# Pretreatment patient-specific IMRT quality assurance: A correlation study between gamma index and patient clinical dose volume histogram

M. Stasi, S. Bresciani,[a)] A. Miranti, A. Maggio, and V. Sapino
*Department of Medical Physics, IRCC: Institute for Cancer Research and Treatment at Candiolo (TO) 10060, Italy*

P. Gabriele
*Department of Radiotherapy, IRCC: Institute for Cancer Research and Treatment at Candiolo (TO) 10060, Italy*

(Received 6 April 2012; revised 31 October 2012; accepted for publication 31 October 2012; published 30 November 2012)

**Purpose:** The aim of this work is to investigate the predictive power of a common conventional intensity modulated radiation therapy (IMRT) quality assurance (QA) performance metric, the gamma passing rate (%GP), through the analysis of the sensitivity and of the correlation between %GP and different dose discrepancies between planned dose-volume histogram (DVH) and perturbed DVH. The perturbed DVH is calculated by using a dedicated software, 3DVH (Sun Nuclear Corporation, Melbourne, FL), which is able to modify the dose distribution calculated by the treatment planning system (TPS) according to the dose discrepancies detected with planar measurements in order to predict the delivered 3D dose distribution in the patient.

**Methods:** Twenty-seven high-risk prostate cancer (PP) patients and 15 head and neck (HN) cancer patients, treated with IMRT technique, were analyzed. Pretreatment verifications were performed for all patients' plans by acquiring planar dose distributions of each treatment field with 2D-diode array. Measured dose distributions were compared to the calculated ones using the gamma index (GI) method applying both global (Van Dyk) and local normalization, and %GP were generated for each pair of planar doses using the following acceptance criteria: 1%/1, 2%/2, and 3%/3 mm. Planar dose distributions acquired during pretreatment verifications, together with patient's DICOM RT plan, RT structure set, and RT dose files from TPS were loaded into the 3DVH software. Percentage dose differences (%DE) between DVHs, obtained by TPS and by 3DVH, were calculated; statistical correlation between %DE and %GP was studied by using Pearson's correlation coefficient ($r$). This analysis was performed, for each patient, on planning target volumes and on some typical organs at risk of the prostatic and head and neck anatomical district. The sensitivity was calculated to correctly identify the pretreatment plans with high dose errors and to quantify the incidence of false negatives, on varying the gamma index method.

**Results:** Analysis of %DE vs %GP showed that there were only weak correlations (Pearson's $r$-values $< 0.8$). The results also showed numerous instances of false negatives (cases where high IMRT QA passing rates did not imply good agreement in anatomy dose metrics) and the reverse, mainly for the 3%/3 mm global gamma passing rate.

**Conclusions:** The lack of correlation between conventional IMRT QA performance metrics gamma passing rates and dose errors in DVHs values and the low sensitivity of 3%/3 mm global gamma method show that the most common published acceptance criteria have disputable predictive power for per-patient IMRT QA. © *2012 American Association of Physicists in Medicine*. [http://dx.doi.org/10.1118/1.4767763]

Key words: IMRT, dosimetry, gamma pass method, pretreatment quality assurance

## I. INTRODUCTION

Intensity modulated radiation therapy (IMRT) is a highly conformal external beam technique characterized by steep infield dose gradients, which allows better conformality of dose to the planning target volume (PTV) and better avoidance of organs at risk (OARs).

The increasing complexity of IMRT treatments along with its widespread adoption require an effective and extensive quality assurance (QA) program, both in terms of treatment machines delivery precision and treatment planning system (TPS) calculation accuracy.[2] Therefore, dedicated pretreatment patient-specific QA procedures are required for the sake of detecting possible mismatches between the dose calculated by TPS and the dose actually delivered by treatment machines.[1]

Conventional IMRT QA is usually performed by applying the patient plan to a phantom and comparing the measured and calculated phantom dose distributions, as recommended in the 2003 "Guidance Document" on IMRT.[3] Commonly, a point dose measurement using an ion chamber as well as a planardose measurement using radiographic film was traditionally implemented for dosimetric QA of treatment plans. However, there are now new products that allow the acquisition of absolute dose distribution measurements using detector arrays, either ion chamber or diode-based, in a 2D or 3D geometry.[4–8]

Phantom dose distributions are first calculated by a TPS and then delivered and measured at the treatment machine. A very common method to quantitatively compare measured and calculated dose maps is the calculation of gamma index (GI). This method, first introduced by Low *et al.*,[9] represents the minimum multidimensional distance between the measurement and calculation points in a space composed of dose and physical distance coordinates, scaled by preselected limits called *acceptance criteria* for distance to agreement (DTA) and dose differences (%DD).[10]

Depending on the normalization value of the dose difference between measured and calculated dose points, we distinguish between the local calculation method and the global calculation method.[11] A GI smaller than unity indicates that the measured absorbed dose agrees with the calculated one within the passing criteria. The goodness of a treatment plan is measured through the evaluation of the gamma passing rate (%GP), which represents the percentage of dose points per plan that comply the acceptance criteria.

The topic of IMRT QA, performed by the gamma method, has been widely investigated.[12] There have been many studies on suggested acceptance/action levels for planar IMRT QA.[14,15] Some of these studies base action levels on retrospective statistical analysis of the performance levels/metrics that have been achieved over many plans and IMRT beams.[14] The recent report of the AAPM Task Group 119 (Ref. 16) and some other papers[13,14,17,18] reported that the 3% dose difference and 3 mm DTA criteria is most commonly used by physicists in pretreatment IMRT QA. TG-119 proposed IMRT action level for %GP was 90% for per-beam planar analysis and 88%–90% for composite irradiations. However, a survey analysis[13] showed that when the institutions used 3%/3 mm criterion, %GP action level most commonly used was 95%. Some papers have suggested that GI has no correlation to clinically relevant criteria and have proposed a need for further study such as that being presented.[13,19] In the literature, suggested %GP which define plans to be *acceptable* for treatment was presented[14] but they only rely on *achievable results* rather than on *clinical* results. This lack of nexus between suggested %GP and its clinical meaning makes it difficult to give an interpretation of the obtained result and does not give any assurance about the goodness and the safety of the plan. Recent studies by Nelms *et al.* and Zhen *et al.* showed no correlation between the %GP and the magnitude of dose discrepancy between the 3D calculated dose and the actual delivered dose.[2] Moreover, a direct relationship between dose errors found during per-beam QA and 3D dose obtained from the composition of all beams cannot be assumed, since, after composition, dose errors of one field may be compensated by dose errors of other fields in an unknown way.[2,20] Furthermore, GI method is limited by the fact that it only determines the number of points out of tolerance without giving any information about their spatial location. Therefore, one cannot assume that the %GP of the entire plan corresponds to the one of the single organ.

