Models for predicting objective function weights in prostate cancer IMRT

Justin J. Boutilier, Taewoo Lee, Tim Craig, Michael B. Sharpe, and Timothy C. Y. Chan

Articles you may be interested in

Interactive prostate segmentation using atlas-guided semi-supervised learning and adaptive feature selection
Med. Phys. **41**, 111715 (2014); 10.1118/1.4898200

Predicting objective function weights from patient anatomy in prostate IMRT treatment planning
Med. Phys. **40**, 121706 (2013); 10.1118/1.4828841

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 (2013); 10.1118/1.4769424

A sensitivity-guided algorithm for automated determination of IMRT objective function parameters
Med. Phys. **33**, 2935 (2006); 10.1118/1.2214171

Clinical applications of IMRT to adenocarcinoma of the prostate: Portal dose verification and intensity modulated neutron radiotherapy
Med. Phys. **32**, 302 (2005); 10.1118/1.1827731

Models for predicting objective function weights in prostate cancer IMRT

Justin J. Boutilier[a)] and Taewoo Lee
Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road,
Toronto, Ontario M5S 3G8, Canada

Tim Craig
Radiation Medicine Program, UHN Princess Margaret Cancer Centre, 610 University of Avenue,
Toronto, Ontario M5T 2M9, Canada and Department of Radiation Oncology, University of Toronto,
148 - 150 College Street, Toronto, Ontario M5S 3S2, Canada

Michael B. Sharpe
Radiation Medicine Program, UHN Princess Margaret Cancer Centre, 610 University of Avenue,
Toronto, Ontario M5T 2M9, Canada; Department of Radiation Oncology, University of Toronto, 148 - 150
College Street, Toronto, Ontario M5S 3S2, Canada; and Techna Institute for the Advancement of Technology
for Health, 124 - 100 College Street, Toronto, Ontario M5G 1P5, Canada

Timothy C. Y. Chan
Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road,
Toronto, Ontario M5S 3G8, Canada and Techna Institute for the Advancement of Technology for Health,
124 - 100 College Street, Toronto, Ontario M5G 1P5, Canada

Purpose: To develop and evaluate the clinical applicability of advanced machine learning models that simultaneously predict multiple optimization objective function weights from patient geometry for intensity-modulated radiation therapy of prostate cancer.

Methods: A previously developed inverse optimization method was applied retrospectively to determine optimal objective function weights for 315 treated patients. The authors used an overlap volume ratio (OV) of bladder and rectum for different PTV expansions and overlap volume histogram slopes (OVSR and OVSB for the rectum and bladder, respectively) as explanatory variables that quantify patient geometry. Using the optimal weights as ground truth, the authors trained and applied three prediction models: logistic regression (LR), multinomial logistic regression (MLR), and weighted K-nearest neighbor (KNN). The population average of the optimal objective function weights was also calculated.

Results: The OV at 0.4 cm and OVSR at 0.1 cm features were found to be the most predictive of the weights. The authors observed comparable performance (i.e., no statistically significant difference) between LR, MLR, and KNN methodologies, with LR appearing to perform the best. All three machine learning models outperformed the population average by a statistically significant amount over a range of clinical metrics including bladder/rectum V53Gy, bladder/rectum V70Gy, and dose to the bladder, rectum, CTV, and PTV. When comparing the weights directly, the LR model predicted bladder and rectum weights that had, on average, a 73% and 74% relative improvement over the population average weights, respectively. The treatment plans resulting from the LR weights had, on average, a rectum V70Gy that was 35% closer to the clinical plan and a bladder V70Gy that was 29% closer, compared to the population average weights. Similar results were observed for all other clinical metrics.

Conclusions: The authors demonstrated that the KNN and MLR weight prediction methodologies perform comparably to the LR model and can produce clinical quality treatment plans by simultaneously predicting multiple weights that capture trade-offs associated with sparing multiple OARs. © 2015 American Association of Physicists in Medicine. [http://dx.doi.org/10.1118/1.4914140]

1. INTRODUCTION

Intensity-modulated radiation therapy (IMRT) is the most common method of external beam radiation treatment for prostate cancer patients.[1] Treatment planning optimization can be used in prostate cancer IMRT treatment planning to find a treatment plan that determines the best trade-off between a set of conflicting objectives. One way to formulate such a problem is to assign each objective a weight representing its relative importance and then combine all objectives into a single objective function.

In treatment planning, objective function weights are typically seen as tuneable parameters. Planners usually solve the treatment planning problem iteratively, tuning the weights in

a trial and error approach, until a resulting plan is deemed acceptable.[2–5] This *a priori* approach is widely used in clinical practice, though it can be inefficient and time-consuming. More recently, research has focused on exploring the Pareto surface of the treatment planning problem.[6–11] This *a posteriori* approach allows a treatment planner to choose a final plan from a large database of already optimized plans. Once a plan is chosen, it is possible to automate the planning process by using the reference plan's dose volume histogram (DVH) trade-offs to guide the algorithm.[11] In all cases, treatment planning effort can be reduced by starting from a "good" set of weights.

Recent advances in inverse optimization have demonstrated the potential of deriving objective function weights from historical treatment plans.[12] The inverse optimization method takes a historical treatment plan (i.e., already optimized and delivered) and determines the values of the objective function weights that are needed to recreate the given plan. Using such a method to retrospectively determine weights lets us better understand the relationship between weights and the optimized dose distribution. A recent study provided a proof-of-concept for geometry-driven weight determination in prostate cancer.[13] In particular, they developed a statistical model to relate inversely optimized objective function weights and patient anatomy, and quantify the trade-off between rectum and bladder sparing. Such a model can be used in a predictive manner to determine a set of initial weights for treatment planning for future patients, personalized to their anatomy.

Related studies have considered using geometric features to predict dose objectives or DVHs for IMRT treatment planning and quality control.[14–25] Such methods generally fall into two categories. The first approach, which we will refer to as the "library method," uses a database of previously delivered clinical plans from which the final plan or dose objective for the new patient is chosen as is.[14–19] The second approach, which we will refer to as the "prediction method," uses an internal geometric quantifier like an overlap volume histogram (OVH) to predict the dose for a single organ-at-risk (OAR), such as the rectum.[20–25] Although the predicted dose can be used to guide the planning process, the planner is still required to manually tune the weights in order to achieve the desired plan. The potential of these approaches has been demonstrated for prostate cancer[17–21] as well as for other cancer sites.[14–16,22–25] A limitation of these approaches is the focus on predicting the dose or objective weight for a single OAR. As a result, these methods do not consider the trade-off between sparing multiple OARs in close proximity to the tumor.

