Technology in Cancer Research &
Treatment
Volume 18: 1-10
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1533033819857758
journals.sagepub.com/home/tct
SAGE

Knowledge-Based Statistical Inference Method for Plan Quality Quantification

Jiang Zhang, MS[1], Q. Jackie Wu, PhD[2], Yaorong Ge, PhD[3], Chunhao Wang, PhD[2], Yang Sheng, PhD[2], Jatinder Palta, PhD[4], Joseph K. Salama, MD[2], Fang-Fang Yin, PhD[2], and Jiahan Zhang, PhD[2]

Abstract

Aim: The aim of the study is to develop a geometrically adaptive and statistically robust plan quality inference method. **Methods and Materials:** We propose a knowledge-based plan quality inference method that references to similar plans in the historical database for patient-specific plan quality evaluation. First, a novel plan similarity metric with high-dimension geometrical difference quantification is utilized to retrieve similar plans. Subsequently, dosimetric statistical inferences are obtained from the selected similar plans. Two plan quality metrics—dosimetric result probability and dose deviation index—are proposed to quantify plan quality among prior similar plans. To evaluate the performance of the proposed method, we exported 927 clinically approved head and neck treatment plans. Eight organs at risk, including brain stem, cord, larynx, mandible, pharynx, oral cavity, left parotid and right parotid, were analyzed. Twelve suboptimal plans identified by dosimetric result probability were replanned to validate the capability of the proposed methods in identifying inferior plans. **Results:** After replanning, left and right parotid median doses are reduced by 31.7% and 18.2%, respectively; 83% of these cases would not be identified as suboptimal without the proposed similarity plan selection. Analysis of population plan quality reveals that average parotid sparing has been improving significantly over time (21.7% dosimetric result probability reduction from year 2006-2007 to year 2016-2017). Notably, the increasing dose sparing over time in retrospective plan quality analysis is strongly correlated with the increasing dose prescription ratios to the 2 planning targets, revealing the collective trend in planning conventions. **Conclusions:** The proposed similar plan retrieval and analysis methodology has been proven to be predictive of the current plan quality. Therefore, the proposed workflow can potentially be applied in the clinics as a real-time plan quality assurance tool. The proposed metrics can also serve the purpose of plan quality analytics in finding connections and historical trends in the clinical treatment planning workflow.

Keywords

plan quality assurance, data analytics, knowledge-based planning

Abbreviations

CDF, cumulative distribution function; DDI, dose deviation index; DRP, dosimetric result probability; DVH, dose–volume histogram; gDTH, general distance-to-target histogram; IMRT, intensity modulated radiation therapy; KBP, knowledge-based planning; OARs, organs at risk; PDF, probability density function; PQM, plan quality metric; PTV, planning target volume.

Received: February 27, 2017; Revised: April 29, 2019; Accepted: May 22, 2019.

Introduction

The outcomes of intensity-modulated radiation therapy (IMRT) treatments are affected by various factors,[1] including positioning accuracy, machine delivery precision, and treatment plan quality. While positioning accuracy is actively monitored during each treatment and machine delivery precision is ensured by conducting plan-specific quality assurance procedures, plan quality evaluation is less objective. In current

[1] Division of Medical Physics, Duke Kunshan University, Kunshan, Jiangsu, China
[2] Department of Radiation Oncology, Duke University Medical Center, Durham, NC, USA
[3] College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte, NC, USA
[4] Department of Radiation Oncology, Virginia Commonwealth University, Richmond, VA, USA

Corresponding Author:
Jiahan Zhang, PhD, Duke University Medical Center, DUMC 3295, 200 Trent Dr, Durham, NC 27710, USA.
Email: jiahan.zhang@duke.edu

practice, planners and physicians' judgments play a significant role in plan quality evaluation and improvement. Patient anatomy, especially the geometric shape between planning target volumes (PTVs) and organs at risk (OARs), can greatly affect OARs dose sparing difficulty. Hence, plan quality evaluation is inherently patient-specific and challenging, even for experienced planners and physicians, due to the high spatial variations in PTVs and OARs.

Attempts have been made to develop objective and quantitative plan quality metrics (PQMs). An in-house plan quality score named "plan quality metric" was used by Nelms *et al*[1] to evaluate one prostate treatment plan quality variability. It is a hard-coded continuous scoring function, which evaluates quality-affecting factors such as institutions, treatment planning systems, and planners, and so on. Recently, a set of population-based PQMs were developed by Mayo *et al*,[2] which used statistical analysis of dose–volume histograms (DVHs) of historical head and neck IMRT treatment plans. However, all the historical plans were referenced in the PQMs without considering anatomical variations. Therefore, the metrics are objective but may lack patient-specific evaluation contribution. Knowledge-based planning (KBP) approaches[3-7] explicitly model anatomical variations and generate "best achievable" patient-specific DVH predictions. Any clinically significant dose–volume point on the predicted DVH curve can be directly used as objectives in plan optimization. Recently, it has been proposed to use KBP models for independent quality assessment of the clinical Pinnacle's Auto planning.[8] Although KBP methods are proven to be successful in improving planning efficiency and ensuring consistent plan quality in the clinical workflow,[9-19] the robustness of the current KBP models can be questionable for complicated treatment sites such as head and neck. More specifically, most current KBP models only use the overlap volume histogram (OVH) of one OAR for its DVH prediction without considering the effect from the anatomies of other OARs relative to PTVs, while for complicated plans, the model fails to take into account relative positions of multiple PTVs and tradeoffs of multiple OARs.

We propose a knowledge-based plan quality inference method that references to similar plans in the historical database for patient-specific plan quality evaluation. First, reference plans are selected based on a novel plan similarity metric that includes both anatomical feature difference and dose prescription ratio difference. Anatomical features are characterized by a set of general distance-to-target histogram (gDTH) matrices accounting for potential multiple PTVs. Subsequently, statistical inferences are performed on these reference plans and used to evaluate the quality of the current plan. In this study, head and neck IMRT plans with sequential boost treatments are used to perform the plan quality inference. Two self-developed patient-specific PQMs, dosimetric result probability (DRP) and dose deviation index (DDI), are used to generate plan quality evaluation scores. Experiments are conducted to validate the method and demonstrate the potential application in plan quality analytics.

