A mathematical framework for virtual IMRT QA using machine learning

G. Valdes,[a),b)] R. Scheuermann,[a)] C. Y. Hung, A. Olszanski, M. Bellerive, and T. D. Solberg
Radiation Oncology Department, Perelman Center for Advanced Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19123

(Received 16 March 2016; revised 31 May 2016; accepted for publication 1 June 2016; published 20 June 2016)

Purpose: It is common practice to perform patient-specific pretreatment verifications to the clinical delivery of IMRT. This process can be time-consuming and not altogether instructive due to the myriad sources that may produce a failing result. The purpose of this study was to develop an algorithm capable of predicting IMRT QA passing rates *a priori*.
Methods: From all treatment, 498 IMRT plans sites were planned in ECLIPSE version 11 and delivered using a dynamic sliding window technique on Clinac iX or TrueBeam Linacs. 3%/3 mm local dose/distance-to-agreement (DTA) was recorded using a commercial 2D diode array. Each plan was characterized by 78 metrics that describe different aspects of their complexity that could lead to disagreements between the calculated and measured dose. A Poisson regression with Lasso regularization was trained to learn the relation between the plan characteristics and each passing rate.
Results: Passing rates 3%/3 mm local dose/DTA can be predicted with an error smaller than 3% for all plans analyzed. The most important metrics to describe the passing rates were determined to be the MU factor (MU per Gy), small aperture score, irregularity factor, and fraction of the plan delivered at the corners of a 40 × 40 cm field. The higher the value of these metrics, the worse the passing rates.
Conclusions: The Virtual QA process predicts IMRT passing rates with a high likelihood, allows the detection of failures due to setup errors, and it is sensitive enough to detect small differences between matched Linacs. © *2016 American Association of Physicists in Medicine.* [http://dx.doi.org/10.1118/1.4953835]

Key words: virtual IMRT QA, machine learning, Poisson regression, local analysis

1. INTRODUCTION

It is common practice to measure 2D and 3D dose distributions prior to treating any patient using IMRT. To assess the integrity of the delivery, the measured dose and dose distribution(s) are compared with those predicted by the planning system. A number of different metrics are commonly used to assess agreement of the two distributions, including point-by-point percent dose difference and distance-to-agreement (DTA) and the gamma index which combines both DTA and dose evaluation.[1–3,12] These metrics are very sensitive to the method used to analyze the data. Analysis can be performed per beam or per plan (composite), and can be performed by normalizing the plans or fields with respect to the global maximum dose or local dose, typically with the requirement that 90% of the points pass the particular criterion in order for a plan to be considered clinically acceptable.[1] When 3% dose difference (global normalization) and 3 mm DTA are used, 3%/3 mm (G), the IMRT QA process is generally insensitive and unable to catch significant clinical errors.[4–10] Use of log files and independent dose calculations have been suggested as more efficient substitutes for measurement based IMRT QA.[4,6] However, none of these methods answer the fundamental question of whether a plan can be delivered within an acceptable inaccuracy given the limitations of the Linac and the treatment planning system (TPS). Therefore, to date, and despite its limitations, measurement based IMRT QA using the gamma comparison is the standard in most clinics, though a lack of consensus on how the analysis is performed (per field vs per plan, global vs local dose comparison, dose normalization point/value, % dose difference, DTA, threshold and interpolations) remains. If composite (per plan) local analysis, 2%/2 mm (L), is used more clinically, relevant errors could be detected.[5] Additionally, using a 90% passing rate threshold to decide whether a plan is acceptable has also been identified as a significant limitation to detecting clinically relevant errors, and the development of site specific thresholds has been suggested.[11] The 90% threshold for all plans with composite analysis and global normalization was derived in Task Group 119, assuming that gamma failing rates followed a Gaussian distribution and determining what threshold will include 95% of plans passing QA.[12] With today's level of accuracy, however, this threshold is easily met by most plans if proper commissioning has been performed. Additionally, as it will be shown in this paper, gamma-passing rate does not follow a Gaussian distribution and its value depends on the complexity of the plans, which typically depends on the disease site. Therefore, the derivation of the 90% is not mathematically accurate and does not take into account different plan complexity.

Even within a specific disease site, passing rates can show large variability, thus plan specific passing rates with respective confidence intervals should be the ultimate goal. In order to use a plan specific threshold, an algorithm capable of predicting passing rates with clinically relevant accuracy is needed. Once the dependence of the passing rate on the complexity metric has been removed, any variation in the passing rate would be due to random noise. In pursuit of that goal, different complexity metrics have been proposed to characterize treatment plans, with the intention of correlating them with IMRT QA passing rates. These metrics fall into two general categories: fluence map-based[13–16] or aperture-based.[16–21] Some of these metrics have been incorporated into TPSs, with a goal of producing plans that are more likely to pass QA.[22–24] In all cases, the correlation of the complexity metrics with the passing rate has been performed using a single metric at a time, with the expectation that this metric will explain most of the variance observed on the passing rates. Nevertheless, different aspects of the complexity of the plans that might interact in any given case are associated with plans failing QA. As a result, only weak correlations have been found between passing rates and these metrics.[13–21] Therefore, an algorithm that integrates different complexity metrics and is capable of predicting IMRT QA passing rates has yet to be developed. In the present paper, an algorithm that predicts passing rates *a priori* (Virtual IMRT QA) was developed using machine learning (ML) to maximize the prediction accuracy of the algorithm in out of sample data.

2. MATERIALS AND METHODS

2.A. Dataset

From multiple treatment sites, 498 IMRT plans [Breast (110), Central Nervous System (58), Gastro Intestinal (78), Genitourinary (64), Gynecologic (19), Head and Neck (5), Lung (134), and Pediatrics (30)] were planned using ECLIPSE version 11 (Varian Medical Systems, Palo Alto, CA) and delivered using a dynamic sliding window technique on either TrueBeam (HD MLCs) or one of our four nominally matched Clinac iX Linacs with millennium MLCs (Varian Medical Systems, Palo Alto, CA). Dose were calculated with 2.5 mm grid size. All plans were clinical plans used to treat patients at our institution. The Clinac selected for IMRT QA delivery was chosen randomly in each case depending on the unit that finished treatment first on any given day. 416 plans used 6 MV (6X) only, 32 plans used 15 MV only (15X), and 50 plans were mixed energy (mixed). The IMRT QA measurements were performed as part of our clinical routine using Mapcheck2 with the SNC software version 6.1 (Sun Nuclear, Melbourne, FL) with measurement uncertainty turned off. Additionally, ten random plans with passing rates within 3% of the predicted value were also repeated. The second delivery of these plans also resulted in passing rates that were within 3% of the predicted value. The Mapcheck2 was set up on the couch with 3 cm of solid water on top and the gantry angle over-ridden (delivery was always performed at 0° gantry angle) but collimator rotation was allowed. Every day before IMRT QA delivery, dose calibration was verified for a 100 MU, 10×10 cm reference field and the dose renormalized if there was a disagreement larger than 0.5% in absolute dose on the central axis diode. Only points with doses greater than 10% of the global maximum dose per plan were included in the analysis. Composite local 3% dose/3 mm DTA passing rate was recorded for all plans and saved to a Redcap database (Harvard Catalyst, Boston, MA).