In this paper, we develop machine learning methods that have not been previously explored in the weight or dose prediction literature. Two of our models, K-nearest neighbors (KNN) and multinomial logistic regression (MLR), generalize previous research to the setting of multiple OARs by simultaneously predicting multiple weight values. We also develop a logistic regression (LR) model for comparison. We use a dataset of over 300 prostate cancer patients and clinically validate our models across a wide range of patients representing the largest dataset used in the literature to date for comparable studies. Validation of the advanced models is needed on a well-studied

geometry such as prostate before they can be applied to more complex sites with many OARs.

2. METHODS AND MATERIALS

2.A. Data

We obtained prostate IMRT treatment plans delivered at the Princess Margaret Cancer Center as a part of the PROFIT trial (NCT00304759). Data access was granted under UHN REB 137 11-0107-CE. The historical treatment plans were designed by a gastrourinary group consisting of multiple oncologists and treatment planners using the same protocols. Each treatment plan was delivered with seven equidistant fixed coplanar 6 MV photon beams at gantry angles of 40°, 80°, 110°, 250°, 280°, 310°, and 355°. The CTV consisted of the prostate gland (±10 mm of seminal vesicles) and a 10 mm margin (7 mm posteriorly) was added to the CTV to obtain the PTV. Patients who previously had a prostatectomy or pelvic lymph node irradiation were excluded, along with those patients with abnormal anatomic features such as prosthetic hips or pelvic kidneys. A total of 315 patients satisfied the inclusion criteria and were exported from the PINNACLE treatment planning system (Phillips Radiation Oncology Systems, Madison, WI) in DICOM format and imported into MATLAB via CERR (Computational Environment for Radiotherapy Research).[26]

2.B. Inverse optimization

An inverse multiobjective optimization model was used retrospectively to determine objective function weights for each of the 315 patients. Details about the general inverse optimization approach are available elsewhere.[12] The inverse optimization model (IOM) takes historical treatment plans as input and determines the objective function weights used to generate those plans. The IOM was derived from a forward optimization problem (FOP), which corresponds to the treatment planning optimization problem. The FOP objective was a weighted function composed of five objectives corresponding to the bladder, rectum, left femoral head, right femoral head, and PTV ring, which we will index by B, R, LF, RF, and PR, respectively. PTV ring is an artificial structure comprising a ring around the PTV and is used to achieve better dose conformity at Princess Margaret. Although the treatment planning problem solved in clinical practice has many more objectives, it was shown that our formulation is able to closely replicate clinical quality plans with optimal weights.[12] Therefore, the treatment plans generated by the IOM weights (i.e., IOM plans) will be treated as a proxy for the clinical plans.

We now briefly review the FOP and refer the interested reader to the literature for a detailed description.[13] The FOP penalizes the dose to an OAR k linearly, when it is above some threshold θ_k. We set our threshold to 50 Gy for the bladder, rectum, and PTV ring. We penalize the maximum dose to the left and right femoral heads. The upper bound on the dose delivered to all voxels was set to 81.9 Gy while the lower bounds on the dose delivered to the CTV and PTV were set to 78 and 74.1 Gy, respectively. We used lower bounds for the

dose delivered to the CTV and PTV because they are generally assigned very large weights.

We applied the IOM to all 315 patients resulting in a vector of five weights for each patient, which we refer to as the "IOM weights." All patients had IOM plans with clinical DVH metrics that were within 10% of the historical plans' metrics. The inversely determined objective function weights were then used as the ground truth in order to train and test our prediction models. The majority of the weight is split between the bladder and rectum while both femurs and PTV ring receive very little. The IOM weights for five sample patients, chosen to illustrate a full range of weight values, are shown in Table I along with the population average of the IOM weights across all 315 patients. The five sample patients are constant in the tables throughout this paper and will be referred to as "cohort S."

2.C. Quantifying patient geometry

We used OVHs to quantify patient anatomy.[14] We generated OVH curves for the bladder and rectum using PTV expansions of 0–2.5 cm in 0.1 cm increments. In order to capture the spatial tradeoff between the bladder and rectum, we then computed the ratio of rectum overlap volume to bladder overlap volume for each PTV expansion value $x \in \{0, 0.1, \ldots, 2.5\}$, which we denote as OV_x. This yielded a total of 26 OV_x values for each patient. Next, we computed the slope of the bladder and rectum OVH curves at points interpolated between each pair of consecutive expansion points in $\{0, 0.1, \ldots, 2.5\}$. This resulted in an additional 50 features for each patient, which we denote by $OVSB_{x,x+0.1}$ and $OVSR_{x,x+0.1}$ for the bladder and rectum, respectively. Thus, in total, we derived 76 features from the OVHs for each patient. These features were then used as potential independent variables in our prediction models.

2.D. Weight prediction

We trained three weight prediction models using the IOM weights from Sec. 2.B and the patient features described in Sec. 2.C. We began with a logistic regression model, which captures the primary trade-off between the bladder and rectum. To generalize our approach for more complicated tumor sites with many OARs nearby, we developed a weighted K-nearest neighbor model and a multinomial logistic regression model, which can simultaneously predict any number of weight values.

Due to the dimensionality of our feature space, it was not feasible to consider all possible combinations of features. Instead, for each model, we first considered all possible one- and two-feature models from which the best was chosen using a specified criteria outlined in the individual model sections below. We then employed stepwise forward regression to determine if any further variables improved the model.

If such an approach were to be implemented clinically, we would expect the models to be trained on all past patient data and used to predict a single new patient. To simulate this idea, we employed a leave-one-out approach for each model. In other words, we set aside a single patient and trained our models using the remaining 314 patients. The resulting model is then used to predict the remaining out-of-sample patient. This process was repeated 315 times, once for each possible out-of-sample patient. All statistical work was carried out using the software language R.[27]

2.D.1. Logistic regression

Due to the bicriteria nature of prostate cancer treatment planning, it was intuitive to first apply a logistic regression model that predicts the bladder weight from a set of patient features. The rectum weight was then computed by subtracting the bladder weight from 1. We also considered predicting the rectum weight but found the bladder weight model to be slightly more accurate. Because of the low weight values of the left femoral head, right femoral head, and PTV ring (see Table I), they were fixed to 0 during the regression and then set to the population averages of 0.003, 0.003, and 0.05, respectively, *post-hoc*. Finally, the bladder and rectum weights were scaled down by a factor of 0.944 to ensure that the weights sum to one. This logistic regression approach improves upon a previous approach in the literature[13] by predicting bladder and rectum weights on the full interval [0,1] and scaling post-hoc, instead of training the model on a restricted interval (due to the normalization being done *a priori*), which makes it impossible to predict extreme weight values that lie outside the restricted interval.

Our logistic regression selection criteria consisted of two steps: (1) identify the models with statistically significant independent variables (i.e., features) at the 95% level using the Wald Test,[28] and (2) choose the model with the smallest Akaike information criterion (AIC) value.[28] Using the

TABLE I. The optimal objective function weights determined by the IOM for cohort S.

