Biomedical Physics & Engineering Express

RECEIVED
25 March 2022

REVISED
2 July 2022

ACCEPTED FOR PUBLICATION
20 July 2022

PUBLISHED
2 August 2022

PAPER

# Prediction of VMAT delivery accuracy using plan modulation complexity score and log-files analysis

**Pietro Viola[1], Carmela Romano[1], Maurizio Craus[1], Gabriella Macchia[2], Milly Buwenge[3], Luca Indovina[4], Vincenzo Valentini[5], Alessio G Morganti[3,6], Francesco Deodato[2] and Savino Cilla[1,*]**

[1] Medical Physics Unit, Gemelli Molise Hospital, Campobasso, Italy
[2] Radiation Oncology Unit, Gemelli Molise Hospital, Campobasso, Italy
[3] Radiation Oncology Department, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy
[4] Medical Physics Unit, Fondazione Policlinico Universitario A. Gemelli, Roma, Italy
[5] Radiation Oncology Department, Fondazione Policlinico Universitario A. Gemelli, Roma, Italy
[6] DIMES, Alma Mater Studiorum, Bologna University, Bologna, Italy
[*] Author to whom any correspondence should be addressed.

**E-mail:** savinocilla@gmail.com **and** savino.cilla@gemellimolise.it

**Keywords:** VMAT, complexity index, log-files

## Abstract

The purpose of this study was to develop a predictive model based on plan complexity metrics and linac log-files analysis to classify the dosimetric accuracy of VMAT plans. A total of 612 VMAT plans, corresponding to 1224 arcs, were analyzed. All VMAT arcs underwent pre-treatment verification that was performed by means of the dynamic log-files generated by the linac. The comparison of predicted (by TPS) and measured (by log-files) integral fluences was performed using $\gamma$-analysis in terms of the percentage of points with $\gamma$-value smaller than one ($\gamma$%) and using a stringent 2%(local)/2 mm criteria. This $\gamma$-analysis was performed by a commercial software LinacWatch. The action limits (AL) were derived from the mean values, standard deviations and the confidence limit (CL) of the $\gamma$% distribution. A plan complexity metric, the modulation complexity score (MCS), based on the aperture beam area variability and leaf sequence variability was used as input variable of the model. A binary logistic regression (LR) model was developed to classify QA results as 'pass' ($\gamma\% \geqslant$ AL) or 'fail' ($\gamma\% <$ AL). Receiver operator characteristics (ROC) curves were used to determine the optimal MCS threshold to flag 'failed' plans that need to be re-optimized. The model reliability was evaluated stratifying the plans in training, validation and testing groups. The confidence and action limits for $\gamma$% were found 20.1% and 79.9%, respectively. The accuracy of the model for the training and testing dataset was 97.4% and 98.0%, respectively. The optimal MCS threshold value for the identification of failed plans was 0.142, providing a true positive rate able to flag the plans failing QA of 91%. In clinical routine, the use of this MCS threshold may allow the prompt identification of overly modulated plans, then reducing the number of QA failures and improving the quality of VMAT plans used for treatment.

## 1. Introduction

Volumetric modulated arc therapy (VMAT) is a rotational form of intensity-modulated technique (IMRT), in which highly conformal doses can be realized by a complex interplay between the speed of gantry rotation, the multileaf collimator (MLC) shape, and the linac dose rate [1]. The resulting improvement in target volume conformity and normal tissue sparing resulted in significant reduction acute and late toxicities [2].

Modern planning procedures use complex advanced algorithms for dose optimization and calculation and are vulnerable to several uncertainty sources, including small fields modelling, dose calculation accuracy, tongue-and-groove effect and interleaf leakage and transmission [3, 4].

The solution space is highly degenerate, meaning that a multitude of different output plans may produce similar calculated dose distributions. In some anatomical sites, the complex relations between target volumes and organs-at-risk and the complexity of prescribed dose distributions require highly modulated plans that may affect the delivery accuracy.

Due to this increased complexity, patient-specific quality assurance (PSQA) has been strongly recommended by various professional organizations [5, 6], although the debate on the need to perform dosimetric measurements for each patient is still open [7]. PSQA consists in individualized measurements, usually performed before the first treatment fraction, using a large variety of phantoms and dosimetric systems, including two or three-dimensional arrays of ionization chamber of diodes and gafchromic films, and can be very time-consuming.

Because the agreement between calculated and measured complex dose distributions is expected to decrease as plan modulation increases, a large number of metrics have been defined using plan and machine properties (fluence, MLC aperture, position and displacement, gantry speed and dose rate variations, number of monitor units MU) to quantify plan complexity [8]. One of these metrics, the modulation complexity score (MCS), was initially found highly sensitive to delivery accuracy for both IMRT and VMAT techniques [9, 10]. Successively, several studies [11–13], focused on the predictivity power of these complexity metrics for dose delivery accuracy, provided discordant results and translated in lack of consensus and guidelines.

Recent approaches using machine learning methods (as support vector machine, gradient boosting, random forest, Poisson regression, regression tree analysis and deep learning networks) are reporting preliminary success for the prediction and classification of IMRT/VMAT plan quality [14] and delivery accuracy [15, 16].

On the other hand, machine delivery log-file analysis has been introduced as an alternative, effective and efficient approach for PSQA of IMRT and VMAT delivery accuracy [17–20]. Log-files are dynamic temporal tracking files, written in a proprietary format, containing a high frequency recording of the main parameter's characteristic of the linac during irradiation. In our clinic, we implemented a dedicated software (LinacWatch, Qualiformed, La Roche-sur-Yon, FRA), able to decrypt the linac log-files and to generate an irradiated fluence map that can be compared with the predicted one in terms of gamma-index passing rate.

The aim of this research was to evaluate the ability of the modulation complexity score, together with linac log-file analysis, to successfully predict plan deliverability in a large plan population. Because of the widespread implementation of VMAT in clinical practice, a successful prediction of patient-specific QA outcomes should result in a significant increase in PSQA efficiency.

