ROC analysis in patient specific quality assurance

Marco Carlone[a)]
Department of Medical Physics, Trillium Health Partners, Mississauga, Ontario L5M 2N1, Canada; Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Ontario M5G 2M9, Canada; and Department of Radiation Oncology, University of Toronto, Toronto, Ontario M5S 3S2, Canada

Charmainne Cruje, Alejandra Rangel, Ryan McCabe, and Michelle Nielsen
Department of Medical Physics, Trillium Health Partners, Mississauga, Ontario L5M 2N1, Canada

Miller MacPherson
Department of Medical Physics, Trillium Health Partners, Mississauga, Ontario L5M 2N1, Canada; Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Ontario M5G 2M9, Canada; and Department of Radiation Oncology, University of Toronto, Toronto, Ontario M5S 3S2, Canada

(Received 26 September 2012; revised 1 March 2013; accepted for publication 1 March 2013; published XX XX XXXX)

Purpose: This work investigates the use of receiver operating characteristic (ROC) methods in patient specific IMRT quality assurance (QA) in order to determine unbiased methods to set threshold criteria for γ-distance to agreement measurements.
Methods: A group of 17 prostate plans was delivered as planned while a second group of 17 prostate plans was modified with the introduction of random multileaf collimator (MLC) position errors that are normally distributed with $\sigma \sim \pm 0.5$, ± 1.0, ± 2.0, and ± 3.0 mm (a total of 68 modified plans were created). All plans were evaluated using five different γ-criteria. ROC methodology was applied by quantifying the fraction of modified plans reported as "fail" and unmodified plans reported as "pass."
Results: γ-based criteria were able to attain nearly 100% sensitivity/specificity in the detection of large random errors ($\sigma > 3$ mm). Sensitivity and specificity decrease rapidly for all γ-criteria as the size of error to be detected decreases below 2 mm. Predictive power is null with all criteria used in the detection of small MLC errors ($\sigma < 0.5$ mm). Optimal threshold values were established by determining which criteria maximized sensitivity and specificity. For 3%/3 mm γ-criteria, optimal threshold values range from 92% to 99%, whereas for 2%/2 mm, the range was from 77% to 94%.
Conclusions: The optimal threshold values that were determined represent a maximized test sensitivity and specificity and are not subject to any user bias. When applied to the datasets that we studied, our results suggest the use of patient specific QA as a safety tool that can effectively prevent large errors (e.g., $\sigma > 3$ mm) as opposed to a tool to improve the quality of IMRT delivery. [http://dx.doi.org/10.1118/1.4795757]

Key words: IMRT, quality assurance, ROC, sensitivity, specificity

I. INTRODUCTION

With widespread use of IMRT and VMAT in radiotherapy, patient specific quality assurance (QA) is now a staple of many medical physics departments. Given the complex nature of IMRT/VMAT beam delivery, many institutions rely on a patient specific measurement to assure that the beam fluence delivered by the linear accelerator conforms to the planned beam fluence. Most accepted methods to quantify the patient specific measurement[1,2] are based on comparisons of absolute dose (AD) and distance to agreement (DTA). The method of Low *et al.*[3] is used often, and this technique is typically referred to as the gamma (γ) analysis. For simple, one-dimensional distributions, it is relatively straightforward to compute the probability of two distributions being different using standard statistical methodology.[4] For complex two-dimensional (2D) distributions such as those measured in typical IMRT/VMAT deliveries, obtaining a measure of the difference between the two distributions in a statistically meaningful way is more complicated. The method of Low[3] computes the dose difference at a point and the distance to the nearest point with equivalent dose for all points in a 2D or higher distribution (between measured and calculated distributions). The scaled dose differences and distances to agreement are added in quadrature; the γ-statistic is then created by measuring the percentage of points with a gamma index less than or equal to a threshold value of 1. A decision threshold value of the percentage of points passing the criteria separates accepted from unaccepted plans.

The practice of IMRT QA analysis is thus influenced by the criteria used as well as the decision threshold value. Extensive work has been conducted to frame the limitations and extent of the contribution of patient specific planar measurements to both the quality and the safety of radiation therapy treatments. For example, studies have attempted to evaluate the effectiveness of γ-based tests in detecting a variety of errors in the delivery of IMRT techniques, from detecting large errors such as missing fields[5] to subtle, but important, errors such as the positioning of the multileaf collimator (MLC) leaves.[6–10] For each of these studies, a small combination of gamma

criteria (e.g., usually 2%/2 mm DTA and or 3%/3 mm DTA) has been used with the purpose of (1) reporting the performance of the test in terms of points passing the criteria[11] and/or (2) selecting an achievable tolerance criteria that could separate acceptable plans from unacceptable ones.[12] Tolerance criteria were historically selected based on experience of achievable passing rates[13] and most recently have been related to desirable clinical or biological endpoints.[9,14–16] Other studies have used statistical methods to evaluate the underlying distribution of expected outcomes based on past experience with the purpose of alleviating the lack of reference or baseline to assess the resultant passing rate.[17–19]

