Ready. Set. 2.0!

The NEW SunCHECK™

INDEPENDENT QA.
YOUR WAY.

"Having patient information, having machine information all in one easily accessible place... ***that's basically what you would want in a system.***"

Robert Staton, Ph.D., UF Health Cancer Center at Orlando Health, Orlando, FL

Learn more at:
sunnuclear.com

Platform

Patient

Machine

SUN NUCLEAR corporation

Efficiently train and validate a RapidPlan model through APQM scoring

Marco Fusella,[a),*] Alessandro Scaggion,[*] Nicola Pivato, Marco Andrea Rossato, Alessandra Zorz, and Marta Paiusco
Medical Physics Department, Veneto Institute of Oncology IOV-IRCCS, Padova 35128, Italy

(Received 30 November 2017; revised 20 March 2018; accepted for publication 21 March 2018; published xx xxxx xxxx)

Purpose: The aim of this study was to propose and validate an intuitive method for training and to validate knowledge-based planning (KBP) systems based on a patient-specific plan quality scoring.
Methods: A sample of 80 clinical plans of prostate cancer patients were ranked on the basis of the Adjusted Plan Quality Metric (APQM%). This quality metric was computed normalizing the Plan Quality Metric (PQM%) score to the best possible OAR sparing estimated by the Feasibility DVH (FDVH) algorithm. Two different plan libraries were created, purging all the plans below the first quartile or below the median the APQM% distribution. These libraries were used to populate and train two RapidPlan models: respectively, the $APMQ_{25\%}$ *and the* $APMQ_{50\%}$ models. No further refinements or actions were undertaken on these two models. Their performances were benchmarked against another two RapidPlan models. An *Uncleaned* model, which was populated and trained with the initial sample of 80 plans, and a *Cleaned* model, obtained through the standard iterative cleaning and refinement process suggested by the vendor and in literature. The outcomes of a planning test based on 20 patients within the training library (closed loop) and 20 patients outside of the training library (open-loop) were compared through various DVH metrics and the PQM% score.
Results: The selection through APQM% thresholding roughly preserves the geometric variety of the *Cleaned* model; only the $APMQ_{50\%}$ model showed a modest broadness reduction. The models generated through APQM% thresholding showed target coverage and OARs sparing equal or superior to the *Uncleaned* and *Cleaned* models both for the closed- and the open-loop tests. No significant differences were found between the four models. PQM% analysis ranked the overall plan quality as: $86.5 \pm 6.5\%$ $APQM_{50\%}$, $83.1 \pm 5.9\%$ $APQM_{25\%}$, $80.39 \pm 10.6\%$ *Cleaned* and $79.4 \pm 8.5\%$ *Uncleaned* in the closed-loop test; $84.9 \pm 7.6\%$ $APQM_{50\%}$, $82.6 \pm 7.9\%$ $APQM_{25\%}$, $80.39 \pm 10.6\%$ *Cleaned* and $79.4 \pm 8.5\%$ *Uncleaned* in the open-loop test.
Conclusions: Forward feeding a RapidPlan model through a thresholding selection based on APQM% is proven to produce equal or better results than a model based on a manually and iteratively refined population. A tighter APQM% threshold turns approximately into a higher average quality of plans generated with RapidPlan. A trade-off must be found between the mean quality of the KBP library and its numerosity. The proposed KBP feeding method helps the KBP user, because it makes the model refinement more intuitive and less time consuming. [https://doi.org/10.1002/mp.12896]

Key words: knowledge-based planning, modeling, optimization, quality metric, treatment planning

1. INTRODUCTION

The interest in automated or knowledge-based solutions to radiation treatment planning is rapidly growing because of their capability to improve both the plan consistency and the planning efficiency. RapidPlan, the commercial knowledge-based planning (KBP) solution by Varian Medical Systems (Palo Alto, CA), uses regression analysis to build a model which correlates the geometric relationships between OARs and PTV with the dose-volume histogram of a library of clinical plans.[1] For each new patient, the model predicts a range of achievable OAR dose-volume histograms and a set of dose-volume objectives that are used to drive the inverse planning process. The effectiveness of the proposed optimization objectives is only as good as the quality of the training set.[1] Therefore, a carefully selected group of consistent high quality treatment plans is recommended to feed RapidPlan, together with a labor intensive validation process.[2–4] Indeed, RapidPlan model has to be reviewed, refined and validated through a complex and iterative process that requires great effort by the user.[2–5] Despite the promising results reported for RapidPlan,[2–4,6] its implementation may turn out to be time consuming.

This hurdle has already been tackled by Li et al. who proposed a compelling solution based on script-driven automated planning.[7] However, small centers can hardly rely on such an approach, so a solution based on an easier and more accessible tool would be beneficial to any KBP user.

In 2012, Nelms et al. introduced the "Plan Quality Metric" (PQM) scoring which, through a list of metrics with specific weightings or scores, enables plans to be given an overall quality score for ease of comparison and validation.[8] More recently, Ahmed et al. proposed a tool to estimate a priori the possible OAR sparing: termed as Feasibility DVH (FDVH).[9]

The FDVH measures the difficulty of achieving some DVH objectives considering the patient-specific anatomical challenges and ensuring the complete target coverage to full dose. PlanIQ (Sun Nuclear Corp., Melbourne, FL) implements both of these tools and combines them in a new metric called Adjusted PQM (APQM). It is a plan quality score customized to the patient-specific feasibility analysis and is therefore consistent across all patients. This property makes APQM scoring a complementary tool for the evaluation and refinement of RapidPlan models.[9,10] In fact, unlike PQM, APQM allows to adequately evaluate large OARs-PTV overlap or particularly unfavorable OARS geometry and consequently tailor the judgment of quality. Such a possibility levels out the individual planning hurdles and serves as a patient-specific benchmark to meaningfully compare the quality of hundreds of plans.

