A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance

Seiji Tomori
Department of Radiology, National Hospital Organization Sendai Medical Center, Sendai, Miyagi 983-8520, Japan
Department of Radiation Oncology, Tohoku University Graduate School of Medicine, Sendai, Miyagi 980-8574, Japan

Noriyuki Kadoya[a)], Yoshiki Takayama, and Tomohiro Kajikawa
Department of Radiation Oncology, Tohoku University Graduate School of Medicine, Sendai, Miyagi 980-8574, Japan

Katsumi Shima
Department of Radiology, National Hospital Organization Hakodate National Hospital, Hakodate, Hokkaido 041-8512, Japan

Kakutarou Narazaki
Department of Radiology, National Hospital Organization Sendai Medical Center, Sendai, Miyagi 983-8520, Japan

Keiichi Jingu
Department of Radiation Oncology, Tohoku University Graduate School of Medicine, Sendai, Miyagi 980-8574, Japan

(Received 6 March 2018; revised 18 July 2018; accepted for publication 19 July 2018;
published 17 August 2018)

Purpose: Patient-specific quality assurance (QA) measurement is conducted to confirm the accuracy of dose delivery. However, measurement is time-consuming and places a heavy workload on the medical physicists and radiological technologists. In this study, we proposed a prediction model for gamma evaluation, based on deep learning. We applied the model to a QA measurement dataset of prostate cancer cases to evaluate its practicality.
Methods: Sixty pretreatment verification plans from prostate cancer patients treated using intensity modulated radiation therapy were collected. Fifteen-layer convolutional neural networks (CNN) were developed to learn the sagittal planar dose distributions from a RT-3000 QA phantom (R-TECH.-INC., Tokyo, Japan). The percentage gamma passing rate (GPR) was measured using GAFCHROMIC EBT3 film (Ashland Specialty Ingredients, Covington, USA). The input training data also included the volume of the PTV (planning target volume), rectum, and overlapping region, measured in cm^3, and the monitor unit values for each field. The network produced predicted GPR values at four criteria: 2%(global)/2 mm, 3%(global)/2 mm, 2%(global)/3 mm, and 3%(global)/3 mm. Adam, an algorithm for first-order gradient-based optimization of stochastic objective functions, was used for learning and for optimizing the CNN-based model. Fivefold cross-validation was applied to validate the performance of the proposed method. Forty cases were used for training and validation set in fivefold cross-validation, and the remaining 20 cases were used for the test set. The predicted and measured GPR values were compared.
Results: A linear relationship was found between the measured and predicted values, for each of the four criteria. Spearman rank correlation coefficients in validation set between measured and predicted GPR values at four criteria were 0.73 at 2%/2 mm, 0.72 at 3%/2 mm, 0.74 at 2%/3 mm, and 0.65 at 3%/3 mm, respectively ($P < 0.01$). The Spearman rank correlation coefficients in the test set were 0.62 ($P < 0.01$) at 2%/2 mm, 0.56 ($P < 0.01$) at 3%/2 mm, 0.51 ($P = 0.02$) at 2%/3 mm, and 0.32 ($P = 0.16$) at 3%/3 mm. These results demonstrated a strong or moderate correlation between the predicted and measured values.
Conclusions: We developed a CNN-based prediction model for patient-specific QA of dose distribution in prostate treatment. Our results suggest that deep learning may provide a useful prediction model for gamma evaluation of patient-specific QA in prostate treatment planning. © *2018 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.13112]

Key words: convolutional neural network, deep learning, gamma evaluation, patient QA, radiotherapy

1. INTRODUCTION

Modern radiotherapy techniques, including intensity modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT), allow a complex dose distribution to be applied, minimizing the damage done to healthy tissues.[1] However, these techniques are highly complex in both planning and delivery, and must be thoroughly safety tested prior to implementation.[2] Patient-specific pretreatment quality assurance (QA) is an important step in the treatment process,

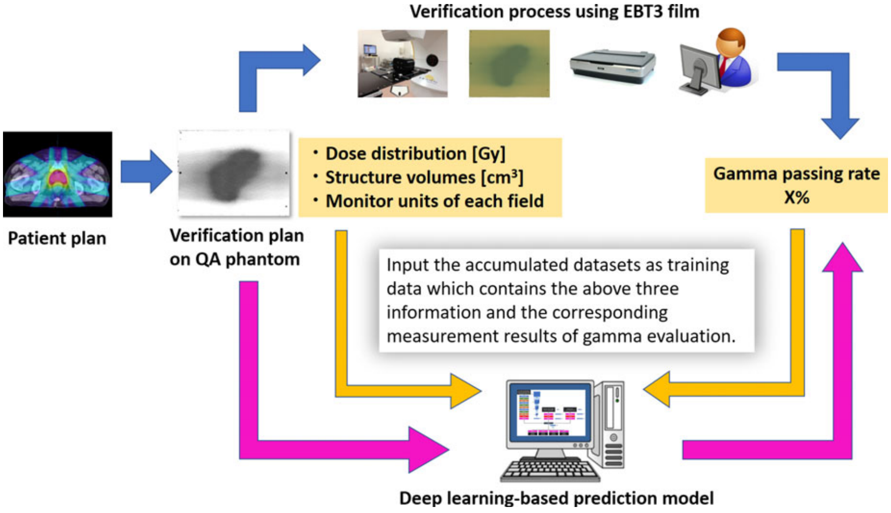
Verification process using EBT3 film
Patient plan
Verification plan on QA phantom
• Dose distribution [Gy]
• Structure volumes [cm³]
• Monitor units of each field
Gamma passing rate X%
Input the accumulated datasets as training data which contains the above three information and the corresponding measurement results of gamma evaluation.
Deep learning-based prediction model

FIG. 1. Schematic diagram of the dose distribution verification process. In this study, EBT3 film was used to evaluate the gamma passing rate. The cumulative results were used as the training input to the deep learning-based prediction model. [Color figure can be viewed at wileyonlinelibrary.com]

and it has been widely rated as essential.[3] The process typically involves comparing the calculated dose distribution with the measured dose distribution, using a gamma evaluation.[4] This form of evaluation has become the mainstay of patient-specific QA, when applied to IMRT and VMAT. It is recommended in a number of institutional guidelines and is a readily available, efficient tool for assessing the accuracy of patient-specific treatment delivery.[3,5,6] Gamma evaluation allows the agreement between two dose distributions to be evaluated in a combined dose-spatial domain.[4] The agreement in percentage of the dose points is referred to as the gamma passing rate (GPR). In most cases, the full distribution of γ values is not considered when calculating the GPR. Instead, only values in the region of interest or values corresponding to a dose that exceeds a user-defined lower threshold are typically considered. The result is therefore influenced by user-defined criteria: the dose difference, ΔD (% global or local), and distance-to-agreement, Δd (mm).

