12-2013

# AN EVALUATION OF THE CONSISTENCY OF IMRT PATIENT-SPECIFIC QA TECHNIQUES

Elizabeth M. McKenzie

# AN EVALUATION OF THE CONSISTENCY OF IMRT PATIENT SPECIFIC QA TECHNIQUES

A

THESIS

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
and
The University of Texas
MD Anderson Cancer Center
Graduate School of Biomedical Sciences
In Partial Fulfillment
of the Requirements
for the

Degree of

MASTER OF SCIENCE

By

**Elizabeth MaryAnn McKenzie**

Houston, TX

December 2013

**AN EVALUATION OF THE CONSISTENCY OF IMRT PATIENT SPECIFIC QA TECHNIQUES**

by

*Elizabeth M. McKenzie*

APPROVED:

_____
Stephen F Kry, Ph.D.
Supervisory Professor

_____
David Followill, Ph.D.

_____
Peter Balter, Ph.D.

_____
Francesco Stingo, Ph.D.

_____
Jimmy Jones, M.S.

---

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

# I. Acknowledgements

This body of research would not have been possible without the support of several people. It would be impossible to thank them for everything, but I would like to say a few words. First, I would like to thank the physics assistants, notably Scott Laneave, Nick Murray, Andrea Ohrt, and Luke Whittlesey. Their willingness to take the time to teach, to answer questions, and to offer general help proved to be an invaluable resource. They helped me learn clinical skills that I would never have been able to acquire in the classroom. I appreciate not only the knowledge they shared, but also the friendships I gained. Thank you, I'll miss you guys. Also, I'd like to thank Jared Ohrt for helping me to navigate the many mysterious error messages of Pinnacle. He also generously helped me learn to do scripting in Pinnacle, a skill-set I am sure to use in the future. Thank you.

Next, I would like to thank my committee members. My advisor, Dr Stephen Kry, allowed me the opportunity to adopt this project, giving me the chance to have thesis work that gave me hands-on experiences in the clinic. He met with me often, and was always willing to discuss new ideas. Every time we met, he would bring up an interesting aspect of the project I hadn't yet considered, helping to deepen the scope of this project. Thank you for your encouragement, guidance, and support.

This project involved a considerable amount of statistics in the data analysis, which I would not have been able to perform without the help of Dr Francesco Stingo. He graciously met with me on several occasions to not only discuss which methods I should employ, but also offered insight into why one would use them and how they work. Thank you for your patience and willingness to teach.

The measurements involved in this research required considerable amount of time in the clinic. Thank you Dr Peter Balter for offering your clinical knowledge, and for helping me in not only obtaining the equipment I needed, but also helping me to understand its implementation. You also brought up many insightful questions during my committee meetings that helped me to think more critically about the aspects of this project. I would also like to thank Jimmy Jones for his help in teaching me to use the diode array, and generously allowing me to borrow it. I also appreciate the clinical perspective that he offered during my committee meetings.

Dr David Followill offered not only his invaluable experience to the research aspect of this project, but also to the practical side. He helped me to navigate through several hurdles of this project, and also helped to lighten the serious sides of this work with his great sense of humor. Thank you for keeping things in perspective.

Last but not least, I would like to thank those that offered much needed friendship, love and support throughout my time here at MD Anderson. My parents, Laura and Devoy McKenzie, were always available to talk to over the phone and helped to keep me grounded with their loving support. My fiancé Patrick Boehnke would listen to my complaints and accomplishments, cheering me up when I was blue and celebrating with me when I made progress. He has also provided invaluable support in helping me to learn the R statistical computer package. Thank you Patrick.

Also, I appreciate the friendship and support of my classmates. We may have had different projects, but we all went through the same trials. It's this bond that I believe

helped us get through those trials. Thank you everyone for your support both inside and outside of school. I had a wonderful two years getting to know all of you.

It would be impossible to thank everyone, and there are plenty of unnamed people who have helped me along this journey. Thank you to all of you for helping me along this road.

# AN EVALUATION OF THE CONSISTENCY OF IMRT PATIENT-SPECIFIC QA TECHNIQUES

**Elizabeth MaryAnn McKenzie**

**Supervisory Professor: Stephen F. Kry, Ph.D.**

## II. ABSTRACT

To ensure the integrity of an intensity modulated radiation therapy (IMRT) treatment, each plan must be validated through a measurement-based quality assurance (QA) procedure, known as patient specific IMRT QA. Many methods of measurement and analysis have evolved for this QA. There is not a standard among clinical institutions, and many devices and action levels are used. Since the acceptance criteria determines if the dosimetric tools' output passes the patient plan, it is important to see how these parameters influence the performance of the QA device. While analyzing the results of IMRT QA, it is important to understand the variability in the measurements. Due to the different form factors of the many QA methods, this reproducibility can be device dependent.

These questions of patient-specific IMRT QA reproducibility and performance were investigated across five dosimeter systems: a helical diode array, radiographic film, ion chamber, diode array (AP field-by-field, AP composite, and rotational composite), and an in-house designed multiple ion chamber phantom. The reproducibility was gauged for each device by comparing the coefficients of variation (CV) across six patient plans. The

performance of each device was determined by comparing each one's ability to accurately label a plan as acceptable or unacceptable compared to a gold standard.

All methods demonstrated a CV of less than 4%. Film proved to have the highest variability in QA measurement, likely due to the high level of user involvement in the readout and analysis. This is further shown by how the setup contributed more variation than the readout and analysis for all of the methods, except film. When evaluated for ability to correctly label acceptable and unacceptable plans, two distinct performance groups emerged with the helical diode array, AP composite diode array, film, and ion chamber in the better group; and the rotational composite and AP field-by-field diode array in the poorer group. Additionally, optimal threshold cutoffs were determined for each of the dosimetry systems. These findings, combined with practical considerations for factors such as labor and cost, can aid a clinic in its choice of an effective and safe patient-specific IMRT QA implementation.

## III. Table of Contents

## IV. List of Figures

## V. List of Tables

# VI. Chapter 1

# INTRODUCTION AND BACKGROUND

## 1.1 Statement of the Problem

The advent of intensity modulated radiation therapy (IMRT) offers a highly conformal therapy with the ability to increase dose to the treatment site while reducing the toxicities associated with radiation therapy (Veldeman 2008). Its use has seen an explosion of growth, with approximately 30% - 60% of cancer patients receiving IMRT treatment in the United States as of 2008 (Das 2008). While there are dosimetric advantages to IMRT, this more complex technology uses intricate three dimensional dose distributions and a dynamic fluence delivery, requiring more rigorous quality assurance (QA) practices to ensure acceptable dose delivery and machine performance (Low 2011).

The patient-specific validation of radiotherapy plans is an integral part of the clinical implementation of IMRT (Ezzell 2003). Despite the importance of patient-specific QA, no standard has yet emerged, and implementation can vary among institutions (Nelms 2007). Many dosimeter approaches have been formulated to address this need, including but not limited to ion chamber, film, diode arrays, and ion chamber arrays (Low 2011). In conjunction with this hardware, analysis techniques such as percent difference for point measurements and gamma analysis for planar measurements have emerged (Low 2003). These quantitative analyses also do not have a standard acceptance criteria threshold, and can vary by the institution (Nelms 2007). Since studies in radiation therapy outcomes often compare treatments among many different radiation therapy departments, it is important to

establish dosimetric certainty to derive meaningful conclusions from these studies (Das 2008). Therefore, a deeper understanding of the patient-specific QA methods currently in use not only contributes to a safe and effective treatment of the patient, but also to progress in the field of radiation therapy as a whole.

Given the myriad of implementations for patient-specific IMRT QA, this research endeavors to differentiate a selection of modalities, investigating the response of not only the mechanics of devices, but also the quantitative analyses applied to their outputs. Since this project aims to reflect on the dosimetric systems in current clinical use, the following devices have been selected for investigation: a planar diode array (MapCheck, Sun Nuclear, Melbourne, Fl), a helical diode array (ArcCheck, Sun Nuclear, Melbourne, Fl) and a combination of a planar and point measurements using film (Kodak EDR2, Carestream, Rochester, NY) and Wellhofer cc04 ion chamber (CNMC, Nashville, TN). Additionally, an ion chamber based dosimetric system made in-house will be used for further verification of the commercial devices. The above devices' differences or similarities in response to performing patient-specific IMRT QA will be investigated, along with the acceptance criteria used for deeming the IMRT plan acceptable for patient delivery.

## 1.2 Background

### 1.2.1 Intensity Modulated Radiation Therapy (IMRT)

Radiation treatments work under the principle of damaging the DNA of cancerous cells through ionizing radiation (Hall 2006). A problem in this approach becomes immediately salient: how to damage the cancerous cells while minimizing the surrounding normal tissue damage? This issue is more challenging in cases of close-lying critical

structures and organs at risk, such as in cancers of the head and neck. The development of inverse planning was an enormous leap for the field of radiation therapy (Bortfeld 2006). Inverse planning allows the planner to input goals of target coverage and structure avoidance (Hartford 2009). Subsequently, the software optimizes the fluence based on the input parameters, developing a plan where the beam can come from several angles and each beam position can contain modulated beamlets using a collimator such as the multileaf collimator (MLC), therefore sculpting the dose distribution within the patient (IAEA 2008). The International Atomic Energy Agency defines IMRT as "a dose plan and treatment delivery that is optimized using inverse planning techniques for modulated beam delivery," (IAEA 2008). With the power of inverse planning, the dose to the normal tissue can be reduced, allowing for complex plans with a potential escalation in dose to the tumor (Mell 2005).

The increased complexity of IMRT introduces new issues to be monitored. These can include MLC accuracy, monitor unit (MU) delivery, beam modeling, and the accuracy of treatment planning algorithms (Ezzell 2003). While the accuracy of 3-D plans can be confirmed with secondary calculation, IMRT plans are individually tested through direct delivery and measurement, thus performing an end-to-end evaluation of the plan's safety and deliverability.

### 1.2.2 Purpose and Description of Patient-Specific IMRT QA

The end-to-end test of an IMRT plan is known as Patient-Specific IMRT QA. It treats a selected dosimeter/or dosimeter-phantom combination as if it was a patient. The IMRT plan that is intended for the patient is copied within the treatment planning software (TPS) to a CT scan of the dosimeter and phantom. The plan may be altered to accommodate

the QA technique being employed. For example, all of the gantry angles may be changed to zero degrees to allow for only normal incidence to a diode array, or the MU may be scaled to appropriately expose a piece of film. To test the plan transfer process, the new hybrid plan created for the dosimeter is transferred from the TPS to the record-and-verify system, just as a patient plan would be. At the linear accelerator (linac), the dosimeter is set up according to the geometric needs of the measurement. The plan is then loaded to the treatment console, and delivered to the dosimeter. Measurements from this treatment are compared to what was expected from the TPS calculations using a set of pre-defined criteria. The quality assurance program has a set of passing and failing criteria that serve to provide a quantitative assessment of the patient plan. If the plan passes this evaluation, it receives additional scrutiny from professionals such as physicians and physicists, before being ultimately delivered to the patient. Patient-specific QA acts as a sentinel against errors which may emerge from any of the myriad steps in radiation therapy delivery. Because it is an end-to-end process, every step of plan delivery is tested. While the primary goal of patient-specific QA is to ensure that the plan can be administered as intended, it has the additional benefit of highlighting any repairs that may be required from the TPS end to the actual radiation delivery. However, it may be difficult to pinpoint the exact cause of the error from a general patient-specific IMRT QA failure. Thus patient-specific QA not only helps to protect the safety and health of the patient, but also the functionality of the clinic.

### 1.2.3 Theory of Measurement

Three dose measurement devices were investigated in this research: ion chamber, diode, and film. An ion chamber functions by collecting ions produced in a cavity surrounded by a tissue-equivalent wall. Because a point measurement is being made with a

finite collecting volume, an ion chamber can be susceptible to volume averaging effects. This can be a particular problem when the ion chamber is placed in high dose gradients or the volume is too large (Low 2011). In this situation the ion chamber active volume collects charge from a region containing a potentially large range of doses, thus the ion chamber reading becomes an average dose, and may not accurately reflect the dose one intended on measuring. An ion chamber provides many attractive features as a dosimeter. It has very little directional dependence, exhibits fantastic stability, has minimal energy dependence, and can be traced to a calibration standard.

Diodes operate on the principle of p-n junctions. The p region is doped to be an electron receptor, while the n region is doped to be an electron donor. When placed together, a depleted region forms between the p-type and n-type region, creating an electric field in an equilibrium state. This depleted region functions as the active region of the diode, with radiation creating electron-hole pairs leading to a radiation-induced current which is measured. Diodes are about 18,000 times more sensitive than an ion chamber because the ionization energy is less in silicon than in air, and can have much smaller active volumes than an ion chamber (Khan 2010). However, diodes can suffer from an energy dependent response due to their higher atomic number. They also have a directional dependence with sensitivity variation up to 3% when irradiated perpendicular to the diode axis (Low 2011).

Radiographic film consists of a clear film base coated in an emulsion containing a small amount of silver bromide. When irradiated and then developed, the silver bromide crystals that were exposed are reduced to metallic silver, leaving a darkened latent image on the film. The non-irradiated areas are left more transparent. Radiographic film has

excellent spatial resolution due to the small grain size in the emulsion coating. However, the silver increases the effective Z of the film, making it more sensitive to lower energy radiation and scatter. Measurements with film are also dependent on the processor conditions and film batch (Low 2011).

### 1.2.4 Theory of Quantitative Analysis

When using a planar measurement in patient-specific IMRT QA, the gamma analysis (Low 1998) is often used. This method combines the concepts of dose difference and distance to agreement to compare a measured plane with a computed expected dose. Dose difference is expressed as a percent difference between two matching points in the measured and calculated distributions. It is particularly useful in regions of low dose gradient. In high dose gradient regions, if the planes are displaced by a small amount, a large percent difference will result which is not representative of how well the planes match. Therefore, in a high dose gradient, the distance from a point on one plane to the closest matching dose in the reference plane is a more appropriate metric. To incorporate these two concepts, a horizontal plane representing the spatial distance from the measured point to the corresponding calculated point ($|r_c - r_m|$) is combined with a vertical plane representing the difference in dose between the measured and calculated points $D_c(r_c) - D_m(r_m)$. A line is drawn in this three dimensional space connecting the measured point and its corresponding calculated point. The value of this line is $\Gamma(r_m, r_c)$, given by the equation

$$\sqrt{\frac{r^2(r_m,r_c)}{\Delta d_M^2} + \frac{\delta^2(r_m,r_c)}{\Delta D_M^2}} \qquad \textbf{Equation 1}$$

Where r is the spatial distance to agreement $|r_c - r_m|$, $\delta$ is the dose difference $D_c(r_c) - D_m(r_m)$, $\Delta d_M$ is the distance to agreement criteria (e.g. 3mm), and $\Delta D_M$ is the dose difference criteria (e.g. 3%). In computing the gamma index value, the calculated point is taken as the point that minimizes $\Gamma$. This minimized $\Gamma$ is expressed as $\gamma(r_m)$. When $\gamma$ is less than or equal to one, then the calculation at that point passes, otherwise it fails. This process is repeated for all the points in the planar measurement, yielding many points labeled either passing or failing. The percent of pixels passing is a summary metric allowing the evaluator to quantify the percentage of the calculated points which passed the gamma analysis.

For point measurements, as with the ion chamber, a simple percent difference between what was measured and what was expected from the treatment planning system is used. A threshold of percent difference is used to discriminate between failing and passing point measurements.

### 1.3 Hypothesis and Specific Aims

In order to ensure the integrity of an intensity modulated radiation therapy (IMRT) patient treatment plan, each plan must be validated through a quality assurance (QA) procedure. Since an IMRT treatment is a complex three-dimensional composition of many beams, hand calculations are inadequate, and measurements must be taken (Wilcox 2008). Many methods of measurement and analysis have evolved for patient specific IMRT QA. There is not a standard among clinical institutions, and many devices (such as ion chambers, diodes, and film) and many action levels are used (Nelms 2007). According to the *AAPM Task Group Report 120* (Low 2011), each detector for IMRT comes with its own advantages and disadvantages. Because the accuracy of the patient's treatment is dependent on these

measurement techniques, it is important to know how these different tools compare in performance. Kruse (Kruse 2010) took measurements of the same plans using an ion chamber array and EPID. From his data, differences in the fraction of pixels passing can be noted across the modalities. However, this paper only investigates two dosimetric tools, while many more exist and are used clinically. Along with every measurement, action levels must be established to evaluate if the plan is passing or failing. Different institutions apply different acceptance thresholds, with 3%/3mm being the most common (Nelms 2007). Since the acceptance criteria determines if the dosimetric tools' output passes the patient plan, it is important to see how these parameters influence the performance of the QA device. While analyzing the results of IMRT QA, it is important to understand the natural variability in the measurements. This variability could result from set-up and the machine's delivery. Due to the different form factors of the many QA methods, this reproducibility could be device dependent.

**Hypothesis**

The sensitivity among five patient specific IMRT QA detectors using clinically relevant action limits of 3%/3mm and 90% of pixels passing for planar, and 3% dose difference for point measurements to detect failing plans will not exceed 0.90 when compared to a multiple ion chamber phantom standard.

**Specific Aims**

The following specific aims will assess this hypothesis:

1. *To determine the reproducibility in the set up and delivery of IMRT plans to the ArcCheck, MapCheck, Multi Ion Chamber Phantom, EDR2 Film, and cc04 Ion Chamber*

2. *To measure and compare the passing rates of the ArcCheck, Mapcheck, ion chamber, and film for a series of IMRT patient plans, using the Multi Ion Chamber Phantom as a gold standard.*

**1.4 Overview of Thesis Structure**

This thesis serves to address the central hypothesis through pursuing the two specific aims given above. Chapters three and four are self contained studies which address specific aims one and two, respectively. Each includes its own introduction, methods, results, discussion, and conclusion sections that address the goals of each specific aim.

Chapter two is an overarching methods section that more rigorously describes the steps taken in the measurement and analysis of this research. This includes a thorough discussion of some of the statistical methods employed in the data analysis. Because receiver operator characteristic (ROC) curve analysis was such a large part of specific aim two, APPENDIX I will separately discuss this aspect.

Chapter five summarizes this thesis work as a whole, discussing and providing conclusions that have been drawn from all of the specific aims. Also, suggestions for future

work are explored. While a selection of relevant figures and data are printed within chapters three and four, a more complete set of data is provided in APPENDIX IV for the reader's reference.

# VII. Chapter 2

# THEORY OF METHODS

## 2.1 Patient Selection

In order to test a suite of IMRT patient-specific QA devices, as well as some of the various methods in which they are employed, a set of IMRT patient plans were selected. In order to fully test the abilities of each IMRT QA system, it was important to incorporate a variety of plans of different sites and challenge levels. To satisfy this first desire, plans were chosen from a variety of treatment sites: genitourinary, head and neck, gynecological, mesothelioma, gastrointestinal, Mantle, Lung, Spine, and Stereotactic treatments. For the second goal, twenty plans that had previously failed film and ion chamber QA were chosen from the database of previous IMRT QA at the author's institution. Additionally five plans were also selected from this same database that had previously passed QA. Since the results of this project aim to be as clinically relevant as possible, it was decided to use only actual plans that were utilized in the clinic. Other studies that have investigated the performance of IMRT QA devices have generated failing plans by deliberately perturbing some part of the plan, such as MLC alignment. In this project, by using only clinical plans that underwent QA the author endeavors to capture only the challenges that are realistically encountered, thus providing a more readily applicable assessment of the capabilities of the devices and methods under discussion.

The 25 plans that were selected were all recalculated in the treatment planning software (TPS) Pinnacle$^3$ Version 9 (Philips Medical Systems, Andover, MA). The ion chamber measurements from the original IMRT QA of the 25 plans were compared to the

mean ion chamber dose that was recalculated in Pinnacle$^3$ Version 9. This ensured that these plans had failed QA by greater than plus or minus 3% according to the most current abilities of dose calculation. Having this additional test helps to select for plans that are challenging and not simply failures due to calculation errors from weaknesses in older versions of a TPS. Furthermore, all plans were calculated in Pinnacle$^3$ Version 9 when they were copied to the various phantoms being investigated.

These plans were required to be delivered on a variety of dosimeters. One of the previously failing plans (a lung stereotactic case) proved to have too small of a field to be adequately measured by all of the detector systems being studied. Therefore, this patient plan was excluded from the analysis, leading to a total of 24 patient plans in this study.

### 2.2 Dosimeters Studied

This work analyzed the performance of several dosimetric systems, necessitating a familiarization with a variety of software and analyses. Chapter 3 and Chapter 4 describe an overview of how these systems were used, but we present here a more thorough discussion of the methods employed.

#### 2.2.1 Diode Array

A diode array dosimetry system containing 1527 diodes with a custom phantom (MapCheck2 with MapPhan, Sun Nuclear Corporation, Melbourne, FL) was the first IMRT QA device to be studied. Measurements were taken in the form of absolute dose. In order to do this, a dose calibration was performed within the accompanying SNC Patient software. Before every measurement session, an output check was performed (according to the procedures of monthly QA) so that the absolute dose calibration would incorporate any daily

fluctuations in output. The output fluctuated by less than 1% across all measurement sessions. Additionally, an array calibration was performed at the beginning of the first measurement session to correct for any non-uniform response from the diodes.

All diode array measurements were taken with the MapCheck inside the MapPhan, providing 5cm of water equivalent buildup and backscatter along the coronal and sagittal sides of the array (Figure 1). The manufacturer discusses in its product documentation that this setup would allow for measurements to be taken from non-normal incidence beams (SunNuclear 2010).


**Figure 1 Irradiation setup for MapCheck placed within MapPhan phantom**

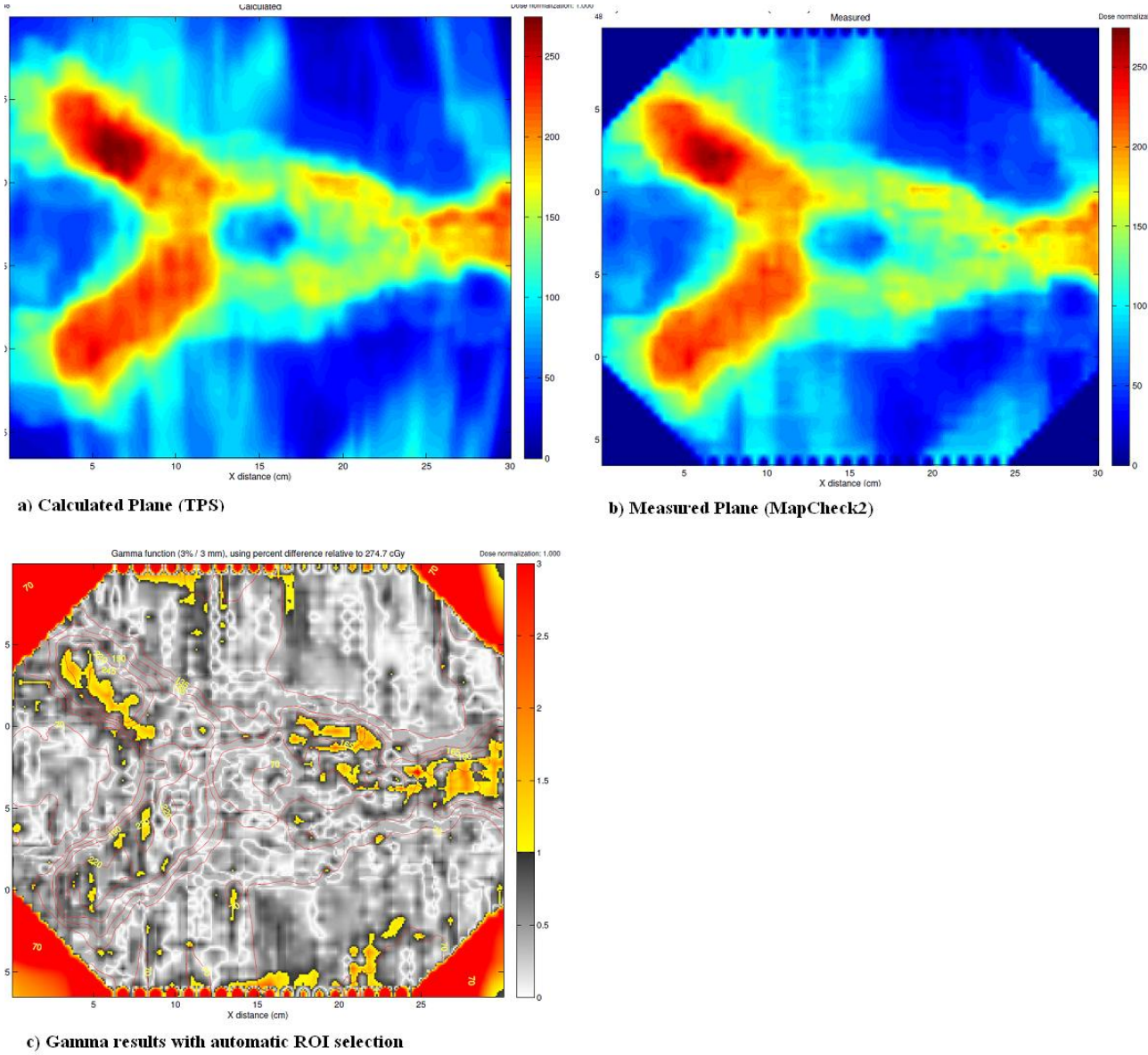
During rotational deliveries, care was taken to avoid irradiating through the couch rails. Therefore for each field, if the beam angle was between 160 – 205 degrees, both rails were moved to their most lateral extent; if the angle was either between 115 – 135 or 225 –

245 degrees, both rails were positioned in their most medial extent (Pulliam 2011). This procedure was followed for all dosimeters which made measurements from plans delivered with original gantry angles.

Two setup geometries were explored for the MapCheck: one in which all beams were delivered perpendicular to the array, and one in which the beams were delivered with their original gantry angles (i.e., rotational delivery). Both types of measurements recorded each field individually. For the rotational deliveries, all array measurements from each beam were summed into a composite planar measurement for comparison with the fluence plane output from the TPS. The plans that were delivered with all beams anterior-posterior (AP) were subdivided into two separate analysis groups: one in which all of the AP fields were summed together into a composite as with the rotational deliveries, and one in which each field was left separate for a field-by-field analysis. Since the final gamma analysis was performed in both SNC Patient and DoselabPro (Mobius Medical Systems, Houston, TX), the composite plans were formed from the raw data in both systems. The output in SNC Patient was a text file similar in format to the original raw measurements. The output from DoselabPro was a tiff file.

Gamma analysis was performed to obtain the percent of pixels passing for each MapCheck measurement for all three methods of measurement (AP, AP composite, and rotational). The gamma analysis was performed in both SNC Patient and Doselab Pro (although in the reproducibility study, only the DoselabPro results were considered). In SNC Patient, a dose threshold of 10% was used with distance to agreement (DTA) and dose difference (DD) criteria set to 2%/2mm, 3%/3mm, and 5%/3mm. The analysis was performed in absolute dose mode, and batch analysis was used to make the process more

efficient. In DoselabPro, no normalization was used. In order to minimize the human dependence on ROI selection, auto-ROI was used. In Doselab, the auto ROI algorithm creates a box bounding the region of the plane containing greater than 30% of the maximum dose. Then a boundary of 10% of the width and height are added to all sides to create the final ROI. Because the geometry of the diode array is not rectangular, in order to guarantee that the rectangular ROI DoselabPro uses would not select regions of the plane lacking diode measurements (Figure 2), a Matlab (MathWorks, Natick, MA) script was written. This script overwrote the reference plan exported from the TPS with zeros in regions where there not any corresponding diodes in the measurement plane. In this way, any discrepancies between the two planes would be due to dosimetric differences, not detector geometry. Auto-registration of the two planes was performed, and then manually fine tuned to provide the most accurate alignment. Additionally, no dose threshold was used for gamma analysis in Doselab, since the software uses relative percent difference instead of local percent difference.




a) Calculated Plane (TPS)

b) Measured Plane (MapCheck2)

c) Gamma results with automatic ROI selection

**Figure 2 When the detector geometry did not match the TPS exported plane geometry, errors in the gamma analysis could result from a large ROI. These regions were therefore excluded with an in-house script.**

Since planes for each beam of the field-by-field analysis had to be exported from Pinnacle, a script was written to facilitate this. Pinnacle does not allow the user to directly export fluence planes of individual beams without normalizing each beam by its associated

number of monitor units (MU). To overcome this limitation, a script was written that zeros the MU for all beams except the one being exported. Now the "composite" exported plane will only contain information from the beam in question, and it will not be normalized by the MU. This was cycled through each beam of each patient plan to generate a reference plane of absolute dose for each field.

### 2.2.2 Radiographic Film and Ion Chamber

The next QA dosimeter system investigated was the EDR2 film (Kodak Carestream, Rochester, NY) and Wellhofer cc04 ion chamber (CNMC, Nashville, TN) combination in the I'mRT body phantom (IBA dosimetry, Schwarzenbruck, Germany) (Figure 3). Since film and ion chamber have different measurement geometries and analysis techniques, each one's performance was considered separately in this work.

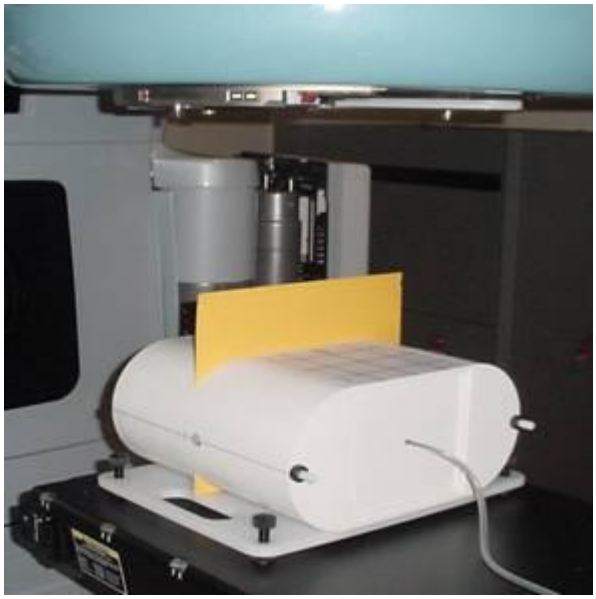


**Figure 3 The irradiation setup for the film and ion chamber dosimeters**

The ion chamber used has a volume of 0.04cc, an inner diameter of 4.0mm, an active length of 3.6mm, and an electrode with a 1mm diameter (CNMC 2009). The film has a responsive range of 24 – 400cGy, with an approximate saturation exposure of 700cGy (Kodak 2001). Its dimensions are 10 x 12 inches. The ion chamber measurements were made with a Model 206 electrometer (CNMC, Nashville, TN). 10 cm x10 cm fields delivered 200MU at 90 and 270 degrees to create a transfer factor converting the ion chamber charge measurements to dose, while incorporating daily fluctuations in output and environment. At the end of a plan delivery, the total accumulated dose was compared to the average dose calculated to the ion chamber ROI in the TPS. Percent difference was calculated by:

$$\frac{Dose_{measured} - Dose_{calculated}}{Dose_{calculated}} \qquad \textbf{Equation 2}$$

EDR2 radiographic film contains less silver than other radiographic film, leading to minimal over-response to low energy radiation (Childress 2005). An eight field calibration was delivered for each batch of film to create a sensitometric curve to account for variations in response from batch to batch. The film was positioned parallel to the beams during plan delivery. Holes were poked into the film while it was still in the phantom after delivery in order to provide marks for geometrical registration, as well as remove potential air gaps. All of the films were developed between one and two days post irradiation. The films were scanned on a VIDAR VXR-16 Dosimetry Pro (VIDAR Systems Corporation, Herndon, VA), at a resolution of 71dpi. The gamma analysis was performed in the Omnipro I'mRT software (IBA Dosimetry, Schwarzenbruck, Germany) at 2%/2mm, 3%/3mm, and 5%/3mm, with a 10% dose threshold. Figure 4 shows a screenshot of what this analysis looked like.

The reference dose was the calculated dose, while the film was the evaluated distribution. The film scan was initially normalized to the maximum dose in the calculated plane, and then manual optimization of the normalization point was performed. The resulting percent of pixels passing the DTA and DD criteria were recorded.

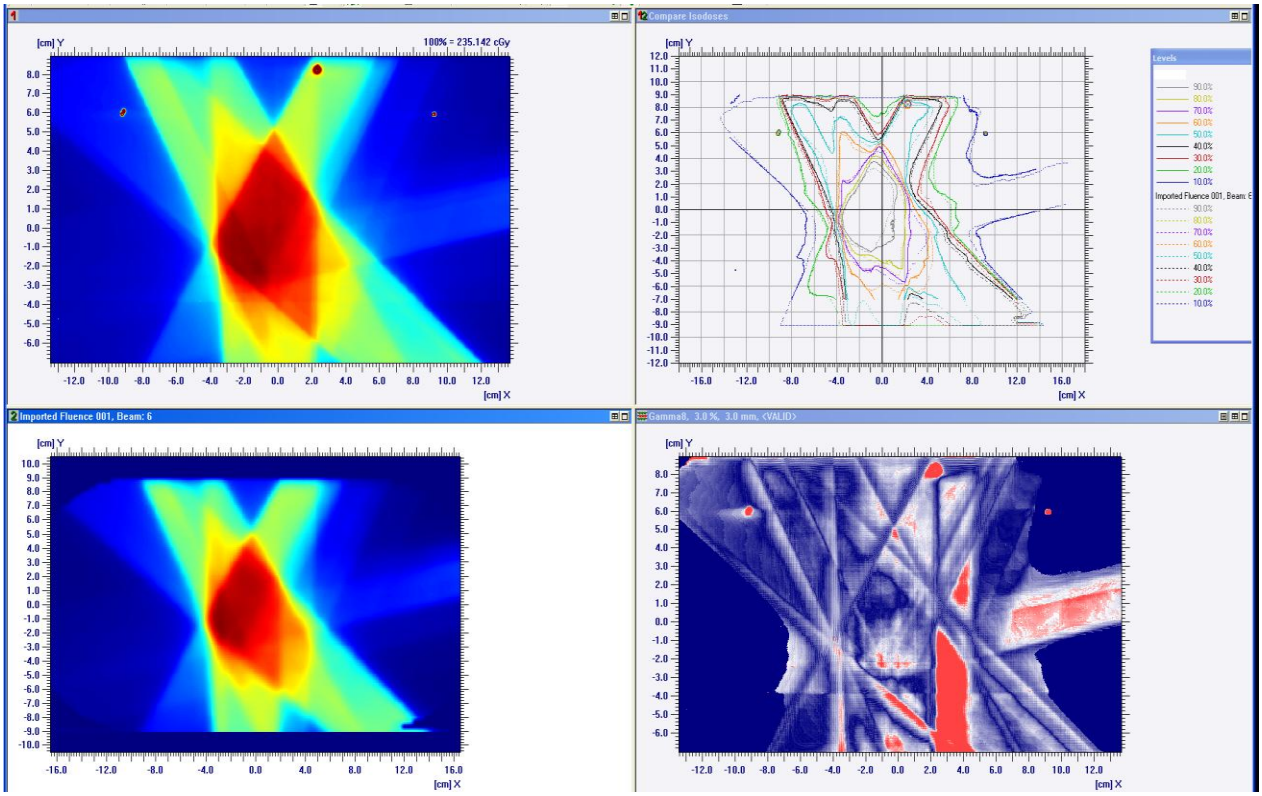


**Figure 4 A screenshot of the gamma analysis for film performed in OmniPro I'mRT software**

### 2.2.3 Helical Diode Array

The cylindrical diode array (ArcCheck, Sun Nuclear Corporation, Melbourne, FL) was developed to have a form factor that is more conducive to measurements performed under rotational delivery (Figure 5). The 1386 diodes are placed in a helical pattern to

reduce the amount of detector overlap, and to obtain a more three dimensional view of the measured dose distribution as compared to a 2D array measurement. Measurements were taken with the accompanying SNC Patient software. As with the MapCheck2 array, dose calibrations were performed to account for a non-uniform diode response and daily fluctuations when converting the measured fluence to dose. Gamma analysis was performed in the SNC Patient software with gamma criteria of 2%/2mm, 3%/3mm, and 5%/3mm with a dose threshold of 10%. The percent of pixels passing were recorded.


**Figure 5 Irradiation setup for the ArcCheck QA dosimetry device.**

### 2.2.4 Multiple Ion Chamber Phantom

A non-commercial dosimeter was also used in this research (Figure 6). The in-house designed multiple ion chamber phantom (MIC) contains 5 five Exradin A1SL 0.057cc ion chambers at 3-dimensionally independent locations. The dose to each ion chamber was calculated with the formula

$$D_w^Q = k_Q * N_{D,w}^{60Co} * M_{raw} * P_{ion} * P_{T,P} * P_{elec} * P_{pol}. \quad \textbf{Equation 3}$$

The $N_{D,W}^{60Co}$ term was obtained from the ADCL calibration on all five ion chambers. $K_Q$, $P_{ion}$, and $P_{pol}$ were estimated from the work by McEwen (McEwen 2010). During the beginning of every session, a 20 cm x20 cm field (so as to not place any of the ion chambers in the penumbra) of 100MU was delivered to the MIC to account for fluctuations in daily output. Temperature and pressure were also measured to correct for environmental changes.

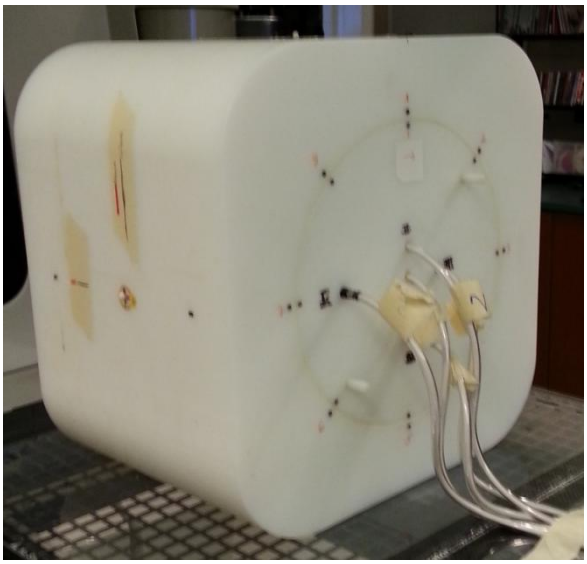


**Figure 6 Irradiation setup for the Multiple Ion Chamber Phantom**

## 2.3 Data Analysis

In order to address specific aim one, an analysis of reproducibility was pursued. To accomplish this, the coefficient of variation (CV) was used as the metric of reproducibility. It is defined as

$$\frac{\textit{sample standard deviation}}{\textit{mean of the sample}} * 100\% \quad \textbf{Equation 4}$$

Because the coefficient of variance is normalized by the mean, this makes it more robust to comparing samples with inherently different values, since samples with a higher average mean will tend to have a higher standard deviation (Rosner 2011). We were interested in seeing if there was a way to place these QA devices into statistically significant groups based on their CV. First an analysis of variance (ANOVA) was performed to see if the underlying mean CV calculated across different patient plans for each QA device was the same. This test compares the variability resulting from between groups and within groups. Once it was determined that at least two of the device group means were not the same from the ANOVA test, a post-hoc Tukey Honestly Significant Difference (HSD) test was done to see which groups differed. The test is performed by comparing the absolute difference in the means of each group to the value of HSD. HSD is defined as

$$\sqrt{\frac{\textit{within group mean square error}}{\textit{number of observations per group}}} * (\textit{studentized range statistic at desired significance level}) \quad \textbf{Equation 5}$$

If the absolute difference in the means of two groups is greater than the HSD value, then the two groups are considered statistically different (Abdi 2010).

In pursuing the research for specific aim one, the opportunity was seen to compare the reproducibility of an ion chamber measurement to the standard deviation in the calculated dose across the ion chamber volume. To see how these two concepts compare, a regression analysis was used to gauge how well the knowledge of the calculated standard

deviation in dose across the ion chamber ROI could predict the reproducibility of the measurement. To account for the uncertainty of the fitted line (i.e. if the experiment was repeated, how the relationship between these two variables might change), confidence intervals were calculated. However, these confidence intervals give information about dose standard deviation predicting the expected, mean value of standard deviation in the measurement. To demonstrate the confidence in predicting the actual values of measurement standard deviation, a wider prediction interval was also calculated to encapsulate the expected values with the error from that observation (Ruppert 2003).

Since this regression analysis took data from five different Exradin A1SL ion chambers making measurements on six different patient plans, the effects of the choice in ion chamber or plan were investigated via looking at the autocorrelation of residuals (for plots of the autocorrelation see the additional figures of APPENDIX IV). Since no effect was found from choice of plan or ion chamber, the regression analysis could be pursued with confidence knowing that these errors were uncorrelated.

For specific aim two, a Receiver Operating Characteristic (ROC) analysis was used. Details of this type of analysis and how it was applied to this body of research is given in APPENDIX I. When the ability of each device to detect acceptable and unacceptable plans was compared, an ANOVA with a post hoc Tukey HSD test was performed, as in specific aim one. This allowed us to see which devices were significantly different in terms of their performance.

# VIII. Chapter 3

# REPRODUCIBILITY IN PATIENT-SPECIFIC IMRT QA

## 8.1 Introduction

Intensity modulated radiation therapy (IMRT) has become ubiquitous in radiation therapy clinics. The increased complexity of IMRT plans necessitates a quality assurance (QA) approach that departs from the traditional hand calculation-based verification. IMRT plans are clinically validated using direct measurement for each patient. To satisfy this need, a number of devices have been developed to measure doses from the IMRT patient plan, which is then compared to the intended dose distribution calculated by the treatment planning system (TPS).

For the sake of convenience, several metrics have been adopted that allow for the sorting of plans as passing or failing, where a passing plan indicates that the delivered dose distribution adequately reflects the intended dose distribution (as calculated by the TPS). Two of these metrics are percent difference and percent of pixels passing the gamma criterion (Low 1998). Percent difference is often used with point measurements, such as with an ion chamber, while the gamma analysis is used for planar measurements such as film or a diode array. The institution chooses a threshold value for these metrics to indicate whether the plan might or might not be suitable to be delivered to a patient. However, the credibility of this sorting rests in part on the reproducibility of the sorting; the reproducibility of the sorting resting in turn on the reproducibility in the delivery of the plan

and the dose measurements. A robust IMRT QA system therefore requires good reproducibility of the measured dose.

Previous studies in the literature have explored the reproducibility of individual measurements. For example, Mancuso et al looked at the reproducibility of ion chamber, film, and 2D diode array measurements in patient specific IMRT QA (Mancuso 2012). However, this work looked at the structure set geometries given in TG119 (Ezzell 2009). Fraser et al. (Fraser 2009) investigated the reproducibility of different ion chambers for IMRT QA. However, no prior study, to the author's knowledge, has explored the reproducibility of IMRT QA results on clinical IMRT plans compared across a wide array of devices. This paper therefore explores the variation in the measured dose for several IMRT QA devices subjected to repeat measurements and analysis.

## 8.2 Methods and Materials

### 3.2.1 Plans

Six different step-and-shoot IMRT patient plans which had previously undergone patient-specific IMRT QA were selected from the authors' institution database. To select for varying degrees of complexity three of the plans were chosen from a pool of plans that previously failed our internal film and ion chamber based QA (Dong 2003), and three of the plans were previously passing. Additionally, different treatment sites were selected: one from thoracic, one from head and neck, one from gynecological, another two from thoracic, and one from gastrointestinal. Henceforth, these six plans will be referred to as THOR1, HN1, GYN1, THOR2, THOR3, and GI1, with the first three coming from the failing pool and the last three coming from the passing pool. All plans were calculated in Pinnacle$^3$

version 9 (Philips Medical Systems, Andover, MA), and all plans were developed with the same TPS beam model with exclusively 6MV photons.

### 3.2.2 Delivery Methods

All plans were delivered on accelerators with matched Varian 21EX beam models (Varian Medical Systems, Palo Alto, CA). For any given dosimeter, all measurements were taken exclusively on the same linear accelerator within a single evening.

Five clinical dosimeters and one in-house designed dosimeter were utilized for this study: EDR2 film ((Kodak Carestream, Rochester, NY)), Wellhofer cc04 ion chamber (CNMC, Nashville, TN), MapCheck2 in the MapPhan phantom (Sun Nuclear Corporation, Melbourne, FL), ArcCheck (Sun Nuclear Corporation, Melbourne, FL), and a multiple ion chamber (MIC) phantom.

For each dosimeter, after an initial setup, each plan was delivered three times. The system was then perturbed, re-setup, and re-irradiated (for a fourth time). The system was perturbed again, re-setup again, and re-irradiated again (for a fifth time). Ultimately, this yielded three irradiations delivered under a single setup ("redelivery"), and three irradiations under an independent setup ("total delivery") for each patient plan on every device, with the third irradiation from the "redelivery" measurements being counted as one of the independent setups for the "total delivery" measurements. The "redelivery" measurements were performed to assess the reproducibility of the machine delivery and device readout, as well as the analysis associated with that QA dosimeter. The "total delivery" measurements incorporate both the variability seen in the "redelivery" measurements, as well as variability

introduced from the setup of the equipment. A flow chart below offers a visual representation of this workflow [Figure 7].

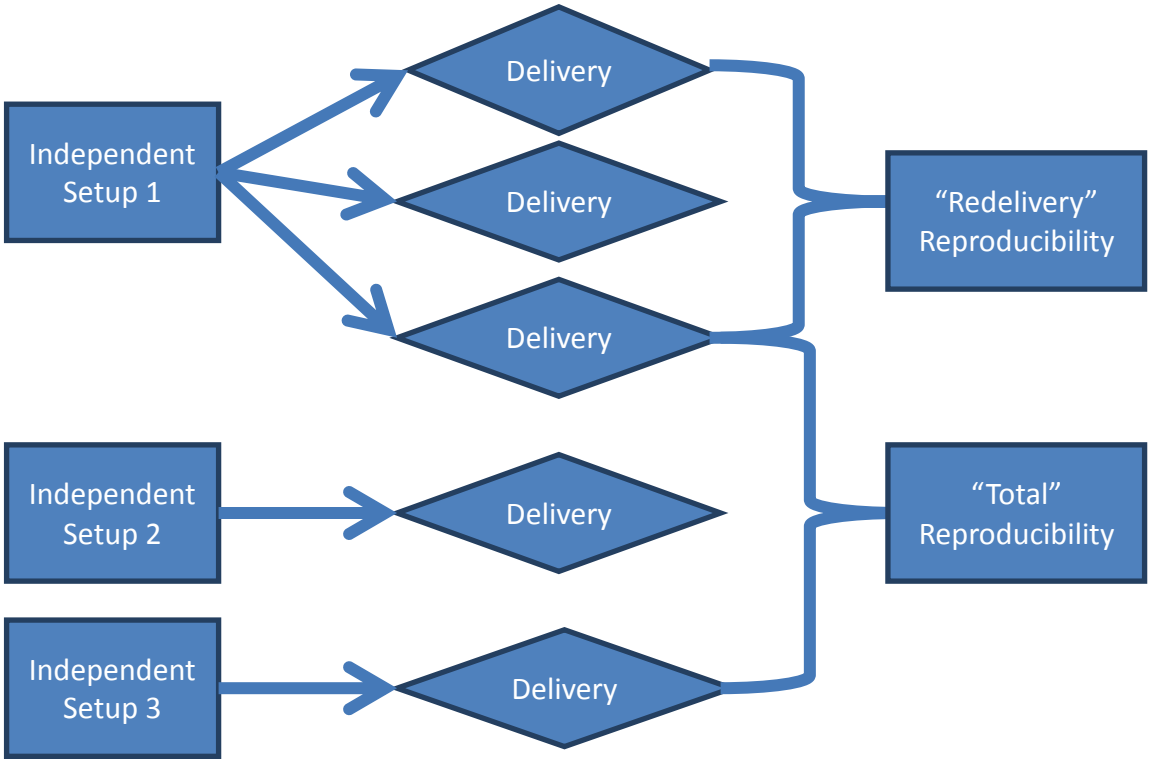



**Figure 7 The workflow used to generate measurements for "total" and "redelivery" reproducibility. This same workflow was performed on all six patient plans on all IMRT QA devices studied**

### 3.2.3 Dosimeters Used

The EDR2 film and cc04 ion chamber were placed in an I'mRT body phantom (IBA Dosimetry, Schwarzenbruck, Germany), with the film placed in the transverse plane. The phantom was shifted to allow the ion chamber to have an average dose of at least 70% of the maximum dose in the plan, and a standard deviation equal to or less than 1% of the mean value dose across the active volume of the ion chamber, in accordance with the practices

performed at the author's institution. This procedure is done with the goal of placing the ion chamber in a high dose, low gradient region. All plans were delivered with their original gantry angles.

Because radiographic film and an ion chamber make fundamentally different measurements with differing geometries, these two dosimeters were considered separately in assessing their reproducibility. For the film, a gamma analysis in OmniPro I'mRT (IBA Dosimetry, Schwarzenbruck, Germany) with a 10% dose threshold and criteria of 3%/3mm was used for the reproducibility analysis.

The MapCheck2 diode array was placed within 5 cm of water equivalent buildup and 5cm for backscatter. The device was then irradiated and analyzed in 3 different configurations. First, it was delivered with gantry angles set to zero degrees for an AP beam delivery. This was further broken down into a (1) field-by-field gamma analysis, and a (2) composite AP analysis. Thirdly, the (3) original gantry rotational gantry angles from the IMRT plan were delivered to the MapCheck in the MapPhan phantom. Per the manufacturer's instructions, if the majority of angles for any plan came from near 90 or 270 degrees the array would have been placed sagittally. However, since this was not the case for any of the plans under investigation (none had a majority of beams coming from either 70-100 degrees or 250-290 degrees), all plans were delivered with the MapCheck flat on the treatment couch. This measurement was followed with a gamma analysis on the rotational composite field. For all three conditions, all gamma analysis was done at 3%/3mm in the DoseLab Pro software (Mobius Medical Systems, Houston, TX), with an automatically selected ROI and using absolute dose. The measured plane was the reference distribution in the gamma analysis.

The ArcCheck was positioned on the couch, with shifts away from or towards the gantry as was necessary to avoid irradiation of the electronics, and to center the treatment plan on the center diodes. Gamma analysis was performed in the SNC Patient software (Sun Nuclear Corporation, Melbourne, FL) at 3%/3mm, with an ROI that encompassed all of the diodes, using absolute dose mode, and using the measured plane as the reference distribution in the gamma analysis.

The in-house designed multiple ion chamber phantom (MIC) consisted of five Exradin A1SL 0.57cc ion chambers set at three-dimensionally independent points (different depths, heights, and lateral positions). The holes for the ion chambers were set in an insert that can rotate to eight different positions. This allowed for the flexibility to position the ion chambers such that the number of ion chambers in a high dose, low gradient region was maximized. This gave five ion chamber measurements per irradiation. Each ion chamber was completely independent, having its own tri-axial cable and electrometer. The measured dose from each ion chamber was used in the reproducibility analysis. The coefficient of variance was calculated for each of the five ion chambers, and then averaged over each patient plan to arrive at a summary statistic.

### 3.2.4 Methods of Analysis

In order to evaluate the reproducibility, the data was divided into "redelivery" and "total delivery" conditions as defined under Delivery Methods above. The three measurements for the "redelivery" and "total delivery" groups were evaluated together to find the standard deviations and coefficients of variation. Ultimately, this yields a coefficient of variation for each patient on each IMRT QA method for both types of

reproducibility ("redelivery" and "total delivery"). The "total delivery" measurement captures the variation introduced by the delivery/readout and the setup, added in quadrature on the reasonable assumption that these errors are uncorrelated because they relate to independent processes. Also these measurements are estimates of the underlying standard deviation in the population, therefore we can use them to solve for an estimate of the population standard deviation in the setup.

**Equation 6**

$$\sqrt{\sigma^2_{redelivery\ measurements} + \sigma^2_{setup}} = \sigma_{total\ delivery\ measurements}$$

Therefore, reproducibility in the setup can be extracted from the "total delivery" measurement by removing the "redelivery" measurement component. This allows us to explore how much variability is derived from the setup and how much from the delivery/readout.

Follow-up analysis for point dosimetry was conducted as related to the reproducibility in a measurement. Ion chambers are accepted as a trusted standard in the field of dosimetry (Low 2011), however in patient specific IMRT QA the ion chamber is often utilized outside of reference conditions (Fraser 2009). There is an assumption that minimizing the gradient across the ion chamber and placing the ion chamber in a high dose region is related to a more reproducible measurement. A common method of determining if the chamber is in an acceptable gradient is by looking at the standard deviation of the dose across the ion chamber volume as calculated by the TPS. By comparing the standard deviation in the dose across the ion chamber active volume with the measured standard deviation in repeated measurements, this work analyzes that assumed relationship. This was

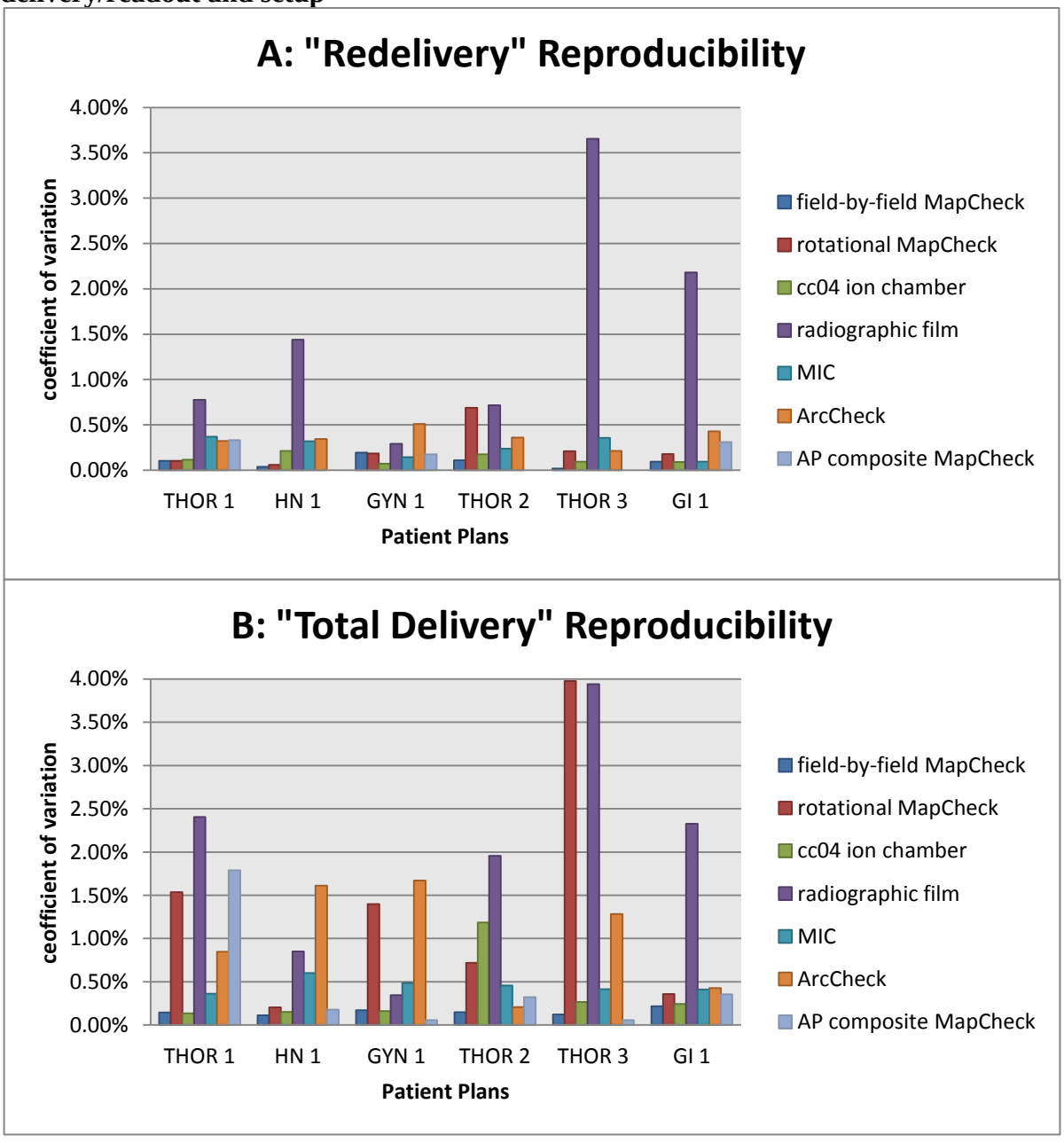
done by plotting the standard deviations of the dose to the ion chamber ROI as calculated in Pinnacle$^3$ 9.0, versus the standard deviations found in the measurements. The R statistical package was used to perform a regression analysis. A minimum dose threshold based on a percentage of the maximum dose in the plan is an additional criteria used to select an appropriate ion chamber measurement point. This method is used to place the ion chamber in an adequately high dose region. A similar regression analysis comparing the standard deviation in the measurement to the percent of the plan maximum dose was performed to analyze this relationship.

### 8.3 Results

Figure 8 A and B shows the "redelivery" and "total" reproducibility for each QA system measured on each patient plan, where the total reproducibility includes the overall uncertainty from setup, delivery, and readout of results. The coefficient of variation describes the variability in the absolute dose measurement (for the ion chamber readings) or the percent of pixels passing gamma (for the planar/array devices). A salient feature of these plots is the heterogeneity in a dosimeter's coefficient of variation (CV) across the different plans. While the plan "Thor 3" showed the overall highest CV for both the "redelivery" and "total delivery", no plan-based statistical difference was found when an ANOVA was run (p-value of 0.88). This indicates that plan-dependent characteristics were not particularly important in terms of device reproducibility. Rather, reproducibility was determined more by the device as detailed below.

Figure 8A, shows the variety of "redelivery" reproducibility exhibited by the different devices due to readout and machine fluctuations, and analysis. This bar plot shows the CV's for each device grouped by IMRT plan. Most devices except film demonstrated consistent QA results, as shown by the relatively small bars. Overall, the AP field-by-field MapCheck displayed the lowest variability in its measurements. The AP composite MapCheck can be seen to have a very low coefficient of variation, and in fact, has a "redelivery" coefficient of variation of 0% for three plans (the exact same percent of pixels passing were obtained for each delivery under the same setup). In contrast, the radiographic film generally had much higher variability in the measurements. Apart from film, all other QA systems showed a less than 1% variation for "redelivery" reproducibility (Figure 8A), indicating little variation in the delivery/readout portion of the QA.

