Title
Quantifying and Improving Error Sensitivity of Intensity Modulated Radiation Therapy Quality Assurance

Author
Steers, Jennifer Mariah

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Quantifying and Improving Error Sensitivity of Intensity Modulated Radiation Therapy Quality Assurance

A dissertation submitted in partial satisfaction

of the requirements for the degree of Doctor of Philosophy

in Biomedical Physics

by

Jennifer Mariah Steers

2018

ABSTRACT OF THE DISSERTATION

Quantifying and Improving Error Sensitivity of

Intensity Modulated Radiation Therapy Quality Assurance

by

Jennifer Mariah Steers

Doctor of Philosophy in Biomedical Physics

University of California, Los Angeles, 2018

Professor Daniel Abraham Low, Chair

Purpose

To quantify and elucidate factors affecting error sensitivity in current IMRT QA comparisons performed with the gamma comparison, to investigate causes for gamma comparison insensitivity, and to utilize these results to develop and validate a new method for analyzing IMRT QA dose distributions in the clinic.

Methods

Over 20,000 gamma comparisons were performed for three detector geometries – ArcCHECK, MapCHECK, and Delta 4 – for a variety of IMRT and VMAT cases in the presence of induced errors. Differences in error sensitivity for each device geometry and delivery technique were studied with the use of 1mm vs. 1mm calculation-only comparisons and in-house MATLAB gamma comparison software developed specifically for this project. Additionally, the effects of spatial sampling for each device were evaluated using gamma comparisons performed at the true

spatial sampling of each detector compared to those at 1mm. Patterns of gamma failures were also investigated in the presence of induced errors of increasing magnitude.

Results from these gamma comparisons were considered in developing a new comparison technique for IMRT QA analysis. A new analysis method that segments the IMRT QA comparisons by different dose and gradient thresholds was developed with the use of known induced errors in a calculation-only scenario for the ArcCHECK, MapCHECK, and Delta 4 measurement geometries. Results from the new method were validated for a separate cohort of patient plans, as well as with the use of real plan measurements with and without intentional errors on the MapCHECK device.

Results

Differences in gamma comparison error sensitivity were observed for the ArcCHECK, MapCHECK, and Delta 4 geometries when removing the effects of different spatial sampling for each device. While sensitivity was error type-specific for most studied gamma criteria, a gamma criterion with a 10% low dose threshold and local dose difference normalization appeared to offer similar sensitivity across the three devices. For more commonly used gamma criteria, the Delta 4 appeared more sensitive for the majority of induced error types. Additionally, error sensitivity was lower for VMAT cases compared to IMRT cases across all detector geometries. Reducing the spatial sampling of each device from 1mm to the true spatial sampling of the device did not noticeably affect gamma comparison error sensitivity. In evaluating patterns of gamma failures and gamma value maps in the presence of induced errors of increasing magnitude, it was observed that high dose gradients likely limit the sensitivity of the gamma comparison, regardless of dose difference normalization or detector geometry. Additionally, for some cases the number of diodes in real measurements not falling along these gradients may be alarmingly low, which may help explain why the gamma comparison can fail to flag large errors for certain cases.

A new method, gradient-dose segmented analysis (GDSA) was developed to allow more clinically meaningful and sensitive IMRT QA comparisons. This method segments the comparison points into regions of high-gradient, high-dose low-gradient, and low-dose low-gradient points. The mean local dose difference in high-dose low-gradient regions of the comparison was found to predict true changes in PTV mean in the patient DVH. The development of GDSA made use of over 180,000 comparisons to select appropriate dose and gradient thresholds for IMRT and VMAT cases on the MapCHECK, Delta 4, and ArcCHECK devices. Predictions for change in PTV mean dose performed best for the MapCHECK and Delta 4 geometries, with a nearly 1:1 correlation between predicted and true change in PTV mean dose. Additionally, as a binary pass/fail metric, GDSA exhibited higher sensitivity and specificity than five studied gamma criteria. GDSA results were validated with a separate cohort of patients as well as real MapCHECK measurements. GDSA is feasible for clinical implementation as it would not require an increase in time spent analyzing the results.

Conclusions

A variety of measurement scenarios were considered in controlled calculation-only comparisons that suggest device-specific and delivery technique-specific gamma criteria may be appropriate in order to achieve similar sensitivity in IMRT QA comparisons across the field. Additionally, the complexity of gradient maps in current IMRT QA appears to be a driving factor in error sensitivity for the gamma comparison. Finally, the gradient-dose segmented analysis (GDSA) method has been developed and validated for the purpose of IMRT QA analysis for three different detector devices. GDSA was shown to predict changes in PTV mean in the patient DVH using only information from the calculations and measurements in the phantom geometry. As a binary pass/fail metric, GDSA was also shown to be more sensitive and specific than the gamma comparison. Results from this new analysis technique can help predict the clinical relevance of dose differences in IMRT QA measurements, thus offering more meaningful IMRT QA results.

The dissertation of Jennifer Mariah Steers is approved.

Dan Ruan

James Michael Lamb

Michael McNitt-Gray

Benedick Andrew Fraass

Daniel Abraham Low, Committee Chair

University of California, Los Angeles

2018

For my parents, Rosemary Barber-Steers and John C. Steers,

who battled cancer with inspiring resolve and quiet strength.

Table of Contents

List of Figures

List of Appendix Figures

List of Tables

List of Appendix Tables

Acknowledgements

There are many to whom I owe an enormous debt of gratitude for their support of this work and my efforts during my time in graduate school. None of this would have been possible without the wisdom and guidance of my adviser, Dr. Benedick A. Fraass, who I have had the extreme honor of working under for the past nine years. As both an undergraduate researcher and a graduate student, Dick never failed to make time for teaching or discussions, and his patient and humble nature is something to be truly admired. All the project meetings that stretched long beyond their scheduled time slots and the impromptu discussions over the Keurig machine will be greatly missed. Thank you for teaching me to always dig deeper into how things work and instilling in me a drive to continually move our field forward. I am also eternally grateful to the members of my committee, Daniel Low, Michael McNitt-Gray, Dan Ruan, and James Lamb who have offered invaluable insight and advice on my project.

The physicists at Cedars-Sinai – John, Tiffany, Yong, Bob, Rob C., Nancy, Wensha, and Kai – have been a tremendous source of knowledge throughout my graduate career. From helping me learn to use clinical equipment, obtain all the necessary measurements for my project, and allowing me to assist with clinical tasks, they welcomed me as a true member of their team from the start. Thank you also to my first office comrades at Cedars – Rob R., Emi, Elizabeth, and Mark – for all the insightful conversations and for your friendship.

My path to medical physics began during my undergraduate years at the University of Michigan, and I am lucky to call many of the researchers who are still there my mentors and friends. Thank you for sharing with me your love for our work and especially to Martha Matuszak and Jean Moran who have remained mentors during my time in graduate school and always supported my endeavors from afar.

There are also many people I would like to thank for their personal support during my graduate years. First, to my classmates at UCLA, especially Victoria, Dan, Anthony, Angelia, and Katelyn, who have become close friends and will remain life-long colleagues. You have been a great support system and resource in navigating the unique experience of graduate school. To my roommate and dear friend Chrisi, and U-M college friends, Sipkje, Liting, Anna, and Margaret, who all have gently and lovingly listened to my struggles and celebrated my triumphs over the years. To Gwen and Foster, who have become a family to me in LA, and to my whole choir at Blessed Sacrament Church - the time we spent together and the music we created was deeply touching and a source of pure joy in my life. And to Mark and Joe, who not only generously offered me unique musical opportunities that were much needed creative releases from work, but who were also sources of loving support during my dissertation years.

My family has been a well-spring of love and encouragement throughout my life and especially during graduate school. Thank you for your compassion and understanding for the moments I missed while working towards my goals. I owe a special thank you to my brother Brian, who, even from a young age, has always pushed me to reach higher by example of his own high ideals. Beyond this, Brian has always managed to stay abreast of my research and career progress and has been a never-failing source of support. Last, but certainly not least, I owe enormous gratitude to my parents, Rosemary and John Steers, who instilled in me a love of music, scientific curiosity, and taught me the value of a strong work ethic. My mother has been a tireless beacon of love and encouragement throughout my life and my most ardent cheerleader, especially these past six years. I would have never made it this far without everything you have done for me.

This work is not just my own, but a collection of efforts from loved ones who have supported me, mentors who have educated me, and researchers who have come before me. Thank you for making this possible.

Vita

Education

2012 – 2015 M.S., Biomedical Physics
University of California – Los Angeles, Los Angeles, CA

2007 – 2012 B.S.E., Nuclear Engineering and Radiological Sciences, Minor in Music, University of Michigan, Ann Arbor, MI

Honors and awards

- UCLA Graduate Student Association Dissertation Year Fellowship for 2017-18
- *Practical Radiation Oncology* (PRO) Journal Peer Review Apprenticeship for 2017
- Cedars-Sinai Phase One Foundation Young Investigator Award in Translational Cancer Research for 2014

Peer-reviewed Publications

1. **Steers, J. M.**, and Fraass, B. A. "IMRT QA: Selecting gamma criteria based on error detection sensitivity". Medical Physics, 43(4), 1982-1994 (2016). PMID: 27036593
2. Matuszak, M. M., **Steers, J. M.**, Long, T., McShan, D. L., Fraass, B. A., Romeijn, H. E., & Ten Haken, R. K. "FusionArc optimization: a hybrid volumetric modulated arc therapy (VMAT) and intensity modulated radiation therapy (IMRT) planning strategy". Medical Physics, 40(7), 071713 (2013). PMID: 23822417
3. Fraass, B. A., **Steers, J. M.**, Matuszak, M. M., & McShan, D. L. "Inverse-optimized 3D conformal planning: Minimizing complexity while achieving equivalence with beamlet IMRT in multiple clinical sites". Medical Physics, 39(6), 3361-3374 (2012). PMID: 2255717

Conference Abstracts

1. **Steers, J.M.,** and Fraass, B.A. "SU-I-GPD-T-266: The Influence of Steep Dose Gradients On the Sensitivity to Errors of IMRT QA Analyzed with Gamma." *Medical Physics* 44.6 (2017). Poster presentation at AAPM annual conference 2017, Denver, CO.
2. **Steers, J.M.,** and Fraass, B.A. "SU-FT-282: Quality Assurance for IMRT/VMAT QA Devices: Issues Affecting the Timing for ArcCHECK Recalibration." *Medical Physics* 43.6 (2016): 3527-3527. Poster presentation at AAPM annual conference 2016, Washington D.C.
3. Boehnke, E.M., DeMarco J., **Steers J.M.,** Fraass, B.A. "SU-G-BRB-03: Assessing the Sensitivity and False Positive Rate of the Integrated Quality Monitor (IQM) Large Area Ion Chamber to MLC Positioning Errors." *Medical Physics* 43.6 (2016): 3631-3631. Oral presentation at AAPM annual conference 2016, Anaheim, CA.
4. **Steers, J.M.**, Fraass, B.A., "WE-D-BRA-02: Dosimetric Impact On Patient Plans From Errors That Can Pass IMRT QA." *Medical Physics* 42.6 (2015): 3667-3667. Oral presentation at AAPM annual conference 2015, Austin, TX.
5. **Steers, J.M.,** Fraass, B.A., "Evaluation of the Potential Clinical Impact of Broad Acceptance Criteria for Radiotherapy IMRT QA", Poster presentation at the Samuel Oschin

Comprehensive Cancer Institute Poster Day 2014, Awarded Phase One Foundation Young Investigator Award in Translational Cancer Research, Los Angeles, CA.

6. **Steers, J.M.**, Fraass, B.A., "TU-C-BRE-08: IMRT QA: Selecting Meaningful Gamma Criteria Based On Error Detection Sensitivity." *Medical Physics* 41.6 (2014): 456-456. Oral Presentation at AAPM annual conference 2014.
7. Fraass, B. A., **Steers, J. M.**, Matuszak, M. M., & McShan, D. L. "SU-E-T-843: Direct Segment Optimization of Conformal Plans: Achieving Equivalence with Beamlet IMRT." *Medical Physics* 38.6 (2011): 3685-3685. Poster presentation at AAPM annual conference 2011.
8. Little, M., Ritter, T., **Steers, J.M.**, Fraass, B.A., "Initial Evaluation of an Inverse-Planned Direct Segment Optimization Technique for Prostate Radiotherapy." *International Journal of Radiation Oncology* Biology* Physics* 81.2 (2011): S901. Poster presentation at ASTRO annual conference 2011.

Seminars

1. **Steers, J.M.**, Matuszak, M.M., Fraass, B.A., **"**Direct Segment Optimization: Optimizing Conformal Plans without IMRT", Presentation at the Great Lakes Chapter of AAPM, Semi-Annual Meeting, December 2010, Troy, MI
2. **Steers, J.M.**, Fraass, B.A., "Direct Segment Optimization: Optimizing Conformal Plans without IMRT," Presentation to the Department of Radiation Oncology, University of Michigan, October 2010, Ann Arbor, MI

Extramural Invited Presentations

1. "Evaluation of the Potential Clinical Impact of Broad Acceptance Criteria for Radiotherapy IMRT QA", **Steers, J.M.**, Cedars-Sinai Medical Center Department of Biomedical Sciences, Presented at the Biomedical and Translational Science Seminar Series, Los Angeles, CA, September 2014
2. "FusionArc: Integrating Arc Therapy and IMRT," **Steers, J.M**., Matuszak, M.M., Presented at the Michigan Section - American Nuclear Society Meeting, Ann Arbor, MI, January 2012

Chapter I.

Introduction

In 2018, it is estimated that approximately 1.7 million new cases of cancer will be diagnosed in the United States alone[1], with roughly 66% of these patients receiving radiation therapy as part of their treatment regime[2]. While radiation therapy can be delivered using a variety of different techniques, patients are most commonly treated with external beam radiotherapy. This treatment utilizes focused beams of high energy particles generated by linear accelerators to induce cell death, with the overall goal of maximizing tumor cell killing while minimizing damage to normal tissues.

Due to the nature of radiation therapy treatments, safety is a main tenet in Radiation Oncology departments. Radiation therapy clinics perform a multitude of quality assurance checks to ensure that patients receive safe, effective, and high-quality treatments free of clinically significant errors. Should treatment errors occur, they can carry devastating consequences – under-dosing the tumor can cause a lack of tumor control, potentially allowing disease spread. Conversely, misadministrations that result in excess dose to the patient can induce unexpected toxicities ranging in severity from atypically severe erythema to death. Patient-specific pre-treatment quality assurance (QA), also colloquially known as intensity modulated radiation therapy (IMRT) quality assurance (QA), is one important pre-treatment check to ensure the plan delivered to the patient matches closely to what was planned in the treatment planning system (TPS). This process, which compares a physical measurement of the patient's plan to a calculation from the TPS prior to treatment, is one crucial component in attempting to prevent large treatment errors. The physical measurement of a patient's plan prior to treatment was introduced partly in response to the increasing complexity of radiotherapy treatment plans over the past few decades. In current radiotherapy deliveries there are many pathways in which a delivered plan may differ from the plan calculated in the TPS, such as improper transfer of plan parameters between the TPS and

the treatment machine, malfunction of the linear accelerator during delivery, or limitations in calculation algorithms.

I.A. The parallel evolution of radiotherapy treatment plans and IMRT QA

The delivery of radiation for therapeutic purposes has developed significantly over the past 30 years, and the ability to spare normal tissues while maintaining dose coverage of the tumor volume has improved greatly. In tandem with the advances seen in delivery techniques, the methods used to ensure that patient plans are safe to deliver have also evolved. Prior to the mid-1980s, radiation therapy treatments often consisted of beams shaped using rectangular high-density jaws in the head of the linear accelerator, and sometimes secondary collimation using a Cerrobend block to shield critical structures. The example in Figure 1-1.a. shows that for these types of simple field shapes, a good deal of normal tissue outside of the tumor volume may be irradiated in order to entirely cover the tumor volume. Since these treatment fields were relatively simple in nature, verifying that plan parameters were properly transferred from the TPS to the delivery machine could be performed with visual verifications or using record-and-verify systems.

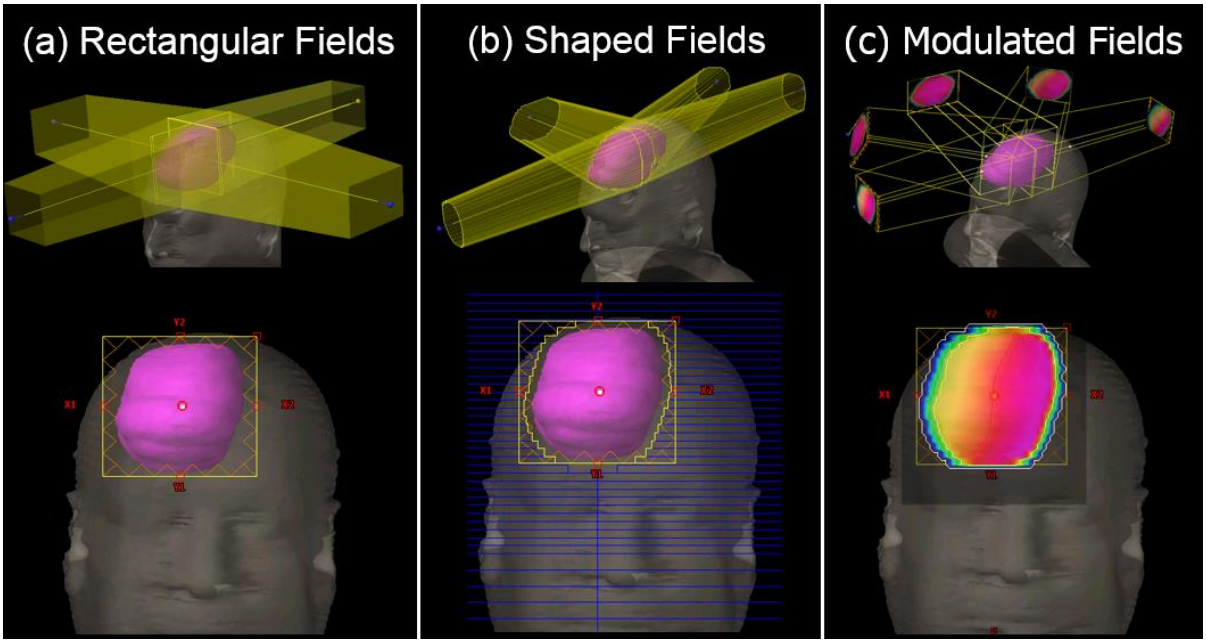

Figure 1-1. The evolution of treatment planning fields, from (a) simple rectangular fields, to (b) fields shaped with multi-leaf collimators (MLCs), and (c) modulated fields created by complex combinations of small MLC-shaped fields.

By the late 1980s and early 1990s, Cerrobend blocks were regularly utilized to create shaped fields shaped to the beam's-eye view (BEV) of the targets[3] and by the mid-1990s multi-leaf collimators (MLCs) were in wide-spread use. MLCs consist of many tungsten "leaves" (illustrated in Figure 1-1.b and Figure 1-2) that can move separately to create unique beam apertures and allow computer-controlled delivery of radiation therapy treatments[4,5]. Since beams shaped with Cerrobend blocks and MLCs were still relatively intuitive relative to the patient anatomy, plan QA checks consisted mainly of visual verification and independent dose calculations (typically a monitor unit calculation).

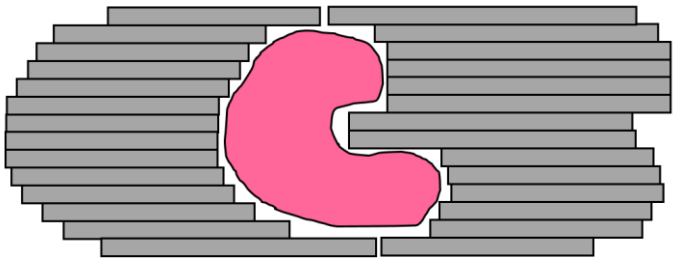

Figure 1-2. An illustration of a multi-leaf collimator from the beam's eye view (BEV)

As advances were made in computing, intensity modulated radiation therapy (IMRT)[6–11] was introduced to radiation therapy. The planning of IMRT makes use of an inverse planning scheme, where users define dose objectives up front, and optimization algorithms attempt to achieve these goals through beamlet, or fluence optimization. Thus, the resulting plans were much more complex in nature, as the dose from a given static beam angle would be modulated as opposed to a uniform intensity, illustrated in Figure 1-1.c (top). The delivery of these modulated fields required algorithms that could convert the fluence maps into many small MLC-shaped apertures, that when summed together resulted in the planned fluence. However, with this development, the delivered apertures became smaller and non-intuitive in shape and one could not simply visually check that the plan that reached the patient was what was intended. Additionally, considerations from other sources of uncertainty, such as beam modeling limitations for small fields and delivery limitations for complex deliveries, led to the use of physical measurements to verify the delivered

dose from the machine prior to a patient's treatment. To ensure all beam segments summed to the expected dose, independent monitor unit (MU) checks[12] were often utilized. In addition to this, clinics began delivering each IMRT plan to a measurement device prior to the patient's treatment. At this time, these measurements were typically made with the use of ion chambers for absolute dose verification and with radiographic film[13–18] to visualize the delivered dose distribution in 2D. These measurements were then compared to TPS calculations of the patient plan superimposed on a model of the measurement geometry in the treatment planning system.

Given the large amount of data available for analysis with 2D film measurements, two new methods to quickly compare these complex dose distributions were introduced – the composite comparison[19] and the gamma comparison[20,21]. It had been previously noted that dose difference comparisons were most useful in low gradient regions of the dose distribution, while distance-to-agreement (DTA), the closest distance between a measurement and calculation point, was well-suited for high gradient regions of dose distribution comparisons[22]. Both the composite and gamma comparison methods made use of this concept, but with slightly different approaches. While the composite comparison was a binary pass/fail test based on a point passing either DTA or the percent dose difference criterion, the gamma comparison combined these parameters into a single metric, γ. These methods offered relatively intuitive ways of evaluating dose distributions, especially for TPS commissioning purposes, and were extended to IMRT QA comparisons.

As film was cumbersome to use, the early 2000s saw the development of 2D planar measurement arrays for IMRT QA which consisted of many diodes or ion chambers[23,24], allowing the use of one measurement device for both absolute dose and dose distribution measurements. A binary pass/fail version of the gamma comparison[25] was also presented, allowing a more efficient method for evaluating IMRT QA comparisons by setting a threshold for the required fraction of comparison pixels that had γ values less than unity. This method of passing or failing IMRT QA comparisons became very common around this time, with many clinics selecting different gamma

criteria to pass or fail a plan in the clinic. Since this binary version of the gamma metric has many different parameters from which the user can select, the variation in gamma comparison settings was not surprising.

Over the course of the following decade new devices for IMRT QA measurements[26,27], new analysis methods[28–30], and even new delivery techniques such as VMAT[31], were introduced clinically. Despite these new technologies, the gamma comparison became the default standard for IMRT QA comparisons and is widely available for use in commercial software packages. In particular, the nearly ubiquitous 3%/3mm TH 10 (global) 90% pixels passing criterion was utilized across the board for new devices and delivery techniques, with the assumption that a high gamma pass rate was indicative of excellent agreement between measurement and calculation. The publication of TG-119 further solidified this gamma criterion as the standard for the field[32], with some institutions reporting the use of even broader acceptance criteria[33] for IMRT QA comparisons. In recent years, it has become apparent that large, clinically significant errors can go undetected when employing this commonly-used gamma criterion for IMRT QA and that gamma passing criteria do not relate well to clinically relevant end-points[23–37]. Due to this issue, and the fact that gamma passing rates may hold different meanings for the various types of detector geometries and delivery techniques,[35,40,44–55] the utility of current IMRT QA comparisons has been called into question.

With the heavy workload IMRT QA places on clinical physics staff, the desire to replace current IMRT QA methods with less onerous plan checks is currently a source of discussion in the field[56,57]. In spite of this, physical measurements for IMRT QA still play an important role in ensuring patient safety and are presently one of the best methods to prevent large misadministration errors, such as those reported in recent New York Times articles[58,59].

I.B. Current measurement methods

A variety of measurement devices exist and are in current clinical use for the purpose of IMRT QA measurements. While film and ion chamber are still in use in some clinics for IMRT QA, some groups have been investigating the use of 3D dosimetry with the use of polymer gels[60–62] and PRESAGE™ materials[63–67]. However, with films, gels, or PRESAGE™, the materials can only be used once and each method can be somewhat labor intensive in terms of handling the materials and processing the measurements. Because of this, detector arrays are the primary measurement devices utilized for the majority of IMRT QA measurements.

Commercially available 2D detector arrays make use of many ion chambers[24,68–70] or diodes[23,26,27,71–74] to measure delivered plans. Arrays come in a variety of geometries from planar, to cylindrical in nature. Three of these differing geometries are shown in Figure 1-3. It is most common for users to deliver the composite plan to these arrays – that is the patient's treatment plan is delivered using the planned gantry angles and the entire measurement is acquired and compared to a composite calculation in the same phantom geometry.

An additional measurement technique makes use of the flat-panel imager, or EPID[75–77] on the linear accelerator for acquiring patient-specific pre-treatment measurements. This method has become increasingly popular in recent years since it allows a reduction in time spent running IMRT QA checks with the added bonus of higher spatial resolution. However, EPID measurements remove an external check on the gantry positioning, so this must be considered when using this as an IMRT QA measurement device. It is important to understand that no measurement method is without limitations, and these limitations must be considered when selecting a measurement device for use in the clinic.

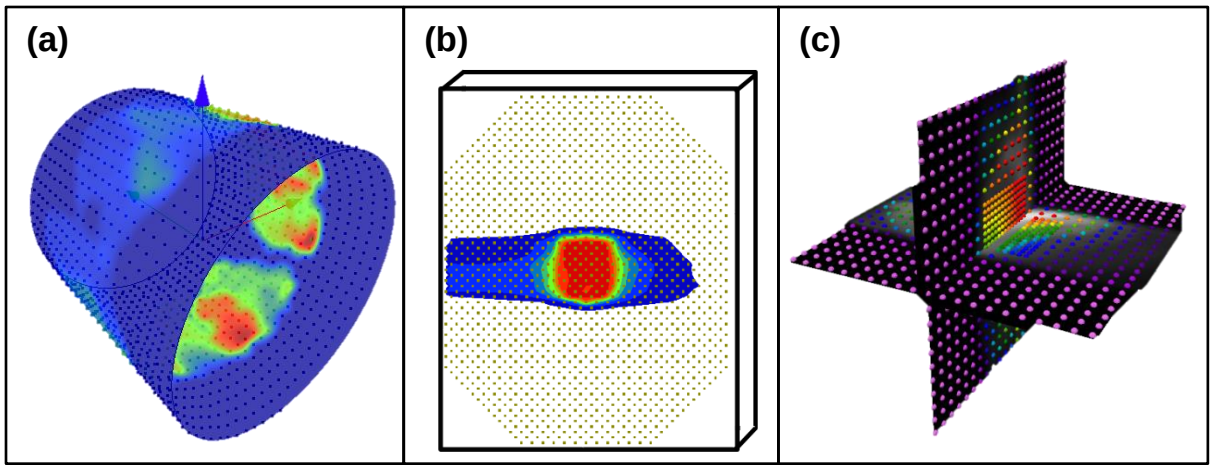

Figure 1-3. Three various commercially available diode array geometries are shown. (a) The ArcCHECK cylindrical array (Sun Nuclear, Melbourne, FL) (b) the MapCheck planar array (Sun Nuclear, Melbourne, FL) and (c) the Delta 4 cross, or bi-planar array (ScandiDos, Uppsala, Sweden).

I.C. Other analysis techniques for IMRT QA

While other analysis techniques have been introduced for comparing measurements and calculations as alternatives to the gamma comparison, none of these are in wide clinical use. Part of this is likely owing to lack of vendor implementation, but also that there is a larger experience among clinics using the gamma comparison. Gradient compensation[28], developed by Moran *et al*., aims to mitigate the effects of geometrical misalignments in dose difference comparisons, which can result from various different sources such as small set-up errors, small delivery uncertainties, or calculational uncertainties, to name a few. Gradient compensation is performed by selecting a geometrical uncertainty parameter (typically 1 mm) and multiplying this by the gradient map of the planned dose and subtracting these dose differences from the original dose difference map. The resulting dose difference map is that which remains after removing these small geometrical uncertainties. While this method is intuitive and easy to apply, it has not been widely implemented commercially or adopted in other clinics.

The normalized agreement test (NAT), presented by Childress and Rosen[29], developed a methodology that removed a large number of points in the analysis if 1) comparison points had dose differences or DTAs less than a specified criterion or 2) the measurement point was lower

than the calculation outside what was assumed to be the PTV region (75% of maximum calculated dose). Only if the measured dose was greater than the calculated dose, or the comparison point fell in the high dose region above 75% maximum was the NAT value computed. The purpose of this was to incorporate regions of the dose distribution in the analysis which are likely to be more biologically meaningful. Bakai *et al.* also introduced a method that modified the gamma comparison and essentially gave similar results to the gamma comparison, but was calculated more efficiently[30].

More recently, the ability to reconstruct 3D dose and dose-volume histograms from measurements performed using 2D detector arrays has been introduced in several commercially available products[78–83]. This allows the user to visualize the dose differences in the patient, as opposed to relying on less clinically meaningful pass/fail metrics. While these methods are highly attractive for IMRT QA plan review, they would be difficult to implement for every IMRT QA case. This is because the time for analysis could cause an unmanageable increase in clinical workload especially if many of the plans did not have concerning dose differences between measurements and calculations. On top of this, the usefulness of these comparisons is limited by the accuracy of the back-projection and dose calculation algorithms utilized.

I.D. Current recommendations and limitations

Despite previous reports on the insensitivity of gamma comparisons, gamma is still the primary method used for IMRT QA analysis. While 3%/3mm TH 10 (G) is the most commonly used gamma criterion[33], the recent publication of TG-218 has suggested the use of a stricter criterion – 3%/2mm TH10 (Global) with 90% pixels passing[84]. TG-218 also cautions against the use of the gamma percent pixels passing as the only metric to evaluate when looking at IMRT QA results, instead suggesting that the location of gamma failures, as well as gamma value histograms be evaluated as well. TG-218 also does not address sensitivity differences between measurement devices and delivery techniques, and has thus recommended the use of a universal gamma

criterion, though it is unclear if this is appropriate. Regardless of stricter gamma criteria and closer analysis of the results, these current IMRT QA methods are limited by the fact that the gamma analysis results are difficult to interpret since there is no direct link to clinically-meaningful dose metrics.

While plans rarely fail QA with the gamma comparison, TG-218 does offer a list of items to double-check in the presence of a failing result. However, there is often no clear explanation for the plan failure, how to improve the passing rate, and most importantly, if the dose differences between measurement and calculation are clinically significant. In many instances, a failing plan is simply re-planned with the hope of obtaining passing results for the new plan or gamma criteria are relaxed to provide a "passing" result. In a survey by Nelms *et al.* a shocking 46% of respondents admitted to changing the gamma criteria to achieve a higher passing rate in the presence of a failing result[33], which does nothing to ensure the patient's treatment is free of errors. It is clear that a better understanding of current limitations in IMRT QA, as well as more clinically useful methods for analyzing the data, are necessary.

I.E. Dissertation overview

This chapter has introduced the evolution of pre-treatment patient-specific QA checks, known colloquially as intensity modulated radiation therapy (IMRT) QA, as well as the current measurement devices and comparison techniques used in the field. While IMRT QA is an important step in ensuring a patient's radiation treatment plan is safe to deliver, several limitations exist in current IMRT QA comparisons, several of which this work aims to address. It is well known that gamma comparisons with commonly used criteria can be insensitive to large errors, should these errors exist in patient treatment plans. Additionally, gamma comparison results do not relate to differences in patient dose metrics, making interpretation of the comparison results difficult.

Chapter II introduces a new method, called the error curve method, that allows us to objectively quantify the error sensitivity of the gamma comparison for any detector device. This method, rather than reporting gamma passing rates in the presence of induced errors, allows us to determine the range of errors that could pass the gamma comparison for a cohort of test patient plans. This chapter specifically studies the range of errors that could pass IMRT QA for the ArcCHECK device for 36 different combinations of gamma criteria, allowing us to determine how sensitivity changes with many different combinations of gamma parameters.

Chapter III makes use of the error curve method presented in Chapter II to explore different aspects of the measurement scenario that may affect error sensitivity in IMRT QA comparisons for the ArcCHECK, MapCHECK, and Delta 4 phantoms. These experiments were performed by developing in-house MATLAB code that allows calculation-only gamma comparisons between error-free and error-induced calculations for two spatial sampling scenarios – 1mm vs. 1mm calculations and 1mm vs. calculations at the real spatial sampling of a given detector geometry. The use of calculation vs. calculation gamma comparisons removes uncertainties and limitations that can confound results when using real measurements. This allows us to elucidate the true effects of measurement array geometry, spatial sampling, and delivery technique on gamma comparison sensitivity for five different gamma criteria. The work in this chapter also allows us to determine if different gamma criteria may be required for different devices and delivery techniques in the clinic, which is important as the same clinical passing criteria are typically applied regardless of measurement device and delivery technique.

In Chapter IV errors of increasing magnitude are induced in plan calculations and compared to plan features to understand if particular features in the planned dose distributions, such as dose gradients, have qualitative or quantitative relationships between gamma value maps and error curve ranges. Additionally, the number of measurement points falling within different sensitive

regions of the plan dose comparisons are investigated to understand if any of these features drive gamma comparison insensitivity.

Based on the work in Chapter IV, Chapter V investigates a method for segmenting IMRT QA comparisons based on gradient and dose thresholds as a new way of comparing IMRT QA dose distributions that offers i) more clinically meaningful results than the gamma comparison ii) higher sensitivity to errors than current comparisons iii) is simple and fast to calculate ***and*** evaluate and iv) is unlikely to significantly inhibit clinical workflow if implemented clinically.

Chapter VI validates the new method presented in Chapter V with a separate cohort of test plans. Additionally, 20 real MapCHECK measurements are evaluated with the new method and with the gamma comparison to compare differences in sensitivity when using real measurements and evaluate potential clinical feasibility. The strengths and weaknesses of the new method are discussed, as well as considerations for clinical implementation. Finally, Chapter VII summarizes the results of this work and discusses future work that can be done to further advance the field of IMRT QA.

References

1. National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program Fact Sheets: All Cancer Sites. (www.seer.cancer.gov). Accessed April 23, 2018.

2. Physician Characteristics and Distribution in the U.S., 2010 Edition, 2004 IMV Medical Information Devision, 2003 SROA Benchmarking Survey.

3. Goitein M, Abrams M, Rowell D, Pollari H, Wiles J. Multi-dimensional treatment planning: II. Beam's eye-view, back projection, and projection through CT sections. *Int J Radiat Oncol Biol Phys*. 1983;9(6):789-797. doi:10.1016/0360-3016(83)90003-2.

4. Brahme A. Optimal usage of multileaf collimators in stationary beam radiation therapy. *Strahlentherapie und Onkol Organ der Dtsch Rontgengesellschaft.[et al]*. 1988;164(6):343-350.

5. McShan DL, Fraass BA, Lichter AS. Full integration of the beam's eye view concept into computerized treatment planning. *Int J Radiat Oncol*. 1990;18(6):1485-1494. doi:10.1016/0360-3016(90)90325-E.

6. Brahme A. Optimization of stationary and moving beam radiation therapy techniques. *Radiother Oncol*. 1988;12(2):129-140. doi:10.1016/0167-8140(88)90167-3.

7. Bortfeld T, Bürkelbach J, Boesecke R, Schlegel W. Methods of image reconstruction from projections applied to conformation radiotherapy. *Phys Med Biol*. 1990;35(10):1423-1434. doi:10.1088/0031-9155/35/10/007.

8. Webb S. Optimization by simulated annealing of three-dimensional conformal treatment planning for radiation fields defined by a multileaf collimator. *Phys Med Biol*. 1991;36(9):1201-1226. doi:10.1088/0031-9155/36/9/004.

9. Mackie TR, Holmes T, Swerdloff S, et al. Tomotherapy: A new concept for the delivery of dynamic conformal radiotherapy. *Med Phys*. 1993;20(6):1709-1719. doi:10.1118/1.596958.

10. Stein J, Bortfeld T, Dörschel B, Schlegel W. Dynamic X-ray compensation for conformal radiotherapy by means of multi-leaf collimation. *Radiother Oncol*. 1994;32(2):163-173. doi:10.1016/0167-8140(94)90103-1.

11. Carol M, Grant III WH, Pavord D, et al. Initial Clinical Experience with the Peacock Intensity Modulation of a 3-D Conformal Radiation Therapy System. *Stereotact Funct Neurosurg*. 1996;66(1-3):30-34. doi:10.1159/000099664.

12. Xing L, Chen Y, Luxton G, Li JG, Boyer a L. Monitor unit calculation for an intensity modulated photon field by a simple scatter-summation algorithm. *Phys Med Biol*. 2000;45(3):N1-N7. doi:10.1109/IEMBS.2000.900697.

13. Low DA, Chao KSC, Mutic S, Gerber RL, Perez CA, Purdy JA. Quality assurance of serial tomotherapy for head and neck patient treatments. *Int J Radiat Oncol Biol Phys*. 1998;42(3):681-692. doi:10.1016/S0360-3016(98)00273-9.

14. Olch AJ. Dosimetric performance of an enhanced dose range radiographic film for intensity-modulated radiation therapy quality assurance. *Med Phys*. 2002;29(9):2159-2168. doi:10.1118/1.1500398.

15. Zhu XR, Jursinic PA, Grimm DF, Lopez F, Rownd JJ, Gillin MT. Evaluation of Kodak

EDR2 film for dose verification of intensity modulated radiation therapy delivered by a static multileaf collimator. *Med Phys*. 2002;29(8):1687-1692. doi:10.1118/1.1493781.

16. Dogan N, Leybovich LB, Sethi A. Comparative evaluation of Kodak EDR2 and XV2 films for verification of intensity modulated radiation therapy. *Phys Med Biol*. 2002;47(22):4121-4130. doi:10.1088/0031-9155/47/22/314.

17. Childress NL, Salehpour M, Dong L, Bloch C, White RA, Rosen II. Dosimetric accuracy of Kodak EDR2 film for IMRT verifications. *Med Phys*. 2005;32(2):539-548. doi:10.1118/1.1852791.

18. Childress NL, Dong L, Rosen II. Rapid radiographic film calibration for IMRT verification using automated MLC fields. *Med Phys*. 2002;29(10):2384-2390. doi:10.1118/1.1509441.

19. Harms WB, Low DA, Wong JW, Purdy JA. A software tool for the quantitative evaluation of 3D dose calculation algorithms. *Med Phys*. 1998;25(10):1830-1836. doi:10.1118/1.598363.

20. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys*. 1998;25(5):656. doi:10.1118/1.598248.

21. Low D a., Dempsey JF. Evaluation of the gamma dose distribution comparison method. *Med Phys*. 2003;30(9):2455. doi:10.1118/1.1598711.

22. Dyk J Van, Barnett RB, Cygler JE, Shragge PC. Commissioning and quality assurance of treatment planning computers. *Int J Radiat Oncol Biol Phys*. 1993;26(2):261-273. doi:10.1016/0360-3016(93)90206-B.

23. Jursinic P a, Nelms BE. A 2-D diode array and analysis software for verification of intensity modulated radiation therapy delivery. *Med Phys*. 2003;30(5):870-879. doi:10.1118/1.1567831.

24. Spezi E, Angelini AL, Romani F, Ferri a. Characterization of a 2D ion chamber array for the verification of radiotherapy treatments. *Phys Med Biol*. 2005;50(14):3361-3373. doi:10.1088/0031-9155/50/14/012.

25. Depuydt T, Van Esch A, Huyskens DP. A quantitative evaluation of IMRT dose distributions: Refinement and clinical assessment of the gamma evaluation. *Radiother Oncol*. 2002;62(3):309-319. doi:10.1016/S0167-8140(01)00497-2.

26. Feygelman V, Opp D, Javedan K, Saini AJ, Zhang G. Evaluation of a 3D Diode Array Dosimeter for Helical Tomotherapy Delivery QA. *Med Dosim*. 2010;35(4):324-329. doi:10.1016/j.meddos.2009.10.007.

27. Feygelman V, Zhang G, Stevens C, Nelms BE. Evaluation of a new VMAT QA device, or the “X” and “O” array geometries. *J Appl Clin Med Phys*. 2011;12(2):3346. doi:10.1120/jacmp.v12i2.3346.

28. Moran JM, Radawski J, Fraass BA. A dose gradient analysis tool for IMRT QA. *J Appl Clin Med Phys*. 2005;6(2):62-73. doi:10.1120/jacmp.2024.25338.

29. Childress NL, Rosen II. The design and testing of novel clinical parameters for dose comparison. *Int J Radiat Oncol Biol Phys*. 2003;56(5):1464-1479. doi:10.1016/S0360-3016(03)00430-9.

30. Bakai A, Alber M, Fridtjof N. A revision of the γ-evaluation concept for the comparison of

dose distributions. *Phys Med Biol*. 2003;48(21):3543. doi:10.1088/0031-9155/48/21/006.

31. Otto K. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Med Phys*. 2008;35(1):310-317. doi:10.1118/1.2818738.

32. Ezzell GA, Burmeister JW, Dogan N, et al. IMRT commissioning: Multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119. *Med Phys*. 2009;36(11):5359-5373. doi:10.1118/1.3238104.

33. Nelms BE, Simon JA. A survey on planar IMRT QA analysis. *J Appl Clin Med Phys*. 2007;8(3):2448. doi:10.1120/jacmp.v8i3.2448.

34. Jin X, Yan H, Han C, Zhou Y, Yi J, Xie C. Correlation between gamma index passing rate and clinical dosimetric difference for pre-treatment 2D and 3D volumetric modulated arc therapy dosimetric verification. *Br J Radiol*. 2015;88(1047). doi:10.1259/bjr.20140577.

35. Templeton AK, Chu JCH, Turian J V. The sensitivity of ArcCHECK-based gamma analysis to manufactured errors in helical tomotherapy radiation delivery. *J Appl Clin Med Phys*. 2015;16(1):4814. doi:10.1120/jacmp.v16i1.4814.

36. Vieillevigne L, Molinier J, Brun T, Ferrand R. Gamma index comparison of three VMAT QA systems and evaluation of their sensitivity to delivery errors. *Phys Medica*. 2015;31(7):720-725. doi:10.1016/j.ejmp.2015.05.016.

37. Yan G, Liu C, Simon TA, Peng LC, Fox C, Li JG. On the sensitivity of patient-specific IMRT QA to MLC positioning errors. *J Appl Clin Med Phys*. 2009;10(1):120-128. doi:10.1118/1.2961902.

38. Nelms BE, Zhen H, Tomé W a. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. *Med Phys*. 2011;38(2):1037-1044. doi:10.1118/1.3544657.

39. Zhen H, Nelms BE, Tomé W a. Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA. *Med Phys*. 2011;38(10):5477. doi:10.1118/1.3633904.

40. Stasi M, Bresciani S, Miranti A, Maggio A, Sapino V, Gabriele P. Pretreatment patient-specific IMRT quality assurance : A correlation study. *Med Phys*. 2012;39(12):7626-7634. doi:10.1118/1.4767763.

41. Nelms BE, Chan MF, Jarry G, et al. Evaluating IMRT and VMAT dose accuracy: practical examples of failure to detect systematic errors when applying a commonly used metric and action levels. *Med Phys*. 2013;40(11):111722. doi:10.1118/1.4826166.

42. Nelms B, Jarry G, Chan M, Hampton C, Watanabe Y, Feygelman V. Real-world examples of sensitivity failures of the 3%/3mm pass rate metric and published action levels when used in IMRT/VMAT system commissioning. *J Phys Conf Ser*. 2013;444:012086. doi:10.1088/1742-6596/444/1/012086.

43. Chan. Using a Novel Dose QA Tool to Quantify the Impact of Systematic Errors Otherwise Undetected by Conventional QA Methods: Clinical Head and Neck Case Studies. *Technol Cancer Res Treat*. 2013;13(1):1-11. doi:10.7785/tcrt.2012.500353.

44. Neilson C, Klein M, Barnett R, Yartsev S. Delivery quality assurance with ArcCHECK. *Med Dosim*. 2013;38(1):77-80. doi:10.1016/j.meddos.2012.07.004.

45. Pulliam KB, Followill D, Court L, et al. A six-year review of more than 13,000 patient-specific IMRT QA results from 13 different treatment sites. *J Appl Clin Med Phys*. 2014;15(5):4935. doi:10.1120/jacmp.v15i5.4935.

46. Bresciani S, Di Dia A, Maggio A, et al. Tomotherapy treatment plan quality assurance: the impact of applied criteria on passing rate in gamma index method. *Med Phys*. 2013;40(12):121711. doi:10.1118/1.4829515.

47. Cruz W, Narayanasamy G, Regan M, et al. Patient specific IMRT quality assurance with film, ionization chamber and detector arrays: Our institutional experience. *Radiat Phys Chem*. 2015;115:12-16. doi:10.1016/j.radphyschem.2015.06.002.

48. Both S, Alecu IM, Stan AR, et al. A study to establish reasonable action limits for patient-specific quality assurance in intensity-modulated radiation therapy. *J Appl Clin Med Phys*. 2007;8(2):1-8.

49. Basran PS, Woo MK. An analysis of tolerance levels in IMRT quality assurance procedures. *Med Phys*. 2008;35(6):2300. doi:10.1118/1.2919075.

50. Pawlicki T, Yoo S, Court LE, et al. Moving from IMRT QA measurements toward independent computer calculations using control charts. *Radiother Oncol*. 2008;89(3):330-337. doi:10.1016/j.radonc.2008.07.002.

51. Sanghangthum T, Suriyapee S, Srisatit S, Pawlicki T. Statistical process control analysis for patient-specific IMRT and VMAT QA. *J Radiat Res*. 2013;54(3):546-552. doi:10.1093/jrr/rrs112.

52. Hussein M, Rowshanfarzad P, Ebert MA, Nisbet A, Clark CH. A comparison of the gamma index analysis in various commercial IMRT/VMAT QA systems. *Radiother Oncol*. 2013;109(3):370-376. doi:10.1016/j.radonc.2013.08.048.

53. McKenzie EM, Balter P a, Stingo FC, Jones J, Followill DS, Kry SF. Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans. *Med Phys*. 2014;41(12):121702. doi:10.1118/1.4899177.

54. Wen N, Zhao B, Kim J, et al. IMRT and RapidArc commissioning of a TrueBeam linear accelerator using TG-119 protocol cases. *J Appl Clin Med Phys*. 2014;15(5):4843. doi:10.1120/jacmp.v15i5.4843.

55. Mancuso GM, Fontenot JD, Gibbons JP, Parker BC. Comparison of action levels for patient-specific quality assurance of intensity modulated radiation therapy and volumetric modulated arc therapy treatments. *Med Phys*. 2012;39(7):4378. doi:10.1118/1.4729738.

56. Siochi RAC, Molineu A, Orton CG. Patient-specific QA for IMRT should be performed using software rather than hardware methods. *Med Phys*. 2013;40(7):070601. doi:10.1118/1.4794929.

57. Smith JC, Dieterich S, Orton CG. It is STILL necessary to validate each individual IMRT treatment plan with dosimetric measurements before delivery. *Med Phys*. 2011;38(2):553-555. doi:10.1118/1.3512801.

58. W. Bogdanich. Radiation Offers New Cures, and Ways to Do Harm. *The New York Times*. 2010:A1.

59. W. Bogdanich. As Technology Surges, Radiation Safeguards Lag,. *New York Times2*.

2010;A1.

60. McJury M, Oldham M, Cosgrove VP, et al. Radiation dosimetry using polymer gels: methods and applications. *Br J Radiol*. 2000;73(873):919-929. doi:10.1259/bjr.73.873.11064643.

61. Gum F, Scherer J, Bogner L, Solleder M, Rhein B, Bock M. Preliminary study on the use of an inhomogeneous anthropomorphic Fricke gel phantom and 3D magnetic resonance dosimetry for verification of IMRT treatment plans. *Phys Med Biol*. 2002;47(7):N67-N77. doi:10.1088/0031-9155/47/7/401.

62. Cheng-Shie W, Y. X. Three-dimensional dose verification for intensity modulated radiation therapy using optical CT based polymer gel dosimetry. *Med Phys*. 2006;33(5):1412-1419. doi:10.1118/1.2188820.

63. Pengyi G, John A, Mark O. A practical three-dimensional dosimetry system for radiation therapy. *Med Phys*. 2006;33(10):3962-3972. doi:10.1118/1.2349686.

64. Y. GP, A. AJ, M. O. Characterization of a new radiochromic three-dimensional dosimeter. *Med Phys*. 2006;33(5):1338-1345. doi:10.1118/1.2192888.

65. Mark O, Harshad S, Pengyi G, John A. An investigation of the accuracy of an IMRT dose distribution using two- and three-dimensional dosimetry techniques. *Med Phys*. 2008;35(5):2072-2080. doi:10.1118/1.2899995.

66. Oldham M, Thomas A, O'Daniel J, et al. A quality assurance method that utilizes 3D dosimetry and facilitates clinical interpretation. *Int J Radiat Oncol Biol Phys*. 2012;84(2):540-546. doi:10.1016/j.ijrobp.2011.12.015.

67. Harshad S, David S, John A, Geoffrey I, Mark O. Investigation of the feasibility of relative 3D dosimetry in the Radiologic Physics Center Head and Neck IMRT phantom using Presage/optical-CT. *Med Phys*. 2009;36(7):3371-3377. doi:10.1118/1.3148534.

68. S. A, A. B, F. B, et al. Dosimetric characterization of a large area pixel-segmented ionization chamber. *Med Phys*. 2004;31(2):414-420. doi:10.1118/1.1639992.

69. Stasi M, Giordanengo S, Cirio R, et al. D-IMRT verification with a 2D pixel ionization chamber: dosimetric and clinical results in head and neck cancer. *Phys Med Biol*. 2005;50(19):4681-4694. doi:10.1088/0031-9155/50/19/017.

70. Poppe B, Blechschmidt A, Djouguela A, et al. Two-dimensional ionization chamber arrays for IMRT plan verification. *Med Phys*. 2006;33(0094-2405; 0094-2405; 4):1005-1015. doi:10.1118/1.2179167.

71. Sadagopan R, Bencomo JA, Martin RL, Nilsson G, Matzen T, Balter PA. Characterization and clinical evaluation of a novel IMRT quality assurance system. *J Appl Clin Med Phys*. 2009;10(2):2928. http://www.ncbi.nlm.nih.gov/pubmed/19458595.

72. Létourneau D, Publicover J, Kozelka J, Moseley DJ, Jaffray D a. Novel dosimetric phantom for quality assurance of volumetric modulated arc therapy. *Med Phys*. 2009;36(5):1813-1821. doi:10.1118/1.3117563.

73. Yan G, Lu B, Kozelka J, Liu C, Li JG. Calibration of a novel four-dimensional diode array. *Med Phys*. 2010;37(1):108-115. doi:10.1118/1.3266769.

74. Al Shukaili K, Petasecca M, Newall M, Rosenfeld AB. A 2D silicon detector array for

quality assurance in small field dosimetry: DUO. *Unpublished*. 2016. doi:10.13040/IJPSR.0975-8232.6(9).3807-20.

75. B. GP, C. PC. Dosimetric properties of an amorphous silicon electronic portal imaging device for verification of dynamic intensity modulated radiation therapy. *Med Phys*. 2003;30(7):1618-1627. doi:10.1118/1.1582469.

76. B. W, S. S, S. R, G. FB. Dosimetric IMRT verification with a flat-panel EPID. *Med Phys*. 2003;30(12):3143-3155. doi:10.1118/1.1625440.

77. Moran JM, Roberts DA, Nurushev TS, Antonuk LE, el Mohri Y, Fraass BA. An Active Matrix Flat Panel Dosimeter (AMFPD) for in-phantom dosimetric measurements. *Med Phys*. 2005;32(2):466-472. doi:10.1118/1.1855012.

78. Boggula R, Lorenz F, Mueller L, et al. Experimental validation of a commercial 3D dose verification system for intensity-modulated arc therapies. *Phys Med Biol*. 2010;55(19):5619-5633. doi:10.1088/0031-9155/55/19/001.

79. Nakaguchi Y, Araki F, Maruyama M, Saiga S. Dose verification of IMRT by use of a COMPASS transmission detector. *Radiol Phys Technol*. 2012;5(1):63-70. doi:10.1007/s12194-011-0137-y.

80. Nelms BE, Opp D, Robinson J, et al. VMAT QA: Measurement-guided 4D dose reconstruction on a patient. *Med Phys*. 2012;39(7):4228-4238. doi:10.1118/1.4729709.

81. Olch AJ. Evaluation of the accuracy of 3DVH software estimates of dose to virtual ion chamber and film in composite IMRT QA. *Med Phys*. 2011;39(1):81-86. doi:10.1118/1.3666771.

82. Opp D, Nelms BE, Zhang G, Stevens C, Feygelman V. Validation of measurement-guided 3D VMAT dose reconstruction on a heterogeneous anthropomorphic phantom. *J Appl Clin Med Phys*. 2013;14(4):70-84. doi:10.1120/jacmp.v14i4.4154.

83. Renner WD, Norton K, Holmes T. A method for deconvolution of integrated electronic portal images to obtain incident fluence for dose reconstruction. *J Appl Clin Med Phys*. 2005;6(4):22-39. doi:10.1120/jacmp.v6i4.2104.

84. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. *Med Phys*. 2018;45(4):e53-e83. doi:10.1002/mp.12810.

Chapter II.

Methods for quantifying the sensitivity of the gamma comparison for IMRT QA

II.A. Motivation

With the advent of increasingly complex treatment deliveries in radiation therapy, patient plan-specific quality assurance (QA) has become an integral component in the process of preparing a treatment plan for delivery. Safety considerations for intensity modulated radiation therapy (IMRT) programs have been addressed by various guidance documents which suggest the implementation of appropriate patient-specific QA programs[1-4]. The purpose of patient-specific IMRT QA is to ensure that all plan parameters are properly transferred from the planning system to the treatment machine and that the measured plan closely matches the plan predicted by the treatment planning system (TPS). This is typically achieved by comparing plan calculations to measurements taken with an ion chamber, film, and/or detector array. The most commonly employed technique for comparing measurements and calculations is the gamma comparison[5, 6], originally introduced by Low *et al*. The gamma comparison combines both distance-to-agreement (DTA) and percent dose difference (%Diff) measures into one metric, γ:

$$\gamma = min \left| \sqrt{\frac{\Delta d^2}{\%dose^2} + \frac{\Delta D^2}{DTA^2}} \right| \qquad [Eq. 2.1]$$

where %dose is the user-defined percent dose difference parameter, DTA is the user-defined distance-to-agreement parameter, Δd is the percent dose difference between measurement and calculation point, and ΔD is the DTA between measurement and calculation point. The minimum value within the user-defined radius is the final gamma value. Two dose difference normalization schemes are possible with the gamma comparison – local and global. The local normalization normalizes the absolute dose difference by the dose value at that comparison point, whereas the

global normalization normalizes the dose difference at each comparison point by the maximum dose in the evaluated distribution.

In commercial implementations a low dose threshold (TH) can be applied to remove dose points with a value below a user-defined dose threshold, with the most commonly used threshold being 10%. Points in the comparison with gamma values less than unity are said to pass the comparison, whereas gamma values greater than one are labeled are failing. The final result is the percentage of comparison pixels “passing”, with gamma values less than one. This result is used to determine if the measurement and calculation agree within an acceptable range.

The gamma comparison was initially developed for use in determining agreement between test fields in commissioning scenarios. However, as IMRT QA measurements began to utilize measurement arrays with many data points to evaluate, the gamma comparison was employed to analyze these complex dose distributions. As the use of the gamma comparison became widespread, gamma criteria and tolerance limits for IMRT QA programs were left to each clinic’s discretion. 3%/3mm, TH=10% became a default standard for many clinics, and the recommendations published in TG-119[7] propagated the use of this criterion for the use of IMRT QA. This criterion was also ubiquitously applied as different measurement devices came into clinical use, often without appropriate sensitivity studies.

Numerous studies have reported clinically observed gamma passing rates and proposed gamma criteria and tolerance limits for their clinics[8-13]. Pawlicki *et al*. also demonstrated the utility of process control charts for IMRT QA as a means of determining more meaningful action levels[14]. More recent studies have reported on the insensitivity of commonly used criteria (namely 3%/3mm) to detect clinically significant errors for different measurement scenarios[15-24], however, most do not identify more appropriate gamma criteria for use. A study by Kry *et al.* retrospectively analyzed over 700 head and neck phantom credentialing results and the corresponding institutional IMRT QA results, obtained by the Houston Imaging and Radiation Oncology Core

(IROC). They found that the reported IMRT QA results were not able to predict whether an institution would pass or fail the IROC credentialing process[25]. Additionally, other studies have shown the inability of 3%/3mm and 2%/2mm to identify plans deemed as unacceptable to treat[26, 27]. However, even in light of these findings, the majority of radiation oncology clinics still employ 3%/3mm TH=10% as their standard criterion, with some clinics reporting the use of even broader criteria[28].

The subject of appropriate gamma criteria and tolerances for IMRT QA is further complicated by the assortment of measurement devices available, delivery techniques employed, and differences in treatment site complexity. Different QA systems have been shown to exhibit dissimilar sensitivities for the same gamma criteria[27, 29], suggesting that gamma criteria should be device-specific. Differences between QA results for IMRT deliveries and VMAT deliveries are less straightforward. Sanghangthum *et al*.[30] reported less variation of random errors and fewer systematic errors in VMAT results compared to IMRT results. However, different measurement devices were used for the two delivery techniques. A study by Wen *et al.* noted statistically significant differences in IMRT and VMAT QA ion chamber measurements for 10 MV and 15 MV plans but not for 6 MV, 6 MV-FFF, or 10 MV-FFF plans[31]. Finally, another study reported no statistically significant differences between IMRT and VMAT deliveries when using ion chamber point doses, radiochromic film, or a 2D diode array[32].

A number of institutions have also reported differences in gamma passing rates for different treatment sites[8, 10, 14, 21, 33-36]. Another study noted trends in gamma passing rates between case types, but the differences were not statistically significant[37]. Garcia-Vicente *et al.* also reported unremarkable differences between site-specific gamma passing rates using the ArcCHECK® device[13]. Additionally, there has been little evidence suggesting an appropriate low dose threshold setting for gamma comparisons. While a 10% threshold is most commonly used, it has been acknowledged that the same dose threshold setting for different device geometries could have

different meanings[12, 38]. Finally, Bailey *et al.* highlighted that some publications fail to report all gamma comparison settings (absolute dose vs. relative dose, global vs. local normalization, 2D vs. 3D, and uncertainty on/off) which can have a large effect on the reported gamma results[39].

Clearly, the variety of measurement devices, treatment techniques, options for gamma comparison criteria, and gamma comparison settings play a large role in gamma results, making direct comparisons between publications somewhat challenging and conclusions about appropriate gamma criteria even more unclear. However, from these publications, it seems likely that standardized gamma settings and IMRT QA plan acceptance criteria should ultimately be specific to the measurement device, delivery technique, and perhaps the treatment site.

The question still remains as to what constitutes appropriately sensitive gamma criteria for clinical use. Here we describe a new method, called the error curve method, which quantitatively evaluates the sensitivity of gamma criteria to induced errors in the TPS calculations. We show how these sensitivities may aid in the selection of appropriate gamma criteria for static beam IMRT cases measured on the ArcCHECK® detector device. Furthermore, the error curve method is straightforward to apply to other combinations of measurement devices, treatment sites, and delivery techniques.

II.B. Methods

II.B.1. Overview

We introduce a method to quantitatively evaluate the sensitivity of many different combinations of gamma criteria. Gamma comparisons are made between measurements without induced errors and plan calculations with known errors of varying magnitudes. For each combination of test case, gamma criterion, and error type, the percent pixels passing are graphed versus the corresponding error magnitude in order to create what we call an “error curve”. The width of each error curve represents the sensitivity of a specific gamma criterion to the studied error. This method is described in detail below.

II.B.2. Plan QA measurements

This study made use of 11 fixed-beam DMLC IMRT plans previously treated at our institution. Cases were chosen to include a range of treatment sites, field sizes, complexity, and fractionation schemes. Composite plans were delivered to the ArcCHECK® detector device[38, 40, 41] (Sun Nuclear Corporation, Melbourne, FL) using 6 MV photons from a Varian Trilogy linear accelerator with the Millennium 120 leaf multi-leaf collimator (MLC). Data were acquired using the vendor-provided ArcCHECK® software, SNC Patient (version 6.2.3).

Briefly, the ArcCheck device, shown in Figure 2-1, consists of a cylindrical array of 1386 diodes arranged helically along the cylindrical surface. The diodes are located at a radius of 10.5 cm away from the center of the device. The phantom is constructed of PMMA and is modeled in the TPS as a homogenous cylindrical structure. Calibration of the ArcCHECK® device was performed prior to the measurements. Additionally, all measurements were performed on the same day to minimize the effects of setup variation and daily output fluctuations. Ten additional plans were measured as a validation dataset approximately four months after the initial dataset. Recalibration of the device was not performed prior to the validation measurements, as the change in diode response over the four months was negligible, as shown in Figure 2-2. As a note, all cases included in this study would have passed QA clinically at 3%/3mm TH=10% 95% PP with gamma passing rates ranging from 95.6% to 100% and a mean gamma passing of 98.8% for the 21 cases.

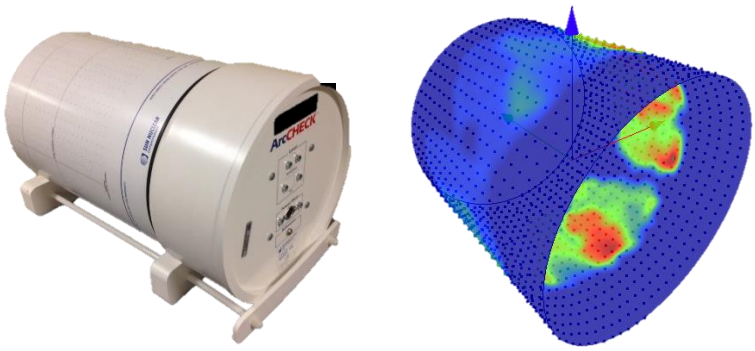

Figure 2-1. (Left) Picture of an actual ArcCHECK device (Sun Nuclear Corporation, Melbourne, FL) and (Right) the diode locations and measured dose for a sample ArcCHECK measurement.

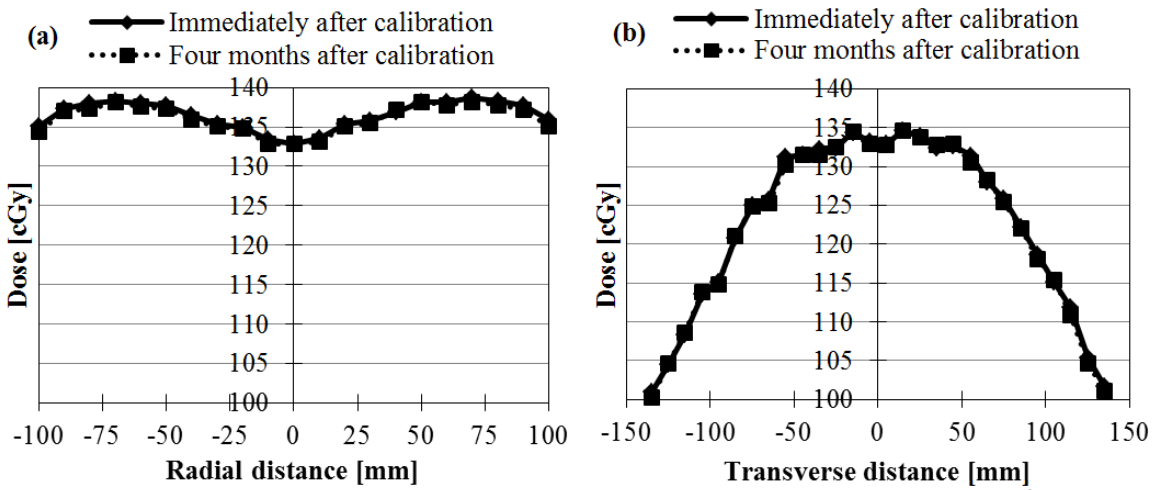

Figure 2-2. Minimal differences in measurements taken four months apart are shown for a 25x25 cm^2 field measured on the ArcCHECK® device for *(a)* a radial profile and *(b)* a transverse profile across the device. The difference in machine output between the two measurement dates was 0.30% and was accounted for by scaling the dotted line by this value. Differences in set-up were not taken into account.

II.B.3. TPS calculations

Patient plan beam arrangements were superimposed on the MVCT ArcCHECK® phantom (provided by Sun Nuclear) in the Eclipse (Varian Medical Systems, Palo Alto, CA) treatment planning system (TPS). Dose calculations on the ArcCHECK® phantom were performed on a 1 mm grid using the Analytical Anisotropic Algorithm (AAA, version 8). It is relevant to note that when calculations with grid sizes larger than 1 mm are loaded into SNC Patient (version 6.2.3), the program automatically interpolates the calculated dose down to a 1 mm grid for comparison to measurements. We noted that the gamma comparison results varied based on grid size for several different gamma criteria. This difference is likely due in part to the inherent differences in the dose calculation at different grid sizes but also due to the interpolation in SNC Patient. The effect of grid size on the gamma comparison results varied from case to case, as shown for two example cases in Figure 2-3. Thus, calculation grid sizes should be carefully considered in gamma comparison studies, as the chosen grid size may greatly affect the results depending on the case type, TPS, and analysis software.

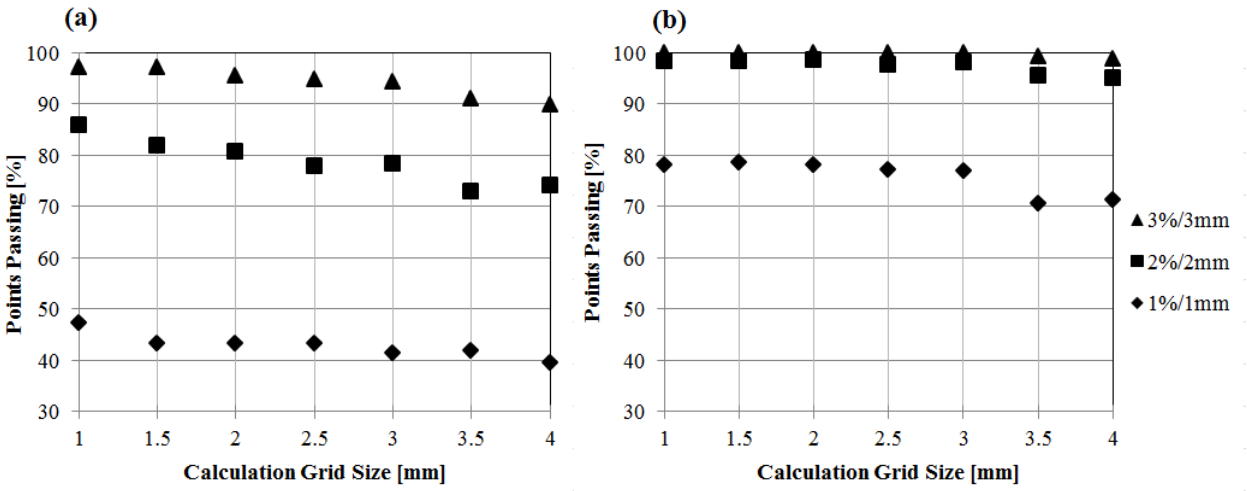

Figure 2-3. Percent pixels passing are shown for three different gamma criteria as a function of grid size for (a) a brain case and (b) a head and neck case. The effect of grid size on gamma comparison results varies by case.

II.B.4. Induced errors

The sensitivity of different gamma criteria was studied for three induced error types: (1) simple MU errors, which are analogous to a dose calibration error, (2) pseudo-random MLC errors which simulate potential errors or inaccuracies in MLC trajectories or control point definitions and (3) changes to the focal spot size in the Eclipse beam configuration, a parameter which affects penumbra width and, to a lesser extent, the output of the beam. For each error type, many plan calculations were created, each with a different error magnitude. MU errors were induced by uniformly scaling the plan MUs by a singular value ranging from -15% to +15% for 19 different error magnitudes. That is, for each case, 19 different MU error-induced TPS calculations were calculated and compared to the measurement.

MLC errors were induced by exporting the original MLC files and using an in-house MATLAB script to apply perturbations to each active leaf for every control point in the beam. Perturbations were chosen in a pseudo-random fashion from a square distribution bounded by varying user-defined ranges, therefore, the majority of leaves experience perturbations smaller than the maximum extent of the user-defined range. The user-defined ranges were both in the positive and negative direction such that when the user specifies a range of 1.0 cm, the square distribution

is sampled from -1.0 cm to +1.0 cm. Eight user-defined MLC error ranges were utilized for this study, ranging from zero, or no induced error, to ±3.0 cm.

Penumbral errors were induced by changing the focal spot size setting in the Eclipse beam configuration, which also causes small changes in the beam output. Penumbra widths (between 20% and 80%) were calculated using different focal spot size settings for a 10x10 cm^2 field at SSD = 90 cm and d = 10 cm and were compared to a diode measurement in water with the same SSD and depth. The focal spot size that produced a calculated penumbra width closest to that measured in water by a diode was labeled as the calculation without induced errors, or error magnitude of zero. Conveniently, the focal spot setting that gave the closest agreement to the diode measurement was focal = 0 mm. Thus, calculations with focal spot size = 0 mm were labeled as the calculations without an induced error for this study. Penumbra errors were induced for nine different focal spot size settings, with a maximum increase in penumbra width of 13.4 mm for the largest setting. Penumbra error magnitudes are denoted as the mismatch in penumbra, or the difference in penumbra width relative to the calculation without induced errors (i.e. calculation with focal = 0 mm). Table 1 in the Appendix shows focal spot sizes and their corresponding calculated penumbra widths and penumbral mismatch from the focal = 0 mm calculation.

II.B.5. Gamma comparisons

All gamma comparisons were performed in SNC Patient (version 6.2.3) between composite ArcCHECK® measurements and TPS calculations with and without induced errors. For this study, absolute dose, and global (Van Dyk) percent dose difference normalization settings were utilized. Neilson *et al*. have also suggested using a global normalization of the data for ArcCHECK® measurements in order to reduce the number of failing low dose points that are likely to have large uncertainties[34]. In another study using planar devices, it was noted that while global

normalization of the data can mask some errors, local normalization has the ability to magnify errors at low doses[12]. The question of normalization settings remains a subject of debate.

Based on the ArcCHECK® device geometry, comparisons were performed in 3D. Since calculation shifts cause the comparison in SNC Patient v.6.2.3 to revert to 2D mode, no shifts, automatic or manual, were used in the comparisons. Uncertainty corrections were turned off, as this is a flat correction added on top of the gamma comparison that reduces the sensitivity of the test by increasing passing rates. Use of the uncertainty correction for comparisons made with global dose normalization has been noted to inflate gamma passing rates anywhere from 0.5% to 5%[22].

II.B.6. Error curves

In this study we present a method for determining the sensitivity of many different gamma criteria to induced errors. Error-induced calculations are compared to measurements without induced errors for 36 different gamma criteria – 1%/1mm, 1%/2mm, 1%/3mm, 2%/1mm, 2%/2mm, 2%/3mm, 3%/1mm, 3%/2mm, 3%/3mm, 4%/4mm, 5%/2mm, and 5%/3mm, each at dose threshold values of 10%, 20%, and 50%. The gamma passing rate for each criterion is recorded as a function of error magnitude and plotted to create error curves. The details of the error curve workflow are illustrated in Figure 2-4.

Each error curve is case specific and is also unique to each class of induced error. For a given error type, calculations with induced errors of different magnitudes are compared to the QA measurement made on the ArcCHECK® without induced errors. Gamma passing rates are then recorded for these comparisons and are plotted against the induced error magnitude. This is repeated for each gamma criterion under investigation and for each case. In this study we investigated the sensitivities of 36 different gamma criteria – 1%/1mm, 1%/2mm, 1%/3mm, 2%/1mm, 2%/2mm, 2%/3mm, 3%/1mm, 3%/2mm, 3%/3mm, 4%/4mm, 5%/2mm, and 5%/3mm,

each at dose threshold values of 10%, 20%, and 50%. The workflow for creating error curves is illustrated in Figure 2-4 and an example error curve is shown in Figure 2-5.

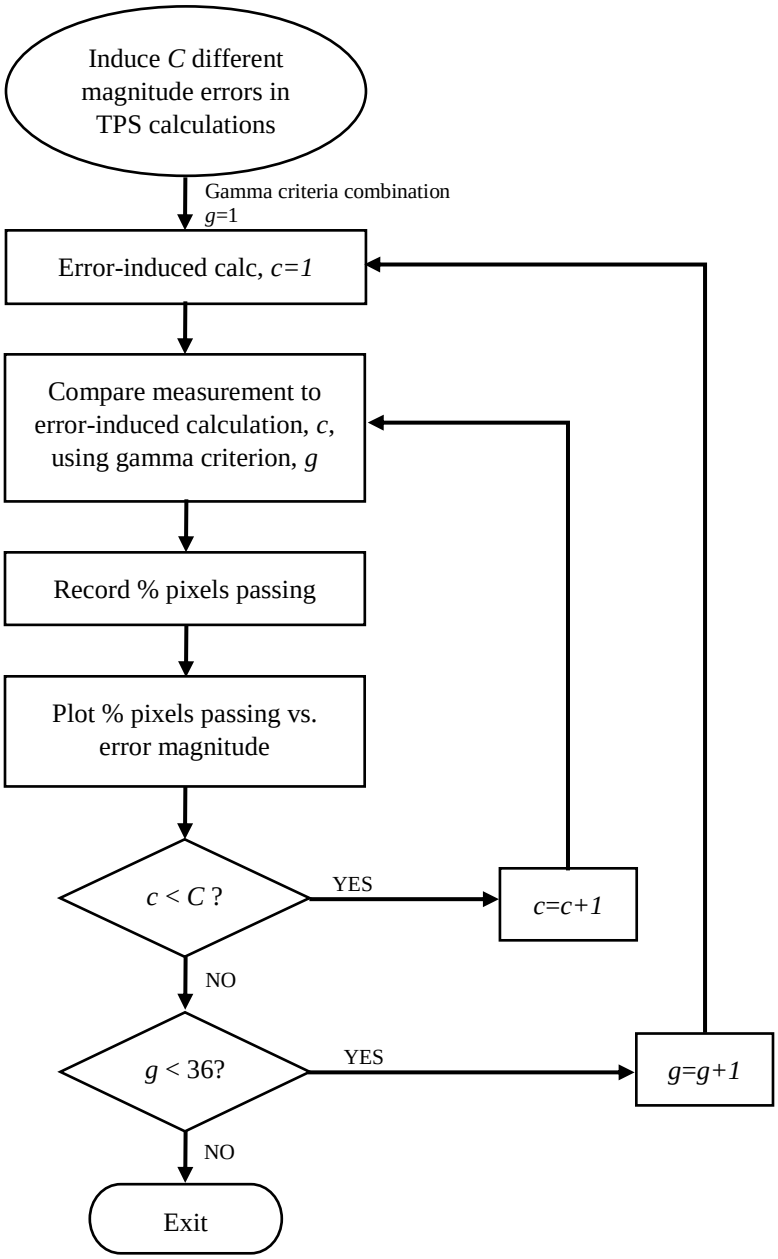

Figure 2-4. Error curve workflow for one case, using g=36 combinations of gamma criteria and an unspecified number, C, of different error-induced calculations of varying magnitude.

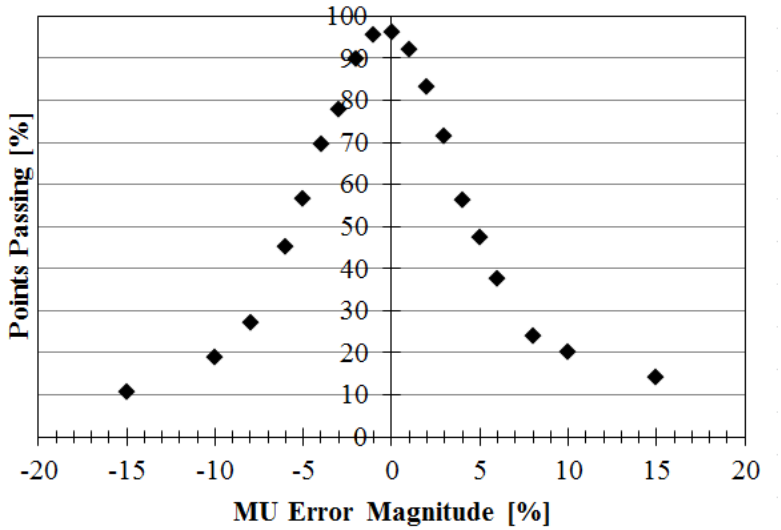

Figure 2-5. An example error curve for gamma criterion 2%/3mm, TH=50% is shown for induced MU errors. Each error curve is specific to the case, error type, and the gamma criterion chosen.

After obtaining each error curve, a sum of 2 to 3 Gaussians is automatically fit to the curve using a script that makes use of the MATLAB curve fitting toolbox, as shown in Figure 2-6.a. The width of each fit error curve is calculated at four different percent pixels passing (PP) values – 80%, 85%, 90%, and 95% (the widths at these PP cut-offs are labeled $W_{80\%}$, $W_{85\%}$, $W_{90\%}$, and $W_{95\%}$, as shown in Figure 2-6.b). These *W* values, or error curve widths, represent the sensitivity of the gamma comparison at a specific gamma criterion for a designated percent pixels passing cut-off. The error curve technique described here is labor-intensive: this study of 21 cases and three error types makes use of over 20,000 gamma comparisons and over 4,000 error curve widths. While some of the analysis was automated, this method would greatly benefit from full automation which would allow the study of more cases in a timely manner.

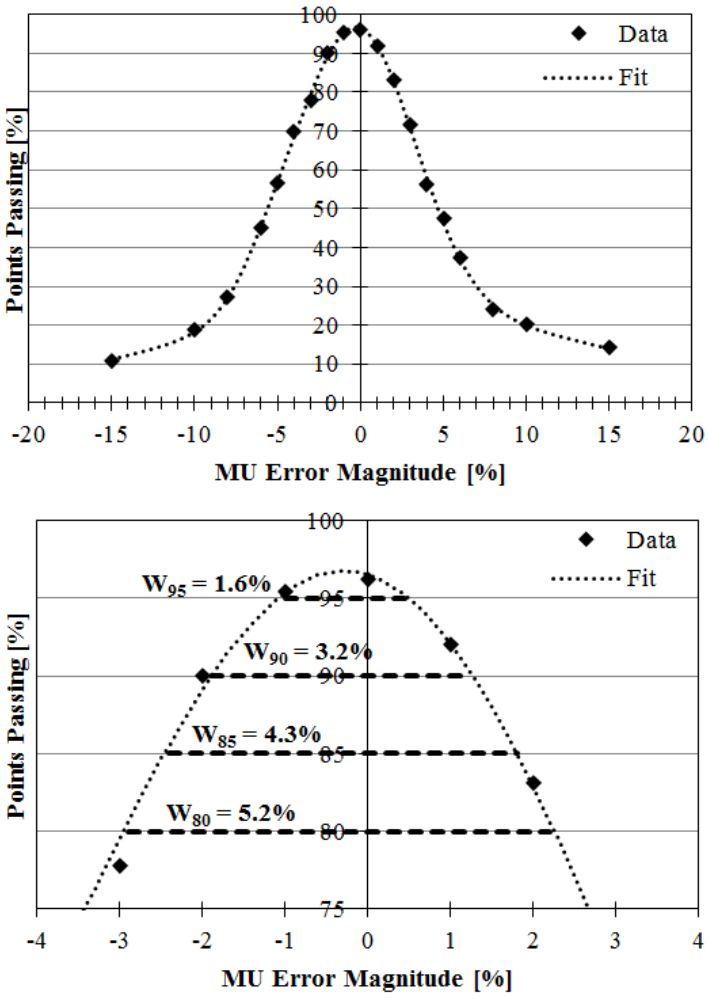

Figure 2-6. (a) Error curve from Figure 2-5 with a sum of Gaussians fit to the data. (b) Zoomed in from (a), the fit line is used to calculate curve widths at 80%, 85%, 90%, and 95% pixels passing.

II.C. Results

II.C.1. Detecting systematic errors with the error curve method

Early in this study we noticed that all MU error curves were initially shifted from zero. An investigation revealed that the radial symmetry on the linear accelerator used to make the measurements needed re-tuning and also that the ArcCHECK® required recalibration. After re-tuning the machine and recalibrating the ArcCHECK® all error curves shifted closer to zero (although select cases were still shifted). An example of this is illustrated in Figure 2-7, which shows an error curve created using a measurement with the correct ArcCHECK® absolute dose

calibration (solid) and an error curve created by applying an outdated dose calibration to the same measurement (dotted).

This example shows that shifted error curves may alert a physicist to the presence of a systematic error in the QA program. Although these shifts do not directly identify the source of the error, it is conceivable that shifts in different directions may be indicative of different types of systematic errors. Since the width of the error curves appear to be stable regardless of shifts, cases with small shifts from zero were not eliminated from this study.

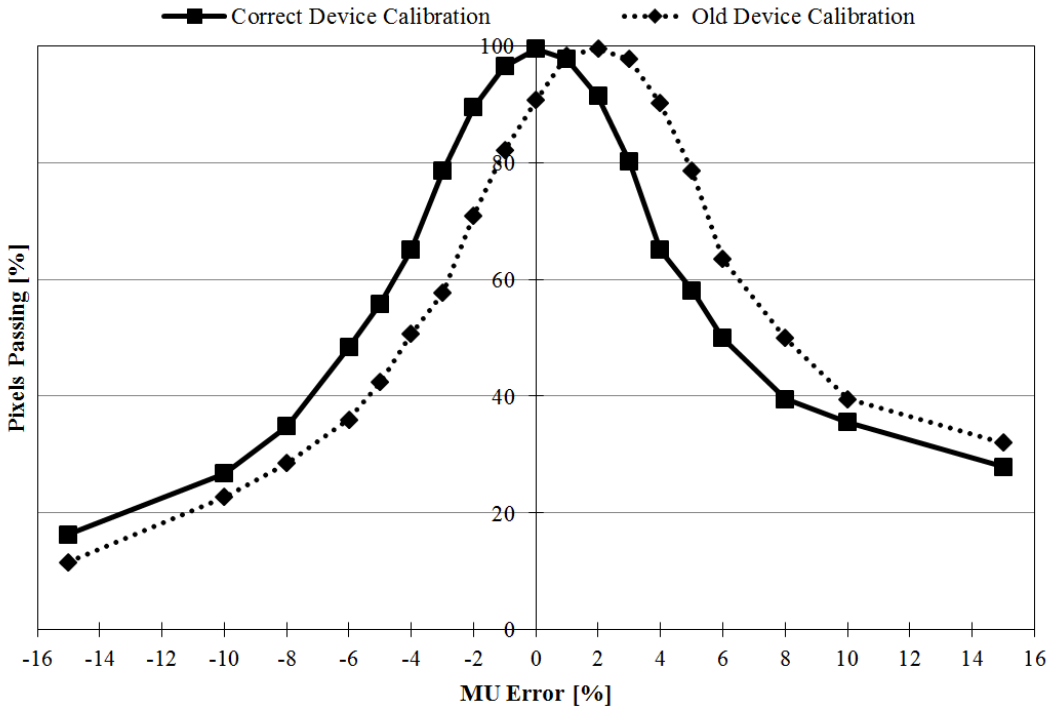

Figure 2-7. Shifting of error curves can be caused by systematic errors. An outdated absolute dose calibration was applied to the measurement and was used to create an error curve (dotted), which is clearly shifted from center. The measurement with the correct absolute dose calibration file applied (solid) corrects the shift from zero.

II.C.2. Interpreting error range plots

Error curve widths, or error ranges, represent the magnitude of an error that could go undetected in the QA process, given a specific gamma criterion and cut-off value. As a note, an error curve range of zero signifies that gamma passing rates were not high enough to compute an error curve width for a given PP cut-off value. As error ranges for $W_{80\%}$ and $W_{85\%}$ may be of less interest to

the general public, and only serve to increase error ranges compared to higher PP cut-offs, these data are available in Appendix A.

Since MU errors were induced in both positive and negative directions and not all curves are centered perfectly around zero, error curve ranges for MU errors are reported as the full width of the curve. However, if a curve is centered about zero, an error curve range of 5% means that errors of ±2.5% could go undetected for that specific case using the specified gamma criterion. Since MLC-induced errors are random leaf perturbations, an error curve range of 0.5 cm represents random errors undetected up to ±0.5 cm.

II.C.3. MU error

MU error ranges are shown in Figure 2-8 and Figure 2-9 for cut-off values of 90% and 95% for the 36 gamma criteria studied. From these graphs, it is clear that large dose errors of up to approximately 15% could potentially go undetected using a gamma criterion of 3%/3mm, TH = 10%, 90PP. The use of broader gamma criteria results in even larger error ranges, in some cases exceeding 20% in MU. It also appears that criteria with a DTA setting = 1 mm may be too strict for clinical use, as the error ranges for these criteria are at, or close to, zero. As expected, lowering the PP cut-off only serves to increase error curve widths and thus is not shown here. Data for $W_{80\%}$ *and* $W_{85\%}$ are available in Appendix A.

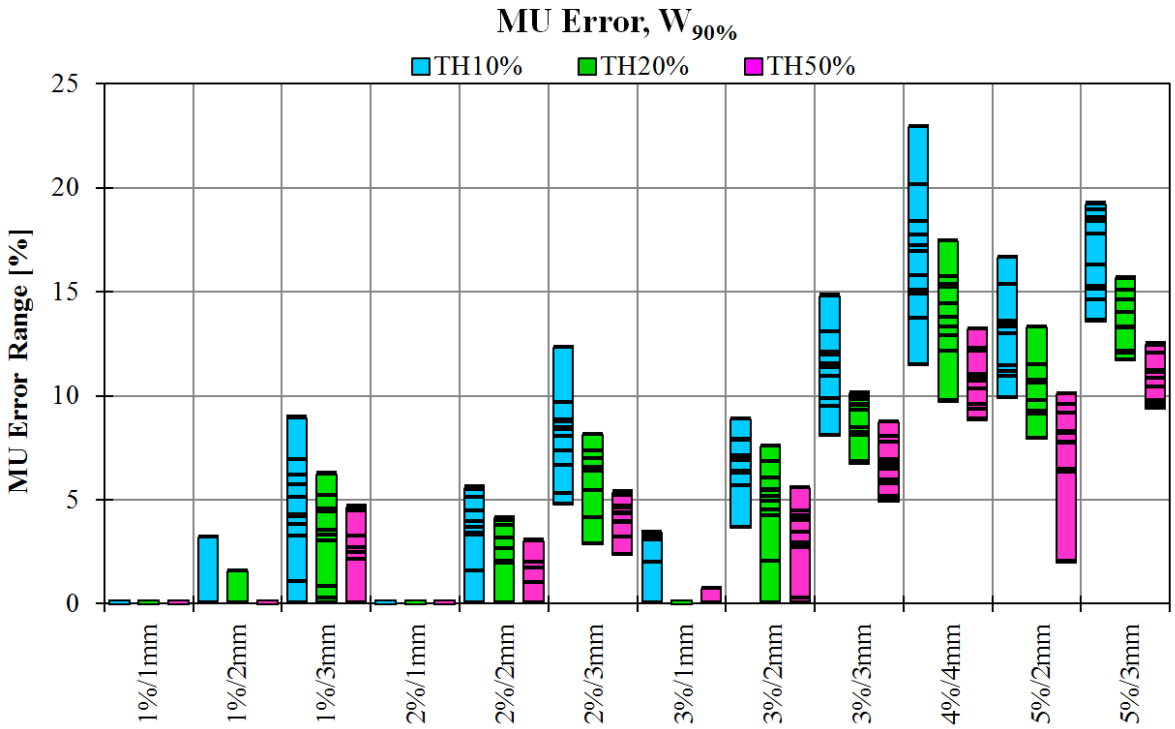

Figure 2-8. MU error ranges at 90PP. Each line represents the error range for one case. Bars are shown to aid in visualizing the error ranges across the studied cases. MU error range is in both directions, thus, an MU error range of 5% is an undetected error of ±2.5% MU if the error curve is centered around zero.

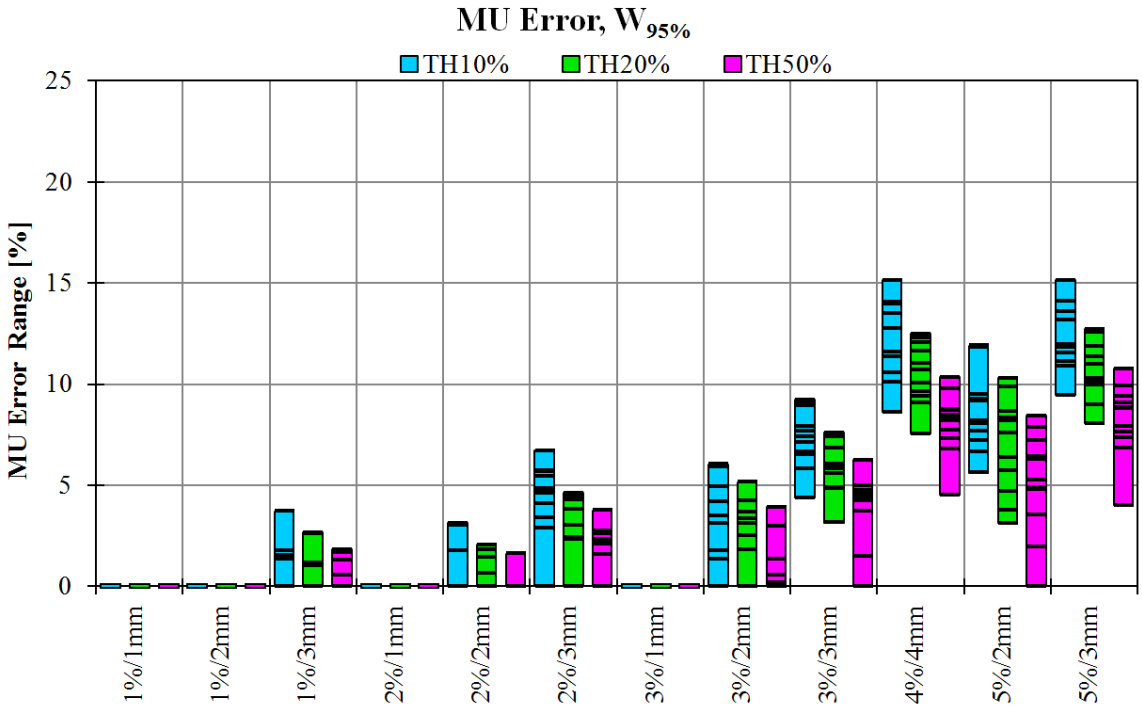

Figure 2-9. MU error ranges at 95PP. Each line represents the error range for one case. Bars are shown to aid in visualizing the error ranges across the studied cases. MU error range is in both directions, thus, an MU error range of 5% is an undetected error of ±2.5% MU if the error curve is centered around zero.

II.C.4. MLC error

MLC error ranges are shown in Figure 2-10 and Figure 2-11 for cut-off values of 90% and 95% for the 36 gamma criteria studied. Similar to the MU error, criteria with DTA = 1 mm generally have error ranges at, or close to, zero indicating that perhaps these criteria may not be appropriate for clinical use with the ArcCHECK®. Additionally, criteria of 3%/3mm and above show large error curve ranges, with the largest exceeding ±2.5 cm for 4%/4mm, TH=10%, 90PP.

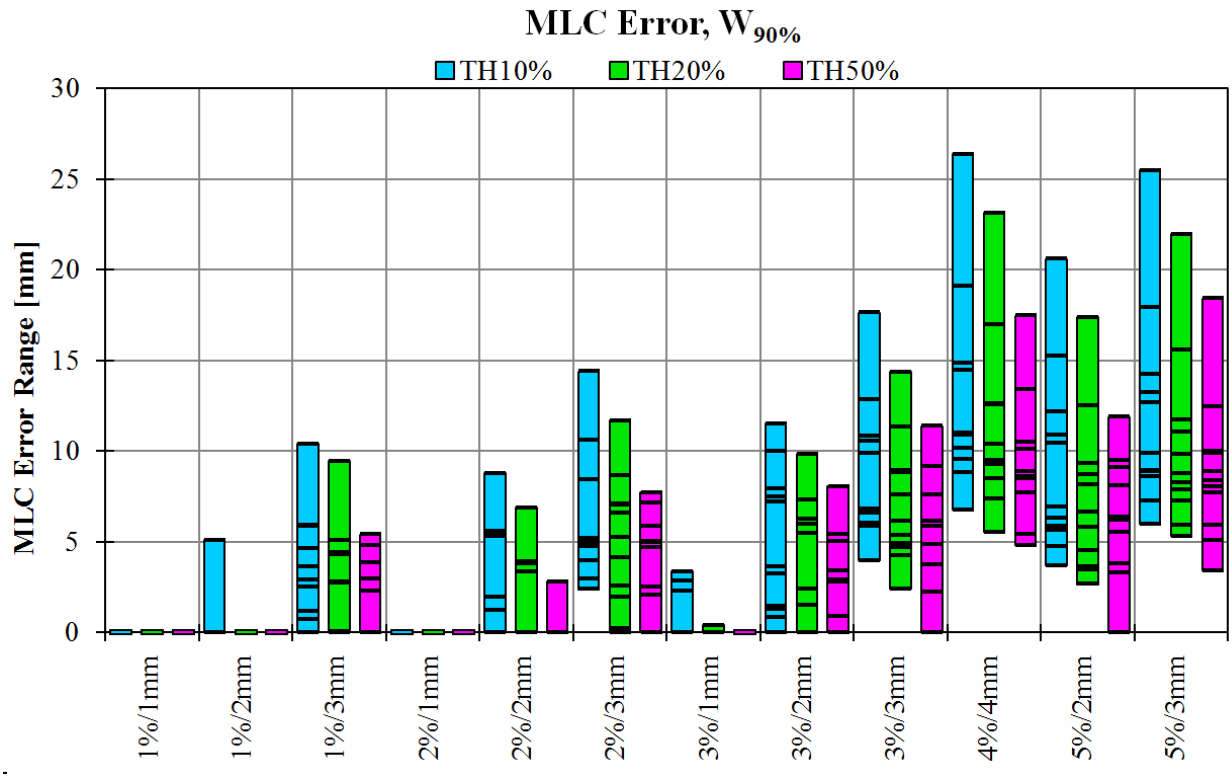

Figure 2-10. MLC error ranges at 90PP. Each line represents the error range for one case. Bars are shown to aid in visualizing the error ranges across the studied cases.

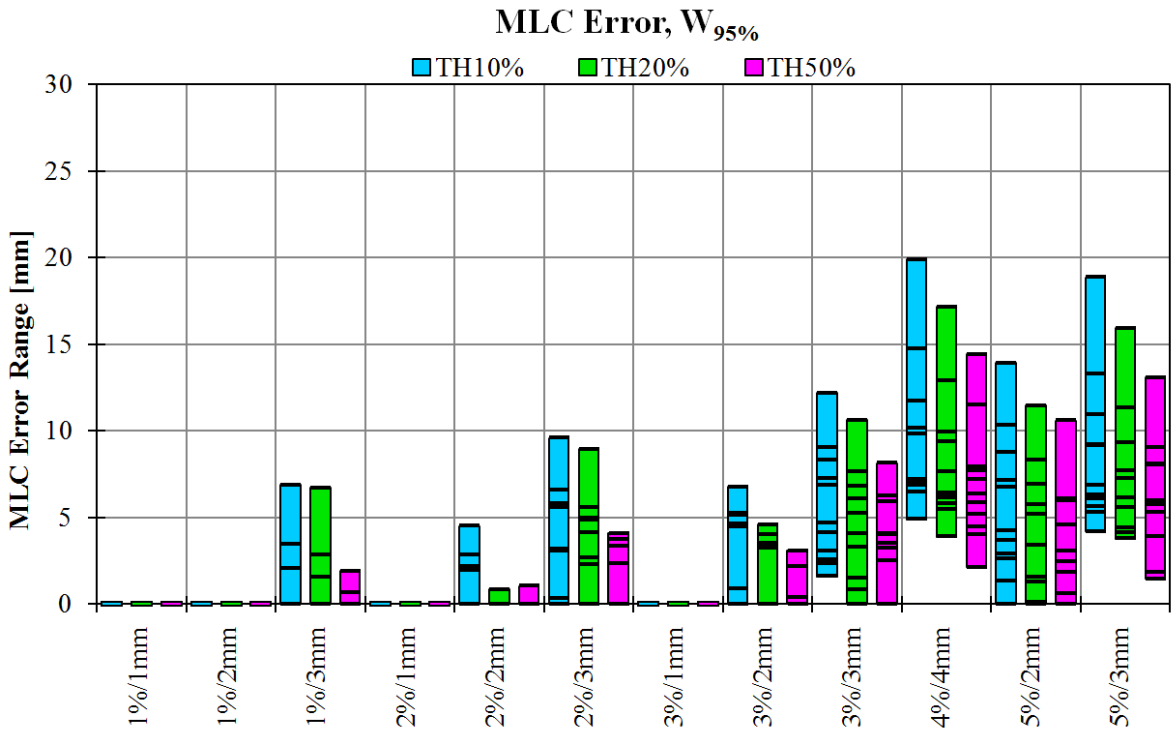

Figure 2-11. MLC error ranges at 95PP. Each line represents the error range for one case. Bars are shown to aid in visualizing the error ranges across the studied cases.

While these errors can be large there are several potential reasons that such large errors could go undetected, such as the fact that in composite analyses the MLC errors from different beams could compensate for one another. Also, this error type is sampled from a uniform distribution meaning that not all leaves will experience large perturbations relative to their original positions. Additionally, each case has a unique sensitivity which is likely affected by the initial location of the MLCs, the jaws, the complexity of the fluence distribution, and the size of the PTV. To illustrate the case-specific responses to the same MLC error range, Figure 2-12.a-c shows PTV DVH responses to a ±1.0 cm MLC error for three different cases. Figure 2-12.a. and Figure 2-12.b. highlight cases that are less sensitive to a ±1.0 cm MLC error, but exhibit different responses in the resulting PTV dose distribution (i.e. underdosing vs. overdosing). Figure 2-12.c. shows a case that is simply more sensitive to a large MLC error, however, this case also had a smaller range of errors that could pass QA than the cases shown in Figure 2-12.a. and Figure 2-12.b.

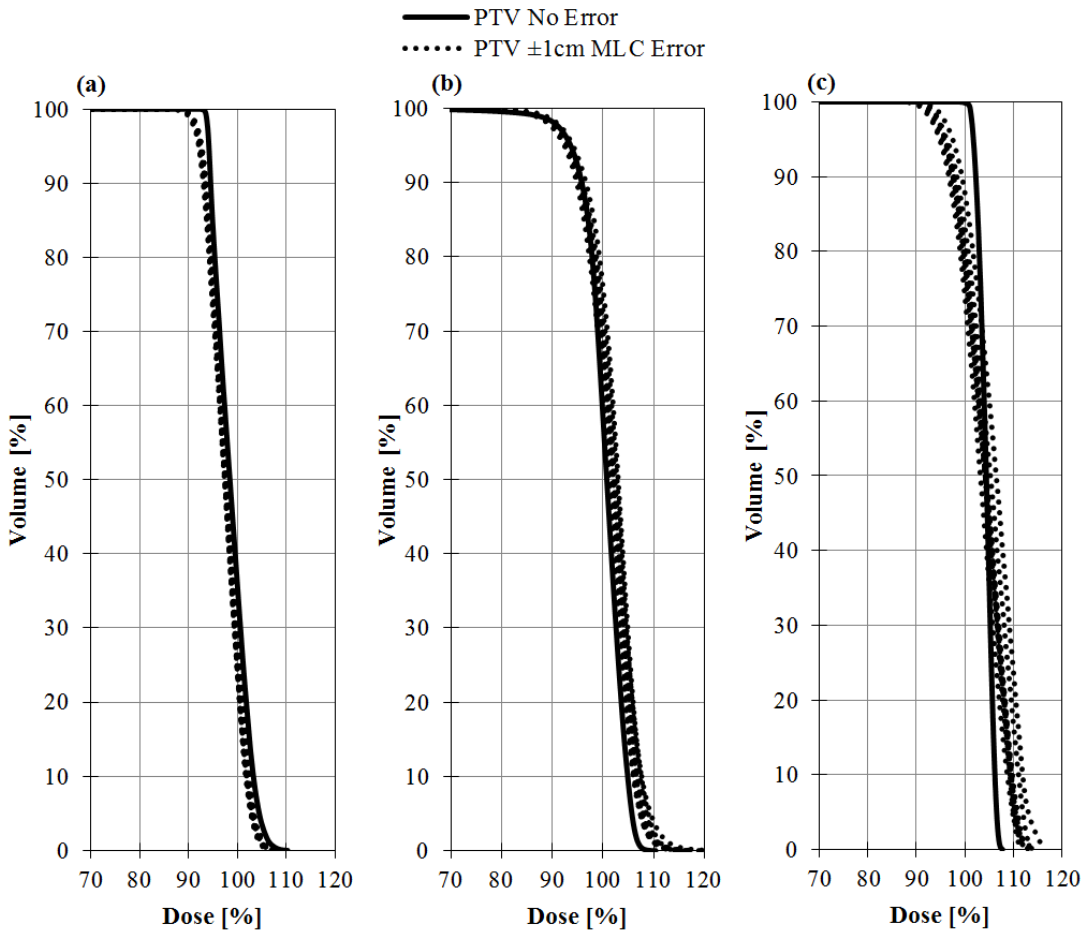

Figure 2-12. PTV DVHs are shown for three cases (a-c) that exhibit different responses to a ±1 cm MLC error. Ten different trials for the MLC error were performed for each case due to the random component of the induced error.

