Radiotherapy and Oncology 124 (2017) 130–138

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

FALCON

# Assessment of the novel online delineation workshop dummy run approach using FALCON within a European multicentre trial in cervical cancer (RAIDs)

Eleonor Rivin del Campo [a,*], Sofia Rivera [a], María Martínez-Paredes [b], Philippe Hupé [c,d,e,f,g], Andrea Slocker Escarpa [h], Isabelle Borget [i,j], Renaud Mazeron [a], Suzy Scholl [k], Amalia Palacios Eito [l], Christine Haie-Meder [a], Cyrus Chargari [a], Eric Deutsch [a]

[a] Department of Radiation Oncology, Gustave Roussy Cancer Campus, Villejuif, France; [b] Department of Radiology and Medical Physics, Medical School, University of Cordoba, Spain; [c] Institut Curie; [d] PSL Research University; [e] Inserm, U900, Paris; [f] Mines Paris Tech, Fontainebleau; [g] CNRS, UMR144, Paris; [h] Department of Radiation Oncology, IUCT Oncopole, Toulouse; [i] Service de Biostatistique et d'Epidemiologie, Gustave Roussy; [j] INSERM U1018, CESP, Université Paris-Sud, Université Paris-Saclay, Villejuif; [k] Department of Medical Oncology, Institut Curie, Paris, France; [l] Department of Radiation Oncology, Reina Sofia University Hospital, Cordoba, Spain

ARTICLE INFO

*Article history:*
Received 29 January 2017
Received in revised form 10 April 2017
Accepted 4 May 2017
Available online 19 May 2017

*Keywords:*
Interobserver variability
Intraobserver variability
Quality assurance
Cervical cancer
Brachytherapy
e-Learning

ABSTRACT

*Background and purpose:* Online delineation workshops (ODW) permit training of geographically dispersed participants. The purpose is to evaluate the methodology of an ODW using FALCON to harmonize delineation within a European multicentre trial on locally advanced cervical cancer (LACC).

*Material and methods:* Two ODW included 46 clinicians (14 centres). Clinicians completed baseline (C1), guideline (C2) and final contours (C3) for external beam radiotherapy (EBRT) and brachytherapy (BT) for LACC. Interobserver and intraobserver variability was evaluated quantitatively (using the DICE index) and qualitatively compared to expert contours.

*Results:* Nine clinicians submitted for EBRT and BT for C1–C3. Thirty-two sent any contour. Interobserver quantitative comparisons for EBRT showed significant improvement for C2 vs. C1 for bowel, CTV node, CTV-p and GTV node with significant detriment for GTV node (C3 vs. C1; C2), CTV-p (C3 vs. C2) and bowel (C3 vs. C2), showing in general an improvement in C2 vs. C1, with a detriment in C3 vs. C2 for two target volumes and an organ at risk. For BT there was significant improvement for C2 vs. C1 for bladder, GTV, HR-CTV and IR-CTV, with significant detriment for bladder (C3 vs. C2), thus overall improvement in C2 vs. C1, with only a detriment in C3 vs. C2 for bladder. Centres using MRI imaging for BT contouring did significantly better in the BT case for HR-CTV than those which used other techniques (C2 vs. C1: $p < 0.005$; C3 vs. C1: $p = 0.02$). Intraobserver quantitative comparisons showed significant improvement contouring a region of interest between C2 vs. C1, C3 vs. C1 and C3 vs. C2 for EBRT and between C2 and C1 for BT.

*Conclusions:* ODW offer training, initial contouring harmonization and allow assessment of centres.

© 2017 Elsevier B.V. All rights reserved. Radiotherapy and Oncology 124 (2017) 130–138

Much has evolved since the first contouring dummy run including distant centres within a multicentre trial, which used CT hard copies [1]. As described in 1995, online education allows participative medical training for geographically dispersed students [2]. Flexibility, essential within e-learning, especially for medical professionals, defined as 'learner control', offers self-task management [3]. Student outcome evaluation is also important, though few report objective internal testing to validate web-based learning tools as a primary outcome [4–7].

Radiotherapy quality assurance has become key to ensure interpretable results within multicentre trials, especially after reports have shown the influence of contouring on patient outcomes [8–11]. Hence the phase III trial of concurrent cisplatin and tirapazamine in head and neck cancer in which when radiotherapy compliance was analysed, a significant reduction of 2 year overall survival and locoregional control was observed when treatment plans were largely deviated from protocol [8].

Proper delineation of target volumes (TV) and organs at risk (OAR) is crucial, allowing optimal oncological treatment and better knowledge of the dose received by surrounding healthy tissue. Thus, several studies have evaluated interobserver and sometimes intraobserver variability between contours [12–15]. Two recent

* Corresponding author at: Department of Radiation Oncology, Gustave Roussy Cancer Campus, 114 Rue Edouard, Vaillant, 94805 Villejuif, France.
*E-mail address:* Eleonor.RIVINDELCAMPO@gustaveroussy.fr (E. Rivin del Campo).

http://dx.doi.org/10.1016/j.radonc.2017.05.008
0167-8140/© 2017 Elsevier B.V. All rights reserved.

reviews addressed this issue, one proposing reporting items for these studies, which this paper will adhere to [16,17]. In locally advanced cervical cancer (LACC) this variability acquires even higher significance. Recent advances in External Beam Radiotherapy (EBRT) and Brachytherapy (BT), namely image guided brachytherapy (IGBT), have shown 3 year local control rates of 92% (tumours > 5 cm) and 98% (tumours 2–5 cm) [18]. This was achieved by applying the Gynaecological GEC-ESTRO (Groupe Européen de Curiethérapie – European Society for Radiotherapy & Oncology) recommendations to the high risk clinical TV (HR-CTV) and dose volume constraints for OAR [19].

The purpose of this study is to validate the methodology of an online delineation workshop (ODW) within a European multicentre prospective study in LACC (Rational molecular Assessments and Innovative Drug Selection: RAIDs), which includes 22 European clinical centres including Eastern and Western Europe [20]. To this aim, participant contours in different periods were reviewed, as well as the participants' personal perception of the knowledge acquired.

## Materials and methods

Before the ODW a general questionnaire about LACC radiotherapy was sent to RAIDs centres for input on their practice (Table 1).

