


# Performance Comparison of Knowledge-Based Dose Prediction Techniques Based on Limited Patient Data

Technology in Cancer Research & Treatment  
Volume 17: 1-10  
© The Author(s) 2018  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1533033818811150  
journals.sagepub.com/home/tct  


Angelia Landers, PhD<sup>1</sup> , Ryan Neph, BS<sup>1</sup>, Fabien Scalzo, PhD<sup>2</sup>, Dan Ruan, PhD<sup>1</sup>, and Ke Sheng, PhD<sup>1</sup>

## Abstract

**Purpose:** The accuracy of dose prediction is essential for knowledge-based planning and automated planning techniques. We compare the dose prediction accuracy of 3 prediction methods including statistical voxel dose learning, spectral regression, and support vector regression based on limited patient training data. **Methods:** Statistical voxel dose learning, spectral regression, and support vector regression were used to predict the dose of noncoplanar intensity-modulated radiation therapy ( $4\pi$ ) and volumetric-modulated arc therapy head and neck,  $4\pi$  lung, and volumetric-modulated arc therapy prostate plans. Twenty cases of each site were used for k-fold cross-validation, with  $k = 4$ . Statistical voxel dose learning bins voxels according to their Euclidean distance to the planning target volume and uses the median to predict the dose of new voxels. Distance to the planning target volume, polynomial combinations of the distance components, planning target volume, and organ at risk volume were used as features for spectral regression and support vector regression. A total of 28 features were included. Principal component analysis was performed on the input features to test the effect of dimension reduction. For the coplanar volumetric-modulated arc therapy plans, separate models were trained for voxels within the same axial slice as planning target volume voxels and voxels outside the primary beam. The effect of training separate models for each organ at risk compared to all voxels collectively was also tested. The mean squared error was calculated to evaluate the voxel dose prediction accuracy. **Results:** Statistical voxel dose learning using separate models for each organ at risk had the lowest root mean squared error for all sites and modalities: 3.91 Gy (head and neck  $4\pi$ ), 3.21 Gy (head and neck volumetric-modulated arc therapy), 2.49 Gy (lung  $4\pi$ ), and 2.35 Gy (prostate volumetric-modulated arc therapy). Compared to using the original features, principal component analysis reduced the  $4\pi$  prediction error for head and neck spectral regression (−43.9%) and support vector regression (−42.8%) and lung support vector regression (−24.4%) predictions. Principal component analysis was more effective in using all/most of the possible principal components. Separate organ at risk models were more accurate than training on all organ at risk voxels in all cases. **Conclusion:** Compared with more sophisticated parametric machine learning methods with dimension reduction, statistical voxel dose learning is more robust to patient variability and provides the most accurate dose prediction method.

## Keywords

radiotherapy, knowledge-based planning, dose prediction, machine learning, automated planning

## Abbreviations

ANN, artificial neural network; CPI, cricoid pharyngeal inlet; CT, computed tomography; DVH, dose–volume histogram; HN, head and neck; IMRT, intensity-modulated radiation therapy; KBP, knowledge-based planning; MAD, mean absolute difference;

<sup>1</sup> Department of Radiation Oncology, University of California, Los Angeles, Los Angeles, CA, USA

<sup>2</sup> Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA

## Corresponding Author:

Angelia Landers, PhD, Department of Radiation Oncology, University of California, Los Angeles, 200 Medical Plaza, Suite B265, Los Angeles, CA 90095, USA.  
Email: angelialanders1@gmail.com



MAE, mean absolute error; OAR, organ at risk; OVH, overlap volume histogram; PC, principal component; PCA, principal component analysis; PTV, planning target volume; RMSE, root mean squared error; SR, spectral regression; SVDL, statistical voxel dose learning; SVR, support vector regression; TMJ, temporomandibular joints; VMAT, volumetric-modulated arc therapy.

Received: August 5, 2018; Revised: September 20, 2018; Accepted: October 12, 2018.

## Introduction

Radiotherapy treatment planning involves considerable interplanner and intraplanner variability, which leads to sub-optimal plans, inconsistent planning results, and time inefficiencies.<sup>1-3</sup> Knowledge-based planning (KBP) and automated planning techniques are actively being developed to address these challenges and facilitate easier or faster means to attain optimal plans.<sup>4-6</sup>

To facilitate automated planning, KBP trains predictive models on a knowledge base to predict the dose of new patients. This study compares the prediction accuracy of 3 learning and dose prediction methods for representative intensity-modulated radiation therapy (IMRT) sites and modalities. Statistical voxel dose learning (SVDL) bins each organ at risk (OAR) voxel according to their respective Euclidean distance to the planning target volume (PTV). Statistical voxel dose learning dose prediction has previously been reported to be fast and accurate.<sup>7</sup> However, it is difficult to incorporate additional geometrical features into SVDL, which could be considered a nonparametric machine learning method. In this study, additional geometrical features are included in the 2 parametric machine learning methods: spectral regression (SR) and support vector regression (SVR).

A key component in KBP is learning the correlation between patient anatomies and planning dose. To achieve this goal, Wu *et al*<sup>8</sup> introduced the concept of the overlap volume histogram (OVH) and established its relationship with the dose-volume histogram (DVH). The OVH represents the relative spatial relationship between an OAR and the PTV. Points on the OVH are correlated with dose-volume points, which can perform predictions. This study evaluated the predicted dose-volume points for the PTV and OARs of 32 head and neck (HN) cases as quality control measures to aid in replanning. Support vector regression was used by Zhu *et al*<sup>9</sup> and later by Yuan *et al*.<sup>2</sup> Principal component analysis (PCA) was performed on the spatial and volumetric input features to identify the relevant principal components (PCs). The PCs were used to train a SVR model using the DVH as output. The training set size for these studies ranged from 18 to 82 patients, demonstrating SVR's capability of mitigating overfitting with relatively small training sets. Appenzoller *et al*<sup>10</sup> predicted the full DVH for prostate and HN cases by binning their distance to the target and performing a skew-normal fit on the voxel doses within each bin. The skew-normal fit was used to estimate the dose of new voxels according to their distance to the target, and

the predicted DVH was calculated over the skew-normal fit of all OAR voxels. We have previously reported a comparison among OVH, skew-normal fitting, and SVDL prediction accuracy.<sup>7</sup> Statistical voxel dose learning is similar to the skew-normal fitting method, but instead of performing a skew-normal fit of voxel doses with the same geometrical feature, the median dose value of each distance bin was used, which resulted in comparable dose prediction and shorter computational time.