## 2. Materials and methods

### 2.1. Treatment plans

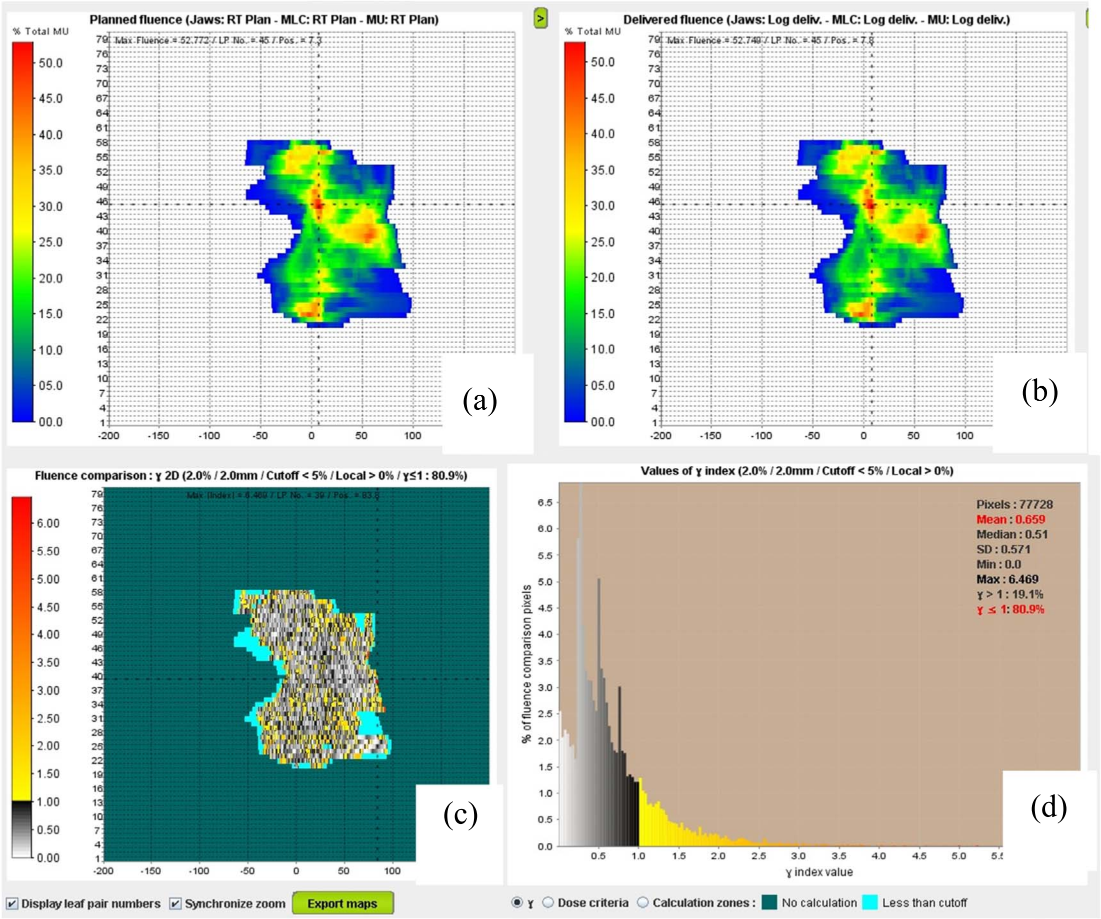
A total of 612 consecutive treatment plans were analyzed during 2021. All plans were optimized in a VMAT 'dual-arc' modality for a total of 1224 VMAT arcs measurements. Plans corresponds to a large number of clinical sites treated in our departments with different complexity, including brain, head and neck, lungs, oesophagus, breast, abdomen, pancreas, prostate, pelvis, and spine tumors.

All plans were generated using the Autoplanning module implemented into Pinnacle[3] treatment planning system version 16.2 (Philips, Medical Systems, Fitchburg, WI). This is a template-based planning engine that uses an iterative approach of progressive optimization that mimic all the steps of experienced and skilled planners, as well described in literature [21]. All dose distributions were calculated using the collapsed cone convolution algorithm with a dose calculation grid of 2.0 mm. Each treatment was delivered by one of the two matched VersaHD linear accelerator (Elekta, Crawley, UK), equipped with the high-definition Agility multileaf collimator (160 leaves with 0.5 cm length at isocenter).

### 2.2. The linacwatch software

The LinacWatch software (Qualiformed, La Roche-sur-Yon, FRA) analyzed the dynamic log-files generated by linacs following the delivery of each radiation beam in IMRT or VMAT treatments. It allows an accurate verification in real time of the compliance of the linac performance during the treatment session with that scheduled by the treatment planning system, covering the position of each moving leaf of the MLC, the position of the jaws, the MU number, the delivered integral fluence, the gantry and collimator rotation angles and the beam off lags. For an Elekta VersaHD linac, all aforementioned data are recorded and transferred to the LinacWatch every 250 ms. Linacwatch is able to calculate the integrated fluence, i.e. the MU delivered per unit of integrated surface over the total duration of the radiation session) at 100 cm from the radiation source in a very short time (less than a second). This calculation is carried out at each control point of the log-file and RT-plan Dicom file from the leaves position and the MU delivered; then the integrated fluence is calculated by LinacWatch by 'painting' the intensity corresponding to the difference in MU delivered between a given control point and the preceding one. In addition, LinacWatch directly supplies the modulation complexity score (MCS), as explained in the next section.

Figure 1 shows the LinacWatch graphic interface reporting (a) the calculated fluence of treatment planning system, (b) the obtained fluence from log-files analysis, (c) the fluence comparison and (d) the histogram of $\gamma$-index values.


Planned fluence (Jaws: RT Plan - MLC: RT Plan - MU: RT Plan)

Max Fluence = 52.772 / LP No. = 45 / Pos. = 7.3

Delivered fluence (Jaws: Log deliv. - MLC: Log deliv. - MU: Log deliv.)

Max Fluence = 52.749 / LP No. = 45 / Pos. = 7.8

Fluence comparison : γ 2D (2.0% / 2.0mm / Cutoff < 5% / Local > 0% / γ≤1 : 80.9%)

Max (Index) = 6.469 / LP No. = 39 / Pos. = 83.8

Values of γ index (2.0% / 2.0mm / Cutoff < 5% / Local > 0%)

Pixels : 77728
Mean : 0.659
Median : 0.51
SD : 0.571
Min : 0.0
Max : 6.469
γ > 1 : 19.1%
γ ≤ 1: 80.9%

☑ Display leaf pair numbers ☑ Synchronize zoom Export maps

● γ ○ Dose criteria ○ Calculation zones : 🟩 No calculation 🟦 Less than cutoff

**Figure 1.** Computer screen showing the LinacWatch user interface for log-files analysis. The screen is subdivided in four areas. From the left: (a) the calculated fluence of treatment planning system, (b) the obtained fluence from log-files analysis, (c) the fluence comparison and (d) the histogram of $\gamma$-index values.