This study analyzes APQM as a tool to expedite and facilitate the configuration and refinement of a RapidPlan model. APQM patient-specific quality score was used to select plans to build a RapidPlan model in a feed-forward process without any further refinement. VMAT plans for prostate cancer were considered. The performance of the obtained model was benchmarked against the performance of a RapidPlan model populated, refined and validated according to the standard iterative methodology proposed in literature.

2. MATERIALS AND METHODS

2.A. Patient sample

One hundred patients treated for radical prostate cancer, between 2015 and 2016, were collected from the database of our Institute. They were all manually planned and treated with Volumetric Modulated Arc Therapy using 1 or 2 full arcs and 6-MV photons. All of the treatments were planned with Eclipse and progressive resolution optimizer (PRO) v. 11 to deliver 78 Gy or 70 Gy (PTV T) over 39 or 28 fractions. The planning goals were to deliver 95% of the prescribed dose to 100% of the PTV, limiting the overdosage to 110% of the prescribed dose. All plans were optimized with the sparing of rectum, bladder, and femoral heads according to our department prostate radical treatment protocol, which is based upon RTOG 0126. Table I reports a descriptive statistic of the dosimetric results of the population. Eighty of these patients were used for model building and closed-loop validation, while the remaining twenty patients were used only for the open-loop validation phase. Closed-loop validation, that is, replanning with RapidPlan treatments that were planned manually and were included in the training library, proves the self-consistency of the model. Open-loop validation, that is, replanning with RapidPlan treatments that were planned manually but were not included in the training library, allows to test the capability and the efficacy of the model to predict the planning outcome of patients unknown to the model itself.

2.B. Plan quality scoring

The "Plan Quality Metric" (PQM) scoring, introduced by Nelms et al.,[8] is a user-defined metric intended to quantify and compare the quality of treatment plans by mimicking the judgment of a clinical team. PQM% quantifies the overall quality achieved by a treatment plan in terms of adherence to a list of planning objectives/endpoints. Indeed, it is built with a collection of sub-components which represents a set of clear and specific treatment plan objectives set by clinicians (DVH points, conformity indices, etc.). Each sub-component is associated with a function that should mathematically describe the clinician judgment criterion. This function translates the achieved value of each submetric into a numerical score. The sum over of submetric scores divided by the total maximum achievable constitutes the composite PQM%. A detailed description of metrics and score functions is given in table I and in fig. 2 in Ref. [8].

The "Feasibility DVH" (FDVH) tool uses the CT images and DICOM RT structure set of the patient to generate a fictitious dose distribution based on first principle assumptions and a series of energy-specific dose-spread calculations.[9,10] This 3D dose distribution is ideal and is built intentionally unachievable, such that each PTV is evenly painted with the prescription dose (the DVH of each PTV will therefore be a perfect corner). A high dose gradient and moderate dose periphery is then added to the PTV dose cloud. Once the dose cloud is generated, for each individually considered OAR the lower possible boundary of its DVH is predicted.[9,10] A detailed explanation of the algorithm can be found in Ref. [9].

The commercial PlanIQ software v2.1 implements both the PQM formalism and FDVH estimation algorithm and integrates them into the so-called Adjusted Plan Quality Metric (APQM). Once the FDVH is used to estimate the feasibility of attaining each treatment plan objective, the APQM% is obtained by normalizing the result of each PQM% submetric to the ideal 3D dose distribution corresponding result. This process takes into consideration the unique challenges of specific patient anatomy when comparing plan quality.[10] To summarize, PQM% measures the clinical acceptability of a treatment plan on the basis of population-based approved standards, while APQM% measures how close a treatment plan is to the best possible achievable result for a particular patient. Moreover, the use of APQM% score can help the planner to understand whether low absolute plan quality metric scores (PQM%) are due, at least in part, to the specific anatomy challenges that cannot be fulfilled, while ensuring target coverage.

In this study, APQM% was used to rank and identify the best possible plan candidates to feed the KBP algorithm, while PQM% was used to compare different plans for the same patient.

The PQM algorithm as introduced in Ref. [8] was herein adopted since its definition closely matches the clinical endpoints used in our clinic, with the only exception of a unique PTV.

TABLE I. Dosimetric endpoints of prostate treatment plans. Number of plans meeting the endpoints, mean ± 1 standard deviation and existence range across the entire sample.

