PV-0531: Multi-centre evaluation of atlas-based and deep learning contouring using a modified Turing Test

Gooding, M.; Smith, A.; Peressutti, Devis; Aljabar, Paul; Evans, E.; Gwynne, S.; Hammer, C.; Meijer, H.J.M.; Speight, R.; Welgemoed, Camarie

Published in:
Radiotherapy and Oncology

DOI:
10.1016/S0167-8140(18)30841-7

Publication date:
2018

Citation for published version (APA):
Gooding, M., Smith, A., Peressutti, D., Aljabar, P., Evans, E., Gwynne, S., Hammer, C., Meijer, H. J. M., Speight, R., Welgemoed, C., Lustberg, T., Soest, J., Dekker, A., & Elmpt, W. (2018). PV-0531: Multi-centre evaluation of atlas-based and deep learning contouring using a modified Turing Test. *Radiotherapy and Oncology*, *127*, S282-S283. https://doi.org/10.1016/S0167-8140(18)30841-7

of the 2015 MICCAI Head and Neck Auto Segmentation Challenge [2], carefully annotated according to clinical guidelines [3]. Dataset *B* contains 467 training and 40 test cases with routine-level clinical annotations. The DNN architecture used is a modified 2D U-Net [1], trained three times on each dataset on image patches in transversal, sagittal and coronal view respectively. We calculate an ensemble prediction by averaging the three individual models' predictions and post-process it by binarization and selection of the largest connected component. Both ensemble models trained on dataset *A* (referred to as model *Ma*) vs. *B* (denoted *Mb*) are evaluated on the test cases of *A* and *B*, using the Dice score as similarity measure to the reference segmentation.

Results

Figure 1 shows box plots of the Dice scores obtained on the test cases of *A* and *B* from both models *Ma* and *Mb*. The results of models *Ma* and *Mb* on a single test dataset are similar. The overall highest median Dice score of 0.887 is obtained when evaluating model *Ma* on the test cases of *A*, the score of *Mb* on *A* is slightly lower at 0.845. However there is a difference between evaluation on test datasets *A* and *B* for both models. On the curated dataset *A*, the median of the Dice score is higher and the variance is significantly lower than on the clinical dataset *B* for both models. This is probably due to the inconsistent references in dataset *B* which makes quantitative evaluation on this dataset difficult.

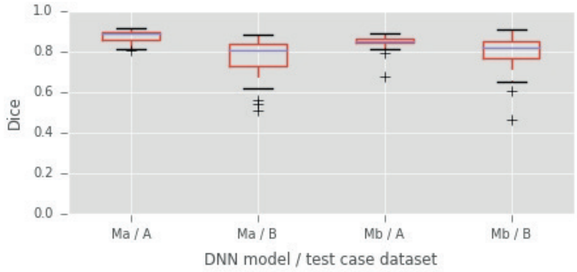


Fig. 1: Dice score of the models *Ma* and *Mb* on the test cases of datasets *A* and *B*.

Conclusion

A main problem of using clinical data for training and testing is the difficulty of quantitative evaluation which is also done in each training step of the DNN. However, on curated testing data, segmentation results after training on clinical vs. curated data seem to be very similar. This suggests that more easily available routine-level clinical data may be sufficient to train high quality segmentation DNNs, but curated data may be helpful for quantitative evaluation. A clinical qualitative evaluation of both models on data independent from both *A* and *B* is work in progress.

[1] Ronneberger O et al., MICCAI LNCS, Vol. 9351, 234-241, 2015

[2] Raudaschl PF et al., Med. Phys., 44(5), 2020-2036, 2017

[3] Sharp GC et al., A Public Domain Database for Computational Anatomy, 2017

PV-0531 Multi-centre evaluation of atlas-based and deep learning contouring using a modified Turing Test

M. Gooding[1], A. Smith[2], D. Peressutti[1], P. Aljabar[1], E. Evans[3], S. Gwynne[4], C. Hammer[5], H.J.M. Meijer[6], R. Speight[7], C. Welgemoed[8], T. Lustberg[9], J. Van Soest[9], A. Dekker[9], W. Van Elmpt[9]

[1]*Mirada Medical Limited, Science and Medical Technology, Oxford, United Kingdom*

[2]*Mirada Medical Limited, Dept. of Engineering, Oxford, United Kingdom*

[3]*Velindre Cancer Centre, Clinical Oncology, Cardiff, United Kingdom*

[4]*South West Wales Cancer Centre, Clinical Oncology, Swansea, United Kingdom*

[5]*University Medical Center Groningen, Department of Radiation Oncology, Groningen, The Netherlands*

[6]*Radboud University Medical Center, Department of Radiation Oncology, Nijmegen, The Netherlands*

[7]*St James University Hospital, Medical Physics and Engineering, Leeds, United Kingdom*

[8]*Imperial College Healtcare NHS Trust, Radiotherapy Department, London, United Kingdom*

[9]*MAASTRO Clinic, Department of Radiation Oncology, Maastricht, The Netherlands*

Purpose or Objective

While quantitative assessment of autocontouring quality is useful, frequently used measures do not necessary indicate clinical acceptability or benefit. In contrast, clinical based assessment metrics, such as time saved with autocontouring or subjective evaluations, are both time consuming to perform and difficult to implement in a multi-centre evaluation. Inspiration is taken from the Artificial Intelligence community to propose an assessment method based on the 'Turing Test". The objective of this study was to perform a multi-centre evaluation of two autocontouring methods using this approach.

