Medical Physics
The International Journal of Medical Physics Research and Practice

Sample size requirements for knowledge-based treatment planning

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

Citation: Medical Physics **43**, 1212 (2016); doi: 10.1118/1.4941363

View online: http://dx.doi.org/10.1118/1.4941363

View Table of Contents: http://scitation.aip.org/content/aapm/journal/medphys/43/3?ver=pdfcov

Published by the American Association of Physicists in Medicine

Sample size requirements for knowledge-based treatment planning

Justin J. Boutilier[a)]
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

(Received 18 August 2015; revised 10 January 2016; accepted for publication 24 January 2016; published 9 February 2016)

Purpose: To determine how training set size affects the accuracy of knowledge-based treatment planning (KBP) models.
Methods: The authors selected four models from three classes of KBP approaches, corresponding to three distinct quantities that KBP models may predict: dose–volume histogram (DVH) points, DVH curves, and objective function weights. DVH point prediction is done using the best plan from a database of similar clinical plans; DVH curve prediction employs principal component analysis and multiple linear regression; and objective function weights uses either logistic regression or K-nearest neighbors. The authors trained each KBP model using training sets of sizes $n = 10$, 20, 30, 50, 75, 100, 150, and 200. The authors set aside 100 randomly selected patients from their cohort of 315 prostate cancer patients from Princess Margaret Cancer Center to serve as a validation set for all experiments. For each value of n, the authors randomly selected 100 different training sets with replacement from the remaining 215 patients. Each of the 100 training sets was used to train a model for each value of n and for each KBT approach. To evaluate the models, the authors predicted the KBP endpoints for each of the 100 patients in the validation set. To estimate the minimum required sample size, the authors used statistical testing to determine if the median error for each sample size from 10 to 150 is equal to the median error for the maximum sample size of 200.
Results: The minimum required sample size was different for each model. The DVH point prediction method predicts two dose metrics for the bladder and two for the rectum. The authors found that more than 200 samples were required to achieve consistent model predictions for all four metrics. For DVH curve prediction, the authors found that at least 75 samples were needed to accurately predict the bladder DVH, while only 20 samples were needed to predict the rectum DVH. Finally, for objective function weight prediction, at least 10 samples were needed to train the logistic regression model, while at least 150 samples were required to train the K-nearest neighbor methodology.
Conclusions: In conclusion, the minimum required sample size needed to accurately train KBP models for prostate cancer depends on the specific model and endpoint to be predicted. The authors' results may provide a lower bound for more complicated tumor sites. © *2016 American Association of Physicists in Medicine.* [http://dx.doi.org/10.1118/1.4941363]

Key words: knowledge-based treatment planning, sample size

1. INTRODUCTION

In recent years, knowledge-based treatment planning (KBP) has garnered significant interest from both the academic and clinical communities. KBP exploits information from historical treatment plans to predict metrics for new treatment plans. These predicted metrics can be used as a reference for creating new treatment plans, as a quality control tool during traditional treatment planning, or as part of an automated treatment planning framework.[1–6]

Most research on KBP leverages machine learning methods in combination with patient anatomical features to predict plan quality metrics or treatment plan parameters. The quantity and quality of treatment plans used to train these models can directly influence the accuracy of the predicted metrics. In general, a larger pool of historical plans increases the observed variations in organ geometries, allowing the model to exploit this information to make more accurate predictions.

Based on the predicted endpoint, KBP methods generally fall into one of three categories. The first approach seeks to predict specific points on an organ-at-risk (OAR) dose–volume histogram (DVH).[7–11] Generally, the predicted DVH points correspond to clinical quality control or acceptability criteria, which are subsequently used as dosimetric goals during the planning process. The second approach attempts to predict the complete DVH curve for each OAR.[12–14] The predicted DVH curve may then be used as a reference plan in standard planning approaches or as a target plan during treatment planning optimization. The final approach predicts optimization objective function weights, which correspond to the relative importance of each OAR.[15,16] These weights are then used to generate a treatment plan that can be used as an advanced starting point during treatment planning.

KBP solutions may be developed in-house, as discussed above, or as commercial software. KBP has garnered substantial interest from commercial treatment planning software companies. For example, Varian medical systems recently developed a KBP tool called RapidPlan. RapidPlan uses patient anatomy to predict DVH curves and optimization objectives for intensity modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT) treatment plans. Our work in this paper was motivated in part by RapidPlan's frequently asked questions, which states, *"The minimum number of plans required for model creation is 20; however, adding additional plans will usually help create a more robust model."*[17] In particular, we aim to quantify the minimum number of plans required to create an accurate KBP model. This fundamental topic has not yet been explored in the academic literature and, given the trend toward clinical implementation of KBP, requires immediate attention.

In this study, we develop an experimental methodology to determine the minimum number of treatment plans required to create an accurate KBP model. We select four models from the three KBP approaches described above and apply our methodology to investigate how the size of the training set (i.e., "sample size") influences the quality of predicted endpoints. The predicted endpoints we analyze are the specific quantities that are output from the various prediction methodologies, which can be used in a subsequent treatment planning step. We use a large dataset of 315 prostate cancer patients from the Princess Margaret Cancer Center to complete our experiments. Statistical testing is used to estimate the minimum sample size needed to achieve consistent accuracy metrics for each KBP model. We believe that because prostate cancer has relatively homogeneous anatomy, our results may provide a lower bound on the required sample size for more complicated tumor sites.

2. METHODS AND MATERIALS

2.A. Data

This study used a previously obtained data set,[16] which contains 315 prostate IMRT treatment plans delivered at the Princess Margaret Cancer Centre as a part of the PROFIT trial (NCT00304759). All patients were obtained from the same trial arm and had a prescribed dose of 78 Gy in 39 fractions. Retrospective data access was granted under UHN REB 137 11-0107-CE. All plans were exported from the clinical treatment planning system (Pinnacle, Phillips Radiation Oncology Systems, Madison, WI) in DICOM format and imported into MATLAB via CERR (computational environment for radiotherapy research).[18]

2.B. Models

We selected models from the three KBP categories outlined in the Introduction. We investigate a single model for both DVH point and DVH curve prediction and we consider two models for objective function weight prediction. We give brief overviews of each model and refer the interested reader to the original papers for more details.