However, the implementation of patient-specific QA using gamma evaluation and point dose measurement is time-consuming.[7] Currently, the patient-specific QA requires a measurement using phantom or device in every treatment plan before it begins. Given the necessity of confirming an IMRT plan for each patient, currently available patient-specific IMRT QA techniques impose a heavy clinical workload burden.

A UK survey of patient-specific QA practice noted that recent approaches have reduced the role of measurement and it reported a move away from individual patient measurement, especially in prostate treatment.[8] To solve the problems related to the workload, a practical and precise predictive method has been reported. The prediction models reported in previous studies have used the 3D predicted dose, with the gamma evaluation of the planar dose distribution conducted using criteria greater than 3%(global)/3 mm.[9,10] Valdes et al. discussed machine learning-based prediction models, in which gamma evaluation was performed using criteria of 3%(local)/3 mm.[11,12] However, recent publications have suggested that criteria stricter than 3%(global)/3 mm be applied for IMRT plan gamma evaluation for error detection.[13,14] Nelms et.al. reported that 2%(local)/2 mm was the most sensitive criterion to detect clinically relevant errors.[14] On the other hand, TG 218 recommended that the global normalization method and true composite delivery method, in which all beams were delivered in the actual treatment setting, should be used for IMRT QA measurement.[15] The criteria should be selected based on the measurement device and situation. A new gamma prediction model is needed to reflect the recent reassessment of the recommended IMRT QA measurement in the true composite delivery and global normalization methods.

Interian et al. investigated the feasibility of using transfer learning with VGG-16 ImageNet model and predicted the GPR value in 3%(local)/3 mm.[16] The use of deep learning for predicting QA results has been recently developed, and it has the potential to accurately predict GPR values. In pursuit of the goal, in this study, we developed a novel prediction model based on a deep learning convolutional neural network (CNN). A model proposed in this study is able to predict GPR values based on multiple criteria along with global normalization and to experimentally test the feasibility of a model using a dataset of patient-specific QA results. Figure 1 shows a schematic diagram of the concept. The study focused on prostate cancer treatment, with the goal of predicting the GPR value on a planar dose distribution in a verification measurement.

2. MATERIALS AND METHODS

2.A. Overview of dose distribution verification

Sixty prostate cancer treatment plans, which were continuously treated from May 2015 to June 2017, were selected from the iPlan Dose ver. 4.1.2 (BrainLAB, Munich,

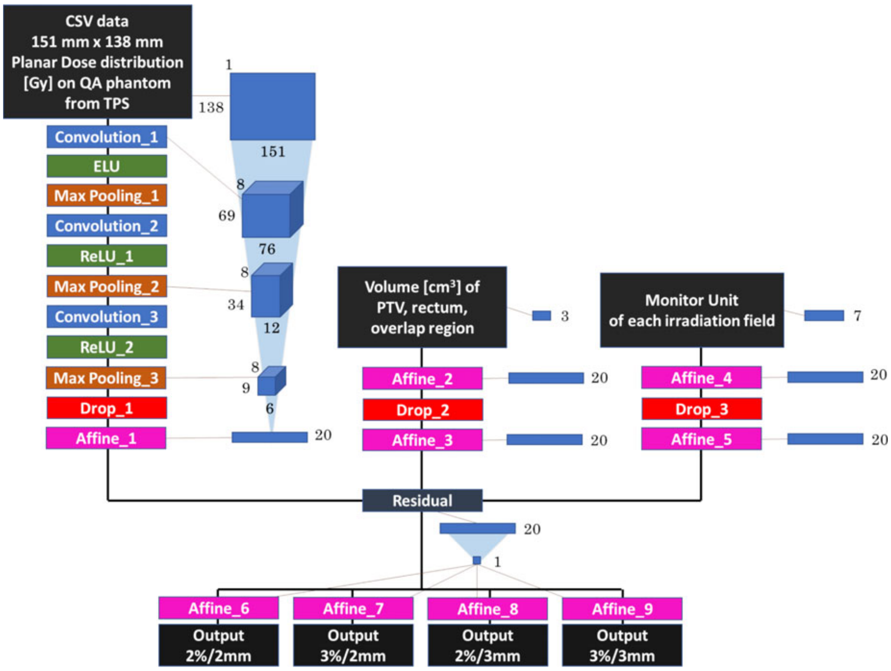

FIG. 2. Schematic diagram of convolutional neural network architecture for prediction of GPR values for the input dose distribution on the QA phantom. The network outputs prediction results for four criteria: 2%(global)/2 mm, 3%(global)/2 mm, 2%(global)/3 mm, and 3%(global)/3 mm. [Color figure can be viewed at wileyonlinelibrary.com]

Germany) treatment planning system (TPS) used by our hospital. All cases underwent prostate and partial seminal irradiation. Some irregular treatment cases, including pelvic nodes or other sites, were excluded. The cases were planned using seven-field IMRT, and almost the same gantry angle was used in all cases. The dose calculation algorithm was implemented using a Monte Carlo approach. All plans were verified using an RT-3000 QA phantom (R-TECH.INC, Tokyo, Japan) and GAFCHROMIC EBT3 film (Ashland Specialty Ingredients, Covington, USA). The patient treatment plans were applied to the QA phantom, which was exposed to the planned fluence. The film was inserted at the center of the phantom, in the sagittal plane. After exposure, the sagittal planar dose distribution on the film was scanned and 2D gamma analysis was performed to compare the calculated and measured dose distributions. The gamma agreement index and percentage GPR were then calculated. The gamma agreement index values were calculated using four gamma criteria pairings: 2%(global)/2 mm, 2%(global)/3 mm, 3%(global)/2 mm, and 3% (global)/3 mm. The GPR was defined as the percentage of points for which the gamma index was less than one, and the dose evaluation region as that in which the dose exceeded more than 30% of maximum dose. This relatively high threshold value was chosen to minimize the effect of noise arising from the use of EBT3 film.[17] The measured GPR results of four criteria were used for the supervised learning.

2.B. Datasets for CNN

The dose distribution was calculated from the TPS and extracted in comma-separated value format, with a height of 138 mm, width of 151 mm, and pixels of 1 mm in height and width. The data were held in a 138 × 151 matrix containing dose distribution information in gray (Gy) units. Sixty sagittal planar dose distributions in the QA phantom were collected in the same format. Three types of information were used as input data for deep learning. The first set comprised the sagittal planar dose distributions on the isocenter, the second set comprised the volumes of the PTV and rectum, and of the overlapping regions between the structures, measured in cm^3, and the third set comprised the monitor unit (MU) values for each field. The three volumes were structured by the same physician. Prostate and base of seminal vesicle were defined as CTV, and PTV was defined as the enlarged structure of CTV to entire perimeter with 7mm, except for posterior with 5mm. Previous studies reported that the GPR was affected by the complexity of the plan. The number of MU was a factor for assessing beam and overall plan complexity and deliverability.[11,18–21] In terms of volume, several studies have reported that patient geometric features were correlated with plan complexity and quality, especially PTV, OAR (rectum), and the overlapping region.[22,23] We therefore used the MU and volume as additional input information. For supervised learning, the measured GPR values of four criteria were used as the training data output values. The 40 cases were used for training and validation set in fivefold cross-validation, and the remaining 20 cases were used for the test set.

