.83 (0.10) | 2.8 |
| | *p*-value | 0.164 | 0.500 | 0.087 | 0.194 | 0.268 | 0.076 | |
| | OOO | | | | | | | |
| | AVE | 0.457 | 8.706 | −29.499 | 0.632 | 0.548 | 0.785 | |
| **P2** | L2 | 0.40 (0.12) | 9.05 (2.70) | −32.33 (20.18) | 0.60 (0.12) | 0.51 (0.14) | 0.76 (0.12) | **1.3** |
| | L3 | 0.43 (0.13) | 8.64 (6.09) | −13.33 (41.50) | 0.64 (0.14) | 0.60 (0.17) | 0.74 (0.16) | 2.2 |
| | L4 | 0.49 (0.08) | 7.78 (1.76) | −32.66 (15.40) | 0.67 (0.07) | 0.57 (0.11) | 0.85 (0.07) | 2.5 |
| | *p*-value | 0.125 | 0.135 | 0.231 | 0.209 | 0.200 | **0.014** | |
| | OOO | | | | | | L4 | |
| | AVE | 0.439 | 8.493 | −26.107 | 0.635 | 0.557 | 0.782 | |
| **P3** | L2 | 0.28 (0.13) | 12.22 (5.80) | 38.21 (53.41) | 0.45 (0.18) | 0.54 (0.24) | 0.40 (0.16) | **1.2** |
| | L3 | 0.35 (0.19) | 11.80 (12.08) | 44.03 (95.12) | 0.51 (0.25) | 0.57 (0.26) | 0.51 (0.29) | 1.8 |
| | L4 | 0.53 (0.13) | 8.28 (4.23) | −18.04 (22.90) | 0.70 (0.12) | 0.63 (0.13) | 0.80 (0.16) | 3.0 |
| | *p*-value | **<0.001** | **0.042** | **0.001** | **0.001** | 0.796 | **<0.001** | |
| | OOO | L4 | L2[a] | L4 | L4 | | L4 | |
| | AVE | 0.383 | 10.768 | 21.402 | 0.550 | 0.580 | 0.570 | |

AVE, average; DSC, DICE similarity coefficient; II, inclusion index; JI, Jaccard index; MDC, mean distance to conformity; OOO, odd-one-out (the one differing from the others in the group); P1, P2, P3, patient 1, 2 and 3; SI, sensitivity index; TS, total score; VD, volume difference.

Standard deviations (SDs) are put in parentheses.

Statistically significant *p*-values (<0.05) are highlighted in bold.

[a] *p*-values 0.07 and 0.01 towards L3 and L4, respectively.

consistent with the previous work and other reports showing that the position of the arm changes the location of the axillary nodes, particularly the central ones, in relation to the movement of the adjacent vessels and muscles.[15,29–32] As a result, the whole CTV of P3 presented the lowest JI, DSC, and II, while showing the highest MDC and SI compared to P1 and P2. In the comparison of the individual nodal level SC-CTVs within each patient, P1 and P2 showed no statistically significant discrepancies with the GS counterparts (except the metric II for P2, which favored L4 against L2-3). However, looking at the TS assigned to axillary levels of these two patients, L2 and L3 presented the worst value, confirming that poor concordance, although small, could remain even with conventional arm set-up. Other authors found that the most critical region is the crossing point from the medial part of infraclavicular nodes to the lateral part of supraclavicular nodes.[26,33] In the current study, the whole supraclavicular fossa, corresponding to L4, contributed to variability to a lesser extent, and high concordance with the GS was observed even for P3. In fact, for all patients, the highest TS has always been assigned to L4. Similar findings emerged also in the previous study.[15]

Comparative analysis of the single nodal levels in each patient with the GS counterparts showed high concordance for P1, the simple case patient, while the worst degree of agreement was observed for P3, once again confirming the results of the first study.[15]