2.B.1. DVH point prediction

We tested the methodology proposed by Wu *et al.* for predicting bladder and rectum achievable dose metrics.[7] The goal of this methodology is to use a database of previous patients with similar anatomy to predict a new patient's achievable dose.

We generated OVH curves[19] for the bladder and rectum using PTV expansions of 0–25 mm in 1 mm increments. OVH curve metrics are used in the database lookup to exclude dissimilar patients.

Let $D_{v,i}$ denote the dose to a specific fractional volume, v, of patient i (i.e., a point on the DVH curve). Let $V_{95,i}$ denote the percentage of the PTV receiving 95% of the prescribed dose for patient i, and let $r_{v,i}$ denote the PTV expansion distance resulting in an overlap percentage volume of v for patient i (i.e., a point on the OVH curve). The predicted dose metric for the bladder (or rectum) of a new patient j was defined as

$$D_{v,j} = \min_{i=1,\ldots,n} \{D_{v,i} | r_{v,j} \geq r_{v,i} \text{ and } V_{95,i} \geq 99\%\}. \quad (1)$$

In this study, we set $v = 30\%$ and $v = 50\%$ for both the bladder and rectum. These four metrics correspond to the clinical acceptability criteria ($V54$ Gy $\leq 50\%$ and $V70$ Gy $\leq 30\%$) at the Princess Margaret Cancer Centre. If the set $\{r_{v,j} \geq r_{v,i} \text{ and } V_{95,i} \geq 99\%\}$ was empty for patient j, we set $D_{30,j} = 70$ Gy and $D_{50,j} = 54$ Gy. DVH point prediction error, defined as the absolute difference between the predicted and clinical dosage, was computed for each patient.

2.B.2. DVH curve prediction

For DVH curve prediction, we tested a hybrid of the methodologies proposed in Zhu *et al.* and Yuan *et al.*, which we believe are similar to the method used by Varian's RapidPlan.[12,13,17] In this methodology, the goal is to predict complete DVH curves for the bladder and rectum with the corresponding DTH curves and all organ volumes as independent variables.

The independent variables for this method are DTH curves and organ volumes. We generated DTH curves[12] for the bladder and rectum using Euclidean distance. The number of points on each DTH curve corresponds to the number of voxels in the corresponding OAR. We also computed the volume in cubic centimeters (cm^3) for the bladder, rectum, and PTV.

We sampled 50 equally spaced points from each patient's bladder and rectum DVH and DTH curves as outlined in Zhu *et al.* We then used principal component analysis (PCA) to further reduce the dimension of each DVH and DTH curve to two. During this process, we obtained the *loading matrix* for each of the four PCAs (bladder/rectum and DTH/DVH), which are used to project between principal component space and feature space.[20]

Let $\text{DVH}_{1,s}$ and $\text{DVH}_{2,s}$ correspond to the first and second DVH principal component of OAR s, $s \in \{b,r\}$ where b denotes the bladder and r denotes the rectum. Let $\text{DTH}_{1,s}$ and $\text{DTH}_{2,s}$ correspond to the first and second DTH principal component of OAR s. Let V_s denote the volume in cm^3 of OAR s and let V_{PTV} denote the volume of the PTV. The four regression models were defined as

$$\begin{aligned}\text{DVH}_{p,s} = {} & \beta_0^{p,s} + \beta_1^{p,s}\text{DTH}_{1,s} \\ & + \beta_2^{p,s}\text{DTH}_{2,s} + \beta_3^{p,s}V_b + \beta_4^{p,s}V_r \\ & + \beta_5^{p,s}V_{\text{PTV}}, \quad p = 1,2,\ s \in \{b,r\}.\end{aligned} \tag{2}$$

For each testing set patient, we used the loading matrices to project the 50-dimensional DTH curves onto the 2D principal component space (i.e., $\text{DTH}_{1,s}$ and $\text{DTH}_{2,s}$). We then used the estimated regression coefficients (i.e., $\beta_0^{p,s},\ldots,\beta_5^{p,s}$) to predict the corresponding DVH principal components (i.e., $\text{DVH}_{1,s}$ and $\text{DVH}_{2,s}$). Finally, given the predicted DVH principal components, we used the corresponding loading matrices to project back to 50-dimensional DVH space.

DVH curve prediction error, defined as the total area difference between the predicted and clinical DVH, was computed for each patient. DVH point prediction error, defined as the absolute difference between the predicted and clinical dosage, was also computed. To accomplish this, we extracted D_{50} and D_{30} (i.e., dose to 50% and 30% volume) from both the bladder and rectum curves.

2.B.3. Objective function weight prediction

At Princess Margaret Cancer Centre, treatment planners apply a predefined template of dose objectives to each patient. Planners then modify the objective function weights corresponding to each objective to personalize the treatment plan. The ability to accurately predict objective function weights removes the need for planners to modify the weights manually.

We applied two weight prediction models from Boutilier *et al.*[16] We chose the logistic regression model as a benchmark because it performed best for prostate cancer, and we chose the K-nearest neighbor model because of its simplicity and its potential for application to more complicated tumor sites. A previously developed inverse optimization method (IOM) was applied to determine five optimal objective function weights for each patient, corresponding to the bladder (b), rectum (r), left femur (lf), right femur (rf), and PTV ring (pr), a structure used to promote dose conformity.[22] IOM takes a previously delivered treatment plan as input and determines which objective function weights are required to recreate the given plan. The logistic regression model predicts the bladder and rectum weight while the K-nearest neighbor (KNN) model simultaneously predicts all five objective function weights. Note that the five objective function weights are normalized so that they sum to one.

Using OVH curves, we computed the ratio of rectum overlap volume to bladder overlap volume for each PTV expansion value $x \in \{0,1,\ldots,25 \text{ mm}\}$, which we denote as OV_x. Next, we computed the slope of the bladder and rectum OVH curves between each pair of consecutive expansion points in $\{0,1,\ldots,25 \text{ mm}\}$. We denote these slopes by $\text{OVSB}_{x,x+1}$ and $\text{OVSR}_{x,x+1}$ for the bladder and rectum, respectively. We used these OVH related metrics as independent variables for weight prediction.