The previous papers[2,19] used simulations and error-inductions to test correlation and sensitivity, while our study used retrospective analysis of real, low-density QA measurements. The novelty and the aim of this paper is right to investigate the correlation between %GP obtained during standard per-beam pretreatment QA tests, based on real retrospective data with a common 2D array (MapCheck - Sun Nuclear Corporation, Melbourne, FL) of actual clinical plans, with different acceptance dose discrepancy, between planned dose-volume-histogram (DVH) and patients' predicted DVH, calculated by 3DVH software (Sun Nuclear Corporation, Melbourne, FL).

## II. MATERIALS AND METHODS

### II.A. Patients

Twenty-seven high-risk prostate cancer patients, treated with 5-fields simultaneous integrated boost (SIB)-IMRT for prostate and pelvic nodes (PP), and 15 head and neck (HN) cancer patients, treated with 7-fields IMRT (12 SIB-IMRT and 3 non-SIB-IMRT), were enrolled in the study, for a total amount of 240 treatment fields tested. In all these cases IMRT is delivered by sliding window multileaf sequencing method. All inverse plans had been created using Varian Eclipse TPS v.8.0 (Varian Medical Systems, Palo Alto, CA) and the treatment was delivered using a 6 MV linear accelerator Clinac 600 C/D (Varian Medical Systems, Palo Alto, CA) equipped with a Millennium 120 multileaf collimator (MLC).

### II.B. Gamma index evaluation and analysis

Pretreatment verifications were performed for all patients' plans by acquiring plane dose distributions of each treatment field. Measurements were taken using the diode array MapCheck model 1175 (Sun Nuclear Corporation, Melbourne, FL), routinely used in our Institute, with absolute dose calibration, and the software MapCheck v. 5.02.00.02 (Sun Nuclear Corporation, Melbourne, FL). MapCheck is a $22 \times 22$ cm$^2$ array made out of 445 *n*-type solid-state diode detectors distributed at 7.07 mm along x and y axes in the $10 \times 10$ cm$^2$ central area, while in the remaining area the distance between diodes is doubled. "Measurement Uncertainty" was on in MapCheck software as per TG119 instructions.[21–23]

The per-beam IMRT QA dose distributions of each treatment plan were analyzed employing the GI method.[24] Adopting the method proposed by Zhen *et al.*,[19] %GP were generated for each pair of planes using both local and global gamma calculation method and the following acceptance criteria: 1%/1, 2%/2, and 3%/3 mm. Dose values below 10% (TH = 10%) of the per-beam normalization maximum dose were ignored. In the analyses and scatterplots we used the average %GP over all beams, per plan.

### II.C. Software

The 3DVH software provides a 3D dose distribution modified by perturbing through the dose errors detected during pretreatment measurements and the 3D calculated dose distribution. This perturbation is achieved by applying the planned dose perturbation (PDP) algorithm.[25] This algorithm uses conventional per-beam planar dose QA methods to feed a sophisticated three-dimensional (3D) perturbation system that "corrects" the original 3D patient dose as generated by the TPS and outputs a 3D patient dose grid that has built into it the manifestation of any errors detected by the planar QA. The PDP method does not introduce new sources of variation or error that may occur with an independent 3D dose algorithm (i.e., variations that might not be errors but just differences of the new algorithm vs the TPS algorithm). PDP will alter dose only if and where dose differences are detected in dosimetry array systems.[19,26] As output, 3DVH software compares the DVHs calculated by TPS to the ones obtained from perturbation, and it shows dose distributions, both alone and as comparison, in the sagittal, coronal, and axial planes of patients.[2]

### II.D. Dose comparison and differences estimation

Planar dose distributions acquired during pretreatment verifications, RTPlan (DICOM file containing all the information about the plan's parameters) and RTDose (DICOM file containing all the information about the dose distribution) exported from TPS, were loaded on the software 3DVH.

We calculated mean DVH values obtained both by 3DVH and by TPS. To verify if our procedures were affected by systematic errors, we performed a *t-test* statistical analysis between the mean values.

In order to evaluate, through a correlation analysis, the actual meaning of the %GP (related to the GI), percentage dose errors were calculated (%DE) according to the following equation:

$$\%\mathrm{DE} = \left|\frac{\mathrm{D}_{\mathrm{3DVH}} - \mathrm{D}_{\mathrm{TPS}}}{\mathrm{D}_{\mathrm{TPS}}}\right| * 100, \qquad (1)$$

where $\mathrm{D}_{\mathrm{3DVH}}$ was the dose value showed by the software 3DVH and $\mathrm{D}_{\mathrm{TPS}}$ was the dose value reported by TPS. This analysis was performed, for each patient, on PTVs and some OARs of the anatomical treated district, as shown in Table I.

TABLE I. Structures analyzed for the two different anatomical regions, head and neck (HN) and prostate and pelvic lymph nodes (PP). To side, analyzed doses are reported for each structure.

| Anatomical region | Structure | Analysed dose |
|---|---|---|
| PP | PTV1 (Pelvic lymph nodes) | $\mathrm{D}_{\mathrm{Mean}}$, $\mathrm{D}_{95}$ |
| | PTV2 (Boost) | $\mathrm{D}_{\mathrm{Mean}}$, $\mathrm{D}_{95}$ |
| | Rectum | $\mathrm{V}_{50}$, $\mathrm{V}_{70}$ |
| | Bladder | $\mathrm{D}_{\mathrm{Mean}}$ |
| | PTV boost | $\mathrm{D}_{\mathrm{Mean}}$, $\mathrm{D}_{95}$ |
| HN | Spinal cord | $\mathrm{D}_{2\%}$ |
| | Parotid R | $\mathrm{D}_{\mathrm{Mean}}$ |
| | Parotid L | $\mathrm{D}_{\mathrm{Mean}}$ |

For each PTV, we took into account as relevant dose values $\mathrm{D}_{\mathrm{Mean}}$ and $\mathrm{D}_{95\%}$ (dose to 95% volume). In the HN patients, for OARs, we considered: $\mathrm{D}_{2\%}$ (dose to 2% volume)[1] for the spinal cord and $\mathrm{D}_{\mathrm{Mean}}$ for parotids. In the PP patients, $\mathrm{V}_{50}$ and $\mathrm{V}_{70}$ (volume that received more than 50% and 70%, respectively) are taken into account for rectum and $\mathrm{D}_{\mathrm{Mean}}$ for bladder.

