Evaluation of plan quality assurance models for prostate cancer patients based on fully automatically generated Pareto-optimal treatment plans

Yibing Wang¹,* Sebastiaan Breedveld¹, Ben Heijmen¹ and Steven F Petit¹

¹ Department of Radiation Oncology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands.

Mail address: P.O. Box 5201, 3008 AE Rotterdam, the Netherlands
Visiting address: Groene Hilledijk 301, 3075 EA Rotterdam, the Netherlands

*Corresponding author: Yibing Wang, MSc

Phone: +31 107042691
Email: y.wang.1@erasmusmc.nl

Running title: Evaluation of treatment planning QA models

Keyword: knowledge based automated planning, overlap volume histogram (OVH), plan quality assurance, prostate radiotherapy IMRT/VMAT, Dose Volume Histogram (DVH) prediction

Abstract

IMRT planning with commercial TPSs is a trial-and-error process. Consequently, the quality of treatment plans may not be consistent among patients, planners and institutions. Recently, different plan quality assurance (QA) models have been proposed, that could flag and guide improvement of suboptimal treatment plans. However, the performance of these models was validated using plans that were created using the conventional trial-and-error treatment planning process. Consequently, it is challenging to assess and compare quantitatively the accuracy of different treatment planning QA models. Therefore, we created a golden standard dataset of consistently planned Pareto-optimal IMRT plans for 115 prostate patients. Next, the dataset was used to assess the performance of a treatment planning QA model that uses the overlap volume histogram (OVH).
115 prostate IMRT plans were fully automatically planned using our in-house developed TPS Erasmus-iCycle. An existing OVH model was trained on the plans of 58 of the patients. Next it was applied to predict DVHs of the rectum, bladder and anus of the remaining 57 patients. The predictions were compared with the achieved values of the golden standard plans for the rectum $D_{\text{mean}}$, $V_{65}$, and $V_{75}$, and $D_{\text{mean}}$ of the anus and the bladder.

For the rectum, the prediction errors (predicted-achieved) were only $-0.2\pm0.9$ Gy (mean±1 SD) for $D_{\text{mean}}$, $-1.0\pm1.6\%$ for $V_{65}$, and $-0.4\pm1.1\%$ for $V_{75}$. For $D_{\text{mean}}$ of the anus and the bladder, the prediction error was $0.1\pm1.6$ Gy and $4.8\pm4.1$ Gy, respectively. Increasing the training cohort to 114 patients only led to minor improvements.

A dataset of consistently planned Pareto-optimal prostate IMRT plans was generated. This dataset can be used to train new, and validate and compare existing treatment planning QA models, and has been made publicly available. The OVH model was highly accurate in predicting rectum and anus DVHs. For the bladder, larger prediction errors were observed.

### 1. Introduction

Intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) have increased the organ sparing potential for prostate patients compared to 3D conformal radiotherapy (3D-CRT) (Palma et al., 2008). Because IMRT/VMAT planning with the current commercial treatment planning systems (TPSs) is a trial-and-error process, based on a series of subjective human decisions, the quality of the IMRT/VMAT treatment plans may not be consistent among patients, planners or institutions with different experience (Nelms et al., 2012).

To ensure consistent high quality treatment plans, several groups have developed treatment plan quality assurance (QA) models, which predict the feasible dose levels for the organs at risk (OARs). The achieved dose of a treatment plan can be compared with the predicted dose by such a plan QA model to determine if the OARs can be spared further. Wu et al. and Kazhdan et al. have proposed a method using Overlap Volume Histogram (OVH), to predict achievable OAR doses (Wu et al., 2009; Kazhdan et al., 2009). The OVH describes the orientation and position of an organ at risk (OAR) with respect to the planning target volume (PTV). This method has been successfully applied to the prostate (Yang et al., 2013; Wu et al., 2014; Wang et al., 2013), head-and-neck (Wu et al., 2009; Wu et al., 2011), pancreas (Petit et al., 2012) and lung cancer patients (Petit and van Elmpt, 2015). Appenzoller et al. and Shiraishi et al. derived the predicted dose volume histogram (DVH) by predicting and summing up the doses of multiple subvolumes of the OAR, which were with different distances to the PTV surface. The prediction of the doses to the OAR subvolumes was based on a skew-normal distribution model (Appenzoller et al., 2012; Shiraishi et al., 2015). Good et al. investigated a mutual information method to select the most similar case in a database of prior treated patients to predict doses for prostate patients (Good et al., 2013). For head-and-neck patients, Moore et al. predicted mean parotid gland
doses from the overlap with the target (Moore et al., 2011). Zhu et al. and Yuan et al. developed a principle component analysis (PCA) based model to investigate the major factors that affect OAR doses in prostate and head-and-neck patients (Zhu et al., 2011; Yuan et al., 2012). For prostate patients, Nwankwo et al. predicted the dose in each OAR voxel, based on the geometric position relative to the target (Nwankwo et al., 2014).