The logistic regression model used two patient features, OV_4 and $\text{OVSR}_{0,1}$, as previously determined.[16] The functional form of the logistic regression equation was defined as

$$\alpha_b = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{OV}_4 + \beta_2 \text{OVSR}_{0,1})}}, \tag{3}$$

where α_b denotes the bladder weight. The rectum weight was computed *post hoc* as $\alpha_r = 1 - \alpha_b$. The bladder and rectum weights were multiplied by 0.944 to accommodate nonzero weights for the left femur, right femur, and PTV ring. The value of 0.944 was determined by setting the lf, rf, pr weights to their population averages as determined by the IOM.[16]

For the KNN model, we first used a 3-means clustering algorithm to partition the objective function weights into three distinct groups, corresponding to patients that are bladder-weighted (b), rectum-weighted (r), and roughly equally weighted or balanced (a). 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. In particular, the KNN model used a triangular kernel function and two patient features, OV_4 and $\text{OVSB}_{13,14}$ with $K = 14$.[16] When the training set size is equal to 10, all neighbors are considered (i.e., $K = 10$). Let p_i denote the probability a patient belongs to cluster i and let $\mathbf{c}_i$ denote the centroid weight vector of cluster i. For each patient, the KNN predicted weight vector, $\boldsymbol{\alpha}_{\text{KNN}}$, was defined as

$$\boldsymbol{\alpha}_{\text{KNN}} = p_b\mathbf{c}_b + p_a\mathbf{c}_a + p_r\mathbf{c}_r. \tag{4}$$

Objective function weight error, defined as the absolute difference between the IOM weights and the model predicted weights, was computed for each patient.

2.C. Experimental setup

We first outline our methodology at a high level, which can be used to determine the minimum required sample size for any KBP model. First, randomly select roughly 1/3 of the patients to serve as a validation set for all experiments. Next, consider a set of n different training set sizes starting from a sample size of 10 patients. For each value of n, randomly select 100 different training sets with replacement from the remaining 2/3 patients (i.e., bootstrap). Use each of the 100 training sets to train each KBP model for each value of n. Finally, for each KBP approach, use each of the models (i.e., 100 for each n) to predict KBP endpoints for all validation set patients.

Now we specialize this general methodology and outline the specifics of our experimental setup. We first set aside 100 randomly selected patients (i.e., 32%) to serve as a validation set for all experiments. We considered eight different training set sizes corresponding to $n = 10$, 20, 30, 50, 75, 100, 150, and 200 patients. For each value of n, we randomly selected 100 different training sets with replacement from the remaining 215 patients (i.e., bootstrap). Each of the 100 training sets was used to train each of the four KBP models for each value of n. Finally, for each KBP approach, we used each of the 800 models (i.e., 100 for each n) to predict the KBP endpoints for the 100 testing set patients.

To evaluate the performance of the models, we compared what we refer to as traditional model error. For each approach, model error is defined as the variation between the predicted model value and the true value for each of the 100 patients in the validation set. For each of the 100 bootstrapped training sets, we computed the median error across all 100 patients in the validation set. For illustration, we plot the distribution of error (across the bootstraps) as a function of sample size using a series of box-and-whisker plots.

To estimate the minimum required sample size, we recommend using the Mann–Whitney–Wilcoxon[21] test to determine when increasing the sample size no longer provides a statistically significant increase in model accuracy at the 99% significance level. To do this, we test the null hypotheses that the median error for each sample size from 10 to 150 is equal to the median error for the maximum sample size of 200. The smallest sample size without a statistically significant difference in median error is characterized as the minimum required sample size for that model. If all sample sizes result in a statistically significant difference in model error, we conclude that a sample size of at least 200 is required.

3. RESULTS

3.A. DVH point prediction

Figure 1 shows the error distribution for each of the eight training set sample sizes for the four DVH objectives.