### 2.3. The modulation complexity score (MCS)

The plan complexity was assessed using the Modulation Complexity Score, MCS, originally introduced by McNiven *et al* [9]. MCS was initially designed for step-and-shoot treatments and it was later adapted by Masi *et al* [10] to VMAT treatments. This score characterises the fluence modulation with two parameters:

- the aperture area variability, AAV, that represents the variability in the shape of segments, i.e. the difference between leaf pair apertures for any segment compared to the maximum leaf separation in the beam, defined as:

$$AAV_{cp} = \left(\frac{\sum_1^N (p_{i,left\_bank} - p_{i,right\_bank})}{\sum_1^N (\max(p_{i,left\_bank}) - \max(p_{i,right\_bank}))}\right)$$

- the leaf sequence variability, LSV, that represents the variability in the area of segments, i.e. the variation between adjacent leaves in the same leaf bank, defined as:

$$LSV_{cp} = \left(\frac{\sum_1^{N-1}(p_{\max} - |(p_i - p_{i+1})|)}{(N-1)\times p_{\max}}\right)_{left\_bank} \times \left(\frac{\sum_1^{N-1}(p_{\max} - |(p_i - p_{i+1})|)}{(N-1)\times p_{\max}}\right)_{right\_bank}$$

where $p_i$ is the coordinate of the $i_{th}$ leaf position, $p_{max}$ is the maximum distance between positions for a givent leaf bank, summed over all control point and N is the number of leaves in the bank.

The MCS for an arc is then the product of LSV and AAV weighted by the relative number of monitor units:

$$MCS_{arc} = \sum_1^N \left[\left(\frac{AAV_{cp,i} + AAV_{cp,i+1}}{2}\right)\right] \times \left(\frac{LSV_{cp,i} + LSV_{cp,i+1}}{2}\right) \times \frac{MU_{cpi,i+1}}{MU_{arc}}$$

where $MU_{cpi,i+1}$ indicates the MUs delivered between two successive control points (cpi and cp(i+1)).

This definition considers that during a VMAT arc, MUs are delivered continuously between adjacent control points and therefore, the computation of the

MCS must consider the product of the mean values between adjacent control point of $LSV_{cp}$ and $AAV_{cp}$. The product is then weighted by the relative number of monitor units delivered between two consecutive control points and then summed over all CP in the arc.

From these definitions, the MCS score uses a fixed range from 0 to 1, where the MCS is 1 for a simple unmodulated field and it approaches 0 for complex, highly modulated fields.

### 2.4. Pre-treatment verification and γ-analysis

The comparison of plan predicted and log-files integral fluences was performed using γ-analysis in terms of the percentage of points with γ-value smaller than one (γ%), using a stringent 2%(local)/2 mm criteria.

Following the recommendations of the AAPM Task Group No. 218 document [5], the action limits (AL) were derived from the mean values and standard deviations of γ%, providing the confidence limit (CL):

$$CL = (100 - \text{mean}) + 1.96\,\sigma$$

Then, the γ% of each plan is requested to be higher than AL = (100−CL).

QA results with γ% less than or equal to AL are defined as 'failed' plans, whereas γ% values larger than AL are defined as 'pass' plans.

### 2.5. Test case

In order to better understand the underlying effect of the linac delivery capability, we performed a more in-depth analysis on a representative prostate case. This case was optimized by six different plans with increasingly tighter constraints on rectal sparing (the main organ-at-risk), dose gradient and MLC motion parameters to force an increase of plan complexity (i.e. a decreased of MCS values). In particular, rectal mean doses and the volume of rectum receiving 50Gy and 60 Gy were progressively reduced, while at the same time the requests for increasingly steep dose gradients have been accentuated. All plans had the same dosimetric objectives on the target volume so that they can be considered clinically effective. The relationship between MCS and γ% was then investigated.

### 2.6. Modelling and statistical analysis

The overall dataset was split into a training and validation set used for model development and cross-validation and a testing set used for evaluation. The training/validation and testing set included 979 and 245 arcs respectively (i.e. 80%/20% split). To ensure that the testing set was representative of the whole population of target values (γ%), the dataset was split using a stratified technique based on the distribution of γ%.

A logistic regression analysis was performed to model the probability of predicting γ% using the MCS metric as input variable. Logistic regression is a classical algorithm that is usually used for binary classification tasks. This model calculates the class membership probability for one of the two categories in the dataset ($y_i = 0$ or 1) using a logistic equation:

$$p_i = \frac{e^{(\beta_0+\beta_1 \cdot x_i)}}{1 + e^{(\beta_0+\beta_1 \cdot x_i)}}$$

This equation can be linearized by the following transformation

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 \cdot x_i$$

where $x_i$ are the explanatory variables.

The left-hand side is termed the logit, i.e. logistic unit, and $\beta$ are the regression coefficients ($\beta_0$ and $\beta_1$ are also known as intercept and rate parameters, respectively). The formula illustrates that the probability of the dependent variable of an interested outcome is equal to the value of the logistic function of the linear regression expression. Because the value of the linear regression expression can vary from negative to positive infinity, after transformation, the resulting expression for the probability $p_i$ ranges between 0 and 1.

The best parameter estimates are the ones that maximize the likelihood of the statistical model actually producing the observed data:

$$\ln(L) = \sum_{1}^{N}\left[\ln(1 - p_i) + y_i\left(\frac{p_i}{1 - p_i}\right)\right]$$

or, rewritten in term of the 'logistic loss' function $L_{log}$:

$$L_{log} = \ln(L)$$
$$= -N\sum_{1}^{N}[-\ln(1 + e^{(\beta_0+\beta_1 \cdot x_i)}) + y_i(\beta_0 + \beta_1 \cdot x_i)]$$

where N is the size of the sample.

