Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning

Chris McIntosh\(^1,3\) and Thomas G Purdie\(^1,2,3\)

\(^1\) Department of Medical Imaging & Physics, Princess Margaret Cancer Centre, University Health Network (UHN), Toronto, ON, Canada
\(^2\) Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada
\(^3\) The Techna Institute, UHN, Toronto, ON, Canada

E-mail: chris.mcintosh@rmp.uhn.on.ca and tom.purdie@rmp.uhn.on.ca

Received 20 July 2016, revised 28 September 2016
Accepted for publication 1 November 2016
Published 20 December 2016

Abstract

Automating the radiotherapy treatment planning process is a technically challenging problem. The majority of automated approaches have focused on customizing and inferring dose volume objectives to be used in plan optimization. In this work we outline a multi-patient atlas-based dose prediction approach that learns to predict the dose-per-voxel for a novel patient directly from the computed tomography planning scan without the requirement of specifying any objectives. Our method learns to automatically select the most effective atlases for a novel patient, and then map the dose from those atlases onto the novel patient. We extend our previous work to include a conditional random field for the optimization of a joint distribution prior that matches the complementary goals of an accurately spatially distributed dose distribution while still adhering to the desired dose volume histograms. The resulting distribution can then be used for inverse-planning with a new spatial dose objective, or to create typical dose volume objectives for the canonical optimization pipeline. We investigated six treatment sites (633 patients for training and 113 patients for testing) and evaluated the mean absolute difference in all DVHs for the clinical and predicted dose distribution. The results on average are favorable in comparison to our previous approach (1.91 versus 2.57). Comparing our method with and without atlas-selection further validates that atlas-selection improved dose prediction on average in whole breast (0.64 versus 1.59), prostate (2.13 versus 4.07), and rectum (1.46 versus 3.29) while it is less important in breast cavity (0.79 versus 0.92) and lung (1.33 versus 1.27) for which there is high conformity and minimal dose shaping. In CNS brain, atlas-selection has the potential to be impactful (3.65 versus 5.09), but selecting the ideal atlas is the most challenging.
Keywords: machine learning, treatment planning, external beam radiotherapy, dose prediction, atlas-selection, multi-atlas, decision forests

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

1. Introduction

The delivery of radiotherapy (RT) is a complicated process that requires both clinical and technical expertise. Manual RT planning often involves multiple planning iterations, creating and updating dose-volume objectives and the creation of regions of interest (ROIs) solely to facilitate planning. The planning process therefore necessitates tremendous human resources often requiring hours to days to plan each patient (Das et al 2009). There is also considerable variability in planning requirements depending on the treatment site and treatment technique, which leads to high inter- and intra-institutional variation in clinical practice (Nelms et al 2012a, 2012b, Ohri et al 2013). In addition, plan quality has been positively correlated with planning experience and the time invested in an individual plan (Batumalai et al 2013).

There has been interest to overcome these challenges by incorporating automation into the conventional treatment planning process. These planning pipelines have traditionally incorporated historical treatment planning data with algorithms to estimate dose-volume objectives based on a limited number of features (Appenzoller et al 2012, Wu et al 2013, Yang et al 2013). In McIntosh and Purdie (2016) we introduced a novel planning pipeline that predicts a probabilistic estimate of the dose at each voxel in the image volume, thus creating a spatial dose-volume objective (SDO). The SDO punishes deviation from a specified dose value at each voxel, instead of trying to achieve a fixed amount to a volume. Our approach uses machine learning and radiomic image features to estimate the dose-per-voxel based on the dose to voxel-feature relationship observed in the most similar patients from a training database. Further, we learn to automatically select the most similar patients based on the most relevant features to dose prediction scored using the gamma metric (Low and Dempsey 2003). We incorporate features at various data scales and learnable contextual information about image appearance from the raw image volume. We aim to accurately match a novel patient to the patients in a gold standard training database with the most similar geometrical anatomy from both contoured and non-contoured structures. We output a probabilistic estimate of dose-per-voxel, enabling future development of new dose objectives. For simplicity, we refer to this as the probabilistic dose distribution. From the probabilistic distribution we previously generated a scalar dose distribution using a maximum-a-posterior estimate, which treats neighboring dose voxels independently and thus cannot constrain the dose to a desired DVH.

Our main contributions in this work are: to extend the method in McIntosh and Purdie (2016) to use a hybrid atlas-learning phase incorporating features based on the spatial dose distribution scored using the gamma metric (Low and Dempsey 2003) and features derived from DVHs in conjunction with a joint dose prior; and, to extend the previous validation to three novel clinical sites: CNS brain, lung, and rectum. Specifically, we introduce a conditional random field (CRF) model (Lafferty et al 2001) to transform the probabilistic distribution of dose into a scalar dose distribution that adheres to a predicted joint probability function...
prior (analogous to a DVH) of dose values for the target and pertinent organs-at-risk (OARs). As a result, this is the first work to balance the complementary goals of generating dose distributions that are spatially appropriate over the entire dose grid while still ensuring the proper distribution of dose within delineated ROIs is achieved. Our results demonstrate that balancing the goals can significantly improve dose predication accuracy in comparison to just the DVH, or just the spatial distribution independently (section 4).

Our hypothesis is that multiple-atlas-based dose prediction with CRF-optimized dose priors can accurately predict a resulting dose distribution without having to specify any objectives. The method can provide the required dose prediction component for a fully automated planning solution and can readily be applied to multiple clinical sites and treatment modalities using a single framework.

2. Methods and materials

2.1. Clinical treatment planning

Automated dose prediction training, testing and analysis was conducted using patients from six treatment sites (table 1). Patients were consecutively selected for each treatment site, with all patients clinically treated at our institution between 2011 and 2014. All patients were part of a retrospectively approved institutional research ethics board approval. Patients were simulated, immobilized and treated consistently within each treatment site according to the respective site-specific protocol table 1.

Clinical treatment plans were generated with the Pinnacle3 treatment planning system (Philips, Madison, WI) as per the site-specific treatment planning protocol. The clinical treatment plans for a given treatment site were generated using either IMRT or VMAT and for a specific dose-fractionation as specified in table 1. Target volumes were delineated by the treating radiation oncologist and organs at risk were delineated by the radiation oncologist and/or treatment planner (dosimetrist) as per the site-specific treatment planning guidelines.