### II.E. Correlation and sensitivity analysis

Statistical correlation between %DE and %GP was studied by using Pearson's correlation coefficient ($r$).

From the database of 27 patients treated in PP, 42 $r$-values, corresponding to the correlation between the six %GP and the seven DVH dose values, were evaluated. From the 15 patients treated in HN, 24 $r$-values, corresponding to the six %GP and the four DVH dose values, were evaluated. %DE was assumed to be correlated with a determined %GP when the $p$-value, obtained from $r$, is $<0.05$. Statistical evaluations were performed using LabFit software.[27]

We compared the strength of correlation, in terms of $r$-values, obtained using the global calculation method with the one obtained using the local GI method. The comparison has been performed by using a paired samples $t$-test statistical analysis.

Further, in order to quantify the sensitivity of gamma index method, we calculated the number of "false negative (FN)" cases (cases where high IMRT QA passing rates did imply large errors in anatomy dose metrics) and of "true positive (TP)" cases (cases where low IMRT QA passing rates did imply large errors in anatomy dose metrics). In particular, we considered "FN" all those structures that have DVH errors higher than 5% (cases where low IMRT QA passing rates did not imply large errors in anatomy dose metrics), or 3% among those patients with %GP higher than 95% (using 3%/3 and 2%/2 mm both local and global method). Instead we considered "TP" all the cases that have DVH errors higher than 5% or 3% and %GP lower than 95%. From the "false negative" rate and the "true positive" rate, we calculated sensitivity that refers to the ability of the gamma index test to correctly identify the plan with dose errors greater than 5% or 3%.

## III. RESULTS

### III.A. Evaluation of the %GP

For all of the patients we evaluated the %GP using two different calculation methods and three different acceptance criteria, as previously described. Table II shows average %GP calculated for PP and HN patients. In the overall sample of 27 PP patients and 15 HN patients, we calculated a total number of 252 %GP data points. Maximum and minimum values were also reported.

### III.B. Dose comparison and differences estimation

The %DE was calculated for each patient, as reported in Table III for the patient number 14 (HN group). The total

TABLE II. Average %GP evaluated over the two data-set of patients (PP and HN). Different %GP are obtained depending on the calculation method and on the different acceptance criteria. Range of variability is also reported.

| Anatomical region | Methods | Acceptance criteria (mm) | Mean %GP ± SD | Range |
|---|---|---|---|---|
| PP | Global | 3%/3 | 97.2 ± 2.0 | 91.0–99.9 |
| | | 2%/2 | 90.3 ± 5.2 | 77.5–98.9 |
| | | 1%/1 | 59.9 ± 9.4 | 44.1–79.5 |
| | Local | 3%/3 | 94.3 ± 3.1 | 86.3–98.9 |
| | | 2%/2 | 85.0 ± 6.2 | 72.1–95.2 |
| | | 1%/1 | 53.6 ± 7.9 | 40.8–67.8 |
| HN | Global | 3%/3 | 98.1 ± 1.3 | 95.8–99.6 |
| | | 2%/2 | 93.8 ± 2.3 | 89.0–97.9 |
| | | 1%/1 | 67.4 ± 5.2 | 57.7–79.4 |
| | Local | 3%/3 | 95.9 ± 2.1 | 90.7–98.7 |
| | | 2%/2 | 89.4 ± 3.6 | 81.1–95.3 |
| | | 1%/1 | 60.4 ± 5.6 | 49.3–72.4 |

number of %DE data points was 189 for PP patients and 75 for HN patients. Mean DVH values and *t*-test *p*-values, are illustrated in Table IVa (PP group) and IVb (HN group). The *t*-test results between the planned and estimated DVH values showed that mean values were comparable ($p > 0.05$); this indicates that there were no systematic errors.

### III.C. Correlation and sensitivity analysis

Results of statistical correlation ($r$ and $r^2$) between %DE and %GP, calculated using six different acceptance criteria and methods, are shown in Tables Va (PP group) and Vb (HN group) along with the respective *p*-values. The *r*-values always being <0.8 indicated weak correlation between the %GP and the absolute percentage dose differences. Furthermore, *r*-values are negative for all correlations except for $V_{70}$ of rectum and $D_{Mean}$ of the parotid gland. These negative *r*-values proved a decrease in clinical metrics with increasing passing rate. All the positive *r*-values are not correlated.

No significant difference between the strength of correlation obtained using local or global %GP calculation method has been observed for PP patients ($p = 0.64$), while a statistically significant difference has been observed for HN patient ($p = 0.0001$). In particular, for HN patients, %GP calculated using the local method and the different acceptance criteria resulted in a stronger correlation with %DE.

The number of correlations using 3%/3 mm was higher (75% of cases) when compared to 2%/2 mm (63% of the cases) and 1%/1 mm (25% of the cases). Nevertheless, the strength of correlation did not improve significantly (*t*-test statistical analysis comparing mean *r*-values) passing from acceptance criterion 3%/3 mm to 2%/2 mm, while it decreased significantly (*t*-test statistical analysis comparing mean *r* values) when acceptance criteria 1%/1 mm was used.

Table VI gives the range and the mean absolute errors (%) observed in DVH dose metrics for plans exceeding ≥95% passing rate. No patients have %GP higher than 95% by using the acceptance criterion 1%/1 mm, and, in general, %GP are too low to establish different acceptance thresholds, therefore, we did not perform sensitivity calculation for this criterion.

Further analyses on the dependence of correlation on the analyzed structures were performed using the 3%/3 mm acceptance criterion, since it gave the highest number of correlations and because until recently, this was the criterion used in our clinical routine. The number of correlations changed for different structures: in Figs. 1 and 2 are showed the correlation between %GP calculated using global method and the %DE for PTV; in Figs. 3 and 4 are showed the correlation between %GP and OARs' dose values. In particular, $D_{Mean}$ for parotid glands (Fig. 4) was never correlated with %GP, as well as $V_{70}$ of rectum. However, the $D_{Mean}$ of bladder and $D_2$ of spinal cord were always correlated.

Table VII gives the number of FN and TP and the resultant sensitivity per 3%/3 mm and 2%/2 mm metric, in global and local gamma-index method. The mean absolute DVH dose errors 3% and 5% were used as thresholds to calculate sensitivity.

## IV. DISCUSSION

The aim of this paper was to evaluate the predictive meaning of the GI, in terms of correlation between the %GP obtained during standard pretreatment QA tests and the dose discrepancy between planned DVH and patients' perturbed DVH. This paper sets out to evaluate if the standard action levels used by most clinics (3%/3 mm with a 95% passing rate and global normalization) are justified.