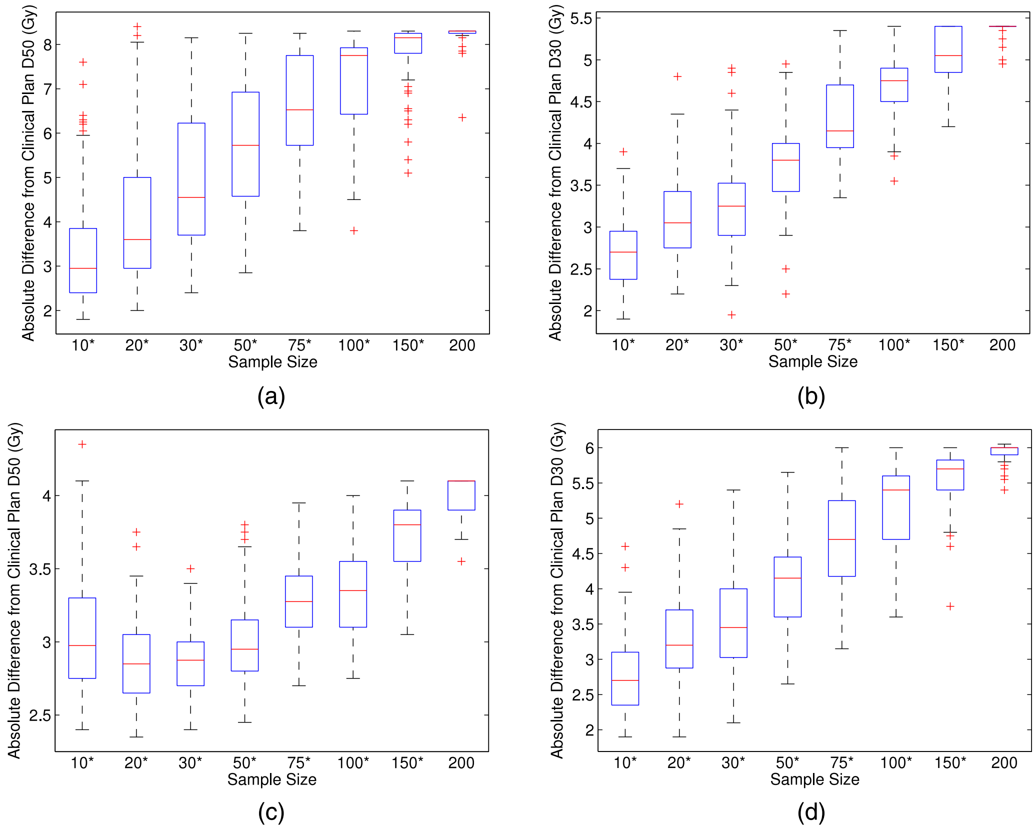


Fig. 1. The distribution of error for each DVH objective. (a) bladder D_{50}^{B}, (b) bladder D_{30}^{B}, (c) rectum D_{50}^{R}, and (d) rectum D_{30}^{R}. The upper and lower edges of the boxes depict the 75th and 25th percentiles, respectively. The horizontal lines inside the box depict the median, and the crosses represent values that are more than 1.5 times the interquartile range (IQR) away from the median (i.e., outliers). A starred x-axis value (i.e., 10*) indicates statistical significance and the smallest value without a star is the minimum required sample size.

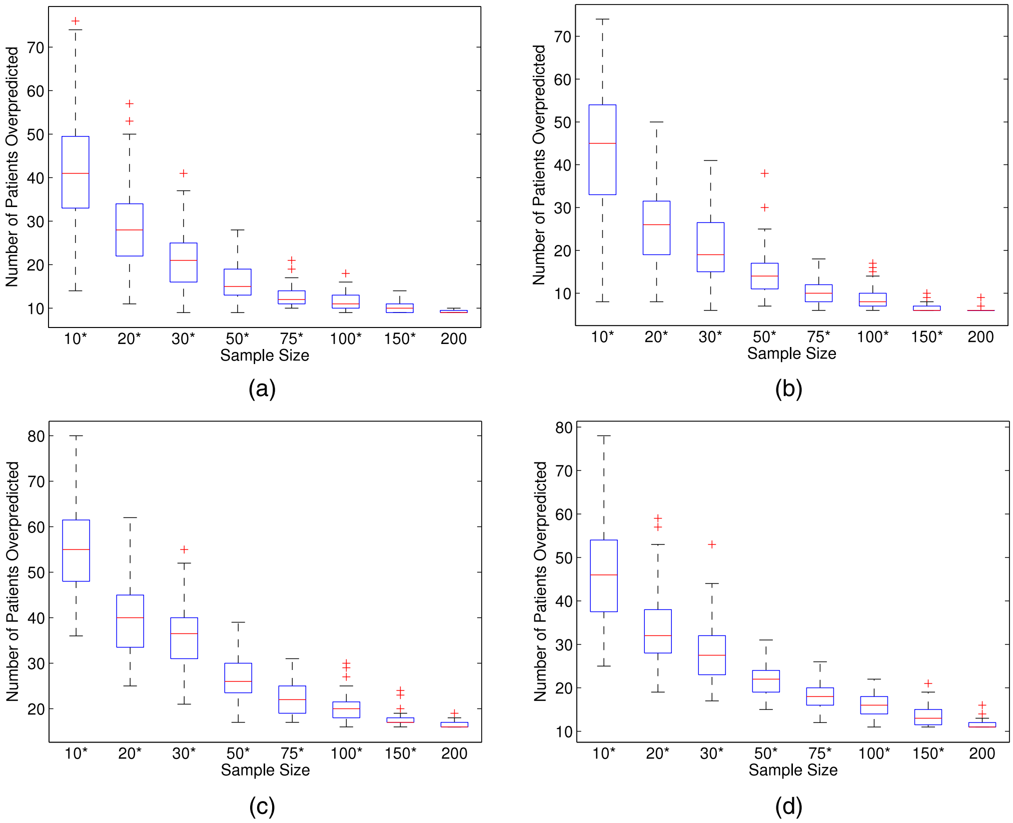


FIG. 2. The distribution of the number of patients overpredicted for each DVH point. (a) bladder D^B_{50}, (b) bladder D^B_{30}, (c) rectum D^R_{50}, and (d) rectum D^R_{30}.

Figures 1(a)–1(d) illustrate the error distributions for bladder D^B_{50}, bladder D^B_{30}, rectum D^R_{50}, and rectum D^R_{30}, respectively.

Figure 1 shows that the prediction error increases as a function of sample size for the DVH point method. The increasing error is a result of predicted dose metrics that are so low they are not realistically achievable for new patients. The DVH point model uses the minimum dose metric across all similar patients [cf., Eq. (1)] causing the lowest delivered dose in the eligible pool of patients to decrease as the sample size increases.

To further investigate the increasing model error, we first define *underprediction* as predicting a dosimetric value better than the clinical (i.e., type I error) and *overprediction* as predicting a dosimetric value worse than the clinical (i.e., type II error). To examine this phenomenon, we counted the number of patients that were overpredicted in each bootstrap sample. Figure 2 shows the distribution of the number of patients overpredicted for each predicted DVH point.

For each of the four DVH point metrics, the number of patients overpredicted at each sample size from 10 to 150 was significantly different from the number of patients overpredicted at a sample size of 200. These results suggest that the number of overpredicted patients is still decreasing and a sample size of 200 or more is required to obtain consistent model predictions.

3.B. DVH curve prediction

Figures 3(a) and 3(b) show the distribution of error as sample size increases for the bladder and rectum DVH, respectively.