In order to avoid overfitting, the logistic loss function was modified adding a penalty term, the L2 norm, which effectively shrinks the estimates of the coefficients toward zero. The new loss function is:

$$L_{log} + \lambda\sum_{1}^{p}\beta_j^2$$

where j is the number of coefficients in the model. This penalized loss function is also called 'Ridge regression'. The optimal value of the regularization parameter $\lambda$ was determined through 10-fold cross-validation in the training set.

The goodness of the logistic regression model fit was evaluated by the Hosmer–Lemeshow test. Specifically, this test calculates if the observed event rates match the expected event rates in population subgroups.

Following the suggestions of Carlone *et al* [22], receiver operating characteristic (ROC) curves were used to determine an unbiased method to set threshold criteria. Plans with a MCS value below a given threshold and γ% below AL (i.e. failing pretreatment QA) are true-positive (TP). Similarly, false-positive

**Table 1.** Results of plan complexity and patient-specific QA with 2%(local)/2 mm $\gamma$-criteria.

| Metric | Training dataset | Testing dataset |
|---|---|---|
| Mean ± SD | 0.243 ± 0.086 | 0.236 ± 0.081 |
| Range | 0.078 – 0.565 | 0.077 – 0.541 |
| $\gamma$% | | |
| Mean ± SD (%) | 87.9 ± 4.1 | 87.8 ± 4.0 |
| Range (%) | 67.3–99.1 | 69.7–98.0 |
| Number of failed plans (N) | 43 | 12 |

plans (FP) are defined as the plans with a MCS value less than a given threshold value, but a $\gamma$% above AL (i.e. passing the QA). Plans with MCS over the threshold and passing QA are a true-negative (TN) while plans with MCS larger than the threshold and failing QA procedures are false negative (FN).

The classification performances are evaluated by calculating the sensitivity, specificity and accuracy as following:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \cdot 100$$

$$\text{Specificity} = \frac{TN}{TN + FP} \cdot 100$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100$$

In this context, the sensitivity of the test is its ability to determine the 'failed' plans correctly. The specificity of the test is its ability to determine the 'pass' plans correctly. Lastly, the accuracy of the test is its ability to differentiate failed and pass plans correctly.

The most optimal threshold for MCS is identified as the value that will optimize the true positive fraction and the true negative fraction. The diagnostic performance is evaluated by the area under the ROC curve (AUC). The closer the AUC is to 1.0, the better its performance; the closer the AUC is to 0.5, less useful the diagnostic test is.

## 3. Results

### 3.1. $\gamma$% and MCS distributions

Table 1 shows the overall results of QA procedure. A Shapiro-Wilk test was performed to test the normally of $\gamma$% and MCS values. Both $\gamma$% and MCS data distributions were found to be normally distributed in both training and testing dataset ($p < 0.05$). The mean ± SD of $\gamma$% were 87.9% ± 4.1% and 87.8% ± 4.0% in the training and testing datasets, respectively; minimum values were 67.3% and 69.7% respectively. There was no significant difference in $\gamma$% between the training and testing datasets ($p = 0.593$). The confidence and action limits were found 20.1% and 79.9%, respectively. For simplicity, a value of 80% was adopted for $\gamma$% action limit. The mean ± SD of MCS were 0.243 ± 0.086 and 0.236 ± 0.081 in the training and testing datasets, respectively; minimum values were 0.078 and 0.077. No significant difference in MCS distributions was found between the training and testing datasets ($p = 0.253$).

Figure 2 presents the scatter plot of the correlation between $\gamma$% and MCS for the training and testing datasets.

Figure 3 presents the scatter plot of $\gamma$% against MCS for the single prostate test case and the dose-volume histograms for the prostate (PTV) and the rectum obtained after six different optimization cycles with increasingly tighter constraints on rectal sparing and dose gradient (to force an increase of plan complexity) but with the same dosimetric objectives on the target volume. A strong correlation between the two variables was observed ($R^2 = 0.83$).

Figure 4 reports the $\gamma$% and MCS values obtained by LinacWatch for different anatomical sites and techniques. The mean ± SD for $\gamma$% was 88.1% ± 3.0%, 84.7% ± 4.3%, 88.2% ± 2.9%, 87.2% ± 3.8%, 88.8% ± 3.8% and 94.3% ± 3.3% for anorectal, head-neck, prostate, gynaecologic, brain and stereotactic body treatments (SBRT), respectively. Similarly, the mean ± SD for MCS was 0.235 ± 0.077, 0.191 ± 0.052, 0.239 ± 0.071, 0.224 ± 0.069, 0.279 ± 0.071 and 0.444 ± 0.080 for anorectal, head-neck, prostate, gynaecologic, brain and SBRT, respectively. A Kruskal-Wallis test reported significant differences among the different groups ($p < 0.001$). The post-hoc test showed that these differences are statistically significant when the head-and-neck and SBRT plans are compared to others groups. Head-and-neck plans reported the lower agreement for $\gamma$%, with 8.2% of plans below the action limit of 80% (14 plans out of 171). Gynaecological and brain tumours sites reported 1.8% (3 out 167) and 0.8% (1 out 118) of plans below the 80% action limit.

All anorectal, prostate and SBRT plans were considered optimal for dosimetric pre-treatment purposes. In particular, despite the stringent 2%(local)/2 mm criterion, SBRT plans reported a very high agreement between predicted and measured fluences (mean $\gamma$% of 94.3%), due to their low degree of modulation, as expressed by the high values of MCS (mean MCS of 0.444).

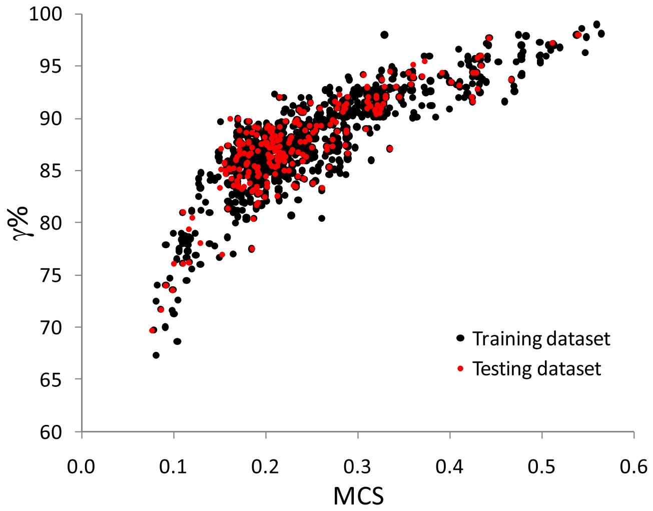


**Figure 2.** Scatter plot of γ% against MCS. Plans with γ% above 80% are those passing QA pretreatment verification. Black and red dots refer to plans within the training and the testing datasets, respectively.