### *ODW structure*

Two to four participants from each centre (proportional to the gynaecological team) were enroled in an ODW in LACC, exceeding its capacity, thus two ODW were planned. A technical partnership was established with ESTRO. The methodology was similar to that used in FALCON (Fellowship in Anatomical deLineation and CONtouring) ESTRO ODW [21]. Live presentations were via WebEx and contouring was done using the FALCON EduCase™ contouring platform.

Training was given by an expert, CHM, with one tutor per 10 clinicians. Tutors were radiation oncologists with experience in LACC, trained to use FALCON EduCase™. Live sessions were completed in 3 weeks and participants delineated EBRT (on Computed Tomography: CT) and subsequent BT (on Magnetic Resonance Imaging: MRI) image sets for the same clinical case. The case and image sets with expert contours were chosen with CHM, from the ESTRO FALCON EduCase™ contouring library.

The ODW were held on June–July 2013 and January 2014, respectively, with an identical structure. The first two live sessions were presented by tutors.

- Session 1 exposed FALCON EduCase™ and the clinical case. Participants were informed (orally and in writing) that their contours would be in a study evaluating the ODW, requesting their conformity, which was not revoked. Clinicians had 6 days for baseline contouring (C1, reflecting daily practice).
- Session 2 presented contouring guidelines for EBRT and BT based on the EMBRACE (An intErnational study on MRI guided BRachytherapy in locally Advanced CErvical cancer) protocol, reviewed baseline contours, and included a question-and-answer session. Recommendations from the Gynaecological GEC-ESTRO working group, EMBRACE protocol, a pelvic nodal atlas and two consensus atlases for pelvic normal tissue were sent to clinicians to aid delineation [19,22–25]. They had 2 weeks to modify contours for the same image sets (guideline contouring: C2).
- In session 3 CHM reviewed baseline and guideline contours and held a question-and-answer session.

Lastly, clinicians performed final contouring (C3) for EBRT and BT 1.5–2 months after session 3, to evaluate the long term teaching impact.

### *Clinical case*

A forty-five year old patient with a FIGO IIIB squamous cell CC was studied. Gynaecological exam: large growth (85x50x60 mm) involving the vagina (all fornices 1 cm, anterior vaginal wall 4 cm). The right parametrium had proximal infiltration, the left one until pelvic side wall. Bladder mucosa was not involved. Abdominopelvic CT showed CC with vaginal involvement, enlarged external, internal, lower common iliac, and pre-sacral nodes. No paraaortic nodes. The response to EBRT and concomitant chemotherapy was good: tumour dimensions of 55x40x30 mm, free right parametrium, induration of half of the left parametrium, and involvement of 1 cm of the anterior vaginal wall at the time of BT.

- Volumes required for contouring exercises (at least specified slices for OAR and whole ROI for TV):
  - EBRT:
    - OAR: Bladder, rectum, bowel, sigmoid.
    - GTV-P (gross tumour volume-P): Cervix, parametria and vaginal gross disease.
    - CTV-nodes: Nodal elective volume.
    - GTV node: Radiologically pathological lymph nodes (to boost).
    - CTV-P: GTV-P, uterus and vagina ($\geq$20 mm below GTV-P).
  - BT:
    - OAR: Bladder, rectum, sigmoid.
    - GTV: Macroscopic tumour at BT.
    - HR-CTV: Macroscopic tumour at BT + whole cervix + presumed extra-cervical tumour extension.
    - IR-CTV (intermediate risk CTV): HR CTV + GTV at diagnosis + $\geq$10 mm margin to residual disease at time of brachytherapy towards potential spread.

### *Contour evaluation methodology*

Intraobserver variability was evaluated between C2 vs. C1, C3 vs. C2 and C3 vs. C1, for EBRT and BT treatments, quantitatively and qualitatively.

Interobserver variability was determined quantitatively by analyses centred on regions of interest (ROI) and on years of experience, and for BT also between centres that used MRI-based IGBT and others.

Contours were quantitatively classified by DICE scores [DICE = 2 $\times$ ($Volume_{expert} \cap Volume_{participant}$)/($Volume_{expert}$ + $Volume_{participant}$)] given by FALCON EduCase™ Output [26]:

DICE references for TV [27,28]:

A: Optimal: >0.81
B: Average: 0.65–0.81
C: Suboptimal: <0.65

DICE references for OAR [29]:

A: Optimal: >0.81
B: Suboptimal: $\leq$0.81

In MRI-based brachytherapy for cervical cancer, Dimopoulos et al. defined a range of 0.5–0.7 using the conformity index for target volumes, which when converted to DICE is roughly 62.5–0.81