In addition to these relatively simple learning methods, artificial neural networks (ANNs) were used to predict brain stereotactic radiosurgery (SRS) and prostate dose distribution for voxels within 30 to 32 mm of the PTV.<sup>4</sup> Like SVR, ANNs are flexible in the number of input features, in this case using spatial and volumetric features as well as the number of fields and the voxel angle from the PTV centroid and principal coordinate system defined by PCA. Small training set sizes of 23 and 43 for prostate and SRS cases, respectively, were used for acceptable dose prediction accuracy (8%-10% average) for the small volume near the target. However, the computational time becomes intractable for larger regions of interest and more complex cases using the method.

A more sophisticated atlas-based dose prediction method was reported for HN patients.<sup>5</sup> Each patient in the training set represents 1 atlas, and computed tomography (CT) radiomics texture features were extracted to characterize each image. Feature extraction and characterization was then performed on CTs of the patients to implement atlas selection. This resulted in probabilistic dose estimates which were then used with a conditional random field to find the most likely voxel dose from similar atlases and evaluate the prediction accuracy. This method required relatively large training sets, with training set sizes ranging from 58 to 144 patients for the various treatment sites. It took 48 hours to train the model and additional 15 minutes for individual case prediction.

A summary of these dose prediction studies is presented in Table 1. In clinical practice, KBP learning and prediction needs to balance the efficiency, sample size requirement, and accuracy. Although increasing the number of plans could increase the training accuracy theoretically, a large high-quality training set is not always attainable. The plan quality heterogeneity often increases with more plans included. Due to the need to repeatedly retrain the model with new or updated cases, the computational speed is also important. In this study, we focus on comparing learning methods that are fast, yet rely on relatively small data sets that are readily available at most clinics. A direct

**Table 1.** Summary of Previous Dose Prediction Studies.

	Input Features	Predicted Output	Learning Method	Training Set Size	Disease Sites	Training Time	Prediction Time
Wu <i>et al</i>	Overlap volume histogram	Dose–volume points	OVH-DVH correlation	32	HN	-	-
Zhu <i>et al</i>	Distance to target	DVH	SVR	18	Prostate	-	-
Yuan <i>et al</i>	Distance to target Structure volumes Overlap volumes Out-of-field volume	DVH	SVR w/PCA	64, 82	Prostate, HN	-	-
Appenzoller <i>et al</i>	Distance to target	DVH	Skew-normal fitting	20, 24	Prostate, HN	-	-
Shiraishi <i>et al</i>	PTV volume Distance from structures  Number of fields Angles from PTV centroid Principal coordinate system	Voxel dose Within 30-32 mm of PTV	ANN	23, 43	Prostate, SRS	4 hours	-
Tran <i>et al</i>	Distance to target	OAR voxel dose	SVDL	20	Liver	2-4 minutes	1 second
McIntosh <i>et al</i>	Radiomics features	Image voxel dose	Atlas-based CRF	97 144 113 144 77 58	Breast cavity Whole breast CNS brain Prostate Lung Rectum	48 hours	15 minutes

Abbreviations: ANN, artificial neural network; CNS, central nervous system; CRF, conditional random field; DVH, dose–volume histogram; HN, head and neck; OAR, organ at risk; OVH, overlap volume histogram; PCA, principal component analysis; PTV, planning target volume; SVDL, statistical voxel dose learning; SVR, support vector regression.

**Table 2.** Patient Cohort Disease Sites and Planning Techniques.

	Head and Neck	Lung	Prostate
4 $\pi$	X	X	
VMAT	X		X

Abbreviation: VMAT, volumetric-modulated arc therapy.

comparison of the dose prediction methods will provide guidance to automated treatment planning.

## Materials and Methods

Dose prediction for both noncoplanar IMRT (4 $\pi$ ) and volumetric-modulated arc therapy (VMAT) was performed for HN, lung, and prostate cases. It is the current standard-of-care modality in IMRT. The 4 $\pi$  radiotherapy was developed to substantially improve dose conformity and normal tissue sparing by optimizing noncoplanar beam angle selection. This improved dosimetry has been reported for liver,<sup>11,12</sup> lung,<sup>13</sup> brain,<sup>14</sup> prostate,<sup>15,16</sup> and HN<sup>17</sup> cases. Both modalities can be automated using a voxel-based optimization without a need to manually select the beams, making them well suited for KBP automated treatment planning.

The patient cohorts and their respective planning techniques for this study were chosen as listed in Table 2. Twenty patients were selected for each treatment site. Predictions were performed using k-fold cross-validation with k = 4, resulting in

training set sizes of 15 patients. These 3 disease sites were chosen as we believe them to be generally representative of IMRT planning cases. Patients' CTs and VMAT plans were retrospectively obtained from the clinic under an institutional review board–approved protocol (IRB# 12-001882). Head and neck and lung cases were replanned using 4 $\pi$  optimization in MATLAB (version 2017a; Mathworks, Natick, Massachusetts).<sup>11</sup>

Although noncoplanar beams have been shown to improve the dosimetry,<sup>11,15,16,13</sup> the delivery of such beams can be challenging due to potential patient collision. Furthermore, prostate dose constraints are typically met with coplanar arc beams. For these reasons, we only included coplanar VMAT plans for prostate cases. On the other hand, dose compactness has a much higher importance in lung stereotactic body radiation therapy (SBRT) with dose prescriptions of 12.5 Gy in 4 fractions.<sup>18,19</sup> Noncoplanar 4 $\pi$  treatment is particularly effective in reducing the high dose spillage and improving dose conformity.<sup>13</sup> The current study focuses on centrally located lung tumors. Lastly, HN cases represent the most challenging disease site, commonly including up to 40 OARs with complex PTV shapes, location, and various prescription dose levels. Head and neck cases are highly diverse based on the origin of the tumors and the laterality. In this study, without losing generality, we focused on oropharyngeal tumors, which are representative of the complex HN anatomy. Both 4 $\pi$  and VMAT plans were included for HN cases as they could both be valuable depending on actual dosimetric requirement and delivery time.

For each site, there are numerous OARs to predict. All relevant OARs with clinical dose constraints were predicted. We tested predictions from models trained independently on each OAR as well as those trained on all OAR voxels collectively. Separate models allow more specialized predictions, but combining all voxels ensures large training data. For the VMAT predictions, voxels were separated to in-beam and out-of-beam groups based on their axial slice location relative to the PTV voxels.