Methods and Materials

In this study, we define a novel plan similarity metric (described in detail in section "Plan Similarity Metric") that includes quantification of organ shape and distance differences between a historical plan and the target plan (ie, the plan being evaluated). Similar plans are selected based on the plan similarity metric and subsequently referenced for plan quality inference. The 2 PQMs are constructed based on the dosimetric statistical inference of the selected similar plans. To automate the plan evaluation process, we have developed a stand-alone application with multiple modules including historical data extraction, target plan selection, plan similarity analysis, and PQM evaluation and visualization. Nine hundred twenty-seven clinically approved head and neck treatment plans with 2 PTVs were exported automatically and used as a historical database for the experiments. Dose–volume points of clinical interest are referred to as dosimetric results in this article.

Historical Data Extraction

To efficiently extract data from the clinical database, we have developed a stand-alone application using Windows Presentation Foundation application based on .NET 4.5 framework. Multiple modules, including historical data extraction, reference plan selection, plan similarity analysis, and PQM evaluation and visualization, have been developed and integrated using enterprise-standards Model–View–ViewModel pattern implemented through Prism library.[20] Eclipse scripting application interface is utilized for accessing and extracting clinical treatment plans. During data extraction, all Health Insurance Portability and Accountability Act (HIPPA) sensitive information is anonymized. Furthermore, the 8 structures used in this application conform to the nomenclature standardization defined in TG263,[21] with standardized IDs of Mandible, Parotid_L, Parotid_R, SpinalCord_PRV05, Cavity Oral, Pharynx, Larynx, and Brainstem. To correctly collecting OAR-specific information, we developed and applied a routine to identify the names of those structures in the treatment planning system (TPS) based on regular expressions and our prior knowledge of the naming convention of our institution. Both segmented structure volumes and the DVHs of each treatment plan are exported. A total of 927 clinically approved head and neck IMRT treatment plans are extracted and anonymized.

Plan Similarity Metric

To evaluate the difference between 2 plans in terms of expected dose distributions, we take into consideration the anatomy difference and the plan prescription difference. In order to quantify the anatomical difference, the anatomical feature of one treatment plan is described by a series of newly developed gDTH matrices. The gDTH is derived from the 1D OAR distance to PTV descriptor DTH. It is an abstraction of the geometric relationship between one OAR and one PTV and represents both the shape of the 2 volumes and the distance between them.[22] The gDTH extends the application of 1D cumulative DTHs to a treatment plan with multiple PTVs by adding additional dimensions. For 2PTV plans, each element in the 2D gDTH matrix with the distance index of t_1 and t_2 is defined as the portion of the OAR volume that has a maximum distance t_1 to primary PTV and maximum additional distance t_2 to boost PTV at the same time among all the voxels inside:

$$gDTH(t) = \frac{|\{p\epsilon OAR|d(p, PTV_{pri}) \leq t_1,\ d(p, PTV_{bst}) - d(p, PTV_{pri}) \leq t_2\}|}{|OAR|},$$

where PTV_{pri} represents the primary PTV and PTV_{bst} represents the boost PTV. Similarly, this definition can be generalized to even higher dimensions for cases with more than 2 PTVs.

The anatomical difference $AD_{m,n}$ between target plan m and reference plan n is therefore quantified as the sum of the weighted gDTH matrix square differences for all the OARs to be analyzed:

$$AD_{m,n} = \sum_{t=1}^{T} \left[p_{m,t}\, q_t \sum_{i,j} (gDTH_{m,t,ij} - gDTH_{n,t,ij})^2 \right],$$

where t denotes each OAR, $p_{m,t}$ is a volume-dependent weighting factor, and q_t is a sparing priority weighting factor (ie, plan being evaluated). The gDTH matrix square difference between plan m and plan n for OAR t is calculated as the sum of square difference of each matrix element indexed at i and j. Prescription dose difference $PD_{m,n}$ between plan m and plan n is quantified as the prescription dose ratio square difference:

$$PD_{m,n} = \left(\frac{D_{m,\,pri}}{D_{m,bst}} - \frac{D_{n,\,pri}}{D_{n,bst}} \right)^2,$$

where the primary prescription dose is denoted as D_{pri} and boost prescription dose is denoted as D_{bst}. Therefore, the plan similarity metric $S_{m,n}$ between plan m and plan n is defined as the addition of a weighted sum of gDTH matrix square difference and weighted prescription dose ratio square difference:

$$S_{m,n} = \sum_{t=1}^{T} \left[p_{m,t}\, q_{m,t} \sum_{i,j} (gDTH_{m,t,ij} - gDTH_{n,t,ij})^2 \right] + l_m \left(\frac{D_{m,\,pri}}{D_{m,bst}} - \frac{D_{n,\,pri}}{D_{n,bst}} \right)^2,$$

where l_m denotes the relative weighting of the dose ratio difference. It is determined as the ratio of minimum anatomical difference averaged by the number of structures and the mean value of prescription dose ratio averaged by both the number of treatment plans and the number of structures.

To balance the contributions of OARs' size differences on the plan similarity matrix, a volume-dependent weighting factor $p_{m,t}$ is assigned to each OAR. Mathematically, a smaller OAR structure volume results in a smaller range of distances to primary PTV, leading to a smaller gDTH matrix. Therefore, small OARs on average yield small gDTH square differences, whereas big OARs tend to have relatively large values. If no volume-dependent weighting factor is applied, a small OAR in the selected similar plans would have more varied geometries. $p_{m,t}$ is designed as the inverse of the number of elements that are larger than 0 and smaller than 1 in the gDTH matrix of the target plan OAR.

The second weighting factor q_t takes both OAR sparing priorities and axial positions into consideration. The OAR sparing priorities are affected by various factors such as OAR's sensitivity to radiation and the impact of its potential complications to the patients' quality of life. Planners pay more attention to structures that have higher radiosensitivity and higher impact on the quality of life during plan evaluation. Therefore, the geometries of more prioritized OARs are more indicative of the plan dose distribution. Since coplanner beam technique is applied to all the treatment plans, the dose to the OARs that locate on the same axial plane with PTVs is more sensitive to their geometries due to direct irradiation. A higher weighting factor needs to be assigned for the OAR that has higher sparing priority and is closer to the PTV along axial direction. The weighting factor q_t for brain stem, mandible, larynx, pharynx, parotids, spinal cord, and oral cavity are set as 20, 50, 20, 20, 500, 20, and 20, respectively.