2.B. Feature extraction and complexity metrics

Five main sources of errors that could lead to disagreement between the treatment planning system and delivery of the plans were identified: MLC leaf transmission, leaf end leakage (Dosimetric Leaf Gap, DLG), transmission through the jaws, tongue and groove effect, and charged particle equilibrium failure. For each plan, 78 different complexity metrics (referred also as features) were defined to characterize these categories. Plans were characterized using aperture-based complexity metrics based on the MLC leaf positions and fractional monitor units delivered per control point. We believe that aperture-based complexity metrics are more direct descriptors of the five main sources of discrepancies described above than fluence-based complexity metrics as they represent the delivery parameters utilized by the treatment machine and as such may offer better insight into the disagreement between the calculated and measured dose. Geometrical complexity metrics as well as those weighted by the fractional monitor units for each aperture were included. Complexity metrics were computed as the average over all control points and beams or as the average of the complete irradiated area outline (CIAO). In some cases, the maximum and minimum values of the complexity metric for a given beam of the respective plan were also included. In general, these complexity metrics (features) are as follows: the fraction of MUs per dose delivered, energy, type of Linac, jaw position, collimator angle, distribution of MLC leaf pair gaps (up to the fifth moment of the distribution), ratio of the area of MLC within the jaws aperture, area, perimeter, aperture irregularity as defined by Du *et al.*,[17] fraction of the area of the plan delivered within circles of different radii centered at isocenter (5, 10, and 20 cm), fraction of MLC leaf gaps with an opening smaller than a certain value average over all beams and control points, 2, 5, 10, and 20 mm (small aperture score, SAS, as defined by Crowe *et al.*[21]) or the maximum value of those for a given beam of the plan (only averaged over the control points), and the fraction of area receiving dose through the jaws and fractional area receiving dose from different number of beams. A detail list of all complexity metrics has been included in the supplementary material.[25] All complexity metrics were calculated using a combination of SQL queries on the ECLIPSE database and using our own MATLAB functions (Matlab R 2014a, Natick, MA). Before being used in the algorithm, complexity metrics were rescaled using their mean values and normalized by their standard deviations. Additionally, the effect of MLC speed was modeled by dividing each opposing MLC gap by the monitor units of that control point and then averaging these values over all control points and all beams for a given plan. This metric was not selected by the algorithm and as it did not improve the accuracy, and therefore the effect of MLC speed is not discussed further.

2.C. Plan visualization

As each plan is characterized by a vector of 78 complexity metrics, direct visualization of the effect of different metrics on the passing rates is prohibitive. Therefore, a principal component analysis (PCA) was developed to find the dimensions where the complexity metrics have the biggest variance.[26] PCA performs this task by finding the dimensions or orthogonal base (vectors) on which if every vector $\mathbf{x}$ is projected from the feature space matrix $\mathbf{X}$, the variance of the final components of each vector is maximized. Each of these dimensions is found as a linear combination of the initial complexity metrics (features). PCA then, transform the vectors $\mathbf{X}$ into this orthogonal base. This allows the mapping of a high dimensional vector such as our complexity metrics to a smaller dimensional space. By selecting the first two dimensions (the two dimensions with the highest variance), a representation of a high dimensional vector can be performed in a 2D plane. In this paper, the first two dimensions of vectors following a PCA transformation were used to graph the distribution of plans on a 2D map. Visual inspection of low passing rates plan clustering was performed as a verification of accurate plan characterization by the complexity metrics. In addition to dimensionality reduction, PCA could also be used for feature selection because redundant complexity metrics can be eliminated. The trade-off, however, is that interpretation of the relevant feature is lost because the final complexity metrics are linear combinations of the initial metrics. Therefore, in the present paper, we used PCA only for visualization and we have used Lasso regularization (see below) to perform feature selection. This method keeps the interpretation of the complexity metrics intact but is not affected by redundant complexity metrics. PCA is not necessary beyond a graphical representation of the plans.

2.D. Model

A ML algorithm () was trained to learn the relationship between the plan characteristics and the passing rates. A generalized linear model (GLM) using Poisson regression with Lasso regularization was used to model the failure rate (100 – passing rate).[27,28] Lasso regularization was selected because of its capability to accommodate highly correlated feature sets.[27,28] A detailed, step by step derivation of our mathematical formulation has been included in the supplementary material for the interested reader.[25] This description is self-contained and it should be very helpful for those less familiar with machine learning. Briefly, using Poisson regression with Lasso regularization we model the failing rate as

$$\text{Failing rate} = 100 - \text{passing rate} = e^{\boldsymbol{\beta}^T \mathbf{x}}, \tag{1}$$

where $\mathbf{x}$ is a 79 dimensional column vector, $(1,x_1,x_2,\ldots,x_{78})$, and each component (except 1) represents one of our complexity metrics. $\boldsymbol{\beta}^T$ is the transpose of a constant vector for all IMRT plans with the same dimensions as $\mathbf{x}$. $\boldsymbol{\beta}$ is estimated as the constant vector that will maximize the conditional probability of obtaining $\boldsymbol{\beta}$ given our dataset of failing rates and complexity metrics. Laplace prior (Lasso) regularization has been assumed for $\boldsymbol{\beta}$ to limit the complexity of the model, perform feature selection and to follow Occam's razor principle that establishes that the most plausible model is the simplest one. In our case, simplicity is measured by the number of components equal to 0 in the estimated $\boldsymbol{\beta}$. It is important to highlight that Lasso regularization performs feature selection by driving the components of the vector $\boldsymbol{\beta}$ that correspond to redundant or unimportant complexity metrics to 0 and effectively eliminating them in the model through the multiplication of the vectors $\boldsymbol{\beta}^T$ with the vector $\mathbf{x}$. In Poisson Regression with Lasso Regularization, $\boldsymbol{\beta}$ is estimated $\left(\hat{\boldsymbol{\beta}}\right)$ by solving the following convex optimization problem:

$$\begin{aligned}&\text{argmin}_{\boldsymbol{\beta}}\,\text{Loss}(\boldsymbol{\beta}|D)\\&\quad=\text{argmin}_{\boldsymbol{\beta}}\left[-\sum_{j=1}^{n}\left(y_j\mathbf{x}_j\boldsymbol{\beta}^T-e^{\boldsymbol{\beta}^T x_j}\right)+\lambda|\boldsymbol{\beta}|\right],\end{aligned}\tag{2}$$

where D represents our dataset: the pair of all complexity metrics of a given plan $j(x_j)$ and the failing rate given by y_j. The summation in Eq. (2) is over all the plans included in the training set. λ is a constant that governs complexity and it is selected as explained below. The higher the value λ, the more components with value 0 the estimated $\hat{\boldsymbol{\beta}}$ will have and the fewer complexity metrics will be used in the model. Once the $\hat{\boldsymbol{\beta}}$ that minimizes Eq. (2) is determined, then for any arbitrary plan, Eq. (1) can be used to calculate its failing rate given its characteristics. All the functions used to solve the optimization problem on Eq. (2) can be downloaded from the source: http://web.stanford.edu/~hastie/glmnet_matlab/.