Ultimately, clinical physicists are expected to make accept/reject decisions based on the results of planar dose comparisons. An IMRT fluence pattern that is indistinguishable from the planned fluence pattern should be identified as a positive test result while fluence patterns that are significantly different should be classified as a negative test result. The ability of the test to detect "abnormal" fluence distributions can be evaluated in terms of the test's sensitivity and specificity. It is, however, difficult to quantify sensitivity and specificity of a test using the γ-statistics alone since previous studies have focused on the physical requirements of the fluence measurement device (dose response, detector spacing, etc.). Further, test results are bounded to a specific threshold value (percentage of points passing), which is subject to user bias. The test accuracy is thus an ineffective means of evaluating its performance since it relies on an arbitrary decision threshold.

Signal detection theory offers statistical tools to help quantify test results where a binary outcome is generated.[20] In diagnostic imaging, there is now extensive literature describing the use of the receiver operating characteristic analysis to quantify the value of a diagnostic imaging test. This method has also been used in other areas of medical testing with binary outcomes.[21–24] Measurements of true positive results and false negative results, plotted in the form of a ROC curve, allow the sensitivity and specificity of a test to be quantified in a manner that is independent of threshold bias. The purpose of this work is to investigate the value of ROC methodology as it is applied to patient specific IMRT quality assurance with the objective of removing user bias in determining the technique's fundamental detectability.

II. METHODS AND MATERIALS

II.A. ROC methodology

In medical imaging, ROC analysis has been used to define the ability of diagnostic tests to discriminate between normal and abnormal images. An important feature is that it evaluates diagnostic performance without being affected by varying decision threshold values.[20] Due to the existence of varying case severities, overlaps between normal and abnormal cases occur. Diagnostic tests that perform well display minimum overlap (Fig. 1, center image) while poor performance tests display significant overlap (Fig. 1, left image). For a good performance test, the most optimal threshold can easily be identified as the value that will optimize the true positive fraction (TPF) and the true negative fraction (TNF). For the rest of the tests, a change in the value of the threshold represents a trade-off between the test sensitivity and specificity. Viewed within the context of ROC analysis, planar dose comparisons using gamma based tests exhibit overlapping distributions of plans, some of them fall within the desired standard of quality while others fall outside of it.

To perform ROC analysis, populations of known normal and abnormal cases are placed through the diagnostic test of interest. The fractions of abnormal cases diagnosed to be abnormal (TPF) and normal cases diagnosed to be abnormal ($1 - \text{TNF}$, or false positive fraction, FPF) are calculated for varying thresholds. TPFs are plotted against corresponding FPFs to produce the ROC curve in the ROC space, which consists of values from 0 to 1 in both axes (Fig. 1, right image). To evaluate diagnostic performance, the area under the ROC curve (AUC) is calculated. The closer the AUC is to 1.00, the better its performance. On the contrary, the closer the AUC is

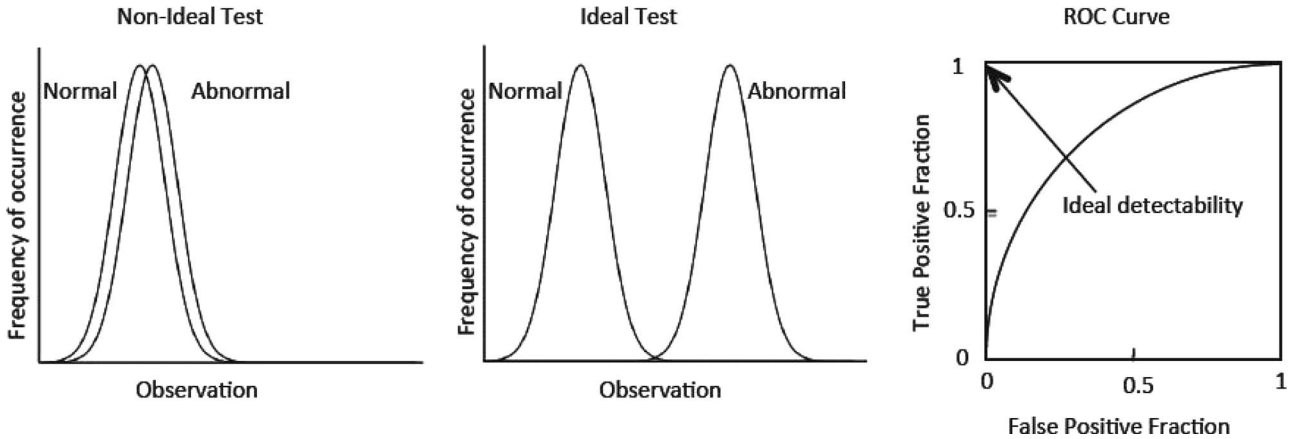

FIG. 1. Illustration of tests whose binary outcome lead to good or poor detectability. Tests where a normal result and an abnormal result share a very similar distribution (left panel) are difficult to discriminate on the basis of measurements below or above a threshold value. Tests whose normal and abnormal distributions have dissimilar distributions, such as in the middle panel, are easier to differentiate using a threshold value. Tests that are more ideal lead to better detectability, where the false positive fraction approaches 0, and the true positive fraction approaches 1 (right panel).

to 0.50, less useful the diagnostic test is. Optimal decision criteria or thresholds may also be determined. The importance of determining optimal criteria or thresholds lies in the tradeoff between test sensitivity and specificity (TNF). Sensitivity and specificity reach a maximum when the selected threshold corresponds to the point on the ROC curve closest to (0, 1).