2.2. Method overview

An overview of the proposed process is presented in figure 1. For simplicity, since all of our image processing is performed in 3D we refer to a 3D computed tomography (CT) image volume simply as an image, and specify a particular slice where appropriate. Training images are first loaded and characterized into image features. The features describe the appearance and texture of each voxel across a variety of scales. Machine learning using atlas regression forests (ARFs) is used to determine which features are relevant to dose prediction. Statistics over the features are used to characterize each complete image (density estimation). The ARFs are then cross-validated across all pairs within the training set, and the dose prediction accuracy of each ARF on all other training images is estimated. A second machine learning step is trained to predict the accuracy of a given ARF for each patient training image using the learned image descriptor (density estimate) as input. A planning image for a novel patient is first characterized into features, which are used to compute a density estimate for the training ARFs. The density estimate is used to predict the accuracy of each ARF on the novel patient image. The ARFs with the highest predicted accuracy are then used to predict dose for the novel patient image. We predict probabilistic dose estimates at each image voxel, i.e. a probability distribution function (PDF) over the range of
potential dose at each voxel. In this work we introduced a joint voxel optimization process through a CRF to find the most likely spatial distribution of voxel under an inferred joint probability dose prior from the most similar atlases.

Complete details of the contextual atlas regression forest planning pipeline are presented in McIntosh and Purdie (2016). In what follows we provide a higher-level summary in an effort to keep this manuscript more self-contained. For ease of reference the mathematical notation is kept consistent. There are four main components: feature extraction; atlas-to-image mapping; atlas-selection; and CRF dose prediction.

### 2.3. Pre-processing

The data outlined in table 1 is divided into two independent sets per treatment site. The training data consists of patient CT planning images, contours of the targets and relevant OARs (figure 3), and corresponding clinical dose distribution pairs.

We begin by defining our notation. Let \( P_j \) be an individual RT plan from a set of \( M \) patients and plans \( \mathbf{P} \). Each plan contains a CT image, a specification of dose-per-voxel, and one or more ROIs. The image is loaded from a DICOM CT image volume, \( I(x) \) where \( x \in \Omega \), the image domain, with \( |\Omega| \) voxels. The set of ROIs is denoted as \( \{C_{k,j}\} \) for plan \( P_j \) with \( k_j \in [1, \infty) \) ROIs. For simplicity, \( C_{k,j} \) will generically refer to the target ROI for a particular

### Table 1. Summary of clinical data for each treatment site.

| Site          | Total dose (cGy) | Fractions | Technique | Patients |
|---------------|------------------|-----------|-----------|----------|
|               |                  |           |           | Total    |
| Breast cavity | 1000             | 5         | IMRT      | 117      |
| Whole breast  | 4240             | 16        | IMRT      | 163      |
| CNS brain     | 6000             | 30        | IMRT      | 130      |
| Prostate      | 7800             | 39        | VMAT      | 164      |
| Lung          | 4800             | 4         | VMAT      | 94       |
| Rectum        | 5000             | 25        | VMAT      | 78       |
| Total:        |                  |           |           | 746      |

CNS: Central nervous system

IMRT: Intensity modulated radiation therapy

VMAT: Volumetric modulated arc therapy

![Figure 1. Flow chart showing training and testing pipelines of proposed voxel-based dose prediction algorithm. Dashed lines indicate learned output models from the training phase being used to predict dose for novel images.](image)

![Table 1. Summary of clinical data for each treatment site.](table)
plan, and multiple targets will appear sequentially in the set. Finally, the dose distribution is specified as $d_j(x)$, though we will equivalently write $d_j(x)$ where it improves clarity.

We define a set of features, $F$, with $N$ features per voxel. An individual feature $F_{j,x}$ is calculated from plan $P_j$ and voxel $x$. Rather than always writing $1...n$ for some set with $n$ elements, we will use $\ast$ to denote taking the entire set over a particular index, or group of indices. Therefore, $F_{j,x,\ast}$ is the same as $\{F_{j,x,H,N}\}$. For ease of reference our notation is summarized in Table 2.

### 2.3.1 Feature extraction

Following McIntosh and Purdie (2016), feature extraction uses both non-contoured and contoured image data. First, the patient external is extracted from the image via thresholding and morphological operations, or obtained from an existing external ROI where available. In order to capture features across a variety of scales we perform convolution with a 3D texture filter bank that is an extension of the Leung–Malik filter bank (Leung and Malik 2001). For every scale and orientation we use one first and one second derivative of an anisotropic Gaussian filter parameterized by the scale $\sigma$ at $\sigma_x = \sigma, \sigma_y = 3 \sigma, \sigma_z = \sigma / 3$ where use a set of scales with $\sigma = \{24, 48, 64\}$. For each scale, the set of orientations and rotations is parameterized by the azimuth, zenith, and angular rotation about the vector in a spherical co-ordinate system with $x, y, z$ aligned to the in-slice imaging plane from the CT image. We focus on the in-slice imaging plane (e.g. axial) with 6 samples, and 4 samples in each of the remaining orthogonal planes. This leads to $6 + 4 + 4 = 14$ filter orientations per scale and filter type (i.e. first or second derivative), for a total of $3 \times 14 \times 2 = 84$ rotational filters. We also use rotationally invariant filters in the form of isotropic Gaussians and Laplacian of Gaussians taken at set scales. The rotationally invariant filters are taken at a scale of 10, leading to a total of 86 filters. Filtering is performed in millimetres in world co-ordinates to account for variable voxel spacing during RT planning image acquisition. Filter parameters were selected through manual iteration on the training data and remain fixed for all experiments. Increased sampling in both scale and orientation space was added until no impressionable change was observed, with selection of relevant features left to the regression forest algorithm. Finally we include 4 target specific features for each target: a signed distance transform of the target $\Phi_x$, and a vector in 3D denoting the direction and distance to the closest point on the target boundary: $(x - c)/c = \arg\min_c \|x - c\|_2 \forall c \in C_1$, where $C_1$ is the target.