Patient (type)	IOM weights				
	Bladder	Rectum	Left femur	Right femur	PTV ring
1 (High Blad.)	0.948	0.009	0.002	0.001	0.080
2 (Med. Blad.)	0.651	0.311	0.002	0.003	0.033
3 (Balanced)	0.446	0.471	0.004	0.003	0.076
4 (Med. Rect.)	0.292	0.627	0.004	0.003	0.074
5 (High Rect.)	0.097	0.849	0.002	0.001	0.051
Pop. Avg.	0.508	0.437	0.003	0.003	0.049

approach described above, the best model used two patient features, $OV_{0.4}$ and $OVSR_{0,0.1}$. The functional form of the logistic regression equation is

$$\alpha_B = \frac{1}{1 + e^{-(\beta_0 + \beta_1 OV_{0.4} + \beta_2 OVSR_{0,0.1})}}, \tag{1}$$

where α_B denotes the bladder weight.

2.D.2. Multinomial logistic regression

Next, we trained a multinomial logistic regression model that simultaneously predicts all five objective function weights from a set of patient features. Our model selection criteria were identical to those of the logistic regression model. The model with the minimal AIC value used only one feature, $OV_{0.4}$. The functional form of the multinomial logistic regression is

$$\begin{bmatrix} \alpha_B \\ \alpha_R \\ \alpha_{LF} \\ \alpha_{RF} \\ \alpha_{PR} \end{bmatrix} = \frac{1}{1 + \sum_{k \in \{R, LF, RF, PR\}} e^{(\beta_{k,0} + \beta_{k,1} OV_{0.4})}}$$

$$\times \begin{bmatrix} 1 \\ e^{(\beta_{R,0} + \beta_{R,1} OV_{0.4})} \\ e^{(\beta_{LF,0} + \beta_{LF,1} OV_{0.4})} \\ e^{(\beta_{RF,0} + \beta_{RF,1} OV_{0.4})} \\ e^{(\beta_{PR,0} + \beta_{PR,1} OV_{0.4})} \end{bmatrix}, \tag{2}$$

where α_B, α_R, α_{LF}, α_{RF}, and α_{PR} denote the bladder weight, rectum weight, left femoral head weight, right femoral head weight, and PTV ring weight, respectively.

2.D.3. K-nearest neighbor

The third approach we considered was a classification-based model that combines k-means clustering and weighted K-nearest neighbors. We used a k-means clustering algorithm[29] to partition the IOM weights into three (i.e., $k = 3$) distinct groups, corresponding to patients that are bladder-weighted, rectum-weighted, and roughly equally weighted or balanced. Each patient was assigned a label corresponding to the cluster to which that patient belonged.

Using the patient labels, we trained a distance-weighted KNN model, which uses a set of patient features to determine the probability that a given patient belongs to each cluster.[30] Once our features were selected, we determined the K-nearest neighbors in feature space using Euclidean distance. We then used a kernel function to value each neighbor according to distance, with closer neighbors receiving a higher value. The probability of a patient belonging to cluster i equaled the summation of values for all neighbors belonging to cluster i (determined via labeling) divided by the summation of values for all K neighbors. The three probabilities and the corresponding centroid weight vectors were then used to assign an expected weight vector to each patient. Let b, r, and n index the bladder, rectum, and balanced clusters, respectively. Let p_i
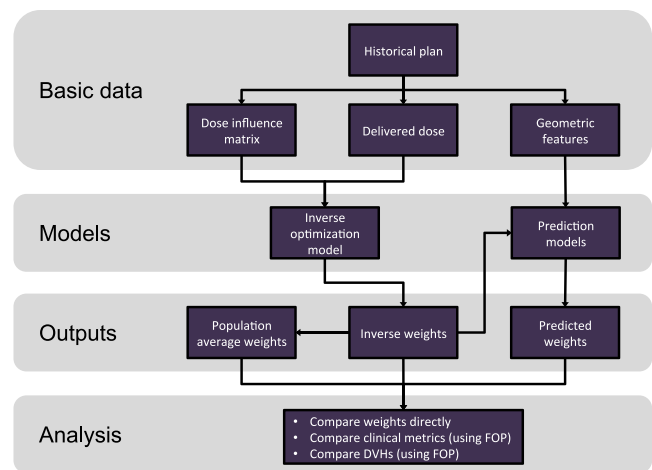

Fig. 1. Overview of our methodology.

denote the probability of belonging to cluster i and $\mathbf{c}_i$ denote the centroid weight vector of cluster i. For each patient, the KNN predicted weight vector, α_{KNN}, was

$$\alpha_{KNN} = p_b \mathbf{c}_b + p_n \mathbf{c}_n + p_r \mathbf{c}_r. \tag{3}$$

Our model selection process consisted of determining the most predictive patient features, choosing the best kernel function, and finding the optimal value of K. For each feature combination, we considered five kernel functions (rectangular, triangular, Epanechnikov, Gaussian, and rank) and all values of $K \leq 50$. From these 250 models, we chose the model with the highest classification accuracy. This process was repeated for each set of features. Classification accuracy was determined by assigning each patient to the cluster for which that patient has the highest probability of belonging (i.e., the largest p_i). These labels were then compared against the original labeling from k-means clustering and the proportion that matched was the classification accuracy. The model with the highest classification accuracy used a triangular kernel function and two patient features, $OV_{0.4}$ and $OVSB_{1.3,1.4}$ with $K = 14$.

2.E. Metrics

To compare the weights from the different models, we computed the absolute difference between each model's predicted weights and the IOM weights for each patient. The Population Average weight vector ("Pop. Avg.") was used as a baseline performance measure.

Table II. The logistic regression predicted weight vectors for cohort S.

Patient	Predicted weights				
	Bladder	Rectum	Left femur	Right femur	PTV ring
1	0.893	0.052	0.003	0.003	0.050
2	0.505	0.440	0.003	0.003	0.050
3	0.144	0.801	0.003	0.003	0.050
4	0.450	0.495	0.003	0.003	0.050
5	0.226	0.719	0.003	0.003	0.050
Pop. Avg.	0.506	0.439	0.003	0.003	0.050

TABLE III. The multinomial logistic regression predicted weight vectors for cohort *S*.

| Patient | Predicted weights | | | | |
	Bladder	Rectum	Left femur	Right femur	PTV ring
1	0.901	0.070	0.003	0.002	0.024
2	0.447	0.479	0.004	0.004	0.065
3	0.149	0.784	0.003	0.003	0.061
4	0.490	0.438	0.004	0.004	0.064
5	0.262	0.663	0.004	0.004	0.067
Pop. Avg.	0.508	0.437	0.003	0.003	0.049

To assess the clinical applicability of our models, five treatment plans and dose distributions were generated for each patient using the weight vectors from the IOM (representing the ground truth), LR model, MLR model, KNN model, and the Pop. Avg. From the dose distributions, we measured the following metrics, which are derived from the clinical DVH acceptability criteria used at Princess Margaret Cancer Center: V54.3Gy and V70Gy for both the bladder and the rectum, and V54.3Gy for the femoral heads. We also computed the voxel-by-voxel absolute dose difference between the IOM dose distribution and the predicted dose distribution in the bladder, rectum, CTV, and PTV. From these difference distributions, we calculated each patient's median and 95th percentile dose difference. All comparisons between different models (either weights directly or dosimetric differences) were tested for statistical significance using the Mann–Whitney–Wilcoxon test.[28] Figure 1 provides a graphical overview of our methodology from input data, to model training, to prediction, to metrics analysis.