II.B. Creation of a beam dataset with known fluence errors

Delivery errors in IMRT can occur due to poor MLC performance,[25] beam and MLC modeling errors,[26–28] algorithm limitations in the treatment planning system,[29] the linear accelerators basic ability to match a rapidly varying spatial fluence pattern, or even data transfer errors.[30] In order to determine the sensitivity and specificity of MLC fluence errors in our IMRT patient specific QA, a set of prostate plans, each with a seven field dynamic (sliding window delivery) treatment, were divided into two groups, one for control and the other for test. The unmodified group (UG) served as the control, without any changes to the MLC plan; the modified group (MG) provided the test case, with predetermined MLC errors to simulate a delivery error. We assumed that our linear accelerator was able to deliver the MLC plan, modified or unmodified, with equal bias between groups, i.e., a MLC delivery error was consistent across the groups, regardless of the introduction of the test errors. This was assured by considering compliance to MLC carriage and leaf gap pair constraints. We only simulated MLC delivery error since this was relatively simple to produce on our linear accelerators. Other types of delivery errors were not simulated in this study.

II.C. MLC perturbation to simulate poor delivery of dynamic IMRT beam fluence

Beam fluences for 34 prostate IMRT plans (Varian Eclipse, version 8.5) were divided into two groups, UG and MG. The 17 plans in UG were delivered as planned; the 17 plans in MG were manipulated using a MATLAB program to introduce random leaf errors that are normally distributed with standard deviation (σ) approximately equal to ± 0.5, ± 1.0, ± 2.0, and ± 3.0 mm. The positions of all closed leaves were not altered. In each plan, each field was perturbed independently by a given magnitude of error (e.g., $\sigma = \pm 0.5$ introduced independently to each of the 7 fields in plan X). Finally, 68 modified plans resulted from four unique modifications to each of 17 plans were created.

The new MLC positions were verified in order to comply with mechanical limitations of the Millenium MLC in the Varian iX linear accelerator. The position of a MLC leaf is limited by its opposite's pair position and carriage position; a minimum gap of 0.5 mm is required by a moving leaf pair, while a maximum travel distance of 150 mm from plan-defined carriage position is permitted. Since carriage position limits maximum and minimum leaf positions, the revision of rule compliance was prioritized. First, leaf positions that violated maximum or minimum positions were replaced by closest limits. All leaf pairs were then checked for a 0.5 mm minimum gap. For leaf pairs that did not satisfy the minimum gap requirement after verification of carriage limit issues, the position of a randomly chosen leaf was placed 0.5 mm away. For leaf pairs that did not violate any limits, no adjustment was done. Through these steps, the deliverability of modified leaf positions was ensured.

II.D. Beam fluence measurement

The MapCHECK2 detector array (Sun Nuclear Corporation, Melbourne, FL) was placed on an isocentric mounting fixture (IMF); planar dose measurements were collected using MapCHECK Software Version 3.5. Five different criteria were used, this included γ analysis (absolute mode, VanDyk and ROI criteria enabled) for 1%/1 mm, 2%/2 mm, 3%/3 mm, 4%/4 mm, and 5%/5 mm.

III. RESULTS

III.A. Resultant MLC position errors

Because of the mechanical restrictions of the Varian MLC, the induction of leaf position errors using $\sigma = \pm 0.50$, ± 1.00, ± 2.00, and ± 3.00 mm did not result in exactly these standard deviations, instead we obtained $\overline{|\sigma|} = 0.41 \pm 0.16$, 1.28 ± 0.18, 2.12 ± 0.12, and 3.13 ± 0.15 mm.

III.B. ROC analysis

Patient specific measurements and comparisons were carried through for each of the 68 modified plans and 17 unmodified plans using each of the five criteria mentioned above. Plots of the fraction of fields with a passing rate greater than a user defined threshold (between 0% and 100%) were binned and plotted against pass rate percentage. Figure 2 shows 4 of the 20 plots generated for each combination of five criteria and four $\overline{|\sigma|}$. From here, we generated a ROC curve by varying the pass rate threshold and for each point calculating:

1. The fraction of failed modified plans, which we designate TPF, and
2. The fraction of passed unmodified plans, which we designate 1-FPF.