**Figure 8 "redelivery" and "total" reproducibility (CV) for each dosimeter system, grouped by patient plan. All methods showed a CV of less than 4%, with film demonstrating the highest variability for both delivery/readout and total reproducibility, where total reproducibility includes variation from both the delivery/readout and setup**




The "total delivery" reproducibility is displayed in Figure 8B, and shows a greater spread in device performance. This relationship shows how the reproducibility of these

devices may have similar consistency in the read out of their results, but their different setups can lead to QA system-based differences in the constancy of QA results.

To examine specifically the devices' performance, the CV's for each patient plan were averaged across each device. Figure 9A displays the "redelivery" average CV per device (blue bars) with the standard error shown as error bars. Figure 9B similarly shows the "total delivery" average CV (red bar) with standard error bars. Figure 9A shows that the radiographic film demonstrates a clearly higher variability in QA results compared to the other dosimeter systems. An ANOVA was performed with a post hoc Tukey HSD test, which indicated that of the CV's from the "redelivery" measurements, radiographic film was the only device that was statistically different from the other dosimeters (p-value of 0.0001). Figure 9B shows the "total" variability among devices, with larger and more variable values being apparent. This indicates how variability from delivery/readout is relatively higher than the setup in film compared to other dosimeters. An ANOVA was performed with a post hoc Tukey HSD test on the "total" variability. Two groups of devices were apparent, with two devices (ArcCheck and MapCheck with original gantry angles delivered) not being significantly different from either group (p-value of 0.004). This is illustrated in a Venn diagram (Figure 10). Film was again the most variable device, but was not statistically different from all other techniques, as was the case in "redelivery" reproducibility.

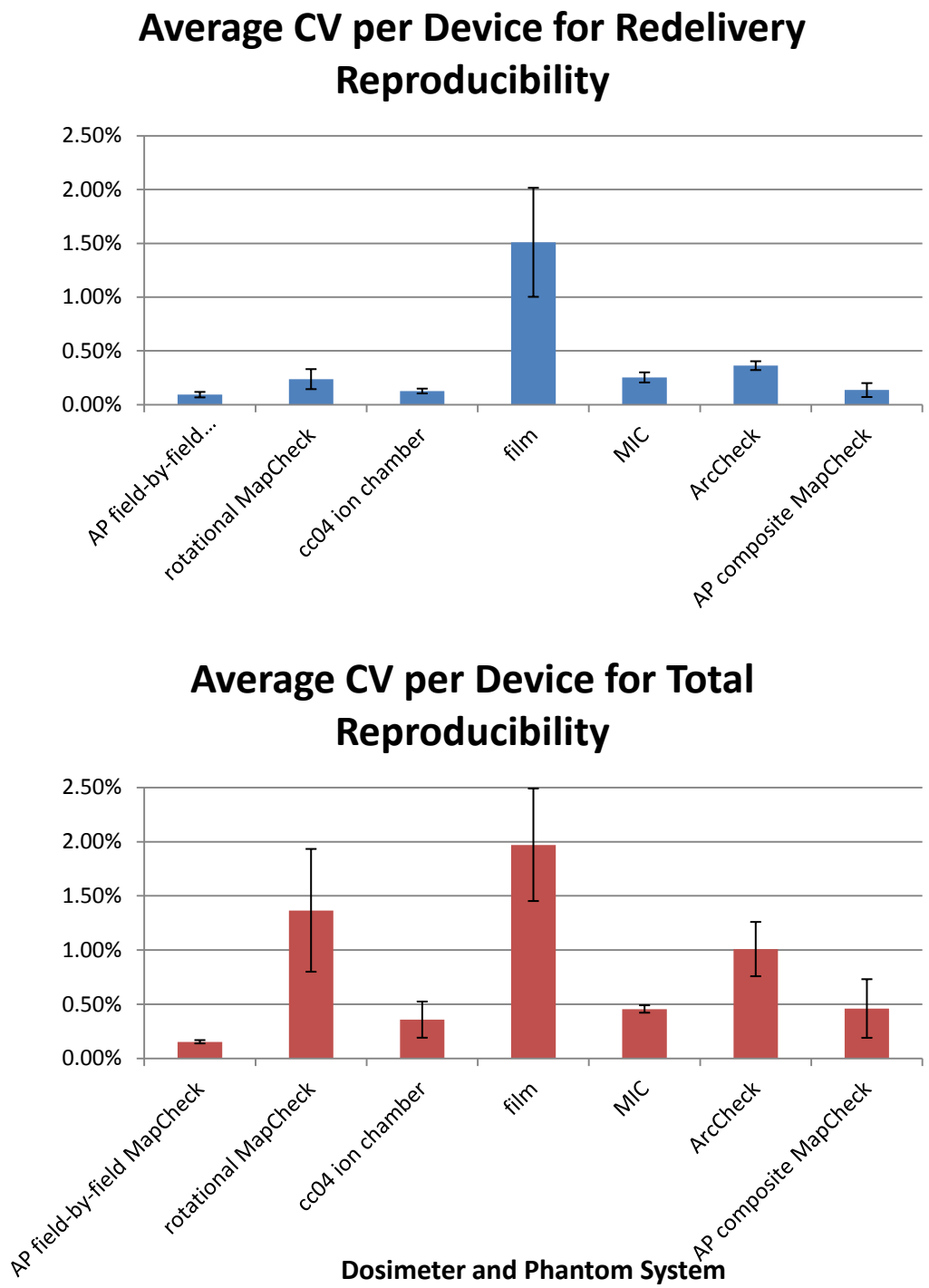



**Figure 9 patient-averaged "redelivery" (blue) and "total" (red) reproducibility (CV) for each device. The errors bars are given as standard error. The film shows the highest variability of all the devices.**

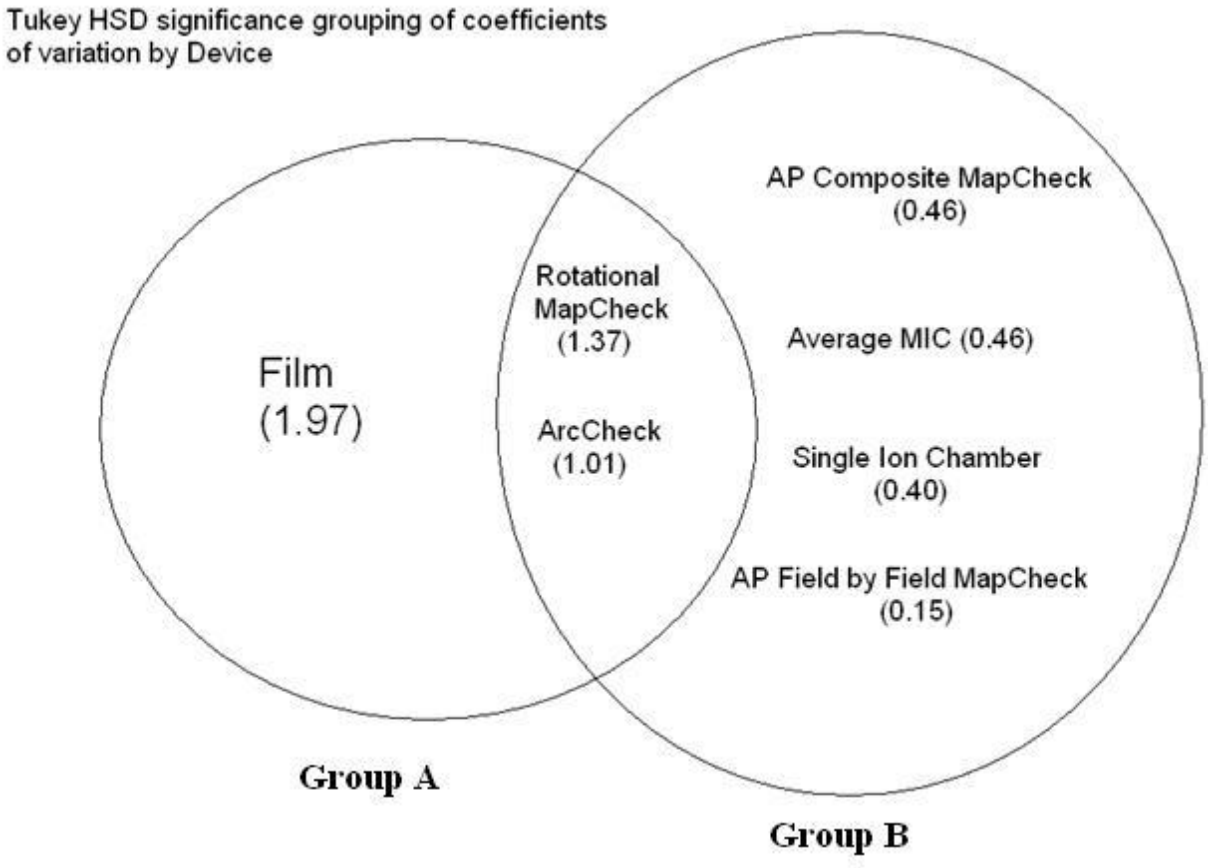



**Figure 10 Venn Diagram showing statistically significant grouping of dosimetric systems based on their "total" reproducibility. The CV of film was significantly different from the cc04 ion chamber, MapCheck with AP beams formed into a composite plane, MapCheck with AP beams analysed in a field-by-field analysis, and the multiple ion chamber phantom (MIC). The CV's of the ArcCheck and the MapCheck with the original rotational gantry angles were not significantly different from either group.**

When assessing the reproducibility of the QA systems, there is a difference in how much variability is contributed from setup and from delivery/readout. Each device showed different sensitivities to these two sources of variability. The "redelivery" measurement

captures the reproducibility from readout/delivery, while the "total" measurement contains contributions from both the setup and readout/delivery. Therefore, the setup reproducibility was calculated through the use of the "total" and "redelivery" measurements in Equation 6. From these values, the degree of influence from delivery/readout and setup can be evaluated. These results are summarized in Table 1. From this table, it can be seen that film is the only dosimeter with a higher delivery/readout contribution than the setup (59% vs. 41%). All other QA systems show a much greater variability stemming from the setup, with the rotational MapCheck demonstrating the highest (97% from setup vs. 3.0% from delivery/readout). It is of interest to note that the same physical dosimeter setup was used for AP field-by-field, rotational, and AP composite MapCheck. However the differences in the way the results were obtained and analyzed has led to different degrees of sensitivity to setup and readout/delivery. For example, the variability in the AP composite MapCheck and MapCheck with original gantry angles delivered shows a higher dependence on the setup than the AP field-by-field MapCheck. Additionally, the values of the CV's averaged across patient plans shows a lower variability in the AP field-by-field MapCheck than in its other two configurations.

**Table 1 Isolating the error caused by setup alone from the redelivery and total coefficients of variation averaged across all patient plans**

| | "redelivery" measurement CV | setup CV | "total delivery" measurement CV | % variation from delivery/read-out | % variation from setup |
|---|---|---|---|---|---|
| **AP field-by-field Mapcheck** | 0.09% | 0.12% | 0.15% | 37% | 63% |
| **rotational MapCheck** | 0.24% | 1.3% | 1.4% | 3.0% | 97% |
| **cc04 ion chamber** | 0.13% | 0.33% | 0.36% | 13% | 87% |
| **radiographic film** | 1.5% | 1.3% | 2.0% | 59% | 41% |
| **MIC IC Avg** | 0.25% | 0.38% | 0.46% | 31% | 69% |
| **ArcCheck** | 0.36% | 0.94% | 1.0% | 13% | 87% |
| **AP composite Mapcheck** | 0.14% | 0.44% | 0.46% | 8.7% | 91% |

Ion chamber measurements are generally the most trusted. In clinical practice, this reliability is assumed to be partially dependent on the standard deviation of the dose as calculated across the ion chamber volume ROI. The common guidance for the placement of an ion chamber measurement is to put it in a high dose, low gradient region. The dose is usually assessed by seeing what percent of the plan maximum dose the average dose across

the ion chamber is. The gradient is assessed by calculating the standard deviation in the dose across the ion chamber ROI. As an additional analysis into the nature of QA reproducibility, this assumption was studied by comparing the percent standard deviation in the reproducibility measurements across the six patient plans with the percent standard deviation in the dose across the ion chamber ROI in the TPS. Using a linear model to fit the percent standard deviation in the measurement data to that found in the ROI, 95% prediction confidence bands were calculated for the ability to predict a percent standard deviation in the measurement based on the percent standard deviation calculated across the ion chamber. Because an arbitrary measurement includes setup uncertainty, this work only evaluated the results of the dose standard deviation compared to the "total delivery" reproducibility. These results are shown in Figure 11. The R-squared value for this model is only 0.36, showing very little linear relationship. This plot also shows that the prediction bands are wider than the plot in some places, revealing a lack of prediction power – that is, this research failed to show an ability to predict the reproducibility of a measurement from the standard deviation in the dose across the ion chamber. The analysis was done again with expansions of the ROI ion chamber volume by 1, 2, 3, and 5mm to attempt to capture any effects from the local dose environment. Again, these showed little relationship between these two features, with R-squared values of 0.39, 0.41, 0.42, and 0.41 for the 1, 2, 3, and 5mm expansions, respectively (Table 2).

**Table 2 The results of the regression analysis for 0, 1, 2, 3, and 5mm expansions around the ion chamber volume compared to resetup standard deviation**

| expansion on IC ROI (mm) | R-sqr | F-statistic | p-value |
|---|---|---|---|
| 0 | 0.36 | 16.04 | 0.0004 |
| 1 | 0.39 | 18.03 | 0.0002 |
| 2 | 0.41 | 19.23 | 0.0001 |
| 3 | 0.42 | 20.3 | 0.0001 |
| 5 | 0.41 | 19.41 | 0.0001 |

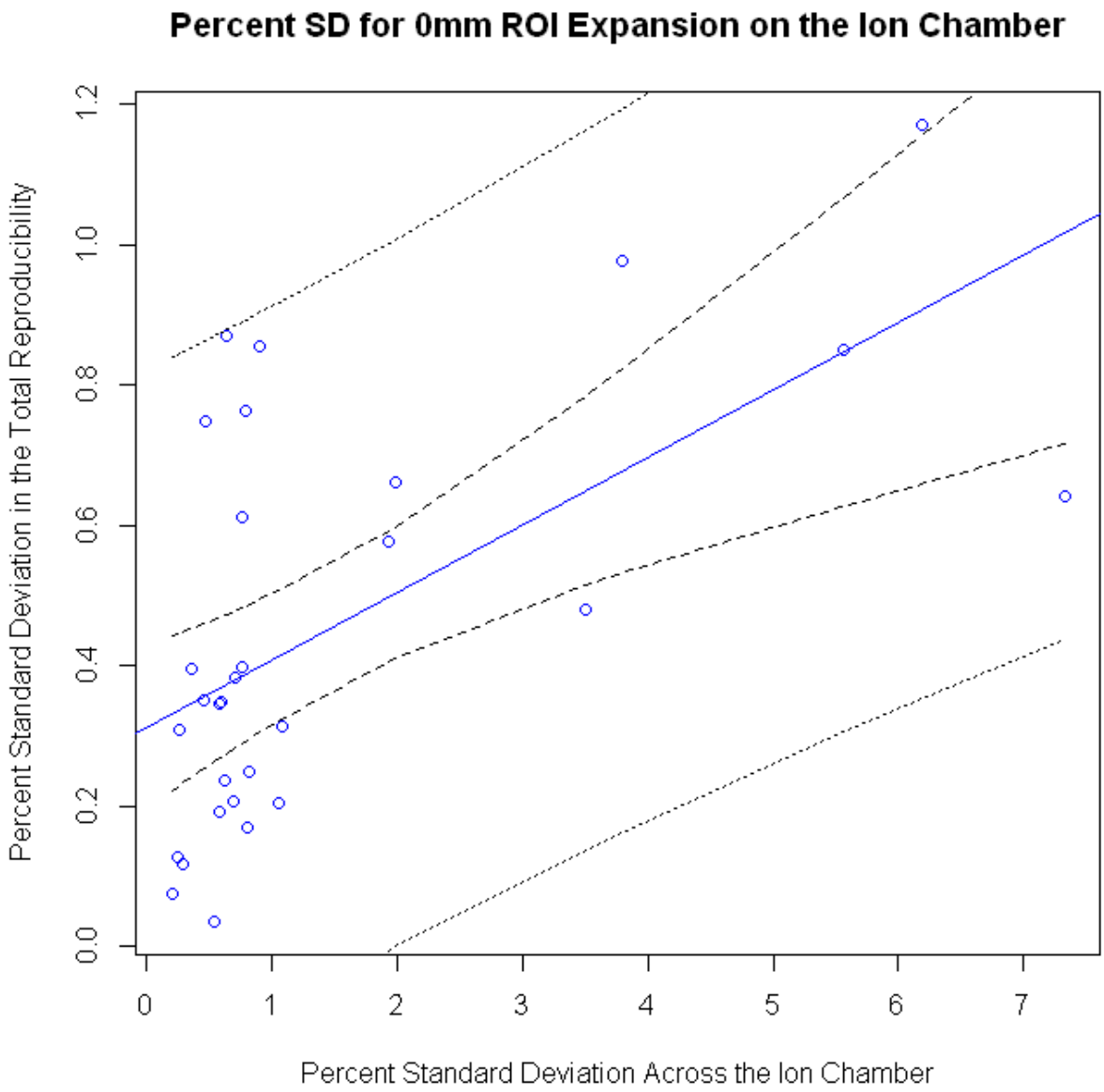



**Figure 11 Regression performed to see the ability to predict percent "total" reproducibility given a known percent standard deviation in dose across the ion chamber. This percent standard deviation in dose is used to assess the severity of the dose gradient across the ion chamber. The inner pair of dotted lines is the confidence interval in the linear fit, while the outer pair of lines is the prediction confidence interval.**

An additional examination of the ion chamber data also failed to show a relationship between total reproducibility and the percent of the plan maximum dose in the ion chamber ROI (Figure 12). This plot shows that the percent standard deviation mildly decreases as the dose increases in the ion chamber active volume. However, this regression line was not shown to be statistically significant (p-value of 0.74), revealing a lack of linear fit in the data. Looking at the data points themselves, it is worth noting that even at a high percent of the maximum dose (i.e. 96% of the plan's maximum dose), there is a possibility of seeing a non-zero percent standard deviation in the total reproducibility (i.e. 0.86%).

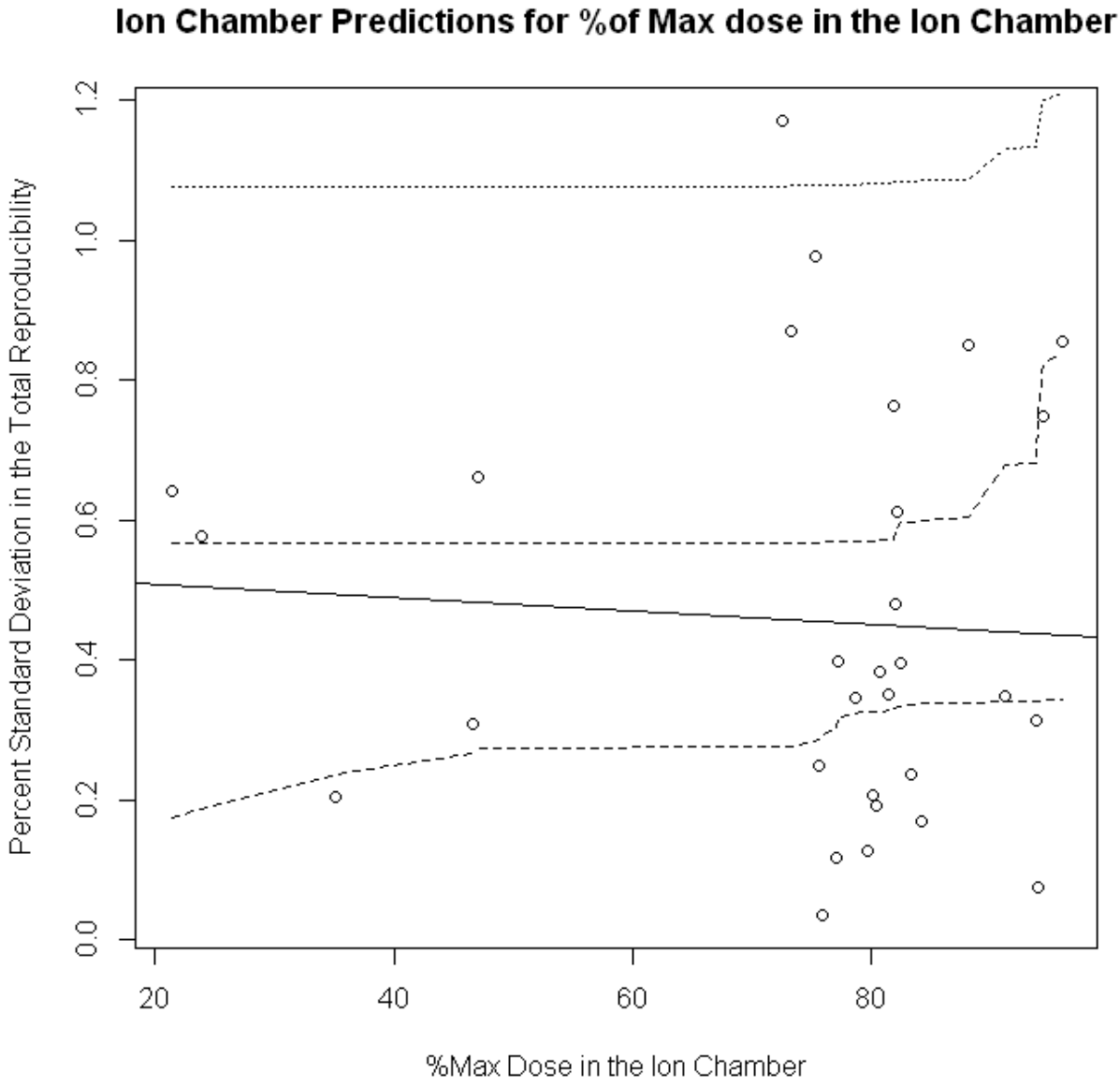



**Figure 12 A regression analysis exploring whether knowing the percent of the plan's maximum dose in the chamber ROI can predict the "total" reproducibility in the measurement is shown. As with the standard deviation in the dose study, the inner pair of dotted lines is the 95% confidence interval in the linear fit, while the outer pair of lines is the prediction confidence interval.**

### 8.4 Discussion

Performing a reproducibility study on different patient specific IMRT QA equipment and methods shows a comparison in the potential for variation in QA results. The most salient finding was the lower reproducibility in EDR2 film, which was significantly different from the other methods in both the "total delivery" (a coefficient of variation of 2.0%) and "redelivery" (a coefficient of variation of 1.5%) studies. This is most likely influenced by the high degree of human interaction involved in the readout and analysis of the film performed at the author's institution. The user has input on aligning the film to the calculated dose plane. The user also selects an ROI for the gamma analysis and can choose a normalization point from any location in the dose plane. No other QA method in this publication's study allowed for as much user latitude. An additional aspect inherent to film is the variation introduced from the film processing. Time before developing and developer conditions can all influence the optical density of film (Childress 2005).

There were two types of reproducibility measured: "redelivery" and "total delivery." When the patient plan CV's are averaged per device, the "total delivery" reproducibility was generally higher than the "redelivery" reproducibility since it includes variation from both delivery/readout and setup. After extracting the setup reproducibility from the "total delivery," it was seen that the setup was more responsible for the variability than the readout/delivery. The only exception to this was the radiographic film, with its more involved readout. The "total delivery" study most closely mimics what a clinic could expect as the equipment is setup from day to day to perform QA. The "redelivery" study represents

what kind of variation would be seen if nothing was moved and only the plan was redelivered. In the clinic, when a plan gives suspicious QA results, sometimes it is redelivered with the assumption that there was some error in the initial delivery. This study shows what kind of variation could be expected from simply retaking the plan again. For example, if radiographic film is your dosimeter and you set it up again for another measurement, you could obtain a 1% higher result in percentage of pixels passing. It is important to notice that on average the coefficients of variation in the redelivery study are non-zero across all devices investigated, showing that there is some variability inherent in the delivery and readout, itself.

Apart from film, none of the other methods are significantly different in their coefficients of variation for the "redelivery" reproducibility. These other 6 methods have immediate readout capabilities, minimizing the degree of human interaction. However, this homogeneity decreases for the "total" reproducibility. The Tukey HSD test divides the methods into two significant groups (labeled group A and B in Figure 10), with two methods belonging to both groups. All of the methods in Group A of Figure 10 consist of a rotational delivery to a planar dosimeter. Because this grouping appears in the "total delivery" reproducibility results, the planar measurements with original gantry angles appear to be more sensitive to setup errors than when the fields are delivered perpendicularly. This may be due to the fact that there is an increase in the degrees of freedom in the motion of the dosimeter relative to the delivery of the beam. All point measurements and AP planar measurements belong in Group B with lower coefficients of variation. Within the Group B exhibiting low variability, the AP composite MapCheck's higher CV may be due to the fact that any discrepancies in each field have been combined together. The AP field-by-field

MapCheck shows the lowest variability. This might be from a combination of the ease of setup and the lower degrees of freedom of motion between the dosimeter and beam delivery.

In the portion of this study investigating the relationship between the variations calculated in the dose across the ion chamber and the variation in the measurement, we failed to find a significant relationship between these two characteristics. In fact, the highest adjusted R squared value showed that the variation in dose across the ion chamber could only predict 42% of the variation witnessed in the "total delivery" measurements. Similar analysis with expansions around the ion chamber ROI of 1, 2, 3, and 5mm showed that even when incorporating the local dose environment, we were unable to prove a significant relationship between standard deviation in the dose across the ion chamber ROI and reproducibility.

Prediction intervals were also calculated to judge the range of expected measurement reproducibility given a certain choice of percent standard deviation in dose across the ion chamber. Because this was done as follow-up analysis and not a primary goal of this study, most of the data is clustered below 2% standard deviation across the chamber. Consequently, the confidence interval is wide and there is little data at large standard deviations. Nevertheless, some information may be gleaned from this data. For a 0mm expansion around the active volume, a standard deviation in dose of less that 2% should lead to a standard deviation in your measurement of less than 1%. Of note, even a point with virtually no gradient across the chamber can still readily show a measurement reproducibility of nearly 1%.

## 8.5 Conclusion

By analyzing the coefficients of variation across seven different methods of patient-specific IMRT QA, the goal of this research was to provide a comparison of the reproducibility of a QA measurement. In terms of reproducibility all methods demonstrated coefficients of variation less than 4%, both for delivery/readout and for the total reproducibility. Film performed the poorest, with an average coefficient of variation across all plans of 1.5% for delivery/readout and 2% for total reproducibility (including delivery/readout and setup). Excluding film, the next largest average CV was 0.36% (ArcCheck) for delivery/readout reproducibility, and 1.4% (planned angle MapCheck) for the total reproducibility. When the setup reproducibility was extracted from the measurements that looked at the total reproducibility, it was seen that the setup contributes a greater amount of variability than the delivery/readout for all QA methods except film. This could be explained by the fact that film's readout is less automated than the other methods. With the data provided from this study, a clinic can have a greater insight into the expected reproducibility of their patient-specific IMRT QA with respect to the dosimeter and method employed.

# I. Chapter 4

## Towards an Optimization of Patient-Specific IMRT QA Techniques

### 4.1 INTRODUCTION

Intensity modulated radiation therapy (IMRT) is a commonly practiced form of radiation therapy. Because of the complexity of this treatment technique, verification of patient plans is performed via direct measurement, called patient-specific IMRT quality assurance (QA). Despite the widespread practice of IMRT QA, its implementation has not been standardized, and many methods and types of equipment exist to accomplish it (Nelms 2007). With such heterogeneity in the field, we asked whether the efficacy among the various methods is equal or whether there is an optimal way to perform IMRT QA with the goal of distinguishing between acceptable and unacceptable plans. This is further complicated by it not only being a question of the detector used in performing IMRT QA; but also of how the data should be is analyzed. Whereas ion chamber measurements typically rely on a percent dose difference cutoff, gamma analysis for planar QA relies on three parameters: percent dose difference, distance to agreement, and percent of pixels passing (Low 1998). Additionally, multiple software packages exist for gamma analysis, which may implement the calculation differently.

Insight into this question can be achieved by evaluating various IMRT QA techniques using receiver operating characteristic (ROC) curves, which can address the question of performance for both hardware and the methodologies used (DeLong 1988). Recently published comments have called attention to the apt application of ROC analysis as a quantitative means of assessing the practice of patient-specific IMRT QA (Gordon 2013). In ROC analysis, a curve of the sensitivity and specificity of a test is plotted as the values of

the cut-off are varied. In this study, *sensitivity* is the ability of a dosimeter to accurately label an unacceptable plan as failing; conversely, *specificity* is the ability to label an acceptable plan as passing. Cutoff values in this study were the percent of pixels passing for gamma analysis and the percent difference for ion chamber measurements. There is an inherent trade-off in these two parameters: as the cutoff is rendered more stringent to increase sensitivity, the specificity decreases. A ROC curve gives the user a convenient, holistic view of these trade-offs across all cutoffs. An example ROC curve is shown in Figure 13, where the vertical axis is sensitivity (range, 0 to 1), and the horizontal axis is 1 – specificity (range, 1 to 0). The ROC curve for an ideal dosimeter that perfectly sorts patient plans, which is also shown in this figure, has an area under the curve (AUC) equal to one. In contrast, a 45 degree diagonal line (AUC equals 0.5) represents a dosimeter that sorts plans completely randomly. The AUC is a useful metric with which to determine the performance of a device over the entire range of cutoff values. This AUC is also equivalent to the probability that for a randomly selected acceptable and unacceptable plan, the dosimeter correctly classifies these two plans as passing and failing. A detailed explanation of ROC techniques is well explained in the literature (Metz 1978).

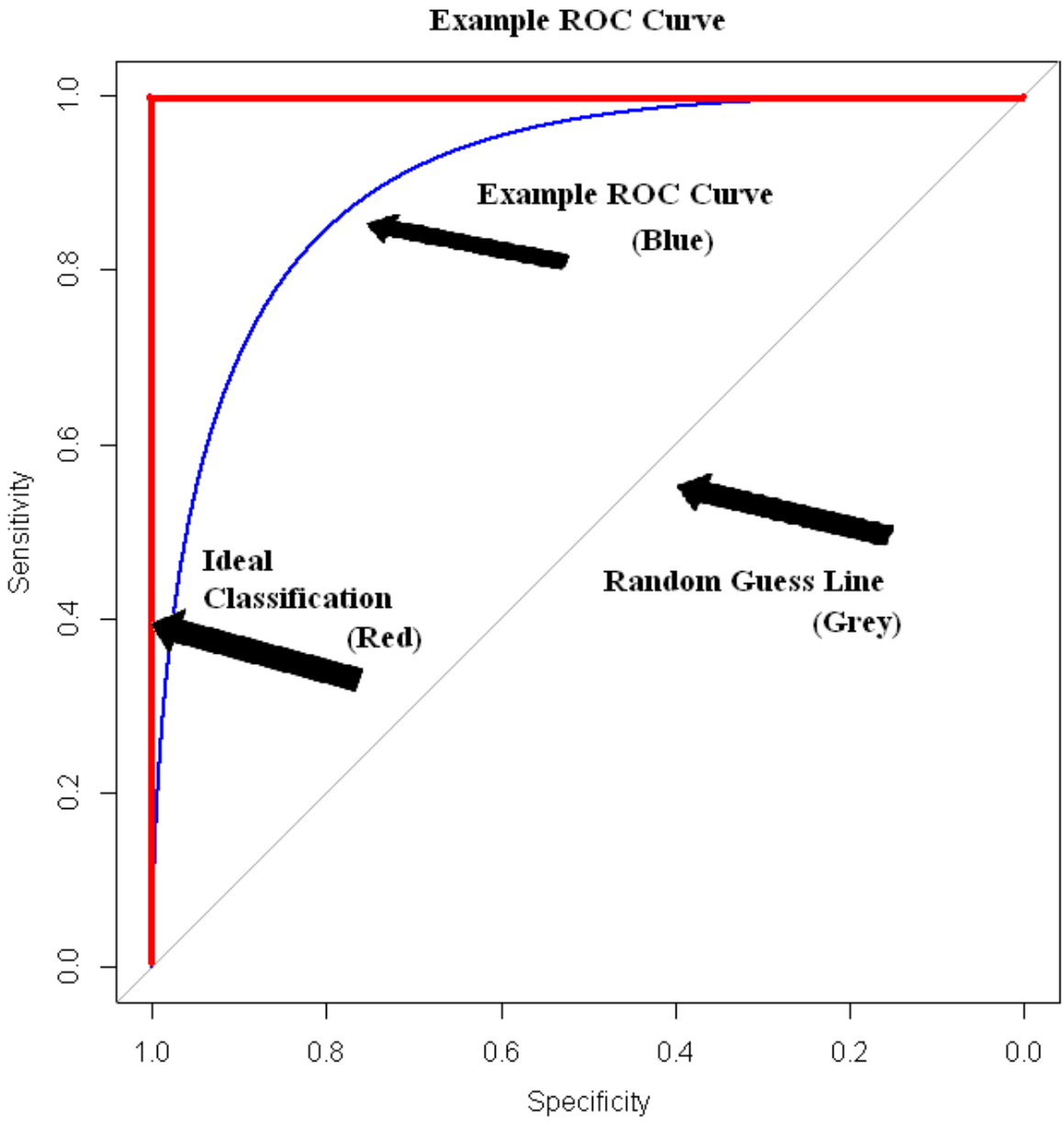



**Figure 13 ROC curve given as an example. This type of plot shows the ability of a test to accurately sort incidents, where the true state is determined by a gold standard. The vertical axis shows sensitivity, whereas the horizontal axis shows specificity. The thicker red line shows a test with perfect classification, whereas the thinner blue line shows what a realistic ROC curve would look like for a test. The diagonal grey line is the ROC curve that would result from a test with random classification.**

One recent study examined a diode array's optimal cutoffs through the lens of ROC analysis (Carlone 2013). Many other studies have also explored this large question of optimal IMRT QA criteria (Nelms 2007; Howell 2008; Wilcox 2008; Kruse 2010). However, none have applied this analysis technique to study a broad range of dosimeters, comparing not only the hardware, but also the protocol used in the setup and analysis. Consequently, the relative performance of various QA techniques remains unclear. To this end, this research uses ROC analysis to discover which, if any, of the most commonly used QA procedures perform most robustly in terms of both their sensitivity and specificity, and what optimal cutoffs can be gleaned from these ROC curves. More specifically, we investigated the abilities of the MapCheck2 in a variety of configurations, ArcCheck, radiographic film, and an ion chamber using original clinical patient plans to generate clinically relevant comparisons.

## 4.2 METHODS

### 4.2.1 Patient selection

Twenty-four clinical patient plans were selected from our database of previously delivered patient-specific IMRT QA plans at our institution. To more rigorously test the performance of various QA dosimeters, most of these plans (19) were selected from a group that had previously failed film and ion chamber QA at the authors' institution (<90% pixels passing at 5%/3mm, or an ion chamber reading of >3% dose difference). These plans were not modified to artificially create failing cases; instead they were true clinical IMRT plans created with the intent of patient delivery. Since it is highly difficult to predict all possible failure modes in IMRT plan delivery, it was believed that actual patient plans would be

more insightful than using induced errors. The remaining 5 plans previously passed IMRT QA. In addition, a variety of treatment sites (thoracic, gastrointestinal, head and neck, stereotactic spine, gynecologic, mesotheliomal, and genitourinary) were selected to ensure that the scope of dosimeter performance would reflect the variety of plans seen in the clinic. All of the plans were calculated in the Pinnacle$^3$ 9.0 treatment planning software (TPS) (Phillips Healthcare, Andover, MA). The clinical acceptability of each plan was determined on the basis of measurements in a multiple ion chamber phantom (described below). Then, each plan was delivered to a variety of dosimeters to assess the performance of each dosimeter.

### 4.2.2 Dosimeters

#### *Multiple ion chamber phantom*

An in-house designed multiple ion chamber phantom (Figure 14) was selected as the gold standard with which to classify a plan as acceptable or unacceptable. This sorting of plans into acceptable or unacceptable was considered the "true" sorting and was based solely on the results of the multiple ion chamber phantom; such sorting was unrelated to the original internal IMRT QA results. The performance of all of the other dosimeters was compared with the classification results of the multiple ion chamber phantom.

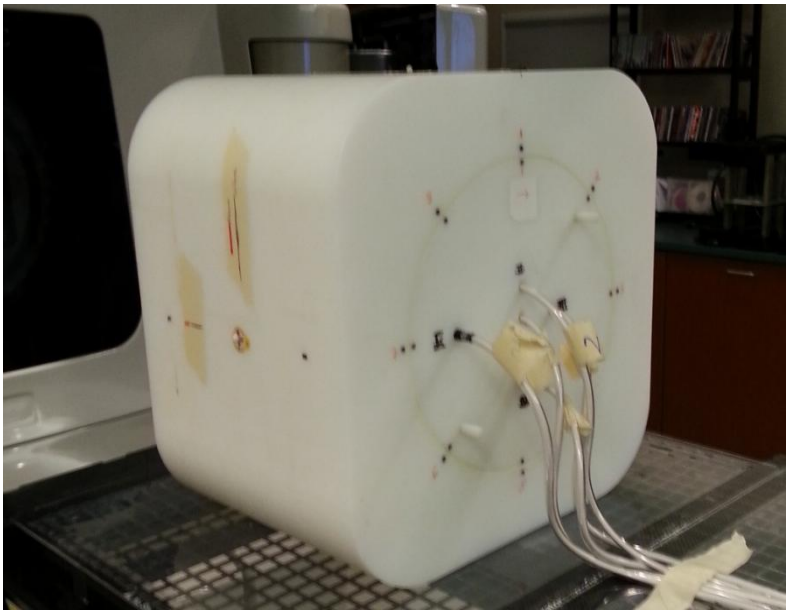


**Figure 14 Multiple ion chamber phantom irradiation setup. This phantom contains five ion chambers placed in an insert that can rotate to eight positions. The ion chambers are located at 3-dimensionally independent locations to better sample the IMRT QA. This phantom was used as the gold standard for this study.**

The ion chamber is accepted as a reliable benchmark in radiation therapy dosimetry (Bogner 2004); however it is only a point measurement. To more fully evaluate each plan, the multiple ion chamber phantom was created with five ion chambers (Exradin A1SL 0.057cc) positioned at unique depths, heights, and lateral positions within a cylindrical insert. This insert can rotate to eight different positions, allowing a large number of points to be three-dimensionally sampled. All 24 patient plans were delivered at the original gantry angles with two different insert rotations, leading to 10 ion chamber readings per patient plan. The phantom's insert rotational positions, along with shifts to the phantom, were made to maximize the number of points that fell within a high-dose, low-gradient region. Each ion chamber was calibrated by an Accredited Dosimetry Calibration Laboratory, and the absolute dose was determined at each measurement location. This dose was then compared

with the dose calculated by the planning system over the volume corresponding to the active volume of the ion chamber.

Although the definition of a truly acceptable versus unacceptable plan is ultimately a matter of clinical judgment, the use of multiple ion chamber measurements as the gold standard with which to classify plans has been previously used in IMRT QA comparisons (Kruse 2010).

Each plan was then also delivered to the dosimeters listed below to assess the sorting performance of each. The algorithm used to transform measurements from the multiple ion chamber phantom to a label of acceptable or unacceptable for each IMRT plan is more fully explained in APPENDIX III.

***MapCheck***

A diode array (MapCheck2, Sun Nuclear Corporation, Melbourne, FL) with 5-cm water equivalent buildup was used to measure the delivered dose distribution in three separate ways. *The first method* was a field-by-field analysis with all of the plans' beams delivered with a gantry angle of zero degrees (anterior-posterior field). The percent of pixels passing per field were combined by using an MU-weighted average to provide a single value of percent of pixels passing for all of the fields. *The second method* combined all of the AP-delivered fields into a composite measurement and compared that with the composite calculated dose plane. *The third method* delivered all of the fields at their original rotational gantry angles with the MapCheck placed in the MapPhan phantom.

Because most of the original gantry angle fields did not enter laterally, the plans were delivered with the diode array flat on the treatment couch (as per the manufacturer's

instructions). For all MapCheck configurations, the diodes were calibrated for absolute dose and corrected for accelerator daily output fluctuations. Plans on all three methods underwent gamma analysis (Low 1998) at 2%/2mm, 3%/3mm, and 5%/3mm using both SNC Patient software (Sun Nuclear Corporation, Melbourne, FL) and DoseLab Pro software (Mobius Medical Systems, Houston, TX). With SNC Patient, the region of interest (ROI) accounted for the known locations of the diodes, whereas with DoseLab Pro, the ROI was automatically selected. In both software packages, the TPS was used as the evaluated distribution in the gamma analysis.

***Film and ion chamber***

Radiographic film (Kodak EDR2) and a single ion chamber (Wellhofer cc04) were placed in an I'mRT body phantom (IBA Dosimetry, Schwarzenbruck, Germany). The plans were all delivered with their original gantry angles. Due to inherent differences in the types of measurement (point vs. planar, absolute dose vs. relative dose), the ion chamber and film were analyzed as two separate dosimeters, although their measurements were taken simultaneously. The ion chamber was placed in a position with a standard deviation across the ion chamber ROI of less than 1% of the mean dose, and a mean dose of greater than 70% of the maximum dose in the plan. Shifts to the phantom were applied if necessary to satisfy these criteria. The absolute dose was determined from a transfer factor applied to the electrometer reading, corrected for daily output of the accelerator, and compared with the dose calculated by the planning system over the volume corresponding to the active volume of the ion chamber.

The film evaluated a transverse plane of the delivered dose distribution. It then underwent gamma analysis in the Omnipro I'mRT software (IBA Dosimetry, Schwarzenbruck, Germany) at 2%/2mm, 3%/3mm, and 5%/3mm, all with a 10% dose threshold. In this software, the ROI was manually selected to be the area of the film contained within the phantom, and the TPS was selected to be the reference distribution in the gamma analysis. The film optical density was converted to dose with use of a batch-specific calibration curve and spatially registered with use of pin pricks. The film was then used as a relative dosimeter with the normalization point manually selected to maximize agreement with the calculated plane.

***ArcCheck***

The ArcCheck (Sun Nuclear Corporation, Melbourne, FL) cylindrical diode array was treated with the electronics facing away from the linear accelerator. As with the MapCheck, the array was calibrated for absolute dose. If necessary, shifts were applied to the ArcCheck to avoid irradiating the electronics. Gamma analysis was performed in the SNC Patient software at 2%/2mm, 3%/3mm, and 5%/3mm, with a 10% dose threshold.

### 4.2.3 Data analysis

First, we defined which plans were acceptable and which were unacceptable based on the multiple ion chamber measurements. A gold standard need not be infallible, but it must be considerably more accurate than, and independent of, the tests being evaluated (Metz 1978). The 10 ion chamber points measured on each patient plan were pruned down to include only those that satisfied the criteria of an ROI mean dose of at least 50% of the

maximum plan dose and a standard deviation of less than 1% of the mean dose across the ion chamber ROI, i.e., points in a relatively high dose and low-gradient region. These ion chamber measurements were then compared with their expected values calculated in the TPS. The deviations at the points showing a greater than 3% dose difference were summed together and then normalized by the number of points in high-dose, low-gradient regions used to assess the plan. This final value is in essence an average deviation metric and was the metric that was used to summarize the ion chamber measurements for each plan into one value, accounting for the varying number of points per plan remaining after the dose and standard deviation criteria.

Based on hierarchical clustering (Hastie 2009), a plan was classified as acceptable if its multi-ion chamber metric was less than 0.30; a plan was considered unacceptable if this metric was greater than 0.30. Although this cutoff threshold was defined on the basis of the arbitrary inclusion criteria of only those points receiving >50% of the maximum dose and <1% standard deviation across the chamber, this threshold was actually quite robust; plans were sorted identically based on any inclusion of points between 25% and 65% of the maximum dose and up to 2% standard deviation. Moreover, multiple ion chamber readings were also used as a gold standard in work by Kruse, in which he classified a plan as failing if any individual ion chamber measurements differed by greater than 4% (Kruse 2010). Interestingly, his methodology sorts our data set's plans in the same way that the method used in our study does. The method used in this research is shown in APPENDIX III.

Once the plans were sorted into acceptable and unacceptable plans, the ability of each alternate dosimeter to correctly sort the plans could be conducted. This was done by using ROC analysis. ROC curves were formed by comparing the passing and failing results

of each dosimeter system on the set of 24 acceptable and unacceptable patient plans. These curves have a staircase-like pattern due to the finite number of cases considered. Because of its independence from the prevalence of unacceptable plans, sensitivity weighting, and specificity weighting, the AUC statistic is commonly used to compare different ROC curves (DeLong 1988). It was therefore used here to compare each device's discriminating capabilities. Confidence intervals were calculated with the use of the bootstrap method implemented in the pROC R package (Robin 2011). Bootstrapping was also applied to compare the AUCs using the "Z" statistic and by obtaining p-values to determine whether pairs of AUCs were significantly different (Robin 2011). An explanation of bootstrapping is found in APPENDIX II.

ROC curves are generated by considering all possible thresholds (e.g., all ion chamber dose difference thresholds, or all percent of pixels passing for a given dose difference and distance to agreement criteria). Once a ROC curve has been generated, a natural follow-up is to find the value on the curve (e.g., what percent of pixels passing threshold) that provides the best discriminatory power. Optimal cutoff criteria and their accompanying confidence intervals were calculated in the R statistical packages pROC (Robin 2011) and ThresholdROC (Skaltsa 2010). The Youden Index method was used, which finds the point along the curve farthest from the diagonal random guess line (Perkins 2006). The Youden Index has been shown to be a more robust optimization method than is finding the point closest to perfect classification (0,1) (Perkins 2006). However, this optimal point may not accurately reflect practical realities. For example, if the prevalence of a failing plan is low, having an overly sensitive cutoff could lead to an excessive number of false positives (i.e., acceptable plans labeled as failing), wasting time in the clinic. Conversely, if

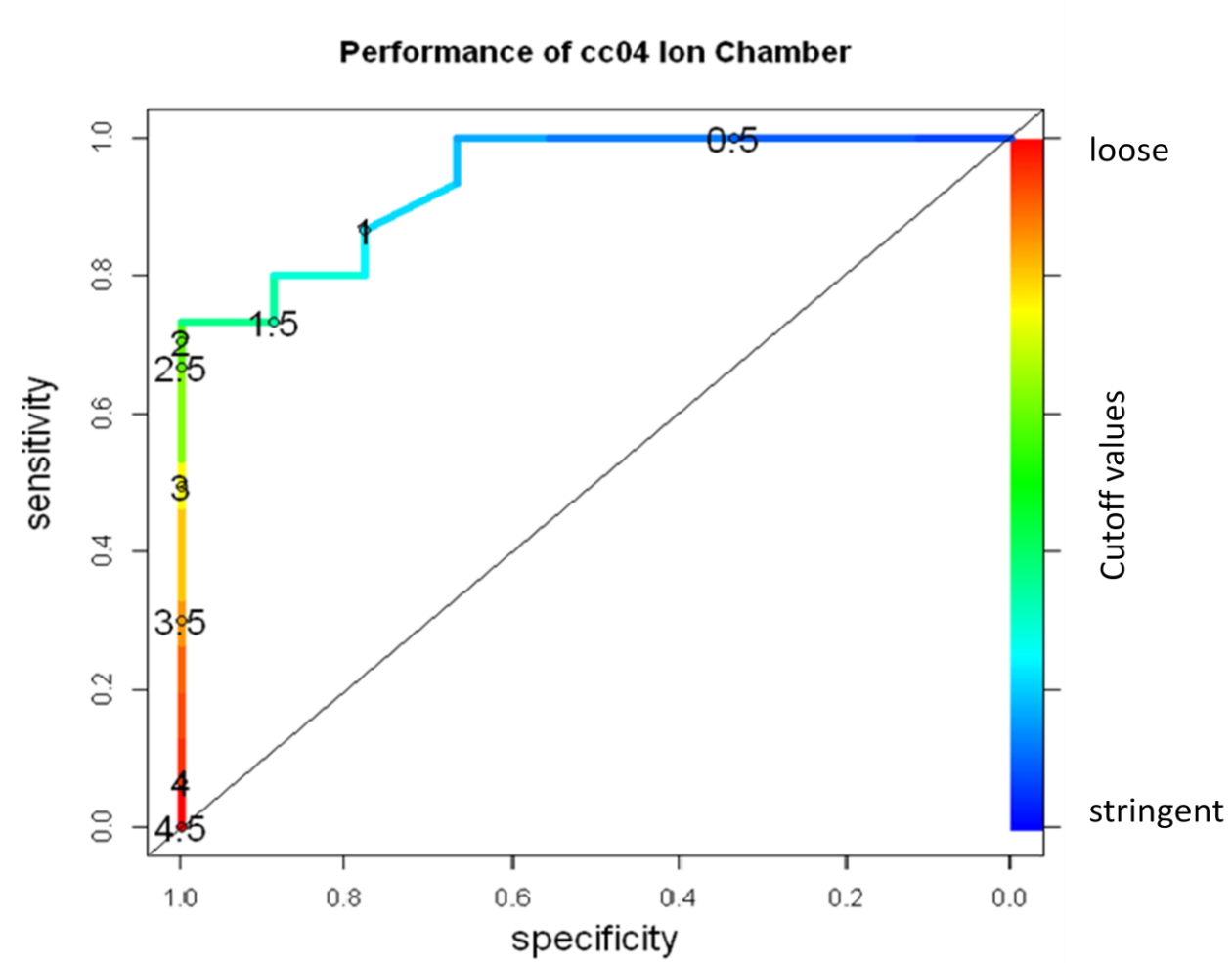
the cost related to passing an unacceptable plan is high, a sensitive cutoff would be favored over one with high specificity. Therefore, the optimal cutoff values were calculated with use of the ThresholdROC package by minimizing a cost function that incorporates the cost of false negatives and prevalence (Skaltsa 2010). The prevalence of a failing plan was estimated at 3% based on the work by Dong (Dong 2003). The cost values are dependent on the situation at a particular clinic and can include such factors as the risks of delivering a failing plan to a patient, tempered by the extra time demanded in the clinic if an acceptable plan is falsely labeled as failing. To estimate a common cost metric reflected in clinical practice (e.g., a 3% acceptance criteria with an ion chamber), the cost was varied until the optimal cutoff of 3% was generated for the cc04 ion chamber (Ezzell 2003). This cost weighting was then used to determine the equivalent threshold (i.e., using the same cost weighting) for the other dosimeters examined.

### 4.3 RESULTS

Each of the 24 plans was delivered to the multi-ion chamber phantom. After pruning data points to exclude those with low doses or high dose gradients, the average number of ion chamber measurements per patient plan was 6, with a minimum of 4 and a maximum of 9. This gold standard ultimately sorted the 24 plans into 9 acceptable and 15 unacceptable plans, yielding a good distribution of plan challenge levels on which to rigorously test the different QA systems.

After delivery of these 24 plans to each QA device, ROC curves were created. As an example, Figure 15 shows the ROC curve generated for the single cc04 ion chamber in the I'mRT phantom (Sing 2005). The numbers printed on the curve are the cutoff values (in %

dose difference). Across the 24 patient plans, the percent difference for the ion chamber ranged from 0% to 4.5%. As would be expected, as the cutoff increased from more liberal (4.5%) to more stringent (0.5%), the sensitivity increased (i.e., the device was better at detecting and failing unacceptable plans). Concurrently, the specificity decreased (i.e., the device was less adept at passing acceptable plans). The curve for this dosimeter lies well above the "random guess" diagonal line, showing an overall strong ability to discriminate between acceptable and unacceptable plans.




68

**Figure 15 ROC curve for the cc04 ion chamber. This plot shows how the ROC curve is generated by varying the cutoff values from more to less stringent. The percent dose difference cutoff values used to create the curve are numerically printed on the curve and also color coded, using a spectrum with red being the least stringent and blue being the most.**

An ROC curve was generated for each QA system, and in the case of the planar measurements, for gamma criteria of 2%/2mm, 3%/3mm, and 5%/3mm, leading to 16 curves shown in Figure 16. The MapCheck curves shown in Figure 16 include only gamma analysis results from the SNC Patient software. Of those curves, the MapCheck with original

rotational gantry angles delivered (Figure 16a) consistently fell close to the diagonal line, indicating poor discriminatory abilities. Similarly, the MapCheck curves for field-by-field AP beam delivery also fell close to the diagonal line (Figure 16b). In contrast, ROC curves that were relatively far from the diagonal line were the MapCheck with all AP fields formed into a composite dose plane (Figure 16c), the cc04 ion chamber (Figure 16d), the ArcCheck (Figure 16e), and film (Figure 16f), indicating a relatively strong ability to discriminate between acceptable and unacceptable plans.

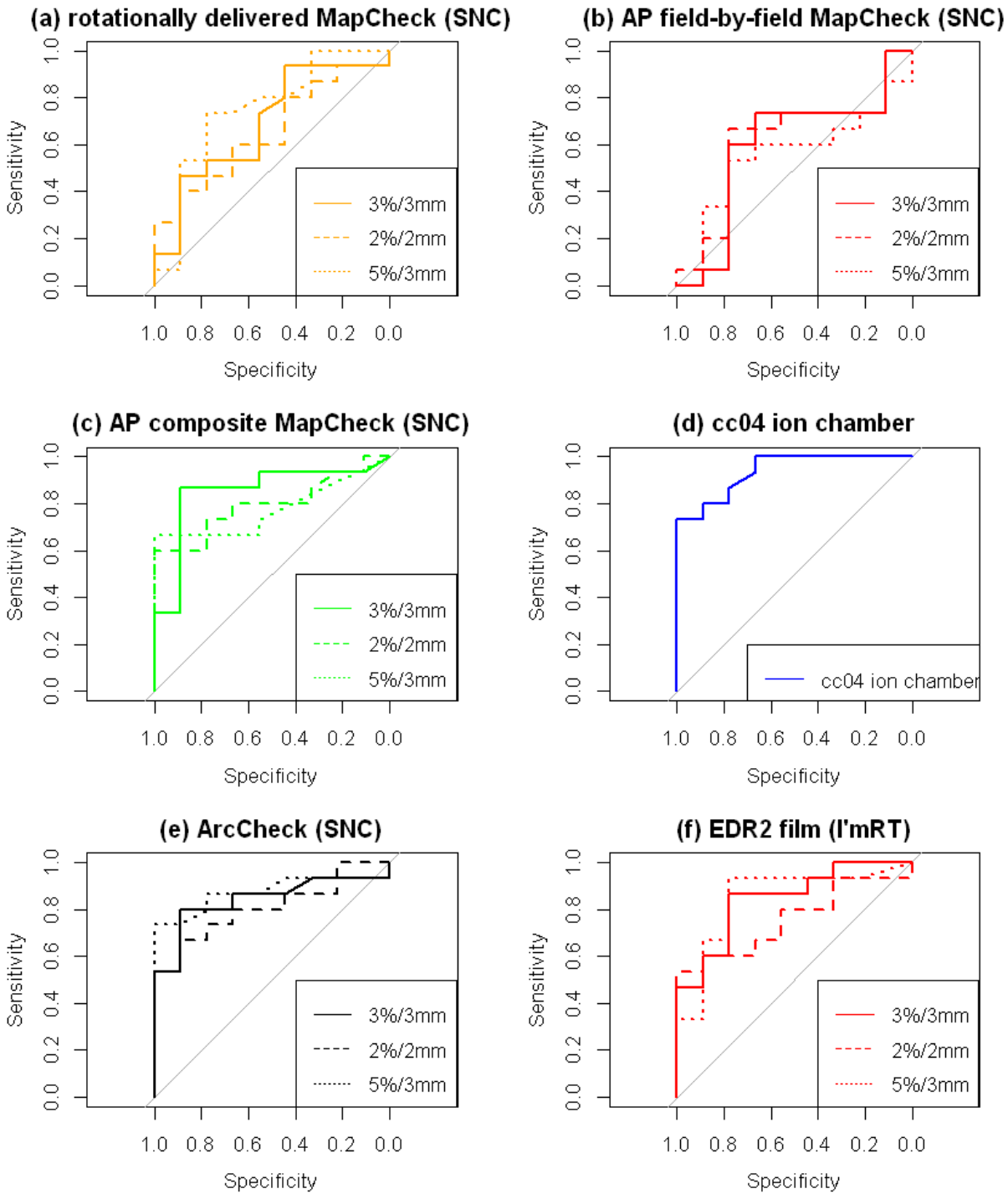



**Figure 16 ROC curves generated for each analysis, grouped by dosimetric system. For each planar dosimeter, each panel contains an ROC curve for 2%/2mm, 3%/3mm, and 5%/3mm as the criteria for the gamma analysis. For this figure, all MapCheck gamma analysis was performed using SNC Patient.**

Each panel in Figure 16 (except the single ion chamber) contains multiple curves, one each for 2%/2mm, 3%/3mm, and 5%/3mm. These three different curves are generated by varying the cutoff criteria (percent of pixels passing for the gamma analysis) from very liberal to very conservative such that the curve begins at the bottom left and ends at the top right, respectively. Each curve is formed from a different range of percent of pixels passing (or percent difference in the case of the ion chamber). A D-test performed with 2000 replicates bootstrapped to the data in the pROC package compared the planar measurements at 2%/2mm, 3%/3mm, and 5%/3mm for each device. Due to the fact that several one-to-one comparisons were performed, there is a probability that one would obtain statistical significance by chance. To correct for this, a False Discovery Rate (FDR) correction was applied to the data. However, even without the correction, none of the devices evaluated showed significant differences (alpha = 0.05) in their AUC among the three dose difference and distance to agreement criteria. This suggests that different gamma criteria can be used at various cutoff values to obtain similar discrimination ability. Clinically, there may still be a practical reason to have a preference between these thresholds. For example, for looser criteria (e.g., 5%/3mm), in order to obtain a similar sensitivity and specificity, the cutoff value may have to be set impractically high (i.e., more than 99% of pixels passing).

