Published text:

Using dose-surface maps to predict radiation-induced rectal bleeding: a neural network approach

Florian Buettner, Sarah L Gulliford, Steve Webb, Mike Partridge

Joint Department of Physics, Institute of Cancer Research and Royal Marsden NHS Foundation Trust, Sutton, Surrey SM2 5PT, UK

E-mail: florian.buettner@icr.ac.uk

Abstract. The incidence of late-toxicities after radiotherapy can be modelled based on the dose delivered to the organ under consideration. Most predictive models reduce the dose distribution to a set of dose-volume parameters and do not take the spatial distribution of the dose into account. The aim of this study was to develop a classifier predicting radiation-induced rectal bleeding using all available information on the dose to the rectal wall. The dose was projected on a two-dimensional dose-surface map (DSM) by virtual rectum-unfolding. These DSMs were used as inputs for a classification method based on locally-connected neural networks. In contrast to fully-connected conventional neural nets, locally-connected nets take the topology of the input into account. In order to train the nets, data from 329 patients from the RT01 trial (ISRCTN 47772397) were split into 10 roughly equal parts. By using 9 of these parts as a training set and the remaining part as an independent test set, a 10-fold cross-validation was performed. Ensemble learning was used and 250 nets were built from randomly selected patients from the training set. Out of these 250 nets an ensemble of expert nets was chosen. The performances of the full ensemble and of the expert ensemble were quantified by using receiver-operator-characteristic (ROC) curves. In order to quantify the predictive power of the shape, ensembles of fully-connected conventional neural nets based on dose-surface histograms (DSHs) were generated and their performances were quantified. The expert ensembles performed better than or equally as well as the full ensembles. The area under the ROC curve for the DSM-based expert ensemble was 0.64. The area under the ROC curve for the DSH-based expert ensemble equalled 0.59. This difference in performance indicates that not only volumetric, but also morphological aspects of the dose distribution are correlated to rectal bleeding after radiotherapy. Thus, the shape of the dose distribution should be taken into account when a predictive model for radiation-induced rectal bleeding is developed.

1. Introduction

Normal-tissue toxicities after radiotherapy are related to a variety of dose and non-dose factors. To date significant efforts have been made to exploit derived relationships and predict the occurrence of complications after radiotherapy. Several studies have correlated dosimetric factors to the incidence of rectal bleeding following pelvic radiotherapy (Boersma et al. 1998, Fiorino et al. 2003, Jackson et al. 2001, Schultheiss et al. 1995, Wachter et al. 2001). Typically, the dose delivered to the rectum is described by measures such as the mean dose, the volume \( V_d \) receiving a dose that exceeds a threshold dose \( d \) or other dose-volume-histogram (DVH) based measures. A number of studies have used uni- and multi-variate analysis to quantify correlations between these measures and different radiation-induced complications. Some authors (Gulliford et al. 2004, Tucker et al. 2006, Chen et al. 2007) have indicated that due to the strong correlations between summary measures like \( V_{55} \) and \( V_{60} \) it is hard to single out the measure with the highest correlation. Gulliford et al. (2004) proposed the use of artificial neural networks, using all bins of a DVH as input, to predict rectal bleeding and nocturia. Different approaches include neural networks with growing/pruning approaches (Chen et al. 2007), the use of the self-organizing map technique to project DVH-data in a low-dimensional space (Chen et al. 2008) and a genetic algorithm for variable selection for modelling radiotherapy outcomes (Gayou et al. 2008). Alternatively, Söhn et al. (2007) proposed the use of principal component analysis and performed a regression analysis based on the first 3 principal components of the DVHs. However, these techniques do not consider the effects of the shape of the dose distribution and morphological information is lost. Tucker et al. (2006) introduced a cluster model which is based on the assumption that the complication probability is correlated with the size of clusters of damaged rectal wall. However, this model reduces the dose distribution to one single parameter and thus only takes the shape of the dose distribution into account to a small degree. More recently, Gianolini et al. (2008) and Munbodh et al. (2007) suggested the extraction of spatial features from DSMs and testing correlations with rectal toxicity. The aim of the present study was to develop a non-linear predictive model for late rectal bleeding which exploits the predictive power of the shape of the dose distribution and makes use of possible synergistic interactions between different aspects of the dose distribution. Therefore, a neural-network-based classifier was generated. As any type of feature extraction is usually accompanied by a loss of information, dose-surface maps (DSMs) were used as the explanatory input-variable for the neural nets. The quality of the classifier was assessed using receiver-operator-characteristic (ROC) curves.