II.C.5. Penumbra error

The nature of the penumbra error in this study simulates an error in the initial fitting of the beam during configuration, or reconfiguration, of a dose calculation algorithm in the TPS. For a scenario in which the dose calculation algorithm has a large mismatch in the calculated penumbra relative to the commissioning measurement, one might hope that an end-to-end IMRT QA test would fail most of these plans. However, we found that if all 21 cases had a 5 mm penumbra mismatch in the plan calculation, 100% of the cases would pass at 3%/3mm, TH=10%, 90PP. Even at 2%/2mm, TH=10%, 90PP, more than half of the cases would pass with a 3 mm mismatch in penumbra. As this error is fundamentally different from the previous two errors, further discussion of the penumbra error can be found in the Appendix B.

II.C.6. Gamma sensitivity as a function of dose threshold

While the most commonly used threshold setting is 10%, this study investigated three threshold values – 10%, 20%, and 50%. Figures 7-10 show that as the dose threshold is increased, the error ranges for MU and MLC errors decrease dramatically for all combinations of gamma criteria studied. The relationship between dose threshold and gamma sensitivity is also consistent across all four *W* values (data for $W_{80\%}$ and $W_{85\%}$ can be found in Appendix A).

II.C.7. Validation cases

In order to determine if the results from the initial 11 cases were representative of the range of errors that could be passed in IMRT QA for a clinic treating a variety of sites, the error ranges for ten additional cases are shown in Figure 2-13 and Figure 2-14 for TH=50%. While differences in the TH=50% results are visible between error ranges for the initial cases (green) and the validation cases (purple), these differences are mostly small and not unexpected given the fact that each dataset consists of only 10-11 cases and were chosen such that a wide range of plan complexity and treatment sites were studied. It is conceivable that if more plans were chosen and matched to have similar characteristics, a smaller variation in error ranges may be observed.

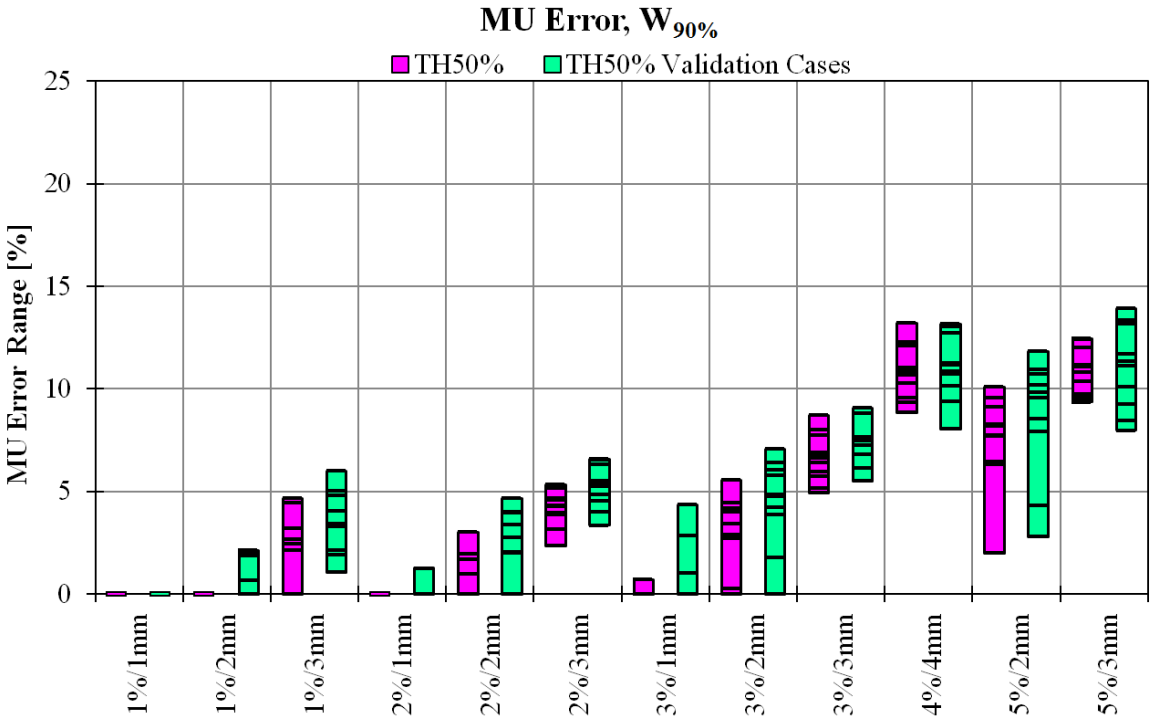

Figure 2-13. MU error ranges at 90PP for initial 11 cases and 10 validation cases show that the error ranges between the two datasets are quite similar. Each line represents the error range for one case and bars are shown to aid in visualizing the error ranges across the studied cases.

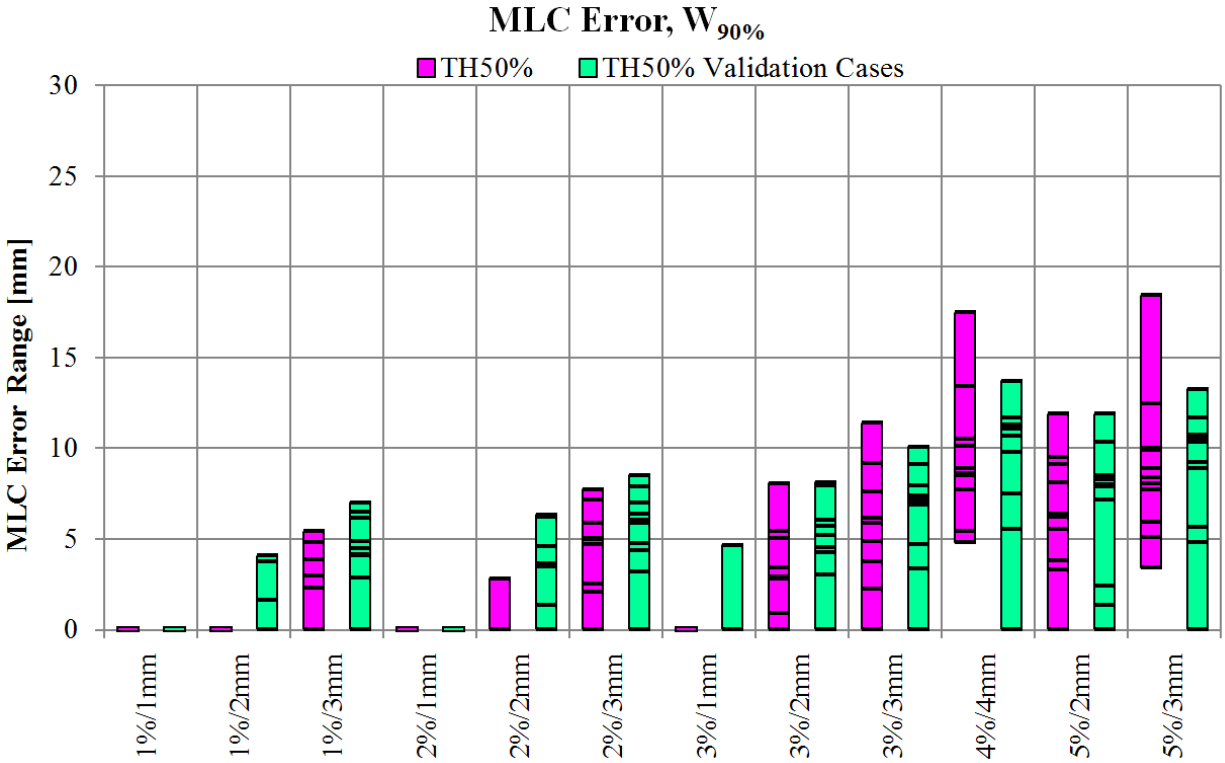

Figure 2-14. MLC error ranges at 90PP for initial 11 cases and 10 validation cases show that the error ranges between the two datasets are quite similar. Each line represents the error range for one case and bars are shown to aid in visualizing the error ranges across the studied cases.

II.C.8. Case Study: A 10% dose error in clinical IMRT QA

A case recently planned in our clinic serves as an example of a 10% dose error that could have gone undetected in the QA process with commonly used gamma criteria. A five fraction fixed-beam SBRT lung case was planned and a verification calculation on the ArcCHECK® phantom geometry was exported for comparison with ArcCHECK® QA measurements. After the export of the verification plan, the physician requested the plan to be renormalized down by 10% in dose and while the correct plan was exported and delivered to the ArcCHECK®, it was compared to the wrong calculation that was 10% higher in dose than the delivered plan. This resulted in a gamma passing rate of 89.3% (3%/3mm, TH=10%, 3D absolute dose, global normalization, uncertainty corrections off; QA comparison shown in Figure 2-15). Though not our clinical practice, TG-119 recommends the use of uncertainty corrections for diode arrays, which would have inflated the passing rate for this case to 91.7%.

For clinics following the TG-119 guidelines that 88%-90% pixels should pass at 3%/3mm TH=10% with uncertainty corrections turned on for diode array composite QA analyses, this case could have passed QA. However, such a low passing rate is typically investigated in our clinic and once corrected, the QA comparison resulted in a passing rate of 99.1% (3%/3mm, TH=10%, 3D absolute dose, global normalization, uncertainty corrections off). Even though the passing of this plan in QA would not have resulted in a misadministration, as the correctly normalized plan was exported to the machine for treatment, this case study highlights the fact that the MU errors presented in the previous section of this paper are not simply of academic interest. It is important for a QA program to be capable of identifying large errors in patient plans in order to prevent potential misadministrations, and this case presents a real-world example of how such dose errors could go undetected. Clearly, there is a pressing need for more sensitive gamma criteria and tighter acceptance tolerances for IMRT QA.

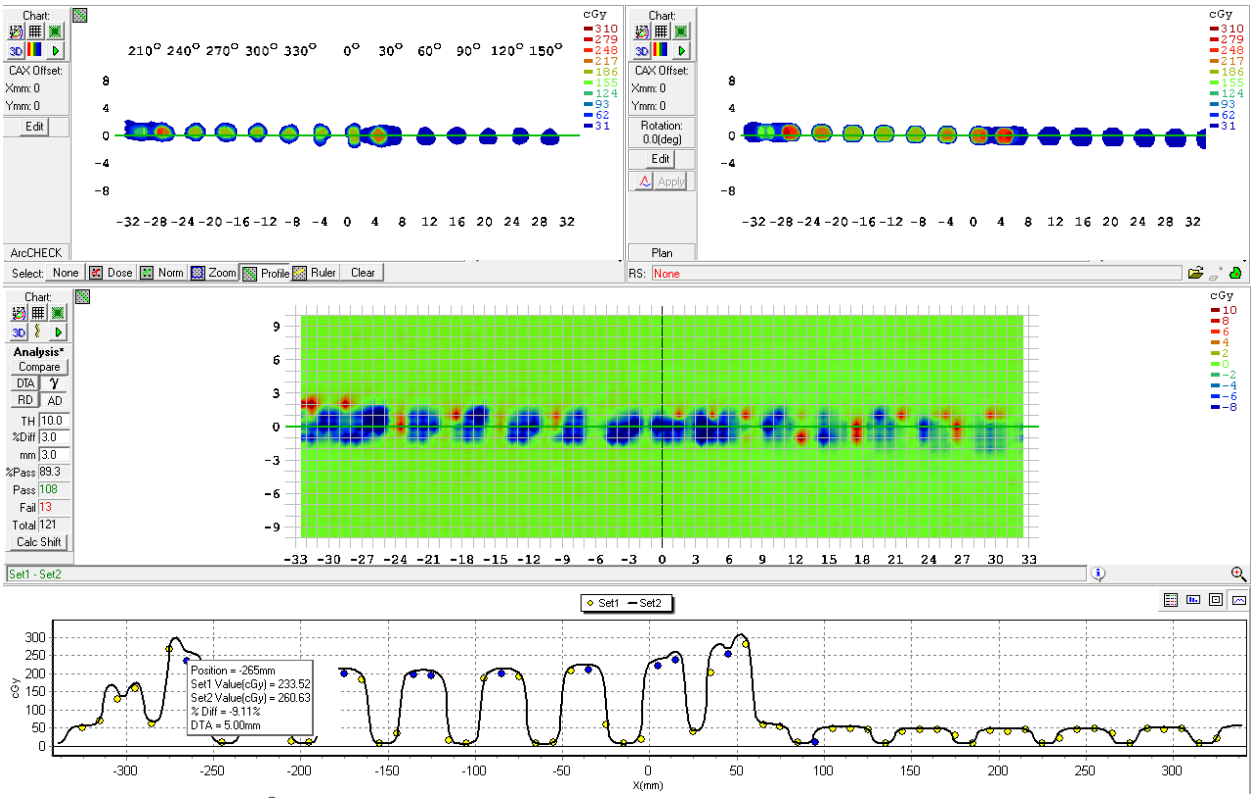

Figure 2-15. ArcCHECK® comparison between measurement and incorrect verification calculation (10% difference in normalization) shows a relatively high passing rate using 3%/3mm, TH=10% (with uncertainty corrections turned off).

II.D. Discussion

The results from this study further support the growing body of evidence that suggests 3%/3mm TH=10% is too broad to detect clinically significant plan errors. In this study we find that a higher dose threshold of 50% paired with 2%/3mm or 3%/2mm at 90%-95% pixels passing may offer decent sensitivity to MLC and MU errors without being overly strict (such as is the case with 1%/1mm). However, for very small fields measured on the ArcCHECK® device, high dose thresholds may not be appropriate due to the device's relatively coarse measurement sampling. A separate study would be required to determine appropriate threshold settings for small field cases measured on the ArcCHECK® device. Additionally, it is unknown if the aforementioned relationship between error sensitivity and dose threshold will exist for comparisons made using local dose normalization or other measurement devices. Further study with more cases is warranted before implementing these criteria clinically.

Several other publications have also suggested the use of dose thresholds other than 10%, albeit somewhat indirectly. Nelms *et al*. hypothesized that decreasing the low dose threshold in conjunction with global normalization of the data may lead to an increase in the passing rate, causing decreased sensitivity of the test, although this was not shown[22]. However, another study reported that gamma passing rates were independent of low dose threshold, but this study only compared two lower value thresholds, TH=10% and TH=5%[36]. We noted that two other studies made use of a higher dose threshold (TH=20%) in their comparisons, but the reasoning for this higher setting was not explained[29, 42]. A study by Wu *et al*. looked at the sensitivity of two values calculated in 3D, one called γ_{PTV}, which represents the gamma passing rate calculated for the points within the PTV volume, and γ_{10}, which represents the gamma passing rate calculated for the points enclosed within the 10% isodose surface. When looking at these two metrics (which are essentially two different thresholds) they found that γ_{PTV} passing rates decrease faster than γ_{10} passing rates with increasing error size[43]. While these studies hint that there is a benefit to using a higher dose threshold, we have explicitly shown that gamma sensitivity to induced errors increases dramatically with increasing threshold.

Another study induced MLC errors in SBRT VMAT cases for five gamma criteria using a method similar to that presented here, concluding that 2%/1mm was a suitable gamma criterion with 90% pixels passing for MapCHECK 2 and 80% pixels passing for EBT2 film[44]. However, the method presented here considers a wide range of gamma criteria combinations, including 12 gamma criteria each studied at three different dose threshold values for four PP cut-off values. While labor intensive, the error curve method allows a more thorough analysis of appropriate gamma criteria for clinical use.

The error ranges presented here will be distinct from those found in IMRT QA programs at different institutions which may use other combinations of measurement device, calculation algorithm, delivery technique, and gamma comparison settings than those used in this study.

Other measurement devices in particular are likely to have gamma sensitivities different from those reported here using the ArcCHECK® device. A thorough study across many devices, stratified by delivery technique and case type would be the most comprehensive way to evaluate gamma sensitivity using the error curve technique. Additionally, the effect of calculation algorithm and grid size must be considered for each TPS and algorithm implementation.

This study has several limitations. First, the small number of cases studied (n=21) offers insight into the sensitivity of many gamma criteria, however, conclusions on exact criteria to be used should involve more cases and perhaps be stratified by case type. While the initial 11 cases had results comparable to those obtained for the validation dataset, error ranges for some gamma criteria showed larger differences between the initial dataset and the validation dataset. However, this is not unexpected given that cases were chosen to include a variety of sites and plan complexity.

It should be noted that this method does not identify particular errors in a case but rather is a relative comparison between plans without intentional errors and plans with induced errors to understand the sensitivity of different gamma criteria. Additionally, this study uses a very specific combination of detector device, software, delivery technique, and gamma comparison settings. For example, if this study were to be performed using local dose normalization or with a planar measurement array, very different gamma sensitivities could be obtained. Results from this study should not be used to make clinical decisions on appropriate gamma criteria for other measurement devices or delivery techniques. The results presented here are also heavily dependent on the dose calculation algorithm (AAA), and the chosen grid size (1 mm).

II.E. Conclusions

In recent years, the appropriateness of commonly used criteria for the gamma comparison in IMRT QA has been brought into question. The error curve method presented here allows the

quantitative determination of gamma criteria sensitivities to induced TPS errors and is straightforward to apply to other combinations of delivery techniques and measurement devices. Use of the error curve method to quantitatively select gamma criteria for other measurement devices may allow the determination of more sensitive metrics for use across the field.

This study shows that large errors, such as 15% MU errors and ±1 cm random MLC errors, can potentially go undetected using a 3%/3mm, threshold=10% at 90% pixels passing for static beam IMRT plans measured on the ArcCHECK® device. This work supports the growing body of evidence on the inability of commonly used gamma criteria to detect large errors in patient-specific QA. The current results also clearly show a gamma %Diff setting of 3% does not necessarily translate to identifying dose errors of 3% or more. While commonly used criteria are clearly too broad to detect clinically significant errors, we have shown that slightly stricter criteria with a higher dose threshold (such as 2%/3mm or 3%/2mm with TH=50%) may offer desirable sensitivity to errors without further increasing the time-intensive nature of IMRT QA.

This work also explicitly shows that gamma sensitivity to errors can be dramatically increased with the use of higher low dose thresholds. However, for small field plans (e.g., typical SBRT or SRS fields), using an increased dose threshold is potentially problematic and should be investigated thoroughly before implementation due to the effects of the limited number of dose points in small field plans evaluated with higher thresholds. Additionally, the error curve method may also serve as a tool to identify systematic errors in the clinic when error curves are consistently shifted from zero. Future work will aim to analyze relevant DVH metrics in order to assess the clinical impact of potentially missed errors in IMRT QA and also to determine if correlations exist between error curve ranges and relevant patient dose metrics.

A.II. Appendix A for Chapter II

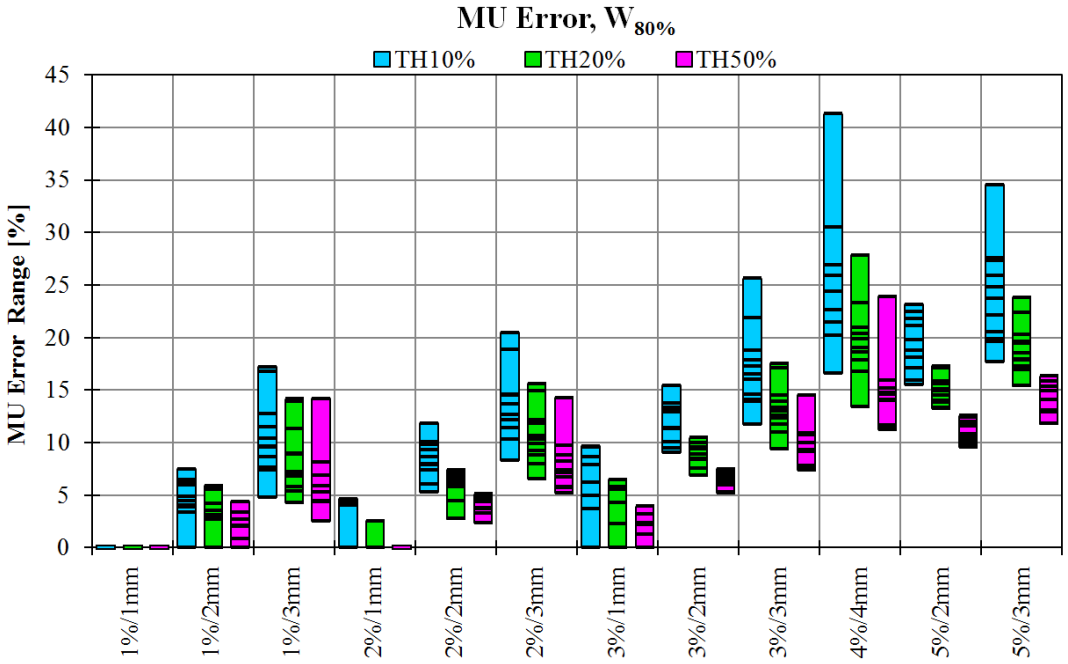

Figure A.2-1. MU error ranges at 80PP. Each horizontal line represents the error range for one case. Color bars are shown to aid in visualizing the error ranges across the studied cases. MU error ranges are in both directions, thus, a MU error range of 5% is an undetected error of ±2.5% MU if the error curve is centered around zero.

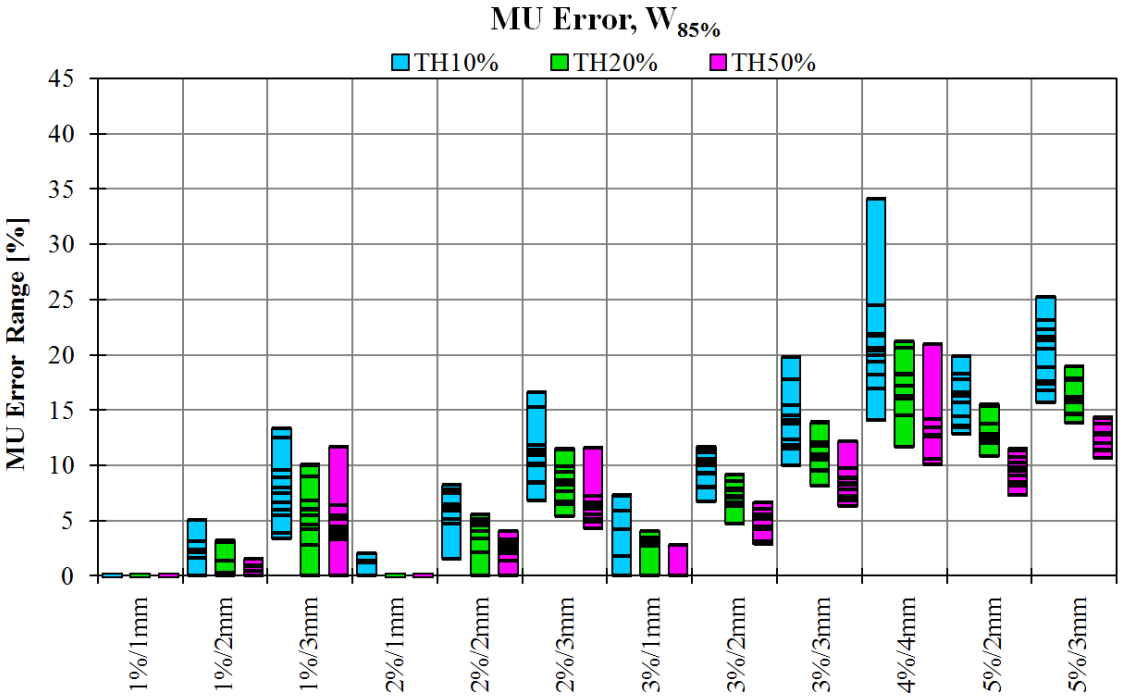

Figure A.2-2. MU error ranges at 85PP. Each horizontal line represents the error range for one case. Color bars are shown to aid in visualizing the error ranges across the studied cases. MU error ranges are in both directions, thus, a MU error range of 5% is an undetected error of ±2.5% MU if the error curve is centered around zero.

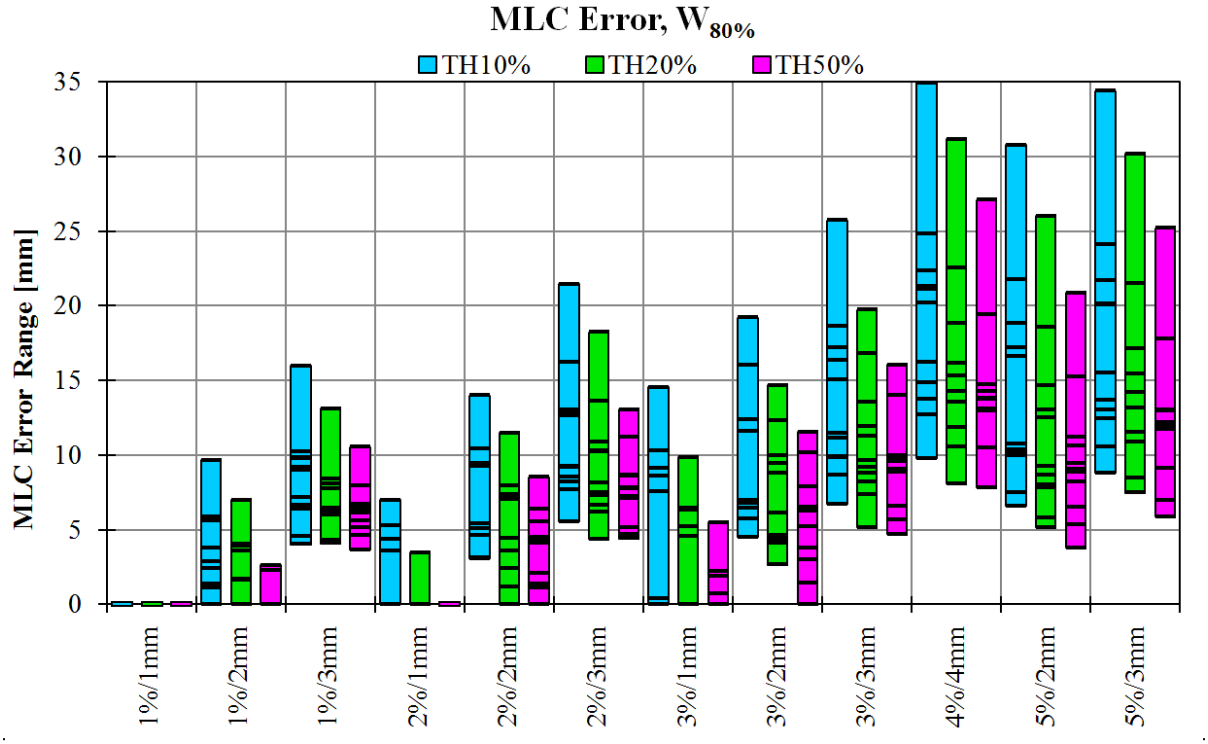

Figure A.2-3. MLC error ranges at 80PP. Each horizontal line represents the error range for one case. Color bars are shown to aid in visualizing the error ranges across the studied cases.

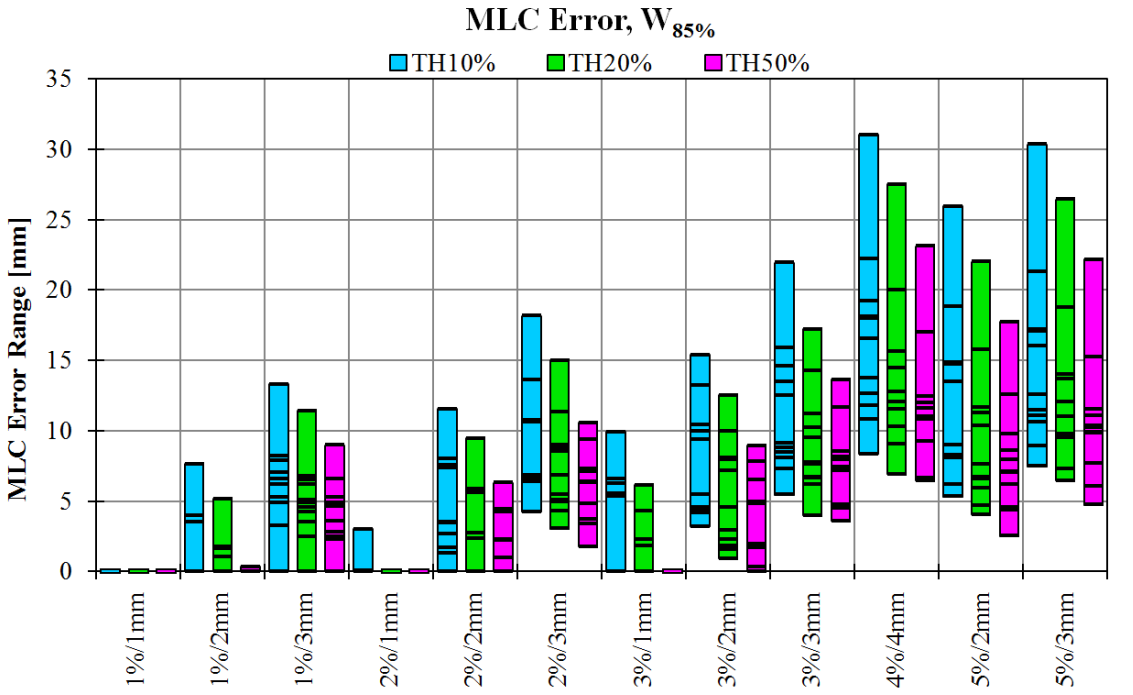

Figure A.2-4. MLC error ranges at 85PP. Each horizontal line represents the error range for one case. Color bars are shown to aid in visualizing the error ranges across the studied cases.

B.II. Appendix B for Chapter II

Penumbra Error

Penumbral errors were induced by changing the focal spot size setting in the Eclipse beam configuration, which also causes small changes in the beam output. Table V.2-1. shows the penumbra widths for calculated profiles at different focal spot size settings, along with the penumbra width from a diode measurement made in water. Since the focal spot size = 0 mm calculation yields a penumbra width closest to that measured with a diode, this calculation is labeled as the calculation without induced errors. In this study we report error ranges as the mismatch in penumbra width between the calculations with induced penumbra broadening and the calculation without an induced error (focal spot size = 0 mmm). The penumbra mismatch is also shown in Table B.2-1.

Table B.2-1. Penumbra widths calculated at different focal spot size settings and compared to the penumbra width measured with a diode in water for a 6 MV 10x10 cm^2 field at SSD = 90 cm and depth = 10 cm. Penumbra mismatch compared to the calculation without induced errors (focal=0 mm) is also shown.

Focal spot size setting [mm]	Calculated penumbra width [mm]	Diode measured penumbra [mm]	Penumbra mismatch from focal=0 calc [mm]
0	2.8	2.6	0
1	3.3		0.5
2	4.5		1.7
3	5.8		3
4	7.2		4.4
5	8.7		5.9
6	10.2		7.4
8	13.2		10.4
10	16.2		13.4

Penumbra errors in this study simulate a fundamental error in the initial configuration of the beam if the focal spot size setting was too large in Eclipse TPS. This could happen, for example, if the configured beam calculations are verified with beam data acquired with an inappropriate detector (i.e. too large of an active volume). The sensitivity of the gamma comparison to detect such errors

is relevant since many institutions use the gamma comparison for commissioning purposes, especially in end-to-end IMRT QA tests. Penumbra error results are presented here as the percentage of cases that would fail to identify different magnitudes of penumbra broadening, instead of the error range graphs presented for the other error types as this error is more relevant in a commissioning scenario and not daily IMRT QA. Figure B.2-1. shows the percentage of cases from our study that would fail to identify 3 mm, 5 mm, 7 mm, and 10 mm mismatches in penumbra (relative to the calculation without an induced error) for five different gamma criteria – three that are more commonly used and two that make use of a higher dose threshold. Even at the stricter criterion of 2%/2mm TH=10%, 90PP, a 3 mm mismatch, or broadening of the penumbra, would fail to be identified in over half of the 21 cases. For 3%/3mm TH=10% at 90PP and 95PP, over 90% of the cases with a 5 mm mismatch would fail to be identified. This illustrates the extreme insensitivity of this type of test to identify issues in the fundamental configuration of the beam in the TPS. While end-to-end tests can be a useful check near the end stages of commissioning a new dose calculation algorithm for use in the clinic, the results should be interpreted with caution and should only be used after a thorough validation with more basic beam data has been completed.

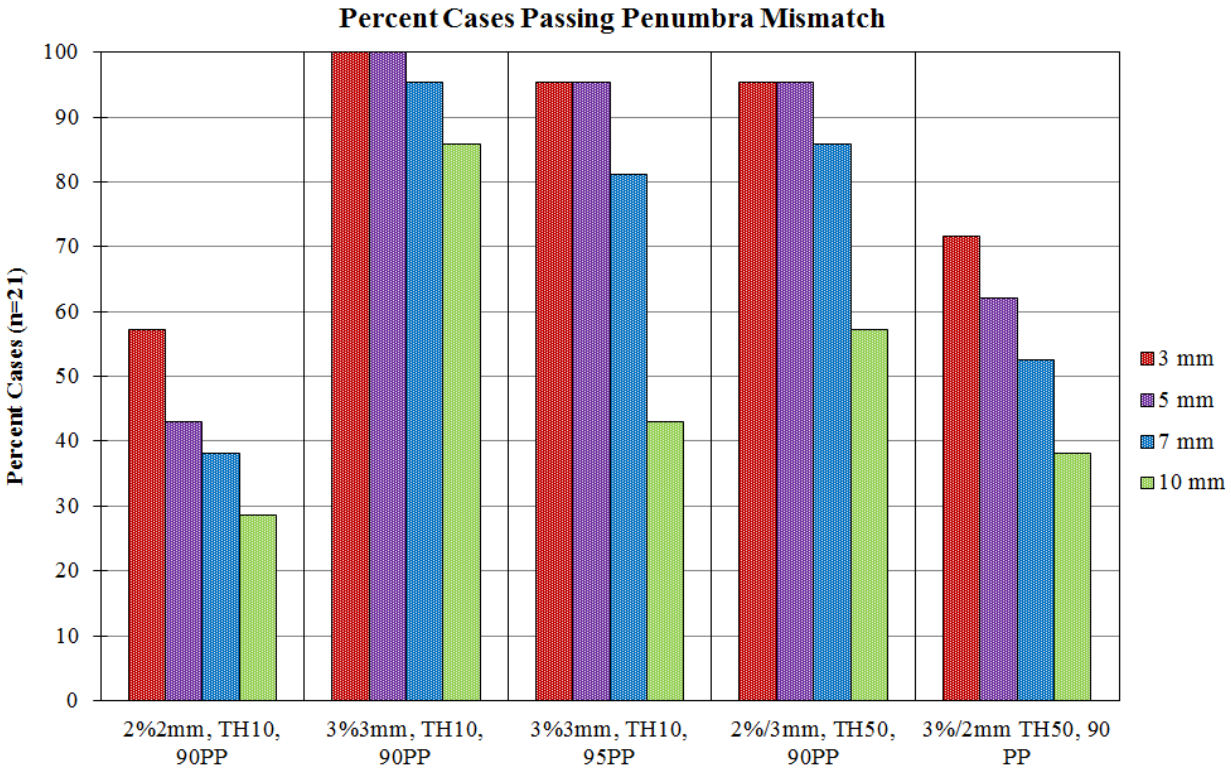

Figure B.2-1. The percent of cases that would pass a given penumbra mismatch in IMRT QA given the specified criteria.

References

[1] G.A. Ezzell, J.M. Galvin, D. Low, J.R. Palta, I. Rosen, M.B. Sharpe, P. Xia, Y. Xiao, L. Xing, X.Y. Cedric, "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**, 2089-2115 (2003).

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

[3] J.M. Moran, M. Dempsey, A. Eisbruch, B.A. Fraass, J.M. Galvin, G.S. Ibbott, L.B. Marks, "Safety considerations for IMRT: Executive summary," Practical Radiation Oncology **1**, 190-195 (2011).

[4] A.C. Hartford, J.M. Galvin, D.C. Beyer, T.J. Eichler, G.S. Ibbott, B. Kavanagh, C.J. Schultz, S.A. Rosenthal, "American College of Radiology (ACR) and American Society for Radiation Oncology (ASTRO) Practice Guideline for Intensity-modulated Radiation Therapy (IMRT)," American Journal of Clinical Oncology **35**, 612-617 (2012).

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

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

[7] G.A. Ezzell, J.W. Burmeister, N. Dogan, T.J. LoSasso, J.G. Mechalakos, D. Mihailidis, A. Molineu, J.R. Palta, C.R. Ramsey, B.J. Salter, "IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119," Medical Physics **36**, 5359-5373 (2009).

[8] S. Both, I.M. Alecu, A.R. Stan, M. Alecu, A. Ciura, J.M. Hansen, R. Alecu, "A study to establish reasonable action limits for patient-specific quality assurance in intensity-modulated radiation therapy," Journal of Applied Clinical Medical Physics **8**, 1-8 (2007).

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

[10] P.S. Basran, M.K. Woo, "An analysis of tolerance levels in IMRT quality assurance procedures," Medical Physics **35**, 2300-2307 (2008).

[11] A.J. Vinall, A.J. Williams, V.E. Currie, A. Van Esch, D. Huyskens, "Practical guidelines for routine intensity-modulated radiotherapy verification: pre-treatment verification with portal dosimetry and treatment verification with in vivo dosimetry," The British Journal of Radiology **83**, 949-957 (2010).

[12] P. Carrasco, N. Jornet, A. Latorre, T. Eudaldo, A. Ruiz, M. Ribas, "3D DVH-based metric analysis versus per-beam planar analysis in IMRT pretreatment verification," Medical Physics **39**, 5040-5049 (2012).

[13] F. Garcia-Vicente, V. Fernandez, R. Bermudez, A. Gomez, L. Perez, A. Zapatero, J.J. Torres, "Sensitivity of a helical diode array device to delivery errors in IMRT treatment and establishment of tolerance level for pretreatment QA," Journal of Applied Clinical Medical Physics **13** (2012).

[14] T. Pawlicki, S. Yoo, L.E. Court, S.K. McMillan, R.K. Rice, J.D. Russell, J.M. Pacyniak, M.K. Woo, P.S. Basran, A.L. Boyer, C. Bonilla, "Process control analysis of IMRT QA: implications for clinical trials," Physics in Medicine and Biology **53**, 5193-5205 (2008).

[15] X. Jin, H. Yan, C. Han, Y. Zhou, J. Yi, C. Xie, "Correlation between gamma index passing rate and clinical dosimetric difference for pre-treatment 2D and 3D volumetric modulated arc therapy dosimetric verification," The British Journal of Radiology **88** (2015).

[16] A.K. Templeton, J.C. Chu, J.V. Turian, "The sensitivity of ArcCHECK-based gamma analysis to manufactured errors in helical tomotherapy radiation delivery," Journal of Applied Clinical Medical Physics **16**, 4814 (2015).

[17] L. Vieillevigne, J. Molinier, T. Brun, R. Ferrand, "Gamma index comparison of three VMAT QA systems and evaluation of their sensitivity to delivery errors," Physica Medica (2015).

[18] G. Yan, C. Liu, T. Simon, L.-C. Peng, C. Fox, J. Li, "On the sensitivity of patient-specific IMRT QA to MLC positioning errors," Journal of Applied Clinical Medical Physics **10** (2009).

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

[20] H. Zhen, B.E. Nelms, W.A. Tomé, "Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA," Medical Physics **38**, 5477-5489 (2011).

[21] M. Stasi, S. Bresciani, A. Miranti, A. Maggio, V. Sapino, P. Gabriele, "Pretreatment patient-specific IMRT quality assurance: A correlation study between gamma index and patient clinical dose volume histogram," Medical Physics **39**, 7626-7634 (2012).

[22] B.E. Nelms, M.F. Chan, G. Jarry, M. Lemire, J. Lowden, C. Hampton, V. Feygelman, "Evaluating IMRT and VMAT dose accuracy: Practical examples of failure to detect systematic errors when applying a commonly used metric and action levels," Medical Physics **40**, 111722 (2013).

[23] B. Nelms, G. Jarry, M. Chan, C. Hampton, Y. Watanabe, V. Feygelman, "Real-world examples of sensitivity failures of the 3%/3mm pass rate metric and published action levels when used in IMRT/VMAT system commissioning," Journal of Physics: Conference Series **444**, 012086 (2013).

[24] M.F. Chan, J. Li, K. Schupak, C. Burman, "Using a novel dose QA tool to quantify the impact of systematic errors otherwise undetected by conventional QA methods: clinical head and neck case studies," Technology in Cancer Research & Treatment **13**, 57-67 (2014).

[25] S.F. Kry, A. Molineu, J.R. Kerns, A.M. Faught, J.Y. Huang, K.B. Pulliam, J. Tonigan, P. Alvarez, F. Stingo, D.S. Followill, "Institutional Patient-specific IMRT QA Does Not Predict Unacceptable Plan Delivery," International Journal of Radiation Oncology • Biology • Physics **90**, 1195-1201 (2014).

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

[27] E.M. McKenzie, P.A. Balter, F.C. Stingo, J. Jones, D.S. Followill, S.F. Kry, "Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans," Medical Physics **41** (2014).

[28] B.E. Nelms, J.A. Simon, "A survey on IMRT QA analysis," Journal of Applied Clinical Medical Physics **8** (2007).

[29] M. Hussein, P. Rowshanfarzad, M.A. Ebert, A. Nisbet, C.H. Clark, "A comparison of the gamma index analysis in various commercial IMRT/VMAT QA systems," Radiotherapy and Oncology **109**, 370-376 (2013).

[30] T. Sanghangthum, S. Suriyapee, S. Srisatit, T. Pawlicki, "Statistical process control analysis for patient-specific IMRT and VMAT QA," Journal of Radiation Research **54**, 546-552 (2013).

[31] N. Wen, B. Zhao, J. Kim, K. Chin-Snyder, M. Bellon, C. Glide-Hurst, K. Barton, D. Chen, I.J. Chetty, "IMRT and RapidArc commissioning of a TrueBeam linear accelerator using TG-119 protocol cases," Journal of Applied Clinical Medical Physics **15** (2014).

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

[33] T. Pawlicki, S. Yoo, L.E. Court, S.K. McMillan, R.K. Rice, J.D. Russell, J.M. Pacyniak, M.K. Woo, P.S. Basran, J. Shoales, A.L. Boyer, "Moving from IMRT QA measurements toward independent computer calculations using control charts," Radiotherapy and Oncology : Journal of the European Society for Therapeutic Radiology and Oncology **89**, 330-337 (2008).

[34] C. Neilson, M. Klein, R. Barnett, S. Yartsev, "Delivery quality assurance with ArcCHECK," Medical Dosimetry **38**, 77-80 (2013).

[35] K.B. Pulliam, D. Followill, L. Dong, M. Gillin, K. Prado, S.F. Kry, "A six-year review of more than 13,000 patient-specific IMRT QA results from 13 different treatment sites," Journal of Applied Clinical Medical Physics **15** (2014).

[36] S. Bresciani, A. Di Dia, A. Maggio, C. Cutaia, A. Miranti, E. Infusino, M. Stasi, "Tomotherapy treatment plan quality assurance: the impact of applied criteria on passing rate in gamma index method," Medical Physics **40**, 121711 (2013).

[37] W. Cruz, G. Narayanasamy, M. Regan, P. Mavroidis, N. Papanikolaou, C.S. Ha, S. Stathakis, "Patient specific IMRT quality assurance with film, ionization chamber and detector arrays: Our institutional experience," Radiation Physics and Chemistry **115**, 12-16 (2015).

[38] V. Feygelman, G. Zhang, C. Stevens, B.E. Nelms, "Evaluation of a new VMAT QA device, or the "X" and "O" array geometries," Journal of Applied Clinical Medical Physics **12**, 3346 (2011).

[39] D.W. Bailey, B.E. Nelms, K. Attwood, L. Kumaraswamy, M.B. Podgorsak, "Statistical variability and confidence intervals for planar dose QA pass rates," Medical Physics **38**, 6053-6064 (2011).

[40] D. Létourneau, J. Publicover, J. Kozelka, D.J. Moseley, D.A. Jaffray, "Novel dosimetric phantom for quality assurance of volumetric modulated arc therapy," Medical Physics **36**, 1813-1821 (2009).

[41] G. Yan, B. Lu, J. Kozelka, C. Liu, J.G. Li, "Calibration of a novel four-dimensional diode array," Medical Physics **37**, 108-115 (2010).

[42] A. Carver, M. Gilmore, S. Riley, J. Uzan, P. Mayles, "An analytical approach to acceptance criteria for quality assurance of intensity modulated radiotherapy," Radiotherapy and Oncology **100**, 453-455 (2011).

[43] C. Wu, K.E. Hosier, K.E. Beck, M.B. Radevic, J. Lehmann, H.H. Zhang, A. Kroner, S.C. Dutton, S.A. Rosenthal, J.K. Bareng, M.D. Logsdon, D.R. Asche, "On using 3D gamma-analysis for IMRT and VMAT pretreatment plan QA," Medical Physics **39**, 3051-3059 (2012).

[44] J.I. Kim, S.Y. Park, H.J. Kim, J.H. Kim, S.J. Ye, J.M. Park, "The sensitivity of gamma-index method to the positioning errors of high-definition MLC in patient-specific VMAT QA for SBRT," Radiation Oncology (London, England) **9**, 167 (2014).

Chapter III.

The impact of detector geometry, device spatial sampling, and delivery technique on gamma comparison sensitivity in the presence of induced errors

III.A. Motivation

Numerous gamma sensitivity studies have been previously performed that aim to quantify the differences in sensitivity between various detector devices in current clinical use for IMRT QA. Prior work has shown conflicting results with regard to the sensitivity of the various measurement devices used across the field. Some studies have shown that different detector devices show dissimilar sensitivities for IMRT QA measurements[1–5] , others suggest that different devices have similar sensitivities[6–8], while some still admit that sensitivity differences between devices in IMRT QA remain unclear[9]. Some of these results, however, are based on the simple fact that high gamma passing rates were achieved for all measurement scenarios.

Furthermore, it is of interest to understand if use of the same gamma criteria for IMRT and VMAT deliveries is appropriate when analyzing QA comparisons. A number of studies have reported on sensitivity differences between IMRT and VMAT cases, but with contradictory results[10–12]. It is conceivable that a universal criterion may be inappropriate for both IMRT and VMAT cases, along with the very different detector array geometries available for use. While a great deal of published data exists on observed gamma passing rates and proposed tolerances for clinical use[13–19], many of these QA comparisons lack the presence of a ground truth, which has been a fundamental issue in understanding IMRT QA sensitivity.

The conflicting nature of published data confounds our understanding of true sensitivity in IMRT QA, and as such the use of a universal gamma criterion remains the standard in the field for IMRT QA comparisons, which has very recently been reinforced by the publication of TG-218[20]. Indeed,

clinically relevant errors can exist in a plan and go undetected with commonly used gamma criteria, as gamma passing rates do not relate to clinical relevant end-points[3,8,21–30]. On top of this, the majority of these reports make use of real measurements to analyze the sensitivity of gamma comparisons. The use of real measurements can confound results due to different compounded sources of uncertainty, such as device limitations, beam modeling limitations, and machine delivery uncertainties.

The work in this chapter aimed to separately quantify the effect of detector geometry and spatial sampling on the sensitivity of the gamma comparison in current IMRT QA comparisons for three detector geometries – ArcCHECK, MapCHECK, and Delta 4. Sensitivity differences were also evaluated separately for IMRT and VMAT cohorts for all measurement geometries. While these delivery techniques are fundamentally different, VMAT and IMRT cohorts were curated to span similar ranges of field size and plan complexity. Confounding factors that are present in real IMRT QA comparisons, such as measurement uncertainty, machine limitations in plan delivery, output fluctuations, beam modeling limitations, as well as device calibration limitations were removed in our work. This was achieved by comparing calculations without induced errors to calculations with simulated errors. Thus, performing these comparisons in a calculation-only scenario allowed us to elucidate the true effects of detector geometry, spatial sampling, and delivery technique in a controlled environment.

In order to remove any vendor-specific differences in gamma comparison implementation, all gamma comparisons were performed using in-house developed code in MATLAB. This also allowed comparisons with various different spatial sampling, which would not be possible with current vendor software. Gamma sensitivity was quantified by calculating error curve widths as described in Chapter II, with error curve widths calculated for all criteria at 95% pixels passing as opposed to simply using the percent pixels passing. This allows us to understand if the actual

magnitude of errors passing IMRT QA for these different scenarios are truly different, which is not possible when simply using the percent pixels passing with γ < 1.

III.B. Methods

III.B.1. Case selection and plan complexity

Treatment plans chosen for this work were selected from a database of previously treated patients treated at our institution. Patient plans were fully anonymized and imported into a non-clinical Eclipse database for all subsequent dose calculations. Plans were chosen to represent a range of treatment sites with varying fields sizes and complexity scores for 20 IMRT and 20 VMAT deliveries. Plan complexity was calculated using the modulation complexity score (MCS), originally described by McNiven *et al*. for step-and-shoot IMRT plans[31] and modified for VMAT/sliding window deliveries by Masi *et al.* (MCSv)[32]. The MCSv metric was chosen as the main descriptor of complexity due to its ability to describe multiple aspects of the delivery scenario. From Masi *et al*., the MCSv value is calculated as:

$$MCSv_{plan} = \sum_{j=1}^{J} MCSv_j \times \frac{MU_j}{MU_{plan}} \qquad [Eq\ 3.1]$$

Where:

- *j* is the beam or arc number
- $MCSv_j$ is the MCVs value for beam or arc number *j*
- MU_j is the total monitor units (MU) for beam/arc *j*
- MU_{plan} is the total plan MUs

$$MCSv_j = \sum_{i=1}^{I} \frac{AAV_{CP_i} + AAV_{CP_{i+1}}}{2} \times \frac{LSV_{CP_i} + LSV_{CP_{i+1}}}{2} \times \frac{MU_{CPi,i+1}}{MU_j} \qquad [Eq\ 3.2]$$

Where:

- *j* is the beam or arc number
- *i* is the control point/segment number
- AAV_{CPi} is the aperture area variability for control point, *i*
- LSV_{CPi} is the leaf segment variability for control point, *i*
- $MU_{CPi,i+1}$ is the MU delivered between segments *i* and *i+1*
- MU_j is the total monitor units (MU) for beam/arc *j*

$$AAV_{CP} = \frac{\sum_{a=1}^{A}(\langle pos_a \rangle_{L\ bank} - \langle pos_a \rangle_{R\ bank})}{\sum_{a=1}^{A}(\langle \max(pos_a) \rangle_{L\ bank \in j} - \langle \max(pos_a) \rangle_{R\ bank \in j})} \quad [Eq\ 3.3]$$

Where:

- *A* is the total number of leaves in the bank
- *a* is the leaf number
- *pos* is the x-position of leaf *a*

$$LSV_{CP} = \left(\frac{\sum_{n=1}^{N-1}(pos_{max} - |(pos_n - pos_{n+1})|)}{(N-1) \times pos_{max}}\right)_{L\ bank} \times \left(\frac{\sum_{n=1}^{N-1}(pos_{max} - |(pos_n - pos_{n+1})|)}{(N-1) \times pos_{max}}\right)_{R\ bank}$$

$$pos_{max}(CP) = \langle \max(pos_{n \in N}) - \min(pos_{n \in N}) \rangle_{leaf\ bank} \quad [Eq\ 3.4]$$

Where:

- *N* is the number of open leaves in the beam
- *n* is the involved leaf

- *pos* is the x-position of leaf *n*

MCSv incorporates measures of leaf position variability between adjacent leaves, the degree of variation of aperture sizes over the maximum aperture area for a beam, as well as weighting the importance of these values for each control point by the fractional MU for a given delivery segment. This is achieved by calculating the leaf sequence variability (LSV) and aperture area variability (AAV) for each control point. The LSV describes the positional differences between adjacent MLC leaves in the same leaf bank for a given control point, while the AAV is a measure of each segment's area, relative to the maximum aperture over all the segments for a given beam or arc. The mean values for LSV and AAV between a given control point and the previous control point are then multiplied together and weighted by the fractional monitor units for a given segment within a beam or arc to obtain $MCSv_{beam/arc}$. The complexity score for the entire plan, $MCSv_{plan}$, is the sum of MCSv over all beams (or arcs), weighted by the fractional MUs for each beam/arc over the entire plan. Values for MCSv range from 0 to 1, with values closer to 0 representing more complex plans, while a MCSv value of 1 would represent the least complex delivery (i.e. a static beam, jaw-defined field).

While plan selection was based primarily on ensuring that the IMRT and VMAT cohorts had similar complexity scores, plans were also selected to encompass a range of field sizes. Plan field size was coalesced into a single metric by averaging the equivalent square field size over all beams in the plan:

$$Avg.Equivalent\ Square\ Field\ for\ Plan = \frac{1}{N}\sum_{n=1}^{N}\frac{(2XY)}{(X+Y)} \qquad [Eq\ 3.5]$$

where *N* is the total number of beams or arcs in a treatment plan, and corresponding jaw coordinates are represented by *X* and *Y*. These equivalent square field sizes are shown for the IMRT and VMAT cohorts in Figure 3-1. It is important to represent a range of field sizes for this work since IMRT QA comparisons may be limited by sparse spatial sampling, and plans with smaller treatment volumes offer fewer high dose, in-field points for analysis, which may have the potential to limit gamma comparison sensitivity.

Complexity values (MCSv) for the IMRT and VMAT cohorts are shown in Figure 3-2. A variety of treatment sites are represented in both cohorts as shown in Figure 3-3, however, matching IMRT and VMAT cohorts by site was not of primary importance when selecting cases. This is because IMRT QA comparisons are unlikely to be directly affected by the specific treatment site, and more so by plan complexity and field size since patient geometry is de-coupled in IMRT QA comparisons.

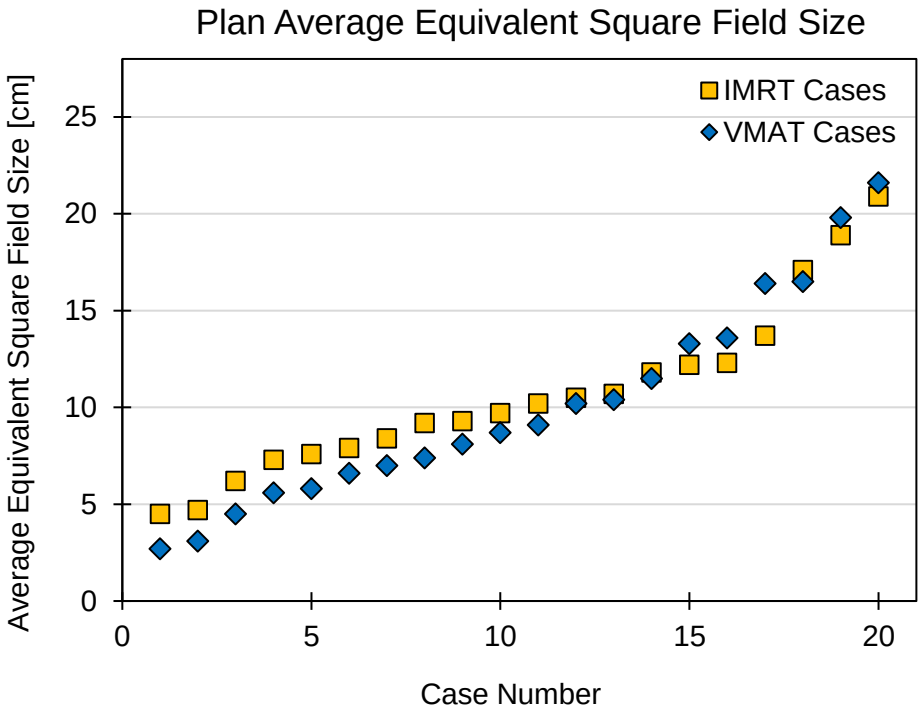

Figure 3-1. Plans with a range of field sizes were chosen for this study, since field size directly affects the number of high dose points available for analysis in IMRT QA comparisons.

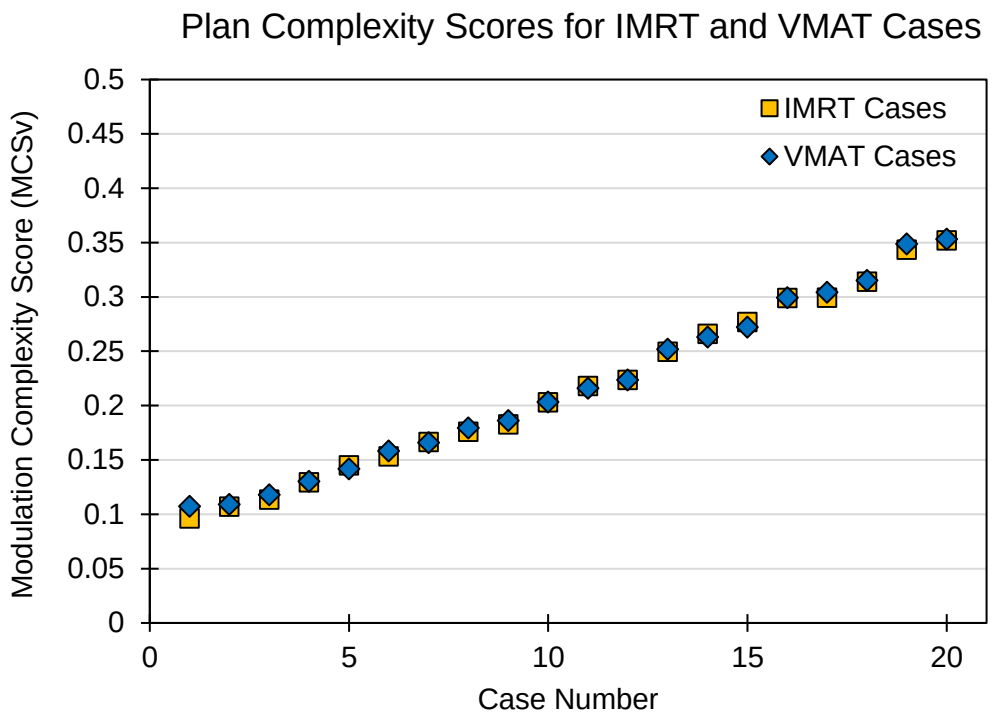

Figure 3-2. Treatment plans were selected such that the MCSv complexity metrics were similar between IMRT and VMAT cohorts.

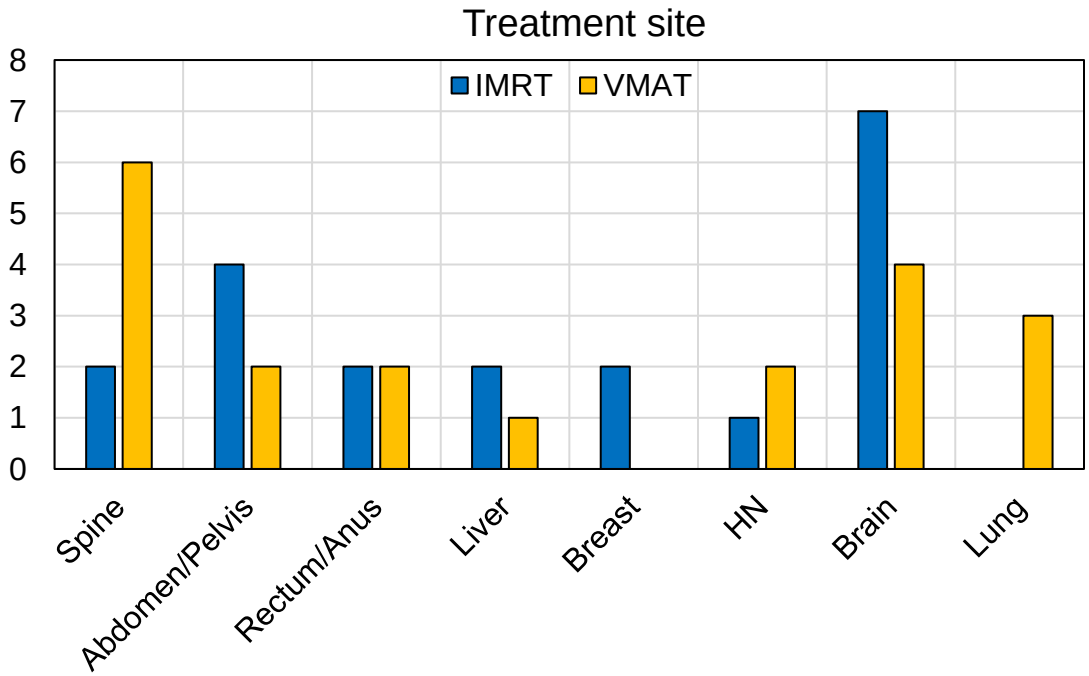

Figure 3-3. The IMRT and VMAT cohorts consist of plans from a range of treatment sites. However, plans were not chosen based on treatment site, so differences in site distribution do exist between the IMRT and VMAT cases.

III.B.2. Detector devices studied

In this work three diode detector array geometries – MapCHECK 2 (henceforth referred to as simply MapCHECK), ArcCHECK, and Delta 4 – were utilized for error sensitivity studies such that results represent a range of measurement geometries in current clinical use. The MapCHECK device is a planar array (Sun Nuclear Corporations, Melbourne, FL)[33] consisting of 1,527 diodes with 7.07 mm spacing and an array size of 32 cm x 26 cm. The Delta 4 device (ScandiDos, Uppsala, Sweden)[34] is a cross-planar array that consists of two planar arrays perpendicular to one another containing a total of 1,069 diodes. The Delta 4 is unique in that the detector spacing is 5mm in the central 6x6 cm^2 area and is less dense with 1 cm spacing outside the central area of the array. The full array size for both planar arrays is 20 cm x 20 cm. The ArcCHECK device was investigated here as well, which consists of 1,368 diodes arranged helically in a cylindrical geometry[35,36]. Detectors sit on a cylindrical surface, 10.5 cm from the center of the device, with 1 cm spacing between adjacent diodes. All studied detector geometries are shown in Figure 1-3 (a-c). For each device, the vendor-provided phantom for TPS dose calculations was obtained and utilized for all dose calculations in the Eclipse TPS. These phantoms are homogenous representations of the actual devices and do not model the internal structure of the devices. However, since all comparisons here are calculation vs. calculation comparisons, any differences or limitations this causes in clinical IMRT QA comparisons are removed in this study.

III.B.3. Simulated errors

A variety of different modifications were made to the cases in this work to simulate treatment errors in plan dose calculations. Plans with simulated errors were recalculated in the Eclipse TPS (AAA v.13, 1mm grid size) and subsequently compared to calculations without induced errors. The errors simulated in this study were chosen by first imagining various failure modes that could

occur at different stages of the treatment planning and delivery process and narrowing this list to a smaller subset of errors for use in this work.

These errors were chosen such that they represented potential failures of various different components of the delivery process, but that may also go undetected using current comparison methods (i.e. the gamma comparison) and have the potential to significantly impact delivered dose distributions. Additionally, it is useful to induce an error that can simulate various different error classes. For example, a shift in the MLC bank positions could result from a miscalibration of the MLC banks on the machine, failure of the carriage mechanism, or inappropriate MLC calibration tables on the linear accelerator. Additionally, this same induced error could also serve to simulate the scenario in which there is a field size mismatch between the measurement and calculation, which could occur during the modeling and commissioning of the dose calculation algorithm in the treatment planning system. A total of five errors were induced in plan calculations:

- MU Errors
- MLC positional errors
    - MLC bank shifts
    - Leading/trailing leaf in center of field
    - Random perturbational error to all involved leaves
- Collimator angle error

The rationale for including these errors are described in detail in the following sections.

MU Errors

An absolute dose scaling error may be one of the most basic yet important errors to catch in the IMRT QA process. While it is unlikely to have a significant output error on a machine, it has also been shown in the previous chapter that current methods on the ArcCHECK device can be

insensitive to large dose scaling errors. Since ensuring the delivered dose closely matches the planned dose is of upmost importance in radiation therapy treatments, this error was chosen as high priority for studying error sensitivity.

Scaling the plan MUs by a percentage of the original MUs serves to simulate errors in machine calibration, plan normalization issues, use of the incorrect calibration files for the measurement phantom, or the QA phantom requiring recalibration. MU errors can also serve to simulate an incorrect machine, since all machines are likely to have slightly different output tuning at any particular time, as well as differences in TPS calculation models. A mismatch in calculation and measurement energies was found to be approximated by MU errors as well. This is illustrated in Figure 3-4 by comparing a calculation for a 6X plan to calculations using the same machine but different energies (6X-FFF and 15X) calculated on the ArcCHECK geometry. Monitor unit (MU) errors were induced in this study by scaling the plan MUs uniformly up or down to make the resulting plan hot or cold compared to the original plan. Induced MU errors ranged from -15% to +15%.

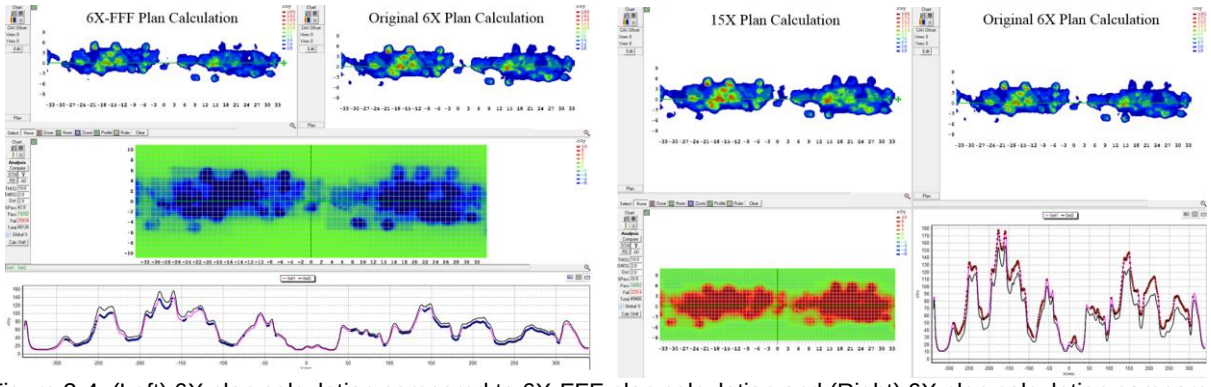

Figure 3-4. (Left) 6X plan calculation compared to 6X-FFF plan calculation and (Right) 6X plan calculation compared to 15X plan calculation. Both show that use of the wrong energy looks similar to an absolute dose scaling error in a typical IMRT QA comparison.

MLC Positional Errors

Three different induced MLC positional errors were included in this work. Each of these errors can aid in simulating delivery on the wrong treatment machine or calculation from the incorrect

machine model since MLC tolerances and performance are likely to differ between machines. Additionally, these errors could also serve to simulate corrupted MLC leaf trajectory files which have been reported in the literature[37].

All leaf errors in this work were induced by exporting the original MLC leaf trajectories and using in-house MATLAB code to modify leaf positions. After modifications were made, the MLC files were imported back into the Eclipse TPS for dose calculation. A minimum leaf gap of 0.05 cm was enforced by the MATLAB code used to modify the MLC files and the "verify leaf positions" option was also utilized in the Eclipse TPS to ensure that the modified MLC files did not violate the delivery constraints of the TrueBeam linear accelerator. When modifications were required by the Eclipse TPS to allow deliverable leaf sequences, these changes were accepted.

MLC bank shifts were induced in plan calculations to cause a widening or narrowing of the field width. This serves to simulate bank calibration issues, a failure in the carriage mechanism, inaccurate MLC calibration tables on the linear accelerator, or improper modeling of the field width in the calculation algorithm. These errors are induced by uniformly shifting the MLC positions in or out by a user-specified symmetric distance.

A trailing/leading leaf error was also induced in this study to simulate scenarios where a particular leaf has difficulty reaching the desired position due to an overly complex delivery or a bad motor for that given leaf. This error is induced by exporting MLC files for a plan and inducing a range of magnitudes by which a leaf in the central portion of the field is leading or trailing. The "centrally located" leaf was chosen automatically by determining the involved leaf numbers for a given plan and then selecting the most centrally located leaf number from the involved leaves. Trailing leaf errors were induced in the centrally located leaf in the left bank of the delivery. Errors were induced for a range of error magnitudes ranging from 0.2 cm to 5.0 cm. For IMRT deliveries, the modified leaf position x_m, for the trailing leaf is:

$$x_m = x_{i-1} - m, \quad for\ x_i - x_{i-1} > 0\ and\ |x_i - x_{i-1}| > m$$

$$x_m = x_{i-1}, \quad for\ |x_i - x_{i-1}| < m \qquad [Eq\ 3.6]$$

Where:

- *i* is the control point index
- *m* is induced error magnitude, or the lag distance
- *x* is the leaf position

This is also illustrated in Figure 3-5, where instead of reaching the position x_i, the leaf only makes it to position x_m. If the difference between leaf positions between the current control point x_i and the previous control point x_{i-1} is smaller than the induced error magnitude, the leaf remains at position x_{i-1}.

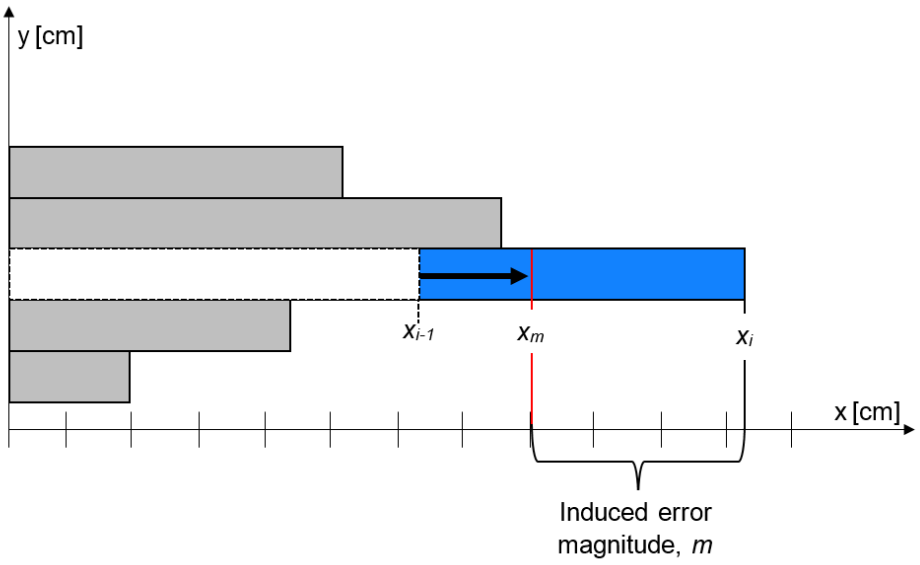

Figure 3-5. The lagging leaf error causes the position of the centrally located leaf in a plan to be modified by the induced error magnitude, m as shown for a leaf moving from left to right.

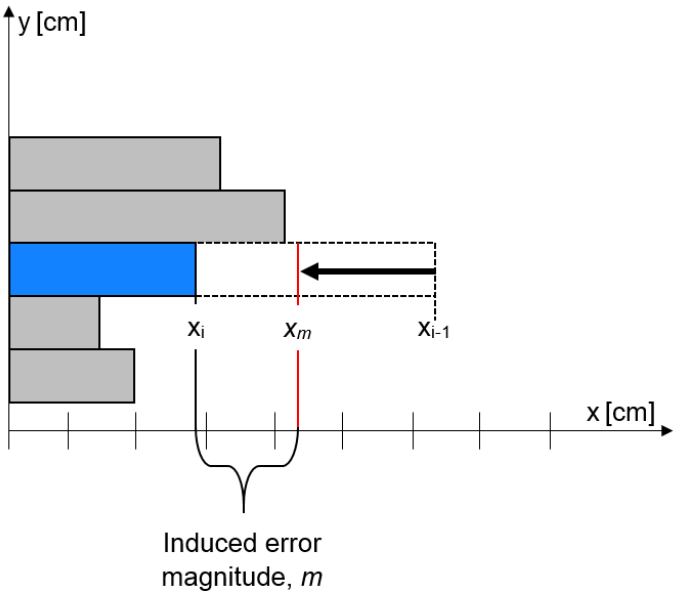

Figure 3-6. The lagging leaf error causes the position of the centrally located leaf in a plan to be modified by the induced error magnitude, m, as shown for a leaf moving from right to left.

For VMAT plans this trailing leaf error is induced in a slightly different manner due to the fact that leaves do not always move in the same direction between adjacent control points for VMAT deliveries. Thus, the directional motion of the leaf in a given control point must be taken into account. This is illustrated visually in Figure 3-5 and Figure 3-6 and defined mathematically as

$$x_m = x_{i-1} - m, \ for \ x_i - x_{i-1} > 0\,, \qquad |x_i - x_{i-1}| > m$$

$$x_m = x_{i-1} + m, \quad x_i - x_{i-1} < 0, |x_i - x_{i-1}| > m$$

$$x_m = x_{i-1}, \quad |x_i - x_{i-1}| < m \qquad [Eq\ 3.7]$$

Where:

- *i* is the control point index
- *m* is the distance of the trailing leaf
- *x* is the leaf position

A third MLC error, which was previously described in Chapter II, applied small, random perturbational shifts to each active leaf in every control point of a beam to simulate inaccuracies in MLC trajectories or control point definitions, as well as corrupted leaf position information, and difficulty delivering highly modulated plans. Similar to previous MLC errors, MLC files were exported and modified using an in-house MATLAB program. Each active leaf in every control point was perturbed by a value chosen pseudo-randomly from a square, two-sided distribution bounded by varying user-defined ranges. The width of this square distribution was varied to induce a range of error magnitudes. For example, a user input of "1.0" represents a square distribution ranging from -1.0 cm to +1.0 cm. This means that while some leaves will experience changes in leaf position up to 1 cm in either direction, most leaves will experience much smaller shifts.

Collimator Errors

Initially, both collimator and gantry angle errors were considered, including errors in IEC angle conventions. The effects of flipped IEC angle conventions were simulated for two measurement geometries (ArcCHECK and MapCHECK) in calculation vs. calculation scenarios using SNC Patient software. For a simple brain case and a complex brain case the gamma passing rates at 3%/3mm TH10 (G) were found to be extremely low. This hints that this error would be easily detected with current methods. Additionally, quality assurance checks associated with an angle conversion in the clinic should likely catch these errors and thus, IEC convention errors were eliminated from the list of errors for this work.

Gantry angle errors were also considered and investigated. For the two sample brain cases previously mentioned, errors of 3° or more would cause a failing IMRT QA result (with the exception of the simple case calculated on the MapCHECK), as shown in Table 3-1. The differences in PTV DVH curves for these small errors are shown in Figure 3-7, illustrating that

smaller gantry angle errors do not result in dramatic DVH differences for PTV coverage for these cases.

Table 3-1. Gamma passing rates calculated at 3%/3mm TH=10% (G) in SNC Patient (v.6.7.3.) for two cases (simple and complex) for the ArcCHECK and MapCHECK geometries. Passing rates are shown as a function of increasing gantry angle error.

	Case 1 - simple		Case 11 - complex	
Gantry error [degrees]	ArcCHECK Gamma Passing [%]	MapCHECK Gamma Passing [%]	ArcCHECK Gamma Passing [%]	MapCHECK Gamma Passing [%]
1°	100%	100%	100%	98.6%
2°	93.7%	100%	86.4%	90.8%
3°	83.3%	99.7%	69%	83%
5°	62.7%	96.2%	50.3%	65.7%
10°	42.4%	79.6%	32.8%	53%

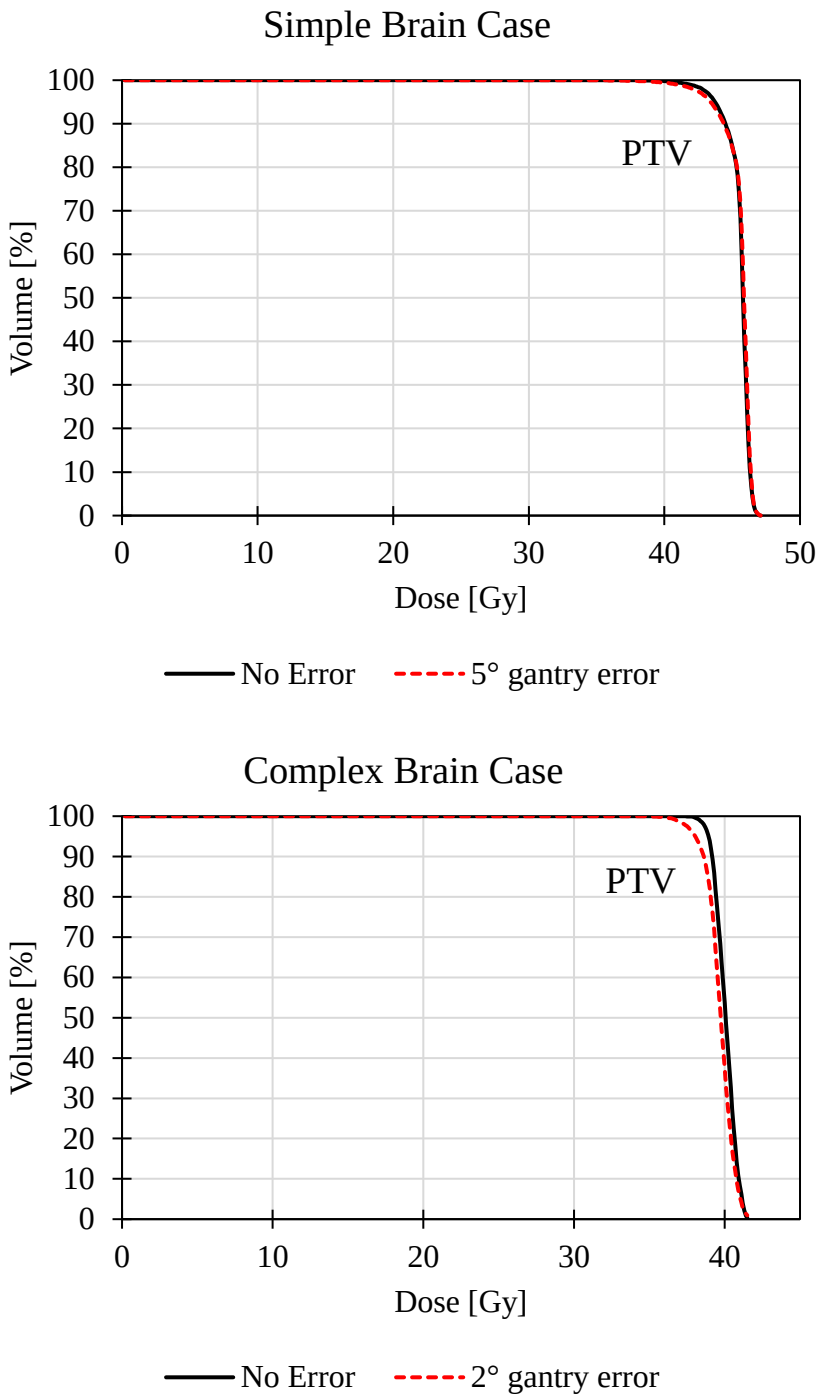

Figure 3-7. Induced gantry angle errors for a simple case (top) and a complex case (bottom) showing that the gantry error magnitude that would cause failures for commonly used gamma criterion 3%/3mm TH10 (G) would cause minimal differences to the PTV coverage.

Alternatively, differences in collimator angles may not be as readily flagged by current methods and could have sizeable effects on dose distributions should errors exist. Table 3-2 shows that large collimator discrepancies could go undetected using typical gamma criteria (3%/3mm TH 10, global normalization). For the simpler brain case, gamma passing rates were above 90% for 10-20° collimator errors, depending on the measurement device for 3%/3mm TH 10 (G). Errors this large could cause concerning DVH dose differences, especially when considering PTV coverage, shown in the top panel of Figure 3-8. For the more complex brain case, smaller collimator errors caused the plan to fail the 3%/3mm TH10 (G) criterion faster. However, for this case, a smaller error such as 3° in collimator rotation could cause concerning dose differences, as seen in the lower panel of Figure 3-8. Compared to gantry angle errors, collimator errors are less likely to be flagged by current methods and have the potential to cause greater DVH dose differences, leading us to select collimator errors only for this work. It should be noted that we are not suggesting gantry angle errors are unimportant, or that there are some cases where a gantry angle error could be similarly detrimental in terms of dose differences. Rather, with the use of current IMRT QA comparison methods, collimator errors are less likely to cause a failing gamma comparison while also having a potentially greater impact on the DVH metrics compared to gantry angle errors of similar magnitude.

Table 3-2. Gamma passing rates calculated at 3%/3mm TH=10% (G) in SNC Patient (v.6.7.3.) for two cases (simple and complex) for the ArcCHECK and MapCHECK geometries. Passing rates are shown as a function of increasing collimator angle error.

	Case 1 - simple		**Case 11 - complex**	
Gantry error [degrees]	ArcCHECK Gamma Passing [%]	MapCHECK Gamma Passing [%]	ArcCHECK Gamma Passing [%]	MapCHECK Gamma Passing [%]
1°	100%	100%	100%	100%
2°	100%	100%	99.8%	99.9%
3°	100%	100%	98.8%	98%
5°	99.7%	100%	88.4%	90.5%
10°	95.3%	98.2%	67.7%	72.9%
20°	78.8%	93%	45.9%	48.8%

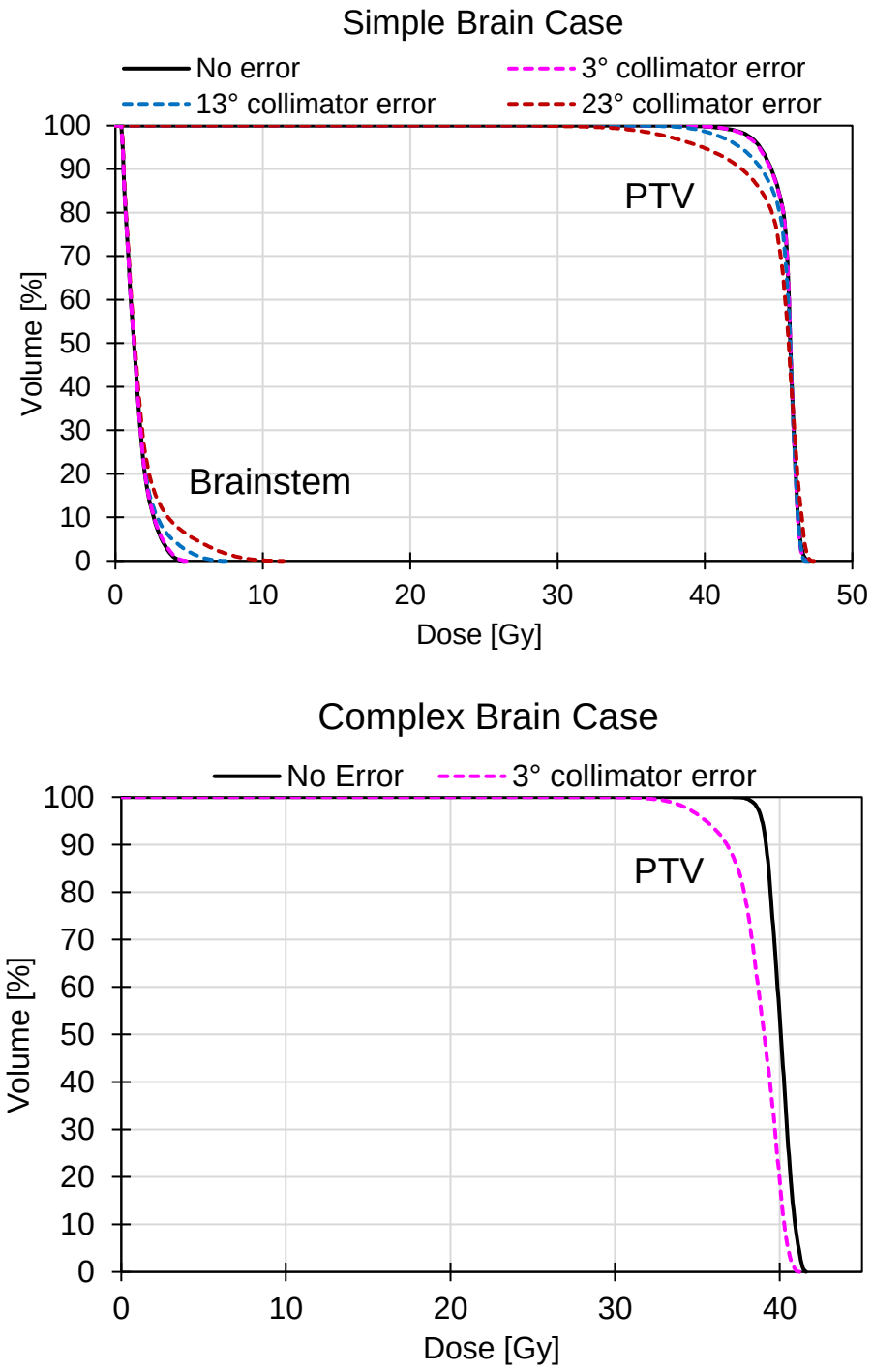

Figure 3-8. Induced collimator angle errors for a simple case (top) and a complex case (bottom). Unlike gantry errors, smaller collimator errors for the complex case could result in passing IMRT QA values, but large differences to the PTV coverage. Similarly, the magnitude of the collimator rotation error required to fail the simple brain case also causes concerning DVH dose differences.

Wrong Patient

The issue of treating John Doe #1 with the plan from John Doe #2 is one type of error that causes concern for some. This scenario is often imagined in the context of prostate patients, where the anatomy and treatment field designs may often be markedly similar. While many safety checks exist to prevent this issue, it is worth considering for this work if such an error would be flagged when comparing John Doe 1's IMRT QA calculation to John Doe 2's IMRT QA measurement. In order to determine the likelihood of this error going undetected in IMRT QA, the IMRT QA calculations for current on-treatment prostate patients from our TrueBeam that treats a large majority of prostate patients were exported. Calculation vs. calculation comparisons between all combinations of current on-treatment prostate patients were compared for both ArcCHECK and MapCHECK geometries. It was found that for all combinations of current on-treatment prostate patients, all calculation vs. calculation comparisons had passing rates below 50% at 3%/3mm TH 10 (G). This means that this type of error would likely be flagged with current IMRT QA techniques should it occur prior to patient treatment and thus this error was not included in this work.

Wrong Plan, Same Patient

Another error to consider would be the circumstance where a patient had both IMRT and VMAT plans generated and the incorrect plan was pushed to the machine, while the other plan calculation was exported for comparison. To consider the likelihood of this error being caught with commonly used IMRT QA metrics, a brain patient was identified that had both clinical IMRT and VMAT plans created for both their initial treatment and boost treatment. The IMRT and VMAT plans were exported and compared in SNC Patient on both the ArcCHECK and MapCHECK geometries. At 3%/3mm TH10 (G) passing rates for the initial plan were 15.9% on the ArcCHECK and 31.5% for the MapCHECK geometry. The boost plan comparison between VMAT and IMRT

plans for the MapCHECK geometry was 36.3%. Since this error is 1) unlikely to occur and 2) very likely to be flagged with current methods, this error was not included in this work.

Device-specific Limitations

Each measurement array has different device-specific limitations either in terms of TPS dose calculations or in the measurement device itself. However, since each phantom has unique weaknesses these errors were not considered in this work. Such limitations include difficulty predicting dose for laterally angled beams in the MapCHECK device, and beams at cardinal gantry angles for the Delta 4 device when using the AAA photon dose calculation algorithm in Eclipse. Additionally, predicting entrance/exit ratios for the ArcCHECK detectors in the Eclipse AAA is limited in accuracy due to difficulty predicting depth doses in the PMMA material with Eclipse heterogeneity corrections.

III.B.4. MATLAB gamma comparison implementation

Gamma comparisons were performed using in-house code written specifically for this work. In order to do this, several steps were required. First, the 3D dose files were exported from the Eclipse TPS and processed using the CERR package[38]. This allows the DICOM dose files to be extracted and placed into a *.mat file easily imported into MATLAB. For some detector geometries, there was no dose point at (0,0,0) in the phantom. When this was the case, the dose cubes were interpolated and re-centered to obtain a dose matrix that was centered at the absolute center of the phantom.

Extensive validation was performed to ensure that the dose matrices were oriented appropriately and that the dose obtained from the CERR extracted MAT files matched that shown in Eclipse. This was performed by spot-checking point doses in Eclipse compared to the MATLAB extracted doses, and also by using select line profiles through the center of each device for a patient test

plan. After this was performed, the test plan was calculated on each phantom geometry and the doses extracted by the vendor software were compared to those extracted at the plane of the detectors using the in-house MATLAB code. For the ArcCHECK device this consisted of extracting doses on a cylindrical plane 10.5 cm from the center of the phantom geometry. For MapCHECK and Delta 4 this consisted of extracting doses at the plane of the diodes in the devices. For the Delta 4 device, the doses from the in-house MATLAB extraction exactly matched the doses extracted by the Delta 4 software. However, for the ArcCHECK and the MapCHECK the dose extractions varied slightly from those extracted by the SNC Patient software. Effort was made to determine the source of these small differences, but it was concluded that the SNC Patient software might be performing some type of interpolation unbeknownst to the user. However, the absolute differences in dose extraction for MapCHECK and ArcCHECK geometries were extremely small in nature and did not result in different gamma passing rates when testing select plans using the vendor software and the in-house software for error-free calculations compared to error-induced calculations.

The in-house MATLAB gamma comparison code was written such that gamma comparisons could be performed for two different spatial sampling scenarios. Error-free calculations were always extracted with 1mm spatial sampling, while error-induced calculations could be extracted either at 1mm sampling or at the true spatial sampling of the detector. Gamma comparisons could then be performed for either sampling scenario between error-free calculations and error-induced calculations. Error-free calculations were considered the “reference” dose distribution, and the error-induced calculations as the “evaluated” distribution. Gamma comparisons were performed using the method originally described by Low *et al.*[39].

III.B.5. Effects of detector geometry on gamma sensitivity

Using the in-house gamma comparison code, the effects of different detector geometries on gamma comparison sensitivity were studied for the IMRT and VMAT case cohorts. This was performed by comparing error curve ranges for the ArcCHECK, MapCHECK, and Delta 4 geometries using 1mm error-free calculations compared to 1mm calculations with induced errors. The comparisons made use of 1mm vs. 1mm calculations in order to remove the effects of spatial sampling so as to decouple device geometry from spatial sampling results. Additionally, calculation-only comparisons remove any potential effects of limitations and uncertainties inherent in the use of real measurements.

Gamma comparisons were calculated for five different gamma criteria in order to understand any criterion-specific behavior. Care was taken in selecting the gamma criteria to be used such that various combinations of normalization strategies, use of different low-dose thresholds, and several percent dose difference and DTA criteria were included. The five studied gamma criteria were:

- 3% / 3mm, TH 10%, global normalization
- 3% / 3mm, TH 10%, local normalization
- 2% / 2mm, TH 10%, global normalization
- 3% / 3mm, TH 50%, global normalization
- 3% / 3mm, TH 50%, local normalization

For each studied gamma criterion, the error curve widths were calculated at 95% pixels passing using the method described in Chapter II.

Statistical Analysis

The assumption of normality was tested using Shapiro-Wilk and Q-Q plots. Shapiro-Wilk results showed that normality could not be assumed, however, the Q-Q plots showed close to normality for some cases. Since the assumption of normality was not obvious and some cases had large error curve ranges for a given error type (that could be considered outliers), non-parametric tests were utilized in this work. Kruskal-Wallis one-way analysis of variance was performed to test for statistically significant differences between the median error curve ranges between the three devices. This is performed separately for each gamma criterion and error type. P-values < 0.05 caused the rejection of the null hypothesis that the median error ranges between the three devices were equal. Post-hoc analysis was performed in the event of a significant result using Dunn's test to determine which devices had statistically significant differences with $\alpha = 0.05$. Since the Dunn's test consists of multiple comparisons, a Bonferonni correction is applied accordingly and we reject H_0 if $p < \alpha / 2$, where H_0 is that the medians between the two compared groups are equivalent. All statistical analyses were performed in R.

III.B.6. Effects of spatial sampling on gamma sensitivity

Using the in-house gamma comparison code, the effects of spatial sampling on gamma comparison sensitivity for each of the three detector devices was studied. This was achieved by computing error curve ranges between 1mm error-free calculations and error-induced calculations downsampled to the real detector locations and subsequently comparing these error ranges to those obtained from comparisons between a 1mm error-free calculation to an error-induced calculation at 1mm sampling (from section III.B.5). Error curve ranges between the two different spatial sampling scenarios are compared separately for each device in order to determine if the resolution of each particular device has significant impacts on the computed error

curve ranges, and ultimately, on the sensitivity of that particular device to induced errors in conjunction with the gamma comparison.

Gamma sensitivity was quantified using the error curve method presented in Chapter II and error curve widths were calculated at the 95% pixels passing level. The same five gamma criteria studied in section III.B.5. – 3% / 3mm, TH10 (G), 3% / 3mm, TH10 (L), 2% / 2mm, TH10 (L) 3% / 3mm, TH50 (G), and 3% / 3mm, TH50 (L) – were also utilized here to elucidate if sensitivity differences are criterion-specific.

Statistical Analysis

Wilcoxon signed rank tests were utilized to compare the differences in error curve ranges for the different spatial sampling scenarios for the same device to determine whether the medians of the two groups are statistically significant. Wilcoxon signed-rank tests were used as the non-parametric alternative to the paired Student's t-test since we cannot assume normality of our data. Significance was defined at the $p < 0.05$ level. All statistical analyses were performed in R.

III.B.6. Sensitivity differences between VMAT and IMRT deliveries

Sensitivity differences between IMRT and VMAT deliveries were studied by comparing the error curve ranges calculated in section III.B.5 (1mm error-free calculation vs. 1mm error-induced calculation) between the IMRT and VMAT cohorts for each studied error type and the five previously mentioned gamma criteria.

Statistical Analysis

Mann-Whitney U tests were utilized to compare the differences in error curve ranges for the different delivery techniques for each device to determine whether the medians of the two groups were statistically significant. Mann-Whitney U tests were used as the non-parametric alternative

to the independent Student's t-test since we cannot assume normality of our data. Significance was defined at the $p < 0.05$ level and all statistical analyses were performed in R.

III.C. Results

III.C.1 Effects of detector geometry on gamma sensitivity

Error curve widths for 1mm error-free calculations compared to 1mm error-induced calculations were evaluated between the three studied detector geometries (ArcCHECK, MapCHECK, and Delta 4) to determine if a particular geometry offers higher sensitivity to errors in conjunction with the gamma comparison. The real spatial sampling of each device is not used in this portion of the work as it could confound the comparisons and make it difficult to separate the independent effects of the actual detector geometry from that of the unique detector arrangements for each device. Kruskal-Wallis p-values are shown in the figures for each gamma criterion, where $p < 0.05$ represents that the medians between the three detector geometries were statistically significantly different. Post-hoc analyses for differences showing statistical significance are available in Appendix A. However, it should be noted that statistical analysis alone neglects the visible trends in our data and sample sizes were relatively small with *n=20* cases in each cohort.

III.C.1.1 IMRT Cases

The median error curve ranges with error bars of ±1σ are shown for the five error different error types in Figure 3-9, Figure 3-10, Figure 3-11, Figure 3-13, and Figure 3-14 for each detector device and each of the five studied gamma critiera. Across the studied gamma criteria, the Delta 4 and MapCHECK devices had similar error sensitivity for the lagging leaf, bank shift, and MU errors, with MapCHECK generally slightly less sensitive than the Delta 4 geometry, while the ArcCHECK had lower sensitivity to these three errors compared to the MapCHECK and Delta 4. While the differences in error sensitivity between ArcCHECK and the Delta 4 and MapCHECK were not always statistically significant, the trends in Figure 3-9, Figure 3-10, clearly show that for

most of the studied gamma criteria, the ArcCHECK generally performs worse in detecting lagging leaf and bank shift errors. For the MU error, the ArcCHECK had considerably worse error sensitivity than the MapCHECK and the Delta 4, as shown in Figure 3-11.

For these three error types the ArcCHECK had lower sensitivity (i.e. larger error curve ranges) compared to MapCHECK and Delta 4 for nearly every gamma criterion, with the exception of 3%/3mm TH 10 (L). The use of 3%/3mm TH 10 (L) resulted in error curve ranges that were markedly similar between all three devices. This is likely owing to the fact that, compared to the planar geometries, there are more low dose measurement points in ArcCHECK comparisons, due to the fact that ArcCHECK measures both entrance and exit doses from any given beam. In a globally normalized setting, these low dose points can serve to inflate gamma passing rates in the presence of an induced error, resulting in larger error curve ranges for the ArcCHECK. However, in a locally normalized setting, these low dose points more easily fail in the presence of an induced error, thus increasing error sensitivity on the ArcCHECK. Since the other devices do not have this same distribution of low-dose measurement points, globally normalized comparisons for MapCHECK and Delta 4 do not result in this same decrease in error sensitivity that is seen on the ArcCHECK device.

In Chapter II we also observed that a 50% low dose threshold for globally normalized gamma comparisons increased error sensitivity on the ArcCHECK when compared to the same criterion with a lower dose threshold – i.e. 3%/3mm TH 10 (G) compared to 3%/3mm TH 50 (G). These same trends are observed here, however, the 50% dose threshold does not equalize the error sensitivity between devices as the locally normalized comparisons did. Additionally, a locally normalized comparison with a higher low dose threshold (3%/3mm TH 50, local) does not equalize error sensitivities as well as 3%/3mm TH 10 (L). This is not surprising given that these low dose points that are available to improve error sensitivity in the 3%/3mm TH 10 (L)

comparisons are removed when applying a 50% dose threshold. For Delta 4 and MapCHECK geometries the use of a 50% dose threshold or local normalization does offer slight gains in relative error sensitivity, but the differences are minimal and much less apparent than for the ArcCHECK device, again most likely owing to the different distribution of low dose points in these planar-type devices.

An illustrative example is also presented in Figure 3-12 showing error curve ranges for each of the 20 IMRT cases for one induced error type (MU errors) for two sample gamma criteria – 3%/3mm TH 10 (G) and 3%/3mm TH 10 (L) – for each of the three devices. This example shows that the local normalization of the data systematically reduces error curve ranges for all cases, and not just a select few, suggesting that this behavior is specific to the device geometry and not a particular set of cases. This further confirms that the treatment of the extra low dose points in the ArcCHECK geometry may be important to consider in gamma comparisons for IMRT cases measured on this device.

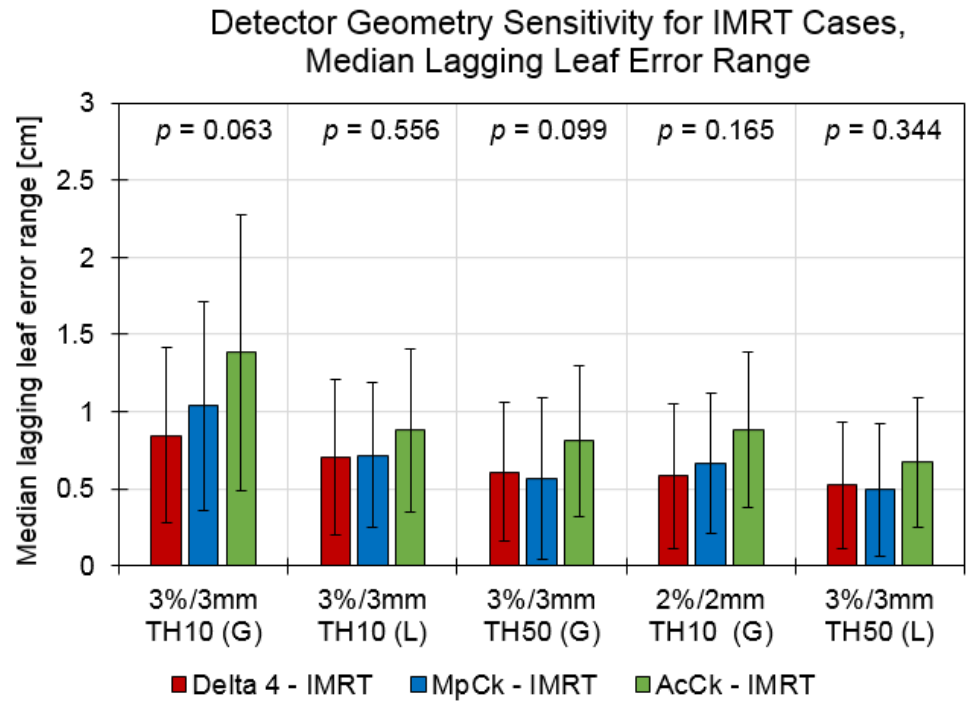

Figure 3-9. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for a lagging leaf error. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown.