A total of 20 standard ROC curves (sensitivity or TPF vs 1-specificity or FPF) were then generated; four of these are plotted in Fig. 3. Those gamma criteria that produced curves with AUC closest to 1 were selected and the corresponding calculated AUC values were plotted against $\overline{|\sigma|}$ (Fig. 4). Uncertainties in AUC were determined by the method described in Lasko[31] and Hanley.[32] Ideal thresholds were determined by finding which threshold corresponded to the point closest to (0.00, 1.00) in the ROC space where sensitivity and specificity are both 100%. These were determined for each of the sizes of error introduced in the modified plans, and plotted in Fig. 5.

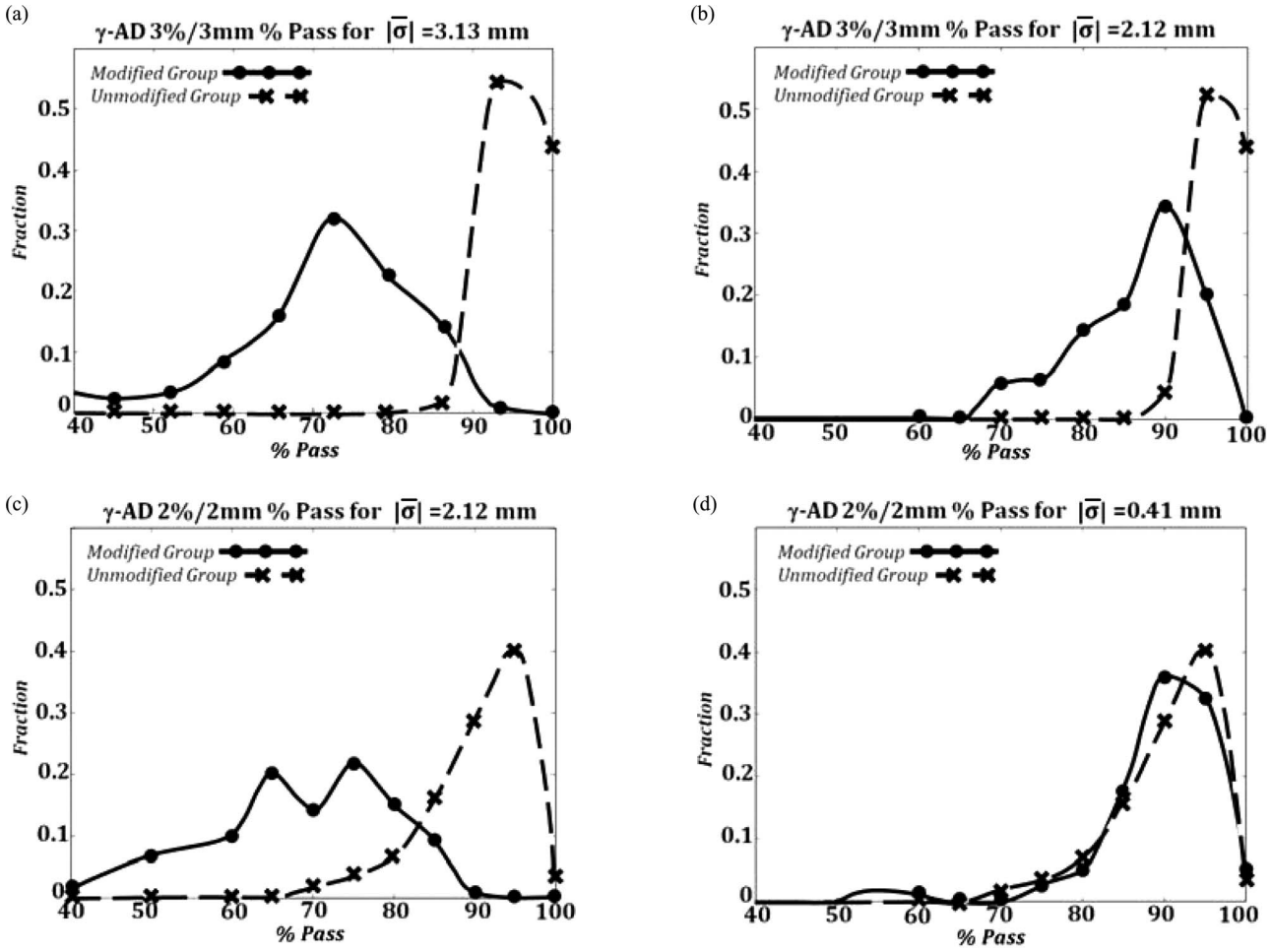

FIG. 2. (a)–(d) Plots of the fraction of fields with a passing rate greater than a user defined threshold (between 0% and 100%). The unmodified MLC group is shown in dashed lines, the group with MLC errors are shown with the solid lines. Separation between the pass rate distribution for the unmodified vs the modified group increases as the size of MLC errors increases and as the γ-AD criterion is decreased.

