ACTA ONCOLOGICA
2022, VOL. 61, NO. 2, 255–263
https://doi.org/10.1080/0284186X.2021.2013530

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

MRI-based automatic segmentation of rectal cancer using 2D U-Net on two independent cohorts

Franziska Knuth[a]*, Ingvild Askim Adde[a]*, Bao Ngoc Huynh[b], Aurora Rosvoll Groendahl[b], René Mario Winter[a], Anne Negård[c,d], Stein Harald Holmedal[c], Sebastian Meltzer[e], Anne Hansen Ree[d,e], Kjersti Flatmark[d,f], Svein Dueland[g], Knut Håkon Hole[d,h], Therese Seierstad[h], Kathrine Røe Redalen[a] and Cecilia Marie Futsaether[b]

[a]Department of Physics, Norwegian University of Science and Technology, Trondheim, Norway; [b]Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway; [c]Department of Radiology, Akershus University Hospital, Lørenskog, Norway; [d]Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [e]Department of Oncology, Akershus University Hospital, Lørenskog, Norway; [f]Department of Gastroenterological Surgery, Oslo University Hospital, Oslo, Norway; [g]Department of Oncology, Oslo University Hospital, Oslo, Norway; [h]Division of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway

ABSTRACT

Background: Tumor delineation is time- and labor-intensive and prone to inter- and intraobserver variations. Magnetic resonance imaging (MRI) provides good soft tissue contrast, and functional MRI captures tissue properties that may be valuable for tumor delineation. We explored MRI-based automatic segmentation of rectal cancer using a deep learning (DL) approach. We first investigated potential improvements when including both anatomical T2-weighted (T2w) MRI and diffusion-weighted MR images (DWI). Secondly, we investigated generalizability by including a second, independent cohort.

Material and methods: Two cohorts of rectal cancer patients (C1 and C2) from different hospitals with 109 and 83 patients, respectively, were subject to 1.5 T MRI at baseline. T2w images were acquired for both cohorts and DWI (b-value of 500 s/mm^2) for patients in C1. Tumors were manually delineated by three radiologists (two in C1, one in C2). A 2D U-Net was trained on T2w and T2w + DWI. Optimal parameters for image pre-processing and training were identified on C1 using five-fold cross-validation and patient Dice similarity coefficient (DSC_p) as performance measure. The optimized models were evaluated on a C1 hold-out test set and the generalizability was investigated using C2.

Results: For cohort C1, the T2w model resulted in a median DSC_p of 0.77 on the test set. Inclusion of DWI did not further improve the performance (DSC_p 0.76). The T2w-based model trained on C1 and applied to C2 achieved a DSC_p of 0.59.

Conclusion: T2w MR-based DL models demonstrated high performance for automatic tumor segmentation, at the same level as published data on interobserver variation. DWI did not improve results further. Using DL models on unseen cohorts requires caution, and one cannot expect the same performance.

ARTICLE HISTORY
Received 27 June 2021
Accepted 21 October 2021

KEYWORDS
Rectal cancer; magnetic resonance imaging; diffusion weighted magnetic resonance imaging; tumor volume; automatic segmentation; deep learning

Background

Defining the tumor volume is an important step in many areas of cancer therapy. The volume is not only needed for radiotherapy and surgical treatment planning but can also be used for treatment response monitoring or extraction of imaging biomarkers. Manual tumor delineation, the current gold standard, is a time and labor-intensive process. In addition, in rectal cancer, high interobserver variations in manual delineations have been reported [1–3].

Magnetic resonance imaging (MRI) is today integral in the diagnostic work-up and staging of rectal cancer, whereas computed tomography (CT) is the basis for radiotherapy treatment planning and regular follow-up after completed treatment. Compared to CT, anatomical T2-weighted (T2w) MRI images have superior soft tissue contrast with the potential for more accurate tumor delineation in the pelvic cavity. Another benefit of MRI is the ability to acquire images depicting functional tissue information, such as diffusion weighted imaging (DWI) that can visualize diffusion restrictions of water molecules in tissues [4]. The use of DWI in combination with T2w images is recommended for rectal cancer diagnostics [5] and may also be valuable for tumor delineation.

Artificial intelligence strategies, in particular deep learning, are increasingly utilized for the purpose of automatic tumor segmentation based on medical images. These strategies have the potential to not only save time, but also decrease the interobserver variations. Several studies on rectal cancer

CONTACT Franziska Knuth franziska.h.knuth@ntnu.no Department of Physics, Norwegian University of Science and Technology, Trondheim, Norway
*Both these authors have equally contributed to this work.

Supplemental data for this article can be accessed here.

segmentation focus on deep neural networks with MR images as basis [1,6–9]. For example, Wang et al. [9] trained a modified, pretrained Resnet50 on T2w images of 461 rectal cancer patients and showed good performance on a hold-out test set from the same cohort as well as on three other external test sets. Trebeschi et al. [1] used a convolutional neural network (CNN) with T2w and DWI to segment locally advanced rectal cancer and evaluated the performance against two independent sets of manual contours. For colorectal cancer, Jian et al. [7] used a CNN for tumor segmentation based on T2w images from 612 patients. However, a common weakness of previous studies is the lack of evaluation on independent datasets.

The aim of this study was to investigate the performance of an MRI-based 2D U-Net deep learning algorithm for automatic segmentation of rectal cancer. First, training and preprocessing parameters were evaluated and optimized before potential improvements of including both anatomical T2w and DWI were assessed. Last, we investigated generalizability by evaluating the optimized U-Net model on an independent dataset.