2.D.1. Determining the complexity parameter λ

Overfitting of the data using a highly complex model can result in poor performance of an algorithm on out-of-sample data (data that the algorithm has not seen). As such, it is important to control the complexity of a model. In our algorithm, the parameter λ controls the complexity. In a process similar to that of human intuition, the value of λ obtained using cross-validation dictates which components of the estimated vector $\hat{\boldsymbol{\beta}}$ are different from 0 and as such, which complexity metrics are included in the model. If λ is too small, the term $\lambda|\boldsymbol{\beta}|$ is less important and more metrics are accepted, resulting in a complex model that over fits the data and will be less capable of predicting future IMRT QA passing results. If λ is too large, all components of the vector $\hat{\boldsymbol{\beta}}$ will be set to 0, and the model will fail to select any complexity metrics and will not explain the data. In this work, an initial investigation is performed to obtain the smallest value of λ that will reduce all components of the vector $\boldsymbol{\beta}$ to 0 (max Lambda) and the maximum value of λ that will not set any component to 0 (min Lambda). Then, 100 values are generated linearly on the log scale and 100 different $\hat{\boldsymbol{\beta}}$ obtained by minimizing Eq. (2) for each λ. The smallest λ within 1 standard deviation of the λ that minimizes a leave-one-out cross-validation (LOOCV) error is chosen to maximize the generalization capability of the algorithm, and the final $\hat{\boldsymbol{\beta}}$ is estimated using that λ. In order to perform LOOCV, a point is set aside and the model is constructed using Eqs. (1) and (2). Once the optimal $\boldsymbol{\beta}$ is determined $\left(\hat{\boldsymbol{\beta}}\right)$, then the complexity metrics of the plans are used together with Eq. (1) to determine the failing rate of the point set aside. This is repeated for all points in the dataset. In order to test the performance of the algorithm, both LOOCV over the whole dataset and a double leave-one-out experiment were performed. In the latter, a data point is put aside and an additional LOOCV is performed over the $n-1$ dataset to select the λ hyperparameter. The λ obtained in this manner is used to determine the $\boldsymbol{\beta}$ coefficients using the $n-1$ training set, and this model is subsequently used to predict the sample left out at the beginning. This process is repeated for each individual sample of the dataset. In that case, λ is not determined using information from the point set aside.

2.D.2. *Poisson regression vs weighted Poisson regression*

Among all possible GLMs, a Poisson regression model was chosen to model the 3%/3 mm DTA (L) failure rate. Several reasons led to this choice of model. First, the failure rate is always a positive number skewed towards 0 and high failure rates are rare events. The probability of failure of each diode within the Mapcheck2 detector array is small and there are a large number of diodes used in determining the passing rate. Predicting failure rate can be formulated as a count problem, specifically the number of diodes that fail out of all diodes considered (those receiving at least 10% of the maximum dose). As described in the literature, counting or rate problems are well modeled using Poisson or weighted Poisson regressions.[29] Figure 1 illustrates why a Poisson distribution provides a good fit to failure rates.

This figure has been presented with the purpose of visual inspection because in reality it could be shown that the failing rate should follow a gamma distribution (a more general distribution than the Poisson distribution) which will be published elsewhere. The constant vector $\boldsymbol{\beta}$ could, though, be obtained using a weighted Poisson regression (supplementary material[25]), where the weight of each observation is proportional to the total number of diodes included in the calculation of the failing rate for each plan. However, modeling the passing rates using a weighted Poisson regression imposes practical challenges. First, the passing rates and not the number of detectors that failed are typically reported. This problem could be overcome if an accurate estimation of the area of the plan included in the analysis, area receiving more than 10% of the maximum dose, is performed. Such endeavor, however, will require more than aperture based complexity metrics. We followed both approaches and found that both methods give very similar results. Therefore, for simplicity purposes the results presented in this paper are those of considering a regular Poisson regression.

2.E. Error handling and redelivery

In a database of 498 IMRT plans, we expect some setup errors, improper dose calibration of the Mapcheck2 device, or other unidentified errors. In evaluating the data, therefore, if a measurement disagreed by more than 3% from the predicted 3% dose/3 mm DTA (L) value, the plan was redelivered and remeasured. The results following redelivery were inserted back into the database. These plans were then represented by two observations in the database. In case the difference between the measurement value on the dataset and the redelivered value was larger than 5%, the initial value was considered in error and replaced (these plans were redelivered multiple times to verify that the initial value was indeed in error). Conversely, if the new measurement was within 5% of that in the database, the new value was entered as an independent measurement without changing the initial value. In total, 43 plans with an

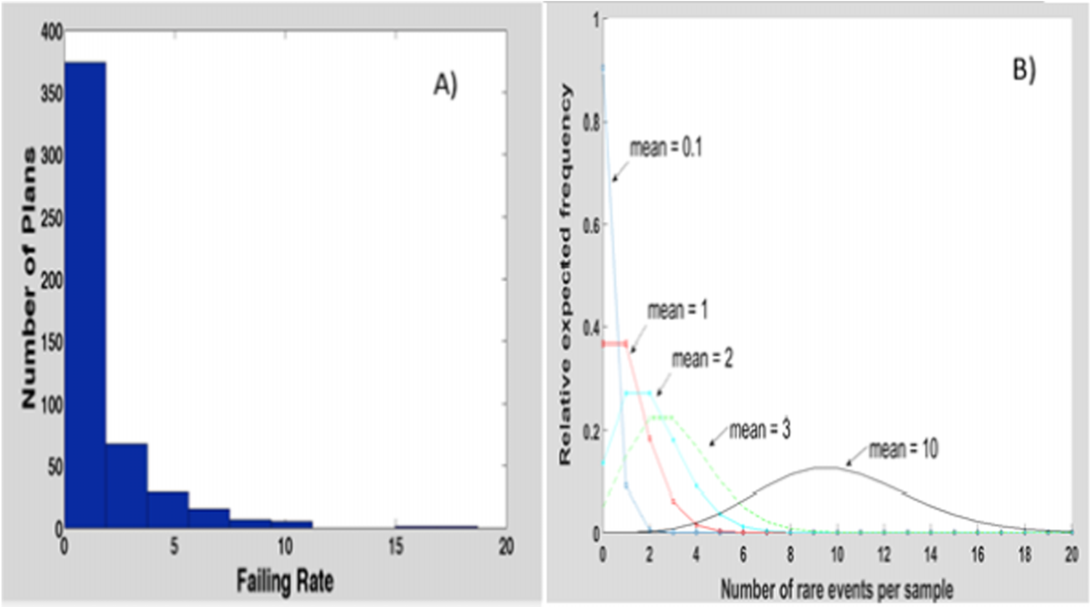

FIG. 1. (A) Histogram showing the measurement/plan failing rate. (B) Poisson distributions for different mean values. Upon visual inspection, the modeling of failing rate as a Poisson distribution is apparent. This assumption is also supported by rigorous statistical tests.

initial disagreement of more than 3% were redelivered. Additionally, 10 random plans whose passing rates were within 3% of the predicted value were also repeated. The second delivery of these plans also resulted in passing rates that were within 3% of predicted value.

3. RESULTS

3.A. Visualization

In order to investigate whether complexity metrics that we extracted were describing passing rates, the principal components of our dataset were obtained. Out of the initial 78 complexity metrics, 21 dimensions retain 95% of the complexity metrics' variance and the first 4 principal components retain 69.67%. Figure 2 shows a schematic representation of the plans on the plane defined by PCA 1 vs PCA 2 (similar results are observed with other principal components). As can be observed, plans with passing rates smaller than 95% seem to cluster in the region of the space where PCA 1 is higher than 0 and PCA 2 is higher than 5, providing confidence that our complexity metrics are describing plan passing rates.