2.C. Architecture of the CNN

Figure 2 shows a schematic diagram of the CNN architecture used in this study. This comprised 15 layers at the longest stream, and contained three convolutional layers, three activation layers (one ELU layer and two ReLU layers), and three

Table I. Specifications of the convolutional neural network depicted in Fig. 2.

Layer	Output	Kernel	Padding	Stride	Number of parameters
Input (dose distribution)	138 × 151 × 1				0
Convolution_1	69 × 76 × 8	7 × 7	3 × 3	2 × 2	400
ELU	69 × 76 × 8				0
Max Pooling_1	69 × 26 × 8	1 × 3	0 × 0	1 × 3	0
Convolution_2	67 × 24 × 8	5 × 5	1 × 1	1 × 1	1608
ReLU_1	67 × 24 × 8				0
Max Pooling_2	34 × 12 × 8	2 × 2	0 × 0	2 × 2	0
Convolution_3	34 × 12 × 8	3 × 3	1 × 1	1 × 1	584
ReLU_2	34 × 12 × 8				0
Max Pooling_3	9 × 6 × 8	4 × 2	0 × 0	4 × 2	0
Drop_1	9 × 6 × 8				0
Affine_1	20				8660
Input [volume (cm^3)]	3				0
Affine_2	20				80
Drop_2	20				0
Affine_3	20				420
Input (MU value)	7				0
Affine_4	20				160
Drop_3	20				0
Affine_5	20				420
Residual	20				0
Affine_6	1				21
Affine_7	1				21
Affine_8	1				21
Affine_9	1				21

Max pooling layers.[24,25] A random dropout layer was employed to mitigate overfitting,[26] and an affine layer with 20 neurons was placed after this. The dose distribution data were analyzed through these layers. The volume values and the MU values were analyzed through relatively simple layers comprising two affine layers with 20 neurons and a random dropout layer. The drop rates of the three drop layers in this CNN were set to 0.25. The three input streams were connected to the residual layer to integrate the information. The output from this residual layer was the sum of the three scalar values from the three upstream affine layers in each neuron. Finally, the residual layer was connected to four final affine single-neuron layers to predict the GPR values for the four gamma criteria. The Huber loss function was applied in the output layer to ensure stable learning,[27] and multi-target regression was used to predict all four results at once. The specifications of the layers are summarized in Table I. The hyper-parameters were determined by a random search method with a training set of 40 cases.

2.D. Setting and environment

The network was optimized using an algorithm called Adam, derived from adaptive moment estimation. This was selected for its robustness. It is applied to a wide range of optimization problems in machine learning, as it requires little memory and is computationally efficient when used for gradient-based optimization.[28] During a training process, the parameters, weights, and biases of the CNN were changed many times through the learning iteration called an epoch. In this study, the weight and bias values were set at the best epoch in which the validation error was a minimum value. If the learning iteration was done with one case, the best epoch depended on a randomly selected single case, and it could result in unstable learning and a lack of generality. To perform a stable training, we used the mini-batch method in which a learning iteration is implemented with several randomly selected data. The mini-batch was set to eight and the number of epochs to 3000. Training was conducted on a PC with an Intel(R) Core(TM) i7–6700K 4.0 GHz CPU and 4 GB of RAM. No GPU was used. Full training of the CNN with fivefold cross-validation described below took approximately 2 h in this environment.

2.E. Model training and evaluation

We used 40 cases for training and validation sets and used the remaining 20 cases for the test set to evaluate the trained models. Figure 3 shows the workflow of the training and evaluation process. In this study, we implemented a fivefold cross-validation method. In each fold, we obtained a trained model and a validation result of eight cases. By collecting each validation result from all five folds, we acquired the validation results of 40 cases. Through this fivefold

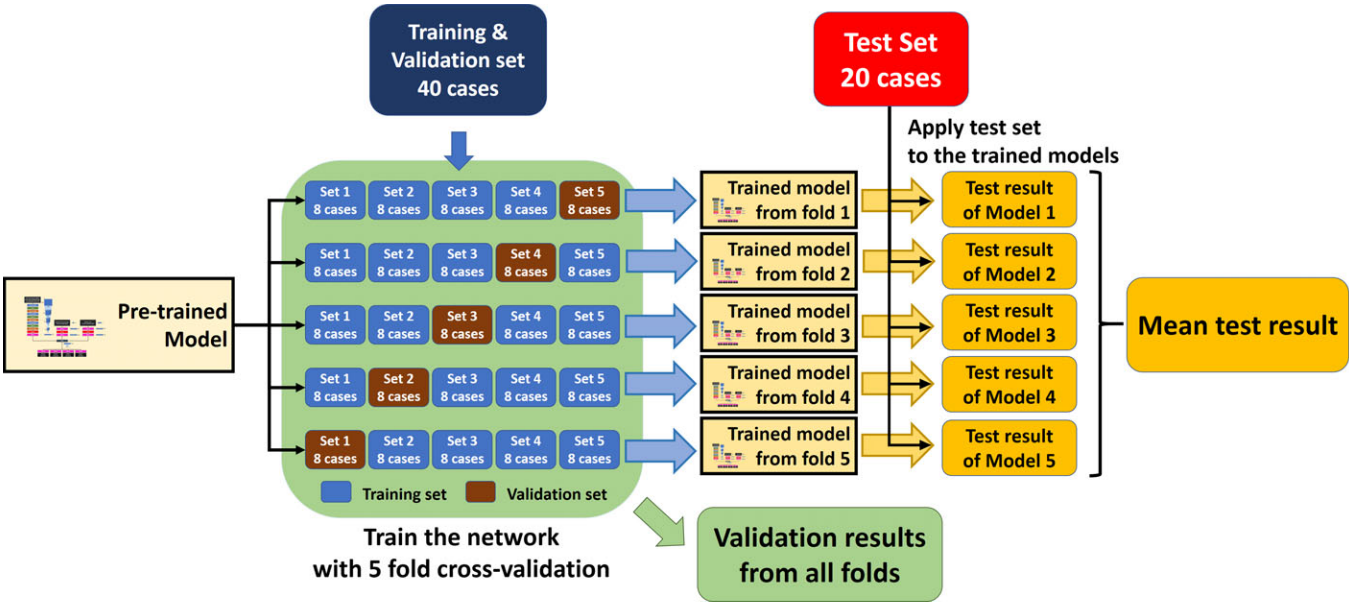

FIG. 3. A schematic diagram of the training and evaluation process with fivefold cross-validation and model averaging. [Color figure can be viewed at wileyonlinelibrary.com]

cross-validation, the five trained models were produced. To reduce generalization error, we used model averaging. We applied the test set to each trained model and then averaged the test results from the five trained models to acquire the mean test result. The validation result and mean test result are summarized in the Results section.