Nine additional MapCheck curves were created for gamma analyses conducted in DoseLab Pro. Figure 17 shows the comparison between SNC Patient analysis and DoseLab Pro analysis with the 3%/3mm criteria (other criteria not shown). For all criteria, a D-test was performed with 2000 replicates bootstrapped to the data. After application of a false discovery rate correction, there were no significant differences between SNC Patient and

DoseLab Pro analysis in terms of AUC for any of the evaluated devices. However, the curves are clearly not superimposed, which is the result of variations in the two software packages, such as different implementations of measured and calculated plane alignment, methods of dose thresholding, and ROI selection. The choice in details of implementing the gamma analysis can lead to different results (Ezzell 2009).

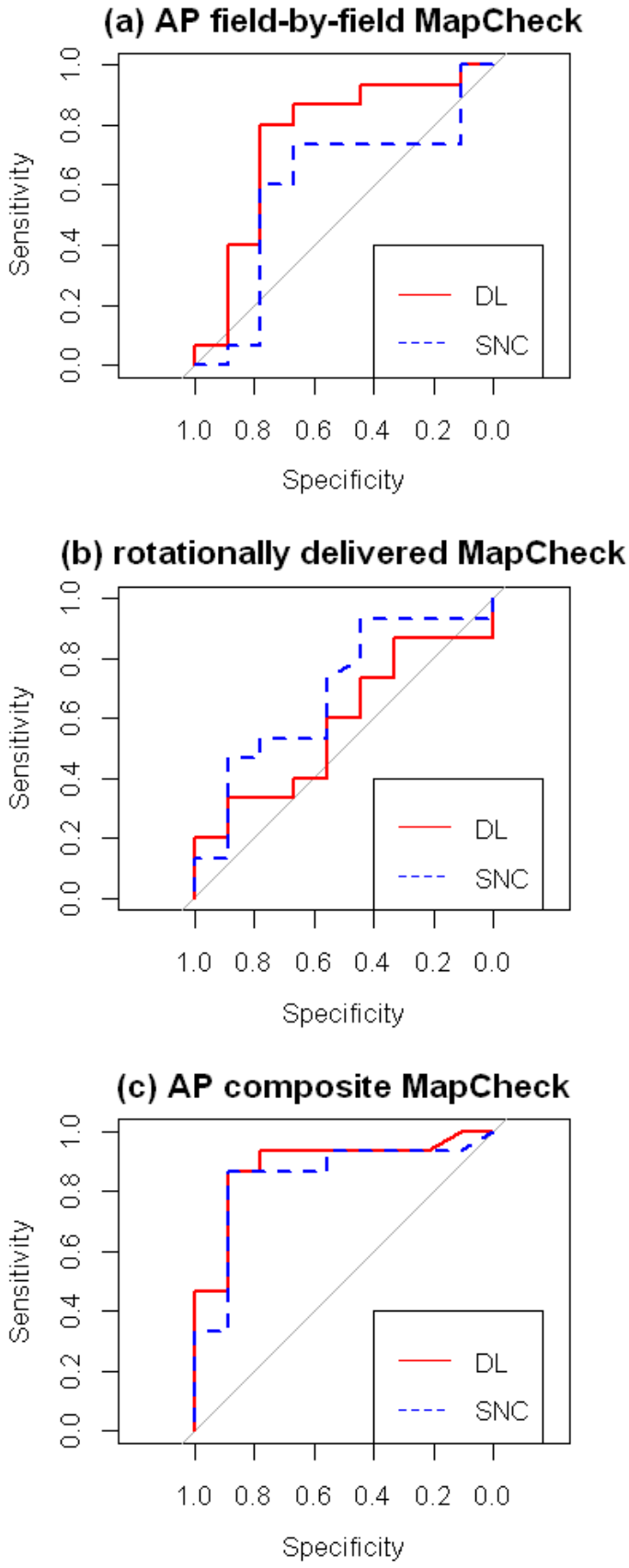



**Figure 17 Comparing gamma calculations between DoseLabPro (solid red line) and SNC Patient (dashed blue line) for ROC curves created from the MapCheck measurements analyzed at 3%/3mm.**

The AUC summarizes the overall ability of a QA system to accurately identify acceptable and unacceptable plans across all dosimeter criteria and cutoff values. The maximum value the AUC can assume is 1 (perfect classification of plans), whereas an AUC of 0.5 is equivalent to a random guess. The best-performing QA system was the single ion chamber with an area of 0.94 (0.82, 1.0), followed by the composite MapCheck at 5%/3mm with all fields delivered AP: 0.93 (0.80, 1.0). The worst performer was the MapCheck field by field at 5%/3mm with all beams delivered AP: 0.55 (0.31, 0.79). All AUC's with confidence intervals are shown in Table 3 in order of decreasing AUC.

**Table 3 Areas under the curves for all dosimetric systems and analysis techniques, with accompanying bootstrapped 95% confidence intervals**

| IMRT QA Method | AUC | C.I. |
|---|---|---|
| **cc04 ion chamber** | 0.94 | ( 0.82 - 1.00 ) |
| **AP composite MapCheck at 5%/3mm (DL)** | 0.93 | ( 0.80 - 1.00 ) |
| **AP composite MapCheck at 3%/3mm (DL)** | 0.89 | ( 0.73 - 1.00 ) |
| **ArcCheck at 5%/3mm (SNC)** | 0.87 | ( 0.71 - 0.99 ) |
| **AP composite MapCheck at 2%/2mm (DL)** | 0.85 | ( 0.67 - 0.99 ) |
| **AP composite MapCheck at 3%/3mm (SNC)** | 0.85 | ( 0.66 - 1.00 ) |
| **EDR2 film at 5%/3mm** | 0.84 | ( 0.66 - 1.00 ) |
| **EDR2 film at 3%/3mm** | 0.84 | ( 0.66 - 0.97 ) |
| **ArcCheck at 3%/3mm (SNC)** | 0.84 | ( 0.67 - 0.99 ) |
| **ArcCheck at 2%/2mm (SNC)** | 0.81 | ( 0.61 - 0.95 ) |
| **AP field-by-field MapCheck at 2%/2mm (DL)** | 0.80 | ( 0.61 - 0.98 ) |
| **AP composite MapCheck at 2%/2mm (SNC)** | 0.80 | ( 0.60 - 0.95 ) |
| **AP composite MapCheck at 5%/3mm (SNC)** | 0.78 | ( 0.57 - 0.92 ) |
| **AP field-by-field MapCheck at 3%/3mm (DL)** | 0.76 | ( 0.51 - 0.97 ) |
| **EDR2 film at 2%/2mm** | 0.76 | ( 0.55 - 0.93 ) |
| **Rotationally delivered MapCheck at 5%/3mm (SNC)** | 0.75 | ( 0.51 - 0.94 ) |
| **Rotationally delivered MapCheck at 3%/3mm (SNC)** | 0.69 | ( 0.44 - 0.89 ) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Rotationally delivered MapCheck at 5%/3mm (DL)** | 0.67 | ( | 0.44 | - | 0.89 | ) |
| **AP field-by-field MapCheck at 5%/3mm (DL)** | 0.65 | ( | 0.38 | - | 0.90 | ) |
| **Rotationally delivered MapCheck at 2%/2mm (SNC)** | 0.65 | ( | 0.41 | - | 0.85 | ) |
| **AP field-by-field MapCheck at 2%/2mm (SNC)** | 0.61 | ( | 0.36 | - | 0.85 | ) |
| **Rotationally delivered MapCheck at 2%/2mm (DL)** | 0.59 | ( | 0.35 | - | 0.83 | ) |
| **AP field-by-field MapCheck at 3%/3mm (SNC)** | 0.59 | ( | 0.35 | - | 0.84 | ) |
| **Rotationally delivered MapCheck at 3%/3mm (DL)** | 0.58 | ( | 0.33 | - | 0.81 | ) |
| **AP field-by-field MapCheck at 5%/3mm (SNC)** | 0.55 | ( | 0.31 | - | 0.79 | ) |

We next compared the capabilities of the IMRT QA systems independent of their data analysis. That is, because there was a lack of significant differences between criteria or analysis software, the AUCs were grouped. All of the AUCs were placed into one of six groups: cc04 ion chamber, AP composite MapCheck, ArcCheck, EDR2 film, AP field-by-field MapCheck, and rotationally (original gantry angle) delivered MapCheck. An ANOVA was performed to look for differences between these groups, including a post hoc analysis using Tukey's Honestly Significant Difference test. The ANOVA analysis found that at least one group was significantly different ($p = 0.0001$), whereas the HSD test was able to group devices at the alpha = 0.05 significance level. The better-performing group contained the cc04 ion chamber, AP composite MapCheck, ArcCheck, and EDR2 film, whereas the AP field-by-field and rotationally delivered MapCheck were in the poorer-performing group. The means of each group are shown in Table 4, with the thick line showing the divide between the two groups. The AUC means in the higher group ranged from 0.815 to 0.937, whereas the poorer group ranged from 0.654 to 0.661.

**Table 4 Average AUC for each device, irrespective of analysis method. The thick line indicates where the devices were significantly grouped based on AUC performance.**

| QA System | Average AUC across all analysis systems |
|---|---|
| **cc04 ion chamber** | 0.94 |
| **AP composite MapCheck** | 0.85 |
| **ArcCheck** | 0.84 |
| **EDR2 film** | 0.82 |
| **AP field-by-field MapCheck** | 0.66 |
| **Rotationally delivered MapCheck** | 0.65 |

Cutoff criteria in the clinic (e.g., a 3% ion chamber criterion) can be based on what has emerged as traditional practice. However, by evaluating the ROC curves, mathematically optimal criteria can be determined. For example, a percent of pixels passing threshold can be selected to provide the optimal sensitivity and specificity for a device at a 3%/3mm criteria. The optimal cutoffs as calculated from the Youden Index for gamma analysis across all planar systems at 2%/2mm ranged from 68.1% to 89.6%; at 3%/3mm it ranged from 85% to 98.1%; and at 5%/3mm it ranged from 96.3% to 99.8%; these findings demonstrated a general trend that a looser gamma criteria requires a more stringent cutoff (and vice versa). The optimal cutoff for each system investigated is shown in Table 5, with 95% confidence intervals, each calculated from 2000 bootstrapped replicates. Some systems, in conjunction with loose gamma criteria, can have clinically unreasonably high "optimal" thresholds. For example, the AP field-by-field MapCheck at 5%/3mm (SNC) and the AP composite MapCheck at 5%/3mm (DL) had optimal cutoffs of 98.7% and 99.7%, respectively. For the AP composite MapCheck at 5%/3mm, three quarters of the plans measured had 99% of pixels passing or higher, leading to an ROC curve that was generated from clinically unreasonable high cutoffs. This will be generally true for liberal dose difference and distance to agreement criteria. Therefore, the performance of the ROC curves and calculated optimal cutoffs must be tempered by clinical realities. When the ROC curve is formed from plans with a higher percent of pixels passing (or lower percent dose difference in the case of the ion chamber), the optimal cutoff value will be generally higher.

**Table 5 Optimal cutoffs given for all dosimetric systems, both with and without weighting by the prevalence of a failing plan and the cost of falsely labeling a failing plan as passing. The calculation of confidence intervals was based on a normal approximation, so there is an opportunity to exceed parameter space.**

| | Youden Index | Empirical cutoffs: Prevalence is 3%, and cost of FN is 0.06 times the cost of FP |
|---|---|---|
| Device | Threshold (no weighting) | Threshold (with weighting) |
| cc04 ion chamber | 1.6 ± 1.1 | 3.0 ± 0.7 |
| **AP composite MapCheck at 5%/3mm (DL)** | 99.7 ± 0.4 | 96.4 ± 2.5 |
| **AP composite MapCheck at 3%/3mm (DL)** | 98.1 ± 2.4 | 81.5 ± 12.7 |
| ArcCheck at 5%/3mm (SNC) | 96.3 ± 1.7 | 92.1 ± 6.1 |
| **AP composite MapCheck at 2%/2mm (DL)** | 89.2 ± 6.6 | 61.6 ± 18.9 |
| AP composite MapCheck at 3%/3mm (SNC) | 97.7 ± 2.0 | 82.5 ± 13.1 |
| EDR2 film at 3%/3mm | 97.0 ± 9.7 | 76.3 ± 8.2 |
| EDR2 film at 5%/3mm | 99.8 ± 1.5 | 91.2 ± 5.7 |
| ArcCheck at 3%/3mm (SNC) | 92.0 ± 7.1 | 69.2 ± 14.0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| ArcCheck at 2%/2mm (SNC) | 74.4 | ± | 14.2 | 49.2 | ± | 11.4 |
| **AP field-by-field MapCheck at 2%/2mm (DL)** | 89.6 | ± | 3.7 | 77.8 | ± | 6.0 |
| AP composite MapCheck at 2%/2mm (SNC) | 82.4 | ± | 9.4 | 66.1 | ± | 10.2 |
| AP composite MapCheck at 5%/3mm (SNC) | 99.6 | ± | 0.4 | 98.5 | ± | 0.7 |
| **AP field-by-field MapCheck at 3%/3mm (DL)** | 98.0 | ± | 2.6 | 92.0 | ± | 4.4 |
| EDR2 film at 2%/2mm | 68.1 | ± | 15.0 | 59.8 | ± | 6.4 |
| Rotationally delivered MapCheck at 5%/3mm (SNC) | 98.5 | ± | 7.2 | 82.7 | ± | 11.9 |
| Rotationally delivered MapCheck at 3%/3mm (SNC) | 94.9 | ± | 9.5 | 70.1 | ± | 12.8 |
| **Rotationally delivered MapCheck at 5%/3mm (DL)** | 98.0 | ± | 5.0 | 84.8 | ± | 9.3 |
| **AP field-by-field MapCheck at 5%/3mm (DL)** | 99.4 | ± | 1.1 | 97.8 | ± | 1.6 |
| Rotationally delivered MapCheck at 2%/2mm (SNC) | 69.2 | ± | 18.5 | 53.7 | ± | 13.2 |
| AP field-by-field MapCheck at 2%/2mm (SNC) | 85.0 | ± | 7.7 | 71.7 | ± | 16.5 |
| AP field-by-field MapCheck at | 96.0 | ± | 4.6 | 89.6 | ± | 7.2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3%/3mm (SNC) | | | | | | |
| **Rotationally delivered MapCheck at 2%/2mm (DL)** | 79.6 | ± | 20.9 | 48.2 | ± | 23.2 |
| **Rotationally delivered MapCheck at 3%/3mm (DL)** | 85.0 | ± | 14.4 | 68.6 | ± | 15.8 |
| AP field-by-field MapCheck at 5%/3mm (SNC) | 98.7 | ± | 1.9 | 96.3 | ± | 2.4 |

The cc04 ion chamber was calculated to have an optimal cutoff of 1.6% via the Youden Index. This is considerably lower than the 3% threshold commonly used in the clinic and would perhaps cause an intolerably high false-positive rate (i.e., failing acceptable plans). However, this value was obtained without consideration of prevalence of unacceptable plans or cost of a false positive. These unrealistic assumptions lead to optimal cutoffs that are considerably more stringent than is expected in the clinic. Part of this is because the actual prevalence of unacceptable plans is fortunately much less than the unweighted prevalence of 50%. Different cutoffs may be obtained by varying these weighting factors. When the cost was manipulated to achieve an optimal threshold of 3% for the ion chamber, this resulted in a cost of passing an unacceptable plan as 0.06 times (about 1/16) the cost of failing a truly acceptable plan. That is, in order to calculate an optimal cutoff of 3% for an ion chamber, this dosimeter must be heavily weighted to preferentially pass both acceptable and unacceptable plans. This is in surprisingly direct opposition to reasonable clinical goals, which would be to err on the side of caution and preferentially fail plans to ensure no unacceptable plans are passed. This same weighting (0.06) was used for all other devices to create the percent of pixels passing criteria that was equivalent (in weighting) to the 3% criteria for the ion chamber (Table 5). The weighted thresholds show lower percent of pixels passing values than the unweighted thresholds. The amount that the thresholds decreased varied among devices. Some showed substantial lowering, whereas others changed only modestly. This is consistent with the fairly large error bars on these thresholds.

## 4.4 DISCUSSION

This research showed that not all of the IMRT QA systems analyzed in this work can equally differentiate between acceptable and unacceptable patient plans. This could be a reflection of the differing measurement geometries, resolution of the measurements, and implementations of the data analyses. In fact, none of the devices sorted the plans in the exact same manner as the gold standard. Despite this, the various QA systems were able to be divided into two groups with significantly different abilities to accurately classify plans. The better performers included the cc04 ion chamber, AP composite MapCheck, radiographic film, and ArcCheck, whereas the field-by-field and rotational MapCheck performed poorer than this group.

The AUC averages in the better-performing group ranged from 0.815 to 0.937. A guideline for assigning a qualitative assessment to the AUC values states that $0.5<AUC\leq0.7$ is "less accurate," $0.7<AUC\leq0.9$ is "moderately accurate," and $0.9<AUC<1$ is "highly accurate" (Greiner 2000). The better-performing group is therefore moderately to highly accurate, whereas the poorer-performing group (ranging from 0.654 to 0.661) would qualify as "less accurate." The single ion chamber showed the highest individual AUC but still showed discrepancies from the gold standard due to differences in the measurement geometry (single versus multiple measurement points). If the local dose around the single ion chamber matches the TPS plan yet deviates in other locations, even the most stringent passing criteria might not be able to label the plan as failing. Task Group 120 (Low 2011)

discussed how an ideal IMRT dosimeter would be able to truly sample the plan three dimensionally; however, such dosimeters have not yet been proven clinically viable.

Some insight is also available on the less well performing QA techniques. The field-by-field MapCheck is particularly interesting because it showed such different abilities to correctly sort plans compared with the composite MapCheck, despite being derived from the same measurement data. The differences in their ability to classify plans stem entirely from the method of analysis. When AP-delivered beams were analyzed field-by-field on the diode array, most fields scored well on a gamma analysis for both failing and passing plans. However, when summed into a composite plane, the small errors compound, leading to a more sensitive test. Publications by Nelms (Nelms 2011) and Kruse (Kruse 2010) have demonstrated some of the shortcomings of field-by-field dosimetry, notably an inability to distinguish between clinically acceptable and unacceptable plans on the basis of percent of pixels passing. Nelms offers the explanation that hot and cold spots may appear per field without deviating enough to be detected using a gamma analysis; however, when summed together these deviations may lead to critical dose errors. Despite these results, a survey (Nelms 2007) showed that 64.1% of clinics use AP field-by-field measurements when performing MapCheck-based IMRT QA, whereas 32.8% use AP composite methods most of the time. Therefore, the question of field-by-field sensitivity is relevant to today's QA practices. Composite diode array dosimetry can also be performed with the original rotational gantry angles, as was done in this study with the MapPhan phantom. However, due to the directional dependence of diodes, such an array will perform better when the beam is perpendicular to array surface (Low 2011). The manufacturer of the diode array cautions that non-normal incidence can lead to errors due to a corruption of 2D information

(the array appears 1D to the beams eye view) and the air cavities perturbing the fluence (SunNuclear 2007). This issue of directional dependence is a possible explanation for the relatively poorer performance of the MapCheck when all beams were delivered at their original gantry angles.

Table 5 shows the optimal thresholds for each device and analysis technique examined in this study. These values establish thresholds to be used for IMRT QA that are founded on the fact that in the clinic, a 3% dose difference threshold is often used for ion chamber–based IMRT plan verification (Ezzell 2003). However, their clinical appropriateness can be questioned because this weighting indicates that the cost of misclassifying an unacceptable plan as acceptable is 1/16 (0.06 times) that of misclassifying an acceptable plan as unacceptable. While this seems like a counterintuitive weighting, the 3% ion chamber threshold was not devised with this cost weighting in mind. This seems to suggest a priority of efficiency in the clinic. Use of this weighting for the planar dosimeters revealed thresholds that are generally consistent with clinical experience. At a 3%/3mm criteria, 90% of pixels passing was often within the confidence interval of the optimal threshold. Some QA methods (such as the ArcCheck at 3%/3mm) showed a weighted threshold that was well below 90% of pixels passing. If a clinic used 90% as its threshold, that could be interpreted as more preferentially weighting failing an unacceptable plan. This is clinically reasonable, and therefore a clinical threshold above the weighted value (or below in the case of dose difference for the ion chamber) in Table 5 is likely a clinically sound decision, whereas a threshold below (or above for ion chamber dose difference) the weighted value is more representative of a liberal cutoff that excessively passes plans, including unacceptable ones.

Future work may be able to expand upon this research by more precisely determining AUC and optimal cutoffs from an expanded set of patient plans with an even greater variety of treatment sites. More patient plan measurements may also lead to tighter confidence intervals. Compared with the wide range of devices and analysis techniques used by the physics community, this work has only measured a small subset of IMRT QA methods. However, the techniques described above could be used to study other methods and determine a clinically relevant cutoff threshold for any particular IMRT QA dosimeter and analysis technique. This could be done to meet the sensitivity, specificity, and financial cost needs of the clinic. Importantly, the results of the IMRT QA analysis act as a guideline for detecting issues with an IMRT plan; it is up to the scrutiny of the clinical team to apply good judgment in determining the acceptability of a plan before treatment.

## 4.5 CONCLUSION

Several commercial patient-specific IMRT QA dosimeters and methods were investigated for their ability to discriminate between acceptable and unacceptable plans on a set of clinical patient plans. An ROC analysis was applied to track the performance of the various methods as a function of the cut-off values (%dose difference for point measurements, % pixels passing for planar measurements). ROC analysis was also used to determine the optimal cut-off values for the various methods being investigated, including when weighted for different costs for falsely failing an acceptable plan versus falsely passing an unacceptable plan.

To compare the methods, the areas under the ROC curve were calculated, revealing that different devices performed significantly poorer or better than others. When averaging

all analysis techniques for each QA method, the ion chamber, AP composite MapCheck, ArcCheck, and radiographic film performed well, whereas the AP field-by-field and rotationally delivered MapCheck performed poorer than this group.

The classification abilities for each device at 2%/2mm, 3%/3mm, and 5%/3mm gamma criteria did not produce statistically different results. That is, a similar level of accuracy in sorting acceptable and unacceptable plans could be expected at different criteria. Naturally, at these different criteria, a different percent of pixels passing would be necessary. For example, at the more liberal 5%/3mm, a very high cutoff would be needed to have an adequate sensitivity.

Without weighting, the optimal cutoff for the ion chamber was a 1.6% dose difference. Although this is a stringent QA threshold, it is calculated with the assumption that half the IMRT plans undergoing QA are failing (without weighting). The QA devices using a gamma analysis show wider confidence intervals for the 2%/2mm criteria, reflecting the nature of the data bunching up against the 100% of pixels passing value with the looser criteria. With a failing plan prevalence of 0.03 and a clinically reasonable cost weighting that placed the ion chamber at an optimal threshold of 3%, the optimal thresholds of the devices migrated to lower sensitivities, indicating that the weighting necessary for a 3% ion chamber cutoff shows a priority towards lowering the incidence of falsely labeling a plan as failing. These values are available to use, but with a cost-benefit analysis balancing the cost of falsely detecting an unacceptable or acceptable plan, an optimal cutoff could be tailored for an individual clinic's needs.

This work shows that depending on the QA system being used, different considerations need to be made. The same cutoff criteria do not yield the same classification

abilities across all devices. Also, this work has shown that QA systems have different abilities to accurately sort acceptable and unacceptable plans. This information can help guide clinics to making more informed decisions when considering how and which patient-specific IMRT QA devices to use in the detection of plan errors.

# II. Chapter 5

## OVERALL DISCUSSION AND CONCLUSIONS

The original goal of this body of research was to compare the sensitivities among several clinical IMRT QA dosimeters, as detailed in the hypothesis. However, in approaching that narrow problem it was found that many more extensive questions could be answered with this data, notably how well a dosimeter system could classify acceptable and unacceptable plans, recommendations for QA cutoff values, the effect of choice of analysis on the QA performance, the reproducibility in patient-specific IMRT QA measurements, and the relationship between dose gradient across the ion chamber and the reproducibility of an ion chamber measurement. All of these topics address current clinical practices and assumptions made in patient-specific IMRT QA, and it is the hope of this author that the body of research outlined in this thesis will provide guidance in clinical decisions, and give the researcher the analysis tools to pursue further study in this area of research.

The hypothesis that none of the QA devices investigated would have sensitivities over 0.90 at 3%/3mm and 90% of pixels passing for planar, and 3% dose difference for point dosimeters, was trivially proven true in the course of this research. In fact, at these conventional passing criteria, the devices with the highest sensitivity were the film and ArcCheck, tied at 0.60. This means that only 60% of the unacceptable plans (as determined by the gold standard) were detected as failing. Although the hypothesis focused on the sensitivities of a dosimeter, it is also important to consider the ability to correctly label acceptable plans as passing (specificity); otherwise the dosimeter may give an excessive

number of false alarms. The ROC analysis allowed this research to balance investigating sensitivity, specificity, and appropriate cutoff values for each dosimeter system under the varying selections of analysis software and criteria. This has considerably expanded the scope of the clinical relevancy of this project.

When comparing the performance of different dosimetric systems, the reproducibility of such measurements is a natural question that arises. While the bulk of this was answered in Chapter 3 while discussing reproducibility, the performance comparisons study reveals another facet of this question. When faced with an ion chamber-based plan failure in IMRT QA, a strategy is to take the measurement again at another point. While there is a chance that the original point measurement failed due to poor measurement geometry, Kruse's research points out that dosimetric accuracy is not necessarily the same throughout a patient plan (Kruse 2010). The multiple ion chamber phantom measurements seem to verify this assertion. It took a three-dimensional sampling of the patient plans investigated in this research, and showed how variation in dose difference between measurement and calculation can occur throughout an IMRT plan. Therefore, when retaking a measurement, the results of the IMRT QA may vary due to detector reproducibility and due to a choice of different measurement location.

Finally, the paramount question is, how do these results affect the treatment that the patient ultimately receives? A study in the literature (Dische 1993) has shown how a ±5% dose difference can lead to loss in tumor control or heighten the risk to normal tissue. Dong et al found that of 751 IMRT plans investigated, 3.1% showed a dose difference of greater than 3.5%. Although this shows a low incidence, it also shows that the incidence of unacceptable plans is nonzero. IMRT involves a complex chain of steps to get from

simulation and planning to delivery. Patient-specific QA offers a safeguard between potential mistakes along that chain, and the dosimetric accuracy of the patient's IMRT treatment. It provides a preemptive, quantitative test to give guidance on the acceptability of an IMRT plan. Therefore, it is important to understand the abilities and limitations of the quantitative test results of dosimeters used in QA, both in terms of accurately sorting acceptable and unacceptable plans and in their reproducibility. Studies have been done to evaluate the sensitivity and reproducibility of detector systems, but this work focuses on the end result seen in the clinic and used for guidance: notably the percent of pixels passing for planar measurements and dose difference for point measurements. While it is ultimately up to the clinician's knowledge and expertise on whether the plan is acceptable, these QA tests provide a valuable metric in the aid of that decision-making.

Future Work

The methods described in this thesis could provide a framework for aiding in the optimization of a patient-specific IMRT QA program. The ROC analysis gives researchers the ability to assess their dosimeters' abilities, and with a clinic-specific cost-benefit analysis, arrive at a quantitatively derived cutoff value. An expansion on this work may be to produce a training set of clinical plans that have already been sorted with a gold standard, and send them out to centers that may wish to determine the performance of any dosimetric system that may not have been addressed in this work.

While the error bars for the AUC's were generated using bootstrapping, this analysis could be expanded by introducing replicates for the generation of ROC curves. Also, the

reproducibility study only had three repeat measurements and looked at six patient plans. Additional repeated measurements and patient plans may expand that investigation.

## III. **References**

Abdi, H. and L. J. Williams (2010). "Tukey's Honestly Significant Difference (HSD) test." Encyclopedia of Research Design. Thousand Oaks, CA: Sage: 1-5.

Bogner, L., J. Scherer, M. Treutwein, M. Hartmann, F. Gum and A. Amediek (2004). "Verification of IMRT: techniques and problems." Strahlentherapie und Onkologie **180**(6): 340-350.

Bortfeld, T. (2006). "IMRT: a review and preview." Physics in Medicine and Biology **51**(13): R363-379.

Carlone, M., C. Cruje, A. Rangel, R. McCabe, M. Nielsen and M. Macpherson (2013). "ROC analysis in patient specific quality assurance." Medical Physics **40**(4): 042103.

Carpenter, J. and J. Bithell (2000). "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians." Statistics in Medicine **19**(9): 1141-1164.

Childress, N. L., M. Salehpour, L. Dong, C. Bloch, R. A. White and Rosen, II (2005). "Dosimetric accuracy of Kodak EDR2 film for IMRT verifications." Medical Physics **32**(2): 539-548.

CNMC. (2009). "Wellhofer Compact Ionization Chambers." from http://www.cnmcco.com/radPhysics/PDFdocs/thimble/CNMC_Wellhofer_compact.pdf.

Das, I. J., C.-W. Cheng, K. L. Chopra, R. K. Mitra, S. P. Srivastava and E. Glatstein (2008). "Intensity-Modulated Radiation Therapy Dose Prescription, Recording, and Delivery: Patterns of Variability Among Institutions and Treatment Planning Systems." Journal of the National Cancer Institute **100**(5): 300-307.

DeLong, E. R., D. M. DeLong and D. L. Clarke-Pearson (1988). "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." Biometrics **44**(3): 837-845.

Dische, S., M. I. Saunders, C. Williams, A. Hopkins and E. Aird (1993). "Precision in reporting the dose given in a course of radiotherapy." Radiotherapy and Oncology **29**(3): 287-293.

Dong, L., J. Antolak, M. Salehpour, K. Forster, L. O'Neill, R. Kendall and I. Rosen (2003). "Patient-specific point dose measurement for IMRT monitor unit verification." International Journal of Radiation Oncology, Biology, Physics **56**(3): 867-877.

Ezzell, G. A., 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 (2009). "IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119." Medical Physics **36**(11): 5359-5373.

Ezzell, G. A., J. M. Galvin, D. Low, J. R. Palta, I. Rosen, M. B. Sharpe, P. Xia, Y. Xiao, L. Xing and C. X. Yu (2003). "Guidance document on delivery, treatment planning, and clinical implementation of IMRT: report of the IMRT Subcommittee of the AAPM Radiation Therapy Committee." Medical Physics **30**(8): 2089-2115.

Fraser, D., W. Parker and J. Seuntjens (2009). "Characterization of cylindrical ionization chambers for patient specific IMRT QA." Journal of Applied Clinical Medical Physics **10**(4): 2923.

Gordon, J. and J. Siebers (2013). "Addressing a Gap in Current IMRT Quality Assurance." International Journal of Radiation Oncology, Biology, Physics.

Greiner, M., D. Pfeiffer and R. D. Smith (2000). "Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests." Preventive Veterinary Medicine **45**(1-2): 23-41.

Hall, E. J. and A. J. Giaccia (2006). Radiobiology for the radiologist. Philadelphia, Lippincott Williams & Wilkins.

Hanley, J. A. (1988). "The robustness of the "binormal" assumptions used in fitting ROC curves." Medical Decision Making **8**(3): 197-203.

Hartford, A. C., M. G. Palisca, T. J. Eichler, D. C. Beyer, V. R. Devineni, G. S. Ibbott, B. Kavanagh, J. S. Kent, S. A. Rosenthal, C. J. Schultz, P. Tripuraneni and L. E. Gaspar (2009). "American Society for Therapeutic Radiology and Oncology (ASTRO) and American College of Radiology (ACR) Practice Guidelines for Intensity-Modulated Radiation Therapy (IMRT)." International Journal of Radiation Oncology*Biology*Physics **73**(1): 9-14.

Hastie, T., R. Tibshirani and J. H. Friedman (2009). The elements of statistical learning : data mining, inference, and prediction. New York, NY, Springer.

Howell, R. M., I. P. Smith and C. S. Jarrio (2008). "Establishing action levels for EPID-based QA for IMRT." Journal of Applied Clinical Medical Physics **9**(3): 2721.

IAEA (2008). Transition from 2-D Radiotherapy to 3-D Conformal and Intensity Modulated Radiotherapy. Vienna, Austria, Internation Atomic Energy Agency.

Khan, F. M. (2010). The physics of radiation therapy. Philadelphia, Lippincott Williams & Wilkins.

Kodak, H. I. (2001). "KODAK Oncology Imaging Guide."

Kruse, J. J. (2010). "On the insensitivity of single field planar dosimetry to IMRT inaccuracies." Medical Physics **37**(6): 2516-2524.

Low, D. A. and J. F. Dempsey (2003). "Evaluation of the gamma dose distribution comparison method." Medical Physics **30**(9): 2455-2464.

Low, D. A., W. B. Harms, S. Mutic and J. A. Purdy (1998). "A technique for the quantitative evaluation of dose distributions." Medical Physics **25**(5): 656-661.

Low, D. A., J. M. Moran, J. F. Dempsey, L. Dong and M. Oldham (2011). "Dosimetry tools and techniques for IMRT." Medical Physics **38**(3): 1313-1338.

Mancuso, G. M., J. D. Fontenot, J. P. Gibbons and B. C. Parker (2012). "Comparison of action levels for patient-specific quality assurance of intensity modulated radiation therapy and volumetric modulated arc therapy treatments." Medical Physics **39**(7): 4378-4385.

McEwen, M. R. (2010). "Measurement of ionization chamber absorbed dose k(Q) factors in megavoltage photon beams." Medical Physics **37**(5): 2179-2193.

Mell, L. K., A. K. Mehrotra and A. J. Mundt (2005). "Intensity-modulated radiation therapy use in the U.S., 2004." Cancer **104**(6): 1296-1303.

Metz, C. E. (1978). "Basic principles of ROC analysis." Seminars in Nuclear Medicine **8**(4): 283-298.

Nelms, B. E. and J. A. Simon (2007). "A survey on planar IMRT QA analysis." Journal of Applied Clinical Medical Physics **8**(3): 2448.

Nelms, B. E., H. Zhen and W. A. Tome (2011). "Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors." Medical Physics **38**(2): 1037-1044.

Perkins, N. J. and E. F. Schisterman (2006). "The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve." American Journal of Epidemiology **163**(7): 670-675.

Pulliam, K. B., R. M. Howell, D. Followill, D. Luo, R. A. White and S. F. Kry (2011). "The clinical impact of the couch top and rails on IMRT and arc therapy." Physics in Medicine and Biology **56**(23): 7435-7447.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez and M. Muller (2011). "pROC: an open-source package for R and S plus to analyze and compare ROC curves." Bmc Bioinformatics **12**.

Rosner, B. (2011). Fundamentals of biostatistics. Boston, Brooks/Cole, Cengage Learning.

Ruppert, D., M. P. Wand and R. J. Carroll (2003). Semiparametric regression. Cambridge ; New York, Cambridge University Press.

Sing, T., O. Sander, N. Beerenwinkel and T. Lengauer (2005). "ROCR: visualizing classifier performance in R." Bioinformatics **21**(20): 3940-3941.

Skaltsa, K., L. Jover and J. L. Carrasco (2010). "Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty." Biomedical Journal **52**(5): 676-697.

SunNuclear (2007). MapCheck for Rotational Dosimetry.

SunNuclear. (2010). "MapPhan: Rotational Dosimetry Delivered." Retrieved June 29, 2013, 2013, from http://www.sunnuclear.com/documents/mapphan.pdf.

Veldeman, L., I. Madani, F. Hulstaert, G. De Meerleer, M. Mareel and W. De Neve (2008). "Evidence behind use of intensity-modulated radiotherapy: a systematic review of comparative clinical studies." The Lancet Oncology **9**(4): 367-375.

Wilcox, E. E., G. M. Daskalov, G. Pavlonnis, 3rd, R. Shumway, B. Kaplan and E. VanRooy (2008). "Dosimetric verification of intensity modulated radiation therapy of 172 patients treated for various disease sites: comparison of EBT film dosimetry, ion chamber measurements, and independent MU calculations." Medical Dosimetry **33**(4): 303-309.

# IV. APPENDIX I

## RECEIVER OPERATOR CHARACTERISTIC CURVE ANALYSIS

In order to grant the reader a more in-depth explanation of a core methodology explored in this thesis, this appendix endeavors to better explain the nuances and mechanics of the receiver operator characteristic (ROC) curve. The ROC curve is a useful depiction of the performance capabilities of a diagnostic test with dichotomous results, that is, either a positive or negative outcome. The appropriate test will have many continuous measurements requiring a cutoff value to separate the binary results. The ROC curve incorporates the concepts of sensitivity and specificity, concatenating these concepts by plotting their changing values as the cutoff is varied from liberal to stringent. Traditionally, the vertical axis is sensitivity, while the horizontal axis is 1 – specificity, such that the ideal curve reaches to the upper left corner. An example ROC curve was shown in Figure 13.

Sensitivity is the ability to label a truly positive result as positive, while specificity is the ability to label a truly negative result as negative. Using the notation of true positive (TP), true negative (TN), false positive (FP), and false negative (FN), the equations for these concepts are shown below.

$$\boldsymbol{sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad \textbf{Equation 7}$$

$$\boldsymbol{specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}} \qquad \textbf{Equation 8}$$

This introduces the need to establish the "truth" by which the test under investigation will be compared. A gold standard is used to determine the veracious classification of each measurement. A pool of several measurements is needed to establish reliable values of sensitivity and specificity for each cutoff. As discussed in the body of this thesis, this work measured 24 patient plans (the pool of measurements), on six different IMRT QA dosimeter techniques (the tests being investigated), and compared the classification of those 24 plans by each dosimeter to the results given by the multiple ion chamber phantom measurements (gold standard). For all plans except the cc04 ion chamber, the gamma index was used, giving a cutoff value in terms of percent of pixels passing. The cc04 ion chamber used percent difference between the plan and the measurement as its cutoff.

Another way to think of ROC analysis is as a binormal distribution of positive and negative measurements and their respective probability density functions (Figure 18). The more overlapped the two distributions, the more difficult it is to accurately sort the positive and negative results. Below is an example of such a distribution. The solid normal curve is the negative test probability density function, while the dotted normal curve is the positive one. The vertical line represents a chosen cutoff, above which a measurement is classified as positive. However, we can see that the upper tail of the negative normal curve will be counted among the positive measurements with this cutoff. This serves to illustrate the inherent tradeoff between sensitivity and specificity. One can make the cutoff more stringent to catch more positive measurements, but at the cost of falsely including some negative measurements.

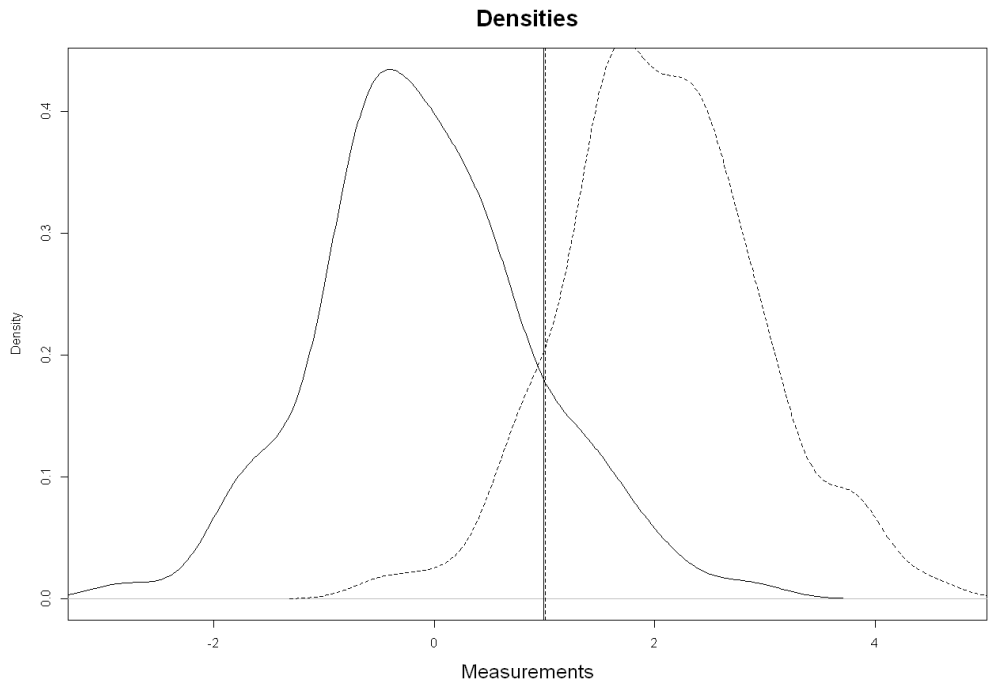



**Figure 18 Binormal curve of probability densities for positive and negative observations**

An ROC curve provides a convenient way to graphically display this tradeoff, allowing the viewer to decide what level of sensitivity or specificity they desire out of their diagnostic test, and then pick the corresponding cutoff value at that point. However, if the user does not know a priori what sensitivity and specificity they desire, but instead wishes to maximize the overall performance of the test, there are many choices available to pick the optimal cutoff value. Two common methods are to find the point along the ROC curve closest to the top left corner (the (0,1) method, Equation 9), or find the point farthest from the line of random guessing (Youden Index, Equation 10). This random guess line is the straight diagonal from the bottom left corner to the top right corner. In the equations

defining the optimal cutoffs below, c is the cutoff, q(c) is the sensitivity as a function of cutoff, p(c) is the specificity as a function of cutoff.

**Equation 9**

$$\min\left\{\sqrt{\left(1-q(c)\right)^2+\left(1-p(c)\right)^2}\right\},\qquad or\ \min\left\{\left(1-q(c)\right)^2+\left(1-p(c)\right)^2\right\}$$

**Equation 10**

$$J=\max\{q(c)+\ p(c)-\ 1\},or\ \max\{q(c)-\ \left(1-p(c)\right)\}$$

However, in practice the additional parameters of cost and prevalence play a role in the selection of an optimal cutoff value. According to a seminal paper on ROC analysis (Metz 1978), if the disease prevalence is low, then the false positive fraction must also be kept low, meaning an ideal point would be pushed more towards the lower left of the ROC curve. This is balanced by looking at the consequences, or cost, of false negatives (missing a failing plan) and false positives (missing a passing plan). If the costs of a false negative outweigh a false positive, then the ideal cutoff moves towards the upper right of the ROC curve, and vice versa. Perkins and Schisterman (2006) (Perkins 2006) showed that of the two methods of finding an ideal cutoff, the Youden index displays a more robust and reliable response when prevalence and cost are taken into account. Equation 11 shows how the Youden optimal cutoff is determined when cost and prevalence are included.

**Equation 11**

$$J=\max\{q(c)+\ r*p(c)-\ 1\};where\ r=\frac{1-prevalence}{(relative\ cost\ of\ FN\ to\ FP)*prevalence}$$

Empirical ROC curves can appear jagged and have a staircase pattern due to the few measurements made due to limitations in time and resources. Using a binormal fit has been shown to be a robust method of smoothing an ROC curve (Hanley 1988). This can be used to give the user a snapshot of how the curve might appear if more data were available. The data is assumed to have a distribution like the one shown in Figure 18. An example of smoothing using a binormal fit is shown in Figure 19.

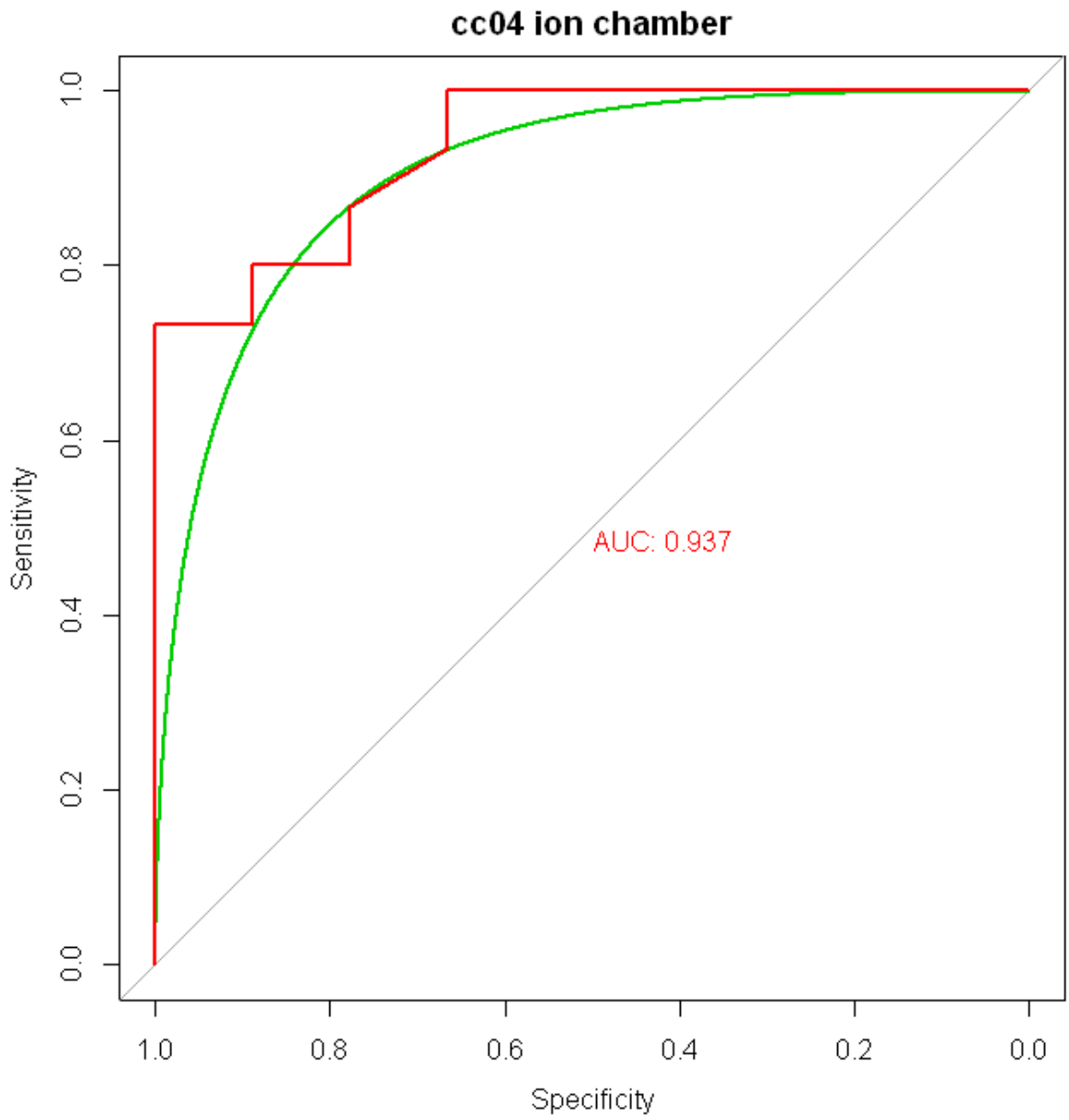



**Figure 19 The empirical and smoothed ROC curves for the cc04 ion chamber**

The jagged nature of the empirical curves can also affect the choice in optimal cutoff, especially if the cost and prevalence are such that the point would occur in a straight part of the lower left or upper right ROC curve. When faced with this situation, the optimization algorithm may never reach a certain point on the curve. To address this issue, the threshold

optimization estimation can be considered not by the geometries of the ROC plot, but through a cost minimization function. This function has been discussed in the literature (Skaltsa 2010). When only the empirical version of this method is considered, no assumptions need to be made about the underlying distributions of the data; therefore in this body of work (Chapter 4), the empirical cost minimization method was used. Optimal thresholds are estimated by setting the ratio of probability distribution functions for positive and negative plans equal to the negative odds ratio times the cost ratio:

**Equation 12**

$$\frac{sensitvity(cutoff)}{specificity(cutoff)} = \left( \frac{1 - prevalence\ of\ a\ failing\ plan}{prevalence\ of\ a\ failing\ plan} * \frac{1}{relative\ cost\ of\ a\ falsely\ classified\ failing\ plan} \right)$$

Empirically, this means that the cutoffs are varied until the value of the ratio of sensitivity to specificity equals the right hand side of Equation 12. The proof for this is shown in the paper by Skaltsa et al (Skaltsa 2010).

# V. APPENDIX II

## BOOTSTRAPPING

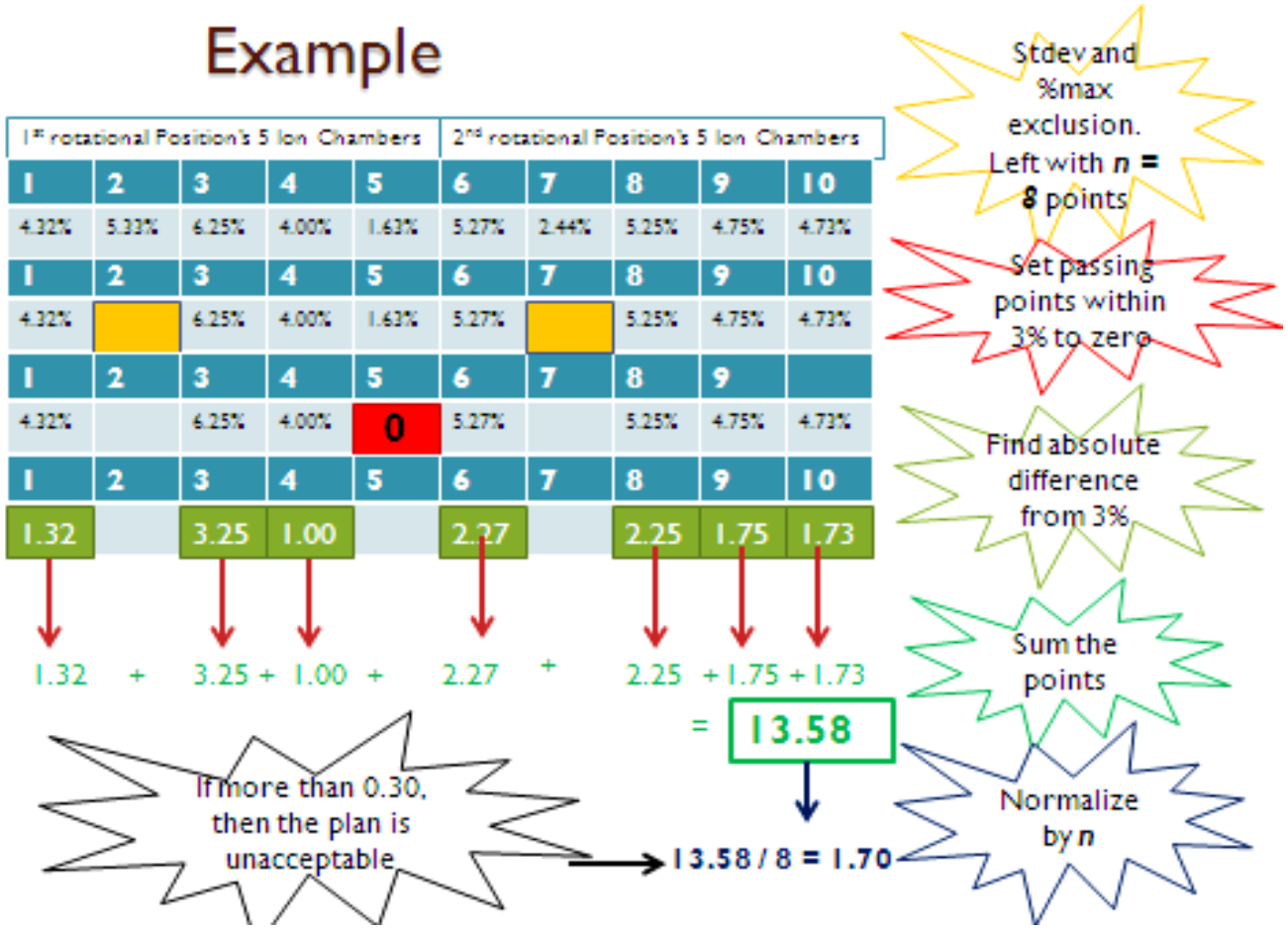
Sometimes in research, due to limitations in resources, multiple replicates are not feasible in the acquisition of the data. However, it is important to be able to gauge the uncertainty in one's measurements. In order to accomplish this, a technique known as bootstrapping has been established to create confidence intervals on datasets with limited replicates. Bootstrapping is a well accepted technique which uses resampling of the data to estimate confidence intervals (Carpenter 2000). In order to estimate a given population parameter, a measurement is made on a random sample taken from that population. Each time a re-measurement is made, it is another random sample from the population. Therefore, one can simulate multiple random samples through randomly re-sampling from actual measurements. These resampled data can be used to calculate multiple values of a population parameter. Bootstrapping thus yields a range of values from which the $95^{th}$ percentile can be selected and applied to the parameter calculated from the original measurement (Carpenter 2000). This technique was applied in this research to calculated confidence intervals for the area under the curves and the optimal cutoff values in Chapter 4.

# VI. APPENDIX III

## MULTIPLE ION CHAMBER PHANTOM PASS/FAIL CRITERIA

Because the multiple ion chamber phantom (MIC) was designed in-house, a metric to determine if the results of its measurement were passing or failing had to be determined. A single value was distilled from incorporating all ion chamber measurements (see Chapter 4). In order to determine the cut off value of this MIC metric that would classify a plan as passing or failing, a study on the robustness of different metrics was conducted. This robustness was gauged by the metric's ability to consistently sort all 24 patient plans across minimum dose threshold cutoffs and percent standard deviations in the dose across the ion chamber volume. To evaluate this, a program was written which iterated through the dose thresholds of 35%, 50%, and 65%, standard deviations of 1%, 2%, and 3%, and an MIC metric of 0.01 – 0.6. It was found that the MIC metrics 0.27 – 0.35 were all the most robust, and all sorted the plans the same way. Therefore, 0.30 was selected in order to be confident that the sorting would not be changed by slight changes in dose level or gradient.