3. RESULTS

3.A. Weight prediction

To illustrate the results of each model, we provide examples of the predicted weight vectors for cohort *S*. Tables II, III, and IV show the predicted weight vectors for cohort *S* along with the population average (over all 315 patients) for the logistic regression, multinomial regression, and *K*-nearest neighbor model, respectively.

TABLE IV. The weighted *K*-nearest neighbor predicted weight vectors for cohort *S*.

| Patient | Predicted weights | | | | |
	Bladder	Rectum	Left femur	Right femur	PTV ring
1	0.881	0.075	0.004	0.003	0.037
2	0.228	0.707	0.003	0.003	0.058
3	0.221	0.719	0.003	0.003	0.054
4	0.365	0.571	0.003	0.003	0.057
5	0.186	0.750	0.003	0.003	0.057
Pop. Avg.	0.512	0.433	0.003	0.003	0.049
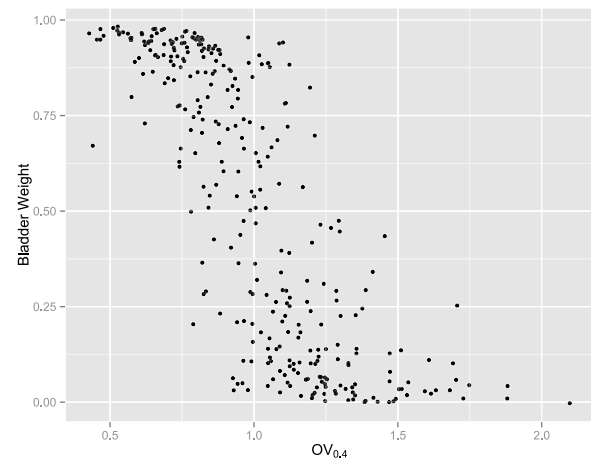
FIG. 2. *S*-shaped relationship between bladder weight and $OV_{0.4}$.

The bladder weight plotted against $OV_{0.4}$, the main predictor in all our models, is shown in Fig. 2. An $OV_{0.4} < 1$ implies that more bladder volume overlaps with the tumor than rectum volume, and hence, the bladder objective should receive more weight. The regression parameters for the logistic model were $\beta_0 = 6.44$, $\beta_1 = -7.41$, and $\beta_2 = 7.26$. For multinomial logistic regression, we report the average values of the regression parameters (across all leave-one-out iterations) for the rectum, which were $\beta_{R,0} = -6.69$ and $\beta_{R,1} = 6.60$ and the PTV ring, which were $\beta_{PR,0} = -6.30$ and $\beta_{PR,1} = 4.27$. In all cases tested, the regression parameters corresponding to the left and right femoral heads were insignificant. The 3-means clustering results are shown in Fig. 3—to present the results in three dimensions, the *y*-axis represents the summation of the left femoral head, right femoral head, and PTV ring weights. The centroid of each cluster, which corresponds to a representative weight vector, is shown as a black "*x*." Table V shows the centroid weight vectors along with the corresponding cluster sizes.

3.B. Weight comparison

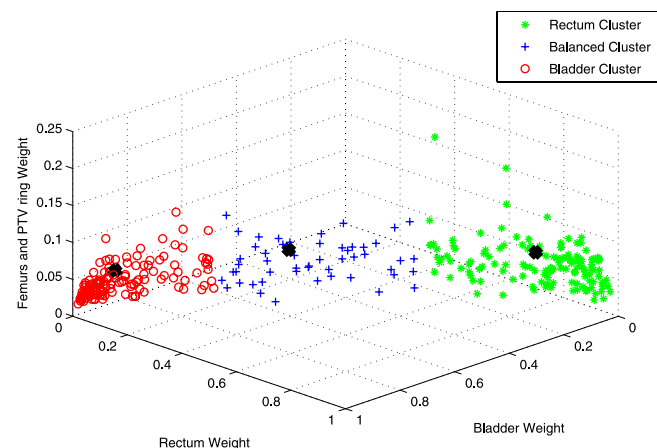For each patient, we computed the absolute difference between each model's predicted weights and the IOM weights to

FIG. 3. 3-means clustering classes with each centroid shown as a black *x*.

TABLE V. The centroids and class sizes determined by the 3-means clustering algorithm.

Cluster	Size	Centroid weights				
		Bladder	Rectum	Left femur	Right femur	PTV ring
Bladder	133	0.888	0.069	0.004	0.003	0.037
Balanced	51	0.528	0.405	0.004	0.003	0.060
Rectum	131	0.114	0.822	0.003	0.003	0.058

measure prediction error between the models. Figure 4 shows the median absolute difference, across all patients, between the IOM weights and the weights derived from the KNN model, LR model, MLR model, and Pop. Avg. The LR model performs the best for the bladder and rectum, with a median weight error that is 0.28 ($p < 0.05$) smaller for both organs than the corresponding error of the Pop. Avg. This represents a relative improvement of 72.7% ($p < 0.05$) and 73.7% ($p < 0.05$) for the bladder and rectum, respectively. The MLR model is not statistically different from the LR model, while the KNN model is statistically different with error that is 20% ($p < 0.05$) larger for the bladder and rectum compared to the LR model. The models were not statistically different for the femoral head weights. The KNN model performs the best for PTV ring, with a 15% ($p < 0.05$) relative improvement over the Pop. Avg.

3.C. Plan comparison

As mentioned in Sec. 2.E, we generated five treatment plans for each patient to compare how well the model predicted plans replicated the IOM plans. We first consider the clinical metrics for each patient's plan, which include bladder V54.3Gy and V70Gy, and rectum V54.3Gy and V70Gy. We compute the absolute difference between the metrics derived from the model predicted plans and the metrics derived from the IOM plans. Figure 5(a) shows the median absolute difference across all patients, while Fig. 5(b) shows the corresponding 95th percentile absolute difference. The femoral head values are not shown because the median differences are zero for all

models. Moreover, more than 75% of the patients achieve a V54.3Gy metric of 0% femoral head volume. All three models have statistically smaller median error than Pop. Avg. across all metrics but are not statistically different from each other at the 95% confidence level. The LR model has the smallest error for all but the median bladder V70Gy metric. The LR model's 0.15% volume ($p < 0.05$) improvement over the Pop. Avg. for the median rectum V70Gy is the largest relative difference, representing an improvement of 35% ($p < 0.05$). Similar improvements are observed for the other metrics.