### 2.4 Atlas-to-image mapping

The first phase of learning defines an image feature set for each pair. Using the characterized features, an ARF is learned for each pair that models the relationship between a patient’s image features and their clinical dose plan on a voxel-by-voxel basis. Each ARF also computes a probability density function estimate to measure the probability of observing the
image features given the training image. For example, the density estimate learns that a particular training image having a target near a given OAR (e.g. in rectum, the dose distribution will pinch in the region of the PTV when the small bowel is in close proximity) was a key feature in generating its dose distribution. The next learning phase will then leverage these density estimates to characterize the image as a whole, and learn to estimate which images will have their corresponding dose distributions accurately predicted by a given ARF (i.e. perform atlas-selection).

First introduced by Leo Breiman in 2001, random forests (RFs) are a generalization of decision trees that use a mode-based voting algorithm over a set or forest of decision trees (Breiman 2001). A novel sample is classified by each decision tree, and then the mode of the output over the entire training set is taken as the final output class. RFs are a non-parametric regression algorithm where a regression tree, estimating a continuous output, is used in place of a classification decision tree. For a complete review the reader is referred to Criminisi et al (2011).

2.4.1 ARF learning. One ARF is learned separately for each training image and dose pair. Learning is performed by constructing a binary tree, \( T_{ij} \), where each node branches to its left- or right-child node based on a learned decision rule. Each node predicts a PDF \( P_{ij}(x) \) over possible dose values, through an empirical estimate (analogous to a histogram over the data samples processed by that node).

The root node begins with subset of data samples randomly chosen from the atlas image, \( I_j(x) \). We use the classical decision rule of \( f \leq t \) for observed feature \( f \), and learned threshold \( t \). The threshold at each node is learned by minimizing the least-squared-error of the predicted dose distribution in the resulting child nodes. For example, splitting the data so that 5 Gy samples are on the left, and 40 Gy on the right. The forest, \( T_i \), is a set of trees, each learned over a different bootstrapped sample from the training image, thus using a combination of weaker learners to build a stronger learner (Breiman 2001).

2.4.2 Inference (atlas-to-dose mapping). Per-voxel inference for a novel image involves calculating the features, and then for a single ARF traversing every tree starting from the root, evaluating \( f \leq t \) for each node, branching accordingly, and then returning the observed dose estimate of \( P_{ij}(d_{ax}|F_{ax} = f) \) at the reached leaf.

In canonical RF, each forest would instead be trained on a random subset of voxels from a random subset of images, and inference would thus be averaged over many images. A key observation is that this type of inference only explicitly considers all voxels independently. The features are taken over sets of voxels via convolution, but a joint observation/estimate of the dose is not made across the set of all observed voxel features \( F_{ax} \). However the dose at each voxel is heavily dependent on its neighborhood, for example being close to the target, i.e. the contextual information. In essence, though we have thus far modelled \( P(d_{ax}|F_{ax}) \), we ultimately wish to model \( P(d_{ax}|F_{ax}, \text{context}) \), i.e. dose given the contextual information about the image appearance as a whole. In what follows we use density estimation and atlas-selection learning to accomplish this.

2.4.3 Density estimation. In McIntosh and Purdie (2016) we introduced using density estimation at the leaf-nodes to estimate the probability of observing a set of features given a particular atlas. Specifically, in addition to a feature threshold and \( P(d_{ax}|F_{ax}) \) each leaf-node further contains a multivariate normal distribution over the observed features on the path from the root to the leaf: \( N(\mu_{a_i,j}, \Sigma_{a_i,j}) \) for tree \( T_{ij} \) from training image \( j \). Means and co-variances
are computed over the data samples at the respective leaf, i.e. $\mu_{l,ij}$ is for leaf $l \in T_{ij}$ from data in $I_j$. Further, define $N(\mu_{l,ia}, \Sigma_{l,ia})$ for some novel data being pushed through the ARF from image $I_a(x)$.

Consider a trivial tree with only a root node and two children, split on the raw CT value, for example. The decision learning determines the optimal dose prediction feature for the sampled voxels, and the density estimates at the leaves produce a two-mode mixture model estimating the likelihood of observing any particular CT value, given the atlas image. Using the same tree, the data in a novel image specifies a new density distribution, and the differences between the learned and observed density distributions indicate the similarity between the novel image and the atlas image. They key is that the density distributions are estimated over the features relevant to dose prediction learned during tree optimization, and thus can ignore superfluous image information. Figure 2 displays an example, where all of the voxels in the image that are predicted by a leaf are encoded by a single colour corresponding to a breadth-first ordered labelling of tree leafs. This is not a visualization of the feature itself, but rather the voxels that have a strong response to the set of features that defines a particular leaf. Thus the colour correspondence of voxels in figures 2(d) and (f) illustrates the set of corresponding voxels for novel images that are used to calculate and compare $N(\mu_{l,ia}, \Sigma_{l,ia})$.

The next step is to establish a distance between the learned leaf model (e.g. lung-like patch in training atlas) and the observed image data (e.g. slightly larger lung-like patch). The difference between the observed word at leaf $l$, and the learned word, is taken as the Bhattacharyya distance for multivariate normal distributions:

$$B(I_j, I_a|T, T_{ia}) = \frac{1}{8}(\mu_{l,ij} - \mu_{l,ia})^T \Sigma^{-1}(\mu_{l,ij} - \mu_{l,ia})$$

$$+ \frac{1}{2} \ln \left( \frac{\det \Sigma}{\sqrt{\det \Sigma_{l,ij} \det \Sigma_{l,ia}}} \right),$$

where $\Sigma = 0.5(\Sigma_{l,ia} + \Sigma_{l,ij})$, where the first term is related to the Mahalanobis distance, and the second term the differences between the co-variances (Bhattacharyya 1946).