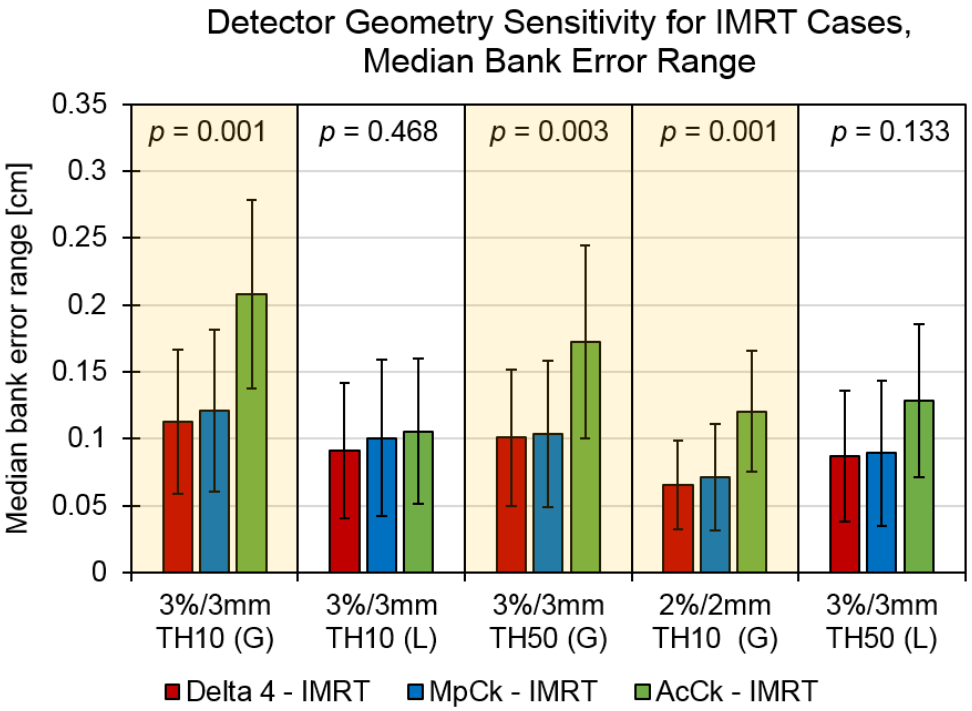

Figure 3-10. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for a bank shift error. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown. P-values < 0.05 signify a significant difference between median error curve ranges for a given gamma criterion and are highlighted in yellow.

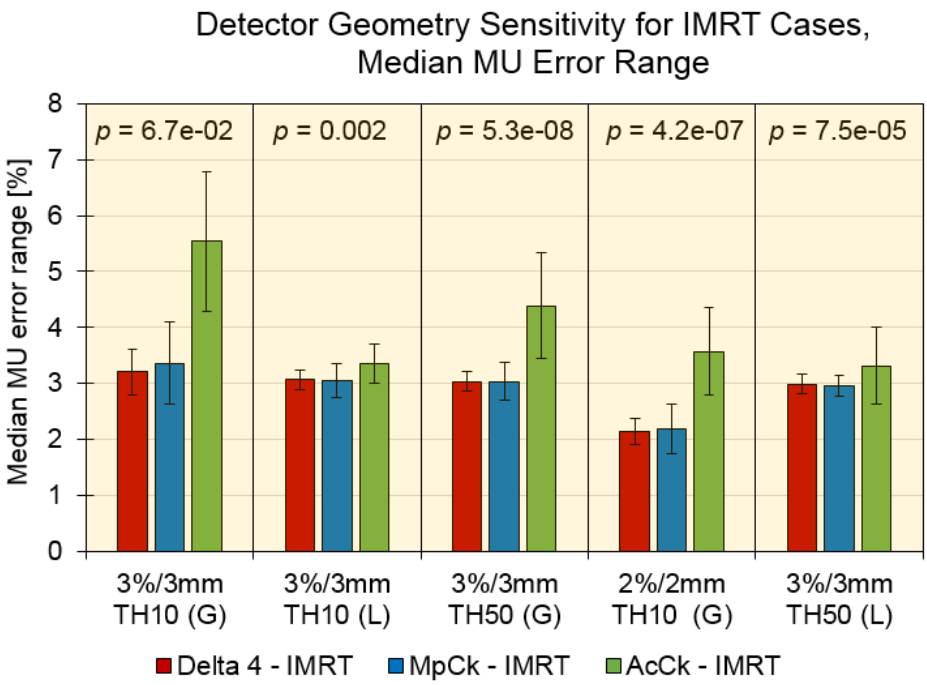

Figure 3-11. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for induced MU errors. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown. P-values < 0.05 signify a significant difference between median error curve ranges for a given gamma criterion and are highlighted in yellow.

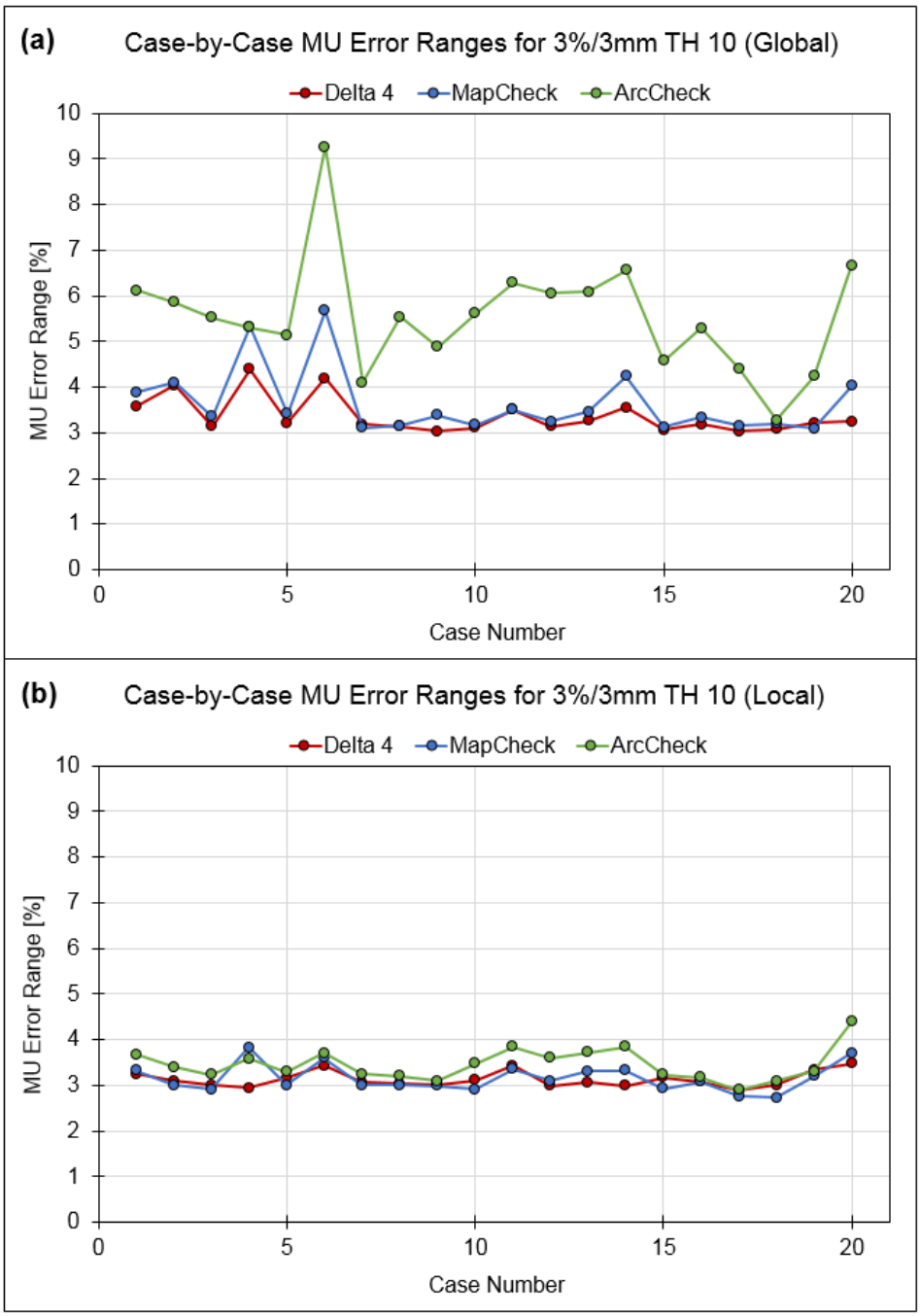

Figure 3-12. Case by case example of the error ranges for the three devices for (a) 3%/3mm TH 10, global normalization and (b) 3%/3mm TH 10, local normalization. This example shows that the local normalization for MapCHECK and Delta 4 reduces error curve ranges marginally, but for the ArcCHECK, error curve ranges are reduced systematically for all cases.

Error curve results for the induced collimator error and the MLC perturbational error exhibit somewhat different trends. Figure 3-13 shows that for the collimator rotation error, the ArcCHECK and Delta 4 devices perform slightly better than the MapCHECK. The results from MLC perturbational error in Figure 3-14 show that the ArcCHECK had the highest sensitivity to this error type, followed by the Delta 4, with the MapCHECK having the lowest error sensitivity for this error type. As with the previous error types, the 3%/3mm TH10 (L) criterion was able to mostly equalize error sensitivities between the different devices. However, for the MLC perturbational error on the ArcCHECK local normalization made error sensitivity superior for this combination of error type and device. These results lead us to conclude that error sensitivity for different devices is markedly different for different types of errors. In general, for globally normalization comparisons, the MapCHECK is generally less sensitive than the Delta 4, whereas the ArcCHECK is less sensitive than these two devices for only specific error types.

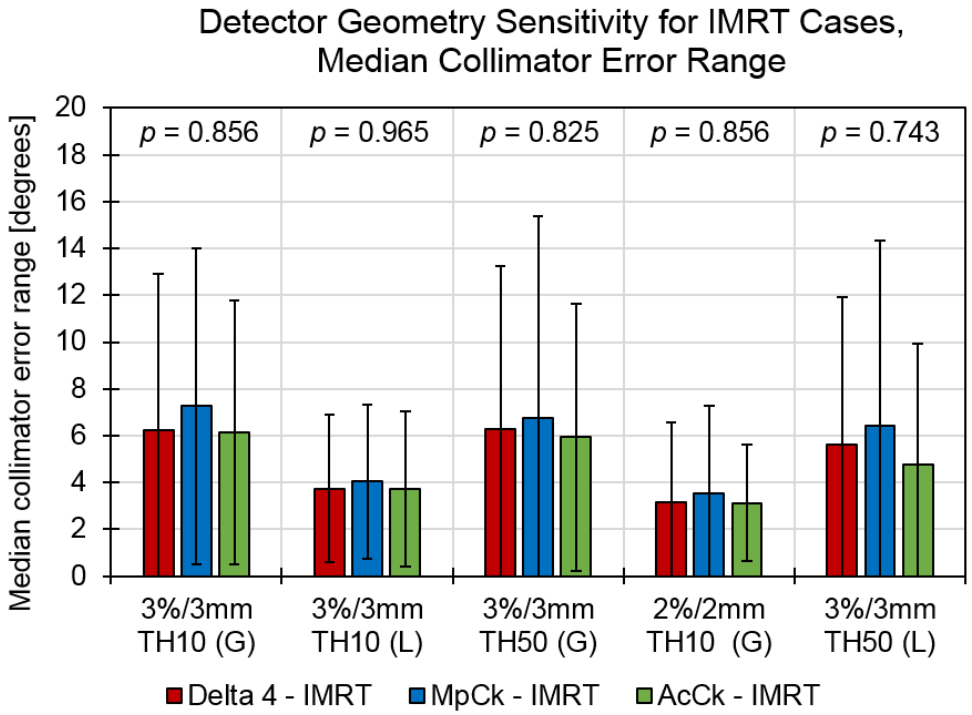

Figure 3-13. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for a collimator rotation error. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown. P-values < 0.05 signify a significant difference between median error curve ranges for a given gamma criterion.

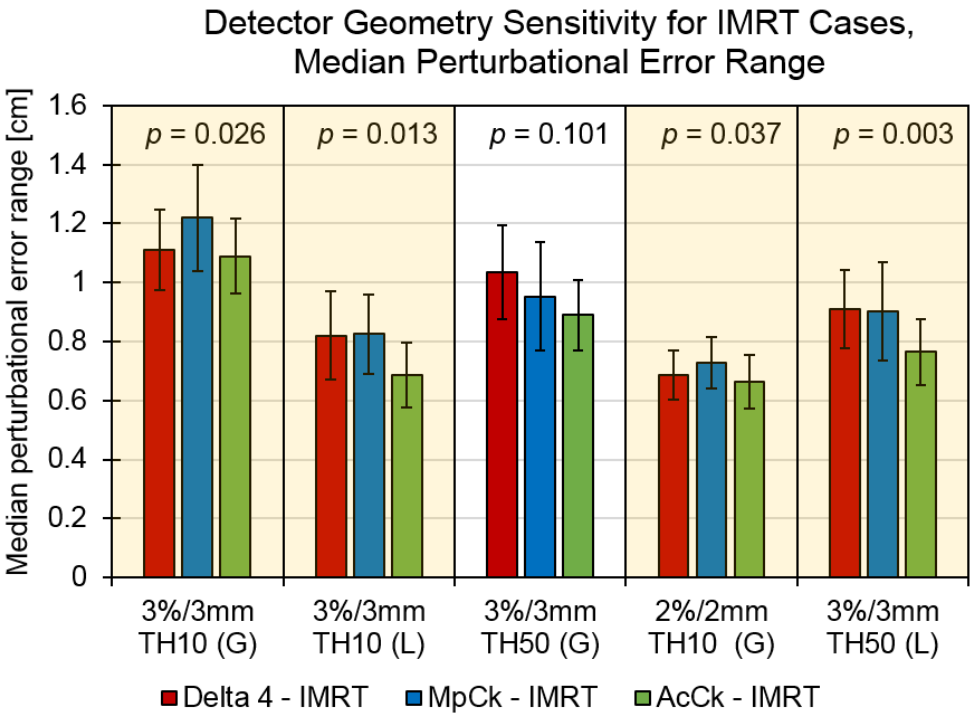

Figure 3-14. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for the MLC perturbational error. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown. P-values < 0.05 signify a significant difference between median error curve ranges for a given gamma criterion and are highlighted in yellow.

III.C.1.2 VMAT Cases

Error sensitivity differences between devices for the 20 VMAT cases in this study are shown in Figure 3-15, Figure 3-16, Figure 3-17, Figure 3-18, and Figure 3-19. The differences between detector geometries for VMAT cases were less pronounced compared to results for IMRT cases in the previous section. Similar to the results for IMRT cases, the ArcCHECK exhibited higher sensitivity to collimator rotation errors and MLC perturbational errors for VMAT cases compared to both Delta 4 and MapCHECK geometries. Likewise, for MU errors, the ArcCHECK performed much worse than the planar geometries across all gamma criteria. As before, a local normalization comparison aided in bringing sensitivity into similar ranges for the three devices.

However, for lagging leaf errors in VMAT cases (Figure 3-15) the Delta 4 had the highest error sensitivity across the studied gamma criteria, whereas the relative sensitivities for ArcCHECK and MapCHECK changed based on the gamma criterion, whereas for IMRT cases the ArcCHECK had considerably lower error sensitivity for the majority of gamma criteria. Figure 3-16 also shows that the three devices performed similarly for bank shift errors, with MapCHECK having slightly lower sensitivity for select gamma criteria, which is also different from the behavior for the IMRT cases.

These slightly different trends in error sensitivity for the different devices between IMRT and VMAT deliveries, especially for the ArcCHECK device, may again be due to the distribution of low dose points in the comparison on the ArcCHECK. Whereas IMRT cases on the ArcCHECK can have a wash of low dose points from the exit doses of IMRT beams, VMAT cases with full arcs typically do not have this low dose wash since the exit doses are masked and summed with entrance doses from opposing control points, resulting in a ring of high dose around the entire circumference of the device. It is thus likely that these low doses on the ArcCHECK explain why it is less sensitive than Delta 4 and MapCHECK for particular errors in IMRT cases, while these relative differences between devices for the same error types are less pronounced for VMAT cases. Regardless of delivery technique, we also note that the ArcCHECK is more sensitive to collimator errors and MLC perturbational errors than the planar-type devices. One possible explanation for this is that these errors may overlap and be obscured at the center of the plan for planar-type devices, whereas sampling the plan some distance away from the isocenter as the ArcCHECK does allows these errors to be more easily flagged with the gamma comparison.

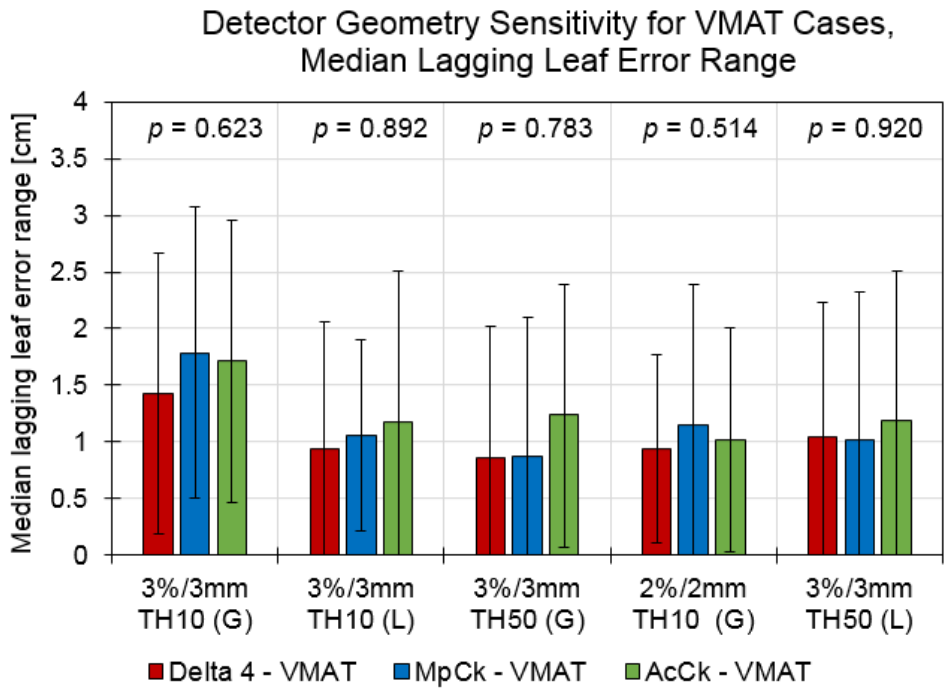

Figure 3-15. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for a lagging leaf error. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown. P-values < 0.05 signify a significant difference between median error curve ranges for a given gamma criterion and are highlighted in yellow.

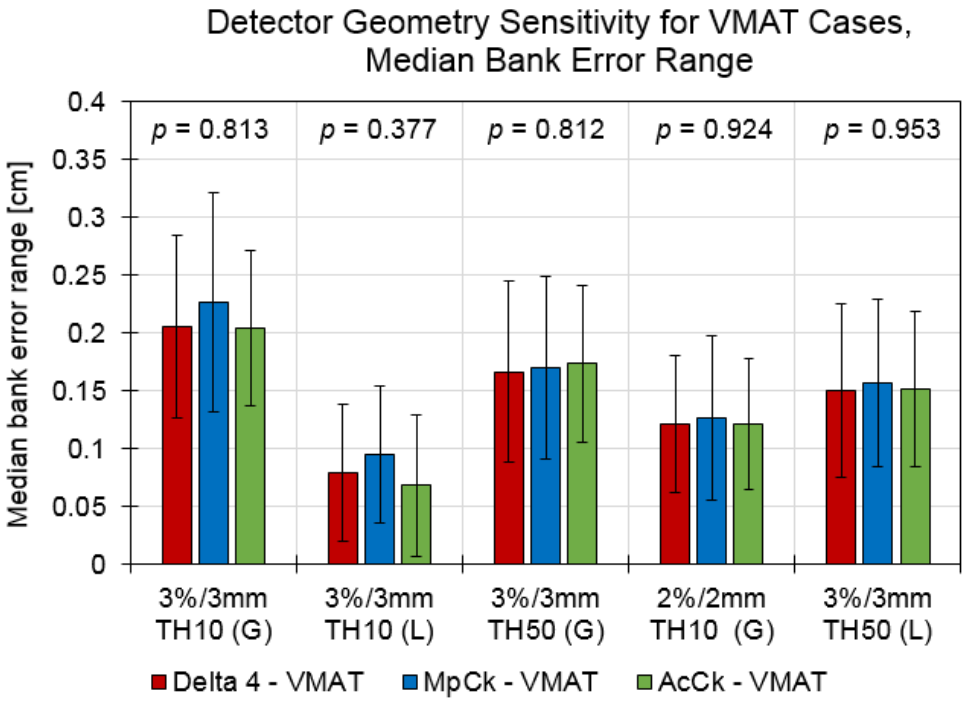

Figure 3-16. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for a bank shift error. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown. P-values < 0.05 signify a significant difference between median error curve ranges for a given gamma criterion and are highlighted in yellow.

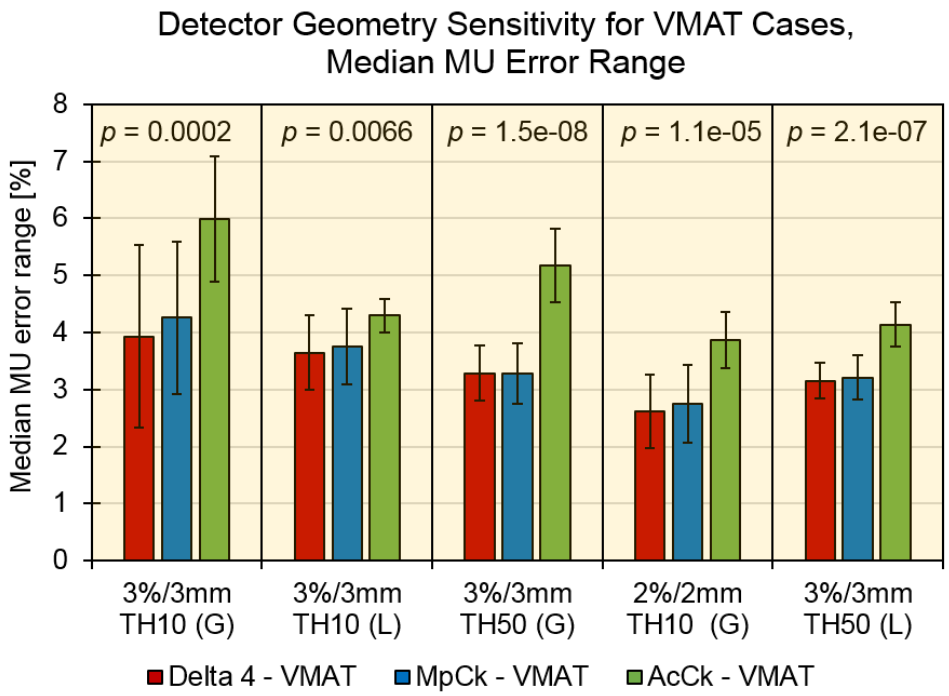

Figure 3-17. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for MU induced errors. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown. P-values < 0.05 signify a significant difference between median error curve ranges for a given gamma criterion and are highlighted in yellow.

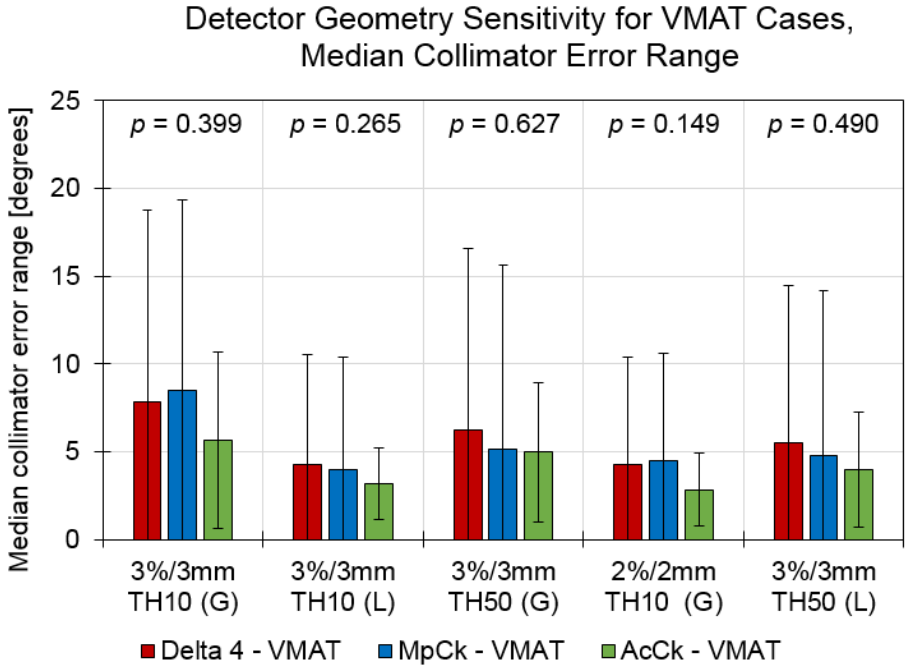

Figure 3-18. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for a collimator rotation error. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown. P-values < 0.05 signify a significant difference between median error curve ranges for a given gamma criterion and are highlighted in yellow.

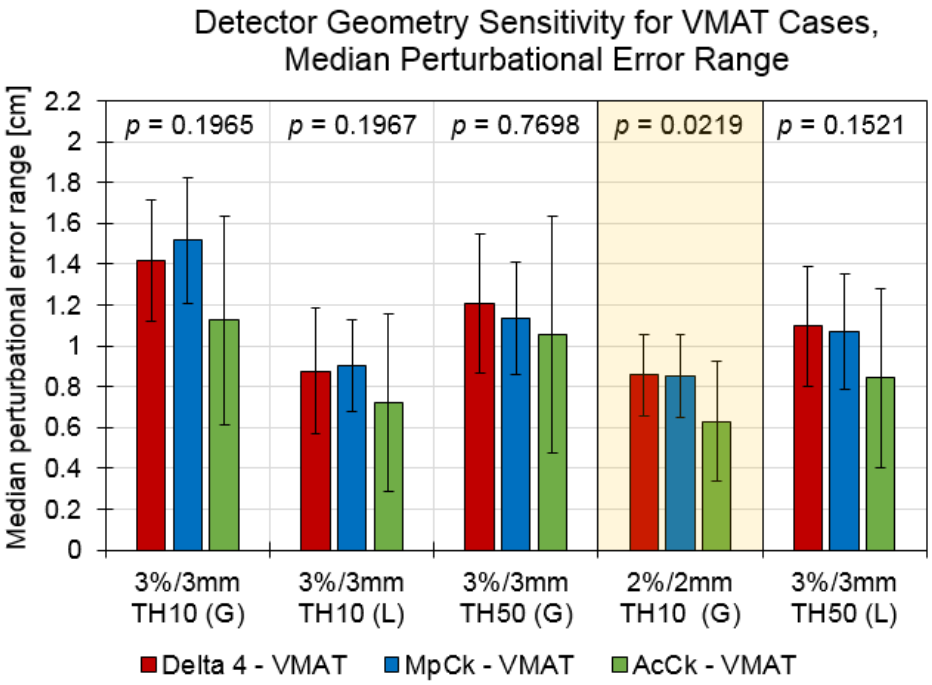

Figure 3-19. Median error curve ranges for Delta 4, MapCHECK, and ArcCHECK devices for the five listed gamma criteria for a MLC perturbational error. Error bars represent one standard deviation and p-values from Kruskal-Wallis analyses are shown. P-values < 0.05 signify a significant difference between median error curve ranges for a given gamma criterion and are highlighted in yellow.

III.C.2 Effects of detector spatial sampling on gamma sensitivity

In order to elucidate the effects of different spatial sampling scenarios for each detector device, calculation-only gamma comparisons and error curve ranges at 95% pixels passing were calculated for the IMRT case cohort for two scenarios:

- 1mm error-free calculations compared to 1mm error-induced calculations
- 1mm error-free calculations compared to error-induced calculations at the real spatial sampling of each device

Results are shown separately for error type and five different gamma criteria for the ArcCHECK (Figure 3-20), MapCHECK (Figure 3-21) and the Delta 4 (Figure 3-22).

For each induced error type on the ArcCHECK device, the differences in error sensitivity for the two spatial sampling scenarios are minimal and rarely significantly different, as shown in Figure

3-20. For select criteria and errors, such as 3%/3mm TH10 (G) criterion for the lagging leaf and perturbational MLC errors, increased spatial sampling does improve the error sensitivity marginally. However, these differences are minimal, and it is sometimes the case that increased spatial sampling actually reduces error sensitivity. While this may initially be counter-intuitive, it is likely that while more points exist in the comparison to aid in sensitivity, there are also more points in the comparison that can similarly inflate gamma passing rates, especially if the increased sampling causes a proportionally higher number of low dose points in the comparison than high dose points.

Results of the spatial sampling study for cases simulated on the MapCHECK geometry are shown in Figure 3-21. Similar to the ArcCHECK results, increasing the spatial sampling did not result in significant changes in the calculated error curve ranges. Conversely, for Delta 4 cases (shown in Figure 3-22), many differences between error curve ranges between the two spatial sampling scenarios were significantly different. For the collimator error, an increase in spatial sampling increased error sensitivity, but for the bank shift and perturbational errors, an increase in spatial sampling decreased error sensitivity. This is similar to the behavior seen in the ArcCHECK and suggests that increasing spatial sampling will not necessarily increase error sensitivity when using the gamma comparison to analyze calculated and delivered dose distributions, and also that spatial sampling is unlikely to be a driver of insensitivity in current IMRT QA comparisons.

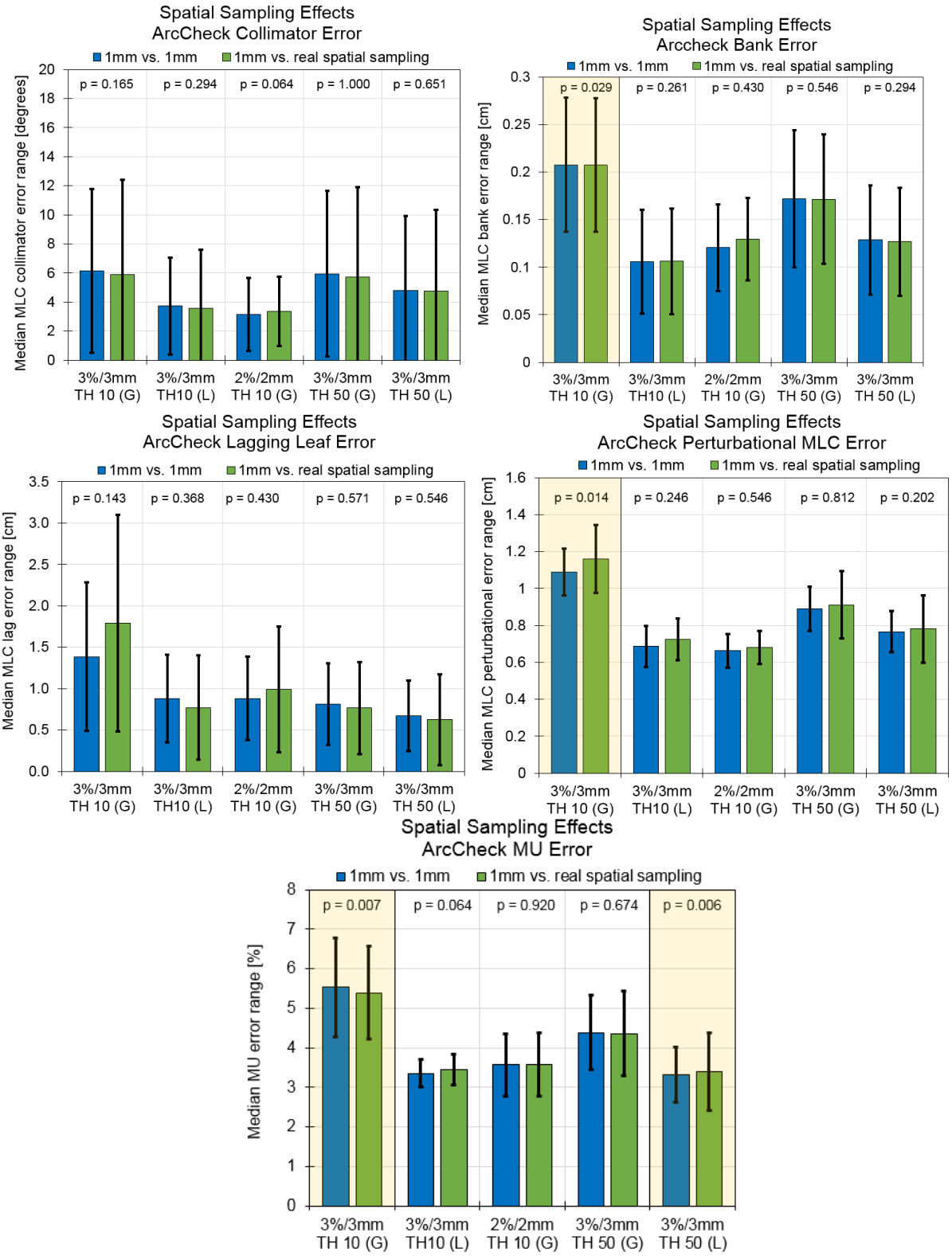

Figure 3-20. Differences in error curve ranges for the ArcCHECK when increasing spatial sampling from the real detector sampling to 1mm vs. 1mm comparisons for the different studied error types. P-values are shown for each induced error and gamma criterion, and differences that are statistically significant are highlighted in yellow.

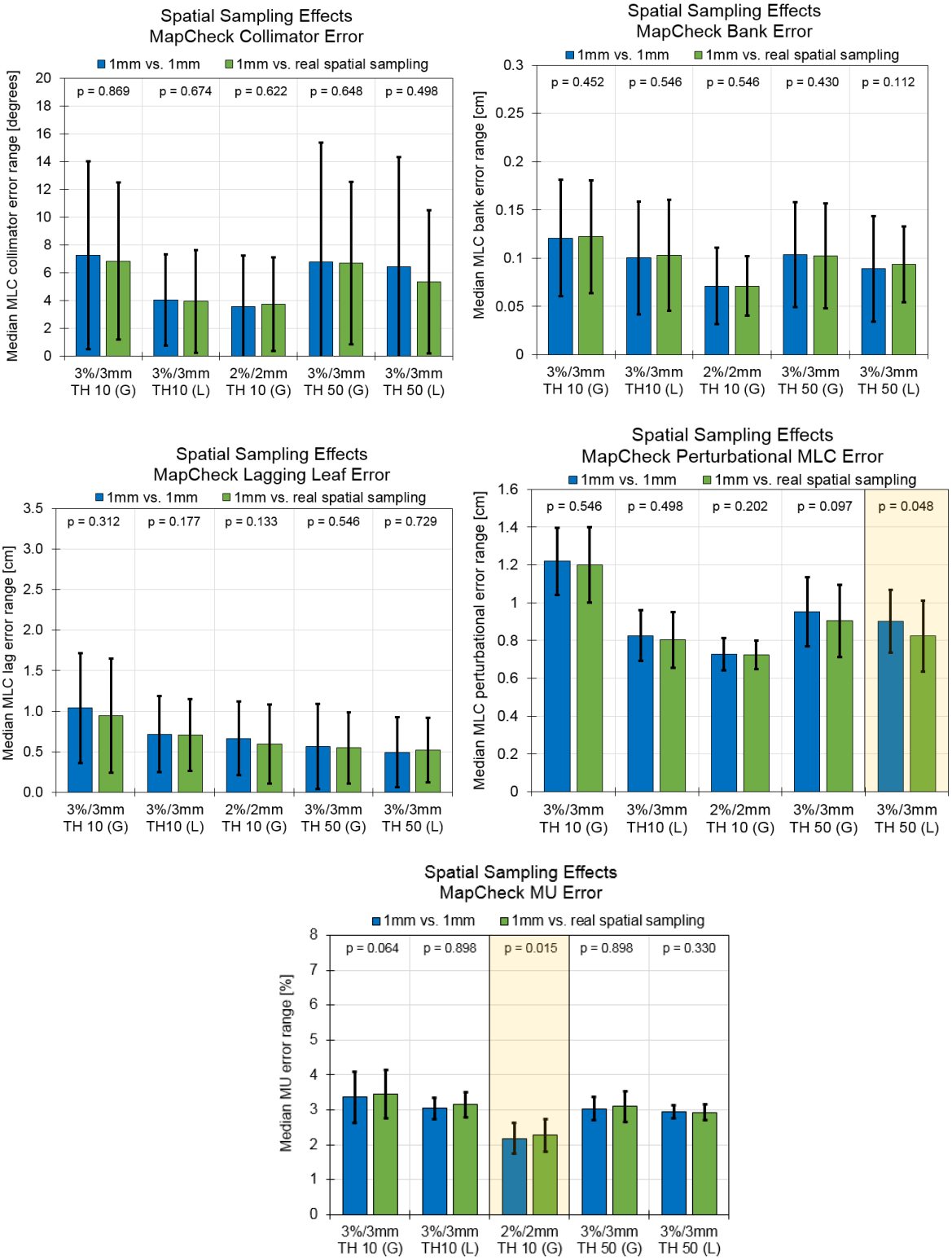

Figure 3-21. Differences in error curve ranges for the MapCHECK device when increasing spatial sampling from the real detector sampling to 1mm vs. 1mm comparisons for the different studied error types. P-values are shown for each induced error and gamma criterion, and differences that are statistically significant are highlighted in yellow.

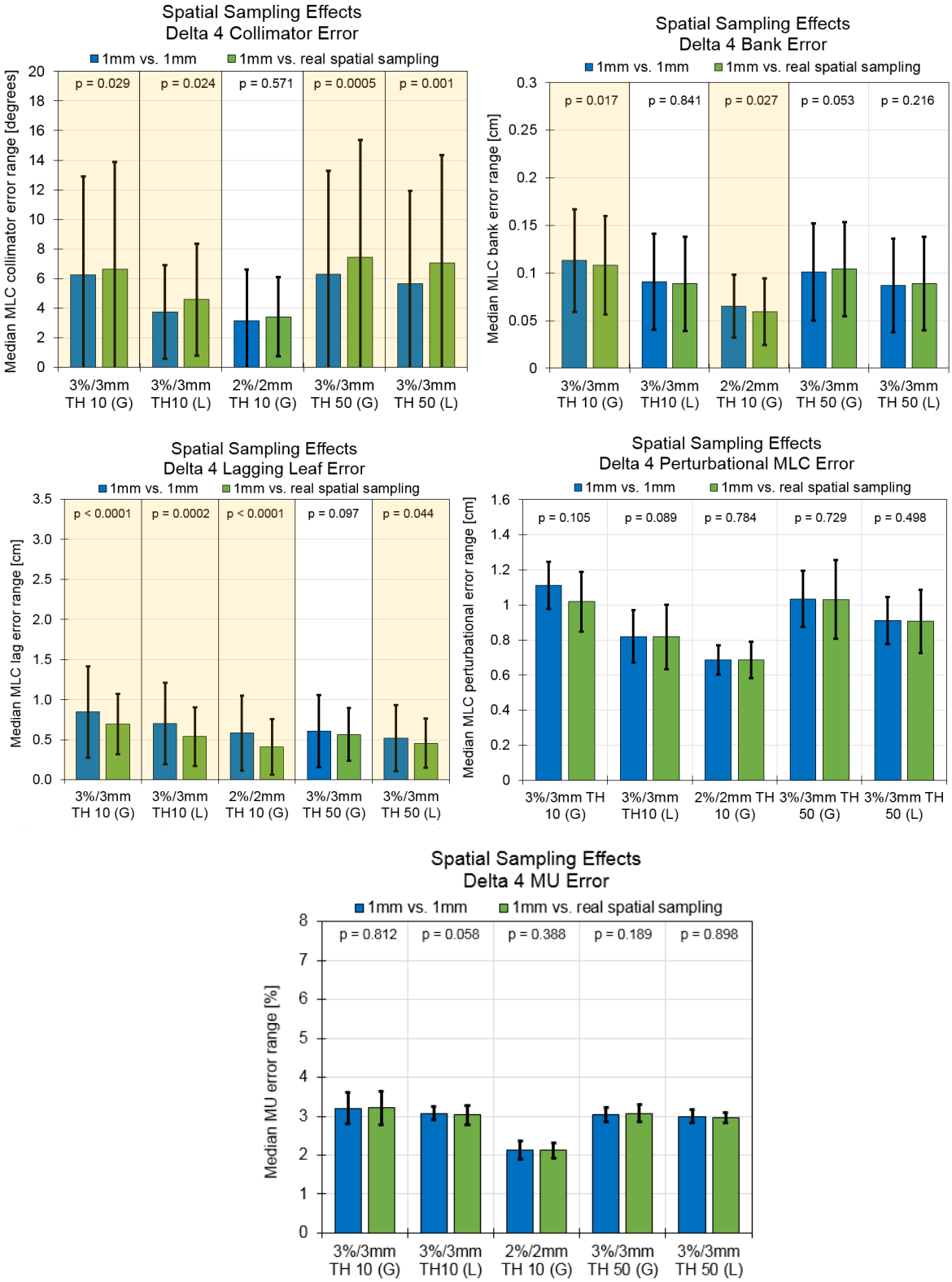

Figure 3-22. Differences in error curve ranges for the Delta 4 device when increasing spatial sampling from the real detector sampling to 1mm vs. 1mm comparisons for the different studied error types. P-values are shown for each induced error and gamma criterion, and differences that are statistically significant are highlighted in yellow.

Case-by-case error curve ranges for one of the induced error types (MLC perturbational) are shown for the three detector geometries in Figure 3-23, illustrating that for each of the studied cases, the error ranges are not systematically higher or lower for a particular spatial sampling scenario, but rather the behavior is case-specific. This case-specific behavior for the different sampling scenarios is likely due to several factors. First, the location of where the plan gradients fall on each device in relation to the measurement points is likely to factor into error sensitivity for the down-sampled comparisons shown here. One can imagine that shifting the isocenter of a given plan by even a few millimeters, especially in the ArcCHECK geometry, could potentially change error curve results.

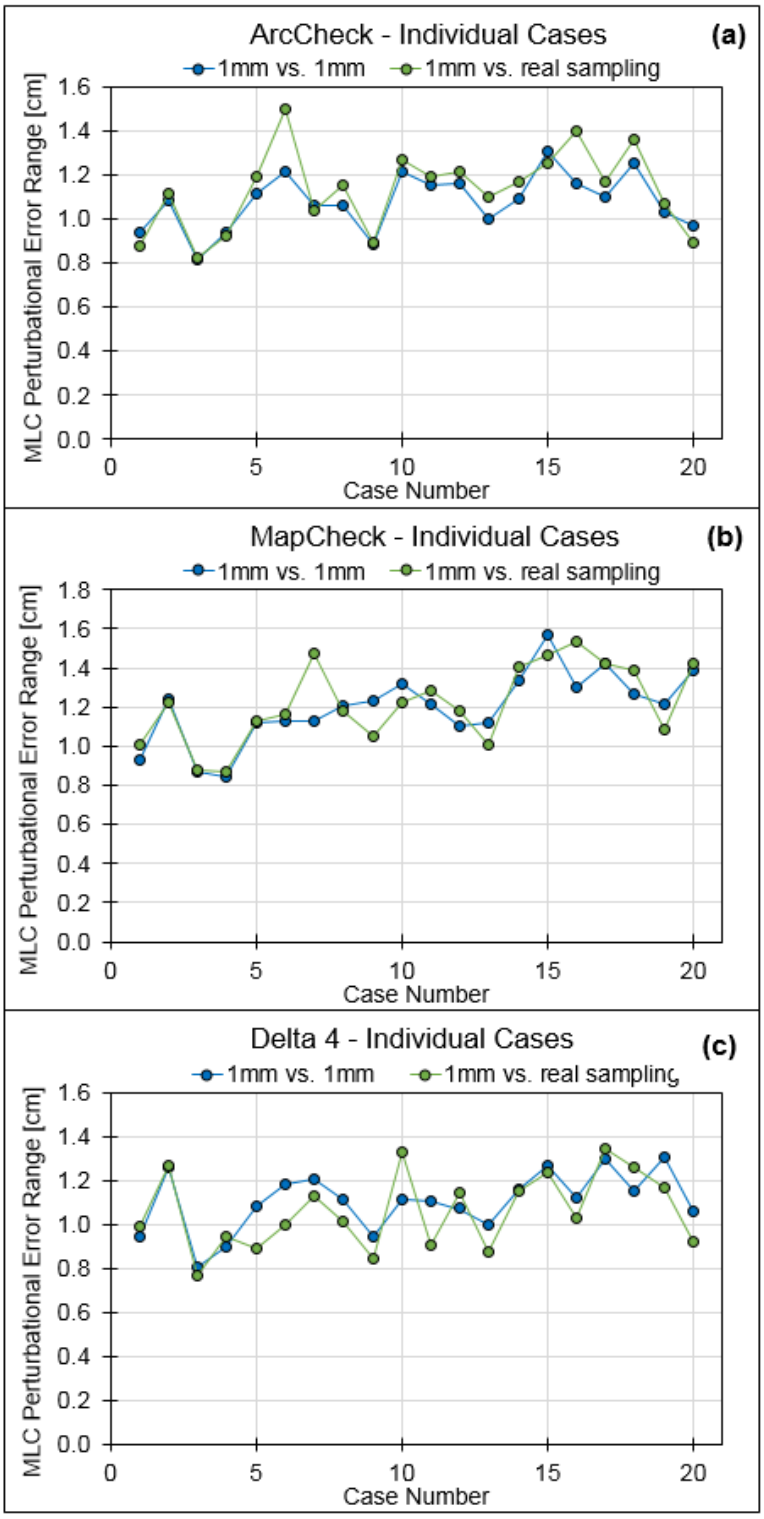

Figure 3-23. Case-by-case error curve ranges for the MLC perturbational error for higher and lower spatial sampling for (a) the ArcCHECK, (b) the MapCHECK, and (c) the Delta 4.

Secondly, while it is generally assumed that higher spatial sampling in IMRT QA comparisons will improve sensitivity to errors, this may be mitigated by the fact that increasing sampling can also increase the number of comparison points in low dose regions. Especially with the use of globally normalized gamma comparisons, this could lead to decreases in error sensitivity. Indeed, higher spatial sampling may give a better representation of the measured plan and may offer more comparison points that could aid in error sensitivity, this also means there are more points in less sensitive regions of the comparison to help mask errors as well. To illustrate this, an IMRT case was selected randomly from our cohort and the number of comparison points falling in high dose and low dose regions of the comparison was evaluated for the MapCHECK and ArcCHECK geometries. For the example case, IMRT 16, the dose maps are shown for the high spatial sampling and real spatial sampling scenarios on the MapCHECK in Figure 3-24 and on the ArcCHECK in Figure 3-25. To partition points into high and low dose, an arbitrary dose threshold of 50% was chosen and points falling above 50% maximum dose were counted as high-dose, whereas all points between 10% (the standard low-dose threshold value) and 50% were labeled as low dose points. The ratio of high dose to low dose points is shown in each figure, illustrating that for the higher-sampling scenario the proportion of high dose points to low dose points is either similar or can even be lower (in the case of the ArcCHECK) which may help explain why error sensitivity is also often similar or lower for the 1mm vs. 1mm comparisons.

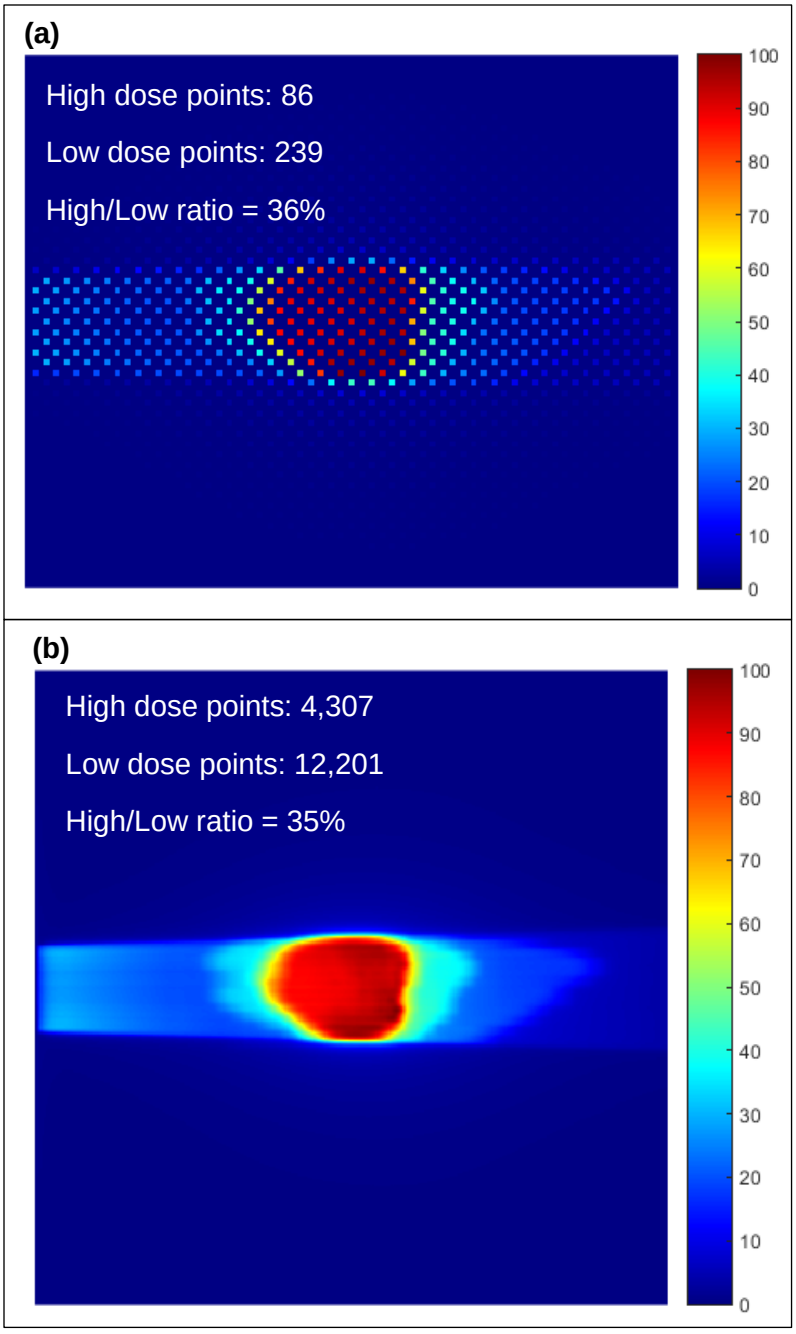

Figure 3-24. The normalized dose maps for IMRT 16 are shown on the MapCHECK device along with the number of high dose and low dose points for (a) the real-spatial sampling of the MapCHECK device and (b) for 1mm spatial sampling.

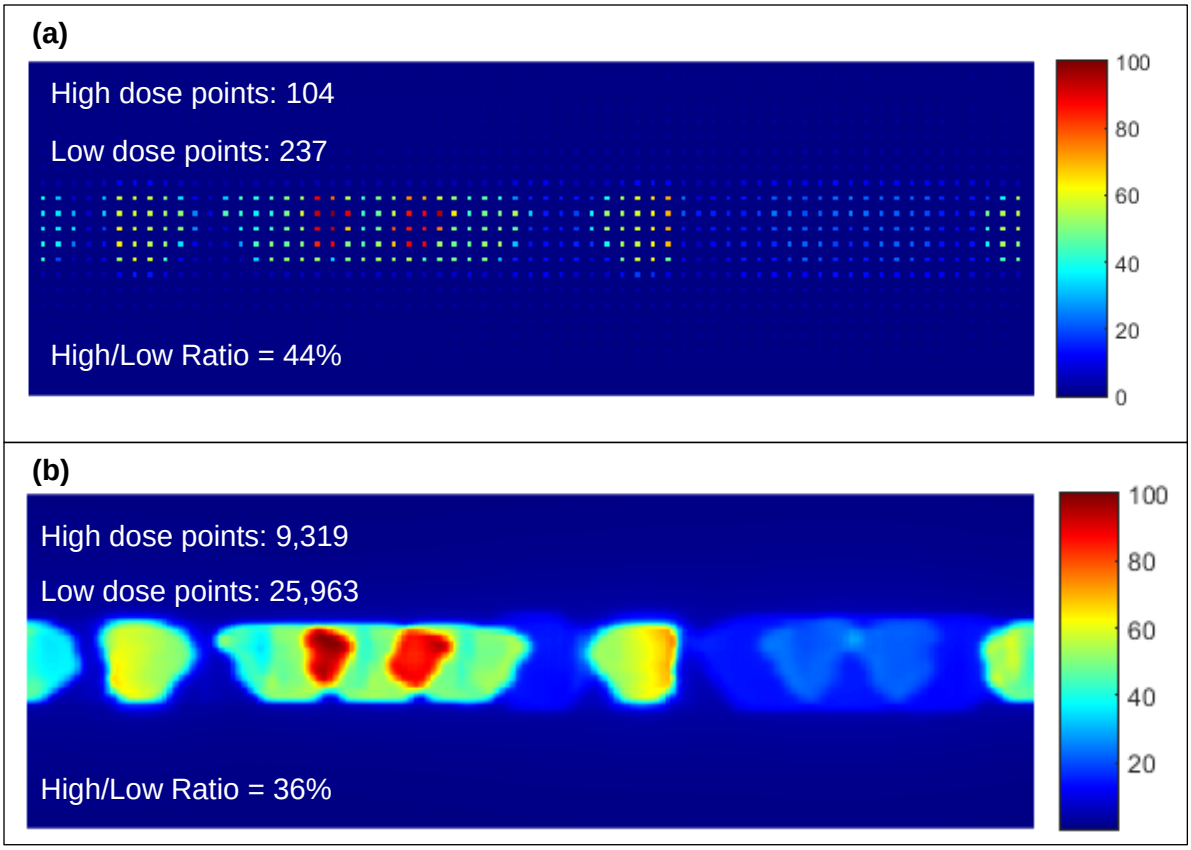

Figure 3-25. The normalized dose maps for IMRT 16 are shown on the ArcCHECK device along with the number of high dose and low dose points for (a) the real-spatial sampling of the ArcCHECK device and (b) for 1mm spatial sampling.

III.C.3. Effects of delivery technique on gamma sensitivity

Using the 1mm only comparisons, the error curve widths were compared between the IMRT and VMAT deliveries for each detector geometry. Figure 3-26, Figure 3-27, Figure 3-28, Figure 3-29, and Figure 3-30 show the median error curve ranges for the lagging leaf error, the bank shift error, the collimator rotation error, the MLC perturbational error, and the MU error, respectively. Though not all differences were statistically significant for a particular device or error type, there is a visible trend that shows gamma comparison error sensitivity for VMAT cases was worse than IMRT cases for MU errors and most MLC errors for all studied devices and almost every studied gamma criterion. Conversely, error sensitivity to collimator errors in VMAT cases was higher than in IMRT

cases for the majority of studied gamma criteria. It is also interesting to note that the sensitivity of IMRT and VMAT cases measured on the ArcCHECK was, in many cases quite similar, while this was not true for the MapCHECK and Delta 4 geometries. This suggests that perhaps the same gamma criteria that are used for IMRT may not be appropriate for VMAT cases, especially for planar-type devices, in order to achieve the same error sensitivity in patient-specific QA.

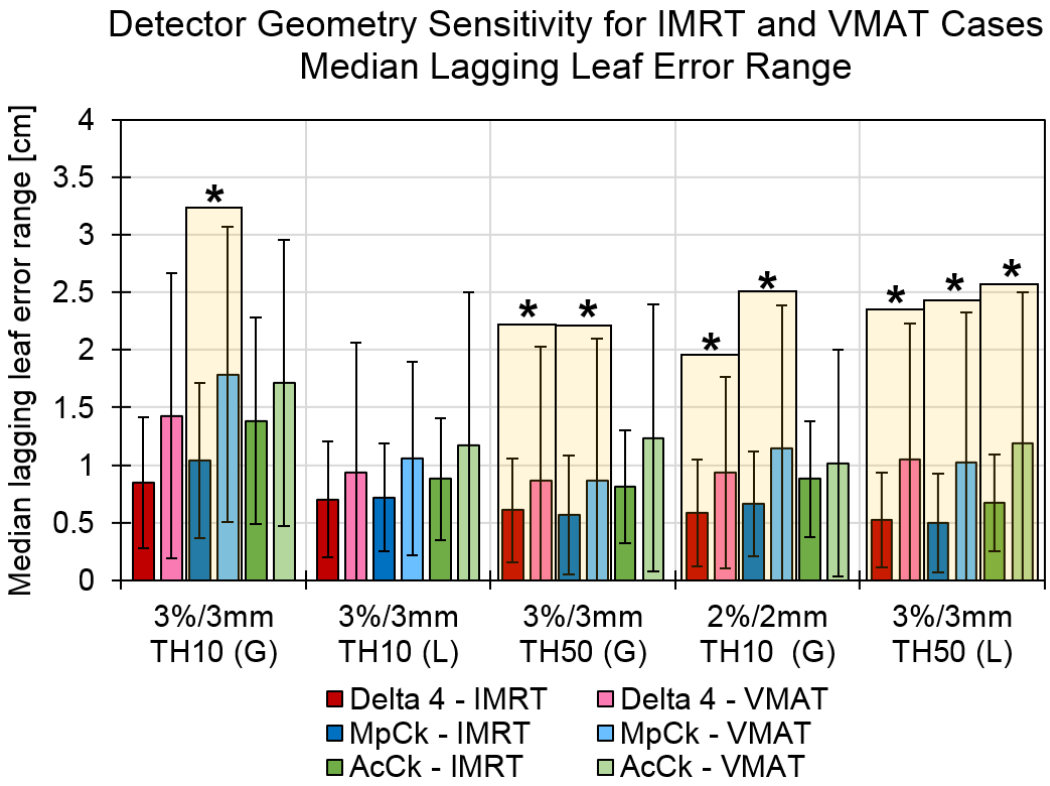

Figure 3-26. Median error curve range for the lagging leaf error for IMRT and VMAT cases for the 1mm vs. 1mm comparisons. Statistical significance is shown by cases highlighted in yellow.

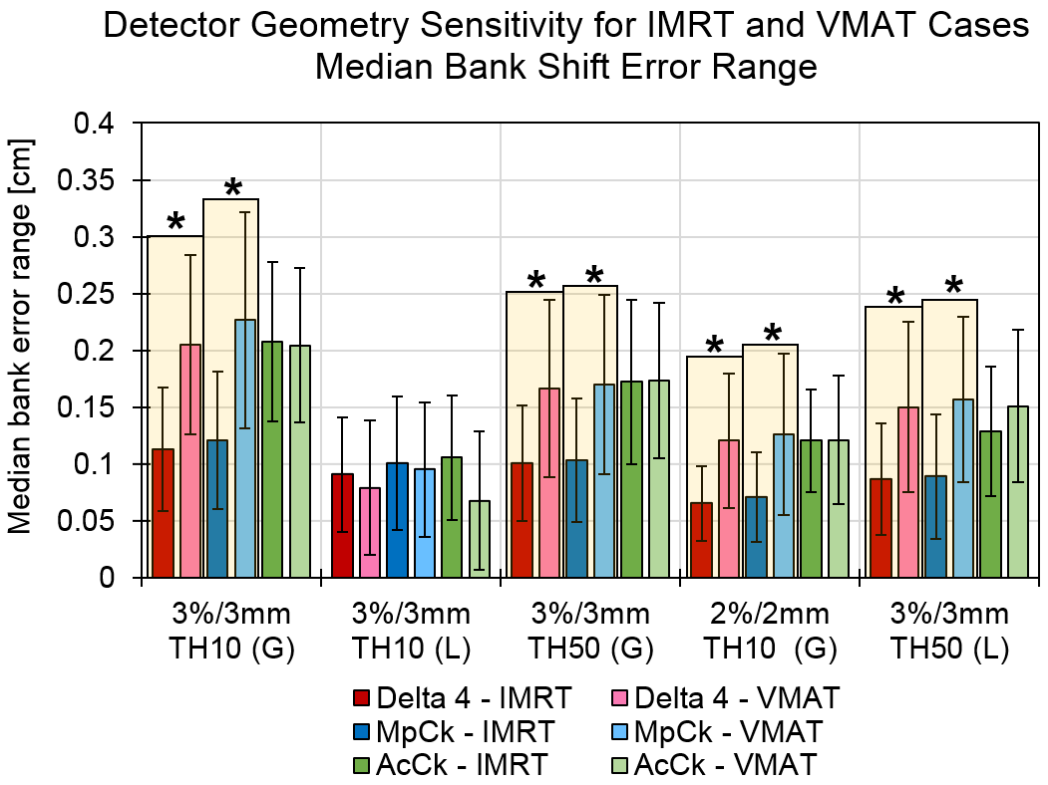

Figure 3-27. Median error curve range for the bank shift error for IMRT and VMAT cases for the 1mm vs. 1mm comparisons. Statistical significance is shown by cases highlighted in yellow.

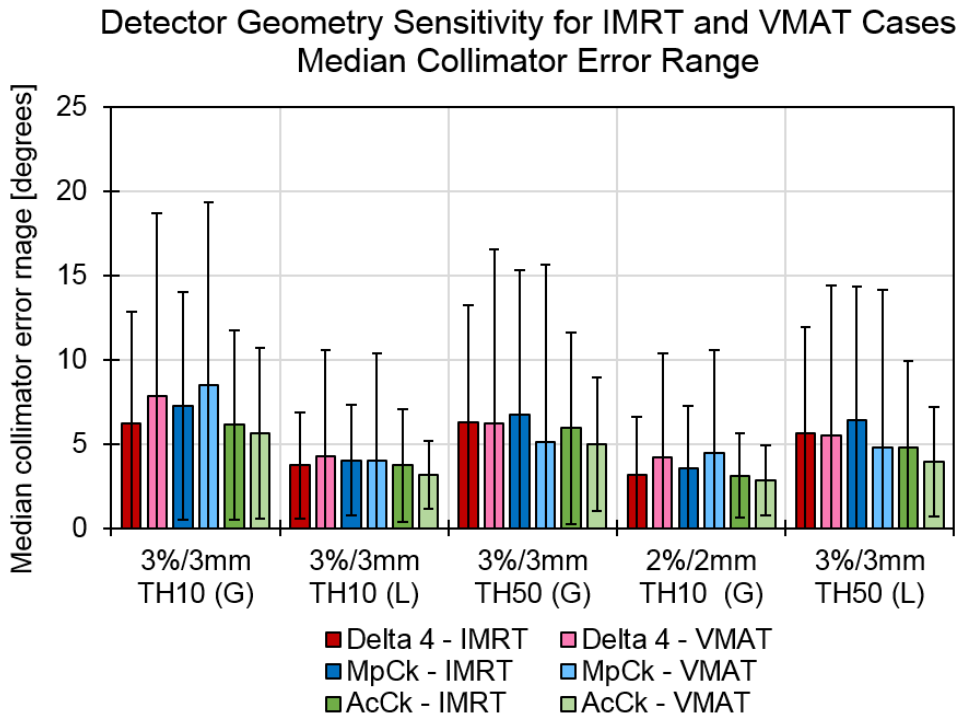

Figure 3-28. Median error curve range for the collimator rotation error for IMRT and VMAT cases for the 1mm vs. 1mm comparisons. Statistical significant is shown by cases highlighted in yellow.

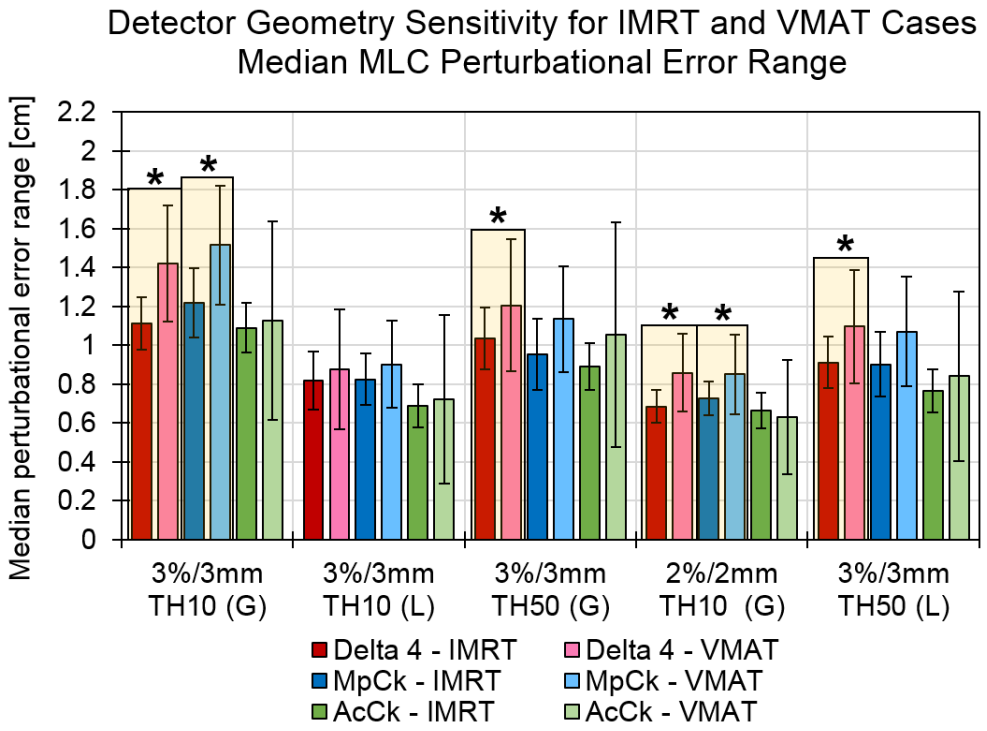

Figure 3-29. Median error curve range for the MLC perturbational error for IMRT and VMAT cases for the 1mm vs. 1mm comparisons. Statistical significance is shown by cases highlighted in yellow.

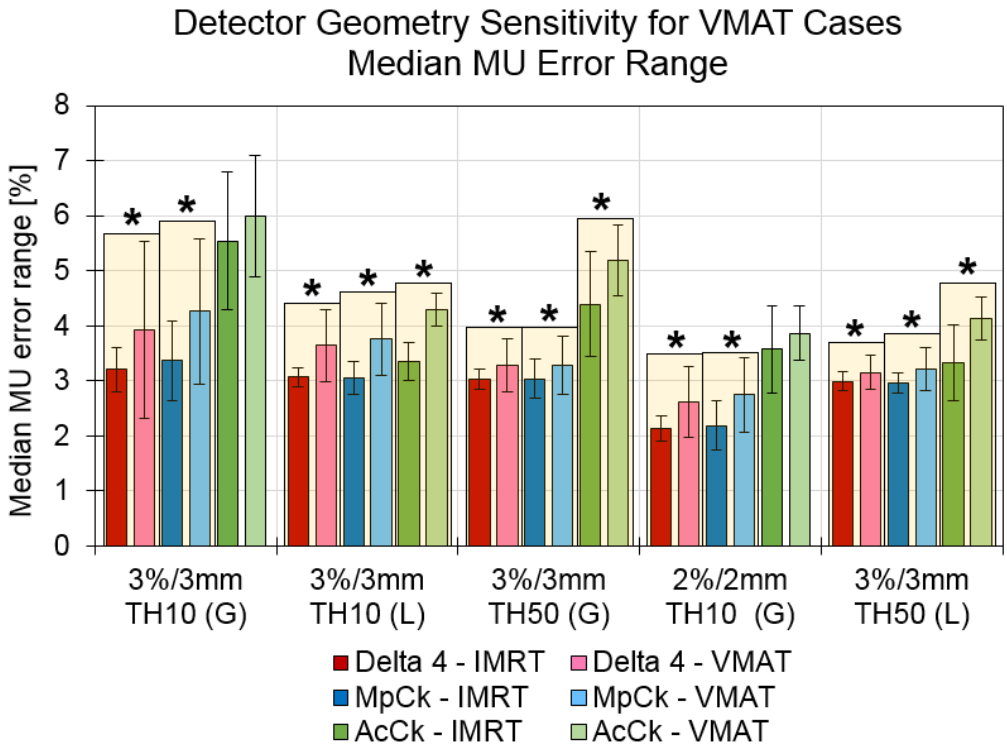

Figure 3-30. Median error curve range for the MU error for IMRT and VMAT cases for the 1mm vs. 1mm comparisons. Statistical significance is shown by cases highlighted in yellow.

One possible explanation for these differences in error curve ranges between VMAT and IMRT cohorts would be if the dosimetric effects from a particular error, say for example a 0.5 cm bank shift, had a lower dosimetric impact on VMAT plans than it did on IMRT plans, thus resulting in the higher error curve ranges for VMAT plans for MLC errors, and higher IMRT error curve ranges for collimator errors. This was investigated by pulling the PTV $D_{95\%}$, PTV $D_{99\%}$, and PTV mean for the MLC and collimator error-induced cases to determine if the distribution of DVH dose differences between VMAT and IMRT cohorts were noticeably different, thus resulting in different error curve ranges. Figure 3-31, Figure 3-32, and Figure 3-33 show that for the MLC errors, the VMAT cohort had similar or sometimes larger DVH dose differences than the IMRT plans, which fails to account for VMAT plans having larger error curve ranges for these error types. Similarly, where IMRT plans had larger error curve ranges for collimator errors, Figure 3-34 shows that IMRT plans also had a larger range of DVH dose differences from collimator errors than the VMAT cohort. Thus, it is unlikely that differences in error sensitivity between IMRT and VMAT cohorts is due to plans experiencing different dosimetric effects from the same induced errors. Although an effort was made to match plan complexity and field sizes between VMAT and IMRT cohorts, VMAT and IMRT deliveries are still fundamentally different and some of these differences in error sensitivity may simply arise from differences between IMRT and VMAT plans other than field size and plan complexity.

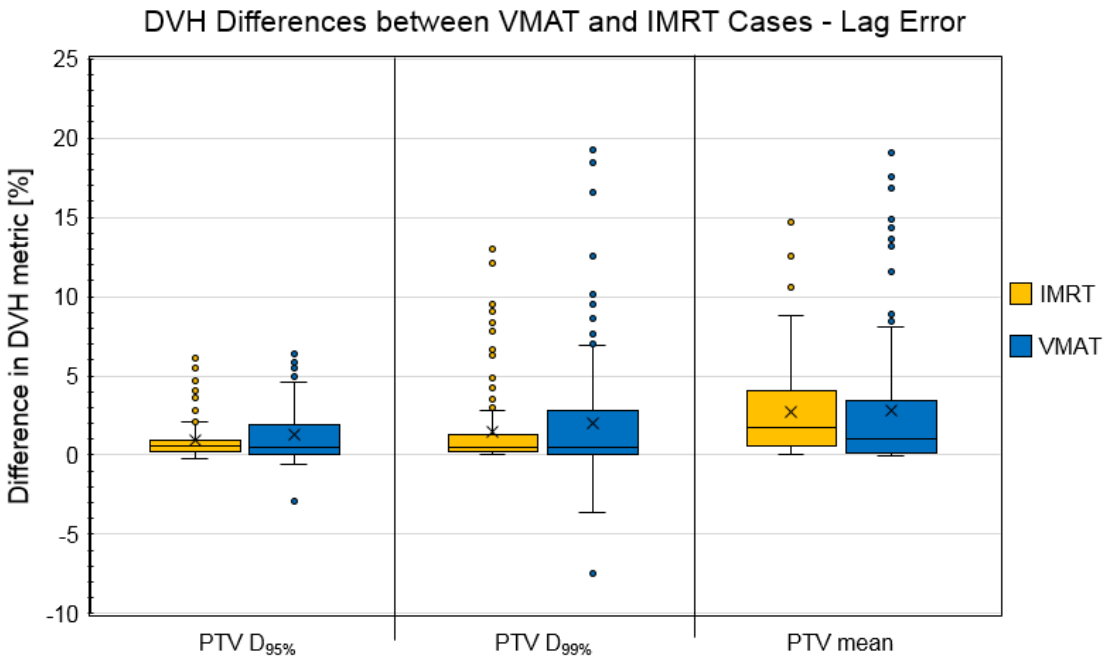

Figure 3-31. PTV DVH metric differences for all lagging leaf error-induced plans between VMAT and IMRT cohorts.

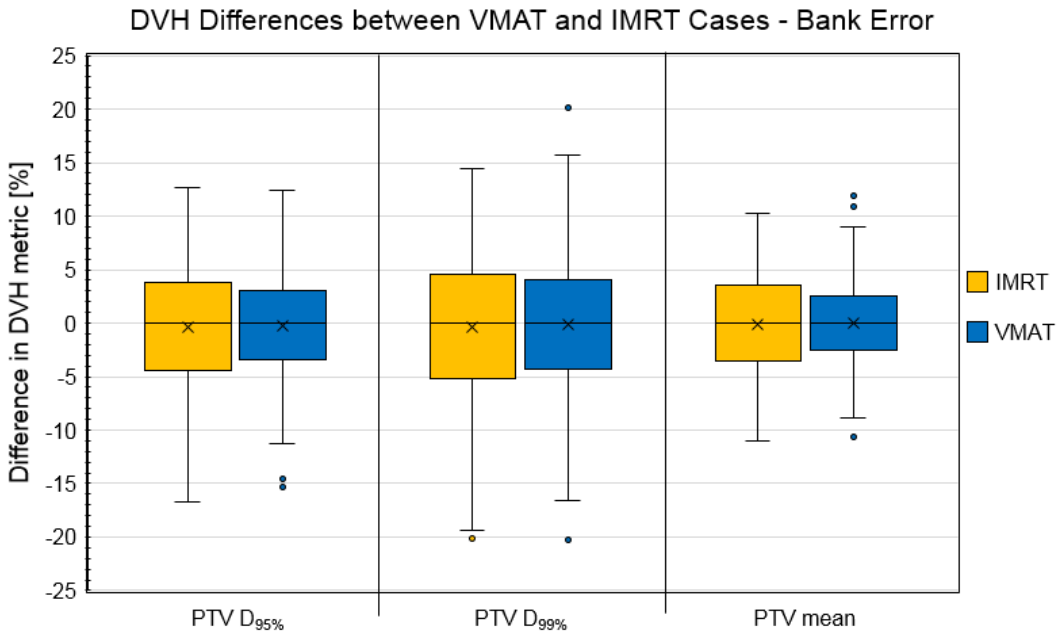

Figure 3-32. PTV DVH metric differences for all bank shift error-induced plans between VMAT and IMRT cohorts.

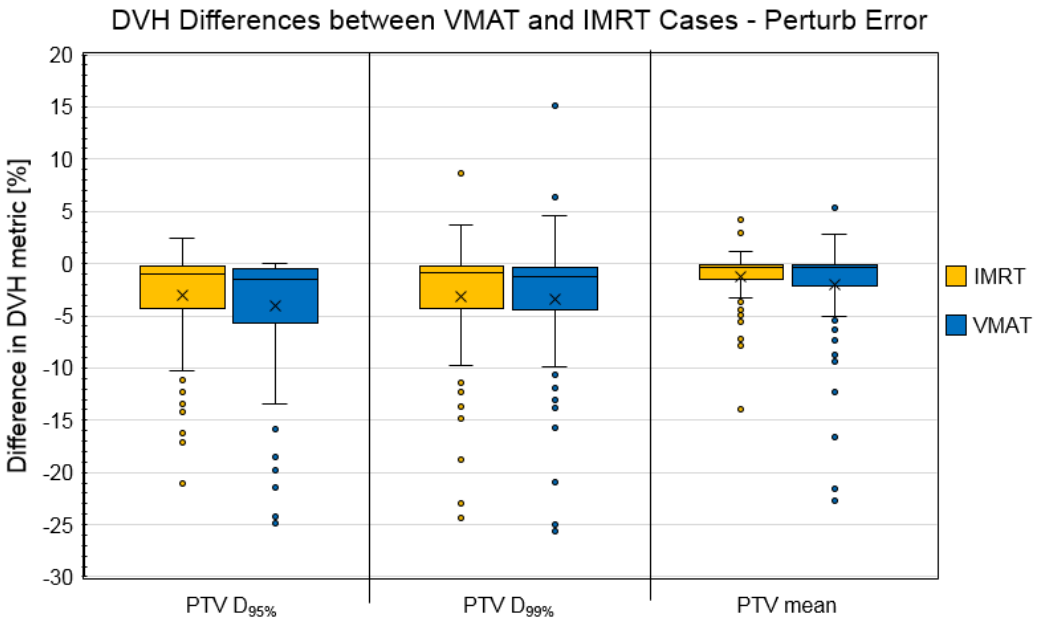

Figure 3-33. PTV DVH metric differences for all MLC perturbational error-induced plans between VMAT and IMRT cohorts.

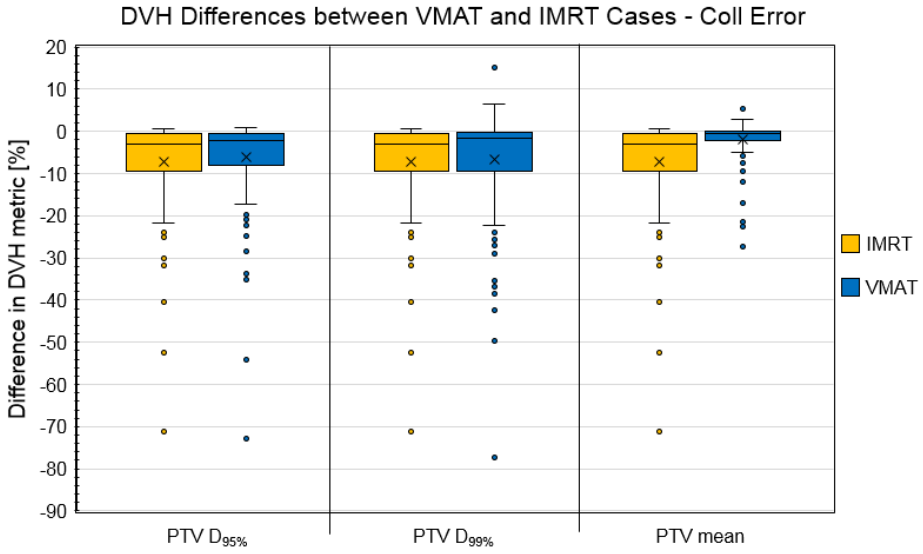

Figure 3-34. PTV DVH metric differences for all collimator rotation error-induced plans between VMAT and IMRT cohorts.

III.D. Discussion

This chapter has elucidated fundamental differences in error sensitivity between many different comparison scenarios that are present in clinical IMRT QA in a manner that removes limitations

that have not been addressed by previous publications. The calculation vs. calculation comparisons have allowed an objective evaluation of sensitivity differences caused by the very different measurement geometries and spatial sampling present in devices currently used across the field.

In general, most current IMRT QA practices apply a universal gamma criterion for passing/failing an IMRT QA comparison, regardless of measurement device or delivery type. While this streamlines clinical workflow, it may also inhibit the ability to flag meaningful errors in a patient's plan, especially depending on what type of error might be present in a plan combined with the specific measurement device. While current linear accelerator technology typically operates within very tight tolerances[40,41], there are scenarios where large errors could make it to treatment. Some examples of this are if a MLC trajectory file were to be corrupted[37], the wrong plan is sent to the machine, or a new version of software has some undiscovered bug in the code, just to name a few. When performing IMRT QA, the purpose of the test is to catch any errors that can cause large dosimetric differences. However, this work has shown that our range of sensitivity to large errors for a variety of different measurement devices is quite different and also dependent on the gamma criterion selected.

IMRT plans measured on the ArcCHECK had noticeably lower error sensitivities to lagging leaf, bank shift, and MU errors compared to Delta 4 and MapCHECK. This may be due to the fact that IMRT cases measured on the ArcCHECK geometry have many more low dose measurement points that fall above the 10% low dose threshold and can serve to increase gamma passing rates in a globally normalized setting in the presence of an induced error. This is also supported by the results in Chapter II which show that a higher low-dose threshold increases error sensitivity for the ArcCHECK device.

This work also shows that the most commonly used gamma criterion (3%/3mm TH10, global) is likely not sufficient to address these sensitivity differences between devices, and suggests either device- and/or delivery-specific gamma criteria, or the use of a different universal criterion. The data in this chapter showed that a 3%/3mm TH10 (L) gamma criteria may be useful as a universal criterion as it offered similar error sensitivity between the studied detector geometries. However, in clinical use, locally normalized comparisons may result in a higher number of IMRT QA failures, and especially if many of these failures were false positives, this could have a dramatic impact on the clinical workflow. Thus, separate passing criteria for individual devices is likely the most conservative and appropriate approach. Further study is required in order to appropriately set individual gamma criteria for different measurement devices.

Differences in error sensitivity for the spatial sampling scenarios for each studied device were not largely different and often showed inconsistent behavior, with some cases having higher error sensitivity with higher spatial sampling, and other cases having higher error sensitivity with sparser spatial sampling. We believe these differences in error sensitivity may thus be driven by the ratio of high-dose to low-dose points available in the gamma comparisons with different spatial sampling, where increased numbers of low dose points can decrease error sensitivity in globally normalized comparisons. Some cases have similar ratios of high- and low-dose points available for comparison between sparse and higher spatial sampling scenarios. Other cases have an even higher number of low dose points in up-sampled comparisons, which may explain why lower error sensitivity can be observed with increased sampling. This leads us to believe that the relatively sparse spatial sampling of detector arrays is not a main driving factor of gamma comparison insensitivity.

For the MLC and MU induced errors, VMAT plans had noticeably lower error sensitivity for the studied gamma criteria when compared to IMRT plans of similar complexity and field sizes.

Conversely, VMAT plans were more sensitive to collimator rotation errors than the IMRT plans in this study. This behavior was consistent across all three measurement devices suggesting there may be real sensitivity differences between these delivery techniques. To ensure that this was not simply due to the same error magnitudes having different dosimetric effects on the different delivery techniques, DVH PTV metrics were extracted from Eclipse for all error-induced plans in this study and plotted between VMAT and IMRT cohorts. While no dramatic differences were noted between PTV dosimetric differences for different delivery techniques, the plan deliveries are fundamentally different. It is important to note that the IMRT and VMAT plans were planned for individual patients (as opposed to an IMRT and VMAT plan for the same patient), which may also explain some of these differences. Regardless, this work suggests that IMRT QA comparisons are much less sensitive to MLC and MU errors in VMAT plans when compared to IMRT plans.

III.E. Conclusions

The work in this chapter made use of a large number of gamma comparisons and error curve ranges (approximately 4,500 error curve widths and roughly 36,000 gamma comparisons) to investigate sensitivity differences and limitations for a variety of different clinical IMRT QA measurement scenarios. Here, we have shown that the sensitivity of the gamma comparison does not largely hinge on the spatial sampling of the device. Noticeable differences in error sensitivity were observed for different detector geometries, but these differences were not consistent across all studied error types and gamma criteria. This suggests that a particular device can not necessarily be labeled “superior” but rather, that different gamma criteria may be required for different devices, particularly for the ArcCHECK array, given its very different geometry. Additionally, the low dose points that appear in ArcCHECK measurements when the array is not irradiated with a high dose around the entire circumference of the device may greatly influence

the gamma comparison sensitivity when using a global normalization of the dose differences. Finally, it is possible there are large error sensitivity differences between IMRT and VMAT cases which may ultimately suggest different gamma comparison settings for different delivery techniques to achieve similar error sensitivity in patient-specific QA.

A.III. Appendix A for Chapter III

Table A.3-1. Dunn's test results for significance between groups for the MLC bank shift error for IMRT cases. For the three gamma criteria that had significant differences between the error curve ranges for the three devices, the ArcCHECK had significantly larger median error ranges while error ranges between MapCHECK and Delta 4 were not significantly different. P-values < 0.025 are highlighted in yellow.

3%/3mm TH 10 (global)			3%/3mm TH 50 (local)			2%/2mm TH 10 (global)		
	D4	MpCk		D4	MpCk		D4	MpCk
MpCk	0.1999		MpCk	0.2400		MpCk	0.1862	
AcCk	0.0002*	0.0037*	AcCk	0.0006*	0.0058*	AcCk	0.0001*	0.0027*

Table A.3-2. Dunn's test results for significance between groups for the MU error for IMRT cases. In general, for gamma criteria that had significant differences between median error curve ranges, it was that the ArcCHECK was significantly more sensitive than the Delta 4 or MapCHECK. For 3%/3mm TH10 (Global) the MapCHECK was significantly less sensitive than the Delta 4 and ArcCHECK. P-values < 0.025 are highlighted in yellow.

3%/3mm TH 10 (global)			3%/3mm TH 10 (local)			3%/3mm TH 50 (global)			2%/2mm TH 10 (global)			3%/3mm TH 50 (local)		
	D4	MpCk		D4	MpCk		D4	MpCk		D4	MpCk		D4	MpCk
MpCk	0.1091		MpCk	0.4711		MpCk	0.3319		MpCk	0.1447		MpCk	0.4711	
AcCk	0.0000*	0.0000*	AcCk	0.0009*	0.0011*	AcCk	0.0000*	0.0000*	AcCk	0.0000*	0.0000*	AcCk	0.0001*	0.0001*

Table A.3-3. Dunn's test results for significance between groups for the MLC perturbational error for IMRT cases. In general, for gamma criteria that had significant differences between median error curve ranges, it was that the ArcCHECK was significantly more sensitive than the Delta 4 or MapCHECK. For 3%/3mm TH10 (Global) the MapCHECK was significantly less sensitive than the Delta 4 and ArcCHECK. P-values < 0.025 are highlighted in yellow.

3%/3mm TH 10 (global)			3%/3mm TH 10 (local)			3%/3mm TH 50 (Local)			2%/2mm TH 10 (global)		
	D4	MpCk		D4	MpCk		D4	MpCk		D4	MpCk
MpCk	0.0242*		MpCk	0.3964		MpCk	0.3519		MpCk	0.0278	
AcCk	0.2691	0.0048*	AcCk	0.0080*	0.0038*	AcCk	0.0010*	0.0033*	AcCk	0.2982	0.0073*

Table A.3-4. Dunn's test results for significance between detector devices for the MLC perturbational error for VMAT cases. The only criterion for which the error ranges between detector arrays reached statistical significance was for 2%/2mm TH 10 (global). In this instance, ArcCHECK was significantly more sensitive, though differences that did not reach significance are still appreciable for other criteria. P-values < 0.025 are highlighted in yellow.

2%/2mm TH 10 (global)		
	D4	MpCk
MpCk	0.3688	
AcCk	0.0136*	0.0055*

Table A.3-5 Dunn's test results for significance between groups for the MU error for VMAT cases. In general, for gamma criteria that had significant differences between median error curve ranges, it was that the ArcCHECK was significantly more sensitive than the Delta 4 or MapCHECK. For 3%/3mm TH10 (Global) the MapCHECK was significantly less sensitive than the Delta 4 and ArcCHECK. P-values < 0.025 are highlighted in yellow.

3%/3mm TH 10 (global)			3%/3mm TH 10 (local)			3%/3mm TH 50 (global)			2%/2mm TH 10 (global)			3%/3mm TH 50 (local)		
	D4	MpCk		D4	MpCk		D4	MpCk		D4	MpCk		D4	MpCk
MpCk	0.3930		MpCk	0.2904		MpCk	0.4964		MpCk	0.4603		MpCk	0.3552	
AcCk	0.0001*	0.0003*	AcCk	0.0014*	0.0076*	AcCk	0.0000*	0.0000*	AcCk	0.0000*	0.0000*	AcCk	0.0000*	0.0000*

References

1. Feygelman V, Zhang G, Stevens C, Nelms BE. Evaluation of a new VMAT QA device, or the “X” and “O” array geometries. *J Appl Clin Med Phys*. 2011;12(2):3346. doi:10.1120/jacmp.v12i2.3346.

2. Hussein M, Rowshanfarzad P, Ebert MA, Nisbet A, Clark CH. A comparison of the gamma index analysis in various commercial IMRT/VMAT QA systems. *Radiother Oncol*. 2013;109(3):370-376. doi:10.1016/j.radonc.2013.08.048.

3. McKenzie EM, Balter P a, Stingo FC, Jones J, Followill DS, Kry SF. Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans. *Med Phys*. 2014;41(12):121702. doi:10.1118/1.4899177.

4. Fredh A, Scherman JB, Fog LS, Munck af Rosenschöld P, Munck af Rosenschöld P. Patient QA systems for rotational radiation therapy: a comparative experimental study with intentional errors. *Med Phys*. 2013;40(3):031716. doi:10.1118/1.4788645.

5. Sjölin M, Edmund JM. Incorrect dosimetric leaf separation in IMRT and VMAT treatment planning: Clinical impact and correlation with pretreatment quality assurance. *Phys Medica*. 2016;32(7):918-925. doi:10.1016/j.ejmp.2016.06.012.

6. Saito M, Sano N, Shibata Y, et al. Comparison of MLC error sensitivity of various commercial devices for VMAT pre-treatment quality assurance. *J Appl Clin Med Phys*. 2018. doi:10.1002/acm2.12288.

7. Varatharaj C, Stathakis S, Ravikumar M, et al. SU-DD-A1-05: Comparison of Four Different Commercial Devices for Rapid Arc and Sliding Window IMRT QA. *Med Phys*. 2010;37(6):3088. doi:10.1118/1.3467986.

8. Vieillevigne L, Molinier J, Brun T, Ferrand R. Gamma index comparison of three VMAT QA systems and evaluation of their sensitivity to delivery errors. *Phys Medica*. 2015;31(7):720-725. doi:10.1016/j.ejmp.2015.05.016.

9. Masi L, Casamassima F, Doro R, Francescon P. Quality assurance of volumetric modulated arc therapy: Evaluation and comparison of different dosimetric systems. *Med Phys*. 2011;38(2):612-621. doi:10.1118/1.3533900.

10. Wen N, Zhao B, Kim J, et al. IMRT and RapidArc commissioning of a TrueBeam linear accelerator using TG-119 protocol cases. *J Appl Clin Med Phys*. 2014;15(5):4843. doi:10.1120/jacmp.v15i5.4843.

11. Mancuso GM, Fontenot JD, Gibbons JP, Parker BC. Comparison of action levels for patient-specific quality assurance of intensity modulated radiation therapy and volumetric modulated arc therapy treatments. *Med Phys*. 2012;39(7):4378. doi:10.1118/1.4729738.

12. Pawlicki T, Yoo S, Court LE, et al. Moving from IMRT QA measurements toward independent computer calculations using control charts. *Radiother Oncol*. 2008;89(3):330-337. doi:10.1016/j.radonc.2008.07.002.

13. Both S, Alecu IM, Stan AR, et al. A study to establish reasonable action limits for patient-specific quality assurance in intensity-modulated radiation therapy. *J Appl Clin Med Phys*.

2007;8(2):1-8.

14. Howell RM, Smith IPN, Jarrio CS. Establishing action levels for EPID-based QA for IMRT. *J Appl Clin Med Phys*. 2008;9(3):2721.

15. Basran PS, Woo MK. An analysis of tolerance levels in IMRT quality assurance procedures. *Med Phys*. 2008;35(6):2300. doi:10.1118/1.2919075.

16. Pawlicki T, Yoo S, Court LE, et al. Process control analysis of IMRT QA: implications for clinical trials. *Phys Med Biol*. 2008;53(18):5193-5205. doi:10.1088/0031-9155/53/18/023.

17. Vinall AJ, Williams AJ, Currie VE, Van Esch A, Huyskens D. Practical guidelines for routine intensity-modulated radiotherapy verification: Pre-treatment verification with portal dosimetry and treatment verification with in vivo dosimetry. *Br J Radiol*. 2010;83(995):949-957. doi:10.1259/bjr/31573847.

18. García-Vicente F, Fernández V, Bermúdez R, et al. Sensitivity of a helical diode array device to delivery errors in IMRT treatment and establishment of tolerance level for pretreatment QA. *J Appl Clin Med Phys*. 2012;13(1):3660. doi:10.1120/jacmp.v13i1.3660.

19. Sanghangthum T, Suriyapee S, Srisatit S, Pawlicki T. Statistical process control analysis for patient-specific IMRT and VMAT QA. *J Radiat Res*. 2013;54(3):546-552. doi:10.1093/jrr/rrs112.

20. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. *Med Phys*. 2018;45(4):e53-e83. doi:10.1002/mp.12810.

21. Kry SF, Molineu A, Kerns JR, et al. Institutional patient-specific IMRT QA does not predict unacceptable plan delivery. *Int J Radiat Oncol Biol Phys*. 2014;90(5):1195-1201. doi:10.1016/j.ijrobp.2014.08.334.

22. Kruse JJ. On the insensitivity of single field planar dosimetry to IMRT inaccuracies. *Med Phys*. 2010;37(6):2516. doi:10.1118/1.3425781.

23. Nelms BE, Zhen H, Tomé W a. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. *Med Phys*. 2011;38(2):1037-1044. doi:10.1118/1.3544657.

24. Zhen H, Nelms BE, Tomé W a. Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA. *Med Phys*. 2011;38(10):5477. doi:10.1118/1.3633904.

25. Stasi M, Bresciani S, Miranti A, Maggio A, Sapino V, Gabriele P. Pretreatment patient-specific IMRT quality assurance : A correlation study. *Med Phys*. 2012;39(12):7626-7634. doi:10.1118/1.4767763.

26. Yan G, Liu C, Simon TA, Peng LC, Fox C, Li JG. On the sensitivity of patient-specific IMRT QA to MLC positioning errors. *J Appl Clin Med Phys*. 2009;10(1):120-128. doi:10.1118/1.2961902.

27. Templeton AK, Chu JCH, Turian J V. The sensitivity of ArcCHECK-based gamma analysis to manufactured errors in helical tomotherapy radiation delivery. *J Appl Clin Med Phys*. 2015;16(1):4814. doi:10.1120/jacmp.v16i1.4814.

28. Jin X, Yan H, Han C, Zhou Y, Yi J, Xie C. Correlation between gamma index passing rate and clinical dosimetric difference for pre-treatment 2D and 3D volumetric modulated arc therapy dosimetric verification. *Br J Radiol*. 2015;88(1047). doi:10.1259/bjr.20140577.

29. Nelms B, Jarry G, Chan M, Hampton C, Watanabe Y, Feygelman V. Real-world examples of sensitivity failures of the 3%/3mm pass rate metric and published action levels when used in IMRT/VMAT system commissioning. *J Phys Conf Ser*. 2013;444:012086. doi:10.1088/1742-6596/444/1/012086.

30. Chan. Using a Novel Dose QA Tool to Quantify the Impact of Systematic Errors Otherwise Undetected by Conventional QA Methods: Clinical Head and Neck Case Studies. *Technol Cancer Res Treat*. 2013;13(1):1-11. doi:10.7785/tcrt.2012.500353.

31. McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. *Med Phys*. 2010;37(2):505-515. doi:10.1118/1.3276775.

32. Masi L, Doro R, Favuzza V, Cipressi S, Livi L. Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy. *Med Phys*. 2013;40(7):071718. doi:10.1118/1.4810969.

33. Jursinic P a, Nelms BE. A 2-D diode array and analysis software for verification of intensity modulated radiation therapy delivery. *Med Phys*. 2003;30(5):870-879. doi:10.1118/1.1567831.

34. Sadagopan R, Bencomo JA, Martin RL, Nilsson G, Matzen T, Balter PA. Characterization and clinical evaluation of a novel IMRT quality assurance system. *J Appl Clin Med Phys*. 2009;10(2):2928. http://www.ncbi.nlm.nih.gov/pubmed/19458595.

35. Létourneau D, Publicover J, Kozelka J, Moseley DJ, Jaffray D a. Novel dosimetric phantom for quality assurance of volumetric modulated arc therapy. *Med Phys*. 2009;36(5):1813-1821. doi:10.1118/1.3117563.

36. Yan G, Lu B, Kozelka J, Liu C, Li JG. Calibration of a novel four-dimensional diode array. *Med Phys*. 2010;37(1):108-115. doi:10.1118/1.3266769.

37. Mans a, Wendling M, McDermott LN, et al. Catching errors with in vivo EPID dosimetry. *Med Phys*. 2010;37(6):2638-2644. doi:10.1118/1.3397807.

38. Deasy JO, Blanco AI, Clark VH. CERR: A computational environment for radiotherapy research. *Med Phys*. 2003;30(5):979-985. doi:10.1118/1.1568978.

39. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys*. 1998;25(5):656. doi:10.1118/1.598248.

40. Stell AM, Li JG, Zeidan O a, Dempsey JF. An extensive log-file analysis of step-and-shoot intensity modulated radiation therapy segment delivery errors. *Med Phys*. 2004;31(6):1593-1602. doi:10.1118/1.1751011.

41. Agnew A, Agnew CE, Grattan MWD, Hounsell AR, McGarry CK. Monitoring daily MLC positional errors using trajectory log files and EPID measurements for IMRT and VMAT deliveries. *Phys Med Biol*. 2014;59(9):N49-N63. doi:10.1088/0031-9155/59/9/N49.

Chapter IV.

The effect of dose gradients on gamma comparison sensitivity

IV.A. Motivation

As the sensitivity of the gamma comparison for IMRT QA has been called into question, recent publications have proposed tighter gamma criteria for clinical IMRT QA[1–4]. While methods other than gamma exist for comparing these complex distributions[5–14], none of these are in routine clinical use. The gamma comparison thus remains the standard for analyzing IMRT QA comparisons[2,15] even though many publications have shown that commonly used criteria can be insensitive to large, clinically meaningful errors[16–25].

While spatial sampling is sometimes blamed for gamma comparison insensitivity, the results from Chapter III showed that increasing the spatial sampling of three different detector arrays – ArcCHECK, MapCHECK, and Delta 4 – did not greatly improve error sensitivity, and in some cases even decreased error sensitivity in conjunction with the gamma comparison. It is unlikely that the driving factor behind low error sensitivity in current IMRT QA comparisons is driven by the measurement device geometry or device spatial sampling.

To further investigate causes for gamma insensitivity, known errors were induced in plan calculations and compared to calculations without induced errors. This was first performed using a manufactured test field, and gamma value maps were studied in order to understand if particular features of the dose distribution prevent comparison points from failing gamma in the presence of induced errors of increasing magnitude. This was extended to real patient plans to understand the relationship between dose distribution features and gamma failures with the use of different detector geometries and gamma criteria.

IV.B. Induced errors in test fields

MU and random MLC perturbational errors were induced in plan calculations for simple test fields. Two test fields, shown in Figure 4-1, is flat rectangular field with intensity spikes and dips in the center of the field. The purpose of this is to introduce simple dose gradients in the center of the field. This test fields were artificially manufactured by creating an optimal fluence file and importing this file into an IMRT test plan in Eclipse. Error-free and error-induced dose calculations were performed using Eclipse AAA (1mm grid size) on the ArcCHECK phantom.

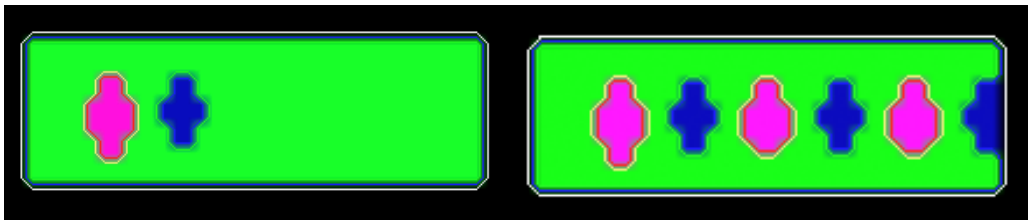

Figure 4-1. The optimal fluence maps for test fields with an intensity spikes and dips. Rectangular field size is 10 cm x 3 cm.

MU scaling and MLC random perturbational errors (previously described in Chapter II) were introduced in the DICOM RT plan files in MATLAB and re-imported into Eclipse for dose calculation. All doses were calculated using Eclipse AAA with a grid size of 1mm on the ArcCHECK phantom. Comparisons for this portion of the work were performed between error-free calculation and error-induced calculations in SNC Patient (v.6.7.3, Sun Nuclear Corporation, Melbourne, FL).

Many different magnitudes of error-induced calculations were studied to observe the progression of failing points as induced errors increased in magnitude from small to large. For the simpler test plan with one intensity spike and dip this is illustrated for induced MU errors in Figure 4-2 with three different induced MU errors shown here. The comparison points failing a gamma criterion of 2%/2mm TH 10 (global) are also shown. As a note, while gamma values are always positive, SNC Patient shows gamma failures as red or blue points based on the sign of the dose difference.

Blue points represent those which fail gamma because the measurement is colder than the calculation, and conversely, red points are those that fail the gamma comparison because the measurement is hotter than the calculation. In this case, the “measurement” is the error-induced calculation.