Next, we used the five treatment plans to compute the dose deposited to every voxel (i.e., dose distribution) in the bladder, rectum, CTV, and PTV for each patient. Using these values, we calculated the voxel-by-voxel absolute difference between
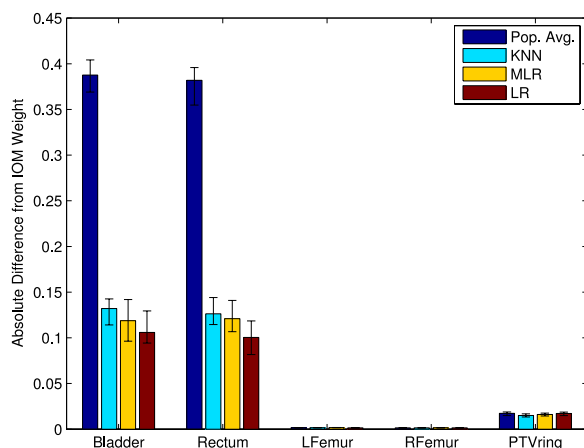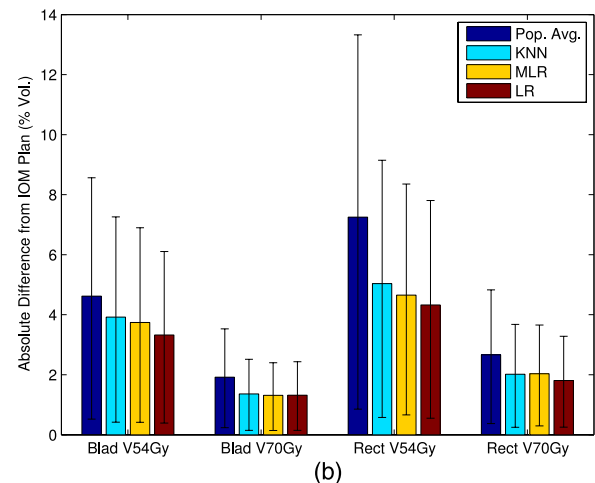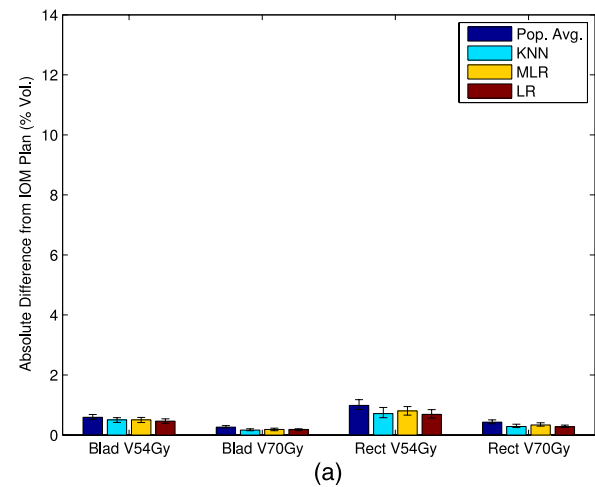

(a)

(b)

FIG. 5. Comparison of error between IOM and model predicted plans for the clinical acceptability criteria. (a) Median error and (b) 95th percentile error.

FIG. 4. Comparison of median weight error between predicted weights and population average weights for all OARs.

the IOM dose distribution and each model's dose distribution. To measure error, we then computed the median and 95th percentile absolute difference in each of the four structures for every patient. The distributions for these two error metrics across all patients will be denoted by the *median voxel error distribution* and *95th percentile voxel error distribution*, respectively. The voxel error distributions for the rectum are shown as histograms in Figs. 6(a) and 6(b). Figures 6(c) and 6(d) compare the same distributions for the CTV.

The median values of the median and 95th percentile voxel error distributions for all four structures are shown in Figs. 7(a) and 7(b), respectively. As with previous comparisons, all models outperform the Pop. Avg. by a statistically significant amount ($p < 0.05$) but do not differ from each other. Similarly, the LR model continues to trend toward the best across all metrics. The plans generated using the Pop. Avg. weights result in 105 patients with a 95th percentile rectum dose error that is 4 Gy larger than the LR model plans. Moreover, the Pop. Avg. plans result in 70 patients with a median rectum dose error that is 1 Gy larger than LR. Again, the model predicted distributions tend to be centered closer to zero with shorter tails. Similar distributions are observed for the bladder and PTV.

Finally, we highlight some observational results related to patient insensitivity to weight values. Figure 8(a) shows a patient who is insensitive to the bladder weight—the IOM plan has a bladder weight of 0.97 while the Pop. Avg. plan has a bladder weight of 0.51—as the bladder DVH curves are nearly identical for the Pop. Avg., LR, and IOM plans. This weight difference allows the Pop. Avg. plan to achieve greater rectum sparing at almost no cost; entirely due to the more efficient utilization of the objective weight. Figure 8(b) displays a patient who is rectum weight insensitive. In this case, the IOM plan has a rectum weight of 0.93 while the Pop. Avg. plan has a rectum weight of 0.44. Figure 8(c) displays a patient who is insensitive to both the bladder and rectum weights. For this patient, the bladder and rectum DVHs are nearly identical, yet the IOM rectum and bladder weights are (0.13, 0.79), the LR rectum and bladder weights are (0.36, 0.58), and the Pop. Avg. rectum and bladder weights are (0.51, 0.44).