In most studies, the prediction accuracy of the plan QA models was assessed by comparing the predictions with the achieved plans generated with the same trial-and-error planning process as used in clinical practice. Therefore, the plans that are used to validate the models may be suboptimal (not Pareto-optimal) and with inconsistent tradeoffs among sparing different OARs while ensuring sufficient PTV coverage. Consequently, it is unclear if an observed prediction error is due to the inherent limitation of the model, or due to the non-optimality of the treatment plans used for validation, or due to different tradeoffs of sparing multiple OARs. As a result, accurate establishment of the inherent prediction accuracy of most plan QA models is lacking. This hampers the comparison of the performance of different treatment planning QA models and may hamper the clinical introduction of any of such models.

In recent years we have developed Erasmus-iCycle, an algorithm for fully automated generation of Pareto-optimal treatment plans with consistent priorities (Breedveld et al., 2012). With Erasmus-iCycle, a protocol specific wish list is used. The wish list reflects the goals of our radiation oncologists with respect to PTV coverage and OAR sparing for a certain protocol and employs hard constraints and objectives in a predefined order of priority. Each treatment plan that was generated using the same wish list, is Pareto optimal and reflects the same balance between the different treatment planning tradeoffs. Therefore Erasmus-iCycle is able to generate a dataset that can be considered as the ground truth to evaluate the performance of different treatment planning QA models.

The purpose of the current study was threefold: (i) to create a dataset of Pareto optimal treatment plans for prostate cancer patients that can be considered as the ground truth for the evaluation of the performance of treatment planning QA models; ii) to make this dataset publicly available to the radiotherapy community so that any groups can train and evaluate the performance of their models on a common dataset; (iii) to validate the performance of one of the published models, namely the OVH model, for prostate cancer patients treated with IMRT/VMAT.

2. Materials and Methods

2.1 Patients and treatment plans
This study includes all prostate cancer patients treated in 2014 at our institute that had an estimated 10%-25% risk of tumor cells in the seminal vesicles. For 115 patients all relevant data could be retrieved and these patients were included in the analyses. The prostate was considered as the high risk CTV. The PTV1 was equal to the CTV expanded with a margin of 5
mm (7 mm in caudal direction). The PTV2 composed PTV1 including the seminal vesicles that were expanded by an isotropic margin of 8 mm (Mutanga et al., 2011). The prescribed dose was 78 Gy to PTV1 and 72.2 Gy to PTV2 in 39 fractions. For all patients VMAT treatment plans were generated using a simultaneously integrated boost technique.

At our institute all VMAT plans are fully automatically generated in a 2-step process (Voet et al., 2014). In the first step, Erasmus-iCycle was used for automated generation of a 23-equiaangular-beam IMRT treatment plan, using hard constraints and prioritized objectives for consistent steering of the multi-criterial plan optimization (see next section). In the next step, the resulting DVH parameters for target and OARs were used to generate for each patient an individualized Monaco template, that was used to produce the final treatment plan using our clinical TPS (Monaco, version 3.3 - 5.0, Elekta AB, Sweden). Generation of the Erasmus-iCycle plan, the Monaco template and the final Monaco plan were fully automated. This procedure allows plans generated by an in-house developed software to be delivered clinically. Previously we demonstrated that the automated process resulted in equal or better plan quality compared to conventional manual planning with the clinical TPS by an expert planner in absence of limitations in planning time (Voet et al., 2014). And deviations between the Erasmus-iCycle plan and the final plan generated with the clinical TPS are generally small (Voet et al., 2014). However, to get a maximally consistent dataset, this study was performed using the original Erasmus-iCycle plans.