III.C. Application to independent sets of prostate plans

We applied the results in Fig. 5 to independent data to verify that the suggested threshold values will effectively detect abnormal MLC delivery. The points of Fig. 5 that correspond to the ideal threshold values to detect 1, 2, and 3 mm random MLC errors were tabulated in Table I. We chose the AP field from a 7 field prostate plan for 20 randomly chosen

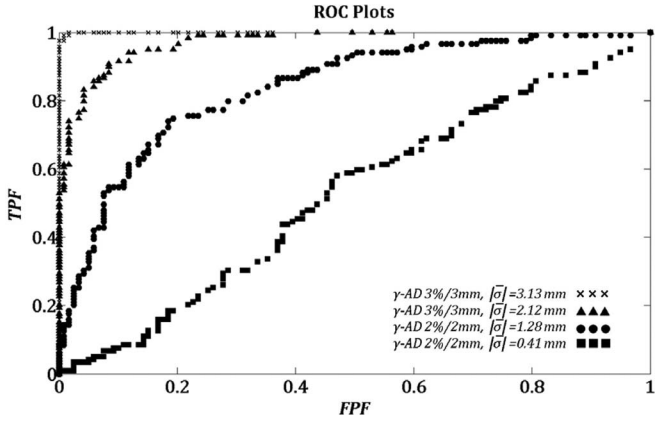

FIG. 3. ROC plots of sensitivity (TPF) vs 1-specificity (FPF) for 4 of 20 curves generated. Curves with highest area have the optimal sensitivity and specificity. Curves along the diagonal, with AUC of 0.5 represent test whose outcome is not significantly different than a random guess.

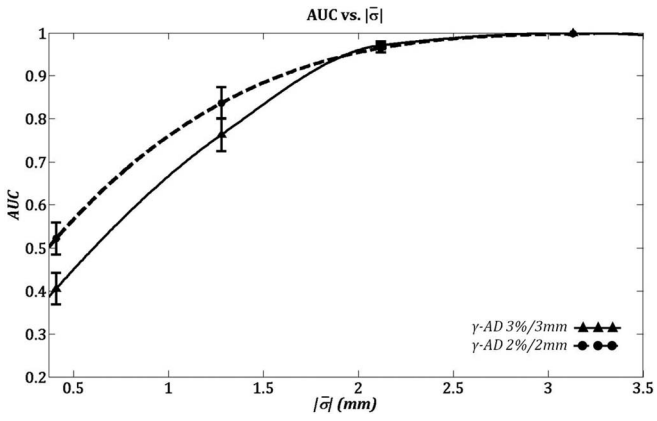

FIG. 4. Measurement of AUC as a function of γ-criterion and size of MLC error. For MLC errors greater than about 2 mm, the detector employed exhibits very good sensitivity and specificity, and hence very good detectability. For smaller MLC errors, sensitivity and specificity decrease to near random results at very small MLC errors (0.5 mm).

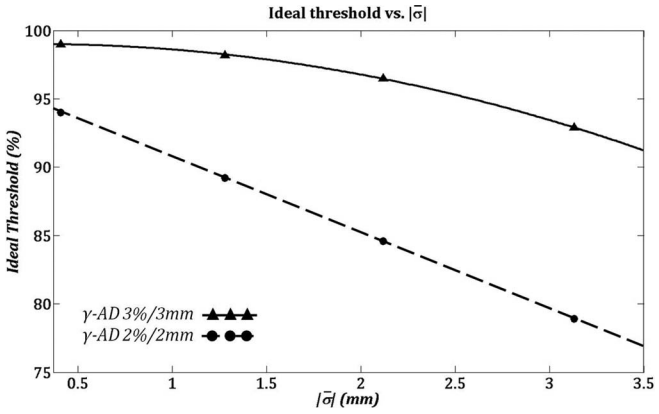

FIG. 5. The ideal threshold value as measured by the point on the AUC curve closest to the point where sensitivity and specificity equal 1.

patients and introduced random errors of 0, 1, 2, 3, 4, and 5 mm for each field. We then measured the beam fluence using the MapCheck2 and applied γ-AD of 2%/2 mm and 3%/3 mm using the threshold points in Table I to detect 3 mm leaf errors. The results are shown in Table II. As expected, our system was able to detect larger errors (4 mm and higher) with 100% accuracy. This accuracy decreased for smaller errors in a manner similar to the trend exhibited by Fig. 4.