It is visible in Figure 4-2 that as error magnitudes increase, the areas with lower gradients fail the gamma comparison first, and by the largest magnitude errors rings around the highest gradient regions of the field without any gamma failures can be visualized. Similar results are visualized for the random MLC perturbational error, shown induced on the test plan with multiple intensity spikes and dips in Figure 4-3. While this error is more complex in nature, it is still clear that as error magnitudes increase, gamma failures generally do not occur first along the highest dose gradients in the plan. This is further visualized with line profiles shown in Figure 4-4 for an induced 10% MU error and in Figure 4-5 for a ±0.8 cm random MLC perturbational error. These comparisons with the test fields led to the hypothesis that there may be a relationship between steep dose gradients and gamma failures in real IMRT QA comparisons.

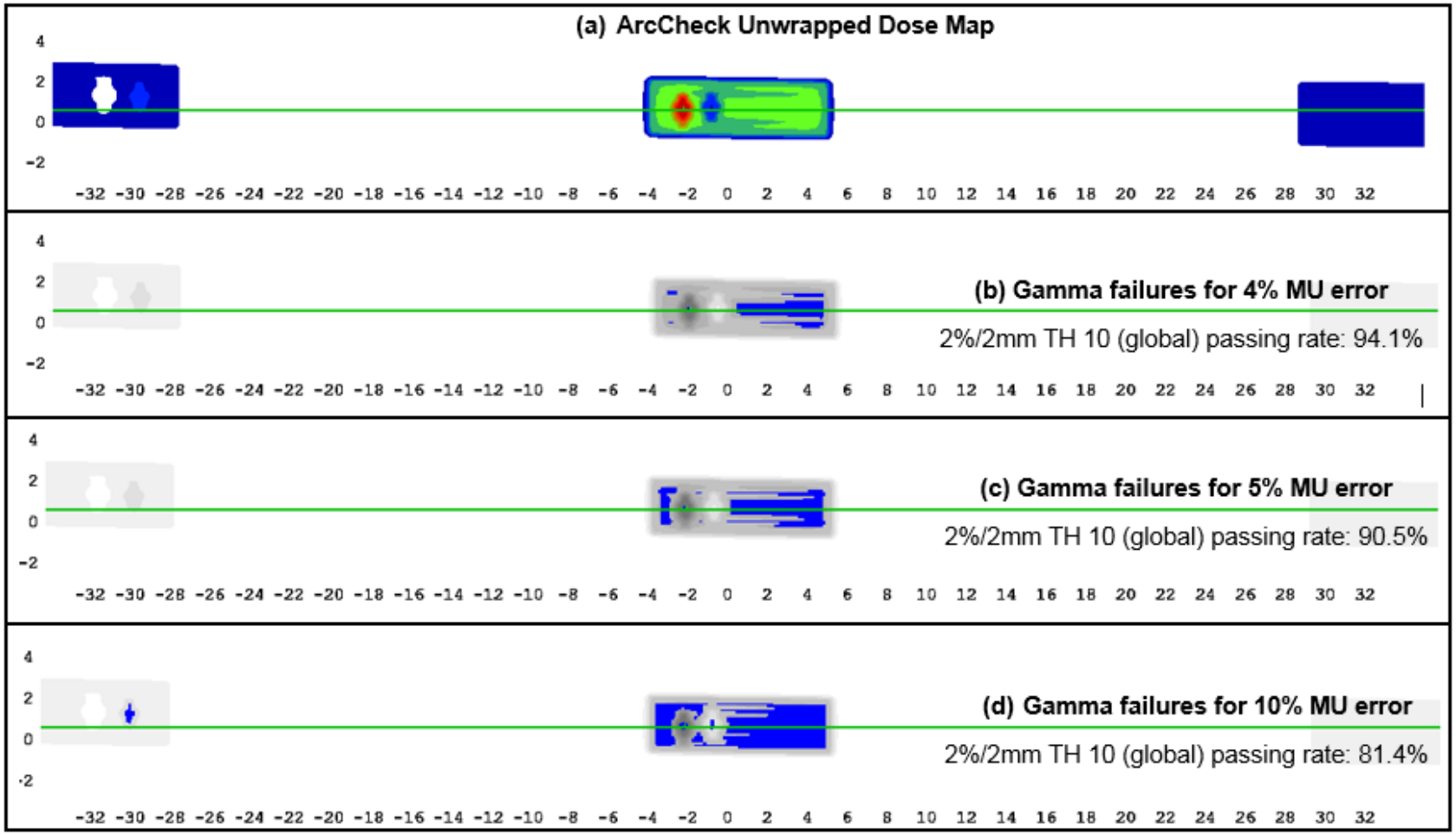

Figure 4-2. For one of the test fields, the (a) ArcCHECK unwrapped dose map is shown. Different magnitudes of MU errors were induced from (b) 4%, (c) 5%, to (d) a 10% scaling of MU values. The gamma passing rates are also shown. Blue points represent gamma failures where the error-induced calculation is colder than the error-free calculation.

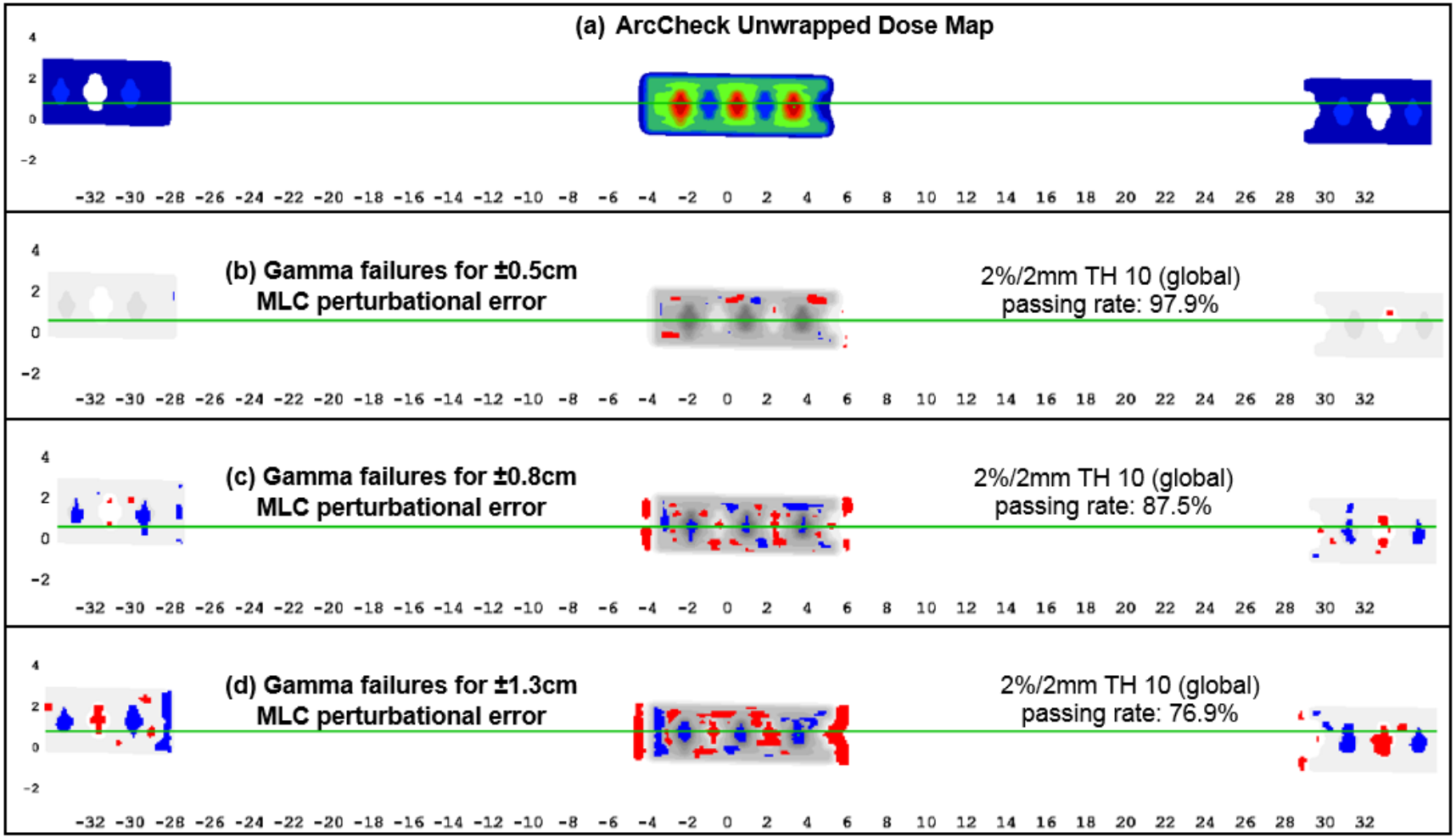

Figure 4-3. For one of the test fields, the (a) ArcCHECK unwrapped dose map is shown. Different magnitudes of MLC random perturbational errors were induced from (b) ±0.5cm, (c) ±0.8cm, to (d) ±1.3cm. The gamma passing rates are also shown. Blue points represent gamma failures where the error-induced calculation is colder than the error-free calculation. Similarly, red points indicate gamma failures where the error-induced calculation is hotter than the error-free calculation.

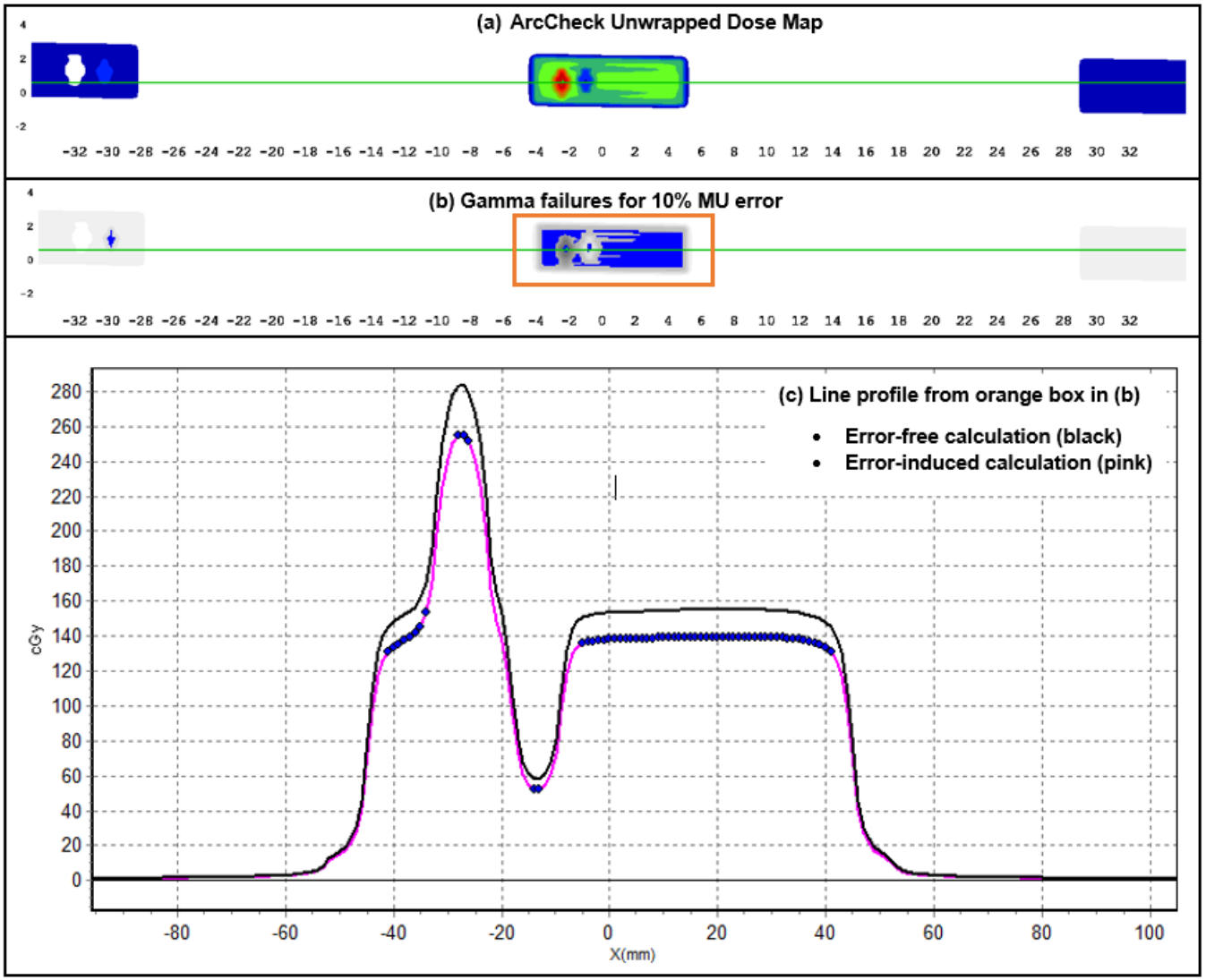

Figure 4-4. *(a)* Unwrapped ArcCHECK dose for the test field with multiple spikes and dips. *(b)* A 10% MU error was induced in one of the plan calculations and the resulting gamma failures at 2%/2mm TH 10 (G) are shown in blue and red. *(c)* A line profile through the center of the field – green line in (a) and (b) – is shown here for the area highlighted in the orange box in (b).

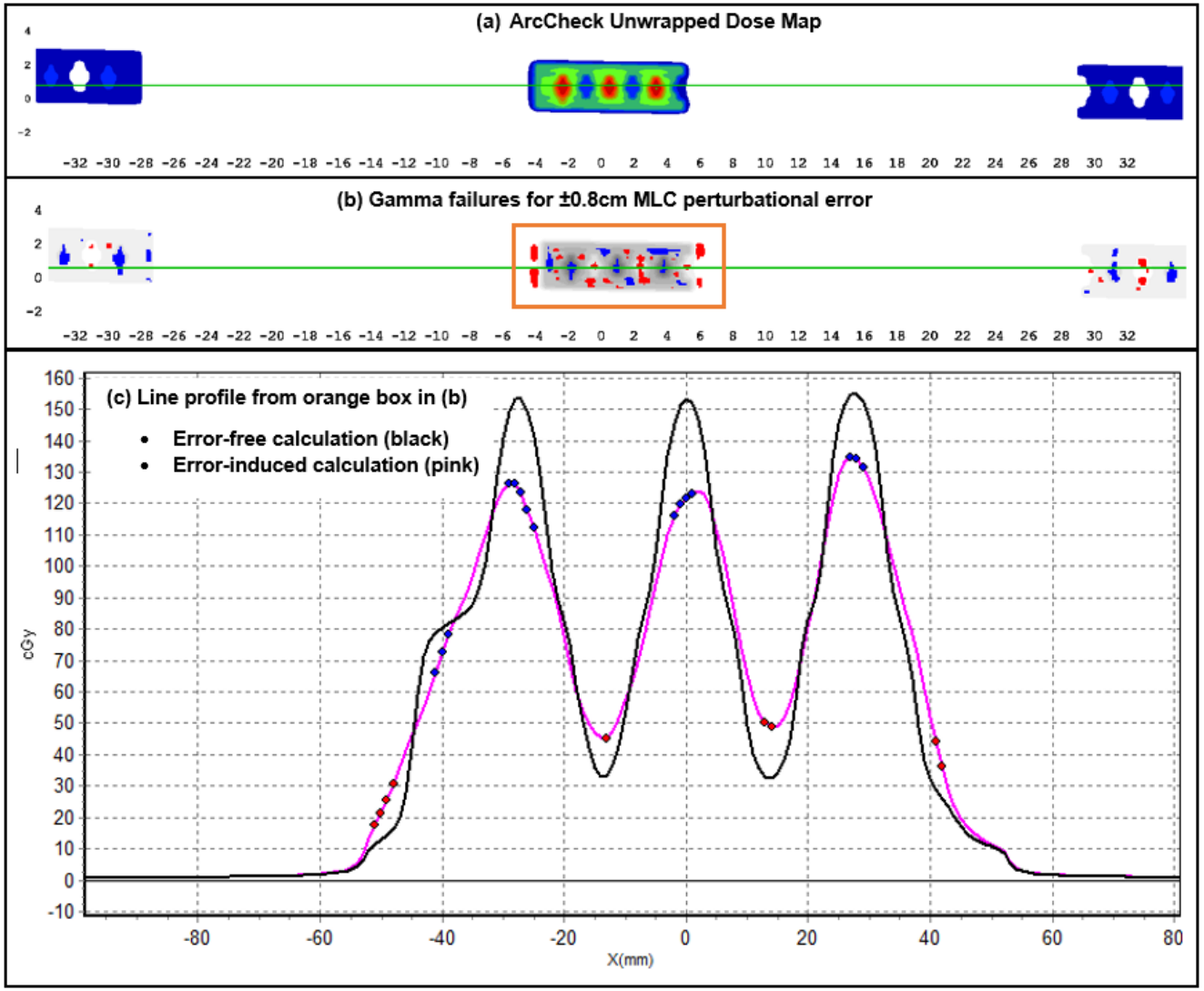

Figure 4-5. (*a)* Unwrapped ArcCHECK dose for the test field with multiple spikes and dips. *(b)* A ±0.8 cm MLC random perturbational error was induced in one of the plan calculations and the resulting gamma failures at 2%/2mm TH 10 (G) are shown in blue and red. *(c)* A line profile through the center of the field – green line in (a) and (b) – is shown here for the area highlighted in the orange box in (b).

IV.C. Dose gradients vs. gamma value maps

IV.C.1. Methods

Based on the preliminary results from the test fields, the relationship between plan dose gradients and gamma values in the presence of induced errors was investigated. The 3D dose gradient maps for the 20 IMRT and 20 VMAT cases from Chapter III were calculated in MATLAB. Based on work published by Moran *et al.*[5] the 3D gradient is defined as

$$G_i = Gradient = \sqrt{\sum \left(\frac{\Delta d_{ij}}{\Delta x_{ij}}\right)^2} \qquad [Eq. 4.1]$$

Where:

- G_i is the generalized gradient at location *i*
- Δd_{ij} is the difference between the doses at grid point *i* and its six nearest neighbors, *j*, (for the 3D gradient)
- Δx_{ij} is the distance between grid points *i* and *j*, which in this case is 1 mm for the calculated dose maps

The gradient maps were calculated in 3D using the dose matrices previously extracted for use in calculating gamma sensitivities in Chapter III. Gradient maps were calculated for the 20 IMRT and the 20 VMAT cases for the ArcCHECK, MapCHECK, and Delta 4 geometries between 1mm error-free calculations and 1mm error-induced calculations. Using patient plans on the various devices, the 3D gamma *value* maps (as opposed to the points with gamma greater than unity) were compared to investigate the relationship between gamma values and plan dose gradients.

Typically, the gamma comparison is used in a binary pass/fail setting, where gamma values greater than unity are labeled as “failing”. However, for this portion of the work the calculated gamma values were computed and stored for the IMRT and VMAT cases studied for the five different induced errors previously described in Chapter III (MU, collimator rotation, bank shift, lagging leaf, and MLC perturbational errors). These gamma value maps were calculated using

the in-house MATLAB gamma comparison software also developed in Chapter III. This software was created to allow calculation vs. calculation gamma comparisons for the ArcCHECK, MapCHECK, and the Delta 4. For some of the devices used here, gamma value maps cannot be obtained from the commercial vendor software for the device. Thus, these comparisons were only possible using the MATLAB in-house developed gamma comparison code for this project.

IV.C.2. Results

Gradient and gamma value maps were first compared for the simplest error type – MU errors – for both locally and globally normalized gamma comparisons on the three detector devices studied. Examples of the relationship between dose, gradient, and gamma values are shown for the same case on the different detector geometries in Figure 4-6 (ArcCHECK), Figure 4-7 (MapCHECK), Figure 4-8 (Delta 4 vertical board) and Figure 4-9 (Delta 4 horizontal board). For each example, the gamma value maps are shown for both local and global gamma comparisons using 3% dose and 3mm distance criteria. The same case is shown here in order to illustrate the fact that the relationship between gradient maps and gamma values appears independent of detector geometry as well as dose difference normalization technique. A more complex IMRT case with an induced MU error is shown on the MapCHECK geometry in Figure 4-10. Of note, the structure in the gamma value maps for both locally and globally normalized comparisons appears to be the negative of the structure in the gradient maps. Additionally, inspecting the gradient and dose maps in tandem with the gamma value maps allows us to see that areas of high dose and low gradient fail the gamma comparison first in the presence of this induced error, regardless of dose normalization strategy. While locally normalized gamma comparisons may increase error sensitivity for the gamma comparison when used as a binary metric, the areas of the comparison that have the lowest gamma values in the presence of induced errors are along high dose gradients, regardless of normalization setting.

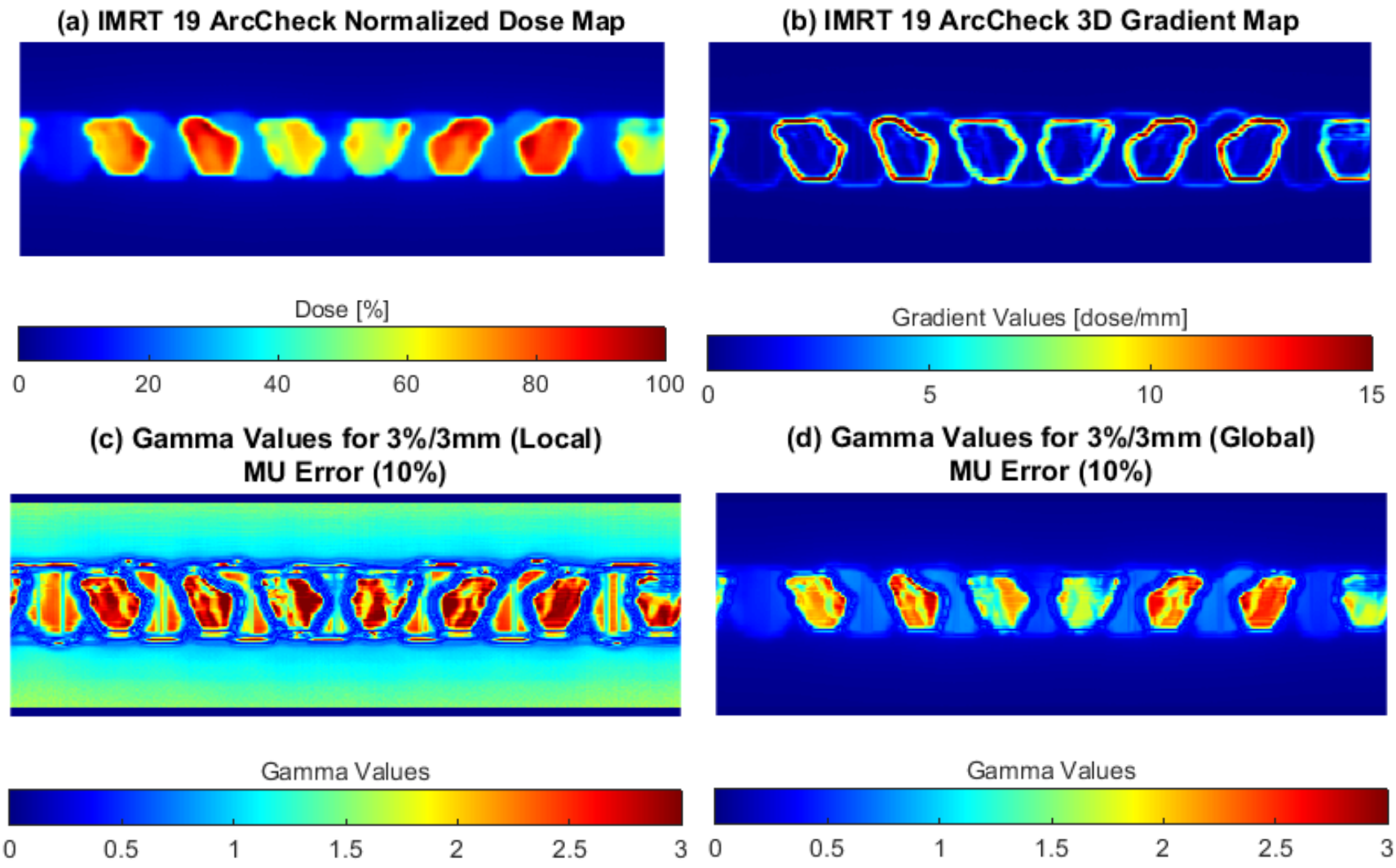

Figure 4-6. *(a)* Unwrapped ArcCHECK dose map and *(b)* calculated 3D gradient map for case IMRT 19. The resulting gamma value maps in the presence of a 10% MU error are shown for *(c)* a locally normalized comparison and *(d)* a globally normalized comparison.

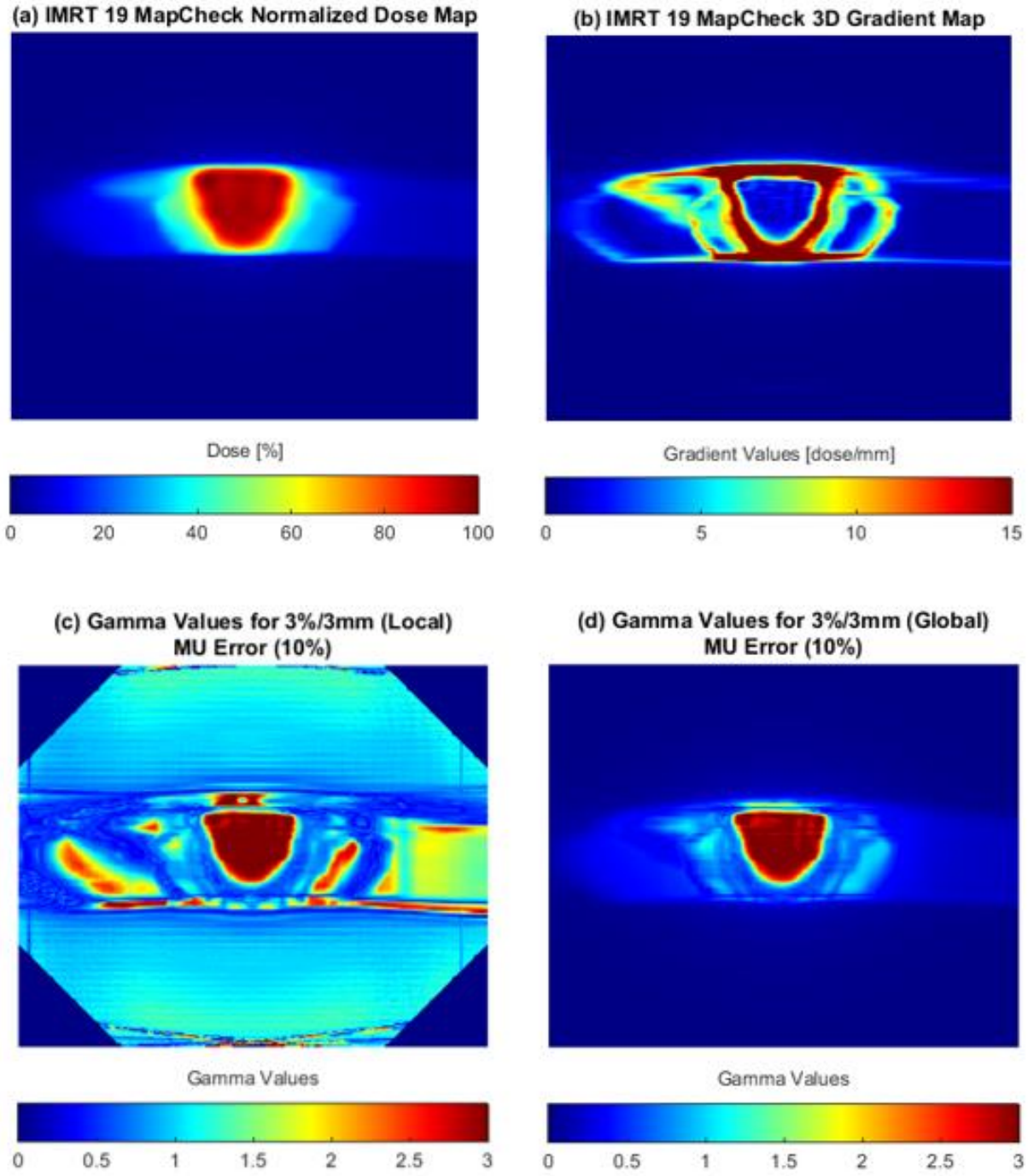

Figure 4-7. *(a)* Error-free calculated dose map on the MapCHECK device and *(b)* corresponding 3D gradient map for the error-free calculation. Gamma value maps are shown for an induced 10% MU error for *(c)* a locally normalized gamma comparison and *(d)* a globally normalized gamma comparison.

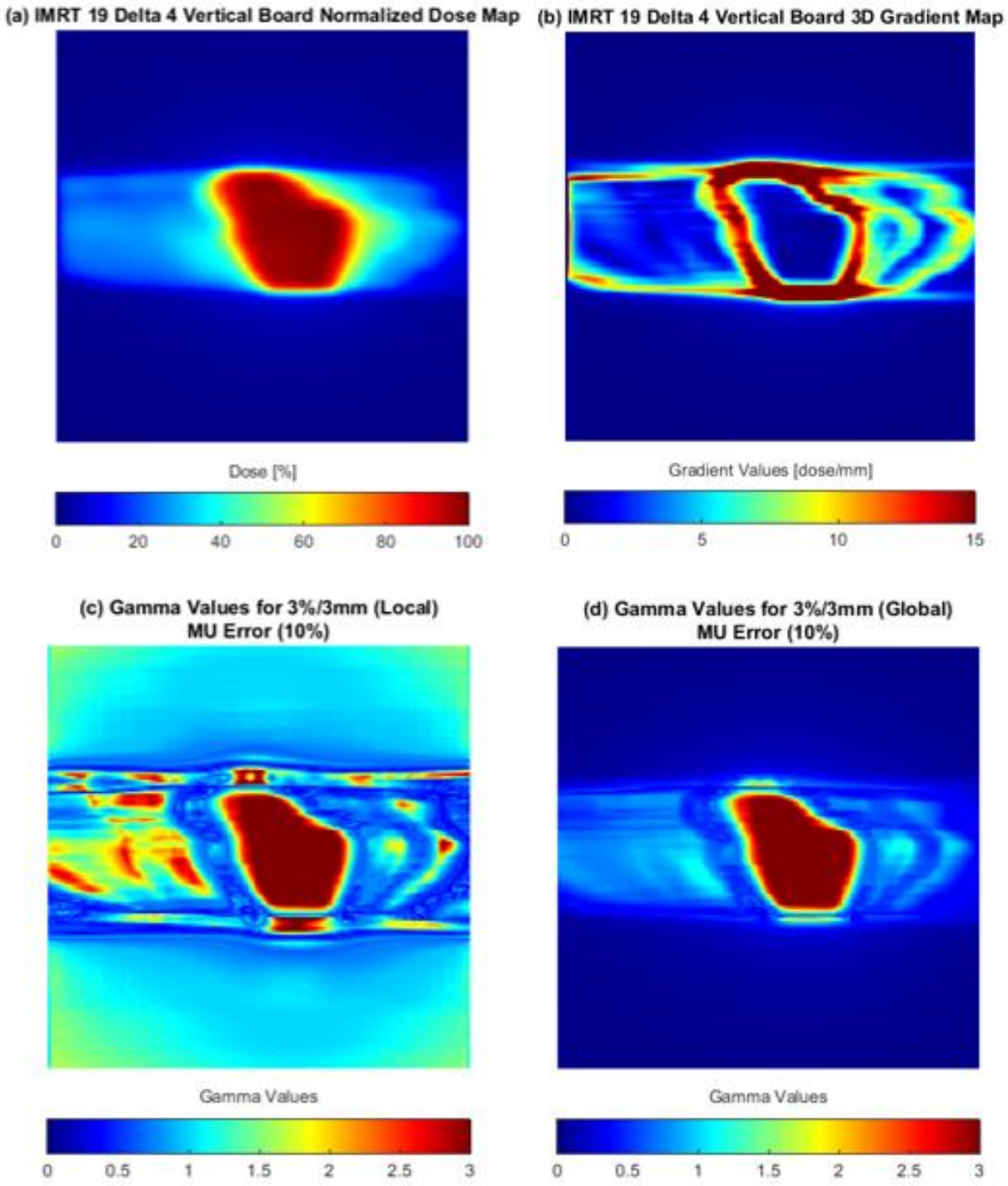

Figure 4-8. *(a)* Error-free calculated dose map on for the vertical board in the Delta 4 and *(b)* corresponding 3D calculated gradient map. Gamma value maps are shown for an induced MU error of 10% for *(c)* a locally normalized gamma comparison and *(d)* a globally normalized gamma comparison.

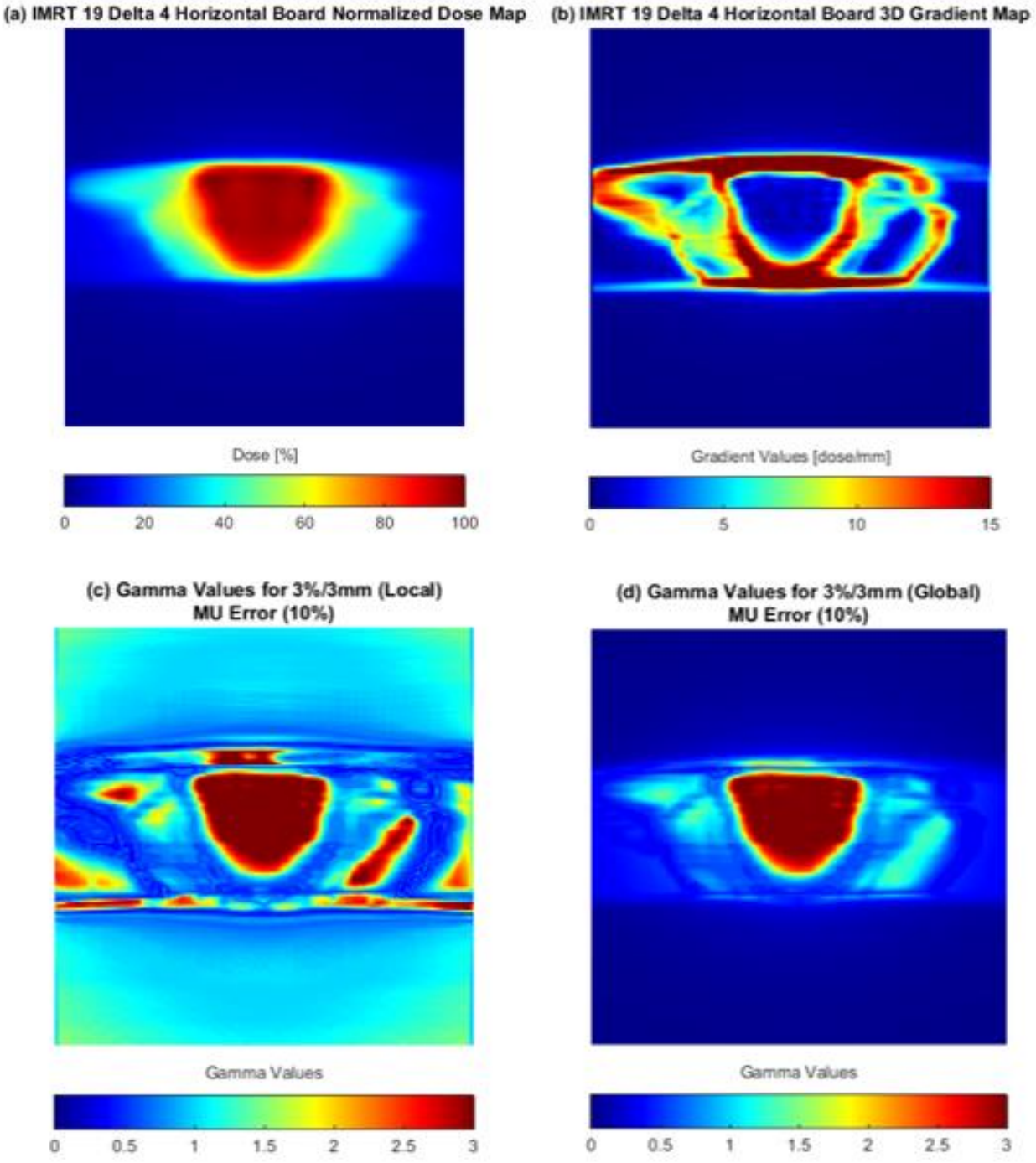

Figure 4-9. *(a)* Error-free calculated dose map on for the horizontal board in the Delta 4 and *(b)* corresponding 3D calculated gradient map. Gamma value maps are shown for an induced MU error of 10% for *(c)* a locally normalized gamma comparison and *(d)* a globally normalized gamma comparison.

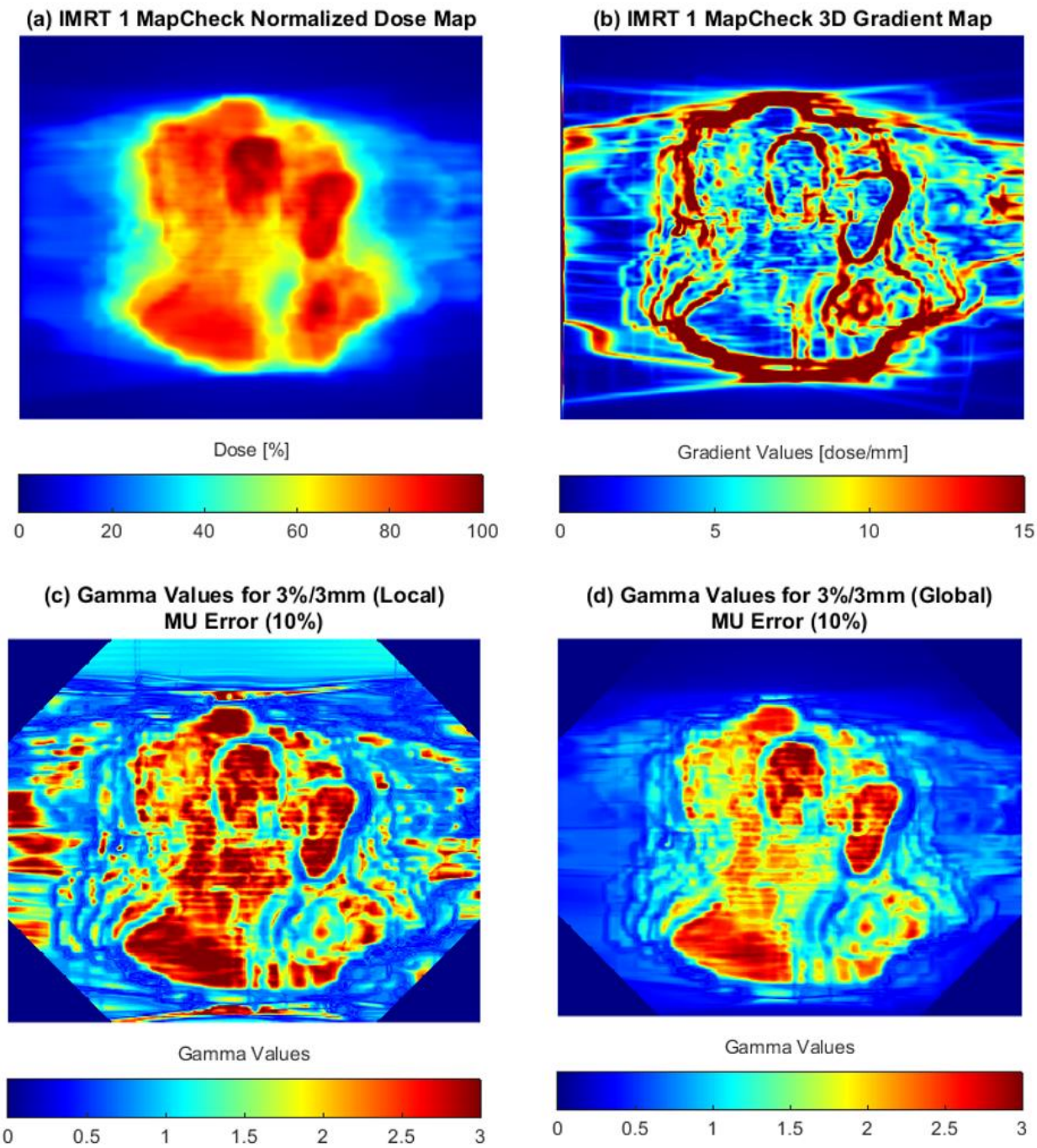

Figure 4-10. The relationship between dose and gradient maps to gamma values in the presence of an induced MU error are shown. *(a)* An error-free calculated dose map for IMRT 1 on the MapCHECK device, where the dose is normalized to the maximum point dose. *(b)* The 3D dose gradient map for IMRT 1 on the MapCHECK *(c)* the calculated gamma values when comparing the dose map in *(a)* to a calculation with an induced 10% MU error using a locally normalized gamma comparison *(d)* the calculated gamma values in the presence of a 10% MU error using a globally normalized gamma comparison.

Representative examples are also shown for a bank error on the ArcCHECK in Figure 4-11, a MLC perturbational error on the MapCHECK in Figure 4-12, and a collimator error on the Delta 4 in Figure 4-14. For the random MLC perturbational and collimator errors, the highest gamma values appear, not surprisingly, more randomly distributed in the comparison due to the nature of these induced errors. For these two error types it is more difficult to visualize from two separate images where the gradients fall in relation to the gamma values. Therefore, colormap overlays of the gamma values on a greyscale image of the gradients are presented in Figure 4-13 and Figure 4-15. These show that even though high gamma values are less uniform for these error types, as one might expect, the gamma values are very low at the location of the highest dose gradients. Thus, regardless of detector device, error type, and dose normalization strategy, high dose gradients appear to prevent comparison points from failing gamma in the presence of known, induced errors.

The majority of case examples shown here are less complex plans for illustrative purposes. However, for more complex plans, the gradient maps may be so intricate that a large number of measurement points fall on these high gradients and subsequently limit error sensitivity in real comparisons. One case example of this is shown in Figure 4-16 for a 3-target 14 field IMRT case. In this example, the ArcCHECK diode map was overlaid on the gradient map in greyscale to illustrate how few measurement points can fall in high-dose, low-gradient regions for highly complex plans. While one might assume that a measurement device with more diodes would increase the sensitivity of IMRT QA comparisons, it is likely that the number of measurement points falling in low gradient regions of the field drive error sensitivity.

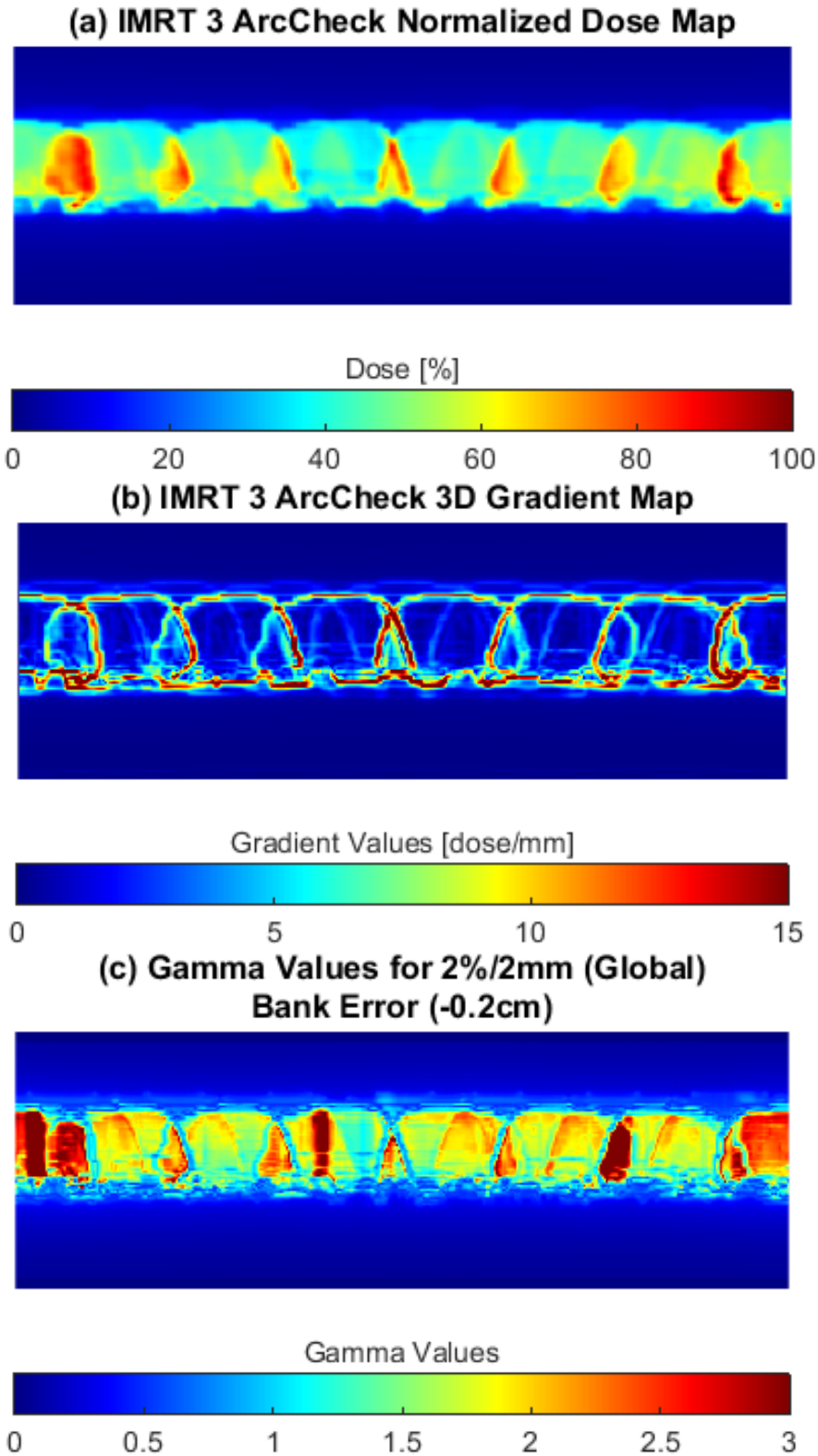

Figure 4-11. *(a)* Error-free unwrapped dose map on the ArcCHECK geometry, *(b)* error-free unwrapped 3D gradient map for IMRT 3 and *(c)* gamma values calculated in the presence of a -0.2 cm bank shift error.

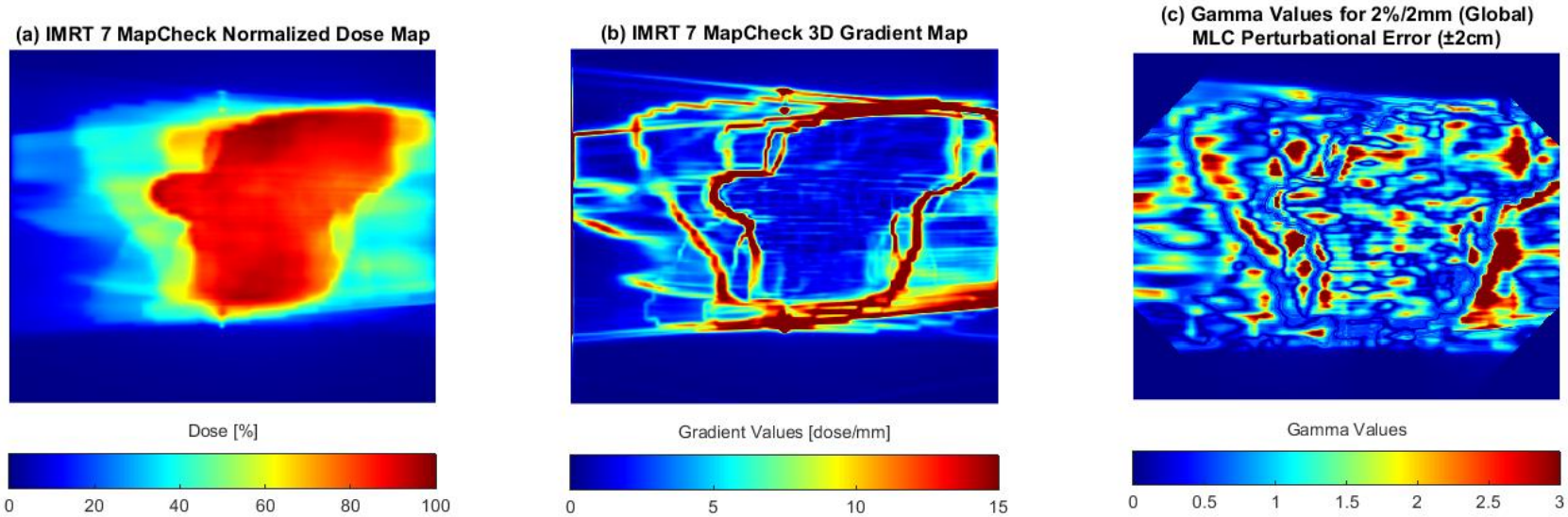

Figure 4-12. *(a)* Error-free MapCHECK dose map, *(b)* error-free MapCHECK gradient map and *(c)* gamma values in the presence of an induced MLC perturbational error of ±2 cm.

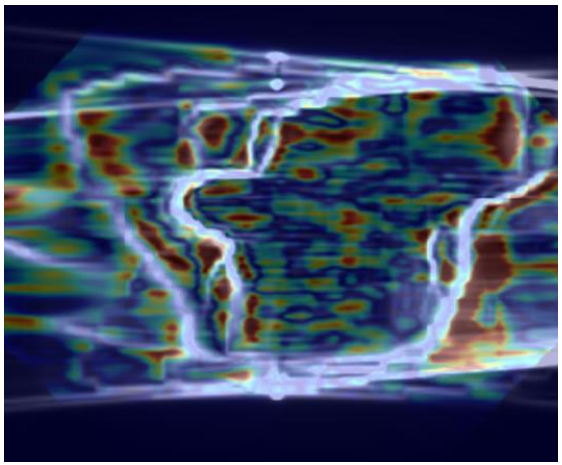

Figure 4-13. The 3D gradient map for IMRT 7 from Figure 4-12 is shown in greyscale, with the highest gradients shown in white. The gamma value map from Figure 4-10c) is overlaid on top of the gradient map to better illustrate the relationship between gradients and gamma values. For this MLC error, the highest gamma values fall just outside the high gradient regions of the plan.

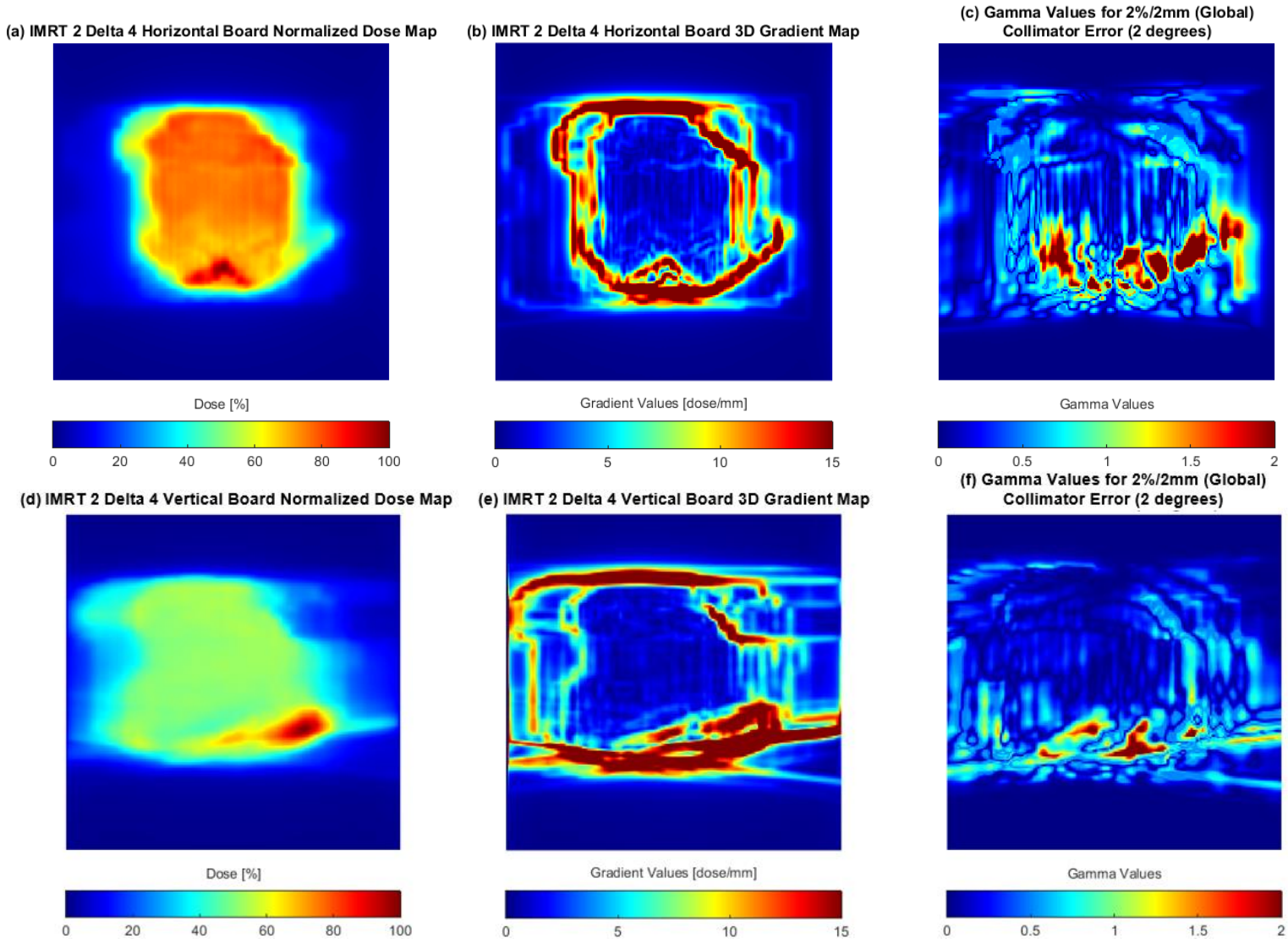

Figure 4-14. *(a)* Error-free Delta 4 horizontal detector board dose map *(b)* error-free Delta 4 horizontal board gradient map and *(c)* gamma values in the presence of an induced collimator error of 2 degrees.

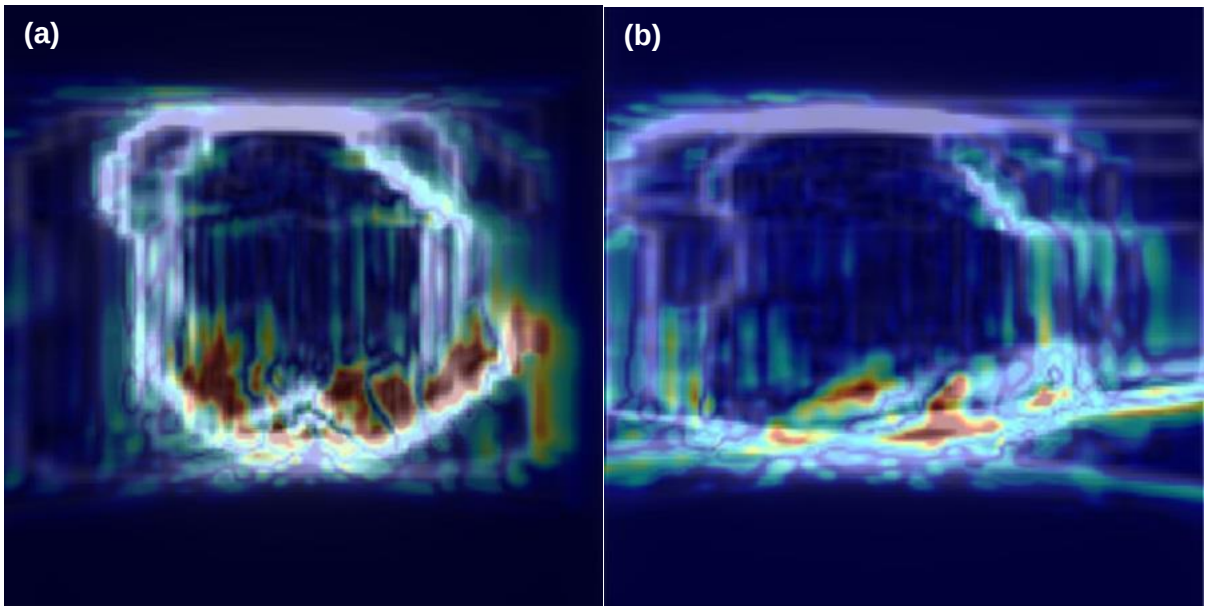

Figure 4-15. The 3D gradient map for IMRT 2 from Figure 4-14 is shown in greyscale for the (a) horizontal Delta 4 board and (b) the vertical Delta 4 board. The highest gradient values are shown in white and the gamma value maps from Figure 4-14 are overlaid on the gradient map to better illustrate the relationship between gradients and gamma values. For this collimator error, the highest gamma values fall just outside the high gradient regions of the plan.

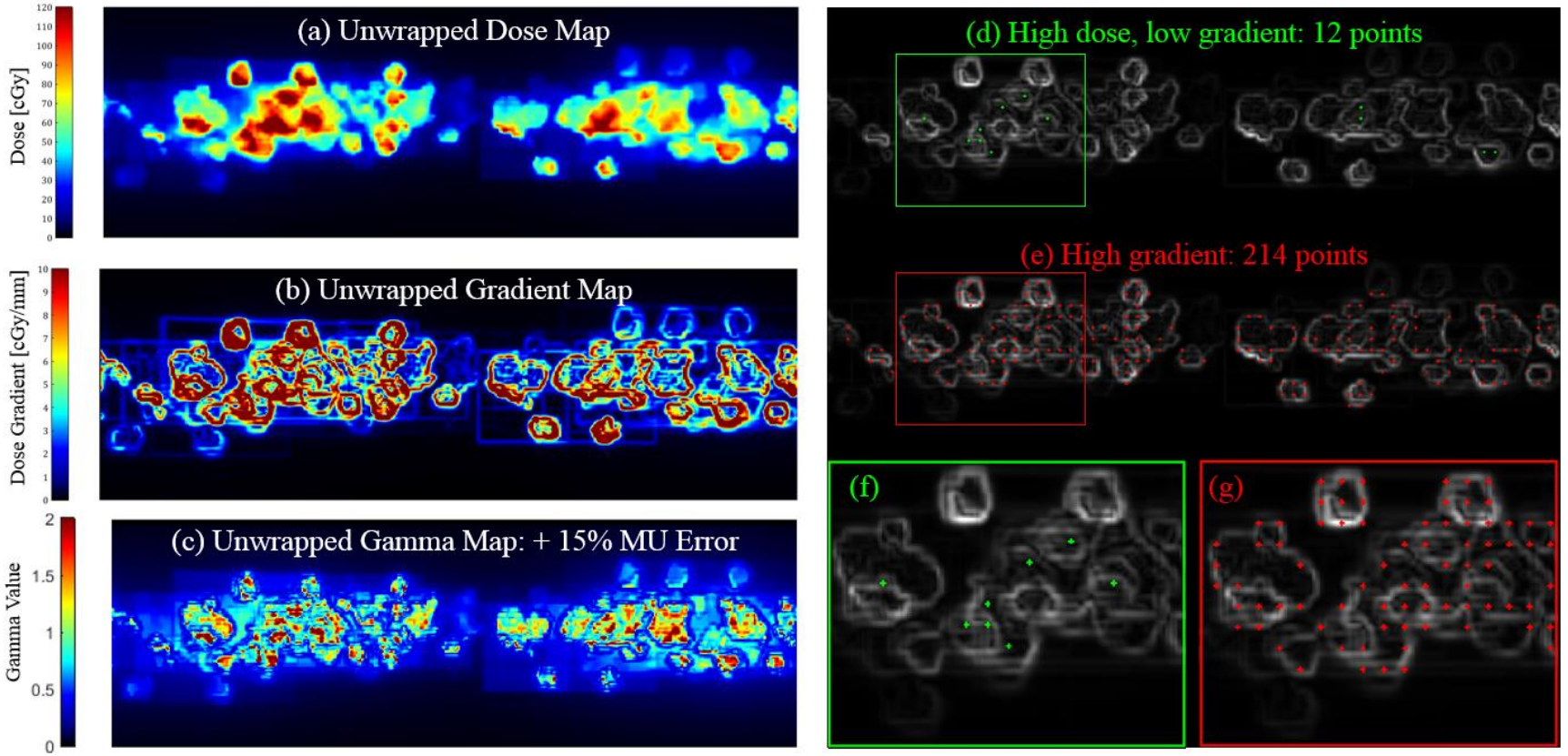

Figure 4-16. *(a)* Unwrapped ArcCHECK dose map *(b)* unwrapped ArcCHECK gradient map and *(c)* unwrapped gamma value map clearly show that points with high gamma values do not fall along the gradients and that higher gamma values occur first in high dose, low gradient regions. The points falling in different gradient and dose regions of the distribution are superimposed on the gradient map to illustrate *(d,f)* the small number of error sensitive points versus *(e,g)* points falling along the high dose gradients.

IV.D. Gamma error sensitivity and number of available measurement points

IV.D.1. Methods

Based on the fact that visible relationships were observed in error-induced comparisons between low gamma values and high dose gradients, we investigated if simple metrics could predict the error sensitivity (i.e. error curve ranges) of the various cases in this study. This was first explored for the 20 IMRT cases on the simplest measurement geometry – MapCHECK. In theory, if the error sensitivity of a particular case can be predicted, it may allow the user to adjust the gamma criterion for a particular measurement in order to achieve the desired error sensitivity.

For the MLC perturbational error, lagging leaf error, bank shift error, and collimator rotation error the error curve ranges for the 1 mm error-free calculation compared to the 1mm error-induced calculation were pulled from the work in Chapter III for the five different gamma criteria studied:

- 3%/3mm TH 10 (G), 95% pixels passing
- 3%/3mm TH 10 (L), 95% pixels passing
- 2%/2mm TH 10 (G), 95% pixels passing
- 3%/3mm TH 50 (G), 95% pixels passing
- 3%/3mm TH 50 (L), 95% pixels passing

The error curve ranges for each combination of error type and gamma criterion were compared to different plan descriptors and Pearson r^2 correlation coefficients were calculated to assess if these plan descriptors could predict the range of errors that could pass QA for the 20 cases in the IMRT cohort. The plan characteristics investigated were the modulation complexity score (MCS)[26,27], the PTV volume, the equivalent sphere diameter of the PTV, and the average aperture area. The average aperture area was calculated in MATLAB using the MLC information from DICOM RT plan files and was defined as:

$$Average\ Aperture\ Area = \frac{1}{n_{Beams}} \sum_{j=1}^{n_{Beams}} \frac{1}{(n_{CP})_j} \sum_{i=1}^{n_{Cp}} A_{ij} \qquad [Eq. 4.2]$$

Where:

- A_i is the open area for a given aperture or control point, *i*, in a given beam*, j*
- n_{CP} is the total number of control points in a given beam or arc
- n_{Beams} is the total number of beams or arcs in a plan

Additionally, we postulated that the number of available points for comparison in different gradient and dose regions of the field may predict gamma comparison error sensitivity. In order to calculate the number of points available in the different segmented regions of the field, 56 different combinations of dose and gradient thresholds were evaluated. For each gradient and dose threshold, the number of points in high-dose low-gradient regions, low-dose low-gradient regions, and high-gradient regions were counted in the 1mm error-free calculations and gradient maps obtained from the work in Chapter III. Gradient thresholds ranged from 4-16 cGy/mm by intervals of 2 cGy/mm and dose thresholds investigated were 20%, 25%, 30%, 40%, 50%, 60%, 70%, and 80% of the maximum error-free calculated dose. The number of points in each region were calculated using an in-house MATLAB script. Pearson correlation coefficients were calculated between the error curve ranges and the number of points in each region for the 56 different dose and gradient thresholds, for a total of 168 (i.e. 56*3) r^2 values for the dose and gradient segmentation alone. Including the four other plan descriptors studied, the four error types, and the error curve ranges for the five different gamma criteria studied, 3,440 r^2 values were calculated to determine if any of these metrics predict gamma error curve ranges.

IV.D.2. Results

Pearson correlation coefficients for the five gamma criteria and the error ranges for the four error types compared to MCSv, PTV volume, equivalent square PTV diameter, and the average aperture area are shown in Table 4-1, Table 4-2, Table 4-3, and Table 4-4. Correlation coefficients less than 0.2 are highlighted in grey, 0.2 – 0.4 in blue, 0.4 – 0.6 in green, 0.6 – 0.8 in yellow, and 0.8 – 1.0 in red. However, all correlation coefficients were less than 0.70. For a particular error type, such as the bank shift error, the modulation complexity score had moderate r^2 values ranging from 0.57 – 0.678 depending on the gamma criterion. However, there was not one plan metric that was able to consistently predict error curve ranges across error types. In terms of predicting error sensitivity for a real measurement, where these errors may be added together in any given way, it clearly would be difficult to predict an error range for a given case and adjust the gamma criteria appropriately.

Table 4-1. Pearson correlation coefficients for MLC perturbational error curve ranges vs. simple plan characteristics

	Pearson r^2 for MLC perturbational error curve ranges vs. plan metrics				
	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
MCS_v	0.3630	0.3260	0.0049	0.2629	0.0675
PTV volume [cm^3]	0.1491	0.0609	0.0371	0.0905	0.0018
Equivalent sphere diameter [cm]	0.1739	0.1057	0.0222	0.1376	0.0042
Average aperture area [cm^2]	0.0678	0.0034	0.0006	0.1264	0.0000

Table 4-2. Pearson correlation coefficients for MLC lagging leaf error curve ranges vs. simple plan characteristics

	Pearson r^2 for MLC lagging leaf error curve ranges vs. plan metrics				
	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
MCS_v	0.0057	0.0201	0.1129	0.0151	0.0907
PTV volume [cm^3]	0.2922	0.6382	0.6174	0.5640	0.6892
Equivalent sphere diameter [cm]	0.3749	0.6357	0.5802	0.5906	0.6184
Average aperture area [cm^2]	0.3086	0.0626	0.0058	0.1208	0.0067

Table 4-3. Pearson correlation coefficients for MLC bank shift error curve ranges vs. simple plan characteristics

	Pearson r^2 for MLC bank shift error curve ranges vs. plan metrics				
	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
MCS_v	0.6143	0.6780	0.5768	0.5885	0.6545
PTV volume [cm^3]	0.1536	0.1106	0.0706	0.1542	0.0726
Equivalent sphere diameter [cm]	0.1416	0.1111	0.0467	0.1468	0.0580
Average aperture area [cm^2]	0.0540	0.0503	0.0742	0.0512	0.0960

Table 4-4. Pearson correlation coefficients for collimator rotation curve ranges vs. simple plan characteristics

	Pearson r^2 for collimator rotation error curve ranges vs. plan metrics				
	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
MCS_v	0.0316	0.0196	0.0614	0.0526	0.0704
PTV volume [cm^3]	0.3127	0.2647	0.2659	0.2690	0.2681
Equivalent sphere diameter [cm]	0.5560	0.4615	0.5135	0.4983	0.5188
Average aperture area [cm^2]	0.0209	0.0087	0.0158	0.0072	0.0161

In studying the number of points in different dose and gradient regions of the plan comparison, weak to moderate correlations were observed between error curve ranges and the number of points falling in high-dose low-gradient, low-dose low-gradient, or high-gradient regions for each of the 56 dose and gradient thresholds. These results can be found in Appendix A. Clearly, the simple plan descriptors studied here are not capable of predicting the gamma comparison error sensitivity across a range of cases and error types. These results suggest that gamma comparison error sensitivity is a more complex relationship that is error type-specific and likely also case-specific, too.

IV.E. Discussion

In this chapter we have shown for a variety of different induced error types that the complex, overlapping gradients in current IMRT QA comparisons may limit gamma comparison sensitivity. This work was performed in calculation vs. calculation scenarios, so any uncertainties introduced in real measurements are not present in the comparisons presented here. While it is clear there is a relationship between high dose gradients and low gamma values, no relationship was

observed between plan metrics or the number of points available in different dose and gradient regions of the field predicted error curve ranges for the variety of studied error types. More importantly, it is clear from the qualitative comparisons of gradient and dose maps compared to gamma value maps that the many overlapping dose gradients in real patient cases are one of the driving factors in gamma comparison insensitivity. Additionally, this observed behavior was independent of gamma comparison criteria, including dose normalization strategy, and also independent of detector geometry.

Another way of visualizing the relationship between dose/gradient and gamma maps is to utilize k-means clustering (i.e. unsupervised learning) to group features in the dose and gradient maps. The k-means clustering algorithm uses features in the data, in this case dose and gradient values, and attempts to assign each data point to a group that has similar features. Clustering analysis was performed using the k-means clustering function in MATLAB with a varying number of *k* clusters. Figure 4-17 shows gamma value maps for four of the induced errors for one of the IMRT cases on the MapCHECK (IMRT 10). The gamma value maps have very different structure based on the type of induced error, as previously mentioned. However, it is clear that for different numbers of clusters, the k-means algorithm is able to separate points into groupings that have markedly similar structure as that seen in the gamma maps, particularly for simpler errors such as MU errors. As few as k=5 clusters were enough to separate dose and gradient points into regions that roughly predict the most sensitive regions of the comparison for bank shift and MU errors, since these errors occur more uniformly throughout the field. For collimator and MLC perturbational errors, which tend to have greater dosimetric effects along the edges of the field, the gamma maps are less uniform. However, the structure seen in the k-means cluster map for 12 clusters mimics the structure in these gamma maps as well.

This type of analysis further supports that a relationship between dose, gradients, and gamma is not purely qualitative, but that more complex unsupervised learning algorithms can also partition dose and gradient maps into regions that are likely to be more sensitive to errors. Gradient and dose maps, either separately or using k-means clustering, may be a way of selecting cases *a priori* that are likely to have small areas of sensitive regions in gamma comparisons and may require additional analysis in IMRT QA. These plans with a large number of overlapping gradients may be limited in error sensitivity, meaning that if errors did exist in the delivered plan, it is possible they could go undetected using conventional IMRT QA comparisons, as the number of points available in the regions of the field that will fail the gamma comparison first (high-dose, low-gradient regions) may contain very few measurement points.

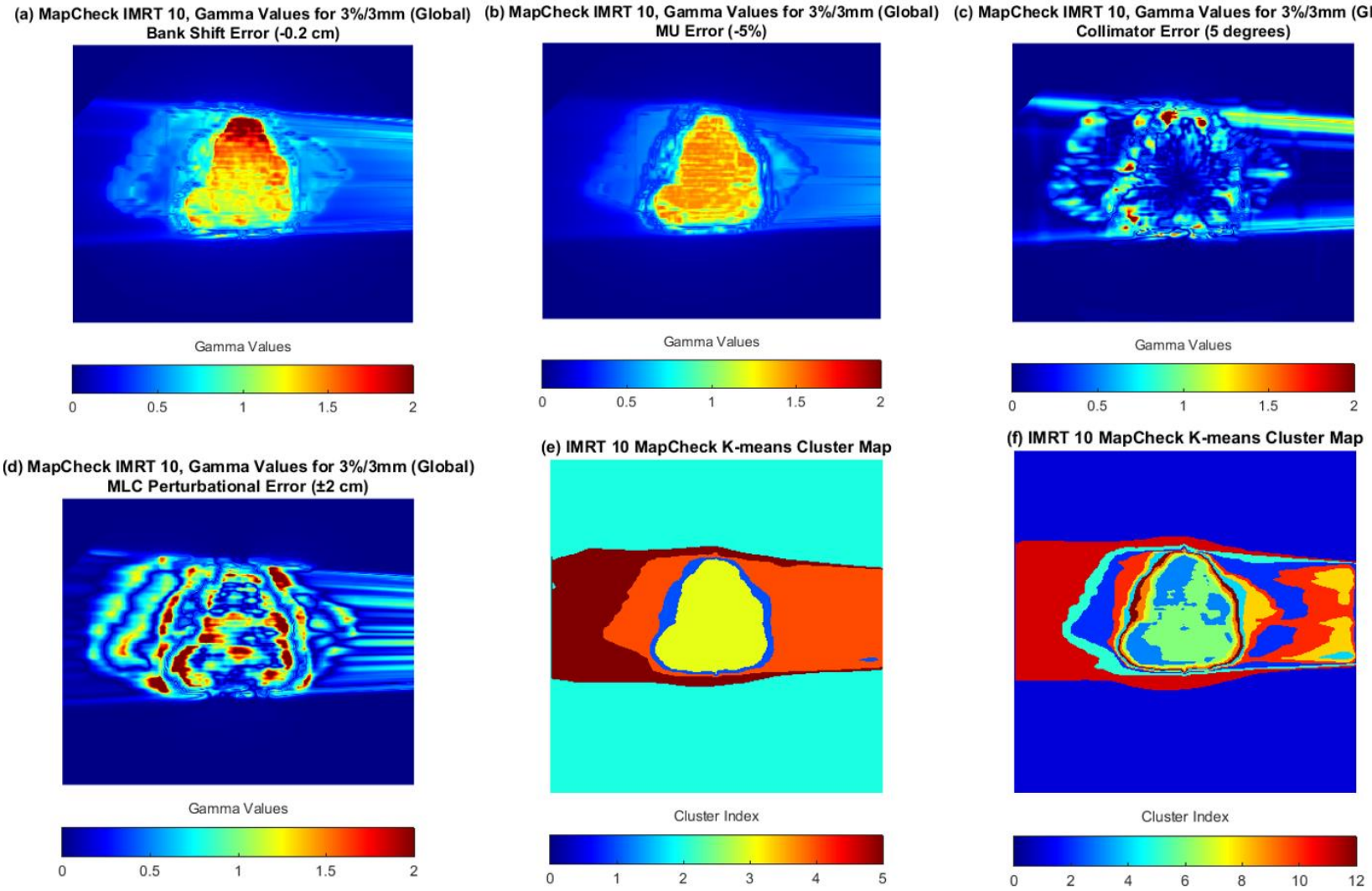

Figure 4-17. For case IMRT 10 on the MapCHECK geometry, 3%/3mm (Global) gamma value maps are shown for *(a)* bank shift error of -0.2cm *(b)* a MU error of -5% *(c)* a collimator error of 5 degrees and *(d)* a MLC perturbational error of ±2 cm. K-means cluster maps are shown for two different values of *k*, *(e)* *k* = 5 and *(f)* *k* = 12.

To evaluate the number of diodes residing in these high-dose low-gradient regions for the plans in this study, the number of diodes was counted for each detector geometry using the true spatial sampling of the device. That is, the 1mm 3D gradient maps and 1mm calculated dose maps were first down-sampled to the true detector locations for each device. Gradient maps were normalized by the dose per fraction for each case to remove the effects of plans with very different fractionation schemes. Next, dose points less than 10% of the maximum dose were excluded, to mimic current clinical gamma comparisons which removes these low dose points from the analysis. Based on qualitative comparisons, we selected 50% as the high dose threshold, and a 5%/mm gradient threshold. Figure 4-18 shows the number of measurement points falling in high-dose low-gradient regions of the field, normalized by the total number of diodes for each device (1386 diodes for ArcCHECK, 1527 diodes for MapCHECK, and 1069 diodes for Delta 4). Normalizing by the number of diodes in each device allows fairer comparisons of available high-dose low-gradient points for analysis between devices with very different spatial sampling.

The first thing to note in Figure 4-18 is that for some cases, there is an alarmingly low number of points that fall within high-dose, low-gradient regions of the field. This is concerning when considering that the cases with the lowest number of high-dose, low-gradient diodes are likely those with smaller target volumes and may very often be cases receiving higher doses per fraction. Thus, these plans may have very few diodes in the measurement that aid in catching errors when using the gamma comparison. Moreover, these are likely the types of cases in which increased error sensitivity is more important.

Another noticeable feature in Figure 4-18 is that the Delta 4 appears to have a higher percentage of diodes falling within high-dose, low-gradient regions of the field, which may in part be due to the unique detector arrangement in this device. In the Delta 4, the central portion of the device

has increased spatial sampling, which also corresponds to the center of most measured plans where the highest doses are likely to be located.

In the setting of continued gamma comparison use, it could prove useful to evaluate plan dose gradients on the phantom prior to measurement to determine if the gradient maps are too complex for appropriate sensitivity in gamma comparisons. Further study is required to determine how best to evaluate these plans with many overlapping gradients for clinical IMRT QA comparisons. However, given the increasing complexity of plans in routine clinical practice, it seems reasonable that new comparison methods that deal separately with these different regions of gradient and dose in the field would be more appropriate moving forward.

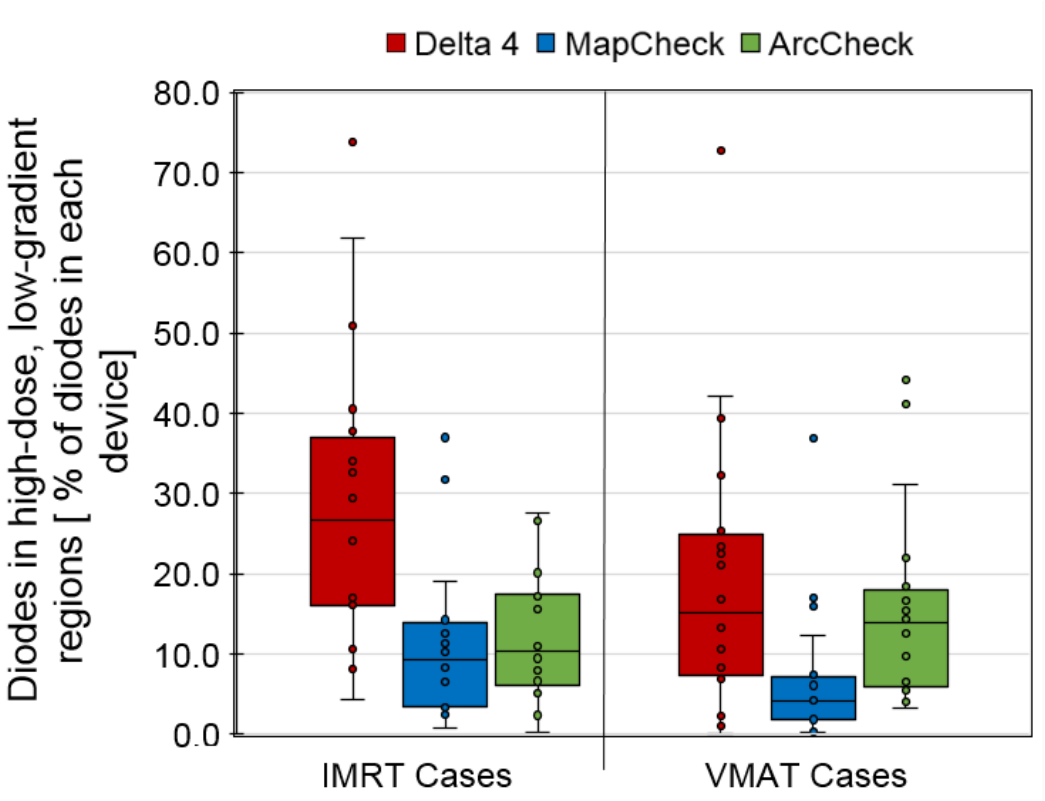

Figure 4-18. The percentage of diodes falling in high-dose, low-gradient regions of the plan measurements for both IMRT and VMAT cohorts and each of the three diode arrays studied.

IV.F. Conclusions

Using a variety of test fields and many patient plans, this chapter has shown that the insensitivity of the gamma comparison may be driven in part by the complexity of dose gradient maps for current IMRT and VMAT deliveries. This was initially discovered by inducing errors in calculation-vs-calculation comparisons on the ArcCHECK for simple test plans and observing, as error magnitudes were increased, where gamma comparison failures occurred first in the presence of these induced errors. These comparisons showed that the gamma comparison failed first in high-dose, low-gradient regions of the field for globally normalized comparisons, while failures occurred last along the highest dose gradients in the plan. 3D gradient maps were calculated for many different patient plans superimposed on the ArcCHECK, MapCHECK, and Delta 4 geometries. Gamma value maps in the presence of known, induced errors were calculated using in-house software and compared to gradient maps. The examples shown here clearly illustrate that the gradient and gamma value maps appear to be negatives of another, suggesting that the lowest gamma values in the presence of an error fall along the highest dose gradients. This behavior was present regardless of the dose difference normalization technique.

Although error curve ranges were not predicted by the number of measurement points falling within different dose and gradient regions of the comparison, there is a clear relationship between gamma sensitivity and the large number of overlapping gradients in current IMRT QA comparisons. Diodes falling on these high gradients can inflate gamma passing rates in the presence of a true error. Additionally, the number of diodes in real comparisons that fall in the error-sensitive regions of the plan (i.e. high-dose, low-gradient regions) can be disturbingly low for cases with more gradients and smaller target volumes. As patient treatments become more complex, increasing error sensitivity in IMRT QA may be desirable and thus new or modified comparison techniques that account for the complexity of plan gradients may be necessary.

A.IV. Appendix A for Chapter 4

Pearson correlation coefficients are shown in the following tables for the five gamma criteria and the error ranges on the MapCHECK for the four error types compared to the number of points available in in different dose and gradient regions of the field based on the listed dose and gradient thresholds.

Table A.4-1. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available in high-dose, low-gradient regions for MLC perturbational errors.

Pearson r^2 for MLC perturbational error curve ranges vs. # high-dose, low-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.0115	0.0311	0.0250	0.0138	0.0075
20%	6	0.0400	0.0539	0.0251	0.0279	0.0032
20%	8	0.0698	0.0727	0.0223	0.0420	0.0008
20%	10	0.0858	0.0794	0.0203	0.0505	0.0002
20%	12	0.0954	0.0826	0.0196	0.0556	0.0001
20%	14	0.1008	0.0842	0.0196	0.0587	0.0000
20%	16	0.1042	0.0853	0.0198	0.0607	0.0000
25%	4	0.0115	0.0099	0.0311	0.0252	0.0135
25%	6	0.0376	0.0256	0.0296	0.0413	0.0070
25%	8	0.0651	0.0410	0.0263	0.0564	0.0030
25%	10	0.0802	0.0481	0.0239	0.0650	0.0015
25%	12	0.0889	0.0516	0.0232	0.0696	0.0010
25%	14	0.0940	0.0535	0.0235	0.0722	0.0008
25%	16	0.0971	0.0547	0.0239	0.0738	0.0008
30%	4	0.0133	0.0031	0.0336	0.0284	0.0144
30%	6	0.0400	0.0136	0.0317	0.0462	0.0079
30%	8	0.0684	0.0265	0.0285	0.0620	0.0037
30%	10	0.0833	0.0329	0.0263	0.0705	0.0021
30%	12	0.0916	0.0363	0.0257	0.0749	0.0015
30%	14	0.0964	0.0382	0.0262	0.0772	0.0013
30%	16	0.0998	0.0396	0.0265	0.0790	0.0012
40%	4	0.0121	0.0013	0.0552	0.0172	0.0210
40%	6	0.0445	0.0110	0.0458	0.0383	0.0097
40%	8	0.0766	0.0239	0.0393	0.0559	0.0042
40%	10	0.0930	0.0302	0.0360	0.0652	0.0024
40%	12	0.1013	0.0333	0.0347	0.0696	0.0017
40%	14	0.1061	0.0350	0.0346	0.0720	0.0015
40%	16	0.1092	0.0362	0.0346	0.0736	0.0013
50%	4	0.0057	0.0006	0.0613	0.0115	0.0261
50%	6	0.0349	0.0102	0.0468	0.0349	0.0105
50%	8	0.0659	0.0236	0.0386	0.0539	0.0040
50%	10	0.0821	0.0302	0.0345	0.0632	0.0020
50%	12	0.0897	0.0328	0.0328	0.0670	0.0013
50%	14	0.0939	0.0340	0.0324	0.0692	0.0011
50%	16	0.0970	0.0351	0.0320	0.0707	0.0009
60%	4	0.0007	0.0000	0.0639	0.0062	0.0339
60%	6	0.0149	0.0036	0.0495	0.0226	0.0171

Table A.4-1 (cont.)

60%	8	0.0349	0.0111	0.0400	0.0383	0.0084
60%	10	0.0473	0.0159	0.0355	0.0470	0.0052
60%	12	0.0551	0.0189	0.0335	0.0514	0.0037
60%	14	0.0595	0.0204	0.0328	0.0538	0.0031
60%	16	0.0623	0.0215	0.0320	0.0552	0.0027
70%	4	0.0000	0.0003	0.0570	0.0033	0.0372
70%	6	0.0047	0.0008	0.0444	0.0137	0.0229
70%	8	0.0121	0.0029	0.0369	0.0230	0.0154
70%	10	0.0167	0.0043	0.0336	0.0281	0.0121
70%	12	0.0197	0.0053	0.0323	0.0307	0.0105
70%	14	0.0217	0.0059	0.0317	0.0321	0.0097
70%	16	0.0231	0.0064	0.0310	0.0331	0.0090
80%	4	0.0164	0.0132	0.0614	0.0035	0.0697
80%	6	0.0090	0.0093	0.0528	0.0005	0.0587
80%	8	0.0059	0.0075	0.0471	0.0000	0.0515
80%	10	0.0048	0.0070	0.0441	0.0000	0.0479
80%	12	0.0044	0.0067	0.0425	0.0001	0.0460
80%	14	0.0041	0.0067	0.0415	0.0001	0.0452
80%	16	0.0040	0.0066	0.0407	0.0001	0.0444

Table A.4-2. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available in low-dose, low-gradient regions for MLC perturbational errors.

Pearson r^2 for MLC perturbational error curve ranges vs. # low-dose, low-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.1681	0.0827	0.0223	0.1624	0.0669
20%	6	0.1408	0.0666	0.0194	0.1503	0.0558
20%	8	0.1265	0.0590	0.0155	0.1418	0.0473
20%	10	0.1168	0.0554	0.0135	0.1346	0.0432
20%	12	0.1109	0.0533	0.0125	0.1303	0.0410
20%	14	0.1079	0.0524	0.0119	0.1282	0.0400
20%	16	0.1064	0.0517	0.0113	0.1274	0.0392
25%	4	0.1599	0.0468	0.0184	0.1815	0.0539
25%	6	0.1279	0.0301	0.0169	0.1714	0.0433
25%	8	0.1089	0.0230	0.0133	0.1607	0.0347
25%	10	0.0978	0.0204	0.0116	0.1524	0.0311
25%	12	0.0909	0.0187	0.0106	0.1468	0.0289
25%	14	0.0876	0.0179	0.0099	0.1441	0.0277
25%	16	0.0857	0.0174	0.0092	0.1426	0.0268
30%	4	0.1702	0.0337	0.0218	0.1911	0.0604
30%	6	0.1305	0.0155	0.0209	0.1805	0.0482
30%	8	0.1077	0.0091	0.0167	0.1679	0.0378
30%	10	0.0945	0.0069	0.0143	0.1585	0.0331
30%	12	0.0862	0.0056	0.0130	0.1523	0.0305
30%	14	0.0822	0.0050	0.0122	0.1492	0.0290
30%	16	0.0803	0.0047	0.0116	0.1478	0.0284
40%	4	0.1862	0.0323	0.0278	0.1828	0.0785
40%	6	0.1365	0.0104	0.0326	0.1701	0.0702
40%	8	0.1024	0.0030	0.0293	0.1530	0.0574
40%	10	0.0843	0.0011	0.0266	0.1413	0.0510
40%	12	0.0727	0.0003	0.0255	0.1332	0.0478
40%	14	0.0670	0.0001	0.0249	0.1290	0.0461
40%	16	0.0641	0.0000	0.0243	0.1268	0.0451
50%	4	0.1769	0.0314	0.0350	0.1772	0.0873
50%	6	0.1166	0.0080	0.0495	0.1605	0.0858
50%	8	0.0724	0.0008	0.0494	0.1341	0.0724
50%	10	0.0509	0.0000	0.0481	0.1168	0.0656
50%	12	0.0379	0.0007	0.0483	0.1049	0.0621
50%	14	0.0314	0.0015	0.0487	0.0985	0.0602
50%	16	0.0285	0.0021	0.0490	0.0957	0.0595
60%	4	0.1588	0.0270	0.0378	0.1677	0.0855

Table A.4-2 (cont.)

60%	6	0.0853	0.0032	0.0541	0.1384	0.0801
60%	8	0.0388	0.0004	0.0548	0.1042	0.0636
60%	10	0.0194	0.0037	0.0548	0.0828	0.0561
60%	12	0.0099	0.0075	0.0566	0.0689	0.0527
60%	14	0.0059	0.0104	0.0582	0.0612	0.0509
60%	16	0.0041	0.0123	0.0595	0.0572	0.0503
70%	4	0.1518	0.0256	0.0464	0.1628	0.0902
70%	6	0.0662	0.0011	0.0687	0.1233	0.0821
70%	8	0.0164	0.0046	0.0690	0.0773	0.0587
70%	10	0.0022	0.0143	0.0670	0.0499	0.0466
70%	12	0.0000	0.0229	0.0670	0.0336	0.0402
70%	14	0.0013	0.0282	0.0675	0.0254	0.0367
70%	16	0.0028	0.0314	0.0680	0.0214	0.0352
80%	4	0.1078	0.0125	0.0489	0.1188	0.0747
80%	6	0.0243	0.0018	0.0701	0.0678	0.0589
80%	8	0.0000	0.0242	0.0661	0.0273	0.0340
80%	10	0.0074	0.0435	0.0612	0.0099	0.0226
80%	12	0.0197	0.0563	0.0588	0.0031	0.0171
80%	14	0.0284	0.0629	0.0579	0.0010	0.0144
80%	16	0.0339	0.0669	0.0575	0.0003	0.0132

Table A.4-3. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available in high-gradient regions for MLC perturbational errors.

Pearson r^2 for MLC perturbational error curve ranges vs. # high-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.2375	0.1275	0.0009	0.0933	0.0294
20%	6	0.1900	0.0917	0.0051	0.0774	0.0318
20%	8	0.1187	0.0526	0.0072	0.0476	0.0224
20%	10	0.0681	0.0320	0.0103	0.0228	0.0174
20%	12	0.0318	0.0176	0.0180	0.0067	0.0158
20%	14	0.0112	0.0083	0.0301	0.0005	0.0158
20%	16	0.0026	0.0035	0.0422	0.0004	0.0165
25%	4	0.2375	0.1275	0.0009	0.0933	0.0294
25%	6	0.1900	0.0917	0.0051	0.0774	0.0318
25%	8	0.1187	0.0526	0.0072	0.0476	0.0224
25%	10	0.0681	0.0320	0.0103	0.0228	0.0174
25%	12	0.0318	0.0176	0.0180	0.0067	0.0158
25%	14	0.0112	0.0083	0.0301	0.0005	0.0158
25%	16	0.0026	0.0035	0.0422	0.0004	0.0165
30%	4	0.2375	0.1275	0.0009	0.0933	0.0294
30%	6	0.1900	0.0917	0.0051	0.0774	0.0318
30%	8	0.1187	0.0526	0.0072	0.0476	0.0224
30%	10	0.0681	0.0320	0.0103	0.0228	0.0174
30%	12	0.0318	0.0176	0.0180	0.0067	0.0158
30%	14	0.0112	0.0083	0.0301	0.0005	0.0158
30%	16	0.0026	0.0035	0.0422	0.0004	0.0165
40%	4	0.2375	0.1275	0.0009	0.0933	0.0294
40%	6	0.1900	0.0917	0.0051	0.0774	0.0318
40%	8	0.1187	0.0526	0.0072	0.0476	0.0224
40%	10	0.0681	0.0320	0.0103	0.0228	0.0174
40%	12	0.0318	0.0176	0.0180	0.0067	0.0158
40%	14	0.0112	0.0083	0.0301	0.0005	0.0158
40%	16	0.0026	0.0035	0.0422	0.0004	0.0165
50%	4	0.2375	0.1275	0.0009	0.0933	0.0294
50%	6	0.1900	0.0917	0.0051	0.0774	0.0318
50%	8	0.1187	0.0526	0.0072	0.0476	0.0224
50%	10	0.0681	0.0320	0.0103	0.0228	0.0174
50%	12	0.0318	0.0176	0.0180	0.0067	0.0158
50%	14	0.0112	0.0083	0.0301	0.0005	0.0158
50%	16	0.0026	0.0035	0.0422	0.0004	0.0165
60%	4	0.2375	0.1275	0.0009	0.0933	0.0294
60%	6	0.1900	0.0917	0.0051	0.0774	0.0318

Table A.4-3 (cont.)

60%	8	0.1187	0.0526	0.0072	0.0476	0.0224
60%	10	0.0681	0.0320	0.0103	0.0228	0.0174
60%	12	0.0318	0.0176	0.0180	0.0067	0.0158
60%	14	0.0112	0.0083	0.0301	0.0005	0.0158
60%	16	0.0026	0.0035	0.0422	0.0004	0.0165
70%	4	0.2375	0.1275	0.0009	0.0933	0.0294
70%	6	0.1900	0.0917	0.0051	0.0774	0.0318
70%	8	0.1187	0.0526	0.0072	0.0476	0.0224
70%	10	0.0681	0.0320	0.0103	0.0228	0.0174
70%	12	0.0318	0.0176	0.0180	0.0067	0.0158
70%	14	0.0112	0.0083	0.0301	0.0005	0.0158
70%	16	0.0026	0.0035	0.0422	0.0004	0.0165
80%	4	0.2375	0.1275	0.0009	0.0933	0.0294
80%	6	0.1900	0.0917	0.0051	0.0774	0.0318
80%	8	0.1187	0.0526	0.0072	0.0476	0.0224
80%	10	0.0681	0.0320	0.0103	0.0228	0.0174
80%	12	0.0318	0.0176	0.0180	0.0067	0.0158
80%	14	0.0112	0.0083	0.0301	0.0005	0.0158
80%	16	0.0026	0.0035	0.0422	0.0004	0.0165

Table A.4-4. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available in high-dose, low-gradient regions for MLC lagging leaf errors.

Pearson r^2 for MLC lagging leaf error curve ranges vs. # high-dose, low-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.2448	0.3524	0.2121	0.3523	0.2776
20%	6	0.2984	0.4595	0.3258	0.4479	0.3940
20%	8	0.3236	0.5242	0.4103	0.5039	0.4745
20%	10	0.3371	0.5632	0.4652	0.5377	0.5251
20%	12	0.3435	0.5886	0.5040	0.5590	0.5606
20%	14	0.3482	0.6047	0.5296	0.5725	0.5834
20%	16	0.3514	0.6166	0.5486	0.5826	0.6011
25%	4	0.1911	0.3681	0.2166	0.3187	0.2875
25%	6	0.2300	0.4529	0.3075	0.3931	0.3809
25%	8	0.2546	0.5108	0.3805	0.4443	0.4512
25%	10	0.2697	0.5475	0.4307	0.4774	0.4978
25%	12	0.2790	0.5726	0.4679	0.4999	0.5320
25%	14	0.2856	0.5889	0.4934	0.5145	0.5550
25%	16	0.2901	0.6012	0.5128	0.5255	0.5731
30%	4	0.1441	0.3568	0.2149	0.2790	0.2848
30%	6	0.1784	0.4359	0.3021	0.3486	0.3739
30%	8	0.2049	0.4956	0.3756	0.4020	0.4453
30%	10	0.2212	0.5318	0.4239	0.4351	0.4901
30%	12	0.2317	0.5569	0.4596	0.4580	0.5231
30%	14	0.2399	0.5742	0.4860	0.4741	0.5470
30%	16	0.2452	0.5868	0.5062	0.4857	0.5657
40%	4	0.1426	0.3587	0.2389	0.2817	0.3066
40%	6	0.1707	0.4426	0.3411	0.3517	0.4079
40%	8	0.2006	0.5105	0.4259	0.4122	0.4883
40%	10	0.2188	0.5497	0.4771	0.4474	0.5348
40%	12	0.2302	0.5746	0.5131	0.4702	0.5668
40%	14	0.2393	0.5924	0.5406	0.4868	0.5910
40%	16	0.2445	0.6046	0.5615	0.4982	0.6099
50%	4	0.1688	0.3919	0.2627	0.3169	0.3338
50%	6	0.1994	0.4861	0.3820	0.3961	0.4489
50%	8	0.2320	0.5585	0.4759	0.4614	0.5349
50%	10	0.2501	0.5972	0.5276	0.4964	0.5801
50%	12	0.2612	0.6199	0.5622	0.5172	0.6092
50%	14	0.2696	0.6348	0.5873	0.5314	0.6302
50%	16	0.2739	0.6453	0.6071	0.5413	0.6477
60%	4	0.2010	0.4251	0.2742	0.3529	0.3497
60%	6	0.2335	0.5205	0.3880	0.4335	0.4598

Table A.4-4 (cont.)

60%	8	0.2613	0.5890	0.4749	0.4942	0.5394
60%	10	0.2758	0.6224	0.5198	0.5242	0.5783
60%	12	0.2863	0.6445	0.5526	0.5446	0.6065
60%	14	0.2947	0.6595	0.5760	0.5587	0.6265
60%	16	0.2992	0.6693	0.5930	0.5680	0.6417
70%	4	0.2038	0.3987	0.2302	0.3378	0.3090
70%	6	0.2190	0.4559	0.2872	0.3838	0.3666
70%	8	0.2302	0.4917	0.3205	0.4130	0.3986
70%	10	0.2375	0.5116	0.3383	0.4295	0.4147
70%	12	0.2463	0.5296	0.3560	0.4456	0.4315
70%	14	0.2546	0.5445	0.3716	0.4593	0.4467
70%	16	0.2605	0.5566	0.3853	0.4703	0.4605
80%	4	0.2126	0.3459	0.1534	0.3081	0.2307
80%	6	0.2041	0.3523	0.1569	0.3090	0.2335
80%	8	0.2040	0.3592	0.1567	0.3131	0.2321
80%	10	0.2037	0.3607	0.1538	0.3137	0.2279
80%	12	0.2052	0.3643	0.1538	0.3166	0.2279
80%	14	0.2057	0.3670	0.1544	0.3184	0.2289
80%	16	0.2063	0.3699	0.1561	0.3205	0.2311

Table A.4-5. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available in low-dose, low-gradient regions for MLC lagging leaf errors.

Pearson r^2 for MLC lagging leaf error curve ranges vs. # low-dose, low-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.2297	0.4685	0.4252	0.4151	0.4684
20%	6	0.2170	0.4393	0.3808	0.3850	0.4301
20%	8	0.2075	0.4231	0.3584	0.3662	0.4115
20%	10	0.1997	0.4068	0.3408	0.3502	0.3952
20%	12	0.1942	0.3929	0.3276	0.3380	0.3827
20%	14	0.1918	0.3851	0.3207	0.3314	0.3761
20%	16	0.1918	0.3818	0.3190	0.3290	0.3745
25%	4	0.1702	0.4470	0.4011	0.3543	0.4450
25%	6	0.1390	0.3909	0.3278	0.2981	0.3778
25%	8	0.1239	0.3611	0.2882	0.2681	0.3418
25%	10	0.1153	0.3398	0.2641	0.2489	0.3186
25%	12	0.1108	0.3249	0.2488	0.2369	0.3040
25%	14	0.1091	0.3167	0.2413	0.2307	0.2968
25%	16	0.1095	0.3136	0.2396	0.2287	0.2951
30%	4	0.1241	0.4123	0.3860	0.3009	0.4235
30%	6	0.0847	0.3365	0.2950	0.2289	0.3372
30%	8	0.0691	0.3010	0.2486	0.1956	0.2946
30%	10	0.0610	0.2777	0.2202	0.1754	0.2676
30%	12	0.0569	0.2618	0.2020	0.1629	0.2503
30%	14	0.0556	0.2536	0.1938	0.1572	0.2425
30%	16	0.0557	0.2497	0.1913	0.1548	0.2398
40%	4	0.0989	0.3543	0.3768	0.2557	0.3960
40%	6	0.0461	0.2450	0.2563	0.1561	0.2777
40%	8	0.0281	0.1947	0.1924	0.1125	0.2171
40%	10	0.0196	0.1647	0.1540	0.0880	0.1801
40%	12	0.0153	0.1433	0.1283	0.0723	0.1550
40%	14	0.0137	0.1325	0.1161	0.0650	0.1429
40%	16	0.0134	0.1276	0.1114	0.0620	0.1382
50%	4	0.0992	0.3425	0.3750	0.2507	0.3881
50%	6	0.0357	0.1991	0.2229	0.1258	0.2340
50%	8	0.0144	0.1282	0.1378	0.0694	0.1499
50%	10	0.0061	0.0906	0.0911	0.0417	0.1038
50%	12	0.0027	0.0665	0.0626	0.0261	0.0748
50%	14	0.0015	0.0541	0.0485	0.0190	0.0604
50%	16	0.0012	0.0482	0.0425	0.0158	0.0540
60%	4	0.1096	0.3528	0.3798	0.2628	0.3928
60%	6	0.0382	0.1898	0.2046	0.1221	0.2151

Table A.4-5 (cont.)

