International Journal of
Radiation Oncology
biology • physics

www.redjournal.org

Physics Contribution

Machine Learning for Patient-Specific Quality Assurance of VMAT: Prediction and Classification Accuracy

Jiaqi Li, MS,* Le Wang, PhD,[†,‡] Xile Zhang, MS,* Lu Liu, MS,* Jun Li, PhD,* Maria F. Chan, PhD,[§] Jing Sui, PhD,[†,‡] and Ruijie Yang, PhD*

**Department of Radiation Oncology, Peking University Third Hospital, Beijing, China; [†]Brainnetome Center & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China; [‡]CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China; and [§]Medical Physics Department, Memorial Sloan Kettering Cancer Center, New York, New York*

Received Dec 27, 2018. Accepted for publication Jul 27, 2019.

Summary

Delivery accuracy of volumetric modulated arc therapy is commonly assessed by measurement-based patient-specific quality assurance; however, this process is labor intensive and time consuming. By using plan complexity metrics as input and gamma passing rate as output, 1 regression model and 1 classification model were developed and validated technically and clinically. The derived regression model could accurately predict gamma passing rate for the majority of volumetric

Purpose: To assess the accuracy of machine learning to predict and classify quality assurance (QA) results for volumetric modulated arc therapy (VMAT) plans.

Methods and Materials: Three hundred three VMAT plans, including 176 gynecologic cancer and 127 head and neck cancer plans, were chosen in this study. Fifty-four complexity metrics were extracted from the QA plans and considered as inputs. Patient-specific QA was performed, and gamma passing rates (GPRs) were used as outputs. One Poisson lasso (PL) regression model was developed, aiming to predict individual GPR, and 1 random forest (RF) classification model was developed to classify QA results as "pass" or "fail." Both technical validation (TV) and clinical validation (CV) were used to evaluate the model reliability. GPR prediction accuracy of PL and classification performance of PL and RF were evaluated.

Results: In TV, the mean prediction error of PL was 1.81%, 2.39%, and 4.18% at 3%/3 mm, 3%/2 mm, and 2%/2 mm, respectively. No significant differences in prediction errors between TV and CV were observed. In QA results classification, PL had a higher specificity (accurately identifying plans that can pass QA), whereas RF had a higher sensitivity (accurately identifying plans that may fail QA). By using 90% as the action limit at a 3%/2 mm criterion, the specificity of PL and RF was 97.5% and 87.7% in TV and 100% and 71.4% in CV, respectively. The sensitivity of PL and RF was 31.6% and 100% in TV and 33.3% and 100% in CV, respectively. With 100% sensitivity, the QA

Corresponding authors: Jing Sui, PhD, and Ruijie Yang, PhD; E-mails: kittysj@gmail.com, ruijyang@yahoo.com

Jiaqi Li and Le Wang made equal contributions to this work.

This work was partly supported by the National Natural Science Foundation of China (grant no. 81071237), Interdisciplinary Medicine Seed Found of Peking University (grant no. BMU20160585), the NIH/NCI P30 Cancer Center Support Grant (grant no. CA008748), and the Strategic Priority Research Program of Chinese Academy of Science (grant no. XDB32040100).

Disclosures: none.

Int J Radiation Oncol Biol Phys, Vol. 105, No. 4, pp. 893−902, 2019

https://doi.org/10.1016/j.ijrobp.2019.07.049

modulated arc therapy plans with high specificity, whereas the classification model had higher sensitivity. Machine learning is a useful tool to reduce quality assurance workload.

workload of 81.2% of plans in TV and 62.5% of plans in CV could be reduced by RF. **Conclusions:** The PL model could accurately predict GPR for most VMAT plans. The RF model with 100% sensitivity was preferred for QA results classification. Machine learning can be a useful tool to assist VMAT QA and reduce QA workload.

Introduction

Over the past 2 decades, there have been rapid developments in radiation therapy delivery techniques. Compared with 3-dimensional conformal radiation therapy, intensity modulated radiation therapy (IMRT), including fixed-gantry IMRT and volumetric modulated arc therapy (VMAT), provides better target coverage and dose sparing to organs at risk.[1,2] Fixed-gantry IMRT and VMAT plans created by inverse planning algorithms consist of highly modulated apertures with increased dosimetric uncertainty and pose a great challenge to treatment planning system (TPS), dose calculation, and linear accelerator (Linac) performance.[3,4] Comprehensive quality assurance (QA) and quality control programs have been developed to assess the reliability of treatment delivery and improve patient safety.[5-8] Although questions and concerns were raised by many studies,[9-12] patient-specific QA with gamma analysis is still the most widely used QA method. Patient-specific QA is labor intensive and time consuming and is unfavorable for busy radiation therapy centers. Recently, researchers have shown an increased interest in using machine learning to predict patient-specific QA results.[13-16]

Valdes et al used the plan complexity metrics and patient-specific QA results of 498 IMRT plans from multiple treatment sites to train a Poisson regression with Lasso regularization (PL) model.[13] The developed model could accurately predict 3%/3 mm gamma passing rate (GPR) with maximum errors smaller than 3%. The generalization performance of this model was then tested in a multi-institutional validation study; 86.33% (120 of 139) plans had a prediction error smaller than 3.5%.[14] The decreased prediction accuracy may be due to different dose verification methods used in different institutions. By using fluence maps of IMRT plans as input, the convolution neural network (CNN) model developed by Interian et al[15] had similar prediction accuracy compared with the previously developed Poisson lasso model. However, more than 15 plans with prediction error higher than 3% were observed in both the CNN and PL models, and the maximum prediction error was higher than 5%. Instead of predicting GPR for plans from multiple treatment sites, Tomori et al[16] trained the CNN model with 60 prostate IMRT plans. Planar dose distributions, geometric features of planning target volume (PTV) and rectum, and MU for each field were used as inputs. The maximum prediction errors were 3.0%, 4.5%, and 5.8% at 3%/3 mm, 3%/2 mm, and 2%/2 mm, respectively.

Previous studies have shown the potential of machine learning models to accurately predict patient-specific QA results. When deciding whether the plan can be delivered accurately enough to be used for patient treatment, it is critical to select the appropriate gamma criteria and tolerance/action limits. The American Association of Physicists in Medicine TG 218 report recommended 95% and 90% as the tolerance and action limits under a 3%/2 mm gamma criterion, respectively.[8] Therefore, the most important function of a machine learning model is to find plans that may fail to pass the tolerance/action limits before QA measurements. However, the classification accuracy of machine learning models under different gamma criteria and tolerance/action limits was not fully investigated and available in literature. Furthermore, previous studies were only based on IMRT plans, and it is still unclear whether the QA results of VMAT plans can be accurately predicted or classified. To the best of our knowledge, this study is the first to report machine learning algorithms used for virtual VMAT QA and the sensitivity and specificity using different gamma metrics.

