Evaluating Pre-Treatment IMRT & VMAT QA Techniques Using Receiver Operating Characteristic (ROC) Analysis

by

Allison Lorraine Mitchell

Graduate Program in Medical Physics
Duke University

Date:_______________________
Approved:

__________________________
Justus Adamson, Supervisor

___________________________
Mark Oldham, Chair

___________________________
Robert Reiman

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Graduate Program in Medical Physics in the Graduate School of Duke University

2013
ABSTRACT

Evaluating Pre-Treatment IMRT & VMAT QA Techniques Quantitatively Using Receiver Operating Characteristic (ROC) Analysis

by

Allison Lorraine Mitchell

Graduate Program in Medical Physics
Duke University

Date:_______________________

Approved:

___________________________

Justus Adamson, Supervisor

___________________________

Mark Oldham, Chair

___________________________

Robert Reiman

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Graduate Program in Medical Physics in the Graduate School of Duke University

2013
Abstract

Purpose: Pre-treatment IMRT and VMAT QA techniques are often commissioned without knowledge of their sensitivity to clinically relevant delivery errors. The purpose of this work is to develop a method to quantify the sensitivity and specificity of pre-treatment IMRT and VMAT QA techniques to treatment delivery errors.

Materials and Methods: To evaluate a QA technique, a population of treatment plans and a population of clinically relevant delivery errors are defined. For each delivery error, a threshold magnitude is determined that induces a substantial change in clinically relevant dosimetric indices. Errors at the threshold magnitude are introduced into the plans and QA is performed with and without intentionally introduced errors. The QA technique is treated as a binary classifier to predict error plans using Receiver Operator Characteristic (ROC) analysis. We applied this technique to evaluate portal imager and 2D ion chamber array based QA for VMAT treatment of brain lesions. Delivery errors included discrepancies in MLC positioning (single leaf and leaf bank); lag of MLC trajectory; and discrepancy in dose rate per control point or gantry angle. The threshold magnitude was determined by achieving a 5% change in target conformity index.
**Results:** The area under the curve (AUC) for the ROC analysis was 0.592 and 0.509 for the ion chamber array and portal imager, respectively, using a gamma index of 3%, 3mm. The AUC increased to 0.632 and 0.777 when 2%, 2mm was used for the ion chamber array and portal imager, respectively. Comparison based on 3% dose agreement resulted in an AUC of 0.557 and 0.693, respectively.

**Conclusion:** For both portal imager and ion chamber array based QA, stricter tolerance than 3%, 3mm is needed to detect clinically relevant delivery errors. This method can be used to quantitatively compare the sensitivity of various QA techniques to clinically relevant dosimetric errors.
Contents

Abstract.................................................................................................................................................. iv

List of Tables.......................................................................................................................................... vii

List of Figures......................................................................................................................................... viii

Acknowledgements .................................................................................................................................... x

1. Introduction .......................................................................................................................................... 1

2. Materials and Methods ...................................................................................................................... 5

  2.1. Concept of Evaluating a QA Technique ...................................................................................... 5

  2.2 Application of evaluating a QA technique .................................................................................... 12

3. Results .................................................................................................................................................. 16

4. Discussion ............................................................................................................................................. 26

5. Conclusion ........................................................................................................................................... 30

References ................................................................................................................................................ 31
List of Tables

Table 1: Population of Brain Lesion Treatment Plans ................................................... 12
Table 2: AUC values for EPID using absolute dose comparison ...................................... 20
Table 3: AUC values for EPID using relative dose comparison ....................................... 21
Table 4: AUC values for MatriXX using absolute and global dose comparison .............. 23
Table 5: The QA passing criteria to achieve a sensitivity ≥ 0.95 and corresponding specificity ............................................................... 24
Table 6: The QA passing criteria to achieve a sensitivity ≥ 0.75 and corresponding specificity ........................................................................................................... 25
Table 7: The QA passing criteria to achieve a sensitivity ≥ 0.5 and corresponding specificity ........................................................................................................... 25
List of Figures

Figure 1: A schematic of the four steps that are used to evaluate a QA program ..........5

Figure 2: A diagram of the first two steps of the process in evaluating a QA technique. The populations are defined by circles and the squares are actions that are taken to go from the one population to the other. The items in red must be pre-determined based on specific clinics and situations. ....................................................................................................................7