Materials and methods

Patients

The patient data in this study was from two clinical studies. The OxyTarget study (NCT01816607) was a prospective observational study that enrolled a total of 192 patients between October 2013 and December 2017. Secondly, the LARC-RRP study (NCT00278694) was a prospective phase II study with intensified neoadjuvant treatment [10] that enrolled a total of 109 patients between October 2005 and March 2010. The current analysis included 109 patients from the former study and 83 patients from the latter study, for which the required images and manual delineations described in the next section were available. The two cohorts are referred to as C1 and C2, respectively. All patients had histologically confirmed rectal adenocarcinoma and successful MRI acquisition with adequate image quality without artifacts and other distortions. Patient and tumor characteristics are summarized in Table 1. For all patients, written informed consent was obtained and the study was performed in accordance with the Helsinki Declaration. Approval was obtained from the Institutional Review Board and the Regional Committee for Medical and Health Research Ethics.

Magnetic resonance imaging

For cohort C1, T2w and DW images with a b-value of 500 s/mm^2 were acquired using a Philips Achieva 1.5T system (Philips Healthcare, Best, The Netherlands). A radiologist R1 with 14 years of experience with pelvic MRI delineated the tumor region of interest on the T2w images with the DW images as guidance. For a subset of 74 patients, a second delineation by a radiologist R2 with 7 years of experience was collected. For cohort C2, T2w images were acquired for all patients and the tumor was delineated by an expert R3 with 12 years of experience in pelvic MRI. Within this cohort either a 1.5T GE Signa® LS scanner (GE Healthcare, Milwaukee, WI, USA) or a 1.5T Siemens Espree scanner (Siemens, Erlangen, Germany) was used, due to a scanner upgrade during the enrollment period. Table 1 lists the imaging parameters for both cohorts. Additional details about the MR sequences can be found in [11,12] for C1 and [10,13] for C2.

Image pre-processing

All images were cropped to a matrix size of 352×352 per slice. This matrix size was the minimal size that still covered

Table 1. Overview of patient and tumor characteristics, and imaging parameters.

Characteristic	Cohort C1	Cohort C2
Patients		
Patients (*n*)	109	83
Sex (male/female; *n*)	73/36	49/34
T-stage (T2/T3/T4; *n*)	20/53/36	5/48/30*
N-stage (NX/N0/N1/N2; *n*)	0/48/37/24	1/12/8/62†
TNM staging edition	7	5
Tumor volume (cm^3; median (range))	22.6 (1.8 − 233.5)	16.5 (1.1 − 293.4)
MRI scanner	Phillips Achieva 1.5 T	GE Signa LS 1.5 T (49/83 cases) Siemens Espree 1.5 T (34/83 cases)
Delineations		
Delineated by Radiologist (R; years of experience)	R1 (14) R2 (7); for 74/109 cases‡	R3 (12)
T2w-MRI		
Repetition time (ms)	2820-3040	3000–4000
Echo time (ms)	80	81–84
In plane image resolution (mm^2)	0.35×0.35	$0.38 \times 0.38 - 0.39 \times 0.39$
Slice thickness (mm)	2.5	4.0
DWI		
Repetition time (ms)	3000	
Echo time (ms)	75	
In plane image resolution (mm^2)	1.25×1.25	
Slice thickness (mm)	4	

*$p = 0.61$ for the two-sided *t*-test between T-stage in C1 and C2; †$p = 0.72$ for the two-sided *t*-test between N-stage in C1 and C2; ‡patient statistics for this subset of C1 and result of the two-sided *t*-test between C1 and the C1 subset for T and N-stage: sex (male/female; *n*): 51/23; T-stage (T2/T3/T4; *n*): 12/38/24 ($p = 0.39$); N-stage (N0/N1/N2; *n*): 31/25/18 ($p = 0.84$); Tumor volume (cm^3, median (range)): 22.6 (2.0-233.5).

the whole tumor for all patients. For patients in C1, the DW images were rigidly registered and resampled toward the T2w image grid before cropping. Details regarding the registration are given in Supplementary S1. Image slices not containing tumor according to the delineation by R1 in C1 and R3 in C2 were discarded, to focus on segmentation rather than detection. Within C1, image slices where the DW image did not fully cover the tumor were also removed. This resulted in 1 826 slices in C1 and 863 slices in C2.

In order to normalize for potential variation in image intensities between scanners, two normalization methods were explored. The first was calculation of the z-score (ZS), and the second was the ZS combined with histogram matching (HM). Both were performed per patient and image type.

For cohort C1, a 15% hold-out test set was created, while the remaining patients were split into five folds used for cross-validation during training. All splits were stratified by sex and T-stage.

Image pre-processing was performed in Python 3.7 [14] using scikit-image 0.18 [15], scikit-learn 0.24 [16] and SimpleITK 2.0 [17].

U-Net architecture and training parameters

A 2D U-Net [18] was used in this study. The learning rate was varied between 1e-4 and 1e-5. Two options for the loss function were explored, namely the Dice loss function (DL) [19] and a modified Dice loss function (mDL) defined as:

$$\mathrm{DL} = 1 - \frac{2\Sigma_i^N p_i g_i}{\Sigma_i^N p_i^2 + \Sigma_i^N g_i^2}$$

$$\mathrm{mDL} = 1 - \frac{2\Sigma_i^N p_i g_i}{\Sigma_i^N p_i + \Sigma_i^N g_i}$$

Here, p and g denote the prediction probability and ground truth, respectively, and the summation runs over all pixels. The modified denominator in mDL compared to DL results in a less penalizing loss function. Details regarding the network architecture and other fixed parameters are listed in Supplementary S2 and the relevant code is available at [20]. Training and model evaluation were performed using deoxys 0.0.8 [21], a framework for running deep-learning experiments with emphasis on tumor auto segmentation. Experiments were run on the Orion computing cluster based at the Norwegian University of Life Sciences.