3.B. Exploration

In the initial exploratory phase, we investigated different questions. First, were separate models needed for each of the four nominally matched Clinacs? Was a separate model needed for the TrueBeam Linac independent of the matched Clinacs? Third, was a model needed for each energy? Fourth, were quadratic terms needed on the vector describing each plan? And finally, does a Poisson distribution describe the results or will assuming this distribution bias the model? In order to address these questions, a model incorporating all data points was initially constructed. The lambda that minimized the LOOCV error was $\lambda_{\min} = 0.0128$, which resulted in 35 degrees of freedom (number of complexity metrics selected) explaining 76.01% of the variance observed in the failing rates. The three most important complexity metrics (the first complexity metrics with coefficients different than 0) were the MU factor (MU per Gy), the MU weighted irregularity factor as described by Du *et al.*[17] and the fraction of MLC segments with opposing leaf gaps smaller than 10 mm (small aperture score, SAS_10 mm). As Fig. 3 shows, "good agreement" between the predicted and the measured failure rate is obtained even in this exploratory model. However, as 14 plans still had residual errors larger than 3%, we looked at the other complexity metrics selected by the algorithm as well as those plans that had a disagreement larger than 3% in order to improve the model. In addition to those mentioned above, three other important features were the energy (6 MV vs other energy), the machine/MLC type (Clinacs equipped with a Millennium MLC vs TrueBeam equipped with an HD-120 MLC), and whether delivery was performed on one specific clinac (Clinac 2).

The first two features were expected and imply that independent models for the machine/MLC type and for each energy should be constructed. The fact that Clinac 2 was an important feature, however, was unexpected as our four clinacs are nominally matched. Further investigation pointed out that plans with a large fraction of the area delivered outside a circle with a radius equal to 15 cm (whole pelvis or three field breast plans) have passing rate of 100% on Clinac 2 but substantially lower passing rates on the other Clinacs. On further inspection it was discovered that off axis profiles for Clinac 2 most closely matched the ECLIPSE model (Fig. 4).

At the corners of the 40×40 cm^2 field, dose differences larger than 3% between Clinacs 1, 3, and 4 and the ECLIPSE model can be observed, while almost perfect agreement is obtained for Clinac 2. These results indicated to us that individual

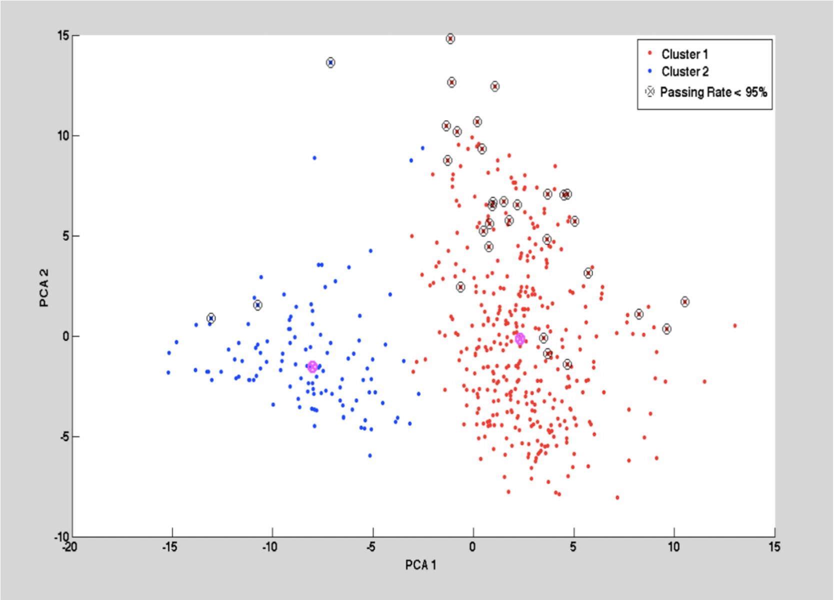

FIG. 2. Clustering of plans with low passing rates along the dimensions PCA 1 and PCA 2. A hierarchical clustering (points reds and blues) has only been used to enhance the visualization representation. Plans with passing rates less than 95% tend to cluster in the center of the plane. In other dimensions, different patterns can be observed. (See color online version.)

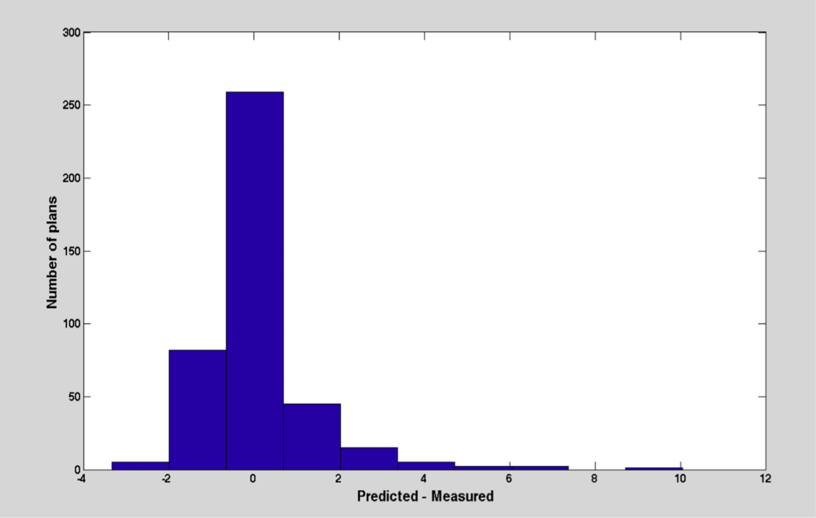

FIG. 3. Residual errors of a model built using all data. Predicted = 3%/3 mm (L) value predicted by the model and measured = 3%/3 mm (L) value measured with Mapcheck2.

models would need to be constructed for TrueBeam, Clinac 2, Clinacs 1, 3, and 4 and for each energy, respectively. In the present paper, we report the models for 6 MV plans for each of the Linacs and the TrueBeam. Models for 15 MV and mixed energy plans with their particular complexity metrics will be evaluated in a future study, particularly because additional data are needed to properly separate these datasets. Additionally, it was also noted that regardless of the Linac, large field plans where multiple Mapcheck2 acquisitions are necessary in order to acquire the full dose distribution are prone to error and the results might depend on the way they are acquired. These results will be published in an independent paper.

3.C. Models for 6 MV plans

On Clinacs 1, 3, and 4, 243 plans were delivered, while 176 plans were delivered on Clinacs 2 and 21 on TrueBeam. Figure 5 shows a figure of the cross-validated Poisson deviance vs the hyperparameter lambda for the 6X model for Clinacs 1, 3, and 4.

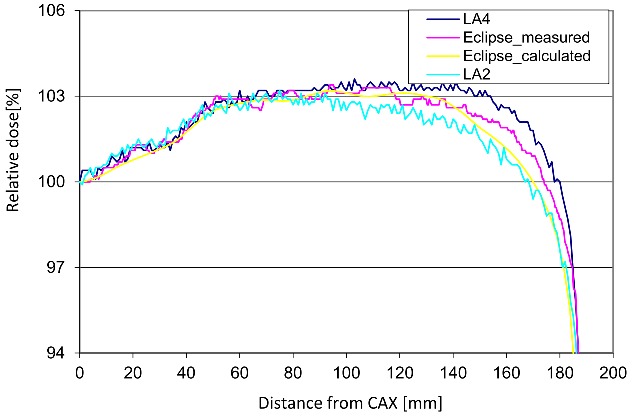

FIG. 4. Diagonal profile of Clinac 4 (LA4) as example of all the clinacs compared to the profile of Clinac 2 (LA2) and the measured and calculated profile from ECLIPSE. Measured = data input into ECLIPSE to build the beam model and calculated = profile generated by ECLIPSE after model reconstruction has been performed.