Figure 3: A schematic of the last two steps of the process in evaluating a QA technique. The QA technique is performed on two sets of plans and a QA score is provided from which an ROC curve can be calculated. ....................................................................................................................8

Figure 4: A plot of an ROC curve displaying the location of the perfect classifier and of the worse classifier, the random prediction line...............................................................10

Figure 5: Conceptual illustration of the typical method used to choose a QA technique’s passing criteria; the QA passing criteria is selected based on an institution’s tolerance for false positives. .................................................................................................................................11

Figure 6: Conceptual illustration of proposed method to choose a QA technique’s passing criteria; the QA passing criteria is chosen based on the institution’s tolerance for both false negatives and false positives. ........................................................................................................11

Figure 7: A plot of the percent change in CI for a single leaf offset versus the error magnitude for various lesion sizes. ........................................................................................................16

Figure 8: A plot of the percent change in CI for a leaf bank offset versus the error magnitude for various lesion sizes ........................................................................................................17

Figure 9: A plot of the percent change in CI for a single leaf delay in MLC trajectory versus the error magnitude for various PTV volumes ........................................................................18

Figure 10: A plot of the percent change in CI for a leaf bank delay in MLC trajectory versus error magnitude for various PTV volumes ........................................................................19

Figure 11: A plot of the ROC curves for the EPID using absolute dose comparison ....20

Figure 12: A plot of ROC curves for EPID using relative dose comparison ..............21
Figure 13: A plot of ROC curves for MatriXX using absolute and global dose comparison. ......................................................................................................................... 22

Figure 14: A plot of ROC curves of the highest AUC for each QA technique tested....... 24
Acknowledgements

Justus Adamson, PhD
1. Introduction

External Beam Radiation Therapy has become increasingly complex with technical advances such as intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT). The complexity of these delivery techniques warrants verification prior to treatment delivery either by measurement or independent calculation. IMRT enables highly conformal non-convex dose distributions. VMAT is a relatively new arc delivery technique which is delivered by means of one or more linac gantry rotations. VMAT delivery combines dynamic multileaf collimators (MLCs) with varying dose rate and gantry speed which allows for high dose conformity and sparing of normal tissue while reducing the number of monitor units (MUs) and the delivery times\(^1\).

A number of methods are routinely used for pre-treatment and patient specific IMRT QA. These methods include film, 2D diode, ion chamber arrays, electronic portal imaging devices (EPIDs), multiple plane 2D detectors, and 3D dosimetry. Film dosimetry permits high spatial resolution; however it has a limited dynamic range and a nonlinear energy dependent dose response and it can also be time consuming, especially for absolute dosimetry\(^2\). There are various 2D detectors commercially available such as the Seven29 and MatriXX (IBA Dosimetry, Germany) 2D ion chamber array or the diode-based device MapCHECK (Sun Nuclear Corporation, US)\(^3\). EPIDs have very high resolution, quick response and have easily available digital analysis tools which
make them useful for analyzing MLC positions at individual control points for VMAT plans. Practical and accurate 3D dosimetry techniques also have potential for pre-treatment IMRT QA and have the benefit of a comprehensive measurement volume.

Although medical physicists continue to use these IMRT QA methods, the sensitivity of a given QA method to clinically relevant delivery errors is often unknown and the choice of comparative measure and tolerance for a given QA technique is not based on the sensitivity to delivery errors.

Even for a given QA measurement device, a variety of comparative measures could be used, and the optimal choice of this comparative measure is often unclear. Gamma analysis is a common comparative measure because it accounts for discrepancies in distance and dose simultaneously. However details such as global/local dose comparison, distance criteria, and choice of threshold vary and optimal choice of these parameters can be unclear. Most institutions perform absolute dose comparisons rather than relative dose comparisons, with the 3% criterion being used most often for the percentage difference analysis, and the 3 mm criterion for the distance-to-agreement (DTA) analysis. A common strategy for analyzing gamma maps from IMRT fields is to calculate the fraction of pixels in each dose plane for which the gamma function returns a value greater than one. A fraction of pixels passing the gamma analysis was found to be a poor predictor of dosimetric accuracy with both planar dosimeters, as well as both sets of gamma criteria. Efficiency in the operation of an IMRT QC program would be
enhanced if patient specific QC were capable of identifying suboptimal MLC performance at a level at which it becomes clinically substantial. A previous theoretical study showed that the most sensitive criteria to systematic MLC offsets of those tested are the 3% AD, 3 mm DTA for MAPCHECK and gamma index with 2% AD and 2 mm DTA for EPID.