Plan Quality Metrics

Fifty (5.3% of all exported plans) reference plans are selected based on plan similarity scores. They serve as the sampled population for statistical inference in the target plan quality analysis. Two newly designed PQMs: DDI and DRP are used to evaluate dose sparing of each selected OAR structure.

Dosimetric result probability. The cumulative distribution function (CDF) of doses at any selected dose volume point retrieved from similar plans is evaluated. Outliers are identified and excluded using *Z*-score method before fitting. It is defined as the difference to the mean value divided by the standard deviation. The threshold for outliers is set to be 2 in this particular study. Nonparametric probability density function (PDF) fitting is performed through kernel density estimation and the corresponding CDF curve is obtained by integrating the fitted PDF curve. Two-order Gaussian kernel is selected for kernel density estimation and the bandwidth is selected based on Silverman's rule of thumb.[23] The CDF value of the target OAR's dose volume point of interest is named as DRP. The DRP provides an intuitive estimate of the dose sparing quality of the selected structure in a pool of plans with similar anatomies: a value smaller than 0.5 indicates a plan has better than median OAR sparing and vice versa.

Dose deviation index. Dose deviation index is defined as the difference between the mean values of the dosimetric results retrieved from selected similar plans and the dosimetric result of the current plan's OAR. Outliers are excluded in the same way as introduced before. A positive DDI indicates a lower dose than average to the target plan OAR and vice versa. The absolute value of DDI provides a numerical estimation of how

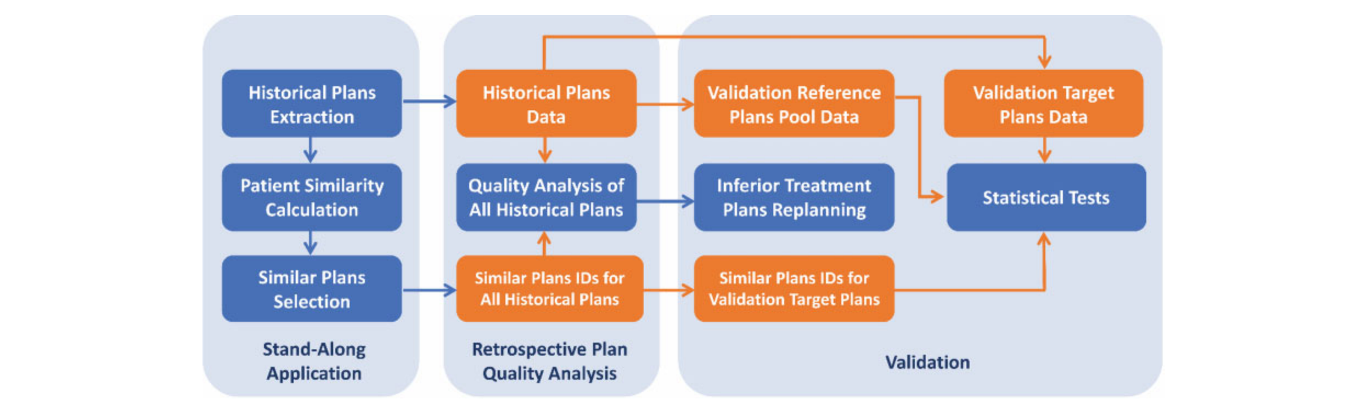
Historical Plans Extraction
Patient Similarity Calculation
Similar Plans Selection
Stand-Along Application
Historical Plans Data
Quality Analysis of All Historical Plans
Similar Plans IDs for All Historical Plans
Retrospective Plan Quality Analysis
Validation Reference Plans Pool Data
Inferior Treatment Plans Replanning
Similar Plans IDs for Validation Target Plans
Validation Target Plans Data
Statistical Tests
Validation

Figure 1. Experimental design workflow. Blue boxes are the actions performed and orange boxes denote the data acquired.

much dose relative to the primary prescription can be reduced or sacrificed to the current OAR while maintaining the overall plan optimality.

Experimental Design

The proposed plan quality inference method is expected to generate real-time robust estimations of relative plan quality scores compared to similar historical plans. The reference plan retrieval process was validated by evaluating the similarities of the selected similar plans using hold-out test data set as target plans. The capability of the proposed metrics in finding plans with low plan quality is demonstrated by replanning of the inferior plans picked out by the proposed PQM DRP. Furthermore, the proposed metrics can be used to analyze plan quality changes in between plan populations, and retrospective plan quality analysis was performed to assess the changes between years within our institution. The workflow of the experimental design is summarized in Figure 1.

Similar-plan retrieval validation. Statistical analysis was performed to validate that reference plans selected by the proposed plan similarity metric are indeed similar in terms of dosimetric results. Twenty-five percent (231) of all the approved plans in the local database were randomly selected as the test target plans and the rest (696) served as the reference plans pool. Median dose was evaluated for Larynx, Pharynx, Parotid_L, parotid_R, and Cavity_Oral, and maximum dose was evaluated for Brainstem, Mandible, and SpinalCord_PRV05. Those dosimetric results were chosen based on the dose constraint protocol in our institution. Intuitively, similar patient anatomies result in similar collective dose distributions for each OAR in high-quality treatment plans. This assumption was validated by statistically comparing the standard deviation of dosimetric results with and without similar plan selection. Note that when a target plan is compared with reference plans without similar plan selection, those reference plans include all the plans in the reference plan pool, similar to the prior research on plan quality evaluation. Our validation comparison directly compares the proposed method to the one in the previous research. Outliers in those reference plans are also excluded with the same criteria as the one used for similar plans.