### Dose Prediction Techniques

We used the median-approximated SVDL method as described by Tran *et al.*<sup>7</sup> Computed tomography and dose arrays were resampled into isotropic  $0.25 \times 0.25 \times 0.25$  cm<sup>3</sup> voxels. The only voxel information used in this SVDL is the Euclidean distance to the PTV surface. Voxels are sorted into bins based on their distance to the PTV and the median dose of each distance bin calculated. The median dose is then used as the predicted dose of each new voxel of the same distance to the PTV. Statistical voxel dose learning can be viewed as a weighted nearest neighbor regression using the median statistic. As such, SVDL is robust to noise on both the input feature vector and mapped dose value.

Spectral regression and SVR are supervised machine learning models that can predict an output value from input features. In this case, the features included the distance to the PTV ( $r = \sqrt{x^2 + y^2 + z^2}$ ),  $r^2$ ,  $r^3$ ,  $r^{-1}$ ,  $r^{-2}$ ,  $y$ ,  $z$ ,  $x^2$ ,  $y^2$ ,  $z^2$ ,  $x^{-1}y^{-1}z^{-1}$ ,  $xy$ ,  $yz$ ,  $xz$ ,  $\sqrt{x^2 + y^2}$ ,  $\sqrt{y^2 + z^2}$ ,  $\sqrt{x^2 + z^2}$ , angle from PTV centroid, angle from linac source, PTV volume, PTV volume,<sup>2</sup> OAR volume, OAR volume,<sup>2</sup>  $\frac{\text{OAR volume}}{\text{PTV volume}}$ , and  $\frac{\text{PTV volume}}{\text{OAR volume}}$ , resulting in a total of 28 features. Spectral regression and SVR predictions were performed using the aforementioned features as well as using PCA. Principal component analysis transforms the feature vectors into orthogonal PCs that are linearly uncorrelated and variance preserving. Principal component analysis allows control of the degree of dimension reduction by simply choosing how many PCs to train on. Dose predictions with PCA were performed with 1 to 28 PCs to find the ideal number of PCs for each case and compared with predictions using the original features. Pearson correlation coefficients ( $R$ ) were evaluated between all pairs of PCs and original features.

Spectral regression is a modified form of ridge regression realized through a subspace learning formulation.<sup>20</sup> The SR algorithm discovers the subspace embedding vectors as the eigenvectors of an eigenproblem involving the similarity matrix from spectral graph theory,<sup>21</sup> with flexible regularization options such as L2 (ridge) regularization. The SR for radiotherapy dose prediction has not been studied before, and it was performed using the open source MATLAB code from Cai *et al.*<sup>20,22,23</sup> The prediction model is based on subspace learning using the input features and corresponding voxel dose labels of the training set. This prediction model produces an embedding vector of the same length as the feature vectors; hence, prediction for new voxels is a simple dot product.

Support vector regression is based on support vector machines, which is a large margin classifier. Support vector machines are used for the classification of new data into separate categories. It trains a model by defining a decision boundary that maximizes the margin separating different categories in labeled training data. Support vector machines use kernel-induced features to specify nonlinear decision boundaries, allowing considerable flexibility in prediction models. Instead of defining decision boundaries to separate categories, SVR trains a regression function that minimizes the error of data points outside of the margin. Support vector regression in combination with PCA has previously been demonstrated to predict specific dose-volume points.<sup>2</sup> It was performed using LIBSVM, an open source software library for support vector optimization.<sup>24</sup> The  $\nu$ -SVR model was chosen after preliminary comparison with  $\epsilon$ -SVR due to more consistent and accurate predictions. Large OARs were downsampled to a maximum of 20 000 voxels due to the substantial time requirements for SVR.

Both SR and SVR require tuning of hyperparameters to ensure the models are properly trained for our specific task. Spectral regression uses the Tikhonov regularization parameter,  $\alpha$ , and SVR requires selection of a kernel and tuning of the penalty parameter,  $C$ , and kernel parameters. Linear, polynomial, radial basis function, and sigmoid kernels were evaluated for the SVR regressions. Parameters were tuned with exhaustive grid searches in exponential increments (eg,  $\alpha = 0.1, 0.3, 1, 3, 10, \dots$ ).

After the optimal parameters for SR and SVR were found, overfitting analysis was performed by predicting one of the 5 patient test cohorts with training set sizes varying from 1 to 15 patients. If the prediction error does not converge, it would indicate that larger training sizes would improve the accuracy and that overfitting is likely a problem with our training set size.

### Prediction Accuracy Analysis

All dose prediction methods in this study are capable of predicting the 3D voxel dose for each OAR. Both root mean squared error (RMSE) and mean absolute error (MAE) were used to evaluate the voxel dose prediction accuracy,

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{d}_i - d_i)^2},$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{d}_i - d_i|,$$

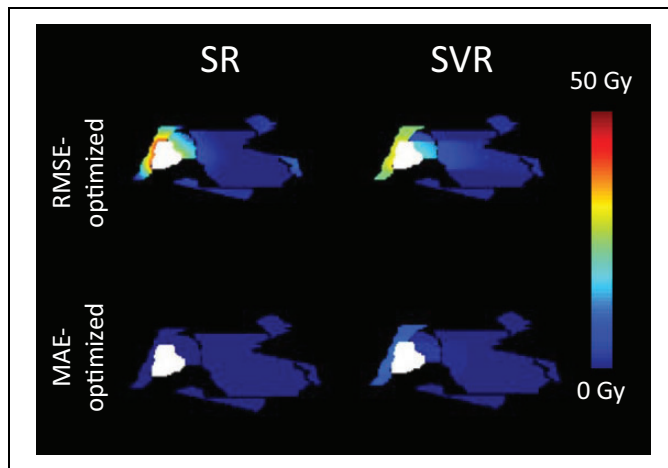
where  $n$  is the number of voxels and  $\hat{d}_i$  and  $d_i$  are the predicted and actual doses, respectively. Both prediction accuracy measures keep the original units (Gy) that can be intuitive to interpret. The errors were calculated for each predicted OAR as well as for all OAR voxels as a whole.

The DVH prediction accuracy was also evaluated, as it still contains relevant clinical information. The mean absolute

**Table 3.** Root Mean Squared Error and Mean Absolute Error Results for RMSE-Optimized and MAE-Optimized Lung  $4\pi$  Spectral Regression and Support Vector Regression Dose Predictions.

(a)	Alpha	RMSE	MAE	
RMSE-optimized	$10^2$	5.497599	3.682	
MAE-optimized	$10^9$	5.812917	1.78	
(b)	C	Gamma	RMSE	MAE
RMSE-optimized	1	0.5	4.0196	1.706
MAE-optimized	0.01	0.5	4.957	1.568

Abbreviations: RMSE, root mean squared error; MAE, mean absolute error.