**Table 1**
Preworkshop questionnaire reflecting daily practice in each centre.

| Centre | Centre Type | # Pat. RCT/Yr. | Type C. Ca. Pat. Treated | BT Applicator | Do Interstitial BT | Dose Rate | BT Imaging Type | Plan To Start 3D IGBT? | BT Prescription | Dose TV | # Fract. | Constraints |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Centre 1* | Academic | <50 | All pat. S.I–IVa, M0 | Ovoids; Interstitial N. | YES | PDR | CT; MRI; US | NA | HR-CTV | 80–84 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 2* | Public | 50–100 | All pat. S.I–IVa, M0 | Ovoids | NO | HDR | CBCT | In 3y | PointA | 65–74 Gy | 4 | Bladder Rectum |
| Centre 3* | Academic; Public | >100 | All pat. S.I–IVa, M0 | Tandem/ring; Ovoids; Tandem/cylinder | NO | HDR | X-ray | NO | PointA | 65–74 Gy | 4 | Bladder Rectum |
| Centre 4* | Academic | <50 | All pat. S.I–IVa, M0 | Ovoids | NO | PDR | CT; MRI | NA | PointA; HR-CTV | 65–74 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 5* | Academic; Public | >100 | All pat. S.I–IVa, M0 | Ovoids | NO | HDR | CT; MRI | NA | HR-CTV | 75–79 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 6* | Academic | <50 | Pos. Pelv./PA LN | Ovoids | YES | PDR | CT; MRI | NA | HR-CTV | 80–84 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 7* | Private | 50–100 | All pat. S.I–IVa, M0 | Ovoids; Interstitial N. | YES | PDR; LDR | CT; MRI; US | NA | HR-CTV | <65 Gy; 65–74 Gy | - | Bladder Rectum Sigmoid |
| Centre 8* | Academic | <50 | Pos. Pelv./PA LN; Neg.LN S. > IIB | Ovoids; Mould; Interstitial N. | YES | PDR | CT; MRI | NA | HR-CTV | <65 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 9* | Public | <50 | Pos. Pelv./PA LN; Neg.LN S. > IIB | Tandem/ring; Mould | NO | PDR | CT; US | NA | HR-CTV | 80–84 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 10* | Academic; Public | <50 | Pos. Pelv./PA LN; Neg.LN S. > IIB | Tandem/ring; Ovoids | NO | PDR | CT; US | NA | PointA; HR-CTV; IR-CTV | 80–84 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 11 | Public | <50 | Pos. PA LN; Neg.LN S. > IIB | Mould | NO | LDR | CT; MRI | NA | IR-CTV | 65–74 Gy | 2 | Bladder Rectum |
| Centre 12* | Public | <50 | All pat. S.I–IVa, M0 | Ovoids | NO | PDR | CT | NA | HR-CTV | 65–74 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 13* | Academic | 50–100 | Pos. Pelv./PA LN; Neg.LN S. > IIB | Ovoids | NO | PDR | CT; MRI | NA | HR-CTV; IR-CTV | <65 Gy; 80–84 Gy | 2 | Bladder Rectum Sigmoid Vagina |
| Centre 14* | Academic | >100 | All pat. S.I–IVa, M0 | Ovoids; Mould; Interstitial N. | YES | PDR | MRI | NA | HR-CTV | >85 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 15 | Public | <50 | Pos. Pelv./PA LN | Ovoids | YES | HDR | MRI | NA | HR-CTV | >85 Gy | 3 | Bladder Rectum Sigmoid |
| Centre 16 | Academic | 50–100 | All pat. S.I–IVa, M0 | Ovoids | YES | PDR | MRI | NA | HR-CTV | 75–79 Gy; 80–84 Gy | 2 | Bladder Rectum Sigmoid |
| Centre 17* | Academic | >100 | All pat. S.I–IVa, M0 | Ovoids; Tandem/cylinder | NO | HDR | CT | NA | PointA; HR-CTV | 80–84 Gy | 2; 3 | Bladder Rectum |

The seventeen centres listed below answered this questionnaire, of which the 14 marked with an asterisk participated in the ODW (the order of the centres does not correspond with the order in the anonymous table).
French centres: Institut Curie*, Gustave Roussy*, Centre Georges François Leclerc*, Centre Léon Bérard*, Centre Alexis Vautrin*, CHU Anne de Bretagne*, Centre Jean Perrin*, Institut Bergonié*, Institut de Cancérologie de l'Ouest*; Centre Oscar Lambret*, Tenon Hospital, Institut Claudius Regaud.
The Netherlands: Academic Medical Centre (AMC)*, Antoni van Leewenhoek Hospital (NKI).
Moldova: Institute of Oncology (IOM)*.
Romania: Emergency County Hospital Oradea*.
Serbia: Institut za onkologiju Vojvodine (IOV)*.
*Abbreviations:* RCT: Radiochemotherapy; yr.: year; C.Ca.: Cervical cancer; Pat.: patients; BT: Brachytherapy; IGBT: Image guided brachytherapy; TV: Target Volume; Fract.: Fractions; S.: Stage; Pos.: Positive; Pelv.: Pelvic; PA: Para-aortic; LN: Lymph Nodes; N.: Needles; LDR: Low Dose Rate; PDR: Pulsed Dose Rate; HDR: High Dose Rate; CT: Computed Tomography scan; MRI: Magnetic Resonance Imaging; CBCT: Cone Beam CT scan; US: Ultrasound; NA: Not applicable; HR-CTV: High Risk CTV; IR-CTV: Intermediate Risk CTV; Gy: Gray.

**Table 2**
Participants enroled in the online delineation workshops which submitted contours for each contouring period.

| | EBRT | | | BT | | |
|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C1 | C2 | C3 |
| Submission of ≥ 1 contour | 28 | 22 | 13 | 30 | 21 | 13 |
| Submission of all contours (OAR and TV) | 14 | 11 | 5 | 24 | 15 | 6 |
| Submission of only TV | 19 | 15 | 7 | 24 | 15 | 9 |
| *Total number of participants (n)* | | | 46 | | | 46 |
| PARTICIPANT POPULATION WHICH SUBMITTED CONTOURS | | | | | | |
| Experienced specialists (>5 years of post-residency experience) | 13 | | | | | |
| Less experienced specialists (≤5 years of post-residency experience) | 8 | | | | | |
| Senior residents (>2 years of experience) | 7 | | | | | |
| Junior residents (≤2 years of experience) | 4 | | | | | |

The participant population which submitted contours, by level of experience.
*Abbreviations:* EBRT: External beam radiotherapy; BT: Brachytherapy; OAR: Organs at risk; TV: Target Volumes.

[27,30]. For OAR, Breunig et al. found an average DICE of 0.61 for volumes <8 cc and of 0.91 for volumes >8 cc, averaging at 0.76 [29]. To simplify the cutoffs and make the study easier to interpret, 0.65 and 0.81 were chosen for TV and 0.81 for OARs. Of note, all statistical analyses performed were independent of the thresholds that were only used to aid interpretation and to display the results.

For the objective qualitative intraobserver assessment, the EduCase™ contour error distance tool showed on axial slices where the participant contour was 3 mm larger or smaller than the expert contour, based on the scalar assessment in the transverse plane for HR-CTV by Petric et al., in 8 directions (anterior, posterior, right, left, anterolateral right and left and posterolateral right and left) to detect the most prevalent areas of uncertainties [13].

Qualitative Classification:

- "Correct": Participant contour ≤3 mm smaller/larger than the expert contour in a given direction.
- "Incorrect": Participant contour >3 mm smaller/larger than the expert contour in a given direction without a probable clinical impact.
- "Very incorrect": Participant contour >3 mm smaller/larger than the expert contour in a certain direction which for that particular ROI will have a probable clinical impact (worse coverage of TV/ higher dose to OAR).

As part of the outcome of e-learning courses depends on participant perception, an anonymous satisfaction questionnaire adapted from FALCON-ESTRO ODW was administered to clinicians (Appendix 3).

### *Statistical analyses*

DICE scores have been transformed using the *logit* function, logit(x)=x/(1-x), so they asymptotically follow a Gaussian distribution [31,32].