Having more accurate PQMs with similar-plan selection also requires the right location of dosimetric result distribution. It was validated by statistically comparing the absolute values of the high-quality target plans' DDIs calculated with and without similar plans selection. Similarly, Student paired *t*-test was performed to compare the mean value of the 2 DDIs and both Wilcoxon signed rank test and sign test were performed to compare the median of the 2 DDIs. If all 3 tests are passed with P values less than .05[24] for all the OARs with the alternative hypothesis of having smaller DDIs with similar plans, a statistically smaller DDI with similar plans can be validated. Since small DDIs are expected for high-quality target plans, it can be concluded that the location of dosimetric results distribution based on similar plans for plan quality analysis is more accurate.

Plan quality assurance validation. Twelve inferior plans for replanning were randomly selected based on the criteria of having the DRP of both parotids' median dose larger than 0.9. Treatment plans with high DRPs for both parotids are more likely to be true inferior plans by reducing the chance of selecting plans with high DRPs solely caused by their unique geometries. Parotids were selected because they are the structures that have the most varied geometric configurations relative to the 2 PTVs. For example, both parotids in most bilateral head and neck treatment plans have negative minimal distances to the primary PTV, whereas only 1 parotid in most ipsilateral treatment plans has negative minimum distance to the primary PTV. Therefore, DRPs calculated from parotid median dose of the selected similar plans have more distinctive differences from the values calculated from all the reference plans, and superiority of applying the proposed PQM with similar plans selection can be emphasized through DRP comparison. The selected inferior plans can be validated to be indeed inferior if they have improved plan quality (lower DRP) after replanning. The proposed metrics can be validated to be more reasonable on suboptimal plans if most selected inferior plans have higher initial DRPs calculated with similar

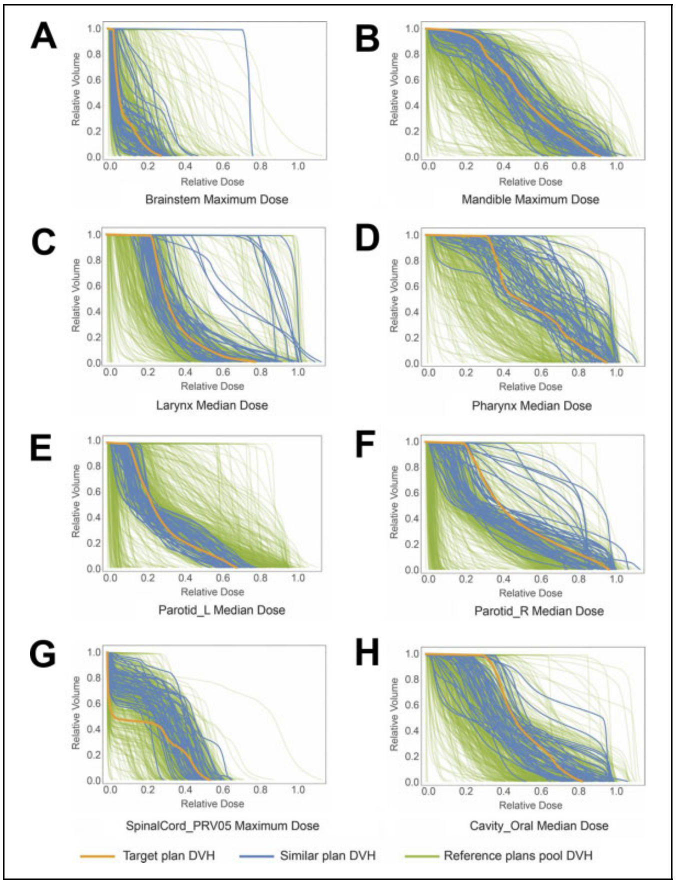

Figure 2. Similar treatment plans dose–volume histogram (DVH) curves of the 8 OARs for one target plan in validation. The blue curves in each plot are the DVHs of the selected similar plans OAR and the yellow curve in each plot is the target plan OAR DVH curve. The DVHs of all the plans in the reference plans pool are shown as green curves in the back. Most similar plans show similar overall DVH distributions with a few outliers.

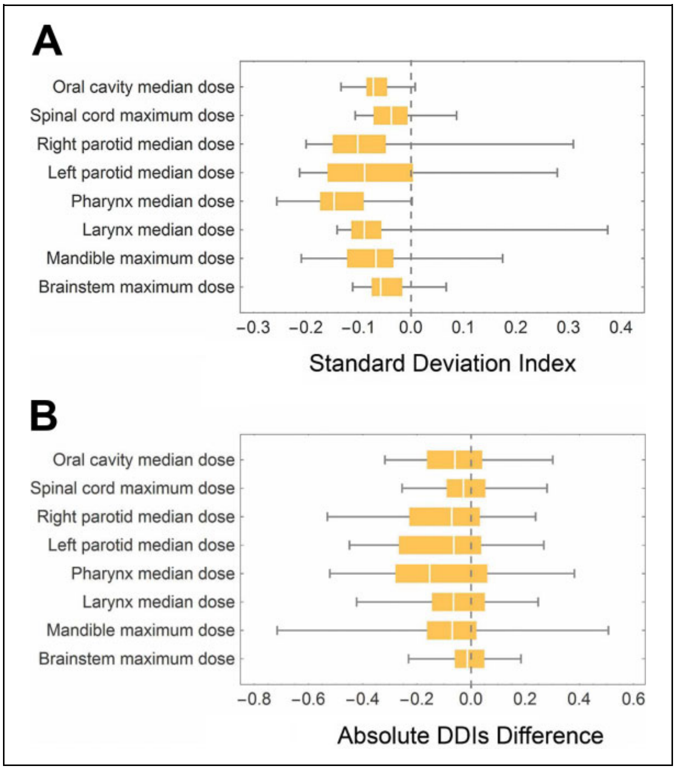

Figure 3. Distribution of standard deviation difference of dosimetric results and absolute dose deviation index (DDI) difference of randomly selected target treatment plans for validation. A, Box and whisker plot of dosimetric result standard deviation difference with and without similar plan selection. Each OAR with the corresponding dosimetric result is labeled on the left side. The central band is the median. The left and right edge of each box represent 25% and 75% quantile, respectively. The ends of the whiskers are the minimum and maximum of the data. All the 75% quantiles of the standard deviation difference are smaller or very close to 0. B, Box and whisker plot of absolute DDI differences with and without similar plan selection. Similarly, the central band is the median value and the box edges are 25% and 75% quantiles. The edges of whickers are minimum and maximum value of absolute dose deviation index (DDI) differences.

plan selection compared with the initial DRPs calculated without similar plan selection.