### 2.2 Multi-criterial plan generation with Erasmus-iCycle

The treatment planning goals were: 99% of both PTVs needed to receive at least 95% of the prescribed dose (78 Gy for PTV1 and 72.2 Gy for PTV2); the rectum volumes that received more than 65 Gy (V<sub>65</sub>) and 75 Gy (V<sub>75</sub>) needed to be lower than 30% and 10%, respectively; the mean dose to the anus had to be lower than 45 Gy; and the maximum dose to the femoral heads could not exceed 55 Gy.

Table 1 shows how these goals and their order of priority were translated into a wish list using hard constraints and objectives with assigned priorities. This wish list was generated in close collaboration between physicists and radiation oncologists and is used as well in clinical practise. All the cost functions that are used are convex and therefore the global optimum is guaranteed to be found (Breedveld et al., 2009). So the resulting plans are not just of high quality but are the best plans achievable given the priorities in the wish list of ensuring PTV coverage and sparing different OARs. Details on multi-criterial plan generation with such a list, based on lexicographic optimization, can be found in (Breedveld et al., 2009; Breedveld et al., 2012; Voet et al., 2013; Voet et al., 2014). Apart from the common mean and maximum dose cost functions, Erasmus-iCycle also uses the Logarithmic Tumor Control Probability (LTCP) (Alber and Reemtsen, 2007) and the generalized equivalent uniform dose (gEUD) (Niemierko, 1999) cost functions. The LTCP cost function is defined as:

\[
LTCP = \frac{1}{N} \sum_{i}^{N} e^{-\alpha(D_i - DP)}
\]
Where $N$ is the number of voxels in the volume, $D_i$ is the dose to the $i^{th}$ voxel, $D^p$ is the prescribed dose and $\alpha$ is a predefined cell sensitivity parameter. The LTCP cost function controls PTV coverage. If the dose in each PTV voxel is equal to the prescribed dose, the LTCP equals to 1. Voxels with a dose lower than the prescribed dose add to the penalty exponentially. If the dose is higher than the prescribed dose, then the LTCP value will slowly drop to 0. In our experience, the cell sensitivity parameter $\alpha = 0.8$ results in acceptable coverage for the PTV for the far majority of patients.

For the rectum, two gEUD cost functions were applied, described by

$$gEUD = \left( \frac{1}{N} \sum_{i}^{N} D_i^\alpha \right)^{\frac{1}{\alpha}}$$

where $\alpha$ is a user-defined parameter, $D_i$ is the dose to the $i^{th}$ voxel in the OAR, and $N$ is the number of total voxels in the OAR. The applied $\alpha = 12$ and $\alpha = 8$ were used to minimize delivery of high and intermediate-to-high doses to the rectum.

For the bladder and anus, only the mean dose was minimized. To yield conformal dose distributions without hotspots in unspecified tissue, shells around the PTV and a shrinkage of the patient contour were created and applied.
Table 1: The wish list: consisting of hard constraints and prioritized objectives

LTCP = logarithmic tumor control probability, gEUD = generalized equivalent uniform dose, $D_{p_1} = 78$ Gy, $D_{p_2} = 72.2$ Gy. Arrows pointing downwards indicate minimization of that objective.