In this study, 1 regression model and 1 classification model were developed to predict patient-specific QA results of VMAT plans for gynecologic (GYN) and head and neck (H&N) cancer. The aims of this study were (1) to investigate the accuracy of a machine learning model to predict GPR of VMAT plans at different gamma criteria; (2) to evaluate the sensitivity and specificity of machine learning model to classify VMAT QA results using different action limits at different gamma criteria; (3) based on the results, to give recommendations for model clinical application. The findings of this study will provide more insights into the field of automation in patient-specific VMAT QA.

Methods and Materials

Clinical data collection

Three hundred three VMAT plans with dual-arc and 2° control-point spacing were retrospectively collected. Among these plans, 176 were GYN plans and 127 were H&N plans. The prescription dose to PTV for GYN patients was 50.4 Gy (1.8 Gy/28 fractions). For H&N cases, prescription doses of 60.04 Gy (1.82 Gy/33 fractions) and 69.96 Gy (2.12 Gy/33 fractions) were delivered to PTV and

Table 1 Summary of complexity metrics used in this study

Number	Metrics	Reference
1	Modulation index for leaf speed f = 2 (MI_s 2)	17
2	Modulation index for leaf speed f = 1 (MI_s 1)	17
3	Modulation index for leaf speed f = 0.5 (MI_s 0.5)	17
4	Modulation index for leaf speed f = 0.2 (MI_s 0.2)	17
5	Modulation index for leaf acceleration f = 2 (MI_a 2)	17
6	Modulation index for leaf acceleration f = 1 (MI_a 1)	17
7	Modulation index for leaf acceleration f = 0.5 (MI_a 0.5)	17
8	Modulation index for leaf acceleration f = 0.2 (MI_a 0.2)	17
9	Modulation index for total modulation f = 2 (MI_t 2)	17
10	Modulation index for total modulation f = 1 (MI_t 1)	17
11	Modulation index for total modulation f = 0.5 (MI_t 0.5)	17
12	Modulation index for total modulation f = 0.2 (MI_t 0.2)	17
13	Proportion of leaf speed ranging from 0-0.4 cm/s ($S_{0\text{-}0.4}$)	18
14	Proportion of leaf speed ranging from 0.4-0.8 cm/s ($S_{0.4\text{-}0.8}$)	18
15	Proportion of leaf speed ranging from 0.8-1.2 cm/s ($S_{0.8\text{-}1.2}$)	18
16	Proportion of leaf speed ranging from 1.2-1.6 cm/s ($S_{1.2\text{-}1.6}$)	18
17	Proportion of leaf speed ranging from 1.6-2.0 cm/s ($S_{1.6\text{-}2}$)	18
18	Proportion of leaf acceleration ranging from 0-1 cm/s^2 ($A_{0\text{-}1}$)	18
19	Proportion of leaf acceleration ranging from 1-2 cm/s^2 ($A_{1\text{-}2}$)	18
20	Proportion of leaf acceleration ranging from 2-4 cm/s^2 ($A_{2\text{-}4}$)	18
21	Proportion of leaf acceleration ranging from 4-6 cm/s^2 ($A_{4\text{-}6}$)	18
22	Average leaf speed (ALS)	18
23	Standard deviation of leaf speed (SLS)	18
24	Average leaf acceleration (ALA)	18
25	Standard deviation of leaf acceleration (SLA)	18
26	Small aperture score 5 mm (SAS 5 mm)	19
27	Small aperture score 10 mm (SAS 10 mm)	19
28	Small aperture score 20 mm (SAS 20 mm)	19
29	Mean asymmetry distance (MAD)	19
30	Modulation complex score (MCS)	20
31	Leaf sequence variability (LSV)	20
32	Aperture area variability (AAV)	20
33	Plan area (PA)	21
34	Plan irregularity (PI)	21
35	Plan modulation (PM)	21
36	Plan normalized MU (PMU)	21
37	Union aperture area (UAA)	21
38	Edge metric (EM)	22
39	Converted aperture metric (CAM)	23
40	Edge area metric (EAM)	23
41	Circumference/area (C/A)	23
42	Average leaf travel distance (LT)	24
43	Combination of LT and MCS (LTMCS)	24
44	Average leaf gap (ALG)	25
45	Standard deviation of leaf gap (SLG)	25
46	Average dose rate (ADR)	-
47	Standard deviation of dose rate (SDR)	-
48	MU value in first arc (MU 1)	-
49	MU value in second arc (MU 2)	-
50	Prescribed dose to primary target per fraction (dose)	-
51	Field length at X direction in first arc (field X1)	-
52	Field length at Y direction in first arc (field Y1)	-
53	Field length at X direction in second arc (field X2)	-
54	Field length at Y direction in second arc (field Y2)	-

Complexity metrics that have "-" in the reference column can be easily extracted or calculated based on plan information in the treatment planning system.

planning gross target volume, respectively. All plans were generated on Eclipse TPS version 10.0 and delivered with Trilogy Linac and Millennium 120 multileaf collimator (Varian Medical System, Palo Alto, CA). The patient-specific QA measurements were performed with MatriXX ion chamber array and MultiCube phantom (IBA Dosimetry, Schwarzenbruck, Germany). The plan was delivered using the true composite method, recommended by the American Association of Physicists in Medicine TG 218 report.[8] The angular dependence of the detector array was corrected using a gantry angle sensor (IBA Dosimetry, Schwarzenbruck, Germany) during measurement. Gamma criteria of 3%/3 mm, 3%/2 mm, and 2%/2 mm with 10% dose threshold, absolute dose mode, and global normalization were used for gamma evaluations.