Objectives		% cases below contstraint	Mean ± 1 SD	[min;max]
PTV	$V_{100\%}$ >= 95%	73%	95.3 ± 0.5%	[93.9%;96.6%]
	$D_{99\%}$ > 95%	100%	98.5 ± 0.7%	[95.2%;99.3%]
	$D_{2\%}$ < 107%	95%	106.1 ± 1.0%	[102.6%;108.9%]
Bladder	V_{40Gy} <= 40%	73%	34.2 ± 8.0%	[6.2%;64.4%]
	V_{65Gy} <= 25%	98%	9.5 ± 6.6%	[0.8%;37.0%]
	V_{75Gy} <= 10 cc	63%	10.09 ± 1.55	[0.17;10.28]
Rectum	V_{40Gy} <= 40%	75%	33.3 ± 16.2%	[9.6%;94.4%]
	V_{65Gy} <= 20%	95%	14.1 ± 8.8%	[2.2%;56.6%]
	V_{75Gy} <= 10 cc	69%	9.7 ± 0.9	[0.0;51.8]
Femoral head L	V_{30Gy} <= 50%	95%	31.6 ± 7.9%	[6.9%;52.7%]
	D_{1cc} <= 45 Gy	100%	2.4 ± 6.0%	[0.0%;31.9%]
Femoral head R	V_{30Gy} <= 50%	98%	2.8 ± 4.6%	[0.0%;23.4%]
	D_{1cc} <= 45 Gy	100%	2.0 ± 5.5%	[0.0%;35.9%]

2.C. Model building and refinement

The APQM% score was computed for each single plan of the entire sample. Afterwards, the eighty plans selected for the model building phase were ranked from the lowest to the highest APQM% value. Two RapidPlan models were populated purging, respectively, all the plans falling below the first quartile ($APQM_{25\%}$) and the median ($APQM_{50\%}$) of the APQM% distribution. Once populated, the two models were trained without any further refinement and actions, so that model generation consisted in a pure feed-forward process. The choice of the two threshold guarantees, at least, the minimally acceptable number of plans for model building, as suggested in the literature and in the RapidPlan user guide.[1,11,12]

The same eighty plans were used to populate and train two models to be used as benchmark. The population and first training process led to an initial RapidPlan model, referred herein as *Uncleaned.* According to literature and vendor recommendations, this model was subsequently refined and cleaned through an iterative process: influential data points were individually examined and judged, geometric outliers were removed, while dosimetric outliers were replanned only in case the model predictions were largely better than clinical plans.[1,13,14] The outliers' classification was based on the statistical metrics given by RapidPlan at the end of the model training phase.[1,2] This procedure led to the *Cleaned* model.

All the four RapidPlan models were configured to generate the list of optimization objectives given in Table II. This list closely match the optimization protocol used in the clinical practice with which the plans of the entire sample were firstly manually optimized. It is worth noting that no this set of objectives did not underwent any refinement procedure as suggested in literature.[4] Such a refinement was out of the scope of this work that focus on the selection and refinement of the population of plans composing the model's library.

2.D. Model validation and comparison

Model validation and comparison were performed by means of closed-loop and open-loop tests.[2,3,12] The closed-loop test was performed on 20 patients chosen randomly between those common to all the four training groups, while the open-loop test was conducted on the 20 patients left for model validation phase and not included in any of the model libraries. Table III shows the samples of patients used for the validation procedure and the population used to build the initial *Uncleaned* model in terms of absolute volume and APQM% distributions. A two-sided *t* test was used to compare the populations.

For each patient four different plans were obtained feeding the Eclipse optimizer (PO v13.7) with the predicted RapidPlan constraints generated by each model. To compare the performance of the models without any bias, a single automatic optimization process without human intervention was performed; the intermediate and final dose calculations were done with Acuros-XB v13.7 algorithm as explained in Ref. [14].

It must be noted that this decision may have led to an unfavorable comparison between clinical and RapidPlan-generated plans. In fact, a certain amount of skilled manual interventions is needed to achieve high quality results even when RapidPlan-generated objectives drive the optimization.[3,4,6,13,15] Nevertheless, the need of an unbiased and relative comparison between the performances of four RapidPlan models justifies the methodology.

The dosimetric features of the four models were compared on the basis of the following DVH metrics: (a) Dose to the 95% of the PTV volume ($D_{95\%}$), the near minimum dose ($D_{98\%}$), the near maximum dose ($D_{2\%}$); (b) V_{40Gy}, V_{65Gy}, V_{70Gy} and Mean Dose for rectum and bladder; (c) the Mean Dose and the D_{1cc} of the femoral heads. The former metrics were complemented by the Homogeneity index $[(D_{2\%} - D_{98\%})/D_{PRESC}]$ and the Conformation Number

TABLE II. Summary of the optimization objectives generated by RapidPlan. The *gen.* indicates those values generated by RapidPlan on the basis of the trained prostate model. D_{presc} indicates the prescription dose.

	Optimization objective			
ROI	Objective type	D (Gy)	V (%)	Weight
PTV	Lower	0.99 D_{presc}	100	130
	Upper	1.02 D_{presc}	0	120
Rectum	Upper	*gen.*	0, 10, 30, 50, 80	*gen.*
Bladder	Upper	*gen.*	0, 10, 30, 60	*gen.*
Femoral head L	Upper	gen.	0, 50	*gen.*
Femoral head R	Lower	*gen.*	0, 50	*gen.*
Body	Normal tissue objective	DistanceFrom TargetBorder = 0.2 cm StartDose = 100 EndDose = 50 FallOff = 0.2 cm^{-1}		100

TABLE III. Distribution of structure volumes measured in cubic centimeters and APQM% score of clinical plans. Closed- and open-loop test samples are compared to the *Uncleaned* model library population. Reported *P*-values refer to a two-sided *t* test.