The largest value of lambda within 1 standard deviation of the lambda that minimizes the cross-validation, 0.014 for the Clinacs 1, 3, and 4 model, was chosen as the hyperparameter in order to have a more robust model. To describe the failing rate, 35 complexity metrics were selected (their corresponding components of the vector $\hat{\boldsymbol{\beta}}$ were not 0), which in this case explains 87% of the observed null deviance with $p = 5.8 \times 10^{-12}$, indicating that the current model is significantly better than a null model (explaining the data with a constant value). The interested reader can find explanation of the deviance in the supplementary material.[25] Similar results were obtained for the models of Clinac 2 and TrueBeam. The fiive most important complexity metrics for each of these models are shown in Table I.

It is important to highlight that different models will select different complexity metrics and coefficients, further supporting the position that models need to be constructed for each independent Linac. Additionally, the TrueBeam model is not representative of all possible plans that could be delivered on this Linac due to the fact that at our institution, the TrueBeam is primarily used to treat smaller targets such as those encountered in SBRT. Finally, explaining 87% of the variance of the measured failing rate for Clinacs 1, 3, and 4, is sufficient to predict all plans within a 3% error in a leave-one-out cross-validated experiment [Fig. 6(A)]. A similar result is obtained for the Clinac 2 [Fig. 6(B)] and TrueBeam models (data not shown). This 3% error threshold was sufficient to identify delivery problems in a number of different plans.

Table II shows five different plans in which the initial result had a disagreement larger than 3% as compared to the predicted value. On redelivery, all fell within 3% of the model, suggesting a setup or other measurement error during initial delivery.

3.D. Number of plans needed and error on a test set

An important question to answer when a model is constructed is how many plans are needed to generate an accurate model. The answer to this question will always be model

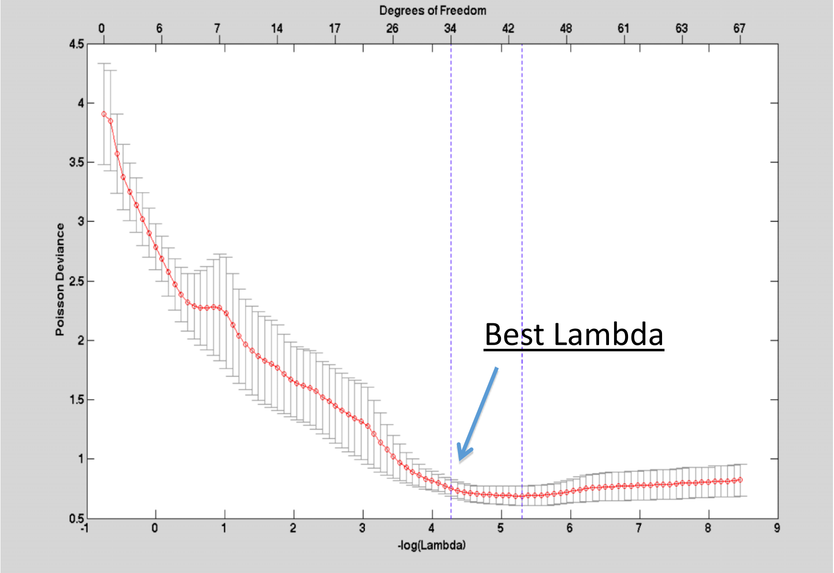

FIG. 5. Leave-one-out cross-validated Poisson deviance for the Clinacs 1, 3, and 4 model. The Lambda value 1 standard deviation from the lambda that minimized the cross-validation was selected as the hyperparameter for the model. Out of the initial 78, 35 features were selected for this lambda.

specific. For instance, if different Linacs are combined in one model, it is plausible to expect that more plans would be needed due to a higher variance of the data than when only one Linac is modeled. Figure 7 shows a learning curve for our Clinacs 1, 3, and 4 model where the normalized deviance of the training and testing sets are plotted as a function of the number of plans in the training set.

As can be observed, the testing deviance plateaus at approximately 200 plans and no further improvement is obtained by including more plans. Finally, the cross-validation error is an overly optimistic estimation of the error that will be made by the algorithm when hyperparameters are selected using the same dataset. This is because the hyperparameter lambda, if chosen using LOOCV as described in Sec. 3.C, contains information about all plans and as such its use in the model can result in an under estimation of the error. In that sense, it is usually recommended the data be split into three sets: the training set to estimate the coefficients of the model, a validation set to estimate the hyperparameters, and a test set to evaluate the error that the model will make. In the present paper, a double LOOCV was performed as explained above. In this manner, the hyperparameter selected to build the model would have not "seen" the result of the plan that would be predicted. The residual errors of the algorithm constructed in this manner should be a very realistic estimation of the errors that the model will make when predicting plans that it has not seen. Figure 8 shows these residual errors for the Clinacs 1, 3, and 4 model.

TABLE I. The five most important features for each model. Features are different for each of the different Linac groups, suggesting that separate models may be needed at each specific institution.

Group	Features
Clinacs 1, 3, and 4	1. Weighted average irregularity factor 2. Weighted out of circle fractional area 40 cm 3. MU factor 4. SAS 5 mm 5. Leaf gap moment 4
Clinac 2	1. MU factor 2. SAS 5 mm 3. Weighted out of circle fractional area 25 cm 4. Out of circle fractional area 30 cm 5. SAS 10 mm
TrueBeam	Average perimeter Infield area 1 Infield area 4 Out of circle fractional area 25 cm Infield area 5

As can be observed, the double leave-one-out error estimation is slightly larger than the cross-validation errors shown in Fig. 5, while still less than 3% for all plans. Finally, a better visualization of the predictive value of the algorithm is shown in Fig. 9, where the predicted passing rates are plotted against the measured values. A linear regression is obtained, as expected, where the observed dispersion is representative of the intrinsic noise within the measurements that is not explained by the model.

4. DISCUSSION

In the present paper, we have demonstrated that it is possible to predict IMRT QA passing rates *a posteriori* within a 3% error using a Poisson regression model with Lasso regularization that combines different complexity metrics. This accuracy was shown for Linacs with different underlying behaviors and as a result, different models. In all cases, at least 87% of the variance observed in the passing rates is explained by the models. This is a substantial improvement over current approaches that analyze the correlation between complexity metrics and passing rates individually. Further improvement of the accuracy may be possible by using a zero-inflated Poisson