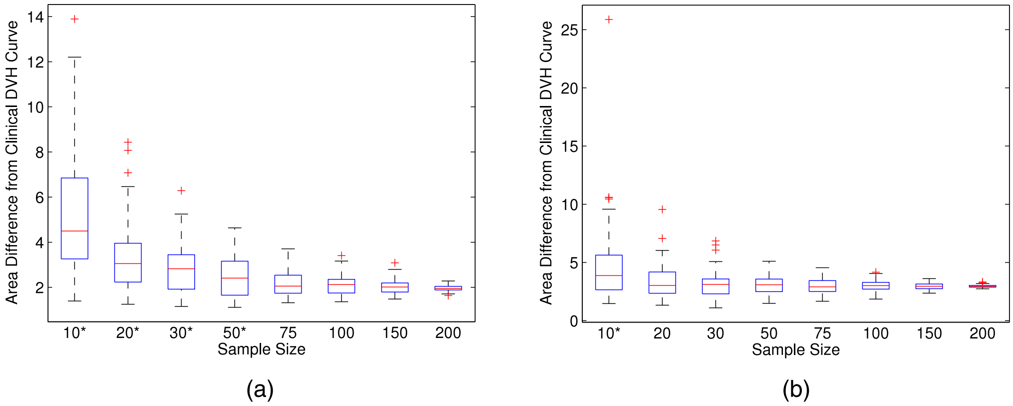


FIG. 3. The distribution of DVH curve error for the (a) bladder and (b) rectum.

For the bladder [i.e., Fig. 3(a)], the median error at sample sizes of 10, 20, 30, and 50 is significantly different than the median error at a sample size of 200. These results suggest that a sample size of at least 75 is required to obtain accurate and consistent bladder DVH predictions. For the rectum [i.e., Fig. 3(b)], only the median error at a sample size of 10 is significantly different when compared to 200. These results suggest that a sample size of at least 20 is required to obtain accurate and consistent rectum DVH predictions. Overall, for DVH curve prediction, a sample size greater than 75 should be used to train KBP models.

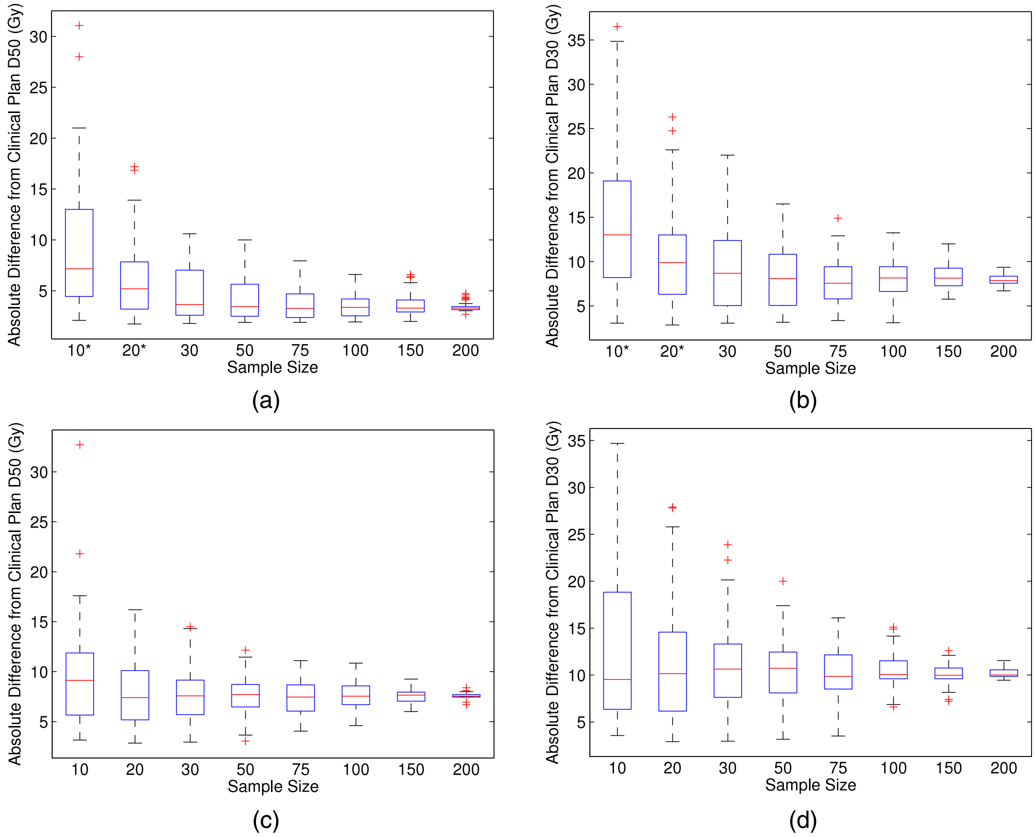
Next we investigate the performance of extracting DVH points from DVH curves. Figures 4(a)–4(d) display the extracted DVH point error distributions for bladder D_{50}^{B}, bladder D_{30}^{B}, rectum D_{50}^{R}, and rectum D_{30}^{R}, respectively. The methodology that explicitly predicts DVH points (Sec. 3.A) outperforms the idea of extracting DVH points after predicting DVH curves (compare Figs. 1 and 4).

To further illustrate the improvement in model accuracy as sample size increases, we plot DVH curves corresponding to four sample sizes for three patients. We chose one patient with poor prediction accuracy, one with average prediction accuracy, and one with high prediction accuracy. Each figure depicts the clinical DVH curve and predicted DVH curves corresponding to sample sizes of 10, 50, 100, and 200. Figures 5(a) and 5(b) depict poorly predicted DVH curves, Figs. 5(c) and 5(d) depict moderately predicted DVH curves, and Figs. 5(e) and 5(f) depict accurately predicted DVH curves. The realism and smoothness of the predicted DVH curve increase with sample size. DVH curves for sample sizes of 100 and 200 appear indistinguishable across all six figures.

3.C. Objective function weight prediction

The distribution of logistic regression weight error is shown in Figs. 6(a) and 6(b) for the bladder and rectum weight, respectively. For the bladder weight, only a sample size of 30 results in a median error that is significantly different than the median error corresponding to 200 samples. In contrast, none of the sample sizes yields a rectum weight median error that is significantly different from the median error of 200 samples. Overall, our results suggest that a sample size of at least 10 will produce consistent results for logistic regression.