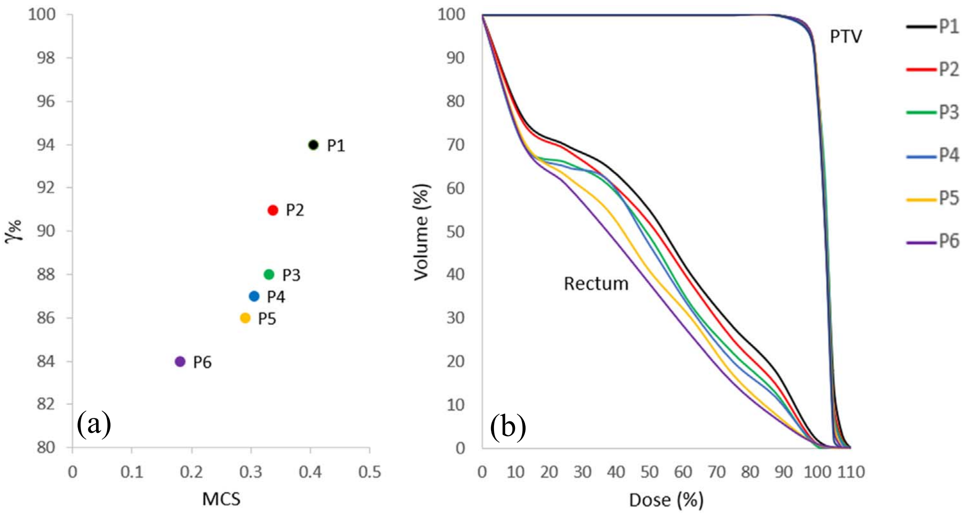


**Figure 3.** (a) Scatter plot of γ% at 2%(local)/2 mm against MCS for six plans of a representative prostate case optimized with increasingly tighter constraints on rectal sparing and dose gradient (from P1 to P6), and (b) dose-volume histograms for the planning target volume (PTV) and the rectum for the six different plans.

### 3.2. Logistic regression and ROC curves

The performance of logistic regression models are reported in figure 5. The Hosmer-Lemeshow test reported p-values of 1.000 and 0.872 in the two datasets, meaning that the logistic regression models provided an optimal fit.

The same values of 0.130 for the MCS metric was identified in both training and testing datasets as the threshold corresponding to a probability of 50% of observing a 'failed' plan. In other words, according to this mathematical model, plans with MCS values less than 0.130 have a much higher probability to fail pre-treatment QA.

However, the value of particular interest is the false positive rate (FP), i.e. the number of plans classified as 'pass' by the model whereas they should be classified as 'failed' plans. Since the false positive rate is more important than false negatives, the model was tuned in order to obtain larger specificity at the cost of lower sensitivity.

Then, new threshold MCS values of 0.142 (CI95%: 0.136–0.152) and 0.142 (CI95%: 0.133–0.159) were considered in the training and testing datasets by the probability analysis to best discriminate the failed plans. Table 2 reports the results as confusion matrices. These are tables used to describe the performance of a classification model on a set of test data for which the true values are known. They are usually applied to binary classification in the form of 2 × 2 tables, representing the counts from predicted and actual values, i.e. the number of negative examples correctly classified (True Negative), the number of positive examples classified accurately (True Positive), the number of

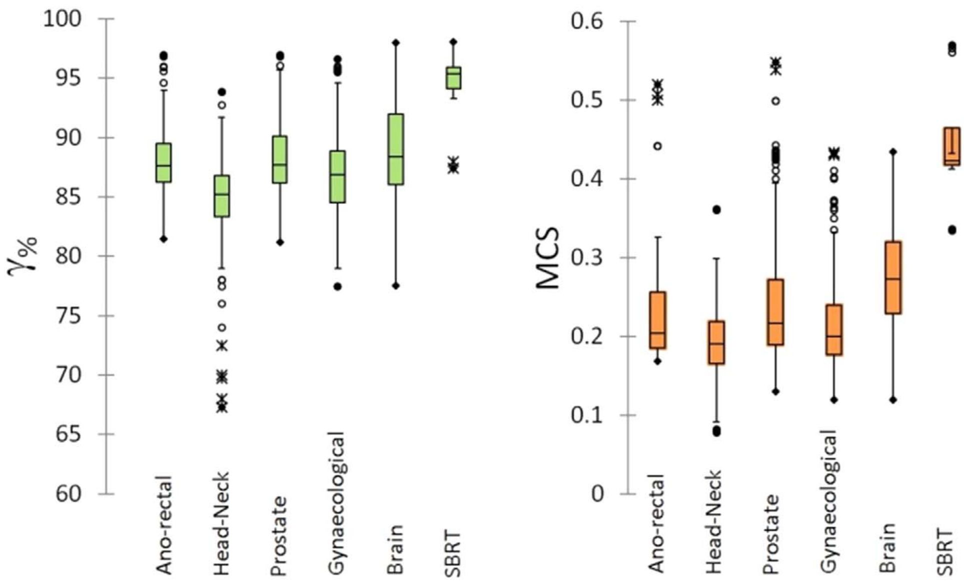


**Figure 4.** Box-plots depicting the $\gamma$% (left figure) and MCS (right figure) values obtained using LinacWatch for various anatomical sites and techniques. The central line marks the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the adjacent values, which are the most extreme data values that are not outliers, and the circles represent the outliers.