Structures volume in cm^3 and APQM% score	*Uncleaned* model		Closed-loop test			Open-loop test		
	Mean ± 1 SD	[min;max]	Mean ± 1 SD	[min;max]	*P*-value	Mean ± 1 SD	[min;max]	*P*-value
PTV	137.7 ± 48.3	[56.0;352.6]	130.5 ± 29.6	[98.4;192.8]	0.401	138.8 ± 30.3	[102.0;195.1]	0.899
Rectum	52.1 ± 23.8	[17.3;126.7]	54.8 ± 27.6	[22.2;100.4]	0.689	53.0 ± 14.4	[39.5;86.3]	0.830
Bladder	291.8 ± 148.5	[56.9;624.6]	331.8 ± 169.1	[124.9;624.6]	0.335	250.6 ± 98.6	[124.5;411.5]	0.139
Femoral head L	179.9 ± 39.8	[101.1;336.6]	175.4 ± 21.0	[126.5;192.5]	0.488	174.6 ± 33.4	[78.4;206.0]	0.544
Femoral head R	182.7 ± 39.9	[126.3;334.8]	174.9 ± 20.8	[132.9;204.2]	0.229	179.7 ± 35.3	[76.2;213.8]	0.741
APQM%	84.9 ± 8.5	[55.1;96.6]	88.2 ± 5.8	[81.2;96.6]	0.098	89.2 ± 5.8	[79.1;99.5]	0.034

$$CN = \frac{TV_{RI}}{TV} \times \frac{TV_{RI}}{V_{RI}}$$

where TV_{RI} indicates the target volume covered by the reference isodose, TV indicates the target volume and V_{RI} indicates the volume of the reference isodose.[16] The overall comparison between the four models was performed through the PQM% score computed for each plan included into the closed- and open-loop tests.

2.5. Statistical analysis

Volume distributions and DVH metrics were compared through a two-sided *t* test with a significance level of 0.05. PQM% values were compared through a Wilcoxon signed rank test, which compares medians of non-normal distributions, with a significance level of 0.05.

3. RESULTS

3.A. Model training statistics

The refinement and cleaning procedure of the *Uncleaned* model led to the removal of 11 plans because they were classified as geometric outliers or influential data points. No dosimetric outliers were identified. As a result, *Cleaned* model was populated with 69 plans. The plan selections based on the APQM% threshold led to 60 and 40 plans library for $APQM_{25\%}$ and $APQM_{50\%}$ models. Table IV presents some of the training quality metrics given in the RapidPlan system report along with the volume distribution of the different samples. R^2 represents the coefficient of determination of regression model parameters, χ^2 is the average chi square of regression model parameters and MSE is the mean squared error between original and estimate values. R^2 and χ^2 measure the goodness-of-fit of the OAR model. MSE describes how well the model is able to estimate the original DVH in a training plan and the closer it is to 0 the better the estimation capability of the model for plans that are not part of the training set. For the library of each model, a two-tailed *t* test compares if the volume distributions of the considered structures were significantly different from the library of the *Uncleaned* model. In Fig. 1, the comparison of the overall plan quality of the trained models is reported as a whisker box-plot of the APQM%.

3.B. Closed-loop validation

Table V depicts the results of the closed-loop validation for the four trained models. In general, plans generated by

TABLE IV. Summary of the model training statistics for each trained OAR and each trained model. Volume distribution descriptive statistics and comparisons are also reported.

Structure	Model	R^2	χ^2	MSE	# structures	# potential outliers	Volume (cc) Mean ± 1sd	Volume (cc) [min;max]	*P*-value
PTV	Uncleaned				80		137.7 ± 48.3	[56.0;352.6]	
	Cleaned				69		132.1 ± 39.1	[65.6;258.7]	0.443
	$APQM_{25\%}$				60		128.3 ± 35.9	[56.0;218.2]	0.211
	$APQM_{50\%}$				40		124.2 ± 32.7	[56.0;193.8]	0.113
Bladder	Uncleaned	0.679	1.065	0.0015	80	6	291.8 ± 148.5	[56.9;624.6]	
	Cleaned	0.540	1.055	0.0014	69	0	292.2 ± 140.6	[90.6;624.6]	0.987
	$APQM_{25\%}$	0.706	1.083	0.0011	60	4	301.6 ± 152.2	[56.9;624.6]	0.704
	$APQM_{50\%}$	0.744	1.133	0.0017	40	4	302.6 ± 149.5	[56.9;624.6]	0.711
Rectum	Uncleaned	0.273	1.028	0.0078	80	3	52.1 ± 23.8	[17.3;126.7]	
	Cleaned	0.310	1.028	0.0088	69	0	51.5 ± 22.0	[17.3;122.0]	0.862
	$APQM_{25\%}$	0.365	1.087	0.0102	60	4	52.5 ± 25.1	[17.3;126.7]	0.921
	$APQM_{50\%}$	0.450	1.068	0.0090	40	1	52.9 ± 24.4	[17.3;122.8]	0.868
Femoral head L	Uncleaned	0.435	1.031	0.0079	80	3	179.9 ± 39.8	[101.1;336.6]	
	Cleaned	0.392	1.048	0.0076	69	2	179.9 ± 40.8	[101.1;336.6]	0.984
	$APQM_{25\%}$	0.448	1.052	0.0095	60	1	177.4 ± 38.7	[102.3;336.6]	0.719
	$APQM_{50\%}$	0.415	1.103	0.0096	40	9	177.0 ± 40.2	[102.3;336.6]	0.708
Femoral head R	Uncleaned	0.397	1.056	0.0049	80	6	182.7 ± 39.9	[126.3;334.8]	
	Cleaned	0.397	1.032	0.0053	69	2	182.4 ± 39.3	[126.3;334.8]	0.961
	$APQM_{25\%}$	0.388	1.083	0.0073	60	15	178.3 ± 37.3	[126.3;334.8]	0.504
	$APQM_{50\%}$	0.548	1.187	0.0141	40	11	176.9 ± 38.9	[128.1;334.8]	0.451