60%	8	0.0120	0.1042	0.1047	0.0555	0.1155
60%	10	0.0027	0.0583	0.0521	0.0238	0.0619
60%	12	0.0002	0.0320	0.0242	0.0089	0.0323
60%	14	0.0000	0.0199	0.0124	0.0035	0.0189
60%	16	0.0003	0.0142	0.0074	0.0015	0.0129
70%	4	0.1045	0.3262	0.3431	0.2444	0.3583
70%	6	0.0253	0.1367	0.1345	0.0841	0.1477
70%	8	0.0016	0.0432	0.0310	0.0168	0.0409
70%	10	0.0015	0.0081	0.0014	0.0002	0.0048
70%	12	0.0084	0.0000	0.0041	0.0045	0.0008
70%	14	0.0144	0.0024	0.0153	0.0128	0.0071
70%	16	0.0179	0.0059	0.0246	0.0192	0.0135
80%	4	0.0987	0.2810	0.2781	0.2148	0.2958
80%	6	0.0141	0.0769	0.0619	0.0449	0.0742
80%	8	0.0003	0.0068	0.0005	0.0005	0.0029
80%	10	0.0111	0.0027	0.0178	0.0111	0.0090
80%	12	0.0257	0.0206	0.0570	0.0362	0.0387
80%	14	0.0365	0.0385	0.0895	0.0567	0.0648
80%	16	0.0425	0.0509	0.1111	0.0700	0.0826

Table A.4-6. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available high-gradient regions for MLC lagging leaf errors.

Pearson r^2 for MLC lagging leaf error curve ranges vs. high-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.2325	0.5459	0.6644	0.5015	0.6184
20%	6	0.1638	0.4360	0.5568	0.3959	0.5044
20%	8	0.1070	0.3313	0.4255	0.2953	0.3804
20%	10	0.0565	0.2181	0.2820	0.1904	0.2477
20%	12	0.0201	0.1129	0.1463	0.0973	0.1243
20%	14	0.0024	0.0424	0.0539	0.0360	0.0432
20%	16	0.0002	0.0114	0.0132	0.0095	0.0087
25%	4	0.2325	0.5459	0.6644	0.5015	0.6184
25%	6	0.1638	0.4360	0.5568	0.3959	0.5044
25%	8	0.1070	0.3313	0.4255	0.2953	0.3804
25%	10	0.0565	0.2181	0.2820	0.1904	0.2477
25%	12	0.0201	0.1129	0.1463	0.0973	0.1243
25%	14	0.0024	0.0424	0.0539	0.0360	0.0432
25%	16	0.0002	0.0114	0.0132	0.0095	0.0087
30%	4	0.2325	0.5459	0.6644	0.5015	0.6184
30%	6	0.1638	0.4360	0.5568	0.3959	0.5044
30%	8	0.1070	0.3313	0.4255	0.2953	0.3804
30%	10	0.0565	0.2181	0.2820	0.1904	0.2477
30%	12	0.0201	0.1129	0.1463	0.0973	0.1243
30%	14	0.0024	0.0424	0.0539	0.0360	0.0432
30%	16	0.0002	0.0114	0.0132	0.0095	0.0087
40%	4	0.2325	0.5459	0.6644	0.5015	0.6184
40%	6	0.1638	0.4360	0.5568	0.3959	0.5044
40%	8	0.1070	0.3313	0.4255	0.2953	0.3804
40%	10	0.0565	0.2181	0.2820	0.1904	0.2477
40%	12	0.0201	0.1129	0.1463	0.0973	0.1243
40%	14	0.0024	0.0424	0.0539	0.0360	0.0432
40%	16	0.0002	0.0114	0.0132	0.0095	0.0087
50%	4	0.2325	0.5459	0.6644	0.5015	0.6184
50%	6	0.1638	0.4360	0.5568	0.3959	0.5044
50%	8	0.1070	0.3313	0.4255	0.2953	0.3804
50%	10	0.0565	0.2181	0.2820	0.1904	0.2477
50%	12	0.0201	0.1129	0.1463	0.0973	0.1243
50%	14	0.0024	0.0424	0.0539	0.0360	0.0432
50%	16	0.0002	0.0114	0.0132	0.0095	0.0087
60%	4	0.2325	0.5459	0.6644	0.5015	0.6184
60%	6	0.1638	0.4360	0.5568	0.3959	0.5044

Table A.4-6 (cont.)

60%	8	0.1070	0.3313	0.4255	0.2953	0.3804
60%	10	0.0565	0.2181	0.2820	0.1904	0.2477
60%	12	0.0201	0.1129	0.1463	0.0973	0.1243
60%	14	0.0024	0.0424	0.0539	0.0360	0.0432
60%	16	0.0002	0.0114	0.0132	0.0095	0.0087
70%	4	0.2325	0.5459	0.6644	0.5015	0.6184
70%	6	0.1638	0.4360	0.5568	0.3959	0.5044
70%	8	0.1070	0.3313	0.4255	0.2953	0.3804
70%	10	0.0565	0.2181	0.2820	0.1904	0.2477
70%	12	0.0201	0.1129	0.1463	0.0973	0.1243
70%	14	0.0024	0.0424	0.0539	0.0360	0.0432
70%	16	0.0002	0.0114	0.0132	0.0095	0.0087
80%	4	0.2325	0.5459	0.6644	0.5015	0.6184
80%	6	0.1638	0.4360	0.5568	0.3959	0.5044
80%	8	0.1070	0.3313	0.4255	0.2953	0.3804
80%	10	0.0565	0.2181	0.2820	0.1904	0.2477
80%	12	0.0201	0.1129	0.1463	0.0973	0.1243
80%	14	0.0024	0.0424	0.0539	0.0360	0.0432
80%	16	0.0002	0.0114	0.0132	0.0095	0.0087

Table A.4-7. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available high-dose, low-gradient regions for MLC bank shift errors.

Pearson r^2 for MLC bank shift error curve ranges vs. high-dose, low-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.1476	0.1075	0.0820	0.1459	0.0591
20%	6	0.1747	0.1363	0.1015	0.1692	0.0791
20%	8	0.1923	0.1564	0.1145	0.1847	0.0933
20%	10	0.2024	0.1669	0.1206	0.1939	0.1004
20%	12	0.2090	0.1737	0.1243	0.2003	0.1052
20%	14	0.2127	0.1775	0.1261	0.2042	0.1079
20%	16	0.2147	0.1794	0.1267	0.2066	0.1094
25%	4	0.1481	0.0912	0.0706	0.1441	0.0494
25%	6	0.1736	0.1186	0.0894	0.1662	0.0678
25%	8	0.1901	0.1378	0.1020	0.1806	0.0810
25%	10	0.1997	0.1483	0.1082	0.1895	0.0881
25%	12	0.2058	0.1551	0.1118	0.1955	0.0927
25%	14	0.2091	0.1589	0.1135	0.1990	0.0952
25%	16	0.2107	0.1607	0.1140	0.2009	0.0964
30%	4	0.1469	0.0901	0.0637	0.1407	0.0473
30%	6	0.1731	0.1172	0.0827	0.1641	0.0654
30%	8	0.1897	0.1364	0.0954	0.1784	0.0787
30%	10	0.1988	0.1464	0.1016	0.1867	0.0856
30%	12	0.2043	0.1524	0.1050	0.1921	0.0900
30%	14	0.2070	0.1558	0.1064	0.1951	0.0922
30%	16	0.2085	0.1576	0.1069	0.1969	0.0934
40%	4	0.1496	0.1000	0.0620	0.1509	0.0509
40%	6	0.1804	0.1323	0.0833	0.1774	0.0728
40%	8	0.1975	0.1518	0.0963	0.1914	0.0869
40%	10	0.2049	0.1599	0.1016	0.1971	0.0930
40%	12	0.2086	0.1642	0.1041	0.2001	0.0964
40%	14	0.2098	0.1661	0.1047	0.2012	0.0977
40%	16	0.2101	0.1668	0.1046	0.2018	0.0983
50%	4	0.1507	0.0973	0.0607	0.1535	0.0505
50%	6	0.1905	0.1365	0.0861	0.1886	0.0771
50%	8	0.2098	0.1586	0.1000	0.2046	0.0931
50%	10	0.2161	0.1664	0.1050	0.2089	0.0992
50%	12	0.2175	0.1692	0.1063	0.2095	0.1014
50%	14	0.2172	0.1695	0.1060	0.2089	0.1017
50%	16	0.2170	0.1697	0.1057	0.2088	0.1019
60%	4	0.1373	0.0803	0.0523	0.1380	0.0421
60%	6	0.1733	0.1127	0.0735	0.1695	0.0641

Table A.4-7 (cont.)

60%	8	0.1925	0.1322	0.0855	0.1859	0.0781
60%	10	0.1997	0.1407	0.0905	0.1918	0.0843
60%	12	0.2021	0.1447	0.0922	0.1938	0.0871
60%	14	0.2022	0.1456	0.0923	0.1936	0.0876
60%	16	0.2023	0.1461	0.0922	0.1937	0.0879
70%	4	0.1201	0.0639	0.0435	0.1207	0.0314
70%	6	0.1489	0.0865	0.0594	0.1458	0.0466
70%	8	0.1638	0.0986	0.0672	0.1589	0.0552
70%	10	0.1701	0.1039	0.0702	0.1645	0.0590
70%	12	0.1726	0.1065	0.0712	0.1668	0.0606
70%	14	0.1727	0.1067	0.0709	0.1669	0.0606
70%	16	0.1728	0.1068	0.0705	0.1670	0.0605
80%	4	0.0531	0.0167	0.0113	0.0578	0.0031
80%	6	0.0651	0.0232	0.0161	0.0689	0.0061
80%	8	0.0706	0.0259	0.0180	0.0739	0.0076
80%	10	0.0721	0.0265	0.0183	0.0752	0.0080
80%	12	0.0715	0.0259	0.0178	0.0745	0.0078
80%	14	0.0706	0.0252	0.0174	0.0735	0.0075
80%	16	0.0701	0.0246	0.0171	0.0730	0.0073

Table A.4-8. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available low-dose, low-gradient regions for MLC bank shift errors.

Pearson r^2 for MLC bank shift error curve ranges vs. low-dose, low-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.0034	0.0004	0.0047	0.0031	0.0034
20%	6	0.0008	0.0003	0.0101	0.0007	0.0093
20%	8	0.0002	0.0011	0.0134	0.0001	0.0133
20%	10	0.0000	0.0018	0.0158	0.0000	0.0160
20%	12	0.0000	0.0022	0.0171	0.0000	0.0177
20%	14	0.0001	0.0025	0.0179	0.0001	0.0187
20%	16	0.0001	0.0025	0.0180	0.0001	0.0191
25%	4	0.0024	0.0001	0.0082	0.0019	0.0064
25%	6	0.0001	0.0030	0.0172	0.0000	0.0161
25%	8	0.0001	0.0062	0.0231	0.0002	0.0231
25%	10	0.0006	0.0083	0.0271	0.0008	0.0274
25%	12	0.0011	0.0098	0.0298	0.0013	0.0308
25%	14	0.0015	0.0108	0.0316	0.0017	0.0331
25%	16	0.0016	0.0111	0.0324	0.0018	0.0342
30%	4	0.0005	0.0011	0.0150	0.0002	0.0104
30%	6	0.0006	0.0080	0.0302	0.0010	0.0254
30%	8	0.0028	0.0144	0.0407	0.0035	0.0365
30%	10	0.0051	0.0188	0.0476	0.0059	0.0436
30%	12	0.0070	0.0223	0.0528	0.0078	0.0494
30%	14	0.0084	0.0246	0.0566	0.0092	0.0536
30%	16	0.0090	0.0255	0.0583	0.0098	0.0558
40%	4	0.0024	0.0073	0.0354	0.0024	0.0227
40%	6	0.0167	0.0296	0.0737	0.0161	0.0556
40%	8	0.0331	0.0525	0.1054	0.0321	0.0849
40%	10	0.0478	0.0704	0.1293	0.0463	0.1068
40%	12	0.0601	0.0849	0.1481	0.0583	0.1252
40%	14	0.0682	0.0943	0.1604	0.0662	0.1376
40%	16	0.0717	0.0981	0.1660	0.0696	0.1438
50%	4	0.0071	0.0143	0.0484	0.0069	0.0316
50%	6	0.0374	0.0547	0.1080	0.0359	0.0816
50%	8	0.0721	0.0974	0.1605	0.0692	0.1284
50%	10	0.1030	0.1318	0.2015	0.0989	0.1651
50%	12	0.1293	0.1607	0.2346	0.1245	0.1967
50%	14	0.1463	0.1793	0.2557	0.1410	0.2178
50%	16	0.1543	0.1879	0.2658	0.1489	0.2288
60%	4	0.0114	0.0215	0.0581	0.0115	0.0394
60%	6	0.0540	0.0785	0.1328	0.0529	0.1033

Table A.4-8 (cont.)

60%	8	0.1006	0.1360	0.1965	0.0978	0.1613
60%	10	0.1441	0.1840	0.2485	0.1394	0.2083
60%	12	0.1822	0.2250	0.2912	0.1762	0.2490
60%	14	0.2072	0.2517	0.3182	0.2004	0.2762
60%	16	0.2203	0.2655	0.3322	0.2135	0.2914
70%	4	0.0183	0.0312	0.0704	0.0185	0.0508
70%	6	0.0826	0.1144	0.1692	0.0808	0.1387
70%	8	0.1507	0.1976	0.2518	0.1459	0.2174
70%	10	0.2078	0.2596	0.3104	0.1999	0.2732
70%	12	0.2530	0.3064	0.3518	0.2430	0.3152
70%	14	0.2787	0.3326	0.3731	0.2676	0.3388
70%	16	0.2918	0.3455	0.3835	0.2806	0.3515
80%	4	0.0438	0.0631	0.1076	0.0421	0.0879
80%	6	0.1523	0.1951	0.2467	0.1445	0.2195
80%	8	0.2462	0.3046	0.3438	0.2326	0.3173
80%	10	0.3110	0.3711	0.3981	0.2929	0.3724
80%	12	0.3521	0.4111	0.4265	0.3319	0.4044
80%	14	0.3710	0.4285	0.4360	0.3501	0.4175
80%	16	0.3792	0.4354	0.4387	0.3585	0.4234

Table A.4-9. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available high-gradient regions for MLC bank shift errors.

Pearson r^2 for MLC bank shift error curve ranges vs. high-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.1806	0.1728	0.1318	0.1689	0.1471
20%	6	0.1365	0.1262	0.0996	0.1282	0.1149
20%	8	0.0965	0.0835	0.0714	0.0914	0.0847
20%	10	0.0571	0.0482	0.0464	0.0549	0.0588
20%	12	0.0254	0.0202	0.0259	0.0247	0.0349
20%	14	0.0064	0.0043	0.0110	0.0064	0.0164
20%	16	0.0007	0.0002	0.0048	0.0007	0.0077
25%	4	0.1806	0.1728	0.1318	0.1689	0.1471
25%	6	0.1365	0.1262	0.0996	0.1282	0.1149
25%	8	0.0965	0.0835	0.0714	0.0914	0.0847
25%	10	0.0571	0.0482	0.0464	0.0549	0.0588
25%	12	0.0254	0.0202	0.0259	0.0247	0.0349
25%	14	0.0064	0.0043	0.0110	0.0064	0.0164
25%	16	0.0007	0.0002	0.0048	0.0007	0.0077
30%	4	0.1806	0.1728	0.1318	0.1689	0.1471
30%	6	0.1365	0.1262	0.0996	0.1282	0.1149
30%	8	0.0965	0.0835	0.0714	0.0914	0.0847
30%	10	0.0571	0.0482	0.0464	0.0549	0.0588
30%	12	0.0254	0.0202	0.0259	0.0247	0.0349
30%	14	0.0064	0.0043	0.0110	0.0064	0.0164
30%	16	0.0007	0.0002	0.0048	0.0007	0.0077
40%	4	0.1806	0.1728	0.1318	0.1689	0.1471
40%	6	0.1365	0.1262	0.0996	0.1282	0.1149
40%	8	0.0965	0.0835	0.0714	0.0914	0.0847
40%	10	0.0571	0.0482	0.0464	0.0549	0.0588
40%	12	0.0254	0.0202	0.0259	0.0247	0.0349
40%	14	0.0064	0.0043	0.0110	0.0064	0.0164
40%	16	0.0007	0.0002	0.0048	0.0007	0.0077
50%	4	0.1806	0.1728	0.1318	0.1689	0.1471
50%	6	0.1365	0.1262	0.0996	0.1282	0.1149
50%	8	0.0965	0.0835	0.0714	0.0914	0.0847
50%	10	0.0571	0.0482	0.0464	0.0549	0.0588
50%	12	0.0254	0.0202	0.0259	0.0247	0.0349
50%	14	0.0064	0.0043	0.0110	0.0064	0.0164
50%	16	0.0007	0.0002	0.0048	0.0007	0.0077
60%	4	0.1806	0.1728	0.1318	0.1689	0.1471
60%	6	0.1365	0.1262	0.0996	0.1282	0.1149

Table A.4-9 (cont.)

60%	8	0.0965	0.0835	0.0714	0.0914	0.0847
60%	10	0.0571	0.0482	0.0464	0.0549	0.0588
60%	12	0.0254	0.0202	0.0259	0.0247	0.0349
60%	14	0.0064	0.0043	0.0110	0.0064	0.0164
60%	16	0.0007	0.0002	0.0048	0.0007	0.0077
70%	4	0.1806	0.1728	0.1318	0.1689	0.1471
70%	6	0.1365	0.1262	0.0996	0.1282	0.1149
70%	8	0.0965	0.0835	0.0714	0.0914	0.0847
70%	10	0.0571	0.0482	0.0464	0.0549	0.0588
70%	12	0.0254	0.0202	0.0259	0.0247	0.0349
70%	14	0.0064	0.0043	0.0110	0.0064	0.0164
70%	16	0.0007	0.0002	0.0048	0.0007	0.0077
80%	4	0.1806	0.1728	0.1318	0.1689	0.1471
80%	6	0.1365	0.1262	0.0996	0.1282	0.1149
80%	8	0.0965	0.0835	0.0714	0.0914	0.0847
80%	10	0.0571	0.0482	0.0464	0.0549	0.0588
80%	12	0.0254	0.0202	0.0259	0.0247	0.0349
80%	14	0.0064	0.0043	0.0110	0.0064	0.0164
80%	16	0.0007	0.0002	0.0048	0.0007	0.0077

Table A.4-10. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available high-dose, low-gradient regions for collimator rotation errors.

Pearson r^2 for MLC collimator rotation error curve ranges vs. high-dose, low-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.2413	0.1811	0.2204	0.2132	0.2250
20%	6	0.2715	0.2029	0.2443	0.2368	0.2493
20%	8	0.2887	0.2165	0.2584	0.2505	0.2637
20%	10	0.2966	0.2237	0.2635	0.2564	0.2689
20%	12	0.2992	0.2266	0.2651	0.2582	0.2706
20%	14	0.3018	0.2293	0.2674	0.2600	0.2730
20%	16	0.3039	0.2312	0.2699	0.2617	0.2755
25%	4	0.2319	0.1712	0.1999	0.1962	0.2034
25%	6	0.2602	0.1926	0.2226	0.2196	0.2265
25%	8	0.2791	0.2077	0.2385	0.2353	0.2426
25%	10	0.2883	0.2160	0.2457	0.2426	0.2501
25%	12	0.2921	0.2199	0.2489	0.2457	0.2533
25%	14	0.2960	0.2235	0.2523	0.2488	0.2568
25%	16	0.2989	0.2261	0.2554	0.2515	0.2600
30%	4	0.2051	0.1476	0.1740	0.1697	0.1774
30%	6	0.2352	0.1715	0.1970	0.1948	0.2009
30%	8	0.2566	0.1889	0.2142	0.2128	0.2184
30%	10	0.2664	0.1971	0.2221	0.2208	0.2265
30%	12	0.2712	0.2014	0.2262	0.2246	0.2307
30%	14	0.2762	0.2059	0.2305	0.2287	0.2350
30%	16	0.2796	0.2088	0.2339	0.2318	0.2384
40%	4	0.1977	0.1392	0.1658	0.1616	0.1693
40%	6	0.2247	0.1607	0.1866	0.1844	0.1910
40%	8	0.2464	0.1784	0.2046	0.2025	0.2096
40%	10	0.2564	0.1866	0.2127	0.2105	0.2179
40%	12	0.2607	0.1903	0.2160	0.2137	0.2212
40%	14	0.2652	0.1944	0.2197	0.2173	0.2248
40%	16	0.2679	0.1967	0.2223	0.2197	0.2274
50%	4	0.2050	0.1446	0.1780	0.1686	0.1813
50%	6	0.2381	0.1709	0.2045	0.1967	0.2093
50%	8	0.2587	0.1874	0.2226	0.2140	0.2284
50%	10	0.2662	0.1934	0.2287	0.2197	0.2348
50%	12	0.2677	0.1946	0.2292	0.2201	0.2353
50%	14	0.2702	0.1969	0.2309	0.2218	0.2369
50%	16	0.2714	0.1978	0.2320	0.2227	0.2380
60%	4	0.2070	0.1488	0.1852	0.1722	0.1886

Table A.4-10 (cont.)

60%	6	0.2361	0.1711	0.2094	0.1968	0.2143
60%	8	0.2533	0.1845	0.2247	0.2110	0.2307
60%	10	0.2605	0.1900	0.2309	0.2165	0.2373
60%	12	0.2638	0.1924	0.2329	0.2184	0.2394
60%	14	0.2666	0.1947	0.2348	0.2202	0.2412
60%	16	0.2674	0.1951	0.2355	0.2207	0.2419
70%	4	0.1868	0.1306	0.1694	0.1552	0.1721
70%	6	0.2068	0.1451	0.1871	0.1729	0.1911
70%	8	0.2170	0.1518	0.1977	0.1815	0.2028
70%	10	0.2230	0.1554	0.2037	0.1862	0.2091
70%	12	0.2277	0.1585	0.2076	0.1894	0.2132
70%	14	0.2317	0.1616	0.2108	0.1922	0.2165
70%	16	0.2340	0.1632	0.2129	0.1937	0.2186
80%	4	0.1088	0.0664	0.1011	0.0853	0.1011
80%	6	0.1137	0.0690	0.1066	0.0905	0.1074
80%	8	0.1164	0.0701	0.1110	0.0931	0.1125
80%	10	0.1176	0.0704	0.1130	0.0941	0.1148
80%	12	0.1184	0.0708	0.1137	0.0945	0.1157
80%	14	0.1186	0.0710	0.1139	0.0944	0.1159
80%	16	0.1184	0.0709	0.1137	0.0940	0.1158

Table A.4-11. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available low-dose, low-gradient regions for collimator rotation errors.

Pearson r^2 for MLC collimator rotation error curve ranges vs. low-dose, low-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.3455	0.3090	0.2557	0.2718	0.2618
20%	6	0.3509	0.3157	0.2550	0.2755	0.2600
20%	8	0.3565	0.3211	0.2569	0.2799	0.2610
20%	10	0.3606	0.3255	0.2593	0.2840	0.2628
20%	12	0.3663	0.3313	0.2624	0.2893	0.2653
20%	14	0.3726	0.3373	0.2666	0.2948	0.2691
20%	16	0.3789	0.3428	0.2714	0.3004	0.2735
25%	4	0.3110	0.2770	0.2185	0.2361	0.2232
25%	6	0.3084	0.2780	0.2101	0.2330	0.2135
25%	8	0.3101	0.2804	0.2081	0.2340	0.2106
25%	10	0.3131	0.2837	0.2101	0.2370	0.2120
25%	12	0.3196	0.2904	0.2137	0.2427	0.2150
25%	14	0.3272	0.2976	0.2183	0.2492	0.2190
25%	16	0.3345	0.3039	0.2234	0.2557	0.2238
30%	4	0.2719	0.2430	0.1844	0.2008	0.1887
30%	6	0.2596	0.2370	0.1673	0.1896	0.1704
30%	8	0.2585	0.2375	0.1622	0.1882	0.1644
30%	10	0.2606	0.2396	0.1630	0.1902	0.1647
30%	12	0.2674	0.2460	0.1665	0.1957	0.1675
30%	14	0.2758	0.2538	0.1712	0.2025	0.1716
30%	16	0.2832	0.2602	0.1757	0.2087	0.1758
40%	4	0.2338	0.2150	0.1501	0.1673	0.1539
40%	6	0.1992	0.1909	0.1167	0.1386	0.1193
40%	8	0.1868	0.1823	0.1035	0.1279	0.1054
40%	10	0.1848	0.1810	0.1008	0.1261	0.1020
40%	12	0.1885	0.1849	0.1009	0.1285	0.1015
40%	14	0.1946	0.1909	0.1031	0.1330	0.1032
40%	16	0.2009	0.1964	0.1063	0.1381	0.1060
50%	4	0.2218	0.2078	0.1426	0.1578	0.1461
50%	6	0.1686	0.1687	0.0966	0.1150	0.0987
50%	8	0.1394	0.1450	0.0733	0.0921	0.0746
50%	10	0.1278	0.1348	0.0646	0.0831	0.0654
50%	12	0.1241	0.1319	0.0603	0.0799	0.0606
50%	14	0.1246	0.1326	0.0591	0.0800	0.0590
50%	16	0.1273	0.1350	0.0598	0.0821	0.0594
60%	4	0.2197	0.2090	0.1430	0.1568	0.1464
60%	6	0.1532	0.1576	0.0882	0.1040	0.0901

Table A.4-11 (cont.)

60%	8	0.1122	0.1220	0.0578	0.0727	0.0588
60%	10	0.0921	0.1035	0.0444	0.0579	0.0449
60%	12	0.0822	0.0945	0.0369	0.0505	0.0369
60%	14	0.0776	0.0904	0.0331	0.0472	0.0328
60%	16	0.0765	0.0891	0.0317	0.0464	0.0312
70%	4	0.2018	0.1929	0.1295	0.1425	0.1325
70%	6	0.1225	0.1292	0.0669	0.0806	0.0682
70%	8	0.0707	0.0816	0.0317	0.0426	0.0322
70%	10	0.0446	0.0557	0.0168	0.0249	0.0167
70%	12	0.0312	0.0419	0.0093	0.0160	0.0091
70%	14	0.0248	0.0350	0.0061	0.0120	0.0057
70%	16	0.0223	0.0321	0.0048	0.0105	0.0044
80%	4	0.1475	0.1413	0.0896	0.0982	0.0910
80%	6	0.0628	0.0699	0.0281	0.0356	0.0281
80%	8	0.0210	0.0290	0.0048	0.0086	0.0046
80%	10	0.0061	0.0119	0.0001	0.0011	0.0001
80%	12	0.0014	0.0052	0.0007	0.0000	0.0009
80%	14	0.0002	0.0027	0.0021	0.0004	0.0025
80%	16	0.0000	0.0018	0.0031	0.0008	0.0037

Table A.4-12. Pearson correlations for error curve ranges for listed gamma criterion vs. the number of points available high-gradient regions for collimator rotation errors.

Pearson r^2 for MLC collimator rotation error curve ranges vs. high-gradient points						
Dose threshold [%]	Gradient threshold [cGy/mm]	3%/3mm TH 10 (G) 95% PP	3%/3mm TH 10 (L) 95% PP	3%/3mm TH 50 (G) 95% PP	2%/2mm TH 10 (G) 90% PP	3%/3mm TH 50 (G) 95% PP
20%	4	0.0631	0.0474	0.0585	0.0459	0.0643
20%	6	0.0219	0.0174	0.0205	0.0128	0.0242
20%	8	0.0011	0.0011	0.0011	0.0000	0.0022
20%	10	0.0079	0.0054	0.0060	0.0137	0.0040
20%	12	0.0459	0.0329	0.0381	0.0571	0.0327
20%	14	0.1031	0.0749	0.0893	0.1172	0.0812
20%	16	0.1586	0.1160	0.1418	0.1750	0.1315
25%	4	0.0631	0.0474	0.0585	0.0459	0.0643
25%	6	0.0219	0.0174	0.0205	0.0128	0.0242
25%	8	0.0011	0.0011	0.0011	0.0000	0.0022
25%	10	0.0079	0.0054	0.0060	0.0137	0.0040
25%	12	0.0459	0.0329	0.0381	0.0571	0.0327
25%	14	0.1031	0.0749	0.0893	0.1172	0.0812
25%	16	0.1586	0.1160	0.1418	0.1750	0.1315
30%	4	0.0631	0.0474	0.0585	0.0459	0.0643
30%	6	0.0219	0.0174	0.0205	0.0128	0.0242
30%	8	0.0011	0.0011	0.0011	0.0000	0.0022
30%	10	0.0079	0.0054	0.0060	0.0137	0.0040
30%	12	0.0459	0.0329	0.0381	0.0571	0.0327
30%	14	0.1031	0.0749	0.0893	0.1172	0.0812
30%	16	0.1586	0.1160	0.1418	0.1750	0.1315
40%	4	0.0631	0.0474	0.0585	0.0459	0.0643
40%	6	0.0219	0.0174	0.0205	0.0128	0.0242
40%	8	0.0011	0.0011	0.0011	0.0000	0.0022
40%	10	0.0079	0.0054	0.0060	0.0137	0.0040
40%	12	0.0459	0.0329	0.0381	0.0571	0.0327
40%	14	0.1031	0.0749	0.0893	0.1172	0.0812
40%	16	0.1586	0.1160	0.1418	0.1750	0.1315
50%	4	0.0631	0.0474	0.0585	0.0459	0.0643
50%	6	0.0219	0.0174	0.0205	0.0128	0.0242
50%	8	0.0011	0.0011	0.0011	0.0000	0.0022
50%	10	0.0079	0.0054	0.0060	0.0137	0.0040
50%	12	0.0459	0.0329	0.0381	0.0571	0.0327
50%	14	0.1031	0.0749	0.0893	0.1172	0.0812
50%	16	0.1586	0.1160	0.1418	0.1750	0.1315
60%	4	0.0631	0.0474	0.0585	0.0459	0.0643

Table A.4-12 (cont.)

60%	6	0.0219	0.0174	0.0205	0.0128	0.0242
60%	8	0.0011	0.0011	0.0011	0.0000	0.0022
60%	10	0.0079	0.0054	0.0060	0.0137	0.0040
60%	12	0.0459	0.0329	0.0381	0.0571	0.0327
60%	14	0.1031	0.0749	0.0893	0.1172	0.0812
60%	16	0.1586	0.1160	0.1418	0.1750	0.1315
70%	4	0.0631	0.0474	0.0585	0.0459	0.0643
70%	6	0.0219	0.0174	0.0205	0.0128	0.0242
70%	8	0.0011	0.0011	0.0011	0.0000	0.0022
70%	10	0.0079	0.0054	0.0060	0.0137	0.0040
70%	12	0.0459	0.0329	0.0381	0.0571	0.0327
70%	14	0.1031	0.0749	0.0893	0.1172	0.0812
70%	16	0.1586	0.1160	0.1418	0.1750	0.1315
80%	4	0.0631	0.0474	0.0585	0.0459	0.0643
80%	6	0.0219	0.0174	0.0205	0.0128	0.0242
80%	8	0.0011	0.0011	0.0011	0.0000	0.0022
80%	10	0.0079	0.0054	0.0060	0.0137	0.0040
80%	12	0.0459	0.0329	0.0381	0.0571	0.0327
80%	14	0.1031	0.0749	0.0893	0.1172	0.0812
80%	16	0.1586	0.1160	0.1418	0.1750	0.1315

References

1. Sjölin M, Edmund JM. Incorrect dosimetric leaf separation in IMRT and VMAT treatment planning: Clinical impact and correlation with pretreatment quality assurance. *Phys Medica*. 2016;32(7):918-925. doi:10.1016/j.ejmp.2016.06.012.

2. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. *Med Phys*. 2018;45(4):e53-e83. doi:10.1002/mp.12810.

3. Nelms BE, Chan MF, Jarry G, et al. Evaluating IMRT and VMAT dose accuracy: practical examples of failure to detect systematic errors when applying a commonly used metric and action levels. *Med Phys*. 2013;40(11):111722. doi:10.1118/1.4826166.

4. Stambaugh C, Ezzell G. A clinically relevant IMRT QA workflow: Design and validation. *Med Phys*. 2018;45(4):1391-1399. doi:10.1002/mp.12838.

5. Moran JM, Radawski J, Fraass BA. A dose gradient analysis tool for IMRT QA. *J Appl Clin Med Phys*. 2005;6(2):62-73. doi:10.1120/jacmp.2024.25338.

6. Childress NL, Rosen II. The design and testing of novel clinical parameters for dose comparison. *Int J Radiat Oncol Biol Phys*. 2003;56(5):1464-1479. doi:10.1016/S0360-3016(03)00430-9.

7. Sauer O a, Wilbert J. Measurement of output factors for small photon beams. *Med Phys*. 2007;34(6):1983. doi:10.1118/1.2734383.

8. Bakai A, Alber M, Fridtjof N. A revision of the γ-evaluation concept for the comparison of dose distributions. *Phys Med Biol*. 2003;48(21):3543. doi:10.1088/0031-9155/48/21/006.

9. Boggula R, Lorenz F, Mueller L, et al. Experimental validation of a commercial 3D dose verification system for intensity-modulated arc therapies. *Phys Med Biol*. 2010;55(19):5619-5633. doi:10.1088/0031-9155/55/19/001.

10. Nakaguchi Y, Araki F, Maruyama M, Saiga S. Dose verification of IMRT by use of a COMPASS transmission detector. *Radiol Phys Technol*. 2012;5(1):63-70. doi:10.1007/s12194-011-0137-y.

11. Nelms BE, Opp D, Robinson J, et al. VMAT QA: Measurement-guided 4D dose reconstruction on a patient. *Med Phys*. 2012;39(7):4228-4238. doi:10.1118/1.4729709.

12. Olch AJ. Evaluation of the accuracy of 3DVH software estimates of dose to virtual ion chamber and film in composite IMRT QA. *Med Phys*. 2011;39(1):81-86. doi:10.1118/1.3666771.

13. Opp D, Nelms BE, Zhang G, Stevens C, Feygelman V. Validation of measurement-guided 3D VMAT dose reconstruction on a heterogeneous anthropomorphic phantom. *J Appl Clin Med Phys*. 2013;14(4):70-84. doi:10.1120/jacmp.v14i4.4154.

14. Renner WD, Norton K, Holmes T. A method for deconvolution of integrated electronic portal images to obtain incident fluence for dose reconstruction. *J Appl Clin Med Phys*. 2005;6(4):22-39. doi:10.1120/jacmp.v6i4.2104.

15. Ezzell G a, Galvin JM, Low D, et al. Guidance document on delivery, treatment planning, and clinical implementation of IMRT: report of the IMRT Subcommittee of the AAPM Radiation Therapy Committee. *Med Phys*. 2003;30(8):2089-2115.

doi:10.1118/1.1591194.

16. Kruse JJ. On the insensitivity of single field planar dosimetry to IMRT inaccuracies. *Med Phys*. 2010;37(6):2516. doi:10.1118/1.3425781.

17. Nelms B, Jarry G, Chan M, Hampton C, Watanabe Y, Feygelman V. Real-world examples of sensitivity failures of the 3%/3mm pass rate metric and published action levels when used in IMRT/VMAT system commissioning. *J Phys Conf Ser*. 2013;444:012086. doi:10.1088/1742-6596/444/1/012086.

18. Nelms BE, Zhen H, Tomé W a. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. *Med Phys*. 2011;38(2):1037-1044. doi:10.1118/1.3544657.

19. McKenzie EM, Balter P a, Stingo FC, Jones J, Followill DS, Kry SF. Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans. *Med Phys*. 2014;41(12):121702. doi:10.1118/1.4899177.

20. Steers JM, Fraass B a. IMRT QA: Selecting gamma criteria based on error detection sensitivity. *Med Phys*. 2016;43(4):1982-1994. doi:10.1118/1.4943953.

21. Kry SF, Molineu A, Kerns JR, et al. Institutional patient-specific IMRT QA does not predict unacceptable plan delivery. *Int J Radiat Oncol Biol Phys*. 2014;90(5):1195-1201. doi:10.1016/j.ijrobp.2014.08.334.

22. Heilemann G, Poppe B, Laub W. On the sensitivity of common gamma-index evaluation methods to MLC misalignments in Rapidarc quality assurance. *Med Phys*. 2013;40(3):031702. doi:10.1118/1.4789580.

23. Stasi M, Bresciani S, Miranti A, Maggio A, Sapino V, Gabriele P. Pretreatment patient-specific IMRT quality assurance : A correlation study. *Med Phys*. 2012;39(12):7626-7634. doi:10.1118/1.4767763.

24. Jin X, Yan H, Han C, Zhou Y, Yi J, Xie C. Correlation between gamma index passing rate and clinical dosimetric difference for pre-treatment 2D and 3D volumetric modulated arc therapy dosimetric verification. *Br J Radiol*. 2015;88(1047). doi:10.1259/bjr.20140577.

25. Templeton AK, Chu JCH, Turian J V. The sensitivity of ArcCHECK-based gamma analysis to manufactured errors in helical tomotherapy radiation delivery. *J Appl Clin Med Phys*. 2015;16(1):4814. doi:10.1120/jacmp.v16i1.4814.

26. McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. *Med Phys*. 2010;37(2):505-515. doi:10.1118/1.3276775.

27. Masi L, Doro R, Favuzza V, Cipressi S, Livi L. Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy. *Med Phys*. 2013;40(7):071718. doi:10.1118/1.4810969.

Chapter V.

A new method for comparing IMRT QA dose distributions: the gradient-dose segmented analysis method

V.A. Motivation

In previous chapters we have explored various factors affecting the sensitivity of current IMRT QA comparisons in conjunction with the gamma comparison, showing that the same gamma criterion across different devices has different implications for error sensitivity. Additionally, large errors have the potential to go undetected with the most commonly used gamma criteria[1–7]. Beyond this, the gamma comparison results are difficult to interpret, and the final result of the comparison, the percent pixels passing, has been shown to lack clinical relevance[2,8–12].

In order to overcome these limitations, the use of "delivered" DVHs have become a subject of interest. These methods make use of back projection calculations that attempt to transform the dose from the measurement in the phantom/array back on to the patient geometry. This allows the user to visualize dose differences between measurement and calculation in both 3D and in DVH format in order to determine if the differences in the delivered plan are clinically meaningful[13–19].

While the use of delivered DVHs offers a unique way to analyze IMRT QA results, these methods have limitations in predicting reconstructed doses and additionally, present a clinical workflow challenge due to the time that would be required if this analysis was performed for each IMRT QA measurement. As IMRT QA already requires a large amount of clinical effort, increasing the amount of time spent performing IMRT QA, and any associated analysis, could prove onerous for the clinical physics team, especially in smaller clinics. The use of the gamma comparison has thus continued to be the standard in the field for IMRT QA analysis[20,21], as it is an efficient comparison technique and has a documented history of use in the field.

While the binary implementation of the gamma comparison has been labeled an “insensitive” metric, it is important to realize this is highly dependent on the gamma criteria chosen to evaluate IMRT QA comparisons. The user has five different parameters to select that can adjust the sensitivity of the test. On top of the many parameters choices available, the interpretation of the gamma passing rate is difficult and high gamma passing rates do not always indicate good agreement between measurement and calculation.

Other than tightening gamma parameters, several other groups have investigated methods to improve our understanding of IMRT QA results by looking at gamma comparisons and gamma values in a slightly different manner. Wootten *et al.* proposed a modified way of analyzing gamma maps by training a machine-learning model on radiomics features in error-induced gamma maps[22], highlighting their method had improved AUC values for ROC analyses when compared to clinical gamma comparisons. Valdes *et al.* also utilized machine learning in an effort to produce “predicted” passing rates and flag plans that had passing rates lower than the predicted, or expected, passing rates[23]. However, these methods still make use of a metric that is fundamentally difficult to interpret when used as a binary metric (i.e. points pass or fail).

In looking to design new analysis techniques it is important to consider the feasibility of any new method for clinical implementation. In addition to this, it is important to ask what size and types of errors we want to be sensitive to in pre-treatment QA. Such a new method should thus have the following characteristics:

- Quick to calculate and easy to interpret such that it will not increase the clinical physics workload associated with IMRT QA.
- Provide meaningful and sensitive results (ex. a plan with a 5% overall dose error ideally would be correctly labeled by the analysis technique as having a 5% dose error).

- Independent of user-selected input parameters/criteria that can modify comparison sensitivity, such as the selection of DTA and percent dose difference criteria, dose thresholds, and normalization criteria

The work in previous chapters has shown not only that sensitivity with a universal gamma criterion between devices can be different, but also that gamma values can be remarkably low along the highest plan dose gradients in the presence of induced errors. Additionally, results from Chapter III suggest that low dose points, especially in the ArcCHECK geometry, may severely inhibit gamma sensitivity. Thus, the number of points available in IMRT QA comparisons that exist in more sensitive regions of the field (i.e. high-dose, low-gradient points) may be outnumbered by the large number of high-gradient and low-dose points that can easily pass the gamma comparison. It is possible that more meaningful comparisons could result by segmenting the regions of the comparison into different point groupings, and analyzing these regions separately. This idea of analyzing only particular regions of the field was utilized by Childress and Rosen in the development of the NAT index, as well[24]. However, many regions of the field were entirely excluded from their method and their exclusion/inclusion criteria appeared to be chosen arbitrarily.

With this is mind, a new method is presented which segments regions of the dose distribution in the phantom geometries based on dose and gradient thresholds. The high-dose low-gradient regions of the field are likely to approximate regions within the PTV in the treatment plan, while the low-dose low-gradient regions of the field may approximate a ring-like structure outside the PTV as shown in Figure 5-1.

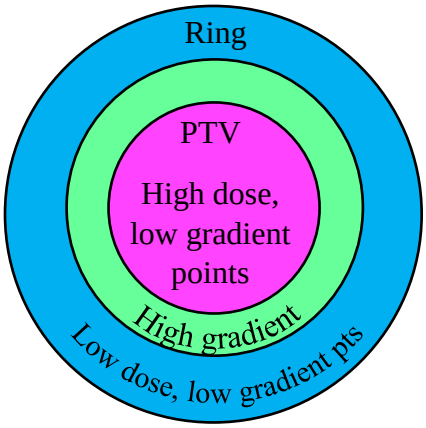

Figure 5-1. Different regions in phantom dose distributions segmented by dose and gradient may approximate the PTV and ring structures outside the high gradient regions of the field.

The aim of this chapter is to investigate if dose difference histogram statistics in these segmented point groupings (illustrated in Figure 5-1) in the phantom geometry can predict changes in DVH metrics in the patient geometry, and in particular if high-dose low-gradient points can predict changes in PTV metrics. This analysis is performed for many different combinations of dose and gradient thresholds in order to segment points into separate groupings (high gradient, high-dose low-gradient, and low-dose low-gradient). This method is investigated for both IMRT and VMAT cases on three unique detector geometries – ArcCHECK, MapCHECK, and Delta 4. It is expected that these thresholds may vary between IMRT and VMAT cases as well as for each studied measurement geometry.

V.B. Methods

As in the previous chapters, the method investigated here is studied using error-free calculations compared to calculations with induced errors of known magnitude. This offers the advantage of being able to remove uncertainties that can be present in real QA measurements, such as delivery uncertainties, machine output fluctuations, device measurement limitations, set-up errors, etc. Additionally, this allows the comparison of DVHs using the patient geometry between error-free and error-induced plans as the ground truth in order to determine the true dose differences to the patient in the presence of a given error.

V.B.1. Ground truth – DVH differences in the patient geometry

For this portion of the work, the previously generated error-induced plans were compared to the error free plans for the same cohort of 20 IMRT and 20 VMAT cases described in Chapter III. DVHs were generated in the Eclipse TPS for error-free and error-induced plans and DVH metrics were extracted from Eclipse utilizing the ESAPI functionality in C# to pull the desired DVH metrics. These endpoints were saved in a text file and parsed by an in-house MATLAB code that concatenates all the DVH metrics into one Excel spreadsheet and calculates the percent dose differences between the error-free and error-induced plans. Since the dose differences between error-free and error-induced plans are simulated in calculations only, the differences between the DVH metrics here are considered the ground truth in our comparisons.

Nine DVH metrics were extracted for every error-free and error-induced IMRT and VMAT plan. Evaluated PTV metrics included the dose to 95% of the PTV volume ($D_{95\%}$), the dose to 99% of the volume ($D_{99\%}$, used as a surrogate for minimum dose), the maximum dose to the PTV volume, and the mean dose to the PTV. Ring structures were created at varying distances from the PTV using expansions of the target volume at 0.5 cm, 1 cm, 2 cm, and 3 cm, as shown in Figure 5-2 for one of the cases from the VMAT cohort. As a note, each ring is an annulus of the previous ring, thus the 3 cm ring is a one centimeter expansion of the 2 cm ring. Ring structure DVH metrics included the maximum point dose for the ring closest to the PTV (0.5 cm), and the mean dose differences for each annulus.

Ring structures were chosen as a simplified approximation of normal tissues and OARs in a plan, as it would be onerous to evaluate dose differences to each OAR in every plan. Several different ring distances were studied since the importance of dose differences at increasing distances away from the PTV changes. In the rings closer to the PTV (0.5 cm and 1 cm), dose differences may be more concerning as any organs-at-risk (OARs) in this region are likely to be receiving higher doses and may already be close to their dose limits. Likewise, the more peripheral rings (2 cm

and 3 cm) are likely to contain OARs that are receiving lower doses and as such, dose differences in these rings are likely to be less critical than in the closer rings. Another consideration in studying dose differences in these various ring structures is that our ability to predict dose changes in any given ring likely also depends on the conformity of the original plan. Thus, it is possible that for highly non-conformal plans dose differences in the closest rings will mimic those seen in the PTV.

We believe that the PTV DVH metrics will be most important and more straightforward to predict in the IMRT QA process and thus study the changes in ring structures here as a secondary metric. The main purpose of IMRT QA is to determine, prior to treatment, if large dose differences exist in a delivered plan that could adversely affect the patient. These larger errors are likely to be more easily flagged by evaluating the high dose regions of the field, as most delivery errors would occur in-field. While doses to normal structures are indeed important, the importance of these differences in the pre-treatment QA measurement, even if they are known, can be difficult to evaluate without knowledge of the patient set-up uncertainties and both inter- and intra-fraction motion.

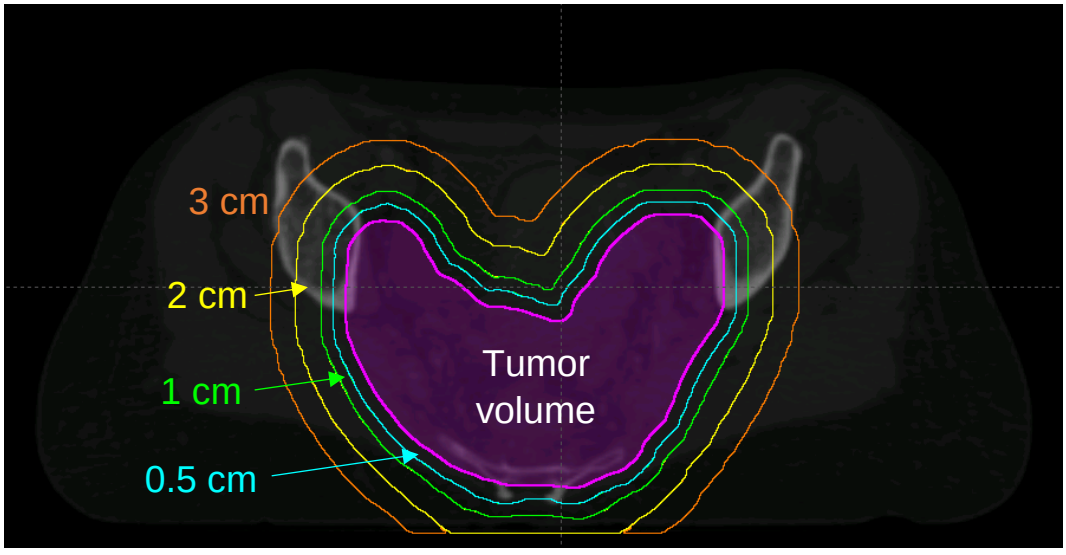

Figure 5-2 Illustration showing the annuli around the tumor volume (or PTV) for one of the patient plans from the VMAT cohort. Rings were created at different distances from the PTV and DVH metrics were calculated in each ring. The CT data in this example is intentionally darkened to better visualize the ring structures.

V.B.2. Predicting changes in the patient DVH using calculations on phantom geometries

As in previous chapters, all work here was performed using three different detector geometries – ArcCHECK, MapCHECK, and Delta 4. From the gamma comparison calculations in previous chapters, gradient and dose maps from both the "calculation" (1mm error-free calculation) and the "simulated measurement" (error-induced calculation down-sampled to the detector locations) were calculated and saved and thus were readily available for analysis in a saved MATLAB file format (*.mat). All subsequent analysis was performed using in-house MATLAB codes written as a part of this project.

Dose and gradient maps were pulled for each comparison between calculation and simulated measurement. This analysis is performed at the current spatial sampling of each device in order to simulate the spatial sampling capabilities of currently available commercial products. For each induced error type (MLC laggy leaf, MLC random perturbational, MLC bank shift, collimator rotations) and error magnitude, the dose difference maps are calculated. The MU error induced plans are not included here as it is expected that predicting dose differences from simple dose scaling errors to be less challenging than predicting changes from MLC and collimator-type errors. It should be noted that dose difference maps were calculated using both a local and global normalization of the dose differences as:

$$\boldsymbol{L}ocal\ \%\ \boldsymbol{D}ose\ \boldsymbol{D}ifference, LDD_{(i,j)} = 100 * \frac{Dcalc_{(i,j)} - Dmeas_{(i,j)}}{Dcalc_{(i,j)}}$$

$$\boldsymbol{G}lobal\ \%\ \boldsymbol{D}ose\ \boldsymbol{D}ifference, GDD_{(i,j)} = 100 * \frac{Dcalc_{(i,j)} - Dmeas_{(i,j)}}{Dcalc_{max}} \qquad [Eq.5.1]$$

Where:

- *(i,j)* is the location of a detector (i.e. a simulated measurement point)
- $Dcalc_{(i,j)}$ is the calculated dose (i.e. error-free plan dose) at point *(i,j)*
- $Dmeas_{(i,j)}$ is the simulated measurement dose (i.e. error-induced plan dose) at point *(i,j)*

– $Dcalc_{max}$ is the maximum dose over all points in the error-free plan

For a given detector geometry and comparison between error-free and error-induced calculation, both the local and global dose differences are calculated and are separately segmented into regions of high gradient and low gradient based on a gradient threshold, T_{grad}. Dose differences falling in low gradient regions are further separated into regions of high and low dose based a given dose threshold, T_{dose}, which is a percentage of the maximum dose in the error-free calculation.

For a given detector geometry and induced error plan, six dose difference histograms are created – both locally normalized and globally normalized histograms for each of the three segmented regions as shown in Figure 5-3.

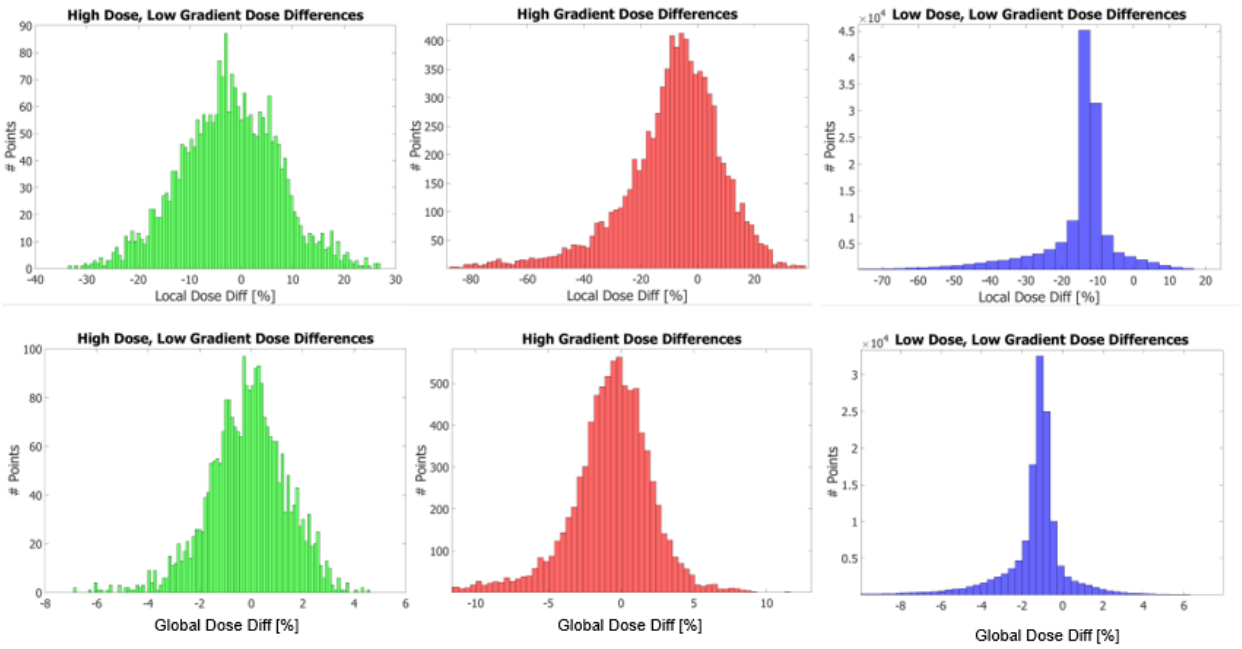

Figure 5-3. Example of both local *(top row)* and global *(bottom row)* dose difference histograms created for each error-induced case.

For each of these histograms in Figure 5-3, 11 histogram metrics are calculated. Dose difference histogram statistics consisted of calculating the mean, median, mode, minimum, and maximum dose differences as well as splitting the dose differences into positive and negative differences

and calculating the mean, median, and mode for these split regions as well, as shown in Figure 5-4.

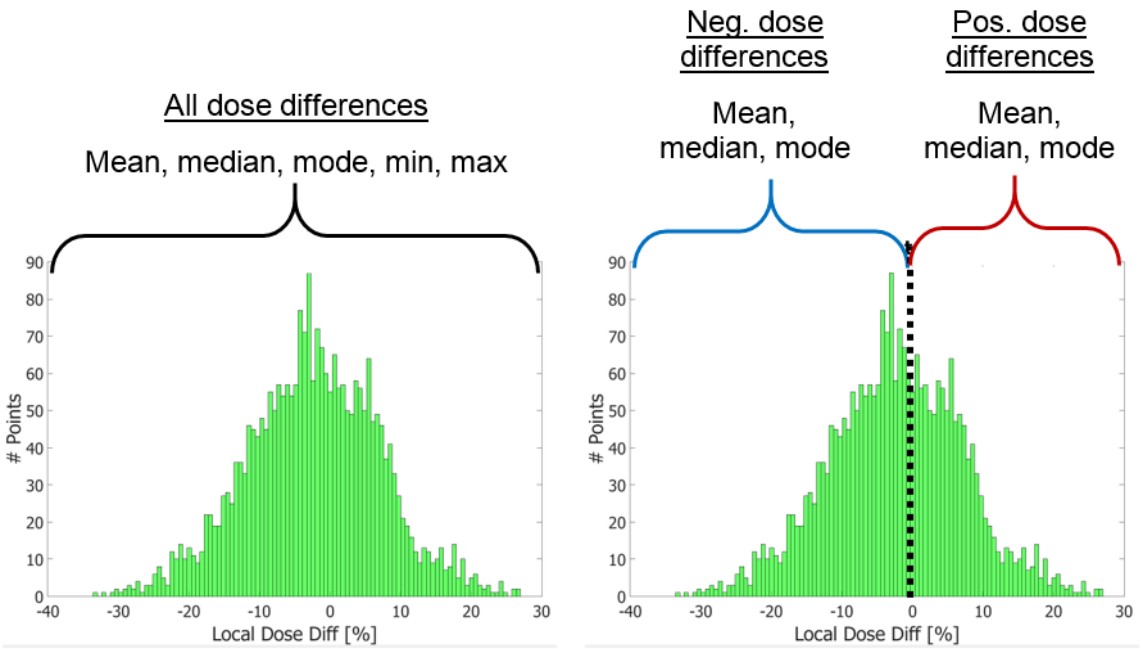

Figure 5-4. 11 histogram descriptors are calculated for each of the 6 histograms shown in Figure 5-3, including the mean, median, mode, min, and max of the entire dose difference histogram *(left)*. The mean, median, and mode are also calculated for both positive and negative dose differences separately so shown *(right)*.

Each of these 11 metrics is then separately compared to the dose differences in PTV and ring structure DVH metrics. High-dose low-gradient dose difference histogram statistics are compared to changes in PTV DVH metrics as it is expected that the high-dose low-gradient regions approximate the dose points that would fall within the PTV. Conversely, the low-dose low-gradient dose difference histogram statistics are compared to changes in ring structure DVH metrics as it is expected these low-dose low-gradient regions to predict the dose changes outside the PTV. For high gradient regions, dose difference histogram statistics are compared to both changes in DVH and ring structure metrics – while it is not expected that gradient dose differences will be as useful here, they are included just in case the high gradient regions could predict changes in ring or DVH metrics since they sit on the edges of these two regions. The general workflow for all the

different combinations of dose and gradient thresholds, as well as for different delivery techniques and different phantom geometries is shown in Figure 5-5.

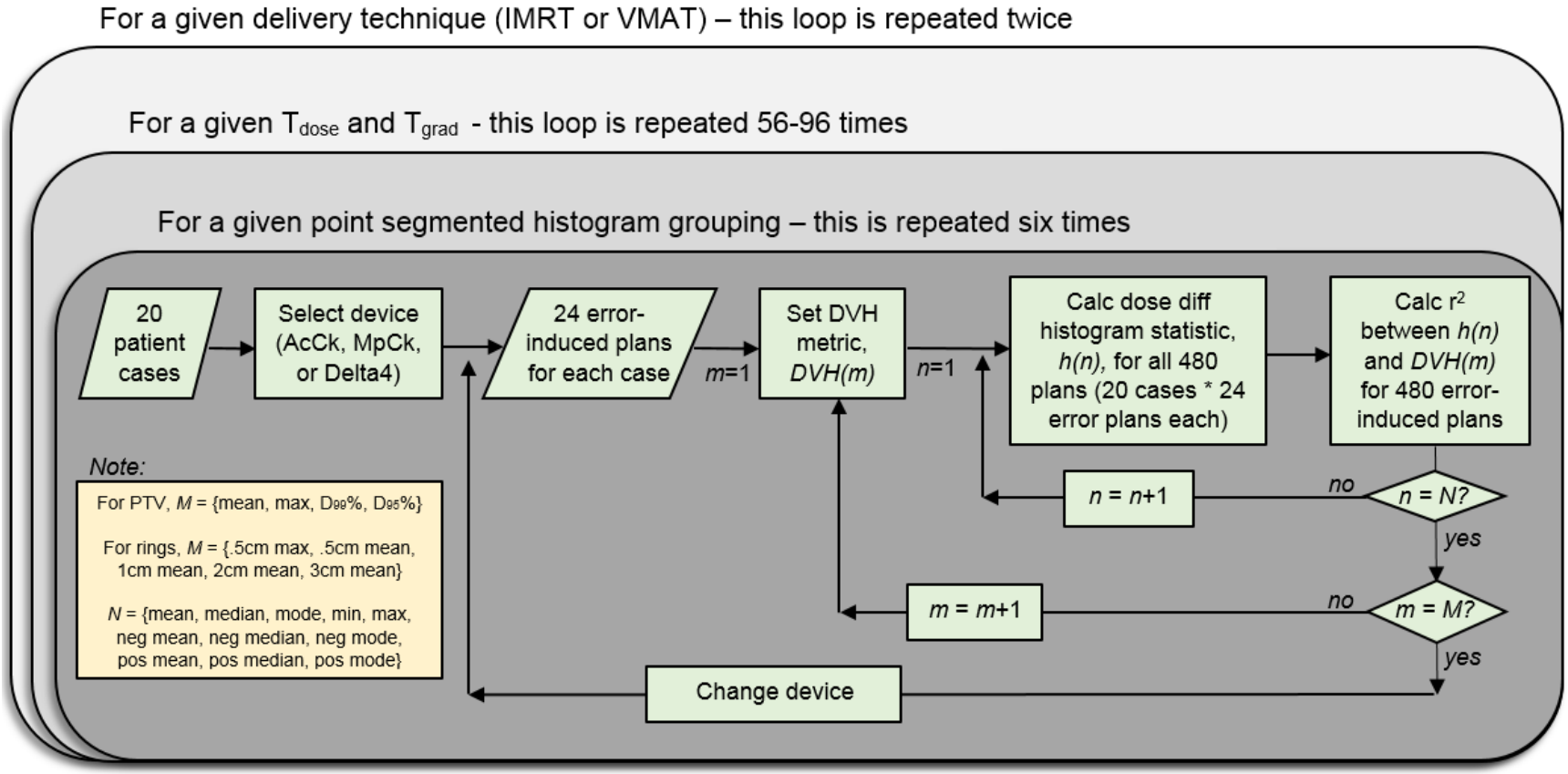

Figure 5-5. The general workflow for testing the new method. This method is repeated for each delivery technique and performed separately for a variety of different dose and gradient thresholds (T_{dose} and T_{grad}, respectively). Using a particular dose and gradient threshold, the innermost loop is repeated six times to obtain the six dose difference histograms shown in Figure 5-3.

The values of T_{dose} and T_{grad} that were investigated in the workflow shown in Figure 5-5 are as follows:

- T_{dose} = {0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8}
- $T_{grad,IMRT}$ = {4, 6, 8, …, 16}
- $T_{grad,VMAT}$ = {4, 6, 8, …, 26}

A larger range of thresholds was investigated for VMAT cases after discovering that many of the VMAT plans had higher maximum gradient values. This was later determined to be the result of several VMAT plans having a higher dose per fraction than in the IMRT cohort. The effects of different fractionation schemes was mitigated by normalizing the gradient maps by the dose per fraction for each case in order to obtain similar gradient values across plans with different fractionation.

Pearson r^2 correlations were calculated between the segmented dose-difference histogram statistics and the changes in the DVH metrics for a given detector geometry and delivery type across all the studied errors to determine if any relationship existed between DVH metrics and dose difference statistics seen in the different point groupings. This can be done across all the different induced errors types since it matters less here what the error type is, but rather that deviations exist in the plan. For each specific combination of delivery technique, detector geometry, dose/gradient threshold, and point grouping (e.g. high-dose low-gradient), nine r^2 values are obtained, evaluating the correlation between histogram statistics from the separate point groupings and differences seen in the four different PTV DVH metrics and the five different ring DVH metrics. Each r^2 correlation is calculated with the use of dose differences from n=480 error-induced patient plans.

Since a large amount of data exists, the final r^2 values are plotted in a heat map as shown in Figure 5-6. Separate heat maps are created for each evaluated DVH metric, but the example here

shows the r^2 values for changes in PTV mean dose. For this example, using the Delta 4 geometry for IMRT cases, each colored box represents the Pearson correlation (r^2 value) between the PTV mean for each of the given segmented dose difference histogram metrics (y-axis) for a specific dose/gradient threshold (listed on the x-axis). This allows quick visualization of the best combinations of dose/gradient threshold and segmented dose difference histogram metrics in predicting the true changes in the patient DVH metrics. Heat maps were created for each DVH metric evaluated showing the correlations between a particular DVH metric – PTV $D_{95\%}$, $D_{99\%}$, mean, max, and ring means, and 0.5 cm ring max – and the segmented dose difference histogram metrics. The heat maps are obtained separately for each device, delivery technique, and DVH metric. Since each heat map contains anywhere from 2400-4200 r^2 values, this analysis across three devices for IMRT and VMAT cases consisted of approximately 20,000 r^2 values for one DVH metric alone. Thus, for 9 DVH metrics the total number of r^2 values calculated for this analysis exceeded 180,000. Pearson correlations were calculated in MATLAB and the process was automated with the final Pearson r^2 values being written to an Excel spreadsheet for data storage.

In order to select the threshold for final use with each detector type, the standard square error (SSE) for each metric, which is the value minimized in the r^2 fitting, is also stored and evaluated for each Pearson correlation. The final dose and gradient thresholds are chosen by first evaluating the heat maps for a subset of thresholds with the highest r^2 values and the final threshold is chosen from this subset based on the smallest SSE.

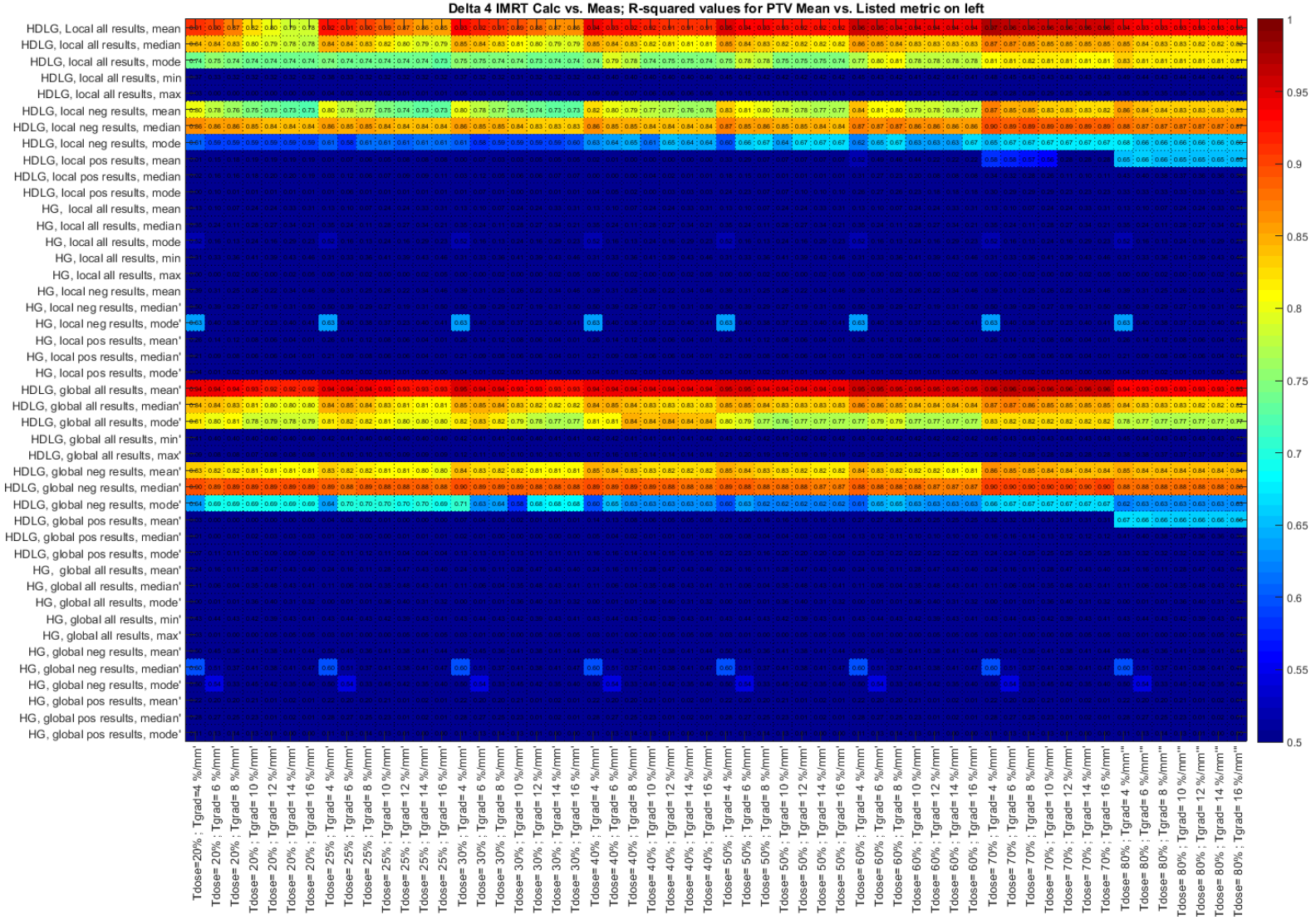

Figure 5-6. Example heat map for predicting true DVH changes to PTV mean dose using calculations on the Delta 4 geometry for IMRT cases. The x-axis shows the different dose and gradient thresholds evaluated for this scenario, and the y-axis shows the different segmented dose difference histogram statistics. Here, for each dose/gradient threshold combination the Pearson correlation is calculated between the change in PTV mean for an error-induced plan and various dose difference histogram statistics for high-dose low-gradient points (HDLG) and high gradient points (HG). The distinction between the high/low gradient and high/low dose is made based on the threshold listed on the x-axis.

V.C. Results

V.C.1. Best predictor for changes in patient PTV DVH metrics – PTV mean

A total of 27 heat maps showing the Pearson correlations between the changes in each of the nine studied DVH metrics for many different dose difference histogram descriptors for all studied combinations of dose and gradient thresholds are shown in Appendix A. In evaluating the heat maps, it was found that for each detector geometry, the PTV metric best predicted by dose difference histogram descriptors was the change in PTV mean dose. While dose difference histogram metrics for other PTV DVH metrics had relatively high r^2 values, sometimes above 0.80, scatter plots of the predicted change in the DVH metric versus the true change in the PTV DVH metrics often showed large variations from the trend line. However, the correlations between the predicted change in PTV mean dose using the dose difference histograms and the true change in PTV mean showed not only the highest Pearson correlations but also considerably less variability in the scatter plots.

The heat map in Figure 5-6, showing correlations for IMRT cases on the Delta 4 geometry for predicting true change in PTV mean, shows the highest r^2 values for “HDLG, local all results, mean” and “HDLG, global all results, mean” which represents the local and global mean dose differences for high-dose, low-gradient points (HDLG), respectively. Since both globally and locally normalized dose differences resulted in similar maximum r^2 values, scatter plots showing the predicted change in PTV mean dose (from the dose difference histograms) versus the true change in PTV mean (from the patient DVH) were also evaluated. Two representative plots are shown in Figure 5-7 for IMRT cases on the Delta 4 geometry. This illustrates that while correlations for globally normalized mean dose differences are very strong, the slope of the fit line for locally normalized comparisons, which have equally strong correlations, is near unity. This nearly 1:1 correlation between the predicted and true change in PTV mean suggests the use of the mean local dose difference in high-dose, low-gradient points for predicting true change in PTV

mean. This 1:1 correlation is a desirable property as this represents a direct relationship between the change in PTV mean predicted by the new method and the change in PTV mean dose from the TPS DVH. A slope much different from unity would require some transformation of the predicted dose difference to understand the real change in PTV mean dose, which may be difficult to ascertain with a high level of certainty.

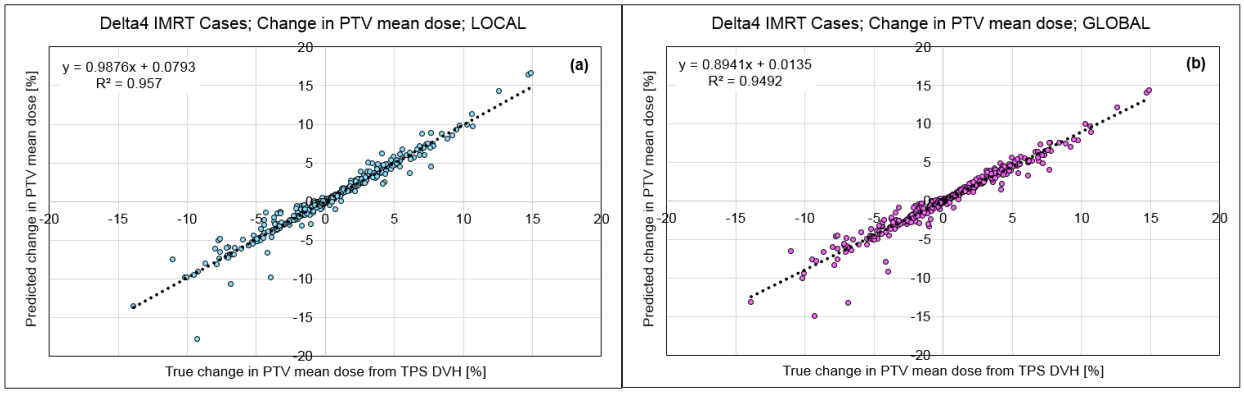

Figure 5-7. Correlations between predicted change in PTV mean dose from the high-dose low-gradient points compared to the true change in the PTV mean in the patient geometry from the TPS DVHs for (a) locally normalized dose differences and (b) globally normalized dose differences.

For all studied detector geometries, the mean local dose difference in high-dose low-gradient points was the best predictor of true changes in PTV mean dose in the patient geometry. The predicted change in PTV mean dose (utilizing local mean dose differences in high-dose, low-gradient comparison points) is shown for IMRT and VMAT cohorts for each detector geometry in Figure 5-8 (a-f). Correlations between the predicted change in PTV mean and the true change in PTV mean were also separated by delivery error type and can be found in Appendix B.

The ability of the doses in the phantom geometry to predict the change in the PTV mean dose depends greatly on the selection of appropriate dose and gradient thresholds used to sort the measurement points. The thresholds that resulted in the best predictive power for PTV mean varied between measurement geometries and also sometimes between delivery techniques for a given detector geometry. The gradient and dose thresholds that were found to be most useful for IMRT and VMAT cohorts for each device are shown in Table 5-1.

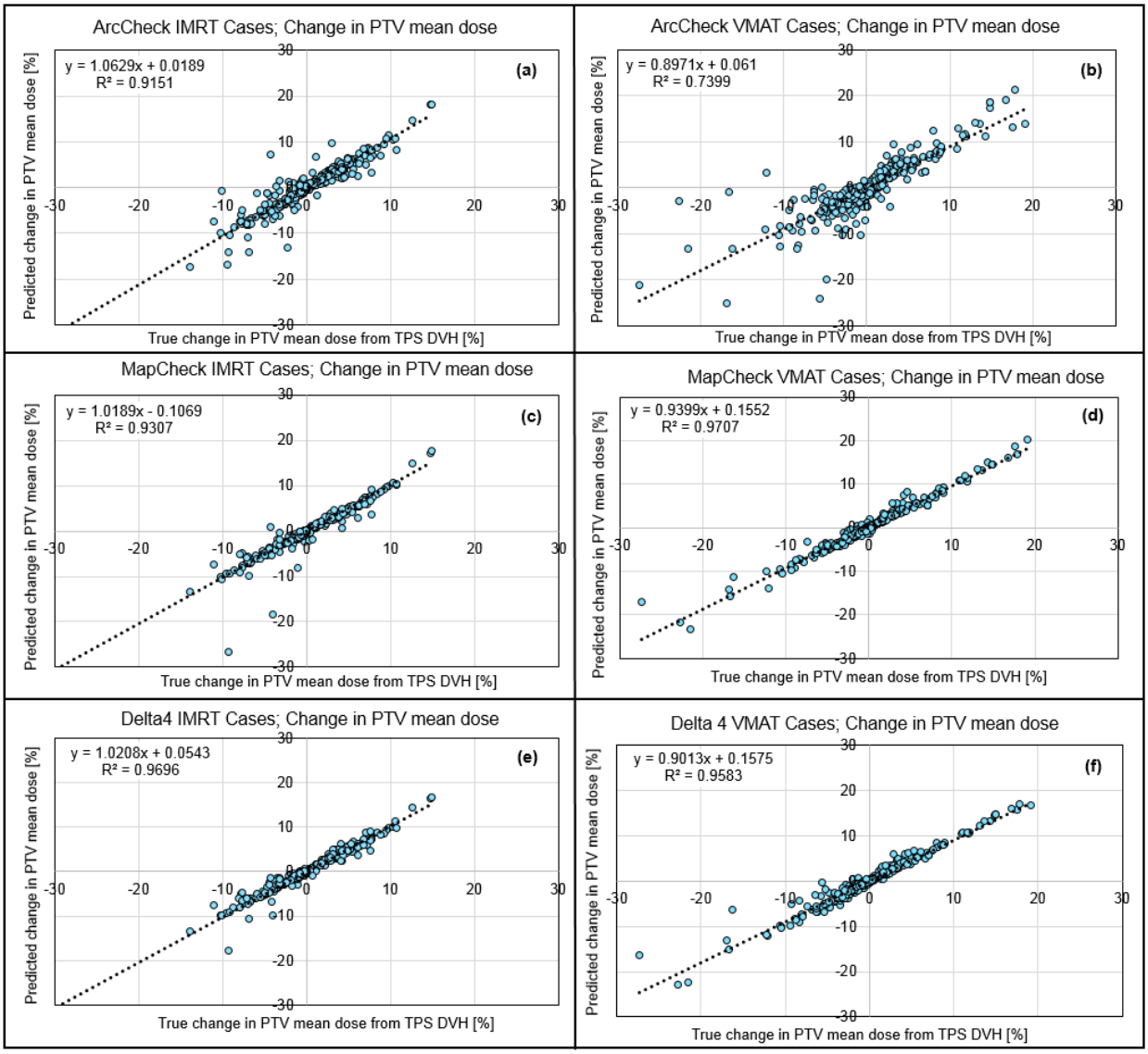

Figure 5-8. Correlations between predicted change in PTV mean dose from the high-dose low-gradient points compared to the true change in the PTV mean in the patient geometry from the TPS DVHs for both IMRT cases *(left)* and VMAT cases *(right)* for each studied device geometry – ArcCHECK *(a-b)*, MapCHECK *(c-d)*, and Delta 4 *(e-f)*.

Table 5-1. Dose and gradient thresholds for each detector geometry and delivery technique that were found to allow the best prediction of the change in PTV mean dose.

	Delta 4		**MapCHECK**		**ArcCHECK**	
	IMRT	VMAT	IMRT	VMAT	IMRT	VMAT
Dose Threshold [%]	70%	80%	60%	60%	40%	70%
Gradient Threshold [%/mm]	4%/mm	8%/mm	4%/mm	6%/mm	4%/mm	6%/mm

The dose thresholds in Table 5-1 are noticeably different between the planar-type devices and the ArcCHECK device. This is not unexpected as the maximum dose measured locally on the ArcCHECK device is lower than that on the MapCHECK or Delta 4 devices for the same IMRT case. This is because the ArcCHECK device samples the dose distribution outside the center of the target volume, whereas the MapCHECK and Delta 4 have detectors in the center of the PTV, where the highest cumulative doses will be for a given plan.

The ability to predict changes to the patient DVH using only information from comparisons in the phantom geometry allows more clinically meaningful results in IMRT QA comparisons in a way that would not require an increase in clinical workload. The utility of predicting changes in PTV mean dose for evaluating an IMRT QA result has interesting implications as well, as ICRU-83 has previously recommended that the mean dose be used as the prescription point[25]. While current compliance rates with this recommendation are low[26], one can imagine that if these guidelines are more widely adopted in the future, the utility of predicting changes in PTV mean dose from IMRT QA results could be even more powerful. One potential limitation in using of this metric, however, is that a large change in the gradient of the PTV DVH could still result in a similar PTV mean dose. In this scenario the change in PTV mean may still be predicted by the new method properly, but the clinical relevance of this type of dose error would likely go undetected if using this metric alone to pass or fail an IMRT QA comparison.

V.C.2. Predicting other PTV DVH metrics

Other PTV DVH metrics studied here (PTV $D_{99\%}$, $D_{95\%}$, and maximum dose) were not as strongly predicted by any of the segmented histogram metrics when compared to the predictions for PTV mean dose. While some correlations for PTV $D_{95\%}$ and maximum dose were moderately strong with r^2 values approaching 0.90 in some cases (such as predictions for $D_{95\%}$) for high-dose low-gradient point regions, the relationship between the predicted and true dose differences was not as strong and trend lines had slopes far from unity, unlike the correlation found previously with

PTV mean dose. This limits the usefulness of predictions for other PTV metrics because attempting to transform the obtained result to the true change in a given DVH metric may be difficult and inaccurate. Heat maps for all PTV DVH metric correlations to those predicted by different segmented dose difference histogram statistics are available in Appendix A.

V.C.3. Predicting DVH dose changes in ring structures

For ring structures, the strongest correlations were observed for the change in mean dose in the 3 cm ring outside the PTV. The change in mean dose in the 3 cm ring was predicted best using the median dose difference in the low-dose, low-gradient points. Using the same dose and gradient thresholds as shown in Table 5-1, the correlations between the predicted and true change in 3 cm ring mean dose are shown in Figure 5-9 for all three geometries for both VMAT and IMRT cases. It is notable that the correlations in Figure 5-9 are strongest for the ArcCHECK and Delta 4 geometries. For the ArcCHECK and Delta 4 devices it may be the case that since they sample the plan in several different planes there are more low-dose points available in the periphery of the plan available for analysis, compared to the MapCHECK.

Additionally, it was also observed that the predictions for changes in the 3 cm ring mean dose were much stronger for VMAT cases than they are for IMRT cases. In considering reasons for the ability to better predict changes in ring doses between IMRT and VMAT cases, we hypothesized that perhaps the conformity of the plans between IMRT and VMAT cohorts were very different. Conformity indices, initially proposed by Shaw *et al.*[27], were calculated for each IMRT and VMAT case based on notation by Feuvret *et al*[28]*.* as:

$$Conformity\ index_{RTOG} = \frac{V_{RI}}{TV} \qquad [Eq. 5.2]$$

Where:

- V_{RI} is the volume of the reference isodose (here set to 90%)
- TV is the total volume of the PTV

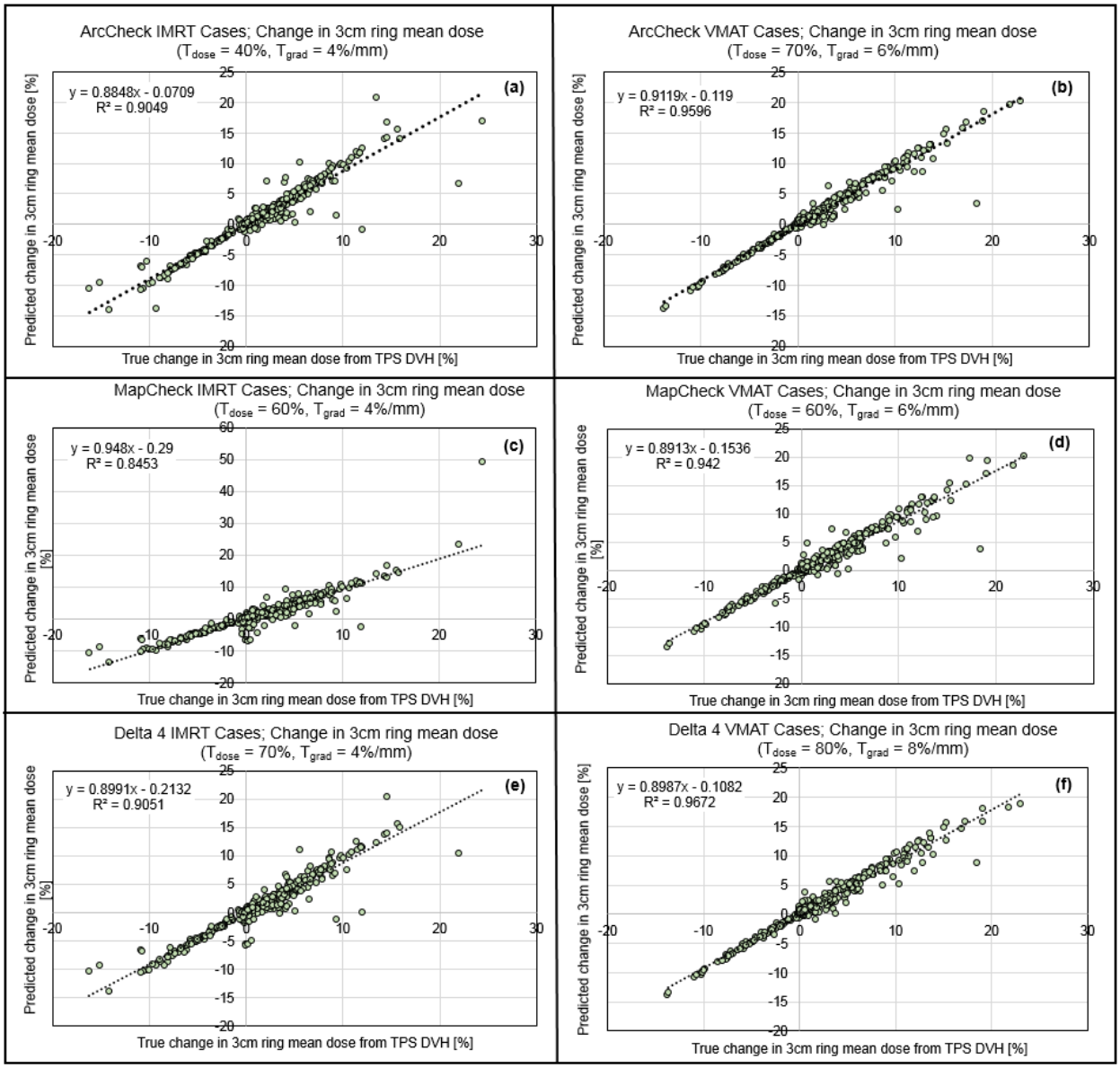

Figure 5-9. Correlations between predicted change in 3 cm ring mean dose from the median dose difference in the low-dose low-gradient points compared to the true change in the 3 cm ring mean in the patient geometry from the TPS DVHs for both IMRT cases *(left)* and VMAT cases *(right)* for each studied device geometry – ArcCHECK *(a-b)*, MapCHECK *(c-d)*, and Delta 4 *(e-f)*.

However, as Figure 5-10 shows, the calculated conformity indices for the two cohorts are very similar. Thus, one possible explanation for why dose differences in ring mean doses are better predicted for VMAT cases is that the low doses are spread more evenly across the devices.

In developing a new QA comparison technique, the change in ring doses would not be used alone to pass or a fail a plan. However, this information could offer some utility if paired with an understanding of changes in the PTV dose distribution. These two pieces together have the potential to give the end-user a sense of what could be happening in the dose distribution on a larger scale. For example, if the PTV mean decreases, but the ring structure mean increases, this may suggest that something is offset in the plan positionally. If the PTV mean increases along with the ring structure mean it may suggest that some type of scaling error may exist.

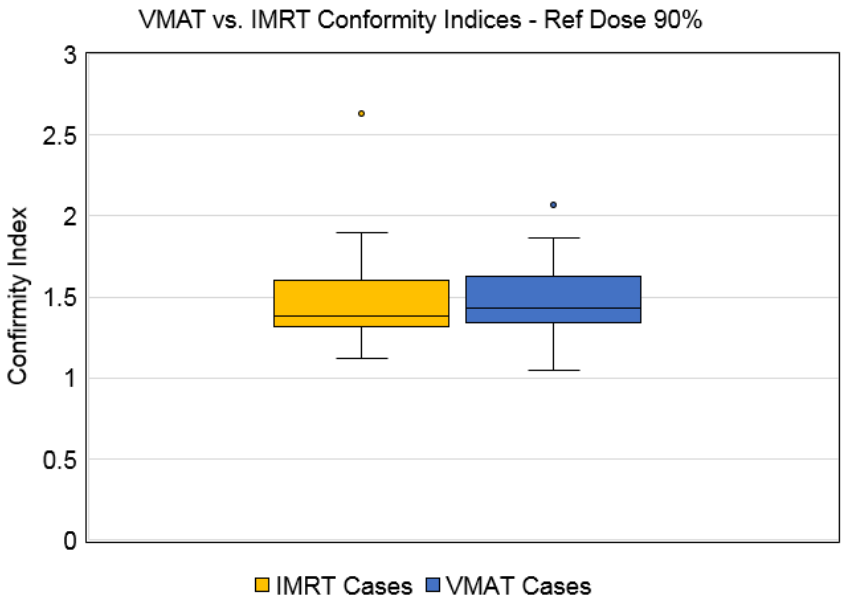

Figure 5-10. Conformity indices between the IMRT and VMAT cohorts show that both plan cohorts have similar dose conformity around the PTV.

V.C.4. Ability of gamma to predict changes in PTV DVH metrics

Previous studies have reported on the lack of relationship between gamma passing rates and clinically meaningful endpoints[2,5,8,10,19,29,30]. Our data also shows that for our case cohort, the

relationship between gamma passing rates and the PTV dose metrics are weak at best. Figure 5-11, Figure 5-12, and Figure 5-13 show the true changes in PTV $D_{95\%}$ and PTV mean dose against the gamma passing rates in our calculation-only scenario for the ArcCHECK, MapCHECK, and Delta 4, respectively. These figures also illustrate the fact that select plans with large changes to the PTV $D_{95\%}$ and PTV mean dose can pass QA using gamma criteria that are actually more stringent than criteria used by many clinics. Additionally, these figures further illustrate the point that when a plan has passing rates even above 99%, there can still be appreciable dose differences in the plan (here dose differences in the PTV $D_{95\%}$ and PTV mean exceeding 2-4%). Results for VMAT cases were similar and are available in Appendix C.

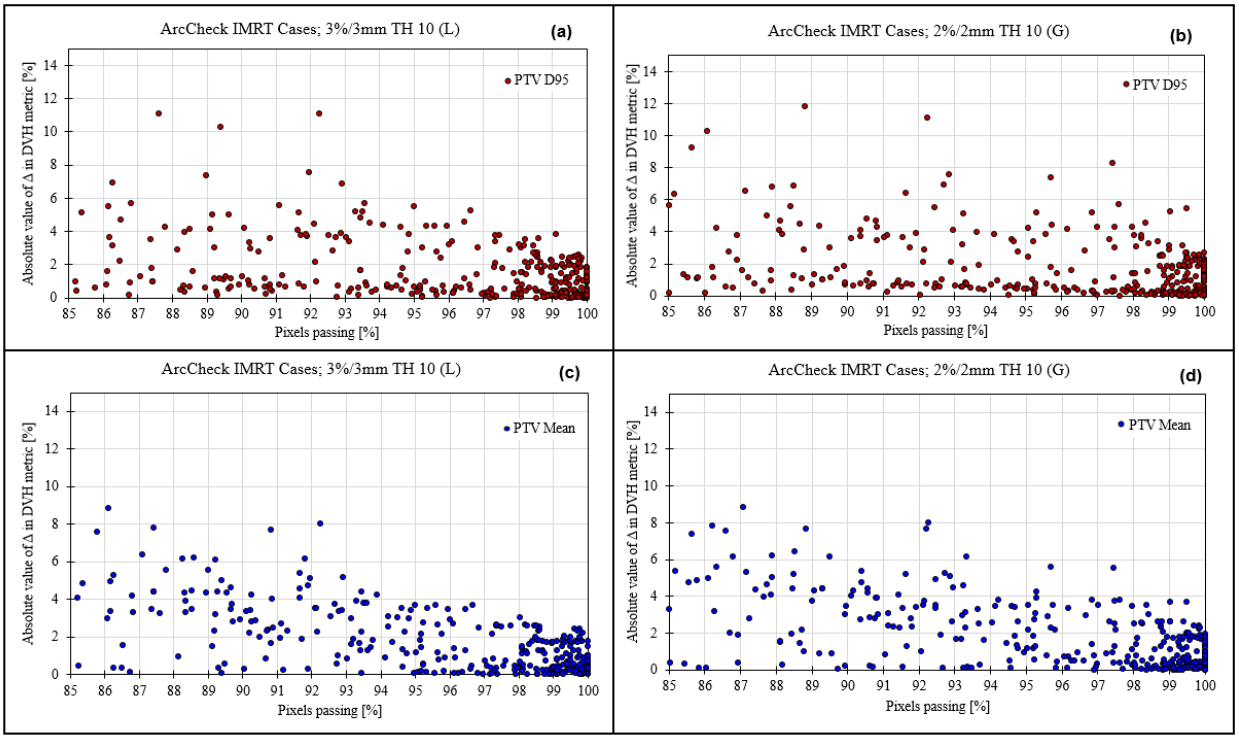

Figure 5-11. The true change in PTV $D_{95\%}$ for *(a)* 3%/3mm TH10 (local), *(b)* 2%/2mm TH10 (global) and the true change in the PTV mean dose for *(c)* 3%/3mm TH10 (local) and *(d)* 2%/2mm TH10 (global) for IMRT cases on the ArcCHECK geometry.

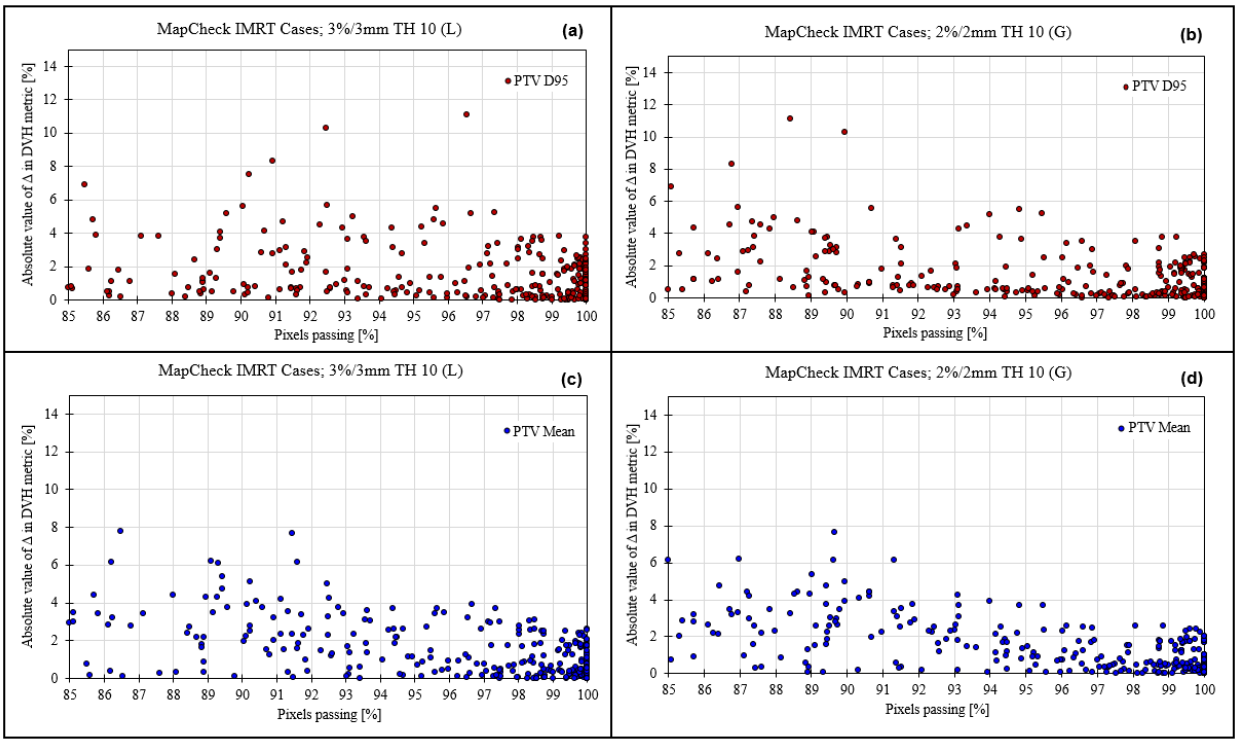

Figure 5-12. The true change in PTV $D_{95\%}$ for *(a)* 3%/3mm TH10 (local), *(b)* 2%/2mm TH10 (global) and the true change in the PTV mean dose for (*c)* 3%/3mm TH10 (local) and *(d)* 2%/2mm TH10 (global) for IMRT cases on the MapCHECK geometry.

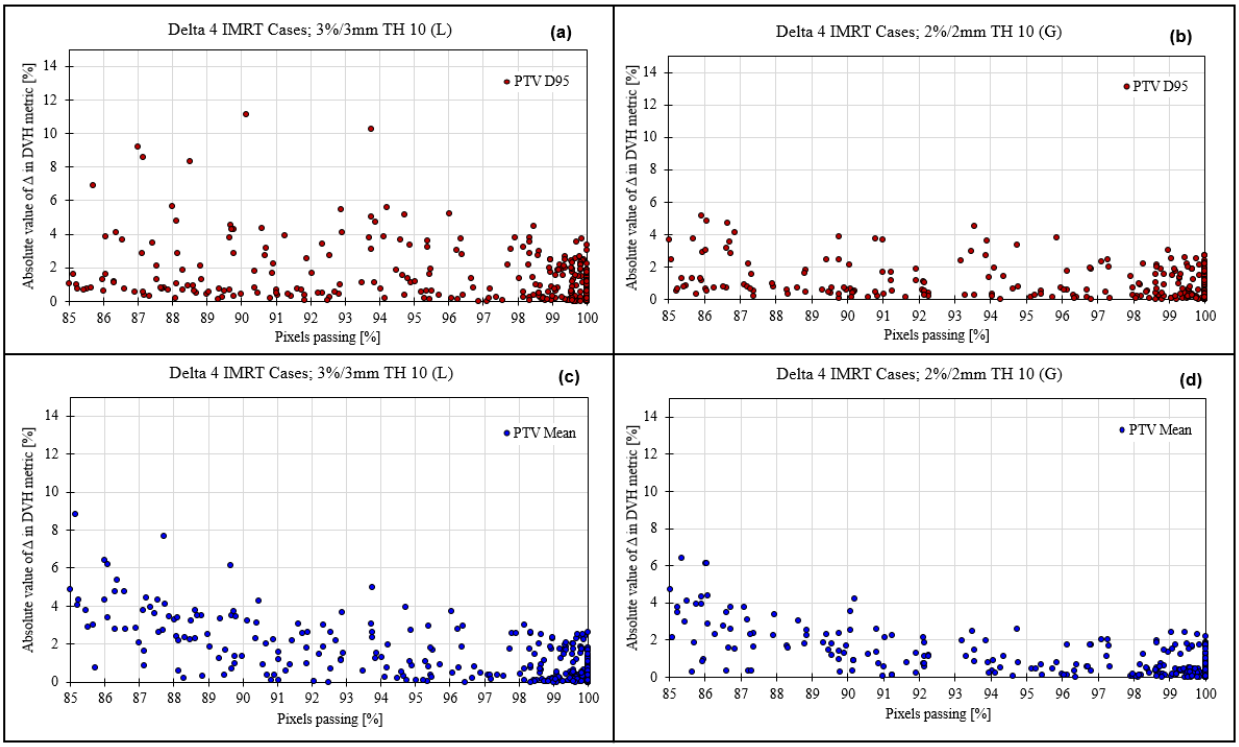

Figure 5-13. The true change in PTV $D_{95\%}$ for *(a)* 3%/3mm TH10 (local), *(b)* 2%/2mm TH10 (global) and the true change in the PTV mean dose for (*c)* 3%/3mm TH10 (local) and *(d)* 2%/2mm TH10 (global) for IMRT cases on the Delta 4 geometry.

V.C.5. The new method: gradient-dose segmented analysis (GDSA)

Based on the results in sections V.C.1-4 and, in particular, the results shown in Figure 5-8, a new method, gradient-dose segmented analysis (GDSA) is introduced for analyzing IMRT QA comparisons. The GDSA method segments QA comparisons based on the dose and gradient thresholds listed in Table 5-1. This method takes as input from the user the dose per fraction for a given case, the calculated 3D dose matrix, and the acquired measurement. First, the 3D gradient map at the plane of the detectors is calculated from the 3D dose map and normalized by the dose per fraction. Next, locally normalized dose difference maps are created between the calculated and measured doses. Dose differences are segmented into regions of high-gradient and low-gradient. If the normalized gradient value at a given detector location is greater than or equal to the specified gradient threshold, the dose difference at that point is labeled “high-gradient”. Similarly, dose differences at detector locations with gradient values less than the gradient threshold are labeled “low-gradient”.

Dose differences in low-gradient regions are further segmented into regions of low- and high-dose based on the dose thresholds in Table 5-1. This dose segmentation step utilizes the calculated dose distribution to determine if a dose difference point should be considered high-dose or low-dose. For example, if the dose threshold is 70%, the method checks to see if the calculated dose at a given detector location is greater than or equal to 70% of the maximum calculated dose. If the calculated dose is greater than or equal to this threshold, the dose difference at that detector location is labeled as “high-dose low-gradient”, otherwise this dose difference is placed in the “low-dose low-gradient” point grouping.

The mean of the locally normalized dose differences from all the high-dose low-gradient points is then computed and output as the final GDSA result. This result directly predicts the change in PTV mean dose between the delivered/measured plan and the calculated plan. Thus, the final output from the GDSA analysis is “predicted change in PTV mean dose [%]”.

V.D Sensitivity of gradient-dose segmented analysis (GDSA) compared to gamma

V.D.1. Methods

Since strong correlations were observed for predicting changes in PTV mean dose using the high-dose low-gradient mean local dose differences, this was labeled our new comparison method, which shall be referred to as the "gradient-dose segmented analysis" or GDSA. In addition to evaluating the ability of the new method to relate to DVH endpoints, the sensitivity and specificity of GDSA as a binary test was compared to that of the five gamma criteria utilized in this work using receiving operating characteristic (ROC) curves.

A 3% change in PTV mean dose was chosen as the pass/fail metric for the ROC analyses. The ability of GDSA to properly identify plans with a PTV mean dose difference greater than ±3% was evaluated and compared to the ability of the five different gamma criteria to fail a plan with a PTV mean dose difference greater than ±3%. ROC analyses were carried out in *R.* The area under the ROC curves (AUC) was also calculated for GDSA compared to the five studied gamma criteria.