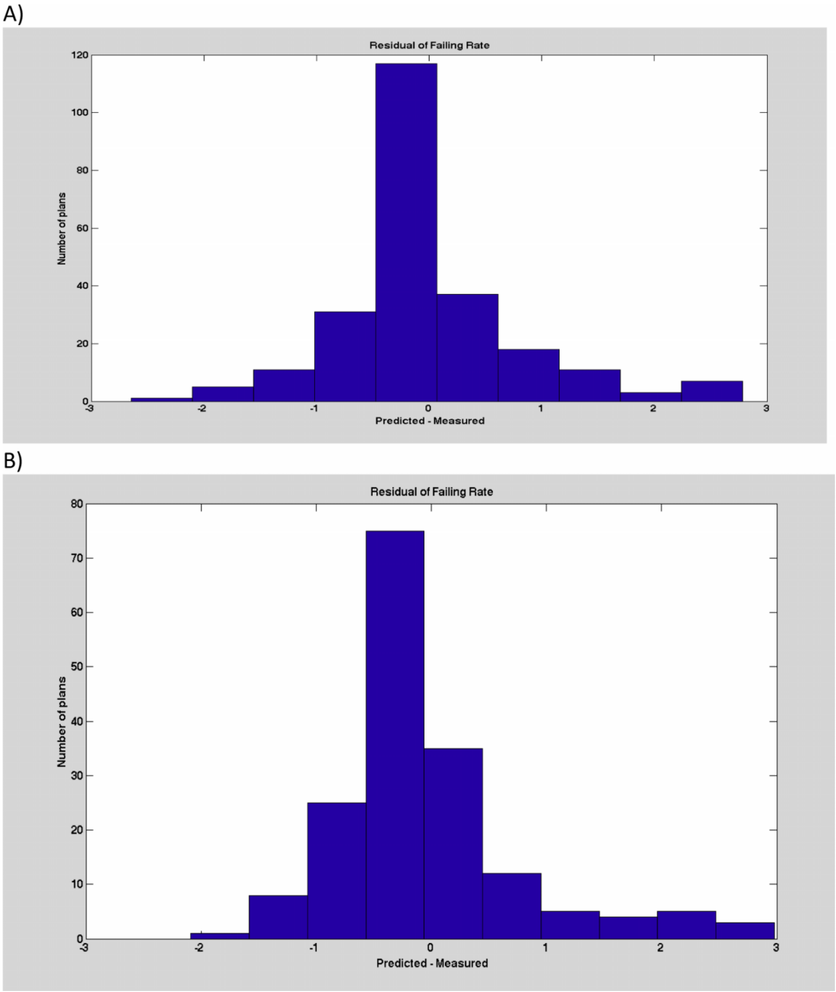

FIG. 6. (A) Cross-validated residual error for the Clinacs 1, 3, and 4 model. (B) Cross-validated residual error for the Clinac 2 model. In both cases a cross-validated residual error smaller than 3% was obtained. Predicted = 3%/3 mm (L) value predicted by the model and measured = 3%/3 mm (L) value measured with Mapcheck2.

or negative binomial regression model, including quadratic terms, ensemble methods, redelivering plans in order to reduce noise within the existing data (variations on the order of 2% are not uncommon after redelivery of plans) or predicting passing rates for individual beams. However, using a 3% error to predict passing rates with was found to be relevant in that it allowed us to determine that the profiles of one Linac was different from our beam model at the periphery of a 40×40 cm^2 field. As a result, we do not consider that further refinement of the model is necessary at this time. We should clarify that we have decided to choose a constant error (3%) around the prediction for all plans as our threshold despite the fact that the error of the prediction, according to a Poisson or gamma distribution, is plan specific and depends on the mean value predicted. However, Poisson cumulative distribution cannot be inverted easily and a simple equation for the error upper and lower bounds, as in the case of the Gaussian distribution, cannot be found. As such, we have decided to use the maximum empirical error that we observed for all plans as our constant error for simplicity purposes. However, the square root of the failing rate (the standard deviation of the Poisson distribution) can also be used as the upper and lower bounds. The standard deviation has the interesting property that plans with larger failing rates will tend to have a higher variability; this is supported by empirical evidence.

Additionally, passing rates have been predicted using aperture-based metrics, which allow easy identification of the different categories leading to disagreement between the treatment planning and delivery platforms. In that sense, the most important complexity metrics across the different models are: MU factor (a measurement of the overall complexity of the plan), different small aperture scores (DLG modeling), the irregularity factor (tongue and groove modeling) and the fraction of beams delivered outside circles of specific radii (associated with beam profile disagreement). It is likely,

TABLE II. Example of five plans whose measurements disagree on more than 3% with the predicted values and after re delivery they all fell within the 3% error.

Database	Predicted	Remeasured
93.6	97.2	98.2
96.60	93.5	95.5
91.8	95.8	93.4
92.4	96.1	94.6
90	93.6	95.2

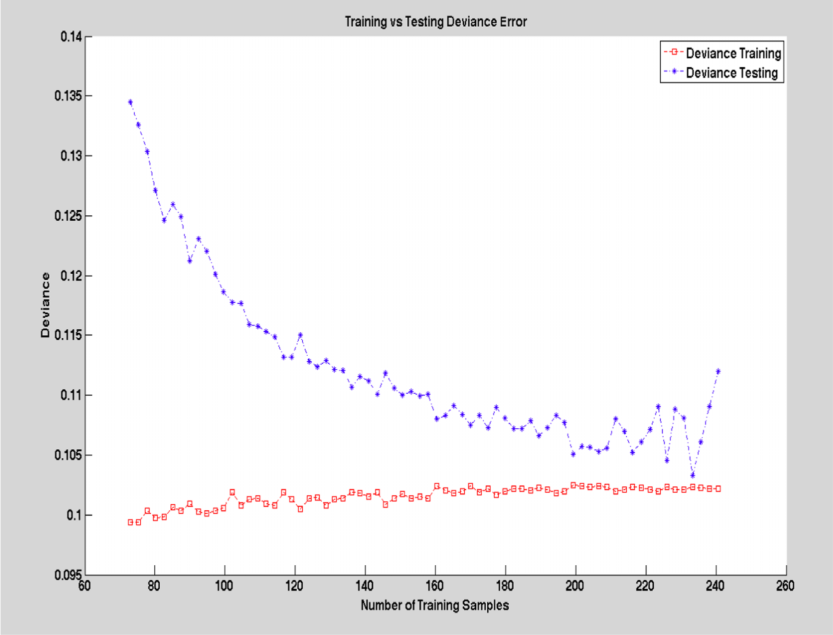

FIG. 7. Learning curve. Deviance of the model on the testing sample and the training sample. Beyond 200 plans, increasing the number of plans in the training sample produces no further improvement in the model.

however, that different delivery systems will be described by different metrics. This highlights one of the key advantages of the Virtual IMRT QA approach, as it allows the identification and troubleshooting of different sources of error.

Experienced physicists know subjectively that plans of varying "complexity," tend to have different passing rates. Head and neck plans, for example, are often highly modulated, and lower passing rates are not uncommon. This is reflected in the RPC credentialing results from IMRT irradiation of the head and neck phantom, in which 18.4% of institutions failed to pass a 7% dose/4 mm DTA criteria (considerably less rigorous than that used in the current study); when the criteria were changed to 5% dose/4 mm, more typical of that used in the institutional setting, the failing rate doubled.[30] The Virtual IMRT QA process has the potential to change how IMRT QA is viewed and evaluated. For instance, if after measurement only 89% percent of pixels meet the passing criteria, but the model predicts 90.5%, the result may be considered