Retrospective plan quality analysis. Retrospective plan quality analysis is performed to assess the change of plan quality over time based on the average DRPs within every 2 years for each OAR in our institution. Median dose was evaluated for Larynx, Pharynx, Parotid_L, parotid_R, and Cavity_Oral, and maximum dose was evaluated for Brainstem, Mandible, and SpinalCord_PRV05. The DRPs of the 8 analyzed OARs in every historical treatment plan were calculated with similar plans selected from the same database. All DRPs calculated for each OAR were averaged in every 2-year period ranging from 2005 and 2018.

Results

Similar Plan Retrieval Validation

The DVH curves of all 8 OARs of the 50 selected similar plans from the reference plans pool for 1 randomly selected validation target plan are shown in Figure 2 as examples. The blue curves are the DVHs of the selected similar plans and the red curves are the target plan's DVH curves. It can be observed that similar plans' OARs show similar overall DVH distributions. The distributions of the differences between the standard deviation of the similar plans' dosimetric results for all the 8 OARs and the constant standard deviation of the reference plans pool are demonstrated as a box plot as shown in Figure 3A. At least 75% of the randomly selected target plans show a smaller standard deviation of dosimetric results for all the analyzed OARs. Both mean and median value of the standard deviation differences for all the 8 kinds of dosimetric results are negative. Similarly, the box plot for the absolute DDIs difference is plotted in Figure 3B. The median values of all dosimetric results are below 0. All the statistical tests comparing the absolute DDIs with and without similar plans selection except those for Brainstem median dose show a *P* value smaller than the significance level .05 (Table 1). Both the median and mean

Table 1. Statistical Tests of Absolute Dose Deviation Indexes (DDIs) With and Without Similar Plan Selection.

		With Similar Plan Selection		Without Similar Plan Selection			95% Confidence Interval of the Difference in Mean		Statistical Test		
Structure	Dosimetric Result	Mean |DDI| ± SE	Median	Mean |DDI| ± SE	Median	Difference in Mean			Paired *t*	Signed rank	Sign
Brainstem	Maximum dose	0.143 ± 0.013	0.091	0.150 ± 0.127	0.100	−0.007	−0.019	0.005	.132	.108	.056
Mandible	Maximum dose	0.175 ± 0.015	0.105	0.257 ± 0.016	0.213	−0.081	−0.102	−0.061	<.001	<.001	<.001
Larynx	Median dose	0.200 ± 0.023	0.103	0.260 ± 0.026	0.136	−0.060	−0.078	−0.041	<.001	<.001	<.001
Pharynx	Median dose	0.222 ± 0.014	0.186	0.342 ± 0.017	0.304	−0.120	−0.151	−0.090	<.001	<.001	<.001
Parotid_L	Median dose	0.162 ± 0.013	0.098	0.249 ± 0.018	0.153	−0.086	−0.110	−0.062	<.001	<.001	<.001
Parotid_R	Median dose	0.140 ± 0.015	0.061	0.222 ± 0.018	0.164	−0.083	−0.104	−0.062	<.001	<.001	<.001
SpinalCord_PRV05	Maximum dose	0.157 ± 0.011	0.112	0.179 ± 0.012	0.124	−0.023	−0.036	−0.009	<.001	<.001	.007
Cavity_Oral	Median dose	0.147 ± 0.011	0.107	0.205 ± 0.012	0.163	−0.058	−0.076	−0.040	<.001	<.001	<.001

Table 2. Dosimetric Result Probabilities (DRPs) for Left and Right Parotid Before and After Replanning.

With Similar Plan Selection				Without Similar Plan Selection			
Parotid_L		Parotid_R		Parotid_L		Parotid_R	
Before Replanning	After Replanning	Before Replanning	After Replanning	Before Replanning	After Replanning	Before Replanning	After Replanning
1.000	0.595	1.000	0.883	0.921	0.319	0.999	0.561
0.958	0.364	0.989	0.472	0.739	0.254	0.768	0.306
0.983	0.338	0.972	0.840	0.164	0.082	0.995	0.940
1.000	0.950	0.996	0.859	0.917	0.821	0.985	0.909
1.000	0.901	0.941	0.892	0.722	0.564	0.905	0.869
0.910	0.590	0.929	0.942	0.973	0.869	0.669	0.694
1.000	0.672	1.000	0.764	0.368	0.115	1.000	1.000
0.998	0.997	0.977	0.983	0.762	0.753	1.000	1.000
0.972	0.690	0.934	0.802	0.719	0.438	0.798	0.662
1.000	0.404	0.973	0.936	0.380	0.091	1.000	0.999
0.989	0.940	0.941	0.468	0.864	0.785	0.701	0.331
1.000	1.000	1.000	1.000	0.976	0.990	0.993	0.997

value of all the dosimetric results of the selected similar plans' DDIs are smaller than the values calculated from the reference plans pool.

Plan Quality Assurance Validation

Twelve inferior plans which have DRPs of both parotids larger than 0.9 with similar plans selection are found in the exported plan database. The DRPs with and without similar plans selection before replanning for both left and right parotid are shown in Table 2. All the DRPs with similar plans selection before replanning have DRPs larger than 0.9, which agree with the inferior plan selection criteria. On the contrary, only 2 of the selected inferior plans show DRPs larger than 0.9 for both parotids without similar plans selection. After replanning, 8 of 12 (shaded in blue in Table 2) treatment plans have the DRPs of at least one parotid less than 0.8 with similar plans selection. Only 2 (shaded in orange in Table 2) treatment plans have one DRP after replanning larger than or the same as before but all within 2%, and one of them have the DRP of the other parotid significantly smaller after replanning. On average, left and right parotid median dose are reduced by 31.7% and 18.2%, respectively, and the DRPs are reduced by 28.7% and 15.3%. The difference in the dose reduction of left parotid versus right parotid is due to the relatively small sample size of suboptimal plans. There are plans that should be spared bilaterally and was only spared for single side during initial treatment. By chance, more left parotids should have been spared. Similarly, most of the DRPs for both parotids without similar plans selection are smaller after replanning.