The poor correlation between current QA techniques with clinical dosimetric indicators has led to recent efforts to directly translate measured dose discrepancies in the QA geometry to clinical dose discrepancies. One method is to use the dose distribution measured inside a QA phantom with a relatively low pixel density detector array to guide the treatment planning system (TPS) dose on the patient dataset, resulting in a high voxel-density patient dose grid. The critical dose volume values from the DVH curves can then be easily extracted. A method to derive actual delivered fluence maps from measured portal dose images (PDIs) and to use them to reconstruct the 3D patient dose has also been implemented. The reconstruction eases the estimation of the clinical relevance of observed dose difference in the pretreatment measurements. Also using EPID dosimetry combined with 3D dose reconstruction is a useful procedure for patient-specific QA of complex treatments. DVH parameters can be used to interpret the dose distribution delivered to the patient in the same way as during standard treatment plan evaluation. However these techniques have yet to be adopted on a widespread scale and traditional QA techniques remain prevalent.
In this study we present a method to quantitatively assess the ability of pre-treatment IMRT and VMAT QA techniques to detect clinically relevant discrepancies between the intended and delivered treatment plan. One purpose of this method is to enable physicists to understand the effect on sensitivity/specificity of various details of a given QA technique, such as comparative measure (gamma analysis, dose difference, etc.), comparative measure criteria (3%, 3mm, etc., global/local), and action levels (90%, etc.) As an example we apply this technique for two methods for pre-treatment QA of brain VMAT plans.
2. Materials and Methods

2.1. Concept of Evaluating a QA Technique

We propose evaluating a pre-treatment IMRT and VMAT QA technique by treating it as a binary classifier of treatment plans with and without intentionally added clinically relevant delivery errors. There are four steps in evaluating a QA technique (Figure 1). The first step is to define a population of representative plans. Next, a population of clinically relevant errors is defined. Then the population of clinically relevant errors are introduced at a threshold magnitude into the representative plans. The last step involves performing an ROC analysis on the QA technique.

Figure 1: A schematic of the four steps that are used to evaluate a QA program

The population of plans should include representative treatment plans for a specific treatment site; different treatment sites should be analyzed as separate populations. We separate plans by treatment site because the nature of treatment plans, including number of monitor units (MUs), dose rate, beam angles, and number of fields,
will vary per treatment site. Also the effect of the delivery errors is evaluated based on dosimetric criteria used clinically, which is specific to the treatment site.

Determining the population of clinically relevant errors involves first differentiating the types of discrepancies between an intended and actual treatment delivery that can occur and that are detectable with the QA technique. For both IMRT and VMAT, potential delivery errors include discrepancies in MLC positioning, gantry rotation and dose distribution. Discrepancies that cannot be detected by the QA device, such as gantry angle discrepancies when the gantry is overridden for EPID based pre-treatment IMRT QA, are not included in the population of errors, and must be weighed separately when choosing a QA device. After the types of discrepancies that can be detected are differentiated, then those errors are intentionally introduced into a subset of clinical plans and dose is recalculated to determine the relationship between them and relevant dosimetric indices. This in turn, is used to determine the threshold magnitude at which the dose degradation is unacceptable. Once the threshold is determined the clinically relevant errors at their threshold magnitude are introduced into the representative plans. Figure 2 illustrates this process.
Figure 2: A diagram of the first two steps of the process in evaluating a QA technique. The populations are defined by circles and the squares are actions that are taken to go from the one population to the other. The items in red must be predetermined based on specific clinics and situations.

Once the set of treatment plans with and without errors at a clinically relevant threshold are created, the QA technique is performed for both sets of plans; comparing each measurement to the original plan with no added errors. The QA technique provides a QA score which can be used as a binary predictor in an ROC analysis (Figure 3).
Figure 3: A schematic of the last two steps of the process in evaluating a QA technique. The QA technique is performed on two sets of plans and a QA score is provided from which an ROC curve can be calculated.