2.5. Atlas-selection learning

With the individual ARFs and corresponding density estimates learned, in this learning phase we use cross-validation strictly within the training set to predict the dose with each ARF over all other training images, measure the accuracy of the predicted dose, and then use the density estimates to learn to predicted the observed accuracy. Finally, only the k ARFs with the highest predicted accuracy are used for a novel image. For example, for a novel CNS brain patient with a PTV near the Brainstem, only ARFs from patients with similar PTVs near the Brainstem are selected and used for dose prediction. The atlas-selection step ensures that the dose-per-voxel is not only specified based on the features at a voxel, but based on the contextual information in the image (i.e. the observed features at other voxels as a group). The atlas-selection step enables us to better approximate $P(d_{a,i} | F_{a,i})$ since each atlas inherently models the voxel dose interdependence (see McIntosh and Purdie (2016) for details).

With (1), the distance between two images can be computed as the sum over all the leaves, which is in-turn a sum over the entire image. However, entire areas of the image may not be relevant to contextual dose prediction, and hence atlas-selection. For example, when planning RT for a right-sided atypical lung target, the inferior left lung of the patient can be any shape or size and it will have no impact on the dose distribution. Instead of assuming a linear combination,
we perform regression using a second independent RF to compute the final distance (denoted as a pRF). We define $E_{dd}^{TT},\alpha_{aj}(\tilde{\alpha})^{\ast\ast} + R^{\ast\ast}$ as the difference between the predicted dose distribution $d_{\alpha},\tilde{\alpha}^{\ast}$ using ARF $T_{j}$ and the clinical dose distribution $d_{\alpha},\ast$. In McIntosh and Purdie (2016) we focused on the Gamma metric (Low and Dempsey 2003), but in this work we combine it with a sum of the absolute difference between the DVHs in the clinical and predicted plans to bring both a spatial and a total distribution-oriented context to the metric.

Each pRF is then trained using $B(l_{a},I_{a}|l,T_{j})$ for $l \in T_{j}$ as input features, over $a \in [1..j - 1, j + 1,..M]$ samples with $E(d_{\alpha},\alpha_{aj}|T_{j})$ as the explanatory variable, i.e. we use leave-one-out cross-validation over the training data set. The testing data is kept completely separate. Using each leaf as a separate feature enables the pRF to learn to ignore any leaf, $l \in T_{r,p}$ that represents an area of the image that is not relevant to atlas-selection. For example, a group of rectum-like patches can form an overall rectum shape, and the detection of that shape in the novel image implies a high predictive accuracy for the ARF.

In summary, each ARF learns a feature-based set of descriptors for its training image, and is paired with a pRF that predicts the accuracy of the ARF for any input image as a function of the learned descriptors at the ARF’s leaves. For computational speed, a sub-sampling rate of 1.5% is used to select voxel from the image for the density estimation.

We now write our equation for contextual atlas regression forests as:

$$P(d_{\alpha}|F_{\alpha},\ast) = \sum_{j=1..M} P_{\alpha}(d_{\alpha}|F_{\alpha},\ast,x,T_{j})P(T_{j}|F_{\alpha},\ast),$$

where $P(T_{j}|F_{\alpha},\ast)$ determines the likelihood of a RF $T_{j}$ given the observed features of the test image which is used as the weighting of the forest prediction. Intuitively, the closer the test image features are to the forest representation, the higher the weight. In this work we use an
equal weighting among the $k = 4$ ARFs with minimally predicted error, where 4 was learned through cross-validation on the training data in McIntosh and Purdie (2016).

2.6. Conditional random field dose estimation

The final step is to infer the most likely spatial dose distribution from $p_{\cdot|\cdot}(d_{a,\cdot|x}|F_{\cdot,a,\cdot})$. In McIntosh and Purdie (2016) we proposed to use maximum-a-posteriori (MAP) estimation per voxel. This approach is fast, but assumes independence among voxels, and therefore does not consider the total distribution of dose during inference. The distribution of dose is not to be confused with the dose distribution, the former is analogous to a DVH. In this work we propose to use the atlas-patients to estimate a joint prior for the distribution of dose to the target and ROIs, and then query $F_{\cdot,\cdot}(d_{a,\cdot|x}|F_{\cdot,a,\cdot})$ to find the most likely spatial distribution of dose to each voxel that adhere to the prior.

Incorporating an independence assumption and a uniform dose prior, our previous work estimated scalar dose from the predicted PDF using a MAP estimate of the form:

$$\hat{d}_{a} = \arg\max_{d_{a,\cdot}} \prod_{x} p_{\cdot|\cdot}(d_{a,\cdot|x}|F_{\cdot,a,\cdot}).$$  \hspace{1cm} (3)

A MAP estimate is used as opposed to an average, as is more typical in RF, to avoid degradation of the dose by a single outlier, e.g. [1 1 1 1 0] across five trees should predict a dose of 1, not 0.8.

Conversely, in this work we define a dose prior $P(d_{a,\cdot}|R(x))$ where $R(x)$ is a binary vector denoting the ROI class membership of $x$, from the set of targets and OAR ROIs $C$. Note that in contrast to a typical DVH, which is a cumulative prior distribution per ROI, we employ a joint prior and thus the prior is of dimensionality $|\mathcal{C}|$. The dose prior is computed as the average prior over the most similar atlases as determined by the pRFs, similar to (2).

Finding the most likely spatial assignment of dose-per-voxel according to $p_{\cdot|\cdot}(d_{a,\cdot|x}|F_{\cdot,a,\cdot})$ under the dose prior can be written as a CRF:

$$\hat{d}_{a} = \arg\max_{d_{a,\cdot}} \prod_{x} p_{\cdot|\cdot}(d_{a,\cdot|x}|F_{\cdot,a,\cdot})P(d_{a,\cdot}|R(x)),$$  \hspace{1cm} (4)

which written in this form is a non-convex continuous optimization problem. We solve by transforming to a binary linear program:

$$\hat{d}_{a} = \arg\max_{d_{a,\cdot}} \sum_{x} \sum_{g} \log(p_{\cdot|\cdot}(d_{a,\cdot|x}|F_{\cdot,a,\cdot}))b_{x,g},$$  \hspace{1cm} (5)

subject to the constraints:

$$\sum_{g} b_{x,g} = 1 \quad \text{for} \quad x \in \Omega,$$

$$b_{x,g} \geq 0 \quad \text{for} \quad x \in \Omega, g \in G,$$

$$\sum_{x \in R} b_{x,g} \leq |R| P(d_{a,\cdot|x}|R(x)) \quad \text{for} \quad R \in \mathcal{C}, g \in G$$  \hspace{1cm} (6)

where $g$ is introduced to discretize the dose into $G$ bins, and $b_{x}$ is a binary indicator variable. The first two constraints ensure that the binary indicator specifies at most one dose value-per-voxel. The third constraint enforces that the dose satisfies the required prior, with one such prior per ROI intersection sub-type (e.g. a voxel being just lung, or target and lung).
3. Results