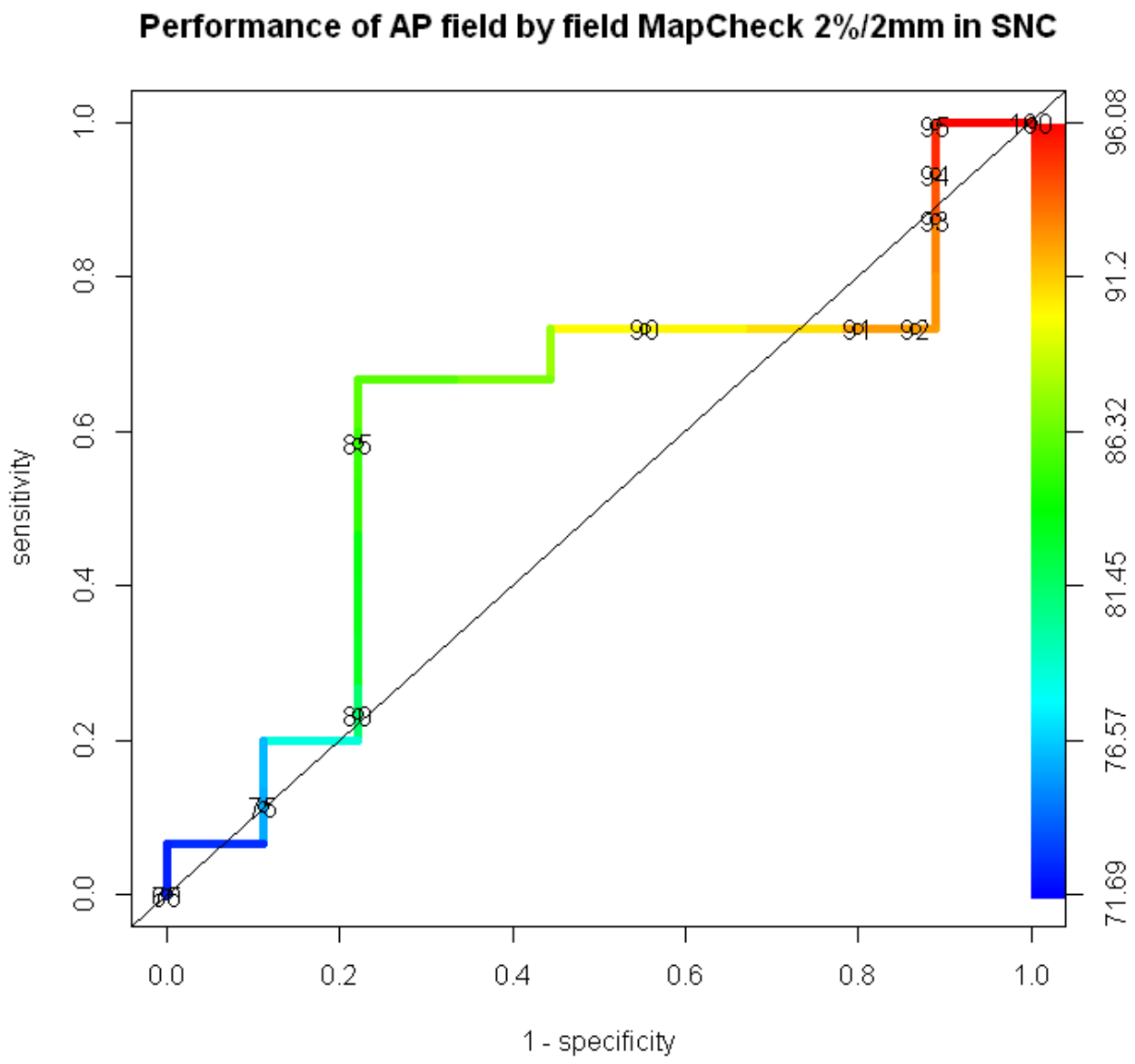
An example of how the MIC uses the 10 ion chamber measurements to sort a plan as acceptable or unacceptable is shown on the next page

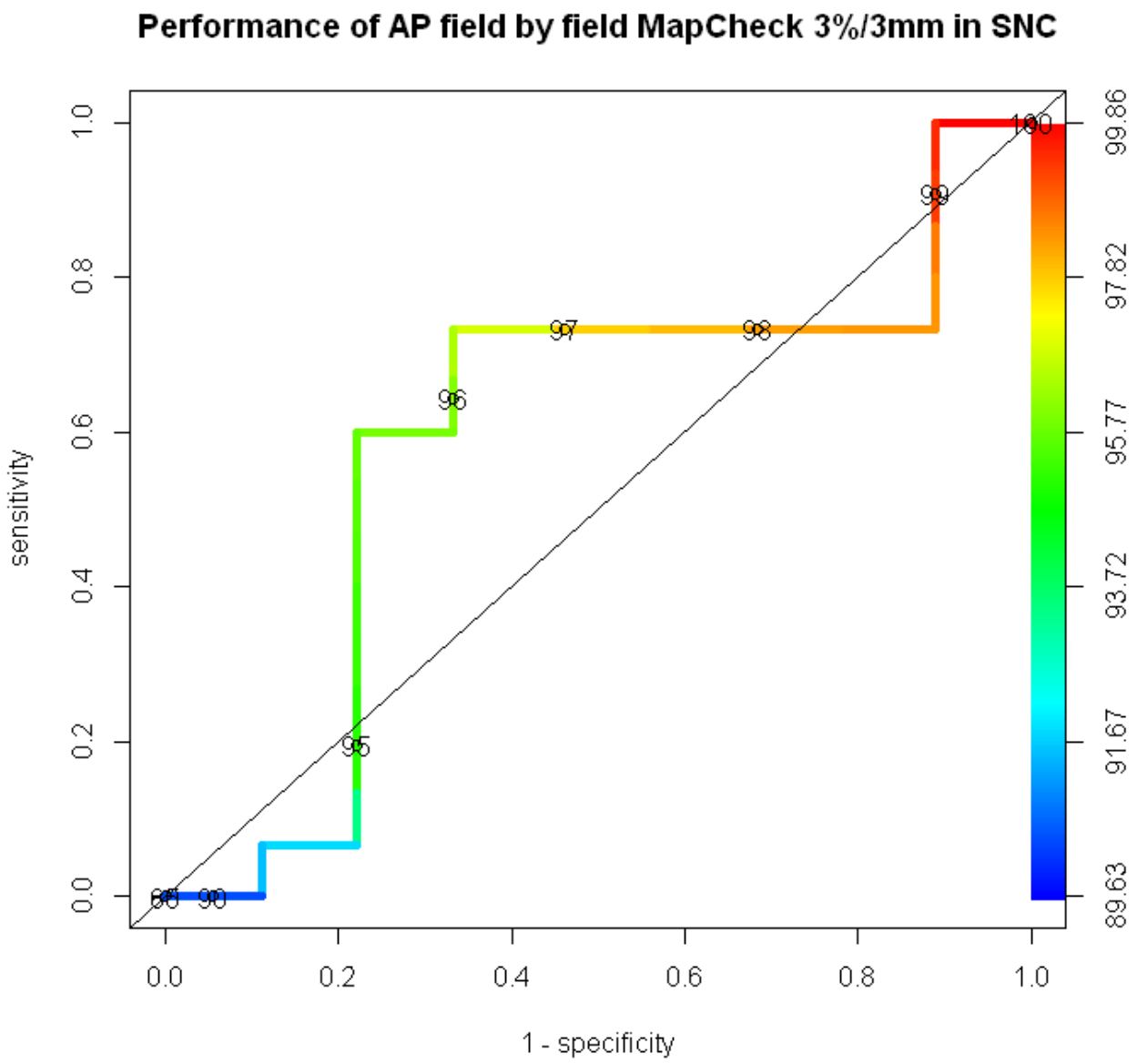

Example
1st rotational Position's 5 Ion Chambers
2nd rotational Position's 5 Ion Chambers
1 2 3 4 5 6 7 8 9 10
4.32% 5.33% 6.25% 4.00% 1.63% 5.27% 2.44% 5.25% 4.75% 4.73%
1 2 3 4 5 6 7 8 9 10
4.32% 6.25% 4.00% 1.63% 5.27% 5.25% 4.75% 4.73%
1 2 3 4 5 6 7 8 9
4.32% 6.25% 4.00% 0 5.27% 5.25% 4.75% 4.73%
1 2 3 4 5 6 7 8 9 10
1.32 3.25 1.00 2.27 2.25 1.75 1.73
1.32 + 3.25 + 1.00 + 2.27 + 2.25 + 1.75 + 1.73
= 13.58
13.58 / 8 = 1.70
Stdev and %max exclusion. Left with n = 8 points
Set passing points within 3% to zero
Find absolute difference from 3%
Sum the points
Normalize by n
If more than 0.30, then the plan is unacceptable

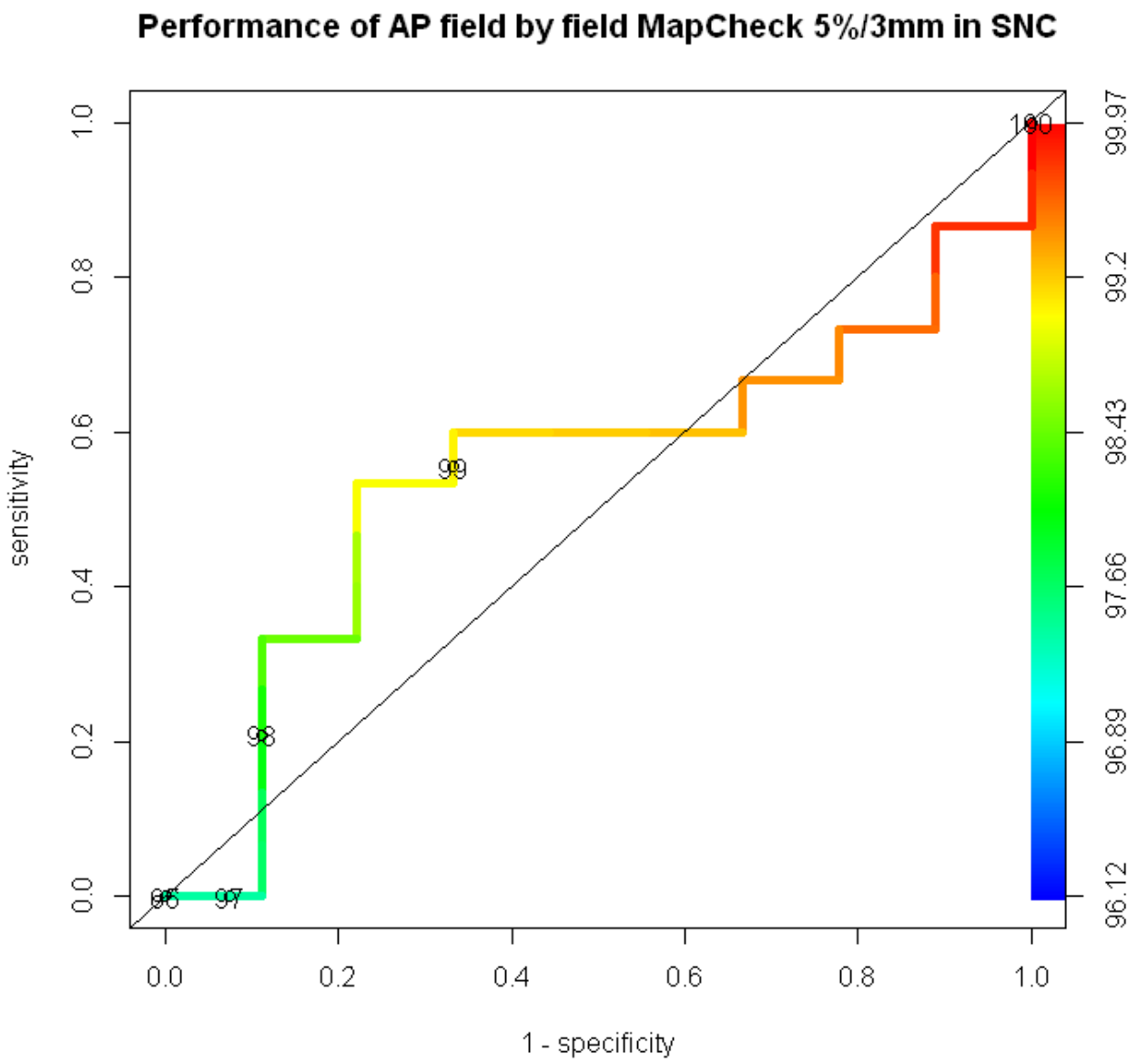
# VII. APPENDIX IV

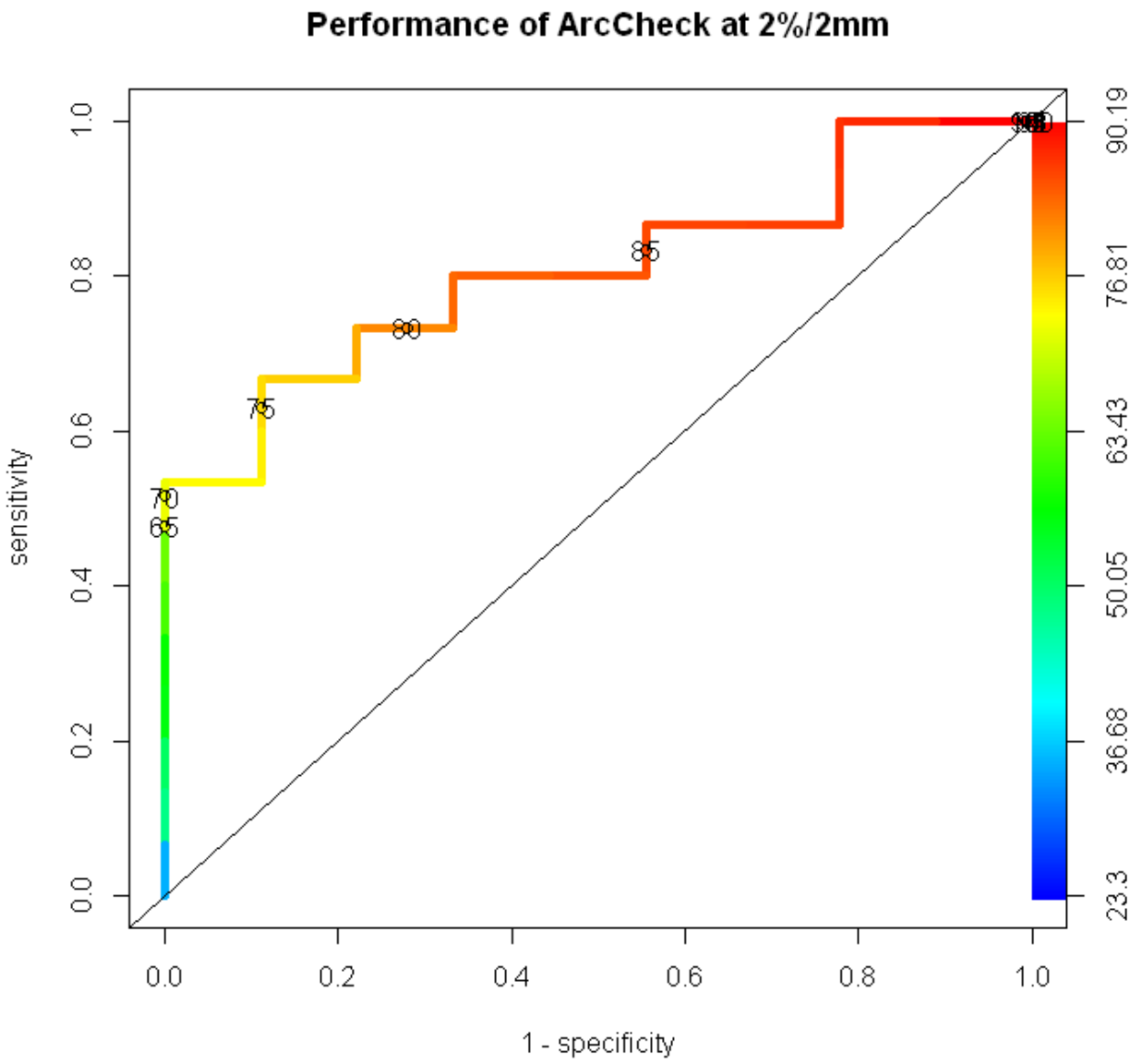
## ADDITIONAL FIGURES

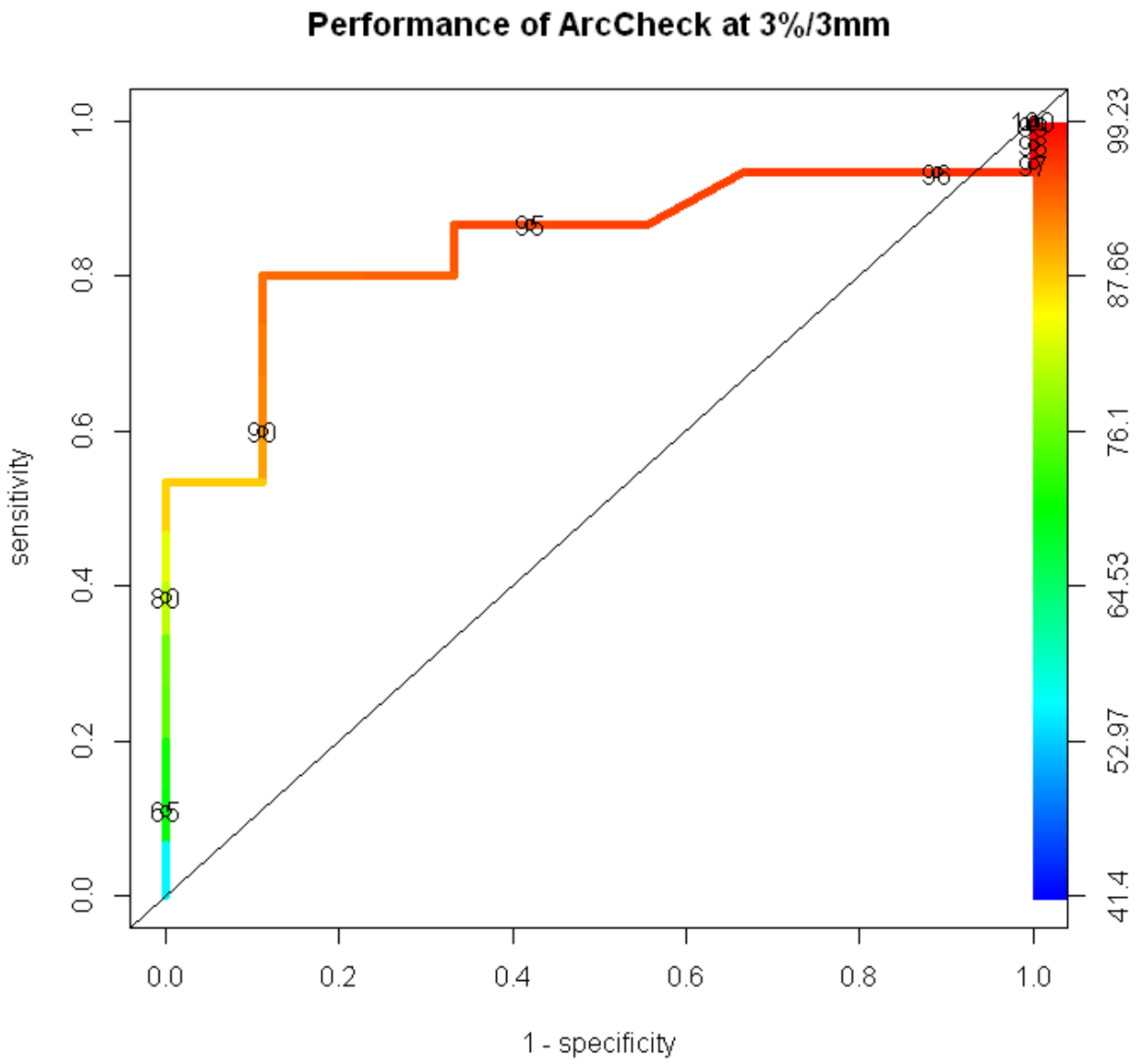
ROC curves created for each dosimeter system (cutoffs are both numerically printed on the curves, and color coded)

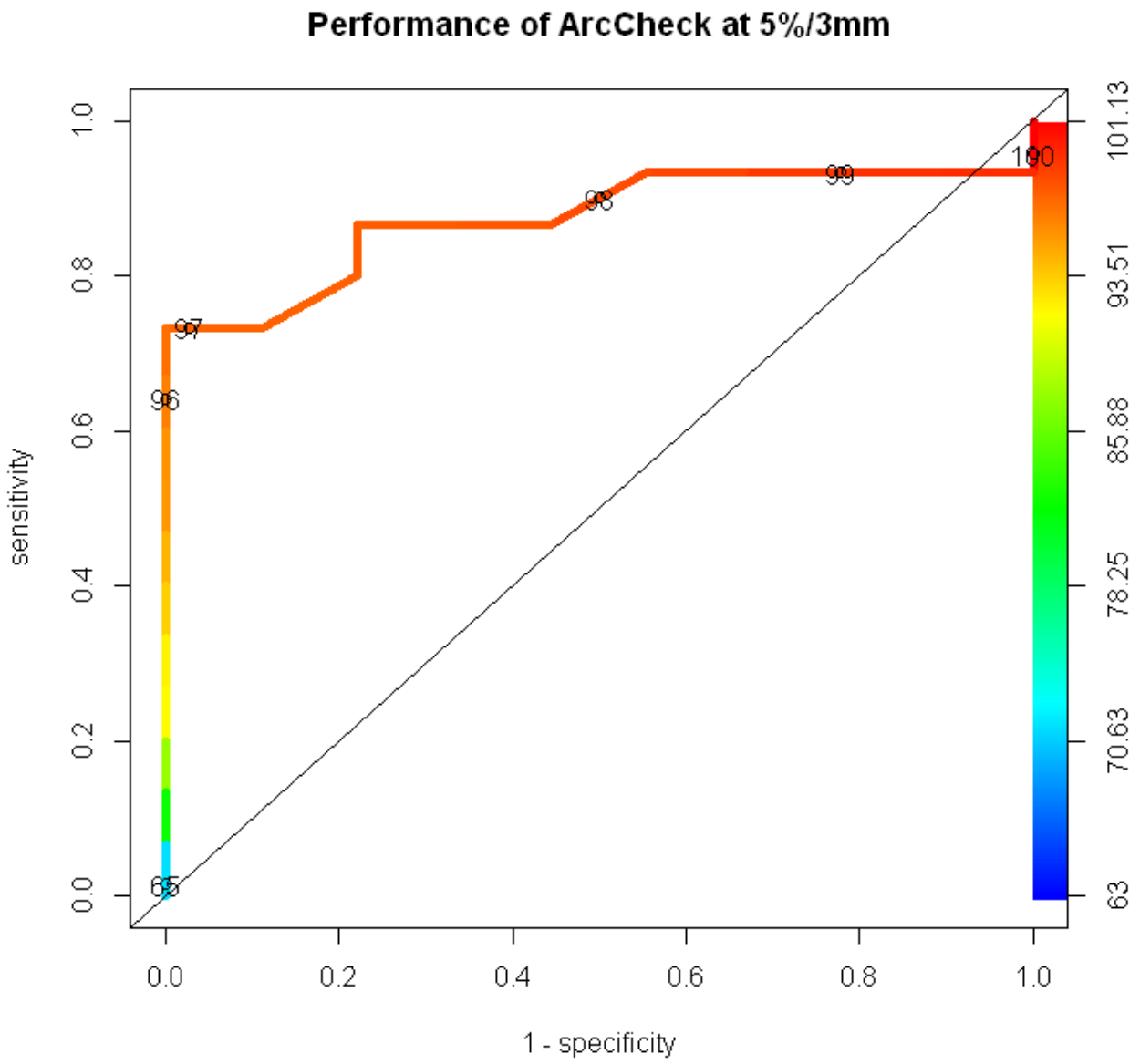


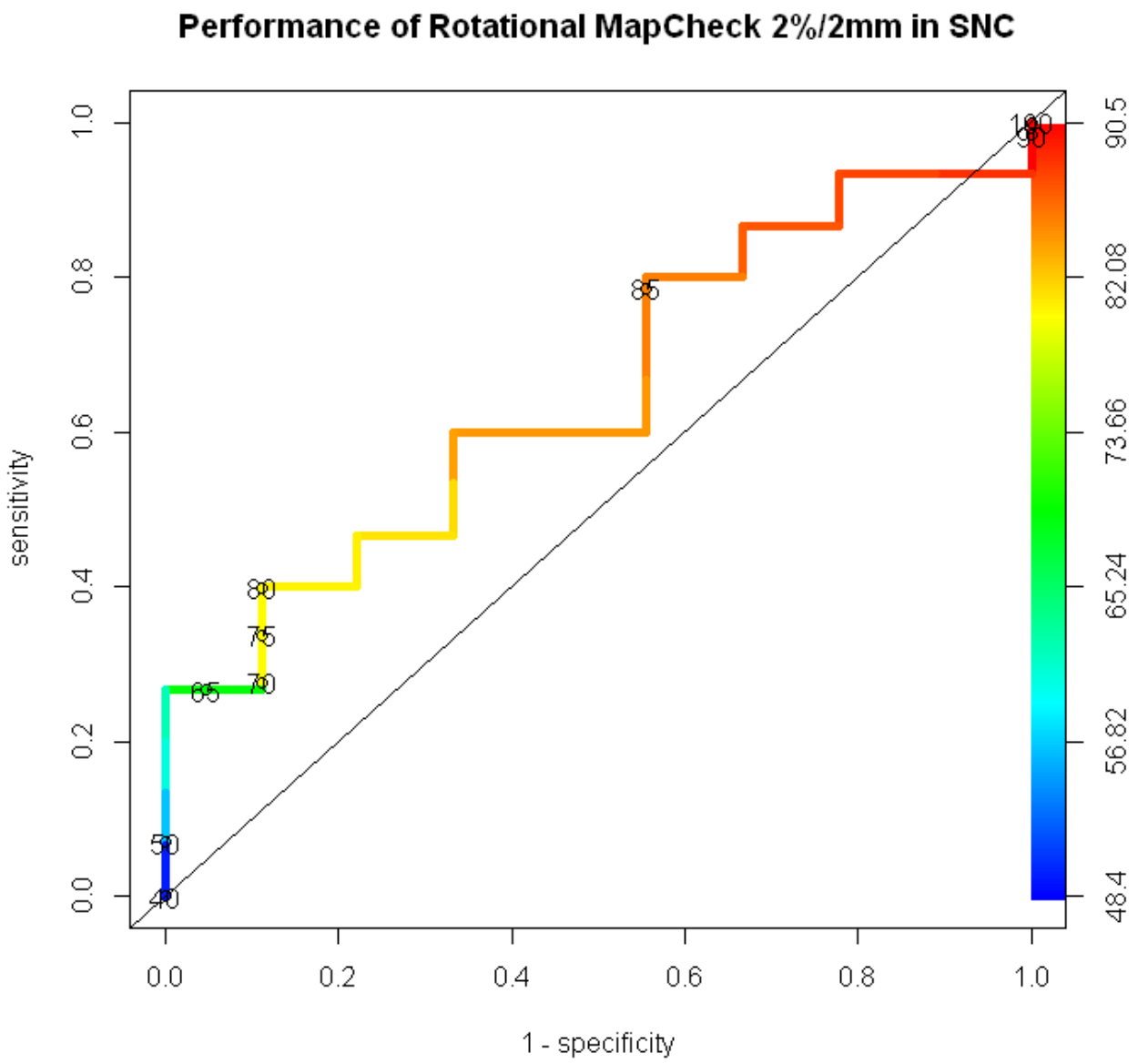

Performance of AP field by field MapCheck 3%/3mm in SNC
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
89.63
91.67
93.72
95.77
97.82
99.86
100
99
98
97
96
95
90

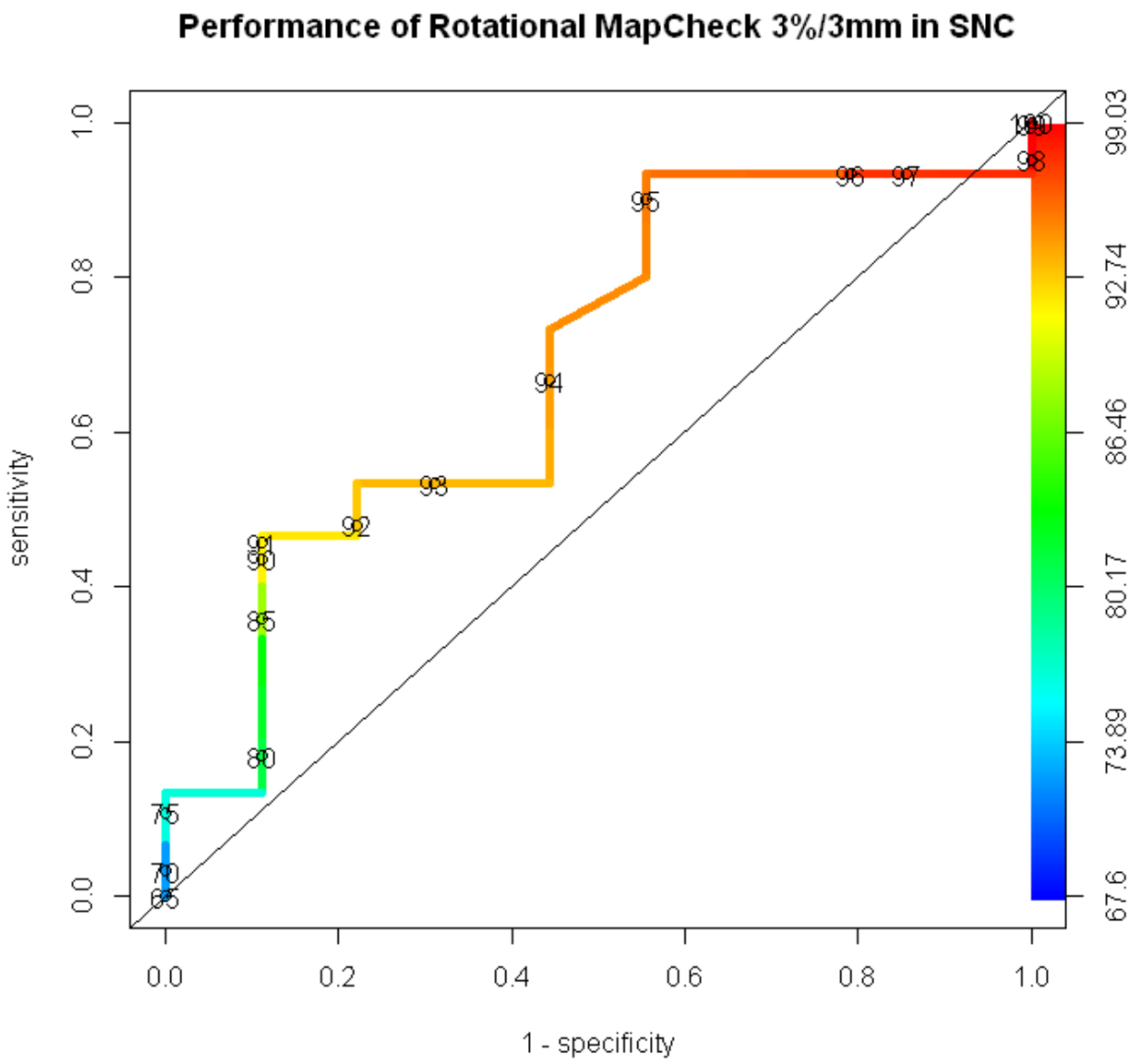

Performance of AP field by field MapCheck 5%/3mm in SNC
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
99.97
99.2
98.43
97.66
96.89
96.12
100
99
98
97
96

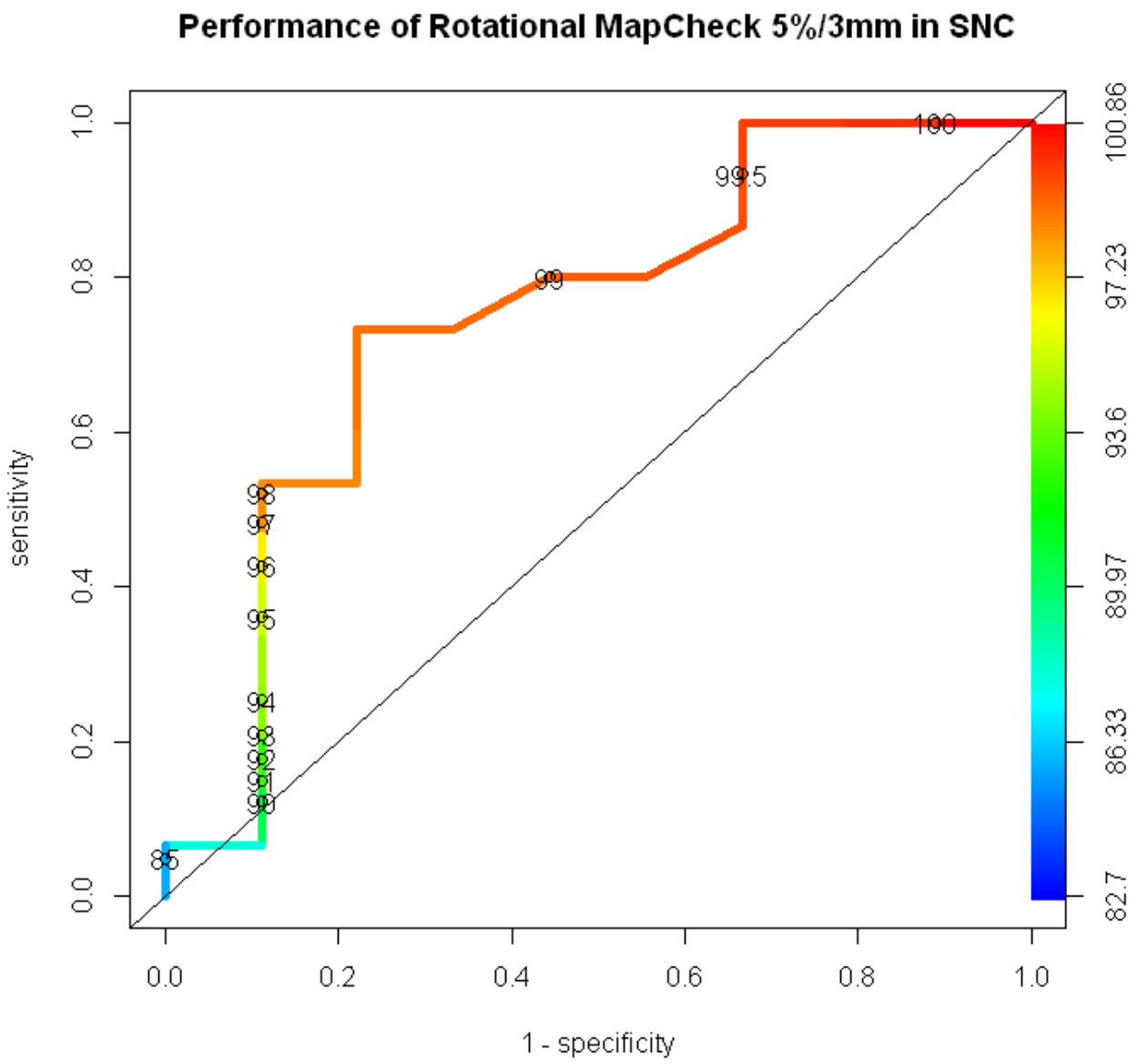

Performance of ArcCheck at 2%/2mm
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
90.19
76.81
63.43
50.05
36.68
23.3
65
70
75
80
85

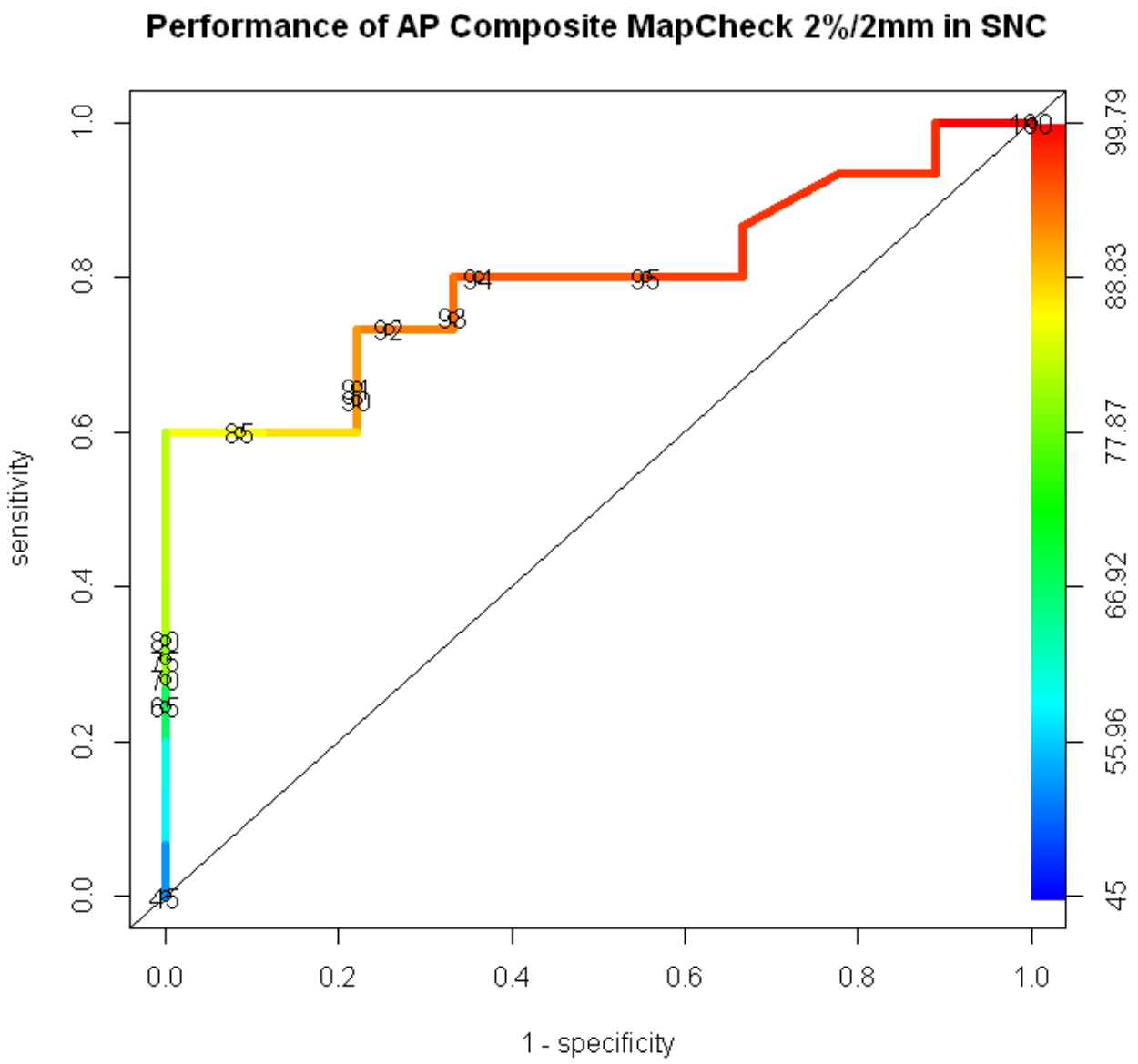

Performance of ArcCheck at 3%/3mm
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
99.23
87.66
76.1
64.53
52.97
41.4
90
95
98
80
65

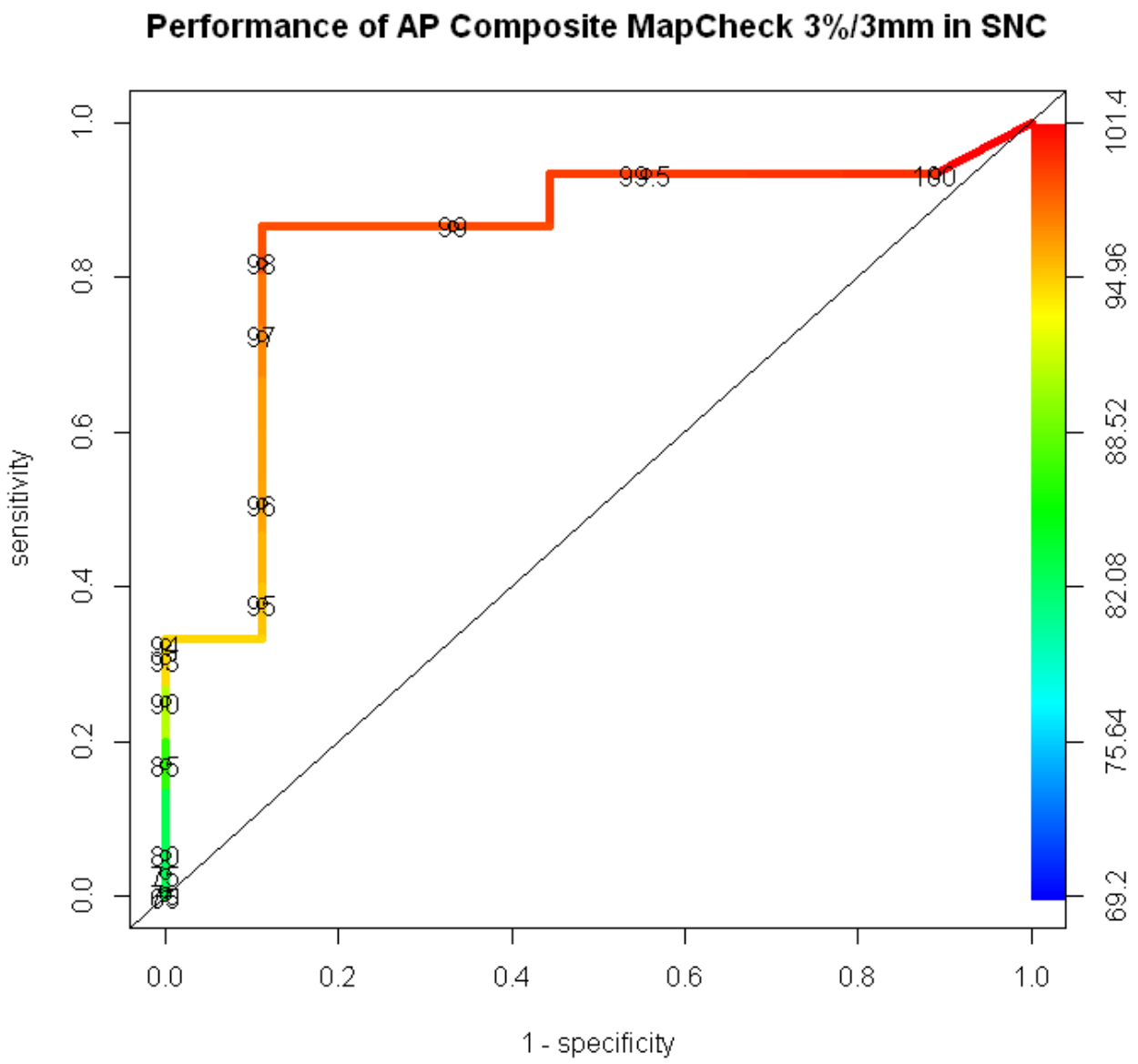

Performance of ArcCheck at 5%/3mm
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
101.13
93.51
85.88
78.25
70.63
63
100
99
98
97
96
65

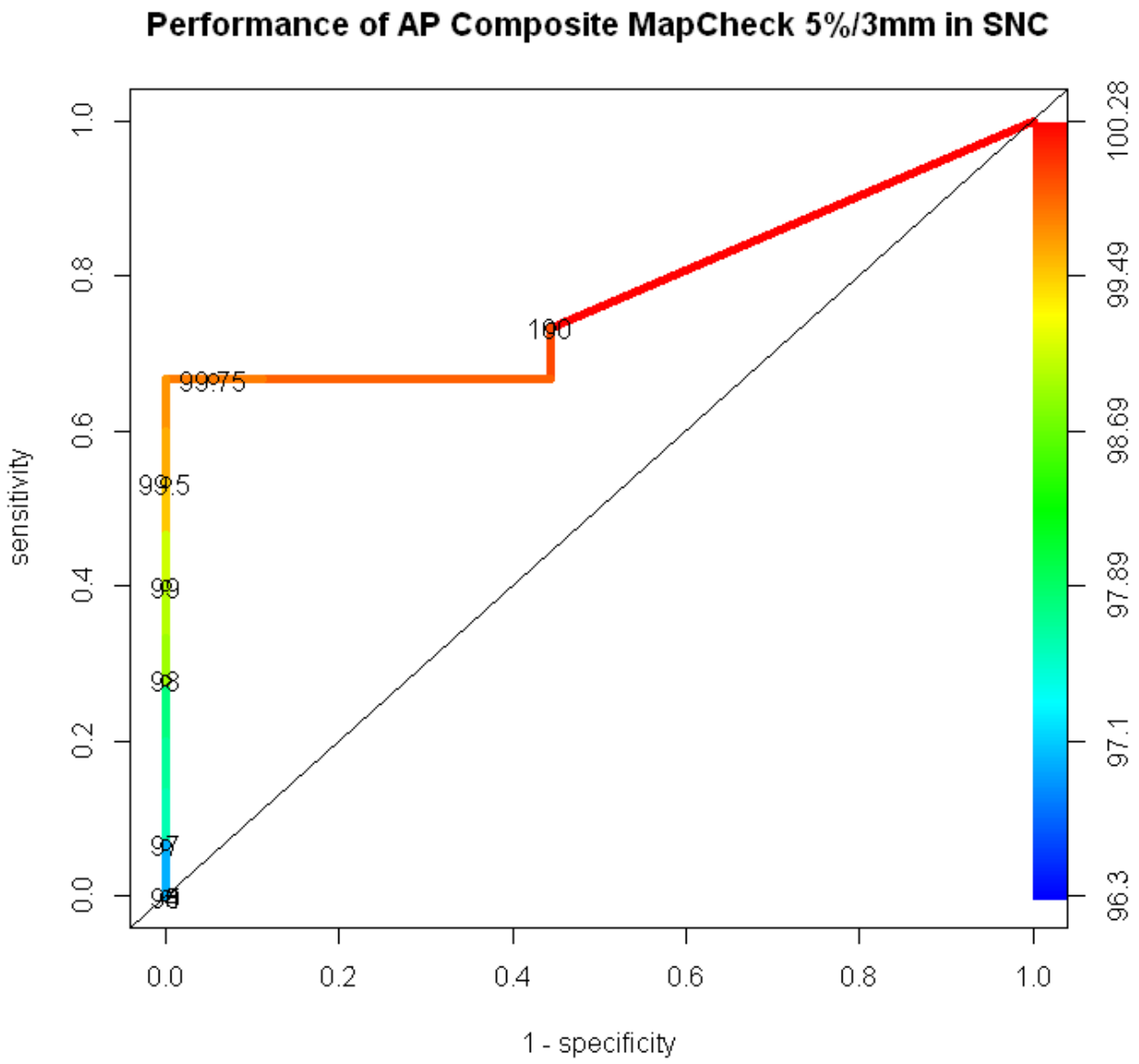

Performance of Rotational MapCheck 2%/2mm in SNC
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
90.5
82.08
73.66
65.24
56.82
48.4
100
90
85
80
75
70
65
50
40

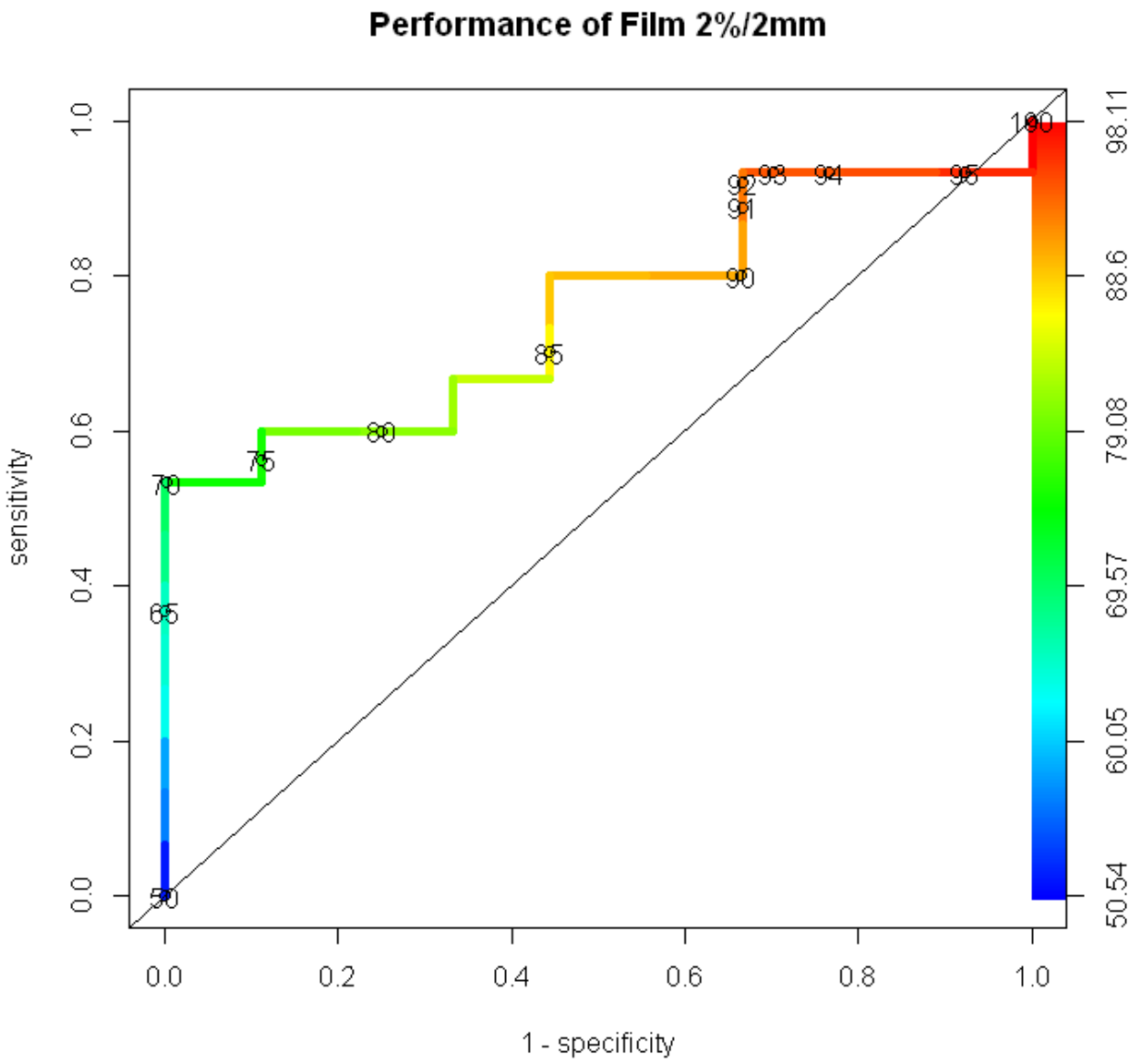

Performance of Rotational MapCheck 3%/3mm in SNC
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
99.03
92.74
86.46
80.17
73.89
67.6

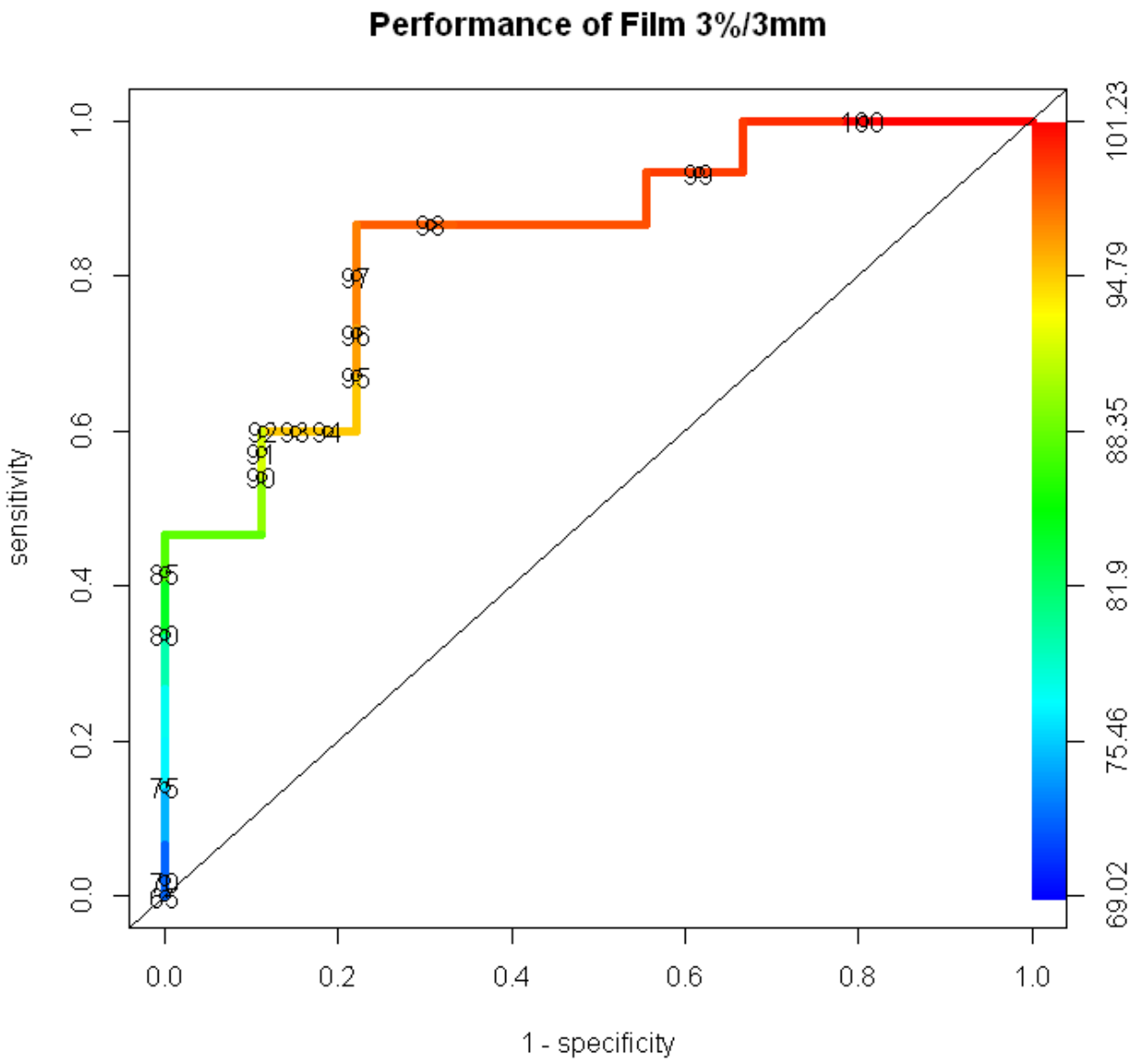

Performance of Rotational MapCheck 5%/3mm in SNC
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
100.86
97.23
93.6
89.97
86.33
82.7
100
99.5
99
85

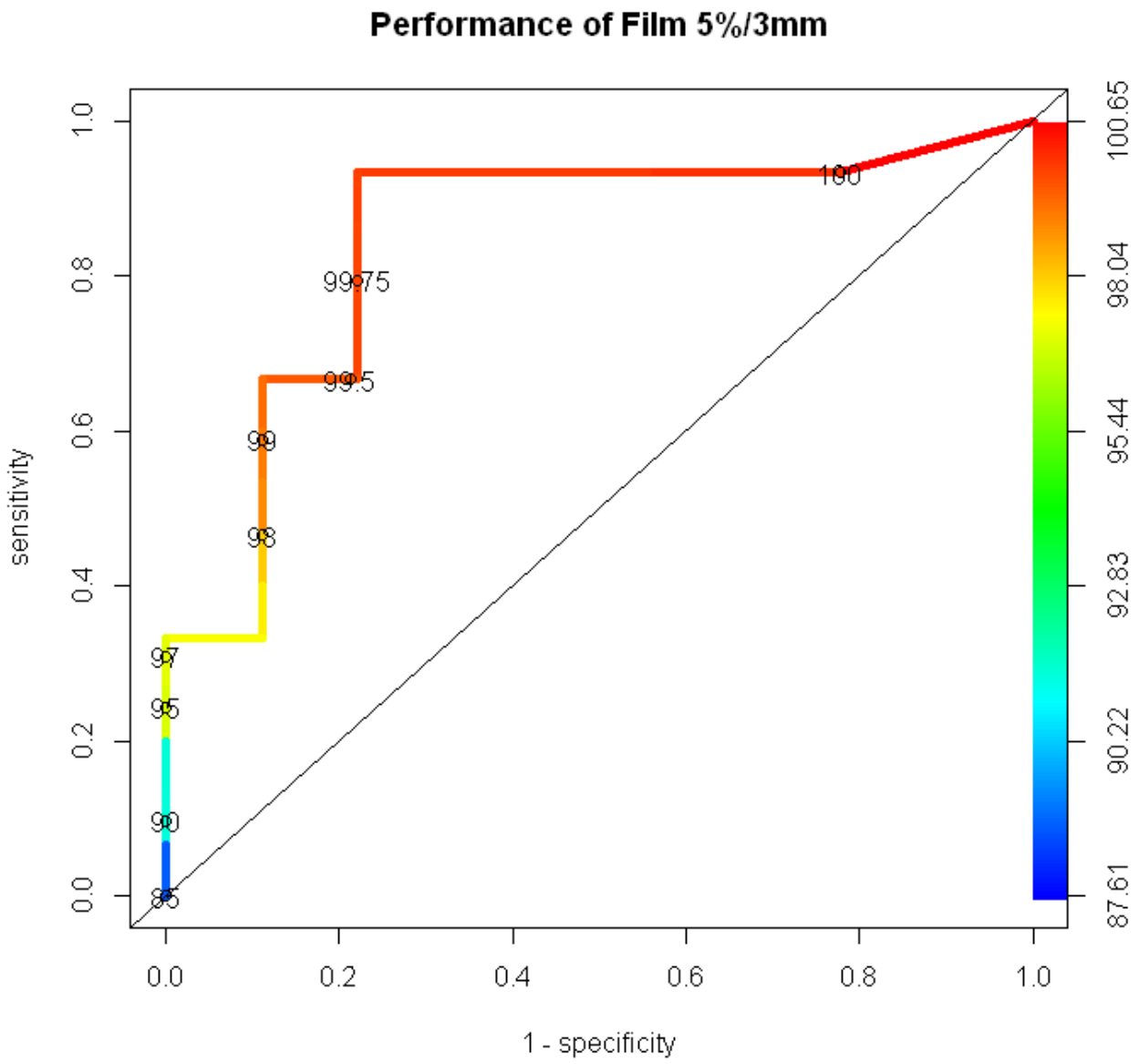

Performance of AP Composite MapCheck 2%/2mm in SNC
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
99.79
88.83
77.87
66.92
55.96
45
85
92
94
95
100

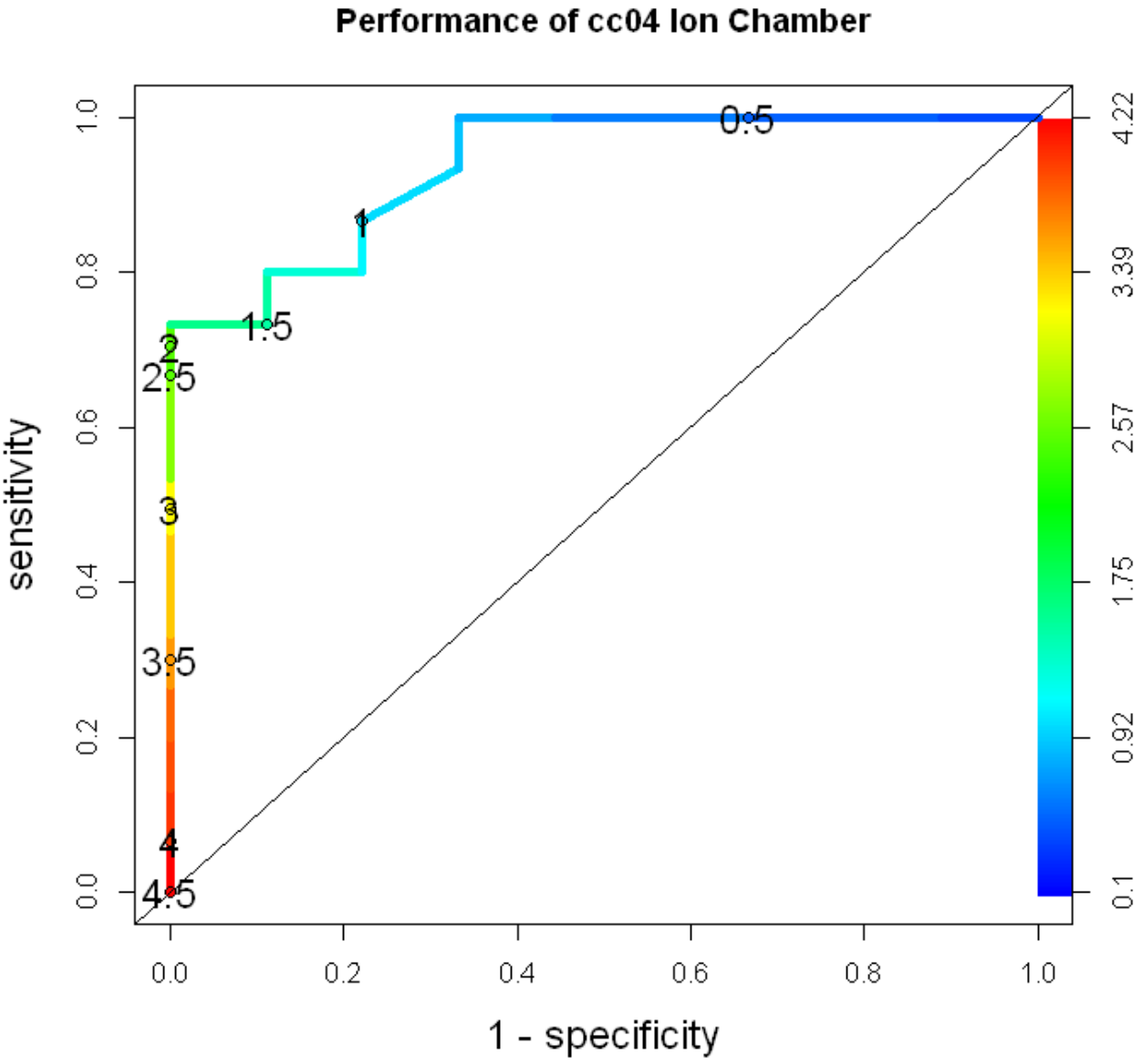

Performance of AP Composite MapCheck 3%/3mm in SNC
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
101.4
94.96
88.52
82.08
75.64
69.2
99.5
100
99
98
97
96
95