A receiver operating characteristic (ROC) curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives (TP) out of the positives versus the fraction of false positives (FP) out of the negatives. In our case a TP is a case with an intentionally added error that fails QA and a true negative (TN) is a case without added error that passes QA. A FP is a case without added error that fails QA and a false negative (FN) is a case with an error that passes QA. An ideal classifier technique would be able to consistently differentiate plans with an error from plans.
without an error. The sensitivity (Equation 1) and specificity (Equation 2) can be calculated from the ROC plot, along with the area under the curve (AUC).

\[
Sensitivity = \frac{TP}{TP + FN} \quad \text{(Equation 1)}
\]

\[
Specificity = \frac{TN}{TN + FP} \quad \text{(Equation 2)}
\]

The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The best possible prediction method, a perfect classification, would yield a point in the upper left corner or coordinate (0, 1) of the ROC space, representing 100% sensitivity and 100% specificity, which case the AUC value would be one. A random guess, line of no-discrimination, would give a point along a diagonal line from the left bottom to the top right corners, for which the AUC value would be 0.5 (Figure 4).
Figure 4: A plot of an ROC curve displaying the location of the perfect classifier and of the worse classifier, the random prediction line.

For many institutions, the QA passing criteria is based solely on the agreement from prior clinical plans with no errors added, as shown in Figure 5. The institution’s choice of QA passing criteria is then based on their tolerance of false positives. More information can be gained with the addition of plans with errors. The distribution of plans with errors includes true positives and false negatives (Figure 6). Ideally this distribution should be lower due to the intentionally added errors. For a given QA technique, an ROC analysis is performed by sliding the passing criteria from the bottom to the top and calculating the sensitivity and specificity at each point.
Figure 5: Conceptual illustration of the typical method used to choose a QA technique’s passing criteria; the QA passing criteria is selected based on an institution’s tolerance for false positives.

Figure 6: Conceptual illustration of proposed method to choose a QA technique’s passing criteria; the QA passing criteria is chosen based on the institution’s tolerance for both false negatives and false positives.
2.2 Application of evaluating a QA technique

We applied the ROC analysis for VMAT treatment of brain lesions. The population of plans included 5 cases which are summarized in Table 1. A VMAT treatment plan provides information about gantry position, MLC leaf position and the number of monitor units to be delivered for a particular arc segment through a series of controls points. In VMAT delivery the linear accelerator automatically synchronizes these three different parameters. The linear accelerator interpolates between these control points and dynamically varies the different parameters in order to deliver a desired dose distribution (Bedford and Warrington 2009, Webb and McQuaid 2009).

<table>
<thead>
<tr>
<th>Patients</th>
<th>Dose/Fx (cGy/fx)</th>
<th>Number of Fx</th>
<th>MU</th>
<th>Volume of Lesion (cm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1400</td>
<td>5</td>
<td>3173</td>
<td>3.34</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>5</td>
<td>846</td>
<td>40.24</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>5</td>
<td>981</td>
<td>41.02</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
<td>5</td>
<td>1138</td>
<td>28.31</td>
</tr>
<tr>
<td>5</td>
<td>180</td>
<td>28</td>
<td>356</td>
<td>48.88</td>
</tr>
</tbody>
</table>

The delivery errors added into the treatment plans included positioning errors of the MLCs, delays in MLC trajectory, and an error in the dose modulation. We applied MLC positioning errors for a single leaf and across the whole leaf bank. The single leaf MLC error represents the clinical situation of a failing MLC motor or calibration issue. The whole leaf bank error represents the clinical situation in which an MLC calibration or commissioning error has occurred. The delay in MLC trajectory was varied per
control point which created a lag of MLC position behind the gantry angle by a certain number of degrees. This represents a clinical situation in which a synchronization issue exists between the linear accelerator and the MLC controller. The error in dose modulation consisted of the delivered MU per control point being constant rather than modulated. This represents the situation in which the record and verify system recognizes the plan as a conformal arc, rather than VMAT, and fails to recognize the dose is to be modulated. This has been reported as a potential error for certain clinical hardware/software combinations. Other potential situations may also cause similar errors that are not discussed here.