Experimental procedure

Figure 1 presents a general overview of the experimental procedure. In the first step, a five-fold cross-validation on the C1 cross-validation set was used to optimize the training parameters (learning rate: 1e-4 vs 1e-5; loss function: DL vs mDL) as well as the normalization of the input images (ZS vs ZS + HM). Using this approach, one model was trained using T2w images and another was trained using both T2w and DW images. Manual delineations made by R1 were used as ground truth. The Sørensen Dice similarity coefficient (DSC) [22] was used as performance measure, defined as

$$\mathrm{DSC} = \frac{2|P \cap G|}{|P| + |G|}$$

Here, P denotes the binary prediction ($p_i > 0.5$) and G the ground truth segmentation.

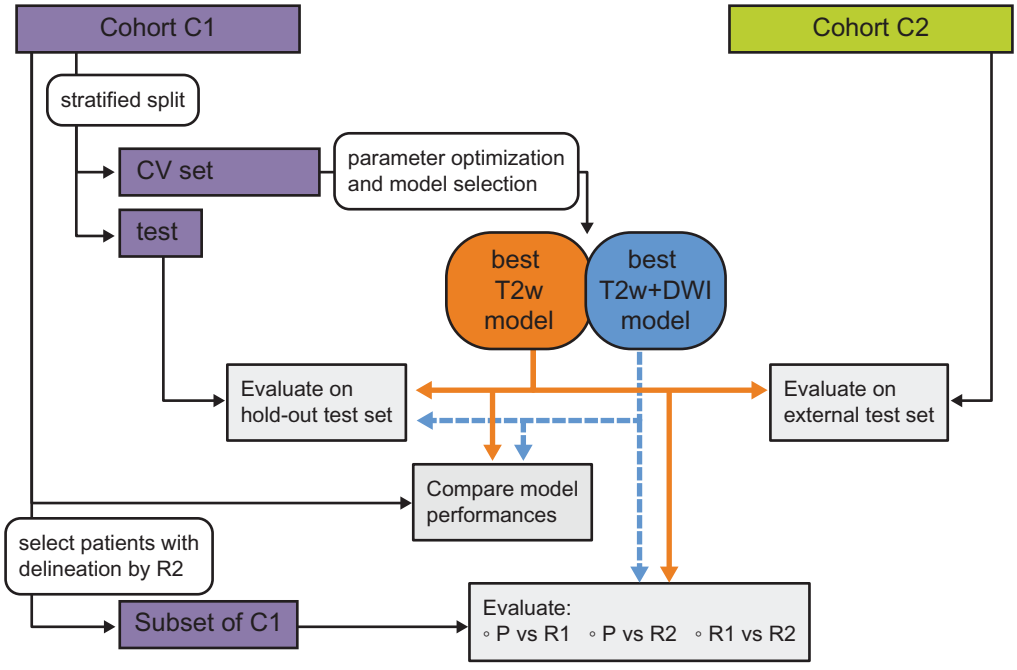


Figure 1. Illustration of the experimental procedure. For both cohorts C1 and C2, T2-weighted (T2w) images were available, whereas diffusion weighted images (DWI) were available for C1 only. Manual delineations were made by radiologist 1 (R1) in C1 and radiologist 3 (R3) in C2. For a subset of C1, delineations by radiologist 2 (R2) were also available. The splits of cohort C1 into five cross-validation (CV) folds and a hold-out test set (test) were stratified by T-stage and sex. For those patients where both T2w images and DWI were available a model on both T2w + DWI was evaluated in parallel to the T2w model. The per patient Sørensen Dice similarity coefficient (DSC_p) was used to quantify performance when evaluating the predicted segmentations (P).

Differences in the per patient DSC (DSC_p) under variation of training and normalization parameters were evaluated using the Friedman test for repeated measurement [23]. If the test was significant, the Nemenyi *post hoc* test [24] was used for a pairwise comparison of the individual parameter combinations. The parameter combination resulting in the best model, defined as having the highest rank sum, was chosen for the subsequent analysis. The selected T2w and T2w + DWI models were evaluated on the test sets, where majority voting was used to combine the models from each of the five individual cross-validation folds. The potential benefit of including DWI in addition to T2w images was tested using a two-sided Wilcoxon signed rank test on the DSC_p scores of the predicted segmentations for patients in C1.

Next, the generalization capability of the best T2w-based model was evaluated using the independent cohort C2. As for the C1 hold-out test set, automatic segmentations were generated for patients in C2 using the best model based on T2w images (majority voting) and the performance was quantified by the the DSC_p. The C2 cohort was selected as it had T2w images with delineations available and followed comparable national guidelines.

Lastly, for C1 the predicted segmentations made by the best T2w and T2w + DWI-based models were evaluated against the delineation by R2 using the DSC_p and tested for statistical significance. This was done to investigate the utility of the model when used by a new user, who was not involved in the training process.

Throughout the statistical analysis, a significance level of 0.05 was used. The statistical analysis was performed using Python 3.7 [14] with SciPy 1.6 [25] and scikit-posthocs 0.6 [26]. Matplotlib 3.4 [27] was used for visualization.