Retrospective Plan Quality Analysis

All the target plans DRPs are averaged every 2 years in order to discover the year-based change of plan quality. The DRPs are calculated with both relative doses normalized the primary prescription doses and absolute dose. The results for all the 8 kinds of dosimetric results are shown in Figure 4. The error bar for each point indicates the standard error of the mean DRP results. The curves for the 2 kinds of DRPs (normalized dose

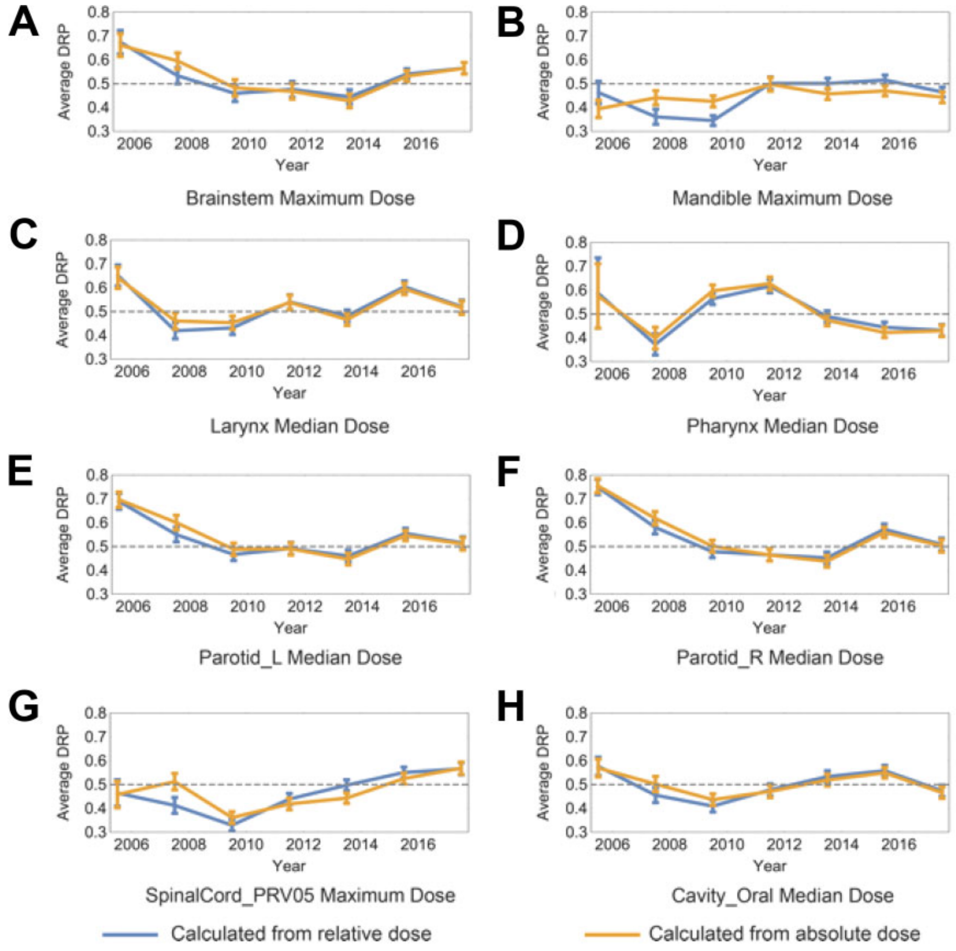

Figure 4. Trends of plan quality metric (PQM) DRP over years. Each plot contains average DRP of the corresponding dose–volume point for each OAR within each 2 years ranging from 2005 and 2018. The blue curve represents the DRP calculated from relative dose normalized by the primary prescription dose and the yellow curve presents the DRP calculated from absolute dose. The first 4 years starting from 2005 shows a distinctive decrease of average DRP on both curves for most of the OARs. Distinctive decreasing trends of dose to both parotids can be observed from (E) and (F), especially during the first 4 years.

or absolute dose) are close and share the same trend throughout the years besides mandible. For both kinds of DRPs, there is a trend of decreasing average DRP for most of the OARs within the first 4 years from year 2005, and it is strongly correlated with the monotonic increasing averaged primary prescription dose and averaged prescription dose ratio during the 4 years shown in Figure 5A and B. The average DRPs are relatively stable and fluctuating around 0.5 after year 2010 without any obvious trend of change for most structures. There is an overall increasing trend of average number of treatment plans that contain each OAR demonstrated in Figure 5C.

Discussion

Both DRP and DDI are defined under similar plans' statistics. The impact of different planning difficulties and varying physicians' preferences are effectively minimized by referencing similar plans instead of the whole plan population. The DRP measures the relative plan quality of the current plan in the context of historical plans. Intuitively, the DRP value of a plan informs the planner the fraction of historical plans the current plan surpasses in terms of OAR sparing. The DDI measures the difference between the dosimetric result of the target plan and that of the averaged historical plans. Applying similar plans selection will have minimum effect in the PQMs with the statistical inference of historical plans if the number of historical plans is limited. In this case, treatment plans with minimum plan similarity scores may not be similar enough to have a dosimetric result distribution exhibiting the past knowledge. In other words, there is not enough knowledge to be gathered if the historical plans database does not contain sufficient quantities of plans. Therefore, the PQMs that depend on similar historical plans may not be applicable when the treatment plan data for a specific treatment site and modality are limited. However, the database, which only consists of anonymized contours and dosimetric parameters, can be potentially transferred and shared for precise and robust plan quality assurance and plan guidance.