To assess interobserver variability, a linear mixed model (**ModelINTER.PART)** was used, with the fixed effects *ROI*, *contouring period*, *experience*, their interactions, the linear and quadratic effects of the *slice* and their interactions with the *ROI* (for BT: the effect *imaging technique* and the interaction *ROI*imaging technique* was added), and the random effect of interparticipant variability, considered different for OAR and TV. To assess intraobserver variability (difference of DICE scores between contouring periods), a paired comparison by participant and ROI was performed using a linear model (**modelINTRA.PART**) with the fixed effect *participant*, *ROI* and their interaction (*participant*ROI*). The average DICE score by participant and ROI for each contouring period was assessed by a similar model (**modelSCORE.PART**). The significance of the fixed effects was computed using Fisher's test for all models. The proportion of pairs *participant*ROI* declared as performing better (or worse) from **modelINTRA.PART**, and their association with other covariates (experience, ROI type, institution or imaging technique) were assessed with Fisher's exact test. For the qualitative analysis, to compare the proportions of correct contours between different contouring periods, we used the test of McNemar [33]. The statistical analysis was performed using R software (R Core Team, 2016) and can be automatically reproduced using the scripts and data in Supplementary information.

## Results

### *Participant population*

Participants from 14 of 22 RAIDs centres submitted contours (Table 1).

Of the 46 enroled participants, nine submitted delineations for all contouring periods for EBRT and BT (Table 2). The description by level of experience of the participant population which submitted contours is in Table 2.

There is no significant relationship between the participants who dropped out after C1 with the initial DICE scores on C1 (whether they were low or high) nor with the years of experience or centre (Appendix 1: Tables 26–28; Appendix 2: Tables 29–31).

The adequacy of the cutoff points for this study was confirmed by the distribution of the pooled data (for EBRT and BT over all contouring attempts). The first quartile for OAR is 0.8 and for TV the first quartile is 0.6 and the third quartile is 0.86, which is mostly consistent with the chosen cutoff points (0.65 and 0.81).

All of the results of all models per contouring period are summarized in Table 3.

### *Results of ModelINTER.PART (interobserver variability)*

All interactions were highly significant ($p < 0.001$), all effects had an impact on DICE scores (Table 2 in Appendix 1, 2). The model captures the quadratic relationship between DICE score and slice number (Fig. 1).

Pairwise comparisons for EBRT and BT between contouring periods by ROI are reported in Table 2 (details in Appendices 1–2, Table 5). For both EBRT and BT in C2 vs. C1 there was a significant improvement, mostly for TV, with no significant decrease. For C3 vs. C1 in EBRT and BT there was also a significant increase observed in certain TV with only a significant decrease for GTV node. However, in C3 vs. C2 for both image sets there was a significant decrease for 2 TV in EBRT and 2 OAR (no decrease for TV in BT), with a significant increase for CTV node.

For EBRT (Appendix 1, Table 6), regarding the experience effect, experienced specialists performed significantly better than junior

**Table 3**
Results for the interobserver and intraobserver quantitative and qualitative analyses. All results reported were statistically significant $p < 0.05$.

| | EBRT | | | BT | | |
|---|---|---|---|---|---|---|
| INTEROBSERVER QUANTITATIVE | | | | | | |
| | C2 vs. C1 | C3 vs. C1 | C3 vs. C2 | C2 vs. C1 | C3 vs. C1 | C3 vs. C2 |
| Comparisons between contouring periods by ROI* | ↑Bowel<br>↑CTV node<br>↑CTV-p<br>↑GTV node | <br>↑CTV node<br><br>↓GTV node<br>↑GTV-p | ↓Bowel<br>↑CTV node<br>↓CTV-p<br>↓GTV node | ↑Bladder<br>↑GTV<br>↑HR-CTV<br>↑IR-CTV | <br><br>↑HR-CTV<br>↑IR-CTV | ↓Bladder |
| Comparisons between experience by ROI* | Sigmoid: Exp. Spec. > Junior Res.<br>Less exp. Spec. > Junior Res.<br>Less exp. Spec. > Senior Res.<br>GTV node: Exp. Spec. < Less exp. Spec.<br>Exp. Spec. < Senior Res.<br>Less exp. Spec. > Junior Res.<br>Senior Res. > Junior Res.<br>GTV-p: Exp. Spec. < Junior Res.<br>Junior Res. > Senior Res. | | | Sigmoid: Exp. Spec. > Senior Res. | | |
| Comparison between imaging techniques* | | | | HR-CTV: C. MRI-IGBT > Others | | |
| INTRAOBSERVER QUANTITATIVE | | | | | | |
| | C2 vs. C1 | C3 vs. C1 | C3 vs. C2 | C2 vs. C1 | C3 vs. C1 | C3 vs. C2 |
| Do participants improve between contouring periods? | Yes | Yes | Yes | Yes | | |
| Is the improvement associated to ROI.type (TV/OAR)? | Yes | | | Yes | | |
| Is the improvement associated to institution? | Yes | | | | | |
| INTRAOBSERVER QUALITATIVE (comparison of the % of correct contours between contouring periods) | | | | | | |
| TV Posterolat. right | ↑ | | | ↑ | ↓ | ↓ |
| TV Anterolat. right | | ↓ | ↓ | | | |
| TV Posterior | | | ↓ | | | |
| TV Posterolat. left | | | ↓ | | | |
| TV Right | | | | ↑ | | |
| OAR Posterolat. right | | | | | ↑ | |

*Abbreviations:* EBRT: External beam radiotherapy; BT: Brachytherapy; C1: baseline contouring; C2: guideline contouring; C3: final contouring; exp.: experience; Spec.: Specialist; Res.: Resident; C. MRI-IGBT: centres using MRI-image guided BT; Posterolat. : Posterolateral; Anterolat. : Anterolateral.
↑: Improvement; ↓: Decrease.
* $p < 0.05$ after correction for multiple testing FDR.