2. Methods and Materials

2.1. Patient cohort and dose-surface maps

The patient database used in this analysis has been described previously (Gulliford et al. in press). It consists of 843 men with localized prostate cancer who were treated
with 3D conformal radiotherapy as part of the MRC RT01 multicentre randomized controlled trial (ISRCTN 47772397). In this trial men were treated with 3D conformal radiotherapy. 421 patients were randomized to receive prostate radiotherapy of 64 Gy and 422 of an escalated dose of 74 Gy. Further details about the implementation and results of the trial can be found in (Sydes et al. 2004) and (Dearnaley et al. 2007). Planning data were available for a subgroup of 388 patients (Gulliford et al. in press). Previous analyses of these data have been performed with focus on the relations between dose, volume and late effects on sexual function (Mangar et al. 2006) as well as a variety of late rectal toxicities (Gulliford et al. in press, Buettner et al. 2009). The endpoint considered in this analysis was late rectal bleeding. Therefore, only patients who were free of rectal bleeding before treatment were considered. Amongst a total of 329 patients, 53 reported late rectal bleeding post-treatment. The RMH grading scheme used in the original trial was dichotomized. Thus, patients who reported no or occasional rectal bleeding were classified versus patients for whom simple outpatient management or transfusion surgery was necessary.

The spatial distribution of the dose was used as the input for the neural-network-based classifier in the form of dose-surface maps. These DSMs reflect the dose delivered to the surface of the rectal wall and are a well-known tool for analyzing radiation-induced rectal toxicity as well as organ motion (Sanchez-Nieto et al. 2001, Hoogeman et al. 2004, Tucker et al. 2006). Different algorithms to generate DSMs from the 3D dose distribution in the rectum exist (Sanchez-Nieto et al. 2001, Hoogeman et al. 2004, Tucker et al. 2006, Munbodh et al. 2008). As a slice-wise unfolding technique has been used successfully for analysing the shape of rectal-surface dose distributions (Gianolini et al. 2008) and resulted in intuitively interpretable maps we chose to apply it in this study. Thus, DSMs were constructed by virtually unfolding the rectum in a slice-wise manner following the methods reported in Sanchez-Nieto et al. (2001),Booth (2002) and Tucker et al. (2006): For every CT slice, the dose at 21 points, which were equally spaced along the contour, was determined by interpolation. This number of points was chosen as it corresponded to the resolution of the dose calculation-algorithms (5 mm) used in the RT01 study. Then, the contour was cut at the posterior-most position on each CT-slice and unwrapped to a flat $21 \times \nu$ array, with $\nu$ being the number of outlined slices. These maps were normalized in the longitudinal direction by interpolation to maps of $21 \times 21$ pixels, as the rectum had been outlined typically over a length of 20-22 slices of 0.5 cm. This was implemented using in-house software GUINESS (Mangar et al. 2006) and MATLAB (MathWorks, Inc., Natick, MA).

2.2. Feed-forward neural networks

An artificial neural network is a non-linear statistical model which extracts linear combinations of the inputs $X$ and uses these features in order to model the outcome
Using dose-surface maps to predict rectal bleeding

\[ f(X) \] as a non-linear function. Usually this model is represented by a diagram as in figure 1. Each node in the hidden and output layers performs a weighted sum over the inputs from the nodes in the previous layer, which is then passed to an activation function \( \sigma \). For increased flexibility a bias can be added to the weighted sum before passing it to \( \sigma \) (Hornik 1993). The activation function used in this study is the hyperbolic tangent sigmoid function

\[
\sigma(\nu) = \frac{2}{1 + \exp(-2 \cdot \nu)} - 1
\]

where \( \nu \) is the input to the function.

