IOP Publishing | Institute of Physics and Engineering in Medicine Physics in Medicine & Biology

Phys. Med. Biol. **62** (2017) 8246–8263 https://doi.org/10.1088/1361-6560/aa8d09

Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study

Xin Zhen[1,2], Jiawei Chen[2], Zichun Zhong[3], Brian Hrycushko[1], Linghong Zhou[2], Steve Jiang[1], Kevin Albuquerque[1] and Xuejun Gu[1]

[1] Department of Radiation Oncology, The University of Texas, Southwestern Medical Center, Dallas, TX 75390, United States of America
[2] Department of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong 510515, People's Republic of China
[3] Department of Computer Science, Wayne State University, Detroit, MI 48202, United States of America

E-mail: xuejun.gu@utsouthwestern.edu (X Gu)

Received 25 July 2017, revised 7 September 2017
Accepted for publication 15 September 2017
Published 11 October 2017

Abstract

Better understanding of the dose-toxicity relationship is critical for safe dose escalation to improve local control in late-stage cervical cancer radiotherapy. In this study, we introduced a convolutional neural network (CNN) model to analyze rectum dose distribution and predict rectum toxicity. Forty-two cervical cancer patients treated with combined external beam radiotherapy (EBRT) and brachytherapy (BT) were retrospectively collected, including twelve toxicity patients and thirty non-toxicity patients. We adopted a transfer learning strategy to overcome the limited patient data issue. A 16-layers CNN developed by the visual geometry group (VGG-16) of the University of Oxford was pre-trained on a large-scale natural image database, ImageNet, and fine-tuned with patient rectum surface dose maps (RSDMs), which were accumulated EBRT + BT doses on the unfolded rectum surface. We used the adaptive synthetic sampling approach and the data augmentation method to address the two challenges, data imbalance and data scarcity. The gradient-weighted class activation maps (Grad-CAM) were also generated to highlight the discriminative regions on the RSDM along with the prediction model. We compare different CNN coefficients fine-tuning strategies, and compare the predictive performance using the traditional dose volume parameters, e.g. $D_{0.1/1/2cc}$, and the texture features extracted from the RSDM. Satisfactory

prediction performance was achieved with the proposed scheme, and we found that the mean Grad-CAM over the toxicity patient group has geometric consistence of distribution with the statistical analysis result, which indicates possible rectum toxicity location. The evaluation results have demonstrated the feasibility of building a CNN-based rectum dose-toxicity prediction model with transfer learning for cervical cancer radiotherapy.

Keywords: convolutional neural networks, deformable image registration, transfer learning, rectum toxicity prediction, rectum surface dose maps

(Some figures may appear in colour only in the online journal)

1. Introduction

Cervical cancer is the third most common cause of female cancer mortality worldwide (Torre *et al* 2015). Locally advanced cervical cancer, a common presentation, (Al-Mansour and Verschraegen 2010) is usually treated with external beam radiotherapy (EBRT) followed by brachytherapy (BT) with and without chemotherapy. Randomized studies have shown excellent treatment outcomes in early stage cervical cancer (Gray 2008, Haie-Meder *et al* 2010). However, treatment results are inferior in advanced stages, with a five-year overall survival rate of 60–65% in stage IIB (Green *et al* 2001), 25–50% in stage IIIB (Montana *et al* 1986, Horiot *et al* 1988), and 20–35% in stage IV disease (Rose *et al* 2011).

Mounting clinical evidence correlates tumor control rate with dose. For instance, a retrospective study, RetroEMBRACE (Tanderup *et al* 2016), initiated by the GEC-ESTRO group, has shown that high risk clinical target volume (CTV_{HR}) dose $\geqslant$85 Gy ($EDQ2_{10}$, D90) delivered in 7 weeks provides a 3-year local control rate >94% in limited-size (20 cm^3), >93% in intermediate-size (30 cm^3), and >86% in large-size CTV_{HR} (70 cm^3) tumors. It also concluded that increasing CTV_{HR} volume by 10 cm^3 requires an additional 5 Gy for equivalent local control. However, a high dose may substantially increase toxicity risks to nearby organs at risk (OARs), such as the rectum, sigmoid, bladder, and vagina. Late rectum morbidity is associated with high rectum volume dose ($D_{2cc} > 75$ Gy) (Mazeron *et al* 2016). Better understanding of the relationship between OAR toxicity and dose is critical for safe dose escalation to improve local control of large-size advanced-stage cervical cancer tumors.

A common limitation of current image-guided EBRT-BT techniques is that they utilize a static image during the computer tomography (CT) simulation scan and ignore anatomic variations throughout the treatment course. Even though such variations are captured by BT fractional CT scans, which are used for BT planning, the hotspots are assumed to be static. Relying on this static assumption, clinicians apply the worst-case addition method to evaluate OARs' $D_{0.1cc}$, D_{1cc}, and D_{2cc} (most exposed 0.1, 1, and 2cm^3 volume) accumulative dose for toxicity prediction. Though the worst-case addition method protects OARs by overestimating OAR dose, it potentially prevents the delivery of possible high dose prescriptions to the target volume. Moreover, the reported dose volume values discard dose distribution information that may be crucial for adaptive BT planning, e.g. local characteristics of dose distribution on the rectum are related to rectum toxicity (Buettner *et al* 2009, Lee *et al* 2012, Wortel *et al* 2015, Drean *et al* 2016).

To study the correlation between received dose and induced OAR toxicity, three important problems must be addressed: (1) how to account for inter-fractional organ deformations and accurately acquire the accumulative dose received by the OARs; (2) how to utilize the OARs' spatial dose distribution information instead of merely using one dimensional (1D)

dose volume parameters such as $D_{0.1/1/2cc}$; and (3) how to build an effective dose-toxicity prediction model with a limited patient sample size.

A typical cervical cancer radiotherapy treatment course is composed of ~25 fractions EBRT treatments and ~4–6 fractions BT treatments. The substantial inter-fractional organ motion that occurs in BT treatment makes reporting the accurate cumulative dose over an entire treatment course a challenging task. It was reported that the OARs intra- and inter-fraction D_{2cc} uncertainties were 20–25% (including a small fraction of 5–11% contouring uncertainties), indicating that organ motion is the major contribution to OAR dose uncertainties (Dinkla *et al* 2013, Lobefalo *et al* 2013, Tanderup *et al* 2013). Intensity-based approaches (e.g. Demons-based) (Thirion 1998, Christensen *et al* 2001, Wang *et al* 2005), may not provide an optimal method for deformable image registration (DIR) in the context of cervical radiotherapy, because of the potential for large deformations in the rectum, bladder, etc, as well as the poor contrast between these OARs and their surrounding tissues. In contrast, the feature-based DIR methods (e.g. finite element model-based) (Brock *et al* 2005, Xiong *et al* 2006, Kaus *et al* 2007, Vasquez Osorio *et al* 2009, Bondar *et al* 2010, Andersen *et al* 2012, Wognum *et al* 2013), aided by organ contours or delineated features, are more favorable for accurate anatomical mapping, especially for hollow organs with substantial deformation, such as the rectum and bladder. Our recent work (Chen *et al* 2016) proposed an improved non-rigid point matching algorithm based on the 'thin plate splines-robust point matching' (TPS-RPM) framework (Chui and Rangarajan 2003). This novel approach was validated on a porcine bladder phantom embedded with fiducial markers as baseline, and satisfactory DIR accuracies of 3.7 ± 1.8 mm and 1.6 ± 0.8 mm were achieved for bladders with large and small deformation. It can be served as a practical tool for accurate surface matching for hollow organs.