We divide the data into independent training and testing sets (table 1). Breast, and prostate sites were used for algorithm development and parameter tuning using a 90%-10% cross-validation split on the training set in McIntosh and Purdie (2016), and those same parameter values are used here. The lung, rectum, and brain data use the same parameters without re-tuning. The results for CNS brain, lung, and rectum are thus entirely without tuning, and demonstrate the method’s parameter insensitivity as evidenced in McIntosh and Purdie (2016). The parameters and their values are as follows. The ARFs are bagged on 40% of the voxels in each image, have a maximum tree depth of 10, use $\sqrt{N}$ features per forest, and 48 trees per forest. Atlas selection is performed using $k = 4$ nearest neighbors. To limit the number of learned features per ARF, they are evaluated to a depth of 7 for atlas selection which yields 128 features. The pRFs use all data in each tree, are trained to no deeper than 5 samples per leaf, have 24 trees per forest, and use the recommended $\sqrt{128}$ features per tree (Breiman 2001).

We evaluate variants of both the atlas-selection algorithm and the atlas-dose to novel image mapping. For atlas-selection we examine using: no atlas-selection, the overlap-volume-histogram (OVH) distance between patients (Kazhdan et al 2009), and the proposed DVH-Gamma hybrid. For mapping the dose onto the novel image we experiment with: deformable image-based atlas-registration (DAR), and our ARFs. For DAR, the CT images are first registered with translation only since all patients are in the same pose for treatment. Deformable registration is carried out between the CT images with the suggested scale and smoothing parameters from Yang et al (2011). Registration is performed in 4 multi-resolution phases at scales of: 1/32, 1/16, 1/8, and 1/4. The dose from the atlas image is warped onto the novel image via the same transformations (translation, then deformable). We also include a generic OVH method where the DVH is taken directly from the nearest patients under the OVH metric.

Approximate upper-bounds for ARFs and DAR are computed by testing all training atlases on all test images. In order to compute a fast upper-bound approximation the images were sub-sampled to 1/4 size in all dimensions. Negligible differences were observed at full image resolution. However, note that the ARF-CRF can outperform the upper-bound, as the bound focuses on atlas-selection and is computed without the aid of a CRF.

Quantitative results are presented in table 3. The error is measured as the mean absolute difference (MAD) between the clinical DVH and the DVH calculated from the predicted dose distribution averaged over the set of ROIs, and the tested patients. The minimum value is 0%, and the maximum is 100%. Standard deviation is calculated between patients. The 95% confidence interval ($CI$) is calculated from a paired two-tailed t-test with the MAD-DVH as the dependent variable, comparing the proposed ARF-CRF with the other methods. We corrected for multiple-comparison testing with Bonferroni correction carried out over the 6 tests per site. The upper bounds are excluded from this result since they are not actually realizable for a novel patient. Figure 3 displays plots of the MAD-DVH error for each ROI in each site individually, including the number of observed instances of the ROI. ROIs were not contoured in plans where they appear outside the field and/or receive little to no dose. Figure 4 displays example dose overlays for both clinical and automated plans under the proposed ARF-CRF pipeline.

The algorithm is reasonably fast, with a total run-time including feature calculation, atlas-selection, and CRF optimization of approximately 15 min for a novel image volume. For speed the CRF is solved on the dose-grid, as opposed to the image grid. Novel sample prediction is carried out independently across trees in the ARFs and thus performance scales well with the system specifications. Training the system depends on the number of atlases used, but takes approximately 48 h per site. Each ARF takes approximately 15 min to train, with the
remaining time being spent on the pARF cross-validation training for atlas-selection within the training set. However, this can be run completely in parallel per atlas, and thus could be made much faster with additional computational resources. Furthermore, training is performed off-line, and only needed once per treatment site initially, or needed when there is an update to the training database.

4. Discussion and related work

The current paper extends the voxel-based dose prediction method previously presented (McIntosh and Purdie 2016), by improving the features for atlas selection to include DVH estimation accuracy, and incorporating CRFs to optimize a joint prior for dose prediction. We have also provided a more comprehensive clinical evaluation for six treatment sites and two treatment modalities and provided evaluation based on DVHs for targets and relevant OARs. The method presented requires only a limited number of ROIs figure 3 (e.g. four for breast cavity and ten for lung) to generate a dose distribution. By modeling the image appearance within ROIs and the appearance and geometrical relationships of non-delineated structures in the images we can better patient similarity than approaches using only ROI data. In this way, we can generate a SDO automatically, that does not require any dose objectives to be specified explicitly; the SDO captures all the necessary information to drive dose prediction.

There are a number of related works in automated planning (Kazhdan et al 2009, Appenzoller et al 2012, Wu et al 2013, Yang et al 2013, Shiraishi et al 2015, Shiraishi and Moore 2016). Most of the work in this area focuses on using high level shape descriptors of contoured structures to retrieve relevant DVHs from a database, and then infer corresponding dose volume objectives (Appenzoller et al 2012, Wu et al 2013, Yang et al 2013). After specifying the dose-volume objectives, planning proceeds as normal. However, recent independent work by Shiraishi and Moore (2016), and simultaneously our group (McIntosh and Purdie 2016) has introduced the notion of predicting the dose on a voxel-by-voxel basis. Shiraishi and Moore used groups of previous patient image-dose pairs to train two dose prediction models: one inside the target and one outside. They performed validation of the predicted dose distributions for 12 prostate and 23 stereotactic radiosurgery plans, demonstrating improved OAR sparing in comparison to DVH models (Appenzoller et al 2012, Shiraishi et al 2015).