For two-class classification problems usually neural networks with two output-nodes are used, each node representing one class (Hastie et al. 2003). As the classes 'rectal bleeding' and 'not rectal bleeding' are mutually exclusive, it is also possible to use a less complex model with only one output node. However, for the specific case of using neural networks for predicting radiation-induced rectal bleeding based on dose-information, Gulliford et al. (2004) have shown that using 2 output nodes instead of one output node leads to superior results. For this reason the neural nets in this study were constructed with 2 output nodes. The output node corresponding to the known outcome is assigned a value of one while the other is zero. The error \( R \) was calculated using the mean log squared error

\[
R = \frac{1}{2} \sum_{k=1}^{2} \sum_{i=1}^{\eta} (\log(1 + \frac{(y_{ki} - f_k(x_i))^2}{2}))
\]

where \( \eta \) is the number of cases used for training and \( x_i \) the input-vector, i.e. the pixels of the DSM of the \( i \)th patient in the training set. \( y_{ki} \) encode the true outcome of the \( i \)th patient such that for patients who reported grade 2 rectal bleeding \( y_{1i} = 1 \) and \( y_{1i} = 0 \) for all other patients and vice versa for \( y_{2i} \). \( f_k(x_i) \) is the outcome for the \( i \)th case as predicted by the net. The error measure was chosen as it is more robust to outliers than the simple least mean-square measure (Liano 1996). Although \( R \) is non-convex and possesses multiple local minima, this does not affect the final classification as the latter is the average over a number of nets trained with different starting weights and training data (Hastie et al. 2003).

The minimization of the error was performed by using the back-propagation learning procedure (Rumelhart et al. 1986) which iteratively updates the weights and biases \( \beta_k \). At the \((r + 1)\)th iteration the updated parameters are calculated as follows:

\[
\beta_k^{(r+1)} = \beta_k^{(r)} + m(\beta_k^{(r)} - \beta_k^{(r-1)}) - (1 - m)\gamma \frac{\partial R}{\partial \beta_k^{(r)}}
\]

\( \beta_k \) denote the parameters in the net with \( k = 1, \ldots K \), \( K \) being the total number of weights and biases in the net. \( \gamma \) is the learning rate and \( m \) a momentum parameter which was used in order to accelerate the learning (Vogl et al. 1988). Learning rate and momentum were varied over a wide range to find the best training parameters.
Using dose-surface maps to predict rectal bleeding

A subset of $M$ patients from the total number of $N = 329$ patients was used for training, the remaining $N-M$ patients were used for independent testing. For the training of the net only a subset of $m$ cases from the total number of $M$ cases available in the training set was used. These cases form the training-construction set and the $M-m$ cases in the training set which were not used for the actual training of the net form the training-validation set. Details on the selection of this subset are provided in section 2.5. In order to prevent over-fitting an early stopping mechanism was implemented using the training-validation set. 20% of the patients in the training-validation set who did not report rectal bleeding and 75% of the patients in the training-validation set who reported rectal bleeding were randomly selected and the validation-error of these cases was computed using equation (2). This was repeated 100 times with different patients sampled from the training-validation set and the mean validation-error was determined. This error gives a more realistic estimate of the error of the classifier than the training-error of the $m$ cases in the training-construction set and was used to monitor the evolution of the error during training. Thus, the classifier was trained by minimizing the error of the training-construction set as long as the error of the training-validation set continued decreasing.

![Figure 1. Typical feed-forward neural network.](image)

**Figure 1.** Typical feed-forward neural network. $X^1 - X^5$ denote the 5 inputs; the arrows between the nodes are associated with weights. $\Sigma$ denotes the weighted sum over the inputs and the respective bias $b_i$ corresponding to the $i$th node. $\sigma$ is the non-linear activation function and $f(X)$ the output of the net.