IV. DISCUSSION

Previous work in this area focused on two principle areas. Initial work examined the impact of machine delivery errors, or known errors in the planning system on the measured fluence map. For instance, Tatsumi and colleagues[6] determined leaf position tolerances for VMAT by calculating the effect of leaf errors using different treatment planning systems. Wijesooriya and colleagues[7] examined the effect machine performance (gantry speed, leaf speed, etc.) on the accuracy of RapidArc delivery by recomputing a 3D dose distribution of plans delivered with known errors and comparing to the original 3D plan. Rangel and colleagues[8] examined the effect of systematic MLC errors on patient specific QA and found that it was not effective at detecting these types of systematic errors. Basran[12] examined the decision tree in IMRT QA, including results of monitor unit second check calculations and different fluence map detectors. These authors suggest threshold values for head and neck and nonhead and neck plans based on the 95% confidence intervals of observed gamma values. Finally, Palta and colleagues[13] reviewed the precision requirements for IMRT delivery at the subsystem level and stressed that each subcomponent of IMRT delivery must be as precise as possible, and more precise, in general, than for non-IMRT deliveries.

TABLE I. Ideal threshold parameters as determined from Fig. 5.

	Criteria	
$\langle\bar{\sigma}\rangle$ (mm)	2%/2 mm (%)	3%/3 mm (%)
1	89.2	98.2
2	84.6	96.5
3	78.9	92.9

A more recent and different approach is to examine the impact on IMRT delivery on clinically relevant parameters such as a DVH or a radiobiological metric, such as the generalized equivalent uniform dose (gEUD). Zhen and colleagues[14] introduced four different types of IMRT errors and examine the impact on DVH. They reported weak correlation between gamma passing rate and critical patient DVH errors. Rangel *et al.*[15] generated random and systematic leaf errors and examined the impact on EUD, and found a small impact. Finally, Moiseenko *et al.*[16] reported that planar fluence measurements were more sensitive to detect changes in gEUD to organ at risk than ion chamber measurements for plans with small amounts of beam modulation, such as for non-head and neck IMRT.

The current study aims to describe a more fundamental method of identifying nonconformal beam fluences by providing a general method to assess the inherent "detectability" of a detector. In medical imaging, an imager must identify images that are abnormal; similarly, in IMRT QA, the process should be optimized to identify plans where the delivered fluence is identifiably different than the planned fluence. Our study is intended to provide a framework for the user of a detector to determine unbiased γ-DTA thresholds for that detector in a specific application. These threshold values maximize the ability (sensitivity and specificity) of the detector to discriminate between fluence patterns that are known to be correct and known to be incorrect, and thus provide a method to determine baseline parameters for clinical use.

To achieve this, we applied ROC methodology. These methods are designed to maximize the outcome of a binary decision by choosing a decision threshold based on measured and optimized detectability. As in medical imaging, the context of use is important in identifying the decision threshold value. For instance, the system requirements to optimize an imaging system to detect abnormal chest x-ray images are different from that used to detect bone fractures. Similarly, we expect that the operating parameters would be different for an IMRT detector based on the type of IMRT delivery (VMAT vs planar) and the treatment site. In this work, we studied beam fluences for prostate IMRT, however, it is likely that different results would be obtained for other sites such as lung or head and neck. In head and neck IMRT in particular, where beam modulation is high, we would expect different results than those we found here for prostate cancer IMRT. Specifically, the ideal threshold percentages for head and neck cancer IMRT may be lower than those for prostate cancer. The purpose of this investigation was to define a method to determine unbiased γ-DTA threshold criteria for *any* disease site, and thus has value as a commissioning tool. We intend on reporting on our experience with this method as a tool to commission an IMRT program for different disease sites (prostate, lung, head and neck, upper GI) in a future publication. The following observations illustrate the features of a ROC analysis that we believe are important to understand if this method is to be used in the commissioning of an IMRT detector.

TABLE II. Effect of applying the ideal threshold pass rates to an independent set of measurements. Using the AP field from a 7 field prostate plan for 20 randomly chosen patients, we introduced random errors of 1, 2, 3, 4, and 5 mm for each field. The number of field that would be rejected based on the ideal threshold points from Table I were then determined.

MapCHECK criteria	2%/2 mm						3%/3 mm					
MLC average leaf error	0 mm	1 mm	2 mm	3 mm	4 mm	5 mm	0 mm	1 mm	2 mm	3 mm	4 mm	5 mm
Ideal threshold point for 3 mm error detection	78.90%						92.90%					
Average pass rate	87.4	81.8	70.4	56.4	50.2	45.1	98.6	95.9	87.1	72.5	68.4	62.0
Standard deviation	5.6	8.6	9.4	8.8	9.4	10.0	2.1	3.9	7.5	7.8	9.4	11.2
Number of points above threshold point	17	14	2	1	0	0	20	18	4	0	0	0
Number of points below threshold point	3	6	18	19	20	20	0	2	16	20	20	20
Rejection percentage	15	30	90	95	100	100	0	10	80	100	100	100

Highest sensitivity and specificity for a test is demonstrated by the largest areas under the ROC curve, Fig. 4 shows the impact on test sensitivity and specificity (in terms of the AUC) as the magnitude of leaf error is varied using two γ-based criteria (lines in Fig. 4 have been drawn for guidance only). These results indicate that for beam delivery systems where MLC errors $\overline{|\sigma|}$ are greater than about 2 mm, the choice of γ criterion (e.g., 2%/2 mm vs 3%/3 mm) has little effect on test performance, while for $\overline{|\sigma|}$ below 2 mm, the maximal AUC increase is approximately 10%, which indicates the magnitude of test performance improvement one can expect as the gamma criterion is varied from 3%/3 mm to 2%/2 mm. However, using the method of Hanley[33] to calculate the difference in AUCs between 2%/2 mm and 3%/3 mm criterion, we found this difference not to be significant ($p > 0.7$).