Fifty-four metrics affecting dose delivery accuracy of VMAT were selected by expert clinical physicists and used to characterize the modulation complexity of VMAT plans. A full summary is given in Table 1.[17-25] Radiation therapy plan files were exported from the Eclipse system and converted into ASCII format. An in-house−developed

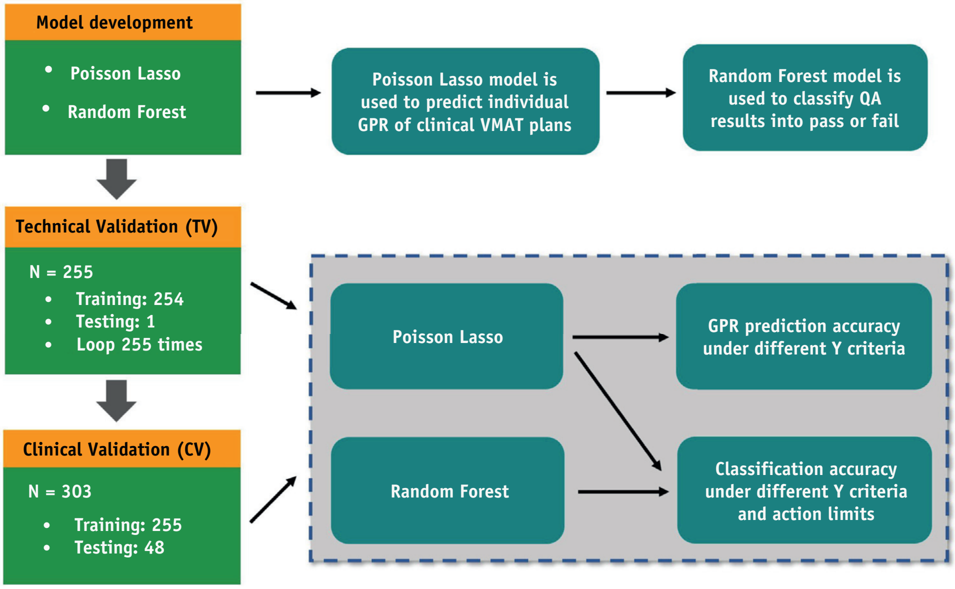
Model development
Poisson Lasso
Random Forest
Poisson Lasso model is used to predict individual GPR of clinical VMAT plans
Random Forest model is used to classify QA results into pass or fail
Technical Validation (TV)
N = 255
Training: 254
Testing: 1
Loop 255 times
Clinical Validation (CV)
N = 303
Training: 255
Testing: 48
Poisson Lasso
Random Forest
GPR prediction accuracy under different Y criteria
Classification accuracy under different Y criteria and action limits

Fig. 1. Flow diagram of the model development, technical validation, and clinical validation.

MATLAB script was used to extract the multileaf collimator positions and MU weights of all control points in the VMAT plans and calculate the complexity metrics.

Machine learning model design and validation

This study contained 3 sections, including model development, technical validation (TV), and clinical validation (CV) (Fig. 1). Two machine learning models were established and validated in this study: one is a PL model, and the other is a random forest (RF) model. In TV, 255 VMAT plans (143 GYN and 112 H&N) with nested cross validation were used to explore the model performance under different gamma criteria and action limits. In CV, an independent cohort of 48 VMAT plans (33 GYN and 15 H&N) without cross validation were used to further validate the reliability and feasibility of machine learning models as a clinical actionable tool for reducing QA workload.

The PL model has been previously used in individualized prediction studies.[13,14] It consists of a generalized linear regression model with Poisson prior and least absolute shrinkage and selection operator regularization. In this study, Poisson regression is used to model the nonnegative data with count attribute, solving the problem via maximum a posteriori estimation and trying to identify the optimal beta weights β by cross validation. Note that instead of predicting passing rate pr(X), Poisson regression predicts the failing rate, which equals $100 - \text{pr}(X)$. This was implemented using the open-source Python package, Statsmodels.

The TV workflow for a PL model with nested 10-fold and leave-one-out cross validation is shown in Figures 1 and 2A. To train the regression model with as much data as possible and test the model with the remaining data, leave-one-out cross validation was used. The data were divided into 254 plans for training and a single remaining plan for testing. In the model-training phase, 10-fold cross validation was used for PL to achieve the optimal hyperparameters, which were used for training the optimal regression models to achieve better generalization (white block in Fig. 2A). After the training process, the model derived from the training data was used to predict the GPR of the remaining test plan. This procedure looped 255 times so that the GPR of every VMAT plan was predicted. The classification performance of the PL model at different gamma criteria and action limits was further evaluated (Fig. 3). The sensitivity and specificity of the PL model was calculated at 3%/2 mm with action limits ranging from 90% to 99% and at 2%/2 mm with action limits ranging from 80% to 90%.

To further improve the classification accuracy, ensemble RF classification models with dimension reduction and balanced sampling techniques were developed and used to classify QA results. The TV workflow for the RF model is shown in Figures 1 and 2B. Principal component analysis (PCA) was adopted to reduce the feature number before classification; this is an orthogonal linear transformation to transform the data from high-dimensional space to a desired low-dimensional level. Here, the data are converted from 54 dimensions to 15 dimensions. Because the VMAT plans were heavily unbalanced between positive and negative plans (number of plans with relative higher GPR vs number of plans with relative lower GPR), leading to undesired classification results, a tendency to divide the samples into the majority class is possible. To avoid the minority class being neglected by the RF classifiers, a random undersampling strategy was applied to balance the