2.2.1. *Locally-connected networks* Conventional neural nets consist of one input layer, one hidden layer and one output layer. As shown in figure 1, every node in the input layer is connected to each node in the hidden layer. When classifying spatial objects such as DSMs or other images, which often consist of hundreds of pixels, this technique has various disadvantages. First, due to the large number of pixels, which can exceed several hundreds, fully-connected networks have a very large number of weights. When a network with one hidden layer of 100 nodes is used, ten thousands of weights have to
be learned and a very large number of training cases is needed in order to be able to
effectively search the space of solutions. Secondly, a fully-connected architecture ignores
the topology of the input entirely as it can be presented in any fixed order. A solution
to this problem is to use so-called locally-connected neural nets. Local connectivity
means that each node in the hidden layer is only connected to a small number of nodes
in the previous layer. As these recipient fields of the hidden nodes are local, not only
is the number of weights considerably reduced, but also local features are extracted.
A possible enhancement of the locally-connected neural nets is weight sharing. In this
technique the same operation is performed on different parts of the image by using the
same weights. The motivation for this approach is that feature detectors (i.e. the set of
weights corresponding to one node in a hidden layer) that are useful in one part of the
image are also useful in a different part of the image. Therefore, the input nodes in the
hidden layers are organized in planes. All nodes within each plane share the same set of
weights and therefore perform the same feature-extracting operation on different parts
of the image. By constructing a set of planes corresponding to different sets of weights,
different features can be extracted from the image. The outputs of all nodes in such a
plane are often referred to as feature map. The feature maps of the first hidden layer
can be used as inputs for a second hidden layer constructed by the same technique in
order to extract higher-order features. The nodes in the second hidden layer are then
fully-connected to the output nodes. This type of specialised architecture can further
improve the performance of the classifier (LeCun et al. 1989, LeCun et al. 1998, Hastie
et al. 2003).

When using shared weights, the back-propagation algorithm has to be adapted. While
in networks without weight-sharing the weights and biases are adapted iteratively by
calculating the partial derivative of the loss function $\frac{\partial R}{\partial \beta_k}$ (see eq. 2.2), in nets with shared
weights the sum of the partial derivatives of the connections sharing the parameter $\beta_k$
has to be calculated:

$$\frac{\partial R}{\partial \beta_k} = \sum_{(i,j) \in V_k} \frac{\partial R}{\partial u_{ij}}$$

with $u_{ij}$ being the weight of the link between the nodes $i$ and $j$. $V_k$ is the set of paired
indices $(i, j)$ such that

$$u_{ij} = \beta_k \ \forall (i, j) \in V_k$$

Different architectures of locally-connected nets were evaluated. Each locally-
connected net had 2 hidden layers. In total 4 locally-connected nets were implemented,
one of them using weight-sharing. All nodes in the second hidden layer were fully-
connected with 2 output nodes (Hastie et al. 2003) which represent the respective classes
rectal bleeding and no rectal bleeding.
The architectures are illustrated in figure 2 and were designed as follows:

- **Architecture 1**: The hidden layers were designed as a $21 \times 10$ array of nodes and
  a $10 \times 10$ array of nodes respectively. The input layer was connected row-wise to
  the first hidden layer so that 3 neighbouring pixels were linked to each node in
the hidden layer. The nodes of the first hidden layer were linked column-wise to the nodes in the second hidden layer in groups of 3 nodes. The receptive fields of adjacent nodes in the hidden layers overlapped by one node/pixel.

- **Architecture 2:** The hidden layers were designed as a $10 \times 10$ and a $3 \times 3$ array of nodes respectively. The pixels in the input layer were linked to the nodes in first hidden layer in groups of $3 \times 3$ pixels; the nodes in the first hidden layer were linked to nodes in the second hidden layer in groups of $4 \times 4$ nodes. The receptive fields of adjacent nodes in the hidden layers overlapped by one node/pixel.

- **Architecture 3:** The hidden layers were designed as a $7 \times 7$ and a $1 \times 7$ array of nodes respectively. The nodes in the input layer were linked to the first hidden layer in groups of $3 \times 3$ pixels. These nodes were linked row-wise to the nodes in the second hidden layer. The receptive fields of adjacent nodes in the hidden layers had no overlap.

- **Architecture 4:** The first hidden layer had four $10 \times 10$ arrays and the second hidden layer had ten $3 \times 3$ arrays. The nodes in the 4 planes of the first hidden layer were connected to the input nodes as in architecture 2. Each node in a plane had the same set of weights so that the same operation was performed on every patch of $3 \times 3$ pixels in the input layers. The resulting feature maps were connected to the planes in the second hidden layer such that it had input connections from one or two feature maps in the previous hidden layer. The receptive fields of each node in the second hidden layer were composed of one or two $4 \times 4$ patches of nodes located at the same position in the respective planes in the first hidden layer. The connections between the hidden layers are listed in table 1.