4. DISCUSSION

Our models can predict new patient weight vectors that are comparable to the clinically acceptable weight vectors
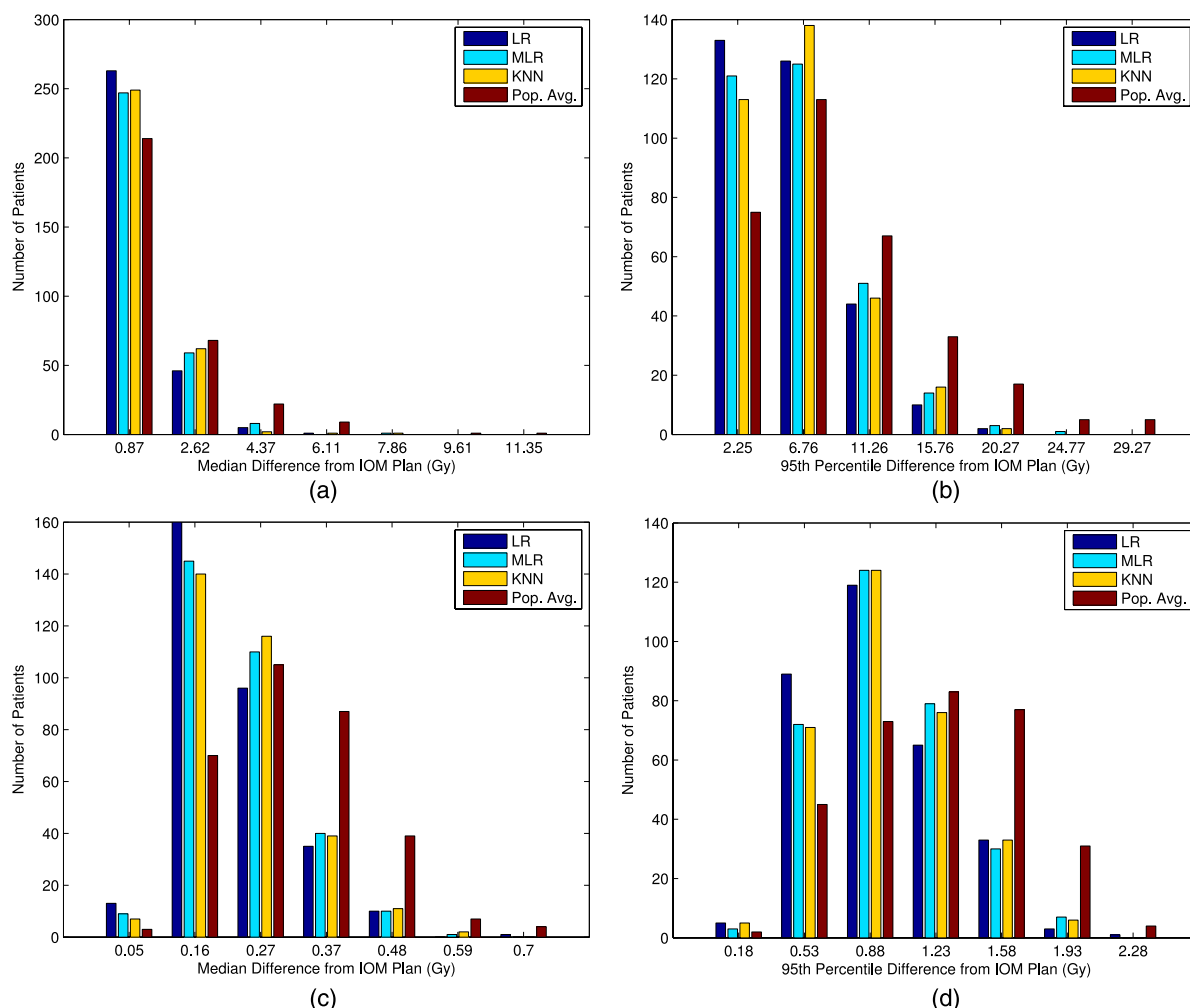

FIG. 6. Comparison of voxel error distributions. (a) Median rectum dose, (b) 95th percentile rectum dose, (c) median CTV dose, and (d) 95th percentile CTV dose.
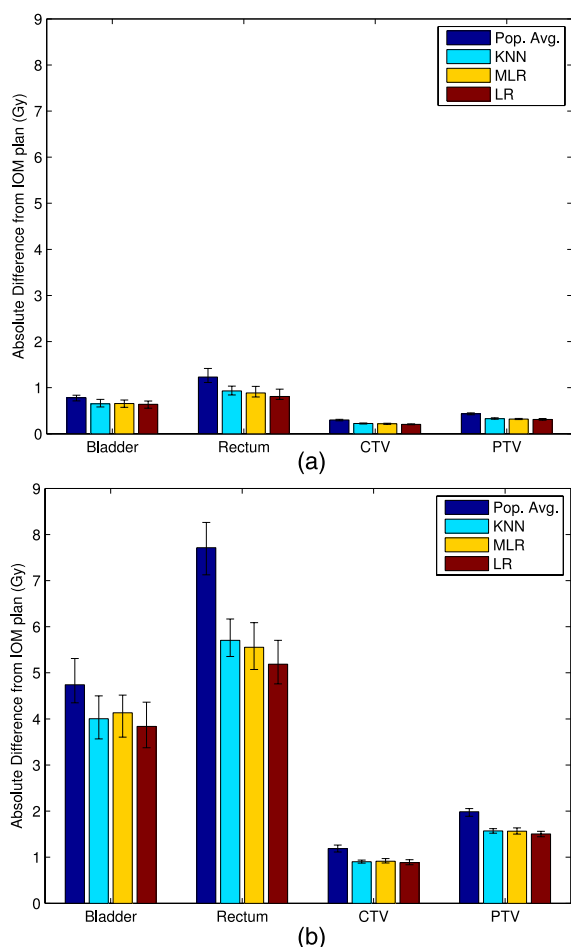
FIG. 7. Comparison of median error between IOM and model predicted plans for voxel error distributions. (a) Median voxel error distribution and (b) 95th percentile voxel error distribution.

determined by the IOM. Our aim was to replicate the IOM weights and thus the IOM treatment plans. The idea is to provide the planner with a good set of starting weights, from which only minor adjustment would be needed ideally. Interestingly, we observed that for 20% of our patients, the predicted weight vectors lead to treatment plans that are strictly better than the IOM treatment plans for all four bladder and rectum clinical metrics. This implies that for 20% of patients, no further personalization or work by the planner is needed to achieve clinically acceptable plans.

As more treatment plans are processed and used to train our models, there is potential for increased prediction accuracy. Since our models "learn" from historical plans, we can expect future plans to exhibit similar trade-offs, which may lead to a method of plan standardization. Furthermore, because our models predict objective function weights, each planner will begin the treatment planning process at an informed position. This may cause a reduction in overall planning time and may potentially lead to improved clinical plans. Weight prediction may also be used to strengthen existing automated treatment planning techniques by allowing them to consider and predict trade-offs between OARs.