The 3DVH validity was demonstrated in previous papers.[2,26] In our study, we started by verifying the consistency of 3DVH with our clinical IMRT plans. For the selected DVH dose values, we did not find a significant difference between mean doses calculated by TPS and 3DVH (Tables III, IVa, and IVb), demonstrating that 3DVH is working properly, measurements are correctly acquired, and the working QA sequence is correctly set up. This was checked first to verify that

TABLE III. Example of data collection and evaluation of the %DE for HN patient number 14.

| Structure | Volume | DVH parameter | (TPS dose ± 0.01) (Gy) | (3DVH dose ± 0.01) (Gy) | %DE |
|---|---|---|---|---|---|
| PTV boost | 309.9 | $D_{Mean}$ | 60.54 | 59.69 | 0.64 |
| | | $D_{95}$ | 62.41 | 62.01 | 1.40 |
| Spinal cord | 14 | $D_{2\%}$ | 42.93 | 42.83 | 0.23 |
| R-parotid | 21.6 | $D_{Mean}$ | 34.97 | 36.53 | 4.46 |
| L-parotid | 21.9 | $D_{Mean}$ | 37.31 | 36.71 | 1.61 |

TABLE IV(a). PP: Volume of the structures and DVH parameters calculated by the TPS Eclipse and by the software 3DVH for PTVs and the two selected OARs.

| Structure | V ± SD (cm$^3$) | DVH parameter | TPS dose ± SD (Gy) | 3DVH dose ± SD (Gy) | Mean dose difference [range] (Gy) | *p* |
|---|---|---|---|---|---|---|
| PTV1 | 608 ± 188 | $D_{95}$ | 52.28 ± 1.93 | 51.26 ± 2.29 | −1.02 [−3.19/0.63] | 0.08 |
| | | $D_{Mean}$ | 55.56 ± 1.96 | 55.22 ± 2.10 | −0.35 [−1.95/1.03] | 0.54 |
| PTV2 (boost) | 163 ± 35 | $D_{95}$ | 64.30 ± 2.73 | 62.19 ± 3.40 | −2.11 [−4.25/0.56] | 0.23 |
| | | $D_{Mean}$ | 67.37 ± 2.61 | 65.59 ± 3.16 | −1.78 [−3.00/0.61] | 0.12 |
| Rectum | 61 ± 21 | $V_{50}$ | 35.44 ± 8.03 | 33.12 ± 8.14 | −0.44[4.66/0.42] | 0.30 |
| | | $V_{70}$ | 0.71 ± 1.60 | 0.31 ± 1.30 | −1.32 [−3.81/1.09] | 0.32 |
| Bladder | 286 ± 159 | $D_{Mean}$ | 45.39 ± 4.91 | 44.70 ± 5.14 | −0.69 [−1.96/1.04] | 0.62 |

TABLE IV(b). HN: Volume of the structures and DVH parameters calculated by the TPS Eclipse and by the Software 3DVH for PTV and the three selected OARs.

| Structure | V ± SD (cm$^3$) | DVH parameter | TPS dose ± SD (Gy) | 3DVH dose ± SD (Gy) | Mean dose difference [range] (Gy) | *p* |
|---|---|---|---|---|---|---|
| PTV1 | 147 ± 103 | $D_{Mean}$ | 62.61 ± 9.37 | 61.70 ± 9.41 | −0.91 [−2.67/0.75] | 0.79 |
| (Boost) | | $D_{95}$ | 59.97 ± 9.03 | 58.38 ± 9.27 | −1.59 [−4.58/0.18] | 0.64 |
| Spinal cord | 21 ± 8 | $D_{2\%}$ | 39.23 ± 10.58 | 38.81 ± 10.49 | −0.43 [−2.01/0.34] | 0.91 |
| Parotid R | 16 ± 6 | $D_{Mean}$ | 28.56 ± 9.62 | 28.31 ± 9.70 | −0.25 [−4.55/1.56] | 0.94 |
| Parotid L | 16 ± 6 | $D_{Mean}$ | 28.93 ± 8.69 | 28.33 ± 8.48 | −0.60 [−2.84/0.94] | 0.85 |

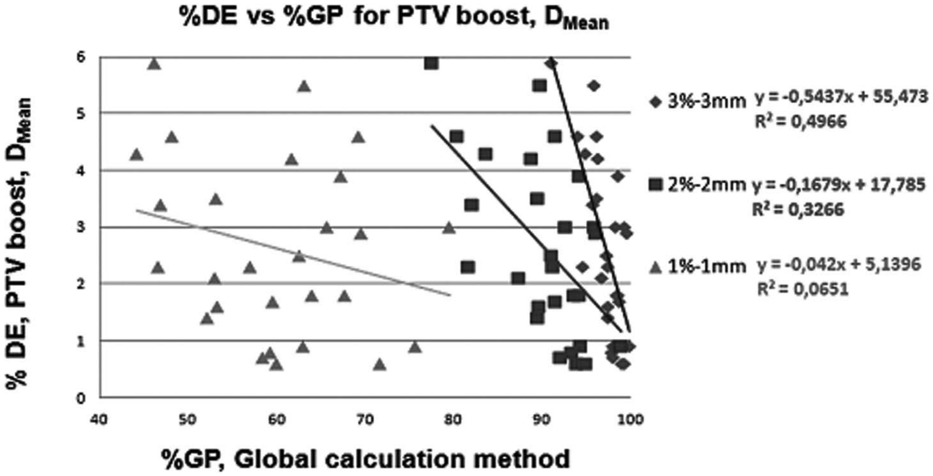

FIG. 1. PP, correlation between %GP calculated using global method and 1%/1mm criterion, 2%/2mm criterion and 3%/3mm criterion and the %DE for DMean of the PTV2 (Boost).