Results

Figure 2(A) summarizes the resulting DSC_p of the five-fold cross-validation on C1, where learning rate, loss function and input normalization were varied. The Friedman test detected an effect of training parameter variation ($p < 0.0001$) for models based on T2w or T2w + DW images. Thus, a *post hoc* Nemenyi test was used, and the results are shown in Figure 2(B).

For both image inputs, the models using ZS normalization and 1e-4 as learning rate achieved the best results, defined as highest rank sum of DSC_p. The optimal choice of loss function varied. For T2w images, the DL function was optimal, which resulted in median (interquartile range) DSC_p of 0.76 (0.16). The combination of T2w and DW images achieved the highest performance when the mDL function was used, giving a DSC_p of 0.77 (0.14). However, the difference in performance between models using either mDL or DL was not significant.

Figure 3(A) shows the C1 cross-validation and test set DSC_p for the best two models based on T2w and T2w + DW images. Test set model performance was similar with DSC_p of 0.77 (0.21) for T2w and 0.76 (0.18) for T2w + DWI as input. Figure 3(B) shows the DSC_p of all patients in C1 for the two models. The two-sided Wilcoxon signed rank test showed no significant difference in performance after the inclusion of DWI. Resulting segmentations for individual image slices in the test set can be found in Figure 4(A).

Figure 3(C) shows the performance of the best model based on T2w images when applied on the independent cohort C2. Compared to the C1 test set results, the performance on C2 was lower giving a DSC_p of 0.59 (0.28).

Variations in DSC_p when the predictions were evaluated against manual delineations (R2) not included as the ground truth are shown in Figure 3(D). These DSC_p's were similar to those obtained between model predictions and the R1 ground truth delineations. The two-sided Wilcoxon signed rank test showed no significant difference (DSC_p for T2w inputs: 0.77 (0.15) for R1 and 0.74 (0.11) for R2; DSC_p for T2w + DWI inputs: 0.78 (0.14) for R1 and 0.76 (0.13) for R2). For comparison, the DSC_p comparing the manual delineations from R1 and R2 was 0.81 (0.07) for the same set of patients.

Discussion

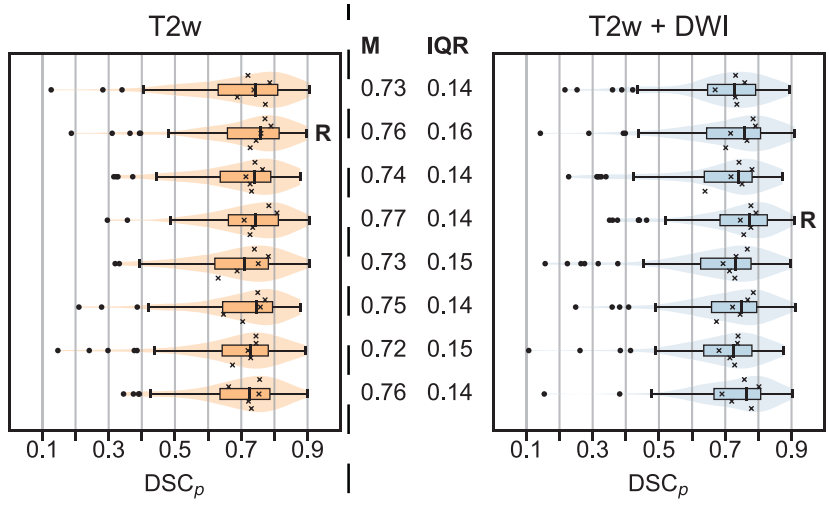
In this study, a 2D U-Net deep learning algorithm was used for automatic segmentation of rectal cancer in MRI images. In the first cohort of 109 patients, the potential benefit of combining T2w and DWI was studied, where our analysis found no significantly improved performance compared to a model based only on T2w images as input. This suggests that T2w images alone contain sufficient information for adequate segmentation, which is beneficial in the context of the clinical workflow and eliminates the need for co-registration. Both models were also evaluated using manual delineations by a second expert on the same cohort and the T2w-based model was evaluated on a second, independent cohort. Our results suggest that a trained model is useful for a new user, whereas applying the model to a new cohort requires more caution.

Using T2w and DWI as basis, Trebeschi et al. [1] used a CNN to segment rectal cancer. They achieved a mean DSC of 0.68 and 0.70 when evaluating their predicted segmentation against two independent readers, respectively. The T2w-based model presented in our study achieved a median (interquartile range) per patient DSC_p of 0.78 (0.14) and 0.76 (0.13) for the two readers. As they [1] calculated DSC on a per slice basis while our DSC was calculated on a per patient basis, the numbers are not directly comparable. Using the per patient DSC_p may give a larger influence to smaller tumors, since they contribute relatively more to the dataset than in a per slice DSC approach where larger tumors dominate the resulting score as they extent over many slices. Another difference is that Trebeschi et al. designed their network for both detection and segmentation, while we focused on segmentation only.