Several groups have also investigated spatial dose distribution and rectal toxicity in prostate cancer radiotherapy (Meijer *et al* 1999, Heemsbergen *et al* 2005, Tucker *et al* 2006, Munbodh *et al* 2008, Buettner *et al* 2009, 2012, Wortel *et al* 2015). Most of these reported studies utilized a three-dimensional (3D) to two-dimensional (2D) mapping approach to obtain a rectum surface dose map (RSDM) for spatial dose feature extraction. For example, Munbodh *et al* (2008) sought to identify dosimetric and anatomic indicators of late rectal toxicity by using the RSDM in prostate cancer patients treated with intensity modulated radiation therapy (IMRT). Buettner *et al* (2012) studied the dose response of the anal sphincter region from the constructed RSDM and correlated 3D dose distributions with various side effects. Wortel *et al* (2015) found substantial relationships between acute rectal toxicity and local dose distributions through RSDM in prostate cancer patients who received IMRT and 3D-conformal radiotherapy (3D-CRT). Recently, an expert system based on machine learning technic was developed to address dosimetric uncertainties caused by motion of hollow organs (e.g. rectum, bladder) for prostate cancer radiation therapy(Guidi *et al* 2017). These limited but innovative studies have demonstrated the great potential of employing RSDM for rectum dose-toxicity relationship analysis.

Although most current studies focus on statistical analysis of the underlying relationship between the OARs' toxicity and the extracted 1D dose volume parameters (ICRU 2013), or 2D/3D localized dose distribution features (Wortel *et al* 2015, Drean *et al* 2016), our ultimate interest lies in the possibility of employing the dose distribution information for induced OAR toxicity prediction. The recent revival of deep learning techniques highlighted by the success of deep convolutional neural networks (CNN) (LeCun *et al* 2015) has brought more alternatives and opportunities for dose-toxicity prediction modeling. The CNN first gained popularity in the computer vision community and is now thriving in the medical image processing domain (Tajbakhsh *et al* 2016). However, training the CNN from scratch is difficult not only because of the extensive computational requirements for the training process, but also because

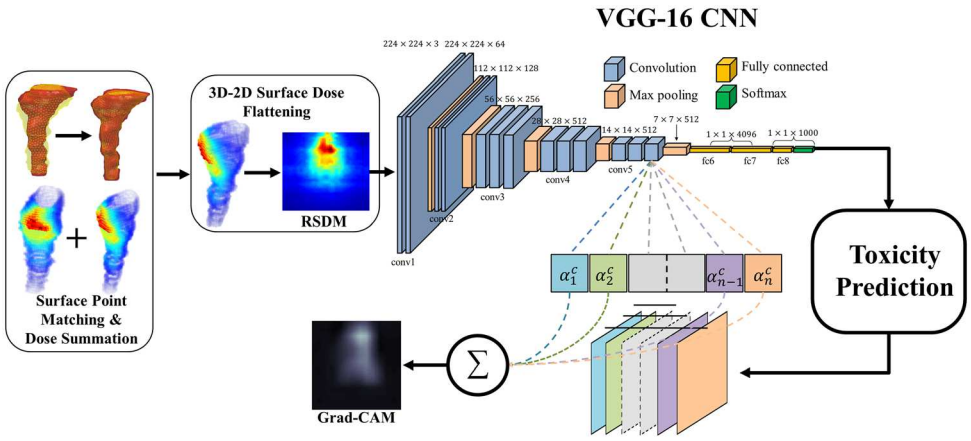

Figure 1. Algorithm workflow.

of the lack of large-scale annotated medical image training datasets and the potential risk of over-fitting. In contrast with training the CNN from scratch, transfer learning fine-tunes the CNN which is pre-trained on a large labeled dataset from a different application. Recently, promising results have been reported that transfer learning can have superior, or at least, the same performance as the CNN trained from scratch for medical image classification (Shin *et al* 2016, Tajbakhsh *et al* 2016).

In this study, we introduced a CNN model to analyze rectum dose distribution and predict rectum toxicity. We adopt a transfer learning strategy to overcome the limited patient data issue by pre-training a VGG-16 CNN on a large-scale natural image database, and then fine-tuned with the patient RSDMs. A gradient-weighted class activation method that highlights the discriminative regions on the RSDM was used to correlate dose distribution and rectum toxicity when the VGG-16 CNN completes a prediction. The general algorithm workflow is depicted in figure 1.

2. Methods and materials

2.1. Patient cohort

We retrospectively collected data of 42 cervical cancer patients treated with 25 fractions (2 Gy per fraction) EBRT followed by 4 (7 Gy per fraction) or 5 fractions (6 Gy per fraction) BT. The collected data include planning images and treatment plans. The patient was scheduled for follow-up examine every 2–3 months after treatment. Patients complaining of hematochezia were further examined by colonoscopy. Patients' rectum toxicity Grades were collected as a clinical parameter in this study, where twelve with Grade $\geqslant$2 rectum toxicity were characterized as toxicity patients and thirty patients with Grade 0–1 toxicity as non-toxicity patients. To account for biologic effects of different fractionation schemes, both the BT and EBRT physical doses were converted to EQD2 doses using a linear quadratic model (Bentzen *et al* 2012) with an α/β ratio of 3 for rectum (Michalski *et al* 2010, Moulton *et al* 2016).

2.2. Rectum surface meshing and DIR

A rectum surface mesh was generated using rectum contours in consecutive CT slices via a particle-based surface meshing approach (Zhong *et al* 2013). Given the initial rectum contour

points, a high-quality isotropic triangular surface meshing was obtained by solving an inter-particle energy function with the quasi-Newton L-BFGS optimizer.

For each patient, the rectum surface meshes in each BT fraction were generated using the physician delineated rectum contours via the above particle-based surface meshing approach. The obtained rectum surface mesh in the first BT fraction was used as a reference, while those rectum surface meshes from other fractions were registered to the reference domain via a recently developed topography-preserved point-matching deformable image registration (TOP-DIR) algorithm (Chen *et al* 2016). The TOP-DIR is a local topography-preserved robust point-matching algorithm designed for accurate dose accumulation on a hollow organ wall. TOP-DIR can find point-to-point correspondence on an organ wall by introducing local anatomic information into the iterative update of correspondence matrix computation in the 'thin plate splines-robust point matching' (TPS-RPM) scheme (Chui and Rangarajan 2003). Once the deformation vector fields (DVFs) on the rectum surface were generated by TOP-DIR, final DVFs defined on each voxel of the dose matrix were estimated by B-spline approximation (Zhen *et al* 2015).

We used the final DVFs to deform and sum all the fractional BT doses to the reference domain (the first BT fraction) to yield a deformable cumulative BT dose. Since patients in the study cohort received a homogenous EBRT dose in the pelvic region, the EBRT dose was added to the BT cumulative dose without deformation. The final accumulative EBRT + BT rectum surface dose was employed for the subsequent dose-toxicity prediction study.

2.3. Rectum unfolding

For each evaluated patient, we generated a 2D RSDM to represent dose distribution on the rectum surface. We used a mapping procedure proposed by Tucker *et al* (2006) to unfold the rectum surface onto a rectangle in a plane. Specifically, the rectum contour centroid on each CT slice was calculated, and n rays were emitted from the centroid at evenly spaced intervals. We recorded the intersection coordinates between the rectal contour and the emanating rays and extracted the dose on each intersection point of the rectum wall. We then cut the rectum at the posterior-most position on all the zCT slices and unfolded it to form a flat rectangular $n \times z$ matrix representing dose distribution on the rectum surface. Rectum unfolding is illustrated in figure 3.

2.4. Convolutional neural network (CNN)

We employed the flattened 2D RSDMs to train the CNN as the prediction model. The CNN is a deep learning architecture with multiple convolutional layers responsible for detecting local features of the inputs. A convolution layer is composed of several convolution kernels that compute different feature maps. To enable local feature detection, each neuron of a feature map is connected to a small neighborhood of outputs from the previous layer. The feature map is generated by convolving the input with a learned kernel and connected to an element-wise nonlinear activation function. A pooling layer is usually placed between two convolution layers to reduce computational complexity and achieve shift-invariance. The pooling layer takes a small neighborhood from the convolutional layer and subsamples it to produce a single output (e.g. average or maximum) from that neighborhood.