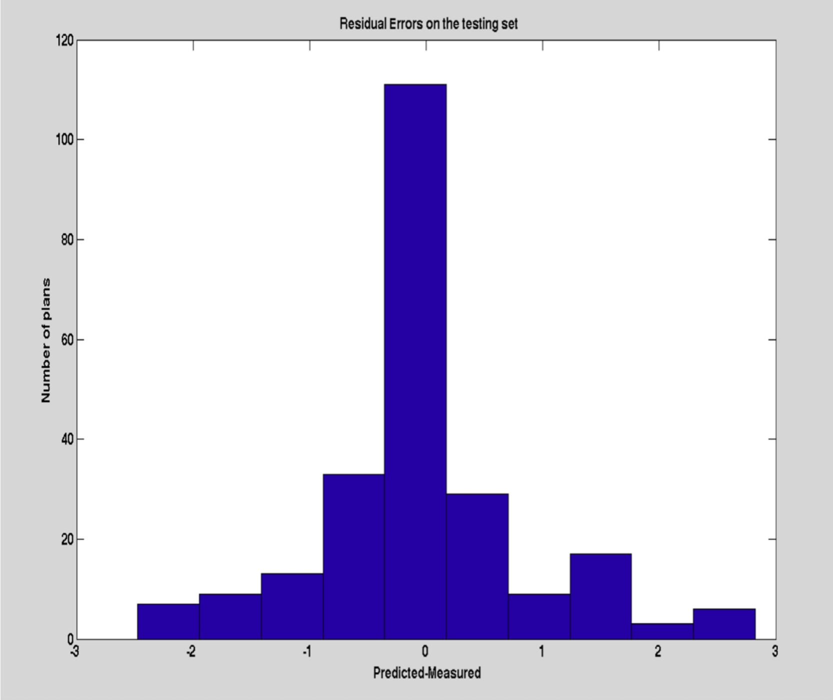

FIG. 8. Residual error of all plans predicted using out of sample data to estimate both lambda and the model coefficients. Predicted = 3%/3 mm (L) value predicted by the model and measured = 3%/3 mm (L) value measured with Mapcheck2.

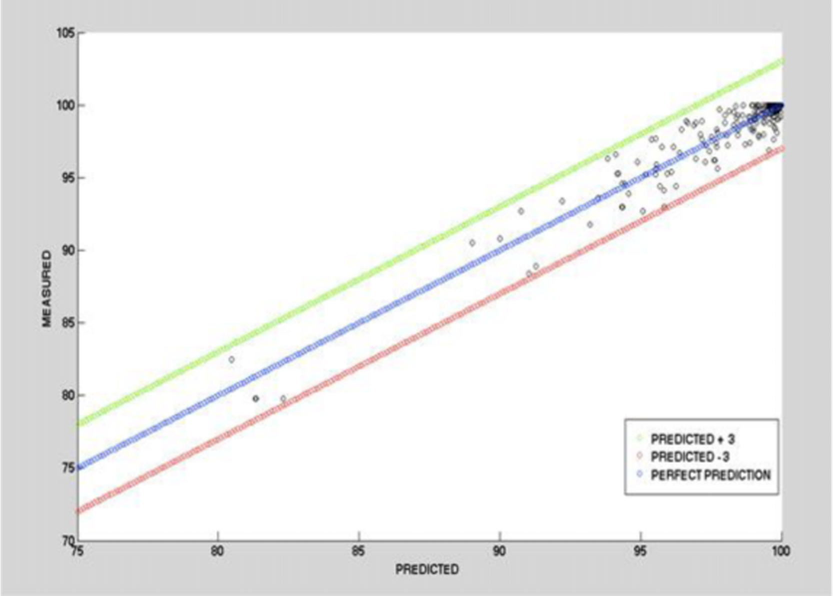

Fig. 9. Measured vs predicted passing rates, with the blue line representing a perfect prediction. All points lay are within ±3% (the green and blue lines). Predicted = 3%/3 mm (L) value predicted by the model and measured = 3%/3 mm (L) measured with Mapcheck2. (See color online version.)

acceptable. Agreement of measured and predicted passing rate, however, is not a guarantee against clinical errors, but rather, an indication that the delivery system has not deviated from past results. In the example described, a potentially better course of action may be to replan before waiting for the IMRT QA result. Additionally, a comparison of the passing rate for this specific plan with results that would have been obtained at other institutions could also be valuable; indeed this is an application that Virtual IMRT QA enables. Moreover, the common 90% threshold for determining whether a plan is acceptable is arbitrary, mathematically inaccurate and very inefficient at detecting clinically relevant errors. If a clinically relevant threshold is found (which can only be obtained using outcome data and it is likely to depend on the treatment site), plan specific predictions could be used to detect errors and delivery limitations, not only when this threshold is exceeded, but also when a significant deviation has occurred from the expected value. Once the treatment planning dose calculation model has been constructed with satisfactory accuracy, the purpose of patient-specific IMRT measurements can also become an exercise in testing whether delivery systems have deviated from their state at commissioning. Virtual IMRT QA improves on the current IMRT QA process by providing QA predictions and thresholds for each individual plan (as shown by the red and green lines in Fig. 9). Once the dependence of the passing rate on the complexity of the plan has been removed through Virtual IMRT QA, then random noise around that value can be expected. In this sense, our models proved useful in detecting both setup and dose calibration errors, and also in pointing out that the characteristics of one of our Clinacs had different profile characteristics in the outer corners of a 40 × 40 field. At the present time, we are investigating whether Virtual IMRT QA is able to detect clinically relevant errors reported in the literature which have been impossible to detect with the current IMRT QA process and the one fits all 90% threshold approach. If this hypothesis proves correct, Virtual IMRT QA could significantly change the IMRT QA process. As reported by Nelms *et al.* however, 2%/2 mm local was superior to 3%/3 mm local or global in detecting clinically relevant errors.[5] In fact, it is possible that in order to make IMRT QA relevant a combination of local normalization, dose threshold values and plan specific passing rates (Virtual IMRT QA) are needed. If the 3%/3 mm local metric were to prove insensitive, even with plan specific predictions, a noiseless database of 2%/2 mm Local passing rates should be constructed and the Virtual IMRT QA method applied to predict these values.

In order for this process to become a reality, however, there are several remaining hurdles. The number of plans needed to construct a model needs to be reduced to a few dozen, as it is not realistic to expect the delivery of hundreds of plans. In that sense, the use of a golden beam data model together with a database of selected plans that could assess the different sources of errors is a promising approach. An alternate solution to this problem may be to predict passing rates for individual beams (instead of plans), as these passing rates will have a stronger dependence on the complexity metrics currently calculated.

In addition to plan specific thresholds, Virtual IMRT QA has other interesting applications. For instance, knowing that a plan is unlikely to pass would reduce the delays that occur when a plan fails QA. By incorporating Virtual IMRT QA predictions within the optimization process, failing plans could potentially be eliminated. Additionally, different departments will be able to compare the accuracy of their commissioning and TPS data by knowing what passing rate would be obtained for the same plan at a different institution. This will require building models for each combination of equipment or methodology for pretreatment verification used by different departments or standardizing them across the field. Thus this application has the potential to standardize dose delivery accuracy across different institutions by bringing all deliveries within the inaccuracy observed at selected institutions whose commissioning and delivery systems have been carefully verified. Moreover, Virtual IMRT QA will be important in adaptive radiation therapy applications, highlighting those plans likely to pass or fail before, and allowing for the QA measurements to be performed after treatment. For example, one Linac could be dedicated to adaptive radiation therapy, and on that Linac, a higher level daily QA would be performed where a set of preselected plans (5 or 6) that test the different factors (identified with Virtual IMRT QA) leading to disagreements between TPS and Linac delivery could be measured. Once the pre treatment measurements for these plans are confirmed within the levels of confidence indicated by Virtual IMRT QA, we could move to treat patients with plans developed in that session. Virtual IMRT QA could then be used to predict whether these plans would pass QA and to the discretion of each department, treat patients. To confirm the prediction of Virtual IMRT QA, the plans used to treat the patients could be measured after treatment. We must state, however, that there are failure modes that cannot be modeled by Virtual IMRT QA, such as those associated with the transferring of a plan from the TPS to the Record and Verify System. These and other failure modes should be tested independently, and different applications for automatic detection could be envisioned.