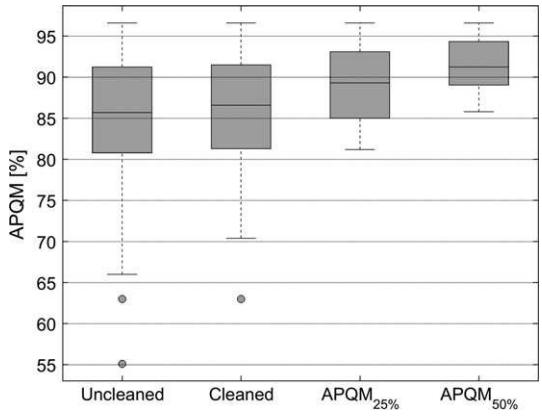

FIG. 1. Comparison of the APQM% distribution for the four trained models.

RapidPlan showed a more uniform and better covered PTV at the expense of a lower sparing of OARs with respect to the clinical plans. No difference can be noted in the target conformality. While the differences in PTV coverage were statistically significant, the lower OARs sparing were only seldom significant. Generally, ranking the models in terms of OAR sparing, $APQM_{50\%}$ was the closest to clinical plans while the others were ranked as $APQM_{25\%}$ > *Cleaned* > *Uncleaned*. The larger differences were seen in the low dose region of rectum (V_{40Gy}). Fig. 2 confronts, in terms of PQM%, the overall plan quality of the four trained models to the clinical plans. PQM% values confirmed the results seen with DVH metrics, even if no statistically significant differences were found. A detailed representation of the planning outcome of the closed-loop validation is reported in Fig. S1 in the supplementary material.

3.C. Open-loop validation

Results of the open-loop validation for the four trained models are reported in Table VI. The general observations made for the closed-loop test remained valid for the open-loop comparison. RapidPlan-generated plans showed a better covered and more homogeneous PTV at the cost of generally less spared OARs with reference to the clinical plans. In this case, the $APQM_{25\%}$ model showed also a better target conformance than the clinical plans. In terms of OAR sparing, $APQM_{50\%}$ equaled the results obtained by $APQM_{25\%}$ and *Cleaned* model and outperformed the *Uncleaned* model. The overall plan quality, compared in Fig. 3 by means of PQM%, confirmed the ranking. A detailed representation of the planning outcome of the open-loop validation is reported in Fig. S2 in the supplementary material.

It is worth noting that all the 20 patients selected to perform the open-loop test laid within the geometrical domain of three of the four trained models. In fact, four patients fell outside the domain of the $APQM_{50\%}$ model only: because of the OAR in two cases (large rectum, small bladder) and because of large PTV-OAR overlaps for the other two patients. When this happened, RapidPlan sent a warning message and discouraged the user to use its DVH predictions because they might be not trustworthy. To allow an unbiased comparison,

TABLE V. Closed-loop test. Comparison of DVH endpoint between the clinically approved plans and the RapidPlan-generated plans related to the four models. Each value is reported as mean value ±1 SD and is accompanied by the *P*-value of a paired *t* test against the clinical plans. Statistically significant comparisons are marked by "*".

Structure	Metric	Clinical	Unclean	Clean	$APQM_{25\%}$	$APQM_{50\%}$
PTV	$V_{95\%}$ (%)	99.2 ± 0.3	99.6 ± 0.2	99.5 ± 0.2	99.5 ± 0.2	99.5 ± 0.2
			0.002*	0.004*	0.005*	0.005*
	$D_{98\%}$ (%)	98.6 ± 0.3	99.1 ± 0.2	99.1 ± 0.2	99.2 ± 0.2	99.1 ± 0.3
			<0.001*	0.001*	0.002*	0.002*
	$D_{2\%}$ (%)	105.5 ± 0.9	105.4 ± 0.2	105.4 ± 0.2	105.4 ± 0.2	105.4 ± 0.4
			0.554	0.652	0.505	0.701
	HI	6.9 ± 1.0	6.2 ± 0.3	6.3 ± 0.3	6.3 ± 0.5	6.4 ± 0.7
			0.029*	0.047*	0.068	0.116
	CN	0.832 ± 0.026	0.830 ± 0.027	0.828 ± 0.027	0.832 ± 0.025	0.832 ± 0.022
			0.729	0.451	0.972	0.982
Rectum	V_{40Gy} (%)	26.0 ± 8.8	28.1 ± 10.4	27.9 ± 10.0	26.9 ± 10.0	24.9 ± 9.4
			0.397	0.431	0.695	0.641
	V_{65Gy} (%)	6.9 ± 3.4	7.6 ± 4.5	7.8 ± 4.5	7.5 ± 4.4	7.1 ± 4.2
			0.128	0.085	0.15	0.563
	V_{75Gy} (cc)	2.8 ± 2.2	3.1 ± 2.7	3.2 ± 2.8	3.1 ± 2.7	3.0 ± 2.6
			0.098	0.046*	0.076	0.252
Bladder	V_{40Gy} (%)	14.2 ± 8.0	14.8 ± 8.3	15.1 ± 9.5	14.8 ± 8.9	14.7 ± 8.2
			0.308	0.241	0.275	0.33
	V_{65Gy} (%)	5.9 ± 3.5	5.9 ± 3.6	6.0 ± 3.8	5.9 ± 3.6	6.0 ± 3.6
			0.412	0.238	0.76	0.307
	V_{75Gy} (cc)	12.3 ± 4.6	12.4 ± 4.4	12.3 ± 4.3	12.3 ± 4.5	12.4 ± 4.4
			0.822	0.992	0.947	0.885
Femoral head L	D_{1cc} (Gy)	23.75 ± 7.28	22.28 ± 4.81	21.66 ± 5.56	22.36 ± 4.34	21.09 ± 3.93
			0.475	0.322	0.594	0.305
	Mean dose (Gy)	10.58 ± 4.31	10.00 ± 2.53	10.01 ± 2.87	9.89 ± 2.44	9.54 ± 2.30
			0.581	0.574	0.531	0.342
Femoral head R	D_{1cc} (Gy)	24.22 ± 3.42	22.70 ± 3.32	22.80 ± 3.68	22.94 ± 3.55	21.64 ± 2.83
			0.191	0.236	0.298	0.035*
	Mean dose (Gy)	10.82 ± 2.73	10.65 ± 2.00	10.60 ± 2.08	10.37 ± 2.06	10.36 ± 2.07
			0.757	0.71	0.45	0.426