For the analysis presented in this work, we restricted the dataset to only include slices containing tumor defined by the manual delineations by R1 and R3. This was done to separate the tasks of detection and accurate segmentation. These are both highly relevant clinical applications, but have vastly different requirements for their performance metric

A

Norm	LF	LR	M	IQR	M	IQR
ZS	DL	1e-05	0.74	0.18	0.73	0.14
ZS	DL	1e-04	0.76	0.16	0.76	0.16
ZS	mDL	1e-05	0.74	0.15	0.74	0.14
ZS	mDL	1e-04	0.74	0.15	0.77	0.14
ZS+HM	DL	1e-05	0.71	0.16	0.73	0.15
ZS+HM	DL	1e-04	0.75	0.15	0.75	0.14
ZS+HM	mDL	1e-05	0.73	0.14	0.72	0.15
ZS+HM	mDL	1e-04	0.73	0.15	0.76	0.14



R: highest rank sum

Friedman test: $p < 0.0001$

Friedman test: $p < 0.0001$

B

		Norm	ZS				ZS+HM			
		LF	DL		mDL		DL		mDL	
Norm	LF	LR	1.E-05	1.E-04	1.E-05	1.E-04	1.E-05	1.E-04	1.E-05	1.E-04
ZS	DL	1.E-05	-	ns	ns	ns	ns	ns	ns	ns
		1.E-04	**	-	ns	ns	ns	ns	**	**
	mDL	1.E-05	ns	***	-	*	ns	ns	ns	ns
		1.E-04	***	ns	***	-	*	ns	***	*
ZS+HM	DL	1.E-05	ns	***	ns	***	-	ns	ns	ns
		1.E-04	***	ns	ns	ns	**	-	ns	ns
	mDL	1.E-05	ns	***	ns	***	ns	**	-	ns
		1.E-04	**	ns	***	ns	***	ns	***	-

T2w

T2w + DWI

ns : $p > 0.05$

$*$: $p \leq 0.05$

$**$: $p \leq 0.01$

$***$: $p \leq 0.001$

$****$: $p \leq 0.0001$

Figure 2. (A) Results of five-fold cross-validation for models trained on T2-weighed (T2w) images alone and in combination with diffusion-weighted imaging (DWI) (T2w + DWI) in cohort C1. Image normalization (Norm) was done by either using the z-score (ZS) or the z-score with histogram matching (ZS + HM). Training parameter options for loss function (DL: dice loss; mDL: modified dice loss) and learning rate (LR) were explored. Median (M) and inter-quartile-range (IQR) of the per patient Sørensen Dice similarity coefficient (DSC_p) are listed. The combined box and violin plot shows the DSC_p for the combined five folds, with outliers shown as dots. The x marks the median DSC_p for the individual folds. The combinations with the highest rank sum are marked with R. (B) Result of *post hoc* Nemenyi pairwise comparison for the results presented in A. Values above the diagonal (orange) state results for models based on T2w images. Values below the diagonal (blue) give the results for models using T2w + DW images as input.

[28]. Another benefit of restricting the analysis to tumor-containing image slices, is that the classes (tumor versus normal tissue) in the dataset were more balanced. This also made the two cohorts, C1 and C2, more comparable. While C1 initially had 71% image slices containing tumor, it was only 29% for C2. In a clinical setting, such a preprocessing step could be performed by a user that selects the slices with visible tumor. Compared to the task of manual delineation, this would still have a large impact on expenditure of time. Another possibility would be to train an independent network that filters the image slices to those containing tumor before these slices are used as input to the segmentation algorithm.

The interobserver variation for cohort C1 with a DSC_p of 0.81 is in line with previously published results [3]. As the network uses the manual delineations as ground truth, the interobserver variation can be interpreted as an upper limit to the performance of an automated algorithm. Thus, if the performance of a trained model is within the range of the inter-observer variation, the model is useful for the user. The use of consensus delineations as ground truth has the potential to make the model more robust and eliminate the influence of a single observer on the trained model. Based on the similar model performance observed when using R1 and R2 for evaluation (cf. Figure 3(D)), we would also expect a similar result if consensus delineations of R1 and R2 would have been used during training.

Often, the automatic segmentations tended to be smoother than the manual ones. Examples of this can be found in Figure 4, especially in the first and second case in Figure 4(A). If we recognize the delineation as a part of a multistep process, these differences might potentially have little effect on the final result. For example, in the case of radiotherapy planning, the small variation in contour smoothness might have little effect on the final planned dose distribution. In addition, other commonly used measures of tumor size and volume, such as the longest diameter, are largely unaffected by a smoother contour.

The performance achieved on the external test set (C2) was lower compared to the C1 test set, with a DSC_p of