<table>
<thead>
<tr>
<th>Hard constraints</th>
<th>Volume</th>
<th>Type</th>
<th>Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTV1</td>
<td>maximum</td>
<td></td>
<td>104% × $D_{p_1}$</td>
</tr>
<tr>
<td>PTV2 – PTV1</td>
<td>maximum</td>
<td></td>
<td>104% × $D_{p_2}$</td>
</tr>
<tr>
<td>PTV2 shell 50 mm</td>
<td>maximum</td>
<td></td>
<td>50% × $D_{p_1}$</td>
</tr>
<tr>
<td>Rectum</td>
<td>maximum</td>
<td></td>
<td>$D_{p_1}$</td>
</tr>
<tr>
<td>Bladder</td>
<td>maximum</td>
<td></td>
<td>$D_{p_1}$</td>
</tr>
<tr>
<td>Femoral heads</td>
<td>maximum</td>
<td></td>
<td>40 Gy</td>
</tr>
<tr>
<td>Unspecified tissue</td>
<td>maximum</td>
<td></td>
<td>104% × $D_{p_1}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Priority</th>
<th>Volume</th>
<th>Type</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 PTV1</td>
<td>1</td>
<td>↓LTCP ($\alpha = 0.8$)</td>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>2 PTV2</td>
<td>2</td>
<td>↓LTCP ($\alpha = 0.8$)</td>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>3 Rectum</td>
<td>3</td>
<td>↓gEUD ($a = 12$)</td>
<td></td>
<td>30 Gy</td>
</tr>
<tr>
<td>4 Rectum</td>
<td>4</td>
<td>↓gEUD ($a = 8$)</td>
<td></td>
<td>20 Gy</td>
</tr>
<tr>
<td>5 Rectum</td>
<td>5</td>
<td>↓mean</td>
<td></td>
<td>10 Gy</td>
</tr>
<tr>
<td>6 External ring</td>
<td>6</td>
<td>↓maximum</td>
<td></td>
<td>40% × $D_{p_1}$</td>
</tr>
<tr>
<td>7 PTV shell 5 mm</td>
<td>7</td>
<td>↓maximum</td>
<td></td>
<td>93% × $D_{p_1}$</td>
</tr>
<tr>
<td>8 Anus</td>
<td>8</td>
<td>↓mean</td>
<td></td>
<td>10 Gy</td>
</tr>
<tr>
<td>9 PTV shell 15 mm</td>
<td>9</td>
<td>↓maximum</td>
<td></td>
<td>70% × $D_{p_1}$</td>
</tr>
<tr>
<td>10 PTV shell 25 mm</td>
<td>10</td>
<td>↓maximum</td>
<td></td>
<td>50% × $D_{p_1}$</td>
</tr>
<tr>
<td>11 Bladder</td>
<td>11</td>
<td>↓mean</td>
<td></td>
<td>40 Gy</td>
</tr>
<tr>
<td>12 Femoral heads</td>
<td>12</td>
<td>↓maximum</td>
<td></td>
<td>20 Gy</td>
</tr>
</tbody>
</table>

2.3 Treatment plan QA model

In this study we used overlap volume histograms (OVHs) to predict the DVHs of the OARs. A detailed description of the OVH method was given previously (Wu et al., 2009; Kazhdan et al., 2009). In brief, the OVH model assumes that dose to points (voxels) of an OAR decreases with increasing distance from the PTV. To predict the DVH dose $D_v$ at fractional volume $v$ for a given patient, the method uses the distance from the PTV, $r_v$, within which a the closest fraction $v$ of the OAR is located.

So:

$$r_{v,PTV} = \max(d(p_m,PTV\ surface) | p_m \in V)$$

Where $p_m$ is an OAR point, belonging to the OAR fraction $V$; $V$ is the closest fraction of the OAR to the PTV surface, with the volume of $v$; and $d(p_m,PTV\ surface)$ represents the Euclidean distance of the point $p_m$ to the PTV surface.
Then the prediction of $D_v$ is the minimal dose among all prior patients for which $r_{v,PTV}$ is smaller than $r_{v,PTV}$ or mathematically:

$$D_v = \min_i \left( D_{v,i} \mid r_{v,i,PTV} < r_{v,PTV}, i \in N \right)$$

Where $v$ is the volume of interest; $D_v$ is the dose to the OAR at volume $v$; $i$ is the index of a prior patient and $N$ is the total number of prior patients.

The assumption that the dose decreases with distance from the PTV is valid when the mean dose to an OAR (or the dose to a large subvolume of an OAR) is planned (optimized) to be as low as possible, as is the case for the rectum, bladder and anus, but e.g. not for the spinal cord (Petit and van Elmpt, 2015). Since points further away from the PTV can be spared more easily than points closer to the PTV, lowering the mean dose to the OAR will in general lead to a strong decreasing dose-distance relation.