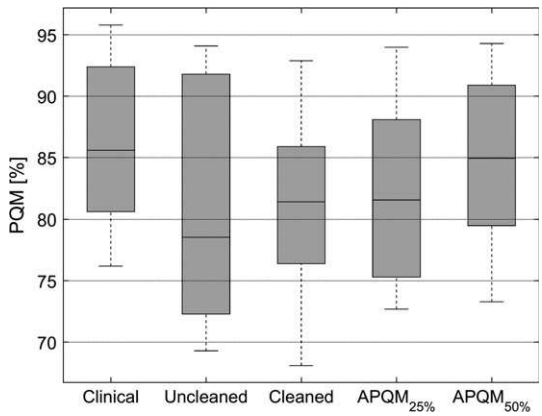

FIG. 2. Closed-loop test. Whiskers box plot of PQM%. The four models are compared to the 20 clinical plans included in the model and considered for the closed-loop test.

the overall quality comparison has been repeated removing the four patients: the results are depicted as white boxes in Fig. 3. As can be seen from the graphs, the four excluded patients were characterized by the lowest PQM% values, and purging them from the analysis shifted the PQM% distribution upwards for all the models.

4. DISCUSSION

The considerable commitment of human resources required to implement a KBP model, can be troublesome in resource-constrained departments.[7] This work investigated the effectiveness of the APQM score in facilitating and increasing the efficiency of the KBP training process.[8–10] The APQM% scoring was used to filter the training set prior to the model training and the performance of the KBP models was evaluated with the PQM% score.

Volume distributions, reported in Table IV, showed that the selection based on APQM% substantially preserves the extent of the geometrical domain of the model population. Besides the removal of the patients with largest PTV, no significant differences were noted comparing the *Uncleaned* and *Cleaned* models to the APQM%-filtered ones. Nevertheless,

TABLE VI. Open-loop test. Comparison of DVH endpoint between the clinically approved plans and the RapidPlan-generated plans related to the four models. Each value is reported as mean value ±1 SD and is accompanied by the *P*-value of a paired *t* test against the clinical plans. Statistically significant comparisons are marked by "*".

Structure	Metric	Clinical	Unclean	Clean	$APQM_{25\%}$	$APQM_{50\%}$
PTV	$V_{95\%}$ (%)	99.2 ± 0.4	99.8 ± 0.3	99.7 ± 0.3	99.8 ± 0.3	99.8 ± 0.3
			<0.001*	0.001*	0.001*	<0.001*
	$D_{98\%}$ (%)	98.6 ± 0.3	99.3 ± ± 0.2	99.2 ± 0.3	99.3 ± 0.3	99.3 ± 0.3
			<0.001*	0.001*	<0.001*	<0.001*
	$D_{2\%}$ (%)	105.5 ± 0.7	104.7 ± 0.8	104.9 ± 0.8	104.8 ± 0.8	104.8 ± 0.5
			0.023*	0.022*	0.031*	0.027*
	HI	6.9 ± 0.9	5.4 ± 0.7	5.7 ± 1.1	5.5 ± 0.8	5.4 ± 0.8
			0.004*	0.004*	0.005*	0.004*
	CN	0.856 ± 0.028	0.847 ± 0.026	0.830 ± 0.032	0.867 ± 0.025	0.843 ± 0.038
			0.185	0.258	0.010*	0.239
Rectum	V_{40Gy} (%)	26.0 ± 12.6	33.2 ± 8.2	31.4 ± 11.2	30.3 ± 5.8	28.8 ± 7.1
			0.09	0.151	0.242	0.462
	V_{65Gy} (%)	6.9 ± 5.0	9.2 ± 4.8	8.3 ± 5.5	8.4 ± 4.4	8.1 ± 4.5
			0.044*	0.159	0.142	0.253
	V_{75Gy} (cc)	2.8 ± 2.4	3.5 ± 2.3	3.3 ± 2.4	3.3 ± 2.2	3.2 ± 2.2
			0.005*	0.024*	0.017*	0.053
Bladder	V_{40Gy} (%)	14.9 ± 18.2	16.8 ± 18.1	17.1 ± 19.7	17.0 ± 20.0	16.9 ± 18.6
			0.011*	0.023*	0.085	0.033*
	V_{65Gy} (%)	5.6 ± 9.9	5.9 ± 9.4	6.0 ± 10.1	6.1 ± 10.1	6.0 ± 9.8
			0.423	0.159	0.189	0.275
	V_{75Gy} (cc)	12.0 ± 9.0	12.9 ± 9.0	12.9 ± 9.3	13.1 ± 9.4	13.0 ± 9.2
			0.007*	0.008*	0.007*	0.003*
Femoral head L	D_{1cc} (Gy)	23.60 ± 4.61	23.77 ± 4.43	21.86 ± 4.60	21.46 ± 2.29	19.52 ± 3.41
			0.916	0.226	0.269	0.115
	Mean dose (Gy)	10.51 ± 2.87	10.98 ± 1.69	10.09 ± 2.19	10.28 ± 1.44	9.64 ± 1.46
			0.494	0.458	0.757	0.325
Femoral head R	D_{1cc} (Gy)	24.09 ± 6.74	19.68 ± 3.24	19.55 ± 3.99	19.47 ± 2.94	18.75 ± 4.20
			0.024*	0.025*	0.086	0.009*
	Mean dose (Gy)	10.75 ± 4.03	10.22 ± 1.89	9.64 ± 2.12	9.61 ± 1.81	9.68 ± 2.25
			0.61	0.233	0.36	0.335