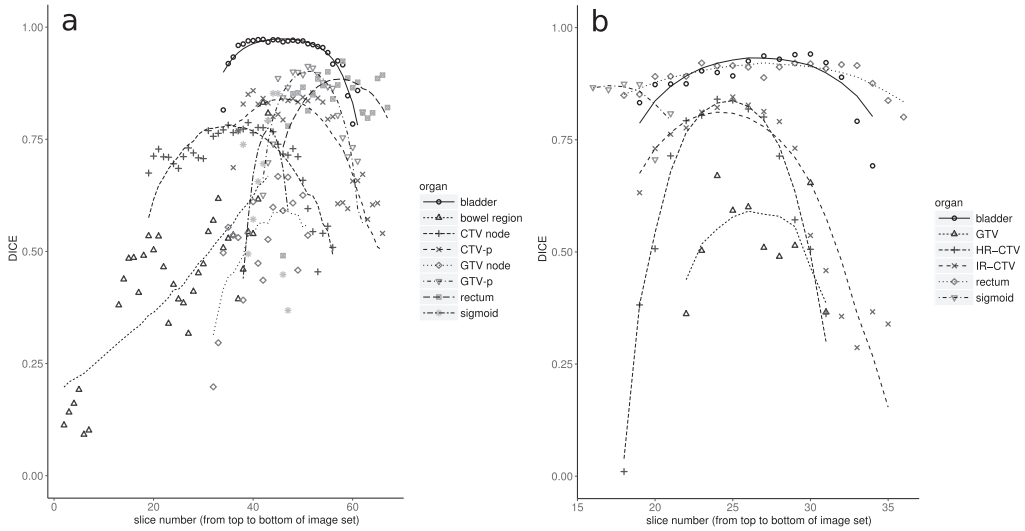


**Fig. 1.** Means of DICE scores (1 indicating perfect concordance between participant and expert; 0 indicating no concordance) by ROI according to slice number. The lines represent the mean of the DICE score predicted by the mixed model, capturing the parabolic effect of the slice number on the DICE scores. (a) Means of DICE scores by ROI according to slice number for EBRT. (b) Means of DICE scores by ROI according to slice number for BT.

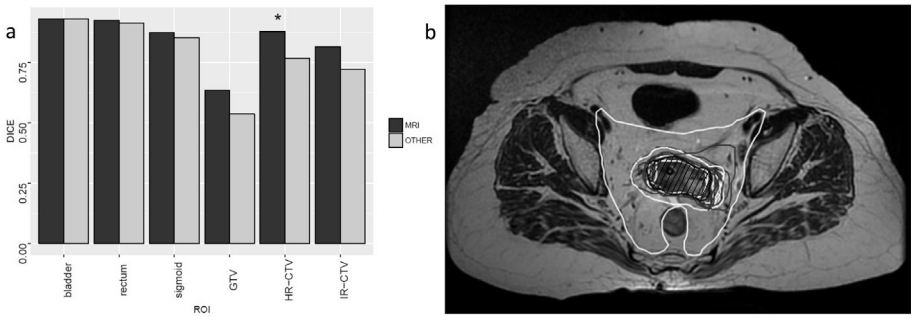


**Fig. 2.** (a) The interaction *ROI*imaging technique* in the BT treatment for centres using MRI-IGBT (black) and those not using it (light grey) (**ModelINTER.PART**). (b) Baseline contours (C1) of GTV for centres using MRI-IGBT (dark grey) and those not using it (white). Expert contour in black, with diagonal lines.

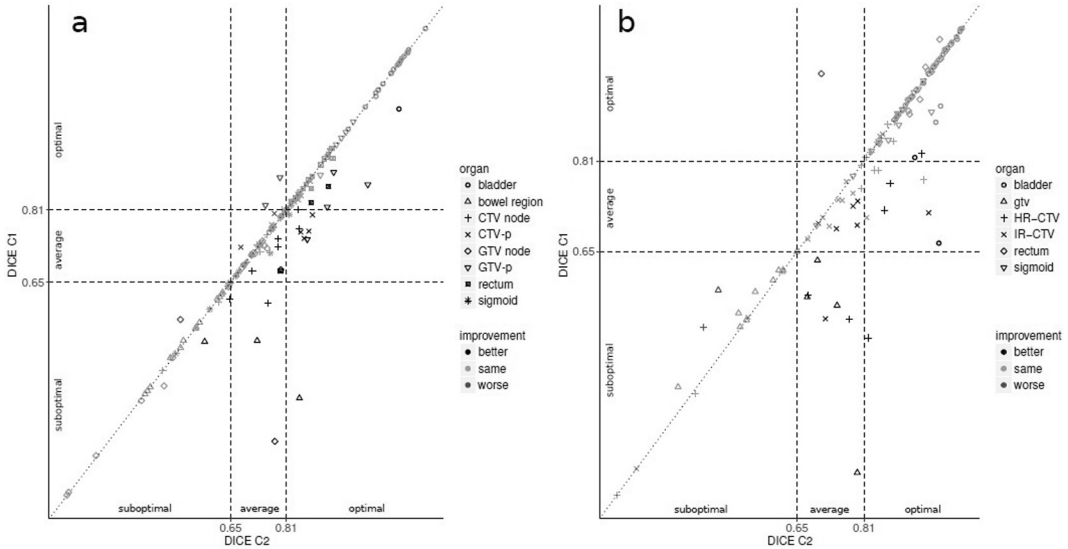


**Fig. 3.** The average score for each participant and each ROI for the C2 (*x*-axis) and C1 (*y*-axis) for (a) EBRT and (b) BT estimated from the **modelSCORE.PART** and whether the delineation improved or not (the difference is significantly better, equal or worse, from **modelINTRA.PART**, intraobserver variability). Examples (Ex.): Ex. 1.: This participant's DICE index did not vary significantly (same) between C2 vs. C1 for GTV, staying within the suboptimal category. Ex. 2.: This participant's DICE index was in significant detriment (worse) between C2 vs. C1 for rectum, changing from the optimal to the average category. Ex. 3.: This participant's DICE index improved significantly (better) between C2 vs. C1 for IR-CTV, changing from the average to the optimal category.

residents for sigmoid and significantly worse than less experienced specialists and senior residents for GTV node and than junior residents for GTV-p. Less experienced specialists did significantly better than junior residents for GTV node and sigmoid, and than senior residents for sigmoid. Between senior and junior residents there were only significant differences for GTV node and GTV-p. For BT the only significant difference was that experienced specialists performed better than senior residents for sigmoid (Table 3; Appendix 2, Table 6).

Regarding the imaging technique, centres that used MRI based IGBT did significantly better than those which used other techniques (CT, X-ray, US) for HR-CTV (Fig. 2, Table 3).