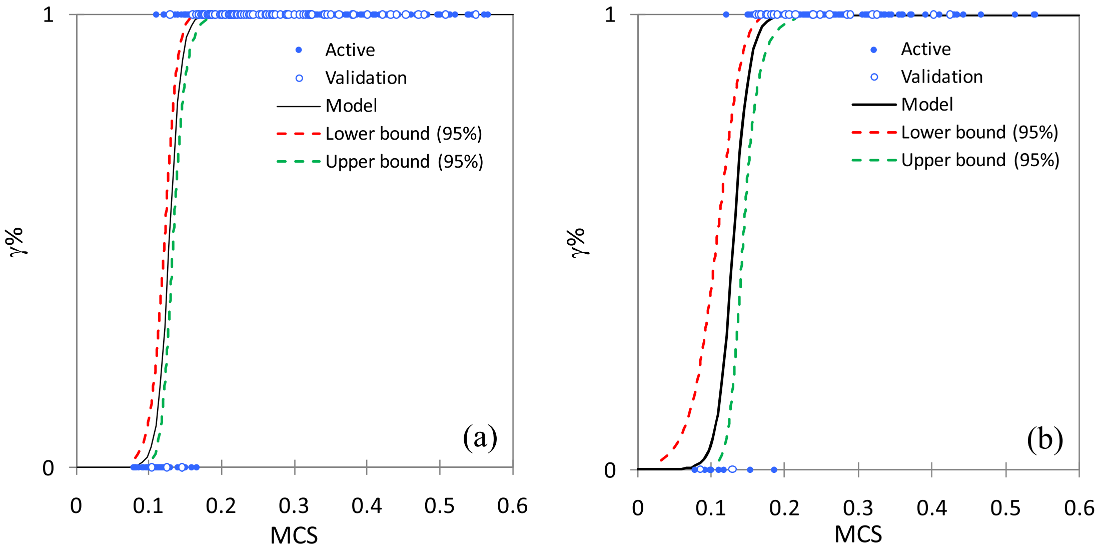


**Figure 5.** Sigmoid curves obtained by logistic regression analysis plotting $\gamma$% against MCS for (a) training and (b) testing datasets.

actual negative examples classified as positive (False Positive) and the number of actual positive examples classified as negative (False Negative).

The optimal value of the regularization parameter $\lambda$ minimizing the 10-fold cross-validation error was 0.025.

For 2%(local)/2 mm $\gamma$% classification using 80% as action limit, the sensitivity, specificity and accuracy of the model for the training dataset were 95.3%, 97.5% and 97.4%, respectively. Similarly, the sensitivity, specificity and accuracy of the model for the testing dataset were 91.7%, 98.3% and 98.0%, respectively. In particular, this means that with a threshold of 0.142, the MCS score correctly flagged more than 90% of plans that failed pretreatment PSQA, while incorrectly flagged 1.6% of plans that passed pretreatment verification.

Figure 6 depicts the ROC curves generated by varying the MCS threshold and plotting the true-positive rate versus the false-positive rate. The AUC was 0.996 and 0.978 for the training and testing datasets, respectively.

## 4. Discussion

In this study we demonstrated that log files-based gamma passing rates can be predicted with high

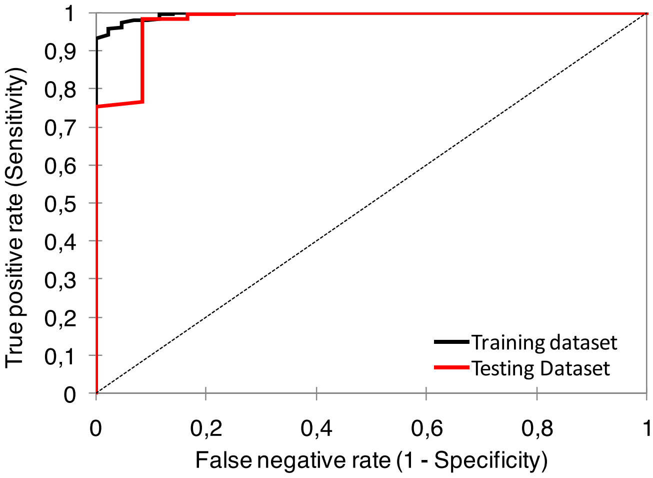


**Figure 6.** ROC plots of sensitivity versus (1-specificity) for the training (black line) and testing (red line) datasets. The diagonal line with AUC of 0.5 represent test whose outcome is not significantly different than a random guess.

**Table 2.** Confusion matrices used to describe the performance of the LR model for the $\gamma$% classification, for both the training and the testing datasets.

| *Training dataset* | | | | |
|---|---|---|---|---|
| from \ to | Fail (Predicted) | Pass (Predicted) | Total | % correct |
| Fail (Actual) | 41 | 2 | 43 | 95.3% |
| Pass (Actual) | 23 | 913 | 936 | 97.5% |
| Total | 64 | 915 | 979 | 97.4% |
| *Testing dataset* | | | | |
| from \ to | Fail (Predicted) | Pass (Predicted) | Total | % correct |
| Fail (Actual) | 11 | 1 | 12 | 91.7% |
| Pass (Actual) | 4 | 229 | 233 | 98.3% |
| Total | 15 | 230 | 245 | 98.0% |

accuracy using a logistic regression model. This approach is particularly advantageous because the dosimetric inaccuracy of a VMAT plan delivery may be accurately anticipated without the need of time-consuming dosimetric measurements. Therefore, given the widespread use of VMAT in clinical practice, a successful prediction of potential failures in patient-specific QA may translate in a significant increase in QA efficiency. For example, if a specific plan is unlikely to pass PSQA after gamma passing rate prediction, its plan complexity may be decreased during the optimization process in order to improve deliverability. Unacceptable plans could then be potentially removed 'a-priori', avoiding treatment delays due to the failure of VMAT QA.

This strategy may be highly beneficial in (a) resource-constrained nations, where machine time is often not available for PSQA, (b) after the implementation of adaptive radiotherapy techniques, where plans must be modified while the patient is still on treatment table or (c) in critical logistic conditions as the actual COVID-19 pandemic, where clinical and dosimetric workflows are deeply altered to reduce risks [23].

The logistic regression model developed in the present study for the 'a-priori' prediction of PSQA failure reported a very high accuracy >0.95 at 2%-2 mm in both training and testing datasets. Moreover, an optimal threshold was defined for MCS: a value of 0.14 was able to flag 'failed' plans, i.e. unacceptable plans that need to be re-optimized.