**Table 1.** Connections between the 4 planes in the first hidden layer and the 10 planes of the second hidden layer of architecture 4. If a plane in the second hidden layer takes inputs from a plane in the first hidden layer, the respective plane is marked with a cross.

<table>
<thead>
<tr>
<th>Hidden layer 1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hidden layer 2</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>4</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

These architectures were chosen as they correspond to different numbers of weights varying over a broad range between 500 and 1500, while not spanning an unreasonably huge space of solutions. The different connectivities are related to features extracted from different types of neighbourhoods. Thus, architecture 1 realizes a row-wise feature extraction, while architecture 2 extracts the first order features from neighbourhoods of $3 \times 3$ pixels with overlap and architecture 3 without overlap. The preference for
the cranio-caudal direction in architectures 1 and 3 was motivated by evidence for the relevance of the longitudinal position of the dose distribution reported in previous studies (Heemsbergen et al. 2005).

2.2.2. Conventional fully-connected ANNs

Conventional fully-connected neural nets were also constructed and their performance was compared to locally-connected neural nets. The DSMs were thresholded at 35 dose levels between 5 Gy and 73 Gy and the fraction of the DSM receiving more than the threshold dose was calculated. Thus, 35 inputs $S_i$, $i = 1, 2, \ldots, 35$, were determined for the threshold doses $\alpha$ [Gy], $\alpha \in \{5, 7, \ldots, 73\}$. The $S_i$ can be interpreted as the bins of DSM-based DSHs. These inputs were connected to 30 nodes in a hidden layer. Again, the output layer consisted of 2 nodes.

All neural nets were implemented using the programming language C, the R project for statistical computing (R Development Core Team 2007) and the package AMORE (Limas et al. 2007).

2.3. Data partitioning for training and testing: cross-validation

Cross-validation is a widely used technique in order to perform a representative testing of predictive models (Hastie et al. 2003). In this study a 10-fold cross-validation was performed and the data were split into 10 roughly equal sets. Then, for $l = 1, \ldots, 10$ a model was fitted to the 9 other parts of the data forming the training set and the error on the $l$th part of the data forming the test set was calculated. The generalization error of the model can be determined by combining these errors.

2.4. Validation of the presented classifier

In order to assess the performance of the classifiers, an ROC-analysis was performed. The ROC curve plots the sensitivity of a classifier versus 1-specificity and describes the performance of a classifier more thoroughly than a simple mis-classification-rate. In order to quantify the performance of the classifiers presented in this work the area under the ROC curve (AUC) was used. By comparing the performances of the DSH-based and the DSM-based classifiers the predictive power which lies in the spatial information of the dose distribution was quantified.

In order to further assess the performance of the classifiers, the effect of the splitting of the data in 10 parts was investigated. Therefore the cross-validation described in section 2.3 was repeated 100 times with different compositions of the 10 sets.

2.5. Ensemble learning

In machine learning, ensemble methods have proven to be an effective approach to make classifiers more efficient (Freund & Schapire 1996, Breiman 2001, Dietterich 2003):
Using dose-surface maps to predict rectal bleeding

(a) Architecture 1: A 21 \times 10 \text{ array} and a 10 \times 10 \text{ array}.

(b) Architecture 2: A 10 \times 10 \text{ array} and a 3 \times 3 \text{ array}.

(c) Architecture 3: A 7 \times 7 \text{ array} and a 1 \times 7 \text{ array}.

(d) Architecture 4: A locally-connected net with shared weights. The hidden layers consisted of 4 10 \times 10 \text{ arrays} and 10 3 \times 3 \text{ arrays}.