To determine the magnitude of the clinically relevant errors, each type of errors was intentionally added to a subset of clinical plans with varying magnitude, and the effect of conformity index (CI) was determined. Since the effect a change in MLC position had varied mostly with PTV volume we introduced single leaf and leaf bank positioning errors with varying magnitudes for conformal arcs surrounding a phantom with lesions of various volumes ranging from 0.25 cc to 50 cc. The conformity index (CI) and the maximum dose and minimum dose of the PTV volumes were measured for each lesion size and for each error.

\[
CI = \frac{\text{Prescription Dose Volume}}{\text{PTV Volume}}
\]

(Equation 3)
The threshold magnitude for each error was determined by achieving at least a 5% change in the conformity index. Treatment errors at this threshold magnitude were introduced into the plans and the plans were prepared for QA.

A Novalis linear accelerator (Varian Medical System, Palo Alto, CA), operating at 15 MV and equipped with 120-leaves (2.5 mm for inner 32 leaf pairs, 5 mm for outer 26 pairs, 7 mm for 2 outermost leaf pairs), dynamic MLC, was the radiation source used in the investigation of two QA techniques. The first QA technique performed was an electronic portal imaging device (EPID), which captures an integrated image from the beams eye view (BEV). EPIDs are mounted on the linear accelerator, providing real-time, digital feedback to the user. It has an active imaging area of 40 x 30 cm$^2$ and the image matrix is created from an array of 1024 x 768 pixels. The resolution of the EPID is approximately 0.4 mm$^{14}$. The treatment plan that was to be delivered was converted to an IMRT plan and the expected fluence was calculated using Varian Portal Dosimetry algorithm. The second QA technique used was a 2D ion chamber array (32x32 detectors, 0.7 cm spacing) called MatriXX (IBA Dosimetry, Germany), which measures an integrated dose plane in a phantom. The resolution of the MatriXX is 7.62 mm. For both EPID and MatriXX, the QA was performed on both the treatment plans with and without intentionally added errors. To improve statistics the QA was performed three times for plans with no errors and once for each error plan, since there are 5 plans with intentional errors for each original unaltered plan.
The ROC analysis was performed using both dose difference and gamma index to compare measured and predicted dose for each QA technique. Matlab was used to perform the gamma index for the EPID while OmniPro I’mRT was used for the MatriXX analysis. For the dose difference, we calculated the percent of pixels with agreement within 3% and 5%. The gamma index was calculated using criteria of 3% dose difference (DD), 3 mm (DTA) and 2% (DD), 2 mm (DTA). Both global and local, and absolute and relative dose comparisons were calculated for the EPID. However, for the Matrixx only global and absolute dose comparisons were used since those are the parameters set for the gamma analysis in OmniPro I’mRT.
3. Results

Figure 7 shows the percent change in the CI for an offset in a single leaf for five PTV volumes. It is clear from Fig. 7 that as the magnitude in error is increased the percent change in CI also increases. The effect is also largely dependent on PTV volume, with little change in CI for the two largest PTV volumes. The magnitude of 0.5 cm was selected for the single leaf offset error due to the majority of the change in CI occurring below this value for all PTV volumes.

![Single Leaf Offset](image)

**Figure 7**: A plot of the percent change in CI for a single leaf offset versus the error magnitude for various lesion sizes.

Figure 8 displays the percent change in CI for an offset in the whole leaf bank for four PTV volumes. For most of the PTV volumes the percent change in CI had increased above 5% by the time the error magnitude of 0.1 cm was reached, with the largest
percent change in CI being 15% and the smallest percent change being approximately 3%. Hence a magnitude of 0.1 cm was chosen for the offset error in a leaf bank.

![Leaf Bank Offset](image)

**Figure 8**: A plot of the percent change in CI for a leaf bank offset versus the error magnitude for various lesion sizes

Figure 9 shows the percent change in CI for a delay in a single leaf trajectory for four PTV volumes. A change in CI for each PTV volume began between an error magnitude of 5 and 10 degrees. A magnitude of 8 degrees was used for the error delay in a single leaf.
Figure 9: A plot of the percent change in CI for a single leaf delay in MLC trajectory versus the error magnitude for various PTV volumes.