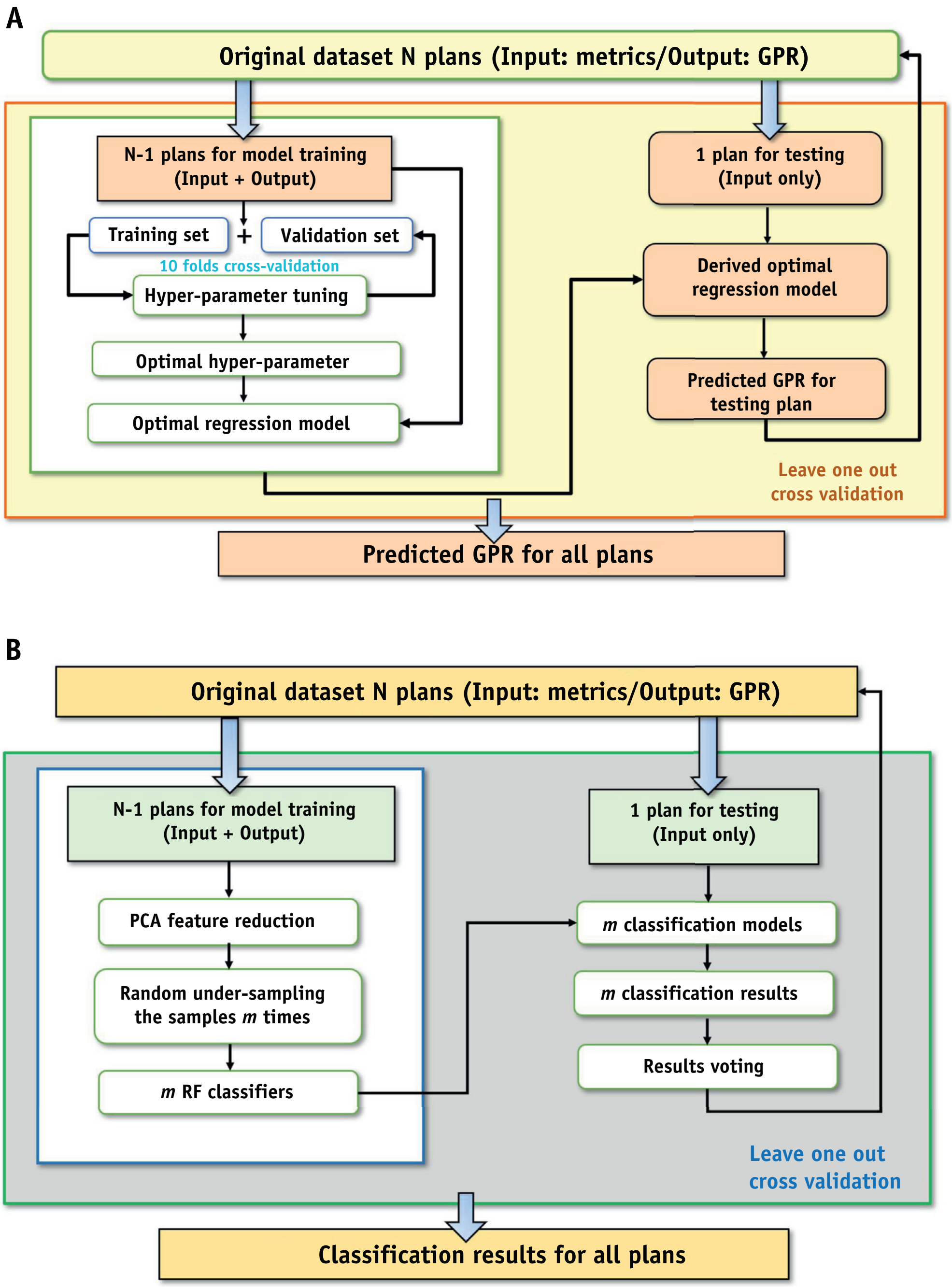
A
Original dataset N plans (Input: metrics/Output: GPR)
N-1 plans for model training (Input + Output)
Training set
+
Validation set
10 folds cross-validation
Hyper-parameter tuning
Optimal hyper-parameter
Optimal regression model
1 plan for testing (Input only)
Derived optimal regression model
Predicted GPR for testing plan
Leave one out cross validation
Predicted GPR for all plans
B
Original dataset N plans (Input: metrics/Output: GPR)
N-1 plans for model training (Input + Output)
PCA feature reduction
Random under-sampling the samples m times
m RF classifiers
1 plan for testing (Input only)
m classification models
m classification results
Results voting
Leave one out cross validation
Classification results for all plans

Fig. 2. The workflow chart for the Poisson lasso regression model (A) and random forest classification model (B) in technical validation. *Abbreviations:* GPR = gamma passing rate; PCA = principal component analysis; RF = random forest.

2 unbalanced classes by down sampling the majority class to the same size as the minority class. RF was used to provide the basic classifiers because RF achieves better performance and generalization by using column sampling to avoid overfitting. These RF classifiers were then ensembled to reduce variance and achieve the final stable classification results.

Similar nested cross validation was adopted to choose the optimal hyperparameters for PCA and RF classifiers, and the data were divided into training plans and testing plans. As shown in Figure 2B, PCA was first used for dimension reduction of training plans (N = 254). Random undersampling of majority classes was performed *m* times to balance the sample distribution, resulting in *m* RF classifiers. For the remaining 1 testing plan, classification was then made by ensemble voting of all the *m* classifiers, aiming to achieve more robust and unbiased results. Here we choose $m = 1000$.

In CV, 255 VMAT plans used in TV were used for model training. Another independent cohort of 48 VMAT

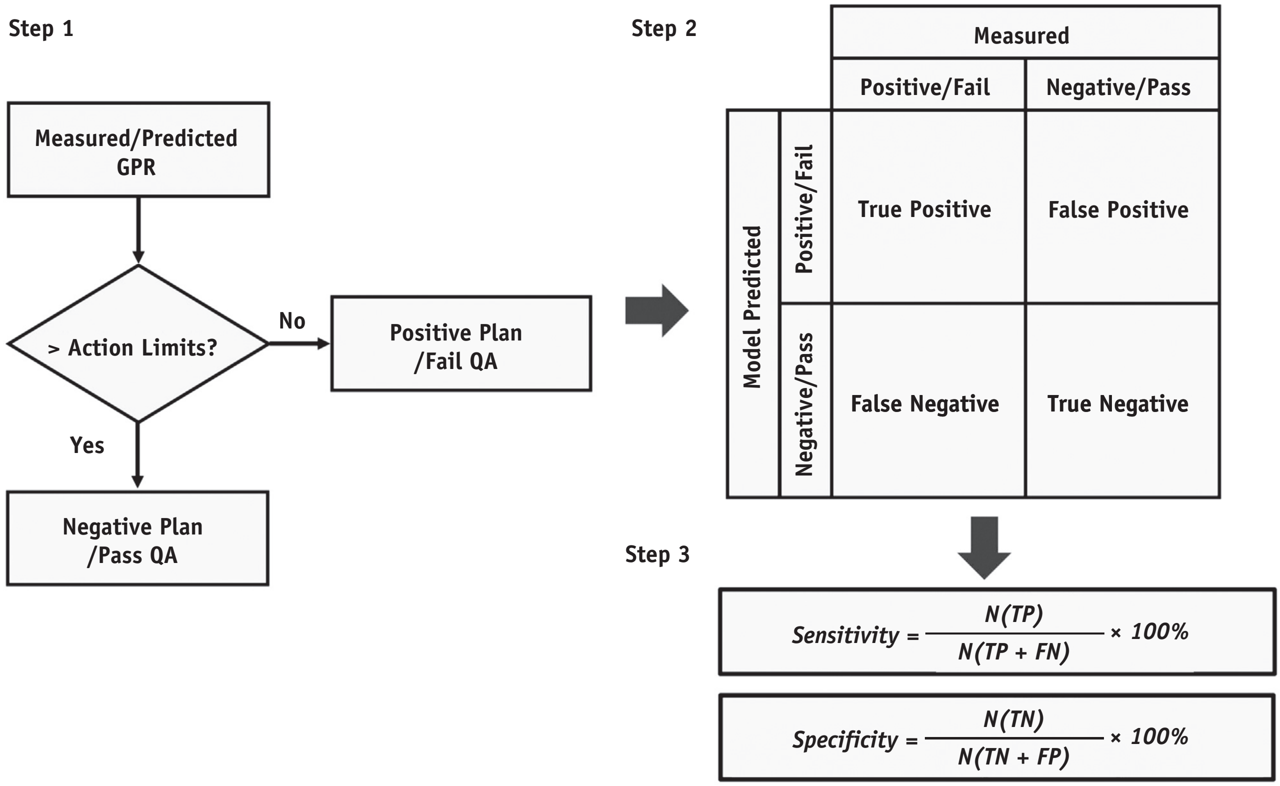

Fig. 3. The workflow chart for evaluating model performance in quality assurance results classification. *Abbreviations:* FN = false negative; FP = false positive; GPR = gamma passing rate; TN = true negative; TP = true positive.

plans, including 33 GYN and 15 H&N plans, was used for external testing (Fig. 1). No cross validation was performed in CV. The prediction and classification process of PL and RF model were the same as described.