An important interpretation of Fig. 4 is that our local patient specific QA program (i.e., γ criteria of 3%/3 mm) is not able to efficiently detect random MLC errors below 0.5 mm since we measured AUC of approximately 0.5 for this magnitude of error. This implies the test behaves more like a random guess of "pass" or "fail." If our center required the detection of random MLC positioning errors in the order of 0.3 mm, Fig. 4 indicates that the devices used in our patient specific QA program cannot meet this requirement. However, from Fig. 4, we also note that test sensitivity and specificity increase rapidly for random MLC positioning errors above 2 mm and reaches near perfect detectability (AUC = 1) for errors above 3 mm. This result indicates that all unsafe deliveries (large errors present) will be detected when adequate patient specific QA is conducted, thus suggesting the use of patient specific QA as a safety tool rather than a tool to ensure high quality treatments.

Figure 5 shows the ideal threshold values for 2 γ-based criteria used in the detection of random MLC errors. The results show that as the stringency of the criteria is increased (e.g., from 3%/3 mm to 2%/2 mm γ-based criterion), the optimal pass rate (to reach maximum sensitivity and specificity) becomes more dependent on the size of error to be detected. For γ criteria of 3%/3 mm, the ideal pass rate in the detection of random errors above 3 mm is approximately 92% (which produces the highest sensitivity and specificity). Detection of smaller errors (e.g., 2 mm) requires a higher pass rate.

V. CONCLUSION

ROC methods can be applied to evaluate patient specific IMRT QA programs. A method has been demonstrated where non-ideal irradiation conditions were simulated by introducing random errors in MLC position during beam delivery. Beam fluences similar to those in prostate IMRT were studied using several criteria. Distributions of true negative and true positive test results were generated. These were compiled as ROC plots which allowed some quantifiable measures to be applied to the patient specific IMRT tests. To the authors knowledge, this is the first demonstrated use of ROC methodology applied to IMRT patient specific QA.

ROC analysis may be useful to understand the extent and limits to detect errors with an IMRT QA program. From the analysis, we conclude that the predictive power of patient specific QA is limited by the size of error to be detected; for the equipment used in our center, we were able to attain nearly 100% sensitivity and specificity in the detection of random MLC errors with a standard deviation >3 mm, which we feel defines a safety component. Sensitivity and specificity decrease rapidly for all gamma and measurement criteria as the size of error to be detected decreases below 2 mm. The predictive power of our patient specific QA program is null (test result is a random guess) regardless of criteria used in the detection of random MLC errors with a standard deviation <0.5 mm.

ACKNOWLEDGMENTS

The authors would like to thank Mike Sharpe and Bill Simon for their critical review of the paper, and for their helpful comments.

[a]Author to whom correspondence should be addressed. Electronic mail: marco.carlone@rmp.uhn.on.ca

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

[2]G. A. Ezzell *et al.*, "IMRT commissioning: Multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119," Med. Phys. **36**(11), 5359–5373 (2009).