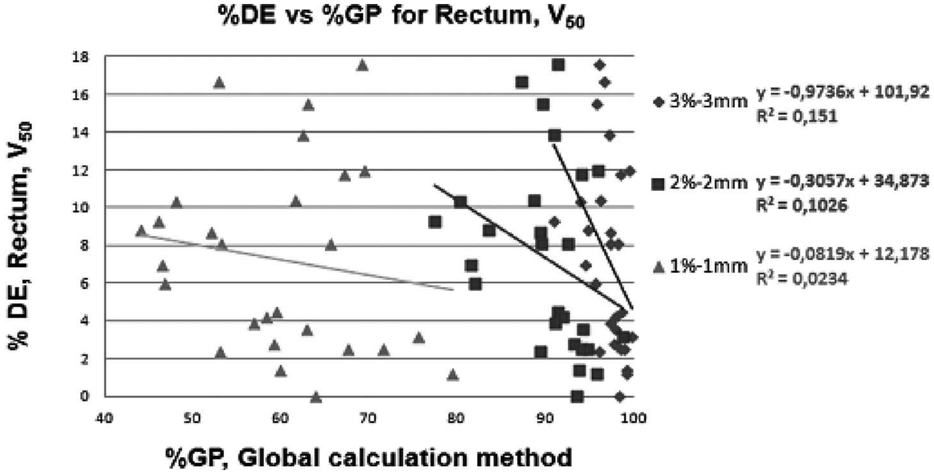

FIG. 2. PP, correlation between %GP calculated using global method and 1%/1mm criterion, 2%/2mm criterion and 3%/3mm criterion and the %DE for D50 of the rectum.

TABLE V(a). PP: indexes of correlation between %GP evaluated with different acceptance criteria and %DE for different structures and DVH parameters.

| Acceptance criteria (mm) | Structure | DVH parameter | Correlation indexes global gamma method | Correlation indexes local gamma method |
|---|---|---|---|---|
| 3%/3 | PTV1 (pelvic lymph nodes) | $D_{95}$ | $r^2 = 0.18\ r = -0.43\ p = 0.03$ | $r^2 = 0.18\ r = -0.42\ p = 0.03$ |
| | | $D_{Mean}$ | $r^2 = 0.36\ r = -0.60\ p < 0.01$ | $r^2 = 0.34\ r = -0.58\ p < 0.01$ |
| | PTV2 (boost) | $D_{95}$ | $r^2 = 0.44\ r = -0.66\ p < 0.01$ | $r^2 = 0.41\ r = -0.64\ p < 0.01$ |
| | | $D_{Mean}$ | $r^2 = 0.50\ r = -0.71\ p < 0.01$ | $r^2 = 0.41\ r = -0.64\ p < 0.01$ |
| | Rectum | $V_{50}$ | $r^2 = 0.15\ r = -0.39\ p = 0.05$ | $r^2 = 0.18\ r = -0.42\ p = 0.03$ |
| | | $V_{70}$ | $r^2 = 0.02\ r = 0.14\ p = 0.50$ | $r^2 = 0.01\ r = 0.12\ p = 0.54$ |
| | Bladder | $D_{Mean}$ | $r^2 = 0.30\ r = -0.55\ p < 0.01$ | $r^2 = 0.26\ r = -0.51\ p = 0.01$ |
| 2%/2 | PTV1 (pelvic lymph nodes) | $D_{95}$ | $r^2 = 0.14\ r = -0.37\ p = 0.06$ | $r^2 = 0.14\ r = -0.38\ p = 0.05$ |
| | | $D_{Mean}$ | $r^2 = 0.26\ r = -0.51\ p = 0.01$ | $r^2 = 0.23\ r = -0.48\ p = 0.01$ |
| | PTV2 (boost) | $D_{95}$ | $r^2 = 0.34\ r = -0.58\ p < 0.01$ | $r^2 = 0.26\ r = -0.51\ p = 0.01$ |
| | | $D_{Mean}$ | $r^2 = 0.33\ r = -0.58\ p < 0.01$ | $r^2 = 0.20\ r = -0.45\ p = 0.02$ |
| | Rectum | $V_{50}$ | $r^2 = 0.10\ r = -0.32\ p = 0.10$ | $r^2 = 0.06\ r = -0.23\ p = 0.24$ |
| | | $V_{70}$ | $r^2 = 0.03\ r = 0.18\ p = 0.37$ | $r^2 = 0.01\ r = 0.11\ p = 0.58$ |
| | Bladder | $D_{Mean}$ | $r^2 = 0.29\ r = -0.54\ p < 0.01$ | $r^2 = 0.27\ r = -0.52\ p = 0.01$ |
| 1%/1 | PTV1 (pelvic lymph nodes) | $D_{95}$ | $r^2 = 0.16\ r = -0.40\ p = 0.04$ | $r^2 = 0.12\ r = -0.35\ p = 0.08$ |
| | | $D_{Mean}$ | $r^2 = 0.23\ r = -0.48\ p = 0.01$ | $r^2 = 0.20\ r = -0.45\ p = 0.02$ |
| | PTV2 (boost) | $D_{95}$ | $r^2 = 0.13\ r = -0.36\ p = 0.07$ | $r^2 = 0.09\ r = -0.30\ p = 0.12$ |
| | | $D_{Mean}$ | $r^2 = 0.07\ r = -0.26\ p = 0.20$ | $r^2 = 0.04\ r = -0.19\ p = 0.35$ |
| | Rectum | $V_{50}$ | $r^2 = 0.02\ r = -0.15\ p = 0.45$ | $r^2 < 0.01\ r = -0.07\ p = 0.72$ |
| | | $V_{70}$ | $r^2 < 0.01\ r = 0.04\ p = 0.85$ | $r^2 < 0.01\ r = 0.03\ p = 0.87$ |
| | Bladder | $D_{Mean}$ | $r^2 = 0.25\ r = -0.50\ p = 0.01$ | $r^2 = 0.20\ r = -0.45\ p = 0.02$ |

there were no significant systematic errors in our clinical QA practices.

Our findings showed that mean %GP, calculated by using both $\gamma$-local and $\gamma$-global normalization, and the acceptance criteria 3%/3 mm were within the thresholds level proposed by previous study.[14] In contrast, %GP significantly decreased when using 2%/2 mm and 1%/1 mm acceptance criteria. However, the low %GP obtained by setting these stringent acceptance criteria may be misleading, since the low passing rates do not imply large errors in anatomy dose metrics (there are a lot of false positives). The results of correlation analysis showed that all $r$-values were low. This proved the weak correlation between the %GP and the absolute percentage dose differences (%DE).