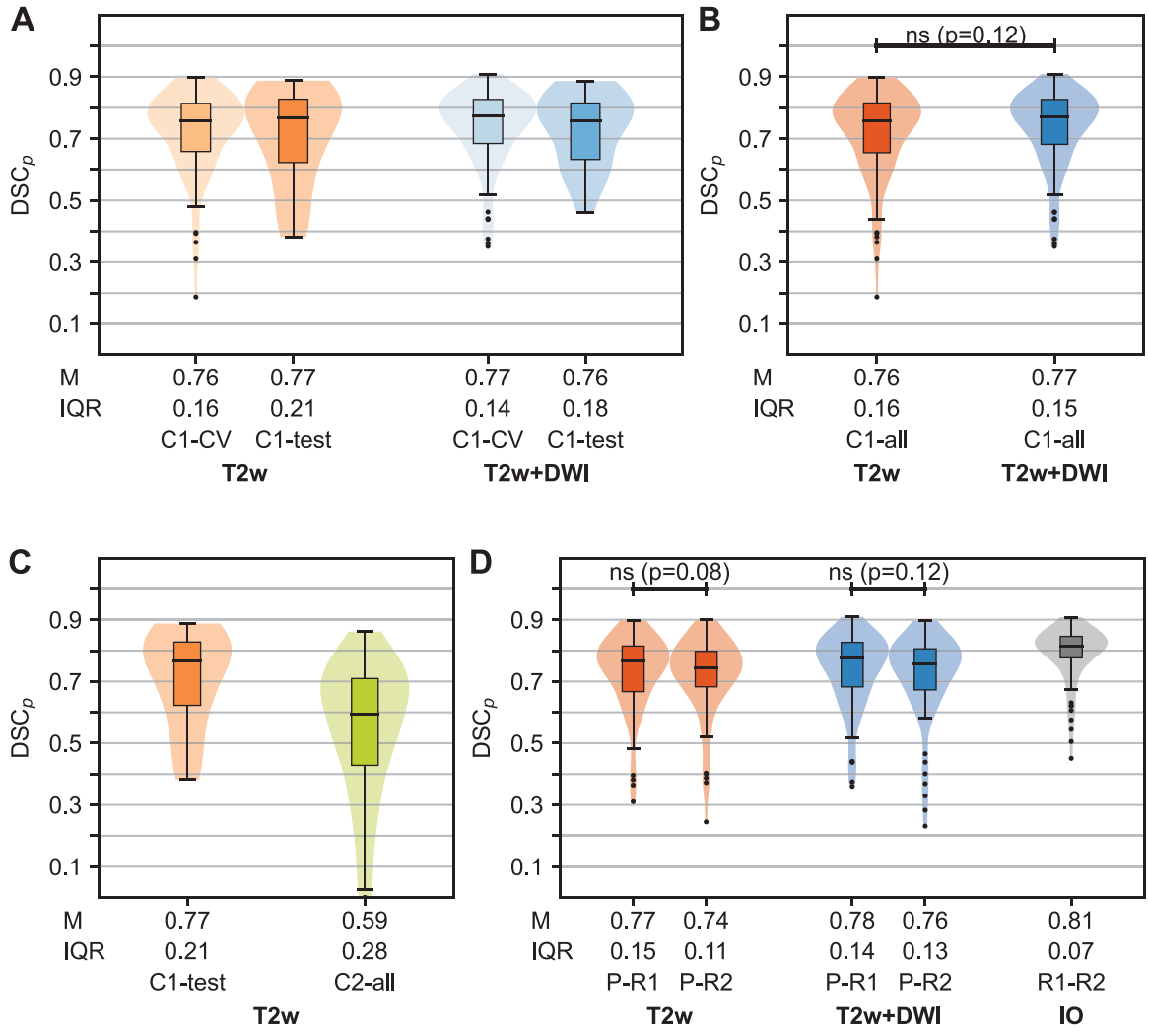


Figure 3. (A) Comparison of the per patient Sørensen Dice similarity coefficient (DSC_p) for the five-fold cross-validation (CV) ($n = 92$) and test set evaluation (majority voting, $n = 17$) of cohort C1. Median (M) and interquartile range (IQR) are stated. Results are shown for the superior models based on T2-weighted (T2w) and both T2w and diffusion weighted imaging (DWI). For each input set, the model with the highest sum rank in the previous analysis was chosen (T2w: Z-score normalization, Dice loss function, 1e-4 learning rate; T2w + DWI: Z-score normalization, modified Dice loss function, 1e-4 learning rate). (B) Comparison of the DSC_p of T2w and T2w + DWI-based models for all patients in C1 (CV and test set) with result of the two-sided Wilcoxon signed rank test. (C) Comparison of the internal and external test set results (DSC_p) for the T2w-based model. (Internal: C1-test; External: C2, $n = 83$). (D) DSC_p between predictions (P) on C1 and manual delineations made by radiologist R1, whose delineations were used during the training, and radiologist R2, respectively. C1 patients were restricted to those with two manual delineations ($n = 74$). For comparison, the interobserver (IO) DSC_p is shown for the same set of patients. For both input combinations (T2w, T2w + DWI), two-sided Wilcoxon signed rank test was used to compare the performance that the two radiologists would experience.

0.59 compared to 0.77. A potential reason for this could be that while C1 was quite homogeneous and acquired at the same scanner, C2 contained images from two different scanners with potentially larger variation in the imaging sequence parameters, such as slice thickness. Furthermore, a systematic difference between R3 versus R1 and R2 cannot be ruled out. It can also be noted, that C1 included both early-stage and locally advanced rectal cancer, while C2 included only locally advanced rectal cancer, which is potentially more complex to delineate. In the context of our study, the work of Wang et al. [9] is relevant to mention. They used a ResNet50-based model to segment rectal cancer and evaluated their model on multiple test sets. These test sets were acquired using MR scanners from multiple vendors. They did not observe a similar drop in performance as we found when using their proposed model to predict an external test set. One potential explanation for this could be that the training dataset used was much larger with data from 461 patients, and that the training images contained more variations in imaging sequence parameters. Thus, there was a larger overlap in the data domain for the independent test sets to the training dataset.

In our study, only parameters such as the loss function, learning rate and image normalization were optimized. In a follow-up study, it would be interesting to further optimize the architecture of the U-Net or to test other architectures such as a ResNet [29] or DenseNet [30]. Moreover, the inclusion of other cohorts would be beneficial for a more thorough test of the model. In addition, training the model on a multicenter cohort could more easily generate a robust model covering a larger data domain which increases the generalizability of the model.