True positives and negatives, and false positives and negatives were defined as

- True negative (TN): An error-induced plan has <3% change in PTV mean dose and gamma/GDSA correctly labels the plan as passing
- True positive (TP): An error-induced plan has ≥3% change in PTV mean dose and gamma/GDSA correctly labels the plan as failing
- False negative (FN): An error-induced plan has ≥3% change in PTV mean dose, but gamma/GDSA incorrectly labels the plan as passing
- False positive (FP): An error-induced plan has <3% change in PTV mean dose, but gamma/GDSA method incorrectly labels the plan as failing

Rates of TN, TP, FN, and FP were also evaluated for 3%/3mm TH10 (G), 2%/2mm TH10 (G), and 3%/3mm TH10 (L) against GDSA. The purpose of this was to determine if any dramatic increases in rates of false negatives or false positives might be seen with GDSA compared to gamma criteria currently in clinical use. This is important to evaluate as any dramatic increases in FNs or FPs for GDSA compared to gamma could inhibit the utility of GDSA in the clinic.

V.D.2. Results

ROC analyses were carried out in *R* to evaluate sensitivity/specificity of both gamma and GDSA as a binary pass/fail metric. The pass/fail metric to perform ROC analyses was chosen to be >3% change in PTV mean dose from the TPS DVHs. This analysis thus tested the ability of both the gamma comparison (for the five studied criteria here) and the GDSA method to correctly flag plans that have a PTV mean dose difference greater than ±3%. ROC curves are shown in Figure 5-14 and corresponding AUC values are shown in Table 5-2. Most notably, for almost all studied scenarios GDSA offers improved accuracy to identify errors in PTV mean >3% compared to the five studied gamma criteria. The only scenario for which this is not true is for VMAT cases on the ArcCHECK, which show only slightly improved error detection with GDSA compared to gamma.

With the exception of VMAT cases on the ArcCHECK, the AUC of GDSA was noticeably higher than the AUC of the five studied gamma criteria for all studied scenarios. Of note, MapCHECK and Delta 4 had very similar and also high AUC values for GDSA. While ArcCHECK performed less well with GDSA, the AUC values for GDSA with ArcCHECK are still higher than those for the gamma comparison on this device. GDSA not only allows a more clinically meaningful result, but also offers improvements in sensitivity and specificity compared to the studied gamma criteria.

From the ROC analysis, it is also clear that the gamma comparison in conjunction with Delta 4 is typically more sensitive than MapCHECK, which is more sensitive than ArcCHECK with the gamma comparison, which is consistent with the error curve analysis of the three devices in Chapter III.

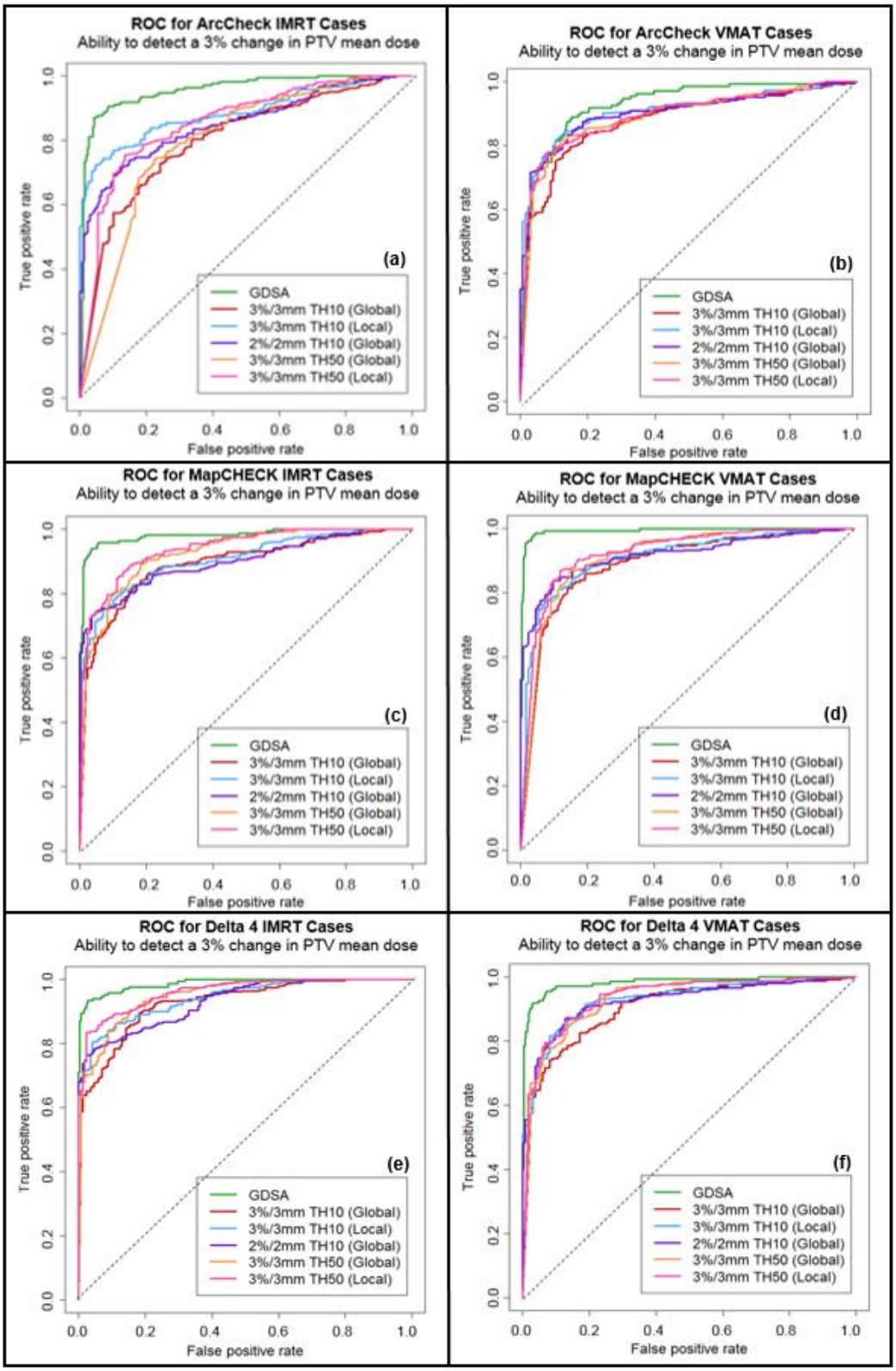

Figure 5-14. ROC curves showing the ability of the new GDSA method (i.e. using high-dose low gradient points to predict changes in PTV mean) compared to the listed gamma criteria to flag a plan as failing in the presence of a PTV mean dose difference greater than ±3%. Results shown for ArcCHECK *(a-b)*, MapCHECK *(c-d)*, and Delta 4 *(e-f)* geometries.

Table 5-2. Area under the ROC curve (AUC) values for the three devices and two delivery techniques are shown for the new GDSA method and the three listed gamma criteria.

	ArcCHECK		**MapCHECK**		**Delta 4**	
	IMRT	VMAT	IMRT	VMAT	IMRT	VMAT
GDSA	0.957	0.930	0.981	0.993	0.988	0.982
3%/3mm TH 10 (G)	0.802	0.877	0.891	0.890	0.928	0.905
3%/3mm TH 10 (L)	0.885	0.907	0.907	0.911	0.942	0.929
2%/2mm TH 10 (G)	0.853	0.901	0.899	0.919	0.932	0.926
3%/3mm TH 50 (G)	0.797	0.891	0.925	0.913	0.951	0.931
3%/3mm TH 50 (L)	0.857	0.894	0.940	0.925	0.962	0.937

The number of true negatives, true positives, false negatives, and false positives was also evaluated for a pass/fail threshold of ±3% change in PTV mean dose for both the GDSA method and the three gamma criteria that are more prevalent in clinic use – 3%/3mm TH 10 (global), 3%/3mm TH 10 (local), and 2%/2mm TH 10 (global). The purpose of this was to determine if this new method would cause a large increase in the number of false positives or false negatives compared to gamma comparisons in the clinical setting. This is important to consider when studying a new technique for clinical use – an increase in false positives could greatly impact the clinical workflow, requiring extra time to investigate these failures, whereas too many false negatives implies the test would fail to catch plans with important dose differences. Most notably, the results, shown in Figure 5-16, illustrate that the GDSA method generally has similar or lower false negative rates than the three gamma criteria, and dramatically lower false positive rates than the listed gamma criteria. Therefore, the rate of false positives/negatives with the use of GDSA is unlikely to be a barrier to clinical implementation.

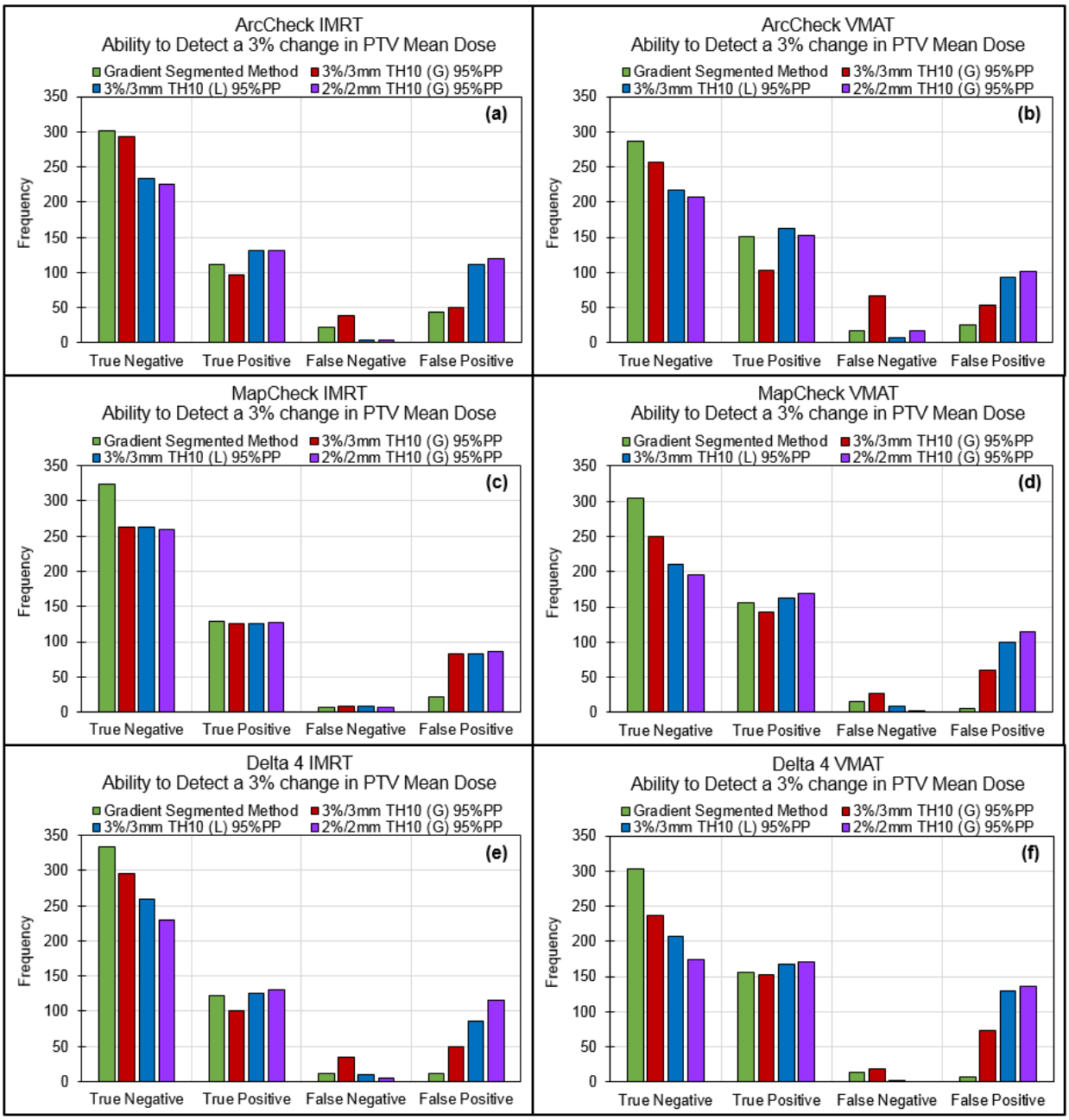

Figure 5-15. True negatives/positives and false negatives/positives for a pass/fail of ±3% change in PTV mean dose for three gamma criteria and the new, gradient segmented method for *(a)* IMRT cases on the ArcCHECK, *(b)* VMAT cases on the ArcCHECK, *(c)* IMRT cases on the MapCHECK, *(d)* VMAT cases on the MapCHECK, *(e)* IMRT cases on the Delta 4, and *(f)* VMAT cases on the Delta 4 device.

V.E. Discussion

In this chapter, gradient-dose segmented analysis (GDSA) has been introduced as a new method for performing IMRT QA comparisons which segments comparison points based on dose and gradient thresholds. The GDSA method was developed using calculation-only comparisons which

made use of an error-free calculation at 1mm spatial resolution compared to many error-induced calculations which were down-sampled to each detector's spatial sampling. Thus, no noise or other uncertainties were introduced in the process of developing and evaluating this method. Proper selection of dose and gradient thresholds allowed the prediction of true changes in PTV mean dose in the patient geometry using only the dose and gradient information in the phantom geometry. The mean local dose difference in the high-dose low-gradient points was able to predict changes in PTV mean dose in the patient. As a secondary result, changes in ring structure mean doses (3 cm from the PTV) were also predicted with moderately strong correlations to the median local dose difference in low-dose, low-gradient comparison points. While the change in mean ring dose 3 cm outside the PTV is not the most important metric when evaluating a plan, the change in PTV mean dose paired with the change in the ring structure may give the user a hint at what type of dose differences are present in the plan – for example increased dose to the ring mean and the PTV mean could signal a calibration issue. Similarly, if the PTV mean was not different and the mean ring dose increased one can imagine that certain scenarios – such as leaves open where they should not be, could cause this type of dose error.

Using the high-dose low-gradient points to predict changes in PTV mean is a feasible metric with which to quickly evaluate a IMRT QA comparison result that not only gives more meaningful results in relation to the actual patient dose distribution, but also is a more sensitive metric than the gamma comparison in the binary pass/fail setting, as shown in the ROC analyses in Figure 5-14. The GDSA method is also feasible to implement clinically in that it would not cause a foreseeable increase in the amount of work required to perform the analysis. The only additional piece of information required by the user would be the dose per fraction in order to normalize the gradient maps. Additionally, the GDSA method does not require the user to select large number of different parameters that can modify the sensitivity of the test. Rather, the user can select a

threshold for change in PTV that matters most in their clinical setting, which would also easily allow clinics to select site-specific thresholds.

With GDSA the changes in PTV mean dose were well predicted for the MapCHECK and Delta 4 geometries, and moderately well for IMRT cases on the ArcCHECK. However, VMAT ArcCHECK cases had similar AUC values compared to the gamma comparison as well as much lower Pearson correlations between predicted and true change in PTV mean dose. To rule out spatial sampling as a cause for lower ArcCHECK correlations, the change in PTV mean dose predicted by the new method for the same dose and gradient threshold values in Table 5-1 was calculated for 1mm vs. 1mm comparisons. Results in Figure 5-16 show that increased spatial sampling does not improve results for the new method in the ArcCHECK geometry.

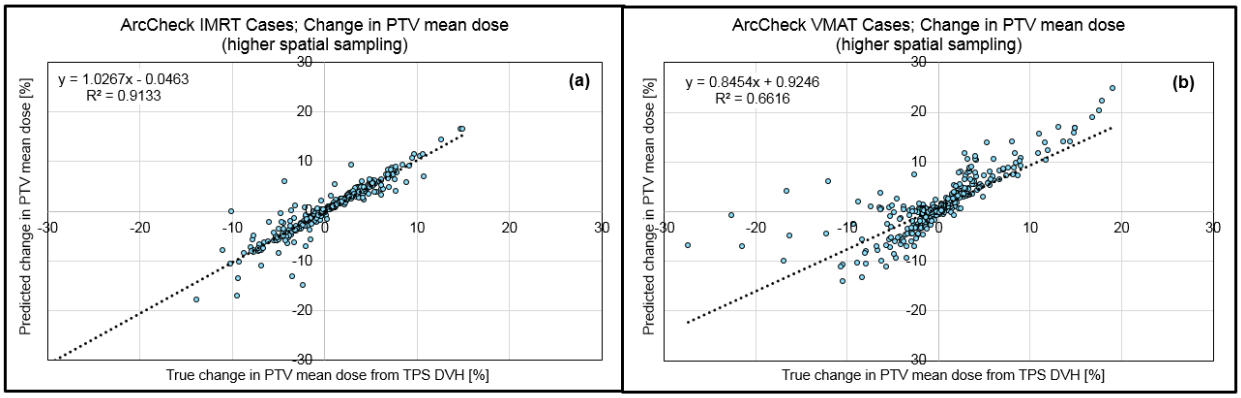

Figure 5-16. Correlations between predicted change in PTV mean dose from the high-dose low-gradient points compared to the true change in the PTV mean in the patient geometry from the TPS DVHs for ArcCHECK geometry with higher spatial sampling (1mm vs. 1mm) for *(a)* IMRT cases and *(b)* VMAT cases.

One reason for this may be the fact that GDSA makes use of gradient and dose maps as a way to predict which measurement/calculation points are inside the PTV volume on the phantom geometry. When looking at the same IMRT plan measured on the ArcCHECK and the Delta 4 in Figure 5-17, it is easy to see that the general shape of the PTV is visible on both these devices. However, looking at a sample VMAT case for these two same devices shows a much different

story. Figure 5-18 shows that the gradient maps for VMAT plans look somewhat random on the ArcCHECK device compared to those on the Delta 4 geometry.

Even though the gradient map for the Delta 4 geometry for the VMAT case is more complex than that of the IMRT case on the Delta 4, it is still easy to predict which points in the measurement likely represent measurement/calculation points inside the PTV volume. However, the same is not true for the ArcCHECK, and in particular VMAT cases on the ArcCHECK. For this comparison scenario it is difficult to discern which comparison points in the ArcCHECK likely lie within the PTV volume based on the dose and gradient maps alone. This is because the plan gradients come together much differently in the center of the plan than on a cylindrical surface some distance away from the target.

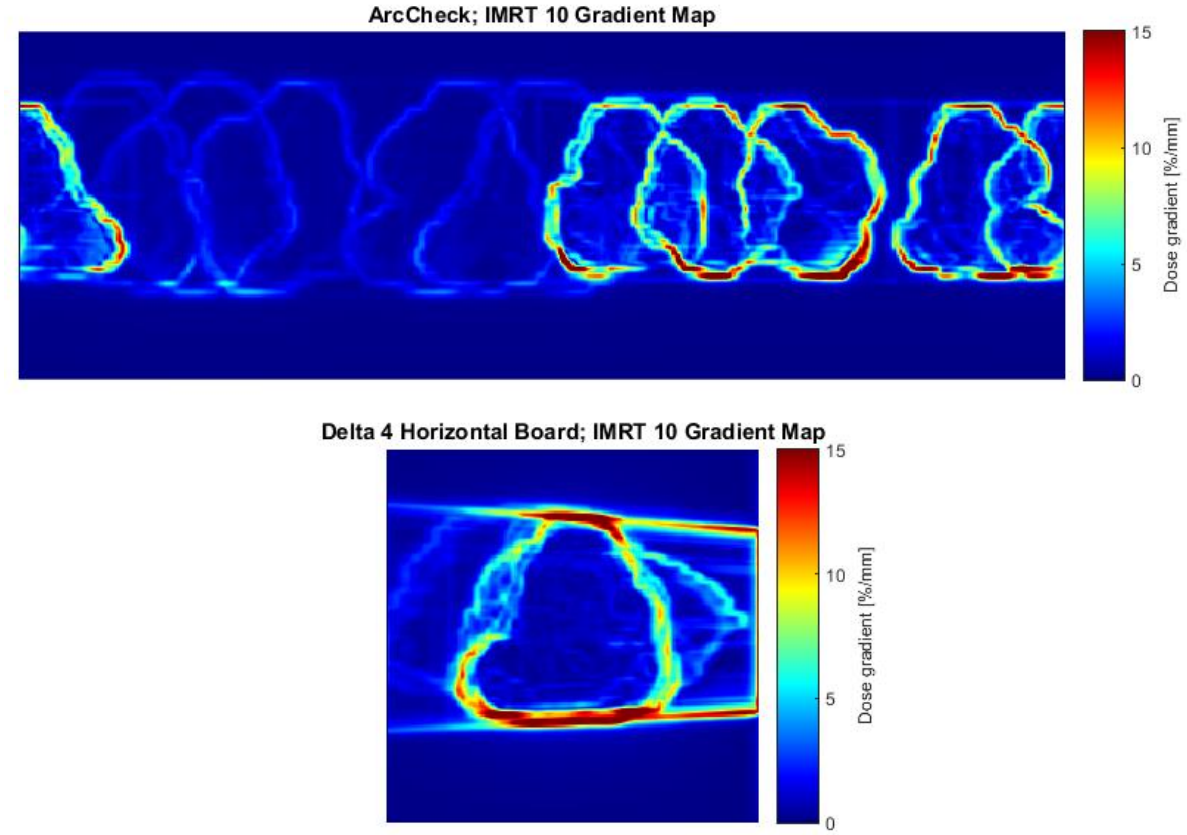

Figure 5-17. For IMRT cases on the ArcCHECK the gradient maps outline the shape of the PTV on both the *(top)* ArcCHECK and *(bottom)* Delta 4.

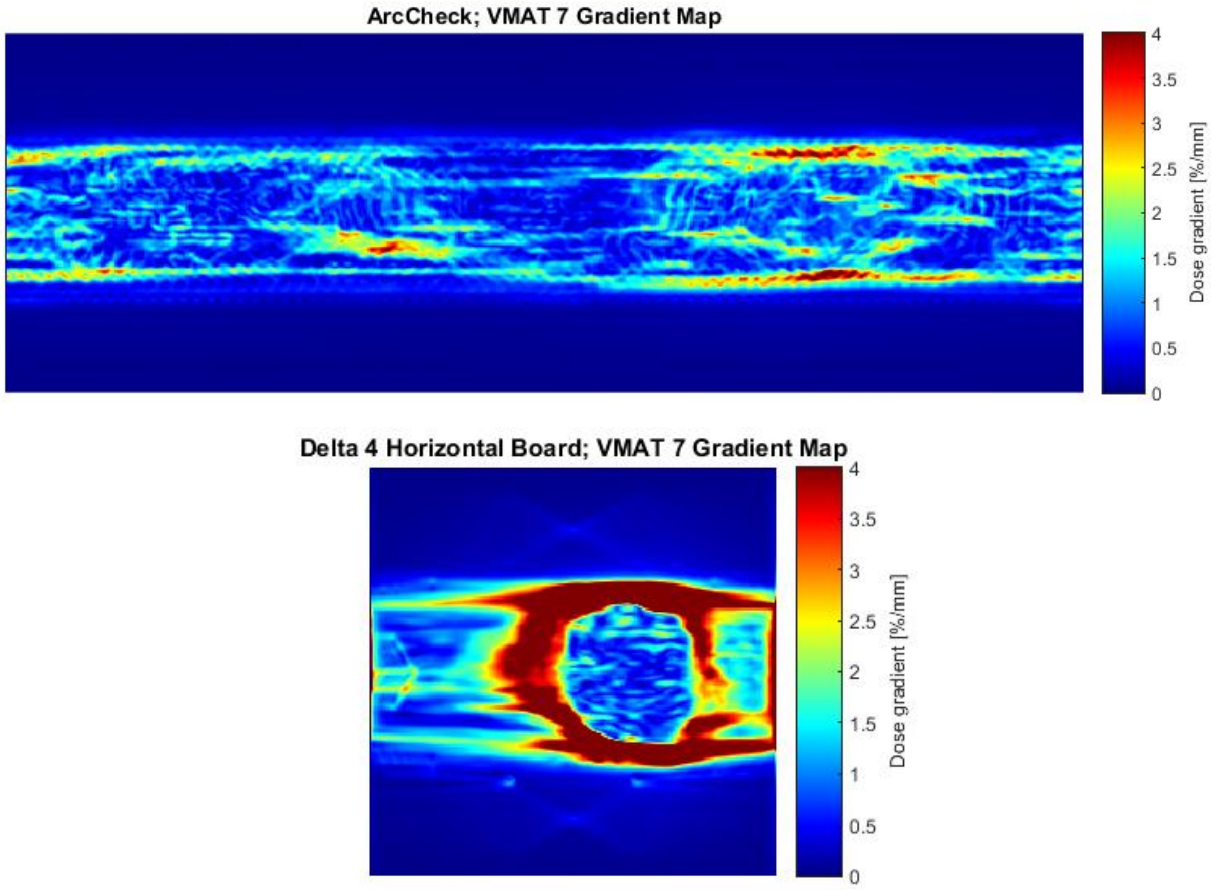

Figure 5-18. *(Top)* ArcCHECK VMAT gradient map shows how the gradients for VMAT cases are non-intuitive relative to the shape of the PTV volume. *(Bottom)* Conversely, the Delta 4 VMAT gradient map still shows the relative shape of the gradient map.

The histograms of the normalized gradient maps from Figure 5-17 and Figure 5-18 were also plotted in Figure 5-19 for IMRT 10 and Figure 5-20 for VMAT 7 to further illustrate the fundamentally different gradient distribution for ArcCHECK VMAT cases. The histograms in Figure 5-19 for the IMRT case show that the range and distribution of gradient values are similar between the ArcCHECK and Delta 4 geometries. However, when studying the histograms of gradient values for the example VMAT case in Figure 5-20, the histogram of gradient values for the ArcCHECK shows that the maximum gradient is much smaller than that for the same case on the Delta 4 geometry. These results illustrate that for VMAT cases on the ArcCHECK it is likely that the swath of low gradients around the entire device limits our ability to sort out which measurement points relate to those located within the PTV volume, and subsequently reduces the accuracy of the GDSA method for the ArcCHECK geometry.

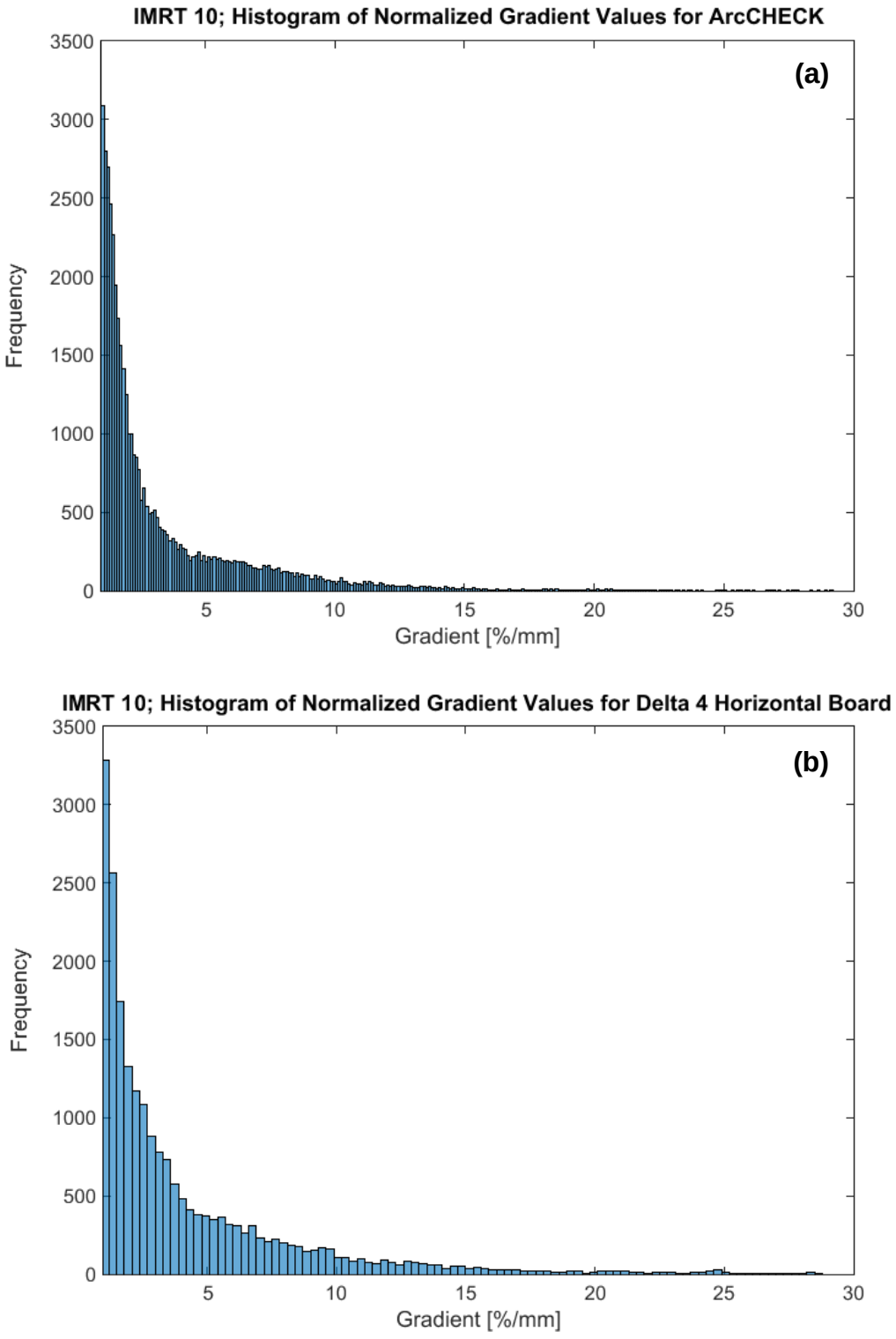

Figure 5-19. Histograms of normalized gradient values for IMRT 10 shown for (a) ArcCHECK and (b) Delta 4 horizontal board

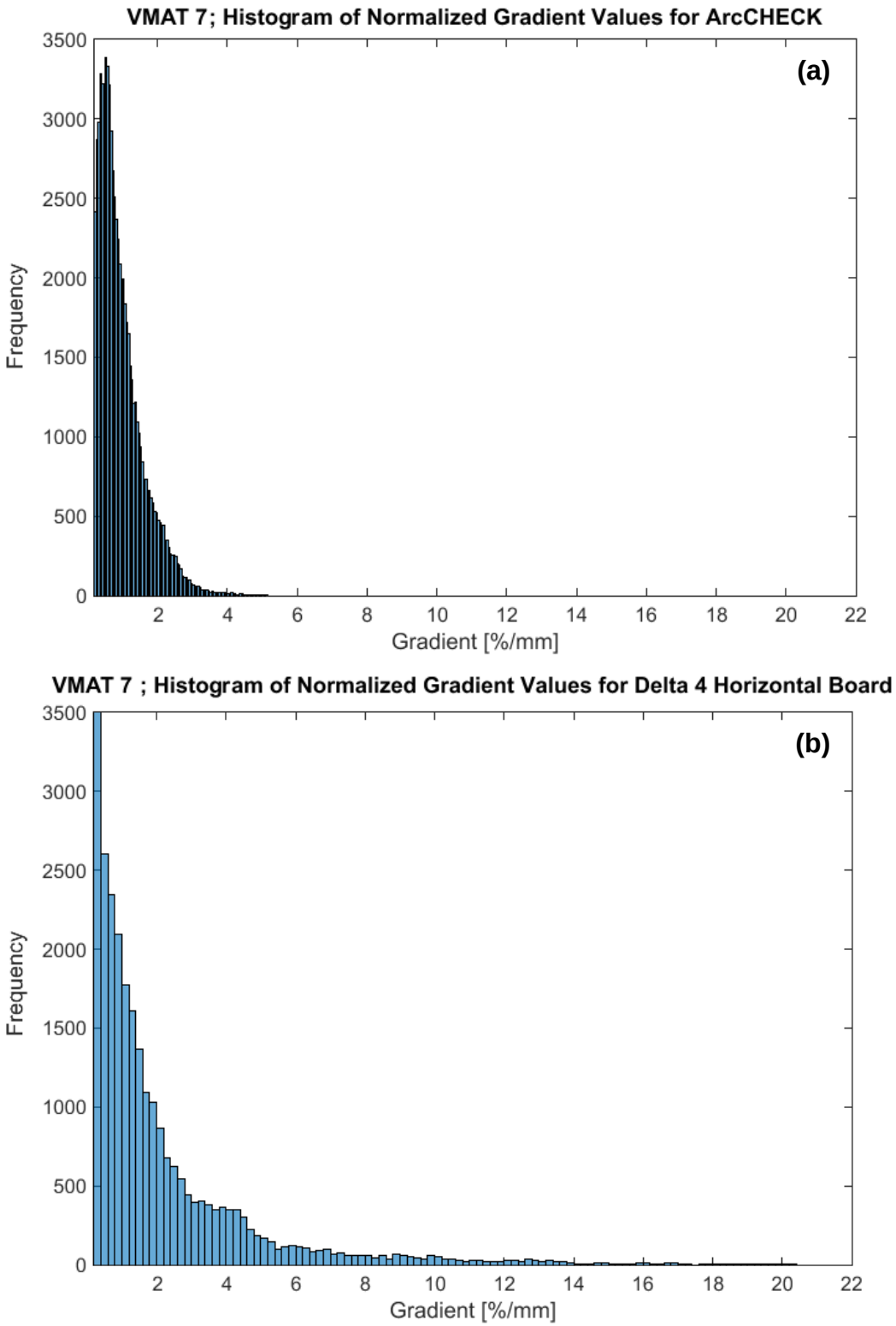

Figure 5-20. Histograms of normalized gradient values for VMAT 7 shown for (a) ArcCHECK and (b) Delta 4 horizontal board

While the current GDSA method did not perform as well for the ArcCHECK geometry as it did with the Delta 4 and MapCHECK geometries, it was also not worse than the studied gamma criteria, and perhaps ArcCHECK-specific modifications are required to improve performance. This is not surprising when considering that the ArcCHECK samples the plan at fundamentally different locations in space (and outside the PTV volume) compared to the planar-type arrays. These challenges, both with the gamma comparison and the GDSA method presented here, point to the fact that a different method of analyzing ArcCHECK data may be necessary.

V.F. Conclusions

A new method, gradient-dose segmented analysis (GDSA), for comparing IMRT QA dose distributions was presented and evaluated. This method relies on calculated dose matrices to obtain gradient maps for the GDSA analysis, but requires only information that is already exported for current IMRT QA comparisons (i.e. the exported DICOM dose files). These dose and gradient maps are then utilized to segment the comparison into:

- High-dose, low-gradient comparison points
- Low-dose, low-gradient comparison points
- High-gradient comparison points

The mean local dose difference in high-dose, low-gradient comparison points was found to predict true changes in PTV mean dose in the patient geometry for error-induced plans with known errors. The predicted change in PTV mean dose was strongly correlated with the true change in PTV mean dose from the patient DVHs, with a nearly 1:1 correlation. This suggests that this GDSA metric can be utilized in evaluating an IMRT QA comparison using more clinically meaningful results, compared to the percent pixels passing result obtained from the gamma comparison.

The GDSA method performed well for both IMRT and VMAT cases on the MapCHECK and Delta 4 geometries and was shown to be more sensitive and specific than gamma comparisons using

current clinical criteria. Performance of the GDSA method on the ArcCHECK device was noticeably worse compared to planar-type devices, though GDSA sensitivity/specificity on ArcCHECK was still slightly higher than that of the gamma comparison. GDSA showed similar rates of false negatives compared to stricter gamma criteria such as 2%/2mm TH 10 (global) and even lower rates of false positives than this same criterion when used as a binary metric with a 3% cut-off for passing/failing plans. In addition to not increasing time spent analyzing data in the clinic, GDSA gives the user a more clinically meaningful end-result. Additionally, users could select any desired threshold for change in PTV mean that they are willing to accept in their clinic for IMRT QA comparisons and would also easily allow the selection of site-specific thresholds on allowable differences in PTV mean between calculation and measurement.

A.V. Appendix A for Chapter V

The heat maps presented here show Pearson r^2 values and were utilized to determine if any of the studied DVH metrics could be predicted by the histogram metrics obtained from dose differences from various dose/gradient segmented regions of the field. A large number of dose and gradient thresholds were studied, as it was not obvious *a priori* which thresholds would offer the best predictions for changes in DVH metrics for each detector device.

A.V.1 Delta 4 IMRT PTV Correlations

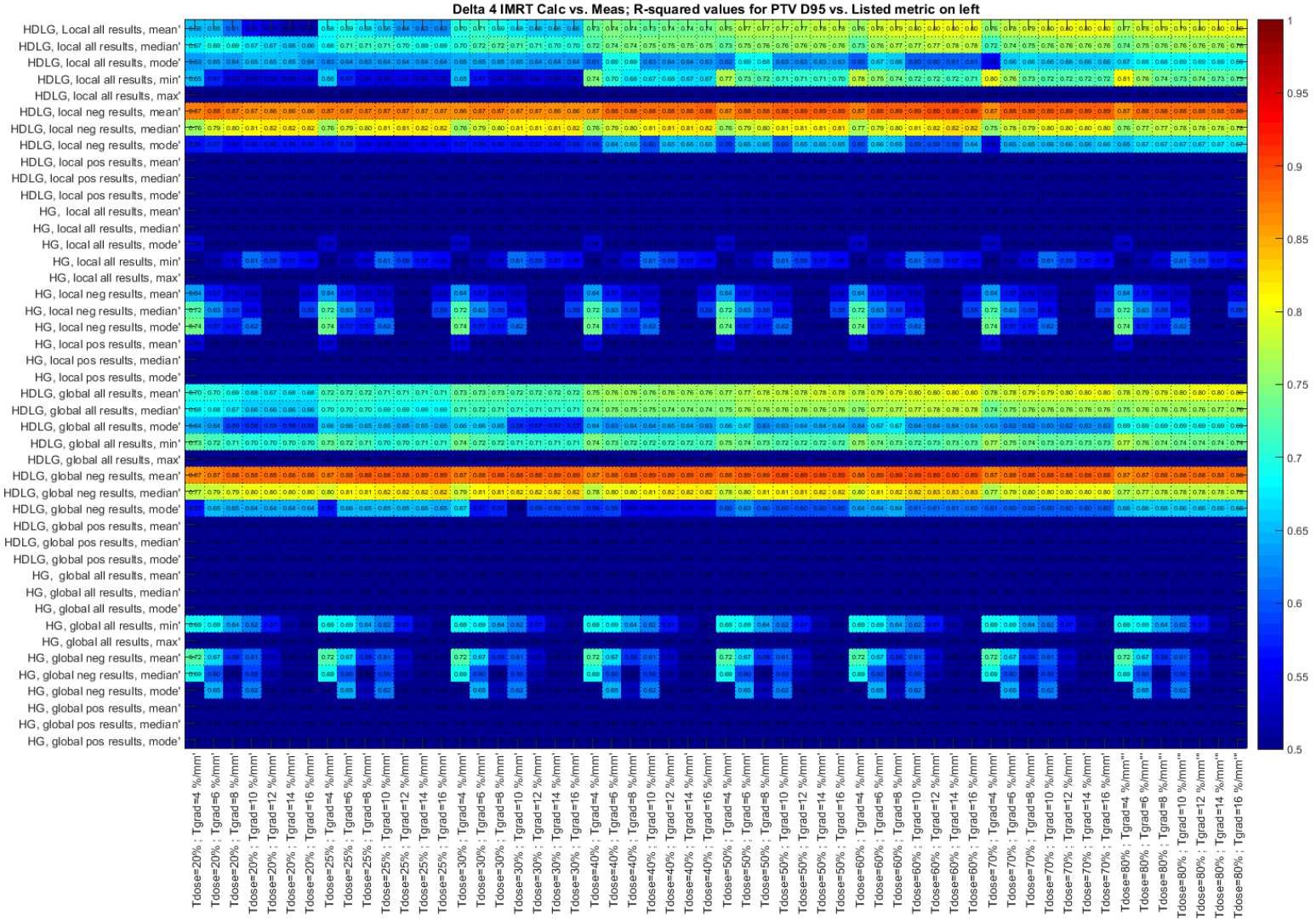

Figure A.5-1. Pearson r^2 heat map for IMRT cases on the Delta 4 for predicting the change in PTV $D_{95\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

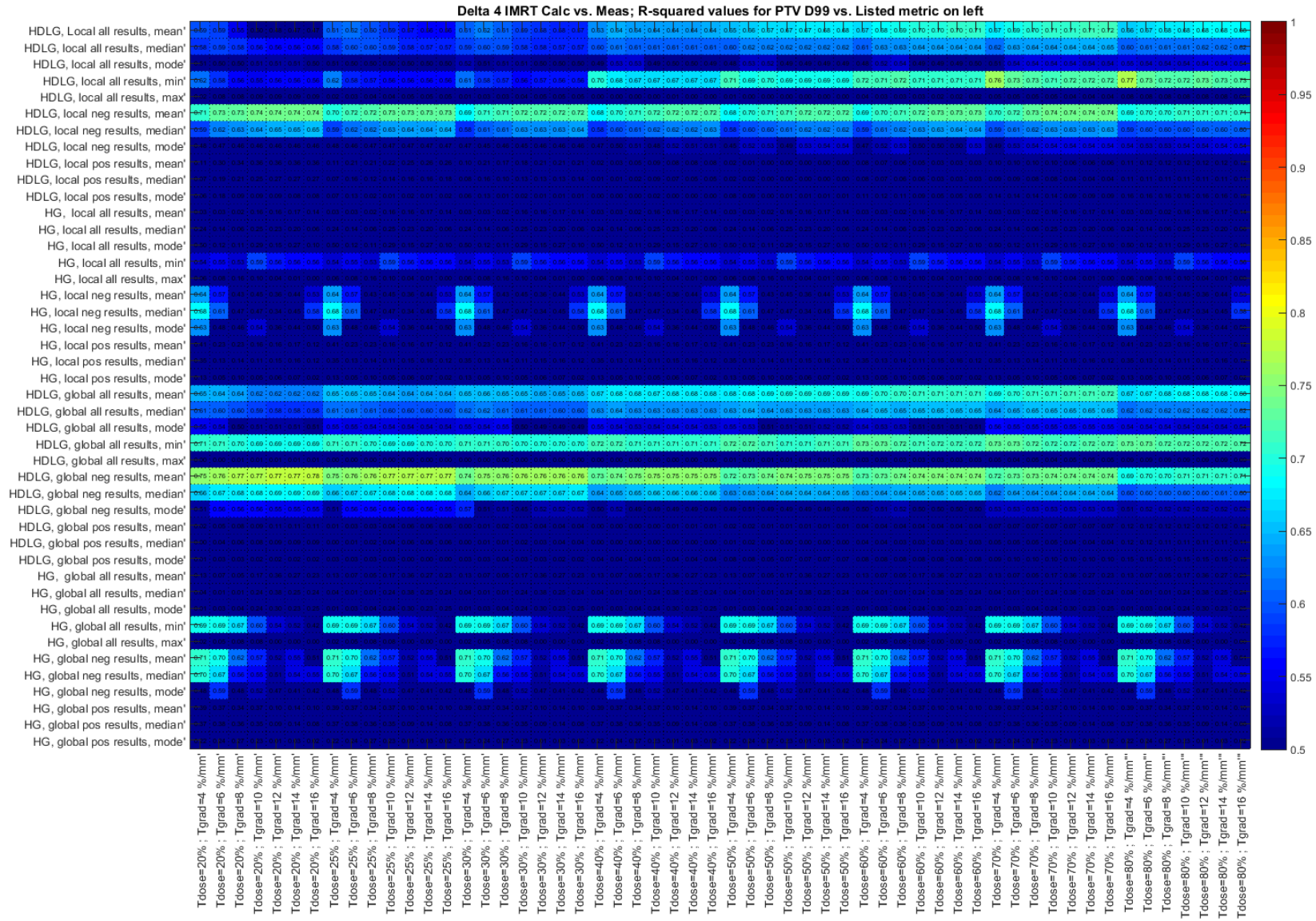

Figure A.5-2. Pearson r^2 heat map for IMRT cases on the Delta 4 for predicting the change in PTV $D_{99\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

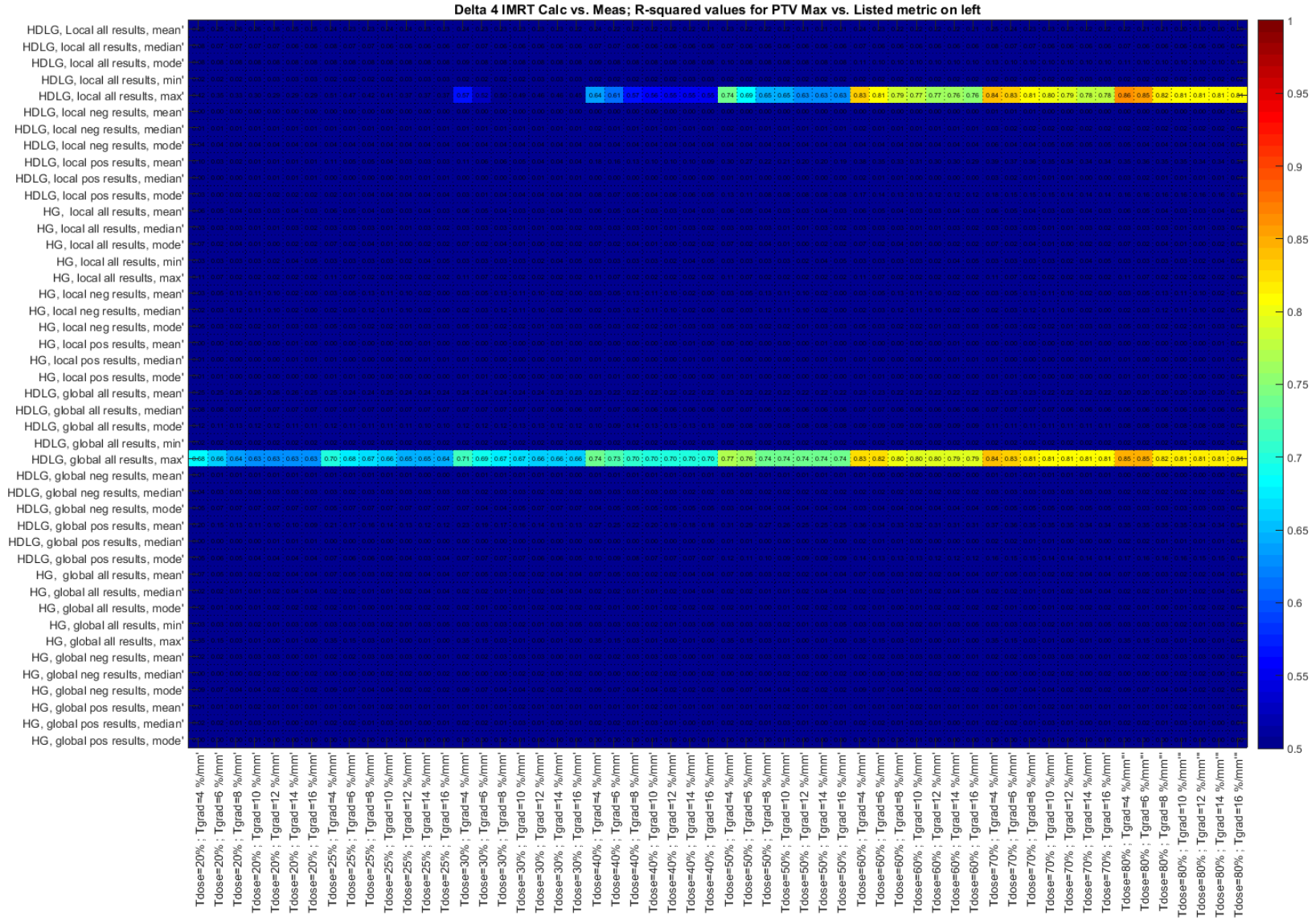

Figure A.5-3. Pearson r^2 heat map for IMRT cases on the Delta 4 for predicting the change in PTV maximum dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

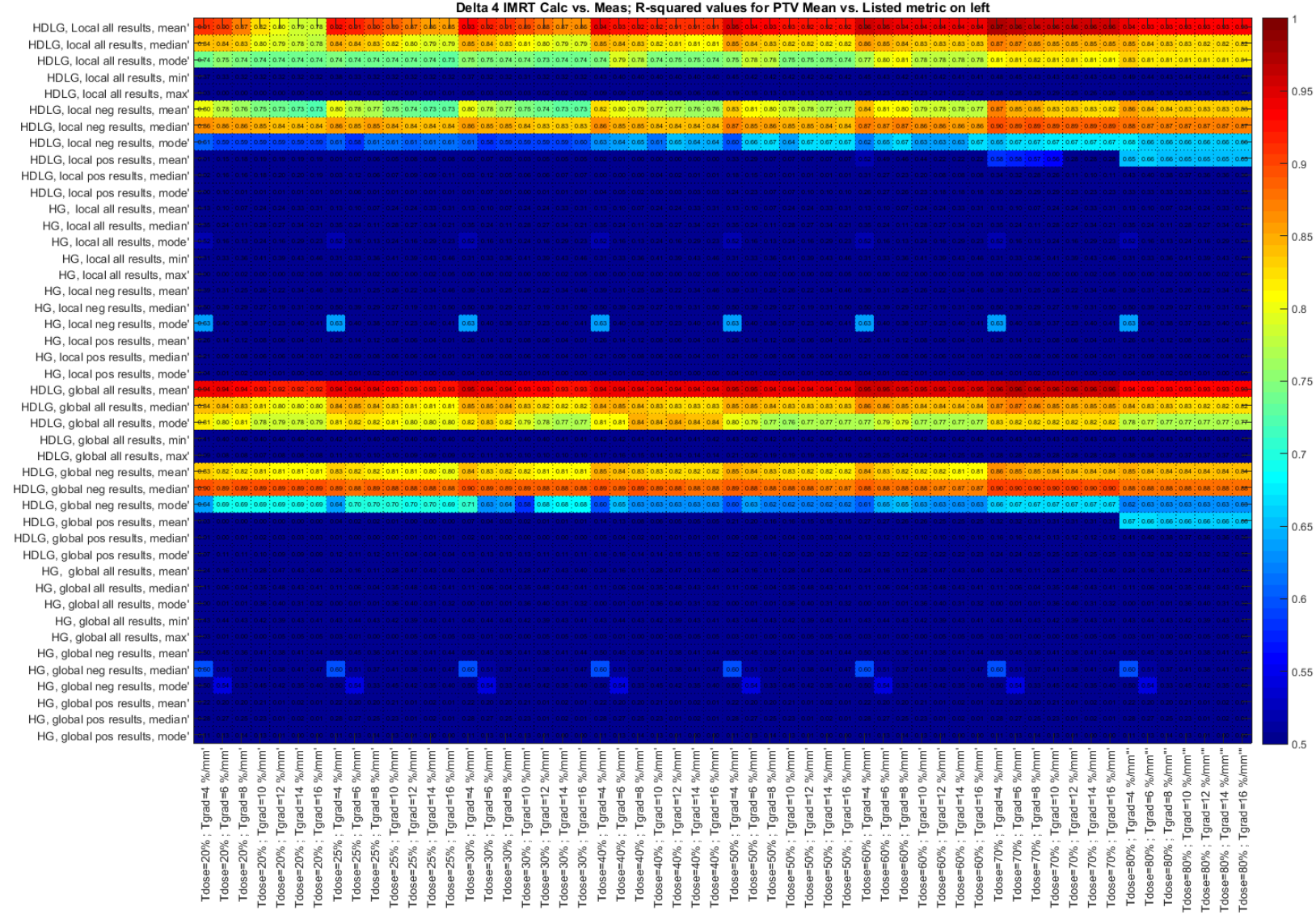

Figure A.5-4. Pearson r^2 heat map for IMRT cases on the Delta 4 for predicting the change in PTV mean dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

A.V.2 Delta 4 IMRT Ring Correlations

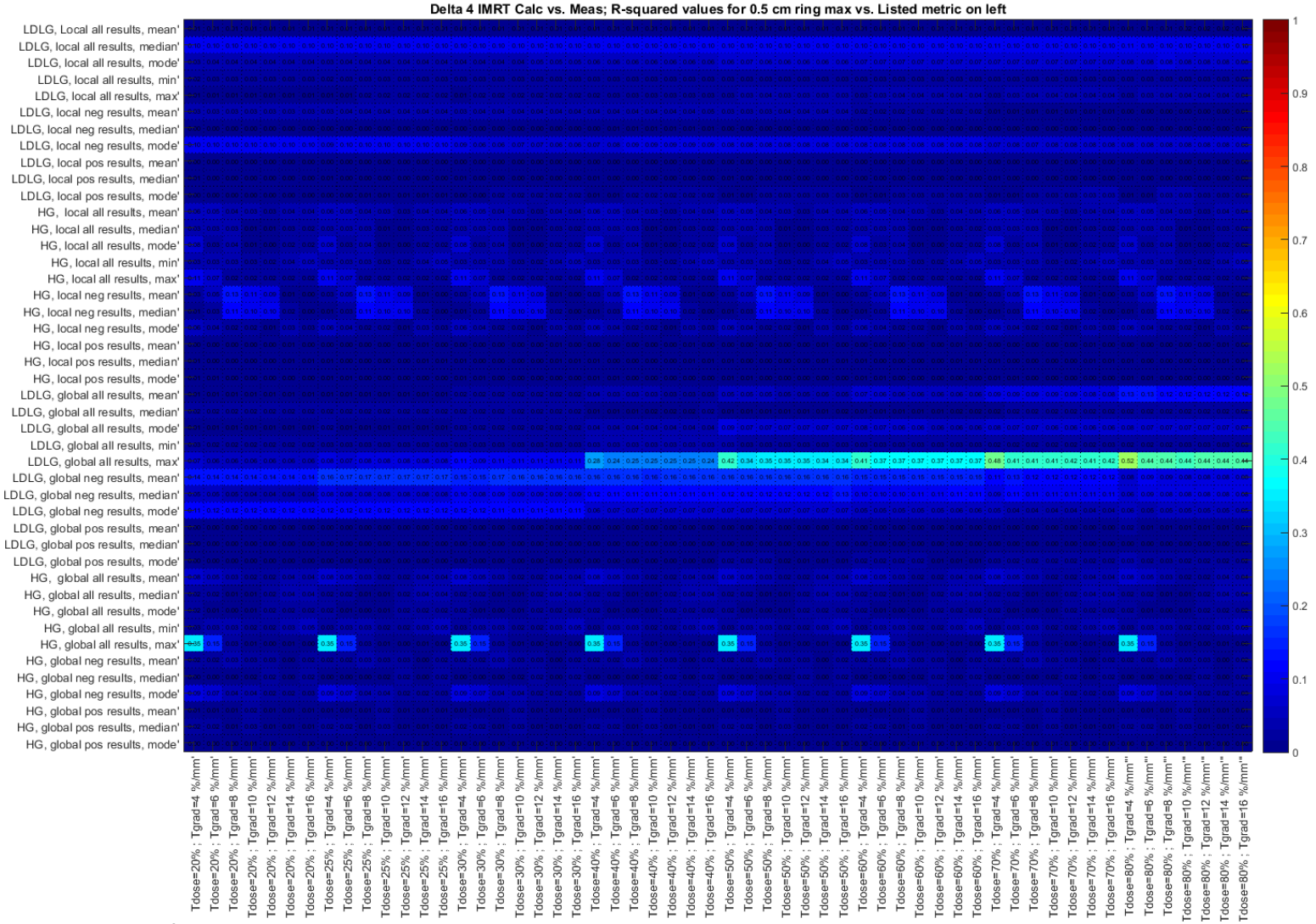

Figure A.5-5. Pearson r^2 heat map for IMRT cases on the Delta 4 for predicting the change in 0.5 cm ring maximum dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

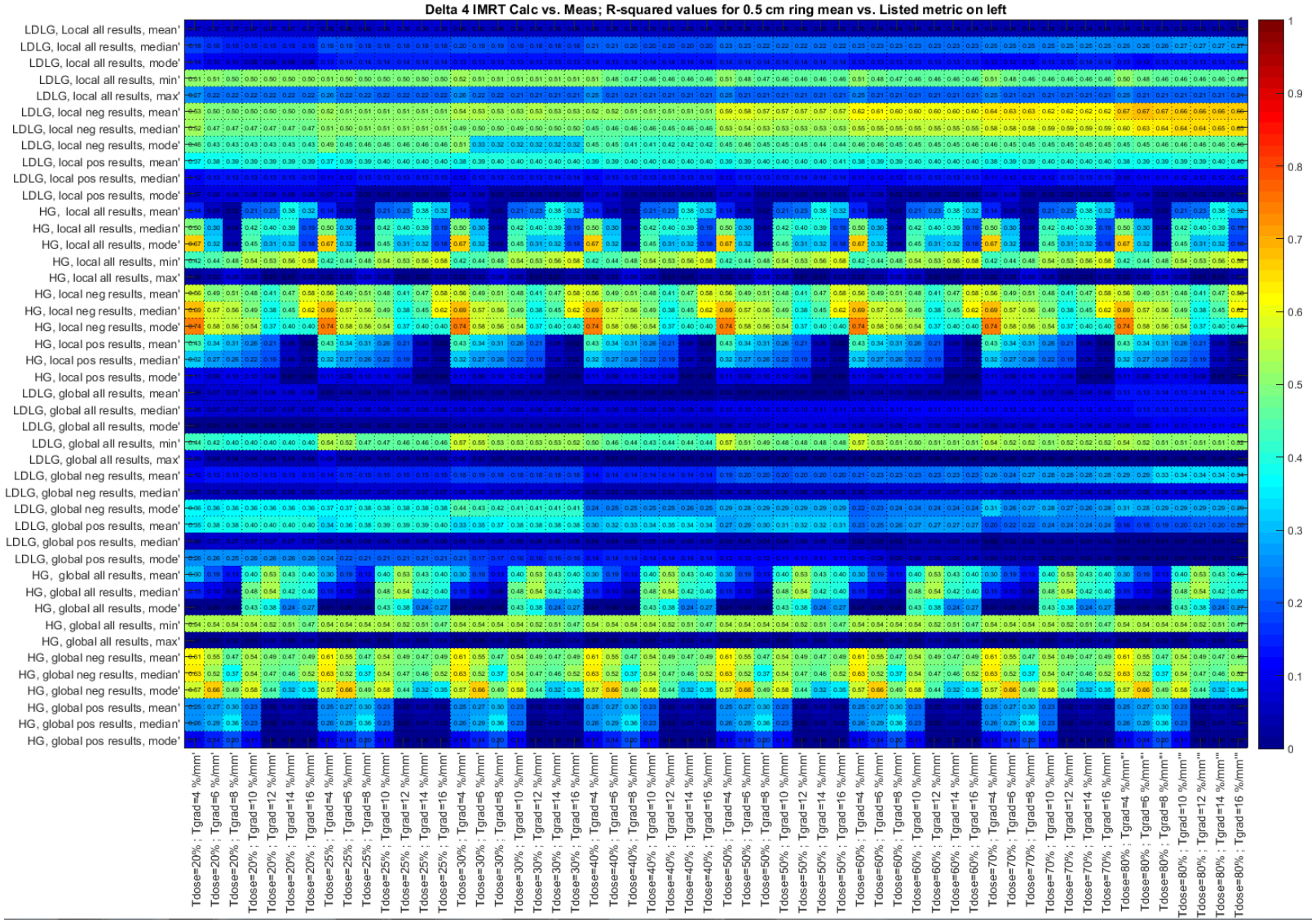

Figure A.5-6. Pearson r^2 heat map for IMRT cases on the Delta 4 for predicting the change in 0.5 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

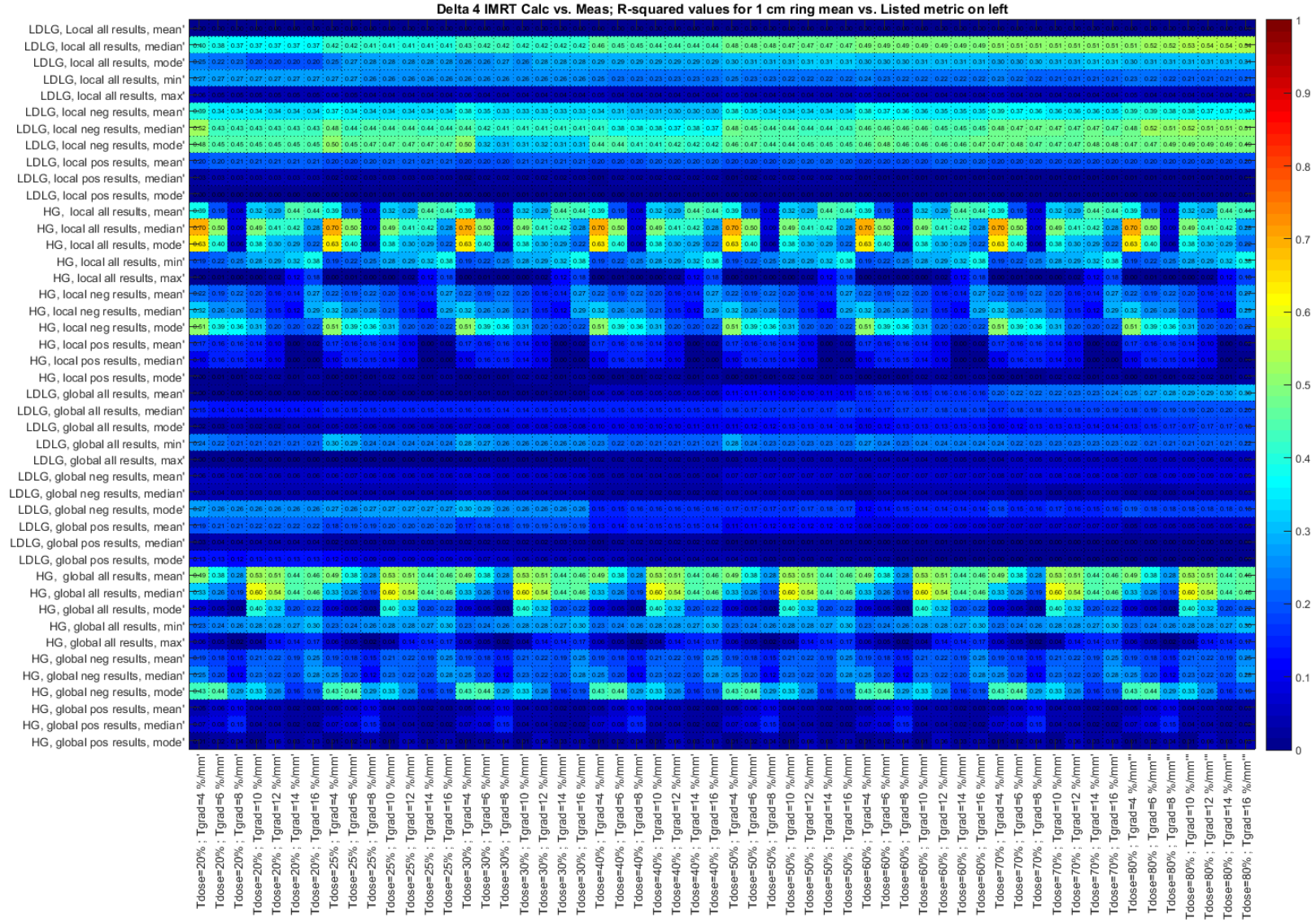

Figure A.5-7. Pearson r^2 heat map for IMRT cases on the Delta 4 for predicting the change in 1 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

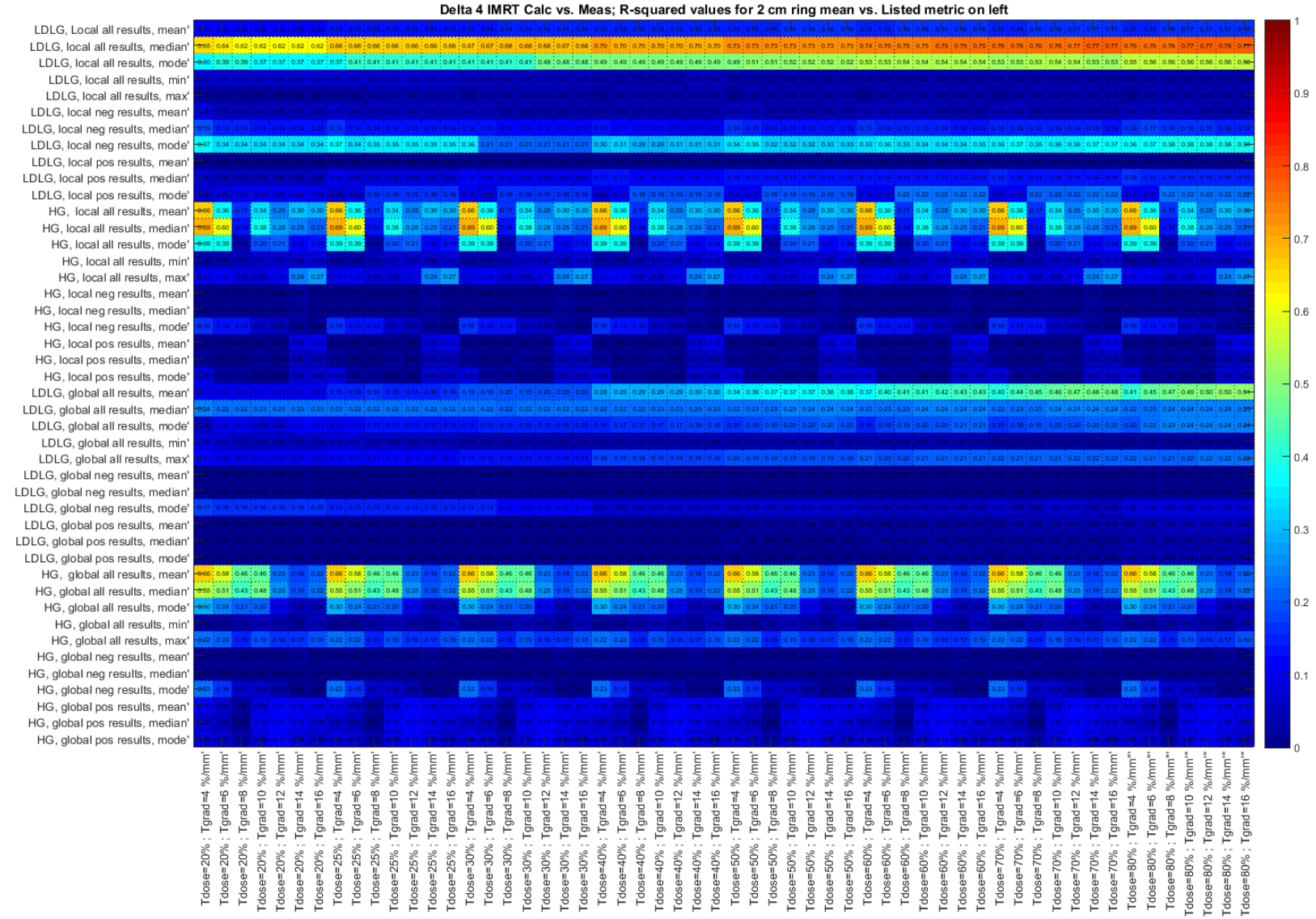

Figure A.5-8. Pearson r^2 heat map for IMRT cases on the Delta 4 for predicting the change in 2 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

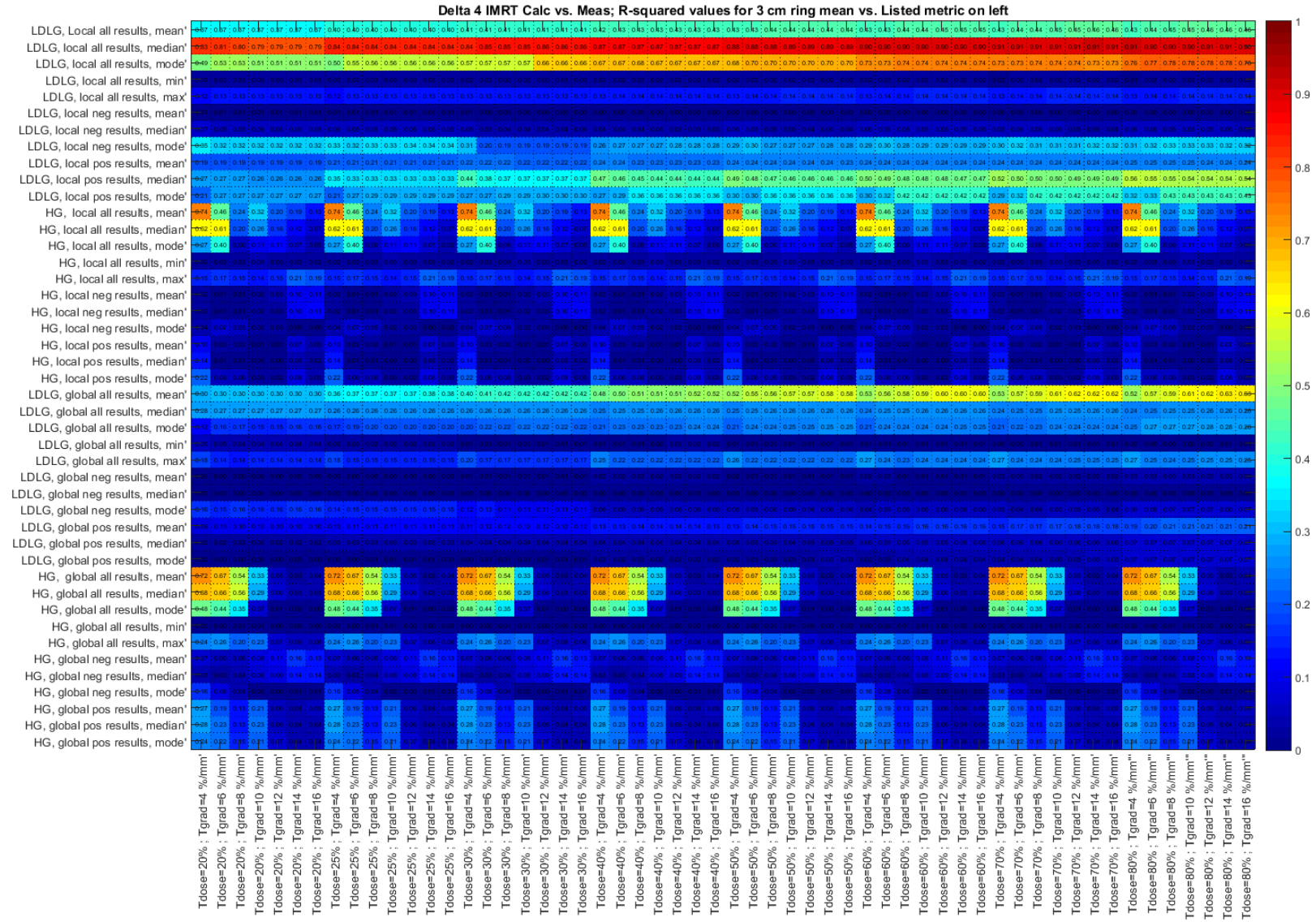

Figure A.5-9. Pearson r^2 heat map for IMRT cases on the Delta 4 for predicting the change in 3 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

A.V.3 ArcCHECK IMRT PTV Correlations

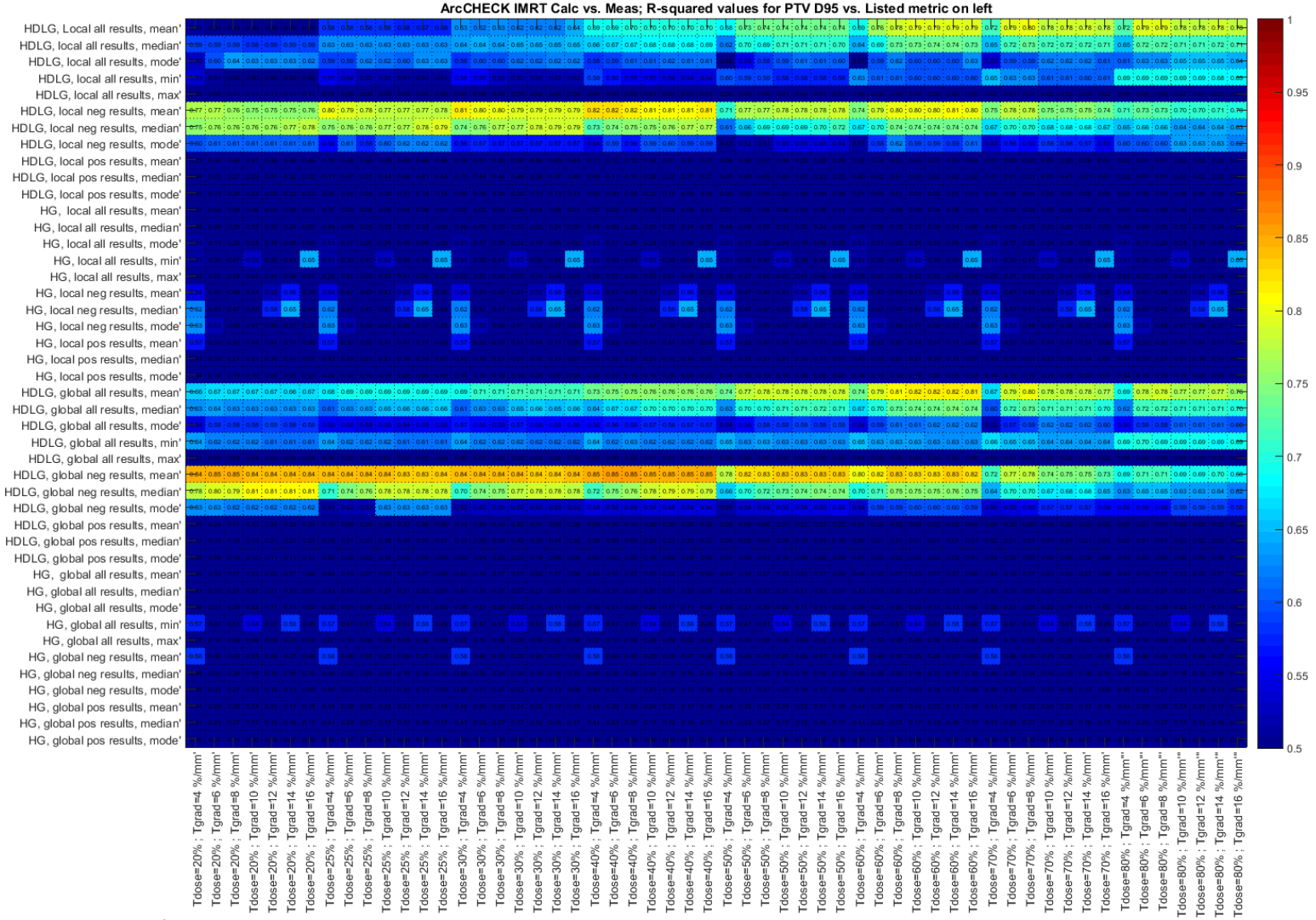

Figure A.5-10. Pearson r^2 heat map for IMRT cases on the ArcCHECK for predicting the change in PTV $D_{95\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

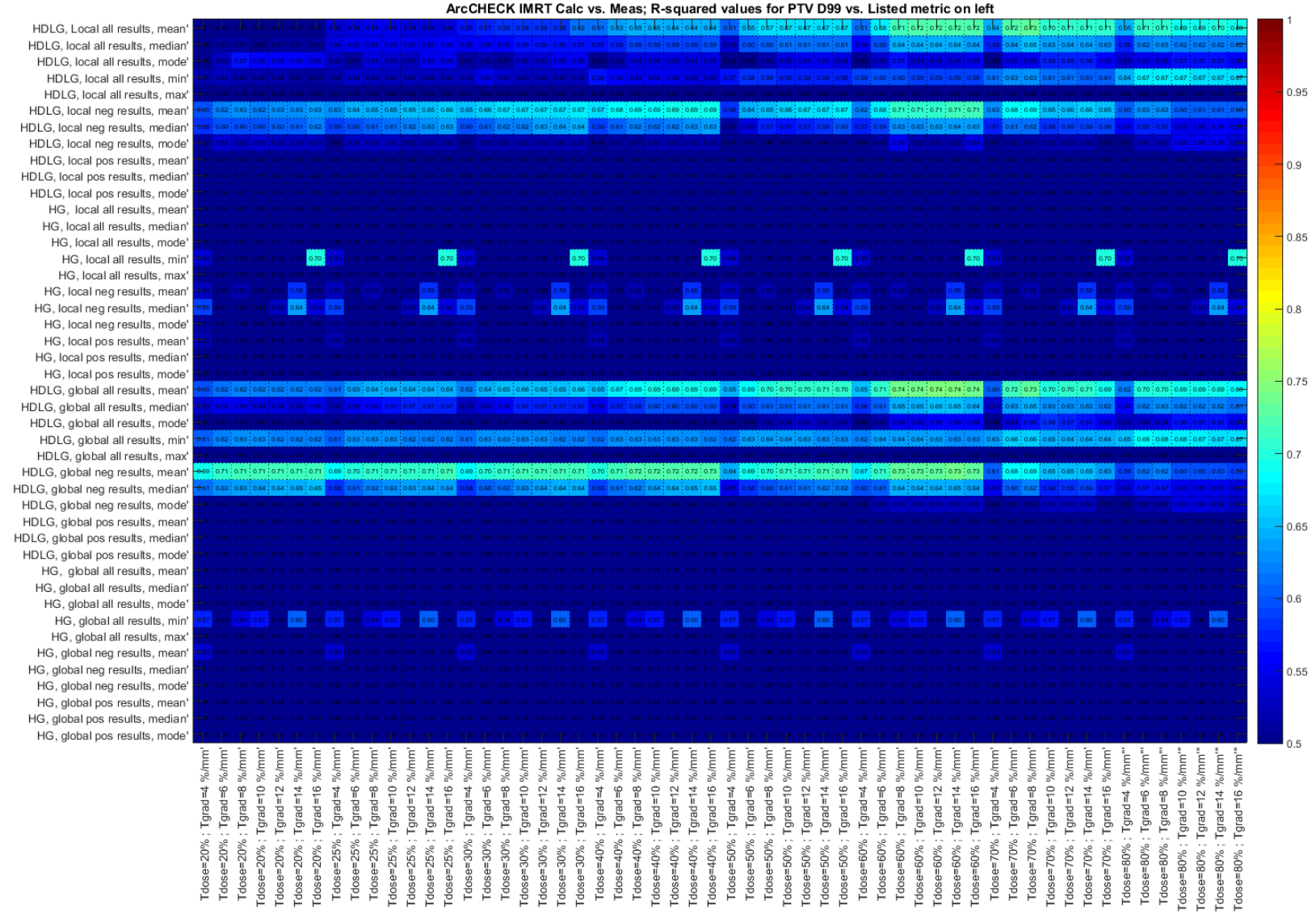

Figure A.5-11. Pearson r^2 heat map for IMRT cases on the ArcCHECK for predicting the change in PTV $D_{99\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

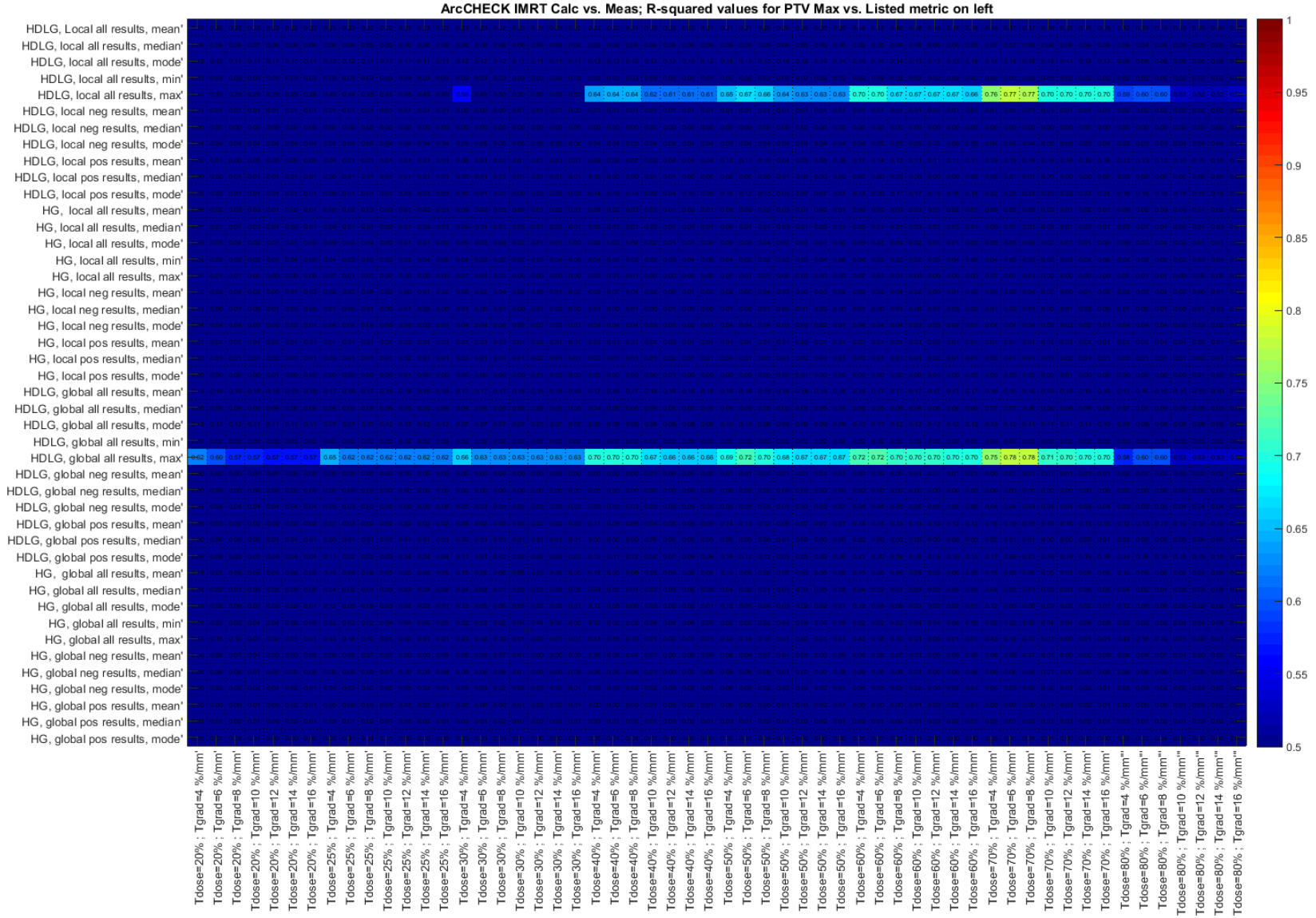

Figure A.5-12. Pearson r^2 heat map for IMRT cases on the ArcCHECK for predicting the change in PTV maximum dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

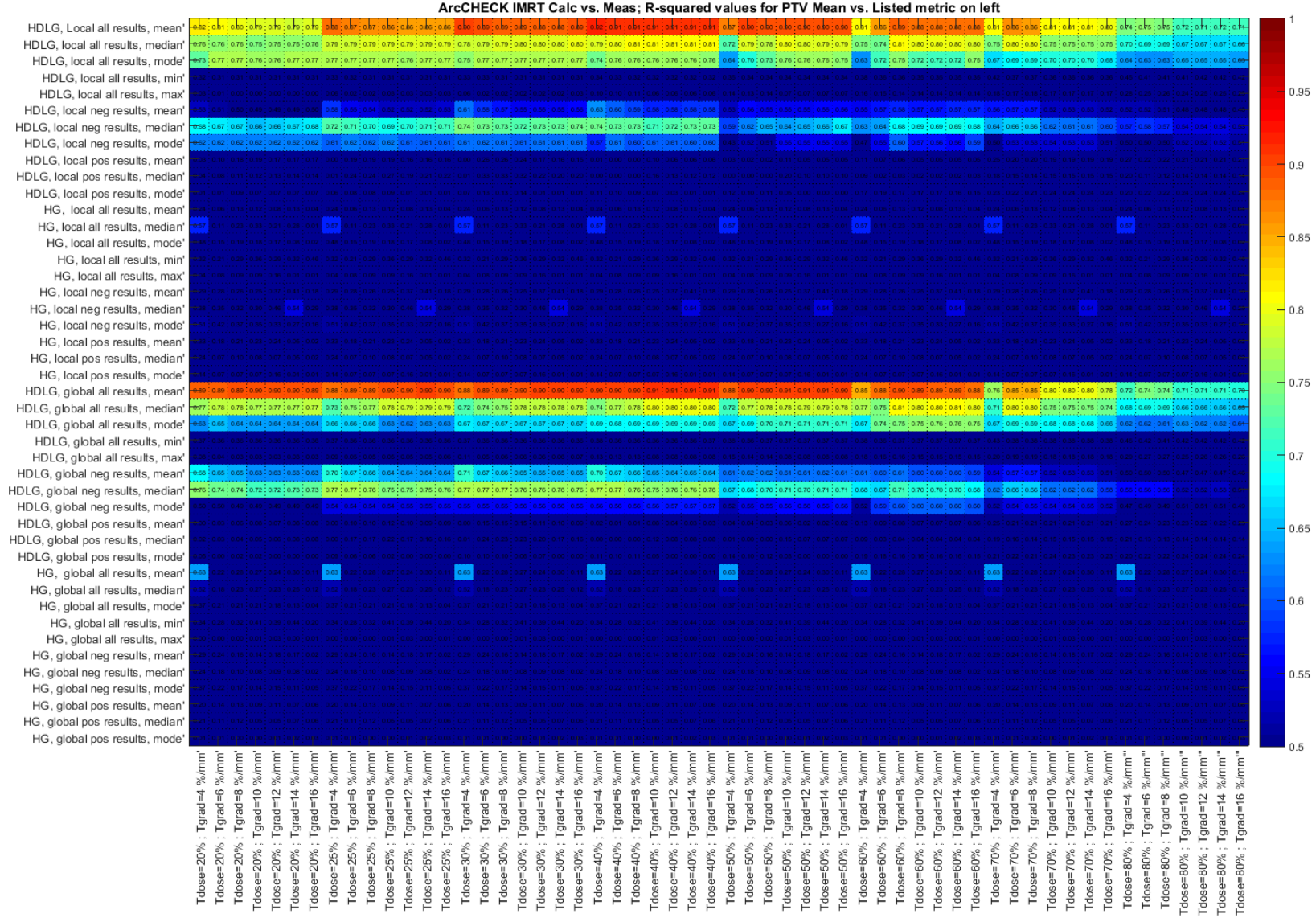

Figure A.5-13. Pearson r^2 heat map for IMRT cases on the ArcCHECK for predicting the change in PTV mean dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

A.V.4 ArcCHECK IMRT Ring Correlations

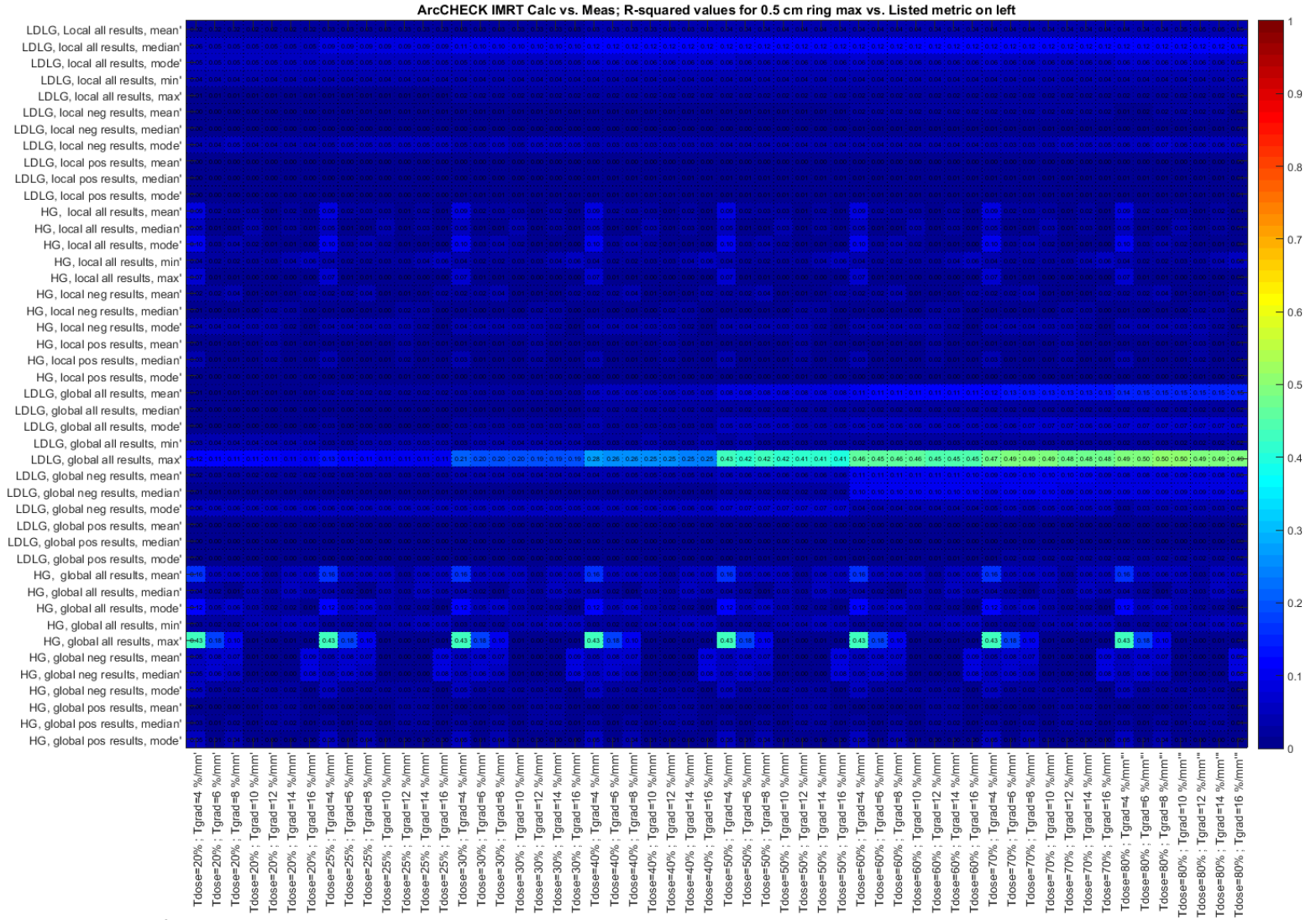

Figure A.5-14. Pearson r^2 heat map for IMRT cases on the ArcCHECK for predicting the change in 0.5 cm ring maximum dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

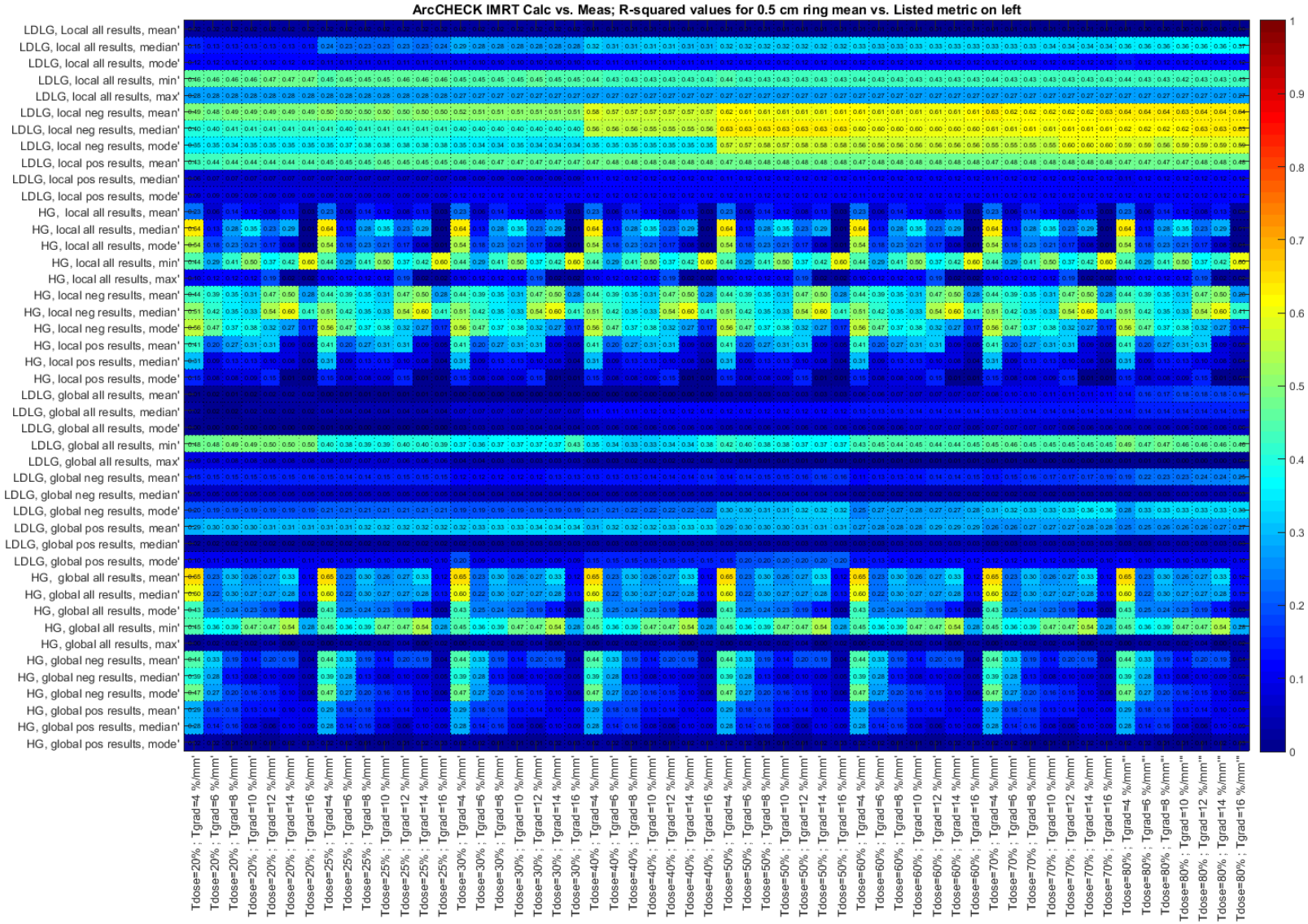

Figure A.5-15. Pearson r^2 heat map for IMRT cases on the ArcCHECK for predicting the change in 0.5 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

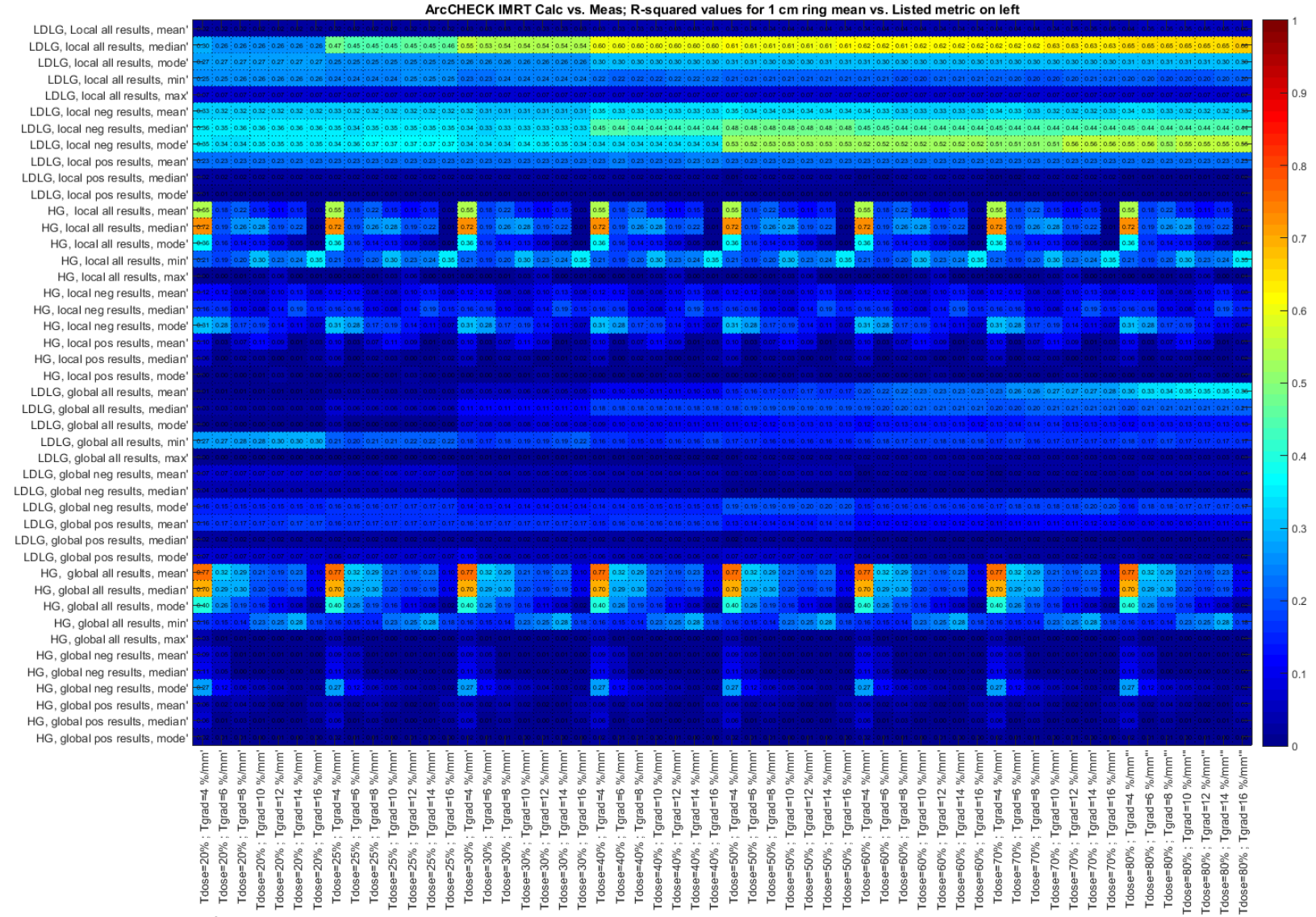

Figure A.5-16. Pearson r^2 heat map for IMRT cases on the ArcCHECK for predicting the change in 1 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

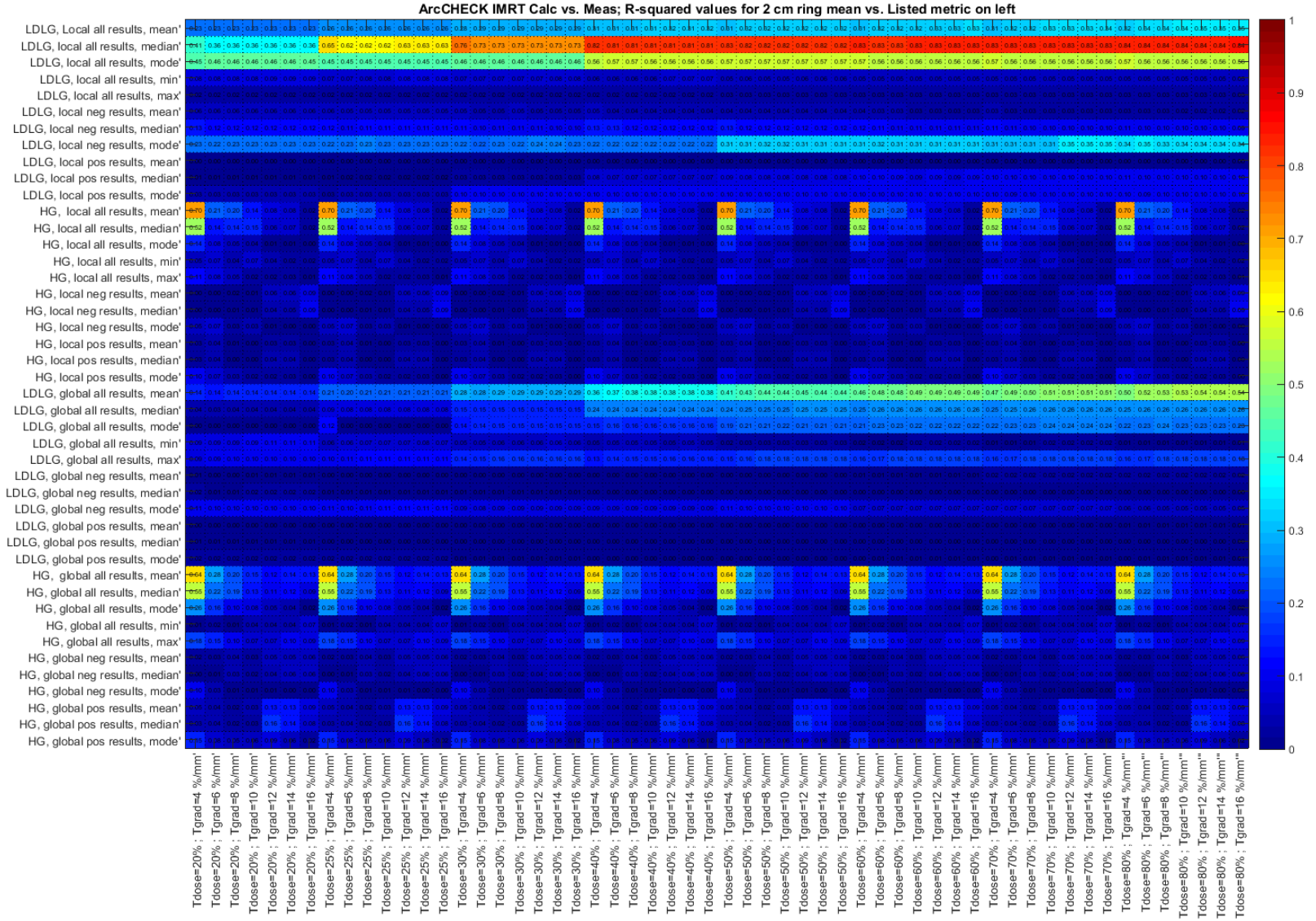

Figure A.5-17. Pearson r^2 heat map for IMRT cases on the ArcCHECK for predicting the change in 2 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