Finally, further testing of our algorithms should be performed. This will allow us to assess the accuracy of Virtual IMRT QA across other plan/delivery/measurement combinations and at other institutions.

5. CONCLUSIONS

Virtual IMRT QA is capable of predicting IMRT QA passing rates within 3% for different delivery platforms and with different underlying sources of errors. This process proved to be clinically significant in detecting small setup errors in the measurement process as well as a mismatch of one of four otherwise identical Linacs. By providing plan specific thresholds, improved efficiency and reduced replanning, standards to which departments can compare their results, safe implementation of adaptive radiotherapy and potentially eliminating failing QA altogether, Virtual IMRT QA may have profound implications for the current IMRT QA process.

CONFLICT OF INTEREST DISCLOSURE

The authors have no COI to report.

[a] G. Valdes and R. Scheuermann contributed equally to this work.

[b] Author to whom correspondence should be addressed. Electronic mail: gilmer.valdes@uphs.upenn.edu

[1] G. A. Ezzell, J. M. Galvin, D. Low, J. R. Palta, I. Rosen, M. B. Sharpe, P. Xia, Y. Xiao, L. Xing, C. X. Yu, I. Subcommittee, and A. R. T. Committee, "Guidance document on delivery, treatment planning, and clinical implementation of IMRT: Report of the IMRT subcommittee of the AAPM radiation therapy committee," Med. Phys. **30**, 2089–2115 (2003).

[2] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," Med. Phys. **25**, 656–661 (1998).

[3] J. Van Dyk, R. B. Barnett, J. E. Cygler, and P. C. Shragge, "Commissioning and quality assurance of treatment planning computers," Int. J. Radiat. Oncol., Biol., Phys. **26**, 261–273 (1993).

[4] R. A. Siochi, E. C. Pennington, T. J. Waldron, and J. E. Bayouth, "Radiation therapy plan checks in a paperless clinic," J. Appl. Clin. Med. Phys. **10**, 43–62 (2009).

[5] B. E. Nelms, M. F. Chan, G. Jarry, M. Lemire, J. Lowden, C. Hampton, and V. Feygelman, "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 (15pp.) (2013).

[6] N. Childress, Q. Chen, and Y. Rong, "Parallel/opposed: IMRT QA using treatment log files is superior to conventional measurement-based method," J. Appl. Clin. Med. Phys. **16**, 5385 (2015).

[7] B. E. Nelms, H. Zhen, and W. A. Tome, "Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors," Med. Phys. **38**, 1037–1044 (2011).

[8] S. Stojadinovic, L. Ouyang, X. Gu, A. Pompoš, Q. Bao, and T. D. Solberg, "Breaking bad IMRT QA practice," J. Appl. Clin. Med. Phys. **16**, 5242 (2015).

[9] E. M. McKenzie, P. A. Balter, F. C. Stingo, J. Jones, D. S. Followill, and S. F. Kry, "Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans," Med. Phys. **41**, 121702 (15pp.) (2014).

[10] J. J. Kruse, "On the insensitivity of single field planar dosimetry to IMRT inaccuracies," Med. Phys. **37**, 2516–2524 (2010).

[11] G. Palaniswaamy, R. Scott Brame, S. Yaddanapudi, D. Rangaraj, and S. Mutic, "A statistical approach to IMRT patient-specific QA," Med. Phys. **39**, 7560–7570 (2012).

[12] G. A. Ezzell, J. W. Burmeister, N. Dogan, T. J. LoSasso, J. G. Mechalakos, D. Mihailidis, A. Molineu, J. R. Palta, C. R. Ramsey, B. J. Salter, J. Shi, P. Xia, N. J. Yue, and Y. Xiao, "IMRT commissioning: Multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119," Med. Phys. **36**, 5359–5373 (2009).

[13] M. Nauta, J. E. Villarreal-Barajas, and M. Tambasco, "Fractal analysis for assessing the level of modulation of IMRT fields," Med. Phys. **38**, 5385–5393 (2011).

[14] S. Webb, "Use of a quantitative index of beam modulation to characterize dose conformality: Illustration by a comparison of full beamlet IMRT, few-segment IMRT (fsIMRT) and conformal unmodulated radiotherapy," Phys. Med. Biol. **48**, 2051–2062 (2003).

[15] M. M. Coselmon, J. M. Moran, J. D. Radawski, and B. A. Fraass, "Improving IMRT delivery efficiency using intensity limits during inverse planning," Med. Phys. **32**, 1234–1245 (2005).

[16] S. B. Crowe, T. Kairn, N. Middlebrook, B. Sutherland, B. Hill, J. Kenny, C. M. Langton, and J. V. Trapp, "Examination of the properties of IMRT and VMAT beams and evaluation against pre-treatment quality assurance results," Phys. Med. Biol. **60**, 2587–2601 (2015).

[17] W. Du, S. H. Cho, X. Zhang, K. E. Hoffman, and R. J. Kudchadker, "Quantification of beam complexity in intensity-modulated radiation therapy treatment plans," Med. Phys. **41**, 021716 (9pp.) (2014).

[18] A. L. McNiven, M. B. Sharpe, and T. G. Purdie, "A new metric for assessing IMRT modulation complexity and plan deliverability," Med. Phys. **37**, 505–515 (2010).

[19] K. C. Younge, M. M. Matuszak, J. M. Moran, D. L. McShan, B. A. Fraass, and D. A. Roberts, "Penalization of aperture complexity in inversely planned volumetric modulated arc therapy," Med. Phys. **39**, 7160–7170 (2012).

[20] C. K. McGarry, C. D. Chinneck, M. M. O'Toole, J. M. O'Sullivan, K. M. Prise, and A. R. Hounsell, "Assessing software upgrades, plan properties and patient geometry using intensity modulated radiation therapy (IMRT) complexity metrics," Med. Phys. **38**, 2027–2034 (2011).

[21] S. B. Crowe, T. Kairn, J. Kenny, R. T. Knight, B. Hill, C. M. Langton, and J. V. Trapp, "Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results," Australas. Phys. Eng. Sci. Med. **37**, 475–482 (2014).

[22] T. Bortfeld, J. Burkelbach, R. Boesecke, and W. Schlegel, "Methods of image reconstruction from projections applied to conformation radiotherapy," Phys. Med. Biol. **35**, 1423–1434 (1990).

[23] J. Llacer, T. D. Solberg, and C. Promberger, "Comparative behaviour of the dynamically penalized likelihood algorithm in inverse radiation therapy planning," Phys. Med. Biol. **46**, 2637–2663 (2001).

[24] S. V. Spirou and C. S. Chui, "A gradient inverse planning algorithm with dose-volume constraints," Med. Phys. **25**, 321–333 (1998).

[25] See supplementary material at http://dx.doi.org/10.1118/1.4953835 for derivations and additional information.

[26] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York, NY, 2009).

[27] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," J. Stat. Software **33**, 1–22 (2008).

[28] J. Bien, R. Tibshirani, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," J. R. Stat. Soc. Series B Stat. Methodol. **74**, 245–266 (2010).

[29] T. Harris, J. M. Hilbe, and J. W. Hardin, "Modeling count data with generalized distributions," Stat. J. **14**, 562–579 (2014).

[30] A. Molineu, N. Hernandez, T. Nguyen, G. Ibbott, and D. Followill, "Credentialing results from IMRT irradiations of an anthropomorphic head and neck phantom," Med. Phys. **40**, 022101 (8pp.) (2013).