Material and Methods

A website was set up to facilitate multi-centre comparison. For each assessment, participants were shown single slice CT images including an OAR contour, and were asked one of three questions; 1) whether they thought the contour was drawn by autocontouring or a human, 2) whether they would accept or reject the contour for use in clinical practice, and 3) which contour they preferred when shown two OAR contours. The CT slice, OAR and question were chosen randomly from a database.

The database consisted of 60 clinical cases from a single institution (40 thoracic, 20 prostate). Participants selected a body region based on their expertise. In addition to the clinical contours, OARs were created using atlas-based contouring [ABC] WorkflowBox 1.4, Mirada Medical, Oxford, UK) and deep learning-based contouring [DLC] (WorkflowBox 2.0 alpha, Mirada Medical, Oxford, UK). Both ABC and DLC were trained using other cases from the same institution.

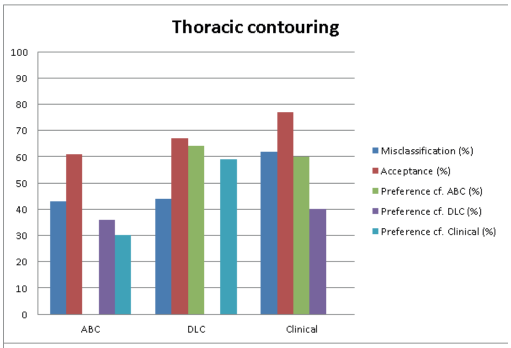
Each participant was asked 100 questions for each anatomic region. For the thoracic evaluation; 15 clinical participants (clinicians, dosimetrist or technicians) from 5 institutions participated, with 5 from the institution providing the contours. For the prostate evaluation; 6 clinical participants from 3 institutions participated, with 4 from the institution providing the contours.

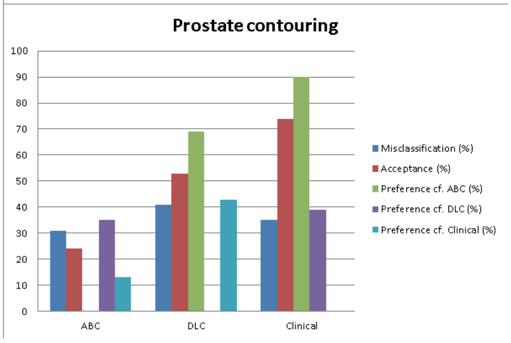
Results

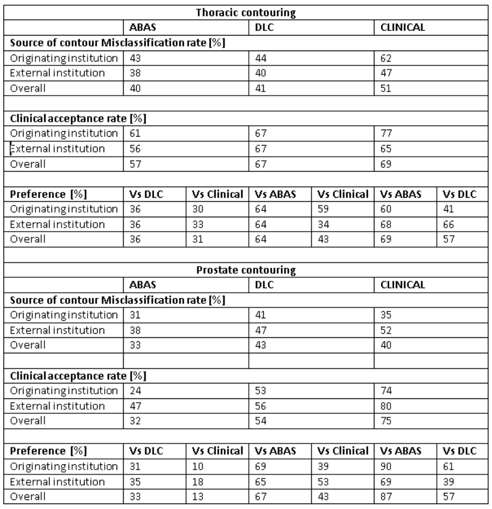
The figure and table show the results summarised over all organs for each contouring method.

For the thoracic evaluation, participants found it hard to identify the source of contours. The overall acceptance of DLC was higher than that of ABC, approaching the same level of acceptance as the clinical contours. Both DLC and Clinical are preferred to ABC, with Clinical being preferred slightly more than DLC.

For the prostate evaluation, participants found it easier to identify the source of contours, but with greater misclassification being caused by DLC. Acceptance of DLC was higher than that of ABC, but still below that of the original clinical contours. Users expressed a preference for DLC and Clinical over ABC, with Clinical being marginally preferred to DLC.