![Figure 3. Distribution of mean average difference (MAD)-DVH errors across patients for all sites and ROIs with error bars representing one standard deviation.](image-url)
However, in preface to discussion of voxel-based dose prediction methods it is important to note that the resulting dose distributions from McIntosh and Purdie (2016), Shiraishi and Moore (2016) and those presented here are not necessarily achievable, and thus a stricter evaluation criteria than improved sparing is warranted. Although Shiraishi and Moore (2016) suggested predicted dose distributions are deliverable since they are generated based on deliverable plans, three-dimensional dose distributions do not form a vector space because dose is purely additive; a negative weighted linear combination of two dose distributions does not necessarily create an achievable dose distribution. In addition the models perform inference per-voxel, and thus are incapable of ensuring spatial consistency across voxels. Therefore, predicting increased sparing or coverage may not actually be achievable when

Figure 4. Visualizations of dose overlays for one example of each site. Dose is normalized to a percentage of prescription for each case (with red indicating 100% of prescription). (Top) Clinical dose. (Bottom) ARF-CRF predicted dose distribution. This figure is best viewed in colour.
Table 3. Mean average difference (MAD) between automated and clinical plan dose volume histograms for all regions of interest (ROIs). Proposed method indicated in bold.

| Map      | Atlas       | MAD-DVH | CI          | MAD-DVH | CI          | MAD-DVH | CI          |
|----------|-------------|---------|-------------|---------|-------------|---------|-------------|
|          |             |         | Breast cavity | Whole breast | CNS brain |         |             |
|          |             |         |              |          |              |         |              |
| u.b. ARF | DVH+Γ       | 0.87 ± 0.46 | —           | 1.23 ± 0.57 | —           | 3.65 ± 1.20 | —           |
| ARF      | DVH+Γ       | 0.96 ± 0.53 | [-0.4, 0.0] | 1.19 ± 0.67 | [-0.8, -0.3] | 5.35 ± 2.26 | [-1.2, 0.7] |
| ARF-CRF  | DVH+Γ       | 0.79 ± 0.39 | —           | 0.64 ± 0.43 | —           | 5.11 ± 2.74 | —           |
| Generic  | OVH         | 1.12 ± 0.53 | [-0.6, -0.1] | 0.88 ± 0.29 | [-0.4, -0.1] | 8.01 ± 3.20 | [-5.1, -0.7] |
| ARF      | OVH         | 1.36 ± 0.56 | [-0.9, -0.3] | 2.01 ± 2.07 | [-2.6, -0.1] | 6.78 ± 3.01 | [-4.5, 1.2] |
| RF       | None        | 0.92 ± 0.51 | [-0.3, 0.1]  | 1.59 ± 0.66 | [-1.2, -0.6] | 5.09 ± 2.09 | [-1.4, 1.5] |
| u.b. DAR | DAR         | 7.52 ± 6.55 | —           | 1.22 ± 0.40 | —           | 4.98 ± 3.44 | —           |
| DAR      | DAR         | 9.54 ± 11.79 | [-16.4, -1.1] | 2.46 ± 1.43 | [-2.6, -1.0] | 8.19 ± 4.54 | [-6.3, 0.1] |
| DAR      | OVH         | 19.93 ± 16.26 | [-29.7, -8.6] | 5.34 ± 9.76 | [-11.1, 1.7] | 9.32 ± 4.34 | [-7.8, -0.6] |
|          |             |         | Prostate     | Lung      | Rectum      |         |             |
| u.b. ARF | DVH+Γ       | 2.53 ± 0.67 | —           | 1.33 ± 0.93 | —           | 2.40 ± 1.01 | —           |
| ARF      | DVH+Γ       | 3.21 ± 1.16 | [-1.9, -0.3] | 1.62 ± 0.85 | [-0.6, 0.0] | 3.09 ± 1.26 | [-2.6, -0.7] |
| ARF-CRF  | DVH+Γ       | 2.13 ± 0.75 | —           | 1.33 ± 0.74 | —           | 1.46 ± 0.54 | —           |
| Generic  | OVH         | 2.53 ± 1.21 | [-1.1, 0.3]  | 2.68 ± 1.19 | [-1.9, -0.8] | 4.34 ± 1.21 | [-3.6, -2.2] |
| ARF      | OVH         | 6.38 ± 3.96 | [-6.9, -1.6] | 3.22 ± 2.49 | [-3.4, -0.4] | 4.42 ± 2.72 | [-4.8, -1.1] |
| RF       | None        | 4.07 ± 2.39 | [-3.7, -0.2] | 1.27 ± 0.69 | [-0.3, 0.4] | 3.29 ± 1.62 | [-3.0, -0.7] |
| u.b. DAR | DAR         | 2.74 ± 0.76 | —           | 7.96 ± 3.31 | —           | 2.92 ± 0.81 | —           |
| DAR      | DAR         | 4.92 ± 1.94 | [-4.0, -1.6] | 10.53 ± 4.13 | [-11.8, -6.6] | 5.45 ± 1.84 | [-5.2, -2.7] |
| DAR      | DAR         | 5.01 ± 2.26 | [-4.5, -1.3] | 13.84 ± 3.81 | [-14.9, -10.1] | 6.18 ± 2.57 | [-6.3, -3.1] |

Notes:
- Abbreviation Definition
- u.b. Upper bound (approximate)
- ARF Atlas regression forest
- CRF Condition random field
- RF Regression forest
- DAR Deformable atlas registration
- OVH Overlap volume histogram
- Γ Gamma analysis metric

* Mean ± standard deviation calculated across patients
* 95% confidence interval for mean difference in MAD-DVH in comparison to ARF-CRF
plans are generated to emulate the predicted dose. That said, this caveat should not detract from work in this direction. As the limit in the absolute difference between the predicted dose distribution, and the clinical dose distribution tends to zero, the predicted dose will by definition be achievable. Hence, in contrast to signed deviation, absolute value deviation penalizes any discrepancy and the more closely the predicted and clinical dose distribution match, the greater the predicted dose can readily be emulated. Thus evaluation of these model should focus on the absolute value deviations from clinical dose, as is done here.

Similar to Shiraishi and Moore (2016) we do observe trends in increased sparing in cases where the absolute error is larger, and it will be very interesting to examine inverse-planning on a large scale with inverse-planning and SDOs to learn if this is achievable. However, the best mechanism through which to do so remains open, and is a separate question from how best to infer the proposed spatial distribution itself.