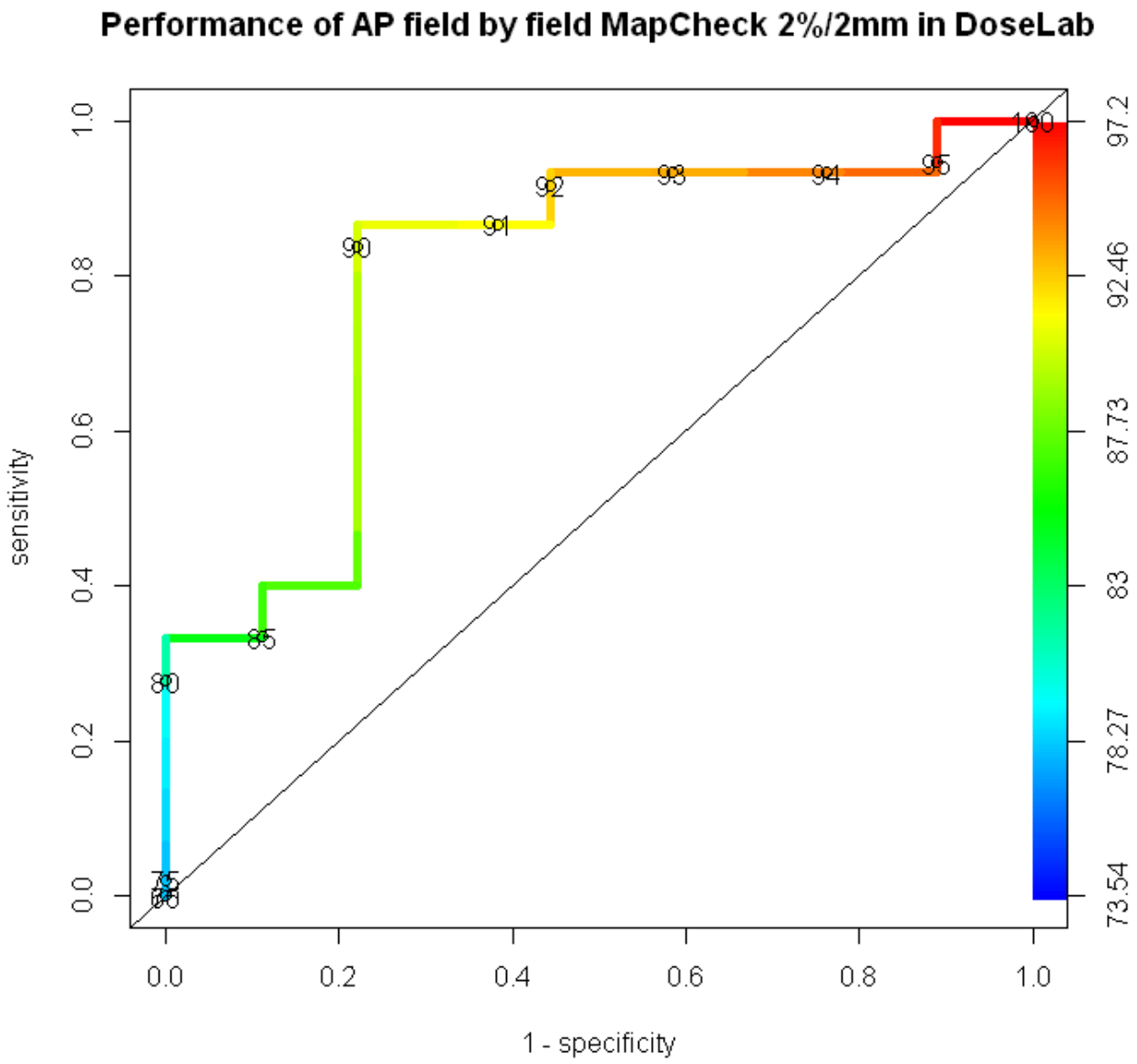

Performance of AP Composite MapCheck 5%/3mm in SNC
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
100.28
99.49
98.69
97.89
97.1
96.3
100
99.75
99.5
99
98
97

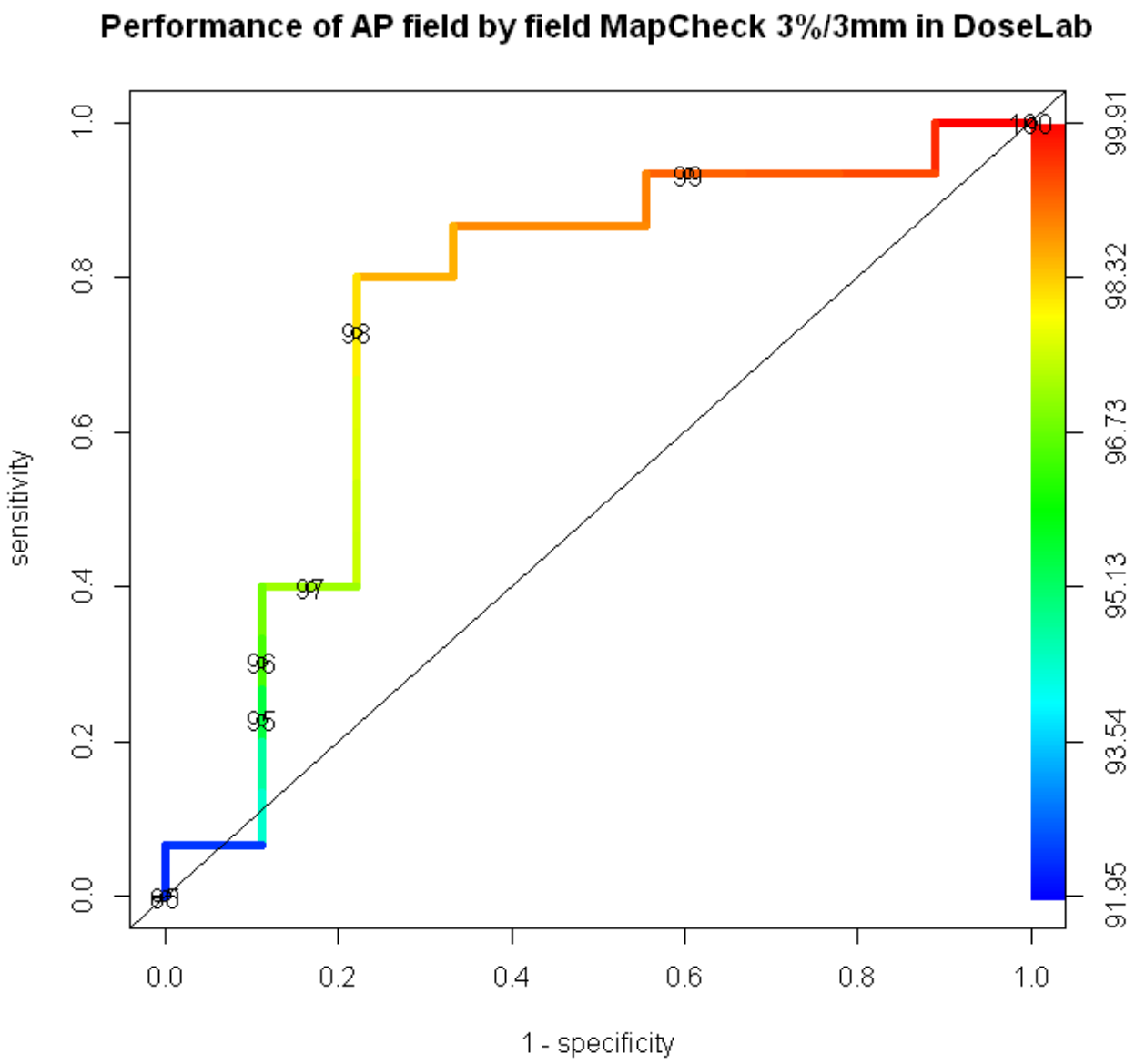

Performance of Film 2%/2mm
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
98.11
88.6
79.08
69.57
60.05
50.54
50
65
70
75
80
85
90
92
93
94
95
100

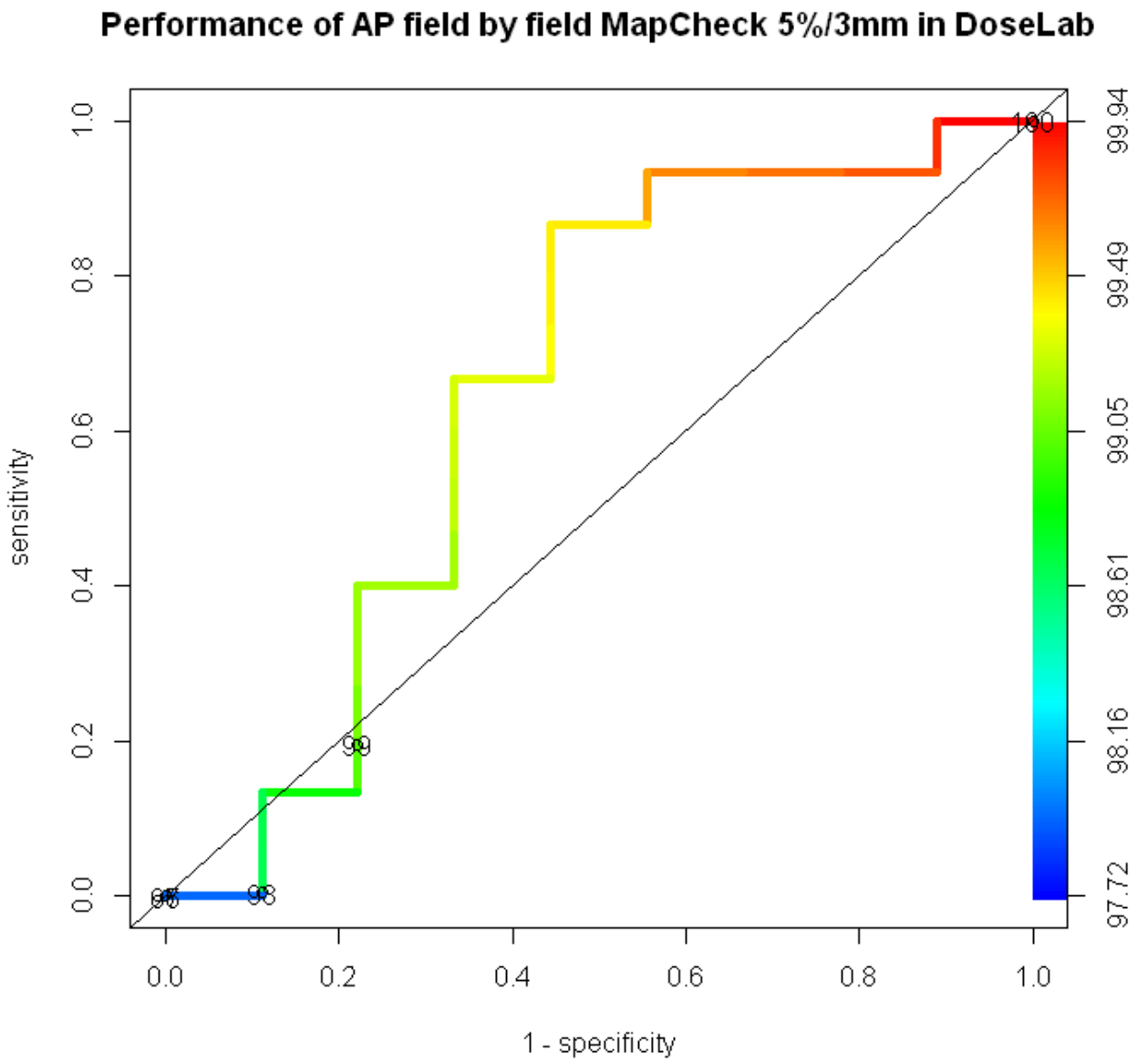

Performance of Film 3%/3mm
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
101.23
94.79
88.35
81.9
75.46
69.02

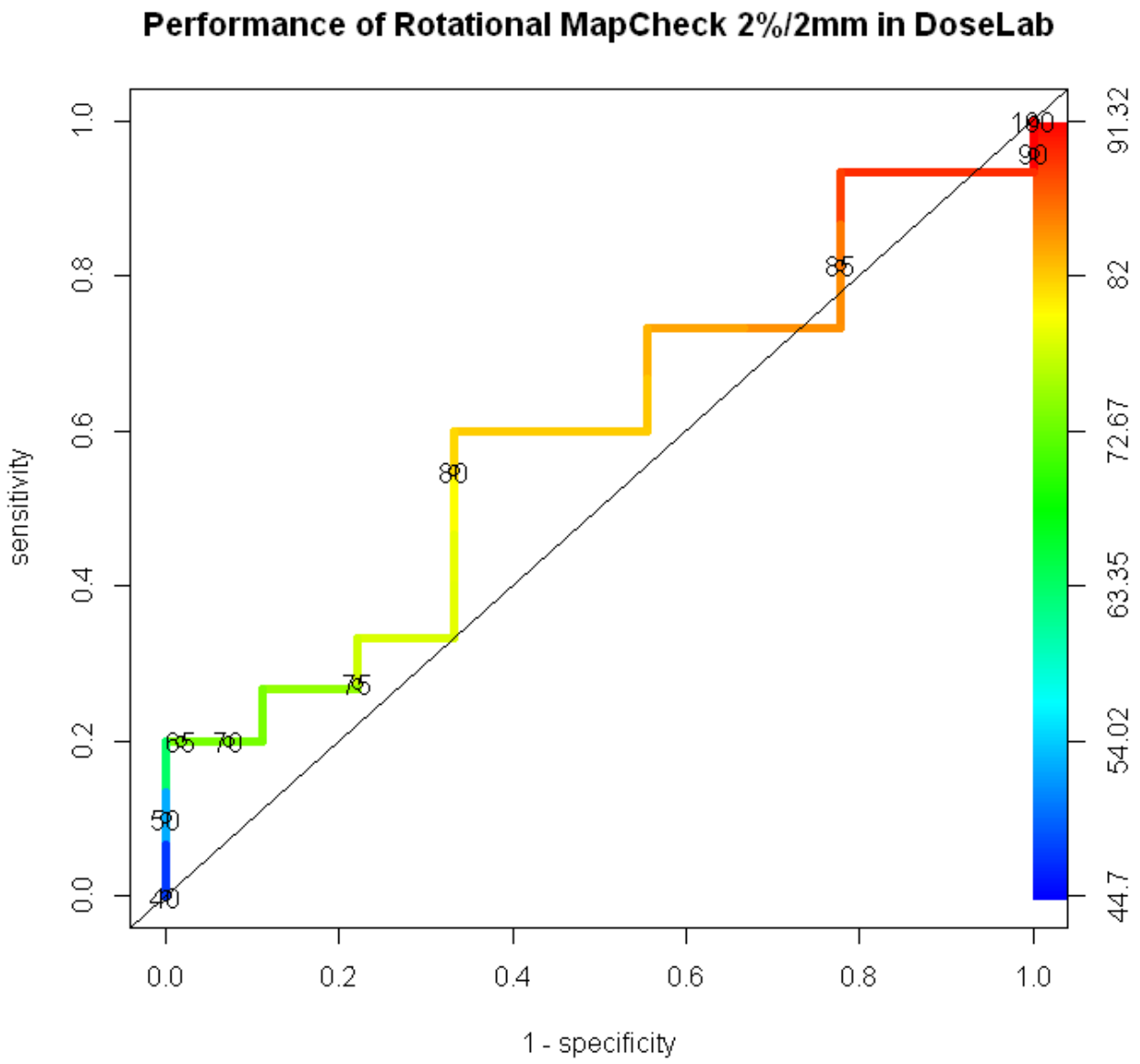

Performance of Film 5%/3mm
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
100
99.75
99.5
99
98
97
95
90
85
100.65
98.04
95.44
92.83
90.22
87.61

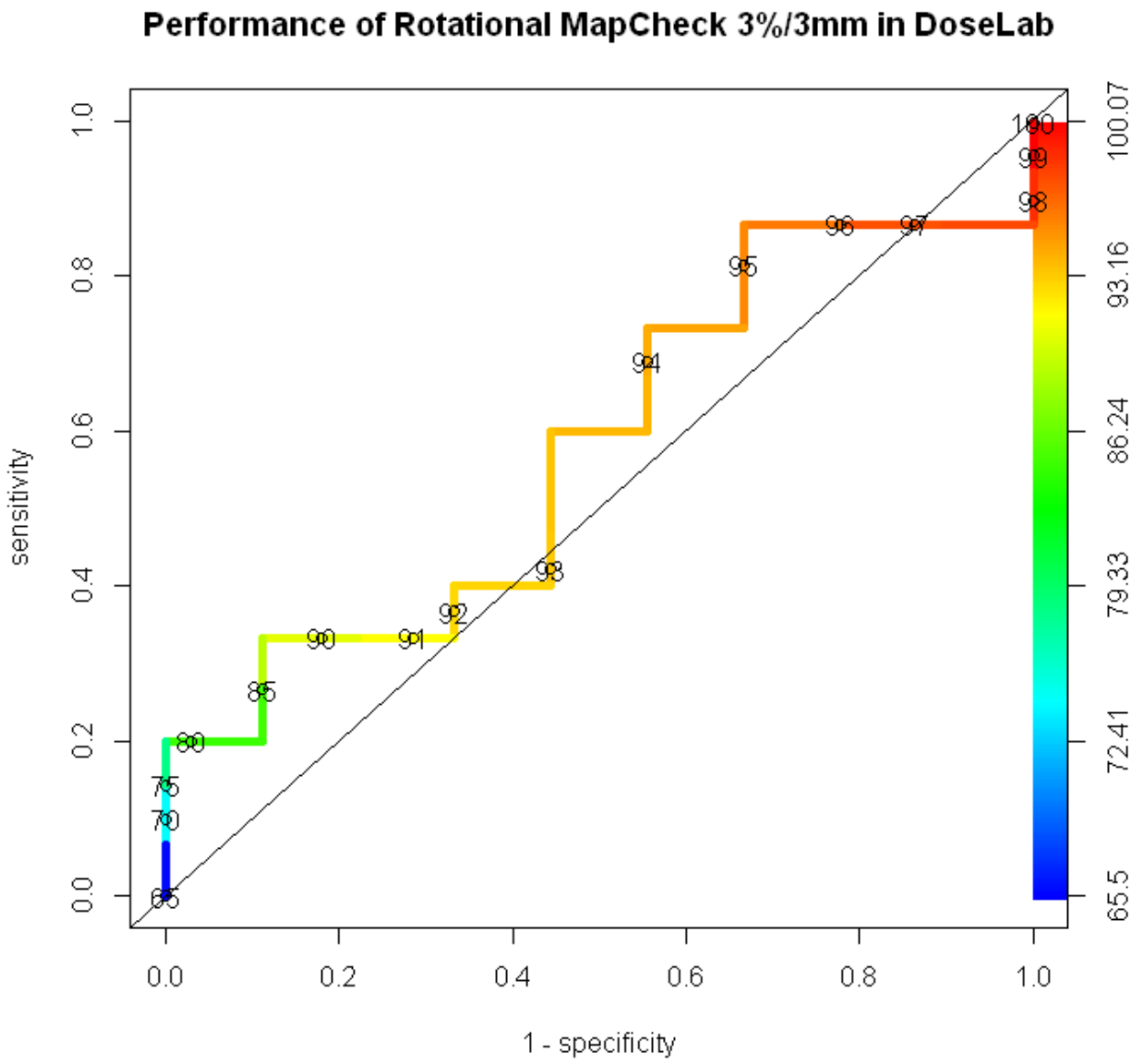

Performance of cc04 Ion Chamber
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.92
1.75
2.57
3.39
4.22
0:5
1
1:5
2
2:5
3
3:5
4
4:5

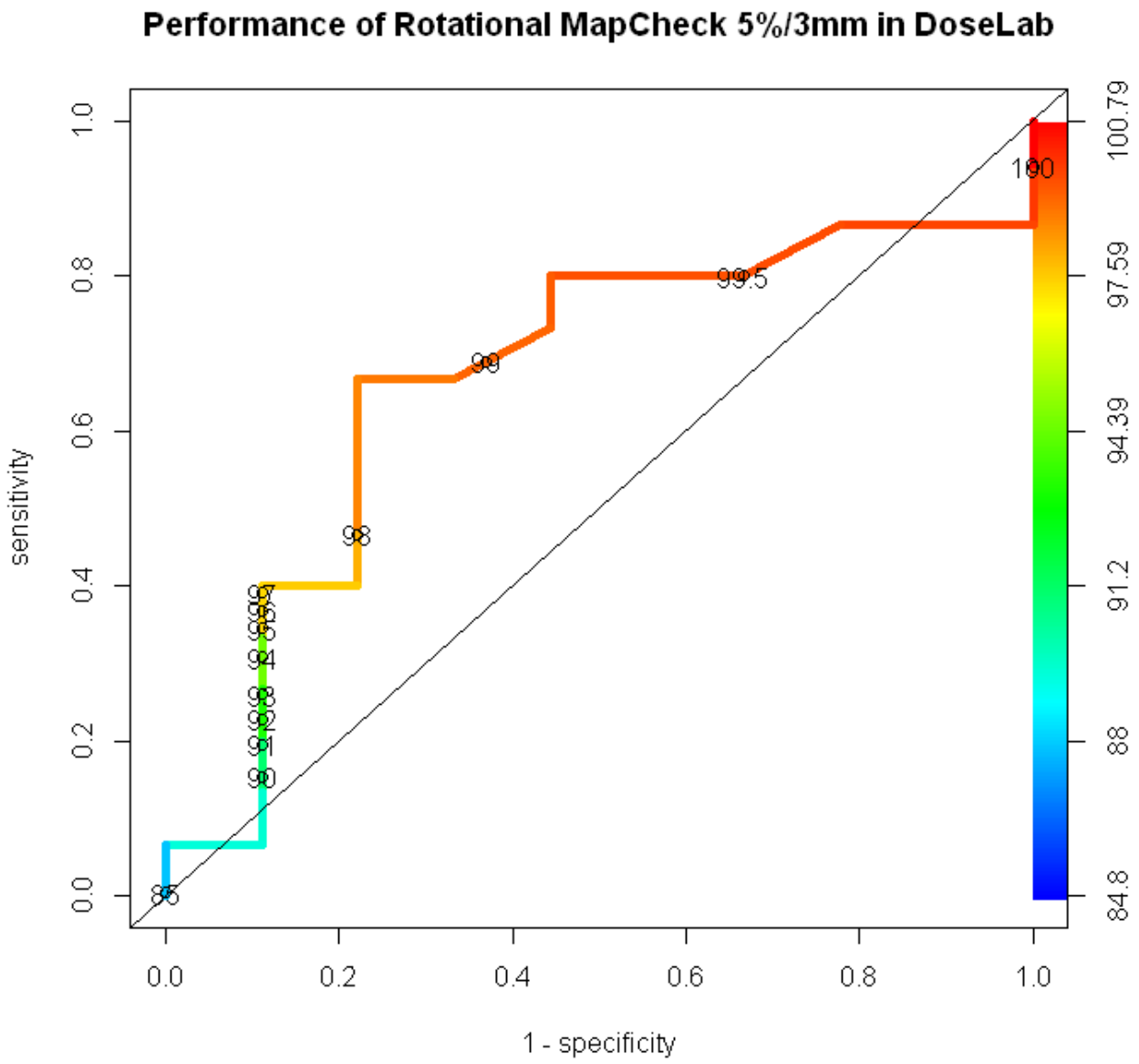

Performance of AP field by field MapCheck 2%/2mm in DoseLab
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
97.2
92.46
87.73
83
78.27
73.54

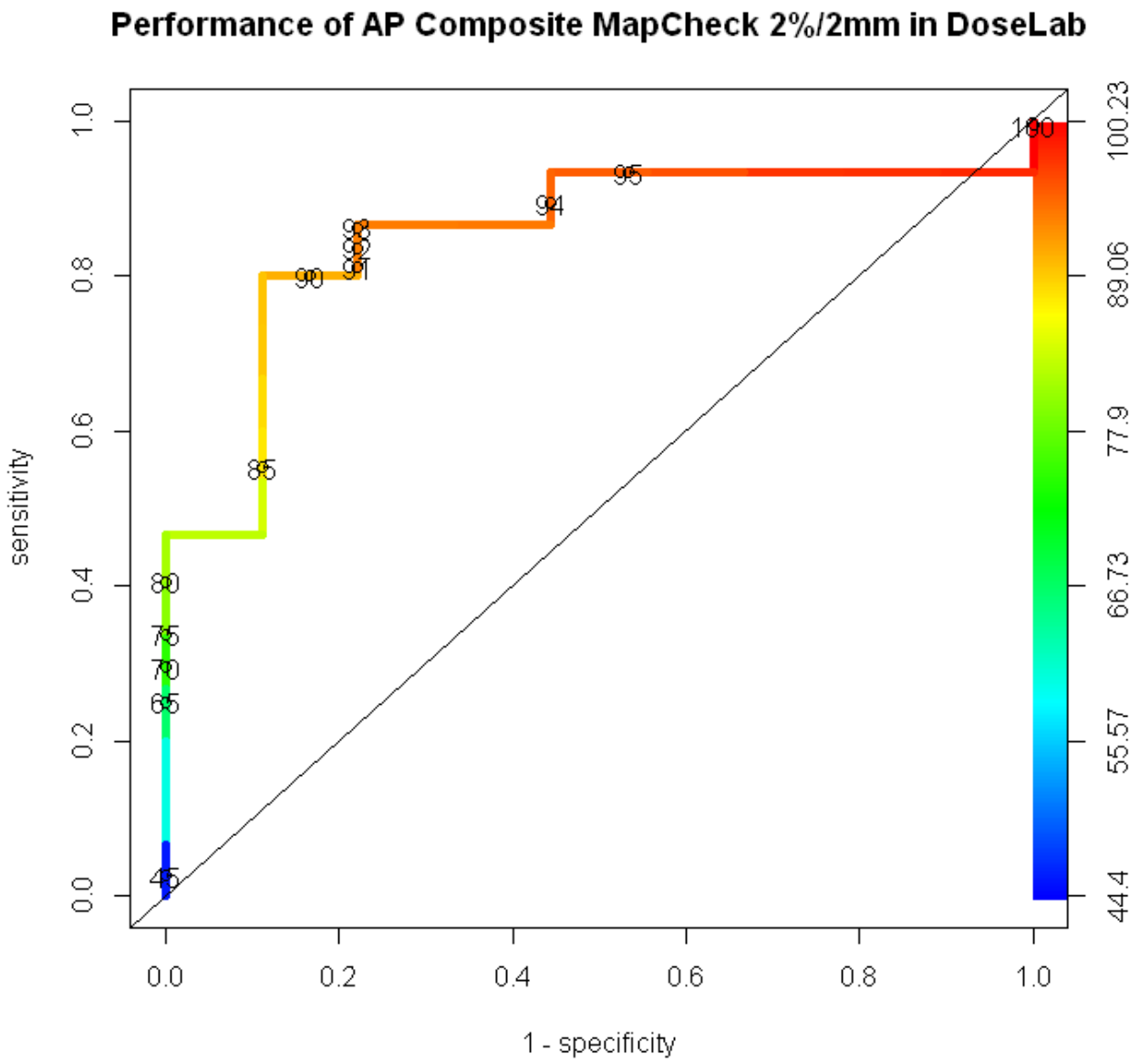

Performance of AP field by field MapCheck 3%/3mm in DoseLab
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
91.95
93.54
95.13
96.73
98.32
99.91
100
99
98
97
96
95

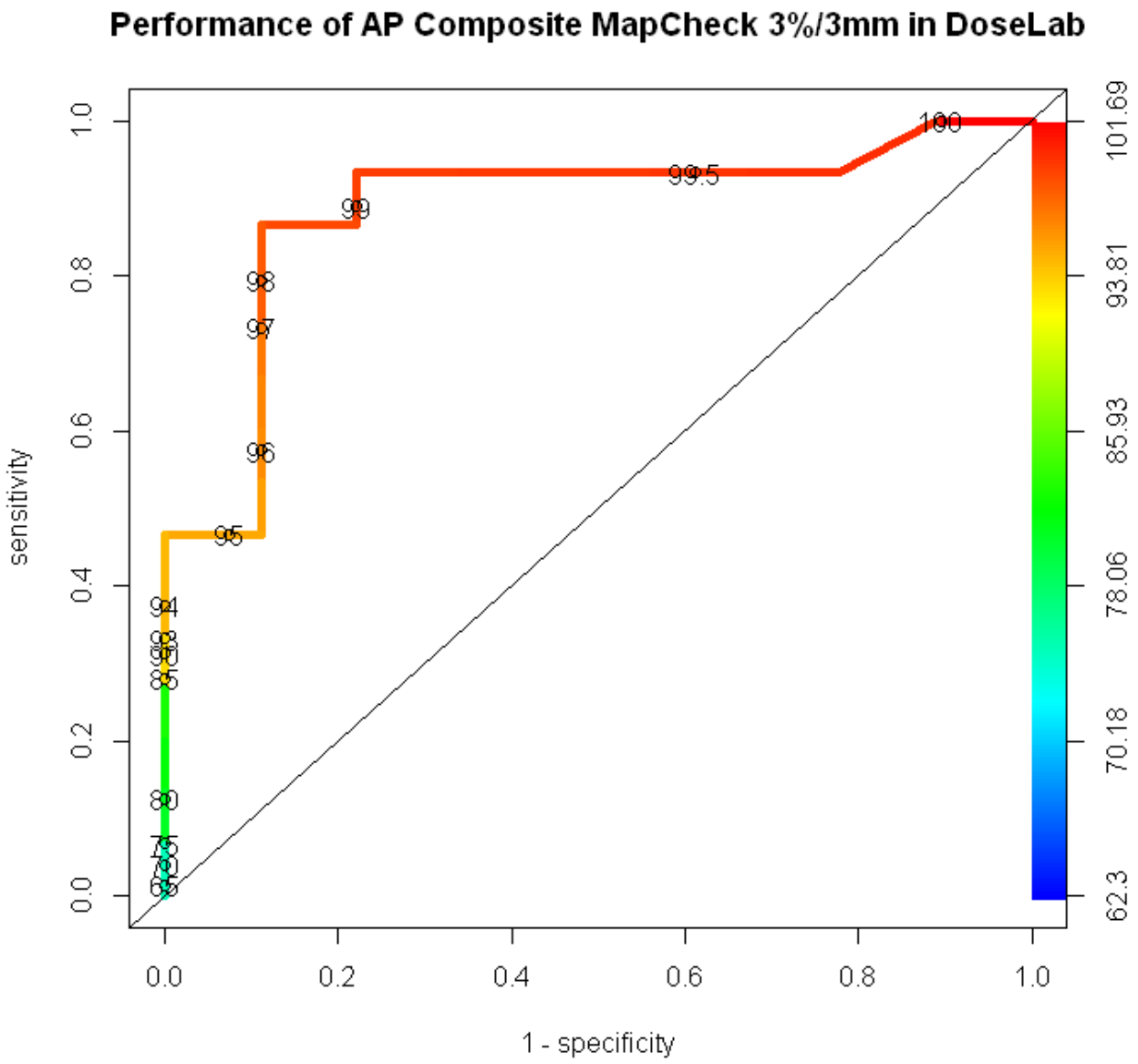

Performance of AP field by field MapCheck 5%/3mm in DoseLab
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
99.94
99.49
99.05
98.61
98.16
97.72

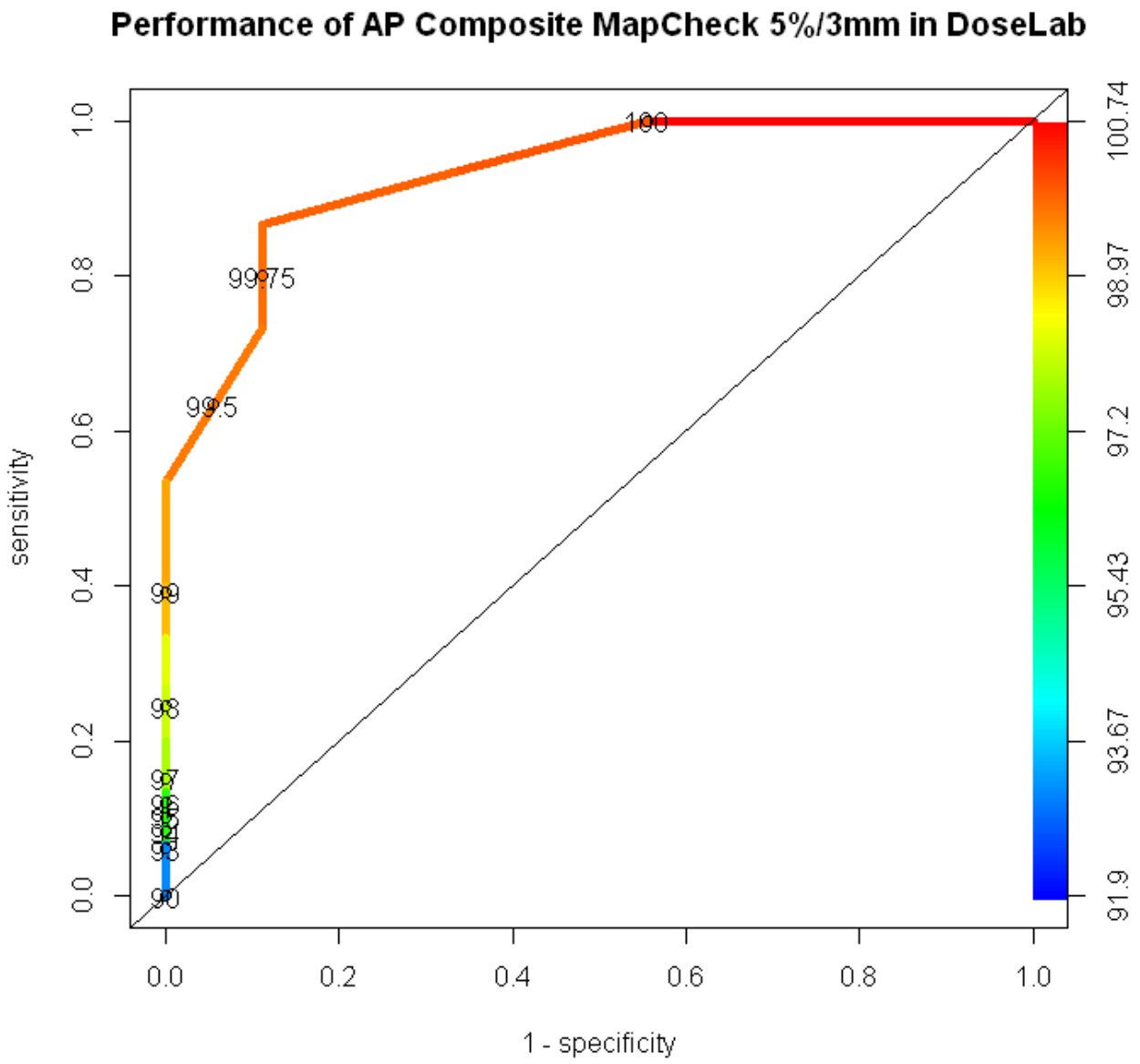

Performance of Rotational MapCheck 2%/2mm in DoseLab
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
91.32
82
72.67
63.35
54.02
44.7
100
90
85
80
75
70
65
50
40


Performance of Rotational MapCheck 3%/3mm in DoseLab
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
100.07
93.16
86.24
79.33
72.41
65.5

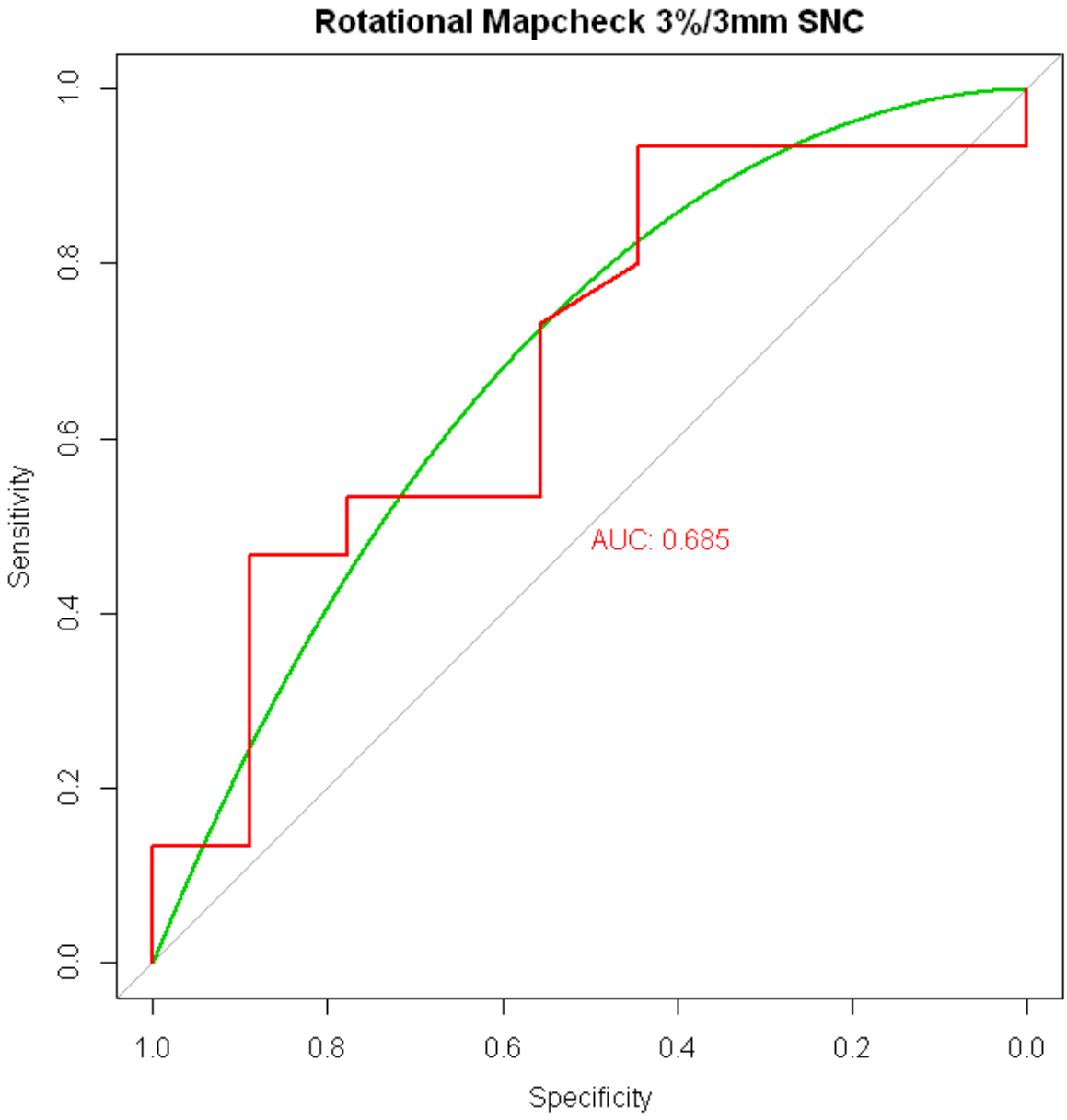

Performance of Rotational MapCheck 5%/3mm in DoseLab
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
100.79
97.59
94.39
91.2
88
84.8
100
99.5
99
98
88

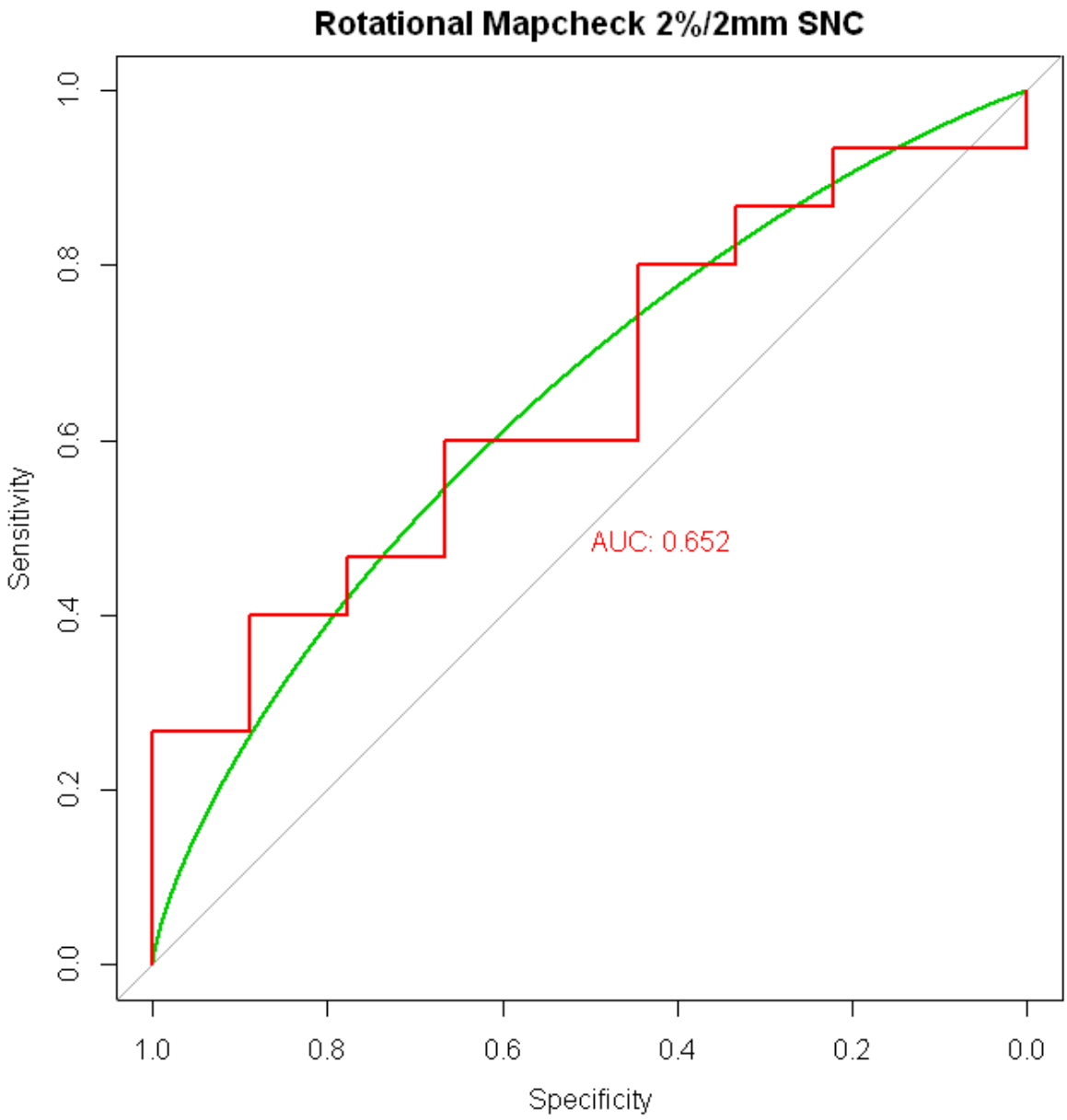

Performance of AP Composite MapCheck 2%/2mm in DoseLab
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
100.23
89.06
77.9
66.73
55.57
44.4
100
95
94
90
85
80
75
70
65
45

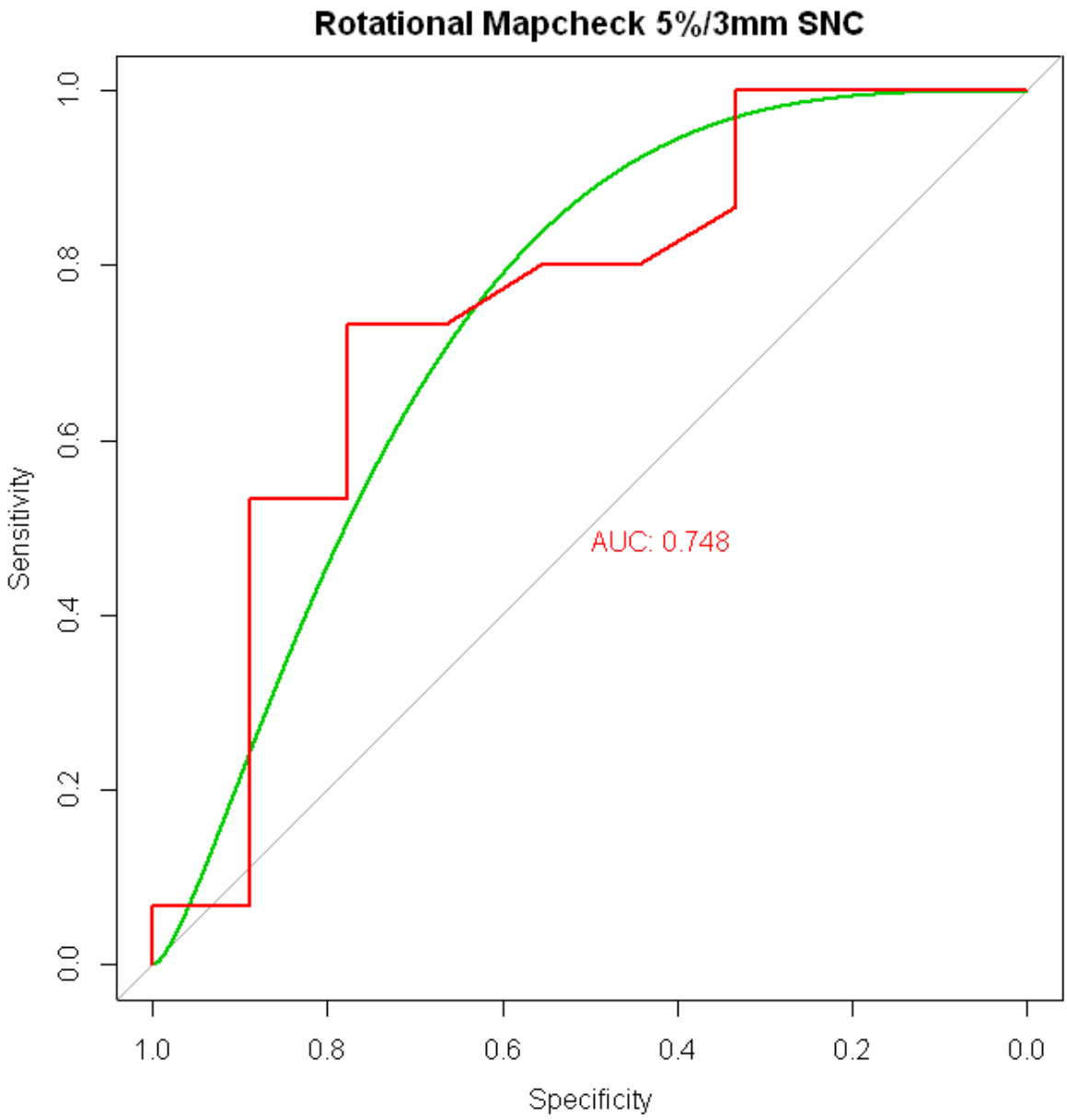

Performance of AP Composite MapCheck 3%/3mm in DoseLab
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
101.69
93.81
85.93
78.06
70.18
62.3
100
99.5
99
98
97
96
95
94
80

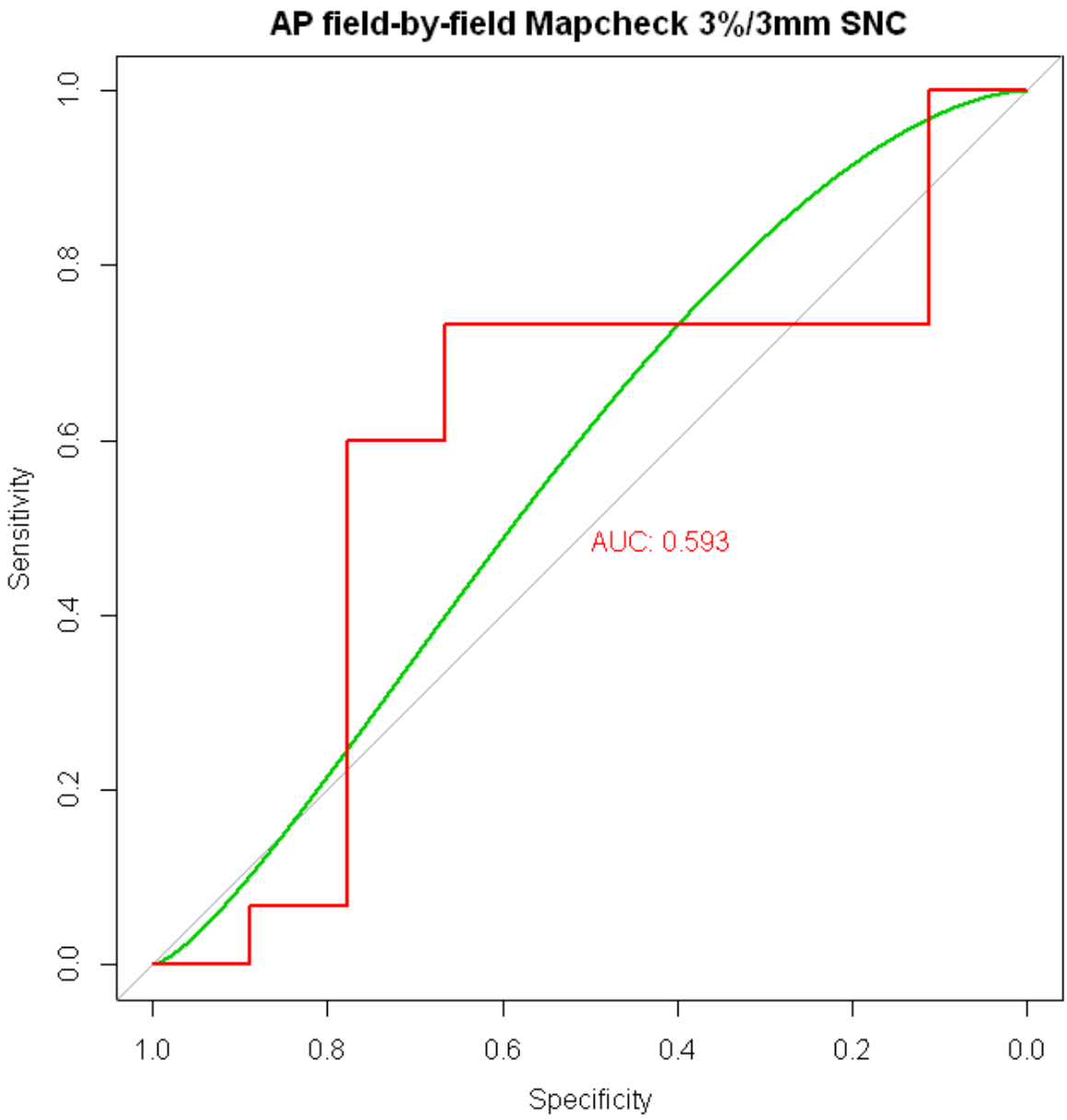

Performance of AP Composite MapCheck 5%/3mm in DoseLab
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
100
99.75
99.5
99
98
100.74
98.97
97.2
95.43
93.67
91.9

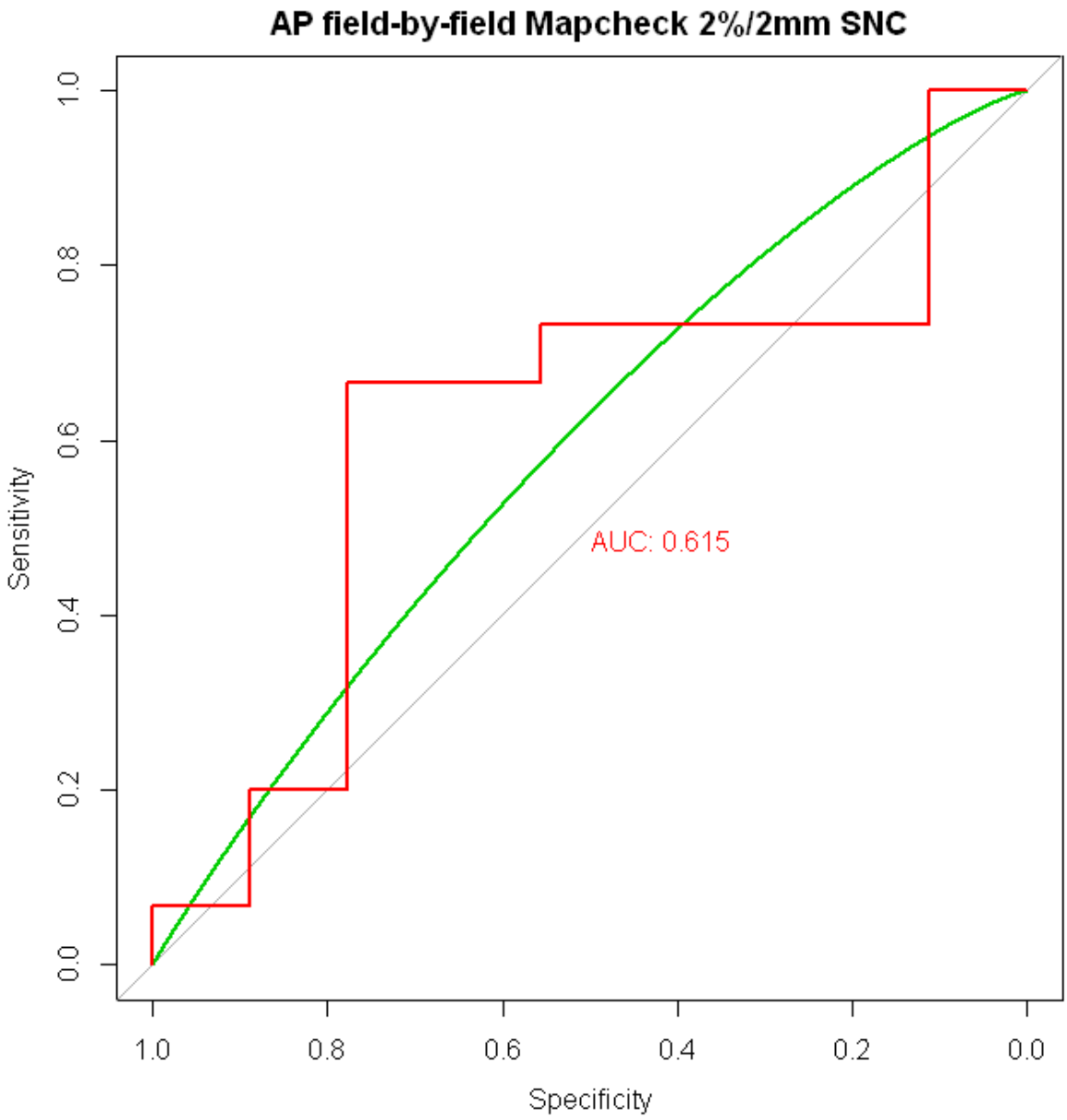
The empirical ROC curves (red) are plotted with their smoothed counterparts (green, binormal smoothing). The Area Under the Curve for each empirical curve is printed on the plot.