2.F. Comparison between CNN architecture patterns and the linear regression model

As depicted in Fig. 2, we produced a CNN model that had three input branches for accepting dose distribution, volumes, and MU. To reveal the contributions from each branch, we implemented several architecture patterns and then evaluated them. Figure 4 presents the list of architecture patterns. Pattern 1 had three branches, which is the same as Fig. 2. Patterns 2 and 3 had two branches, both of which dealt with dose distribution. The difference between them was that another branch of pattern 2 dealt with volume and pattern 3 dealt with MU values. Pattern 4 had only one branch and dealt with dose distribution.

In addition, we implemented a machine learning model. Several prior studies have reported that machine learning models, using information from TPS, were useful for prediction of various factors.[11,29,30] In this study, we also implemented a machine learning-based linear regression model to predict GPR. The volume of three structures and MU values of each field were input into the linear regression model. The fivefold cross-validation method was implemented, and after acquiring the trained model, the test set was applied. The evaluation results of these CNN architecture models and the linear regression model are summarized in the Results section.

3. RESULTS

The GPR values predicted by the model are plotted against the measured values in Fig. 5. If the prediction accuracy had been perfect, the points would form a diagonal line on the graph. In practice, the predicted GPR values deviated from the measured values, though linearity between the two was observed for each of the four gamma criteria.

Figure 6 shows a boxplot of each group of four criteria results in validation set and test set, and Table II lists mean values, standard deviations, median values, mean absolute error (MAE), root mean squared error (RMSE), and Spearman rank correlation coefficients. In all four criteria groups, the difference of mean values and median values between predicted and measured values in both the validation and test set was less than 1%, except for 2%/2 mm in the test set (<2%). The Spearman rank correlation coefficients between measured and predicted GPR values at four criteria in the validation set were 0.73 at 2%/2 mm, 0.72 at 3%/2 mm, 0.74 at 2%/3 mm, and 0.65 at 3%/3 mm, respectively ($P < 0.01$). On the other hand, the Spearman rank correlation coefficients in the test set were 0.62 ($P < 0.01$) at 2%/2 mm, 0.56 ($P < 0.01$) at 3%/2 mm, 0.51 ($P = 0.02$) at 2%/3 mm, and 0.32 ($P = 0.16$) at 3%/3 mm. This confirmed a strong or moderate correlation between the measured and predicted values.

Figure 7 shows the difference between measured and predicted GPR values for each case. The standard deviations in the validation set were 2.2% at 2%/2 mm, 1.5% at 3%/2 mm, 1.2% at 2%/3 mm, and 0.9% at 3%/3 mm, with maximum differences of 5.8%, 4.5%, 4.7%, and 3.0%, respectively. Test set standard deviations were 1.9% at 2%/2 mm, 1.3% at 3%/2 mm, 1.4% at 2%/3 mm, and 1.1% at 3%/3 mm, with

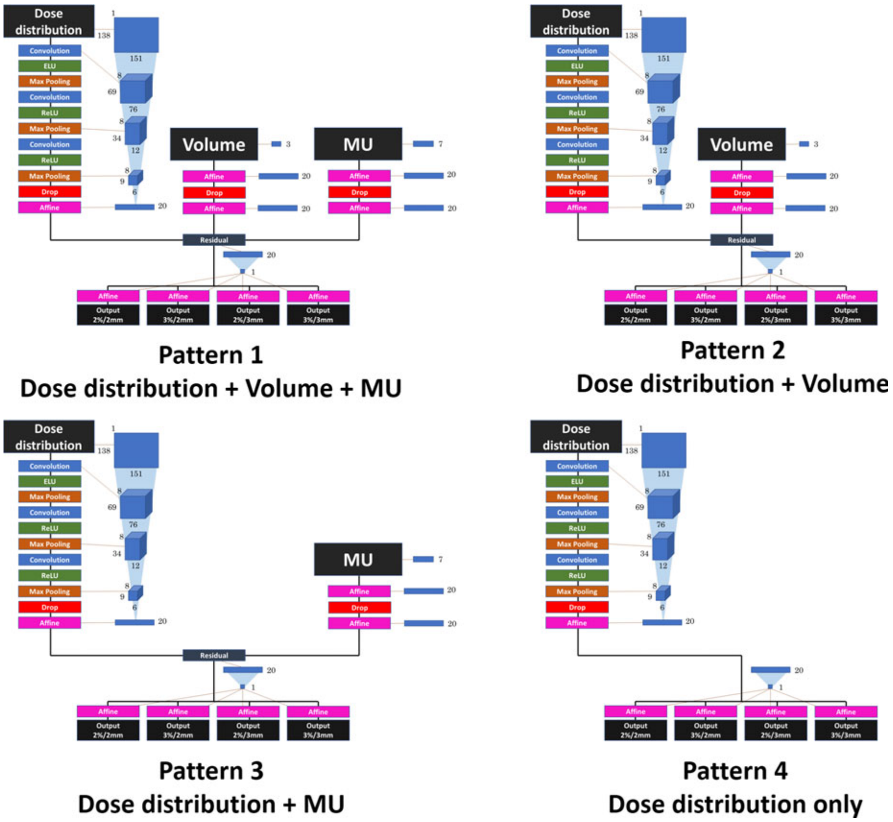

FIG. 4. CNN architecture patterns. Pattern 1 had three branches and was the same as Fig. 2. Patterns 2 and 3 had two branches, both of which dealt with dose distribution. The difference between them was that another branch of pattern 2 dealt with volume, and pattern 3 dealt with MU values. Pattern 4 had only one branch and dealt with dose distribution. [Color figure can be viewed at wileyonlinelibrary.com]

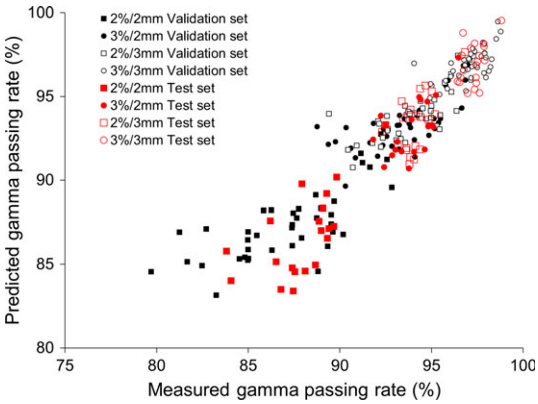

FIG. 5. Scatter plot of predicted against measured GPR values. The predicted values were derived from the CNN-based model (pattern 1; Fig. 4) using fivefold cross-validation. The results for 3%(global)/3 mm, 2%(global)/3 mm, 3%(global)/2 mm, and 2%(global)/2 mm are represented by open circles, open squares, closed circles, and closed squares, respectively. The black points show the validation results, and the red (online version only) points show the test results. [Color figure can be viewed at wileyonlinelibrary.com]