For image classification, the unknown network weights W of a given CNN can be learned with a labeled training image set $S = \{(x_i, y_i)\}, i = 1, \ldots, m$ by solving the optimization problem: $\min_W \sum_{i=1}^{m} \ell(f(W; x_i), y_i)$, where x_i is an input image, y_i is the corresponding label,

$f(W; x_i)$ is the prediction with network weight W for input x_i, and ℓ is a negative log-likelihood loss function defined in terms of the normalized soft-max probability. This optimization problem is typically solved by a stochastic gradient descent scheme. In each iteration, given a shuffled fixed size mini-batch $I_t \in S$, the weight update is given by $W^{(t+1)} = W^{(t)} + Z^{(t+1)}$ where $Z^{(t+1)} = \mu Z^{(t)} - \lambda \nabla_W \left(\sum_{i \in \Gamma_t} \ell(f\left(W^{(t)} + \mu Z^{(t)}; x_i\right), y_i) \right)$. The free parameter μ denotes the momentum that indicates the contribution from the previous weight update, and λ denotes the learning rate.

2.5. Transfer learning

Considering the limited number of patient data in this study, training a large CNN from scratch with random initialization would be impractical and could easily lead to over-fitting. As an alternative, we opted to utilize the learned knowledge from a pre-trained network and apply the trained parameters to a new classification task in a process called *transfer learning* (Shin *et al* 2016, Tajbakhsh *et al* 2016). Transfer learning begins by initializing the network with pre-trained weights from a network of the same architecture and then *fine-tunes* the parameters to accommodate the target application. Depending on the class number of the new classification task, the last fully connected layer is usually replaced with as many neurons as the new class number. In this study, we substituted the last fully connected layer with two neurons, corresponding to toxicity and non-toxicity.

We used the state-of-the-art VGG-16 (Simonyan and Zisserman 2014) as our network architecture. This CNN achieved substantially improved performance over the other networks in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 (Russakovsky *et al* 2015). The VGG-16 network (figure 1) is a deep CNN that consists of 16 layers, including 13 convolution layers and 3 fully-connected layers (termed fc6–fc8). All the convolution layers are built with fixed filters with a size of 3×3, and the stride and padding are fixed at 1. There are 5 groups of convolution layers (termed conv1~conv5), each of which is followed by a max pooling layer with a window size of 2×2 that down-samples the images to reduce computational burden and control over-fitting. The filter numbers of conv1–conv5 are 64, 128, 256, 512, and 512. We pre-trained the VGG-16 CNN with the ImageNet (Deng *et al* 2009), which is currently the largest image dataset, with more than 1.2 million natural images of 1000 object categories. These 1000 object categories represent objects that we come across in our day-to-day lives, such as species of dogs, cats, various household objects, vehicle types etc. Given an input test image, the ImageNet-trained VGG-16 CNN generates the test image's probabilities (or label) belonging to each of the 1000 object class. Options to utilize the pre-trained parameters in VGG-16 include: (1) 'shallow tuning': fine-tune only the last few fully connected layers; and (2) 'deep tuning': fine-tune all the network layers. In this study, considering the substantial difference between the source application (natural image) and target application (dose image), we opted to fine-tune all layers in VGG-16 and compare its gains over the 'shallow tuning'.

2.6. Gradient-weighted class activation map (Grad-CAM)

In addition to predicting rectum toxicity using VGG-16 CNN, we would like to understand what features have been learned and where these features are located. To identify and locate the learned features that distinguish toxicity from non-toxicity, we used the gradient-weighted class activation mapping (Grad-CAM) (Ramprasaath *et al* 2016) to highlight

discriminative regions on the RSDM when VGG-16 completes prediction. Denote the last convolutional layer produce K feature maps $A^k \in R^{u\times v}$ of size $u \times v$ and a Grad-CAM map $L^c_{\text{Grad-CAM}} \in R^{u\times v}$. To obtain the Grad-CAM map, the gradient of score y^c with respect to feature maps A^k of a convolutional layer, i.e. $\frac{\partial y^c}{\partial A^k_{ij}}$, is flowed back and global-average-pooled to obtain the weights a^c_k by $a^c_k = 1/Z \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$. The Grad-CAM map $L^c_{\text{Grad-CAM}}$ is calculated as $L^c_{\text{Grad-CAM}} = ReLU\left(\sum_k a^c_k A^k\right)$, where $ReLU(\cdot)$ is the rectified linear unit. The final Grad-CAM map is obtained by up-sampling $L^c_{\text{Grad-CAM}}$ to the size of the input image.

2.7. Strategies to avoid over-fitting

Though training the large VGG-16 from scratch is resolved by utilizing transfer learning, we still need to fine-tune the million-parameter network, which is challenging. To reduce the chance of over-fitting, two issues, including data imbalance and data scarcity, need to be addressed. The training patient cohort in this study is limited and not balanced, where the toxicity cases constitute only a small minority of the data. Learning from imbalanced data is risky because the network may tend to predict more often on the majority non-toxicity class to score an overall high accuracy, and scarify the sensitivity of identifying the minority toxicity cases.

As for data imbalance, we employed the adaptive synthetic sampling approach (ADASYN) (He *et al* 2008) to generate synthetic minority toxicity data to balance the training data set. The ADASYN determines a weighted distribution for different minority class samples according to their level of difficulty in learning. For each minority class sample, the more adjacent majority samples it has in a certain neighboring range, the more synthetic examples will be created. Specifically, given the labeled training set $S = \{(x_i, y_i)\}, i = 1, \ldots, m$ with a minority class $S_{\min} \in S$ and a majority class $S_{maj} \in S$, the number of data needed to synthesize is determined by $G = (m_l - m_s) \times \beta$, where m_l and m_s represent the numbers of minority and majority class samples, respectively, and $\beta \in [0, 1]$ is a parameter used to specify the desired balance level after generating the synthetic data. For each $x_i \in S_{\min}$, we can find K nearest neighbors based on the Euclidean distance and calculate the density distribution $r_i = \frac{\Delta_i/K}{Z}$, $i = 1, \ldots, m_s$, where Δ_i is the number of examples in the K nearest neighbors of x_i that belongs to the majority class, and Z is a normalized constant so that $\sum r_i = 1$. The number of synthetic samples needed for each minority sample $x_i \in S_{\min}$ is given by $g_i = r_i \times G$. For each of the g_i synthetic data of $x_i \in S_{\min}$, one minority example x_{zi} is chosen at random from the K nearest neighbors of x_i, and the synthetic data can be generated by $s_i = x_i + (x_{zi} - x_i) \times \lambda$, where λ is a random number $\lambda \in [0, 1]$.

As for data scarcity, data augmentation was employed to artificially increase the size of the training dataset. Random geometric image transformation, including translation, rotation, scaling, flipping, as well as augmenting image intensity values, such as blurring and noise addition, were applied to the ADASYN balanced data prior to each fine-tuning iteration to increase the training image dataset to 10 times of its original size. This data augmentation strategy has been shown to be helpful for avoiding over-fitting and successful generalization (Ronneberger *et al* 2015, Milletari *et al* 2016, Kayalibay *et al* 2017).

To even further reduce the chance of getting over-fitting, we opted to only 'slightly' fine-tune the VGG-16 CNN with the augmented training dataset, i.e. a minimal number of training epoch of one was used for fine-tuning, intending to maintain the classification capability that mostly learned from the large natural image dataset ImageNet, and also adapt the network with the flexibility to the new application on the dose image.

2.8. Implementation details

For all patients, the rectum was extracted between the level of the ischial tuberosity and the rectosigmoid junction, with rectum length ranging from 6 cm to 9 cm in the patient cohort. For rectum surface meshing, a region of interest (ROI) with a size of $200 \times 200 \times 100$ (234 mm $\times$ 234 mm $\times$ 200 mm) encompassing the rectum was extracted, and the segmented rectums in each BT fraction were converted to surface meshes with 1500 vertices for the subsequent TOP-DIR surface point matching. For rectum surface unfolding, $n = 30$ rays were emitted from the centroid at a spaced interval of 12°. The flattened RSDM had a size of $30 \times z$, where z is a sampling number along the superior-inferior direction. We set $z = 35$, which approximates the resolution of ~0.5 cm in the SI direction. To accommodate the VGG-16 architecture designed for color image inputs, the flattened grey-scale 2D RSDMs were coded to RGB images with a red-blue colormap, where the two extrema, red and blue, correspond to the maximum 90 Gy and the minimum 50 Gy in EQD2 doses (biologic equivalent dose in 2 Gy fractions). The colorized RSDMs were resampled to 224×224 before being fed to VGG-16. For all of the evaluations, a learning rate $\lambda = 0.0001$ were empirically chosen with a training mini-batch size of 5.