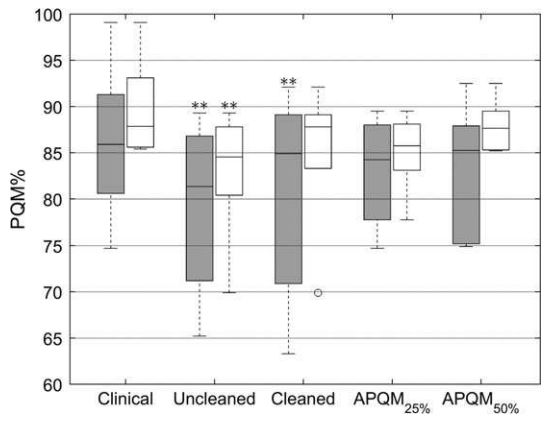

FIG. 3. Open-loop test. Whiskers box plot of PQM%. The four models are compared to the clinical plans selected for the open-loop test. Gray boxes are related to 20 patients while white boxes are related to 16 patients. Statistically significant comparisons are highlighted.

the twofold reduction in sample numerosity of the $APQM_{50\%}$ model resulted in the fact that four out of twenty patients selected for the open-loop test were outside of its geometrical domain. This underlines the need of a proper balance between the size of the sample and the choice of a strict consistency and quality of plans. So, before the clinical implementation of each KBP model, a clear and precise decision about its scope and its intended use is mandatory.

DVH analysis (Tables V and VI) showed that APQM%-filtered models outperformed or equaled the results of the *Cleaned* model both in terms of PTV coverage and OAR sparing. Moreover, $APQM_{50\%}$ model was the closest to the quality of the clinical plans used as reference. These results were confirmed by the analysis of the overall plan quality performed through PQM% comparison (Figs. 2 and 3).

The lack of human intervention during plan optimization may seem to generate low quality plans, it improved the PTV

coverage and homogeneity at the cost of lower OAR sparing. In fact, the relative weight of planning objectives, automatically generated by RapidPlan and left unchanged during the whole optimization procedure, favored the PTV over OARs.

APQM% scoring represents an intuitive and clinical oriented way to judge the overall plan quality with a patient-specific approach, and, as we showed herein, it may be extremely useful during the validation and refinement of KBP models. First of all, introducing a plan selection method based on APQM% ranking helps the user choose the best candidates to feed RapidPlan, improving the consistency of the model library. Moreover, we illustrated that a narrower APQM% threshold results in better planning outcomes. Considering a relative comparison between the planning outcome of the four RapidPlan models here described, the better planning outcomes might be due the fact that a KBP model can produce outcomes only as good as the mean quality of its training set.[1,17] In both closed- and open-loop tests, the differences in mean plan quality between the training sets of the four trained models were qualitatively transferred to the differences in mean plan quality of their outcomes. To summarize, the APQM% selection performed prior to KBP feeding can speed-up the model building, because low quality plans are easily identified during the refinement phase, and higher quality outcomes emerge because APQM% intrinsically increases the quality of the KBP library. Interestingly, this also means that APQM% represents a self-consistent measure of the overall plan quality. Whatever way a model is refined and generated, measuring the quality of its training library with APQM% allows the prediction of the quality of its outcome in terms of PQM% or APQM%.

Within this study, no further refinements were undertaken after the APQM% thresholding, so the lengthy and iterative process of model refinement was reduced to a feed-forward process. Even if this work proved the efficacy of APQM% as plan selection method prior to KBP feeding, this does not diminish the need for a systematic validation of a KBP model prior to its clinical employment and does not allow users to approach KBP solutions in a hasty and simplistic way.

It should be noticed that the general better target coverage and the lower OAR sparing showed by RapidPlan could be due to a possible suboptimal choice of the optimization objectives in the model. This set of objectives could have been refined as indicated in literature to obtain a better comparison with clinical plans.[4] This task was outside the scope of the present work but, in principle, the optimization objectives refinement could be undertaken as a following steps following a procedure similar to the presented one and based on APQM% scoring.