**Figure 1.** Representative isodose color wash comparison between root mean squared error (RMSE)- and mean absolute error (MAE)-optimized spectral regression and support vector regression dose predictions for an example lung  $4\pi$  case. The planning target volumes (PTVs) are shown in white.

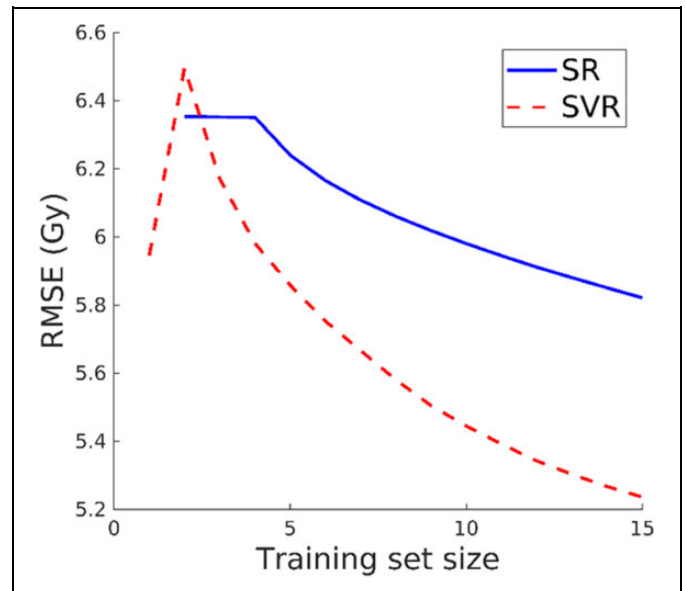
difference (MAD) of the actual and predicted DVHs was calculated for each prediction,

$$MAD = \frac{1}{m} \sum_{j=1}^m |DVH_j^{\text{actual}} - DVH_j^{\text{predicted}}|,$$

where  $m$  is the number of DVH bins, separated by 0.05 Gy. Since the DVH is conveyed as a fractional volume, MAD is reported in units of percentage. Mean absolute difference is compared between the DVH predictions of SVDL and the SR and SVR models with the lowest voxel dose error.

## Results

Exhaustive grid search resulted in different optimal SR and SVR hyperparameters depending on the accuracy measure, RMSE or MAE. To illustrate the problem, the lung  $4\pi$  RMSE and MAE results are shown in Table 3. The optimal hyperparameters would ideally match, whether we optimized the regression parameters based on RMSE or MAE. However, since they resulted in differing parameters, we visually compared the results to determine whether they are representative of a



**Figure 2.** Dose prediction root mean squared error (RMSE) using spectral regression and support vector regression with varying training set size.

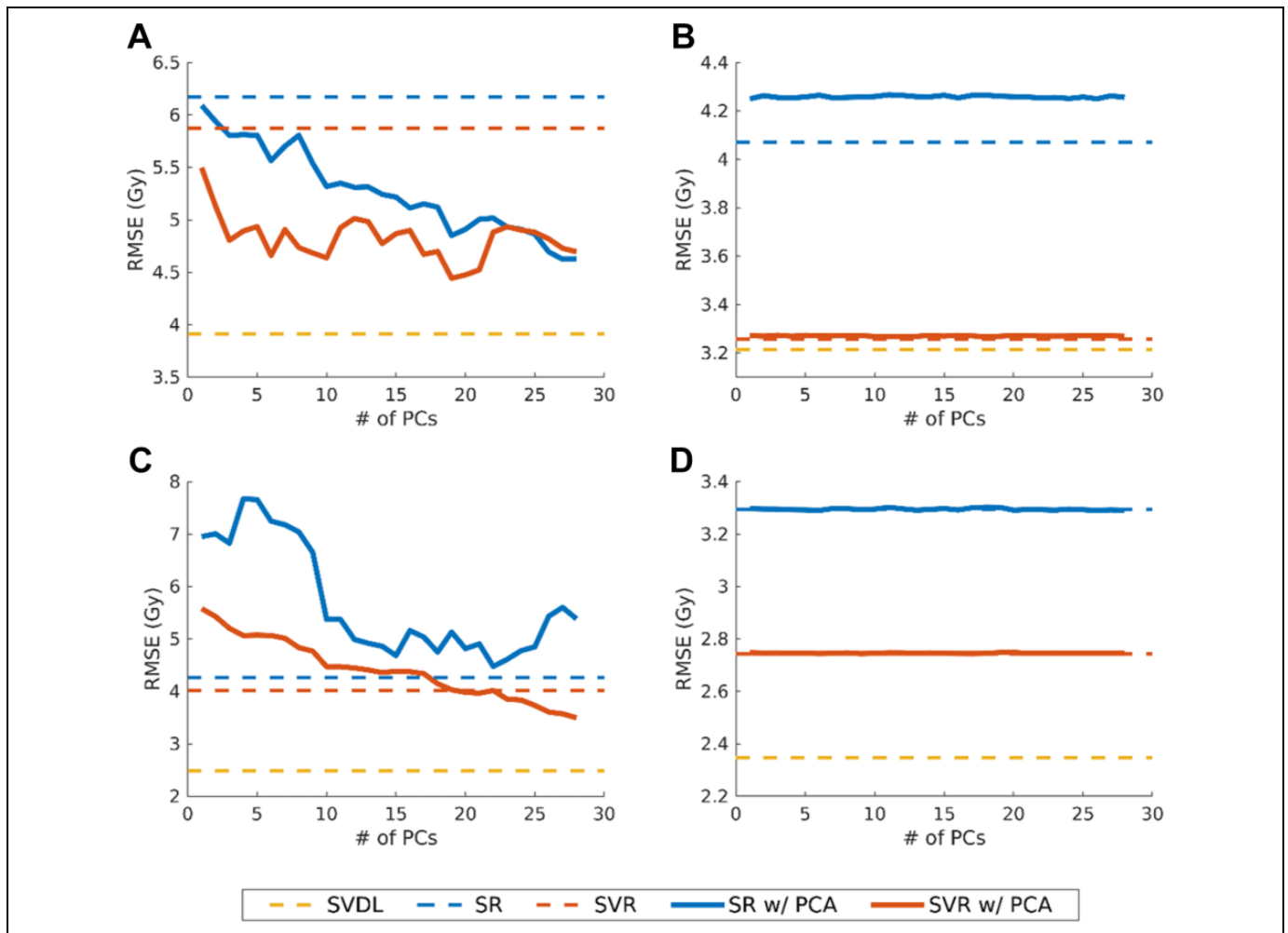
realistic dose prediction. Isodose color wash comparisons are shown in Figure 1. Root mean squared error-optimized dose predictions better approximate the sharp dose gradient of the PTV surface; hence, RMSE was used for the remainder of the study for evaluations of prediction accuracy. In fact, the prediction optimized for MAE resulted in piecewise constant dose predictions, where most of the voxel doses were equal to zero.