The rectum surface meshing was programmed under the Microsoft Visual C++ 2010 platform. The TOP-DIR was implemented on the compute unified device architecture (CUDA) programming environment, and the rectum unfolding and the ADASYN algorithm were coded in Matlab R2015b. The VGG-16 was built on Python 2.7 equipped with two machine learning libraries: Lasagne (Dieleman *et al* 2015) and Theano (Theano Development Team 2016).

2.9. Prediction quantification and comparisons with predictions via $D_{0.1/1/2cc}$ and texture features

To quantitatively evaluate the registration accuracy of the rectum surface, we employed four similarity metrics (Chen *et al* 2015b, 2016): the Dice's coefficient (DC), the percent error (PE), the mean vertex-to-vertex distance (VVD), and the Hausdorff distance (HD). Higher DC or lower PE, VVD, and HD indicate better results.

The prediction performance was quantified by the mean accuracy ($\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})$), sensitivity ($\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$), specificity ($\mathrm{SPE} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$), and the area under the receiver operating characteristic curve (AUC), where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. The repeated stratified 10-fold cross validation (CV) and the leave-one-out cross validation (LOOCV) method were used to assess the prediction performance.

For comparison, we evaluated the advantage of the proposed model over using the dose volume parameters, i.e. $D_{0.1\mathrm{cc}}$, $D_{1\mathrm{cc}}$, and $D_{2\mathrm{cc}}$ (most exposed 0.1, 1, and 2 cm^3 volume), for toxicity prediction. The $D_{0.1/1/2\mathrm{cc}}$ were first computed from the EBRT + BT EQD2 dose by the 'worst-case scenario' (WS) addition method (ICRU 2013), and a logistic regression was performed on the extracted $D_{0.1/1/2\mathrm{cc}}$ to calculate the toxicity probability.

We also extract 43 texture features from the RSDM, including 3 first-order gray level statistical features, 9 gray level co-occurence matrix (GLCM) texture features, 13 gray level run-length matrix (GLRLM) texture features, 13 gray level size zone matrix (GLSZM) texture features, and 5 neighborhood gray-tone difference matrix (NGTDM) texture features (Vallieres *et al* 2015). Statistical analysis was first performed to screen out those features with statistical significance, on which logistic regression was performed to estimate toxicity.

We used the Mann–Whitney test to perform the statistical analyses in this study. Results were considered to be statistically significant if $p < 0.05$.

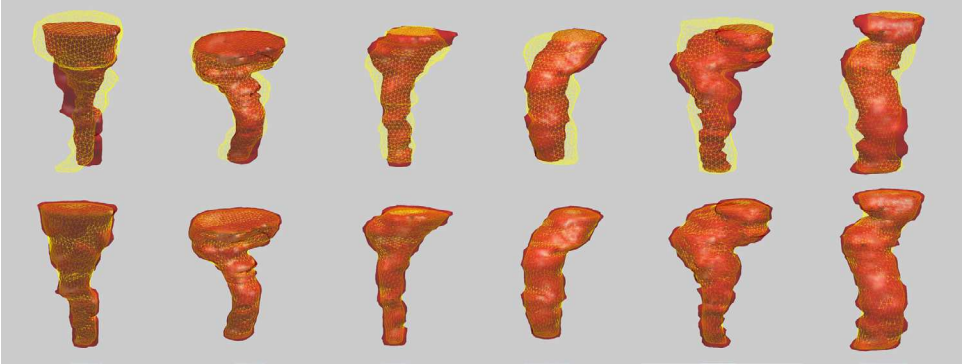

Figure 2. Six example cases of rectum surface meshes before (first row) and after (second row) TOP-DIR. Red: rectum in reference fraction; yellow: rectum in other fractions.

Table 1. The mean (±SD) DC, PE, VVD and HD before and after rectum surface registration by TOP-DIR.

	DC	PE	VVD (mm)	HD (mm)
Before TOP-DIR	0.71 ± 0.04	0.62 ± 0.12	1.5 ± 0.2	7.0 ± 1.3
After TOP-DIR	0.86 ± 0.02	0.26 ± 0.03	0.7 ± 0.1	3.9 ± 0.7

3. Results

3.1. Rectum DIR

Though substantial deformation usually occurs between two fractional rectums, the TOP-DIR can still produce satisfactory registration results (as example cases shown in figure 2). For all 42 evaluated patients, there are 198 BT fractions in total, and 156 DIRs have been performed. The average DC, PE, VVD, and HD before and after TOP-DIR registration over the patient group are summarized in table 1. The mean DC increases from 0.71 to 0.86, and the mean PE, VVD, and HD decrease from 0.62, 1.5 mm, and 7.0 mm to 0.26, 0.7 mm, and 3.9 mm, respectively. These quantitative results indicate that the TOP-DIR based rectum surface registration achieved high DIR accuracy.

3.2. Rectum unfolding

An example case of a rectum surface dose rendered in 3D and the corresponding unfolded 2D RSDM may be seen in figures 3(a) and (b), respectively. The cut is positioned along the posterior-most end of the rectum, and therefore, the horizontal axis of the flattened surface dose map denotes the position along the contour circumference, starting clockwise (view from rectum top to bottom) in the directional order of 'posterior', 'left', 'anterior', 'right', and 'posterior'. The vertical axis lies along the superior and inferior direction.

The averaged RSDMs for patients with and without toxicity are shown in figures 6(b) and (c), respectively. These figures show that the high dose region is at the top part of the rectum in patients who developed toxicity. This observation is analogous to those reported by Heemsbergen *et al* (2005) and Munbodh *et al* (2008), who studied rectum toxicity in prostate cancer patients with EBRT and also observed similar upward shift of the high dose region. We

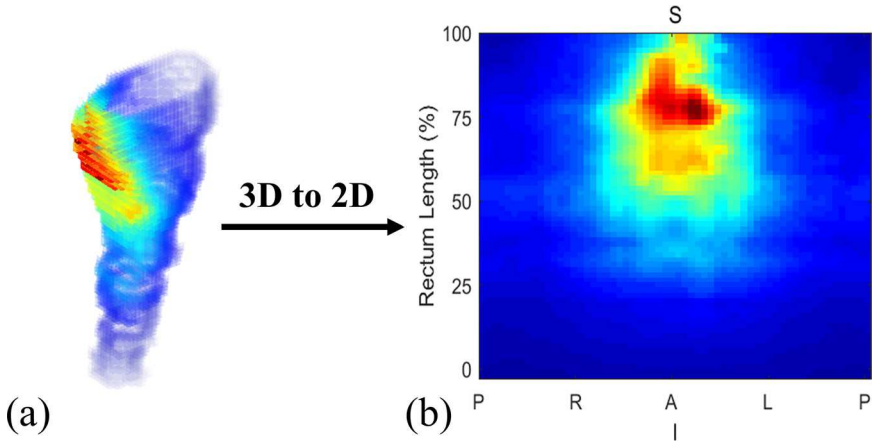

Figure 3. Example of unfolding (a) the 3D rectum surface dose to (b) the 2D RSDM. The abbreviations in (b) indicate the directions of 'superior', 'inferior', 'posterior', 'anterior', 'right', and 'left'.

can also see from figures 6(b) and (c) that the average dose is higher in the high dose region of the toxicity group and covers a larger rectum region. However, as pointed out by Munbodh *et al* (2008), the high dose region in the average RSDM should be interpreted with caution: the larger high dose region found in the average toxicity group RSDM does not necessarily mean that individual RSDMs of each toxicity patient will show a larger high dose region. One possible explanation for this phenomenon is that the high dose regions in the non-toxicity group are more dispersed and therefore appear smaller in the averaged RSDM.

A pixel-wise Mann–Whitney test performed on the RSDMs between the toxicity and non-toxicity groups generated a pixel-wise p-value map; only small p values (< 0.05) are shown in figure 6(a). The region that shows an evident statistical difference ($p < 0.01$, arrow in figure 6(a)) is located on the upper high dose region of the rectum.