However to ensure sufficient PTV coverage, for points of an OAR that are inside the PTV (i.e. $r < 0$), the primary goal is to ensure sufficient dose, instead of as little as possible. From this point of view, it is more difficult to deliver high dose to a point (within the PTV) closer to the PTV boundary than to a point more centrally located in the PTV. So for a certain $v$ that $r_v < 0$ (i.e., within the PTV), more difficult patients have larger (i.e. less negative) $r$. Therefore for points within the PTV, the predicted dose is given by:

$$D_v = \max_i \left( D_{v,i} \mid r_{v,i,PTV} > r_{v,PTV}, i \in N \right)$$

For simultaneous integrated boosts (SIB) plans, the doses to points of the OAR are related to the distances to the surfaces of both the low dose PTV (PTV1) and the boost PTV (PTV2), especially for the subvolumes that are inside PTV2 but outside PTV1. For a point in such subvolume, it becomes more difficult to deliver sufficient dose as the distance to the PTV2 boundary decreases and the distance from the PTV1 boundary increases (See Figure 1). For the subvolumes that are outside the PTV, we only consider the doses to PTV2 surface.
Figure 1. For simultaneous integrated boosts (SIB) plans, if a point is inside PTV2 (low dose PTV), it is more difficult to deliver sufficient dose if it is closer to the PTV2 boundary and further from the PTV1 (boost PTV) boundary. So as shown in the sketch, it is more difficult to deliver sufficient dose to point P than point P', for $r'_{v,PTV1} < r_{v,PTV1}$ and $r'_{v,PTV2} < r_{v,PTV2}$. Note that $r$ is negative if it is inside the PTV.

Therefore, the dose prediction model can be described as:

$$D_v = \min(D_{v,i} \mid r_{v,i,PTV2} < r_{v,PTV2}, i \in N) \text{ if } r_{v,PTV2} > 0 \text{ and } r_{v,i,PTV2} > 0$$

$$D_v = \max(D_{v,i} \mid r_{v,i,PTV1} > r_{v,PTV1} \text{ and } r_{v,i,PTV2} > r_{v,PTV2}, i \in N) \text{ if } r_{v,PTV2} < 0 \text{ and } r_{v,i,PTV2} < 0$$

This more detailed specification of the OVH was not described previously, because in prior studies the volumes of interests were not completely located within the PTV (Wu et al., 2009; Wu et al., 2011), i.e., $r_v > 0$. In the current study we attempt to predict the entire DVH. Because the rectum and bladder (can) partly overlap with the PTV, this specification was necessary to predict the dose to the high dose volumes of the DVH.

Entire DVHs were predicted by repeating the procedure described above for relative volumes ranging from 0 to 100% in steps of 0.5%. For the data points that could not be predicted (no candidates for prediction found in training cohort), the predicted doses were linearly interpolated by the surrounding predictable DVH points. Because the predicted DVH points of the current patient may be derived from data points of different prior patients, the predicted DVHs might be noisy and not always monotonically decreasing. Therefore, finally the predicted DVHs were smoothed three times by average filters with window widths of 8%, 5% and 2% of the relative volumes, respectively. With the larger windows large peaks could be reduced while the smaller windows reduced smaller peaks. The windows widths were selected based on empirical observations. A typical example is shown in Figure S1 of the supplementary materials.
2.4 Model evaluation
The 115 patients were randomly divided into a training group (N=58) and a validation group (N=57). For each patient in the validation group, the DVHs of the rectum, anus, and bladder were predicted by the QA model, based on the patients in the training group. The performance of the prediction model was assessed by comparing predictions of rectum $D_{\text{mean}}$, $V_{65}$, and $V_{75}$, and the $D_{\text{mean}}$ of the anus and bladder with the corresponding values in the Erasmus-iCycle plans.

To investigate whether the size of the training dataset affects the performance of the model, the DVHs of each patient in the validation group were also predicted by a leave-one-out method, i.e., for each of these 57 patients, the training group consisted of all other 114 patients.

2.5 Influence of inconsistent treatment planning priorities
The balance between different treatment plan tradeoffs affects the final plan. If this balance varies among patients in the training set, the model may be over optimistic. For instance, consider the example where for one patient in the training cohort the rectum is sacrificed to optimally spare the bladder, and for another vice versa. The OVH model may select the first patient to predict the bladder dose and the second to predict the rectum dose, while in reality it is impossible to optimally spare both simultaneously. This effect is likely to occur when the prioritization of the different planning goals is not consistent among patients, which is often the case in clinical practice.