A representative example of the overfitting analysis is shown in Figure 2. All other treatment sites exhibited the same trend. Both SR and SVR prediction accuracy did not converge using the training set sizes we had available. The RMSE was still decreasing using our largest training set of 15 patients.

The voxel dose prediction RMSEs are shown in Figure 3. Between the parametric machine learning methods, SVR consistently achieved lower prediction errors compared to SR, both with and without PCA. Nevertheless, both parametric machine learning methods are inferior to SVDL in RMSE. The SVDL prediction RMSE for HN is greater (3.91 Gy  $4\pi$  and 3.21 Gy VMAT) than for the prostate (2.35 Gy) and the lung (2.49 Gy). The actual and predicted OAR voxel dose of an example case is illustrated in Figure 4 as isodose color washes.

Combining all OAR voxels into the same prediction model to enlarge the training set size did not offset the advantage of specialized models for each OAR. For all cases, collectively training on all OAR voxels resulted in increased prediction error of 11.7% (SVDL), 6.7% (SR), and 19.0% (SVR).

Principal component analysis reduced the SR and SVR error for all  $4\pi$  cases, except for lung  $4\pi$  with SR, in which case using PCA yielded worse accuracy than SR without PCA ( $-4.74\%$ ). Dimension reduction with PCA had no effect in improving VMAT dose prediction. For HN VMAT prediction, using PCA with SR also resulted in worse performance than SR using the



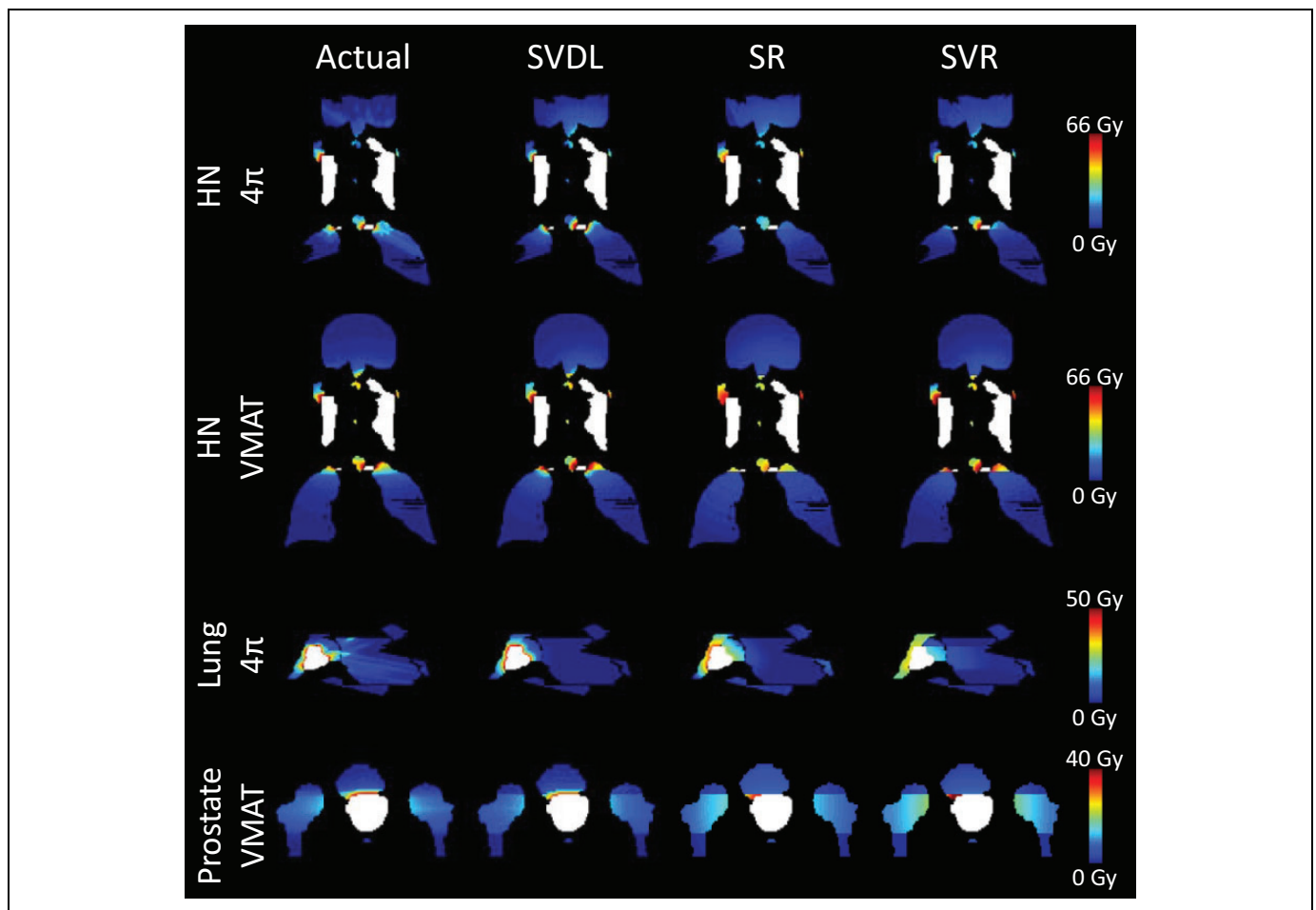
**Figure 3.** Root mean squared error of dose predictions for (A) head and neck  $4\pi$ , (B) head and neck volumetric-modulated arc therapy (VMAT), (C) lung  $4\pi$ , and (D) prostate VMAT. All results shown are for predictions using separate models for each organ at risk (OAR). Errors for spectral regression (SR) and support vector regression (SVR) using principal component analysis are shown with varying numbers of principal components (PCs). Root mean squared error for statistical voxel dose learning (SVDL), which only uses the voxel distance to the planning target volume (PTV), and SR and SVR using the original geometrical features are shown as dashed lines.

original features ( $-4.20\%$ ). With the inclusion of more PCs in  $4\pi$  prediction, RMSE outcome was noisy but typically reduced the error. As shown in Figure 3, SVR and SR for HN  $4\pi$  and SVR for lung  $4\pi$  had general prediction improvement with more PCs. In the case of lung  $4\pi$ , the inclusion of more PCs exhibited a valley trend, seen with 10 to 27 PCs in Figure 3C as the solid blue line. In the analysis of correlations between the PCs and original features, Pearson correlation coefficients ( $R$ ) greater than 0.7 were observed between PC1 and  $r^2$  ( $R = 0.96$ ), PC4 and the PTV volume ( $R = 0.78$ ), and PC7 and  $\frac{PTV \text{ volume}}{OAR \text{ volume}}$  ( $R = 0.90$ ). All other PCs had no strong correlation with the original features. From observation of the resulting coefficients, 0.7 was chosen as the cutoff for strong correlation as most other PC-feature combinations had coefficients less than 0.4.