Next, we examine the KNN results. The distribution of weight error for each of the weights is shown in Fig. 7 as a function of sample size. The model accuracy appears to increase for the bladder [Fig. 7(a)], rectum [Fig. 7(b)], and PTV ring [Fig. 7(e)], while remaining almost constant for the femoral heads [Figs. 7(c) and 7(d)]. For the bladder and rectum weights, all sample sizes less than 150 result in median error that is significantly different when compared to the median error of 200 samples. For the PTV ring, sample sizes of 10, 20, and 30 result in a significantly different median error and for the femoral heads, no sample sizes were determined to be significantly different. Overall, these results suggest that a



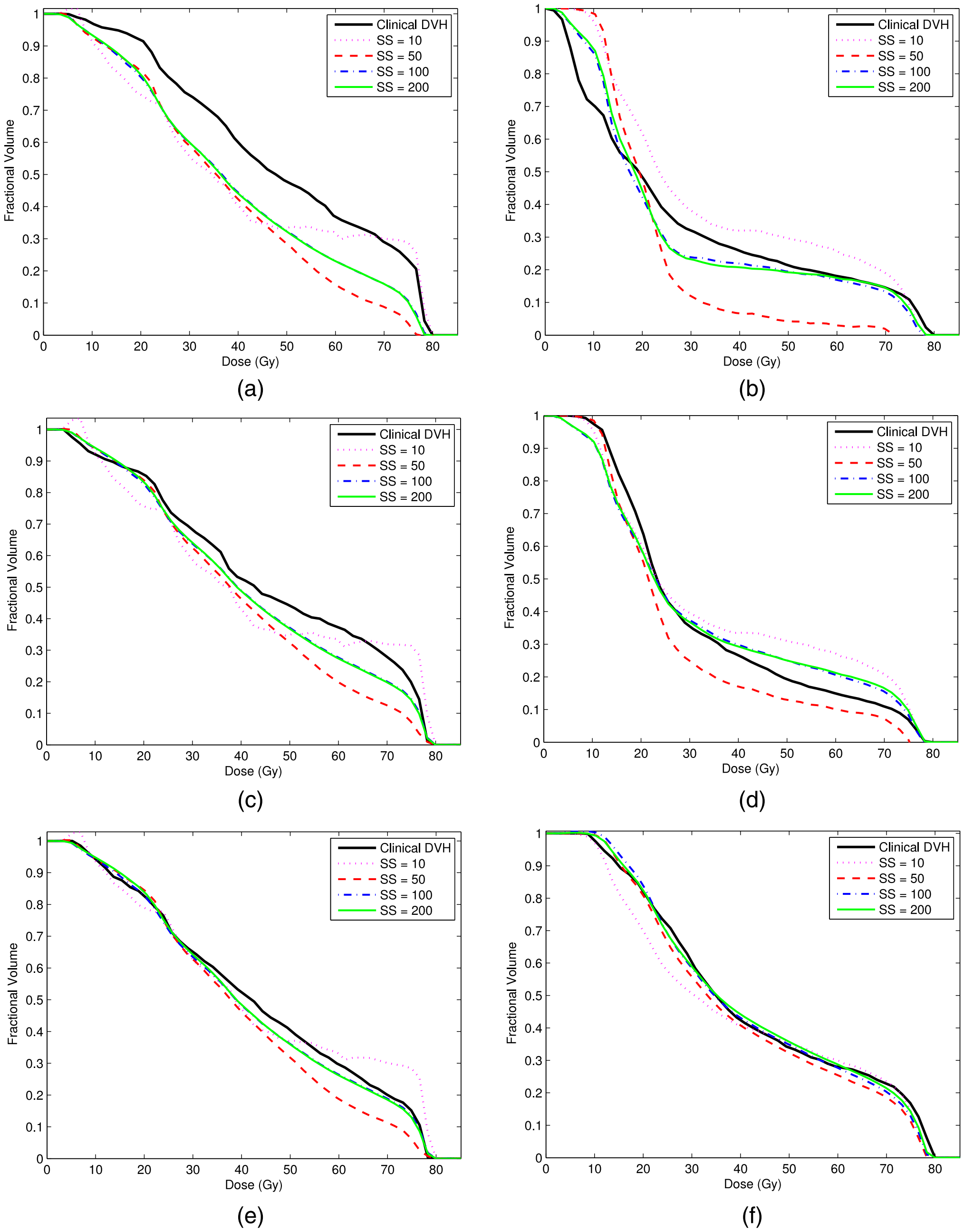
(a)

(b)

(c)

(d)

FIG. 4. The distribution of error for each DVH objective. (a) Bladder D_{50}^{B}, (b) bladder D_{30}^{B}, (c) rectum D_{50}^{R}, and (d) rectum D_{30}^{R}.



(a)

(b)

(c)

(d)

(e)

(f)

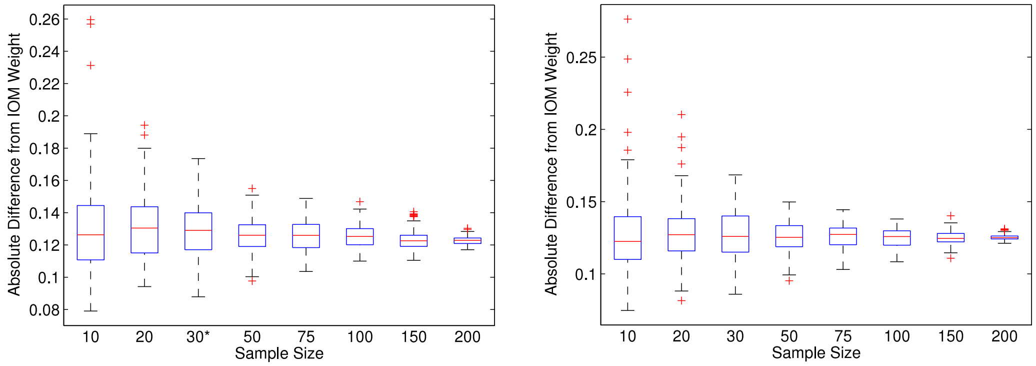
FIG. 5. The progression of DVH curve shape as sample size increases. (a) Poor accuracy bladder, (b) poor accuracy rectum, (c) moderate accuracy bladder, (d) moderate accuracy rectum, (e) high accuracy bladder, and (f) high accuracy rectum.

sample size greater than 150 is required to accurately train a KNN prediction model.