Although all three models (LR, KNN, MLR) generated treatments that were not statistically different in any clinical
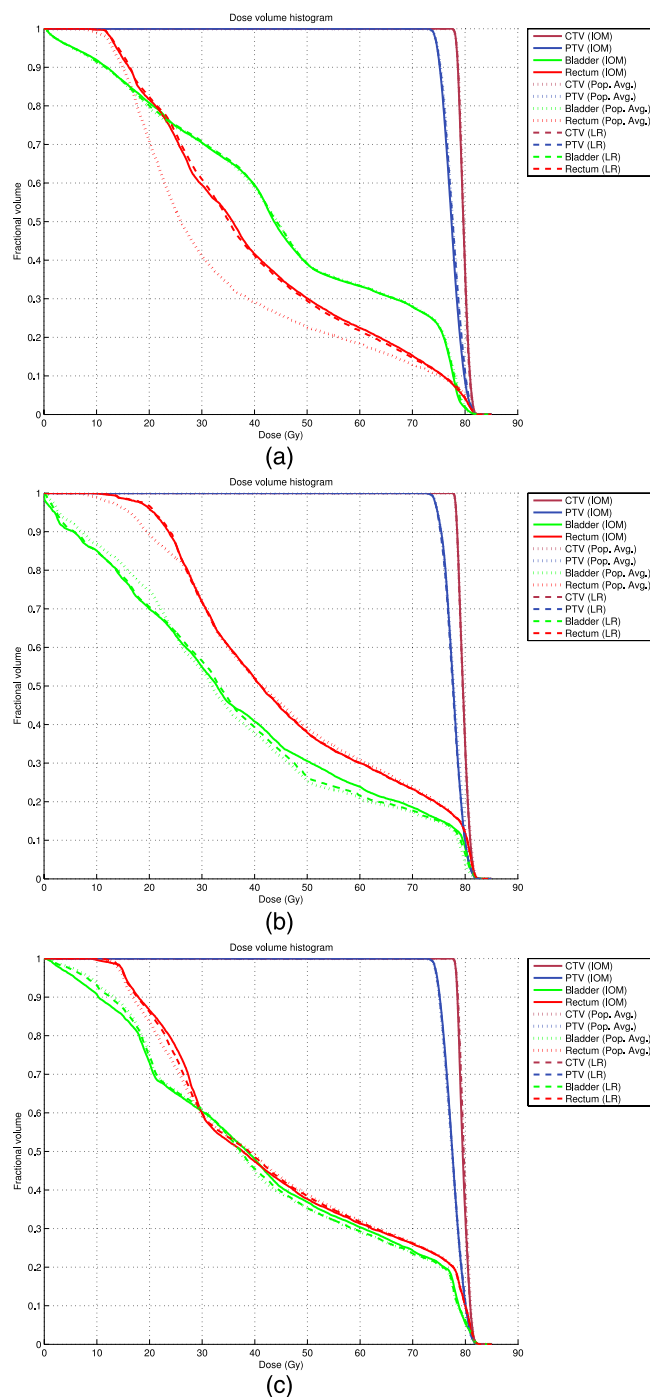

FIG. 8. DVH curves for patients exhibiting weight insensitivity. (a) Bladder insensitive, (b) rectum insensitive, and (c) bladder and rectum insensitive.

metrics, our results suggest that the LR model trended toward the most accurate for prostate cancer. We believe this result is likely due to low weight values for the femurs and PTV ring, which is a result of prostate treatment planning being predominately binomial in OAR trade-offs. However, a major limitation of the LR model is that it is only applicable to tumor sites that have a bicriteria trade-off. Fortunately, the MLR and KNN models, which can predict any number of objective weights, performed similarly. Thus, we believe that these advanced models have much broader applicability. Application of the KNN and MLR methods to head-and-neck

data, where many complex trade-offs are present, is the subject of future work.

Comparing MLR and KNN, we found that they were not statistically different for any clinical metric. The MLR model performed slightly better than the KNN model in rectum, CTV, and PTV voxel dose, but the KNN model performed better for the bladder voxel dose and the DVH metrics used at Princess Margaret. The trade-off between the MLR and KNN models is a matter of personalized weight prediction versus simplicity. The KNN model is a simple classification-based model with minimal model tuning and clear model interpretation. In contrast, the MLR model is not easily interpretable and requires careful model tuning. It is not clear which model is more valuable for weight prediction in more complicated tumor sites, such as head-and-neck or prostate cancer with lymph nodes. In either case, application to a more complicated tumor site will likely require the development of new patient features.

We acknowledge that there may exist other approaches for weight prediction and discuss one such method here. In this approach, we would use historical plans and patient features to train machine learning models that can predict achievable dose for multiple OARs. We would then use these dose metrics as input to an inverse optimization model to determine the weights needed to recreate the achievable dose distribution. In other words, this approach flips the order of inverse optimization and machine learning compared to the methodology in this paper. To implement such an approach, new methodological developments are needed. In particular, the machine learning model would need to simultaneously predict achievable doses for multiple OARs and the inverse optimization model would need to work with only DVH data. Such developments are beyond the scope of this paper and are left for future work. We hypothesize that the performance of this approach will be similar to ours if the historical plans are close to what is best achievable with respect to OAR trade-offs.

The idea of patient insensitivity is shown in Fig. 8. Approximately 20% of patients were insensitive to weight values, meaning that the FOP will result in nearly identical DVHs for a range of weights. To classify insensitivity, we computed 20 bladder and 20 rectum DVH curves corresponding to 20 different weight vectors for each patient. For these weight vectors, LF, RF, and PR were set to the population average and (bladder, rectum) weights ranged from (0,0.95), (0.05,0.9),…, (0.9,0.05), (0.95,0). We plotted all 20 bladder (rectum) DVH curves on the same axis to obtain a bladder (rectum) "DVH cloud." We then computed the area of each patient's rectum and bladder DVH cloud. Patients with a bladder (rectum) DVH cloud area below the 10th percentile of all bladder (rectum) DVH cloud areas were classified as insensitive. For weight insensitive patients, personalized weight prediction is less valuable and it makes sense to employ a more streamlined prediction method such as KNN. Moreover, for highly insensitive patients, it may be possible to simply apply the Pop. Avg. weights.

Determining a metric for classifying patient sensitivity is an important topic for further research. The ability to identify insensitive patients *a priori* will allow the treatment planner

to manage their time and effort put into sparing an OAR that cannot be improved. In addition, if an OAR is insensitive, the treatment planner may choose to give more weight to another OAR where the marginal benefit in dose reduction will be much greater. In combination, the ability to predict patient sensitivity may significantly increase plan quality while simultaneously reducing planning time.

5. CONCLUSION

In this paper, we developed weight prediction methodologies that can simultaneously predict multiple weights and capture the trade-offs associated with sparing multiple OARs near the tumor. By applying these methods to the largest patient data set in the literature, we demonstrated that multinomial logistic regression and *K*-nearest neighbors perform similarly to logistic regression and can produce treatment plans that closely replicate clinical plans. Most importantly, the MLR and KNN methods have the potential to be applied to more complicated tumor sites.

The potential clinical benefits of this work are the ability to standardize and streamline the treatment planning process by reducing time spent manually tuning objective function weights, while at the same time maintaining personalization of the treatments. We see this work as supporting and possibly providing a preliminary method for automated treatment planning of prostate cancer.

Finally, this study introduces the idea of patient weight sensitivity, an important direction for future research, which may further reduce planning time while increasing plan quality through OAR sparing. Future work involves developing a methodology to determine patient sensitivity *a priori* and classify patients accordingly.