Figure 5 shows the RMSE for dose prediction of each OAR for the most accurate method (SVDL in all cases). The box plots illustrate the range of error for each of the 20 patients' OAR prediction. For HN  $4\pi$  predictions, there are consistently

large errors for the cricoid pharyngeal inlet (CPI) and temporomandibular joints (TMJs). These errors are also present in HN VMAT predictions, but not to the same scale. The HN VMAT predictions have more intermediately sized errors for other OARs, namely the parotids, mandible, lips, cochlea, larynx, pharynx, esophagus, oral cavity, and submandibular glands. Lung  $4\pi$  predictions were mostly consistent through each OAR. All OARs had a few outliers except for the lung, esophagus, and liver. The prostate VMAT predictions are overall scaled lower than the other sites and even the outliers have considerably lower error.

The prediction error for DVHs was also evaluated using MAD between the actual DVH and the 3 predicted DVHs. Box plots of the MAD results for the 20 patients of each treatment site are shown in Figure 6. A representative DVH with the predicted DVHs is shown in Figure 7. Mean absolute difference in the DVH fractional volumes across the tabulated dose bins is reported in units of percentage. However, it is important to note



**Figure 4.** Isodose images of the actual and predicted organ at risk (OAR) voxel dose for representative example head and neck (HN), lung, and prostate cases. The planning target volumes (PTVs) are shown in white.

that MAD is not a percent difference metric. Overall, SVDL produced an average MAD of 4.33, compared to 7.37 and 5.65 for SR and SVR, respectively. For reference, the atlas-based dose prediction method introduced by McIntosh and Purdie<sup>5</sup> resulted in an average MAD of 1.33 and 2.12 for lung and prostate, respectively. However, it is important to note the differences both in training set size and computational time with their method, which used 77 lung and 144 prostate plans and took 48 hours for training.

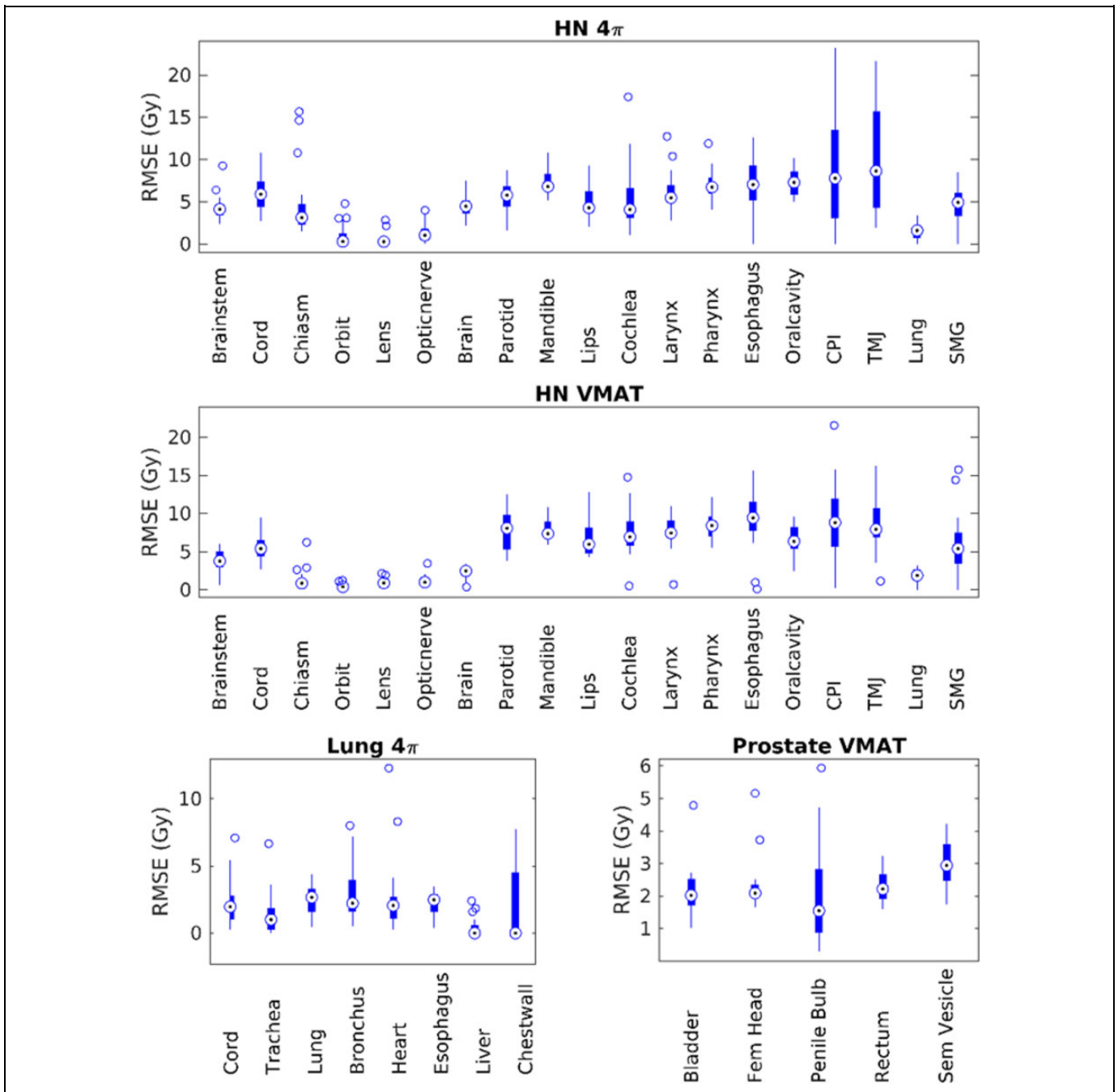
All prediction methods are relatively fast compared to the typical planning time for voxel-based  $4\pi$  and VMAT optimization. Average training times were 24.0 seconds (SVDL), 0.06 seconds (SR), and 87.7 seconds (SVR) for each OAR training set. Average prediction times were 0.03 seconds (SVDL), 0.01 seconds (SR), and 0.89 seconds (SVR) for each new patient OAR.