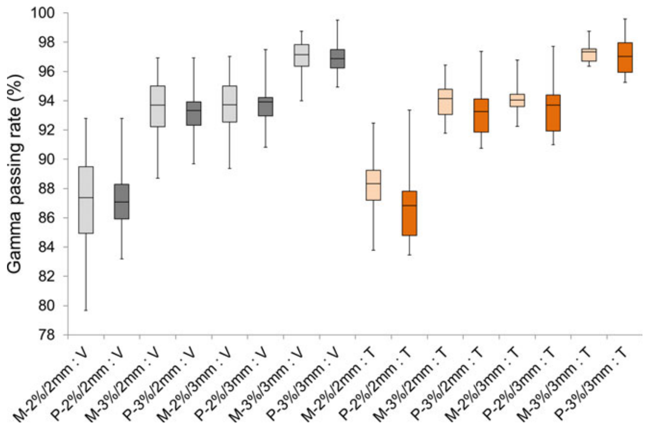

FIG. 6. Box plots of measured and predicted GPR values for each criterion (pattern 1; Fig. 4). M and P on the horizontal axis denote Measured and Predicted. V and T denote the validation set and the test set, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

maximum differences of −3.6%, −3.0%, −2.8%, and −2.3%, respectively. The standard deviations of the differences between validation and test results were almost the same.

Figure 8 shows the relationships between the GPR values from this study at the criteria of 2%/2 mm, 3%/2 mm, 2%/3 mm, and 3%/3 mm. The solid and dotted lines represent linear approximations based on the least mean squares method. The linear relationships between the criteria were observed.

Figure 9 compares results among several CNN architecture patterns and the linear regression model. The results of the 2%/2 mm criterion in those model patterns are summarized in Table III. The architecture pattern 1 was the most

TABLE II. Results of predicted and measured GPR values (%) for each criterion (pattern 1; Fig. 4).

	Mean	SD	Median	MAE	RMSE	S_r
Measured 2%/2 mm validation set	87.13	3.24	87.39	1.62	2.17	0.73 (P < 0.01)
Predicted 2%/2 mm validation set	87.38	2.11	87.07			
Measured 3%/2 mm validation set	93.40	2.12	93.70	1.05	1.47	0.72 (P < 0.01)
Predicted 3%/2 mm validation set	93.37	1.49	93.32			
Measured 2%/3 mm validation set	93.74	1.77	93.74	0.84	1.20	0.74 (P < 0.01)
Predicted 2%/3 mm validation set	93.80	1.38	93.91			
Measured 3%/3 mm validation set	96.98	1.12	97.14	0.64	0.87	0.65 (P < 0.01)
Predicted 3%/3 mm validation set	96.92	1.01	96.87			
Measured 2%/2 mm test set	88.03	2.00	88.35	1.93	2.24	0.62 (P < 0.01)
Predicted 2%/2 mm test set	86.78	2.55	86.83			
Measured 3%/2 mm test set	93.97	1.18	94.15	1.28	1.50	0.56 (P < 0.01)
Predicted 3%/2 mm test set	93.15	1.68	93.25			
Measured 2%/3 mm test set	94.11	0.92	94.04	1.31	1.53	0.51 (P = 0.02)
Predicted 2%/3 mm test set	93.44	1.78	93.71			
Measured 3%/3 mm test set	97.25	0.59	97.34	0.89	1.11	0.32 (P = 0.16)
Predicted 3%/3 mm test set	96.99	1.24	97.03			

SD: standard deviation; MAE: mean absolute error; RMSE: root mean squared error; S_r: Spearman rank correlation coefficient.

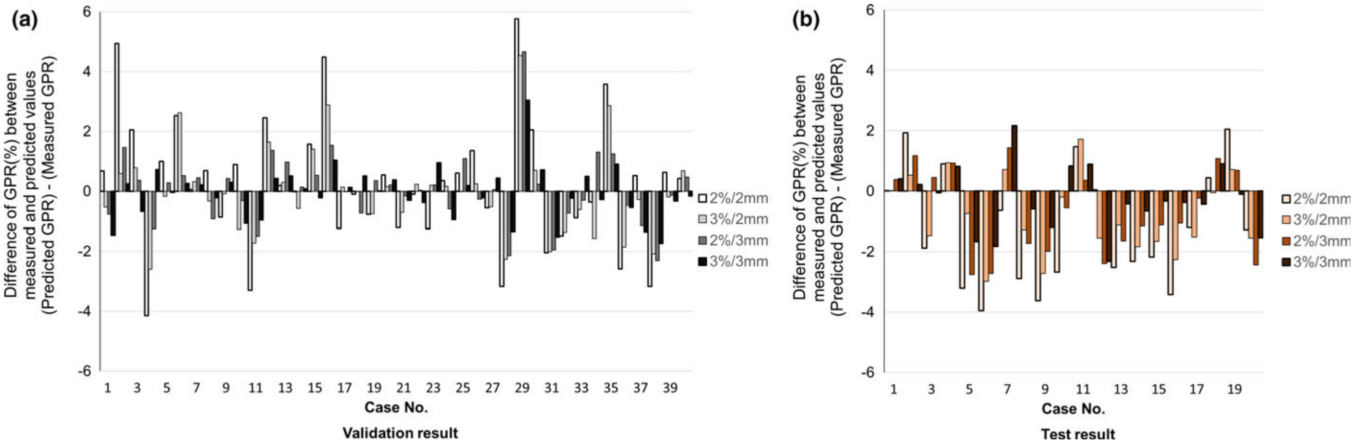

FIG. 7. Difference between measured and predicted GPR values in each case (pattern1; Fig. 4). The left figure shows the validation set results of 40 cases, and the right figure shows the test results of 20 cases. [Color figure can be viewed at wileyonlinelibrary.com]

accurate result in MAE, RMSE, and Spearman rank correlation coefficient. Comparing the results of patterns 1, 2, and 3, MAE and RMSE improved by inputting volume and MU values. The results of the linear regression model were also presented and the MAE and RMSE, in both the validation and test sets, were almost the same as those of patterns 1 and 2. However, the Spearman rank correlation coefficient decreased in the test set.