The ICC (intraclass correlation) for interobserver variability was excellent for OAR in BT (0.92; 95% CI: 0.86–0.96), OAR in EBRT (0.96; 95% CI: 0.93–0.98) and TV in EBRT (0.78–95% CI: 0.68–0.88) while it was fair for TV in BT (0.51; 95% CI: 0.39–0.68) (Table 4 in Appendix 1, 2). The low ICC for TV in BT highlights the difficulty of participants to agree on contours, whether they usually contour on MRI or not (the imaging technique was taken into account in modelINTER.PART).

*Results of modelINTRA.PART and modelSCORE.PART (intraobserver variability)*

Fig. 3 (Fig. 7, Appendixes 1 and 2) represents the average score for each participant and each ROI for C2 and C1 for EBRT and BT estimated from the **modelSCORE.PART** and whether the difference is significantly better, equal or is worse (from **modelINTRA.PART**). The Fisher's exact tests show that participants improved significantly between all contouring periods for EBRT (Appendix 1: Tables 8, 13 and 18). For BT, participants improved significantly between C2 vs. C1 (Appendix 2, Table 8). For EBRT, the improvements were significantly associated to ROI.type (TV vs. OAR) and institutions (C2 vs. C1). For BT, the improvements are significantly associated to ROI.type between C2 vs. C1 (Table 3). Interestingly, the number of participants who performed worse between different contouring periods was never significant (Fisher's exact test).

*Results of qualitative data (intraobserver variability)*

For EBRT, the percentage of "correct" contours was only significantly better between C2 vs. C1 for posterolateral right in TV. It was significantly worse in TV for anterolateral right for C3 vs. C2 and in three directions for C3 vs. C2 (Table 3; Appendix 1, Fig. 15).

For BT, the percentage of "correct" contours was significantly better between C2 vs. C1 for posterolateral right and right in TV and between C3 vs. C1 for posterolateral right in OAR. It only was significantly worse between C3 vs. C1 and C2 for posterolateral right in TV (Table 3; Appendix 2, Fig. 15).

*Results of the satisfaction questionnaire*

The scores over the 20 Organization and Content items for the 20 participants who responded of the 32 that submitted contours, on a scale of 1–5, 5 being excellent, range from 3.95 to 4.60 with an average of 4.358 (Table 3, Appendix 3). When asked whether they would attend another online workshop, 80% of participants answered affirmatively, and 85% would recommend one.

## Discussion

For the first time this recent modality of ODW has been used for assessment of contouring skills as a dummy run within a multicentre trial [21]. Recently, other trials have used ODW within their quality assurance programmes, such as HYPO-G-01 [34]. Other authors, like Fokas et al., advocated training programmes within radiotherapy quality assurance protocols [35]. Our results have shown the ODW feasibility and capability in identifying centres that manifest baseline and subsequent average to optimal contours and are ready to include patients, while offering an effective educational tool for others. The added value of this study is that it reports the participants' point of view, which in light of the post-ODW satisfaction questionnaire results is extremely favourable.

Petric et al. have described graphically how the largest uncertainties in contouring are on the cranial and caudal slices of a volume, which coincides with our results (Fig. 1; Fig. 5 Appendixes 1 and 2) [15]. The bowel follows a particular pattern since the expert contour consisted of individual bowel loops and most participants contoured a bowel bag, but both contouring techniques are valid [36].

An interesting aspect of this study is that the evaluation of interobserver variability allowed assessment of overall improvement/detriment of the participants in the workshop group between them, and not only individual variability versus the expert contour (which affects the comparison of ROI, and has certain flaws) [16,37]. As could be expected, for interobserver comparisons in both EBRT and BT, there was overall more improvement between C2 and C1 than between C3 and C1, and the worse results were mostly between C3 and C2. This suggests that participants gained contouring skills after presentation of the guidelines, and retained part of this knowledge 1.5–2 months later. However, from the intraobserver point of view, only improvements were significant between contouring periods.

When considering interobserver variability with respect to experience in EBRT, experienced specialists did significantly worse than less experienced specialists and senior residents for GTV node. This finding is to be interpreted with caution, as there was a borderline significant paraaortic lymph node (though the clinical case states: 'no positive paraaortic lymph nodes') and a suspicious lymph node in the left groin, deemed as inflammatory by CHM during live sessions. Thus this may simply highlight that less experienced specialists and senior residents were more focused on the clinical information provided. Logically, less experienced specialists, with more experience, did better than junior residents for GTV node and than junior and senior residents for sigmoid, as senior residents did better than junior residents for GTV node (all significant). Surprisingly, junior residents did significantly better than senior residents for GTV-p. For BT the only significant difference was experienced specialists which contoured the sigmoid better than senior residents.

Concerning MRI guided IGBT, MRI allows better visualization of the vagina and uterus than of the rectum or bladder [38]. This may explain our results of interobserver improvement in HR-CTV and IR-CTV for C2 and C3 vs. C1, as opposed to a detriment for the bladder (C3 vs. C2). It is also interesting to note the significant improvement for contouring of HR-CTV for centres doing MRI-IGBT, showing the impact of specific training in MRI-based contouring.

Qualitatively, Petric et al. did not find significant interobserver differences along the 8 directions of space for HR-CTV contours [13]. Our intraobserver differences were significant for certain directions (better or worse) for EBRT, with no obvious explanation as there was no clear clinical impact due to these differences. However, for BT, the significant improvement towards the right and posterolateral right for TV in C2 vs. C1 most probably is because the left parametrial invasion made participants focus more on the left portion of the TV than on the right during C1, and they improved after the guideline session. But they went back to their old ways in C3, doing significantly worse for posterolateral right TV in C3 vs. C1 and C2.

Considering the e-learning educational experience, this ODW allows a self-directed path, each clinician may attend live online sessions, within a blended learning model (with support and interaction with tutors) or follow recordings. This flexibility adapted to the physicians' heavy workload [3]. But this was not always effective, 14 of the initial 46 enroled participants did not submit contours.

Initial limitations of this study were organizational: difficulties to locate radiotherapy professionals, as the ODW was performed before opening RAIDs in clinical centres. Further, the first ODW was held in June–July. Many clinicians could not participate, or only could attend some sessions, with fewer contour sets submitted during July (C2), and August–September (C3). Thus, only 13 contours were submitted for C3, limiting statistical significance and with a less representative population. Another limitation was the use of the DICE index, the only contouring conformity index available as FALCON EduCase™ output at the time. It is less reliable in small ROI volumes, such as GTV node, GTV or sigmoid in our study, showing lower concordance simply because slight delineation discrepancies have more impact on the score. Conversely, in very large ROI volumes, as CTV node or bowel, it seems to lack the sensitivity to identify divergences from the reference contour (Fig. 1) [29,39]. This is due to the duplication of the overlapping volume, which may inaccurately show considerable agreement in these large ROI. The strongpoint of this index is the simplicity of calculation, it is the most used in automatic segmentation studies [40]. It may also be converted into other concordance indexes using certain ratios [30].