The choice of reference volume for comparison is always a matter of contention. The previous study[15] used the STAPLE contours as a reference for comparison, because the aim was not to evaluate the accuracy of the contours, but the concordance between observers. In the current one, the GS contour was created as necessary step to correlate the volume variations with the dosimetric differences in the treatment planning. Since every method used for comparison can be subject to criticism,[34] the choice of the GS is both a strength and a limit. The GS-CTV was created by the median of all the CTVs drawn by the experts involved in the first study for each patient and subsequently validated by the STAPLE algorithm to further reduce differences, therefore, providing the most probable volume. Finally, it underwent adjustments and refinements by three independent ROs and one diagnostic radiologist until agreement was reached. In such a way, the GS-CTVs were considered to be reliable in defining the true extension of nodal levels. However, it can be argued that the "true" CTV is not a real entity.[34] As Allozi and coll.[35] stated, although experts could agree on identification of a certain region of interest on CT images, that does not mean it is necessarily correct. This statement is particularly true for nodal axillary drainage, which is basically a “clinical” target:[12] the landmarks centred on vessels and muscles on non-enhanced CT images are not sufficiently solid to define clear borders, being heavily affected by individual patient variability and physicians’ interpretation. Such uncertainties can generate different CTVs and make it challenging to guarantee the standardization of uniform delineation. To make comparison with the GS, only one SC-CTV for each institution was extracted. As a rule, the choice fell on the SC-CTV drawn by the expert or, as an alternative, by the senior, of each participating centre. On the other hand, the four junior ROs of the newly joined institutions were given their own CTVs, to avoid any duplication. The selection was arbitrarily made and might represent a limitation of the study, although several reasons lay behind this decision. The GS-CTVs were initially created by the median of the SC-CTVs drawn only by the experts; therefore, it seemed reasonable to follow suit. Moreover, by analsing only one SC-CTV for each patient (in total 51 compared 126 CTVs of the first study), the results were streamlined in order to simplify the subsequent study on the dosimetric impact.

Even if we restricted the analysis, the number of participants and datasets remains a strength of the study, being larger than the median of 7 and 9, respectively, reported in the literature by Vinod and coll..[10] The optimal number of observers for this kind of investigation is unknown, but similar studies showed that, as the number of observers increased, the statistical significance of the results increased.[7,8,36] Although there is no consensus on the appropriate metrics for contours comparison,[37] in this study the number of the indexes was expanded compared to the previous one, in order to better quantify the contour variability. The use of multiple metrics allows more in-depth analysis of the contours,[37] as each index contributes by adding a useful piece of information. The definition of a total score, which encompasses all the metrics, can be considered as one of the main strength points of the study, as it gave a clearer and immediate evaluation of the degree of consistency between SC and GS-CTVs. In this way, geometric differences in shape, volume and position between contours are evaluated as a whole, with each of parameters having the same weight. However, in some respects, the TS might be too restrictive and not reflecting the complexity of clinical reality. In fact, depending on the clinical situation, ROs can give higher priority to some parameters rather than others, resulting in discrepancy between the ideal TS and the appropriateness of contouring for that specific case.

To conclude, substantial differences were observed between the SC-CTV and the GS-CTV of the three representative patients, especially for the one with altered arm set-up (P3). The central levels (L2 and L3) were the most penalized. Improving level of accuracy is vital to high-precision RT. Identification of regions where variability is more marked can raise awareness and focus the attention of educational interventions to avoid potential pitfalls in breast target volume contouring. Future developments include the assessment of the dosimetric impact of such delineation variability.

## ACKNOWLEDGMENT

This work was partially supported by the Italian Ministry of Health with Ricerca Corrente and 5x1000 funds. MP and EM were supported by a research grant from Accuray Inc. entitled “Data collection and analysis of Tomotherapy and CyberKnife breast clinical studies, breast physics studies and prostate study”. SGG was partially supported by Associazione Italiana per la Ricerca sul Cancro (AIRC), project IG-14300 “Carbon ions boost followed by pelvic photon intensity modulated radiotherapy for high-risk prostate cancer”, registered at ClinicalTrials.gov (NCT02672449), approved by IEO R86/14- IEO 98. LJI is a PhD student at the European School of Molecular Medicine (SEMM), Milan, Italy. The sponsors did not play any role in the study design, collection, analysis and interpretation of data, nor in the writing of the manuscript, nor in the decision to submit the manuscript for publication. The authors thank the Scientific Committee and Board of the Italian Association of Radiotherapy and Clinical Oncology (AIRO) for the critical revision of the paper. We would like to thank the following for their contributions to the work: Anna Morra, Maria Alessia Zerella, Johannes Lars Isaksson, Delia Ciardo.

## CONFLICTS OF INTEREST

MCL, FC, BAJF received honorarium fee from Accuray Inc. outside the current article. The remaining authors declared no conflict of interest.

## FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## REFERENCES

1. 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* 2019; **137**: 1–8. doi: https://doi.org/10.1016/j.radonc.2019.04.012
2. Chang ATY, Tan LT, Duke S, Ng W-T. Challenges for quality assurance of target volume delineation in clinical trials. *Front Oncol* 2017; **7**: 221. doi: https://doi.org/10.3389/fonc.2017.00221
3. Segedin B, Petric P. Uncertainties in target volume delineation in radiotherapy - are they relevant and what can we do about them? *Radiol Oncol* 2016; **50**: 254–62. doi: https://doi.org/10.1515/raon-2016-0023
4. Sethi RA, No HS, Jozsef G, Ko JP, Formenti SC. Comparison of three-dimensional versus intensity-modulated radiotherapy techniques to treat breast and axillary level III and supraclavicular nodes in a prone versus supine position. *Radiother Oncol* 2012; **102**: 74–81. doi: https://doi.org/10.1016/j.radonc.2011.09.008
5. Dogan N, Cuttino L, Lloyd R, et al. Optimized dose coverage of regional lymph nodes in breast cancer: the role of IMRT. *Int J Radiat Oncol Biol Phys* 2007; **68**: 1238–50.