Comparing the results obtained with the different gamma criteria, the number of correlations obtained using the acceptance criteria of 3%/3 mm was higher (75% of cases) than the number of correlations found by using the acceptance criteria 2%/2 mm (63% of the cases) and 1%/1 mm (25% of the cases). Moreover, the strength of correlation decreased

TABLE V(b). HN: indexes of correlation between %GP evaluated with different acceptance criteria and calculation method and %DE for different structures and DVH parameters.

| Acceptance criteria (mm) | Structure | Dosimetric parameter | Correlation indexes global gamma method | Correlation indexes local gamma method |
|---|---|---|---|---|
| 3%/3 | PTV1 (boost) | $D_{Mean}$ | $r^2 = 0.42\ r = -0.65\ p = 0.02$ | $r^2 = 0.52\ r = -0.72\ p = 0.01$ |
| | | $D_{95}$ | $r^2 = 0.42\ r = -0.65\ p = 0.02$ | $r^2 = 0.62\ r = -0.79\ p < 0.01$ |
| | Spinal cord | $D_{2\%}$ | $r^2 = 0.34\ r = -0.58\ p = 0.02$ | $r^2 = 0.55\ r = -0.74\ p < 0.01$ |
| | R-Parotid | $D_{Mean}$ | $r^2 = 0.06\ r = 0.24\ p = 0.41$ | $r^2 = 0.01\ r = -0.09\ p = 0.76$ |
| | L-Parotid | $D_{Mean}$ | $r^2 < 0.01\ r = -0.03\ p = 0.93$ | $r^2 = 0.10\ r = -0.32\ p = 0.27$ |
| 2%/2 | PTV1 (boost) | $D_{Mean}$ | $r^2 = 0.52\ \mathrm{r} = -0.72\ p = 0.01$ | $r^2 = 0.51\ r = -0.71\ p = 0.01$ |
| | | $D_{95}$ | $r^2 = 0.44\ r = -0.66\ p = 0.02$ | $r^2 = 0.61\ r = -0.78\ p < 0.01$ |
| | Spinal cord | $D_{2\%}$ | $r^2 = 0.37\ r = -0.61\ p = 0.02$ | $r^2 = 0.55\ r = -0.74\ p < 0.01$ |
| | R-Parotid | $D_{Mean}$ | $r^2 < 0.01\ r = 0.04\ p = 0.89$ | $r^2 = 0.09\ r = -0.30\ p = 0.30$ |
| | L-Parotid | $D_{Mean}$ | $r^2 = 0.03\ r = -0.17\ p = 0.57$ | $r^2 = 0.21\ r = -0.46\ p = 0.10$ |
| 1%/1 | PTV1 (boost) | $D_{Mean}$ | $r^2 = 0.18\ r = -0.42\ p = 0.17$ | $r^2 = 0.19\ r = -0.44\ p = 0.15$ |
| | | $D_{95}$ | $r^2 = 0.18\ r = -0.43\ p = 0.17$ | $r^2 = 0.28\ r = -0.53\ p = 0.08$ |
| | Spinal cord | $D_{2\%}$ | $r^2 = 0.25\ r = -0.50\ p = 0.06$ | $r^2 = 0.32\ r = -0.57\ p = 0.03$ |
| | R-Parotid | $D_{Mean}$ | $r^2 < 0.01\ r < 0.01\ p = 0.99$ | $r^2 = 0.04\ r = -0.21\ p = 0.47$ |
| | L-Parotid | $D_{Mean}$ | $r^2 = 0.03\ r = -0.16\ p = 0.58$ | $r^2 = 0.12\ r = -0.35\ p = 0.22$ |

TABLE VI. Range (in brackets) and average absolute errors (%) for clinically relevant metrics in the case of all plans (N) meeting a specific %GP threshold (≥95%) (for 3%/3 mm and 2%/2 mm criteria).

| | | | Observed errors (%) in DVH metrics for plans exceeding ≥95% GP (3%/3 mm and 2%/2 mm criteria) | |
|---|---|---|---|---|
| | Structures | Mean absolute error (%) | 3%/3 mm global $\gamma$(N = 41)/local $\gamma$(N = 26) | 2%/2 mm global $\gamma$(N = 7)/local $\gamma$(N = 1) |
| PP patients | PTV1 | $D_{Mean}$ | (−5.07, 1.18)/(−1.64, 1.18) 1.20/1.12 | (−0.06, 0.88)/no range only 1 pts 0.78/0.55 |
| | | $D_{95}$ | (−2.71, 1.84)/(−7.86, 0.81) 1.59/1.44 | (0.55, 1.10) 0.36/0.06 |
| | PTV2 | $D_{Mean}$ | (−3.45, 0.85)/(−3.01, 0.85) 2.79/2.28 | (−4.64/−1.64) 2.28/0.92 |
| | | $D_{95}$ | (−11.95, 3.61)/(−11.95, 4.23) 3.36/3.16 | (−3.01, −0.92) 3.33/1.64 |
| | Rectum | $D_{Mean}$ | (−8.06, 6.69)/(−2.57, 6.69) 2.97/3.03 | (−2.56, −1.41) 1.82/1.50 |
| | Bladder | $D_{Mean}$ | (−3.83, 2.16)/(3.83, 2.16) 2.09/2.82 | (−0.62, 0.46) 0.43/0.62 |
| HN patients | PTV boost | $D_{Mean}$ | (−4.35, 1.09)/(−2.37, 1.09) 1.43/1.00 | (−0.64, 0.611)/no range only 1 pts 0.44/0.64 |
| | | $D_{95}$ | (−7.10, 0.27)/(−6.18, 0.27) 2.25/1.59 | (−1.40, −0.20) 0.72/1.40 |
| | Spinal cord | $D_{2\%}$ | (−4.40, 0.71)/(−1.53, 0.71) 1.22/0.42 | (−0.90, 0.18) 0.42/0.23 |
| | R-Parotid | $D_{Mean}$ | (−15.38, 4.46)/(−15.38, 4.46) 2.64/1.78 | (−0.28, 4.46) 1.78/4.46 |
| | L-Parotid | $D_{Mean}$ | (−11.033, 4.32)/(−11.03, 4.32) 2.53/1.28 | (−1.81, −0.23) 1.28/1.61 |

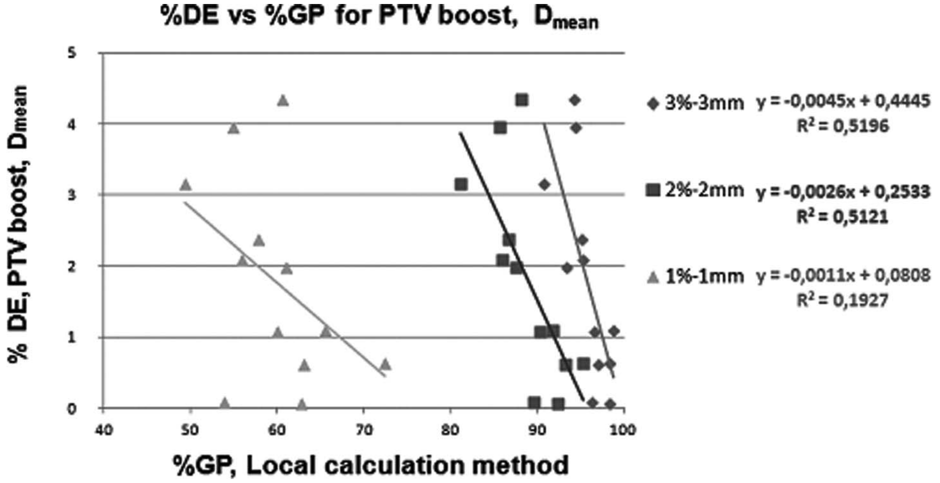

FIG. 3. HN, correlation between %GP calculated using local method and 1%/1mm criterion, 2%/2mm criterion and 3%/3mm criterion and the %DE for DMean of the PTV1 (Boost).