Results

Prediction accuracy

The distributions of measured GPR of VMAT plans in TV and CV are shown in Table 2. In both data sets, majority of the measured GPR was distributed in the range of 95% to 100% at 3%/3 mm, 90% to 100% at 3%/2 mm, and 85% to 95% at 2%/2 mm gamma criteria. In the TV data set, 9 (3.5%) plans had GPR lower than 90% at 3%/3 mm; 19 (7.5%) plans had GPR lower than 90% at 3%/2 mm; and 23 (9.0%) plans had GPR lower than 80% at 2%/2 mm. In CV data set, 1 (2.1%) plan had GPR lower than 90% at 3%/3 mm; 6 (12.5%) plans had GPR lower than 90% at 3%/2 mm; and 5 (10.4%) plans had GPR lower than 80% at 2%/2 mm. The GPR prediction accuracy of PL decreased as the gamma criteria became more stringent (from 3%/3 mm to 2%/2 mm) (Fig. 4).

The differences of absolute prediction errors between PL model in TV and CV were not statistically significant. In TV, 239 (93.7%) plans had absolute prediction error lower than 5% at 3%/3 mm; 233 (91.4%) plans had absolute prediction error lower than 5% at 3%/2 mm; and only 174 (68.24%) plans had absolute prediction error lower than 5% at 2%/2 mm (Table 3). In CV, 45 (93.8%) plans had absolute prediction error lower than 5% at 3%/3 mm; 36 (75.0%) plans had absolute prediction error lower than 5% at 3%/2 mm; and only 29 (60.4%) plans had absolute prediction error lower than 5% at 2%/2 mm. The prediction accuracy of PL model was also affected by measured GPR, as shown in Table 4.

In both technical and CV, plans with measured GPR higher than 95% had significantly lower prediction errors than plans with measured GPR lower than 95% at 3%/3 mm and 3%/2 mm gamma criteria (3%/3 mm TV: 1.36% ± 1.39% vs 4.39% ± 3.37%, $P < .001$; 3%/3 mm CV: 1.38% ± 1.21% vs 4.12% ± 2.63%, $P = .021$; 3%/2 mm TV: 1.57% ± 1.16% vs 4.02% ± 3.28%, $P < .001$; 3%/2 mm CV: 1.71% ± 1.60% vs 4.15% ± 3.35%, $P = .003$); In 2%/2 mm GPR prediction, plans with measured GPR ranging from 85% to 95% had significantly lower prediction errors than plans with measured GPR higher than 95% or lower than 85%. (TV: 2.79% ± 2.51% vs 5.94% ± 4.32%, $P < .001$; CV: 3.13% ± 2.86% vs 8.39% ± 5.39%, $P = .001$).

Classification accuracy

The classification performances of the PL and RF models in TV are shown in Figure 5. In general, PL had higher specificity and RF had higher sensitivity. As the action limit

Table 2 Summary of measured GPR under different gamma criteria

	3%/3 mm		3%/2 mm		2%/2 mm	
Measured GPR	TV, n (%)	CV, n (%)	TV, n (%)	CV, n (%)	TV, n (%)	CV, n (%)
100-95	217 (85.1)	40 (83.3)	170 (66.7)	23 (47.9)	50 (19.6)	2 (4.2)
95-90	29 (11.4)	7 (14.6)	66 (25.9)	19 (39.6)	80 (31.4)	15 (31.3)
90-85	7 (2.7)	1 (2.1)	12 (4.7)	4 (8.3)	62 (24.3)	15 (31.3)
85-80	0	0	3 (1.2)	2 (4.2)	40 (15.7)	11 (22.9)
80	2 (0.8)	0	4 (1.6)	0	23 (9)	5 (10.4)

Abbreviations: CV = clinical validation; GPR = gamma passing rate; TV = technical validation.

value decreased, the sensitivity of PL was remarkably reduced; the opposite trend was observed for RF. For 3%/2 mm GPR classification using 90% as the action limits, the specificity of PL and RF was 97.46% (230/236) and 87.71% (207/236), respectively; the sensitivity of PL and RF was 31.57% (6/19) and 100% (19/19), respectively. For 2%/2 mm GPR classification using 80% as the action limit, the specificity of PL and RF was 96.55% (224/232) and 80.60% (187/232), respectively; the sensitivity of PL and RF were 43.48% (10/23) and 100% (23/23), respectively. The classification performances of the PL and RF models were further validated in CV (Table 5). For 3%/2 mm GPR classification using 90% as the action limits, the specificity of PL and RF was 100% (42/42) and 71.43% (30/42), respectively; the sensitivity of PL and RF was 33.33% (2/6) and 66.67% (4/6), respectively. For 2%/2 mm GPR classification using 80% as the action limit, the specificity of PL and RF was 100% (43/43) and 44.19% (19/43), respectively; the sensitivity of PL and RF was 60% (3/5) and 100% (5/5), respectively.

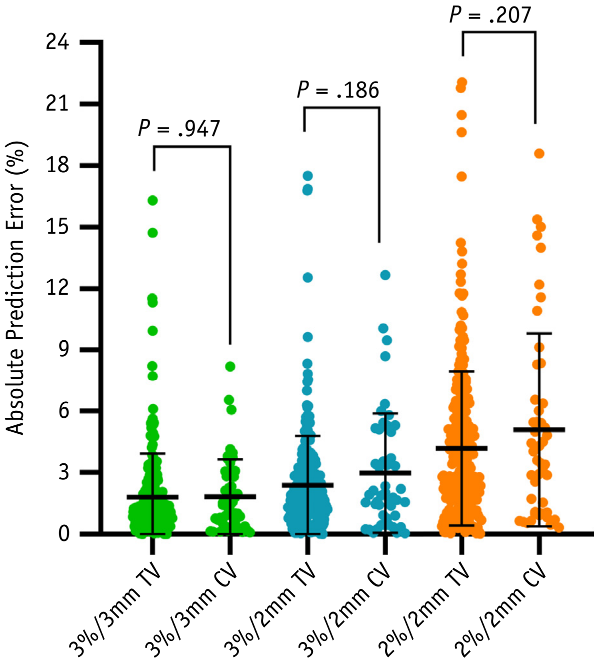

Fig. 4. The distribution of prediction errors of volumetric modulated arc therapy plans at different gamma criteria. *Abbreviations:* CV = clinical validation; TV = technical validation. Error bar = mean ± standard deviation. Student *t* test was performed.