6. Landis DM, Luo W, Song J, Bellon JR, Punglia RS, Wong JS, et al. Variability among breast radiation oncologists in delineation of the postsurgical lumpectomy cavity. *Int J Radiat Oncol Biol Phys* 2007; **67**: 1299–308. doi: https://doi.org/10.1016/j.ijrobp.2006.11.026
7. Struikmans H, Wárlám-Rodenhuis C, Stam T, Stapper G, Tersteeg RJHA, Bol GH, et al. Interobserver variability of clinical target volume delineation of glandular breast tissue and of boost volume in tangential breast irradiation. *Radiother Oncol* 2005; **76**: 293–9. doi: https://doi.org/10.1016/j.radonc.2005.03.029
8. Batumalai V, Koh ES, Delaney GP, Holloway LC, Jameson MG, Papadatos G, et al. Interobserver variability in clinical target volume delineation in tangential breast irradiation: a comparison between radiation oncologists and radiation therapists. *Clin Oncol* 2011; **23**: 108–13. doi: https://doi.org/10.1016/j.clon.2010.10.004
9. Petersen RP, Truong PT, Kader HA, et al. Target volume delineation for partial breast radiotherapy planning: clinical characteristics associated with low interobserver concordance. *Int J Radiat Oncol Biol Phys* 2007; **69**: 41–8.
10. 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* 2016; **121**: 169–79. doi: https://doi.org/10.1016/j.radonc.2016.09.009
11. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004; **23**: 903–21. doi: https://doi.org/10.1109/TMI.2004.828354
12. Castro Pena P, Kirova YM, Campana F, Dendale R, Bollet MA, Fournier-Bidoz N, et al. Anatomical, clinical and radiological delineation of target volumes in breast cancer radiotherapy planning: individual variability, questions and answers. *Br J Radiol* 2009; **82**: 595–9. doi: https://doi.org/10.1259/bjr/96865511
13. Gentile MS, Usman AA, Neuschler EI, Sathiaseelan V, Hayes JP, Small W. Contouring guidelines for the axillary lymph nodes for the delivery of radiation therapy in breast cancer: evaluation of the RTOG breast cancer atlas. *Int J Radiat Oncol Biol Phys* 2015; **93**: 257–65. doi: https://doi.org/10.1016/j.ijrobp.2015.07.002
14. Jing H, Wang S-L, Li J, Xue M, Xiong Z-K, Jin J, et al. Mapping patterns of ipsilateral supraclavicular nodal metastases in breast cancer: rethinking the clinical target volume for high-risk patients. *Int J Radiat Oncol Biol Phys* 2015; **93**: 268–76. doi: https://doi.org/10.1016/j.ijrobp.2015.08.022
15. Ciardo D, Argenone A, Boboc GI, et al. On the behalf of the breast Working group of the Italian association of radiation oncology (AIRO). variability in axillary lymph node delineation for breast cancer radiotherapy in presence of guidelines on a multi-institutional platform. *Acta Oncol* 2017; **56**: 1081–8.
16. Italian association of radiation oncology (AIRO). La Radioterapia dei Tumori DELLA Mammella. *Indicazioni e Criteri Guida* 2013;: 75. −–81.
17. Cucciarelli F, Kirova YM, Palumbo I, Aristei C. Supraclavicular and infraclavicular lymph node delineation in breast cancer patients: a proposal deriving from a comparative study. *Tumori* 2015; **101**: 478–86. doi: https://doi.org/10.5301/tj.5000330
18. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM radiation therapy Committee task group No. 132. *Med Phys* 2017; **44**: e43–76. doi: https://doi.org/10.1002/mp.12256
19. Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimization for medical image segmentation: theory and practice when evaluating with dice score or Jaccard index. *IEEE Trans Med Imaging* 2020; **39**: 3679–90. Epub 2020 Oct 28. doi: https://doi.org/10.1109/TMI.2020.3002417
20. Wardman K, Prestwich RJD, Gooding MJ, Speight RJ. The feasibility of atlas-based automatic segmentation of MRI for H&N radiotherapy planning. *J Appl Clin Med Phys* 2016; **17**: 146–54. doi: https://doi.org/10.1120/jacmp.v17i4.6051
21. Dewas S, Bibault J-E, Blanchard P, Vautravers-Dewas C, Pointreau Y, Denis F, et al. Delineation in thoracic oncology: a prospective study of the effect of training on contour variability and dosimetric consequences. *Radiat Oncol* 2011; **6**: 118. doi: https://doi.org/10.1186/1748-717X-6-118
22. McCall R, MacLennan G, Taylor M, et al. Anatomical contouring variability in thoracic organs at risk. *Int J Radiat Oncol Biol Phys* 2007; **69**: 41–8.
23. Madu CN, Quint DJ, Normolle DP, Marsh RB, Wang EY, Pierce LJ. Definition of the supraclavicular and infraclavicular nodes: implications for three-dimensional CT-based conformal radiation therapy. *Radiology* 2001; **221**: 333–9. doi: https://doi.org/10.1148/radiol.2212010247
24. Ciardo D, Gerardi MA, Vigorito S, Morra A, Dell'acqua V, Diaz FJ, et al. Atlas-Based segmentation in breast cancer radiotherapy: evaluation of specific and generic-purpose atlases. *Breast* 2017; **32**: 44–52. doi: https://doi.org/10.1016/j.breast.2016.12.010
25. Kosztyla R, Olson R, Carolan H, Balkwill S, Moiseenko V, Kwan W. Evaluation of dosimetric consequences of seroma contour variability in accelerated partial breast irradiation using a constructed representative seroma contour. *Int J Radiat Oncol Biol Phys* 2012; **84**: 527–32. doi: https://doi.org/10.1016/j.ijrobp.2011.11.060
26. Li XA, Tai A, Arthur DW, Buchholz TA, Macdonald S, Marks LB, 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**: 944–51. doi: https://doi.org/10.1016/j.ijrobp.2008.10.034
27. National Comprehensive Cancer Network (NCCN)Clinical Practice Guidelines in Oncology on Breast Cancer Version 2. 2020-February 05. 2020. Available from: https://www.nccn.org [accessed 7 February 2020].
28. Martinez-Monge R, Fernandes PS, Gupta N, Gahbauer R. Cross-Sectional nodal atlas: a tool for the definition of clinical target volumes in three-dimensional radiation therapy planning. *Radiology* 1999; **211**: 815–28. doi: https://doi.org/10.1148/radiology.211.3.r99jn40815
29. Saito AI, Vargas C, Morris CG, Lightsey J, Mendenhall NP. Differences between current and historical breast cancer axillary lymph node irradiation based on arm position: implications for radiation oncologists. *Am J Clin Oncol* 2009; **32**: 381–6. doi: https://doi.org/10.1097/COC.0b013e318191718d
30. Klages HT, Szafinski F, Makoski HB. Variation in "supraclavicular" lymph node depth is partly determined by treatment position. *Strahlenther Onkol* 2000; **176**: 315–8. doi: https://doi.org/10.1007/s000660050013
31. Dijkema IM, Hofman P, Raaijmakers CPJ, Lagendijk JJ, Battermann JJ, Hillen B. Loco-Regional conformal radiotherapy of the breast: delineation of the regional lymph node clinical target volumes in treatment position. *Radiother Oncol* 2004; **71**: 287–95. doi: https://doi.org/10.1016/j.radonc.2004.02.017
32. Kirova YM, Servois V, Campana F, Dendale R, Bollet MA, Laki F, et al. Ct-Scan based localization of the internal mammary chain and supra clavicular nodes for breast cancer radiation therapy planning. *Radiother*