## Discussion

Counterintuitively, despite more geometrical features being included, the parametric machine learning methods were less accurate than the simple SVDL for both voxel dose and DVH predictions. Feature reduction using PCA was not able to

substantially improve the results. Using some or all of the PCs from PCA improved the SR or SVR prediction accuracy over using the original features, but only for 3 of the four  $4\pi$  predictions. This is likely because the VMAT predictions already separated in-beam and out-of-beam voxels, making the Euclidean distance to the PTV the primary relevant feature for VMAT cases. Conversely,  $4\pi$  dose distributions are much more dependent on the relative 3D spatial differences than just the axial in-beam and out-of-beam categories. By organizing the features into high variance, orthogonal PCs, the parametric machine learning methods can better utilize the data to improve the prediction accuracy. Strong correlations were found between PCs and  $r^2$ , PTV volume, and  $\frac{\text{PTV size}}{\text{OAR size}}$ . This indicates that these particular features are mostly orthogonal with each other and an increase in model complexity could be beneficial. In general, the increase in error with reduction in PCs suggests that dimension reduction unnecessarily excludes uncorrelated information. The valley trend for PCs in lung  $4\pi$  SR suggests that for this particular case, the full set of features is unnecessary to predict the dose.

Additionally, prediction accuracy was not improved by training a model using all OAR voxels, which resulted in even



**Figure 5.** Root mean squared error of statistical voxel dose learning (SVDL) dose predictions for each organ at risk (OAR).

higher error. A possible reason is that SR and SVR are more dependent on their respective training sets, with variable dose gradients for the different OAR predictions. A training set size analysis was performed to understand the performance of SR and SVR. The results indicate that overfitting exists for the current training set size. Larger training sets are likely required to fully take advantage of the machine learning methods. However, when the training set is practically limited, the study demonstrates that SVDL can better predict the 3D dose distribution and DVH. Statistical voxel dose learning collapses the

available data into one estimated number for each distance bin. This makes it more naive to the variations of PTV and OAR geometry, resulting in site-dependent prediction error. This is evident in the isodose images in Figure 4. Realistic voxel dose predictions require sharp dose falloffs near the PTV boundary but otherwise smooth transitions. Statistical voxel dose learning consistently maintains this sharp dose gradient by independently estimating each distance bin. In contrast, the SR and SVR methods are unable to achieve these gradient qualities, instead having abrupt dose gradients between OARs and at the



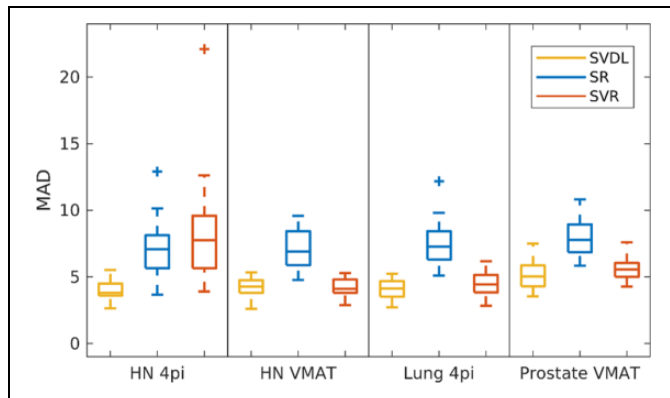
primary beam plane. Statistical voxel dose learning is able to consistently produce such dose distributions because the same predicted dose is applied to all voxels of the same distance to the PTV.

An interesting observation for the machine learning methods is the apparent discrepancy in the RMSE and MAE trends for the same regularization parameter tuning. The discrepancy between MAE- and RMSE-optimized dose predictions is due to their different preferences to small and large errors. Mean absolute error allows larger outliers as long as most voxel residuals are minimal. However, RMSE penalizes large residuals more than small errors. In the case where the MAE is minimized by heavily regularizing the predicted dose distribution, a nearly piecewise constant dose distribution resulted that is drastically different from realistic dose distributions because MAE is dominated by the majority of voxels with near zero dose and forgiving toward the few voxels with substantial dose. Therefore, the RMSE metric is better correlated with realistic dose distributions and was chosen as the sole accuracy measure for the rest of the study.

Although only a subset of HN cases were included, that is, oropharyngeal, these cases are still highly variable in relative

organ geometry, dose prescription levels, and PTV. This high variance contributed to the overall higher HN prediction error compared to lung and prostate cases. Although the lung tumor locations vary, fewer and larger OARs are involved, resulting in more consistent prediction than HN. Prostate dose prediction RMSE had the least amount of variability as would be expected due to the more consistent geometry in prostate cases.

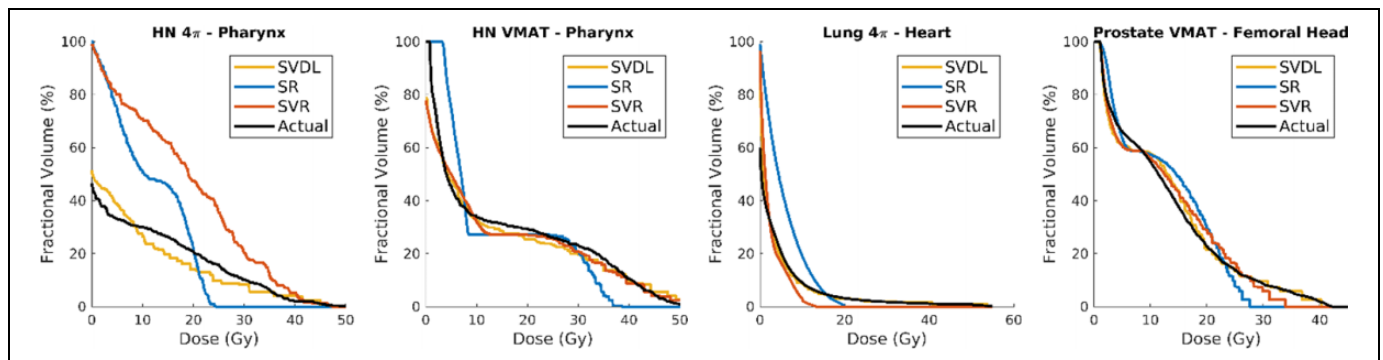
The comparison of the OAR prediction accuracy between HN  $4\pi$  and VMAT highlights interesting differences between the 2 modalities. The VMAT prediction inherently accounts for beam orientation, as in-beam and out-of-beam voxels are separate. However, since  $4\pi$  performs simultaneous beam angle selection with fluence map optimization, it is impossible to preemptively sort in-beam and out-of-beam voxels for  $4\pi$ . Despite this difference,  $4\pi$  prediction had relatively low errors for most OARs, except for the CPI and TMJs, which are small organs near the PTV and can be particularly affected by primary beam paths. In contrast, for HN VMAT prediction, organs at the same level of the PTV had higher RMSEs than OARs out of the VMAT arc plane. This is most likely because they are close to the PTV and in the high-dose gradient region. These organs are very small ( $<3\text{ cm}^3$ ) and are often near the PTV, leading to a greater degree of dosimetric variability and thus inferior dose prediction accuracy.