**Figure 2.** The architectures of 4 locally-connected neural nets are presented. The lines describe the connectivity of the different architectures. For example, in architecture 2 the patch of top-left 3 \times 3 pixels from the DSM is connected to the top-left node in the first hidden layer with 9 different weights. The patch of bottom-right 4 \times 4 nodes in the first hidden layer is connected to the bottom-left node in the second hidden layer with 16 different weights. The second hidden layer is fully-connected to the 2 output nodes with 18 different weights.
Using dose-surface maps to predict rectal bleeding

Instead of using only one classifier to find one best hypothesis \( f \) and output solution of the learning problem, a set of \( n \) hypotheses \( f_i \), with \( i = 1, \ldots, n \) is generated. These hypotheses are combined by voting for an outcome \( f_i(X) \) of an input \( X \). In this study the hypotheses \( f_i \) were aggregated to a combined hypothesis \( f \) by simple summation over the votes: \( f(X) = \sum_{i=1}^{n} f_i(X) \). The use of ensembles can help to overcome several problems which often arise in supervised learning (Dietterich 2003). Thus, when only a small number of cases is available for training and the space in which the classifier searches for solutions is very large (e.g. span by a large number of weights in a neural net), it is possible that several solutions yield similar performances on a training set. Instead of choosing one of these solutions to predict future data-points, a better performance can be achieved when the classification is based on a vote from all equivalent solutions. Furthermore, the parameter space of the neural net algorithm is so large that it is not possible to perform an exhaustive search which would guarantee the optimal solution and methods used for optimizing the net can get caught in a local minimum. Using a combination of solutions reduces the risk of choosing an unrepresentative local minimum. Finally, it is possible that in the space of solutions no hypothesis exists which describes the problem well and only weak classifiers can be constructed. However, it has been shown that the aggregation of many weak classifiers can result in a strong classifier with a good performance (Breiman 2001).

In this work, an ensemble consisting of 250 of artificial neural networks was built. Randomness was injected into the different members of the ensemble by setting the initial weights of the net randomly. Also, each member of the ensemble was trained with different input-data by sampling from the training set. As described in section 2.2 these input-data formed the training-construction set. As the incidence rate of rectal bleeding was considerably lower than 50%, this imbalance in the class prior was addressed by choosing the training-construction set as follows: 20% of the patients in the training set who did not report rectal bleeding and 75% of the patients in the training set who reported rectal bleeding were randomly selected for training.

To investigate potential improvement of the performance of the ensemble, a subset of nets with a good ability to generalize was selected and the performance of this expert-ensemble was compared to the performance of the full ensemble of 250 nets. As a selection of too small a number of expert-nets would counteract the benefits from ensemble-learning as described above, the expert-ensemble was constrained to consist of at least 50 nets. The nets forming the expert ensemble were chosen according to the following algorithm:

(i) for \( i=1 \) to 3

(a) randomly select 30% of the patients having reported rectal bleeding and 10% of the patients not having reported rectal bleeding from the training set

(b) use the first member \( f_1 \) in the ensemble to predict the outcomes, calculate AUC and and set \( \text{AUC}_{\text{max}} = \text{AUC} \)
Using dose-surface maps to predict rectal bleeding

(c) declare it first member of the $i$th expert ensemble
(d) for $j=2$ to 250
1. use $j$th member $f_j$ in the ensemble to predict the outcomes
2. add these predictions to the predictions of the members already in the $i$th expert ensemble
3. calculate AUC
4. if $AUC \geq AUC_{max}$ declare the $j$th member of the ensemble a member of the $i$th expert ensemble and set $AUC_{max} = AUC$

(ii) select nets which occur in all 3 expert ensembles and declare them the final expert ensemble
(iii) if this number is less than 50 repeat steps (i) and (ii).

The parameters in the selection-algorithm were chosen such that at least 50 nets could be selected which not only performed well on the training-construction set but also on other randomly-selected subsets of patients from the training set.

3. Results

3.1. 10-fold cross validation using DSMs

The training parameters of each architecture were optimized independently by varying them over a wide range. The optimal values are listed in table 2.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>learning rate</th>
<th>momentum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$10^{-3}$</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>$10^{-4}$</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>$10^{-4}$</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>$10^{-5}$</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 2. Optimal training parameters for the different architectures

The performance of each architecture was evaluated by calculating the areas under the ROC curves. In figure 3 the dependence of the performance on the number of neural nets in the ensemble is illustrated for architecture 2. First, the performance increases when new members are added to the ensemble, then at 100 members starts to converge to an optimum. In table 3 the number of weights are listed together with the respective performances of the nets. The ROC curve of architecture 2 is shown in figure 4 for the expert ensemble. The AUCs between the different folds varied with a standard deviation of 0.04 when expert selection was performed. Without expert selection the standard deviation was 0.07.
Table 3. Performances of all locally-connected nets. AUC_{all} denotes the AUC calculated from all nets in the ensemble, AUC_{exp} the AUC derived from the experts only.