Figure 10 shows the percent change in CI for a delay in the whole leaf bank trajectory for four PTV volumes. A 10% change in CI occurred for two of the PTV volumes between 5 and 7 degrees while the other two volumes did not have much change in CI as the error magnitude was increased. An error magnitude of 6 degrees for the leaf bank was determined to cause a 5% change in CI.
Figure 10: A plot of the percent change in CI for a leaf bank delay in MLC trajectory versus error magnitude for various PTV volumes.

Figure 11 displays the ROC curves for the EPID using absolute dose comparison. A curve for 3% dose difference, 5% dose difference, 3%, 3 mm, and 2%, 2mm were included. A global and local dose difference was displayed for each gamma criteria. Both the 3%, 3 mm global and local ROC curves lied closely to the random prediction line at 50%. However the 2%, 2 mm global, and 3% dose difference lied further away from the random prediction line. Table 2 displays the area under the curve (AUC) values for each of the criteria. For the EPID using absolute dose comparison the highest AUC was 0.676 for 2%, 2 mm global criteria.
Figure 11: A plot of the ROC curves for the EPID using absolute dose comparison.

Table 2: AUC values for EPID using absolute dose comparison

<table>
<thead>
<tr>
<th>QA Technique</th>
<th>Gamma Index</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPID</td>
<td>3% dose difference</td>
<td>0.664</td>
</tr>
<tr>
<td>EPID</td>
<td>5% dose difference</td>
<td>0.653</td>
</tr>
<tr>
<td>EPID</td>
<td>3% (global), 3 mm</td>
<td>0.509</td>
</tr>
<tr>
<td>EPID</td>
<td>3% (local), 3 mm</td>
<td>0.522</td>
</tr>
<tr>
<td>EPID</td>
<td>2% (global), 2 mm</td>
<td>0.676</td>
</tr>
<tr>
<td>EPID</td>
<td>2% (local), 2 mm</td>
<td>0.578</td>
</tr>
</tbody>
</table>

The ROC curves for the EPID using relative dose comparison are plotted in

Figure 12. A curve for 3% dose difference, 5% dose difference, 3%, 3 mm, and 2%, 2mm
were included. A global and local dose difference was displayed for each gamma criteria. Both the 3%, 3 mm global and local ROC curves lied closely to the random prediction line at 50%. However the 2%, 2 mm global and local gamma criteria, and 3% dose difference lied further away from the random prediction line. Table 3 displays the area under the curve (AUC) values for each of the criteria. For the EPID using relative dose comparison the highest AUC was 0.777 for 2%, 2 mm global criteria.

![ROC Curve-EPID Relative](image)

Figure 12: A plot of ROC curves for EPID using relative dose comparison

Table 3: AUC values for EPID using relative dose comparison

<table>
<thead>
<tr>
<th>QA Technique</th>
<th>Gamma Index</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPID</td>
<td>3% dose difference</td>
<td>0.693</td>
</tr>
<tr>
<td>EPID</td>
<td>5% dose difference</td>
<td>0.637</td>
</tr>
<tr>
<td>EPID</td>
<td>Description</td>
<td>Value</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>3% (global), 3 mm</td>
<td></td>
<td>0.540</td>
</tr>
<tr>
<td>3% (local), 3 mm</td>
<td></td>
<td>0.501</td>
</tr>
<tr>
<td>2% (global), 2 mm</td>
<td></td>
<td>0.777</td>
</tr>
<tr>
<td>2% (local), 2 mm</td>
<td></td>
<td>0.731</td>
</tr>
</tbody>
</table>

Figure 13 displays the ROC curves for the MatriXX using absolute dose comparison. The curves for 3% dose difference, 5% dose difference, 3%, 3 mm, and 2%, 2 mm were included. Only a global dose difference criterion was used. The 5% dose difference curve lied closest to the random prediction line while the 2%, 2 mm gamma criteria lied furthest away from the random prediction line. The area under the curve (AUC) values for each of the criteria is displayed in table 4. For the MatriXX using absolute dose comparison the highest AUC was 0.632 for 2%, 2 mm global criteria.