3.3. Prediction performance comparisons

Table 2 lists the prediction results by logistic regression on the cumulative $D_{0.1/1/2cc}$ EQD2 doses extracted from the rectum surface via the WS addition method. Though the mean $D_{0.1/1/2cc}$ on the toxicity group are generally higher than the non-toxicity group, the prediction capability of the logistic regression on the $D_{0.1/1/2cc}$ is only moderate, with SEN: 56.7% and 52.8%, SPE: 50% and 60.0%, and AUC: 0.47 and 0.58 in LOOCV and 10-fold CV, respectively. These results suggest that the 1D dose volume parameters $D_{0.1/1/2cc}$, which exclude spatial dose information, are not powerful rectum toxicity predictors.

Forty-three texture features were extracted from the RSDM, and only four features were found to be statistical significant ($p < 0.05$), including two GLCM (contrast, correlation) and two NGTDM (coarseness, complexity), which were further used for toxicity prediction by logistic regression. As summarized in table 3, texture features from the RSDM can generate better predictive performance than the $D_{0.1/1/2cc}$, with SEN: 72.1% and 70.8%, SPE: 59.0% and 58.0%, and AUC: 0.7 and 0.7 for 10-fold CV and LOOCV, respectively. This implied that features from the 2D RSDM might offer more useful dosimetric information to correlate rectal toxicity and dose distribution.

Table 2. The accumulative $D_{0.1/1/2cc}$ EQD2 doses (Gy, mean (μ) ± SD (σ)) obtained by the 'worst-case scenario' addition method and the corresponding prediction by logistic regression.

		Patient Cohort		Predictive model evaluation method	Logistic regression		
		Toxicity	Non-toxicity		SEN	SPE	AUC
μ	$D_{0.1cc}$	85.6 ± 11.9	81.8 ± 8.1	10-fold CV	56.7%	50.0%	0.47
(σ)	D_{1cc}	72.8 ± 8.1	70.2 ± 4.7	LOOCV	52.8%	60.0%	0.58
	D_{2cc}	66.8 ± 6.9	64.4 ± 3.8				

Table 3. Prediction on statistical significant texture features by logistic regression.

Texture features with $p < 0.05$	Predictive model evaluation method	Logistic Regression		
		SEN	SPE	AUC
GLCM(contrast, correlation);	10-fold CV	72.1%	59.0%	0.70
NGTDM (coarseness, complexity)	LOOCV	70.8%	58.0%	0.70

For the propose model, we have compared different fine-tuning strategies and validated the gains of transfer learning over training the VGG-16 from scratch. The comparison results are summarized in table 4 and figures 4 and 5. The VGG-16 network was either 'shallowly' fine-tuned on the last few fully connected layers (e.g. only fc8, or fc7/fc6–fc8), or fine-tuned 'deeper' on all the layers (e.g. conv1–fc8). The VGG-16 performs moderately when trained from scratch or with only the last few fully connected layers fine-tuned. Incremental performance may be observed when more convolutional layers are included in fine-tuning. The ROC analysis for LOOCV (figure 4) also favors the inclusion of more layers for fine-tuning, and similar trends were also observed in 10-fold CV. Furthermore, decreased prediction performances were seen when more training epochs were used (figure 5), implying a tendency of getting over-fitting. Therefore, we opted to fine-tune all layers of VGG-16 with a small training epoch number of one for the rest of the evaluations in this study. Fully fine-tuning all layers of VGG-16 achieves satisfactory prediction results, with an SEN: 61.1% and 75%, SPE: 70% and 83.3%, and AUC: 0.70 and 0.89 respectively for 10-fold CV and LOOCV.

Better performance was seen in the LOOCV than the 10-fold CV, this might attribute to the smaller size of training dataset used in each fold of validation in the10-fold CV, where such performance discrepancy can be evident in a small sample scenario. Though 10-fold CV is a well-accepted validation procedure in the data mining field with large sample size, the LOOCV can also produce unbiased validation when the available data are rare, especially in the medical domain where only limited data samples are available (Refaeilzadeh *et al* 2009).

3.4. Grad-CAM map

In LOOCV, for each testing RSDM input into the final trained VGG-16, a Grad-CAM map is generated along with the output prediction. The Grad-CAM map is resampled to the size of the input image and normalized to [0, 1], with higher values representing greater importance in making the prediction decision. The Grad-CAM map can help understand how the network scores the input RSDM with learned features (e.g. dose distribution pattern) from different regions of the input RSDM. Generally, regions of the Grad-CAM map with larger values

Table 4. Prediction performance by training from scratch or fine-tuning different layers of VGG-16.

	10-fold CV			LOOCV		
Fine-tuning strategy	SEN	SPE	AUC	SEN	SPE	AUC
Train from scratch	55.6%	67.7%	0.67	33.3%	63.3%	0.56
Fine-tune only fc8	41.7%	77.4%	0.66	66.7%	43.3%	0.52
Fine-tune fc7–fc8	41.7%	80.4%	0.66	50.0%	46.0%	0.51
Fine-tune fc6–fc8	43.5%	82.2%	0.71	50.0%	50.0%	0.55
Fine-tune conv1~fc8	**61.1%**	**70%**	**0.70**	**75%**	**83.3%**	**0.89**

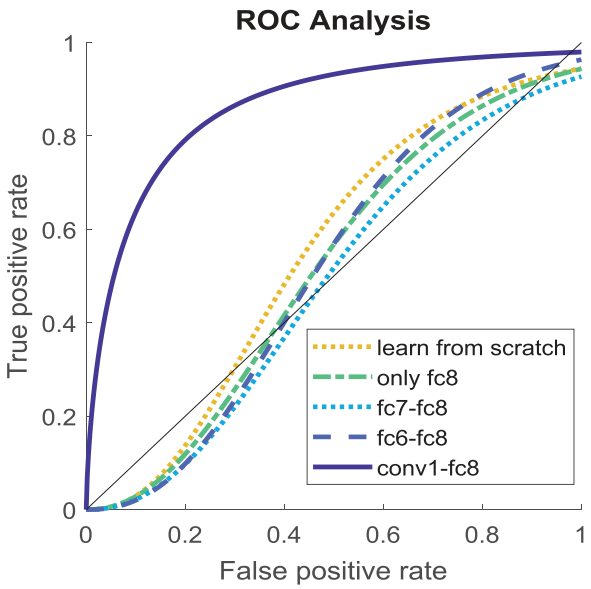

Figure 4. ROC analysis of VGG-16 trained from scratch or fine-tuned with different layers using LOOCV.

correspond to regions in the RSDM whose learned features are more important for toxicity prediction.

Figures 6(d) and (e), respectively, show the average Grad-CAM maps of the toxicity and non-toxicity patients. For the toxicity group, the salient regions of the Grad-CAM are located on the upper rectum, which corresponds to the high dose regions in the RSDM. In contrast, the salient regions in the non-toxicity group are mostly located in the low dose regions in RSDM. The mean dose comparisons of different salient regions between the two groups are shown in figure 6(f). The Grad-CAM maps suggest that the VGG-16 network utilizes learned features from the high and low dose regions in the input RSDM to yield a final prediction score. In addition, the salient regions of the average Grad-CAM of the toxicity group (figure 6(d)) have a similar distribution with the p-value map (figure 6(a)), especially in the regions with small $p < 0.01$ (arrow in figures 6(a) and (d)), suggesting that dose features from these regions in the RSDM are critical for discriminating toxicity from non-toxicity. When comparing the average toxicity Grad-CAM with the average non-toxicity Grad-CAM, the salient regions of the toxicity group may be found mainly in the high dose regions. In contrast, the salient