To investigate this effect, another set of Erasmus-iCycle plans were generated. In this dataset, half of the plans in both training and validation cohort were generated with the clinical wish list (Table 1), and the other half was generated with enhanced bladder sparing at the expense of rectum sparing, achieved by increasing the priority of the bladder mean dose objective to 3, and lowering the priority of all remaining cost functions by one. As a result, both training and validation cohort consisted of plans yielded from two different wish lists. Next the new training plans were used to train the OVH model, and the predicted rectum $D_{\text{mean}}$, $V_{65}$ and $V_{75}$ were validated on the new validation plans.

2.6 Statistics
Pearson’s correlation coefficients ($r$) were calculated to investigate the correlation between the predicted and achieved DVH metrics. Results are expressed as mean ± 1 standard deviation (SD). All analyses were performed with Erasmus-RTStudio in combination with python 2.7.

3. Results
The median volumes of PTV1 and PTV2 were 115.6 cc (range: 62.4 cc – 263.3 cc) and 173.4 cc (range: 105.0 cc – 295.5 cc), respectively
3.1 Erasmus-iCycle plans
A dataset of Pareto optimal treatment plans with consistent treatment planning priorities was created and made publicly available (Petit, 2015).

Figure 2 shows DVHs for the Erasmus-iCycle plans of all 115 study patients. The mean doses varied considerably amongst all patients (rectum: 10.3 – 43.4 Gy; bladder 15.9 Gy - 67.1 Gy; anus: 1.7 Gy - 44.2 Gy). This large spread indicated a large heterogeneity in organ sparing potential among the different patients, depending on their anatomy. For 98.2% (N = 113) of the patients, all clinical goals for all OARs were satisfied. For the remaining two (1.7%), the clinical objective for rectum $V_{75}$ was not achieved. This could be explained by the overlap between the rectum and PTV2 that was as large as 14.8% and 15.1%, prohibiting adherence to the rectum $V_{75}$ goal without a severe reduction of PTV coverage (which had the highest priority in plan generation).

![Figure 2](image-url)

Figure 2. DVHs of all patients in the training and validation groups, demonstrating the large variety of DVHs.

3.2 Plan QA model validation
Figure 3 shows the predicted and achieved DVHs for one randomly selected patient of the validation group. The predicted and achieved DVH of all the patients and OARs are presented in Figure S2 of the supplementary material. For the far majority of the patients, the entire DVHs could be predicted well for rectum and anus, but larger prediction errors were observed for the bladder. Figure 4 compares specific DVH parameters for all patients. For rectum, the mean differences (predicted - achieved) were small ($D_{\text{mean}}$: -0.2±0.9 Gy (r=0.99); $V_{65}$: -1.0±1.6% (r=0.97); $V_{75}$: -0.4±1.1% (r=0.93)). For $D_{\text{mean}}$, the difference was within 1 Gy and 2 Gy for 72% and 96% of patients in the validation cohort, respectively. The predictions were clearly less accurate for patients with high $V_{65}$ and $V_{75}$ ($V_{65} > 25\%, V_{75} > 14\%$).
Figure 3. For an example patient, the predicted rectum, anus and bladder DVHs compared with the achieved DVHs. See Figure S2 of the supplementary materials for the DVHs of all patients.
Figure 4. Comparison of the predicted and achieved DVH values (uneven rows) and difference plots (even rows) for the 57 patients in the validation cohort. The upper half shows the rectum $D_{\text{mean}}$, $V_{65}$, and $V_{75}$ and the lower half shows the anus $D_{\text{mean}}$ and the bladder $D_{\text{mean}}$.

For anus, the difference in $D_{\text{mean}}$ was small (predicted - achieved), as $0.1 \pm 1.6$ Gy ($r=0.99$), but the predicted bladder $D_{\text{mean}}$ systematically underestimated the achievable dose by $4.8 \pm 4.1$ Gy ($r=0.94$).
3.3 Influence of the size of the training set
Increasing the training data set from 58 to 114 patients did not lead to a considerable reduction in prediction errors, except in \( V_{65} \) for patients with large achieved \( V_{65} \) values (above 20%), as is shown in Figure 5. The prediction accuracy for the other DVH parameters are presented in Table S1 in the supplementary materials.

![Graph showing the relationship between predicted and achieved \( V_{65} \) for the rectum with two different training data sizes.](image)

Figure 5. For the 57 validation patients, comparison of predicted \( V_{65} \) with achieved values. The results with two different training data sizes are shown, i.e., using the original training dataset (N = 58) and using leave-one-out method (N = 114).