4. DISCUSSION

The GPR values predicted by the CNN-based model proposed in this study were strongly or moderately linearly correlated with the measured values for both the validation and test results. These results demonstrated the CNN to have successfully learned the features in the input data, producing predictions of GPR values. The CNN-based model output predicted GPR values at four criteria: 3%/3 mm, which is a widely and conventionally used criterion; 3%/2 mm, which is the criterion recommended in the TG218 report; and the other strict criteria of 2%/3 mm and 2%/2 mm. The model was demonstrated to adapt to these recent trends in gamma evaluation. The multi-institution study of gamma evaluation conducted by Scott et al. reported linear relationships between the above criteria.[31] Our own results presented in Fig. 8, also suggested linear relationships between the criteria. Our predicted GPR values were therefore in agreement with the results from previous studies. The comparison results shown in Fig. 9 and Table III revealed that the prediction accuracy improved after adding the volume and MU branches. Although the pattern 4 model, where only the dose distribution branch was able to predict GPR values, the MAE and

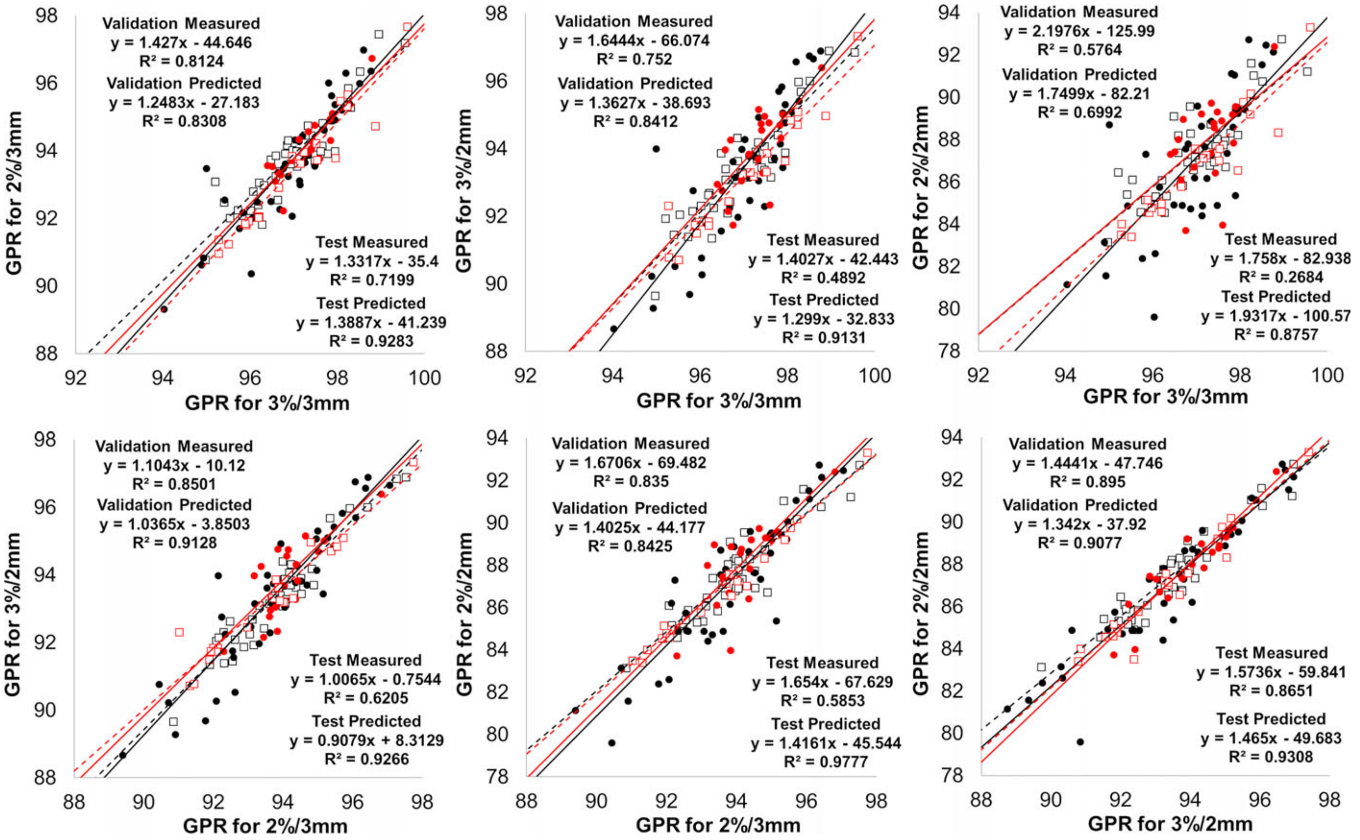

FIG. 8. Relationships between GPR values for four criteria (pattern 1; Fig. 4). Closed circles and open squares denote measured and predicted GPR values, and black and red (online version only) points denote validation and test prediction results, respectively. The solid and dotted lines represent the linear approximation of least squares regression for the measured and predicted GPR values, and the black and red lines represent the validation and test results, respectively. R^2 is the determination coefficient. [Color figure can be viewed at wileyonlinelibrary.com]

RMSE in the test set were larger than those in the other models. On the other hand, linear regression model could predict GPR values of the validation set accurately, in terms of MAE and RMSE. However, the Spearman rank correlation coefficient of the linear regression model in the test set was lowest in all other models in Fig. 9. The Spearman rank correlation coefficient for the test result obtained using the linear regression model was 0.36, which decreased from the validation result of 0.70. As shown by the linear regression analysis plot indicated in Fig. 9, some outlier test results were observed, implying that this linear regression model could not adapt to a specific test case and that the trained linear regression model lacked robustness to new data. Alternatively, the outliners improved in test results using CNN models. This indicated that the CNN model with dose distribution was able to acquire robustness. As a result, the Spearman rank correlation coefficient for the test result of the CNN model was improved compared with that for the test result of the linear regression model. Given these results, the CNN model with dose distribution can be useful for predicting GPR values with additional information helping to improve the prediction accuracy, and may help to acquire the robustness for analyzing new data.

However, the match between measured and predicted GPR was not perfect, and large differences were observed especially in cases where the measured GPR value was low. In Fig. 7, the maximum differences in validation set arose from a single case, which suggested either or both that the model used in the study was unable to adapt to this particular case and/or that the amount of noise present in the case produced an incorrect measured value. In all four criteria groups, The CNN-based model proposed in this study tended to overestimate the GPR in the validation set, especially in cases where the GPR value was low. On the other hand, the model tended to underestimate the GPR in the test set. These differences suggested that the model did not possess perfect generality, and produced a bias in both results. Furthermore, differences in prediction accuracy between the validation and test results were also observed. In Table II, The MAE and RMSE of the test results were larger than those of the validation results, and the Spearman rank correlation coefficients of the test results were lower than the validation results. Thus, the trained model could adopt to the test set, albeit imperfectly. These differences were assumed to be caused by specific factors that reduced the learning accuracy of the CNN. The first factor was the small volume of training data used. The accuracy with which a CNN learns is directly related to the number of training data in the input. However, the balance of the input data, that is, the balance between cases with high GPR values and cases with low GPR values, was also plays an important role in the accuracy of learning. If cases with high GPR values are over-represented in the input, the accuracy of