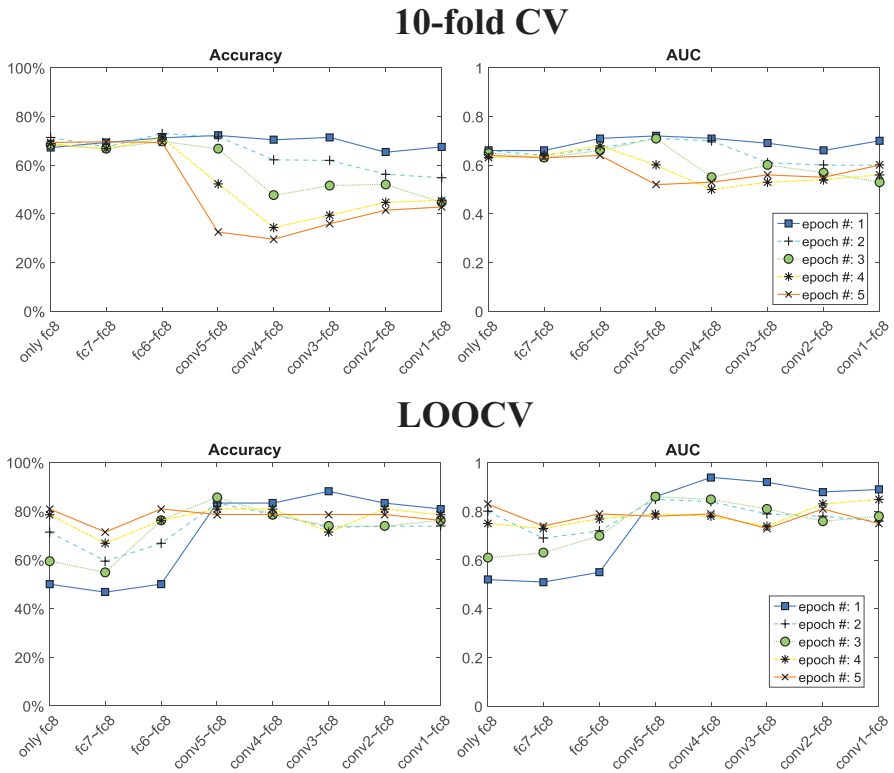

Figure 5. Prediction performance versus fine-tuning on different layers and using different number of training epochs.

regions of the non-toxicity group spread out across the low dose regions. This implies that the VGG-16 CNN used dose features from the high and low dose regions to make prediction decisions, and such features could be geometric features, such as shape, location, area, perimeter (or perhaps their ratios), etc, from the high and low dose regions of the rectum.

4. Discussion and conclusions

In this paper, we have successfully developed a rectum toxicity prediction model by applying a pre-trained CNN network to unfolded RSDMs from cervical cancer patients treated with combined BT and EBRT. A state-of-the-art VGG-16 CNN pre-trained on a large-scale natural image dataset, ImageNet, was used as the network structure and was further fine-tuned on the training RSDMs cohort to adapt to our medical application. Satisfactory prediction results achieved in this work have demonstrated the feasibility of transferring the learned CNN knowledge from natural image to medical image, even though the substantial difference between the two applications suggests that such transfer may be impossible. To the best of our knowledge, this is the first attempt to apply the transfer learning of CNN to radiation dose distribution analysis.

The first step toward estimating an accumulative EBRT + BT dose is to accurately align the underlying anatomy. Therefore, the total EBRT + BT dose is ideally obtained by DIR, which provides means for accumulating dose at tissue voxel level and theoretically can predict

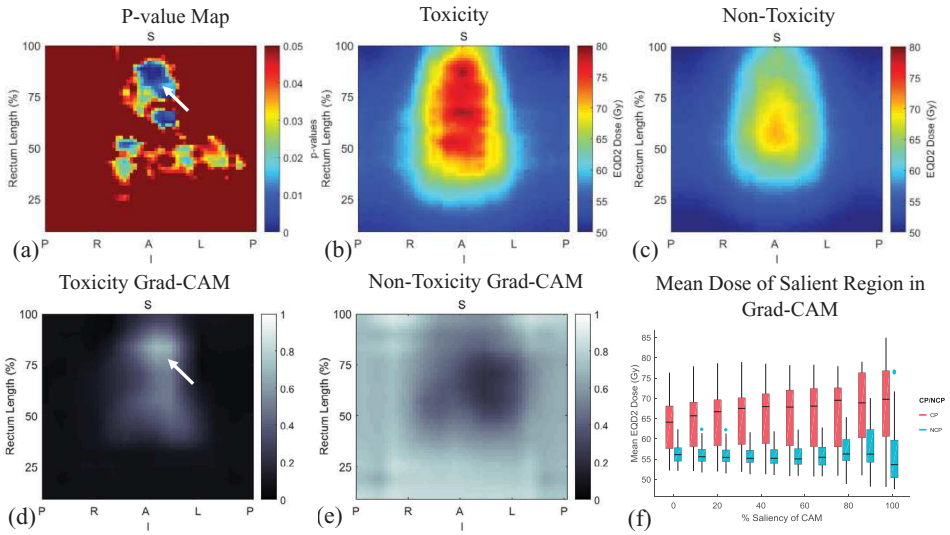

Figure 6. (a): Pixel-wise p-value map (Mann–Whitney test) with small $p < 0.05$ between (b) and (c): the average rectum RSDM of the toxicity and non-toxicity patients. The arrow in (a) indicates regions with $p < 0.01$. (d) and (e): Average Grad-CAM map of the toxicity and non-toxicity groups. The arrow indicates the salient region. (f): Box plot of the mean dose in different salient regions extracted from the Grad-CAM map. The boxes run from the 25th to 75th percentile; the two ends of the whiskers represent the 5% and 95% of the data, and the horizontal line in the box represents the median value. The circles represent outliers.

accurate dose absorption and patient's response to a course of treatment. In our EBRT treatment plan regimen, a homogenous dose distribution (hot spot <107%) often covers the entire pelvic region. Under this scenario, even if the rectum has a large motion, it still receives almost the same dose as planned. On the other hand, DIR between EBRT and BT CTs is a non-trivial task because of the clinical use of the intracavitary applicator in BT. Registering the BT CT image with applicator to the EBRT CT image without applicator (or vice versa) is difficult, if not impossible, since the point-to-point correspondence assumption is usually violated in most DIR algorithms. There are several reported attempts to address this issue (Berendsen *et al* 2013, Vásquez Osorio *et al* 2015), for example, Berendsen *et al* (2013) proposed a DIR with penalty term that minimizes the volume of the missing structure for cervical MR images with and without applicator. Vásquez Osorio *et al* (2015) validated a structure-wise registration with vector field integration to map the largely deformed anatomies between EBRT and BT. However, these novel approaches need comprehensive validations before they can be confidently applied in a clinical setting. Thus adding EBRT to BT without deformation is a good approximation without knowing the uncertainties brought by the EBRT-BT DIR.

The hollow anatomic structure of the rectum makes it possible to take advantage of the CNN, which takes 2D images as input. The flattened 2D RSDM contains the spatial dose distribution information and is intrinsically superior to the 1D dose volume parameters, e.g. $D_{0.1/1/2cc}$, in revealing the dose-toxicity relationship, as shown by the superior prediction performance achieved via the proposed model compared to the prediction using logistic regression on $D_{0.1/1/2cc}$. Furthermore, the proposed model can easily extend its application to other critical organs with hollow structures, such as the bladder, vagina, etc.

In this study, we used VGG-16 as the CNN architecture mainly because of its outstanding performance in the large scale natural images classification challenge (Simonyan and Zisserman 2014). Other state-of-the-art CNN structures, such as CifarNet (Krizhevsky 2009), AlexNet (Krizhevsky *et al* 2012), GoogLeNet (Szegedy *et al* 2014), etc, may yield even better prediction results. Nevertheless, we did not validate these networks, since the purpose of this study was to demonstrate the possibility of transferring learned knowledge between two distinct applications, instead of comparing prediction performances among the currently available CNNs.

Though recent studies have demonstrated the successful application of transfer learning in medical fields with different image modalities, such as x-ray, CT, ultrasound, colonoscopy, etc, (Bar *et al* 2015, Chen *et al* 2015a, Ginneken *et al* 2015, Shin *et al* 2016, Tajbakhsh *et al* 2016), the rationale for its effectiveness is still under investigation. The 'transferability' of knowledge imbedded in a pre-trained network depends on the difference between the two databases: the one on which the CNN was trained and the one to which the learned knowledge will be transferred (Tajbakhsh *et al* 2016). The difference between natural images and textured medical images (e.g. CT, MR) is considerable, and the difference between natural images and radiation dose images is even greater. As a result, fine-tuning all of the CNN layers might be necessary (Tajbakhsh *et al* 2016). The experimental results in this study support the claim that the CNN tends to behave more stably when lower convolutional layers are also fine-tuned.