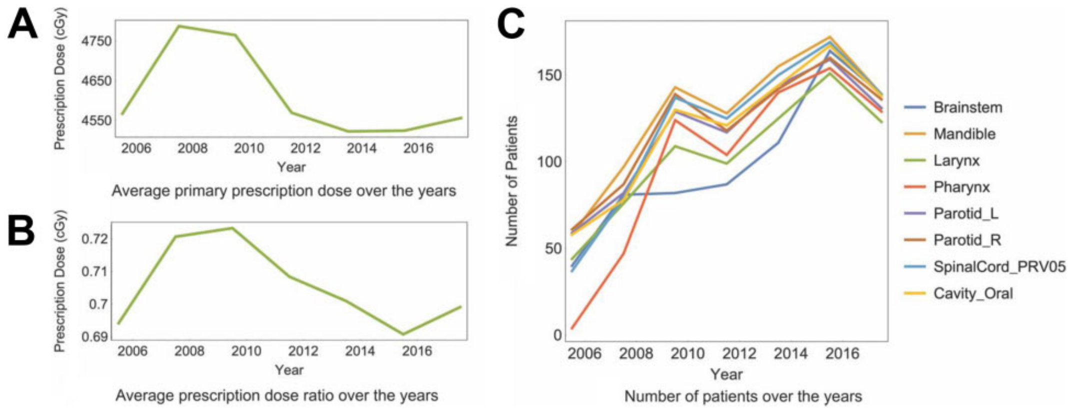

Figure 5. A, Primary prescription dose averaged within every 2 years. B, Prescription dose ratios (D_{pri}/D_{bst}) averaged within every 2 years. The distinctive decrease of average DRP for the first 4 years from 2005 in Figure 4 correlates with the increasing average prescription dose ratio within the same period of time. C, Numbers of plans that contain each OAR averaged within each 2 years.

After similar plan retrieval validation, all structures except brain stem show statistically significant smaller DDIs with similar plans selection for test target plans. The standard deviation of the dosimetric result distribution of the selected similar plans is statistically smaller than that of the reference plans pool for all the analyzed structures. Hence, similar plans have closer dosimetric results than the reference plans pool and their distributions are more statistically significant. Since both arguments are validated, it can be concluded that the application of plan similarity metric calculation and similar plans selection in the calculation of 2 PQMs DDI and DRP for high-quality treatment plans yields a more accurate estimation of patient-specific treatment plan quality.

The absolute DDI of the target plan with similar plans selection can be close to the value calculated from all the reference plans pool. It mostly happens when the target plan has geometries that are similar to most plans in the historical database, resulting in very close mean dosimetric result of similar plans to all the plans in the reference plans pool. It can also be contributed by the lack of similarities of the selected similar plans. In other words, the proposed plan similarity metric fails for those cases. What we discovered is that most of those target plans (plans being evaluated) are ipsilateral plans where one parotid can have overlap with one PTV and the other parotid is far away from it. Other OAR structures which have very small gDTH square differences may compensate for the large differences of one parotid structure and result in overall small plan similarity score for bilateral cases. Those bilateral plans have significantly different dosimetric results of one parotid compared with ipsilateral cases, and they should not be used as reference plans for the ipsilateral target plan. Reducing the cutoff threshold of gDTH matrixes may improve the dosimetric similarities by emphasizing the geometric differences of the OAR volumes that are closer to PTV surfaces. The design of weighting factors for gDTH square differences needs to be improved for more balanced geometric difference contributions of different OARs. The small and positive DDI differences can also be contributed by the unique anatomy of the target plan which has limited truly similar plans in the database. This problem can be solved by expanding the database that can cover most of the PTV anatomy scenarios in clinic. The small DDI differences can also stem from the variations of the OAR's axial position relative to the PTVs. It is especially obvious for brain stem which shows no statistical difference in DDIs calculated with and without similar plan selection. There is a minimum portion of brain stem volume being on the same axial plan with either PTV. Since the overlap volume could be irradiated directly by beams if coplanar technique is applied, the dose to brain stem can be easily reduced. Therefore, the variation of brain stem position, shape, and volume has minimum effect in dose sparing. The inaccurate estimation of DDI can also be contributed from other factors such as the error from kernel density estimation and outlier identification. The accuracy of the *Z*-score-based dosimetric outlier identification method applied in our study requires further assessment. A recent study from Sheng *et al* demonstrate the workflow in determining dosimetric outliers and geometric outliers for prostate cases.[25] Geometric outliers for those head and neck cases could be developed based on Yang's research in future studies.

After plan quality assurance validation, most of the selected inferior treatment plans show reduced median dose to either one of the parotids by having smaller DRPs after the replanning. Hence, most of the selected inferior plans are indeed suboptimal in parotid dose reduction and applying the proposed dose metric has a high true positive rate in detecting inferior plans by customized criteria. In comparison, only 3 of the selected inferior plans would be detected based on our criteria if DRPs without similar plans selection are used. Therefore, the proposed PQM DRP with similar plans selection gives a more accurate score for inferior plans and is effective in picking out inferior plans to assure optimal plan quality in clinical practice.

The retrospective plan quality analysis shows that dose sparing of both parotids have improved significantly over the years (Figure 4). In detail, DRPs calculated from normalized dose for both parotids (Figure 4E and F) have reduced by 21.7% from year 2006 to 2007 to year 2016 to 2017 on average. The DRP calculated from absolute dose shows similar reduction (23.8%) over those years. Since the average primary prescription dose (Figure 5A) does not change too much between those 2 timestamps, dose normalization can be ruled out from the main contributors to the DRP reduction. Instead, factors such as stricter dose constraint or improved planning expertise may play a more important role in the observed parotid dose sparing improvident. A decrease of the DRP averaged every 2 years for most OARs can also be observed within the first 4 years from the year 2005. It strongly correlates with the monotonic increase in the average prescription dose ratio (the most common dose ratio D_{pri}/D_{bst} changes from 44/70 to 50/60). This results in a smaller DRP calculated from similar plans that the selected OARs have overlap with the boost PTV and the prescription dose ratios are smaller than the target plan. Therefore, having an increasing average prescription dose ratio but limited number of plans (Figure 5C) for the first 4 years and a minimum average prescription dose ratio for the rest years cause the decreasing average DRP for the first 4 years. The increasing primary prescription dose during the first 4 years makes the change of prescription dose ratio the main contributing factor to the decrease of averaged DRP. This demonstrates that DRP can not only evaluate patient-specific treatment plan quality on a specific dose–volume point but also reveals the collective change of planning conventions such as prescription dose ratio in retrospective plan quality analysis.