3.4 Influence of the consistency of the treatment plans in the training dataset
Figure 6 shows the prediction accuracy for the rectum when the priority of the bladder is increased compared to the rectum for half of the plans in the training and validation dataset. The difference between prediction and validation increased by 2 Gy (on average), to \(-2.2\pm3.1\) Gy for \( D_{\text{mean}} \), \(-2.1\pm2.6\) % for \( V_{65} \) and \(0.4\pm1.5\) % for \( V_{75} \). This suggests that the expected,
apparent accuracy of the OVH model decreases when the treatment plans in the training and validation database have not been generated with consistent planning priorities.

Figure 6. Predicted vs. achieved DVH metrics ($D_{\text{mean}}$, $V_{65}$, and $V_{75}$) for the rectum when half of the training and validation database consisted of plans with a higher bladder sparing priority than the priority of the rectum, and another half with the clinical prioritizations.

4. Discussion

The quality and consistency of IMRT/VMAT plans may vary among patients and dosimetrists. Plan QA models can help to ensure a high consistency and high quality for treatment plans. Different models have been proposed (Wu et al., 2009; Appenzoller et al., 2012; Shiraishi et al., 2015; Good et al., 2013; Moore et al., 2011; Zhu et al., 2011; Yuan et al., 2012; Nwankwo et al., 2014), but the validation data that were used to evaluate the performance of the QA methods, were generated by the same trial-and-error treatment planning process that may lead to suboptimal plans, as for which treatment plan QA is intended in the first place. Therefore, in the current study, treatment plans for 115 prostate cancer patients were fully automatically generated with lexicographic multi-criteria optimization, resulting in Pareto optimal treatment plans with a consistent balance between the various treatment planning tradeoffs. This dataset can be used as golden standard to validate, compare and develop new treatment plan QA methods and is made publicly available (Petit, 2015). In the current study we used it to evaluate the performance of the overlap volume histogram based plan QA model (Kazhdan et al., 2009; Wu et al., 2009) for prostate cancer patients.
The performance of the OVH model to predict the rectum and anus DVHs was excellent, as clearly indicated by the low systematic errors, standard deviations (SD) and visual agreement (supplementary materials Figure S2). The bladder $D_{\text{mean}}$ on the other hand was systematically underestimated by the OVH model by 4.8 Gy with a larger standard deviation. A systematic error can be corrected simply by increasing/decreasing all predictions by the size of that error. The size of the SD has more impact on the clinical suitability of a model, because it determines the width of the confidence interval to distinct sub-optimal plans from prediction errors. E.g. with a confidence of 95%, only differences larger than a threshold of 1.96 times the SD can be attributed to sub-optimality of the plans, given that systematic errors are predicted for. These thresholds were 1.7 Gy (rectum), 3.1 (anus) and 8.0 Gy (bladder) in the current study which makes the model very suitable for the rectum and anus but less for the bladder.

The large prediction errors of the bladder compared to the rectum and anus can partly be explained by their priorities (high for rectum, low for bladder) in the planning process. Because of the prioritized optimization the organ sparing potential depends on PTV coverage and on all other organs that have a higher priority in the wish list. The OVH model, on the other hand, does not consider inter-organ dependencies other than the PTV. However, increasing the bladder priority in both the validation and training cohort led only to a modest gain in prediction accuracy for the bladder of 2.4±3.3 Gy (not shown in result section). Another explanation could be the differences in dose fall off in the superior-inferior direction compared to the in the axial direction. This could have an impact on the accuracy of the OVH model for organs with various orientations with respect to the edges of the treatment fields (Petit et al., 2012).

When the training and validation groups consist of a mix of plans generated with two different wish lists (high-prioritized and low-prioritized bladder sparing), the apparent accuracy of the OVH model substantially decreased. This finding indicates the dependence of the accuracy of the model on the consistency of the training dataset. In addition, it shows that by using inconsistent planning priorities among the plans in the validation set, the apparent accuracy of the model is lower than the real accuracy, which confirms the need for Pareto optimal, highly consistent validation data to assess the performance of planning QA models. Our previous study (Wang et al., 2013) also showed that if the training dataset were not consistent, the performance of the planning QA model could be misinterpreted. These findings indicate that in clinical practice, the performance of the OVH model, when is used with conventionally optimized treatment plans, is likely less good than what we have found in this study.