Turning to our results, we compared a variety of different atlas-selection and atlas-novel image dose mapping algorithms. Averaged across all treatment sites, the best performer was the proposed CRF-ARF method (1.91) followed by our previous ARF method without CRF (McIntosh and Purdie 2016) (2.57) and RF without atlas-selection (2.71). Examining individual sites and comparing the upper-bound DVH+Γ atlas-selection to RF without atlas-selection indicates that atlas-selection is impactful for whole breast, CNS brain, prostate and rectum, but negligible for breast cavity and lung. This result is logical given that breast cavity and lung have high conformity with little sacrifice of target dose due to OARs, and hence the dose is a consistent function of the distance to the target. While atlas-selection should help in CNS brain (3.65 versus 5.09), in practice that is where we observed the largest performance gap between the ARF and its upper-bound, leading to a negligible improvement over the RF without atlas-selection. This indicates that future feature development could improve atlas-selection, and thus algorithm performance for this treatment site. Another potential improvement for this site is to calculate new features for OARs to help with atlas-selection. It is also feasible that there are many clinically acceptable plans with varying DVHs created by small perturbations in the IMRT beam geometry, thereby confounding atlas-selection.

The proposed ARF-CRF algorithm on average out-performed the ARF algorithm for all sites, but lacked statistical significance indicated by the CI in breast cavity, lung, and brain. We believe the reasoning is the same as above. If atlas-selection is less impactful (or realizable) than inferring the dose-prior, then adhering to it is less impactful as well. The CIs are mostly negative, and so a larger evaluation set might also improve significance. Interestingly, examining the individual ROIs in figure 3 we note that atlas-selection always improves target performance. Comparing with and without atlas-selection in lung RT we observed increased accuracy on the targets, but increased error on the lung dose which is a potential for increased sparing.

DAR is the overall worst performer (9.93), with reasonable prostate and whole breast performance but the worst for breast cavity and lung. Interestingly, the DAR upper bound performance is inline with the ARF methods for whole breast, CNS brain, prostate, and rectum. This validates that both the registration algorithm, and associated parameters, are appropriate and able to obtain a good dose prediction given the ideal atlas, but that atlas-selection is more challenging for the DAR method than the ARF. This result is similar to that observed in McIntosh and Purdie (2016). Prostate and whole breast have a very fixed target geometry, and are thus the best suited to DAR, where as lung and breast cavity have targets that are free to translate around within a larger region of the patient. Large translation of only the target position between an atlas and a novel patient force tearing in the deformation field, which is discouraged and resisted by DAR methods in-place of smooth image warps. As a result the DAR approaches are not as well suited to these sites. In contrast, the RF approaches can learn
to predict dose as a function of voxel-to-target distance which easily adapts to moving the
target a few inches laterally.

The generic OVH method serves as mostly a baseline, since it directly infers a DVH instead
of a spatial dose distribution. It is interesting that the proposed atlas-selection is able to out-
perform OVH without having access to the OAR contours. We believe this can be attributed to
the automatic sub-segmentation of the image into crucial regions (figure 2), which given the
voxel-distance-target feature already included can implicitly calculate an OVH-like metric.
When the ARFs are used for dose mapping, DVH+Γ enables the addition of spatial dose infor-
mation and thus improved performance over OVH, as it was not designed for this purpose.

Further, the ARF-CRFs demonstrate the increased ability of the joint distribution dose
prior, over inferring a set of uni-variate dose priors (DVHs). In Boutilier et al (2016) they
demonstrated that such a method needs a very large number of atlases. This is a result of inter-
sections between ROIs creating high degrees of variability in the DVH, a problem not suffered
by the joint distribution. For example, changing the size of the target by 10% in lung changes
the lung DVH, but not the joint distribution.

Lastly, in McIntosh and Purdie (2016) we observed that the number of needed training
atlas was a function of training site complexity and variability. We found similar results in this
study, for example, that atlas-selection was not crucial for lung, and therefore requires fewer
training images.

When discussing and developing knowledge-based techniques, training data quality is an
important consideration. Since our method, as with other knowledge-based planning meth-
ods, drive solutions that are based on the underlying training data, there is a possibility for
low quality plans in a training database to be propagated onto new patients; though this is the
same risk with an erroneously trained person. Training data will inevitably include variability
in patient anatomical features but also plan quality. Anatomical variation in the training data
is important and ultimately results in a more robust, generalized planning solutions capable
of emulating a greater range of patients and ensuring novel patients to be planned can be
properly matched. In terms of plan quality, our method has been designed with a built-in
outlier removal mechanism to prevent low quality atlas patients from being used for novel
patients. Atlas plans of low quality in the training database will predict low quality plans dur-
ing atlas-selection cross-validation training and thus be automatically scored as inferior plans
for the other cross-validation patients in comparison to higher quality atlases. This will cause
the outlier atlas to receive a low atlas-quality score, and thus it will not be selected for novel
patients. A future area of investigation would be to develop a curated database from multiple
institutions over-time that represented an even higher gold standard.

By extension, erroneous plans have the potential to be used for generating plans using
any knowledge-based method. Erroneous plans are best avoided by having a highly curated
training database. All of the plans used for training here were clinically approved by multiple
experts as part of the clinical quality assurance process with the majority of patient reviewed
in multi-disciplinary peer-review case rounds. From a methodological perspective, we use
plans from multiple patient atlases, which mitigates against erroneous plans if only a minority
of the atlases have the error. As with plan quality, a future strategy and area of investigation
would be to refine the training database over time and use data from multiple institutions to
create a higher gold standard.

In order to generate deliverable dose distributions the method described can be extended
and applied to include fluence optimization (Shepard et al 1999). The dose prediction method
can provide estimates of the beam geometry from patient atlases selected for doing the dose
distribution prediction. Therefore, it is possible to have a completely automated treatment
planning solution without requiring any user interaction to specify beam geometry or setting any dose-volume objectives as in traditional optimized based treatment planning. Although not presented in the current paper, the method can be readily applied to other treatment modalities, such as particle therapy, as the method estimates dose directly from planning images without any planning information being used. Therefore, we can learn how dose distributions are shaped. In addition, there are clear applications for applying the methodology to achieve rapid planning to facilitate adaptive planning strategies.