<table>
<thead>
<tr>
<th>Architecture</th>
<th># hidden and output nodes</th>
<th># weights and biases†</th>
<th>AUC_{all}</th>
<th>AUC_{exp}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>312</td>
<td>1442</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>2</td>
<td>111</td>
<td>1173</td>
<td>0.61</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>58</td>
<td>562</td>
<td>0.59</td>
<td>0.62</td>
</tr>
<tr>
<td>4</td>
<td>492</td>
<td>964</td>
<td>0.56</td>
<td>0.57</td>
</tr>
</tbody>
</table>

† The number of unique weights and biases was calculated by adding up the unique connections between the nodes: 1442 = 210 × 3 + 100 × 3 + 2 × 100 + 312; 1173 = 100 × 9 + 9 × 16 + 2 × 9 + 111; 562 = 49 × 9 + 7 × 7 + 2 × 7 + 58; 964 = 4 × 9 + 16 × 4 + 32 × 6 + 2 × 90 + 492

Figure 3. AUC depending on the number of nets in the ensemble. The classifiers were built by aggregating a varying numbers of nets. All nets were of architecture 2.
Figure 4. ROC curve of the architecture 2 with expert selection. A horizontal averaging over the different folds of the cross validation was performed. The error bars indicate the standard deviation between the different runs.

3.2. 10-fold cross-validation using DSHs

The same split of the data which was used to obtain the results presented in section 3.1 was used to train DSH-based nets. A learning rate of $10^{-5}$ and a momentum of 0.2 were used. The ROC curve of the DSH-based expert ensemble is illustrated in figure 5. The associated area under this curve is 0.59. Without expert selection the same AUC was achieved. The standard deviation of the AUCs of the different folds was 0.09 with expert selection and 0.11 without expert selection.
Using dose-surface maps to predict rectal bleeding

3.3. Impact of data partitioning

In order to quantify the effect of assigning a patient to one of the 10 groups, the data partitioning was repeated 100 times randomly resulting in different compositions of the sets. Each of these 100 datasets was used for a 10-fold cross-validation using architecture 2 based on DSMs and the respective AUC of the classifier was calculated. This resulted in a mean AUC of 0.635 ± 0.018.

4. Discussion

A classification method was presented that was able to predict the incidence of radiation-induced rectal bleeding based on the dose to the rectal wall with an area under the ROC curve of 0.64 for our dataset. In contrast to other predictive models, no feature extraction, such as reducing the dose distribution to DVHs or feature selection was performed. Instead, the neural nets were trained to extract information from a complete representation of the dose to the rectal wall in form of DSMs. Using 10-fold cross-validation, a classifier with an area under the ROC curve of 0.64 was constructed. Although there is a considerable variation of the ROC curves between the different folds,
Using dose-surface maps to predict rectal bleeding

the mean AUC remained constant for different splits of the data. This was related to the fact that the dose was the only information available for the classifier. It is well-known that non-dose factors such as age, diabetes, neo-adjuvant androgen deprivation, previous abdominal/pelvic surgery and a genetic dispositions to a high/low radiosensitivity can impact the incidence of rectal bleeding after radiotherapy (Skwarchuk et al. 2000, Fiorino et al. 2008, Sanguineti et al. 2002). As the number of patients who reported rectal bleeding in the test sets of each fold could be as low as 4, the mis-classification of one or two cases due to one of these confounding factors results in a considerable change of the performance of the classifier on that specific test-set and thus a high variability of performance between the different folds of the cross-validation. The high variance is also an inherent problem of the cross-validation technique (Hastie et al. 2003). On one hand, a high number of folds leads to an unbiased estimation of the true prediction error, but results in a high variance. On the other hand, a low number of folds is related to a low variance, but due to the smaller size of the training sets the error estimation will be biased upward. However, as the mean AUC hardly depends on a specific split of the data into 10 sets, it can be considered a representative measure of performance.

An expert selection algorithm was implemented which resulted in equal or improved performance for all implemented architectures. Furthermore it resulted in more stable classifiers yielding less variation of the AUCs between different runs of the cross-validation.