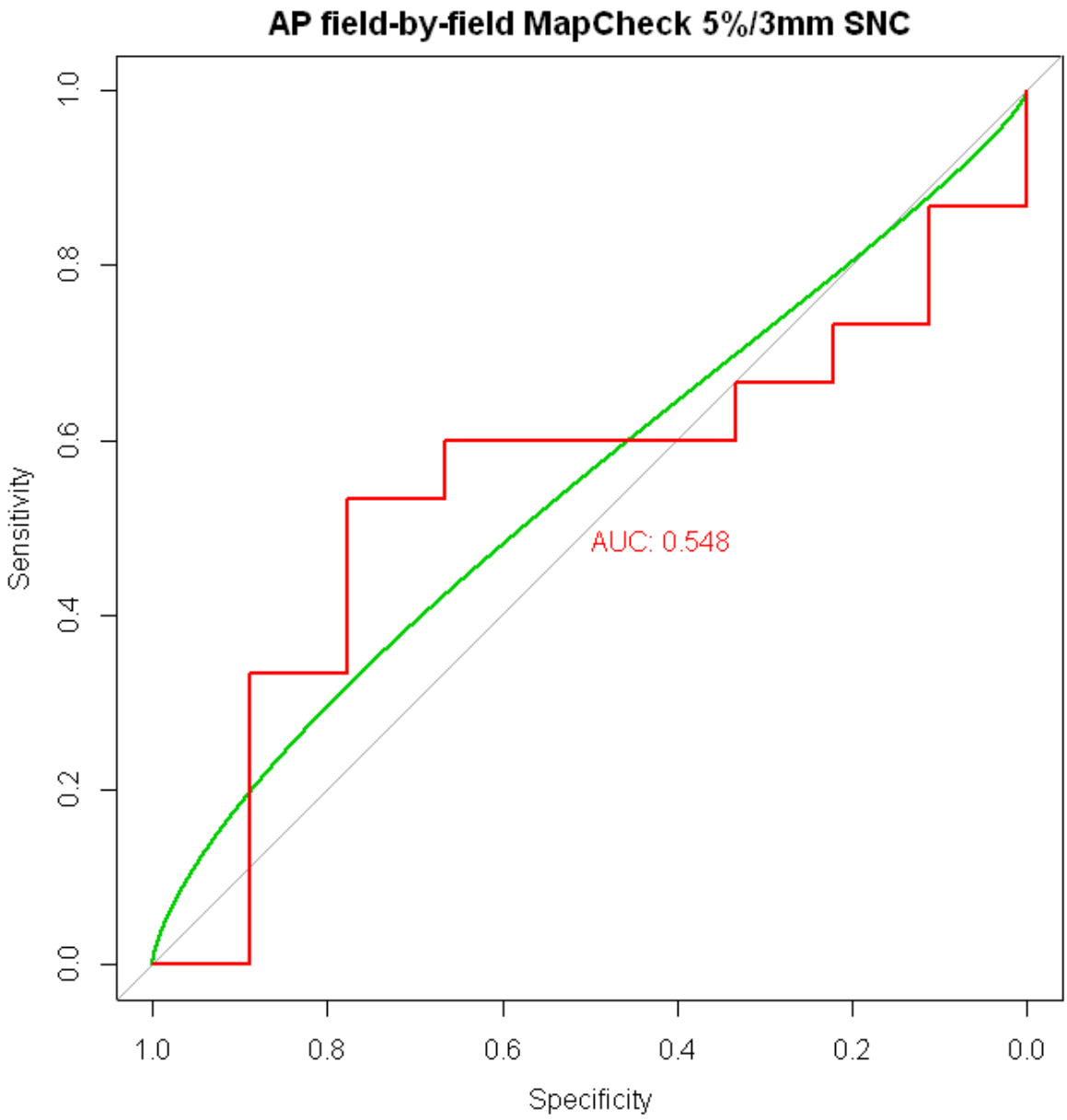


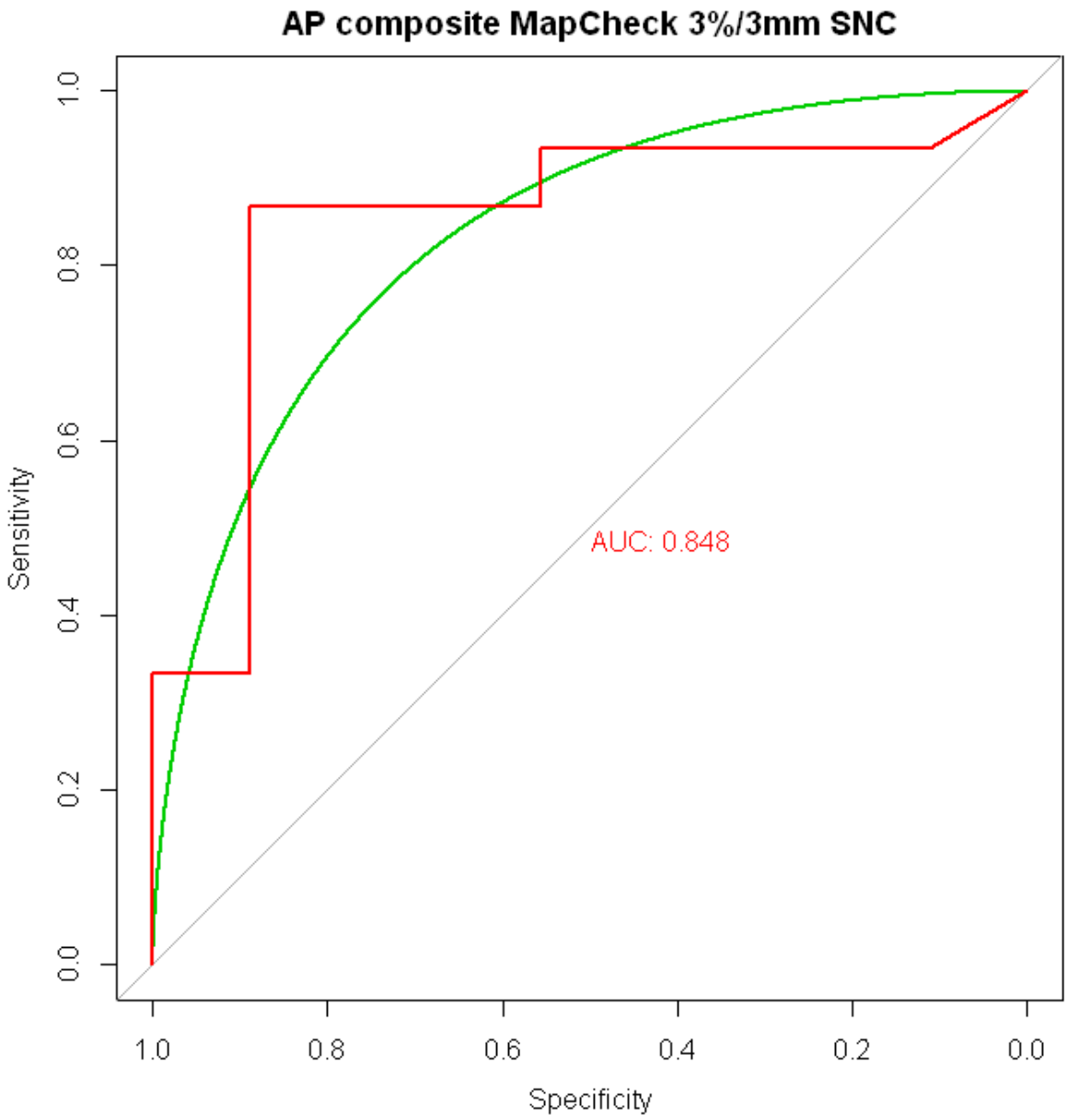

Rotational Mapcheck 2%/2mm SNC
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.652
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

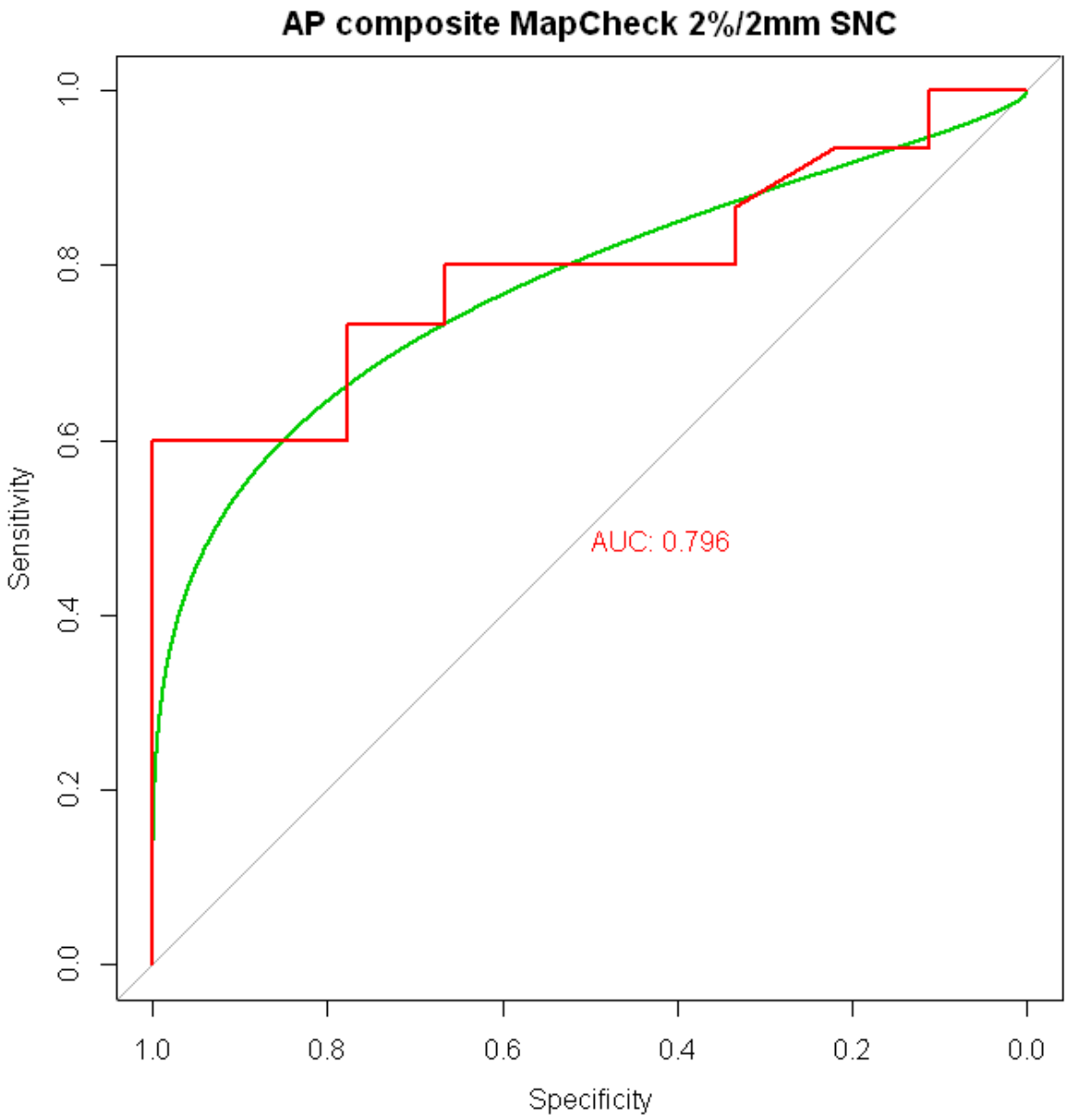

Rotational Mapcheck 5%/3mm SNC
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.748
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

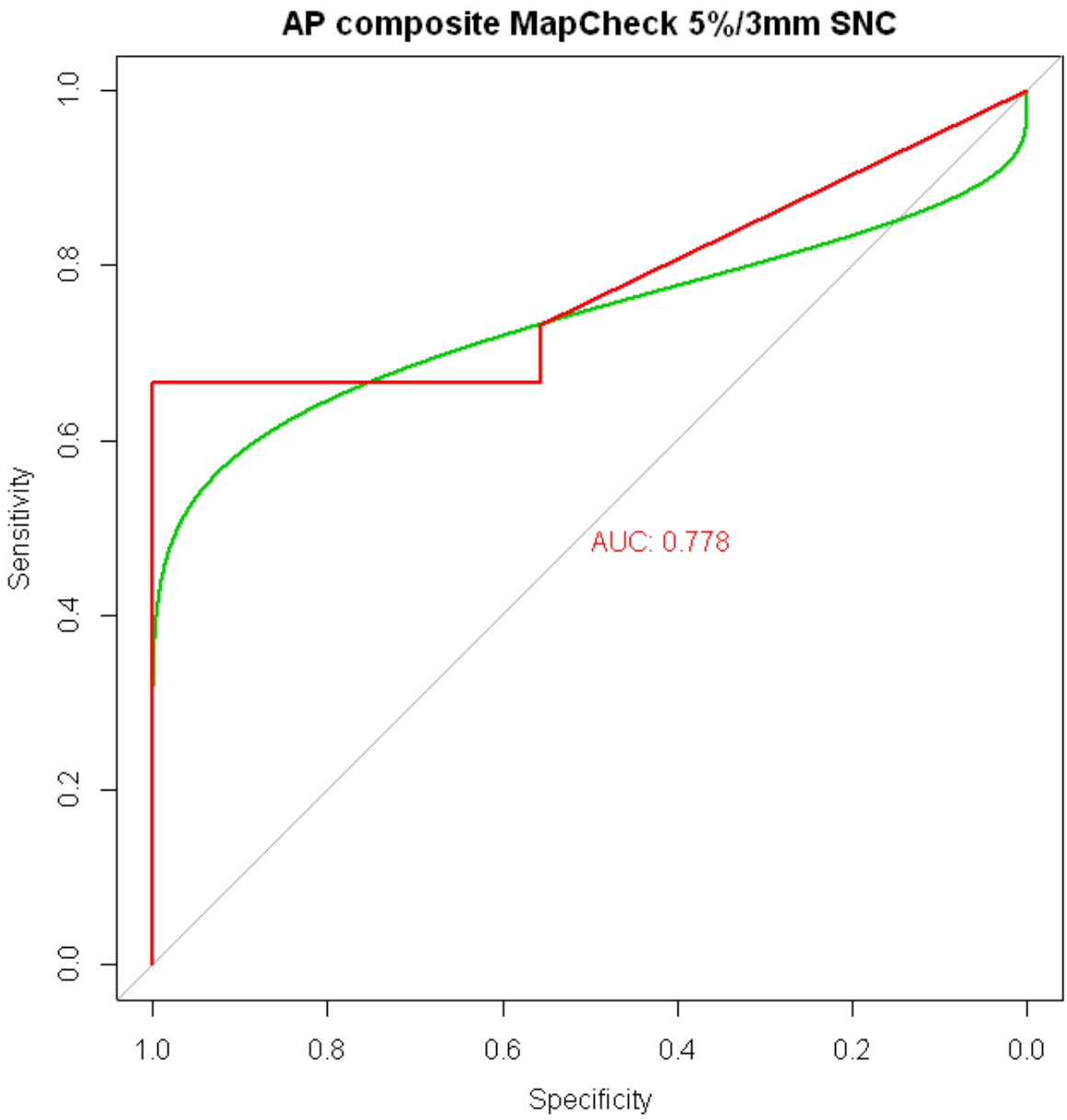

AP field-by-field Mapcheck 3%/3mm SNC
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.593
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

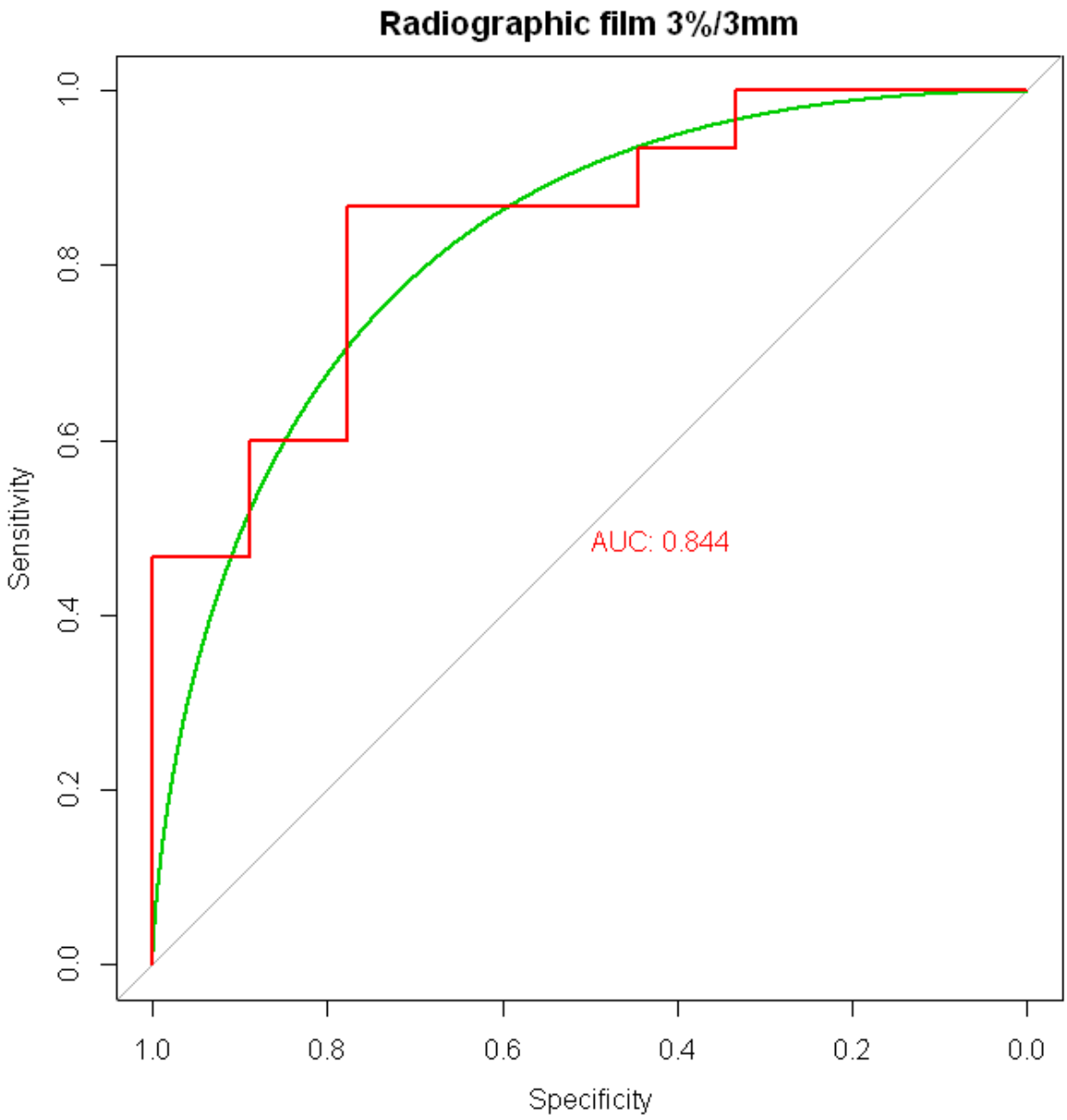

AP field-by-field Mapcheck 2%/2mm SNC
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.615
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

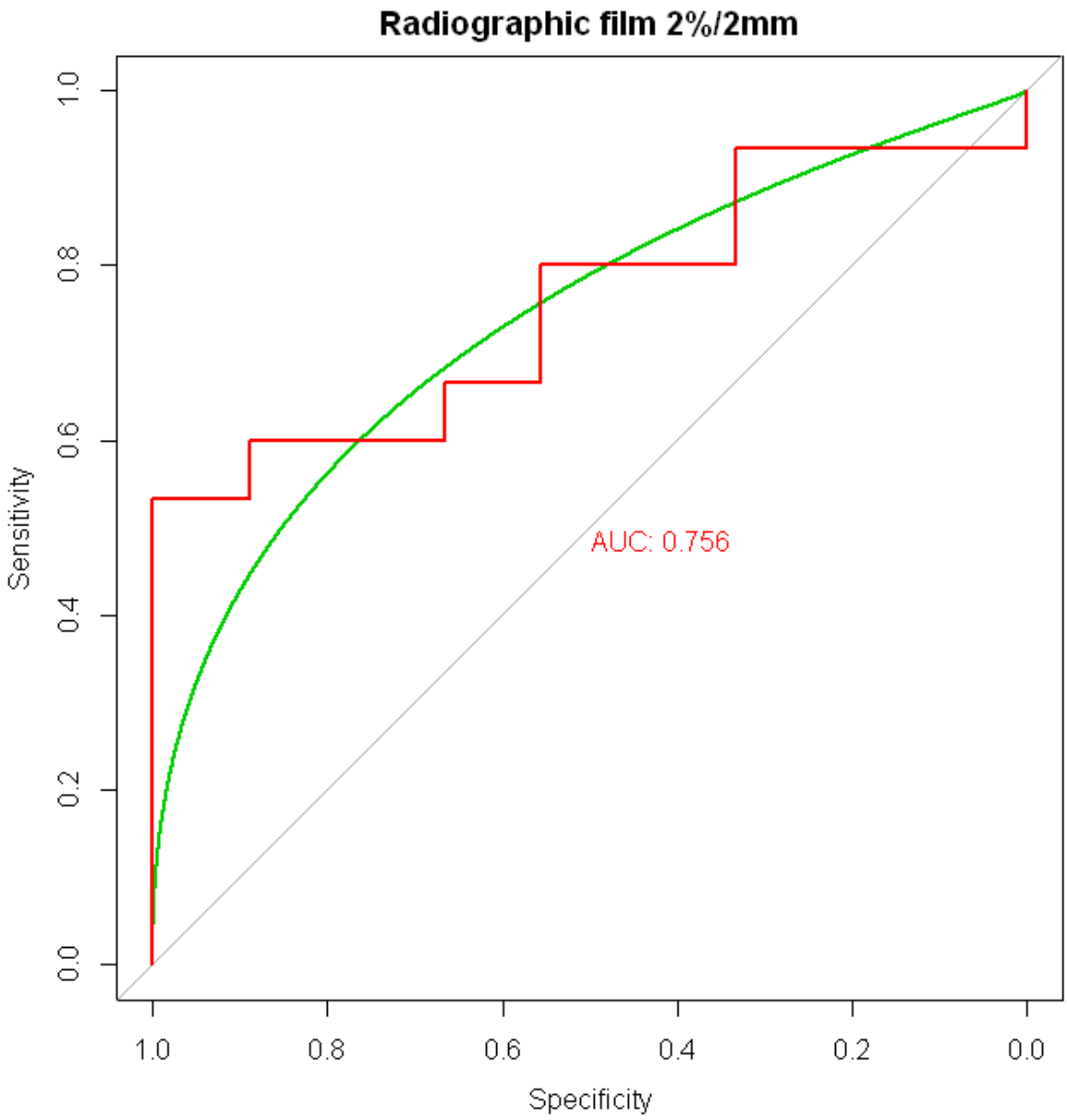

AP field-by-field MapCheck 5%/3mm SNC
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.548
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

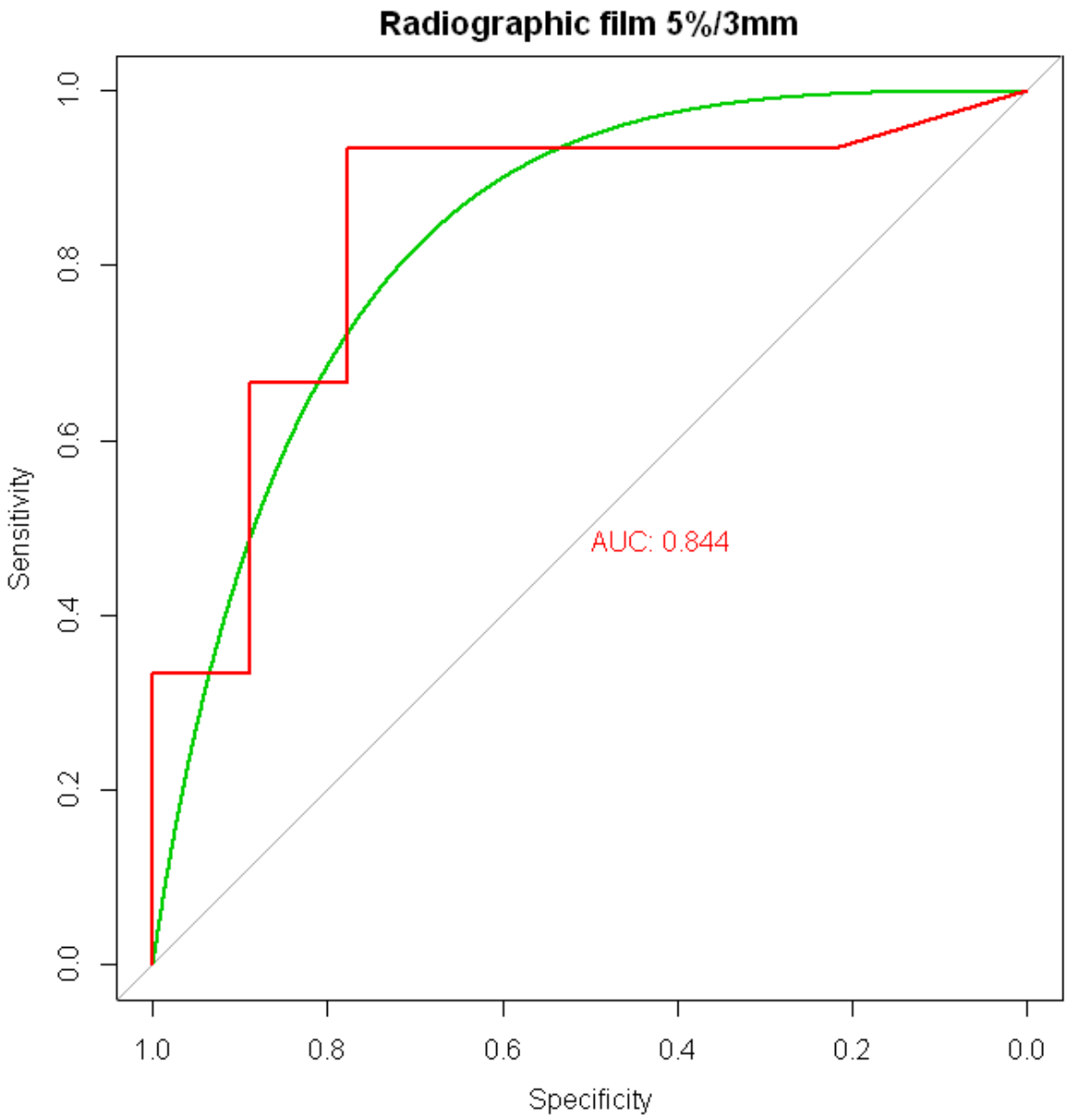

AP composite MapCheck 3%/3mm SNC
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.848

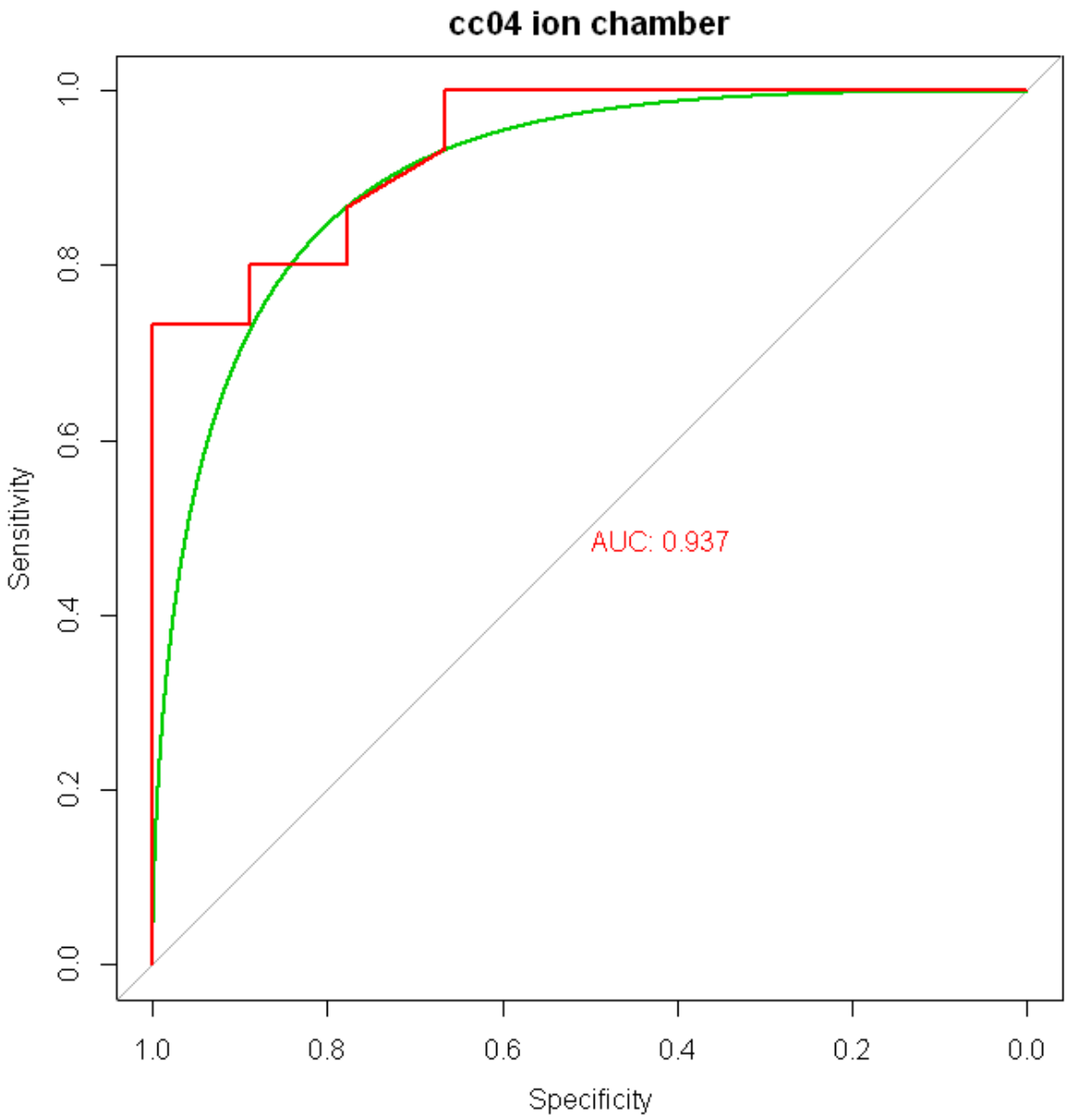

AP composite MapCheck 2%/2mm SNC
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.796

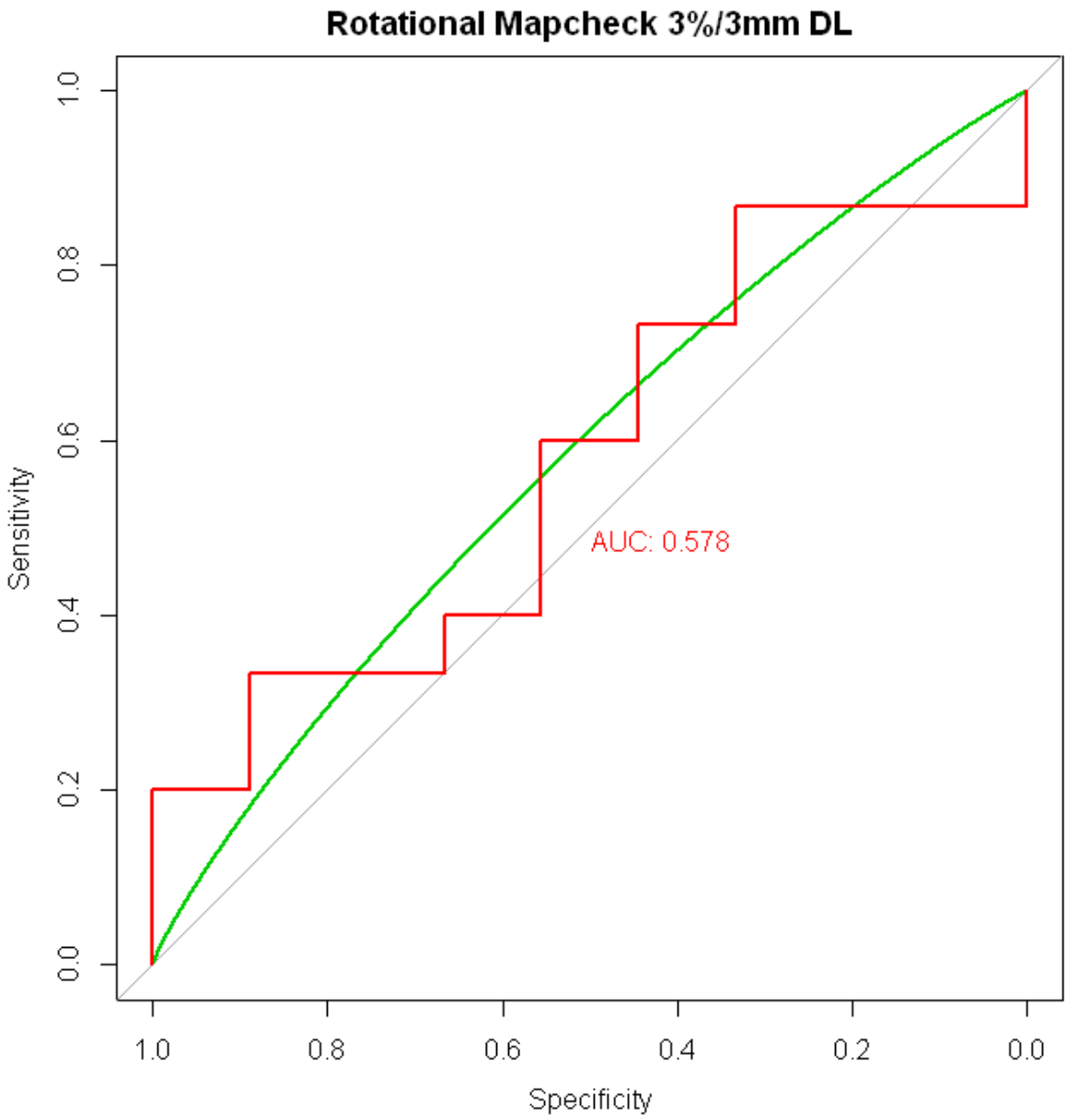

AP composite MapCheck 5%/3mm SNC
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.778
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

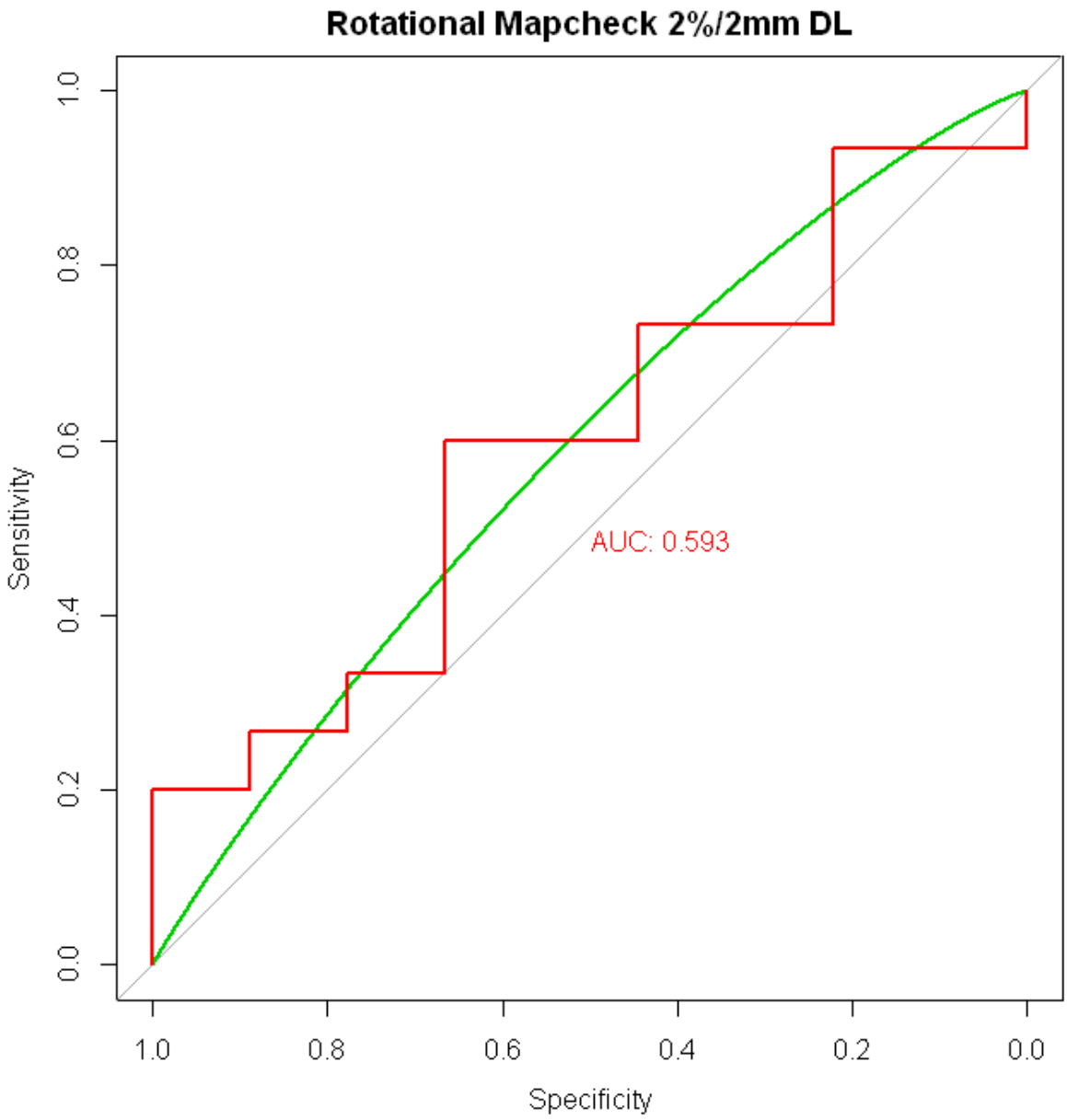

Radiographic film 3%/3mm
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.844

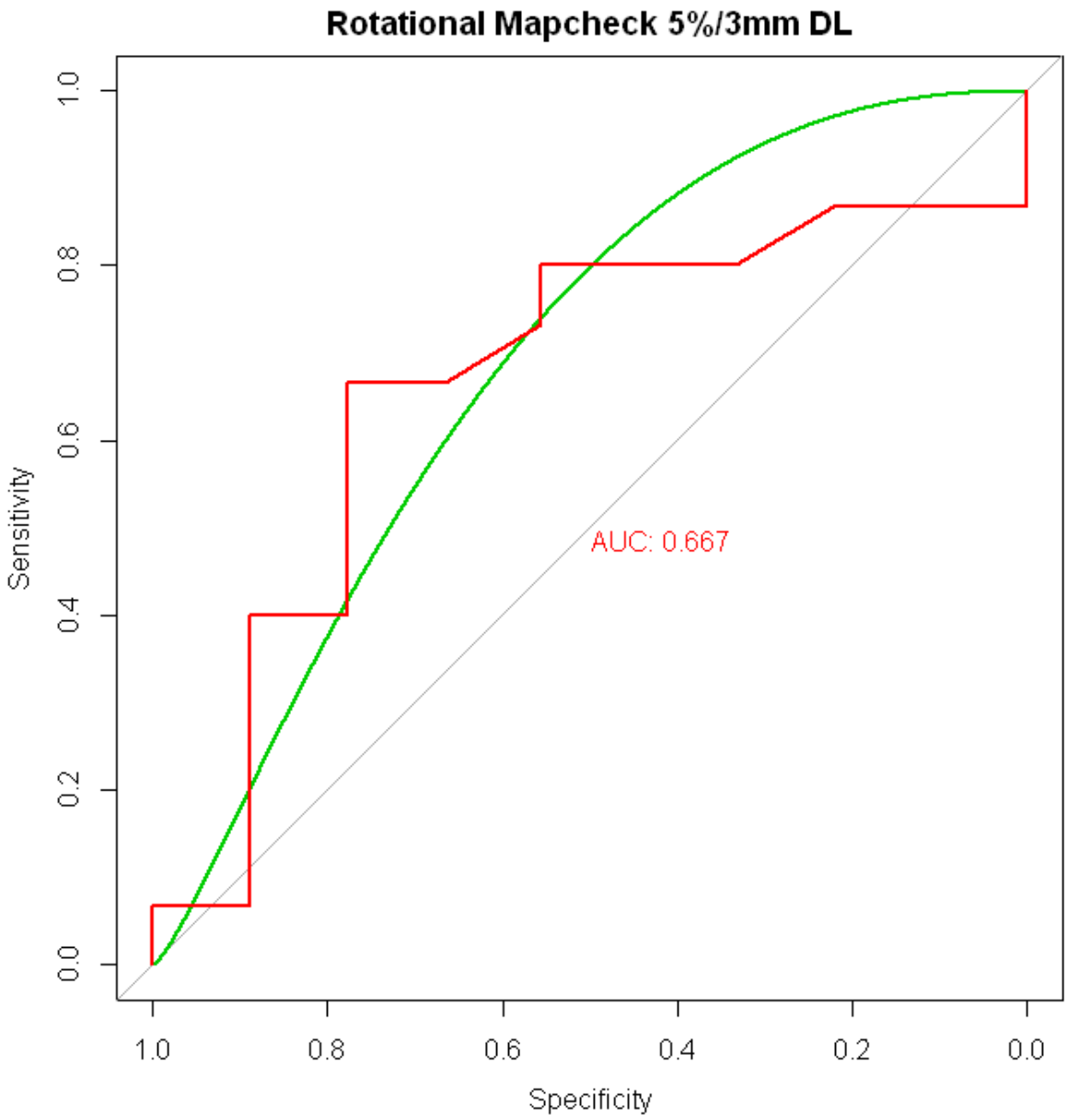

Radiographic film 2%/2mm
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.756

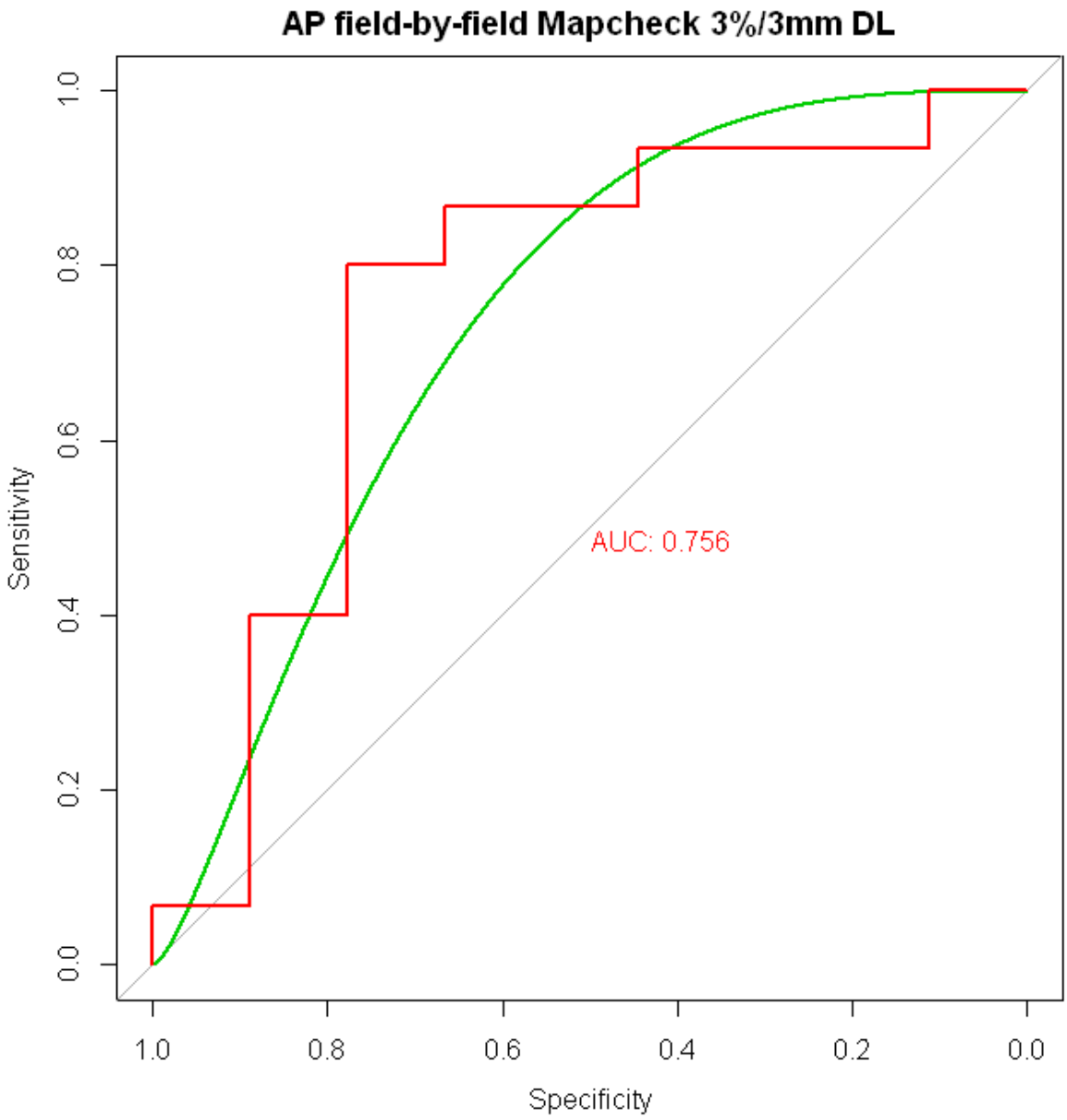

Radiographic film 5%/3mm
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.844

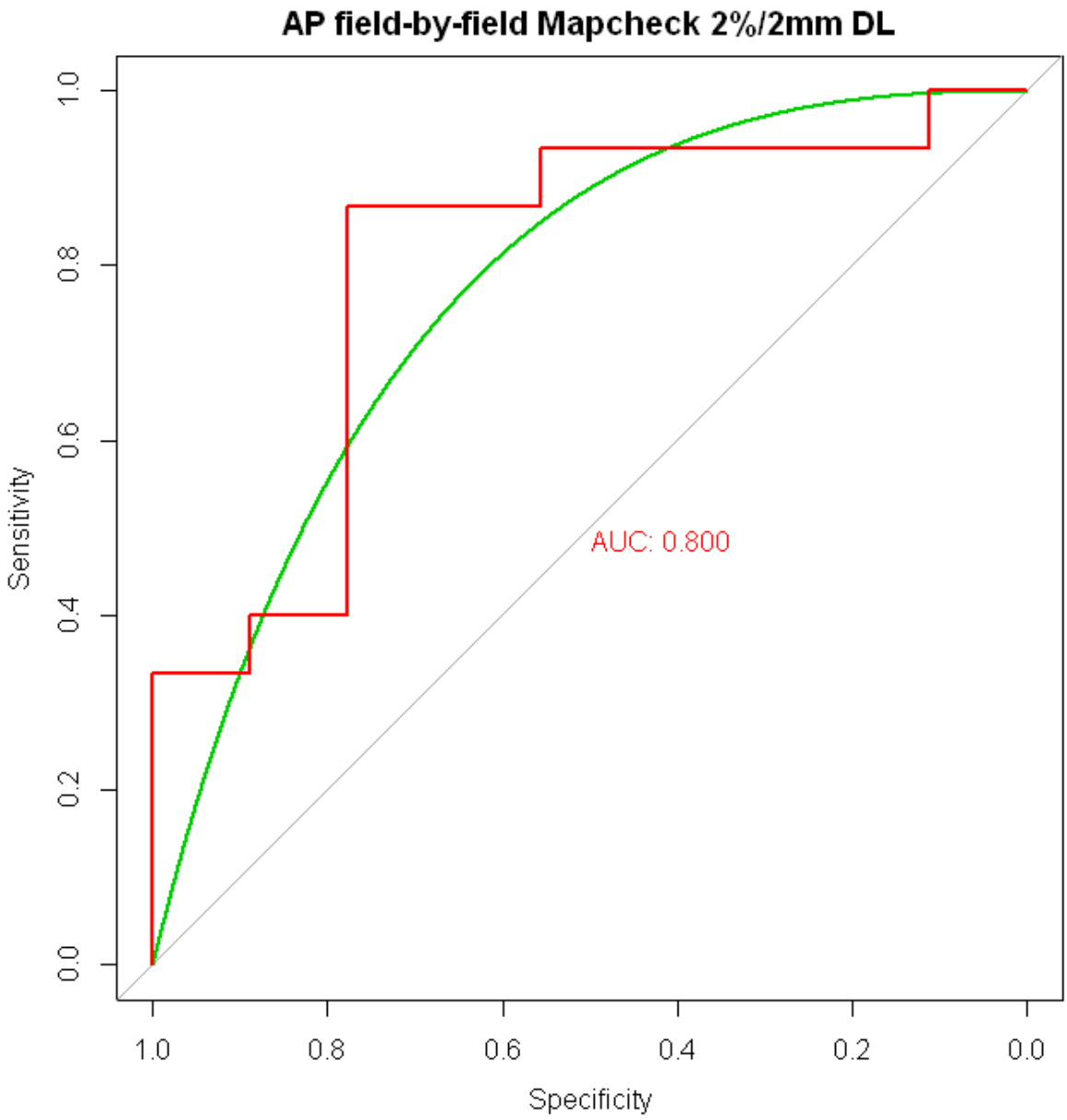

cc04 ion chamber
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.937

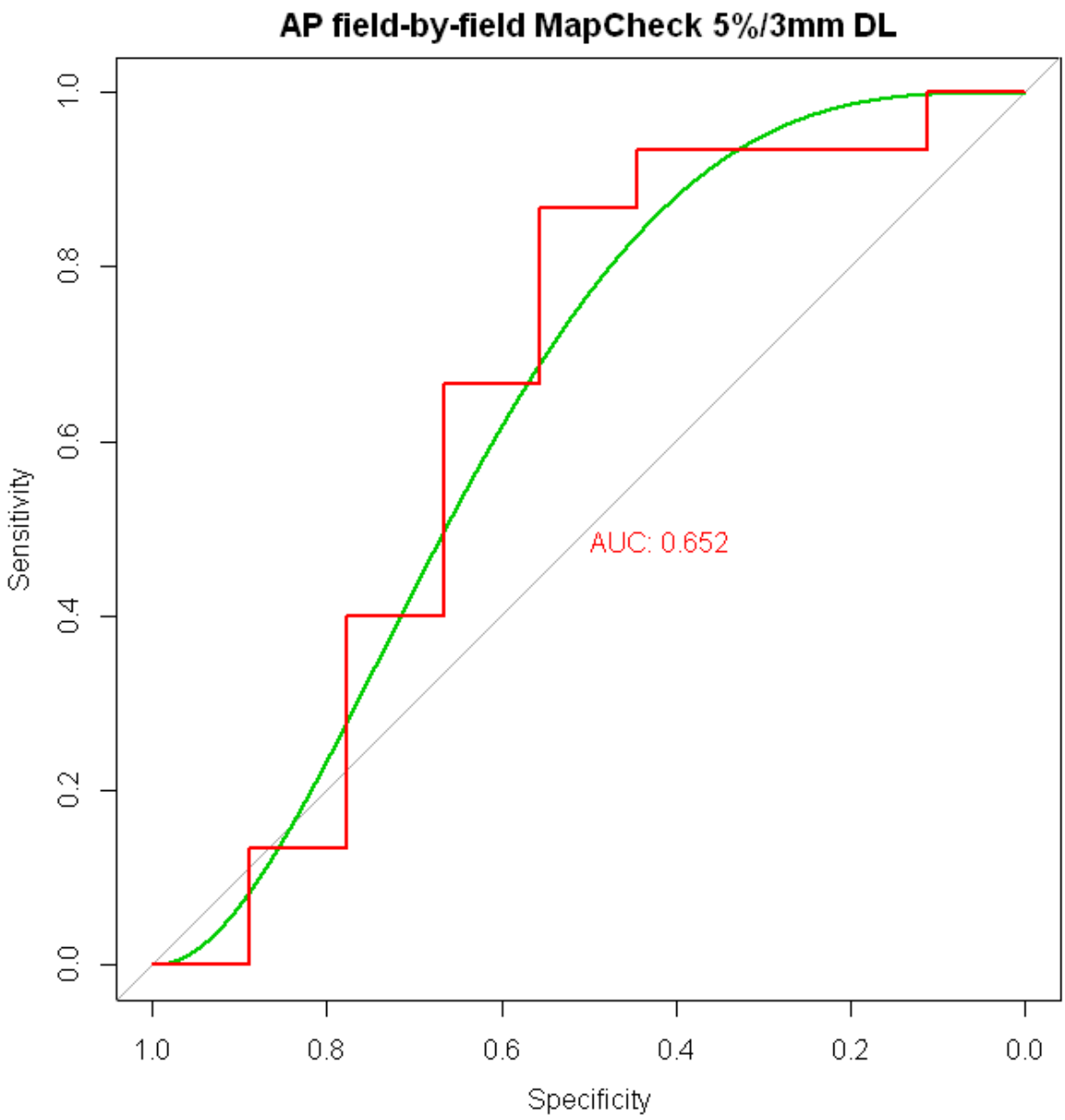

Rotational Mapcheck 3%/3mm DL
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.578
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

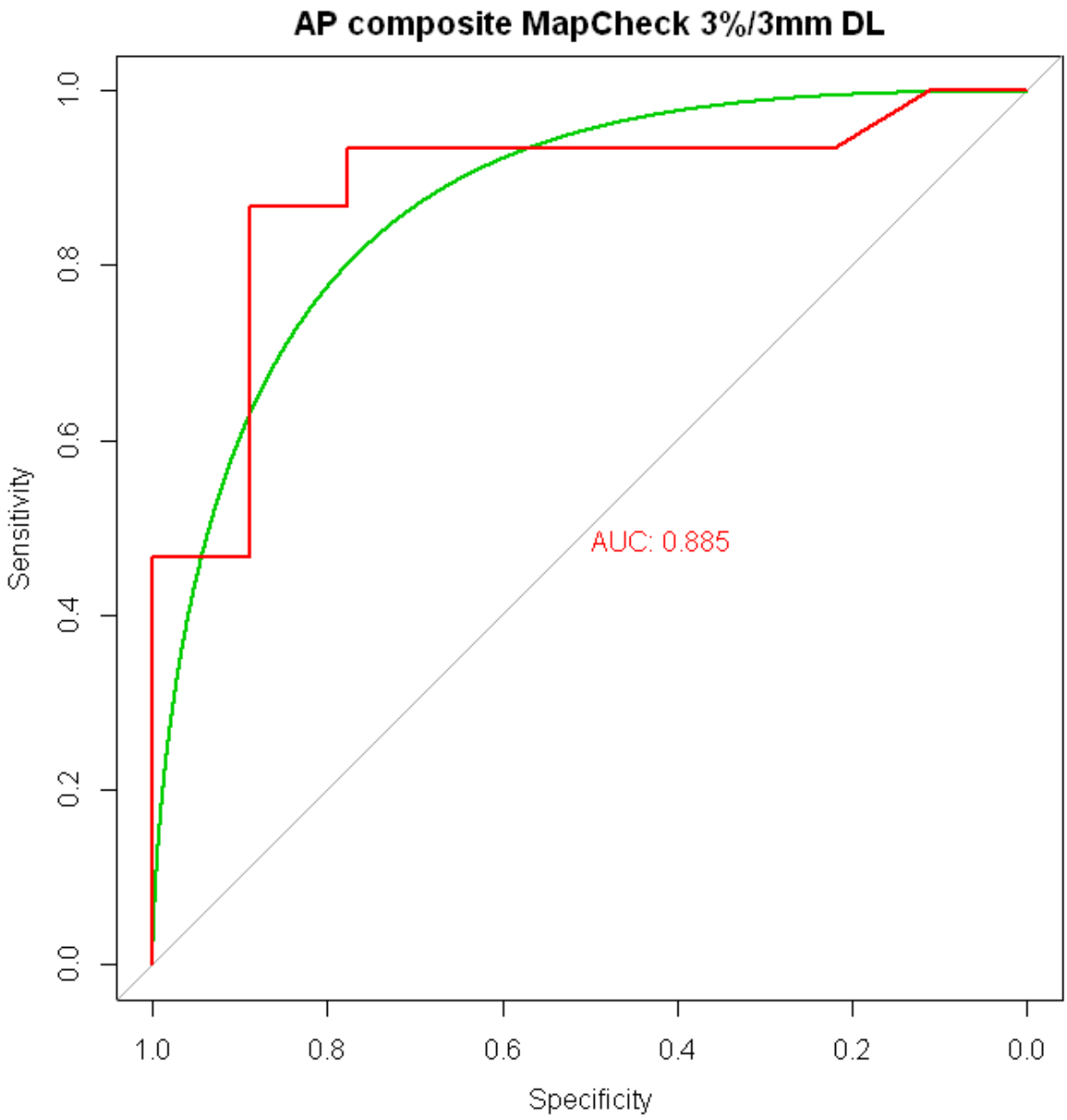

Rotational Mapcheck 2%/2mm DL
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.593
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

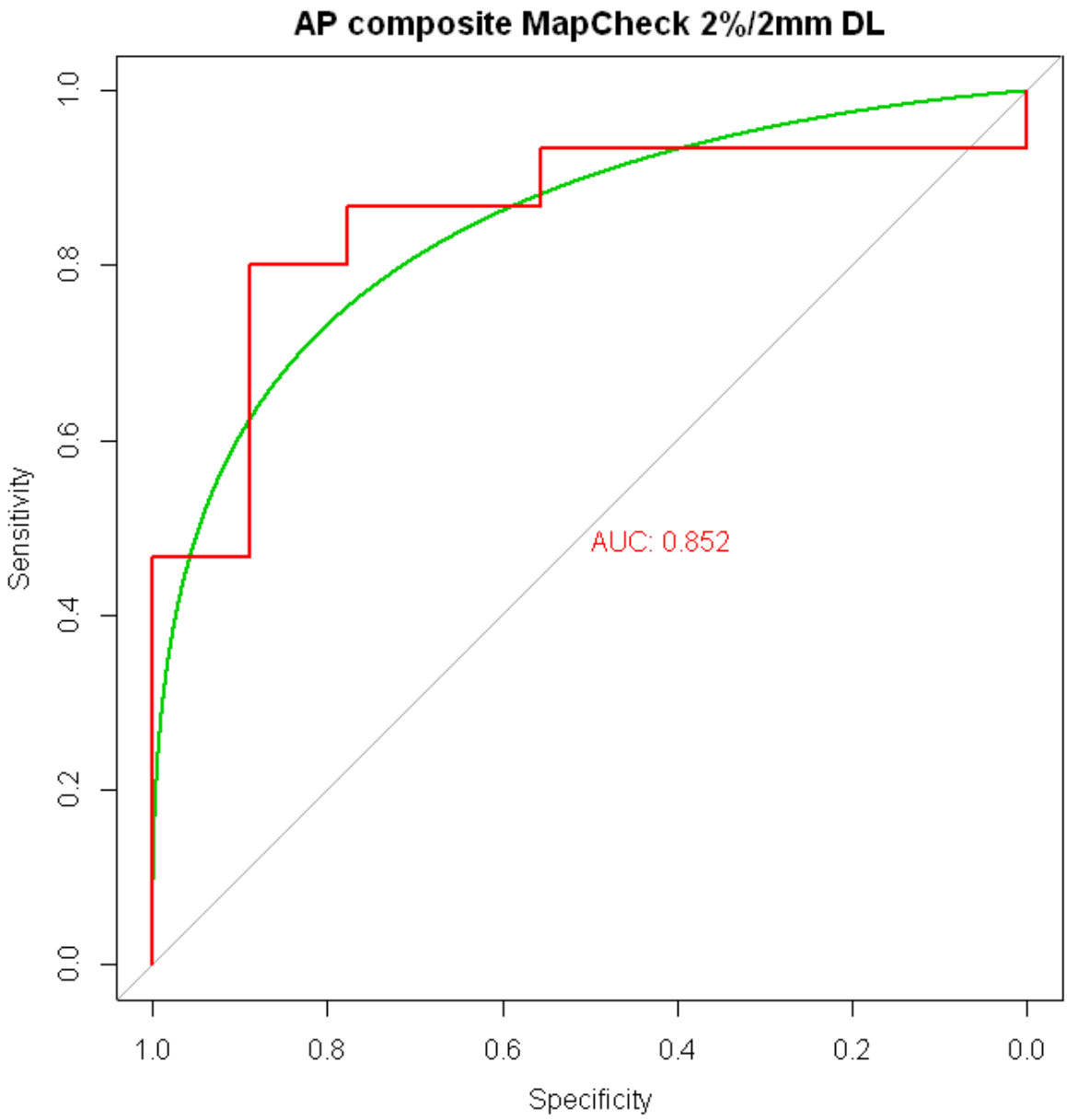

Rotational Mapcheck 5%/3mm DL
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.667
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