Discussion

The main advantage of adopting machine learning into patient-specific QA is that the machine learning model could inform the physicist which treatment plan may fail QA before QA measurements. With the use of machine learning models, patient-specific QA could be narrowed and concentrated on a few plans instead of all plans. Given the large-scale adoption of VMAT in the clinic, machine learning models that are able to predict patient-specific QA results for VMAT plans will be useful for improving the efficiency of VMAT QA. After saving time and resources, physicists can devote more time to the failed plans and identify the cause of failure. In TV, we have demonstrated that the PL model could precisely predict the GPR for more than 90% of VMAT plans within 3.5% accuracy at 3%/3 mm and within 5% accuracy at 3%/2 mm. PL model also maintains the same prediction accuracy during CV. The prediction accuracy of PL models was greatly affected by gamma criteria and measured GPR.

The number of plans with low GPR differed across studies, and this will have a huge impact on prediction accuracy. In the study by Valdes et al,[13] about 8% of IMRT plans (40/498) had measured 3%/3 mm GPR lower than 95%. In the study by Tomori et al,[16] very few IMRT plans had measured 3%/3 mm GPR lower than 95%, and the lowest 3%/3 mm GPR was 94%. In this study, 15.2% of VMAT plans (46/303) had measured GPR lower than 95%, and the lowest 3%/3 mm GPR was 74.63%. In a previous study,[16] 3 plans had measured GPR lower than 90% at 3%/2 mm; 15 plans had measured GPR lower than 85% at 2%/2 mm. In this study, 25 plans had measured GPR lower than 90% at 3%/2 mm; 79 plans had measured GPR lower than 85% at 2%/2 mm. Although there are differences in measured GPR among studies, all previous studies and this study had heavily unbalanced data distribution. It is very difficult for a single institution to collect adequate amounts of low GPR plans for model training. To improve the

Table 3 Summary of prediction errors under different gamma criteria

	3%/3 mm		3%/2 mm		2%/2 mm	
Metrics	TV, n (%)	CV, n (%)	TV, n (%)	CV, n (%)	TV, n (%)	CV, n (%)
Abs Err ≤3.5%	230 (90.2)	40 (83.3)	205 (80.4)	33 (68.8)	133 (52.2)	23 (47.9)
Abs Err ≤5%	239 (93.7)	45 (93.8)	233 (91.4)	36 (75.0)	174 (68.2)	29 (60.4)
Abs Err ≤10%	251 (98.4)	48 (100)	251 (98.4)	46 (95.8)	238 (93.3)	40 (83.3)
MAE (SD)	1.81% (2.12)	1.83% (1.82)	2.39% (2.41)	2.98% (2.91)	4.18% (3.76)	5.10% (4.71)

Abbreviations: Abs Err = absolute prediction error; CV = clinical validation; MAE = mean absolute error; SD = standard deviation; TV = technical validation.

prediction accuracy of plans with low GPR, a multi-institutional collaborative research is warranted.

The prediction accuracy of the machine learning model is also affected by the accuracy and consistency of patient-specific QA measurement. Hussein et al[26] reported that the variability of measurement results among different commercial QA devices increased with tightening gamma criteria. The largest difference for mean GPR and minimum GPR at 2%/2 mm was 8.4% and 15%, respectively. Agnew et al[27] found that less than 0.5% variation existed among GPR at 3%/3 mm with different gamma analysis software; however, the variation increased to 10% at 1%/1 mm. If the measurement results are inaccurate or inconsistent, the machine learning model would struggle to find correlations between input features and GPR and make an accurate prediction; in both this study and previous study,[16] the 2%/2 mm GPR prediction accuracy was much worse compared with 3%/3 mm and 3%/2 mm GPR prediction. Many researchers have shown that 2%/2 mm criterion was more sensitive in detecting clinically relevant errors compared with 3%/3 mm[9,10]; thus, machine learning models that can accurately predict 2%/2 mm GPR are more favorable. Future studies should focus on improving the model prediction accuracy at 2%/2 mm.

Instead of the differences between measured and predicted GPR, the ability of machine learning model to accurately classify plans into "pass" or "fail" based on action limits used is the most important indicator to evaluate the clinical feasibility of machine learning models, as suggested in the TG 218 report.[8] To our best knowledge, this is the first study to discuss the clinical usability of machine learning models based on a clinical decision-making process and to give actionable recommendations for the clinical application of machine learning models. There are 2 types of classification errors. The first is false positive, which will reduce the model specificity and add unnecessary QA workload. The second is false negative, which will reduce the model sensitivity and bring hidden dangers to patient safety and should be avoided. The ability of the PL to detect plans that may fail patient-specific QA (sensitivity) decreased significantly with decreasing value of action limits, whereas the opposite trends were observed for the RF model. This may be because the PL model tended to overestimate the GPR of VMAT plans, especially for those plans that had low measured GPR. To avoid overfitting and classification bias caused by unbalanced training data, data preprocessing strategies such as PCA, random undersampling, and ensemble voting were applied to an RF model for the first time. In TV, a much better sensitivity of the RF model was observed at the cost of a relatively small decrease in specificity. These results are valuable because model sensitivity is more important than specificity in virtual patient-specific QA using machine learning. These results also emphasize the importance of choosing the machine learning algorithm based on its inherent characteristics.