In conclusion, ODW provide feasible and convenient means for initial assessment of contouring practices in geographically dispersed centres, as well as additional training in contouring within the setting of quality control for a multicentre trial. Future studies should focus on improving this training, and developing the optimal sequence of further training for centres which need more improvement (further online training, or combined with specific onsite programmes).

## Conflict of interest statement

The first author of this manuscript has received a DUERTECC/EURONCO grant.

There are no other financial and personal relationships to disclose, on behalf of all authors.

## Acknowledgements

"This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 304810".

"Acting the recipient for the grant for DUERTECC/EURONCO (Diplôme Universitaire Européen de Recherche Translationnelle Et Clinique en Cancérologie)".

The RAIDs consortium, and particularly the participants of the ODW from the RAIDs centres:: Institut Curie, Gustave Roussy, Centre Georges François Leclerc, Centre Léon Bérard, Centre Alexis Vautrin, CHU Anne de Bretagne, Centre Jean Perrin, Institut Bergonié, Institut de Cancérologie de l'Ouest; Centre Oscar Lambret, Academic Medical Centre (AMC), Institute of Oncology, Moldova (IOM), Emergency County Hospital Oradea and the Institut za onkologiju Vojvodine (IOV).

ESTRO for granting us access to FALCON EduCase™ and technical support (Miika Palmu, Christine Verfaille) and Scott Kaylor and Arthur Boyer from EduCase™ (RadOnc eLearning Center, Inc. Fremont, CA, USA) for collaborating by incorporating a development within EduCase™.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.radonc.2017.05.008.

## References

[1] Seddon B, Bidmead M, Wilson J, Khoo V, Dearnaley D. Target volume definition in conformal radiotherapy for prostate cancer: quality assurance in the MRC RT-01 trial. Radiother Oncol 2000;56:73–83. http://dx.doi.org/10.1016/S0167-8140(00)00191-2.

[2] McEnery KW, Roth SM, Kelley LK, Hirsch KR, Menton DN, Kelly EA. A method for interactive medical instruction utilizing the World Wide Web. Proc Symp Comput Appl Med Care 1995:502–7.

[3] Booth A, Carroll C, Papaioannou D, Sutton A, Wong R. Applying findings from a systematic review of workplace-based e-learning: Implications for health information professionals: review article. Health Info Libr J 2009;26:4–21. http://dx.doi.org/10.1111/j.1471-1842.2008.00834.x.

[4] Kronz JD, Silberman MA, Allsbrook WC, Epstein JI. A web-based tutorial improves practicing pathologists' Gleason grading of images of prostate carcinoma specimens obtained by needle biopsy: validation of a new medical education paradigm. Cancer 2000;89:1818–23.

[5] Ridgway PF, Sheikh A, Sweeney KJ, Evoy D, McDermott E, Felle P, et al. Surgical e-learning: validation of multimedia web-based lectures. Med Educ 2007;41:168–72. http://dx.doi.org/10.1111/j.1365-2929.2006.02669.x.

[6] Foroudi F, Pham D, Bressel M, Tongs D, Rolfo A, Styles C, et al. Learning methods in radiation oncology The utility of e-Learning to support training for a multicentre bladder online adaptive radiotherapy trial (TROG 10.01-BOLART). Radiother Oncol 2013;109:165–9. http://dx.doi.org/10.1016/j.radonc.2012.10.019.

[7] Pham D, Hardcastle N, Foroudi F, Kron T, Bressel M, Hilder B, et al. A multidisciplinary evaluation of a web-based elearning training programme for SAFRON II (TROG 13.01): a multicentre randomised study of stereotactic radiotherapy for lung metastases. Clin Oncol 2016. http://dx.doi.org/10.1016/j.clon.2016.03.005.

[8] Peters LJ, Giralt J, Fitzgerald TJ, Trotti A, Bernier J, Bourhis 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 n. d.; 28:2996–3001. doi:http://dx.doi.org/10.1200/JCO.2009.27.4498.

[9] Weber DC, Tomsej M, Melidis C, Hurkmans CW. QA makes a clinical trial stronger: evidence-based medicine in radiation therapy. Radiother Oncol 2012;105:4–8. http://dx.doi.org/10.1016/j.radonc.2012.08.008.

[10] Abrams RA, Winter KA, Regine WF, Safran H, Hoffman JP, Lustig R, et al. Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704-a phase iii trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas. Int J Radiat Oncol Biol Phys 2012;1:809–16. http://dx.doi.org/10.1016/j.ijrobp.2010.11.039.

[11] Ohri N, Shen X, Dicker AP, Doyle LA, Harrison AS, Showalter TN. Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials. J Natl Cancer Inst 2013;105:387–93. http://dx.doi.org/10.1093/jnci/djt001.

[12] Rasch C, Barillot I, Remeijer P, Touw A, Van Herk M, Lebesque JV. Definition of the prostate in CT and MRI: a multi-observer study. Int J Radiat Oncol Biol Phys 1999;43:57–66. http://dx.doi.org/10.1016/S0360-3016(98)00351-4.

[13] Petric P, Dimopoulos J, Kirisits C, Berger D, Hudej R, Pötter R. Inter- and intraobserver variation in HR-CTV contouring: Intercomparison of transverse and paratransverse image orientation in 3D-MRI assisted cervix cancer brachytherapy. Radiother Oncol 2008;89:164–71. http://dx.doi.org/10.1016/j.radonc.2008.07.030.

[14] van Mourik AM, Elkhuizen PHM, Minkema D, Duppen JC, van Vliet-Vroegindeweij C. Multiinstitutional study on target volume delineation variation in breast radiotherapy in the presence of guidelines. Radiother Oncol 2010;94:286–91. http://dx.doi.org/10.1016/j.radonc.2010.01.009.

[15] Petrič P, Hudej R, Rogelj P, Blas M, Tanderup K, Fidarova E, et al. Uncertainties of target volume delineation in MRI guided adaptive brachytherapy of cervix cancer: a multi-institutional study. Radiother Oncol 2013;107:6–12. http://dx.doi.org/10.1016/j.radonc.2013.01.014.