**Figure 6.** Mean absolute difference in fractional volume for the dose-volume histogram (DVH) predictions of 20 patients of each site for (in order from left to right) statistical voxel dose learning (SVDL), spectral regression (SR), and support vector regression (SVR).

## Conclusion

This study compared 3 learning methods using limited training set sizes for the dose prediction of  $4\pi$  noncoplanar IMRT and coplanar VMAT plans for HN, lung, and prostate patients. Statistical voxel dose learning not only was found to be a simpler approach compared to SR and SVR but also produced the lowest prediction error for all cases. Principal component analysis was useful in improving SR and SVR dose prediction accuracy for  $4\pi$  prediction in most cases, but still could not reach the accuracy of SVDL. Training set size analysis found that the parametric machine learning predictions could be improved with larger training sets due to overfitting. Among the 3 sites, prostate plans had the lowest prediction error, with HN producing the highest error. Using separate OAR-specific prediction models was more beneficial than trying to increase



**Figure 7.** Comparison of the predicted and actual dose for representative example dose-volume histogram (DVHs). The structures shown in each of these examples are the pharynx (head and neck), heart (lung), and femoral head (prostate).

data set size by including all OAR voxels in the training set. All training methods utilize clinically feasible time, but SVR is considerably slower than the other 2 methods.


### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported in part by DOE DE-SC0017687, DE-SC0017057 and NIH R44CA183390, R43CA183390 and R01CA188300.

### ORCID iD

Angelia Landers, PhD  <https://orcid.org/0000-0002-3057-9629>

### References

- Moore KL, Brame RS, Low DA, Mutic S. Experience-based quality control of clinical intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys.* 2011;81(2):545-551.
- Yuan L, Ge Y, Lee WR, Yin FF, Kirkpatrick JP, Wu QJ. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. *Med Phys.* 2012; 39(11):6868-6878.
- Nelms BE, Robinson G, Markham J, et al. Variation in external beam treatment plan quality: an inter-institutional study of planners and planning systems. *Pract Radiat Oncol.* 2012;2(4):296-305.
- Shiraishi S, Moore KL. Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy. *Med Phys.* 2016;43(1):378.
- McIntosh C, Purdie TG. Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Phys Med Biol.* 2017;62(2):415-431.
- Ziemer BP, Shiraishi S, Hattangadi-Gluth JA, Sanghvi P, Moore KL. Fully automated, comprehensive knowledge-based planning for stereotactic radiosurgery: preclinical validation through blinded physician review. *Pract Radiat Oncol.* 2017;7(6): e569-e578.
- Tran A, Woods K, Nguyen D, et al. Predicting liver SBRT eligibility and plan quality for VMAT and  $4\pi$  plans. *Radiat Oncol.* 2017;12(1):70.
- Wu B, Ricchetti F, Sanguineti G, et al. Patient geometry-driven information retrieval for IMRT treatment plan quality control. *Med Phys.* 2009;36(12):5497-5505.
- Zhu X, Ge Y, Li T, Thongphiew D, Yin FF, Wu QJ. A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Med Phys.* 2011;38(2):719-726.
- Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore KL. Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Med Phys.* 2012;39(12):7446-7461.
- Dong P, Lee P, Ruan D, et al.  $4\pi$  non-coplanar liver SBRT: a novel delivery technique. *Int J Radiat Oncol Biol Phys.* 2013; 85(5):1360-1366.
- Woods K, Nguyen D, Tran A, et al. Viability of noncoplanar VMAT for liver SBRT compared with coplanar VMAT and beam orientation optimized  $4\pi$  IMRT. *Adv Radiat Oncol.* 2015;1(1):67-75.
- Dong P, Lee P, Ruan D, et al.  $4\pi$  noncoplanar stereotactic body radiation therapy for centrally located or larger lung tumors. *Int J Radiat Oncol Biol Phys.* 2013;86(3):407-413.
- Nguyen D, Rwigema JC, Yu VY, et al. Feasibility of extreme dose escalation for glioblastoma multiforme using  $4\pi$  radiotherapy. *Radiat Oncol.* 2014;9:239.
- Dong P, Nguyen D, Ruan D, et al. Feasibility of prostate robotic radiation therapy on conventional C-arm linacs. *Pract Radiat Oncol.* 2014;4(4):254-260.
- Tran A, Zhang J, Woods K, et al. Treatment planning comparison of IMPT, VMAT and  $4\pi$  radiotherapy for prostate cases. *Radiation Oncology.* 2017;12(1):10.
- Rwigema JC, Nguyen D, Heron DE, et al.  $4\pi$  noncoplanar stereotactic body radiation therapy for head-and-neck cancer: potential to improve tumor control and late toxicity. *Int J Radiat Oncol Biol Phys.* 2015;91(2):401-409.
- Chang JY, Balter PA, Dong L, et al. Stereotactic body radiation therapy in centrally and superiorly located stage I or isolated recurrent non-small-cell lung cancer. *Int J Radiat Oncol Biol Phys.* 2008;72(4):967-971.
- Chang JY, Li Q-Q, Xu Q-Y, et al. Stereotactic ablative radiation therapy for centrally located early stage or isolated parenchymal recurrences of non-small cell lung cancer: how to fly in a "no fly zone". *Int J Radiat Oncol Biol Phys.* 2014;88(5):1120-1128.
- Cai D, He X, Han J. Semi-supervised regression using spectral techniques. Department of Computer Science, University of Illinois at Urbana-Champaign; 2000. Report No. UIUCDCS-R-2006-2749.
- Ng AY, Jordan MI, Weiss Y. On spectral clustering: analysis and an algorithm. *Paper presented at: Advances in Neural Information Processing Systems*, 9–14 December 2002, Vancouver, British Columbia, Canada.
- Cai D, He X, Han J. Spectral regression: a unified approach for sparse subspace learning. *Paper presented at: Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 28–31 October 2007, Omaha, NE. ISSN: 1550-4786.
- Cai D, He X, Han J. Speed up kernel discriminant analysis. *VLDB J.* 2011;20(1):21-33.
- Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *Acm T Intel Syst Tec.* 2011;2(3):27.