One challenge we face when applying the quality analysis tool to a large data set is to properly homogenize the data to reduce variability without losing information. The cohort consists of plans made under different conditions. In our institution, it is still the standard of practice to treat with IMRT technique. The change of delivery technique is not considered as a variable for plan quality. Factors such as machine change, contours, and margins do not change the relationship between planning difficulty and geometries. They can be separated from beam design and independently analyzed and controlled. Evolution of optimization and dose calculation algorithm haven't been homogenized in our study due to the difficulty in quantification. More studies regarding their contributions will be performed in the future. Other variances such as physicians dose goals and target subsites are assumed to be intrinsic to the data set since they are correlated with patient anatomy and prescriptions. Similar plan selection aims to reduce the cohort variation implicitly using those 2 features. For instance, treatment plans of different subsites have different dose distributions. We did not explicitly divide the data set into smaller patches based on subsites. Instead, we utilized our anatomical feature (gDTH) to identify cases with similar expected dose distributions. It is expected that cases from the same subsites will be more likely to be referenced.

Conclusion

Our proposed PQMs DDI and DRP are objective measurements of dose sparing optimality of each individual OAR structure. They are acquired from statistical analysis of similar historical treatment plans selected by the plan similarity metric that considers both geometric and dosimetric differences. The proposed plan quality inference method has been shown to yield more accurate estimations of plan optimality than their counterparts which do not consider patient geometry variability. The retrospective analysis using the proposed metrics not only demonstrates the systematic change of dose sparing convention of one OAR but also reveals changes in other planning conventions such as prescription dose ratio to multiple PTVs.

Authors' Note

This study did not require an ethical board approval because it did not contain human or animal trials.

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 work is partially supported by an NIH grant (#R01CA201212) and a master research grant by Varian Medical Systems.

ORCID iD

Jiang Zhang, MS https://orcid.org/0000-0001-5807-1686
Q. Jackie Wu, PhD https://orcid.org/0000-0001-8235-2322
Chunhao Wang, PhD https://orcid.org/0000-0002-6945-7119
Jiahan Zhang, PhD https://orcid.org/0000-0002-4288-6503

References

1. 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.
2. Mayo CS, Yao J, Eisbruch A, et al. Incorporating big data into treatment plan evaluation: development of statistical DVH metrics and visualization dashboards. *Adv Radiat Oncol*. 2017;2(3):503-514.
3. Zhu X, Ge Y, Li T, et al. A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Medical Physics*. 2011;38(2):719-726.
4. Yuan L, Ge Y, Lee WR, et al. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. *Medical Physics*. 2012;39(11):6868-6878.
5. Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore K. Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Medical Physics*. 2012;39(12):7446-7461.
6. Zhang J, Wu QJ, Xie T, Sheng Y, Yin FF, Ge Y. An ensemble approach to knowledge-based intensity-modulated radiation therapy planning. *Front Oncol*. 2018;8(57). doi: 10.3389/fonc.2018.00057

7. Wu B, Ricchetti F, Sanguineti G, et al. Data-driven approach to generating achievable dose–volume histogram objectives in intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys*. 2011;79(4):1241-1247.
8. Janssen TM, Kusters M, Wang Y, et al. Independent knowledge-based treatment planning QA to audit Pinnacle autoplanning. *Radiother Oncol*. 2019;133:198-204.
9. 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.
10. Good D, Lo J, Lee WR, Wu QJ, Yin FF, Das SK. 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*. 2013;87(1):176-181.
11. Berry SL, Ma R, Boczkowski A, Jackson A, Zhang P, Hunt M. Evaluating inter-campus plan consistency using a knowledge based planning model. *Radiother Oncol*. 2016;120(2):349-355.
12. Chang ATY, Hung AWM, Cheung FWK, et al. Comparison of planning quality and efficiency between conventional and knowledge-based algorithms in nasopharyngeal cancer patients using intensity modulated radiation therapy. *Int J Radiat Oncol Biol Phys*. 2016;95(3):981-990.
13. Delaney AR, Tol JP, Dahele M, Cuijpers J, Slotman BJ, Verbakel WF. Effect of dosimetric outliers on the performance of a commercial knowledge-based planning solution. *Int J Radiat Oncol Biol Phys*. 2016;94(3):469-477.
14. Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WF. Evaluation of a knowledge-based planning solution for head and neck cancer. *Int J Radiat Oncol Biol Phys*. 2015;91(3):612-620.
15. Tol JP, Dahele M, Delaney AR, Slotman BJ, Verbakel WF. Can knowledge-based DVH predictions be used for automated, individualized quality assurance of radiotherapy treatment plans? *Radiat Oncol*. 2015;10:234.
16. Fogliata A, Nicolini G, Clivio A, et al. A broad scope knowledge based model for optimization of VMAT in esophageal cancer: validation and assessment of plan quality among different treatment centers. *Radiat Oncol*. 2015;10:220.
17. Fogliata A, Belosi F, Clivio A, et al. 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*. 2014;113(3):385-391.
18. Wu H, Jiang F, Yue H, Zhang H, Wang K, Zhang Y. Applying a RapidPlan model trained on a technique and orientation to another: a feasibility and dosimetric evaluation. *Radiat Oncol*. 2016;11(1):108.
19. Hussein M, South CP, Barry MA, et al. Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy. *Radiother Oncol*. 2016;120(3):473-479.
20. Garofalo R. *Building Enterprise Applications with Windows Presentation Foundation and the Model View ViewModel Pattern*. United States: Microsoft Press; 2011:224.
21. Mayo CS, Moran JM, Bosch W, et al. American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. *Int J Radiat Oncol Biol Phys*. 2018;100(4):1057-1066.
22. 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.
23. Silverman BW. *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. Boca Raton, FL: Chapman & Hall/CRC; 1998:ix, 175.
24. Dickhaus T. *Theory of Nonparametric Tests*. New York, NY: Springer Berlin Heidelberg; 2018:128.
25. Sheng Y, Ge Y, Yuan L, Li T, Yin FF, Wu QJ. Outlier identification in radiation therapy knowledge-based planning: a study of pelvic cases. *Med Phys*. 2017;44(11):5617-5626.