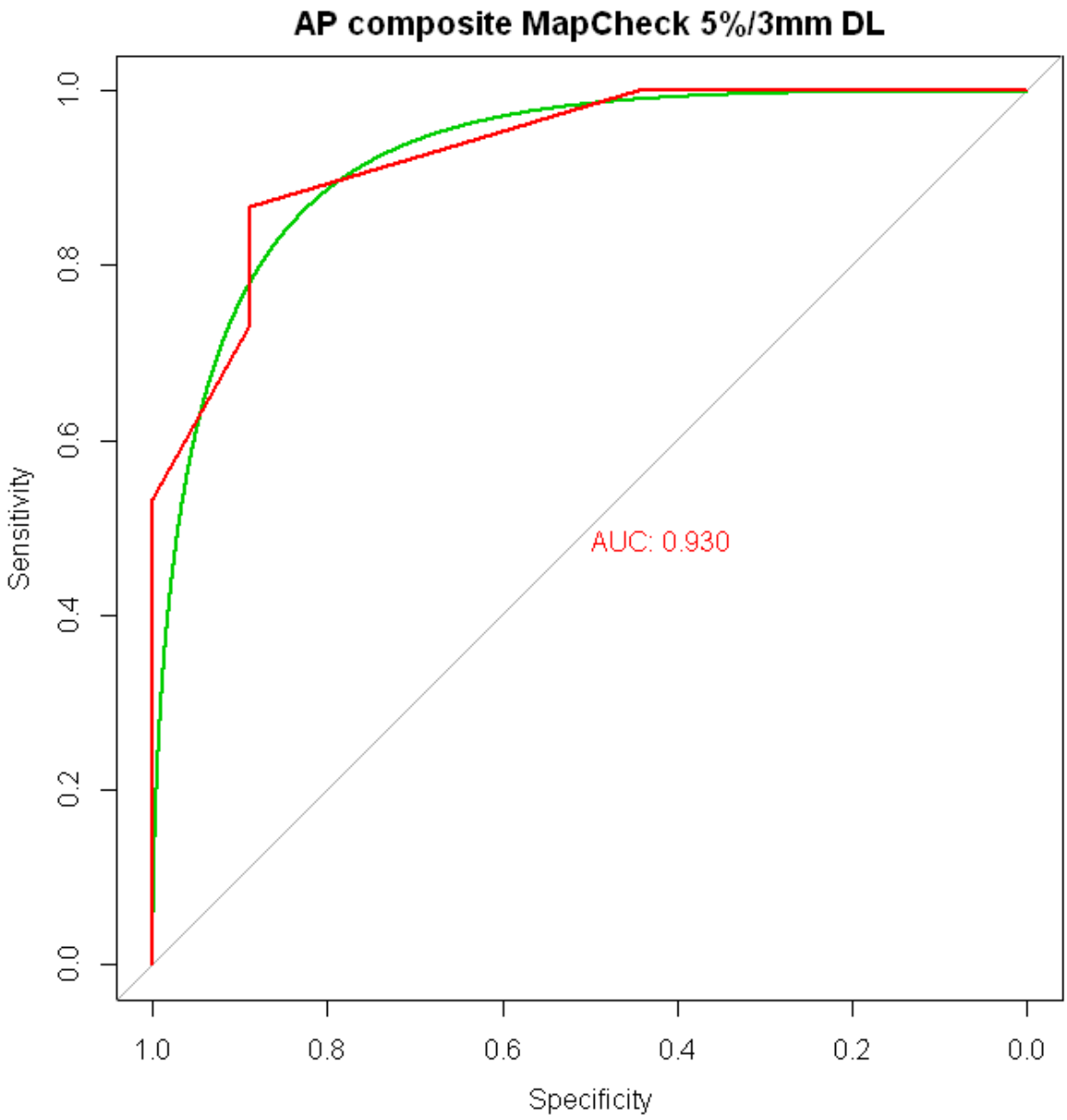

AP field-by-field Mapcheck 3%/3mm DL
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.756
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

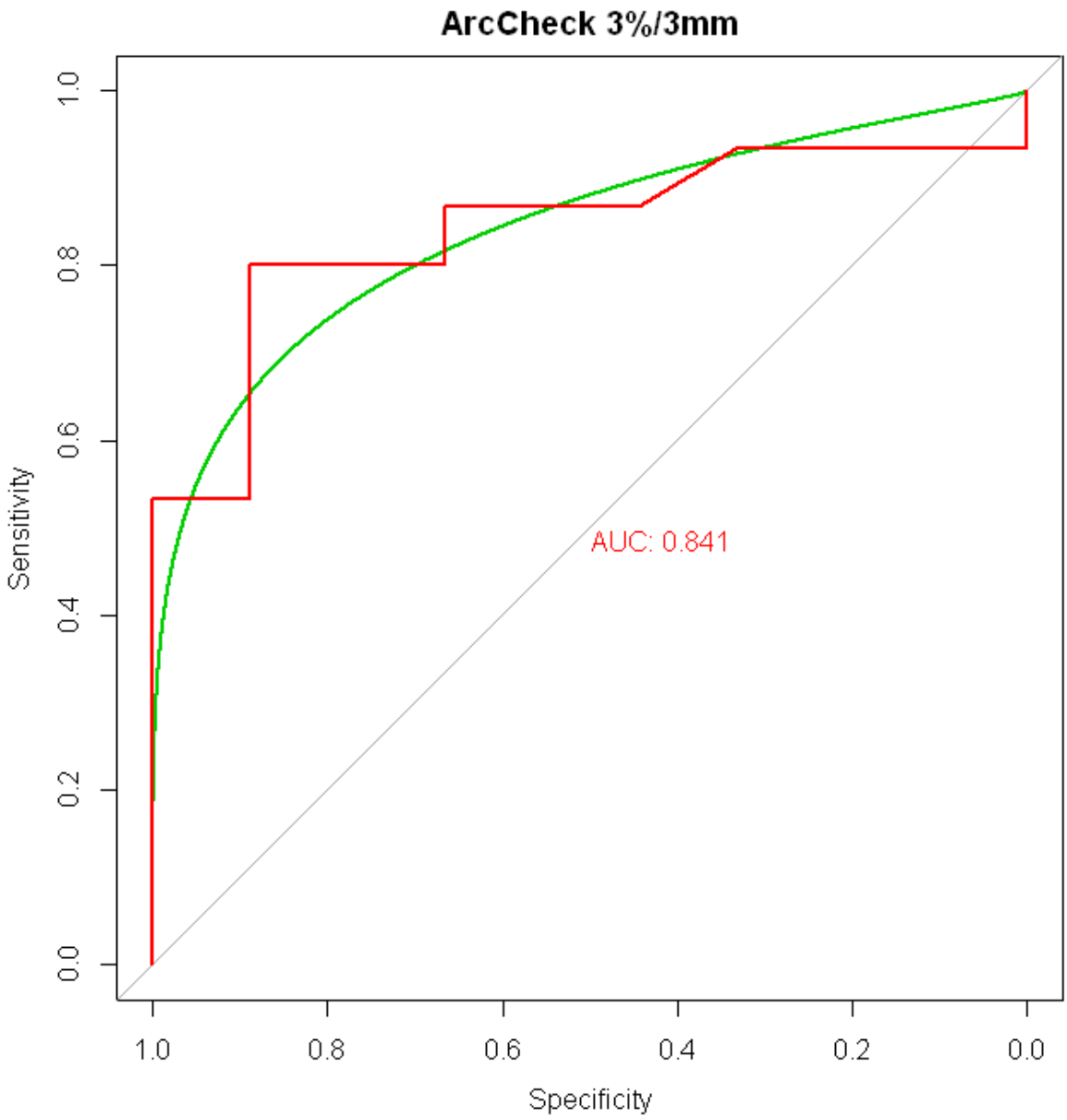

AP field-by-field Mapcheck 2%/2mm DL
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.800
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

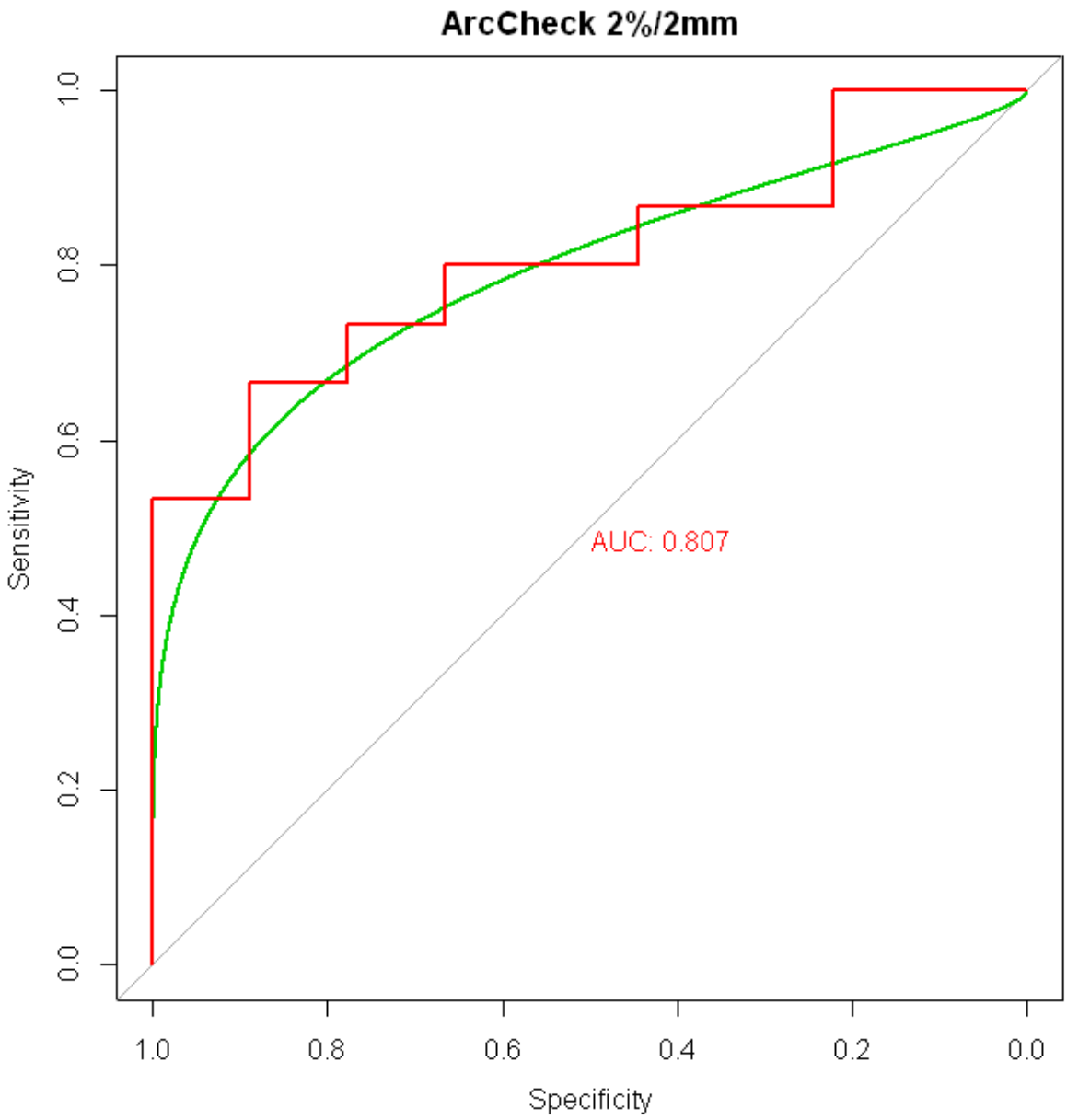

AP field-by-field MapCheck 5%/3mm DL
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.652
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

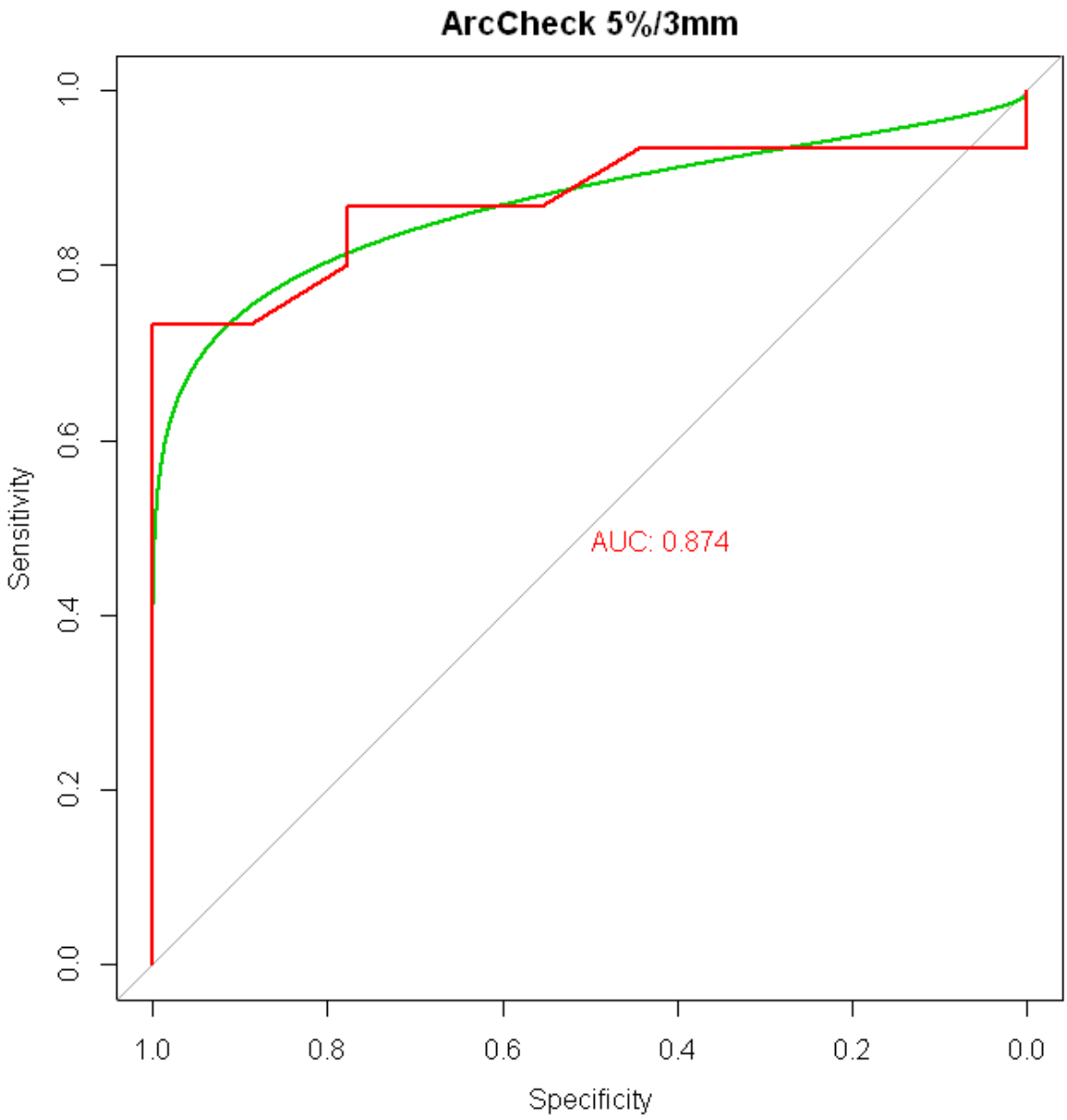

AP composite MapCheck 3%/3mm DL
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.885


AP composite MapCheck 2%/2mm DL
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.852


AP composite MapCheck 5%/3mm DL
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.930


ArcCheck 3%/3mm
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.841
1.0
0.8
0.6
0.4
0.2
0.0
Specificity


ArcCheck 2%/2mm
Sensitivity
Specificity
AUC: 0.807
0.0
0.2
0.4
0.6
0.8
1.0


ArcCheck 5%/3mm
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.874

In order to assess whether a particular dosimetric system's AUC was significantly different from another's, a D-test was performed for each pair of AUC's. The test statistic for the D test is the difference in AUC's divided by the standard deviation of the differences between each bootstrapped AUC. This yielded the following table of p-values. From this table, it was found that there was not a statistically significant difference between any pair of gamma criteria for a given device (i.e. the AUC for film at 2%/2mm was not different from film at 3%/3mm, or 5%/3mm). However, these p-values are for pair-wise comparisons. With so many individual comparisons, there is a non-trivial probability that you will achieve significance in a pair by chance. After adjustment of p-values for this effect, it was found that the AUC results between Doselab and SNC Patient were not significantly different. The findings from this table are discussed on page 65.

D test results comparing AUC of between each device. Results are from bootstraped method from pROC using roc.test

| | $D = \frac{AUC\ 1 - AUC\ 2}{s}$ | D test formula from pROC | s is the standard deviation of the bootstrap differences |
|---|---|---|---|
| | rotationally delivered MapCheck at 3%/3mm (SNC) | rotationally delivered MapCheck at 2%/2mm (SNC) | rotationally delivered MapCheck at 5%/3mm (SNC) |
| rotationally delivered MapCheck at 3%/3mm (SNC) | NA | 0.48795 | 0.41066 |
| rotationally delivered MapCheck at 2%/2mm (SNC) | 0.48795 | NA | 0.30915 |
| rotationally delivered MapCheck at 5%/3mm (SNC) | 0.41066 | 0.30915 | NA |
| AP field-by-field MapCheck at 3%/3mm (SNC) | 0.55291 | 0.70296 | 0.35156 |
| AP field-by-field MapCheck at 2%/2mm (SNC) | 0.66184 | 0.81327 | 0.43339 |
| AP field-by-field MapCheck at 5%/3mm (SNC) | 0.33607 | 0.47665 | 0.18152 |
| AP composite MapCheck at 3%/3mm (SNC) | 0.22831 | 0.12603 | 0.46950 |
| AP composite MapCheck at 2%/2mm (SNC) | 0.42780 | 0.30833 | 0.74878 |
| AP composite MapCheck at 5%/3mm (SNC) | 0.54514 | 0.40724 | 0.84754 |
| EDR2 film at 3%/3mm | 0.25614 | 0.14526 | 0.51440 |
| EDR2 film at 2%/2mm | 0.64709 | 0.50546 | 0.95920 |
| EDR2 film at 5%/3mm | 0.23706 | 0.13853 | 0.53639 |
| cc04 ion chamber | 0.03871 | 0.02043 | 0.08576 |
| rotationally delivered MapCheck at 3%/3mm (DL) | 0.21356 | 0.37106 | 0.14299 |
| rotationally delivered MapCheck at 2%/2mm (DL) | 0.31070 | 0.50306 | 0.22270 |
| rotationally delivered MapCheck at 5%/3mm (DL) | 0.83062 | 0.88677 | 0.33716 |
| AP field-by-field MapCheck at 3%/3mm (DL) | 0.60688 | 0.45417 | 0.96120 |
| AP field-by-field MapCheck at 2%/2mm (DL) | 0.37053 | 0.23983 | 0.70813 |
| AP field-by-field MapCheck at 5%/3mm (DL) | 0.76289 | 1.00000 | 0.47616 |
| AP composite MapCheck at 3%/3mm (DL) | 0.11869 | 0.06513 | 0.32389 |
| AP composite MapCheck at 2%/2mm (DL) | 0.19582 | 0.10472 | 0.46055 |
| AP composite MapCheck at 5%/3mm (DL) | 0.03713 | 0.01622 | 0.11995 |
| ArcCheck at 3%/3mm (SNC) | 0.24336 | 0.15683 | 0.50228 |
| ArcCheck at 2%/2mm (SNC) | 0.40674 | 0.28388 | 0.69010 |
| ArcCheck at 5%/3mm (SNC) | 0.17077 | 0.10963 | 0.37556 |
| Min | 0.03713 | 0.01622 | 0.08576 |
| Max | 0.83062 | 1.00000 | 0.96120 |

D test results comparing AUC of between each device. Results are from bootstraped method from pROC using roc.test

| | AP field-by-field MapCheck at 3%/3mm (SNC) | AP field-by-field MapCheck at 2%/2mm (SNC) | AP field-by-field MapCheck at 5%/3mm (SNC) |
|---|---|---|---|
| rotationally delivered MapCheck at 3%/3mm (SNC) | 0.55291 | 0.66184 | 0.33607 |
| rotationally delivered MapCheck at 2%/2mm (SNC) | 0.70296 | 0.81327 | 0.47665 |
| rotationally delivered MapCheck at 5%/3mm (SNC) | 0.35156 | 0.43339 | 0.18152 |
| AP field-by-field MapCheck at 3%/3mm (SNC) | NA | 0.51182 | 0.65850 |
| AP field-by-field MapCheck at 2%/2mm (SNC) | 0.51182 | NA | 0.52219 |
| AP field-by-field MapCheck at 5%/3mm (SNC) | 0.65850 | 0.52219 | NA |
| AP composite MapCheck at 3%/3mm (SNC) | 0.01501 | 0.01863 | 0.00402 |
| AP composite MapCheck at 2%/2mm (SNC) | 0.02124 | 0.02680 | 0.02075 |
| AP composite MapCheck at 5%/3mm (SNC) | 0.10792 | 0.11683 | 0.02760 |
| EDR2 film at 3%/3mm | 0.02251 | 0.02693 | 0.00633 |
| EDR2 film at 2%/2mm | 0.23076 | 0.24807 | 0.07986 |
| EDR2 film at 5%/3mm | 0.05337 | 0.07047 | 0.01760 |
| cc04 ion chamber | 0.01540 | 0.01489 | 0.00274 |
| rotationally delivered MapCheck at 3%/3mm (DL) | 0.92720 | 0.81720 | 0.83241 |
| rotationally delivered MapCheck at 2%/2mm (DL) | 1.00000 | 0.87001 | 0.72825 |
| rotationally delivered MapCheck at 5%/3mm (DL) | 0.65845 | 0.75860 | 0.36778 |
| AP field-by-field MapCheck at 3%/3mm (DL) | 0.01304 | 0.04728 | 0.11748 |
| AP field-by-field MapCheck at 2%/2mm (DL) | 0.00972 | 0.02071 | 0.04933 |
| AP field-by-field MapCheck at 5%/3mm (DL) | 0.54182 | 0.74328 | 0.41504 |
| AP composite MapCheck at 3%/3mm (DL) | 0.00581 | 0.00479 | 0.01188 |
| AP composite MapCheck at 2%/2mm (DL) | 0.01205 | 0.01716 | 0.02422 |
| AP composite MapCheck at 5%/3mm (DL) | 0.00465 | 0.00723 | 0.00047 |
| ArcCheck at 3%/3mm (SNC) | 0.02329 | 0.02800 | 0.02824 |
| ArcCheck at 2%/2mm (SNC) | 0.08379 | 0.09448 | 0.06497 |
| ArcCheck at 5%/3mm (SNC) | 0.02267 | 0.02824 | 0.01298 |
| Min | 0.00465 | 0.00479 | 0.00047 |
| Max | 1.00000 | 0.87001 | 0.83241 |

D test results comparing AUC of between each device. Results are from bootstraped method from pROC using roc.test

| | AP composite MapCheck at 3%/3mm (SNC) | AP composite MapCheck at 2%/2mm (SNC) | AP composite MapCheck at 5%/3mm (SNC) |
|---|---|---|---|
| rotationally delivered MapCheck at 3%/3mm (SNC) | 0.22831 | 0.42780 | 0.54514 |
| rotationally delivered MapCheck at 2%/2mm (SNC) | 0.12603 | 0.30833 | 0.40724 |
| rotationally delivered MapCheck at 5%/3mm (SNC) | 0.46950 | 0.74878 | 0.84754 |
| AP field-by-field MapCheck at 3%/3mm (SNC) | 0.01501 | 0.02124 | 0.10792 |
| AP field-by-field MapCheck at 2%/2mm (SNC) | 0.01863 | 0.02680 | 0.11683 |
| AP field-by-field MapCheck at 5%/3mm (SNC) | 0.00402 | 0.02075 | 0.02760 |
| AP composite MapCheck at 3%/3mm (SNC) | NA | 0.49012 | 0.43077 |
| AP composite MapCheck at 2%/2mm (SNC) | 0.49012 | NA | 0.80114 |
| AP composite MapCheck at 5%/3mm (SNC) | 0.43077 | 0.80114 | NA |
| EDR2 film at 3%/3mm | 0.94916 | 0.51152 | 0.42794 |
| EDR2 film at 2%/2mm | 0.28579 | 0.64319 | 0.80605 |
| EDR2 film at 5%/3mm | 0.96231 | 0.61555 | 0.46740 |
| cc04 ion chamber | 0.33030 | 0.16948 | 0.09873 |
| rotationally delivered MapCheck at 3%/3mm (DL) | 0.04014 | 0.11569 | 0.16278 |
| rotationally delivered MapCheck at 2%/2mm (DL) | 0.03267 | 0.09544 | 0.17985 |
| rotationally delivered MapCheck at 5%/3mm (DL) | 0.20737 | 0.36581 | 0.43247 |
| AP field-by-field MapCheck at 3%/3mm (DL) | 0.32492 | 0.66816 | 0.86195 |
| AP field-by-field MapCheck at 2%/2mm (DL) | 0.48920 | 0.96546 | 0.84471 |
| AP field-by-field MapCheck at 5%/3mm (DL) | 0.08781 | 0.24246 | 0.40870 |
| AP composite MapCheck at 3%/3mm (DL) | 0.64179 | 0.25108 | 0.28450 |
| AP composite MapCheck at 2%/2mm (DL) | 0.96280 | 0.46348 | 0.47799 |
| AP composite MapCheck at 5%/3mm (DL) | 0.11613 | 0.10628 | 0.06194 |
| ArcCheck at 3%/3mm (SNC) | 0.93894 | 0.58294 | 0.51122 |
| ArcCheck at 2%/2mm (SNC) | 0.72616 | 0.90562 | 0.76373 |
| ArcCheck at 5%/3mm (SNC) | 0.75390 | 0.32542 | 0.31843 |
| Min | 0.00402 | 0.02075 | 0.02760 |
| Max | 0.96280 | 0.96546 | 0.86195 |

D test results comparing AUC of between each device. Results are from bootstraped method from pROC using roc.test

| | EDR2 film at 3%/3mm | EDR2 film at 2%/2mm | EDR2 film at 5%/3mm | cc04 ion chamber | rotationally delivered MapCheck at 3%/3mm (DL) |
|---|---|---|---|---|---|
| rotationally delivered MapCheck at 3%/3mm (SNC) | 0.25614 | 0.64709 | 0.23706 | 0.03871 | 0.21356 |
| rotationally delivered MapCheck at 2%/2mm (SNC) | 0.14526 | 0.50546 | 0.13853 | 0.02043 | 0.37106 |
| rotationally delivered MapCheck at 5%/3mm (SNC) | 0.51440 | 0.95920 | 0.53639 | 0.08576 | 0.14299 |
| AP field-by-field MapCheck at 3%/3mm (SNC) | 0.02251 | 0.23076 | 0.05337 | 0.01540 | 0.92720 |
| AP field-by-field MapCheck at 2%/2mm (SNC) | 0.02693 | 0.24807 | 0.07047 | 0.01489 | 0.81720 |
| AP field-by-field MapCheck at 5%/3mm (SNC) | 0.00633 | 0.07986 | 0.01760 | 0.00274 | 0.83241 |
| AP composite MapCheck at 3%/3mm (SNC) | 0.94916 | 0.28579 | 0.96231 | 0.33030 | 0.04014 |
| AP composite MapCheck at 2%/2mm (SNC) | 0.51152 | 0.64319 | 0.61555 | 0.16948 | 0.11569 |
| AP composite MapCheck at 5%/3mm (SNC) | 0.42794 | 0.80605 | 0.46740 | 0.09873 | 0.16278 |
| EDR2 film at 3%/3mm | NA | 0.18040 | 1.00000 | 0.30536 | 0.04448 |
| EDR2 film at 2%/2mm | 0.18040 | NA | 0.34659 | 0.03257 | 0.23171 |
| EDR2 film at 5%/3mm | 1.00000 | 0.34659 | NA | 0.33455 | 0.03373 |
| cc04 ion chamber | 0.30536 | 0.03257 | 0.33455 | NA | 0.00538 |
| rotationally delivered MapCheck at 3%/3mm (DL) | 0.04448 | 0.23171 | 0.03373 | 0.00538 | NA |
| rotationally delivered MapCheck at 2%/2mm (DL) | 0.03236 | 0.20055 | 0.02893 | 0.00783 | 0.80031 |
| rotationally delivered MapCheck at 5%/3mm (DL) | 0.22141 | 0.53155 | 0.23762 | 0.02766 | 0.24299 |
| AP field-by-field MapCheck at 3%/3mm (DL) | 0.42413 | 1.00000 | 0.45951 | 0.12723 | 0.24541 |
| AP field-by-field MapCheck at 2%/2mm (DL) | 0.64075 | 0.72270 | 0.68902 | 0.20641 | 0.11410 |
| AP field-by-field MapCheck at 5%/3mm (DL) | 0.14008 | 0.52122 | 0.15417 | 0.03687 | 0.60516 |
| AP composite MapCheck at 3%/3mm (DL) | 0.63926 | 0.24834 | 0.65712 | 0.54613 | 0.03256 |
| AP composite MapCheck at 2%/2mm (DL) | 0.93811 | 0.41368 | 0.93985 | 0.35130 | 0.05167 |
| AP composite MapCheck at 5%/3mm (DL) | 0.19420 | 0.06965 | 0.29849 | 0.90337 | 0.00362 |
| ArcCheck at 3%/3mm (SNC) | 0.97097 | 0.46292 | 0.96962 | 0.30144 | 0.04848 |
| ArcCheck at 2%/2mm (SNC) | 0.74997 | 0.67557 | 0.72491 | 0.16838 | 0.09625 |
| ArcCheck at 5%/3mm (SNC) | 0.73796 | 0.27092 | 0.72449 | 0.47116 | 0.03764 |
| Min | 0.00633 | 0.03257 | 0.01760 | 0.00274 | 0.00362 |
| Max | 1.00000 | 1.00000 | 1.00000 | 0.90337 | 0.92720 |

D test results comparing AUC of between each device. Results are from bootstraped method from pROC using roc.test

| | rotationally delivered MapCheck at 2%/2mm (DL) | rotationally delivered MapCheck at 5%/3mm (DL) | AP field-by-field MapCheck at 3%/3mm (DL) |
|---|---|---|---|
| rotationally delivered MapCheck at 3%/3mm (SNC) | 0.31070 | 0.83062 | 0.60688 |
| rotationally delivered MapCheck at 2%/2mm (SNC) | 0.50306 | 0.88677 | 0.45417 |
| rotationally delivered MapCheck at 5%/3mm (SNC) | 0.22270 | 0.33716 | 0.96120 |
| AP field-by-field MapCheck at 3%/3mm (SNC) | 1.00000 | 0.65845 | 0.01304 |
| AP field-by-field MapCheck at 2%/2mm (SNC) | 0.87001 | 0.75860 | 0.04728 |
| AP field-by-field MapCheck at 5%/3mm (SNC) | 0.72825 | 0.36778 | 0.11748 |
| AP composite MapCheck at 3%/3mm (SNC) | 0.03267 | 0.20737 | 0.32492 |
| AP composite MapCheck at 2%/2mm (SNC) | 0.09544 | 0.36581 | 0.66816 |
| AP composite MapCheck at 5%/3mm (SNC) | 0.17985 | 0.43247 | 0.86195 |
| EDR2 film at 3%/3mm | 0.03236 | 0.22141 | 0.42413 |
| EDR2 film at 2%/2mm | 0.20055 | 0.53155 | 1.00000 |
| EDR2 film at 5%/3mm | 0.02893 | 0.23762 | 0.45951 |
| cc04 ion chamber | 0.00783 | 0.02766 | 0.12723 |
| rotationally delivered MapCheck at 3%/3mm (DL) | 0.80031 | 0.24299 | 0.24541 |
| rotationally delivered MapCheck at 2%/2mm (DL) | NA | 0.41211 | 0.20155 |
| rotationally delivered MapCheck at 5%/3mm (DL) | 0.41211 | NA | 0.57270 |
| AP field-by-field MapCheck at 3%/3mm (DL) | 0.20155 | 0.57270 | NA |
| AP field-by-field MapCheck at 2%/2mm (DL) | 0.09022 | 0.38360 | 0.32381 |
| AP field-by-field MapCheck at 5%/3mm (DL) | 0.64067 | 0.91997 | 0.15370 |
| AP composite MapCheck at 3%/3mm (DL) | 0.02286 | 0.14389 | 0.06517 |
| AP composite MapCheck at 2%/2mm (DL) | 0.02888 | 0.22138 | 0.16997 |
| AP composite MapCheck at 5%/3mm (DL) | 0.00273 | 0.02952 | 0.10014 |
| ArcCheck at 3%/3mm (SNC) | 0.04206 | 0.20822 | 0.30279 |
| ArcCheck at 2%/2mm (SNC) | 0.10288 | 0.33765 | 0.59531 |
| ArcCheck at 5%/3mm (SNC) | 0.02257 | 0.17183 | 0.24331 |
| Min | 0.00273 | 0.02766 | 0.01304 |
| Max | 1.00000 | 0.91997 | 1.00000 |

D test results comparing AUC of between each device. Results are from bootstraped method from pROC using roc.test

| | AP field-by-field MapCheck at 2%/2mm (DL) | AP field-by-field MapCheck at 5%/3mm (DL) | AP composite MapCheck at 3%/3mm (DL) |
|---|---|---|---|
| rotationally delivered MapCheck at 3%/3mm (SNC) | 0.37053 | 0.76289 | 0.11869 |
| rotationally delivered MapCheck at 2%/2mm (SNC) | 0.23983 | 1.00000 | 0.06513 |
| rotationally delivered MapCheck at 5%/3mm (SNC) | 0.70813 | 0.47616 | 0.32389 |
| AP field-by-field MapCheck at 3%/3mm (SNC) | 0.00972 | 0.54182 | 0.00581 |
| AP field-by-field MapCheck at 2%/2mm (SNC) | 0.02071 | 0.74328 | 0.00479 |
| AP field-by-field MapCheck at 5%/3mm (SNC) | 0.04933 | 0.41504 | 0.01188 |
| AP composite MapCheck at 3%/3mm (SNC) | 0.48920 | 0.08781 | 0.64179 |
| AP composite MapCheck at 2%/2mm (SNC) | 0.96546 | 0.24246 | 0.25108 |
| AP composite MapCheck at 5%/3mm (SNC) | 0.84471 | 0.40870 | 0.28450 |
| EDR2 film at 3%/3mm | 0.64075 | 0.14008 | 0.63926 |
| EDR2 film at 2%/2mm | 0.72270 | 0.52122 | 0.24834 |
| EDR2 film at 5%/3mm | 0.68902 | 0.15417 | 0.65712 |
| cc04 ion chamber | 0.20641 | 0.03687 | 0.54613 |
| rotationally delivered MapCheck at 3%/3mm (DL) | 0.11410 | 0.60516 | 0.03256 |
| rotationally delivered MapCheck at 2%/2mm (DL) | 0.09022 | 0.64067 | 0.02286 |
| rotationally delivered MapCheck at 5%/3mm (DL) | 0.38360 | 0.91997 | 0.14389 |
| AP field-by-field MapCheck at 3%/3mm (DL) | 0.32381 | 0.15370 | 0.06517 |
| AP field-by-field MapCheck at 2%/2mm (DL) | NA | 0.06094 | 0.14260 |
| AP field-by-field MapCheck at 5%/3mm (DL) | 0.06094 | NA | 0.02487 |
| AP composite MapCheck at 3%/3mm (DL) | 0.14260 | 0.02487 | NA |
| AP composite MapCheck at 2%/2mm (DL) | 0.38848 | 0.05905 | 0.19094 |
| AP composite MapCheck at 5%/3mm (DL) | 0.11972 | 0.01955 | 0.54179 |
| ArcCheck at 3%/3mm (SNC) | 0.62569 | 0.11959 | 0.40023 |
| ArcCheck at 2%/2mm (SNC) | 0.94252 | 0.25291 | 0.30988 |
| ArcCheck at 5%/3mm (SNC) | 0.41032 | 0.08141 | 0.84128 |
| Min | 0.00972 | 0.01955 | 0.00479 |
| Max | 0.96546 | 1.00000 | 0.84128 |

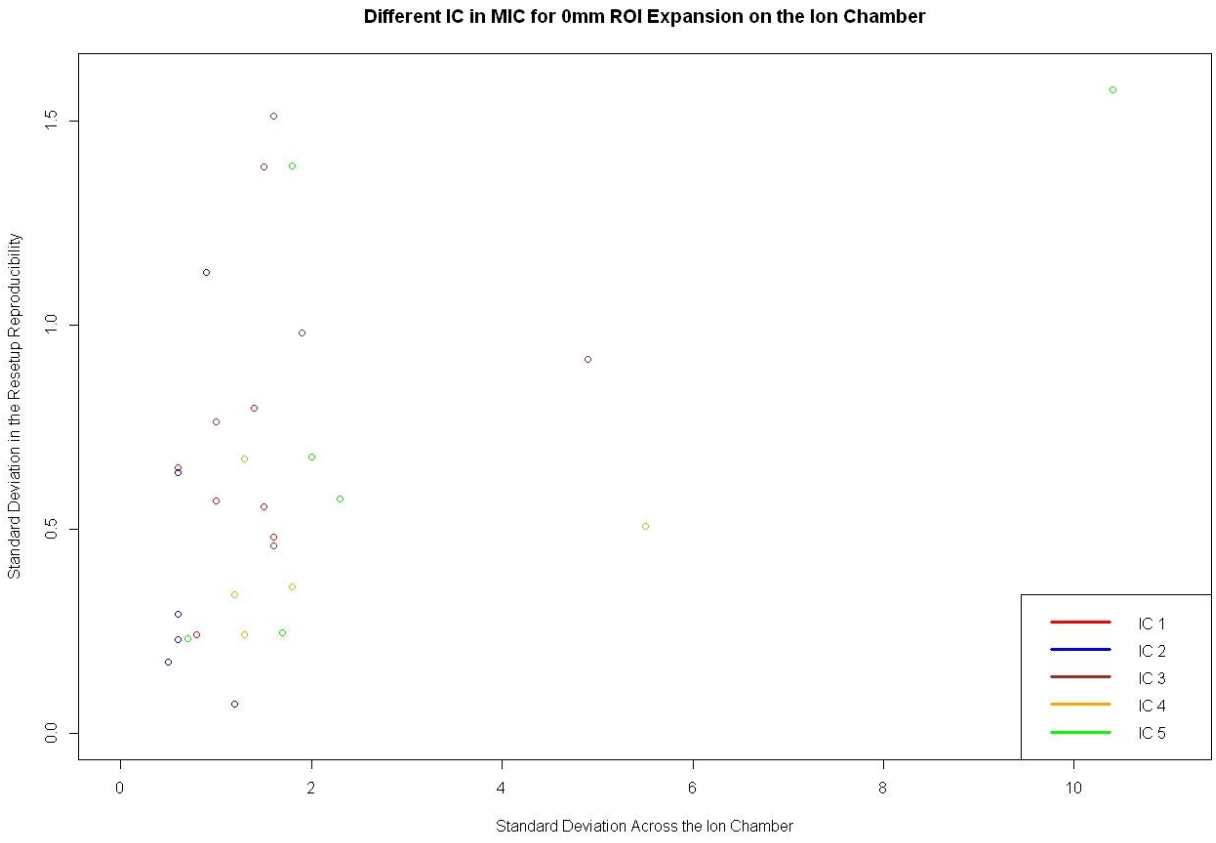
D test results comparing AUC of between each device. Results are from bootstraped method from pROC using roc.test

| | AP composite MapCheck at 2%/2mm (DL) | AP composite MapCheck at 5%/3mm (DL) | ArcCheck at 3%/3mm (SNC) | ArcCheck at 2%/2mm (SNC) |
|---|---|---|---|---|
| rotationally delivered MapCheck at 3%/3mm (SNC) | 0.19582 | 0.03713 | 0.24336 | 0.40674 |
| rotationally delivered MapCheck at 2%/2mm (SNC) | 0.10472 | 0.01622 | 0.15683 | 0.28388 |
| rotationally delivered MapCheck at 5%/3mm (SNC) | 0.46055 | 0.11995 | 0.50228 | 0.69010 |
| AP field-by-field MapCheck at 3%/3mm (SNC) | 0.01205 | 0.00465 | 0.02329 | 0.08379 |
| AP field-by-field MapCheck at 2%/2mm (SNC) | 0.01716 | 0.00723 | 0.02800 | 0.09448 |
| AP field-by-field MapCheck at 5%/3mm (SNC) | 0.02422 | 0.00047 | 0.02824 | 0.06497 |
| AP composite MapCheck at 3%/3mm (SNC) | 0.96280 | 0.11613 | 0.93894 | 0.72616 |
| AP composite MapCheck at 2%/2mm (SNC) | 0.46348 | 0.10628 | 0.58294 | 0.90562 |
| AP composite MapCheck at 5%/3mm (SNC) | 0.47799 | 0.06194 | 0.51122 | 0.76373 |
| EDR2 film at 3%/3mm | 0.93811 | 0.19420 | 0.97097 | 0.74997 |
| EDR2 film at 2%/2mm | 0.41368 | 0.06965 | 0.46292 | 0.67557 |
| EDR2 film at 5%/3mm | 0.93985 | 0.29849 | 0.96962 | 0.72491 |
| cc04 ion chamber | 0.35130 | 0.90337 | 0.30144 | 0.16838 |
| rotationally delivered MapCheck at 3%/3mm (DL) | 0.05167 | 0.00362 | 0.04848 | 0.09625 |
| rotationally delivered MapCheck at 2%/2mm (DL) | 0.02888 | 0.00273 | 0.04206 | 0.10288 |
| rotationally delivered MapCheck at 5%/3mm (DL) | 0.22138 | 0.02952 | 0.20822 | 0.33765 |
| AP field-by-field MapCheck at 3%/3mm (DL) | 0.16997 | 0.10014 | 0.30279 | 0.59531 |
| AP field-by-field MapCheck at 2%/2mm (DL) | 0.38848 | 0.11972 | 0.62569 | 0.94252 |
| AP field-by-field MapCheck at 5%/3mm (DL) | 0.05905 | 0.01955 | 0.11959 | 0.25291 |
| AP composite MapCheck at 3%/3mm (DL) | 0.19094 | 0.54179 | 0.40023 | 0.30988 |
| AP composite MapCheck at 2%/2mm (DL) | NA | 0.30776 | 0.82466 | 0.55836 |
| AP composite MapCheck at 5%/3mm (DL) | 0.30776 | NA | 0.31475 | 0.23868 |
| ArcCheck at 3%/3mm (SNC) | 0.82466 | 0.31475 | NA | 0.43929 |
| ArcCheck at 2%/2mm (SNC) | 0.55836 | 0.23868 | 0.43929 | NA |
| ArcCheck at 5%/3mm (SNC) | 0.70958 | 0.47397 | 0.47319 | 0.33227 |
| Min | 0.01205 | 0.00047 | 0.02329 | 0.06497 |
| Max | 0.96280 | 0.90337 | 0.97097 | 0.94252 |

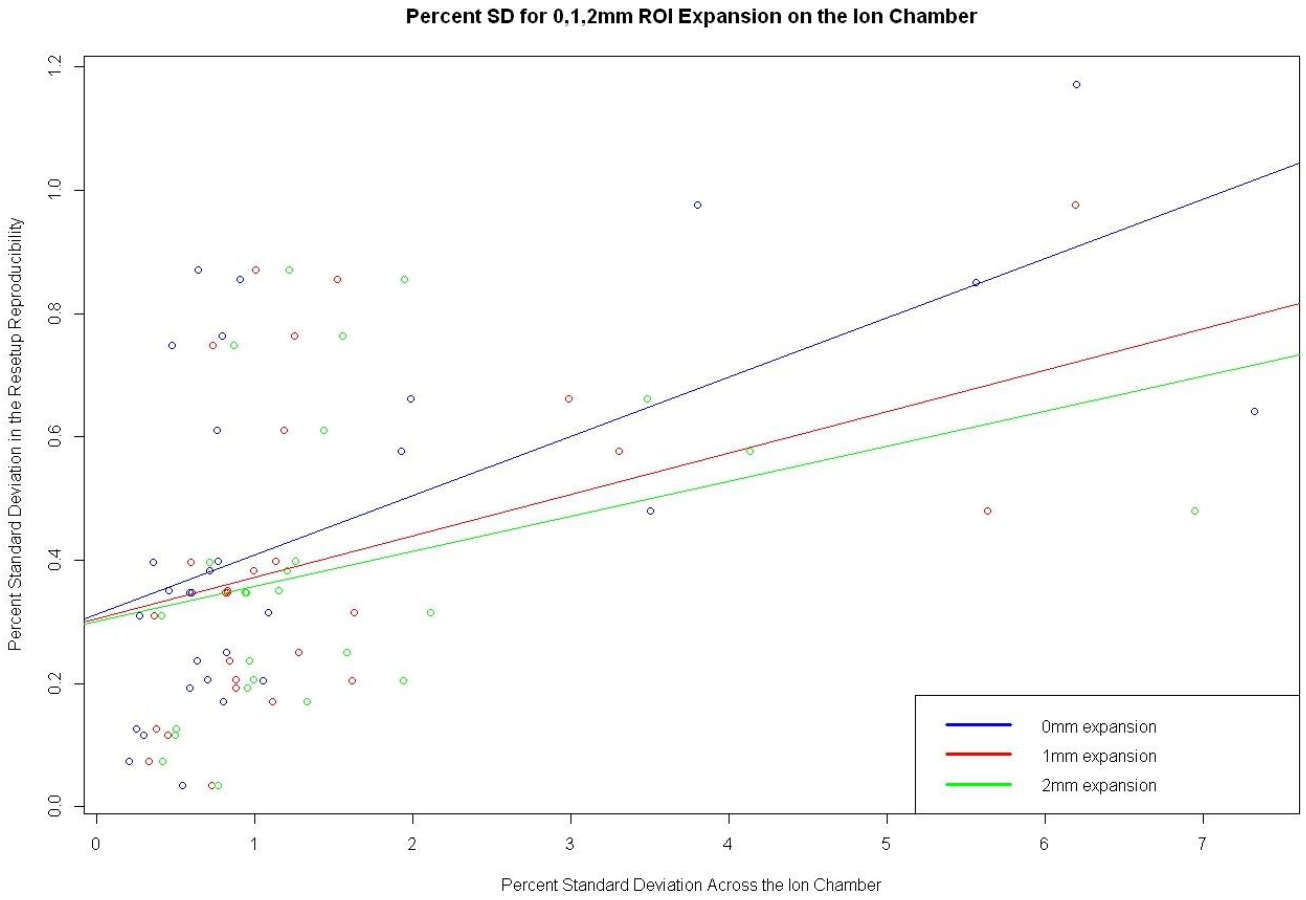
D test results comparing AUC of between each device. Results are from bootstraped method from pROC using roc.test

| | ArcCheck at 5%/3mm (SNC) | |
|---|---|---|
| rotationally delivered MapCheck at 3%/3mm (SNC) | 0.17077 | |
| rotationally delivered MapCheck at 2%/2mm (SNC) | 0.10963 | |
| rotationally delivered MapCheck at 5%/3mm (SNC) | 0.37556 | |
| AP field-by-field MapCheck at 3%/3mm (SNC) | 0.02267 | |
| AP field-by-field MapCheck at 2%/2mm (SNC) | 0.02824 | |
| AP field-by-field MapCheck at 5%/3mm (SNC) | 0.01298 | |
| AP composite MapCheck at 3%/3mm (SNC) | 0.75390 | |
| AP composite MapCheck at 2%/2mm (SNC) | 0.32542 | |
| AP composite MapCheck at 5%/3mm (SNC) | 0.31843 | |
| EDR2 film at 3%/3mm | 0.73796 | |
| EDR2 film at 2%/2mm | 0.27092 | |
| EDR2 film at 5%/3mm | 0.72449 | |
| cc04 ion chamber | 0.47116 | |
| rotationally delivered MapCheck at 3%/3mm (DL) | 0.03764 | |
| rotationally delivered MapCheck at 2%/2mm (DL) | 0.02257 | |
| rotationally delivered MapCheck at 5%/3mm (DL) | 0.17183 | |
| AP field-by-field MapCheck at 3%/3mm (DL) | 0.24331 | |
| AP field-by-field MapCheck at 2%/2mm (DL) | 0.41032 | |
| AP field-by-field MapCheck at 5%/3mm (DL) | 0.08141 | |
| AP composite MapCheck at 3%/3mm (DL) | 0.84128 | |
| AP composite MapCheck at 2%/2mm (DL) | 0.70958 | |
| AP composite MapCheck at 5%/3mm (DL) | 0.47397 | |
| ArcCheck at 3%/3mm (SNC) | 0.47319 | |
| ArcCheck at 2%/2mm (SNC) | 0.33227 | |
| ArcCheck at 5%/3mm (SNC) | NA | |
| Min | 0.01298 | |
| Max | 0.84128 | |

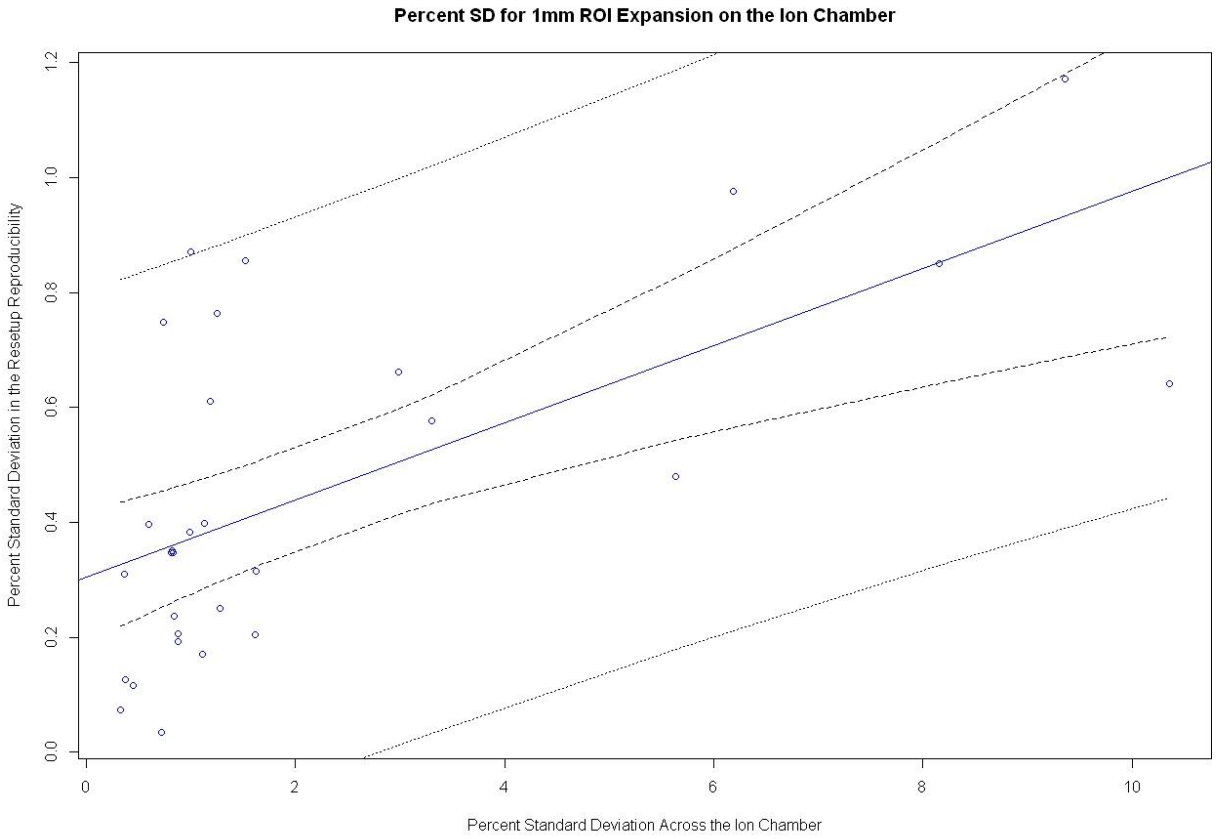
A plot of the standard deviation in the dose across the ion chamber vs the reproducibility in the measurement. Each data point has been color-coded based on which ion chamber was used for its measurement

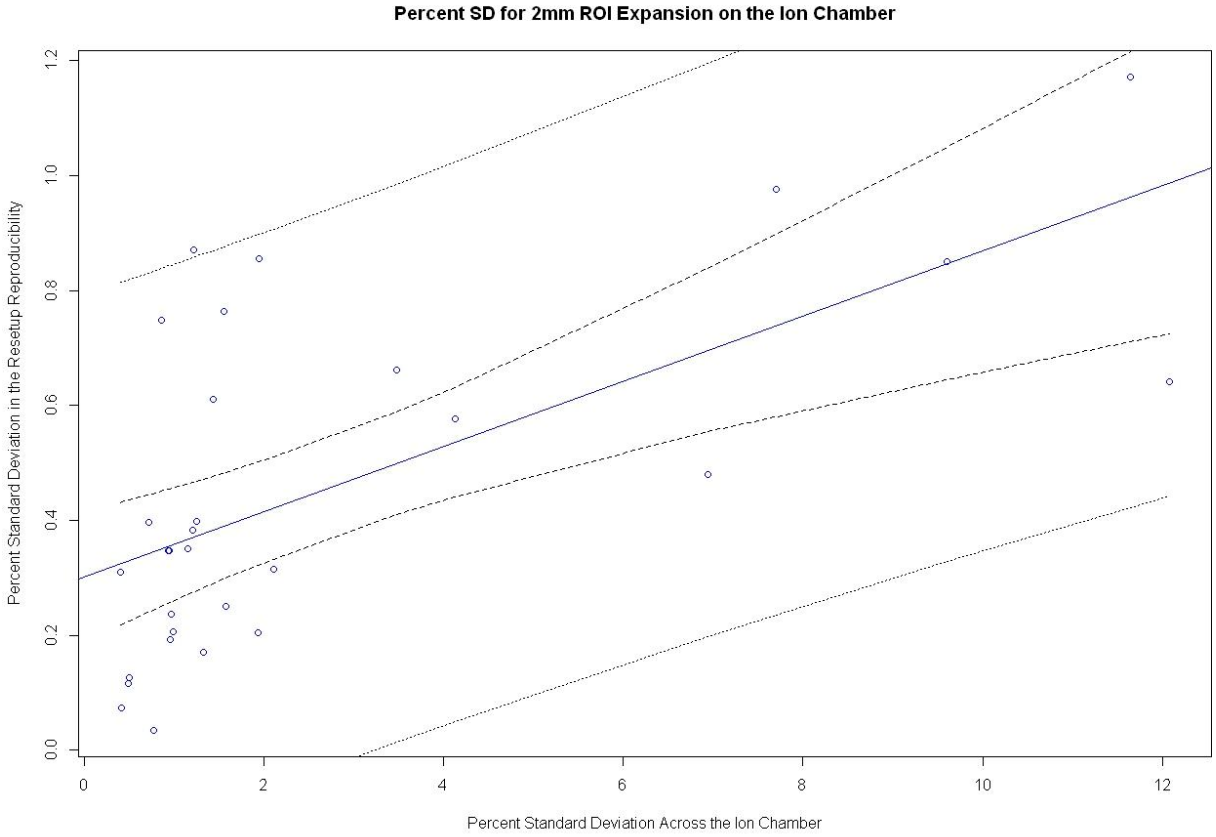


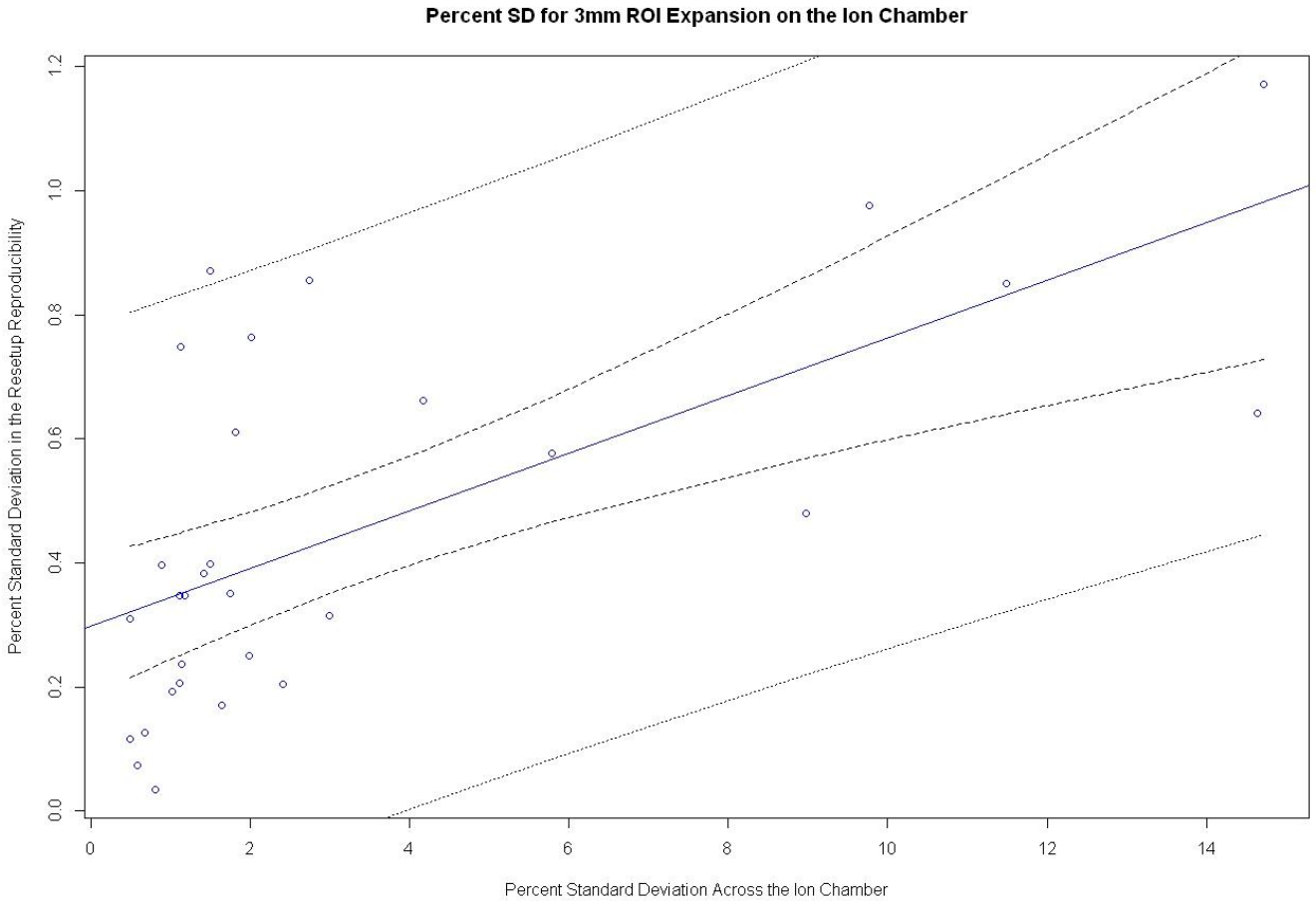
A plot of the different regression lines that resulted from a 0, 1, and 2mm ROI expansion around the ion chamber