[16] 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. http://dx.doi.org/10.1016/j.radonc.2016.09.009.

[17] 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 2016;60:393–406. http://dx.doi.org/10.1111/1754-9485.12462.

[18] Pötter R, Georg P, Dimopoulos JCA, Grimm M, Berger D, Nesvacil N, et al. Clinical outcome of protocol based image (MRI) guided adaptive brachytherapy combined with 3D conformal radiotherapy with or without chemotherapy in patients with locally advanced cervical cancer. Radiother Oncol 2011;100:116–23. http://dx.doi.org/10.1016/j.radonc.2011.07.012.

[19] Haie-Meder C, Pötter R, Van Limbergen E, Briot E, De Brabandere M, Dimopoulos J, et al. Recommendations from Gynaecological (GYN) GEC-ESTRO Working Group (I): Concepts and terms in 3D image based 3D treatment planning in cervix cancer brachytherapy with emphasis on MRI assessment of GTV and CTV. Radiother Oncol 2005;74:235–45. http://dx.doi.org/10.1016/j.radonc.2004.12.015.
[20] Rational molecular Assessments and Innovative Drug Selection: RAIDs. n.d. http://www.raids-fp7.eu/ [accessed December 9, 2016].
[21] Grau Eriksen J, Salembier C, Rivera S, De Bari B, Berger D, Mantello G, et al. ESTRO FALCON e-learning Four years with FALCON – an ESTRO educational project: achievements and perspectives on behalf of ESTRO. Radiother Oncol 2014;112:145–9. http://dx.doi.org/10.1016/j.radonc.2014.06.017.
[22] Taylor A, Rockall AG, Reznek RH, Powell MEB. Mapping pelvic lymph nodes: Guidelines for delineation in intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys 2005;63:1604–12. http://dx.doi.org/10.1016/j.ijrobp.2005.05.062.
[23] Gay HA, Barthold HJ, O'Meara E, Bosch WR, El Naqa I, Al-Lozi R, et al. Pelvic normal tissue contouring guidelines for radiation therapy: a radiation therapy oncology group consensus panel atlas. Int J Radiat Oncol Biol Phys 2012;83. http://dx.doi.org/10.1016/j.ijrobp.2012.01.023.
[24] Gay HA, Barthold HJ, O'Meara E et al. Female Pelvis Normal Tissue RTOG Consensus Contouring Guidelines. n.d. https://www.rtog.org/LinkClick.aspx?fileticket=P5eAjYB90Ow%3D&tabid=355 [accessed December 9, 2016].
[25] An IntErnational study on MRI-guided BRAchytherapy in locally advanced CErvical cancer: EMBRACE. n.d. https://www.embracestudy.dk/.
[26] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302. http://dx.doi.org/10.2307/1932409.
[27] Dimopoulos JCA, De Vos V, Berger D, Petric P, Dumas I, Kirisits C, et al. Inter-observer comparison of target delineation for MRI-assisted cervical cancer brachytherapy: application of the GYN GEC-ESTRO recommendations. Radiother Oncol 2009;91:166–72. http://dx.doi.org/10.1016/j.radonc.2008.10.023.
[28] Petersen RP, Truong PT, Kader HA, Berthelet E, Lee JC, Hilts ML, 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. http://dx.doi.org/10.1016/j.ijrobp.2007.01.070.
[29] Breunig J, Hernandez S, Lin J, Alsager S, Dumstorf C, Price J, et al. A system for continual quality improvement of normal tissue delineation for radiation therapy treatment planning. Int J Radiat Oncol Biol Phys 2012;83:e703–8. http://dx.doi.org/10.1016/j.ijrobp.2012.02.003.
[30] Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R, Georg D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. Strahlenther Onkol 2012;188:160–7. http://dx.doi.org/10.1007/s00066-011-0027-6.
[31] Agresti A. Categorical Data Analysis. vol. 359. 2002. doi:http://dx.doi.org/10.1002/0471249688.
[32] Brock KK. Image processing in radiation therapy. CRC Press; 2013.
[33] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153–7. http://dx.doi.org/10.1007/BF02295996.
[34] HYPOG-01 n.d. https://www.gustaveroussy.fr/en/node/3646 [accessed January 8, 2017].
[35] Fokas E, Clifford C, Spezi E, Joseph G, Branagan J, Hurt C, et al. Comparison of investigator-delineated gross tumor volumes and quality assurance in pancreatic cancer: analysis of the pretrial benchmark case for the SCALOP trial. Radiother Oncol 2015;117:432–7. http://dx.doi.org/10.1016/j.radonc.2015.08.026.
[36] Banerjee R, Chakraborty S, Nygren I, Sinha R. Small bowel dose parameters predicting grade $\geq$3 acute toxicity in rectal cancer patients treated with neoadjuvant chemoradiation: An independent validation study comparing peritoneal space versus small bowel loop contouring techniques. Int J Radiat Oncol Biol Phys 2013;85:1225–31. http://dx.doi.org/10.1016/j.ijrobp.2012.09.036.
[37] Vinod SK, Lim K, Bell L, Veera J, Ohanessian L, Juresic E, et al. High-risk CTV delineation for cervix brachytherapy: Application of GEC-ESTRO guidelines in Australia and New Zealand. J Med Imaging Radiat Oncol 2016:1–8. http://dx.doi.org/10.1111/1754-9485.12509.
[38] Dimopoulos JCA, Schard G, Berger D, Lang S, Goldner G, Helbich T, et al. Systematic evaluation of MRI findings in different stages of treatment of cervical cancer: potential of MRI on delineation of target, pathoanatomic structures, and organs at risk. Int J Radiat Oncol Biol Phys 2006;64:1380–8. http://dx.doi.org/10.1016/j.ijrobp.2005.10.017.
[39] Esthappan J, Ma DJ, Narra VR, Raptis CA, Grigsby PW. Comparison of apparent diffusion coefficient maps to T2-weighted images for target delineation in cervix cancer brachytherapy. J Contemp Brachytherapy 2011;3:193–8. http://dx.doi.org/10.5114/jcb.2011.26470.
[40] Hoang Duc AK, Eminowicz G, Mendes R, Wong S-L, McClelland J, Modat M, et al. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. Med Phys 2015;42:5027–34. http://dx.doi.org/10.1118/1.4927567.