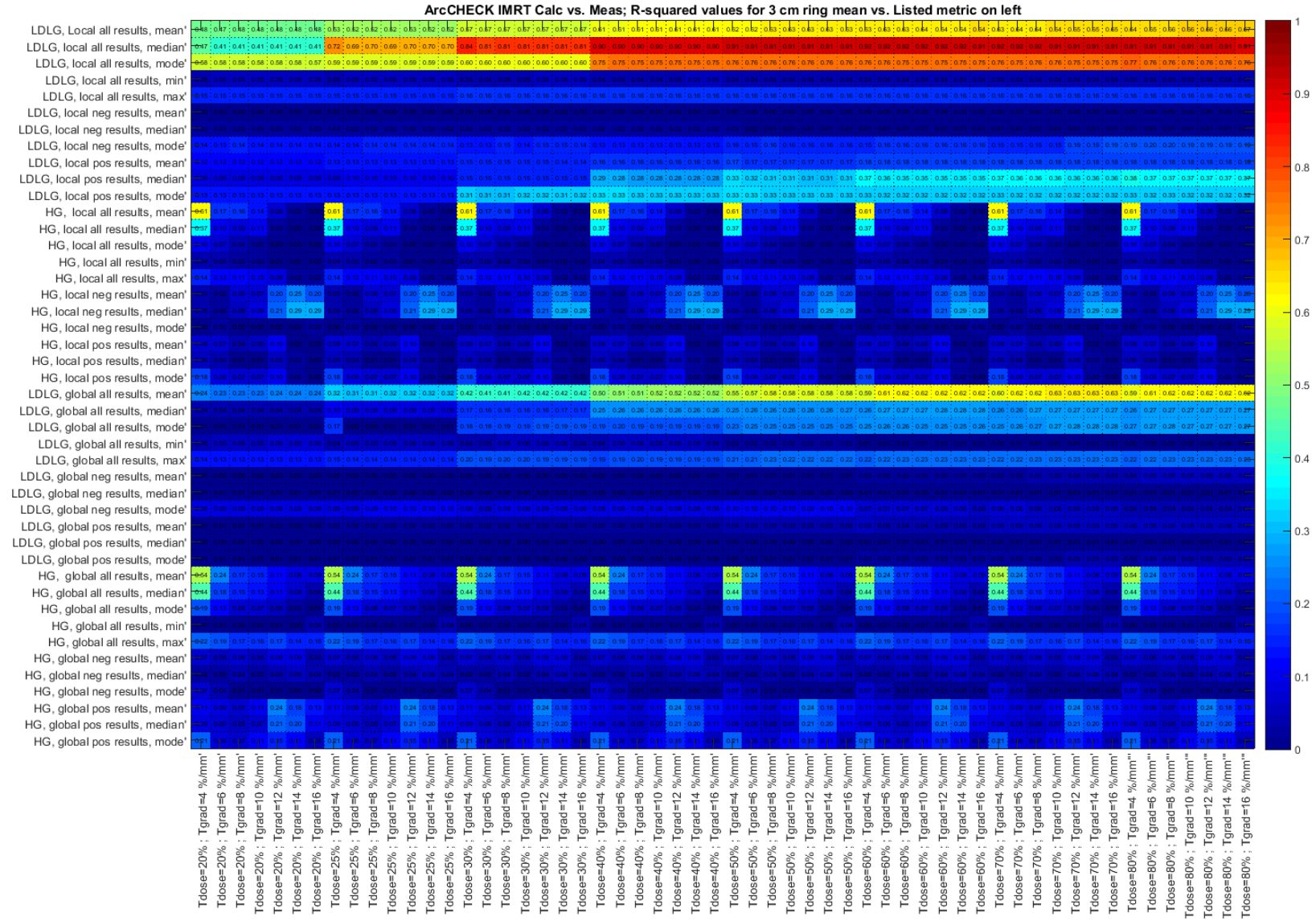

Figure A.5-18. Pearson r^2 heat map for IMRT cases on the ArcCHECK for predicting the change in 3 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

A.V.5 MapCHECK IMRT PTV Correlations

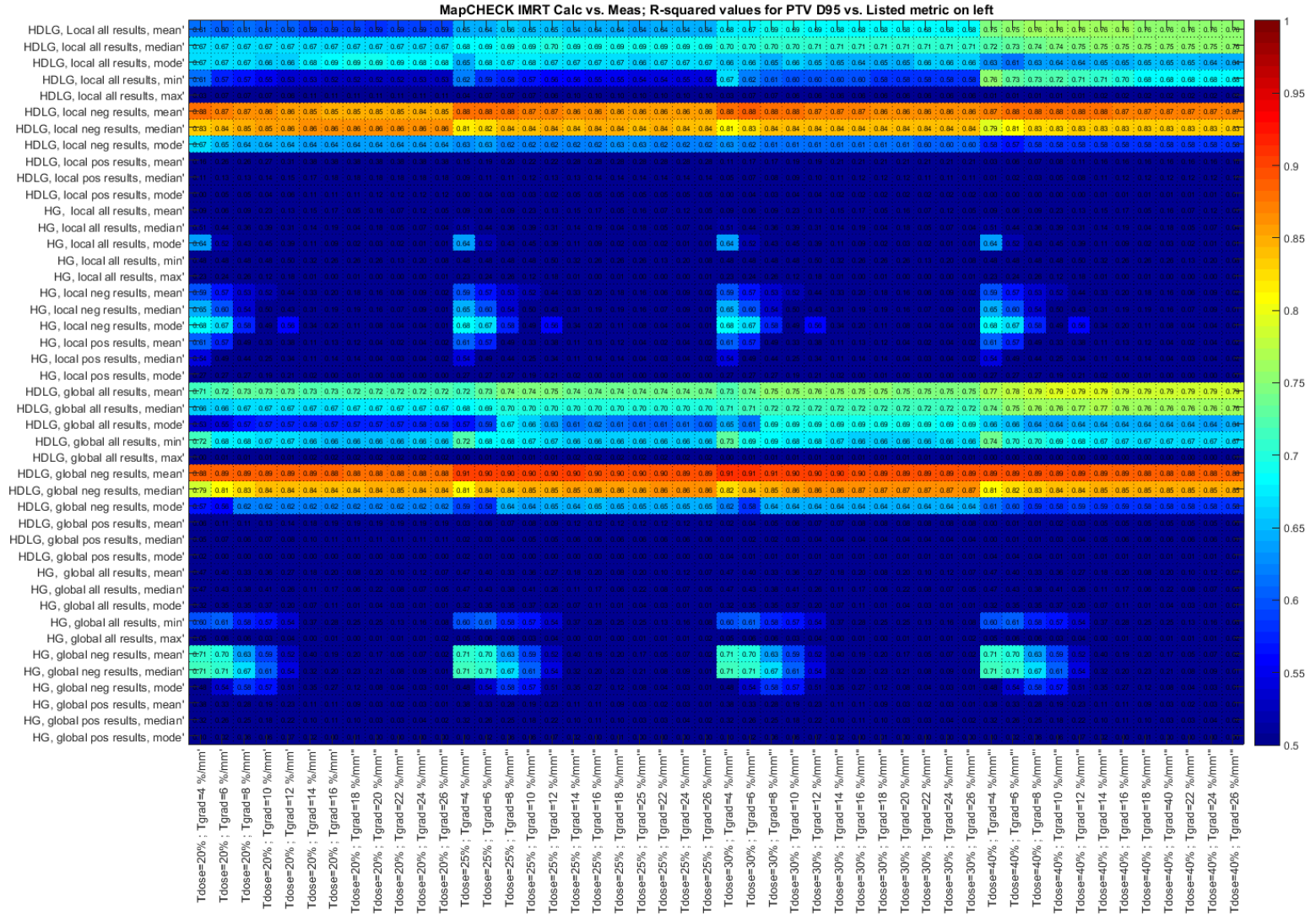

Figure A.5-19. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in PTV $D_{95\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

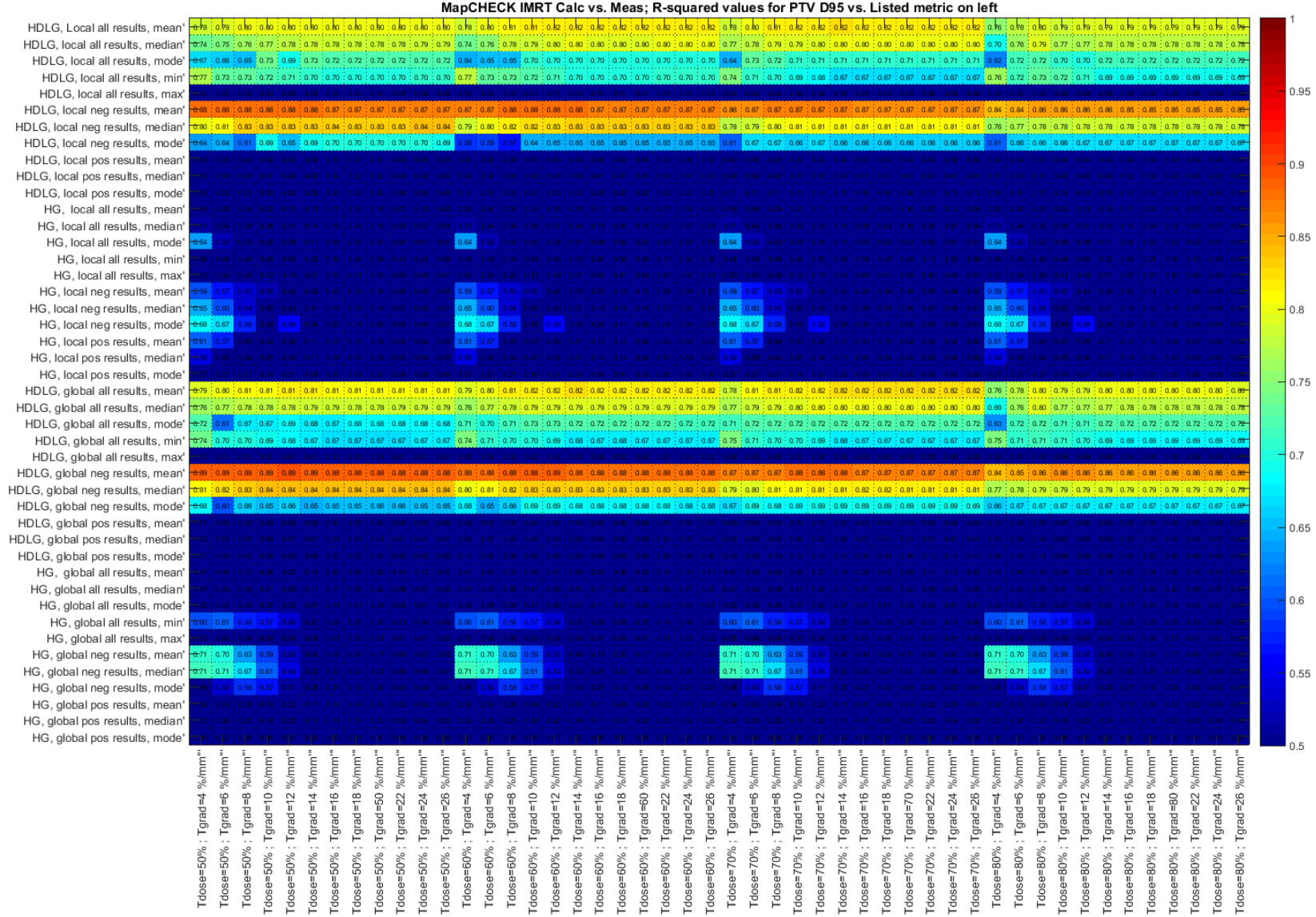

Figure A.5-20. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in PTV $D_{95\%}$ in the patient plan using histogram obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

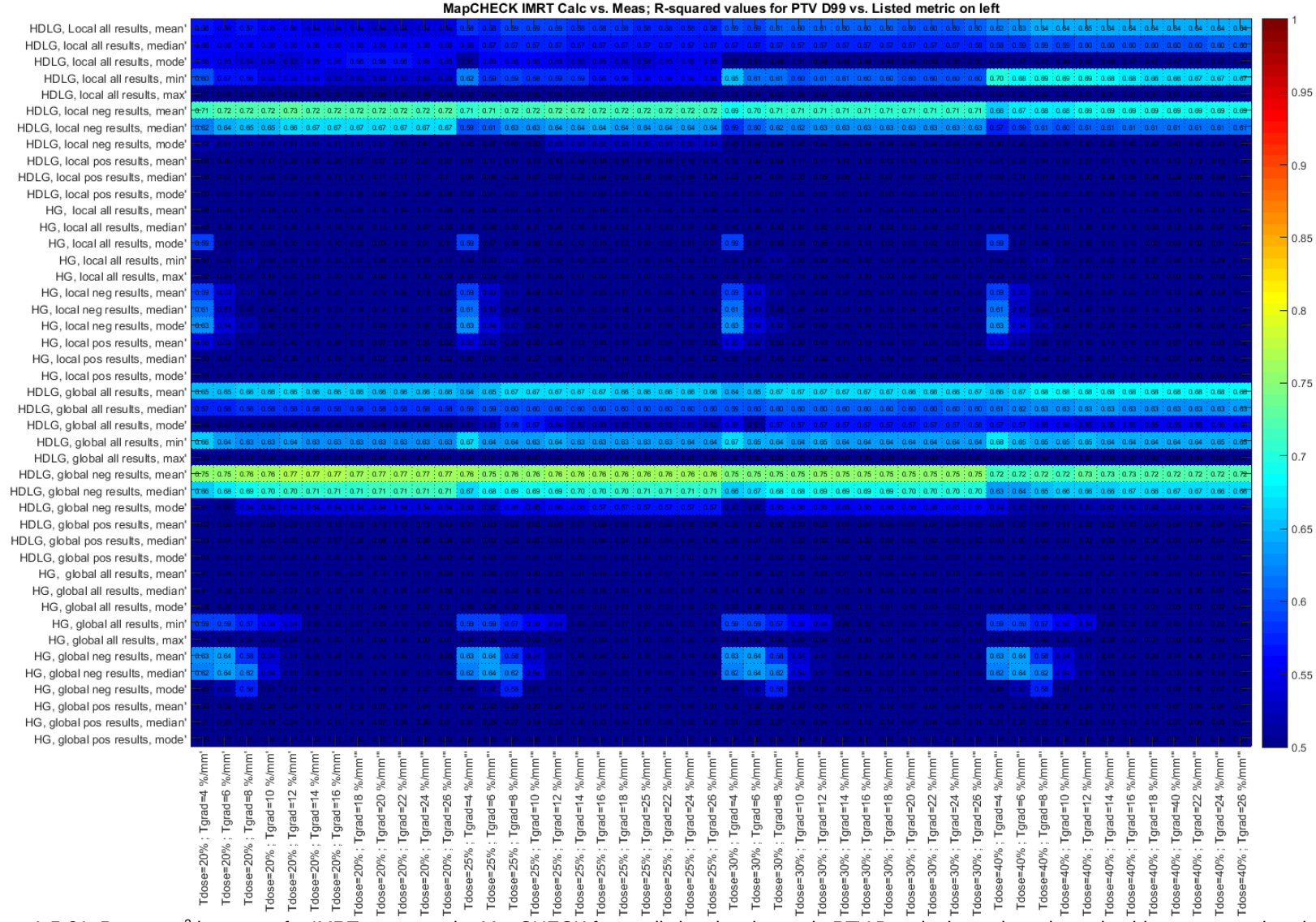

Figure A.5-21. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in PTV $D_{99\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

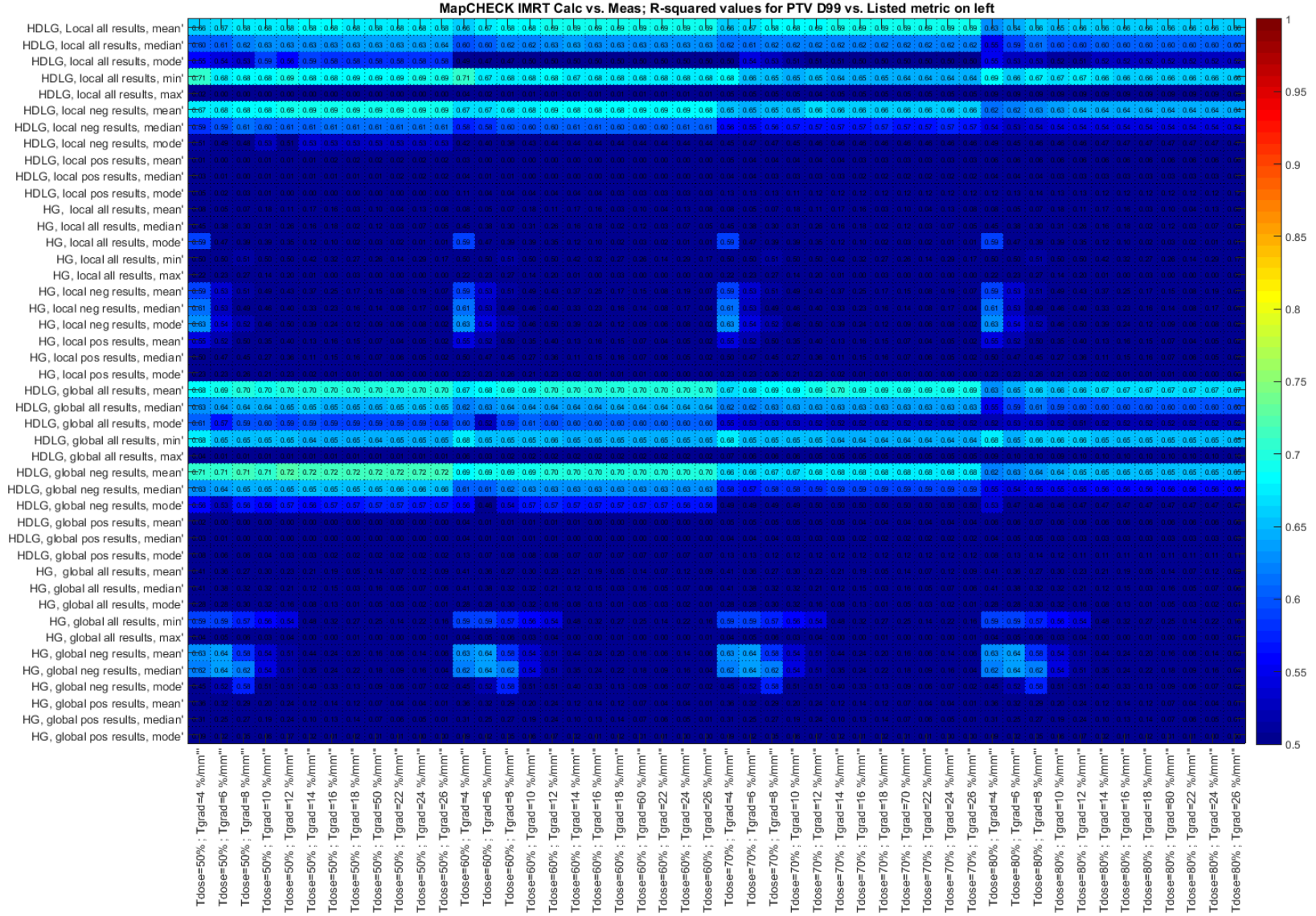

Figure A.5-22. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in PTV $D_{99\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

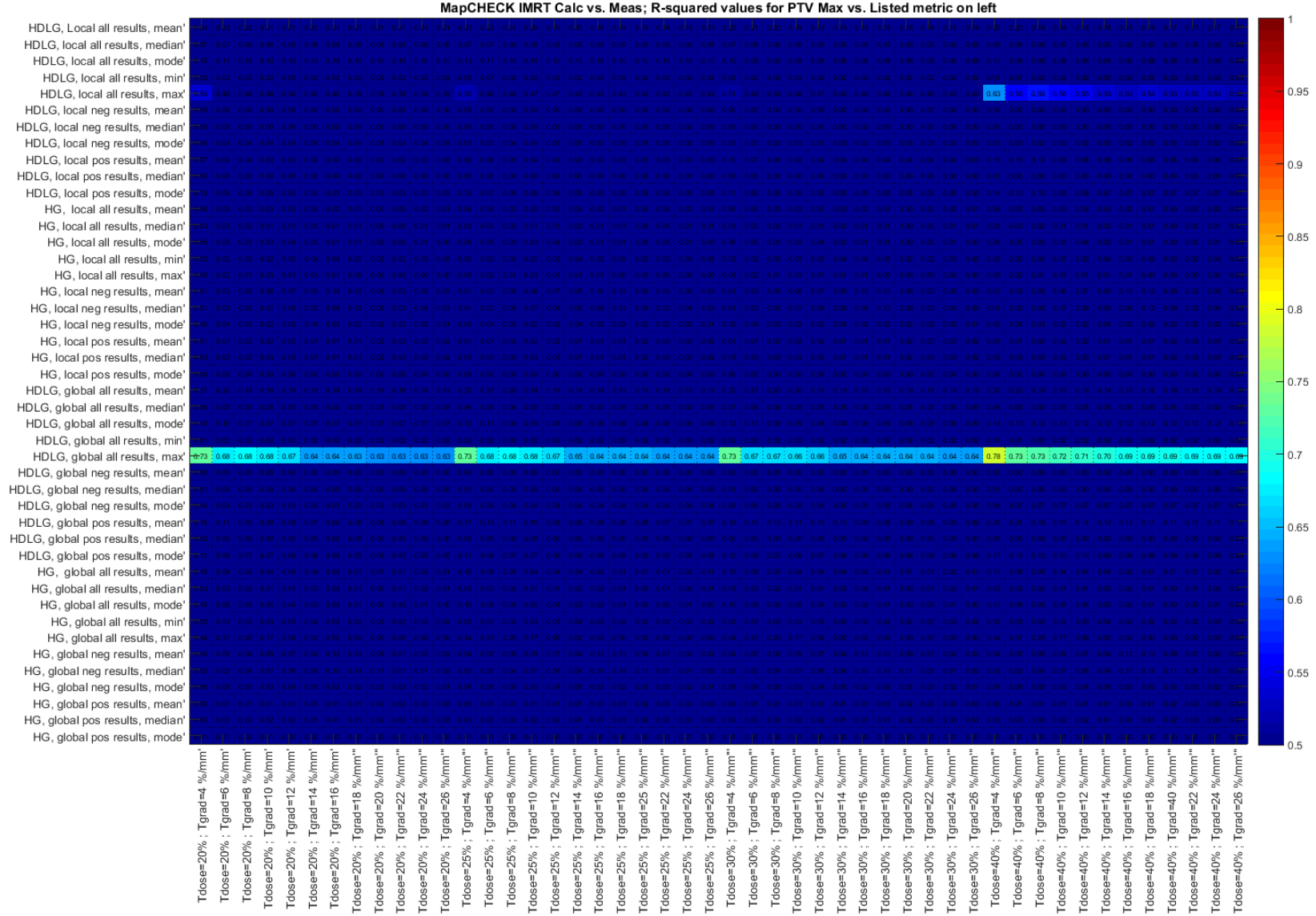

Figure A.5-23. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in PTV maximum dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

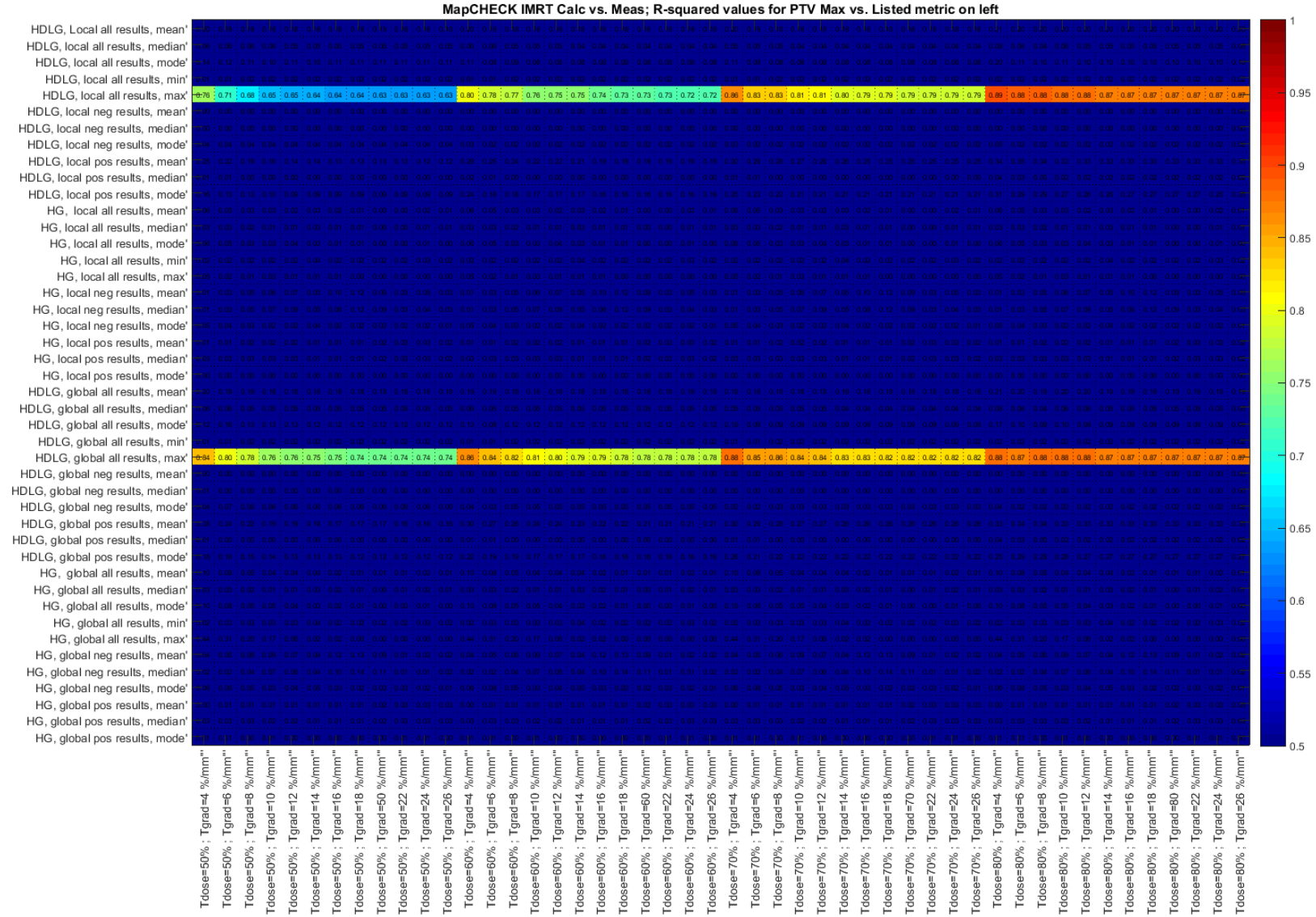

Figure A.5-24. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in PTV maximum dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

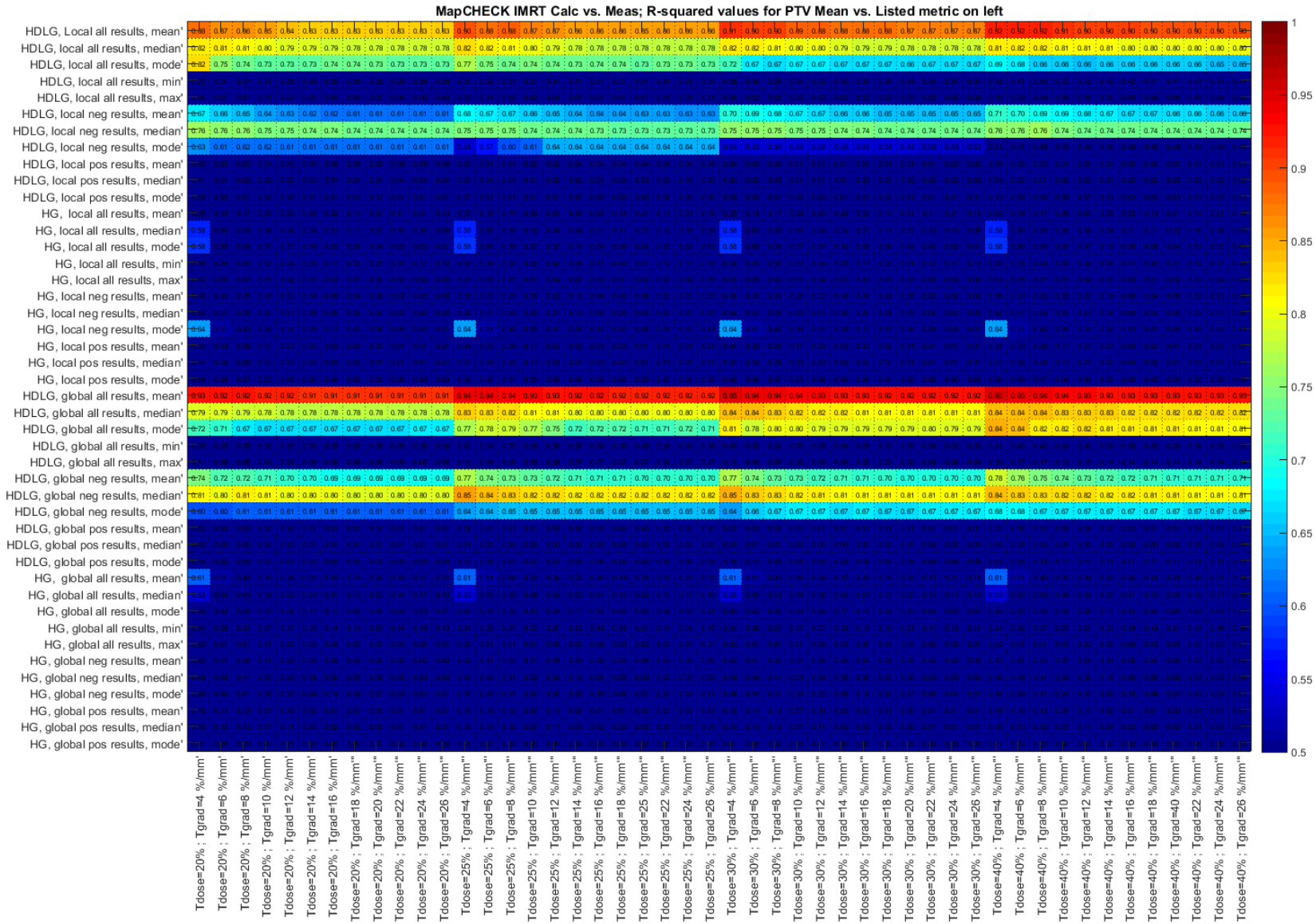

Figure A.5-25. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in PTV mean dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

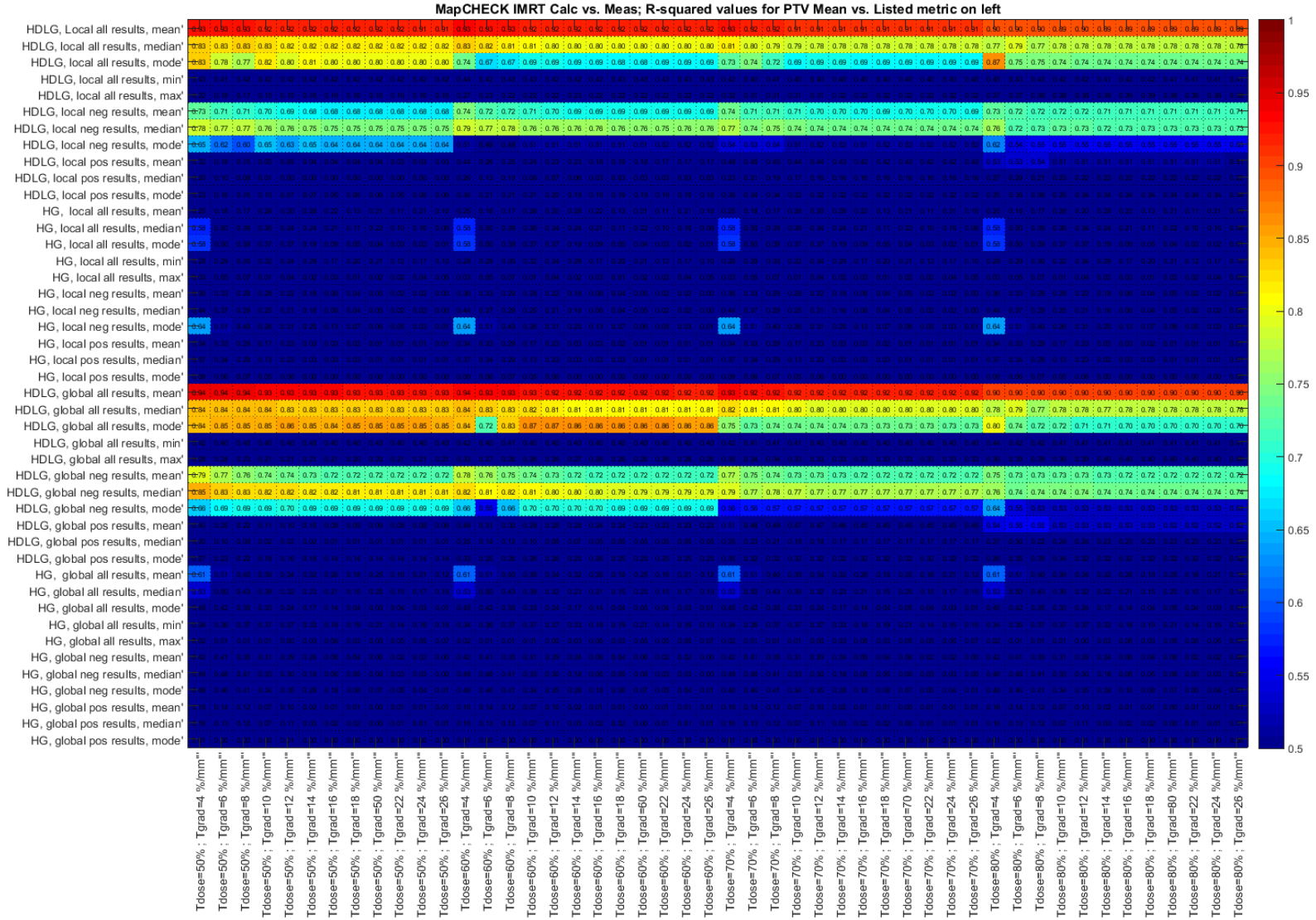

Figure A.5-26. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in PTV mean dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

A.V.6 MapCHECK IMRT Ring Correlations

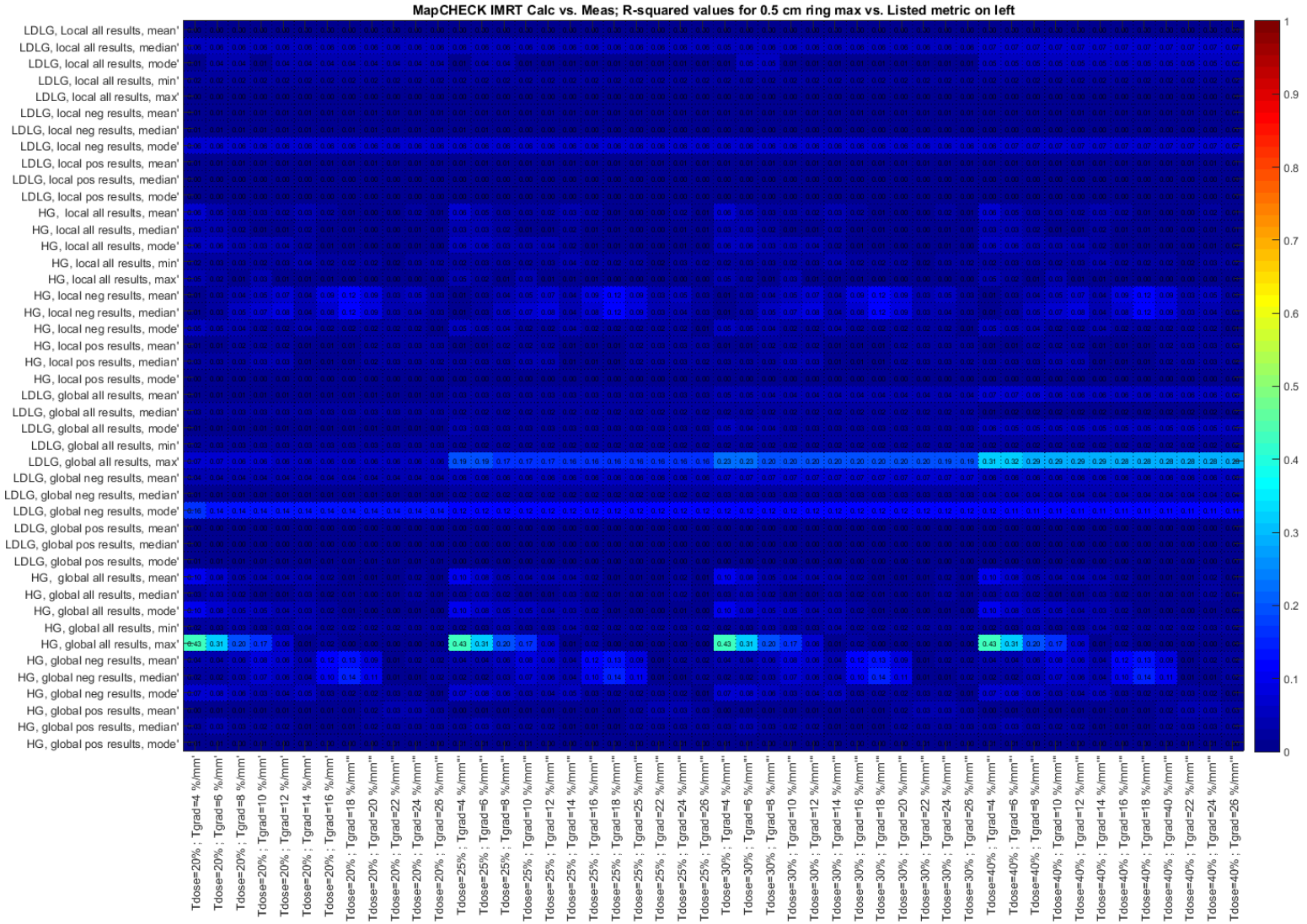

Figure A.5-27. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 0.5 cm ring maximum dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

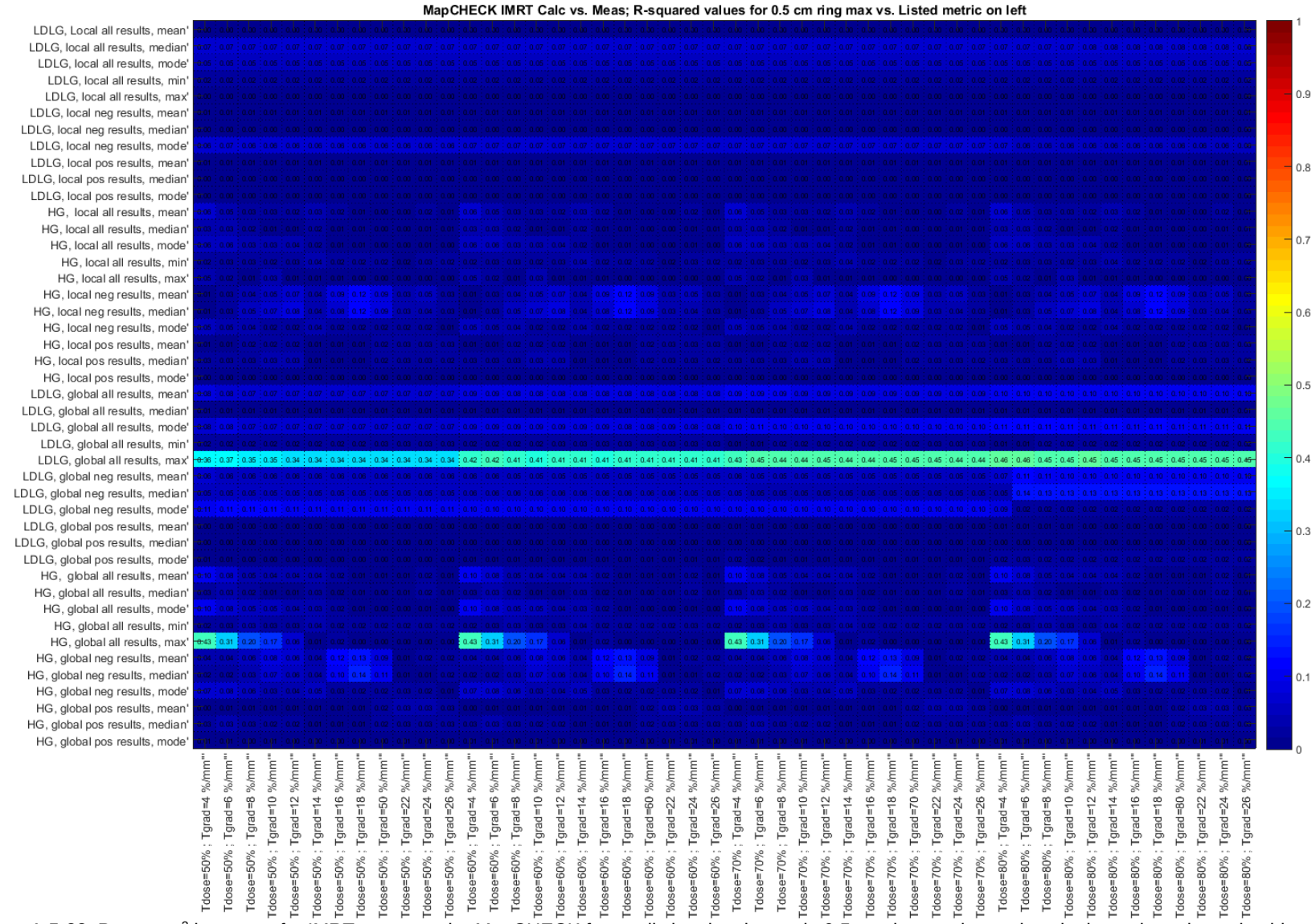

Figure A.5-28. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 0.5 cm ring maximum dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

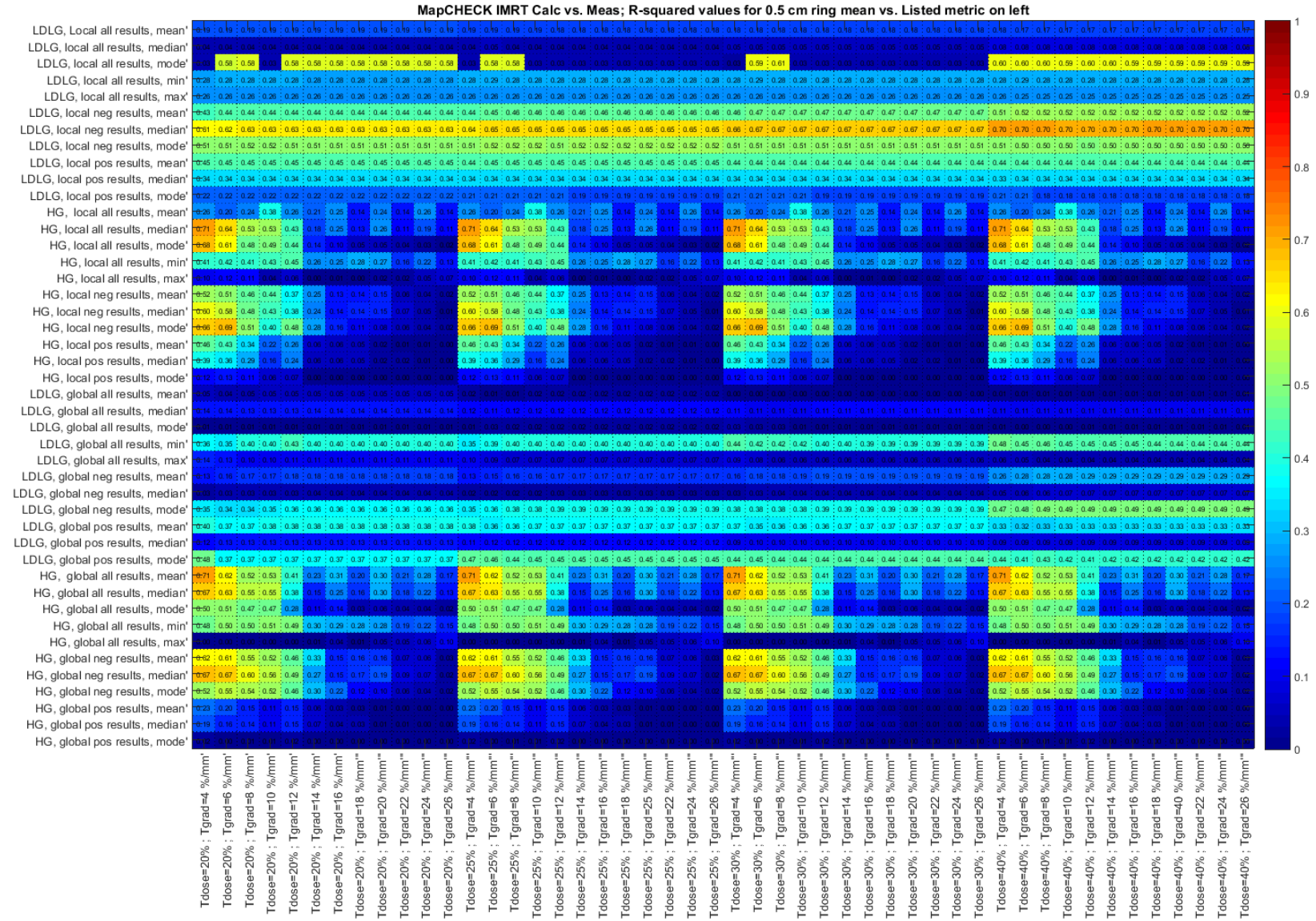

Figure A.5-29. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 0.5 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

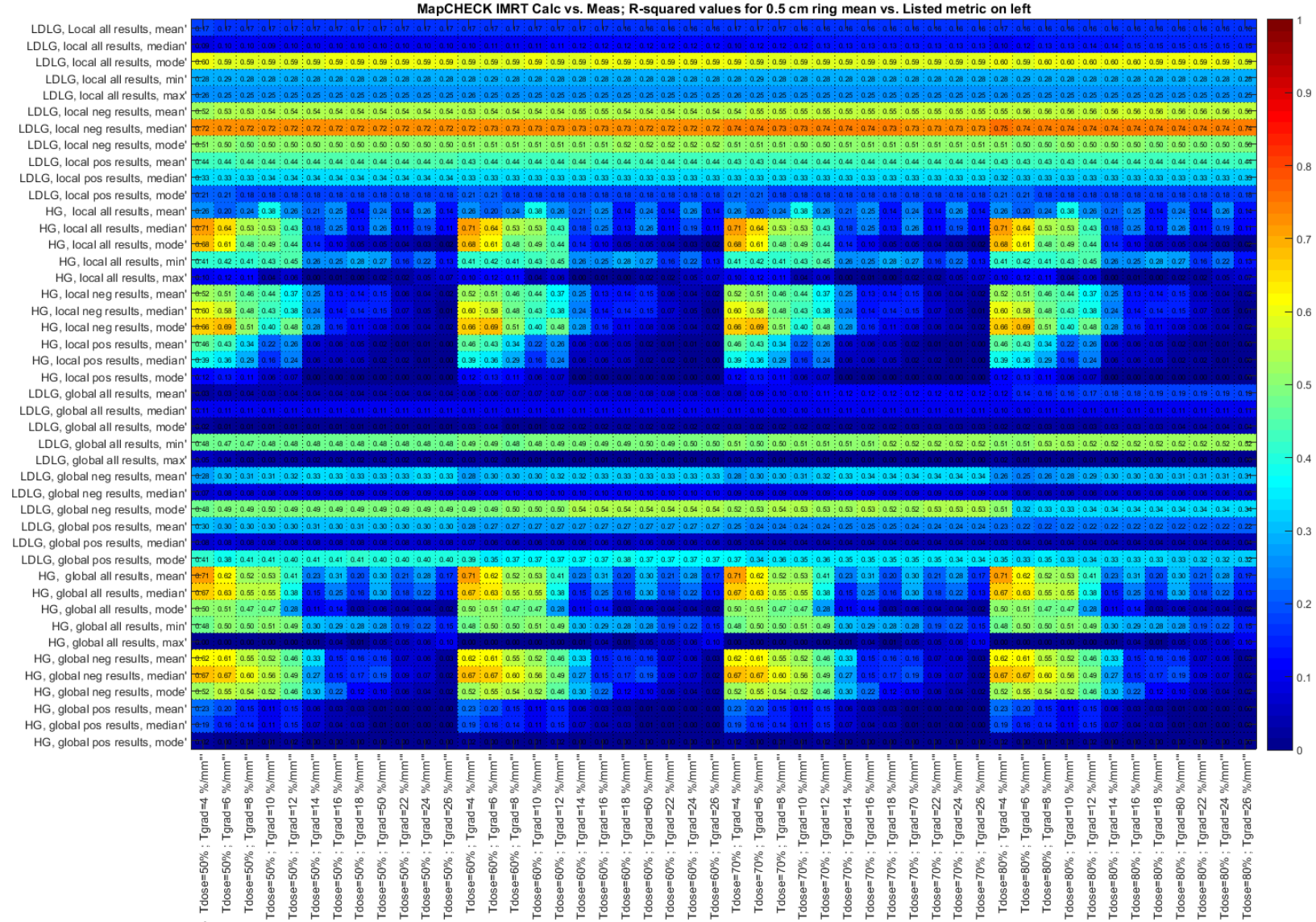

Figure A.5-30. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 0.5 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

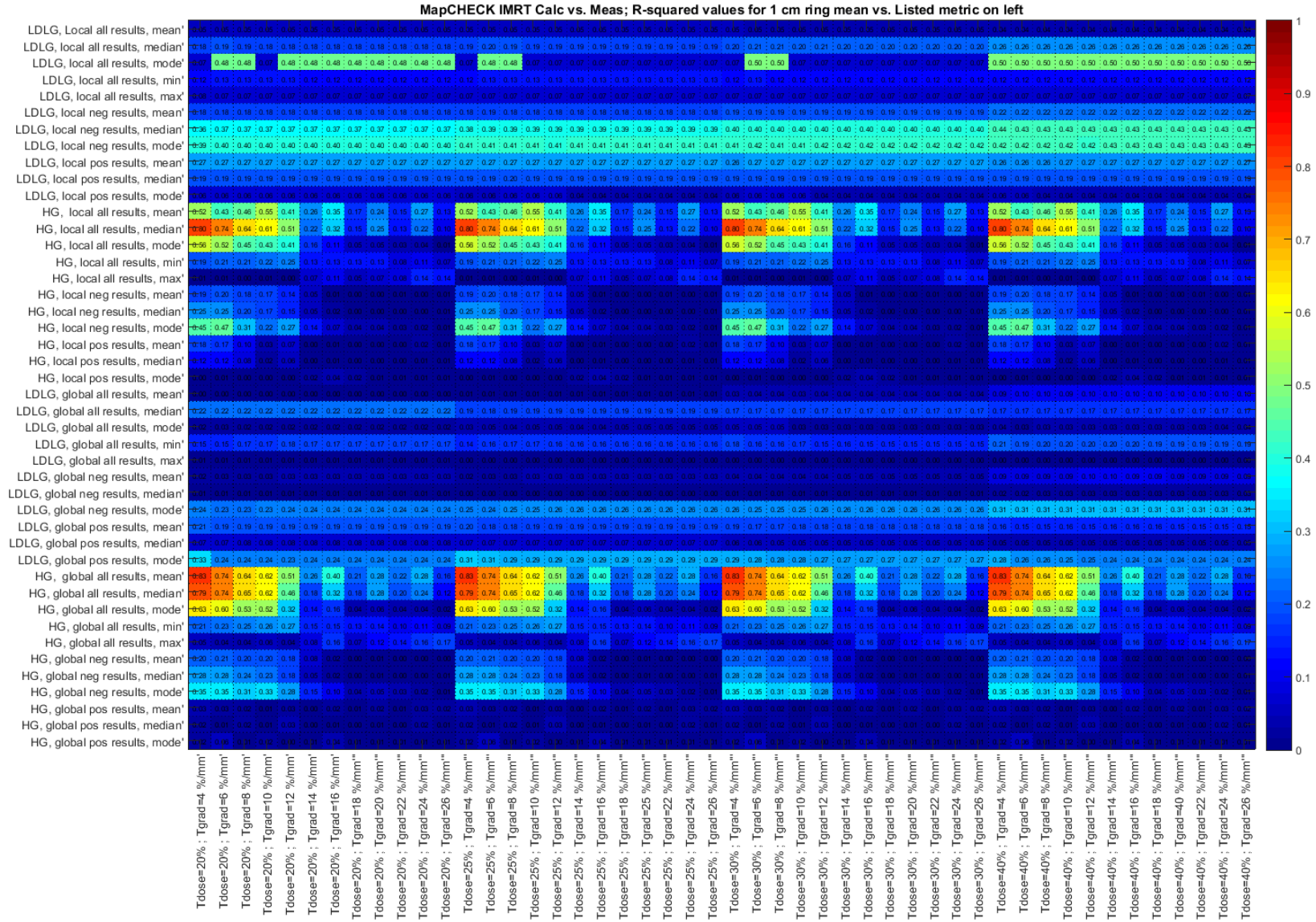

Figure A.5-31. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 1 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

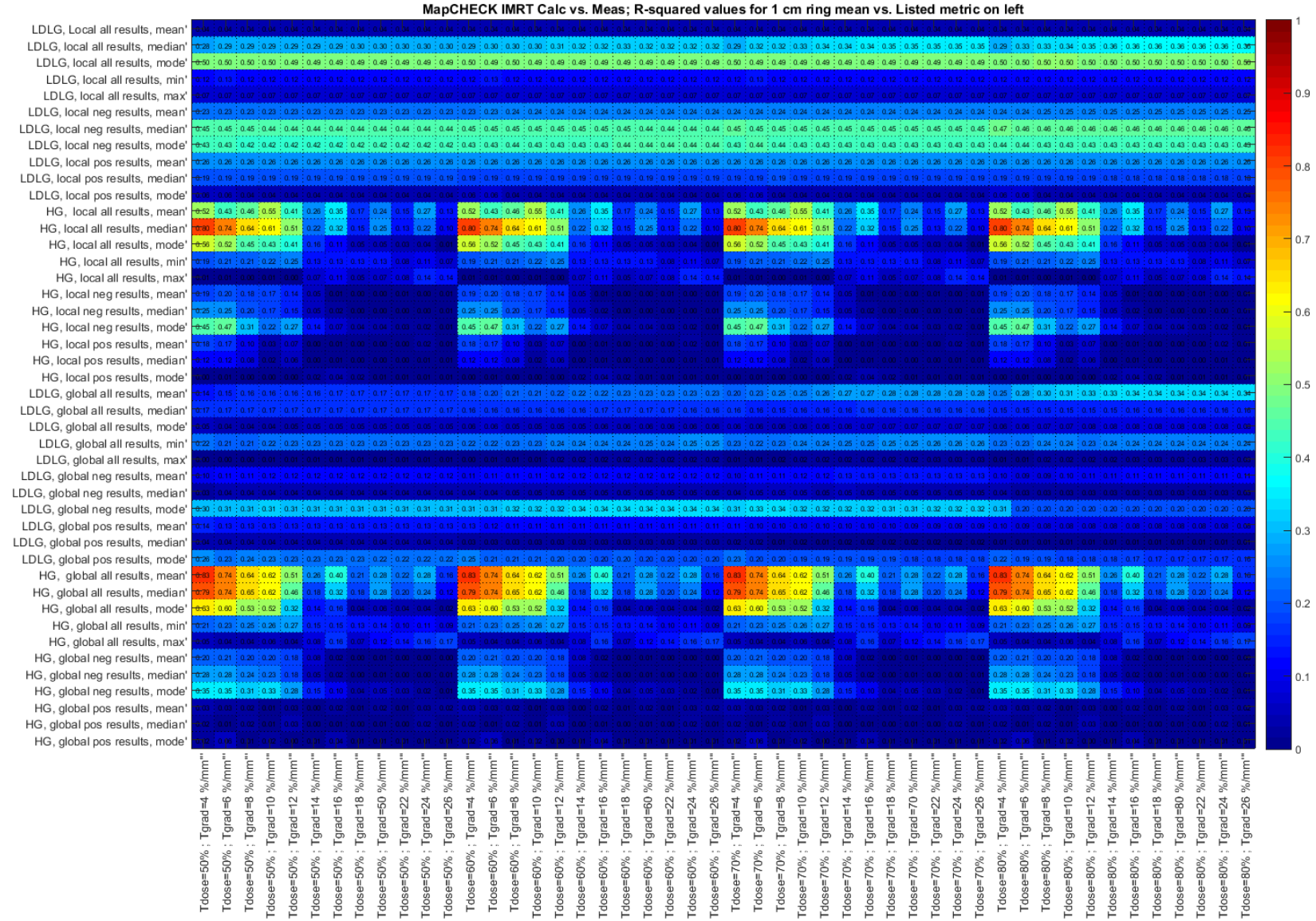

Figure A.5-32. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 1 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

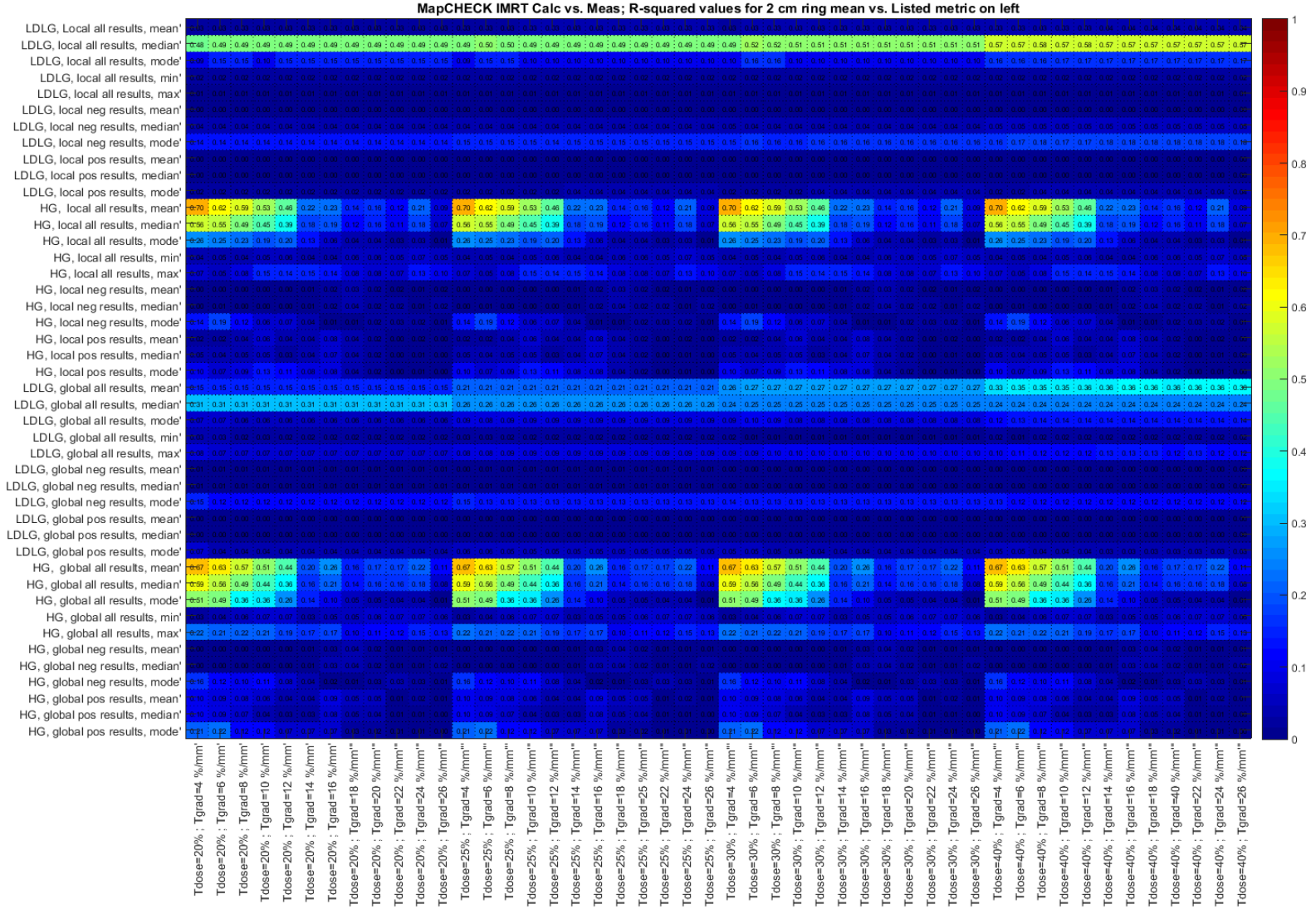

Figure A.5-33. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 2 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

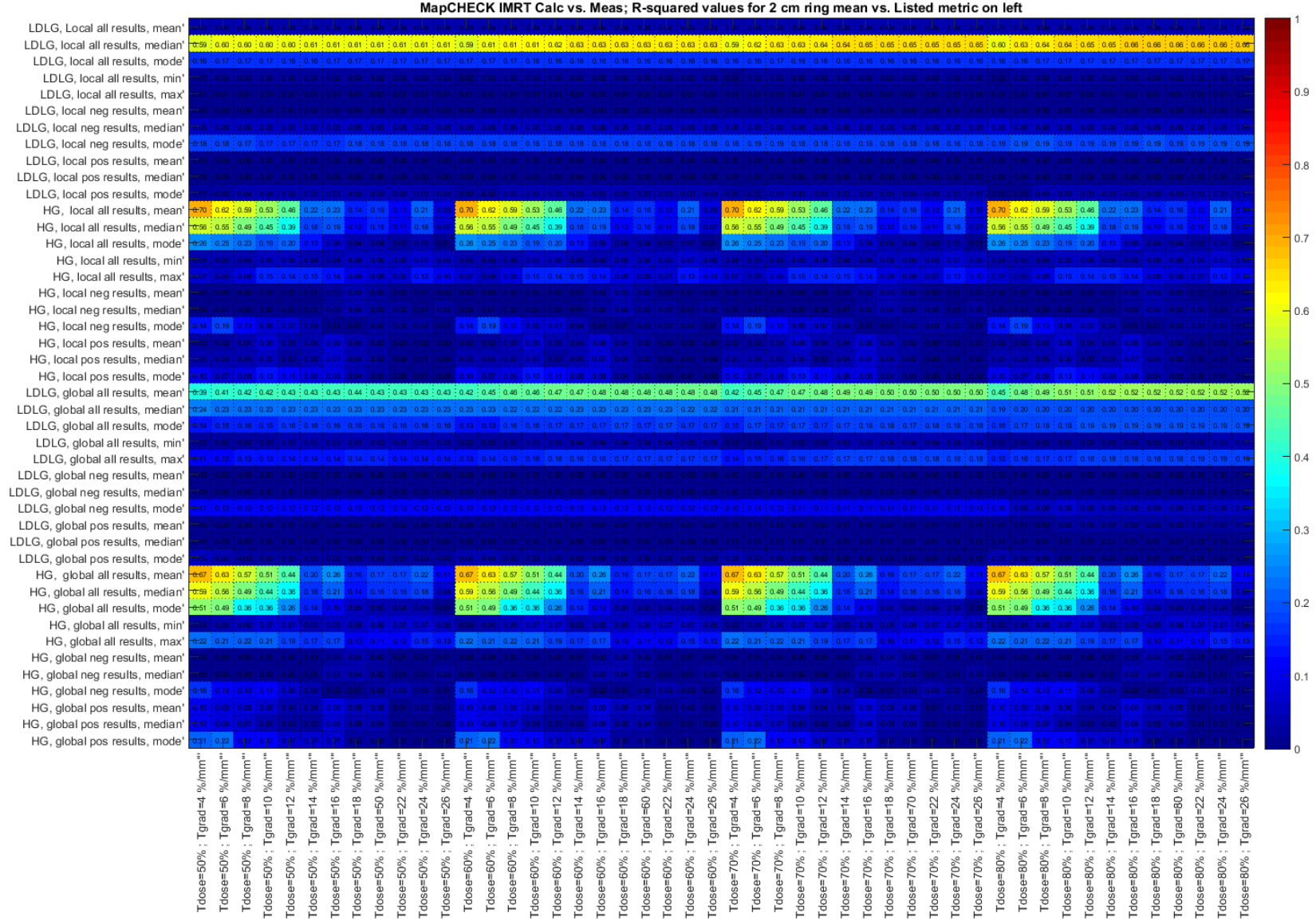

Figure A.5-34. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 2 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

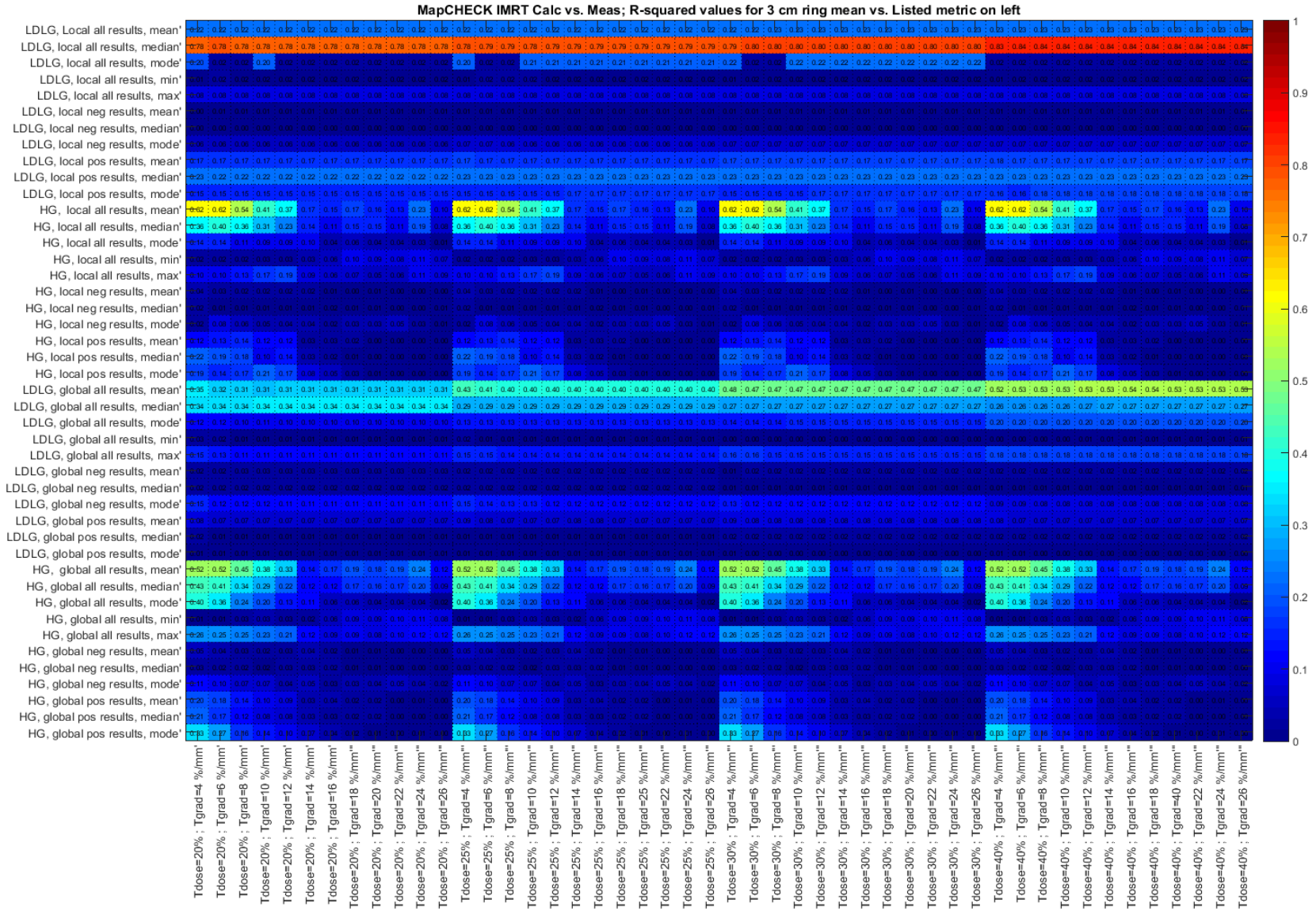

Figure A.5-35. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 3 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

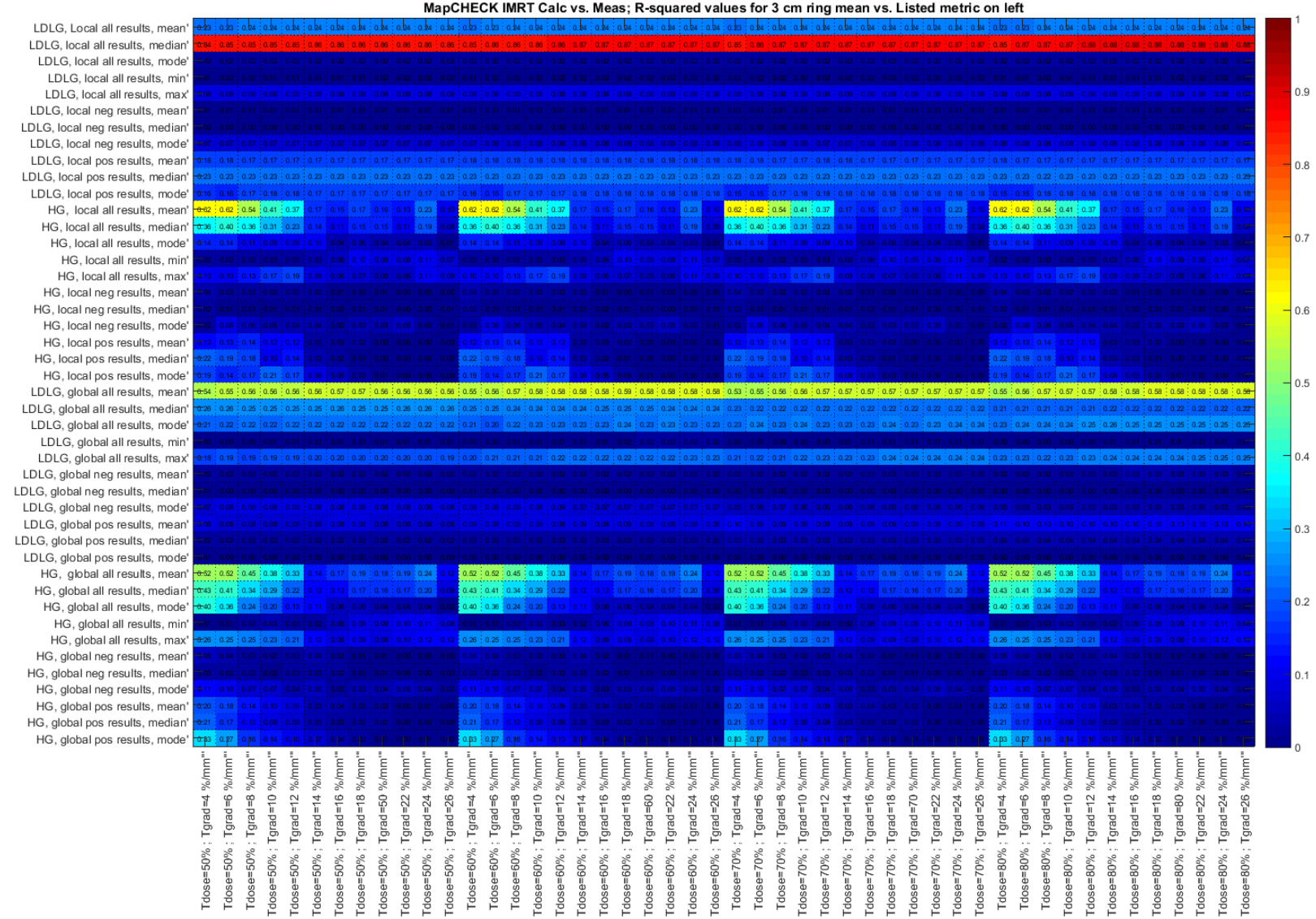

Figure A.5-36. Pearson r^2 heat map for IMRT cases on the MapCHECK for predicting the change in 3 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

A.V.7 Delta 4 VMAT PTV Correlations

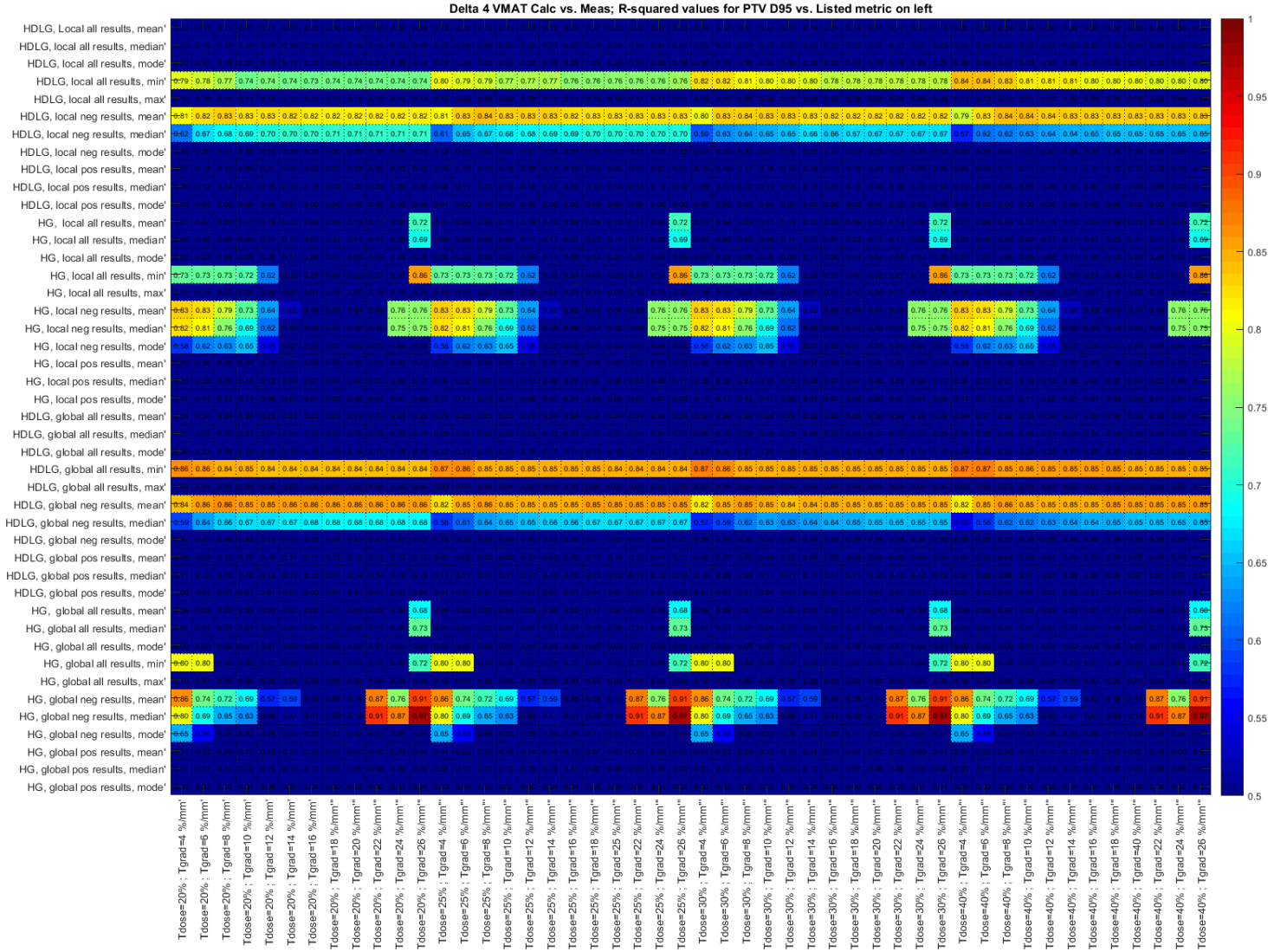

Figure A.5-37. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in PTV $D_{95\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

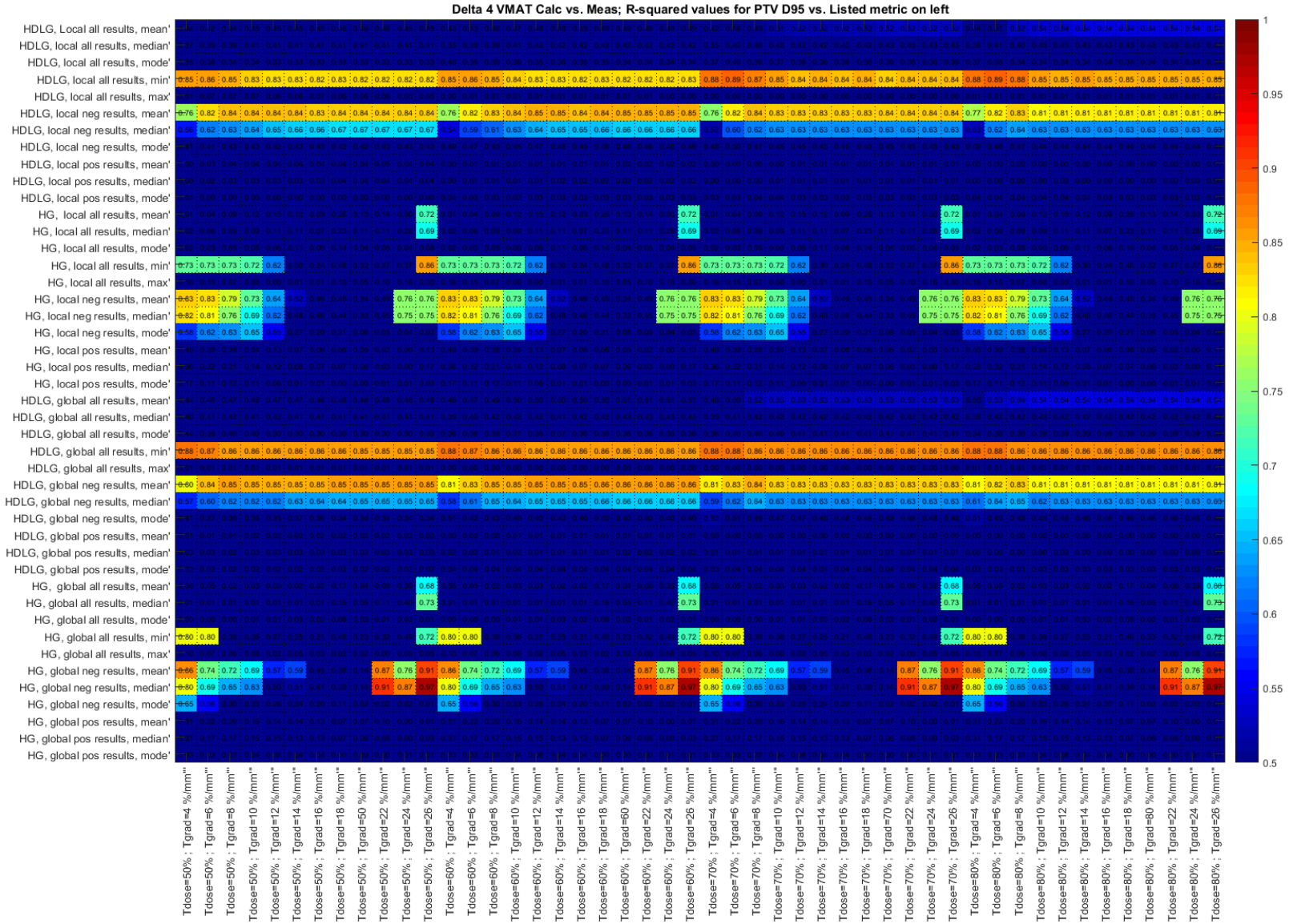

Figure A.5-38. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in PTV $D_{95\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

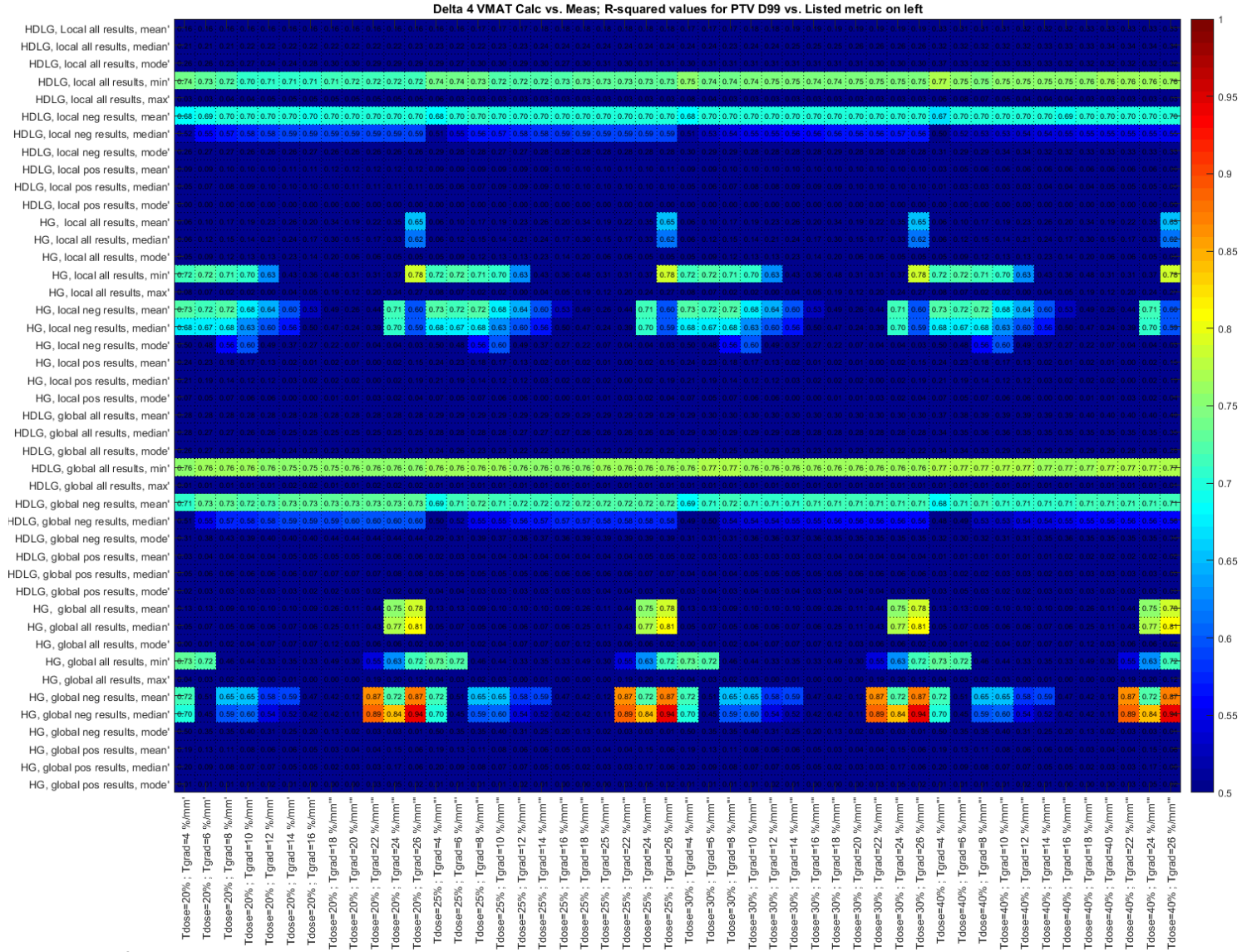

Figure A.5-39. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in PTV $D_{99\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

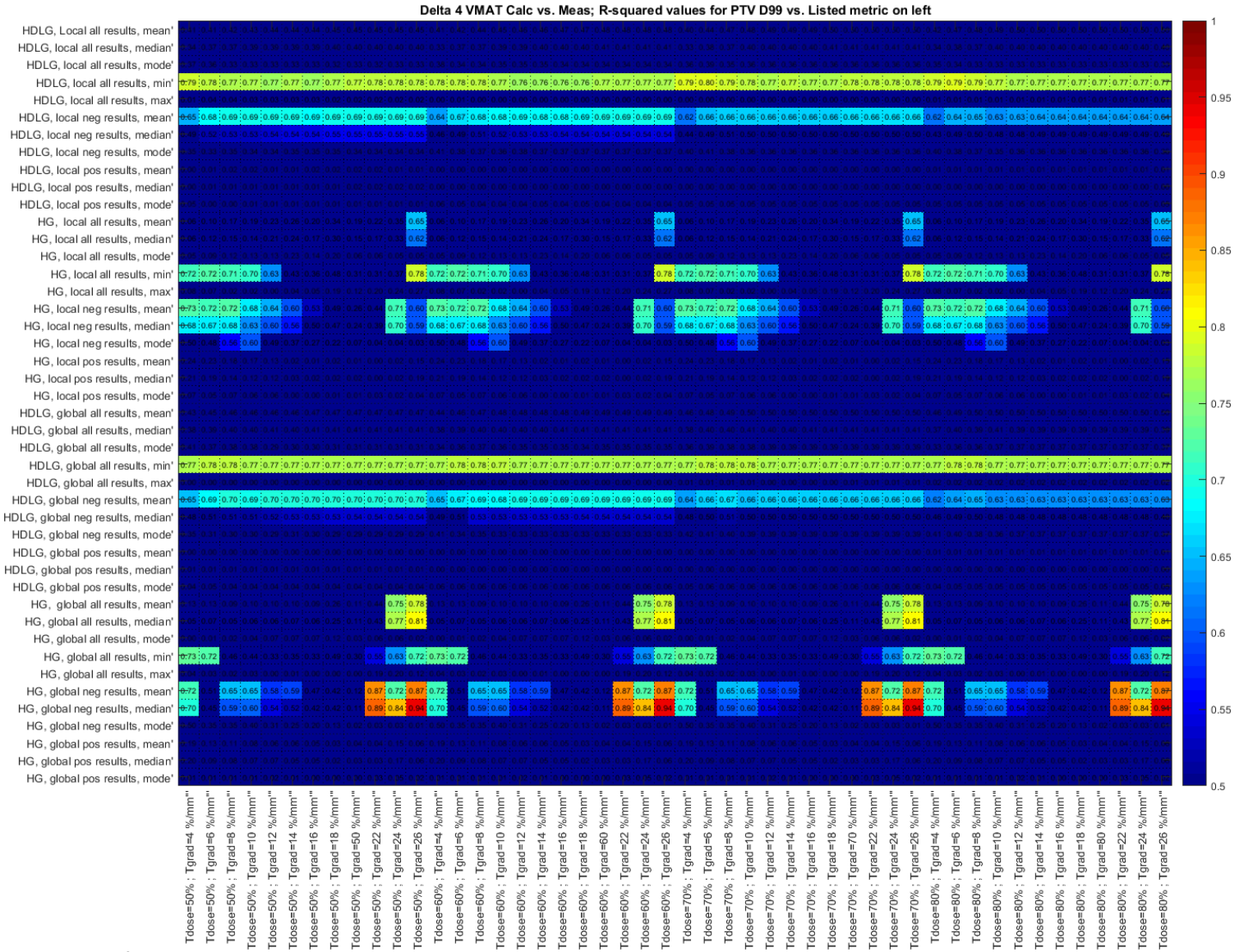

Figure A.5-40. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in PTV $D_{99\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

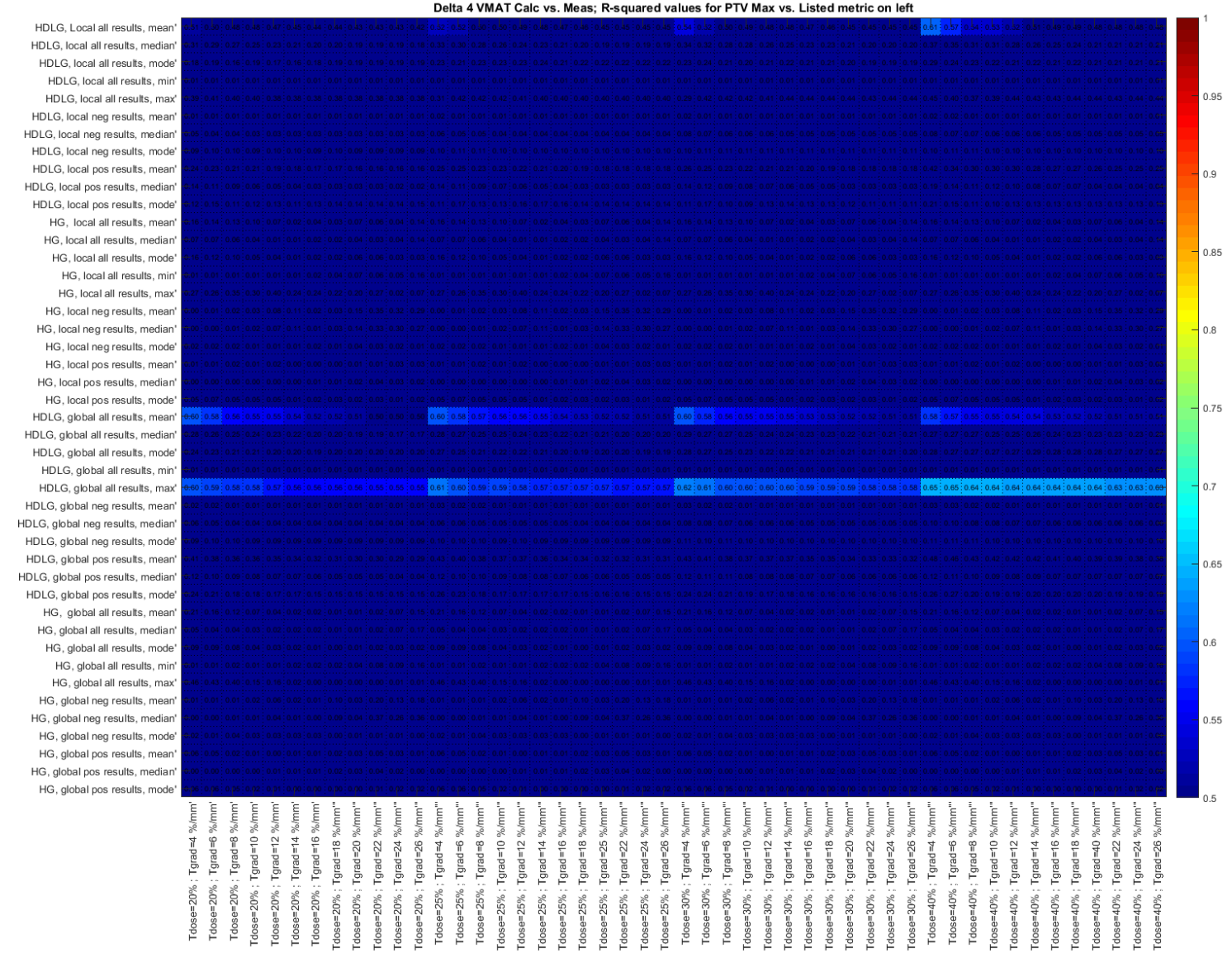

Figure A.5-41. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in PTV maximum dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

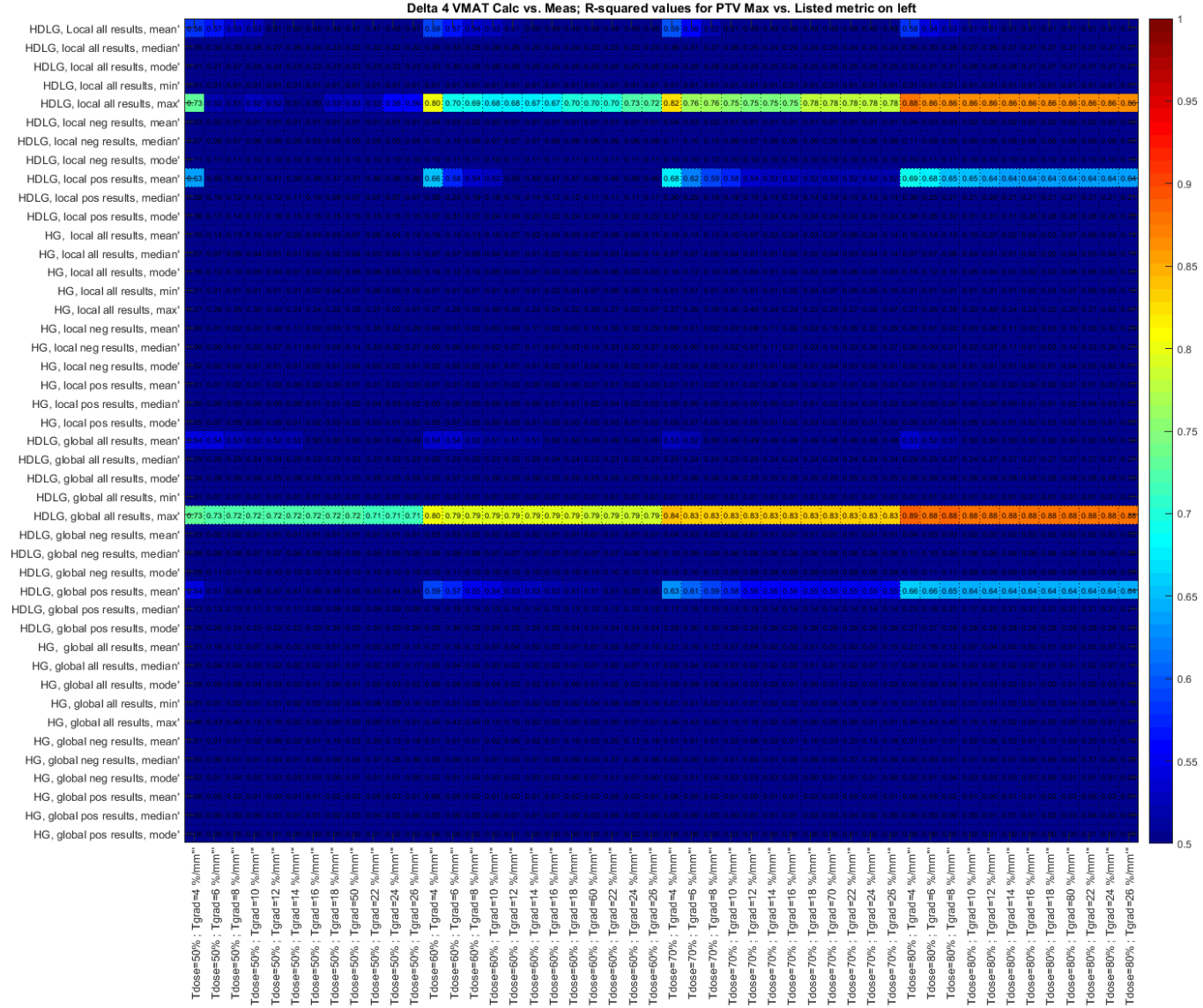

Figure A.5-42. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in PTV maximum dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

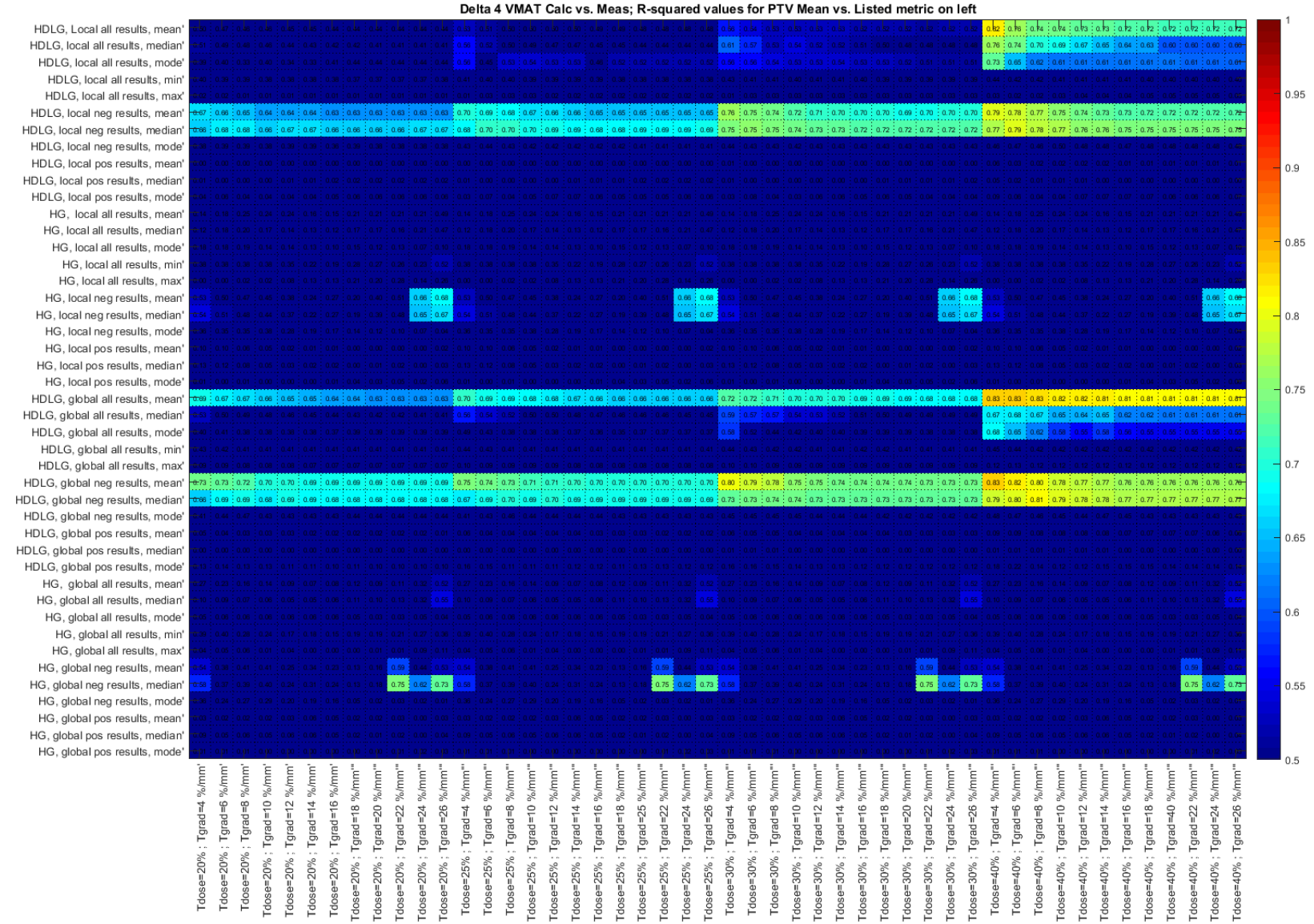

Figure A.5-43. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in PTV mean dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

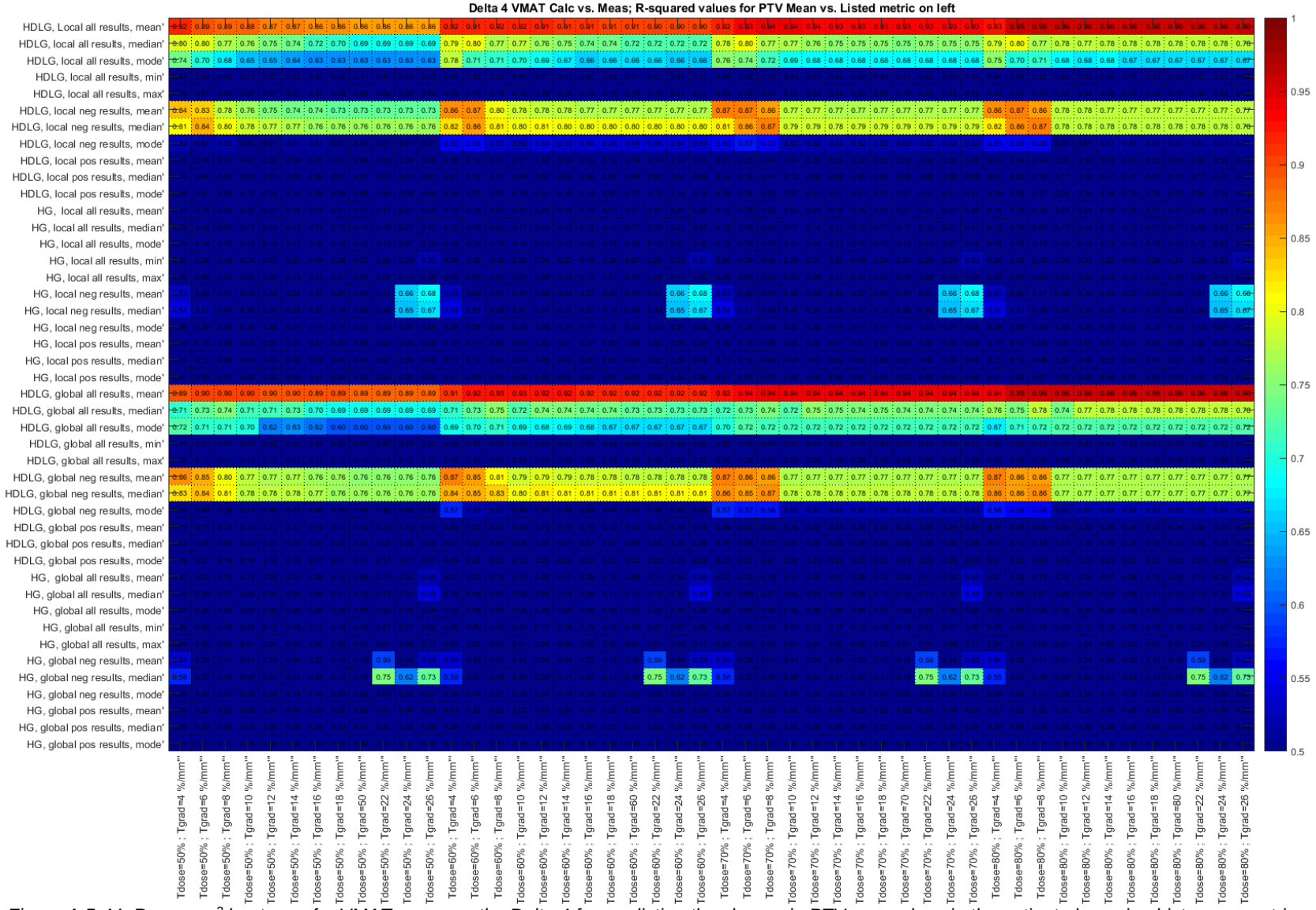

Figure A.5-44. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in PTV mean dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