[3]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), 656–661 (1998).
[4]G. Cowan, *Statistical Data Analysis* (Oxford University Press, Oxford, NY, 1998).
[5]G. Yan, C. Liu, T. A. Simon, L. C. Peng, C. Fox, and J. G. Li, "On the sensitivity of patient-specific IMRT QA to MLC positioning errors," J. Appl. Clin. Med. Phys. **10**(1), 120–128 (2009).
[6]D. Tatsumi *et al.*, "Direct impact analysis of multi-leaf collimator leaf position errors on dose distributions in volumetric modulated arc therapy: A pass rate calculation between measured planar doses with and without the position errors," Phys. Med. Biol. **56**(20), N237–N246 (2011).
[7]K. Wijesooriya, E. Aliotta, S. Benedict, P. Read, T. Rich, and J. Larner, "RapidArc patient specific mechanical delivery accuracy under extreme mechanical limits using linac log files," Med. Phys. **39**(4), 1846–1853 (2012).
[8]A. Rangel, G. Palte, and P. Dunscombe, "The sensitivity of patient specific IMRT QC to systematic MLC leaf bank offset errors," Med. Phys. **37**(7), 3862–3867 (2010).
[9]M. Oliver, I. Gagne, K. Bush, S. Zavgorodni, W. Ansbacher, and W. Beckham, "Clinical significance of multi-leaf collimator positional errors for volumetric modulated arc therapy," Radiother. Oncol. **97**(3), 554–560 (2010).
[10]D. Letourneau, M. Gulam, D. Yan, M. Oldham, and J. W. Wong, "Evaluation of a 2D diode array for IMRT quality assurance," Radiother. Oncol. **70**(2), 199–206 (2004).
[11]K. Krishnamurthy, S. S. Sivakumar, C. A. Davis, R. Ravichandran, and K. El Ghamrawy, "Formulation and initial experience on patient specific quality assurance for clinical implementation of dynamic IMRT," Gulf J. Oncol. **5**, 44–48 (2009).
[12]P. S. Basran and M. K. Woo, "An analysis of tolerance levels in IMRT quality assurance procedures," Med. Phys. **35**(6), 2300–2307 (2008).
[13]J. R. Palta, C. Liu, and J. G. Li, "Quality assurance of intensity-modulated radiation therapy," Int. J. Radiat. Oncol., Biol., Phys. **71**(suppl 1), S108–S112 (2008).
[14]H. Zhen, B. E. Nelms, and W. A. Tome, "Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA," Med. Phys. **38**(10), 5477–5489 (2011).
[15]A. Rangel and P. Dunscombe, "Tolerances on MLC leaf position accuracy for IMRT delivery with a dynamic MLC," Med. Phys. **36**(7), 3304–3309 (2009).
[16]V. Moiseenko, V. Lapointe, K. James, L. Yin, M. Liu, and T. Pawlicki, "Biological consequences of MLC calibration errors in IMRT delivery and QA," Med. Phys. **39**(4), 1917–1924 (2012).
[17]T. Pawlicki *et al.*, "Process control analysis of IMRT QA: Implications for clinical trials," Phys. Med. Biol. **53**(18), 5193–5205 (2008).
[18]T. Pawlicki *et al.*, "Moving from IMRT QA measurements toward independent computer calculations using control charts," Radiother. Oncol. **89**(3), 330–337 (2008).
[19]S. L. Breen, D. J. Moseley, B. Zhang, and M. B. Sharpe, "Statistical process control for IMRT dosimetric verification," Med. Phys. **35**(10), 4417–4425 (2008).
[20]P. M. DeLuca, A. Wambersie, and G. F. Whitmore, "Receiver operating characteristic analysis in medical imaging," Journal of the ICRU **8**(1) (2008), Report 79, Oxford University Press.
[21]P. A. Hoggarth, C. R. Innes, J. C. Dalrymple-Alford, J. E. Severinsen, and R. D. Jones, "Comparison of a linear and a non-linear model for using sensory-motor, cognitive, personality, and demographic data to predict driving ability in healthy older adults," Accid. Anal. Prev. **42**(6), 1759–1768 (2010).
[22]D. H. Kim, L. Sriharsha, C. W. Jung, S. Kamel-Reid, J. P. Radich, and J. H. Lipton, "Comprehensive evaluation of time-to-response parameter as a predictor of treatment failure following imatinib therapy in chronic phase chronic myeloid leukemia: Which parameter at which time-point does matter?," Am. J. Hemat. **85**(11), 856–862 (2010).
[23]S. Chopra *et al.*, "Evaluation of diffusion-weighted imaging as a predictive marker for tumor response in patients undergoing chemoradiation for postoperative recurrences of cervical cancer," J. Cancer Res. Ther. **8**(1), 68–73 (2012).
[24]C. Hoggart *et al.*, "A risk model for lung cancer incidence," Cancer Prev. Res. **5**(6), 834–846 (2012).
[25]T. LoSasso, C. S. Chui, and C. C. Ling, "Physical and dosimetric aspects of a multileaf collimation system used in the dynamic mode for implementing intensity modulated radiotherapy," Med. Phys. **25**(10), 1919–1927 (1998).
[26]X. Mei, I. Nygren, and J. E. Villarreal-Barajas, "On the use of the MLC dosimetric leaf gap as a quality control tool for accurate dynamic IMRT delivery," Med. Phys. **38**(4), 2246–2255 (2011).
[27]B. E. Nelms, H. Zhen, and W. A. Tome, "Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors," Med. Phys. **38**(2), 1037–1044 (2011).
[28]M. J. Williams and P. Metcalfe, "Verification of a rounded leaf-end MLC model used in a radiotherapy treatment planning system," Phys. Med. Biol. **51**(4), N65–N78 (2006).
[29]H. Chung, H. Jin, J. Palta, T. S. Suh, and S. Kim, "Dose variations with varying calculation grid size in head and neck IMRT," Phys. Med. Biol. **51**(19), 4841–4856 (2006).
[30]G. A. Ezzell and S. Chungbin, "The overshoot phenomenon in step-and-shoot IMRT delivery," J. Appl. Clin. Med. Phys. **2**(3), 138–148 (2001).
[31]T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," J. Biomed. Inf. **38**(5), 404–415 (2005).
[32]J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology **143**(1), 29–36 (1982).
[33]J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," Radiology **148**(3), 839–843 (1983).