Automatic segmentation could be implemented in the clinic in several ways. One option is to use the automatic contours for radiotherapy treatment planning. For this

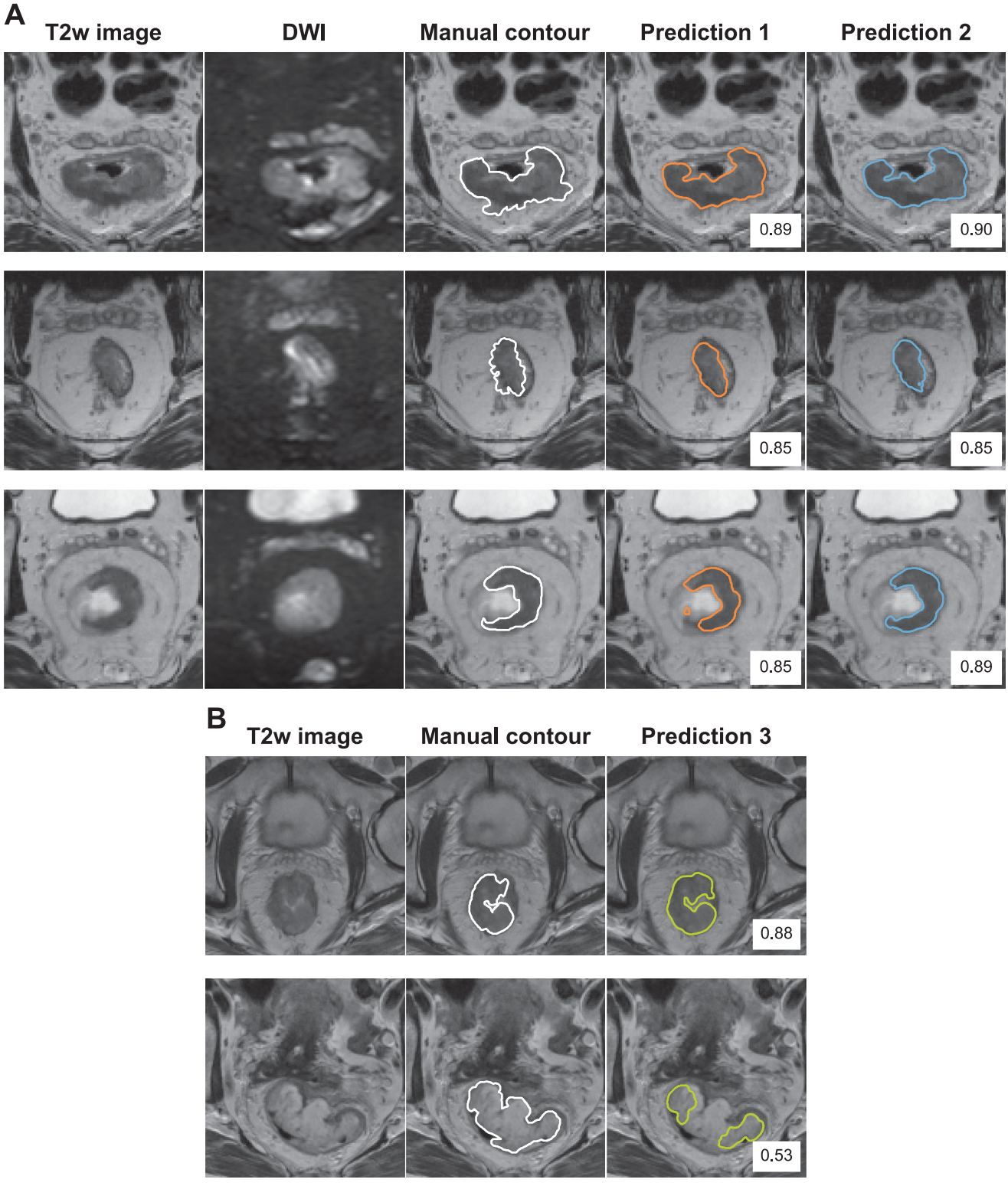


Figure 4. Illustration of predicted contours made by the best models and corresponding DSC value. (A) For three patients in C1, the T2-weighted (T2w) image and the diffusion weighted imaging (DWI) ($b = 500 s/mm^2$) are shown. Manual delineations by radiologist R1 are shown as contour. Predictions made by the best T2w image-based model (Prediction 1) and the best model using T2w + DWI (Prediction 2) are shown (cf. Figure 2). (B) For two patients in C2, the T2w image and manual delineation are shown. The predicted segmentation (Prediction 3) was made using the best T2w-based model trained on C1.

purpose, not only the DSC or other contour-based metrics are important, but also the resulting changes in dose distribution. Another option is to use a fast automatic segmentation to monitor treatment response. Such a volumetric evaluation has the potential to be more accurate than the current standard for treatment monitoring, based on the RECIST system [31]. Compared to using CT, monitoring treatment effects using MRI has the added benefit of no exposure to ionizing radiation and no injection of contrast agent. Furthermore, adequate automatic segmentation could ease the implementation of image biomarker-based decision support systems into the clinical workflow.

In summary, we show high performance of a T2w MR-based deep learning model for automatic tumor segmentation, at the same level as published data on interobserver variation. The addition of DW MR images did not improve results further compared to using T2w MR images alone. Using the model on unseen cohorts requires caution, and one cannot expect the same performance level.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the South-Eastern Norway Regional Health Authority under the [Grant numbers 2013002, 2015048 and 2016050], and the Norwegian Cancer Society under [Grant number 198116-2018].