Figure 13: A plot of ROC curves for MatriXX using absolute and global dose comparison.
Table 4: AUC values for MatriXX using absolute and global dose comparison

<table>
<thead>
<tr>
<th>QA Technique</th>
<th>Gamma Index</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MatriXX</td>
<td>3% dose difference</td>
<td>0.557</td>
</tr>
<tr>
<td>MatriXX</td>
<td>5% dose difference</td>
<td>0.505</td>
</tr>
<tr>
<td>MatriXX</td>
<td>3%, 3 mm</td>
<td>0.592</td>
</tr>
<tr>
<td>MatriXX</td>
<td>2%, 2 mm</td>
<td>0.632</td>
</tr>
</tbody>
</table>

Figure 14 displays the ROC curves for the highest AUC for each QA technique.

The 2%, 2 mm global dose differences were included for all three QA techniques. The 2% (global), 2 mm relative dose comparison for the EPID lied furthest away from the random prediction line. Table 5 displays the area under the curve (AUC) values for each of the QA techniques, the QA passing criteria to achieve a sensitivity ≥ 0.95, and the corresponding specificity. The highest AUC was 0.777 for 2% (global), 2 mm criteria for the EPID using relative dose comparison. Table 6 shows the QA passing criteria to achieve a sensitivity ≥ 0.75 and the corresponding specificity and table 7 shows the QA passing criteria to achieve a sensitivity ≥ 0.5 and the corresponding specificity.
Figure 14: A plot of ROC curves of the highest AUC for each QA technique tested.

Table 5: The QA passing criteria to achieve a sensitivity $\geq 0.95$ and corresponding specificity

<table>
<thead>
<tr>
<th>QA Technique</th>
<th>Gamma Index</th>
<th>AUC</th>
<th>Passing criteria to achieve sensitivity $\geq 0.95$</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPID-Absolute</td>
<td>2% (global), 2mm</td>
<td>0.676</td>
<td>100</td>
<td>0.46</td>
</tr>
<tr>
<td>EPID-Relative</td>
<td>2% (global), 2mm</td>
<td>0.777</td>
<td>100</td>
<td>0.80</td>
</tr>
<tr>
<td>MatriXX</td>
<td>2% (global), 2mm</td>
<td>0.632</td>
<td>97.91</td>
<td>0.20</td>
</tr>
</tbody>
</table>
Table 6: The QA passing criteria to achieve a sensitivity ≥ 0.75 and corresponding specificity

<table>
<thead>
<tr>
<th>QA Technique</th>
<th>Gamma Index</th>
<th>AUC</th>
<th>Passing criteria to achieve sensitivity ≥ 0.75</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPID-Absolute</td>
<td>2% (global), 2mm</td>
<td>0.676</td>
<td>99.98</td>
<td>0.60</td>
</tr>
<tr>
<td>EPID-Relative</td>
<td>2% (global), 2mm</td>
<td>0.777</td>
<td>100</td>
<td>0.80</td>
</tr>
<tr>
<td>MatriXX</td>
<td>2% (global), 2mm</td>
<td>0.632</td>
<td>97.57</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table 7: The QA passing criteria to achieve a sensitivity ≥ 0.5 and corresponding specificity

<table>
<thead>
<tr>
<th>QA Technique</th>
<th>Gamma Index</th>
<th>AUC</th>
<th>Passing criteria to achieve sensitivity ≥ 0.5</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPID-Absolute</td>
<td>2% (global), 2mm</td>
<td>0.676</td>
<td>99.92</td>
<td>0.74</td>
</tr>
<tr>
<td>EPID-Relative</td>
<td>2% (global), 2mm</td>
<td>0.777</td>
<td>99.97</td>
<td>0.94</td>
</tr>
<tr>
<td>MatriXX</td>
<td>2% (global), 2mm</td>
<td>0.632</td>
<td>96.61</td>
<td>0.47</td>
</tr>
</tbody>
</table>
4. Discussion

There is a broad range of possibilities when choosing the clinically relevant errors since data is sparse on what types of errors occur as they are rare events. Many types of errors are extreme, which any QA method should be able to detect; for example if the MLCs were removed from the field. In this research we focused on more subtle errors with the assumption that more drastic errors should also be detectable. The sensitivity, specificity and area under the ROC curve reported here are specific to the population of errors chosen, and are not intended as a measure of a QA technique’s ability to identify any error since inference cannot be made to other types of errors not evaluated. These numbers are more useful as a relative comparison between QA techniques rather than a global estimate of detecting any delivery error. Discrepancies that are not detectable by a QA technique are not considered here, and should be considered separately when assessing a pre-treatment IMRT and VMAT QA technique. For example, for the EPID based QA, the gantry was set to zero for the QA; hence any gantry errors could not be detected.