5. Conclusions

This paper describes a methodology for the rapid estimation of the patient dose distribution without requiring any dose-volume objective specification and requiring only the target and limited OARs as inputs. We have developed a highly general voxel-based dose prediction method incorporating multi-patient atlas training that has applicability across a range of clinical sites and treatment modalities and techniques. We examined variants of our voxel-based dose prediction method with and without atlas-selection, and show that for sites in which there is variability in target coverage and dose distribution shaping between patients, atlas-selection provided substantially better dose prediction results. Whereas sites which have high conformity and minimal shaping, atlas selection was less important.

This paper presents the first work to balance the complementary goals of finding a spatially appropriate dose distribution while adhering to a probabilistic DVH prior. Our results validate that this can greatly improve dose prediction accuracy for a variety of sites. In particular, examining our results in comparison to the generic OVH method shows that adding spatial information to the dose prediction can lead to more accurate DVH predictions for breast cavity, whole breast, CNS brain, lung, and rectum. Comparing our ARF-CRF method to the original ARF, we observe that adding DVH information to augment the spatial distribution improves results for whole breast, prostate and rectum sites.

Based on these results, our future work is to examine our pipeline in conjunction with inverse-planning technologies to determine if atlas-selection can improve OAR sparing. We can apply the presented methodology as a new automated planning strategy or a method that can be readily integrated into existing optimization-based approaches. Although our method outperformed DAR, it is important to note that the DAR algorithms to date have not been designed for dose prediction and are thus ill-suited to the task, as was observed in lung and breast cavity RT. Future work could also include developing customized DAR algorithms specifically for dose prediction.

Acknowledgments

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Institutes of Health Research (CIHR).

References

Abrams R A et al 2012 Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704—a phase III trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas Int. J. Radiat. Oncol. Biol. Phys. 82 809–16
Appenzoller L M, Michalski J M, Thorstad W L, Mutic S and Moore K L 2012 Predicting dose-volume histograms for organs-at-risk in IMRT planning Med. Phys. 39 7446–61
Batmalai V, Jameson M G, Forstner D F, Vial P and Holloway L C 2013 How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case Pract. Radiat. Oncol. 3 e99–106
Bhattacharyya A 1946 On a measure of divergence between two multinomial populations Sankhya: Indian J. Stat. 7 401–6
Boutilier J J, Craig T, Sharpe M B and Chan T C 2016 Sample size requirements for knowledge-based treatment planning Med. Phys. 43 1212–21
Breiman L 2001 Random forests Mach. Learn. 45 5–32
Criminisi A, Shotton J and Konukoglu E 2011 Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning Technical Report MSR-TR-2011-114 Microsoft Research Cambridge
Das I J, Moskvin V and Johnstone P A 2009 Analysis of treatment planning time among systems and planners for intensity-modulated radiation therapy J. Am. Coll. Radiol. 6 514–7
Kazhdan M M, Simari P D, McNutt T, Wu B, Jacques R, Chuang M and Taylor R 2009 A shape relationship descriptor for radiation therapy planning MICCAI (1) (Lecture Notes in Computer Science vol 5762) ed G-Z. Yang et al (Berlin: Springer) pp 100–8
Lafferty J D, McCallum A and Pereira F C N 2001 Conditional random fields: probabilistic models for segmenting and labeling sequence data Proc. of the 18th Int. Conf. on Machine Learning ed C E Brodley and A P Danyluk (San Francisco, CA: Morgan Kaufmann) pp 282–9
Leung T and Malik J 2001 Representing and recognizing the visual appearance of materials using three-dimensional textons Int. J. Comput. Vis. 43 29–44
Low D A and Dempsey J F 2003 Evaluation of the gamma dose distribution comparison method Med. Phys. 30 2455–64
McIntosh C and Purdie T G 2016 Contextual atlas regression forests: multiple-atlas-based automated dose prediction in radiation therapy IEEE Trans. Med. Imaging 35 1000–12
Nelms B E, Robinson G, Markham J, Velasco K, Boyd S, Narayan S, Wheeler J and Sobczak M L 2012a Variation in external beam treatment plan quality: an inter-institutional study of planning and planners systems Pract. Radiat. Oncol. 2 296–305
Nelms B E, Tomé W A, Robinson G and Wheeler J 2012b Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer Int. J. Radiat. Oncol. Biol. Phys. 82 368–78
Ohri N, Shen X, Dicker A P, Doyle L A, Harrison A S and Showalter T N 2013 Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials J. Natl Cancer Inst. 105 387–93
Peters L J, O’Sullivan B, Giralt J, Fitzgerald T J, Trotti A, Bernier J, Bourhis J, Yuen K, Fisher R and Rischin D 2010 Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 0202 J. Clin. Oncol. 28 2996–3001
Shepard D M, Ferris M C, Olivera G H and Mackie T R 1999 Optimizing the delivery of radiation therapy to cancer patients SIAM Rev. 41 721–44
Shiraishi S and Moore K L 2016 Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy Med. Phys. 43 378–87
Shiraishi S, Tan J, Olsen L A and Moore K L 2015 Knowledge-based prediction of plan quality metrics in intracranial stereotactic radiosurgery Med. Phys. 42 908–17
Wu B, Pang D, Simari P, Taylor R, Sanguineti G and McNutt T 2013 Using overlap volume histogram and IMRT plan data to guide and automate VMAT planning: a head-and-neck case study Med. Phys. 40 021714
Yang D, Brame S, El Naqa I, Aditya A, Wu Y, Goddu S M, Mutic S, Deasy J O and Low D A 2011 Technical note: Dirart—a software suite for deformable image registration and adaptive radiotherapy research Med. Phys. 38 67–77
Yang Y, Ford E C, Wu B, Pinkawa M, van Triest B, Campbell P, Song D Y and McNutt T R 2013 An overlap-volume-histogram based method for rectal dose prediction and automated treatment planning in the external beam prostate radiotherapy following hydrogel injection Med. Phys. 40 011709