The validity of this study is limited by the single and relatively simple treatment site considered. A further limitation of this study is represented by the threshold chosen to build the APQM%-based model. A merely ordinal selection was performed: this was intended to demonstrate that a higher threshold implies better results, potentially at the price of reducing the applicability of the model. The authors did not propose a method to select a specific threshold based on APQM% scoring, because it may depend on the APQM% algorithm definition and it should be tailored to the numerosity of the available plan sample. The APQM% models did not undergo any process devoted to outlier detection or removal, and the presence of outliers did not lessen the validity of the presented results. Indeed, the presence of outliers did not prevent the APQM% models from obtaining the best results among the considered models. The usefulness of the iterative cleaning of outliers, that would have narrowed the training sets, has already been debated because it does not ensure a substantial quality increase.[4,12]

It should be emphasized that the entire study is based on the choice of a particular PQM algorithm, which is highly subjective and might not be universally valid. The PQM has nonetheless the advantage of being customized to the single center protocol and habits. This translates into two valuable advantages. First, once the PQM algorithm is defined in accordance with the medical staff and tailored to the center's clinical endpoints, it serves as a quality assurance tool. Furthermore, the availability of a common PQM algorithm allows comparison of KBP models shared between centers or conversely, might help different centers to build a shared KBP model once they agree on the clinical endpoints to be adopted.

5. CONCLUSION

Selecting the plan to be fed to a KBP system on the basis of an overall patient-specific metric of quality is shown to improve or at least equal the quality of KBP model. The use of PQM algorithm complemented by the FDVH estimate allows an intuitive and self-consistent measure of plan quality, which can simplify and reduce the workload related to KBP model training and validation. The proposed methodology is general and can transform a lengthy iterative process into a forward feeding process based on a customizable tool. More work is needed to test the validity of the proposed method to more complex treatments, techniques and sites.

CONFLICT OF INTEREST

None declared.

*Marco Fusella and Alessandro Scaggion equally contributed to this work and should be considered as co-first authors.
[a)]Author to whom correspondence should be addressed. Electronic mail: marco.fusella@iov.veneto.it; Telephone: +39 049 8212967; Fax: +39 049 8212968.

REFERENCES

1. *Eclipse Photon and Electron Reference Guide*. Doc. ID. P1012333-001-A. Varian Medical System; 2015.
2. 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;113:385–391.
3. Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WFAR. Evaluation of a knowledge-based planning solution for head and neck cancer. *Int J Radiat Oncol Biol Phys*. 2015;91:612–620.
4. 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;120:473–479.
5. Wu H, Jiang F, Yue H, Li S, Zhang Y. A dosimetric evaluation of knowledge-based VMAT planning with simultaneous integrated boosting for rectal cancer patients. *J Appl Clin Med Phys*. 2016;17:78–85.
6. 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;95:981–990.
7. Li N, Carmona R, Sirak I, et al. Highly efficient training, refinement, and validation of a knowledge-based planning quality-control system for radiation therapy clinical trials. *Int J Radiat Oncol Biol Phys*. 2017;97:164–172.
8. 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;2:296–305.
9. Ahmed S, Nelms B, Gintz D, et al. A method for *a priori* estimation of best feasible DVH for organs-at-risk: validation for head and neck VMAT planning. *Med Phys*. 2017;44:5486–5497.
10. *PlanIQ reference guide*. Doc. No 1210611, Rev C. Sun Nuclear Corp.; 2015.
11. Boutilier JJ, Craig T, Sharpe MB, Chan TCY. Sample size requirements for knowledge-based treatment planning. *Med Phys*. 2016;43:1212–1221.
12. Delaney AR, Tol JP, Dahele M, et al. Effect of dosimetric outliers on the performance of a commercial knowledge-based planning solution. *Int J Radiat Oncol Biol Phys*. 2016;94:496–477.
13. Fogliata A, Wang PM, Belosi F, et al. Assessment of a model based optimization engine for volumetric modulated arc therapy for patients with advanced hepatocellular cancer. *Radiat Oncol*. 2014;9:236.
14. 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.
15. Wu B, Kusters M, Kunze-busch M, et al. Cross-institutional knowledge-based planning (KBP) implementation and its performance comparison to auto-planning engine (APE). *Radiother Oncol*. 2017;123:57–62.
16. Van't Riet A, Mak AC, Moerland MA, Elders LH, Van Der Zee W. A conformation number to quantify the degree of conformality in brachytherapy and external beam irradiation: application to the prostate. *Int J Radiat Oncol Biol Phys*. 1997;37:731–736.
17. Fogliata A, Reggiori G, Stravato A, et al. RapidPlan head and neck model: the objectives and possible clinical benefit. *Radiat Oncol*. 2017;12:73.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

Figure S1: Closed-loop validation. On the left the DVH curves averaged over the 20 patients is plotted for the clinical sample and every model. On the right the difference between the clinical mean DVH and the mean DVH of each model is reported. (a) PTV, (b) Bladder, (c) Rectum, (d) Left Femoral Head, and (e) Right Femoral Head.

Figure S2: Open-loop validation. On the left the DVH curves averaged over the 20 patients is plotted for the clinical sample and every model. On the right the difference between the clinical mean DVH and the mean DVH of each model is reported. (a) PTV, (b) Bladder, (c) Rectum, (d) Left Femoral Head, and (e) Right Femoral Head.