This work also demonstrates the gain of transfer learning over training from scratch. Given the relatively small patient sample size, training a CNN from scratch is difficult, if not impossible, because a small sample size might not be sufficient to train a large number of network parameters in a deep CNN. For transfer learning, in contrast, the network parameters have been initialized and only need further fine-tuning. Previous studies have shown that transferring features followed by further fine-tuning can result in networks that generalize better than those trained directly on the target dataset (Yosinski *et al* 2014). One thing we should emphasize is that the dose-toxicity prediction model established in this study is preliminary, since limited patient data was used. A large dataset is a necessity to obtain a more powerful and accurate prediction model.

In this study, we also try the Grad-CAM to reveal the possible location of rectal toxicity. We found that the small *p*-value regions on the RSDM (figure 6(a)), which indicate the difference between the toxicity and non-toxicity RSDMs, coincide with the salient regions of the average Grad-CAM of the toxicity group. Each Grad-CAM map was generated by the prediction model along with every input testing RSDM, so the salient region of the averaged Grad-CAM map informs us where those important dose features come from when the prediction model makes classification decisions. The small *p*-value regions on RSDM and the salient region of the toxicity group's Grad-CAM both point to the superior part of the rectum, suggesting that the superior rectum might imply a possible positional explanation for rectum toxicity. This observation accords with those reported by Heemsbergen *et al* (2005) and Munbodh *et al* (2008), who found that the high dose was shifted upward to the top part of the rectum in their studies of rectum toxicity in prostate cancer patients. However, further in-depth investigations are still needed to provide more solid clinical evidence to support the claim that the superior rectum is mainly responsible for rectum toxicity.

Acknowledgments

This work was supported by Varian Medical Systems, Inc. (#OTD-109235), the National Science Foundation (NSF) ACI-1657364, the National Natural Science Foundation of China (81728016 and 81571771). We thank Dr Jonathan Feinberg for editing the manuscript.

References

Al-Mansour Z and Verschraegen C 2010 Locally advanced cervical cancer: what is the standard of care? *Curr. Opin. Oncol.* **22** 503–12

Andersen E S, Muren L P, Sorensen T S, Noe K O, Thor M, Petersen J B, Hoyer M, Bentzen L and Tanderup K 2012 Bladder dose accumulation based on a biomechanical deformable image registration algorithm in volumetric modulated arc therapy for prostate cancer *Phys. Med. Biol.* **57** 7089–100

Bar Y, Diamant I, Wolf L, Lieberman S, Konen E and Greenspan H 2015 Chest pathology detection using deep learning with non-medical training *2015 IEEE 12th Int. Symp. on Biomedical Imaging (ISBI)* pp 294–7

Bentzen S M, Dorr W, Gahbauer R, Howell R W, Joiner M C, Jones B, Jones D T, van der Kogel A J, Wambersie A and Whitmore G 2012 Bioeffect modeling and equieffective dose concepts in radiation oncology—terminology, quantities and units *Radiother. Oncol.* **105** 266–8

Berendsen F F, Kotte A N T J, de Leeuw A A C, Viergever M A and Pluim J P W 2013 Abdominal imaging. computation and clinical applications *Proc., 5th Int. Workshop, Held in Conjunction with MICCAI 2013 (Nagoya, Japan, 22 September 2013)* ed H Yoshida *et al* (Berlin: Springer) pp 136–44

Bondar L, Hoogeman M S, Vasquez Osorio E M and Heijmen B J 2010 A symmetric nonrigid registration method to handle large organ deformations in cervical cancer patients *Med. Phys.* **37** 3760–72

Brock K K, Sharpe M B, Dawson L A, Kim S M and Jaffray D A 2005 Accuracy of finite element model-based multi-organ deformable image registration *Med. Phys.* **32** 1647–59

Buettner F, Gulliford S L, Webb S, Sydes M R, Dearnaley D P and Partridge M 2009 Assessing correlations between the spatial distribution of the dose to the rectal wall and late rectal toxicity after prostate radiotherapy: an analysis of data from the MRC RT01 trial (ISRCTN 47772397) *Phys. Med. Biol.* **54** 6535–48

Buettner F, Gulliford S L, Webb S, Sydes M R, Dearnaley D P and Partridge M 2012 The dose-response of the anal sphincter region—an analysis of data from the MRC RT01 trial *Radiother. Oncol.* **103** 347–52

Chen H, Ni D, Qin J, Li S, Yang X, Wang T and Heng P A 2015a Standard plane localization in fetal ultrasound via domain transferred deep neural networks *IEEE J. Biomed. Health Inf.* **19** 1627–36

Chen H, Zhen X, Gu X, Yan H, Cervino L, Xiao Y and Zhou L 2015b SPARSE: seed point auto-generation for random walks segmentation enhancement in medical inhomogeneous targets delineation of morphological MR and CT images *J. Appl. Clin. Med. Phys.* **16** 387–402

Chen H, Zhong Z, Liao Y, Pompos A, Hrycushko B, Albuquerque K, Zhen X, Zhou L and Gu X 2016 A non-rigid point matching method with local topology preservation for accurate bladder dose summation in high dose rate cervical brachytherapy *Phys. Med. Biol.* **61** 1217–37

Christensen G E *et al* 2001 Image-based dose planning of intracavitary brachytherapy: registration of serial-imaging studies using deformable anatomic templates *Int. J. Radiat. Oncol. Biol. Phys.* **51** 227–43

Chui H and Rangarajan A 2003 A new point matching algorithm for non-rigid registration *Comput. Vis. Image Underst.* **89** 114–41

Deng J, Dong W, Socher R, Li L-J, Li K and Fei-Fei L 2009 ImageNet: a large-scale hierarchical image database *CVPR09* (www.image-net.org/)

Dieleman S *et al* 2015 Lasagne: First release (https://doi.org/10.5281/zenodo.27878)

Dinkla A M, Pieters B R, Koedooder K, Meijnen P, van Wieringen N, van der Laarse R, van der Grient J N, Rasch C R and Bel A 2013 Deviations from the planned dose during 48 hours of stepping source prostate brachytherapy caused by anatomical variations *Radiother. Oncol.* **107** 106–11

Drean G *et al* 2016 Identification of a rectal subregion highly predictive of rectal bleeding in prostate cancer IMRT *Radiother. Oncol.* **119** 388–97

Ginneken B V, Setio A A A, Jacobs C and Ciompi F 2015 Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans *2015 IEEE 12th Int. Symp. on Biomedical Imaging (ISBI)* pp 286–9

Gray H J 2008 Primary management of early stage cervical cancer (IA1-IB) and appropriate selection of adjuvant therapy *J. Natl Compr. Cancer Netw.* **6** 47–52

Green J A, Kirwan J M, Tierney J F, Symonds P, Fresco L, Collingwood M and Williams C J 2001 Survival and recurrence after concomitant chemotherapy and radiotherapy for cancer of the uterine cervix: a systematic review and meta-analysis *Lancet* **358** 781–6

Guidi G, Maffei N, Vecchi C, Gottardi G, Ciarmatori A, Mistretta G M, Mazzeo E, Giacobazzi P, Lohr F and Costi T 2017 Expert system classifier for adaptive radiation therapy in prostate cancer *Australas. Phys. Eng. Sci. Med.* **40** 337–48

Haie-Meder C, Morice P, Castiglione M and Group E G W 2010 Cervical cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up *Ann. Oncol.* **21** v37–40

He H, Bai Y, Garcia E A and Li S 2008 ADASYN: adaptive synthetic sampling approach for imbalanced learning *2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress on Computational Intelligence)* pp 1322–8

Heemsbergen W D, Hoogeman M S, Hart G A, Lebesque J V and Koper P C 2005 Gastrointestinal toxicity and its relation to dose distributions in the anorectal region of prostate cancer patients treated with radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **61** 1011–8