Thoracic contouring						
	ABAS		DLC		CLINICAL	
Source of contour Misclassification rate [%]						
Originating institution	43		44		62	
External institution	38		40		47	
Overall	40		41		51	
Clinical acceptance rate [%]						
Originating institution	61		67		77	
External institution	56		67		65	
Overall	57		67		69	
Preference [%]	**Vs DLC**	**Vs Clinical**	**Vs ABAS**	**Vs Clinical**	**Vs ABAS**	**Vs DLC**
Originating institution	36	30	64	59	60	41
External institution	36	33	64	34	68	66
Overall	36	31	64	43	69	57
Prostate contouring						
	ABAS		DLC		CLINICAL	
Source of contour Misclassification rate [%]						
Originating institution	31		41		35	
External institution	38		47		52	
Overall	33		43		40	
Clinical acceptance rate [%]						
Originating institution	24		53		74	
External institution	47		56		80	
Overall	32		54		75	
Preference [%]	**Vs DLC**	**Vs Clinical**	**Vs ABAS**	**Vs Clinical**	**Vs ABAS**	**Vs DLC**
Originating institution	31	10	69	39	90	61
External institution	35	18	65	53	69	39
Overall	33	13	67	43	87	57

Conclusion

The web-based assessment method provides an easy way to perform multi-centre validation of autocontouring. This study showed that autocontours may be confused with clinical ones, when reviewed blind, and DLC contours were accepted at a similar rate to clinical ones.

PV-0532 Using deep learning to generate synthetic CTs for radiotherapy treatment planning

M. Bylund[1], J. Jonsson[1], J. Lundman[1], P. Brynolfsson[1], A. Garpebring[1], T. Nyholm[1], T. Löfstedt[1]

[1]*Umeå University, Department of Radiation Sciences, Umeå, Sweden*

Purpose or Objective

MR images are often used in radiotherapy for delineation of treatment volumes and organs at risk. However, electron density information is also required when performing treatment planning. Traditionally, this information comes from CT images of the patient. If synthetic CT (sCT) images are instead generated from MR images, an MR-only workflow can be achieved. This allows for reduced registration errors, and can for instance also pave the way for individualized treatment based on the progression of the tumor during treatment in a combined MR-LINAC.

In this project, we are investigating the generation of sCT images using deep learning. The dosimetric accuracy when using these images for treatment planning is evaluated.

Material and Methods

20 male patients with prostate- or rectal-cancer were imaged in both a CT scanner and a 3T MR camera as part of their regular clinical treatment. A deep convolutional neural network (DCNN), using the U-net architecture, was trained on image data from 15 of the patients, and then used to generate sCTs for the remaining five patients. The network had 13 convolution layers in the encoding part and 14 convolution layers in the decoding part, with interleaved subsampling and upsampling layers. Skip connections were used to pass information from the encoding part to the decoding part at different sampling levels.

Fat and Water images from a 2-point Dixon sequence were used as input to the DCNN. The MR images used 2.4 mm isotropic voxels, and an in-plane resolution of 192x192 pixels. The CT images had a slice thickness of 2.0 mm, an in-plane resolution of 512x512 pixels, and a FOV of 55 cm. Before training, the CT images were registered to the MR images, and downsampled to the same resolution.

Treatment plans were created based on the original unmodified CT images. For the five patients with generated sCTs, the treatment plans were then re-calculated based on the DCNN-created sCTs, and the dose distributions of the two plans were compared.

Results

The error in average dose to the PTV ranged from 0.03% to 0.46% (mean 0.28%). For the CTV, the corresponding range was 0.03% to 0.42% (mean 0.25%). Gamma analysis using a 2%/2-mm global gamma criteria showed a 98.67% to 100.00% (mean 99.60%) pass rate for the PTV, and 97.78% to 99.78% (mean 99.13%) for the volume receiving dose >15% of the prescribed dose.

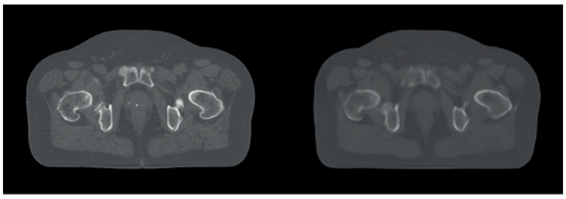


Figure 1: CT (left) and generated sCT (right) showing the same slice for one of the evaluated patients

Conclusion

The results are encouraging, and show that sCTs generated from MR images by a DCNN can be used to calculate treatment plans with dosimetric accuracy comparable to that achieved with sCTs generated by other methods. Using deep learning for sCT generation shows great promise since the method has the potential to robustly handle differences in the input images. Such differences could for instance stem from different MR cameras being used, or a difference in the specific sequences being used as input. This means that the method would not necessarily be site-specific, but could with minor adjustments be used at different sites with varying clinical protocols.

PV-0533 Methods for distortion assessment and correction on the Australian MRI-linac

A. Walker[1,2,3], J. Buckley[3,4], K. Zhang[1,3], B. Dong[1,3], L. Holloway[1,3,4], G. Liney[1,2,3]

[1]*Liverpool and Macarthur Cancer Therapy Centres, Medical Physics, Liverpool BC, Australia*

[2]*University of New South Wales, School of Medicine, Sydney, Australia*

[3]*Ingham Institute for Applied Medical Research, Medical Physics, Liverpool, Australia*