With 100% sensitivity, the patient-specific QA workload of plans labeled "pass" can be reduced. This paradigm shift can both save time and ease the strain on treatment resources. For plans labeled "fail," measurement-based QA still needs to be performed before patient treatment. Based on the results in TV, 81.2% (207/255) and 73.3% (187/255) of QA workload could be reduced by the RF model at 3%/2 mm using a 90% action

Table 4 Summary of absolute prediction error under different measured GPR

	3%/3 mm		3%/2 mm		2%/2 mm	
Measured GPR	TV, mean (SD)	CV, mean (SD)	TV, mean (SD)	CV, mean (SD)	TV, mean (SD)	CV, mean (SD)
100-95	1.36 (1.40)	1.38 (1.21)	1.58 (1.16)	1.72 (1.60)	4.44 (1.92)	9.28 (4.11)
95-90	3.38 (1.35)	3.77 (2.63)	3.17 (1.91)	3.14 (2.44)	1.88 (1.76)	3.28 (2.75)
90-85	5.42 (3.74)	6.55 (NA)	4.21 (2.04)	7.41 (4.86)	3.97 (2.84)	2.98 (3.04)
85-80	0	0	7.34 (2.42)	7.30 (3.10)	5.27 (4.07)	8.85 (4.83)
80	15.52 (1.12)	0	14.88 (4.36)	0	10.36 (5.59)	7.04 (7.62)

Abbreviations: CV = clinical validation; GPR = gamma passing rate; NA = not available; SD = standard deviation; TV = technical validation.

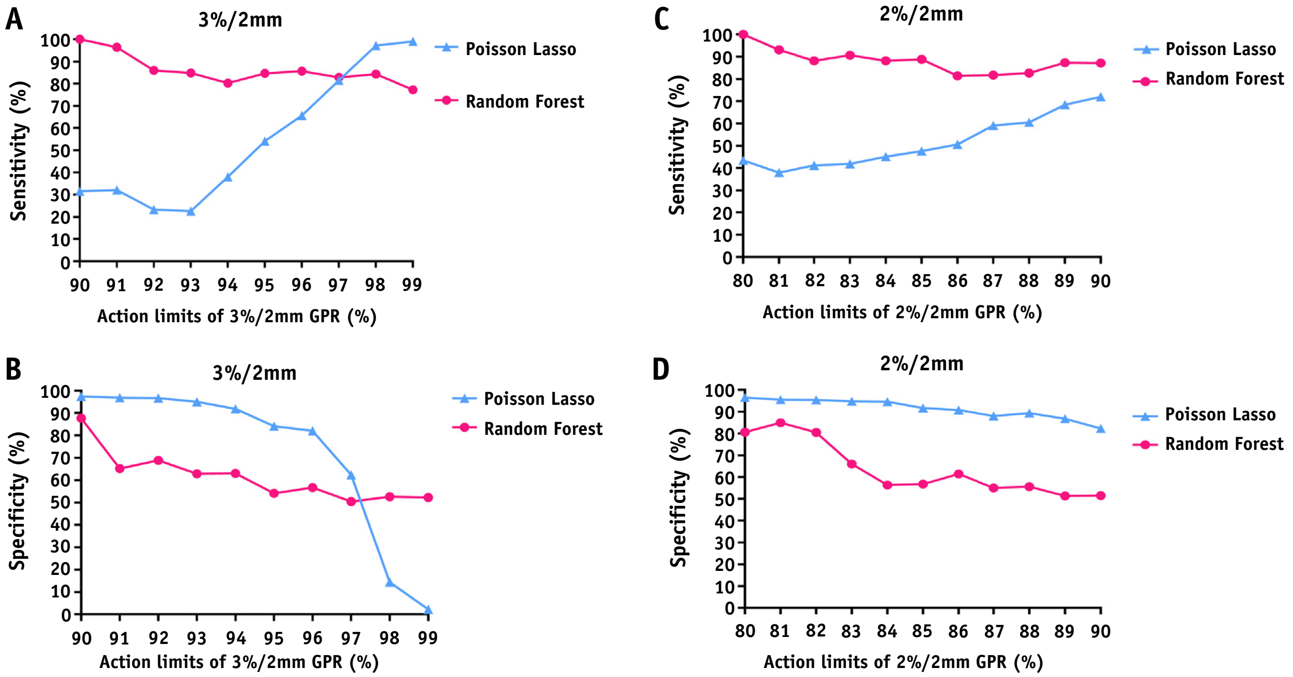

Fig. 5. Classification performance of Poisson lasso model and random forest model in technical validation with various action limits at 3%/2 mm (A, B) and 2%/2 mm (C, D) criteria.

limit and at 2%/2 mm using an 80% action limit, respectively. In CV, the classification performances of PL and RF models were further evaluated. The RF model still has better sensitivity compared with the PL model; only 2 false negative plans were observed at the 3%/2 mm gamma criterion. Interestingly, we found that the 2 false negative plans were labeled "fail" by the RF model at a 2%/2 mm gamma criterion and 80% action limit, indicating that if 1 particular plan were labeled "pass" at 3%/2 mm and labeled "fail" at 2%/2 mm, measurement-based QA still needs to be performed for that plan. Based on the results in CV, 62.5% (30/48) and 39.6% (19/48) of QA workload could be reduced by the RF model at 3%/2 mm using a 90% action limit and at 2%/2 mm using an 80% action limit, respectively. Based on these results, the clinical use of machine learning model is recommended as follows: (1) compared with the PL model, the RF model is safer and more practical to identify plans that may fail QA and thus is a clinically usable tool to reduce VMAT QA workload; (2) classification results of the RF model at both 3%/2 mm and 2%/2 mm criteria should be taken into consideration when physicists decide whether to perform measurement-based VMAT QA; (3) standard QA is no longer necessary only if a treatment plan is labeled "pass" by RF model at both 3%/2 mm and 2%/2 mm criteria using the GPR action limits of 90% and 80%, respectively.

Table 5 Model classification performance in clinical validation

Gamma criteria, Action limits	Model	Predicted	Measured	
			Positive	Negative
3%/2 mm, 90%	Poisson lasso	Positive	2	0
		Negative	4	42
3%/2 mm, 90%	Random forest	Positive	4	12
		Negative	2	30
2%/2 mm, 80%	Poisson lasso	Positive	3	0
		Negative	2	43
2%/2 mm, 80%	Random forest	Positive	5	24
		Negative	0	19

Note that the measurement-based pretreatment QA could only be reduced based on the assumption that accurate TPS commissioning, adequate machine QA, and other dose verification methods such as electronic portal imaging device transit dosimetry or independent dose calculation were followed.[6,7,28-34] Because all input and output data of this study were derived from single Varian Linac and 2D array, it is critical to evaluate the generalization performance of the machine learning model. Plans and corresponding QA results from different institutions with different types of Linac and QA devices will be incorporated in a future study to investigate the effect of Linac or QA device and methods on prediction/classification accuracy and generalization performance of machine learning model. Because this study is an exploratory study, only GYN and H&N VMAT plans were used to train the machine learning model. VMAT plans from other anatomic sites will be evaluated in a future study to investigate the effect of treatment sites on model training and prediction.