ORCID

Franziska Knuth http://orcid.org/0000-0002-6998-8681
Bao Ngoc Huynh http://orcid.org/0000-0001-5210-132X
Aurora Rosvoll Groendahl http://orcid.org/0000-0003-1327-3844
René Mario Winter http://orcid.org/0000-0001-7282-6846
Anne Negård http://orcid.org/0000-0002-7624-5595
Sebastian Meltzer http://orcid.org/0000-0001-6640-3927
Anne Hansen Ree http://orcid.org/0000-0002-8264-3223
Kjersti Flatmark http://orcid.org/0000-0001-7409-0780
Svein Dueland http://orcid.org/0000-0002-6125-6689
Knut Håkon Hole http://orcid.org/0000-0001-6885-8538
Therese Seierstad http://orcid.org/0000-0002-2579-5298
Kathrine Røe Redalen http://orcid.org/0000-0002-1172-4632
Cecilia Marie Futsaether http://orcid.org/0000-0001-7944-0719

References

[1] Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. Sci Rep. 2017;7(1):5301.
[2] Irving B, Franklin JM, Papież BW, et al. Pieces-of-parts for supervoxel segmentation with global context: application to DCE-MRI tumour delineation. Med Image Anal. 2016;32:69–83.
[3] Hearn N, Bugg W, Chan A, et al. Manual and semi-automated delineation of locally advanced rectal cancer subvolumes with diffusion-weighted MRI. Br J Radiol. 2020;93(1114):20200543.
[4] Malayeri AA, El Khouli RH, Zaheer A, et al. Principles and applications of diffusion-weighted imaging in cancer detection, staging, and treatment follow-up. Radiographics. 2011;31(6):1773–1791.
[5] Keller DS, Berho M, Perez RO, et al. The multidisciplinary management of rectal cancer. Nat Rev Gastroenterol Hepatol. 2020;17(7):414–429.
[6] Wang J, Lu J, Qin G, et al. Technical note: a deep learning-based autosegmentation of rectal tumors in MR images. Med Phys. 2018;45(6):2560–2564.
[7] Jian J, Xiong F, Xia W, et al. Fully convolutional networks (FCNs)-based segmentation method for colorectal tumors on T2-weighted magnetic resonance images. Australas Phys Eng Sci Med. 2018;41(2):393–401.
[8] Huang Y-J, Dou Q, Wang Z-X, et al., editors. HL-FCN: hybrid loss guided FCN for colorectal cancer segmentation. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018 Apr 4-7; Washington, DC. IEEE; 2018.
[9] Wang M, Xie P, Ran Z, et al. Full convolutional network based multiple side-output fusion architecture for the segmentation of rectal tumors in magnetic resonance images: a multi-vendor study. Med Phys. 2019;46(6):2659–2668.
[10] Dueland S, Ree AH, Grøholt KK, et al. Oxaliplatin-containing preoperative therapy in locally advanced rectal cancer: local response, toxicity and long-term outcome. Clin Oncol. 2016;28(8):532–539.
[11] Bakke KM, Grovik E, Meltzer S, et al. Comparison of intravoxel incoherent motion imaging and multiecho dynamic contrast-based MRI in rectal cancer. J Magn Reson Imaging. 2019;50(4):1114–1124.
[12] Bakke KM, Meltzer S, Grovik E, et al. Sex differences and tumor blood flow from dynamic susceptibility contrast MRI are associated with treatment response after chemoradiation and long-term survival in rectal cancer. Radiology. 2020;297(2):352–360.
[13] Seierstad T, Hole KH, Groholt KK, et al. MRI volumetry for prediction of tumour response to neoadjuvant chemotherapy followed by chemoradiotherapy in locally advanced rectal cancer. Brit J Radiol. 2015;88(1051):20150097.
[14] Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley, CA: CreateSpace; 2009.
[15] Van Der Walt S, Schönberger JL, Nunez-Iglesias J, Scikit-Image Contributors, et al. Scikit-image: image processing in python. PeerJ. 2014;2:e453.
[16] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Machine Learn Res. 2011;12:2825–2830.
[17] Yaniv Z, Lowekamp BC, Johnson HJ, et al. SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. J Digit Imaging. 2018;31(3):290–303.
[18] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, et al., editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351. Cham: Springer; 2015. DOI:10.1007/978-3-319-24574-4_28
[19] Milletari F, Navab N, Ahmadi S-A, editors. V-Net: fully convolutional neural networks for volumetric medical image segmentation 2016. 4th International Conference on 3D Vision (3DV); 25-28 Oct 2016; Stanford, CA. IEEE; 2016. p. 19. DOI:10.1109/3DV.2016.79
[20] Knuth F. GitHub repository knuth_2021_mri; 2021 [2021-20-21]. Available from: https://github.com/Medical-Radiation-Physics-NTNU/knuth_2021_mri.
[21] Huynh BN. deoxys – Framework for running deep-learning experiments with emphasis on cancer tumor auto-segmentation; 2021 [2021-06-21]. Available from: https://pypi.org/project/deoxys/.
[22] Sorensen TA. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol Skar. 1948;5:1–34.
[23] Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937;32(200):675–701.
[24] Nemenyi PB. Distribution-free multiple comparisons. Princeton (NJ): Princeton University; 1963.
[25] Virtanen P, Gommers R, Oliphant TE, SciPy 1.0 Contributors, et al. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods. 2020;17(3):261–272.
[26] Terpilowski MA. Scikit-posthocs: pairwise multiple comparison tests in python. JOSS. 2019;4(36):1169.
[27] Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9(3):90–95.
[28] Reinke A, Eisenmann M, Tizabi MD, et al. Common limitations of image processing metrics: a picture story. 2021. Available from: https://arxiv.org/abs/2104.05642
[29] He K, Zhang X, Ren S, et al., editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on

computer vision and pattern recognition; 2016 Jun 27-30; Las Vegas, NV. IEEE; 2016. DOI:10.1109/CVPR.2016.90

[30] Huang G, Liu Z, Van Der Maaten L, et al., editors. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017 Jul 21-26; Honolulu, HI. IEEE; 2017. DOI:10.1109/CVPR.2017.243

[31] Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst. 2000;92(3):205–216.