The performance of a classifier based on DSH information only was systematically worse than the performance of the DSM-based classifier. It yielded not only a lower AUC, but also a higher variation of the AUC between different runs of the cross-validation indicating a worse ability of generalization. The increased predictive power of the classifier after including morphological information in the form of the DSMs indicates the predictive power that is inherent in the shape of the dose distribution. As the DSHs were generated based on the DSMs, the difference in the AUCs can be interpreted as an absolute measure for this predictive power. Furthermore the additional information which is gained by considering the shape is being reflected in the lower variance of the ROC curves. However, when the DSM is used as input, the very high number of input-nodes leads to a big space of solutions which the nets have to search. The dependence of the performance on the number of weights (table 3) shows the importance of an adapted architecture of the neural nets: when too many weights span the space of solutions, the additional morphological information the inputs contain cannot be used optimally. Thus, while there is evidence for the predictive power of the spatial distribution of the dose, it is a non-trivial task to process it in a classifier such that this is reflected in the performance. Although the use of DSMs can result in an increase of the performance of a classifier, the error rates are still too high to consider using this technique in a clinical context. Also, if the outputs of the nets were to be used clinically, a rescaling of the outputs would be necessary. As during training the imbalance in the class prior was addressed by choosing certain proportions from each class, the outputs have to be corrected with the corresponding factors. However, this linear transformation does
Using dose-surface maps to predict rectal bleeding

not effect the ROC analysis which is why no correction was made here. For practical applications of the technique it can be desirable to interpret the outputs as probabilities. This can be achieved by using either only one logistic output node or by normalizing the 2 outputs, e.g. by using the softmax function.

The relatively weak performance yielded with the shared-weights technique might be due to the fact that within the DSMs different features are important on different parts of the map. Thus, while a specific feature such as the gradient of the dose might be of some importance on one part of the map, a different feature such as the mean or maximum dose might be more important elsewhere.

It is important to note that this use of DSMs has some limitations. First, different algorithms for projecting the dose to the rectal wall onto a 2D map exist, so that the same dose distribution can result in different maps. However, the slice-wise unfolding technique used for this study has been being used successfully by different groups over the past years (Sanchez-Nieto et al. 2001, Booth 2002, Tucker et al. 2006, Gianolini et al. 2008). Second, organ motion during treatment is not considered. Thus, it is possible that, due to uncertainties in organ positions, the actual dose delivered to the rectal wall differs from the dose calculated based on the planning CT scan.

5. Conclusion

A robust and effective predictive model of rectal bleeding as a result of radiotherapy for prostate cancer has been developed. The predictor based on DSH-information only is inferior to the predictor based on the complete 2D dose map of the rectum. The addition of morphological information in the form of a DSM improved the predictive power which was shown by an ROC analysis. Thus, the DSM-based classification method yielded an area under the ROC curve of 0.64, while the DSH-based classifier resulted in an AUC of 0.59.

Acknowledgments

We acknowledge NHS funding to the NIHR Biomedical Research Centre. We would like to thank the Medical Research Council, especially Matthew Sydes and Rachel Morgan, for kindly providing us follow-up data from the RT01 trial (ISRCTN 47772397) and helpful discussions on the data-set. We would also like to thank Prof. David Dearnaley for insightful discussions on the clinical aspects of this work.

References


Booth J 2002 Modelling the impact of treatment uncertainties in radiotherapy PhD thesis Department of Physics and Mathematical Physics, University of Adelaide.


Buettner F, Gulliford S L, Webb S, Sydes M, Dearnaley D P & Partridge M 2009 Quantifying Correlations between the spatial distribution of the dose to the rectal wall and late rectal toxicity after prostatic radiotherapy; an analysis of RT01 data in preparation.


Hornik K 1993 Some new results on neural network approximation *Neural Netw.* 6(9), 1069–1072.

Jackson A, Skwarchuk M W, Zelefsky M J, Cowen D M, Venkatraman E S, Levegrun S, Burman C M,
Using dose-surface maps to predict rectal bleeding


Wachter S, Gerstner N, Goldner G, Pötzi R, Wambersie A & Ptter R 2001 Rectal sequelae after...