As shown in figure 2, a strong correlation between gamma passing rate and the MCS metric was found. This means that the heavy increase of complexity of a VMAT plan may impact negatively the actual dose delivered to patient. This is a critical argument because the predictivity of complexity metrics of dose delivery accuracy is still today debated, with discordant results translating in lack of consensus and guidelines. Agnew *et al* [11] used the VMAT PSQA results from 711 plans to validate the ability of a complexity metric to predict plan deliverability, reporting a true positive rate for correctly identifying plans failing PSQA of 44% and a false-positive rate was 7%. Nguyen *et al* [24] evaluated VMAT complexity metrics as a means of predicting phantom-based measurement results for treatments delivered on a Varian TrueBeam linacs. The authors reported a moderate correlation of MCS to gamma passing rate, with ROC analysis achieving a 60% true positive rate and a 9% false-positive rate to correctly identify complex plans. Masi *et al* [10] evaluated the effect complexity indexes on 142 VMAT dosimetric accuracy, reporting a high correlation between MCS and gamma index passing rates. On the contrary, Glenn *et al* [13] failed to report any correlations in evaluating different complexity metrics, including MCS, for 343 irradiations on an anthropomorphic head-and-neck phantom, comprising both IMRT and VMAT techniques. The authors reported weak correlations, limiting the predictive utility in assessing plan performance in terms of complexity metrics. Lastly, Xia *et al* [25] analysed the relationship between MCS and gamma pass rates for a total of 275 stereotactic radiosurgery and stereotactic body radiotherapy cases, reporting no conclusive quantifiable correlation between MCS gamma passing rate.

The results of our study highlighted that the MCS plan complexity metric has a strong impact on gamma passing rate, in agreement with the results obtained by Masi [10] and Agnew [11]. This is probably due to the fact that all of the data have been collected in the same institution, from two dosimetrically matched 'twins' LINACs and generated with the Autoplanning optimization engine. This last module demonstrated a major ability to reduce intra and inter-planner planning variability with respect to manual planning [18]. All these aspects had an impact on reducing data collection and QA procedure variability.

This study differs from previous studies for two main aspects. First, this study employed log-files analysis to predict pre-treatment QA results instead of phantom-based measurements. This way, the predicted results are not affected by the well-know variability due to the impact of PSQA measurements in terms of phantoms and dosimetric systems. A study performed by Hussein *et al* [26] evaluated the impact of different commercial dosimetric systems for IMRT and VMAT PSQA on the accuracy of $\gamma$-analysis results. The results shown that using the same pass-rate criteria, the different devices and software combinations exhibit varying levels of agreement with the predicted $\gamma$ analysis. Moreover, the tightening the gamma criteria increased measurement variability among the different QA instruments, with variances in mean and minimum percent up to 15% at 2%/2 mm. This is a crucial point because the accuracy for the prediction of $\gamma$% significantly worsens with more stringent criteria for $\gamma$-index analysis. Several studies [27, 28] have clearly shown that the widespread used 3%/3 mm criteria is insensitive in detecting clinically relevant errors. These authors suggested a retirement of the 3%/3 mm criteria as a primary metric of performance, and the adoption instead of tighter tolerances. Therefore, from this point of view, all predictive models for PSQA results should be trained using tighter criteria. In this study we reported that the use of log-files analysis for PSQA purposes may overcome the limitations of phantom-based systems, allowing the adoption of 2%(local)/2 mm criteria that is more effective in evaluating the accuracy of dose delivery [27, 28]. Obviously, it is very difficult to achieve an appealing passing rate (i.e. >90%) using 2%(local)/2 mm criteria. However, the spread continuum of evaluation results (rather than a cluster next to the maximum 100% value) provide a more useful statistical backdrop to fine-tune the system.

In addition, our analysis reported significant differences among different anatomical sites in terms of $\gamma$ % and MCS. In particular, head-and-neck plans reported the lower MCS values and the lower agreement for $\gamma$%, with 8.2% of plans below the action limit of 80%. This is an expected result due to the major planning challenges for this site, where it is difficult to manage the compromise between tumours irradiation and sparing of healthy tissue. In this case, while the gross and microscopic diseases must be adequately irradiated to doses sufficient for tumour control, a large number of adjacent radiosensitive organs-at-risk must be spared as much as possible to avoid serious long-term sequelae. Therefore, this complex trade-off between competing priorities require increasingly complex plans, i.e. plans with low MCS values [29]. On the other hand, SBRT plans are usually optimized for small convex-shaped lesions that do not require hard constraints for fluence modulation, then translating in less MLC motion complexity and higher MCS values.

In the future, a promising application of these predictive models for QA accuracy will be their direct integration into the treatment planning optimization stage. This way, new 'QA-based' metrics could be used in real-time by the optimizer engine to penalize the solutions that predicts lower QA results. This is an ongoing line of research, and a few investigators are today developing optimization algorithms with the aim to decrease the plan modulation complexity during the VMAT planning optimization process without affecting plan quality [30].

A few limitations of this study should be highlighted. As also reported in other studies [31, 32], the number of failing plans is usually very small and this translates in an unbalanced data distribution in the model training. In this study, we collected 43 and 12 VMAT arcs that failed pretreatment PSQA (i.e. with $\gamma_{\%}$<80% at 2%/2 mm) in the training and testing datasets, respectively. Although small, this number should guarantee the prediction accuracy of the model. In any case, to increase the number of failed plans in the training dataset, a multicentric collaborative research is encouraged. Secondary, our model was developed using treatment plans from the same institution and all generated by the same optimization module (Pinnacle Autoplanning) for two dosimetrically matched Elekta VersaHD linacs. Therefore, our model may not automatically apply to other institutions using different equipments. A future study is needed to evaluate the congruence of our finding using different TPS, linacs and QA devices. Thirdly, the LinacWatch software calculates an integrated fluence, i.e. a fluence integrated over the whole arc, which may potentially mask small delivery errors. Therefore, we adopted a stringent 2%(local)/2 mm gamma in order to increase the sensitivity and to partially compensate this effect. Moreover, in the present study we used a stringent 2%(local)/2 mm $\gamma$-criteria to evaluate the VMAT PSQA. The use of different $\gamma$-criteria (i.e. the widespread used 3%/3 mm) in other institutions may limit the applicability of our model because gamma pass rates are different when different $\gamma$-criteria are used.

Lastly, it must be underlined that the present strategy for patient-specific QA could only be implemented based on the assumption that accurate and adequate TPS and linac QA are performed consistently and continuously.