ACKNOWLEDGMENTS

a)Author to whom correspondence should be addressed. Electronic mail: j.boutilier@mail.utoronto.ca

[1]Prostate Cancer Foundation, *Prostate Cancer Statistics 2014* (Prostate Cancer Foundation, Santa Monica, California, 2014).

[2]Y. Yu, "Multiobjective decision theory for computational optimization in radiation therapy," Med. Phys. **24**, 1445–1454 (1997).

[3]L. Xing, J. G. Li, S. Donaldson, Q. T. Le, and A. L. Boyer, "Optimization of importance factors in inverse planning," Phys. Med. Biol. **44**, 2525–2536 (1999).

[4]C. Cotrutz and L. Xing, "Using voxel-dependent importance factors for interactive DVH-based dose optimization," Phys. Med. Biol. **47**, 1659–1669 (2002).

[5]C. Wu, G. H. Olivera, R. Jeraj, H. Keller, and T. R. Mackie, "Treatment plan modification using voxel-based weighting factors/dose prescription," Phys. Med. Biol. **48**, 2479–2491 (2003).

[6]C. Cotrutz, M. Lahanas, C. Kappas, and D. Baltas, "A multiobjective gradient-based dose optimization algorithm for external beam conformal radiotherapy," Phys. Med. Biol. **46**, 2161–2175 (2001).

[7]H. W. Hamacher and K. H. Küfer, "Inverse radiation therapy planning – A multiple objective optimization approach," Discrete Appl. Math. **118**, 145–161 (2002).

[8]D. L. Craft, T. F. Halabi, H. A. Shih, and T. R. Bortfeld, "Approximating convex Pareto surfaces in multiobjective radiotherapy planning," Med. Phys. **33**, 3399–3407 (2006).

[9]L. Shao and M. Ehrgott, "Approximately solving multiobjective linear programmes in objective space and an application in radiotherapy treatment planning," Math. Methods Oper. Res. **68**, 257–276 (2008).

[10]R. Bokrantz and A. Forsgren, "An algorithm for approximating convex Pareto surfaces based on dual techniques," INFORMS J. Comput. **25**, 377–393 (2013).

[11]M. Zarepisheh, T. Long, N. Li, Z. Tian, H. E. Romeijn, X. Jia, and S. B. Jiang, "A DVH-guided IMRT optimization algorithm for automatic treatment planning and adaptive radiotherapy replanning," Med. Phys. **41**, 061711 (14pp.) (2014).

[12]T. C. Y. Chan, T. Craig, T. Lee, and M. B. Sharpe, "Generalized inverse multi-objective optimization with application to cancer therapy," Oper. Res. **62**, 680–695 (2014).

[13]T. Lee, M. Hammad, T. C. Y. Chan, T. Craig, and M. B. Sharpe, "Predicting objective function weights from patient anatomy in prostate cancer IMRT treatment planning," Med. Phys. **40**, 121706 (9pp.) (2013).

[14]B. Wu, F. Ricchetti, G. Sanguineti, M. Kazhdan, P. Simari, M. Chuang, R. Taylor, R. Jacques, and T. McNutt, "Patient geometry-driven information retrieval for IMRT treatment plan quality control," Med. Phys. **36**, 5497–5505 (2009).

[15]B. Wu, F. Ricchetti, G. Sanguineti, M. Kazhdan, P. Simari, R. Jacques, R. Taylor, and T. McNutt, "Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning," Int. J. Radiat. Oncol., Biol., Phys. **79**, 1241–1247 (2011).

[16]S. F. Petit, B. Wu, M. Kazhdan, A. Dekker, P. Simari, R. Kumar, R. Taylor, J. M. Herman, and T. McNutt, "Increased organ sparing using shape-based treatment plan optimization for intensity modulated radiation therapy of pancreatic adenocarcinoma," Radiother. Oncol. **102**, 38–44 (2012).

[17]D. Good, J. Lo, W. R. Lee, Q. J. Wu, F. F. Yin, and S. K. Das, "A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: an example application to prostate cancer planning," Int. J. Radiat. Oncol., Biol., Phys. **87**, 176–181 (2013).

[18]V. Chanyavanich, S. K. Das, W. R. Lee, and J. Y. Lo, "Knowledge-based IMRT treatment planning for prostate cancer," Med. Phys. **38**, 2515–2522 (2011).

[19]Y. Wang, A. Zolnay, L. Incrocci, H. Joosten, T. R. McNutt, B. Heijmen, and S. Petit, "A quality control model that uses PTV-rectal distances to predict lowest achievable rectum dose, improves IMRT planning for patients with prostate cancer," Radiother. Oncol. **107**, 352–357 (2013).

[20]X. Zhu, Y. Ge, T. Li, D. Thongphiew, F.-F. Yin, and Q. J. Wu, "A planning quality evaluation tool for prostate adaptive IMRT based on machine learning," Med. Phys. **38**, 719–726 (2011).

[21]Y. Yang, E. C. Ford, B. Wu, M. Pinkawa, B. van Triest, P. Campbell, D. Y. Song, and T. R. McNutt, "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 (10pp.) (2013).

[22]R. Lu, R. J. Radke, L. Hong, C.-S. Chui, J. Xiong, E. Yorke, and A. Jackson, "Learning the relationship between patient geometry and beam intensity in breast intensity-modulated radiotherapy," IEEE Trans. Biomed. Eng. **53**, 908–920 (2006).

[23]K. L. Moore, R. S. Brame, D. A. Low, and S. Mutic, "Experience-based quality control of clinical intensity-modulated radiotherapy planning," Int. J. Radiat. Oncol., Biol., Phys. **81**, 545–551 (2011).

[24]L. M. Appenzoller, J. M. Michaski, W. L. Thorstad, S. Mutic, and K. L. Moore, "Predicting dose-volume histograms for organs-at-risk in IMRT planning," Med. Phys. **39**, 7446–7461 (2012).

[25]L. Yuan, Y. Ge, W. R. Lee, F. F. Yin, J. P. Kirkpatrick, and Q. J. Wu, "Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans," Med. Phys. **39**, 6868–6878 (2012).

[26]J. O. Deasy, A. I. Blanco, and V. H. Clark, "CERR: A computational environment for radiotherapy research," Med. Phys. **30**, 979–985 (2003).

[27]R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2013).

[28]M. H. Kutner, C. J. Nachtshiem, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. (McGraw-Hill, New York, NY, 2005).

[29]E. W. Forgy, "Cluster analysis of multivariate data: Efficiency versus interpretability of classifications," Biometrics **21**, 768–769 (1965).

[30]R. J. Samworth, "Optimal weighted nearest neighbour classifiers," Ann. Stat. **40**, 2733–2763 (2012).