Horiot J C, Pigneux J, Pourquier H, Schraub S, Achille E, Keiling R, Combes P, Rozan R, Vrousos C and Daly N 1988 Radiotherapy alone in carcinoma of the intact uterine cervix according to G H Fletcher guidelines: a French cooperative study of 1383 cases *Int. J. Radiat. Oncol. Biol. Phys.* **14** 605–11

ICRU 2013 ICRU REPORT 89: prescribing, recording, and reporting brachytherapy for cancer of the cervix *J. ICRU* **13** 79–88

Kaus M R, Brock K K, Pekar V, Dawson L A, Nichol A M and Jaffray D A 2007 Assessment of a model-based deformable image registration approach for radiation therapy planning *Int. J. Radiat. Oncol. Biol. Phys.* **68** 572–80

Kayalibay B, Jensen G and van der Smagt P 2017 CNN-based segmentation of medical imaging data *CoRR* (arXiv:1701.03056)

Krizhevsky A 2009 *Learning Multiple Layers of Features from Tiny Images* (Toronto: University of Toronto)

Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* vol 25 (Red Hook, NY: Curran Associates, Inc.) pp 1097–105

LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44

Lee R, Chan E K, Kosztyla R, Liu M and Moiseenko V 2012 Dose-distance metric that predicts late rectal bleeding in patients receiving radical prostate external-beam radiotherapy *Phys. Med. Biol.* **57** 8297–307

Lobefalo F *et al* 2013 Dosimetric impact of inter-observer variability for 3D conformal radiotherapy and volumetric modulated arc therapy: the rectal tumor target definition case *Radiat. Oncol.* **8** 176

Mazeron R *et al* 2016 Dose-volume effect relationships for late rectal morbidity in patients treated with chemoradiation and MRI-guided adaptive brachytherapy for locally advanced cervical cancer: results from the prospective multicenter EMBRACE study *Radiother. Oncol.* **120** 412–9

Meijer G J, van den Brink M, Hoogeman M S, Meinders J and Lebesque J V 1999 Dose-wall histograms and normalized dose-surface histograms for the rectum: a new method to analyze the dose distribution over the rectum in conformal radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **45** 1073–80

Michalski J M, Gay H, Jackson A, Tucker S L and Deasy J O 2010 Radiation dose-volume effects in radiation-induced rectal injury *Int. J. Radiat. Oncol. Biol. Phys.* **76** S123–9

Milletari F, Navab N and Ahmadi S-A 2016 V-Net: fully convolutional neural networks for volumetric medical image segmentation *CoRR* (arXiv:1606.04797)

Montana G S, Fowler W C, Varia M A, Walton L A, Mack Y and Shemanski L 1986 Carcinoma of the cervix, stage III. Results of radiation therapy *Cancer* **57** 148–54

Moulton C R, House M J, Lye V, Tang C I, Krawiec M, Joseph D J, Denham J W and Ebert M A 2016 Prostate external beam radiotherapy combined with high-dose-rate brachytherapy: dose-volume parameters from deformably-registered plans correlate with late gastrointestinal complications *Radiat. Oncol.* **11** 144

Munbodh R, Jackson A, Bauer J, Schmidtlein C R and Zelefsky M J 2008 Dosimetric and anatomic indicators of late rectal toxicity after high-dose intensity modulated radiation therapy for prostate cancer *Med. Phys.* **35** 2137–50

Ramprasaath R S, Abhishek D, Ramakrishna V, Michael C, Devi P and Dhruv B 2016 Grad-CAM: why did you say that? visual explanations from deep networks via gradient-based localization *CVPR 2016* (arXiv:1610.02391)

Refaeilzadeh P, Tang L and Liu H 2009 *Encyclopedia of Database Systems* ed L Liu and M T ÖZsu (Berlin: Springer) pp 532–8

Ronneberger O, Fischer P and Brox T 2015 U-Net: convolutional networks for biomedical image segmentation *CoRR* (arXiv:1505.04597)

Rose P G, Ali S, Whitney C W, Lanciano R and Stehman F B 2011 Outcome of stage IVA cervical cancer patients with disease limited to the pelvis in the era of chemoradiation: a gynecologic oncology group study *Gynecol. Oncol.* **121** 542–5

Russakovsky O *et al* 2015 ImageNet large scale visual recognition challenge *Int. J. Comput. Vis.* **115** 211–52

Shin H C, Roth H R, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D and Summers R M 2016 Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning *IEEE Trans. Med. Imaging* **35** 1285–98

Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition *CoRR* (arXiv:1409.1556)

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2014 Going deeper with convolutions arXiv e-prints (arXiv:1409.4842)

Tajbakhsh N, Shin J Y, Gurudu S R, Hurst R T, Kendall C B, Gotway M B and Liang J 2016 Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging* **35** 1299–312

Tanderup K *et al* 2016 Effect of tumor dose, volume and overall treatment time on local control after radiochemotherapy including MRI guided brachytherapy of locally advanced cervical cancer *Radiother. Oncol.* **120** 441–6

Tanderup K, Nesvacil N, Potter R and Kirisits C 2013 Uncertainties in image guided adaptive cervix cancer brachytherapy: impact on planning and prescription *Radiother. Oncol.* **107** 1–5

Theano Development Team 2016 Theano: a python framework for fast computation of mathematical expressions (arXiv:1605.02688)

Thirion J P 1998 Image matching as a diffusion process: an analogy with Maxwell's demons *Med. Image Anal.* **2** 243–60

Torre L A, Bray F, Siegel R L, Ferlay J, Lortet-Tieulent J and Jemal A 2015 Global cancer statistics, 2012 *CA: Cancer J. Clin.* **65** 87–108

Tucker S L, Zhang M, Dong L, Mohan R, Kuban D and Thames H D 2006 Cluster model analysis of late rectal bleeding after IMRT of prostate cancer: a case-control study *Int. J. Radiat. Oncol. Biol. Phys.* **64** 1255–64

Vallieres M, Freeman C R, Skamene S R and El Naqa I 2015 A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities *Phys. Med. Biol.* **60** 5471–96

Vasquez Osorio E M, Hoogeman M S, Bondar L, Levendag P C and Heijmen B J 2009 A novel flexible framework with automatic feature correspondence optimization for nonrigid registration in radiotherapy *Med. Phys.* **36** 2848–59

Vásquez Osorio E M, Kolkman-Deurloo I-K K, Schuring-Pereira M, Zolnay A, Heijmen B J M and Hoogeman M S 2015 Improving anatomical mapping of complexly deformed anatomy for external beam radiotherapy and brachytherapy dose accumulation in cervical cancer *Med. Phys.* **42** 206–20

Wang H, Dong L, Lii M F, Lee A L, de Crevoisier R, Mohan R, Cox J D, Kuban D A and Cheung R 2005 Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **61** 725–35

Wognum S, Bondar L, Zolnay A G, Chai X, Hulshof M C, Hoogeman M S and Bel A 2013 Control over structure-specific flexibility improves anatomical accuracy for point-based deformable registration in bladder cancer radiotherapy *Med. Phys.* **40** 021702

Wortel R C, Witte M G, van der Heide U A, Pos F J, Lebesque J V, van Herk M, Incrocci L and Heemsbergen W D 2015 Dose-surface maps identifying local dose-effects for acute gastrointestinal toxicity after radiotherapy for prostate cancer *Radiother. Oncol.* **117** 515–20

Xiong L, Viswanathan A, Stewart A J, Haker S, Tempany C M, Chin L M and Cormack R A 2006 Deformable structure registration of bladder through surface mapping *Med. Phys.* **33** 1848–56

Yosinski J, Clune J, Bengio Y and Lipson H 2014 How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems* vol 27 (Red Hook, NY: Curran Associates, Inc.) pp 3320–8

Zhen X, Chen H, Yan H, Zhou L, Mell L K, Yashar C M, Jiang S, Jia X, Gu X and Cervino L 2015 A segmentation and point-matching enhanced efficient deformable image registration method for dose accumulation between HDR CT images *Phys. Med. Biol.* **60** 2981–3002

Zhong Z, Guo X, Wang W, Levy B, Sun F, Liu Y and Mao W 2013 Particle-based anisotropic surface meshing *ACM Trans. Graph.* **32** 1–14