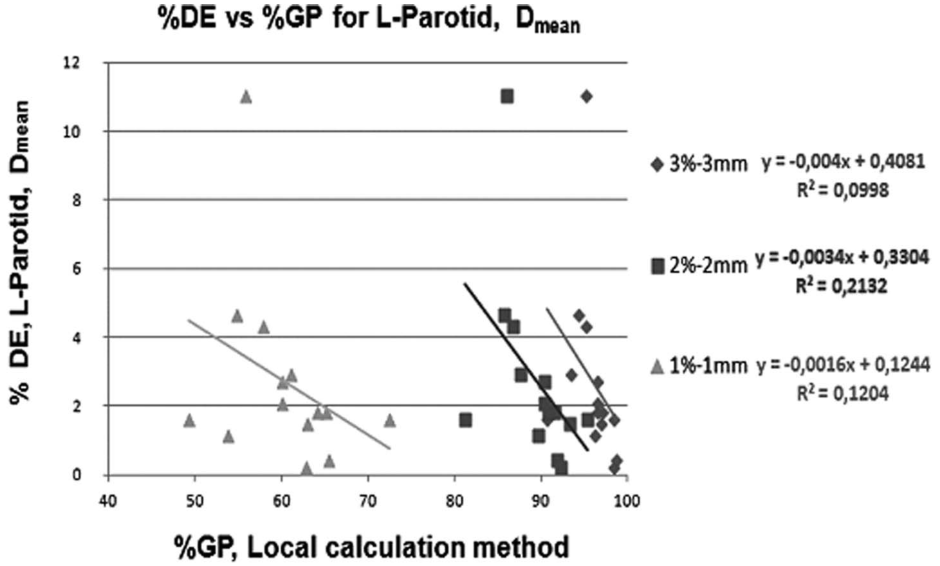

FIG. 4. HN, correlation between %GP calculated using local method and 1%/1mm criterion, 2%/2mm criterion and 3%/3mm criterion and the %DE for DMean of the L-parotid.

TABLE VII. Sensitivity calculation with raw data (number of false negatives and true positives) per metric (3%/3 mm and 2%/2 mm global and local normalization), with 3% and 5% tolerance level.

| | | | # False negatives | #True positives | Sensitivity [C.I.] |
|---|---|---|---|---|---|
| | 3%/3 mm | $\gamma$-local | 17 | 69 | 0.80 [0.76–0.84] |
| Tolerance 3% | | $\gamma$-global | 53 | 33 | 0.39 [0.33–0.44] |
| | 2%/2 mm | $\gamma$-local | 2 | 84 | 0.98 [0.95–0.99] |
| | | $\gamma$-global | 6 | 80 | 0.93 [0.90–0.95] |
| | 3%/3 mm | $\gamma$-local | 5 | 30 | 0.86 [0.81–0.89] |
| Tolerance 5% | | $\gamma$-global | 20 | 15 | 0.43 [0.37–0.49] |
| | 2%/2 mm | $\gamma$-local | 0 | 35 | 1.00 [0.98–1] |
| | | $\gamma$-global | 1 | 34 | 0.97 [0.94–0.99] |

significantly passing from acceptance criterion 3%/3 mm to 1%/1 mm. These findings are in contrast with the results of previous studies,[2,19] probably because different approaches to the same problem have been used. Nelms *et al.*[2] tried to determine the statistical correlation between plan IMRT QA passing rate (%GP) and several clinically relevant, anatomy based dose errors introduced by modifying MLC penumbra and MLC transmission factors. They found that there were only weak to moderate correlations between conventional IMRT QA performance metrics and clinically relevant dose difference metrics, with the moderate correlation/statistically significant cases having a positive slope. This indicated that many of the larger critical errors in patients dose occurred even when QA gamma passing rate was high. Zhen *et al.*,[19] applying a method similar to Nelms' work, had consistent findings for per-beam planar pass rates. Zhen also found that conventional gamma pass rates for three virtual 3D dosimeters with arrays of detectors arranged in common 3D geometries were not consistently correlated to clinically relevant patient DVH-based metrics and that the correlations improved as more stringent criteria were set. Both of these studies were based on inducing beam model errors not subject to the measured uncertainties and inaccuracies presented in a "real" IMRT QA, as performed in our measurements.

In our study, the analysis of the dependence of correlation on the analyzed structures showed that evaluation of %GP through the standard IMRT QA was strongest for the mismatch between calculated and delivered dose in PTVs, while it seemed to be weaker in the detection of possible mismatches between the dose actually delivered to OARs and the calculated one.

The %GP seemed to be more indicative of the actual dose to larger structures' volumes, as shown by the fact that PTV1 (pelvic lymph nodes) and bladder, which are the largest volumes, are correlated in the 83% and in the 100% of the cases, respectively. PTV2 (boost) in PP and the PTV1 in HN, which were the third and fourth larger volumes analyzed in our work, the correlation was found only in 67% of the cases (the percentage was the same for the two volumes). The percentage of correlations decreased to 17% for rectum, while no correlations have been observed for parotid glands. This may be due to the fact that the planar projection of small structures' volumes cover a smaller surface area compared to the one covered by big volumes, resulting in a wider sampled surface, i.e., sampled dose, for larger volumes than the one sampled by small volumes.

The fact that %GP was mostly correlated to %DE of DVH values for PTV structures demonstrates the weakness of the method, since IMRT QA should be applied to ensure the "safety and goodness" of the plan also for OARs.

To evaluate globally the goodness of gamma index, we calculated the sensitivity for different gamma methods.

We found a excellent sensitivity ($>0.93$) for global and local methods with 2%/2 mm and a good sensitivity ($>0.80$) with 3%/3 mm for local normalization.