*Oncol* 2006; **79**: 310–5. doi: https://doi.org/10.1016/j.radonc.2006.05.014

33. Atean I, Pointreau Y, Ouldamer L, Monghal C, Bougnoux A, Bera G, et al. A simplified CT-based definition of the supraclavicular and infraclavicular nodal volumes in breast cancer. *Cancer Radiother* 2013; **17**: 39–43. doi: https://doi.org/10.1016/j.canrad.2012.11.007
34. Eriksen JG, Salembier C, Rivera S, De Bari B, Berger D, Mantello G, et al. Four years with FALCON - an ESTRO educational project: achievements and perspectives. *Radiother Oncol* 2014; **112**: 145–9. doi: https://doi.org/10.1016/j.radonc.2014.06.017
35. Allozi R, Li XA, White J, Apte A, Tai A, Michalski JM, et al. Tools for consensus analysis of experts' contours for radiotherapy structure definitions. *Radiother Oncol* 2010; **97**: 572–8. doi: https://doi.org/10.1016/j.radonc.2010.06.009
36. Pitkänen MA, Holli KA, Ojala AT, Laippala P. Quality assurance in radiotherapy of breast cancer--variability in planning target volume delineation. *Acta Oncol* 2001; **40**: 50–5. doi: https://doi.org/10.1080/028418601750071055
37. Jameson MG, Holloway LC, Vial PJ, Vinod SK, Metcalfe PE. A review of methods of analysis in contouring studies for radiation oncology. *J Med Imaging Radiat Oncol* 2010; **54**: 401–10. doi: https://doi.org/10.1111/j.1754-9485.2010.02192.x