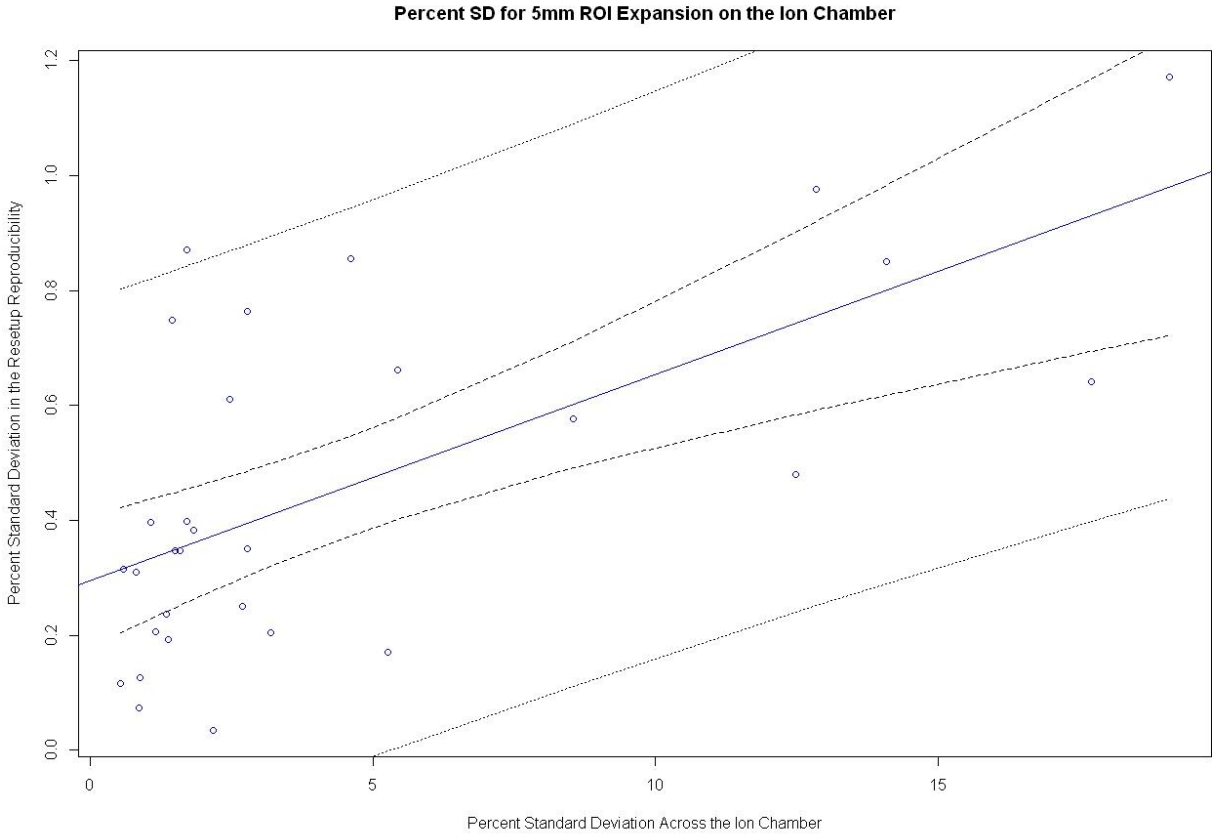


The following four plots are the same as in Figure 11, except a 1, 2, 3, and 5mm expansion around the ion chamber ROI was used to calculate the standard deviation in the dose across the ion chamber

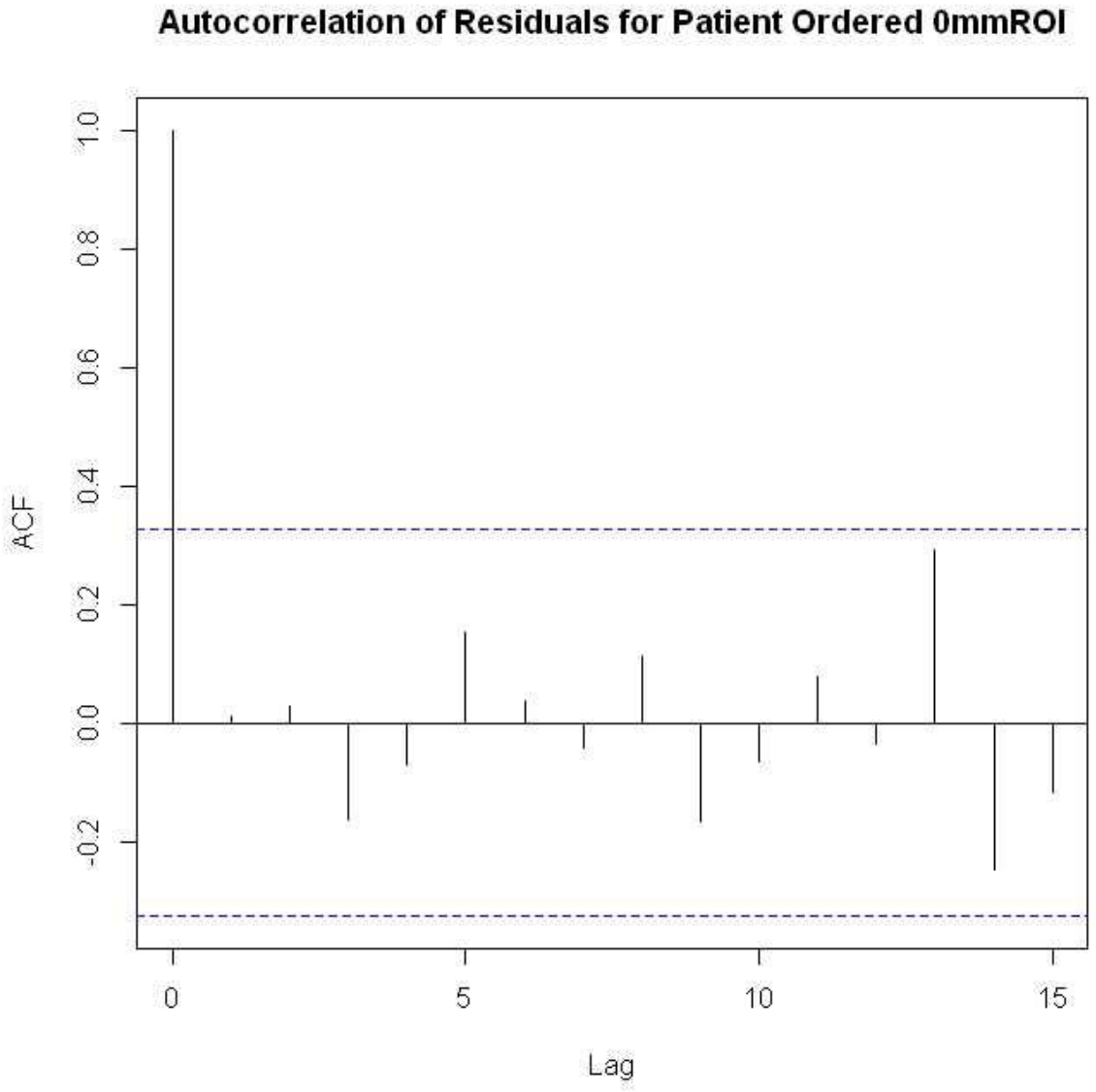


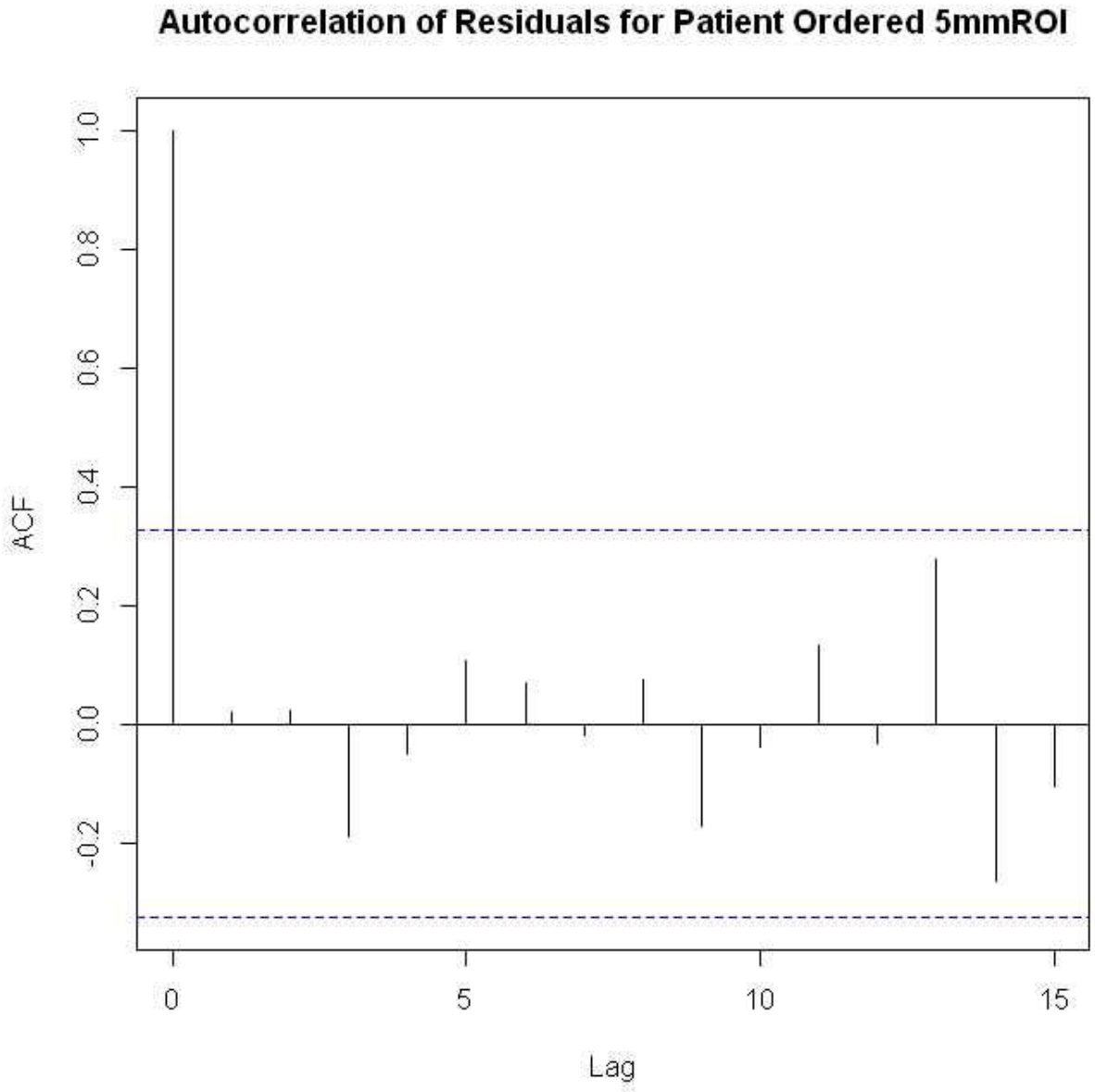

Percent SD for 2mm ROI Expansion on the Ion Chamber
Percent Standard Deviation in the Resetup Reproducibility
Percent Standard Deviation Across the Ion Chamber
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
2
4
6
8
10
12

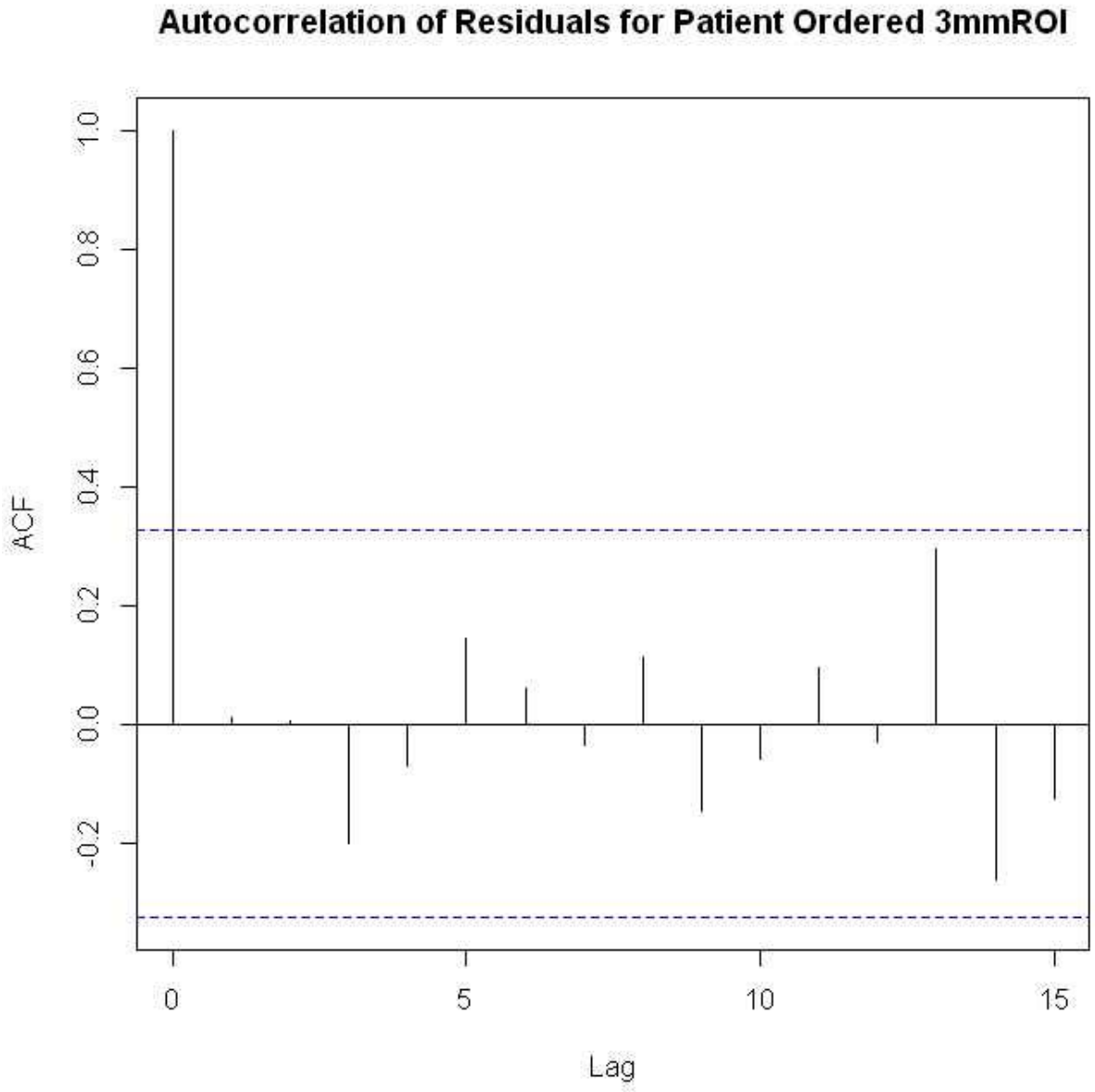

Percent SD for 3mm ROI Expansion on the Ion Chamber
Percent Standard Deviation in the Resetup Reproducibility
Percent Standard Deviation Across the Ion Chamber
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
2
4
6
8
10
12
14

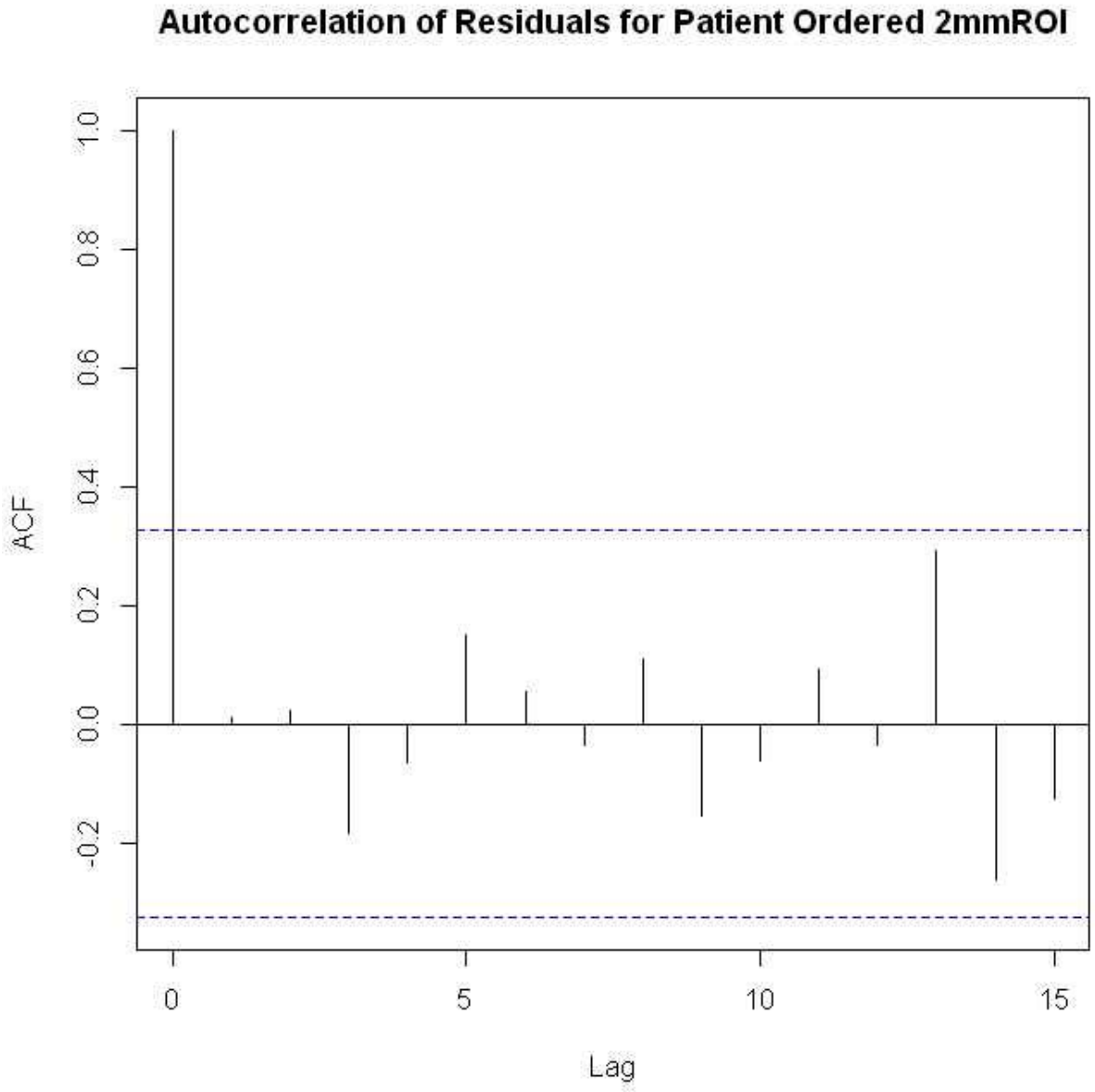

Percent SD for 5mm ROI Expansion on the Ion Chamber
Percent Standard Deviation in the Resetup Reproducibility
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
Percent Standard Deviation Across the Ion Chamber

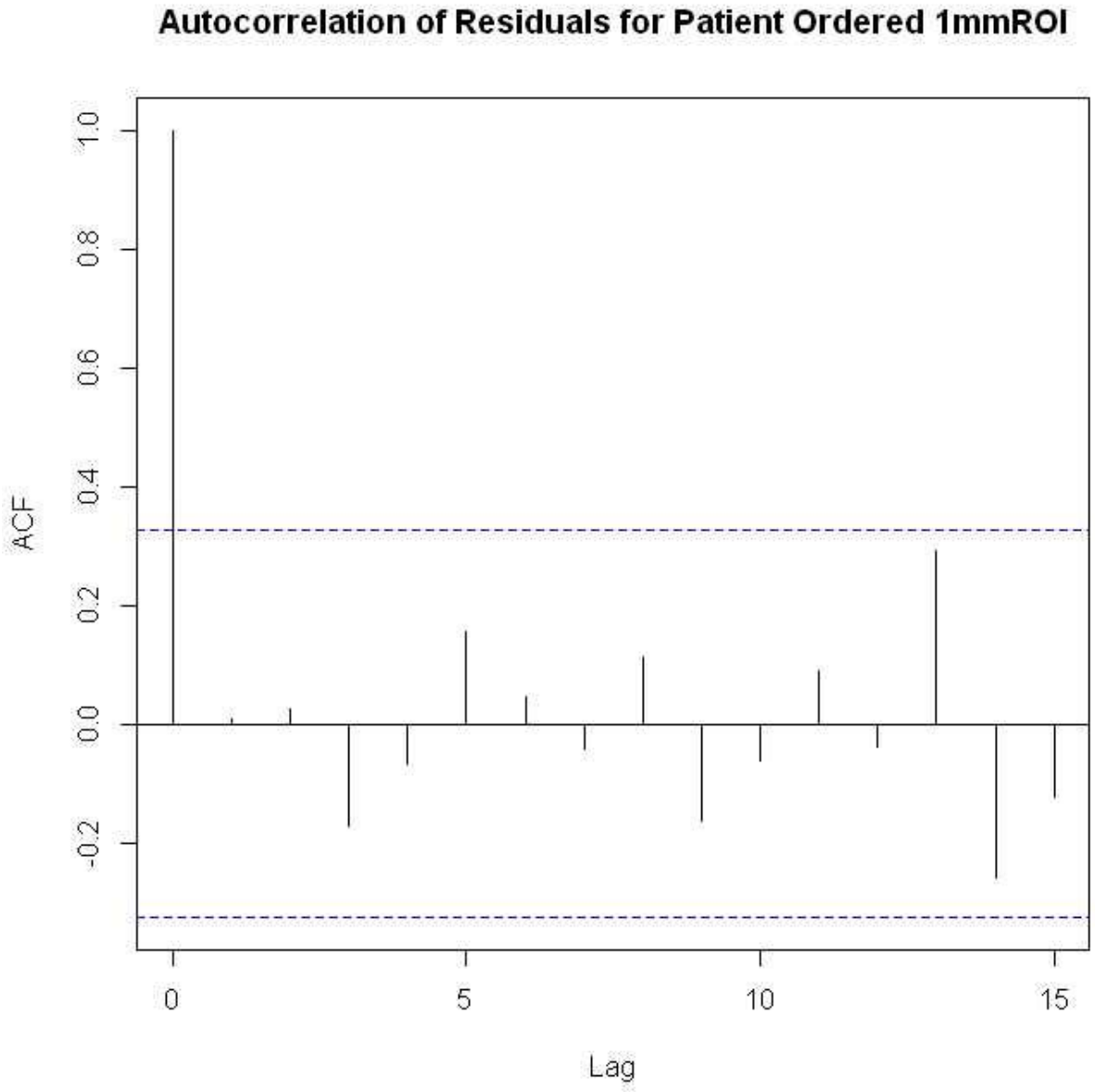
These plots show the results of the autocorrelation of residuals performed in the reproducibility when trying to see if the choice in ion chamber or patient plan would influence the regression.

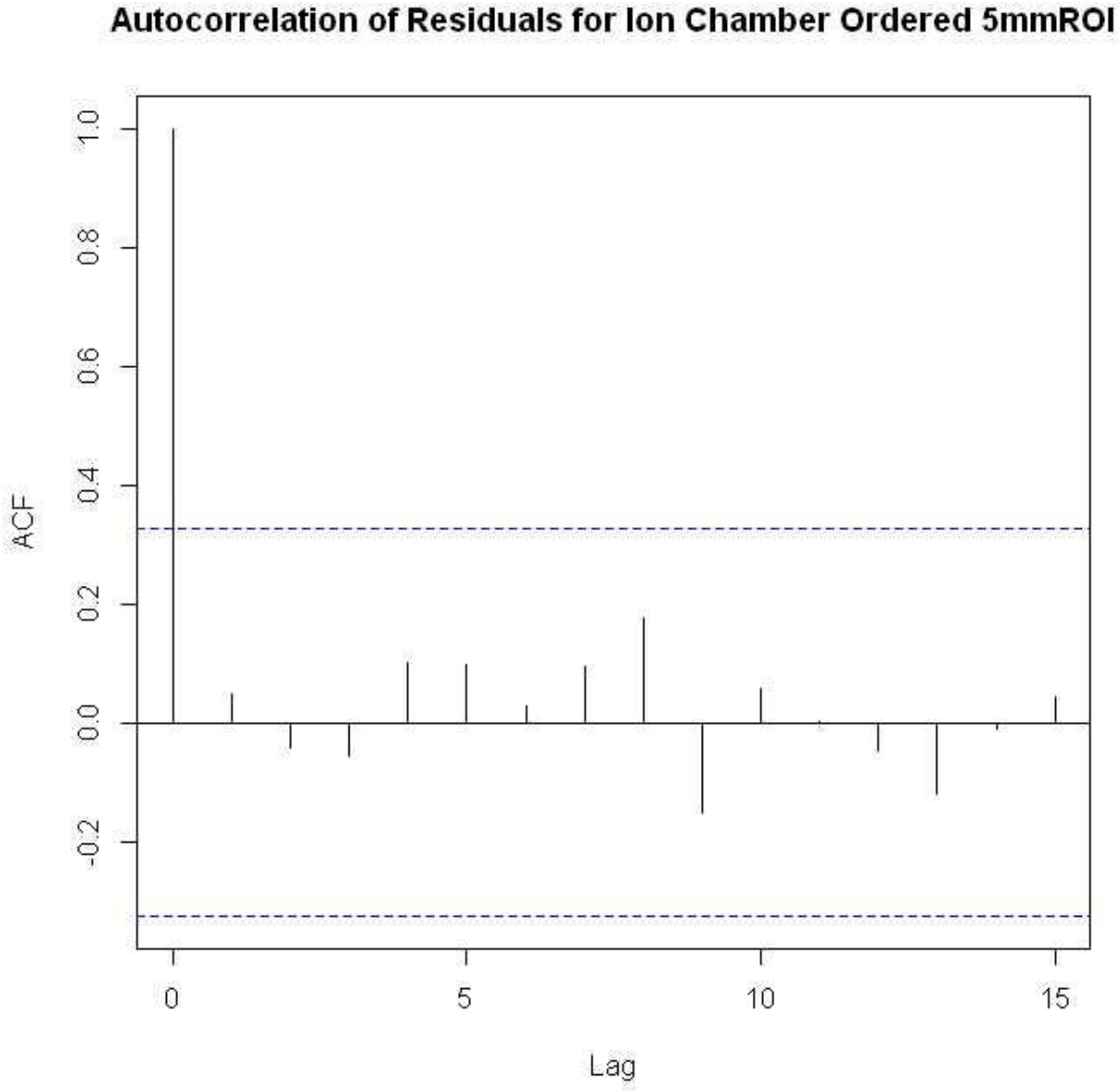


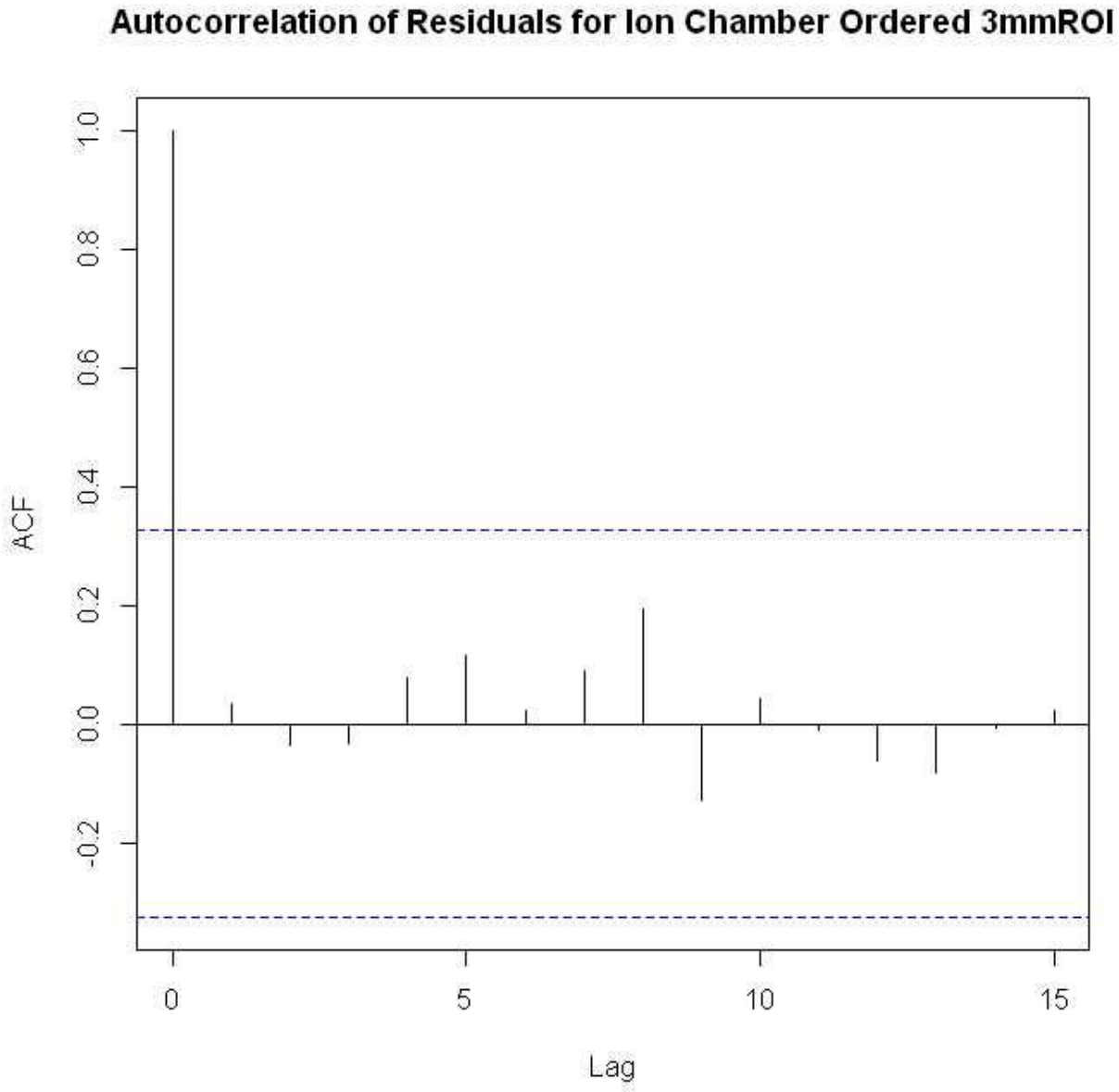

Autocorrelation of Residuals for Patient Ordered 5mmROI
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag

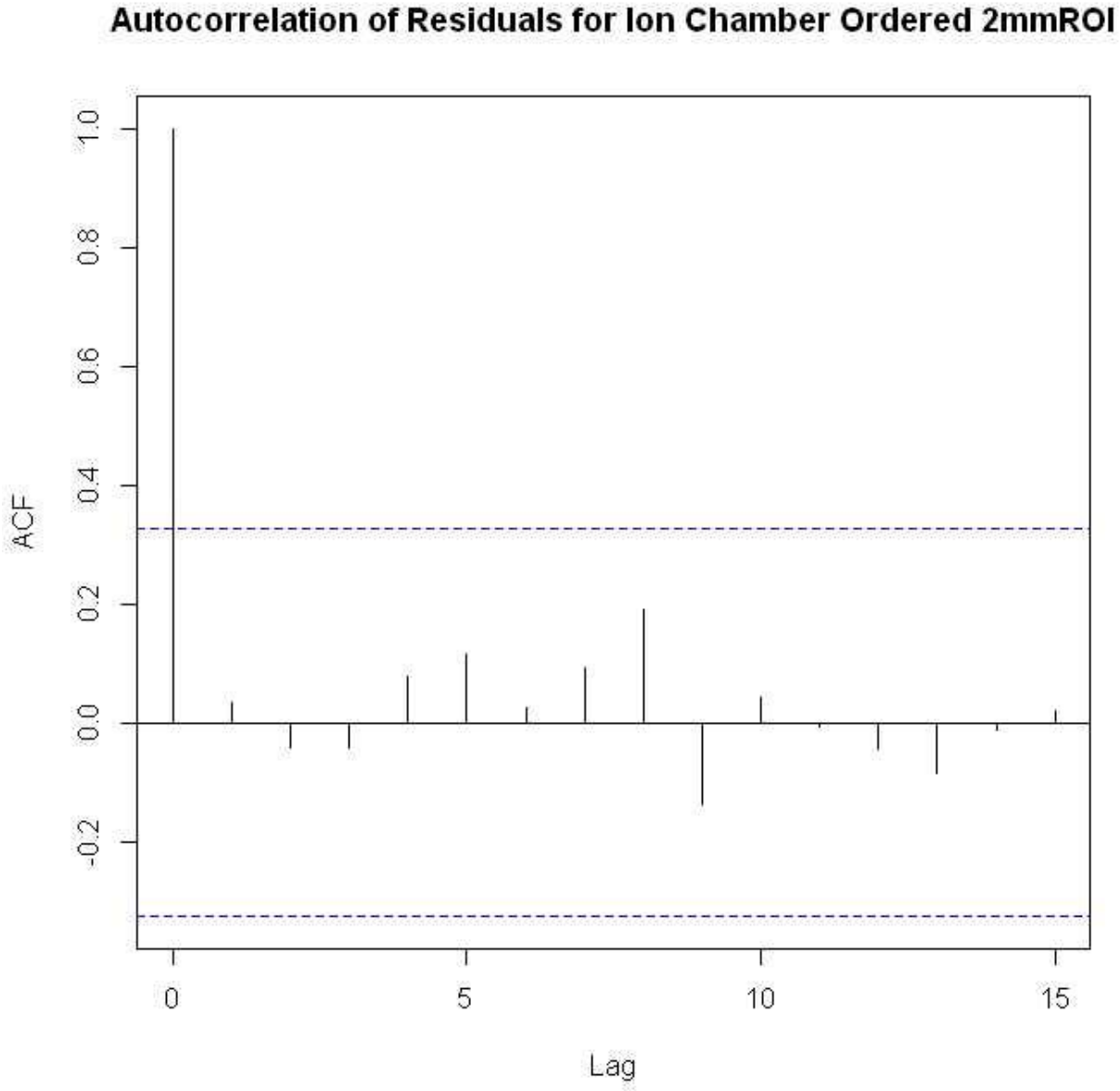

Autocorrelation of Residuals for Patient Ordered 3mmROI
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag

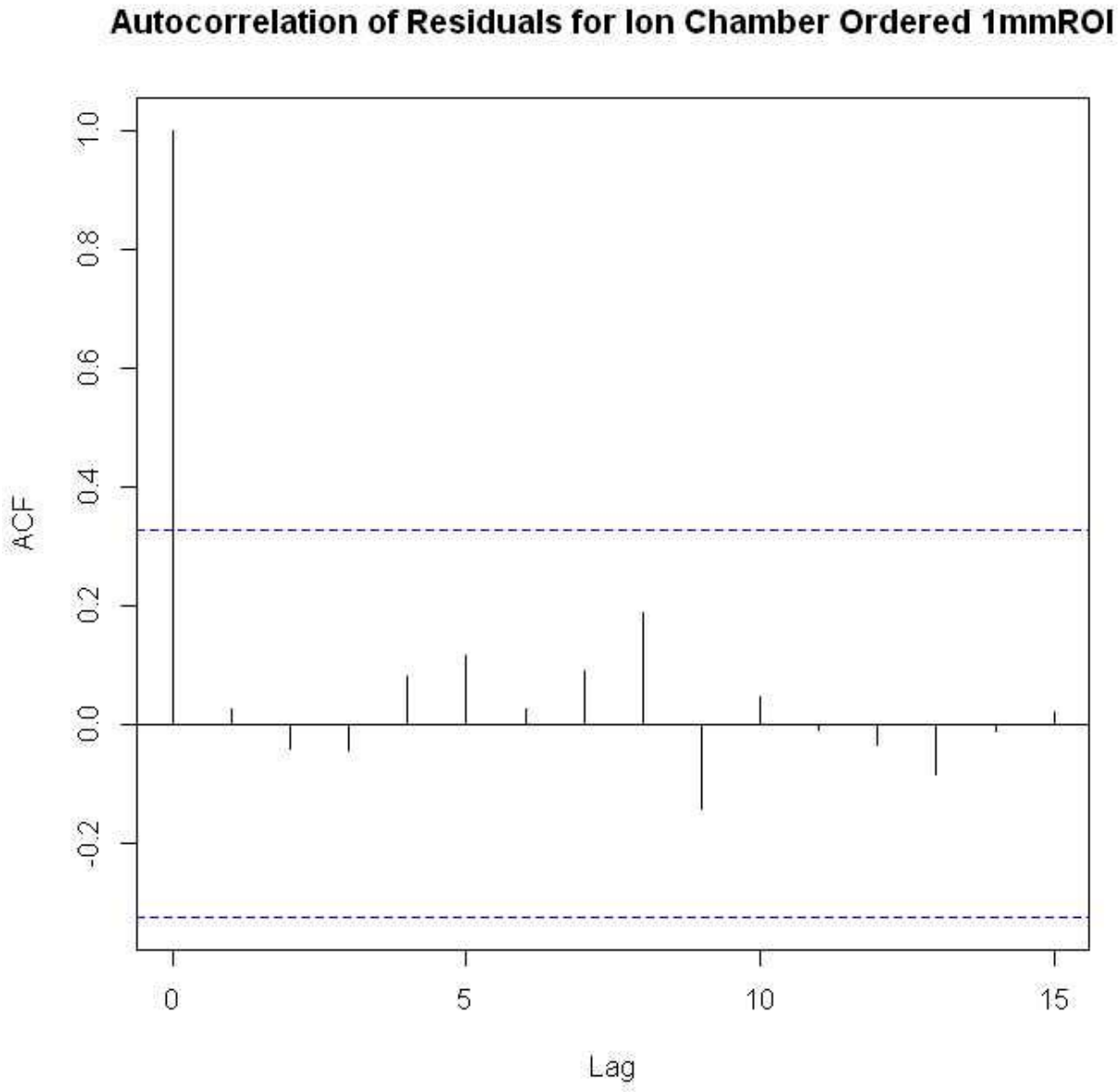

Autocorrelation of Residuals for Patient Ordered 2mmROI
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag

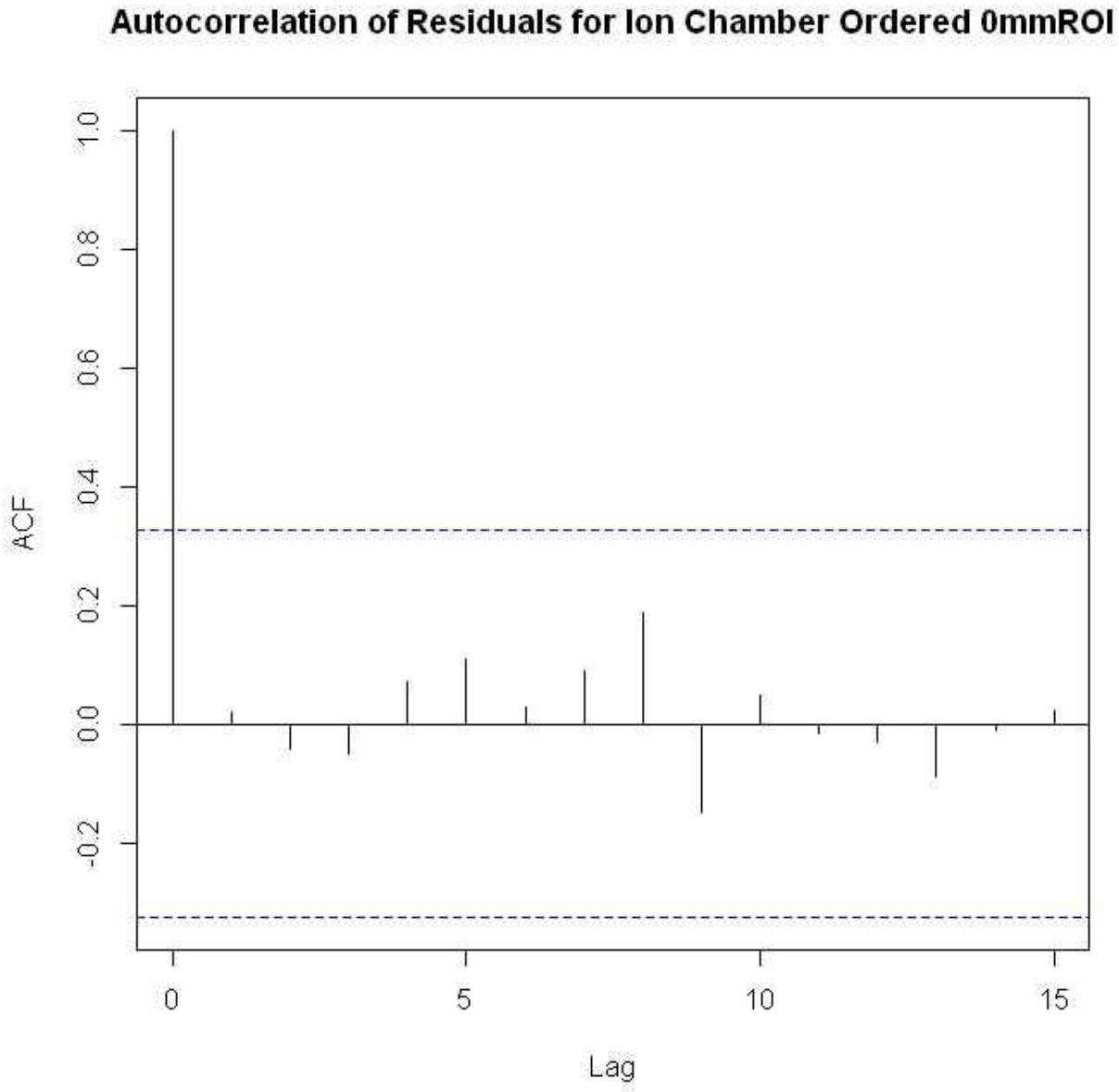

Autocorrelation of Residuals for Patient Ordered 1mmROI
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag


Autocorrelation of Residuals for Ion Chamber Ordered 5mmROI
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag


Autocorrelation of Residuals for Ion Chamber Ordered 3mmROI
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag


Autocorrelation of Residuals for Ion Chamber Ordered 2mmROI
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag


Autocorrelation of Residuals for Ion Chamber Ordered 1mmROI
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag


Autocorrelation of Residuals for Ion Chamber Ordered 0mmROI
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag

Results of ANOVA and Tukey HSD for grouping devices by their CV

```
                                    repr_anova_May1.txt
Analysis of Variance Table For Redelivery

Response: covariance
          Df Sum Sq Mean Sq F value    Pr(>F)
device     6 8.9891 1.49818  6.2985 0.0001462 ***
Residuals 35 8.3252 0.23786
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-----------------Tukey's HSD-------------------------------------------
Study: Redelivery reproducibility among devices

HSD Test for coefficient of variance

Mean Square Error:  0.2378625

device,  means

        covariance    std.err r       Min.      Max.
APcomp  0.13583971 0.06466798 6 0.00000000 0.3329990
APMap   0.09248163 0.02560521 6 0.01741630 0.1972827
arc     0.36268763 0.04110281 6 0.21253985 0.5103091
film    1.50951185 0.50692977 6 0.28951562 3.6535126
IC      0.16865200 0.05614034 6 0.00000000 0.3268020
mic_avg 0.25272939 0.04682981 6 0.09451818 0.3689069
RotMap  0.23735244 0.09304132 6 0.05925593 0.6880233

alpha: 0.05 ; Df Error: 35
Critical Value of Studentized Range: 4.42074

Honestly Significant Difference: 0.8802021

Means with the same letter are not significantly different.

Groups, Treatments and means
a        film            1.51
b        arc             0.3627
b        mic_avg         0.2527
b        RotMap          0.2374
b        IC              0.1687
b        APcomp          0.1358
b        APMap           0.09248

################################################

Analysis of Variance Table for Resetup

Response: covariance
          Df Sum Sq Mean Sq F value   Pr(>F)
device     6 15.265  2.5441   3.934 0.004148 **
Residuals 35 22.634  0.6467
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-----------------Tukey's HSD-------------------------------------------

Study: Resetup reproducibility among devices

HSD Test for coefficient of variance

Mean Square Error:  0.6467004

device,  means
```

repr_anova_May1.txt

| | covariance | std.err | r | Min. | Max. |
|---|---|---|---|---|---|
| APcomp | 0.4599239 | 0.27043751 | 6 | 0.05781211 | 1.7873696 |
| APMap | 0.1515807 | 0.01514462 | 6 | 0.11252482 | 0.2149041 |
| arc | 1.0080696 | 0.25059290 | 6 | 0.20746888 | 1.6697831 |
| film | 1.9706949 | 0.51918545 | 6 | 0.34719205 | 3.9394434 |
| IC | 0.3977391 | 0.16342057 | 6 | 0.17837804 | 1.2081056 |
| mic_avg | 0.4555569 | 0.03469004 | 6 | 0.35905292 | 0.6046549 |
| RotMap | 1.3657074 | 0.56644322 | 6 | 0.20470829 | 3.9768473 |

alpha: 0.05 ; Df Error: 35
Critical Value of Studentized Range: 4.42074

Honestly Significant Difference: 1.451346

Means with the same letter are not significantly different.

Groups, Treatments and means

| | | |
|---|---|---|
| a | film | 1.971 |
| ab | RotMap | 1.366 |
| ab | arc | 1.008 |
| b | APcomp | 0.4599 |
| b | mic_avg | 0.4556 |
| b | IC | 0.3977 |
| b | APMap | 0.1516 |

The raw data of each IMRT QA measurement made on each patient

MIC metric, % pixels passing for planar dosimeters, and % dose difference for ion chamber for each patient plan

| | MIC | Rot MapCheck SNC | | | AP MapCheck SNC (weighted) | | | |
|---|---|---|---|---|---|---|---|---|
| | | 3%/3mm | 2%/2mm | 5%/3mm | 3%/3mm | 2%/2mm | 5%/3mm | |
| thoracic | 2.730971 | 82.2 | 69.2 | 90.4 | 96.03501 | 84.90038 | 98.62346 | |
| meso | 1.697704 | 94.3 | 84.2 | 99.2 | 93.39668 | 75.85528 | 97.98422 | |
| mantle | 3.345321 | 91.8 | 80.4 | 95.6 | 98.31239 | 92.35106 | 99.55921 | |
| lung stereo | 0.891366 | 83.1 | 56.6 | 92.8 | 95.54431 | 86.74466 | 97.43483 | |
| GI | 1.603404 | 97.6 | 88.5 | 99.5 | 95.13647 | 78.6332 | 99.24206 | |
| GI | 0.793442 | 67.6 | 48.4 | 87.9 | 95.12343 | 83.54678 | 98.44285 | |
| rib stereo | 0.929593 | 76.7 | 69.2 | 82.7 | 99.24582 | 94 | 99.80792 | |
| HN | 0.06204 | 76.5 | 61.9 | 85.8 | 95.72042 | 85.9435 | 98.33183 | |
| HN | 0 | 91.5 | 80.1 | 98.3 | 91.55503 | 76.07483 | 96.11779 | |
| GI | 0.536089 | 94.7 | 87 | 98.5 | 95.59646 | 83.0554 | 98.10866 | |
| HN | 0 | 95.8 | 85.1 | 99 | 97.95045 | 90.66777 | 99.40879 | |
| meso | 1.824711 | 94.9 | 81.3 | 99.4 | 95.31672 | 83.3976 | 98.98286 | |
| Gyn | 0 | 94.3 | 83.9 | 98.8 | 98.26314 | 90.07611 | 99.44744 | |
| meso | 0.554139 | 81.3 | 60.3 | 96.6 | 95.93348 | 81.29294 | 99.8 | |
| spine stereo | 1.919981 | 88.2 | 81 | 94.6 | 98.34898 | 92.88732 | 99.58779 | |
| GI | 0 | 92.8 | 83.9 | 98.9 | 89.62603 | 72.51333 | 99.03556 | |
| meso | 1.743721 | 72.5 | 49.6 | 94.3 | 91.85 | 72.82872 | 97.71421 | |
| gyn | 0.635691 | 94 | 84.1 | 98.9 | 90.3896 | 71.69114 | 98.68383 | |
| HN | 0.503247 | 93.3 | 85.2 | 98.5 | 95.26833 | 85.02955 | 97.66064 | |
| GU | 0.036562 | 95.6 | 86.7 | 100 | 96.50268 | 85.91542 | 99.15225 | |
| Thoracic | 0.403836 | 93.7 | 84.2 | 98.4 | 98.59894 | 92.52347 | 99.42114 | |
| HN | 0 | 97.5 | 87.6 | 99.7 | 97.76996 | 90.07194 | 99.19382 | |
| Thoracic | 0.038977 | 95.1 | 87.5 | 99.2 | 99.43651 | 95.05953 | 99.76616 | |
| GI | 0.022858 | 92.8 | 80.6 | 99.8 | 96.86516 | 87.55324 | 98.97647 | |

MIC metric, % pixels passing for planar dosimeters, and % dose difference for ion chamber for each patient plan

| | APMap composite SNC | | | ArcCheck SNC | | | Film | |
|---|---|---|---|---|---|---|---|---|
| | 3%/3mm | 2%/2mm | 5%/3mm | 3%/3mm | 2%/2mm | 5%/3mm | 3%/3mm | 2%/2mm |
| thoracic | 96.2 | 87.6 | 99.6 | 93.7 | 84.6 | 97.8 | 89.75 | 74.86 |
| meso | 86.7 | 60.1 | 99.1 | 76.4 | 61.3 | 87.7 | 77.47 | 56.61 |
| mantle | 99.1 | 97.2 | 100 | 91.7 | 73.6 | 97.5 | 99.24 | 96.13 |
| lung stereo | 95.7 | 80.4 | 97.8 | 87.5 | 76.5 | 91.9 | 96.15 | 90.32 |
| GI | 82.8 | 53 | 97 | 41.4 | 23.3 | 63 | 84.01 | 68.12 |
| GI | 95.2 | 82.4 | 99.5 | 81 | 55.7 | 94.3 | 69.02 | 51.68 |
| rib stereo | 97.7 | 96.4 | 100 | 95.4 | 86.3 | 97.3 | 94.93 | 83.52 |
| HN | 99.7 | 94.9 | 100 | 93.2 | 78.7 | 97.8 | 99.83 | 75.18 |
| HN | 94.4 | 85.7 | 99.8 | 96 | 87.4 | 98.2 | 87.76 | 69.85 |
| GI | 94.6 | 82.4 | 98.9 | 75.4 | 56.8 | 93.3 | 76.83 | 50.54 |
| HN | 99.5 | 96.8 | 99.8 | 95.4 | 85.5 | 98.5 | 98.72 | 88.76 |
| meso | 96.3 | 82 | 99.9 | 83.3 | 63.6 | 95.5 | 72.28 | 58.2 |
| Gyn | 100 | 97.3 | 100 | 92.8 | 75.7 | 97.6 | 99.83 | 94.51 |
| meso | 91 | 67.3 | 97.7 | 69.9 | 50.3 | 82.5 | 74.77 | 63.33 |
| spine stereo | 96.4 | 91.5 | 100 | 92 | 81.2 | 95.4 | 97.01 | 86.41 |
| GI | 98.8 | 82.6 | 100 | 87.2 | 72.4 | 96.9 | 98.6 | 92.44 |
| meso | 69.2 | 45 | 96.3 | 55.6 | 34.8 | 71.4 | 94.9 | 80.72 |
| gyn | 82.6 | 58.3 | 97.6 | 70.8 | 48.1 | 91.7 | 79.7 | 65.36 |
| HN | 97.1 | 92.8 | 99 | 90 | 74.4 | 96.3 | 88.08 | 64.63 |
| GU | 99.4 | 95 | 100 | 95.2 | 86.9 | 97.3 | 98.29 | 89.2 |
| Thoracic | 100 | 96.8 | 100 | 96.6 | 85.9 | 99.4 | 98.69 | 90.01 |
| HN | 98.9 | 91.6 | 99.8 | 95.5 | 83.9 | 99.2 | 99.44 | 94.17 |
| Thoracic | 99.6 | 97.1 | 100 | 95.5 | 85.4 | 99 | 97.09 | 79.75 |
| GI | 99 | 93.7 | 99.7 | 94.3 | 83.2 | 97.7 | 91.9 | 81.63 |

MIC metric, % pixels passing for planar dosimeters, and % dose difference for ion chamber for each patient plan

| | Film | IC | Rot MapCheck DL | | | AP Mapcheck DL(weighted) | | |
|---|---|---|---|---|---|---|---|---|
| | 5%/3mm | | 3%/3mm | 2%/2mm | 5%/3mm | 3%/3mm | 2%/2mm | 5%/3mm |
| thoracic | 97.33 | -2.5 | 85 | 74.7 | 89.5 | 98.52 | 89.51 | 99.27 |
| meso | 96.91 | 3.2 | 93.2 | 78.6 | 99.2 | 94.33 | 78.63 | 99.01 |
| mantle | 99.76 | -4 | 91.3 | 78 | 93.2 | 97.53 | 89.21 | 99.09 |
| lung stereo | 98.85 | 0.8 | 93.2 | 82 | 97.6 | 97.50 | 87.75 | 97.95 |
| GI | 91.37 | 2.5 | 97.6 | 85.8 | 99.7 | 95.46 | 79.73 | 99.38 |
| GI | 87.61 | 2.7 | 65.8 | 47.4 | 84.8 | 96.47 | 84.98 | 98.84 |
| rib stereo | 99.78 | -1.6 | 98.5 | 89.2 | 99.9 | 98.85 | 89.59 | 99.25 |
| HN | 99.83 | -1.1 | 78.3 | 63.4 | 87.9 | 98.16 | 90.82 | 99.16 |
| HN | 99.25 | 0.9 | 89 | 73.5 | 97.4 | 96.83 | 86.71 | 98.67 |
| GI | 97.56 | 3.4 | 94.4 | 82.1 | 98.9 | 98.02 | 91.22 | 99.40 |
| HN | 100 | -0.5 | 96 | 83.1 | 99.3 | 98.89 | 92.26 | 99.69 |
| meso | 88.8 | 3.8 | 93.8 | 79.6 | 99.5 | 97.16 | 87.15 | 99.33 |
| Gyn | 100 | 0.4 | 92.7 | 83.7 | 98.9 | 98.84 | 90.31 | 99.41 |
| meso | 91.39 | 3.2 | 74.4 | 52.6 | 94.5 | 94.67 | 79.15 | 99.42 |
| spine stereo | 99.52 | 0.9 | 84.8 | 73.4 | 91.1 | 97.70 | 89.34 | 98.66 |
| GI | 99.9 | 0.5 | 94.2 | 81.4 | 99.6 | 92.29 | 81.41 | 97.72 |
| meso | 99.52 | 3.9 | 65.5 | 44.7 | 89.4 | 92.38 | 73.54 | 99.19 |
| gyn | 99.03 | 1.3 | 92.9 | 78.4 | 98 | 91.95 | 78.16 | 99.18 |
| HN | 98.01 | -3.6 | 94.4 | 84.6 | 98 | 97.66 | 89.19 | 99.09 |
| GU | 99.93 | 0.4 | 93.4 | 81.4 | 99.5 | 99.14 | 93.14 | 99.58 |
| Thoracic | 100 | -1 | 93.9 | 84.8 | 98 | 99.34 | 94.71 | 99.75 |
| HN | 99.9 | 0.6 | 97.3 | 88.4 | 99.6 | 99.22 | 94.14 | 99.65 |
| Thoracic | 99.99 | 1.5 | 95.4 | 88.4 | 99.3 | 99.58 | 96.21 | 99.84 |
| GI | 97.05 | -0.1 | 90.6 | 77.4 | 98.7 | 98.54 | 92.95 | 99.29 |

MIC metric, % pixels passing for planar dosimeters, and % dose difference for ion chamber for each patient plan

| | APMap comp DL | | | |
|---|---|---|---|---|
| | 3%/3mm | 2%/2mm | 5%/3mm | Truth |
| thoracic | 96.3 | 88.1 | 99.1 | 1 |
| meso | 82.9 | 59.9 | 99.1 | 1 |
| mantle | 96.5 | 82.7 | 99.3 | 1 |
| lung stereo | 95.2 | 84 | 98.3 | 1 |
| GI | 74.9 | 45.9 | 96.7 | 1 |
| GI | 94.5 | 79.9 | 99.3 | 1 |
| rib stereo | 98.1 | 90.6 | 99.7 | 1 |
| HN | 99.7 | 93.3 | 100 | 0 |
| HN | 98.2 | 89.4 | 99.3 | 0 |
| GI | 97 | 89.2 | 99.3 | 1 |
| HN | 99.4 | 95.1 | 99.8 | 0 |
| meso | 93.2 | 74.7 | 99.8 | 1 |
| Gyn | 99.9 | 97.3 | 100 | 0 |
| meso | 81.8 | 60 | 93.1 | 1 |
| spine stereo | 95.5 | 87.5 | 98.1 | 1 |
| GI | 94.6 | 81.3 | 100 | 0 |
| meso | 62.3 | 44.4 | 91.9 | 1 |
| gyn | 80.7 | 66.6 | 97.8 | 1 |
| HN | 98.9 | 93.6 | 99.7 | 1 |
| GU | 99.3 | 95.8 | 99.9 | 0 |
| Thoracic | 99.7 | 97.9 | 99.9 | 1 |
| HN | 99.3 | 93.2 | 100 | 0 |
| Thoracic | 99.6 | 97.5 | 99.8 | 0 |
| GI | 99.2 | 94.6 | 99.9 | 0 |

# Vita

Elizabeth MaryAnn McKenzie was born in Mableton, GA on September 23, 1988, the daughter of Condle Devoy and Laura Ann McKenzie. After completing high school at Penn High School in Mishawaka, Indiana in 2007, she entered Purdue University in West Lafayette, IN. She graduated in 2011 and obtained a Bachelor's of Science degree in honors physics, with minors in mathematics and French. In August 2011 she entered the University of Texas Health Science Center at the Houston Graduate School of Biomedical Sciences. Here she pursued a Master's degree in Medical Physics.

Permanent Address:
1030 Tiverton Ave Apt 325
Los Angeles, CA 90024