## 5. Conclusion

In conclusion, we investigated the ability of a predictive model based on the modulation complexity score and log-files analysis for classification of VMAT patient-specific QA results. A logistic regression was able to accurately predict VMAT PSQA failure results, correctly flagging more than 90% of plans that failed pre-treatment PSQA at 2%(local)/2 mm gamma criteria. This predictive model allows the prompt identification of overly modulated plans, then reducing the number of QA failures and improving the quality of VMAT plans used for treatment.

## Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

## Competing interests

The authors declare no competing interests.

## Conflict of interest statement

All authors declare that they have no conflicts of interest.

## Data sharing statement

Research data are stored in an institutional repository and will be shared upon request to the corresponding author.

## Funding

none.

## ORCID iDs

Gabriella Macchia https://orcid.org/0000-0002-0529-201X
Savino Cilla https://orcid.org/0000-0001-6711-350X

## References

[1] Yu C X and Tang G 2011 Intensity-Modulated Arc therapy: principles, technologies and clinical implementation *Phys. Med. Biol.* **56** R31–54

[2] Teoh M *et al* 2011 Volumetric modulated arc therapy: a review of current literature and clinical use in practice *Br. J. Radiol.* **84** 967–96

[3] Das I J, Ding G X and Ahnesjö A 2008 Small fields: nonequilibrium radiation dosimetry *Med. Phys.* **35** 206–15

[4] Oliver M *et al* 2010 Clinical significance of multileaf collimator positional errors for volumetric modulated arc therapy *Radiother. Oncol.* **97** 554–60

[5] Miften M *et al* 2018 Tolerance limits and methodologies for IMRT measurement based verification QA: Recommendations of AAPM Task Group *Med. Phys.* **45** e53–83

[6] Moran J M *et al* 2011 Safety considerations for IMRT: Executive summary *Pract Radiat Oncol.* **1** 190–5

[7] Smith J C, Dieterich S and Orton C G 2011 It is STILL necessary to validate each individual IMRT treatment plan with dosimetric measurements before delivery *Med. Phys.* **38** 553–5

[8] Antoine M *et al* 2019 Use of metrics to quantify IMRT and VMAT treatment plan complexity: A systematic review and perspectives *Phys Med.* **64** 98–108

[9] McNiven A L, Sharpe M B and Purdie T G 2010 A new metric for assessing IMRT modulation complexity and plan deliverability *Med. Phys.* **37** 505–15

[10] Masi L *et al* 2013 Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy *Med. Phys.* **40** 071718

[11] Agnew C E, Irvine D M and McGarry C K 2014 Correlation of phantom-based and log file patient-specific QA with complexity scores for VMAT *J Appl Clin Med Phys.* **15** 4994

[12] Hernandez V *et al* 2018 Comparison of complexity metrics for multi-institutional evaluations of treatment plans in radiotherapy *Phys Imag Radiat Oncol.* **5** 37–43

[13] Glenn M C *et al* 2018 Treatment plan complexity does not predict IROC Houston anthropomorphic head and neck phantom performance *Phys. Med. Biol.* **63** 20501

[14] Osman A F I and Maalej N M 2021 Applications of machine and deep learning to patient-specific IMRT/VMAT quality assurance *J Appl Clin Med Phys.* **22** 20–36

[15] Carlson J N, Park J M, Park S-Y, Park J I, Choi Y and Ye S-J 2016 A machine learning approach to the accurate prediction of multi-leaf collimator positional errors *Phys. Med. Biol.* **61** 2514

[16] Osman A F, Maalej N M and Jayesh K 2020 Prediction of the individual multileaf collimator positional deviations during dynamic IMRT delivery priori with artificial neural network *Med. Phys.* **47** 1421–30

[17] Chuang K C, Giles W and Adamson J 2020 On the use of trajectory log files for machine & patient specific QA *Biomed. Phys. Eng. Express* **4** 7

[18] Chuang K C, Giles W and Adamson J 2021 A tool for patient-specific prediction of delivery discrepancies in machine parameters using trajectory log files *Med. Phys.* **48** 978–90

[19] Stell A M, Li J G, Zeidan O A and Dempsey J F 2004 An extensive log-file analysis of step-and-shoot intensity modulated radiation therapy segment delivery errors *Med. Phys.* **31** 1593–602

[20] Rangaraj D *et al* 2013 Catching errors with patient-specific pretreatment machine log file analysis *Pract Radiat Oncol.* **3** 80–90

[21] Cilla S *et al* 2020 Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation *Sci Rep.* **10** 423

[22] Carlone M *et al* 2013 ROC analysis in patient specific quality assurance *Med. Phys.* **40** 042103

[23] Khan R *et al* 2020 Evolution of clinical radiotherapy physics practice under COVID-19 constraints *Radioth Oncol.* **148** 274–8

[24] Nguyen M and Chan G H 2020 Quantified VMAT plan complexity in relation to measurement-based quality assurance results *J Appl Clin Med Phys.* **21** 132–40

[25] Xia Y *et al* 2020 Application of TG-218 action limits to SRS and SBRT pre-treatment patient specific QA *Journal of Radiosurgery and SBRT.* **7** 135

[26] Hussein M *et al* 2013 A comparison of the gamma index analysis in various commercial IMRT/VMAT QA systems *Radiother. Oncol.* **109** 370–6

[27] Nelms B E *et al* 2013 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.* **40** 111722

[28] Heilemann G, Poppe B and Laub W 2013 On the sensitivity of common gamma-index evaluation methods to MLC misalignments in rapidarc quality assurance *Med. Phys.* **40** 031702

[29] Cilla S *et al* 2021 Personalized automation of treatment planning in head-neck cancer: A step forward for quality in radiation therapy? *Phys Med* **82** 7–16

[30] Ono T *et al* 2022 Development of a plan complexity mitigation algorithm based on gamma passing rate predictions for volumetric-modulated arc therapy *Med. Phys.* **49** 1793–802

[31] Valdes G *et al* 2016 A mathematical framework for virtual IMRT QA using machine learning *Med. Phys.* **43** 4323

[32] Tomori S *et al* 2021 Systematic method for a deep learning-based prediction model for gamma evaluation in patient-specific quality assurance of volumetric modulated arc therapy *Med. Phys.* **48** 1003–18