4. DISCUSSION

With the increasing clinical interest in KBP, it is critical to understand both the advantages and limitations of KBP models for clinical implementation. In particular, institutions interested in adopting such an approach should ensure they have access to a sufficiently large training database to produce predictions with the required accuracy. Failure to adequately train KBP models may result in poor and inconsistent performance, which may negatively impact future treatment plan quality and limit KBP uptake. The quality and quantity of training set samples can both impact model accuracy. In this paper, we did not consider how the quality of the training set plans impacts model predictions. Instead, we used previously delivered plans of high clinical quality, and we aimed to estimate the quantity needed to achieve consistently accurate model predictions.

The minimum required training set size was found to be different for each of the KBP models we studied in this paper. In particular, DVH point prediction required at least 200 samples, DVH curve prediction required at least 75 samples, LR weight prediction required at least 10 samples, and KNN weight prediction required at least 150 samples. These results imply that the minimum sample size depends on the specific



(a) (b)

FIG. 6. Logistic regression weight prediction error for the (a) bladder and (b) rectum.

model used and hence training set size experiments should be carried out before clinical implementation of a specific KBP approach. In addition to being model-dependent, the minimum required sample size for KBP may also vary depending on the treatment site. In particular, complex treatment sites with more variation between patients (e.g., head and neck) will likely require a larger sample size.

Although DVH curve prediction requires a smaller sample size compared to DVH point prediction, the method requiring fewer samples is not necessarily better. Our results show that, in terms of prediction accuracy, the methodology that explicitly predicts DVH points outperforms the idea of extracting DVH points after predicting DVH curves (compare Figs. 1 and 4). Furthermore, we believe that the DVH point method can be improved by using a less sensitive metric (i.e., fifth percentile instead of minimum).

We observed that logistic regression requires fewer patients compared to the other methods we investigated. We hypothesize that this is because the model is correctly chosen and is able to exploit information regarding the underlying distribution or shape of the data (i.e., S-shaped). In contrast, the DVH point prediction method and KNN weight prediction are nonparametric approaches. These models are not able to exploit information regarding the underlying distribution of the data. In practice, these results suggest that logistic regression and other parametric approaches, if well suited to the data, may require smaller training set sample sizes.

The required training set size needed to predict rectum metrics appears to be smaller than the required training set size for bladder metrics. For example, rectum DVH curve prediction required at least 20 samples, while bladder DVH curve prediction required at least 75 samples. Rectum sparing is generally more difficult and therefore receives more focus at Princess Margaret Cancer Center and other clinical institutions. Variations in bladder filling may also contribute to increased variability in bladder plan metrics. In combination, these factors may result in less relative variability and therefore, smaller required training set sizes, for rectum plan metrics as compared to bladder plan metrics.

In Sec. 3.A, we introduced the concepts of under- and overprediction for the KBP approach of Wu *et al.*, which are closely related to Type I and Type II error, respectively. For a given patient, overprediction occurs when a KBP model predicts a dosimetric value that is worse than what is actually achievable. This is problematic because it may cause the planner to stop before the best possible plan is achieved. In contrast, underprediction, defined as predicting a dosimetric value better than what is achievable, may motivate the planner to strive for the best possible plan. However, underprediction may come at the cost of time and effort on the part of the planner. Due to the predicted value being much lower than what is achievable, the treatment planner may spend extra time and effort trying to improve a treatment plan that is already Pareto optimal.

This study has some limitations. In particular, repeatedly sampling 200 patients with replacement from a total of 215 adds conservatism to our hypothesis testing. Due to the large overlap between samples of size 200, the KBP model predictions will be similar, resulting in artificially low variance. Low variance increases the power of our statistical test, leading to a higher chance of statistical significance. Since our null hypothesis is defined as equal error between two different sample sizes, a higher chance of statistical significance means that the minimum required sample size may actually be overestimated.

Although we only tested our models on prostate cancer patients, we believe our results provide a lower bound on the required number of samples for other tumor sites. Compared to other tumor sites such as head-and-neck, prostate cancer patients are considered relatively homogeneous, with minimal variation between the anatomy and resulting treatment plans for different patients. Moreover, due to being part of a clinical trial, our historical treatment plans are quite uniform and have the same beam angles, prescribed dose, and treatment protocols. As a result, one can expect that a small sample size is able to accurately represent the full spectrum of patients. In contrast, more complicated tumor sites may require a larger pool of patients to represent the complete patient spectrum.

Automated treatment planning is a closely related topic that leverages the results of KBP approaches to automate the treatment planning process. Predicted metrics from each of the approaches tested in this paper can be used as input parameters for an automated treatment planning framework. DVH points (i.e., dosimetric goals) and objective function weights can be