For these reasons, even if the sensitivity is good ($>0.80$), for the 3%/3 mm local method and 2%/2 mm global and local criteria, this approach is not an adequate clinical test because some high clinical differences in dose could be overlooked, even when most severe criterion are applied. Furthermore, a threshold of %GP $> 95\%$ is very difficult to achieve with a 2%/2 mm criteria. So, in this case even if the sensitivity is excellent, it is not applicable in clinical routine because in our experience only 11% of plans pass this threshold.

In general, the local normalization method was always more sensitive than global one (Table VII), as also confirmed by our results. The gamma passing rate, although it provides the quantity of errors, does not specify the magnitude of errors.

## V. CONCLUSIONS

There are weak correlations between gamma passing rates and DVH values for PTV volumes, and a lack of correlation for OAR volumes. The acceptance criteria for which we had the highest frequency of correlations were 3%/3 mm, however, this criterion hid relevant clinical dose metric differences, which is not clinically acceptable. Further investigations are strongly advised concerning the clinical relevance of GI analysis. The findings in this study in combination with those reported in the literature indicate the necessity to integrate IMRT QA analysis results with a methodology that allows clinicians to predict the impact of delivered dose on with DVHs drawn from 3D dose reconstructions on patient anatomy.

## ACKNOWLEDGMENTS

The authors wish to thank Sun Nuclear Corporation and the referees of Medical Physics for their support and their contribution at this work.

[a)] Author to whom correspondence should be addressed. Electronic mail: sara.bresciani@ircc.it

[1] Prescribing, Recording, and Reporting Intensity-Modulated Photon-Beam Therapy (IMRT), ICRU Report No. 83, Journal of the ICRU, Vol. 10(1), Oxford, 2010.

[2] B. E. Nelms, H. Zhen, and W. A. Tomé, "Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors," Med. Phys. **38**(2), 1037–1044 (2011).

[3] G. A. Ezzell, J. M. Galvin, D. Low, J. R. Palta, I. Rosen, M. B. Sharpe, P. Xia, Y. Xiao, L. Xing, and C. X. Yu, "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**(8), 2089–2115 (2003).

[4] V. Feygelman, K. Forster, D. Opp, and G. Nilsson, "Evaluation of a biplanar diode array dosimeter for quality assurance of step-and-shoot IMRT," J. Appl. Clin. Med. Phys. **10**, 3080–3094 (2009).

[5] S. L. Richardson, W. A. Tome´, N. P. Orton, T. R. McNutt, and B. R. Paliwal, "IMRT delivery verification using a spiral phantom," Med. Phys. **30**(12), 2553–2558 (2003).

[6] D. Letourneau, J. Publicover, J. Kozelka, D. J. Moseley, and D. A. Jaffray, "Novel dosimetric phantom for quality assurance of volumetric modulated arc therapy," Med. Phys. **36**(5), 1813–1821 (2009).

[7] D. A. Low, J. F. Dempsey, R. Venkatesan, S. Mutic, J. Markman, E. Mark Haacke, and J. A. Purdy, "Evaluation of polymer gels and MRI as a 3D dosimeter for intensity-modulated radiation therapy," Med. Phys. **26**(8), 1542–1551 (1999).

[8] K. T. Islam, J. F. Dempsey, M. K. Ranade, M. J. Maryanski, and D. A. Low, "Initial evaluation of commercial optical CT-based 3D gel dosimeter," Med. Phys. **30**(8), 2159–2168 (2003).

[9] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," Med. Phys. **25**(5), 656–672 (1998).

[10] J. Salk, M. Kosta, P. Blank, and E. M. Röttinger, "An extensible framework for IMRT verification," Radiother. Oncol. **68**(Suppl. 1), S104 (2003).

[11] 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).

[12] M. Alber *et al.*, Guidelines for the verification of IMRT, ESTRO (2008).

[13] B. E. Nelms and J. A. Simon, "A survey on planar IMRT QA analysis," J. Appl. Clin. Med. Phys. **8**(3), 76–90 (2007).

[14] S. Both, I. M. Alecu, A. R. Stan, M. Alecu, A. Ciura, J. M. Hansen, and R. Alecu, "A study to establish reasonable action limits for patient specific quality assurance in intensity-modulated radiation therapy," J. Appl. Clin. Med. Phys. **8**(2), 1–8 (2007).

[15] T. Pawlicki, S. Yoo, L. E. Court, S. K. McMillan, R. K. Rice, J. D. Russell, J. M. Pacyniak, M. K. Woo, P. S. Basran, A. L. Boyer, and C. Bonilla, "Process control analysis of IMRT QA: Implications for clinical trials," Phys. Med. Biol. **53**(18), 5193–5205 (2008).

[16] 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**(11), 5359–5373 (2009).

[17] P. S. Basran and M. K. Woo, "An analysis of tolerance levels in IMRT quality assurance procedures," Med. Phys. **35**(6), 2300–2307 (2008).

[18] R. M. Howell, I. P. Smith, and C. S. Jarrio, "Establishing action levels for EPID-based QA for IMRT," J. Appl. Clin. Med. Phys. **9**(3), 16–25 (2008).

[19] H. Zhen, B. E. Nelms, and W. A. Tomé, "Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA," Med. Phys. **38**(10), 5477–5489 (2011).

[20] J. J. Kruse, "On the insensitivity of single field planar dosimetry to IMRT inaccuracies," Med. Phys. **37**(6), 2516–2525 (2010).

[21] P. A. Jursinic and B. E. Nelms, "A 2D diode array and analysis software for verification of intensity modulated radiation therapy delivery," Med. Phys. **30**(5), 870–879 (2003).

[22] D. Létourneau, M. Gulam, D. Yan, M. Oldham, and J. W. Wong, "Evaluation of a 2D diode array for IMRT quality assurance," Sci. Direct **70**(2), 199–206 (2004).

[23] TG-119 IMRT Commissioning Test: Instruction for Planning, Measurements, and Analysis, Version 10/21/2009.

[24] T. Liu, P. Rajaguru, G. Dieck, J. Edwards, M. Brewer, P. Mobit, and C. Yang, "SU-E-T-70: Comparison of two 3D gamma index calculation schemes," Med. Phys. **38**(6), 3501 (2011).

[25] http://www.sunnuclear.com/documents/1212W05062010.pdf.

[26] A. J. Olch, "Evaluation of the accuracy of 3DVH software estimates of dose to virtual ion chamber and film in composite IMRT QA," Med. Phys. **39**(1), 81–86 (2012).

[27] W. Pereira da Silva and C. Pereira da Silva, LAB Fit Curve Fitting Software (Nonlinear Regression and Treatment of Data Program), V 7.2.46 (1999-2009), online, available from world wide web: www.labfit.net.