Several studies have investigated and used the overlap volume histogram model for prostate dose prediction (Yang et al., 2013; Wu et al., 2014; Wang et al., 2013). However, none of these studies have predicted the entire DVH, but only a few specific dose volume values. Accurately predicting the dose to the part of the prostate that overlaps with PTV2 but not with PTV1 required some modifications to the original OVH model. Wu et al. used the OVH model to successfully predict the OAR doses for head and neck patients with an integrated boost.
technique with 3 risk levels (Wu et al., 2009; Wu et al., 2011), by considering the OVHs of all PTVs. However, the predicted doses were mostly lower than the lowest prescribed doses, which means outside the PTVs. In this study we have shown that it is also possible to accurately predict doses within the overlap region of an OAR and the PTVs for SIB patients.

In principle, Erasmus – iCycle, as an automated plan generation tool, could also be applied as a planning QA tool for commercial planning systems. It overcomes the limitations of some of the ‘simplistic’ assumptions of the OVH model and therefore provides more sophisticated estimates on the achievable dose to the OARs. However, Erasmus-iCycle can be considered a full TPS, requiring similar costs in setup and maintenance, and computing power as a commercial TPSs. An investment in Erasmus-iCycle just for treatment planning QA may be too large. Therefore here we made Erasmus-iCycle plans available to improve and validated more ‘light weight’ treatment planning QA models.

In the era of automated treatment planning one could expect that the need for treatment planning QA reduces. However, we see a clear role for QA for automated treatment planning. If something in the fully automated treatment planning chain goes awry (mismatched structures, software bugs etc), this could potentially result in a plan that is of significantly lower quality than necessary, but not low enough for the physician to detect during the plan approval phase. Especially if the quality of automatically generated plans is of consistent high and unrivaled quality (i.e. cannot be improved by an experienced planner), criticism could start to impair. The OVH method is totally independent of the Erasmus-iCycle and therefore we are currently evaluating it as a sanity check for routine QA of automatically generated treatment plans.

5. Conclusions

For 115 prostate cancer patients, a Pareto-optimal treatment plan was generated with lexicographic multi-criterial optimization. This plan data set is highly suited for assessing the performance of models for treatment plan QA and has been publicly available for development and validation of treatment planning QA model. We tested the overlap volume histogram (OVH) based QA model, using the generated dataset. It was shown that the OVH model had a very high prediction accuracy for rectum and anus, while for bladder the accuracy was lower.

Acknowledgements

This work is in part funded by an research grant of Elekta AB, Stockholm, Sweden. SP acknowledges financial support from the Dutch Cancer Society (KWF EMCR 2014-6667). Erasmus MC Cancer Institute also has a research collaboration with Accuray Inc, Sunnyvale, USA.
References

Alber M and Reemtsen R 2007 Intensity modulated radiotherapy treatment planning by use of a barrier-penalty multiplier method Optimisation Methods and Software 22 391-411


Breedveld S, Storchi P R M and Heijmen B J M 2009 The equivalence of multi-criteria methods for radiotherapy plan optimization Physics in medicine and biology 54 7199


Niemierko A 1999 A generalized concept of equivalent uniform dose (EUD) Med Phys 26 1100

Nwankwo O, Sihono D S K, Schneider F and Wenz F 2014 A global quality assurance system for personalized radiation therapy treatment planning for the prostate (or other sites) Physics in medicine and biology 59 5575

Petit S F 2015 Radiotherapy treatment planning QA. www.erasmusmc.nl/RadiotherapyTreatmentPlanningQA

Petit S F and van Elmpt W 2015 Accurate prediction of target dose-escalation and organ-at-risk dose levels for non-small cell lung cancer patients Radiotherapy and Oncology


Yuan L, Ge Y, Lee W R, Yin F F, Kirkpatrick J P and Wu Q J 2012 Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans *Medical physics* **39** 6868-78

Zhu X, Ge Y, Li T, Thongphiew D, Yin F-F and Wu Q J 2011 A planning quality evaluation tool for prostate adaptive IMRT based on machine learning *Medical physics* **38** 719-26