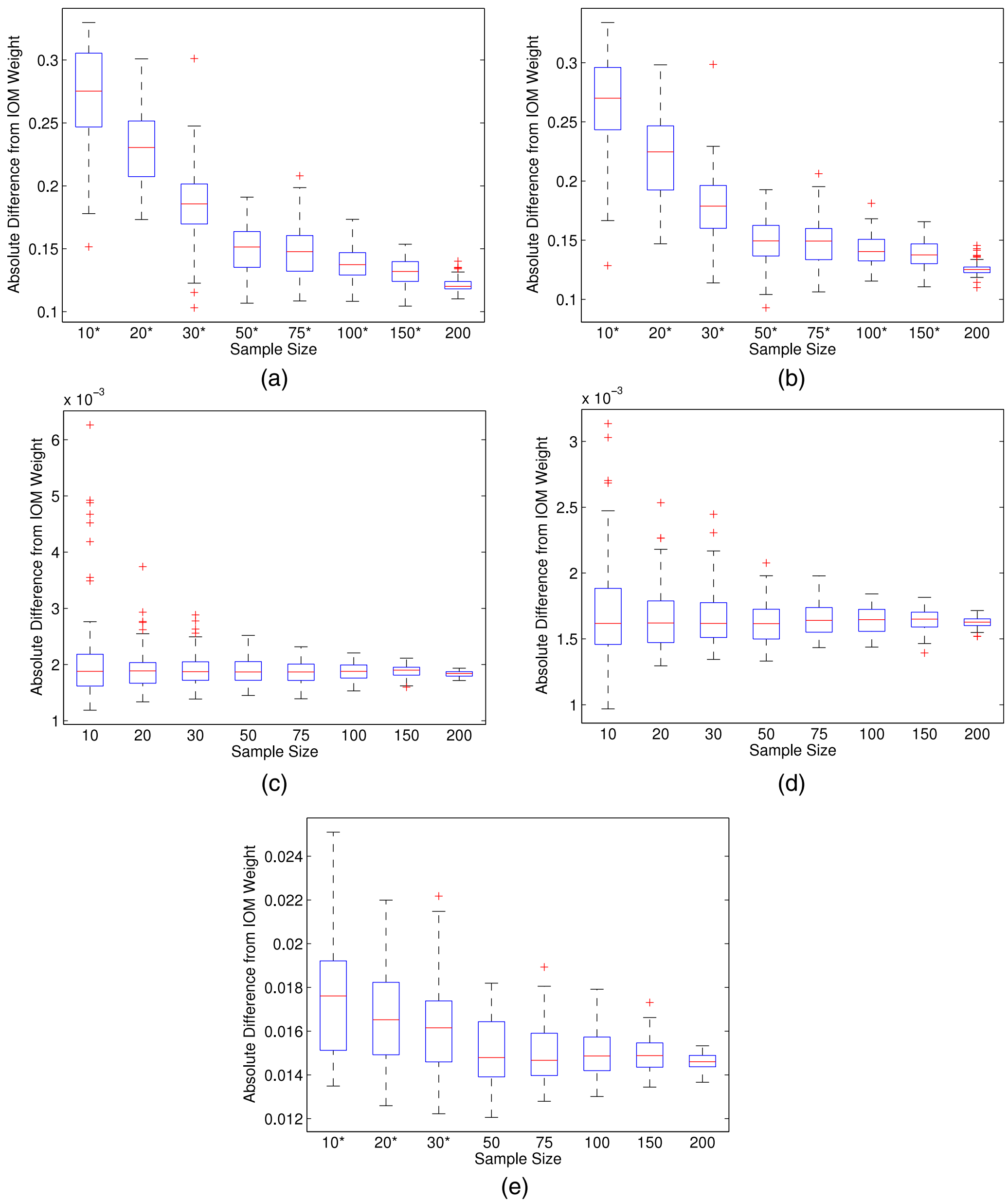


FIG. 7. KNN weight prediction error for the (a) bladder, (b) rectum, (c) left femur, (d) right femur, and (e) PTV ring.

used together or separately as optimization parameters from which an initial treatment plan is generated. DVH curves can also be used to provide input parameters, but are most often used in an alternative optimization approach that seeks to generate a treatment plan with similar DVH curves (e.g., as measured by squared error).

Automated treatment planning solutions may be developed in-house or purchased commercially. Recent work[23,24] has compared the quality of treatment plans generated via Varian's RapidPlan with clinical expert plans. The authors hypothesize that sample sizes of 25–30 plans may produce clinically acceptable treatment plans. However, there is currently no literature that rigorously examines how sample size impacts the resulting treatment plan and we encourage the community to explore this important topic in future research. It is possible that the automated planning portion may be able to mitigate the effects of inaccurate model predictions from KBP approaches, further reducing the required sample sizes determined in this work.

5. CONCLUSION

This paper provided the first scientific analysis of KBP model accuracy as a function of training set size. The minimum required sample size depends on the KBP model. In particular, DVH point prediction required at least 200 samples, DVH curve prediction required at least 75 samples, LR weight prediction required at least 10 samples, and KNN weight prediction required at least 150 samples. Our results suggest that a rule-of-thumb approach for determining the number of training set patients may not be appropriate and that the calculation likely depends on the KBP approach and the treatment site. Although our experiments used data from prostate cancer patients, we believe that our results may provide a lower bound for more complex tumor sites. As more cancer centers and treatment planning companies adopt KBP, future research is needed to determine how predicted endpoints, like those considered in this work, impact the resulting treatment plans after automated planning.

ACKNOWLEDGMENT

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

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

[1] 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 (15pp.) (2014).

[2] 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).

[3] 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).

[4] 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).

[5] P. W. Voet, M. L. Dirkx, S. Breedveld, D. Fransen, P. C. Levendag, and B. J. Heijmen, "Toward fully automated multicriteria plan generation: A prospective clinical study," Int. J. Radiat. Oncol., Biol., Phys. **85**, 866–872 (2013).

[6] P. W. Voet, M. L. Dirkx, S. Breedveld, A. Al-Mamgani, L. Incrocci, and B. J. Heijmen, "Fully automated volumetric modulated arc therapy plan generation for prostate cancer patients," Int. J. Radiat. Oncol., Biol., Phys. **88**, 1175–1179 (2014).

[7] 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).

[8] 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).

[9] 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).

[10] 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).

[11] 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).

[12] 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).

[13] 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).

[14] 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).

[15] 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 (10pp.) (2013).

[16] J. J. Boutilier, T. Lee, T. Craig, M. B. Sharpe, and T. C. Y. Chan, "Models for predicting objective function weights in prostate cancer IMRT," Med. Phys. **42**, 1586–1595 (2015).

[17] RapidPlan knowledge-based planning: Frequently asked questions, Varian Medical Systems, Palo Alto, CA, 2014, available at https://www.varian.com/sites/default/files/resource_attachments/RapidPlanFAQs.pdf, accessed on August 6, 2015.

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

[19] 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).

[20] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemom. Intell. Lab. Syst. **2**, 37–52 (1987).

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

[22] 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).

[23] J. P. Tol, A. R. Delaney, M. Dahele, B. J. Slotman, and W. F. A. R. Verbakel, "Evaluation of a knowledge-based planning solution for head and neck cancer," Int. J. Radiat. Oncol., Biol., Phys. **91**, 612–620 (2015).

[24] A. Fogliata, F. Belosi, A. Clivio, P. Navarria, G. Nicolini, M. Scorsetti, E. Vanetti, and L. Cozzi, "On the pre-clinical validation of a commercial model-based optimisation engine: Application to volumetric modulated arc therapy for patients with lung or prostate cancer," Radiother. Oncol. **113**, 385–391 (2014).