Conclusions

The PL model could accurately predict patient-specific QA results for the majority of VMAT plans at 3%/3 mm and 3%/2 mm gamma criteria. Although the PL model had a higher specificity, a much better sensitivity was achieved by the RF model. The RF model with 100% sensitivity was preferred for QA results classification. Machine learning is proven to be a useful tool to assist measurement-based patient-specific QA and to reduce the QA workload.

References

1. Popescu CC, Olivotto IA, Beckham WA, et al. Volumetric modulated arc therapy improves dosimetry and reduces treatment time compared to conventional intensity-modulated radiotherapy for locoregional radiotherapy of left-sided breast cancer and internal mammary nodes. *Int J Radiat Oncol Biol Phys* 2010;76:287-295.
2. Nicolini G, Ghosh-Laskar S, Shrivastava SK, et al. Volumetric modulation arc radiotherapy with flattening filter-free beams compared with static gantry IMRT and 3D conformal radiotherapy for advanced esophageal cancer: A feasibility study. *Int J Radiat Oncol Biol Phys* 2012;84:553-560.
3. Fog LS, Rasmussen JF, Aznar M, et al. A closer look at RapidArc® radiosurgery plans using very small fields. *Phys Med Biol* 2011;56:1853-1863.
4. Ong CL, Cuijpers JP, Senan S, et al. Impact of the calculation resolution of AAA for small fields and RapidArc treatment plans. *Med Phys* 2011;38:4471-4479.
5. Van Esch A, Huyskens DP, Behrens CF, et al. Implementing RapidArc into clinical routine: A comprehensive program from machine QA to TPS validation and patient QA. *Med Phys* 2011;38:5146-5166.
6. Klein EE, Hanley J, Bayouth J, et al. Task Group 142 report: Quality assurance of medical accelerators. *Med Phys* 2009;36:4197-4212.
7. Smilowitz JB, Das IJ, Feygelman V, et al. AAPM medical physics practice guideline 5.a.: Commissioning and QA of treatment planning dose calculations - Megavoltage photon and electron beams. *J Appl Clin Med Phys* 2015;16:14-34.
8. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. *Med Phys* 2018;45:e53-e58.
9. Heilemann G, Poppe B, Laub W. On the sensitivity of common gamma-index evaluation methods to MLC misalignments in Rapidarc quality assurance. *Med Phys* 2013;40:031702.
10. Nelms BE, Chan MF, Jarry G, et al. Evaluating IMRT and VMAT dose accuracy: Practical examples of failure to detect systematic errors when applying a commonly used metric and action levels. *Med Phys* 2013;40:111722.
11. Nelms BE, Zhen H, Tomé WA. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. *Med Phys* 2011;38:1037-1044.
12. Carrasco P, Jornet N, Latorre A, et al. 3D DVH-based metric analysis versus per-beam planar analysis in IMRT pretreatment verification. *Med Phys* 2012;39:5040-5049.
13. Valdes G, Scheuermann R, Hung CY, et al. A mathematical framework for virtual IMRT QA using machine learning. *Med Phys* 2016;43:4323.
14. Valdes G, Chan MF, Lim SB, et al. IMRT QA using machine learning: A multi-institutional validation. *J Appl Clin Med Phys* 2017;18:279-284.
15. Interian Y, Rideout V, Kearney VP, et al. Deep nets vs expert designed features in medical physics: An IMRT QA case study. *Med Phys* 2018;45:2672-2680.
16. Tomori S, Kadoya N, Takayama Y, et al. A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance [e-pub ahead of print]. Med Phys. https://doi.org/10.1002/mp.13112. Accessed July 31, 2018.
17. Park JM, Park SY, Kim H, et al. Modulation indices for volumetric modulated arc therapy. *Phys Med Biol* 2014;59:7315-7340.
18. Park JM, Wu HG, Kim JH, et al. The effect of MLC speed and acceleration on the plan delivery accuracy of VMAT. *Br J Radiol* 2015;88:20140698.
19. Crowe SB, Kairn T, Middlebrook N, et al. Examination of the properties of IMRT and VMAT beams and evaluation against pre-treatment quality assurance results. *Phys Med Biol* 2015;60:2587-2601.
20. McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. *Med Phys* 2010;37:505-515.
21. Du W, Cho SH, Zhang X, et al. Quantification of beam complexity in intensity-modulated radiation therapy treatment plans. *Med Phys* 2014;41:021716.
22. Younge KC, Matuszak MM, Moran JM, et al. Penalization of aperture complexity in inversely planned volumetric modulated arc therapy. *Med Phys* 2012;39:7160-7170.
23. Götstedt J, Karlsson Hauer A, Bäck A. Development and evaluation of aperture-based complexity metrics using film and EPID measurements of static MLC openings. *Med Phys* 2015;42:3911-3921.
24. Masi L, Doro R, Favuzza V, et al. Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy. *Med Phys* 2013;40:071718.
25. Nauta M, Villarreal-Barajas JE, Tambasco M. Fractal analysis for assessing the level of modulation of IMRT fields. *Med Phys* 2011;38:5385-5393.
26. Hussein M, Rowshanfarzad P, Ebert MA, et al. A comparison of the gamma index analysis in various commercial IMRT/VMAT QA systems. *Radiother Oncol* 2013;109:370-376.
27. Agnew CE, McGarry CK. A tool to include gamma analysis software into a quality assurance program. *Radiother Oncol* 2016;118:568-573.
28. Kerns JR, Childress N, Kry SF. A multi-institution evaluation of MLC log files and performance in IMRT delivery. *Radiat Oncol* 2014;9:176.
29. Ford EC, Terezakis S, Souranis A, et al. Quality control quantification (QCQ): A tool to measure the value of quality control checks in radiation oncology. *Int J Radiat Oncol Biol Phys* 2012;84:e263-e269.
30. Passarge M, Fix MK, Manser P, et al. A Swiss cheese error detection method for real-time EPID-based quality assurance and error prevention. *Med Phys* 2017;44:1212-1223.
31. McCowan PM, Asuni G, van Beek T, et al. A model-based 3D patient-specific pre-treatment QA method for VMAT using the EPID. *Phys Med Biol* 2017;62:1600-1612.
32. Fan J, Li J, Chen L, et al. A practical Monte Carlo MU verification tool for IMRT quality assurance. *Phys Med Biol* 2006;51:2503-2515.
33. Leal A, Sánchez-Doblado F, Arráns R, et al. Routine IMRT verification by means of an automated Monte Carlo simulation system. *Int J Radiat Oncol Biol Phys* 2003;56:58-68.
34. Pawlicki T, Yoo S, Court LE, et al. Moving from IMRT QA measurements toward independent computer calculations using control charts. *Radiother Oncol* 2008;89:330-337.