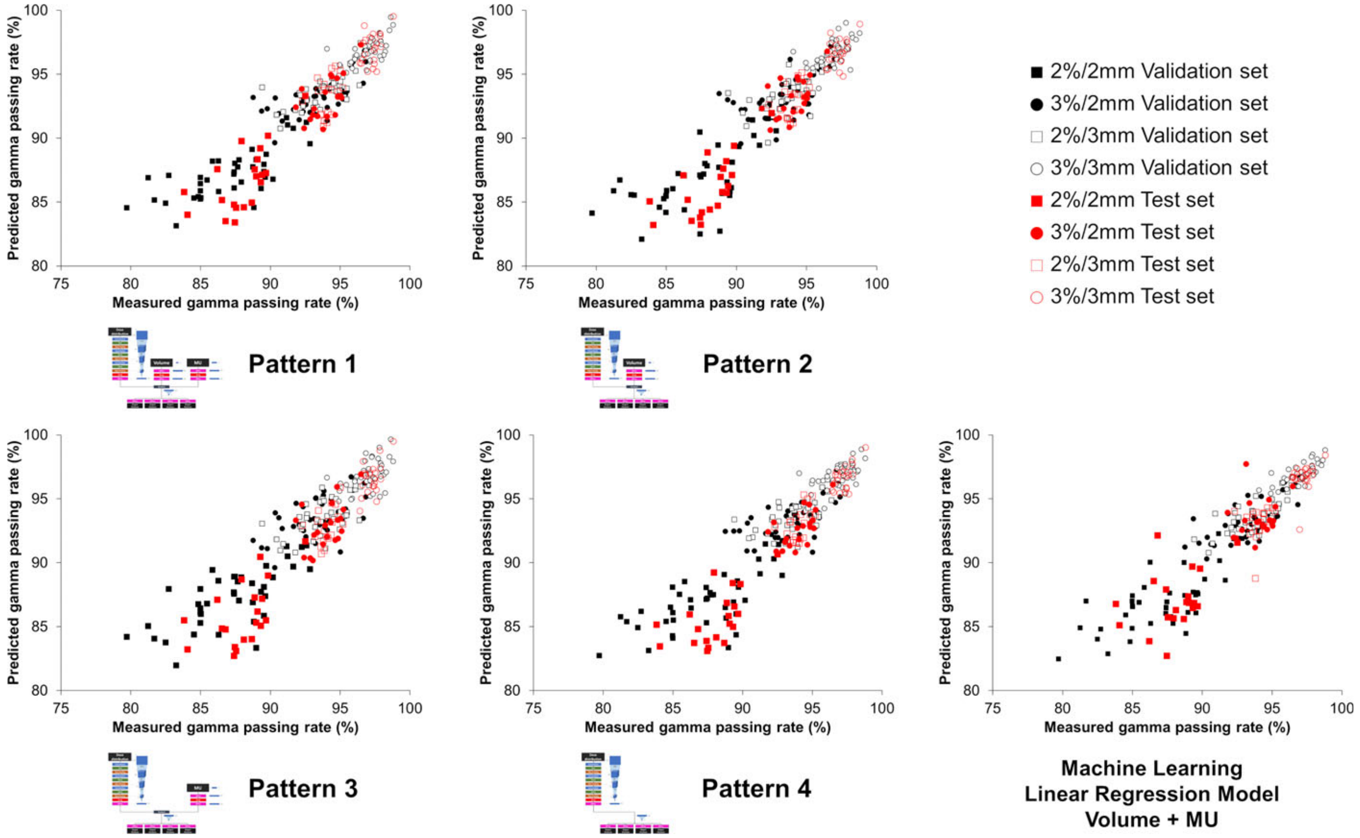

FIG. 9. Comparison of the four CNN architecture patterns and the linear regression model. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE III. 2%/2 mm criterion results of CNN architecture patterns.

	MAE	RMSE	S_r
Pattern 1			
2%/2 mm validation set	1.62	2.17	0.73
2%/2 mm test set	1.93	2.24	0.62
Pattern 2			
2%/2 mm validation set	1.84	2.45	0.66
2%/2 mm test set	2.20	2.53	0.67
Pattern 3			
2%/2 mm validation set	1.82	2.29	0.68
2%/2 mm test set	2.52	2.90	0.55
Pattern 4			
2%/2 mm validation set	1.84	2.36	0.63
2%/2 mm test set	2.59	2.93	0.60
Linear regression model volume + MU			
2%/2 mm validation set	1.79	2.15	0.70
2%/2 mm test set	2.15	2.50	0.36

MAE: mean absolute error RMSE: root mean squared error S_r: Spearman rank correlation coefficient.

learning and prediction will be lower. In the prostate treatment dataset used in this study, only 15 cases had GPR values lower than 85% when the 2%/2 mm criterion was applied. This limited the usefulness of the training data. To improve the prediction accuracy of the model, the number of training cases should be increased, and the balance of the data should be kept. The second factor concerned the accuracy of dose distribution measurement in the input data. Measurements were taken using EBT3 film, and errors may have been introduced by the response homogeneity of the film, the light scanning effect, the calibration of the film, and the setup of the phantom, among other things.[17] These errors would introduce noise into the input data, which in turn would render the predictions less accurate. This effect was especially pronounced in the 2%/2 mm case.

Despite these drawbacks, the CNN-based model proposed in this study successfully learned the features in the input data, and predicted GPR values with more strict criteria than 3%/3 mm. Rather, strict criteria, that is, 2%/2 mm, may be more suitable for learning than the 3%/3 mm criteria. This can be inferred from the results in Table II, in which the Spearman rank correlation coefficient of 3%/3 mm was lower than the other three criteria for both the validation and test results. It is possible that the 3%/3 mm criterion inadequately detected measurement error, resulting in a mismatch between measured and predicted GPR values. This inference is supported by previous studies that reported the insensitivity of the 3%/3 mm criterion.[13–15] Although the MAE and RMSE were minimal in the 3%/3 mm, the stricter criteria, that is, 2%/2 mm, may be more suitable for prediction from the viewpoint of detecting the plan to which we should pay attention. As training was based on the dose distribution in a QA phantom, this may have produced a more direct causal analysis. The dose distribution has spatial information and contains indirect factors that may increase the complexity of dose delivery. As shown in Fig. 9, the CNN architecture pattern 4, which accepted only dose distribution, was also able to predict GPR via a moderate linear correlation. Data on the dose distribution can be suitable for use in predicting gamma evaluation.

The CNN-based model proposed in this study used three types of input information that can be readily collected from any TPS, and it predicted GPR values in four criteria with global normalization. Clinically, the 2%(Local)/2 mm criterion is the most sensitive to clinically relevant errors. However, selecting the local normalization will also cause the low-dose regions to have unrealistic requirements for dose accuracy.[15] We therefore constructed the novel model with the measurement data, including global normalization, to adapt to patient-specific QA. The CNN model has a relatively simple structure, which can mitigate the overfitting caused by a limited amount of training data. This simplicity can be beneficial in clinical situations because accumulating large clinical datasets can be time-consuming, making it difficult to implement a CNN model. In addition, simplicity can reduce the learning period; furthermore, the model used in this study can be run on a standard PC without GPU. The results of this study suggested that a CNN model with dose distribution was able to predict the GPR value as well as a linear regression model in terms of MAE and RMSE and could have a potential to increase the prediction accuracy considering robustness to new data. However, the CNN model is more complex than traditional machine learning, making it difficult to interpret the model. Therefore, carefully verifying a trained CNN model with test data and evaluating the prediction accuracy through various analyses are essential for clinical usage. Although the extent to which measurements can be reduced depends on the accuracy of the CNN output, the prediction model is practical and can help clinical workers reduce the measurement burden of patient-specific QA when verifying the dose distribution.