The choice of QA threshold for clinical effect should be made in conjunction with physicians and could vary greatly by treatment site. In our study the threshold set for a change in conformity index was 5%. There are other possible criteria’s that could be used for the site of brain lesions including the change in maximum dose (Dmax) and minimum dose (Dmin). For other treatment sites including the head and neck cases or
prostate cases there are other options for criteria’s. For example if the treatment site is prostate then the choice of QA threshold may depend on the change in rectum and bladder maximum dose whereas for a head and neck case the QA threshold may be determined from the PTV coverage and the parotid mean dose.

Most institutions use a threshold of 95% and the QA passing criteria is based solely on the agreement from clinical plans with no errors added. The institution will then plan the QA passing criteria depending on how many plans they are willing to accept that didn’t pass. For all QA techniques we evaluated, the threshold value that achieved a sensitivity ≥ 0.95, 0.75, and 0.5 was very high, above 95%. From our results the EPID, using relative dose comparison, had the highest specificity for each calculated sensitivity. If the EPID using relative dose comparison was the QA technique chosen and we accepted 50% of the cases with errors that passed QA, using a gamma index of 2%, 2mm and a threshold of 99.97, then 6% of the cases with no errors will have failed QA.

There is a lack of correlation between conventional IMRT QA performances metrics and dose differences in critical anatomic regions of interest. The most common acceptance criteria and published actions levels therefore have insufficient or at least unproven, predictive power for pre-patient IMRT QA. Moreover, the methodology of basing action levels on prior performance achievements using these conventional methods is unwarranted because meeting these criteria does not ensure that clinically
acceptable dose errors are detected. Most institutions use gamma criteria of 3%, 3 mm and a 90-95% pass rate. Our results showed that a gamma criterion of 2%, 2 mm is more sensitive than a criterion of 3%, 3 mm for both EPID and MatriXX QA techniques. The sensitivity for the 3%, 3 mm criteria was between 0.5 and 0.6 for both QA techniques. In the case of the EPID the sensitivity for the 3%, 3 mm criteria were close to 0.5, a random prediction.

Relative and absolute dose differences were compared for the QA techniques. An absolute dose difference compares the planned dose to a calibration dose while a relative dose difference compares the planned dose to a point in the plan. The relative dose comparison may have performed better due to a slight systematic error in the portal dosimetry calibration. That would lead to a discrepancy for both the original plans and plans with intentionally added errors. For gamma index, global and local dose differences were used in the comparison criteria. A global dose difference compares each point to the max dose while the local dose difference compares each point with the local dose surrounding it. The local dose comparison is often more stringent but according to our results is not necessarily more sensitive than the using the global dose comparison. Often these options are not adjustable in the analysis software as was the case for the analysis of the MatriXX where the gamma index used a global dose comparison.
Comparative measures other than gamma analysis have been developed, but their use has not been widely adopted\textsuperscript{16,17}. The method presented here to evaluate a QA technique could allow for a better understanding of the behavior these measures compared to more well-known measures.
5. Conclusion

For this study we have developed a method that can be used to quantitatively compare the sensitivity of QA techniques to clinically relevant dosimetric errors. The combination of EPID detection and analysis was more sensitive to the clinically relevant errors than the combination of the MatriXX detector and dose difference analysis with respect to the global dose. Our results indicated that a gamma index with 3%, 3 mm criteria is poor indicator of clinically relevant dosimetric errors and a tighter gamma analysis is needed. The most sensitive criteria to the clinically relevant errors are the 2% RD, 2 mm DTA for EPID and gamma index with 2% AD and 2 mm DTA for the MatriXX. Also for a chosen threshold this method can give a good indicator of the probability of failure to detect a subtle but clinically relevant error.
References