5. CONCLUSIONS

In this study, we developed a CNN-based prediction model for patient-specific QA of dose distribution in prostate treatment. As training data, three types of information were used: sagittal planar dose distributions from a QA phantom, the volume of the PTV and rectum and the overlapping region between them, and the MU values for each irradiation field. Strong or moderate correlations were found between the measured and predicted GPR values, and dose distribution can be a useful data for GPR prediction. Our results suggest that deep learning can be applied to the prediction of patient-specific QA dose distributions in prostate treatment.

ACKNOWLEDGMENTS

The authors would like to express our appreciation to the member of Department of Radiation Oncology, Tohoku University School of Medicine for their useful discussion of this manuscript. We would also like thank Radiological Technologists of the Sendai Medical Center for providing us excellent assistance and advice.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

[a]Author to whom correspondence should be addressed. Electronic mail: kadoya.n@rad.med.tohoku.ac.jp; Telephone: +81-22-717-7312; Fax:+81-22-717-7316.

REFERENCES

1. Portelance L, Chao KC, Grigsby PW, Bennet H, Low D. Intensity-modulated radiation therapy (IMRT) reduces small bowel, rectum, and bladder doses in patients with cervical cancer receiving pelvic and para-aortic irradiation. *Int J Radiat Oncol Biol Phys*. 2001;51:261–266.
2. LoSasso T, Chui CS, Ling CC. Comprehensive quality assurance for the delivery of intensity modulated radiotherapy with a multileaf collimator used in the dynamic mode. *Med Phys*. 2001;28:2209–2219.
3. Ezzell G, Burmeister J, Dogan LW, et al. IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119. *Med Phys*. 2009;36:5359–5373.
4. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys*. 1998;25:656–661.
5. ESTRO. *ESTRO Booklet No. 9: Guidelines for the Verification of IMRT*. Brussels: ESTRO; 2008.
6. ICRU 83. Prescribing, recording and reporting intensity-modulated photon-beam therapy (IMRT). *J ICRU*. 2010;10:NP3.
7. Van Esch A, Bohsung J, Sorvari P, et al. Acceptance tests and quality control(QC) procedures for the clinical implementation of intensity modulated radiotherapy (IMRT) using inverse planning and the sliding window technique: experience from five radiotherapy departments. *Radiother Oncol*. 2002;65:53–70.
8. Abolaban F, Zaman S, Cashmore J, Nisbet A, Clark CH. Changes in patterns of intensity-modulated radiotherapy verification and quality assurance in the UK. *Clin Oncol*. 2016;28:e28–e34.
9. Sumida I, Yamaguchi H, Kizaki H, et al. Three-dimensional dose prediction based on two-dimensional verification measurements for IMRT. *J Appl Clin Med Phys*. 2014;15:133–146.
10. Kurosu K, Sumida I, Mizuno H, et al. Curtailing patient-specific IMRT QA procedures from 2D dose error distribution. *J Radiat Res*. 2015;57:258–264.
11. Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning. *Med Phys*. 2016;43:4323–4334.
12. Valdes G, Chan MF, Lim SB, Scheuermann R, Deasy JO, Solberg TD. IMRT QA using machine learning: a multi-institutional validation. *J Appl Clin Med Phys*. 2017;18:279–284.
13. Stojadinovic S, Ouyang L, Gu X, Pompoš A, Bao Q, Solberg TD. Breaking bad IMRT QA practice. *J Appl Clin Med Phys*. 2015;16:154–165.
14. Nelms BE, Chan MF, Jarry G, Lemire M, et al. Evaluating IMRT and VMAT dose accuracy: practical examples of failure to detect systematic errors when applying a commonly used metric and action levels. *Med Phys*. 2013;40:111722.
15. Miften M. TH-A-BRC-03: AAPM TG218: measurement methods and tolerance levels for patient-specific IMRT verification QA. *Med Phys*. 2016;43:3852–3853.
16. Interian Y, Rideout V, Kearney VP, et al. Deep nets vs expert designed features in medical physics: an IMRT QA case study. *Med Phys*. 2018;45:2672–2680.
17. Martisikova M, Ackermann B, Jakel O. Analysis of uncertainties in Gafchromic EBT film dosimetry of photon beams. *Phys Med Biol*. 2008;53:7013–7027.

18. Mohan R, Arnfield M, Tong S, Wu Q, Siebers J. The impact of fluctuations in intensity patterns on the number of monitor units and the quality and accuracy of intensity modulated radiotherapy. *Med Phys*. 2000;27:1226–1237.
19. McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. *Med Phys*. 2010;37: 505–515.
20. Craft D, Suss P, Bortfeld T. The tradeoff between treatment plan quality and required number of monitor units in intensity-modulated radiotherapy. *Int J Radiat Oncol Biol Phys*. 2007;67:1596–1605.
21. Oliver M, Bush K, Zavgorodni S, Ansbacher W, Beckham WA. Understanding the impact of RapidArc therapy delivery errors for prostate cancer. *J Appl Clin Med Phys*. 2011;12:32–43.
22. Lee T, Hammad M, Chan TCY, Craig T, Sharpe MB. Predicting objective function weights from patient anatomy in prostate IMRT treatment planning. *Med Phys*. 2013;40:121706.
23. Boutilier JJ, Lee T, Craig T, et al. Models for predicting objective function weights in prostate cancer IMRT. *Med Phys*. 2015;42: 1586–1595.
24. Clevert DA, Unterthiner T, Hochreiter S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). ICLR 2016;(arXiv:1511.07289)
25. Vinod N, Geoffrey EH. Rectified linear units improve restricted Boltzmann machines. ICML; 2010:807–814.
26. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–1958.
27. Huber PJ. Robust estimation of location parameter. *Ann Math Stat*. 1964;35:73–101.
28. Diederik PK, Jimmy B. ADAM: A method for stochastic optimization. ICLR; 2014 (arXiv:1412.6980)
29. Zhu X, Ge Y, Li T, et al. A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Med Phys*. 2011;38:719–726.
30. Carlson JNK, Park JM, Park SY, et al. A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. *Phys Med Biol*. 2016;61:2514–2531.
31. Scott BC, Bess S, Rachael W, et al. Technical note: relationships between gamma criteria and action levels: results of a multicenter audit of gamma. *Med Phys*. 2016;43:1501–1506.