A.V.8 Delta 4 VMAT Ring Correlations

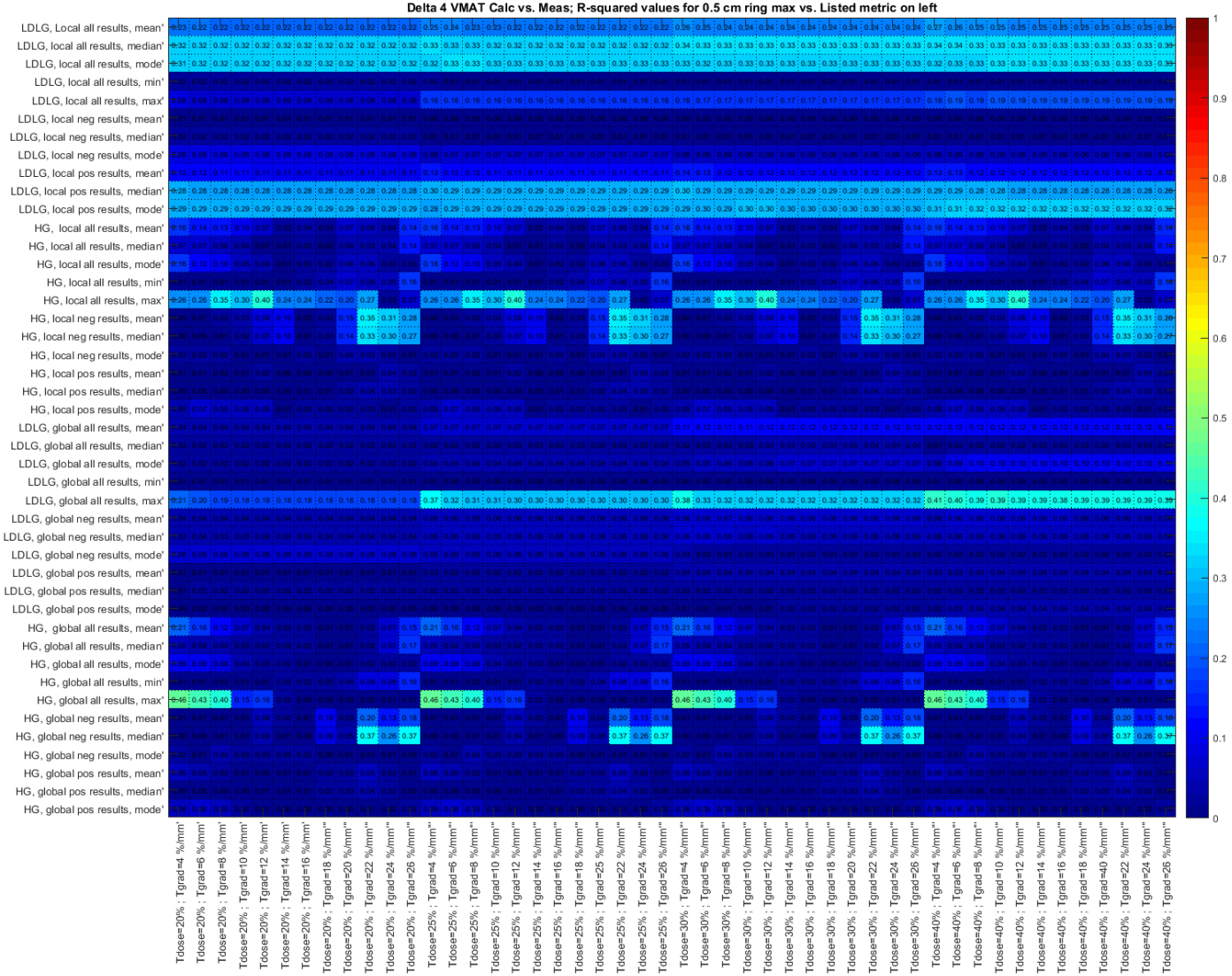

Figure A.5-45. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 0.5 cm ring maximum dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

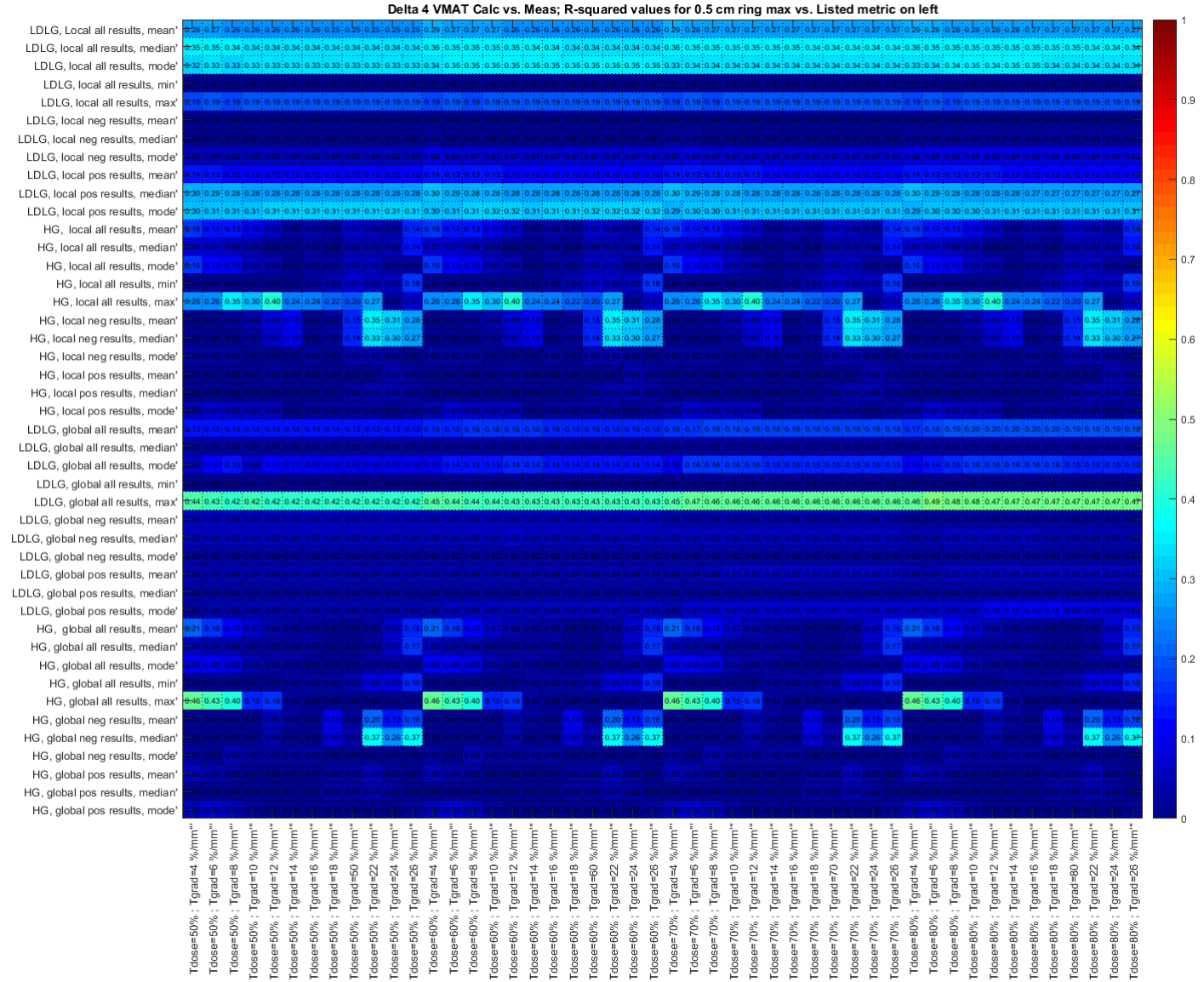

Figure A.5-46. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 0.5 cm ring maximum dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

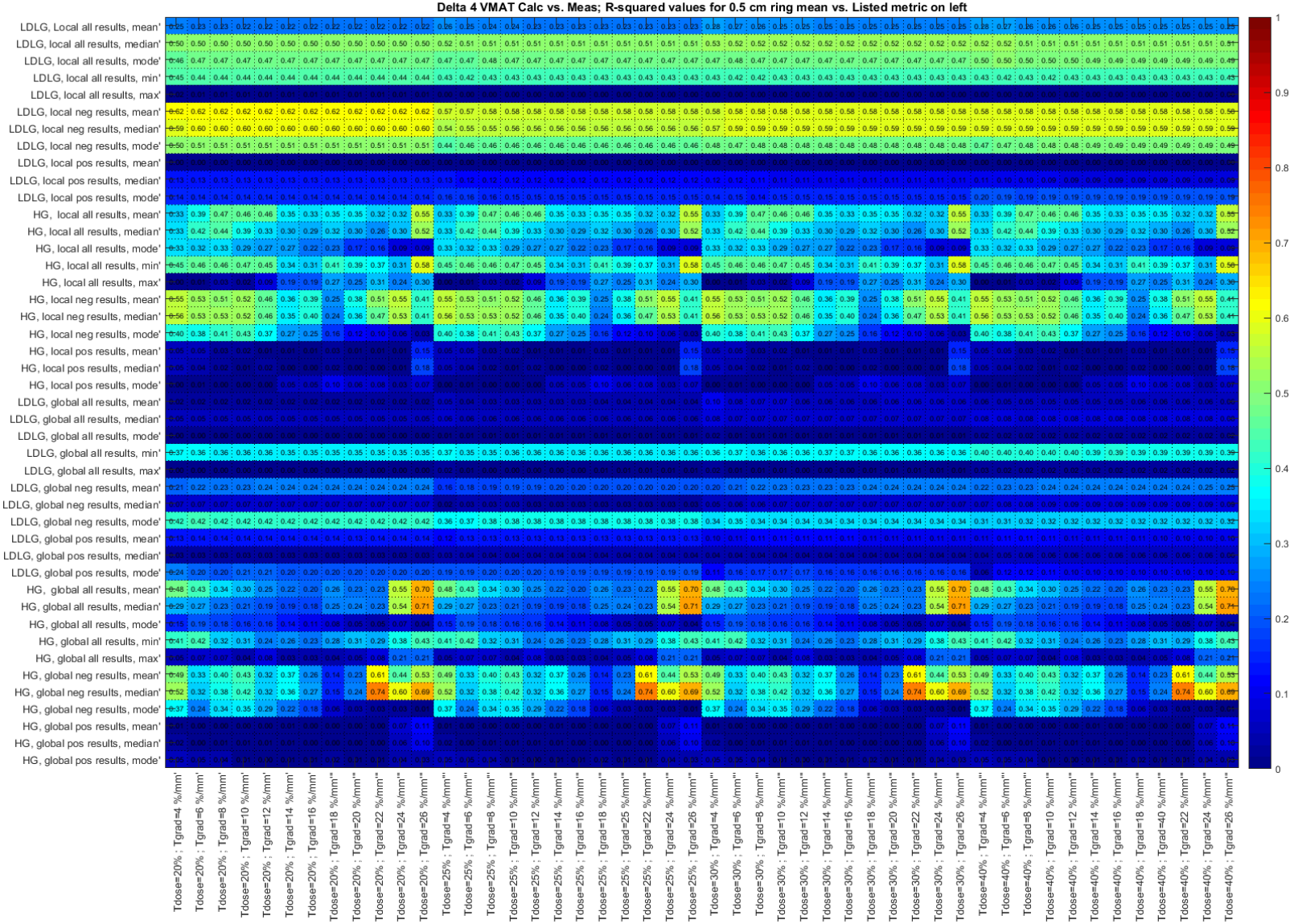

Figure A.5-47. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 0.5 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

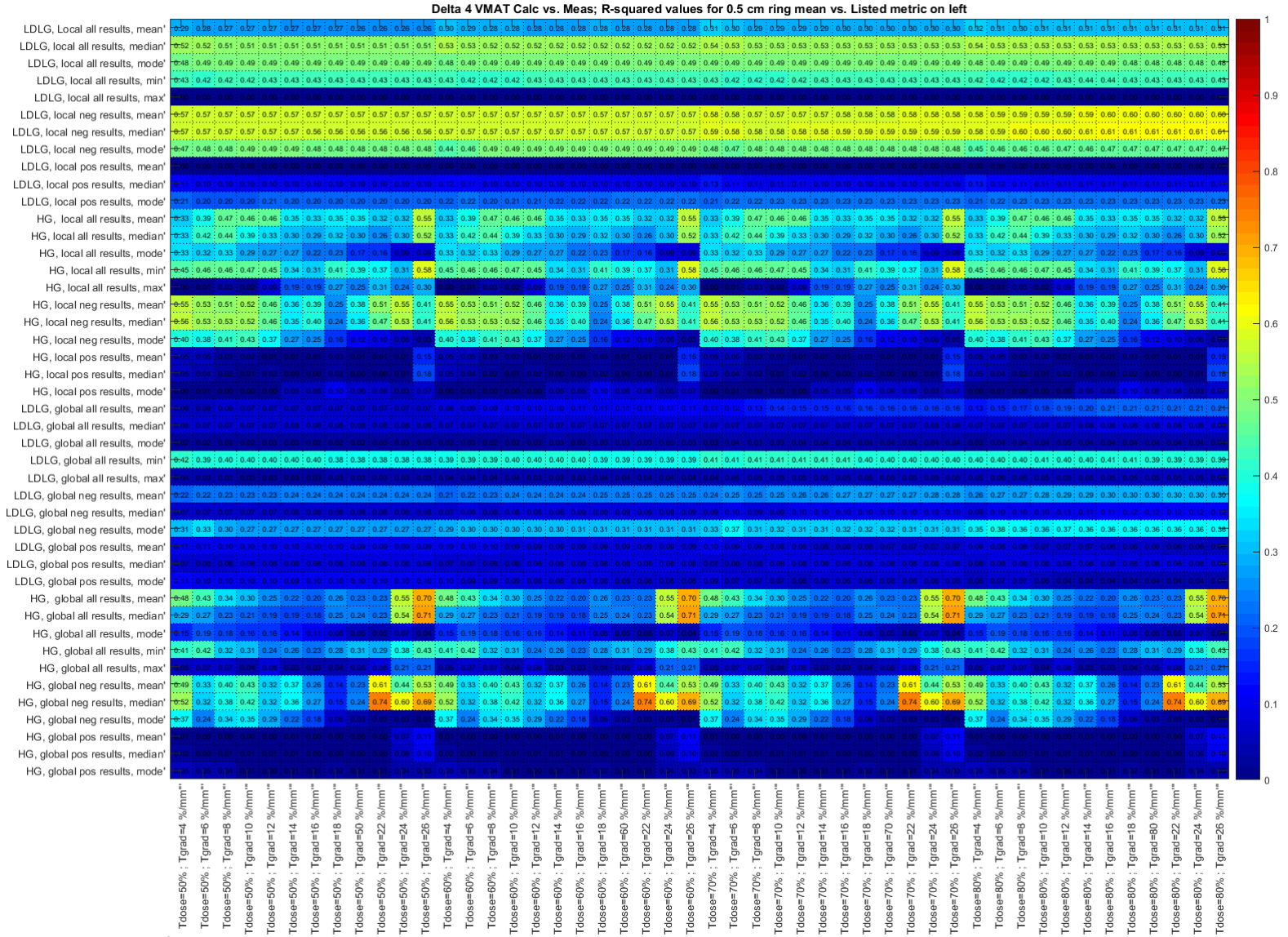

Figure A.5-48. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 0.5 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

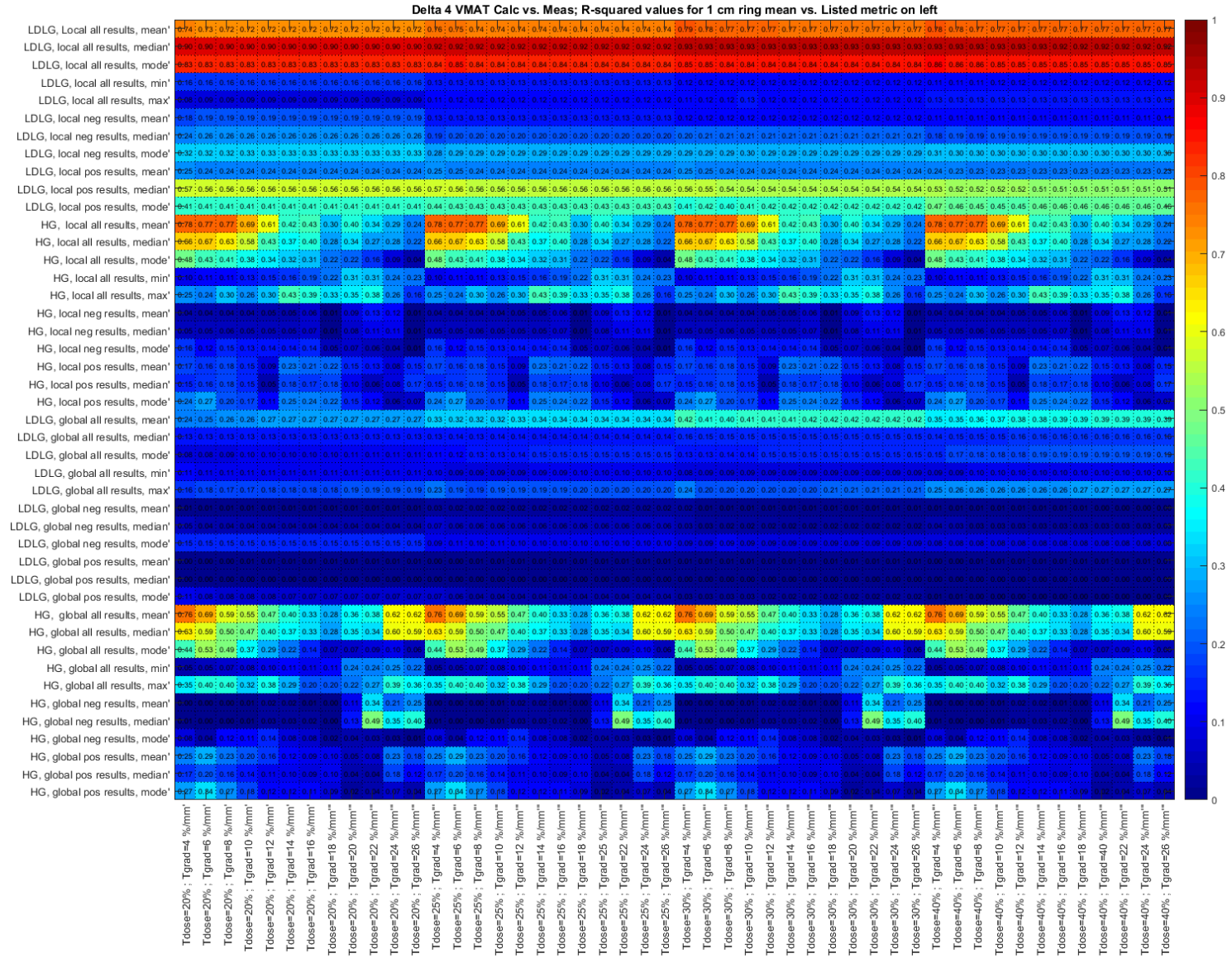

Figure A.5-49. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 1 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

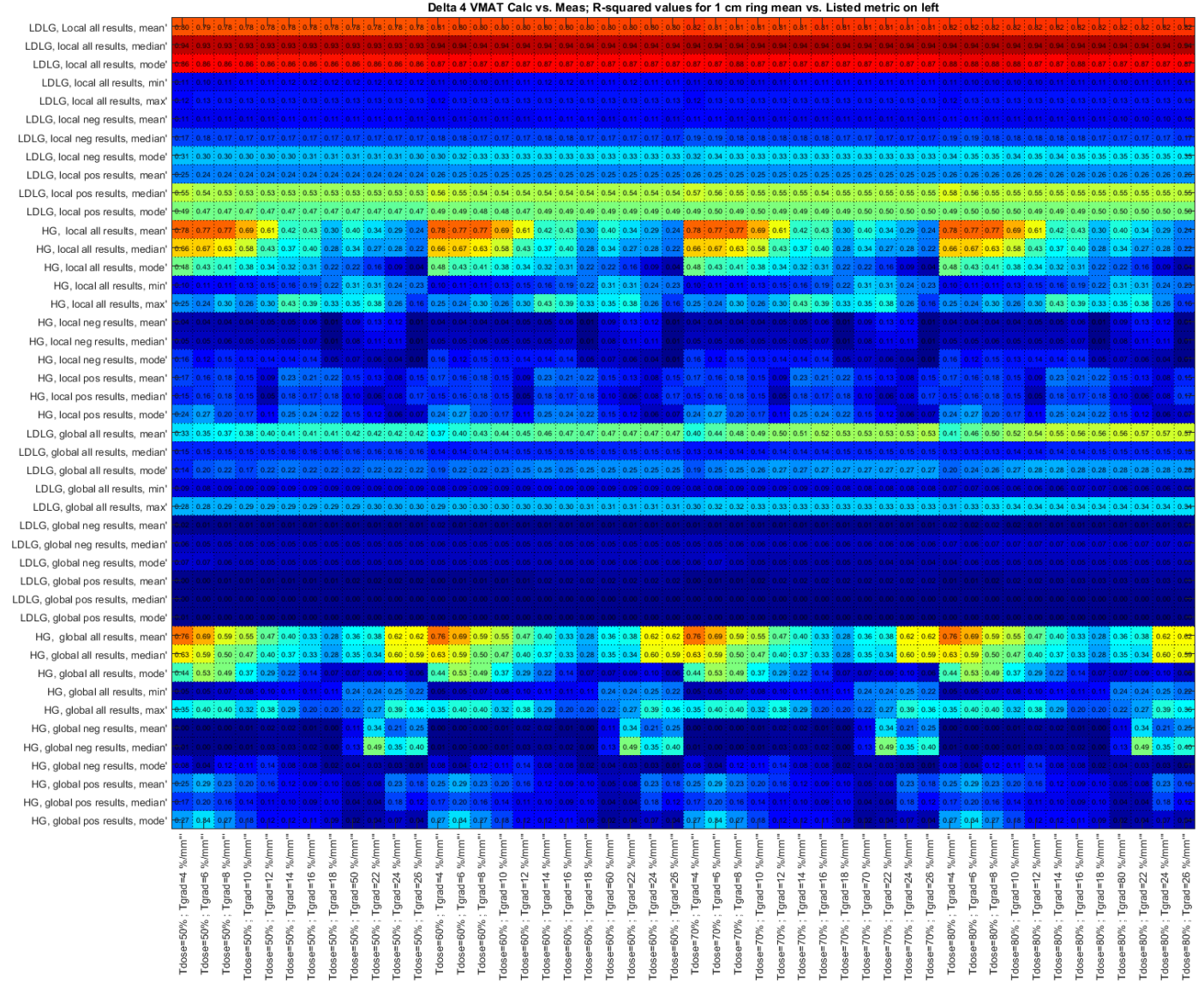

Figure A.5-50. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 1 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

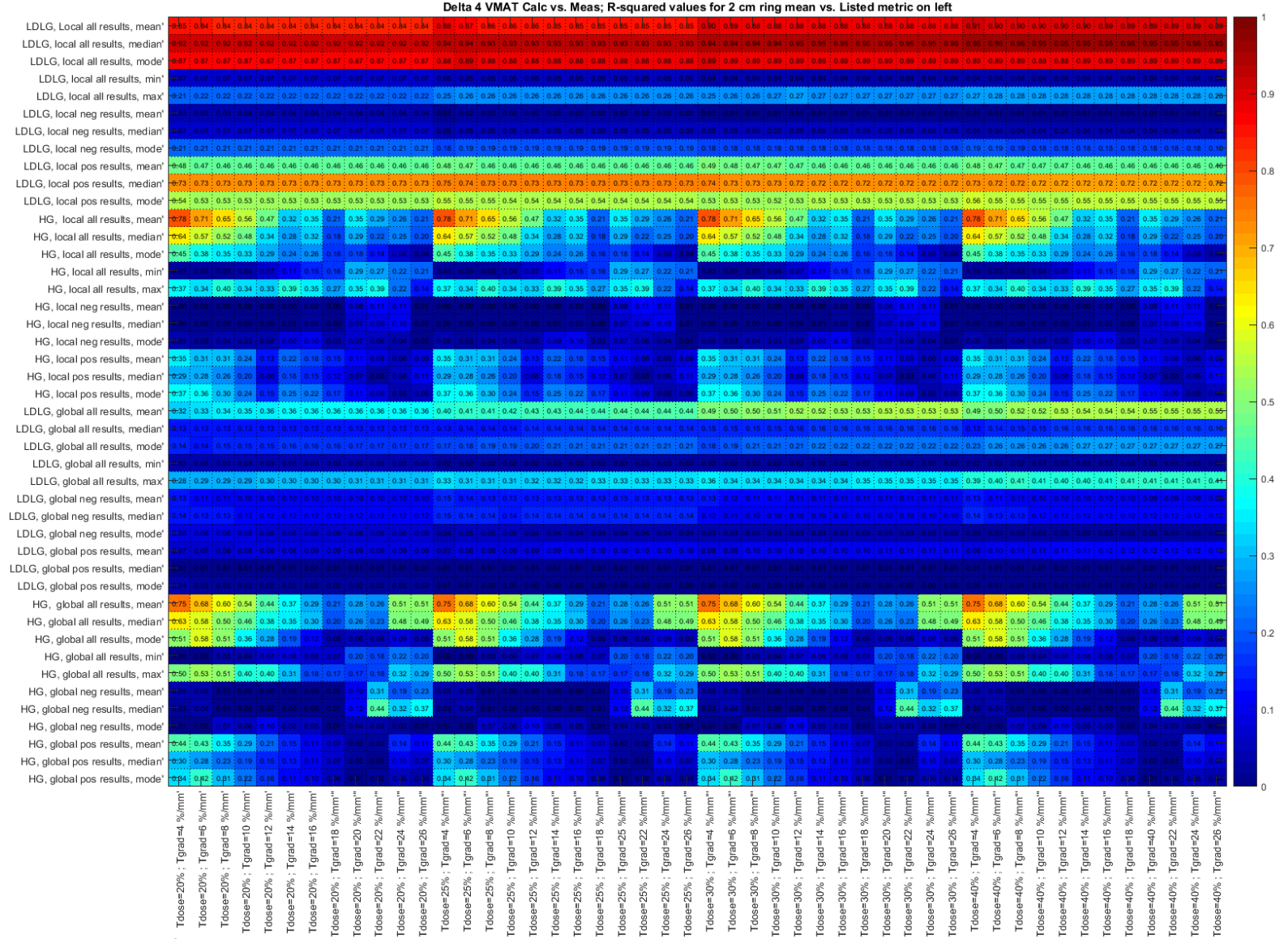

Figure A.5-51. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 2 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

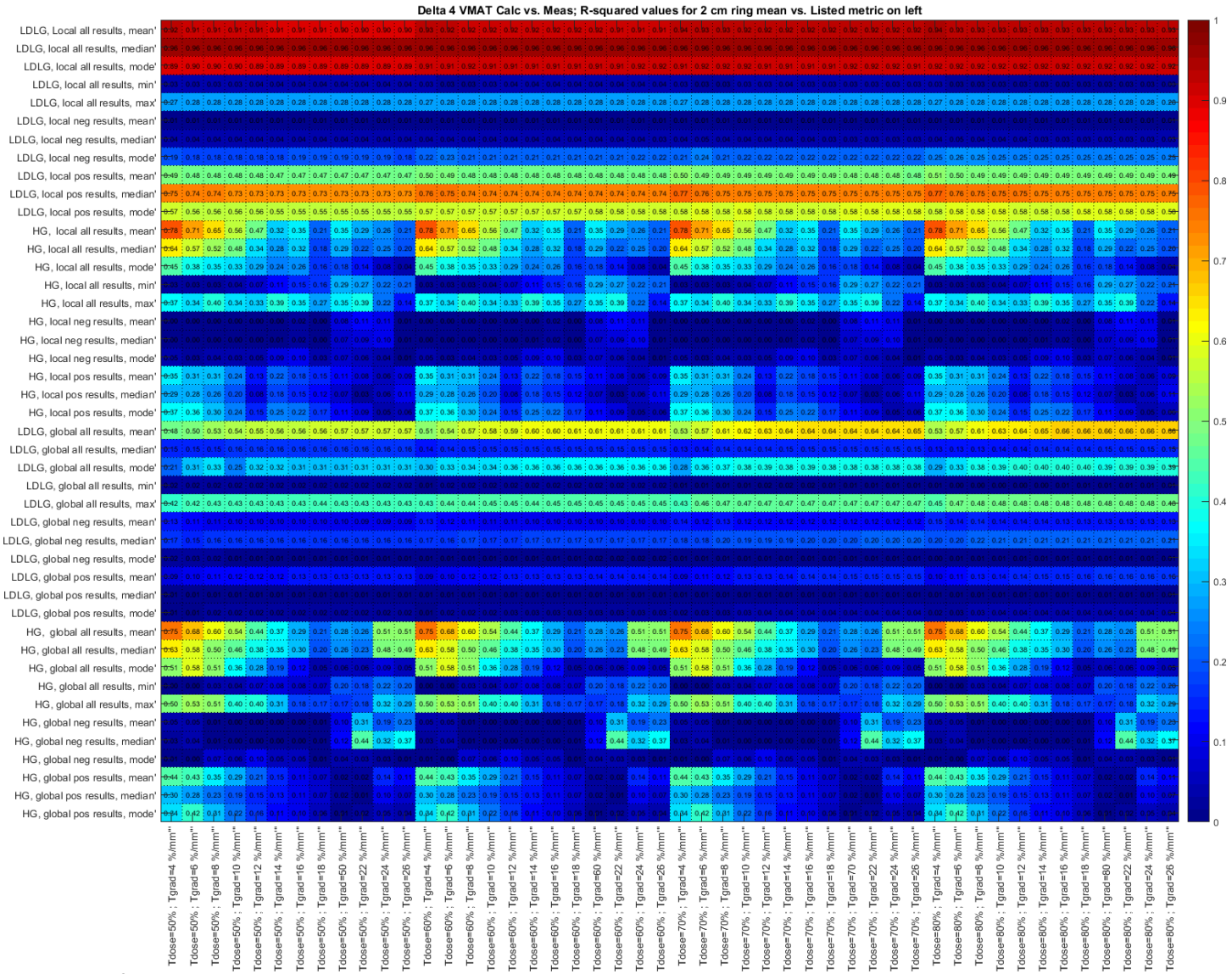

Figure A.5-52. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 2 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

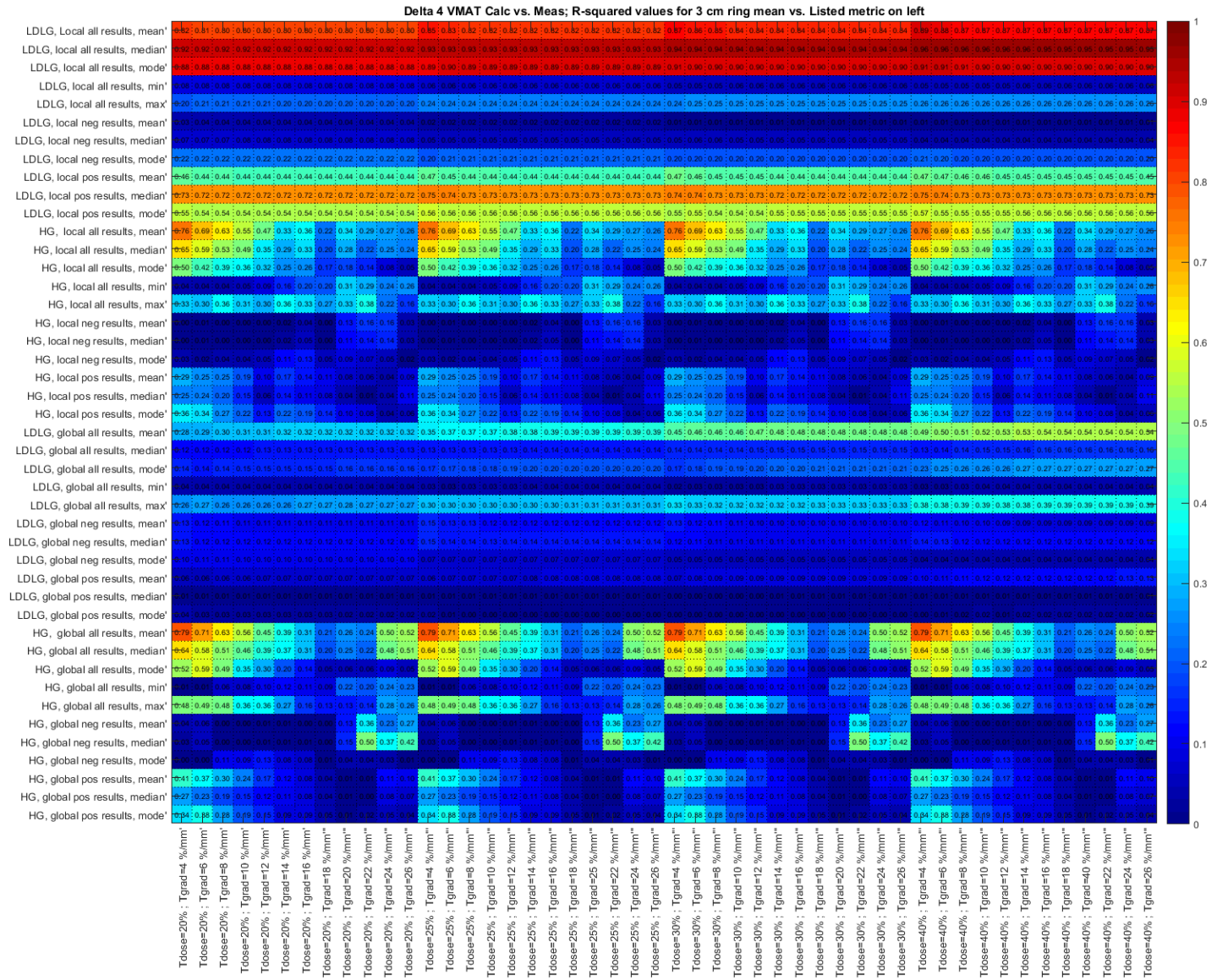

Figure A.5-53. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 3 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

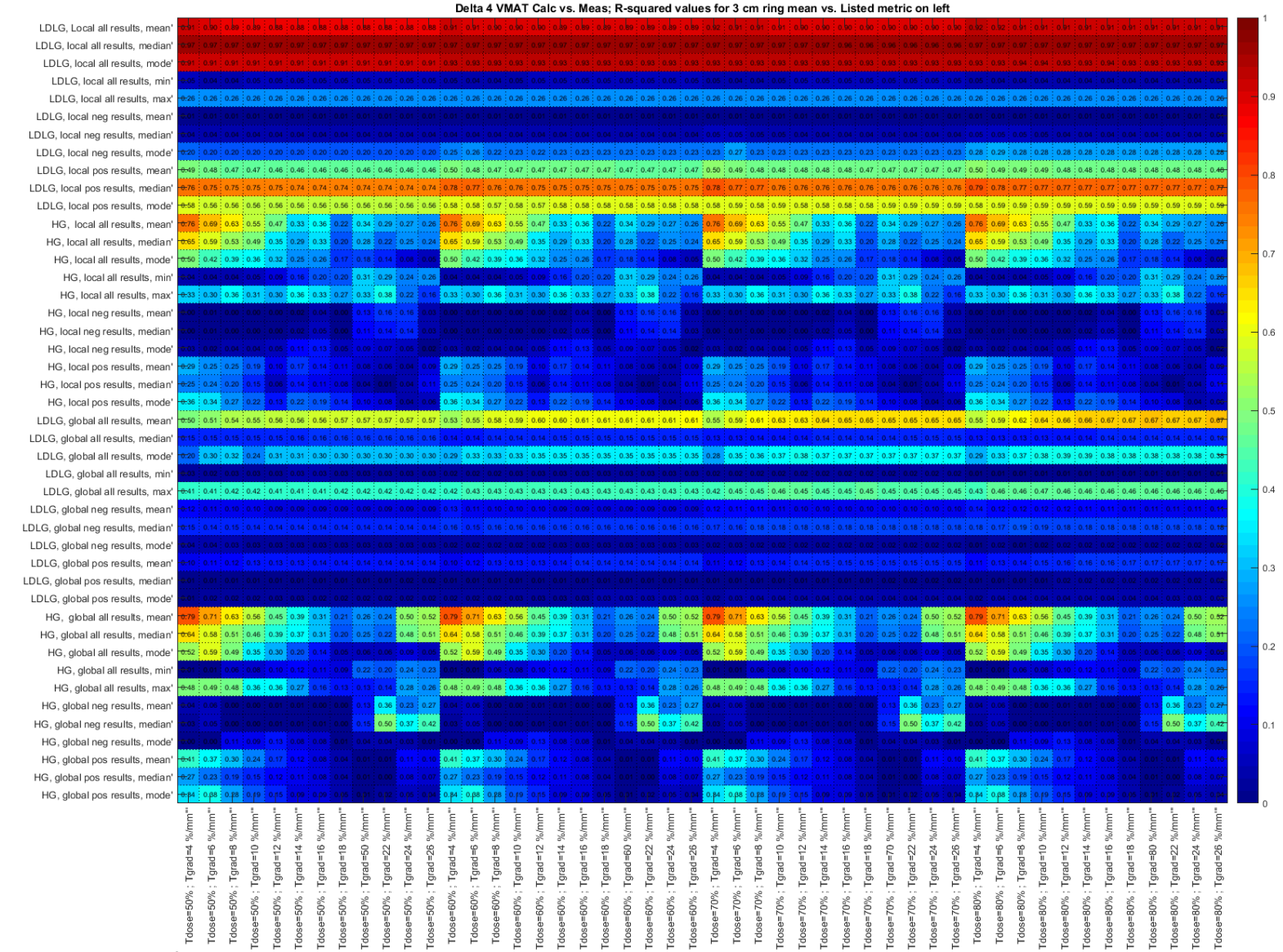

Figure A.5-54. Pearson r^2 heat map for VMAT cases on the Delta 4 for predicting the change in 3 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

A.V.9 ArcCHECK VMAT PTV Correlations

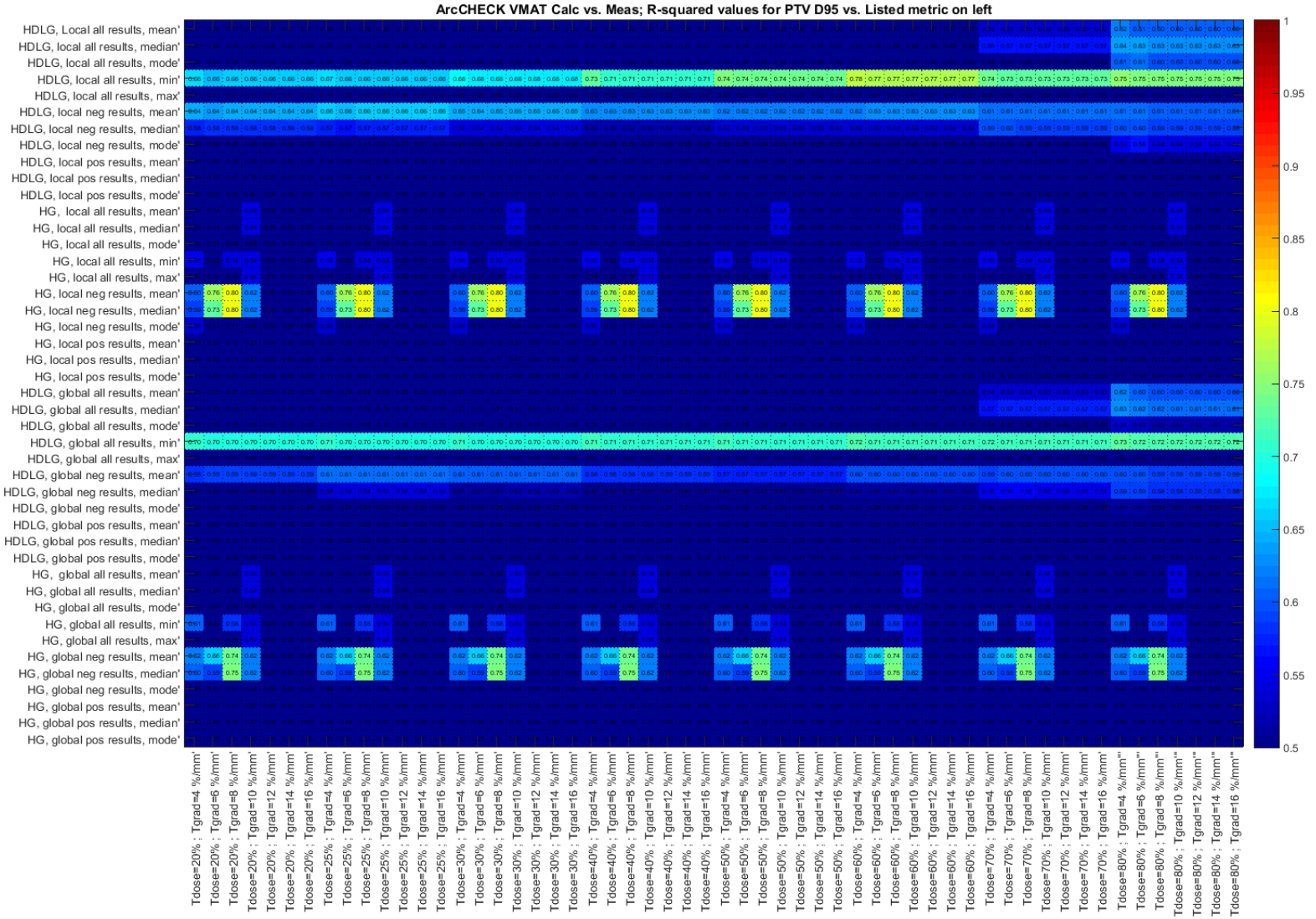

Figure A.5-55. Pearson r^2 heat map for VMAT cases on the ArcCHECK for predicting the change in PTV $D_{95\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

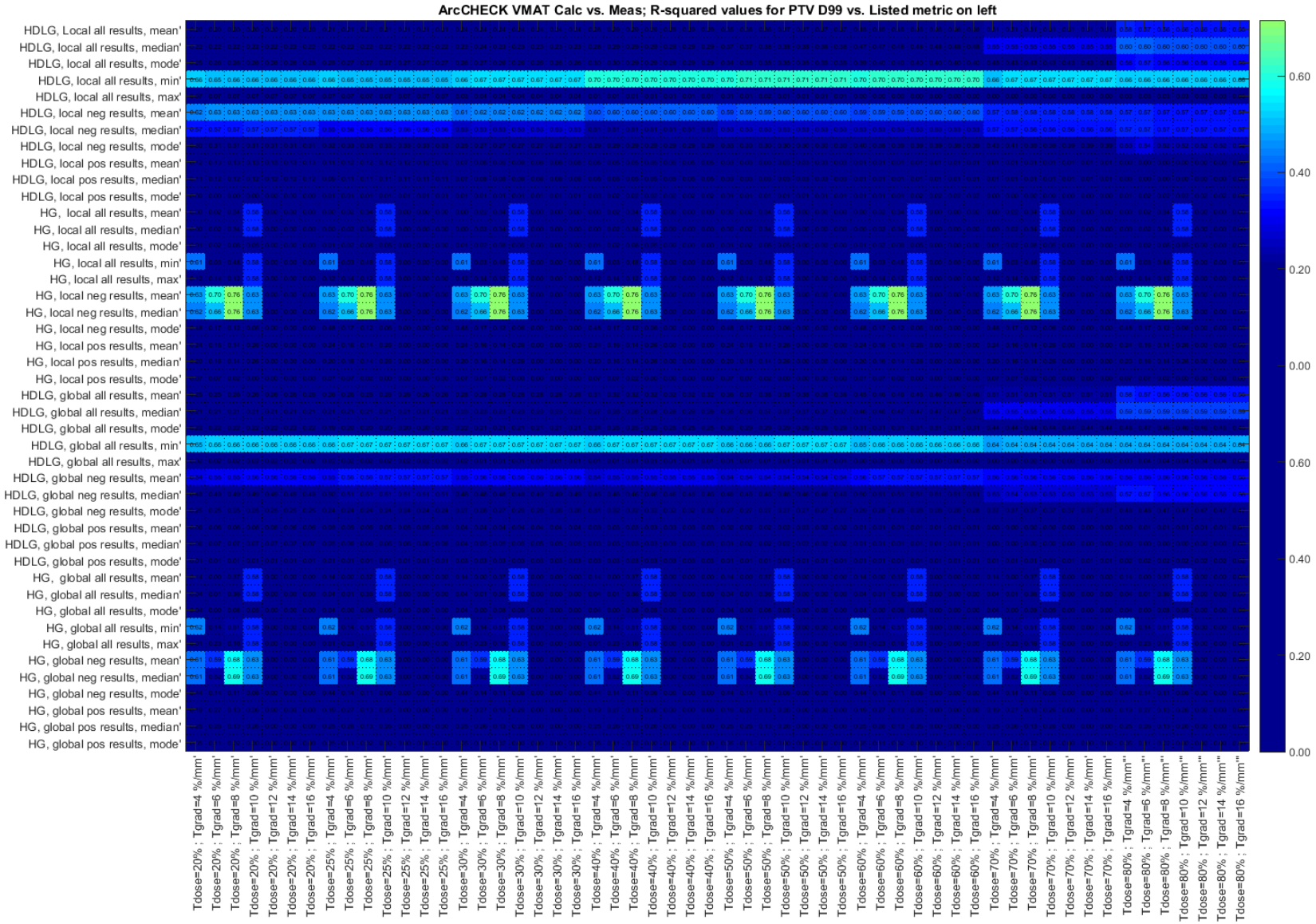

Figure A.5-56. Pearson r^2 heat map for VMAT cases on the ArcCHECK for predicting the change in PTV $D_{99\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

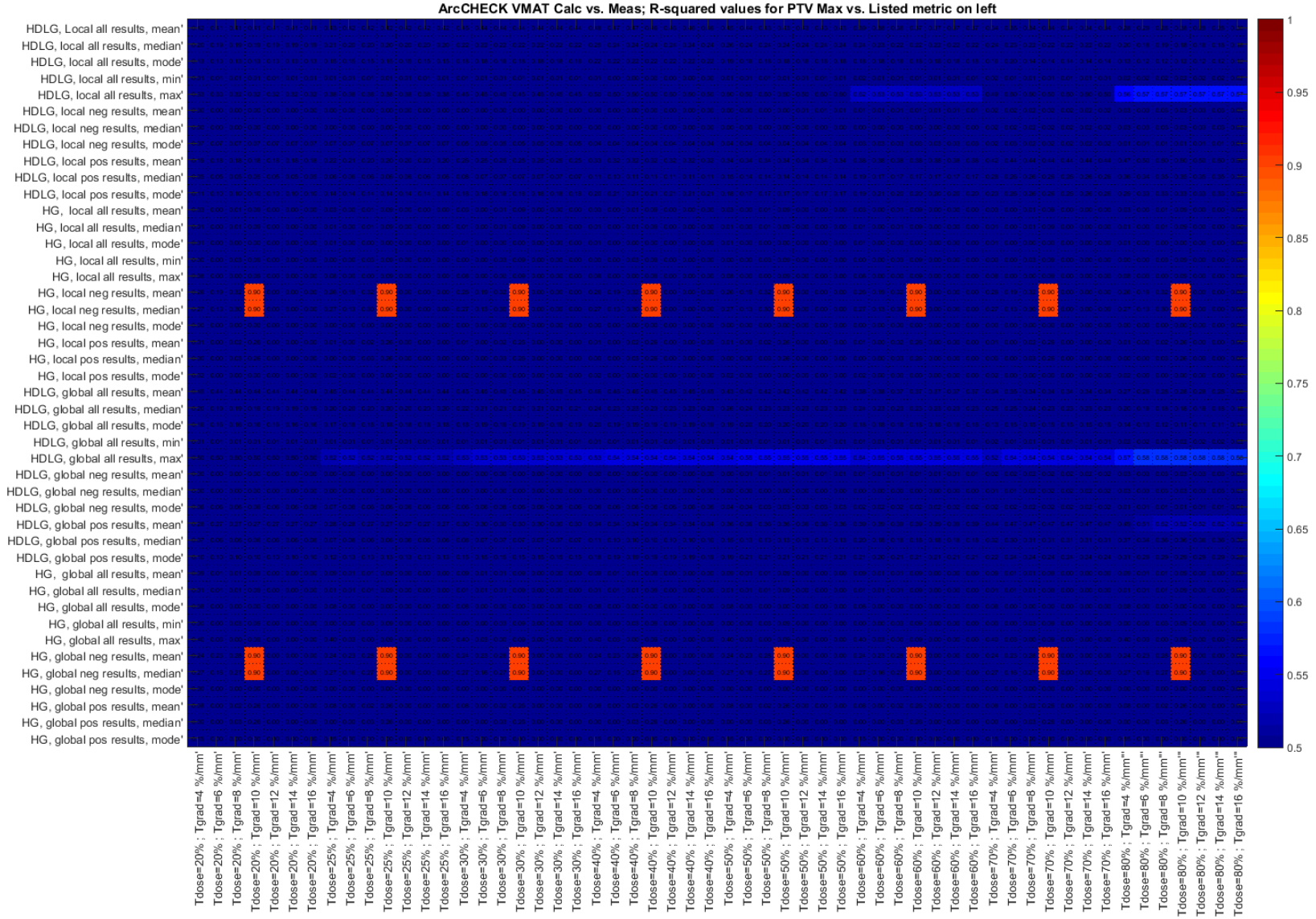

Figure A.5-57. Pearson r^2 heat map for VMAT cases on the ArcCHECK for predicting the change in PTV maximum dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

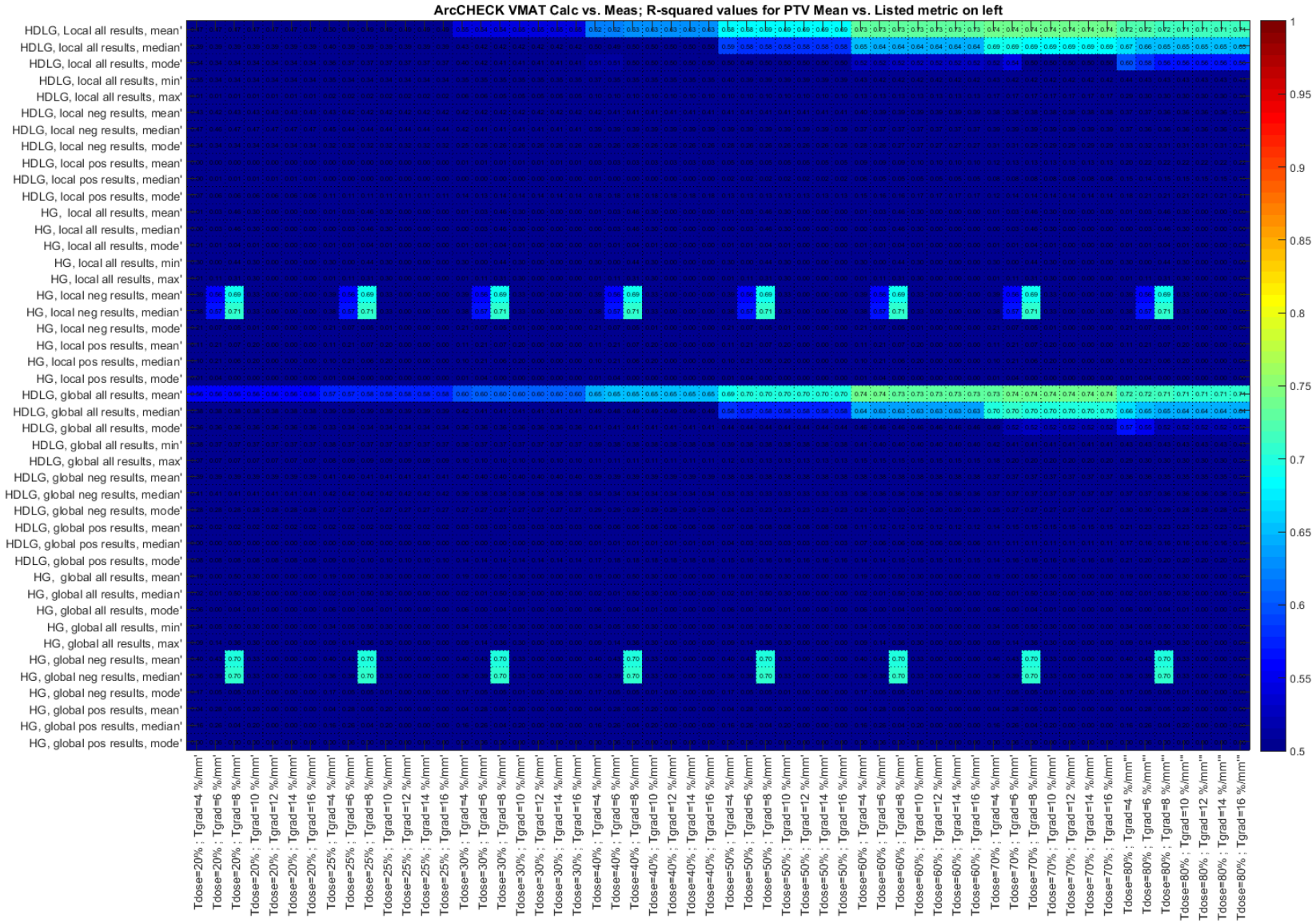

Figure A.5-58. Pearson r^2 heat map for VMAT cases on the ArcCHECK for predicting the change in PTV mean dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

A.V.10 ArcCHECK VMAT Ring Correlations

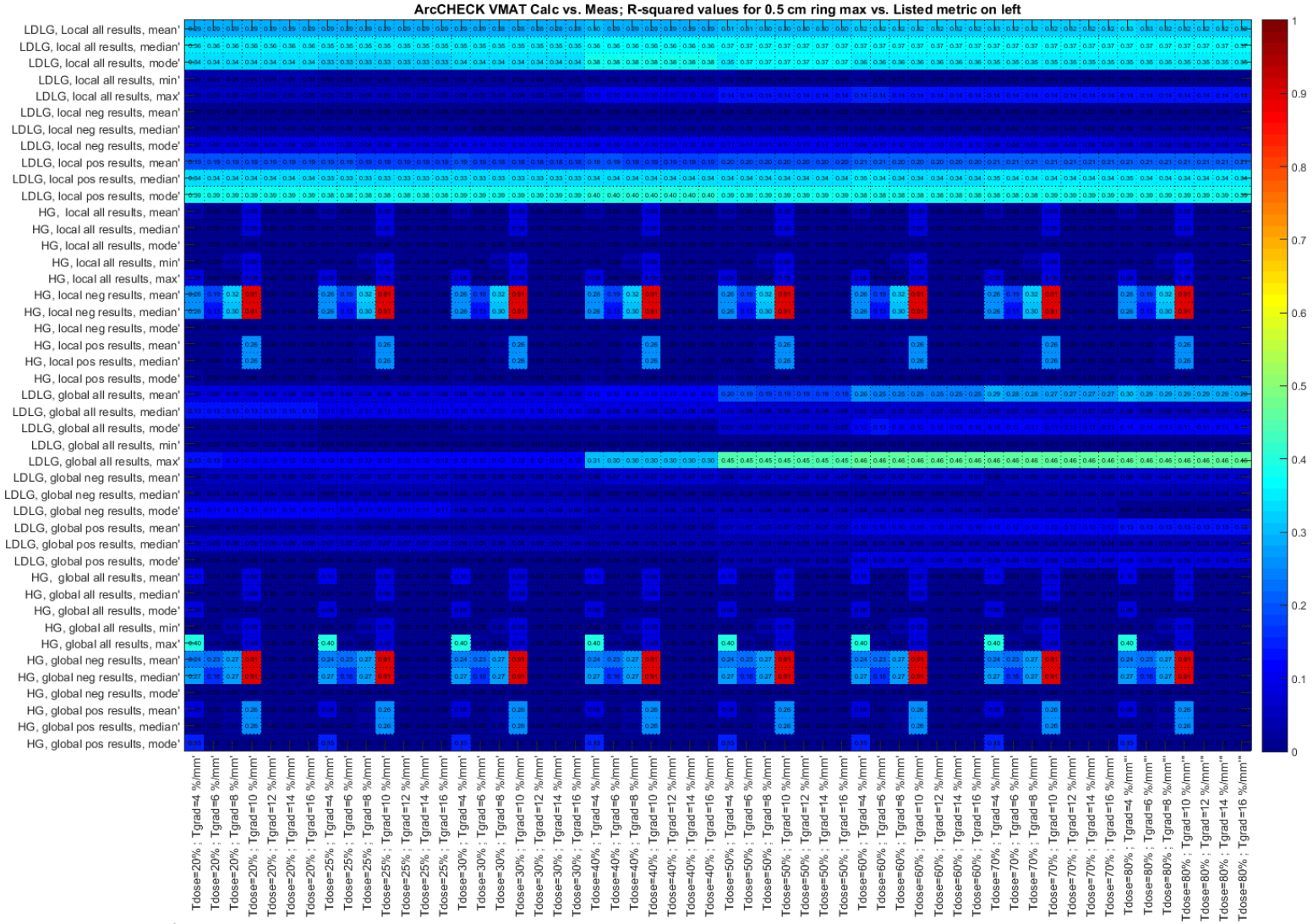

Figure A.5-59. Pearson r^2 heat map for VMAT cases on the ArcCHECK for predicting the change 0.5 cm ring maximum dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

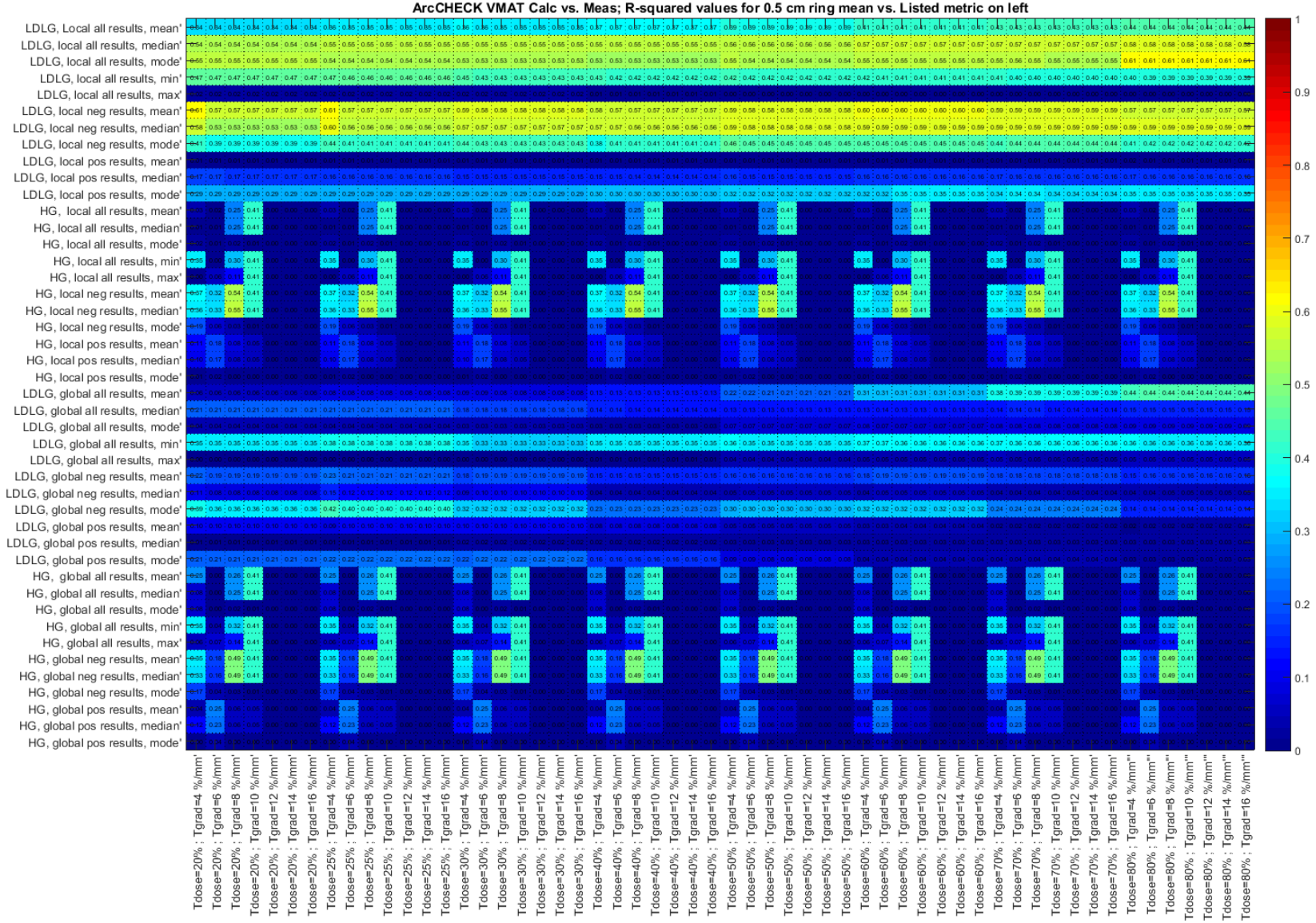

Figure A.5-60. Pearson r^2 heat map for VMAT cases on the ArcCHECK for predicting the change 0.5 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

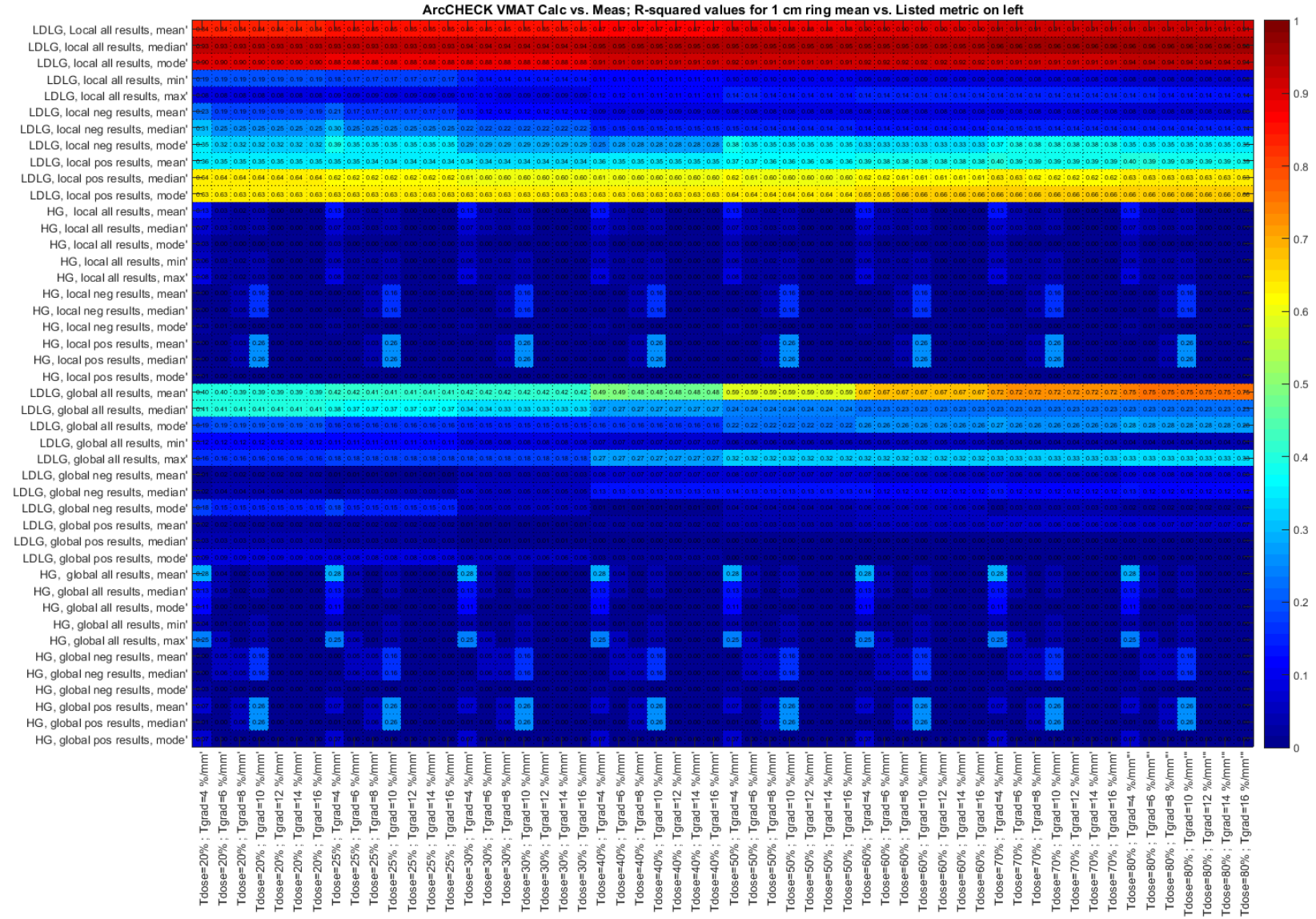

Figure A.5-61. Pearson r^2 heat map for VMAT cases on the ArcCHECK for predicting the change 1 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

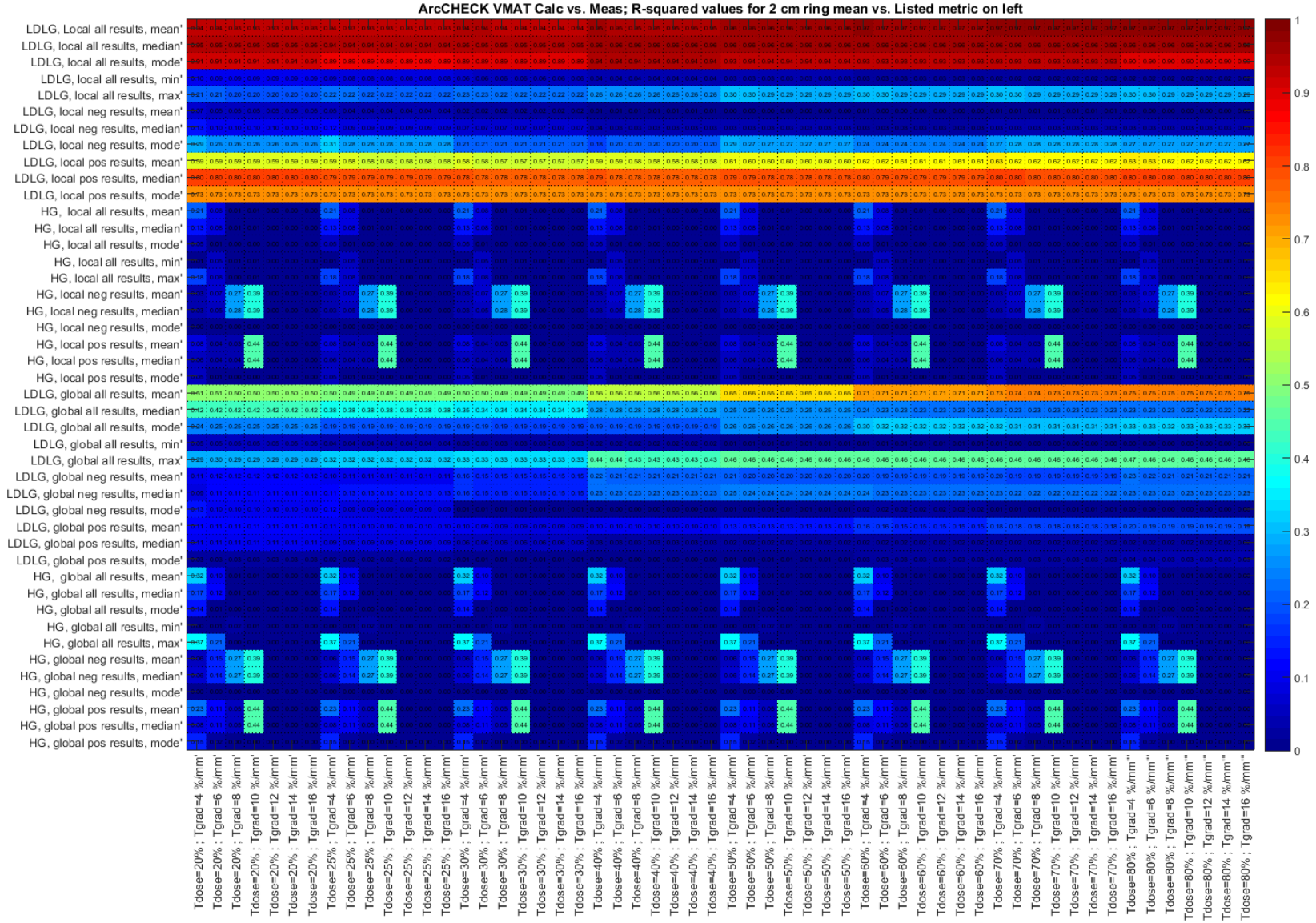

Figure A.5-62. Pearson r^2 heat map for VMAT cases on the ArcCHECK for predicting the change 2 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

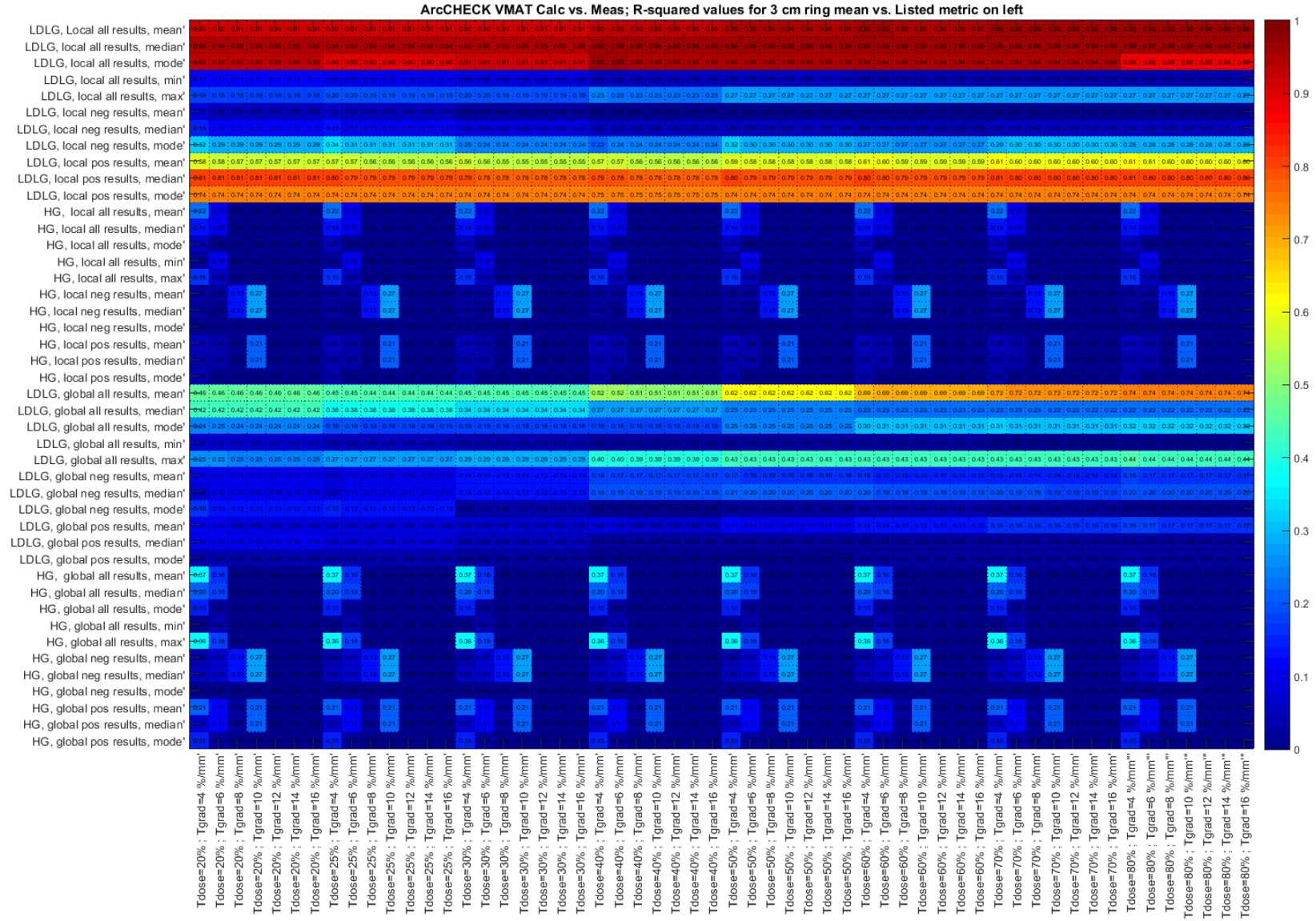

Figure A.5-63. Pearson r^2 heat map for VMAT cases on the ArcCHECK for predicting the change 3 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. The many different dose and gradient thresholds used for segmentation are shown on the x-axis.

A.V.11 MapCHECK VMAT PTV Correlations

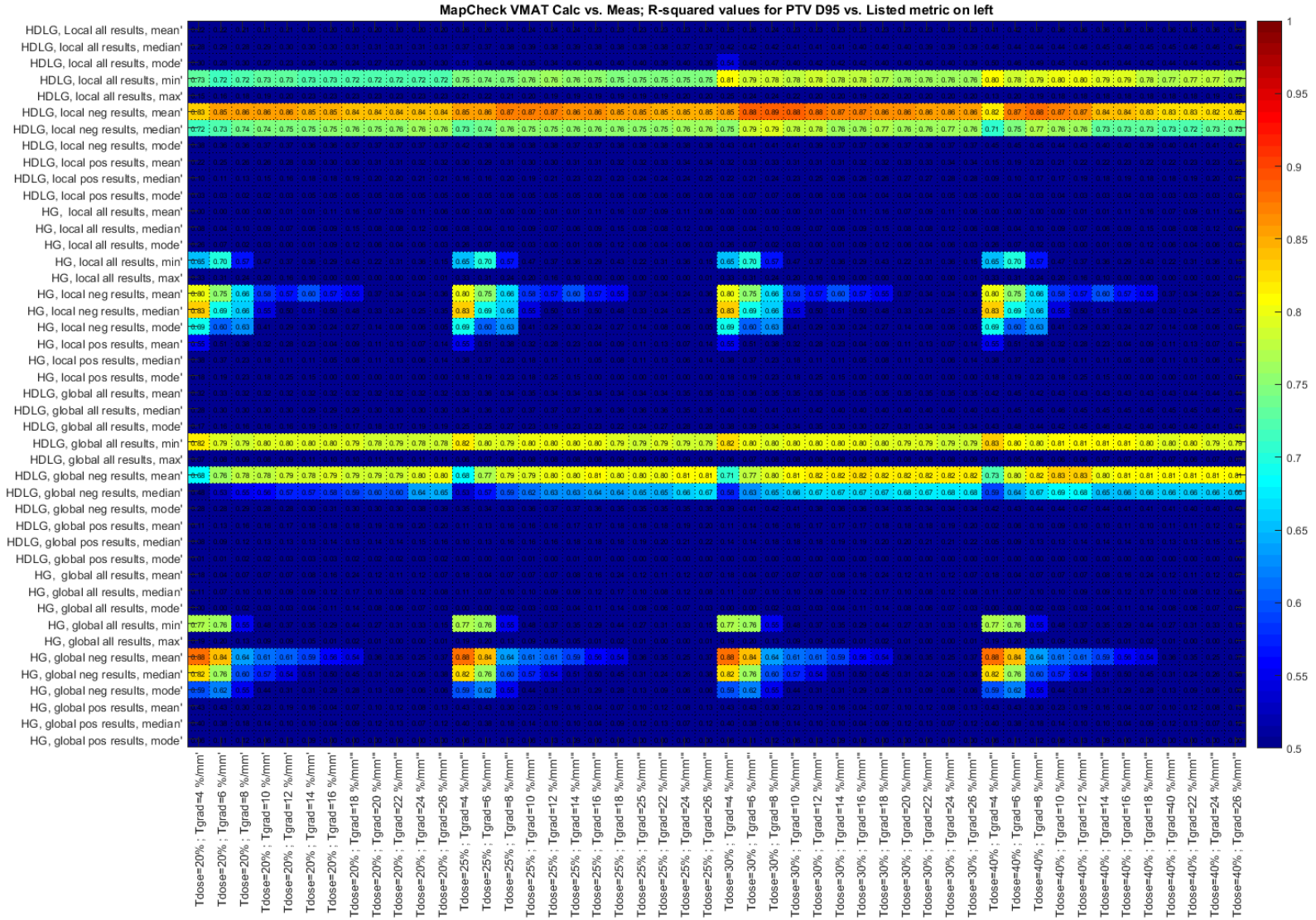

Figure A.5-64. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in PTV $D_{95\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

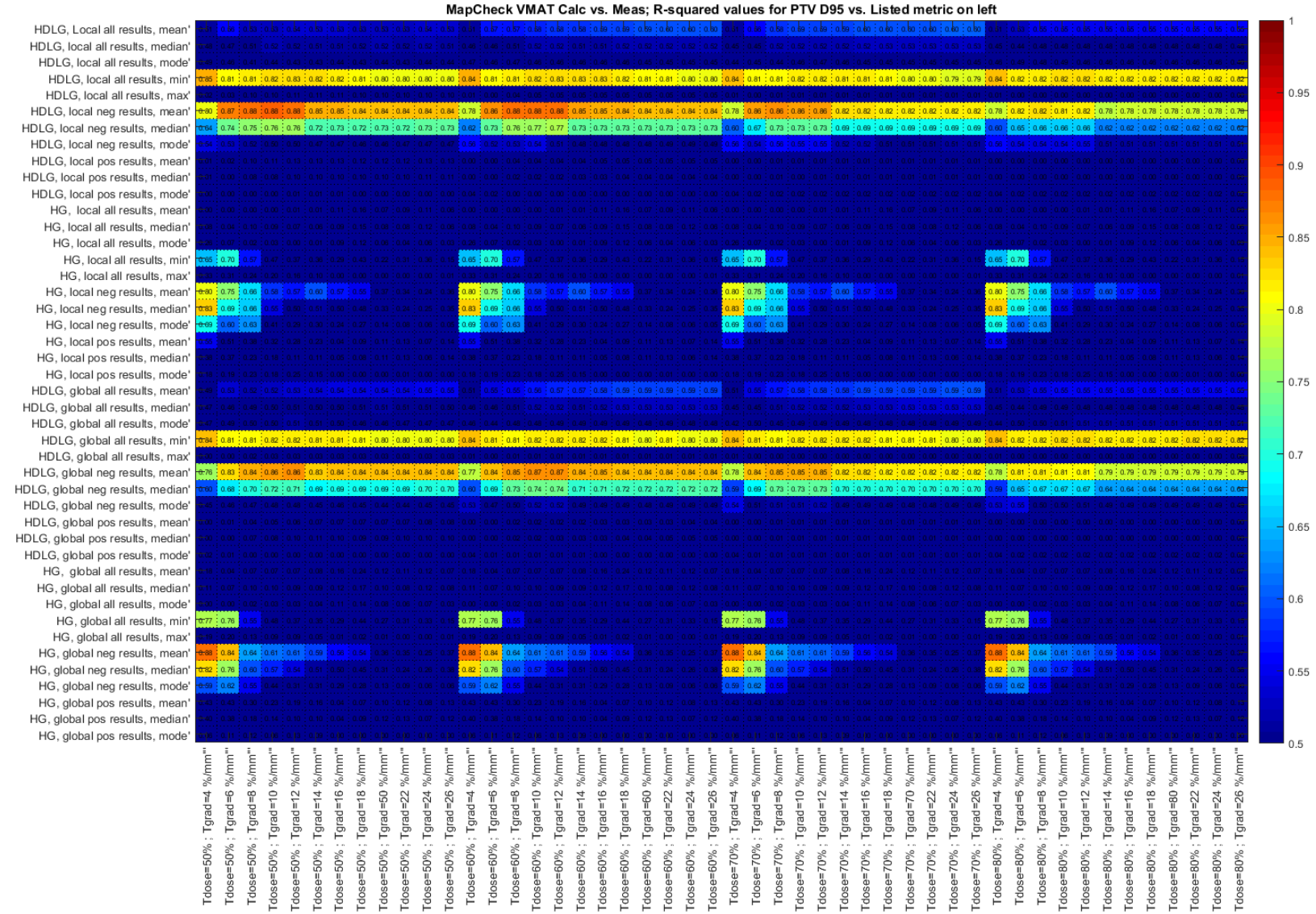

Figure A.5-65. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in PTV $D_{95\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

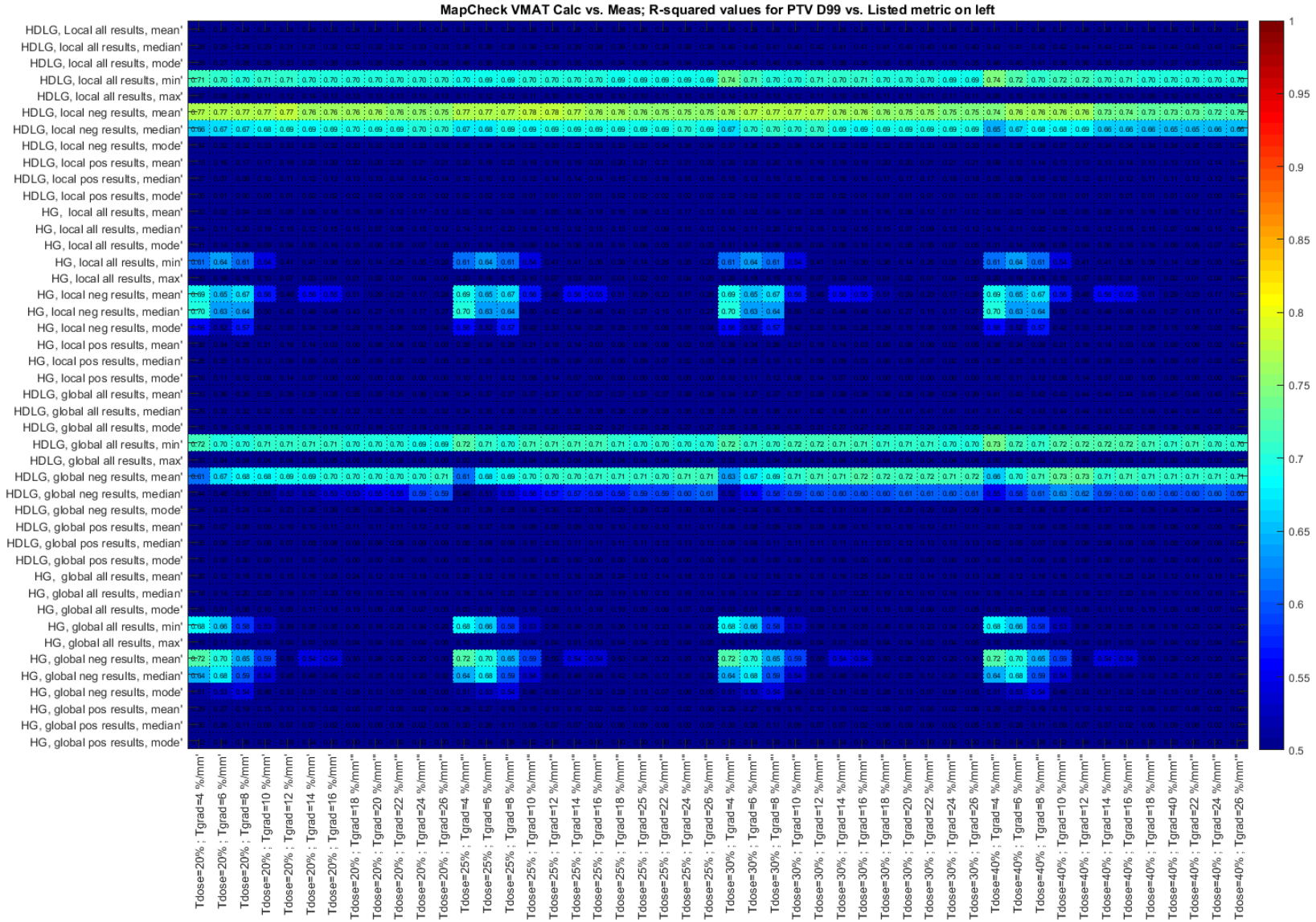

Figure A.5-66. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in PTV $D_{99\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

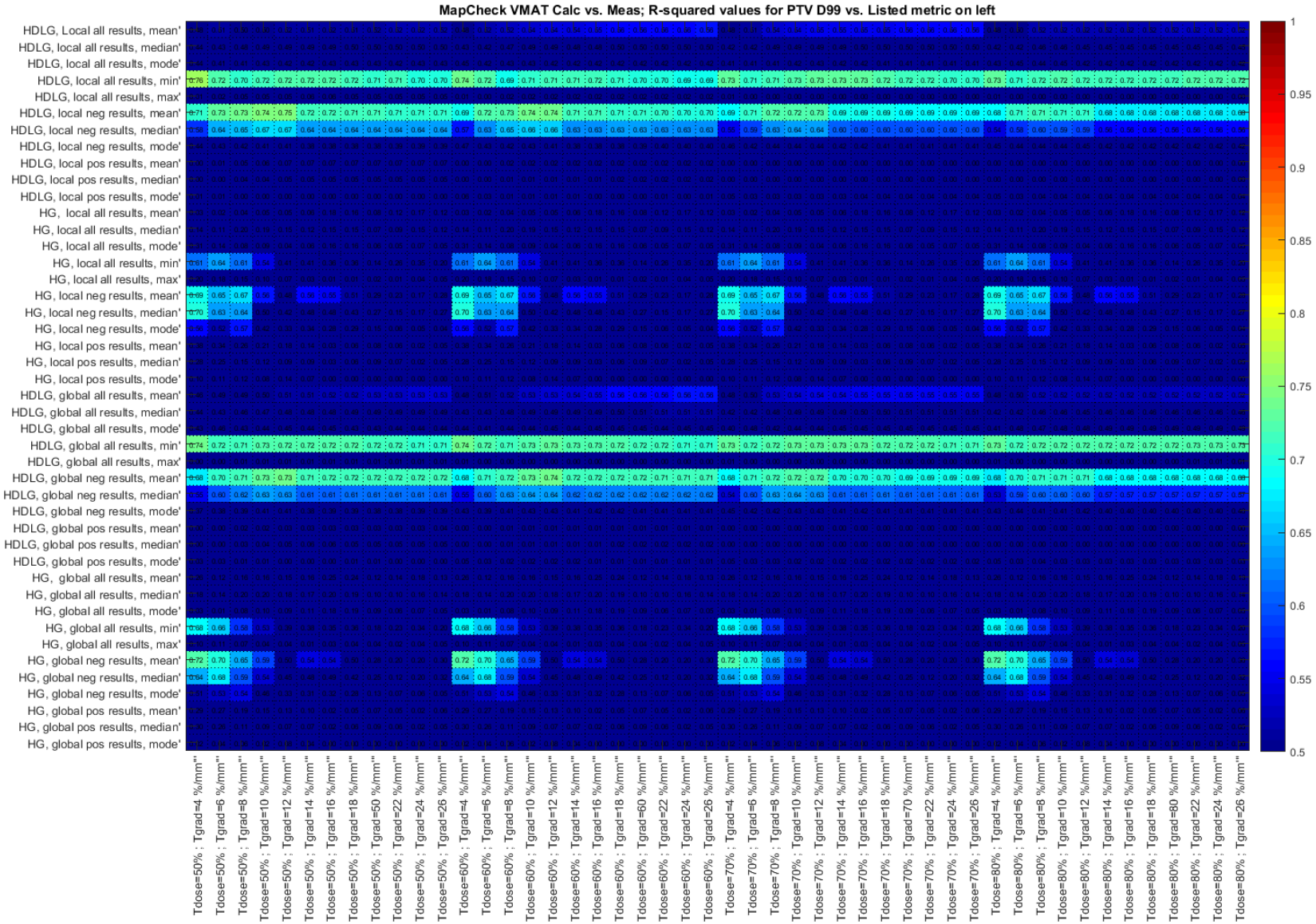

Figure A.5-67. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in PTV $D_{99\%}$ in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

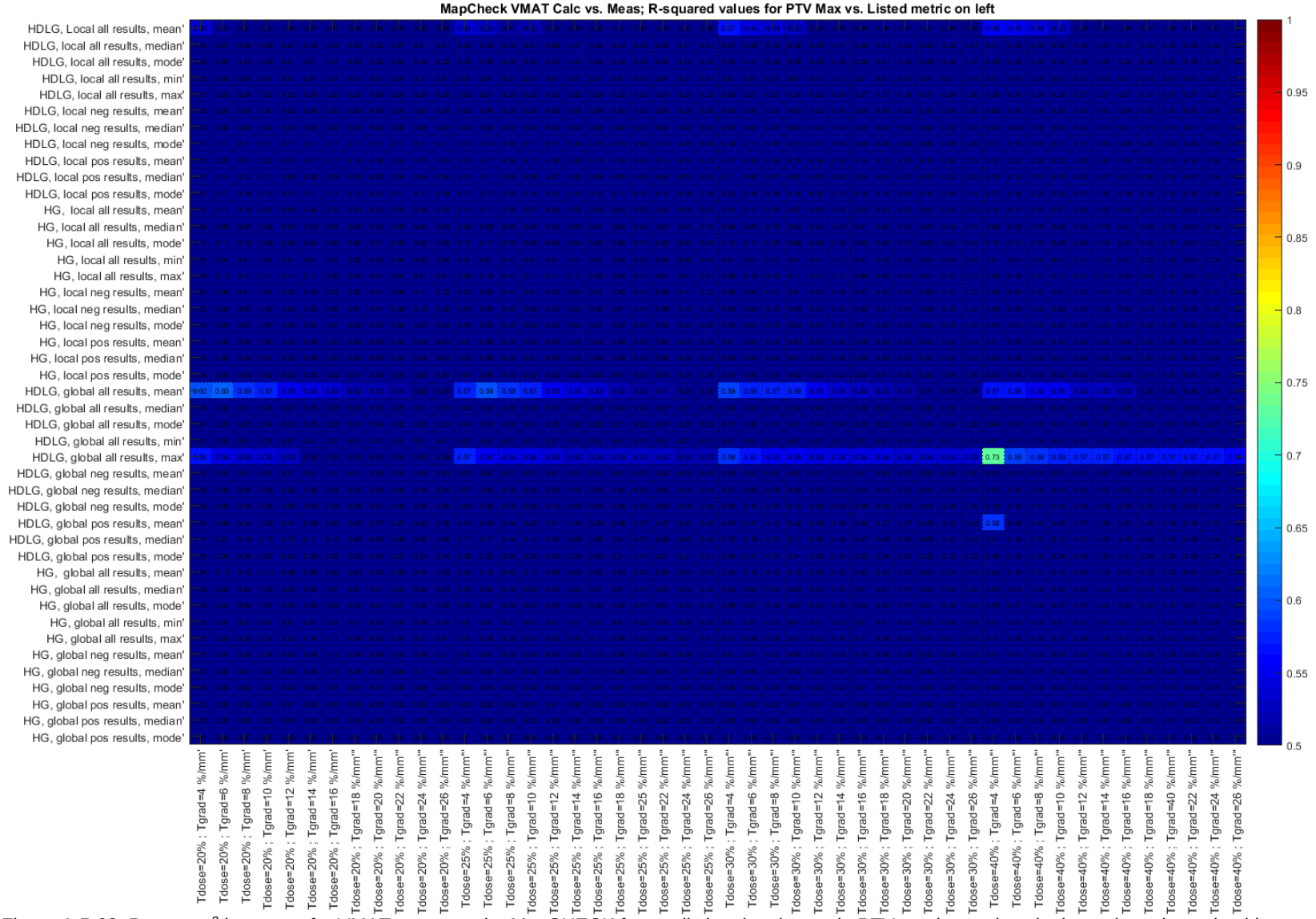

Figure A.5-68. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in PTV maximum dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

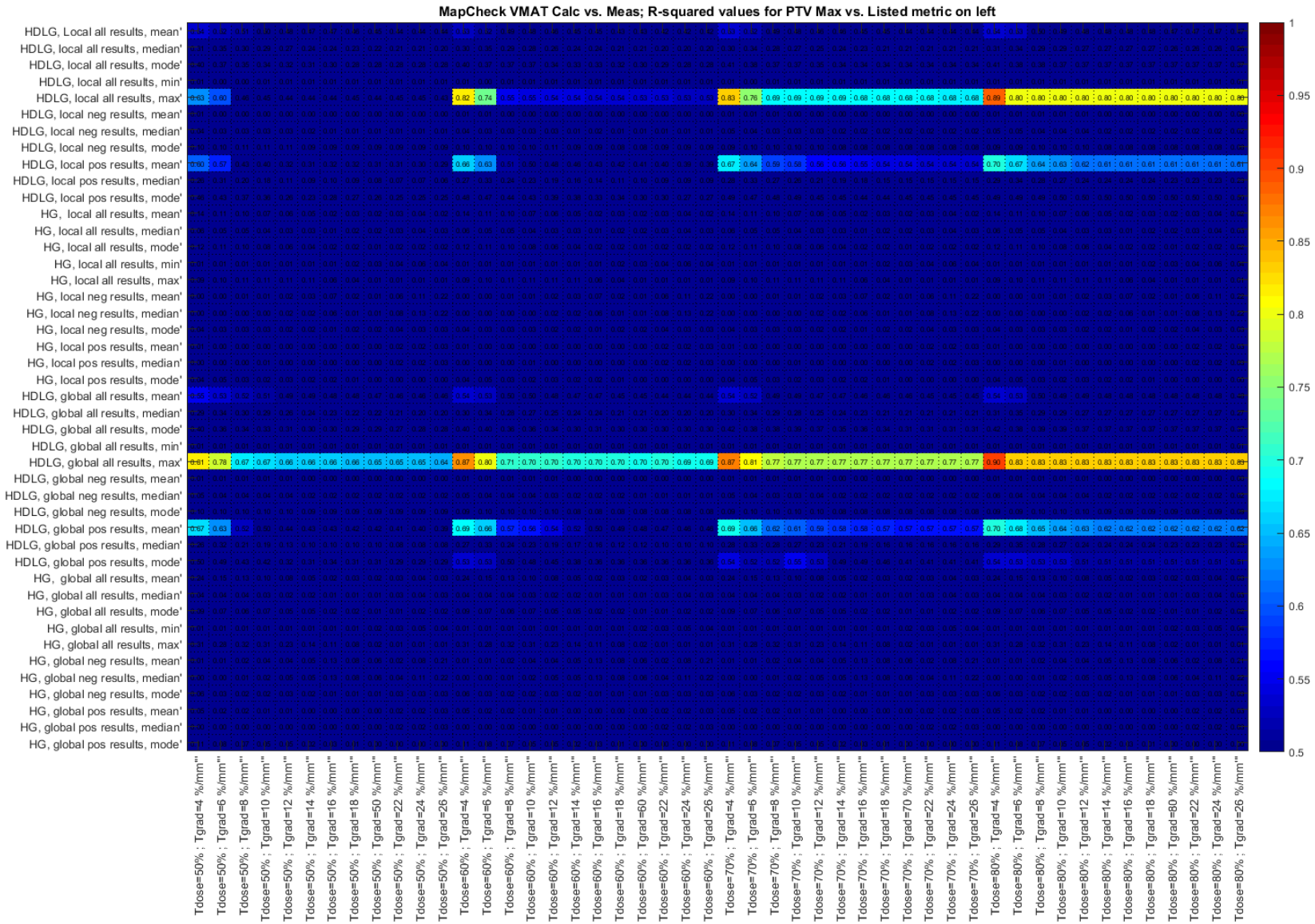

Figure A.5-69. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in PTV maximum dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

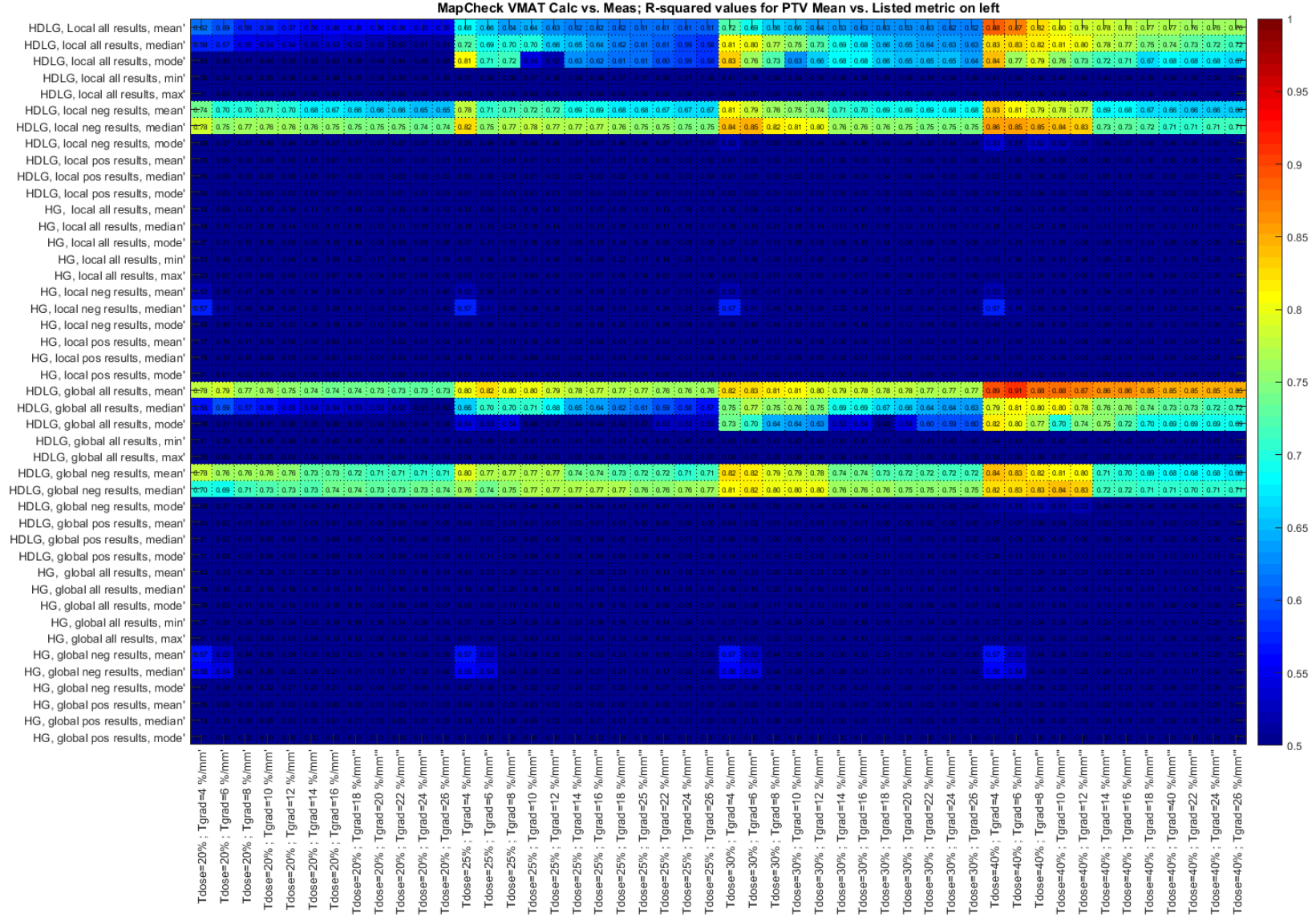

Figure A.5-70. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in PTV mean dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

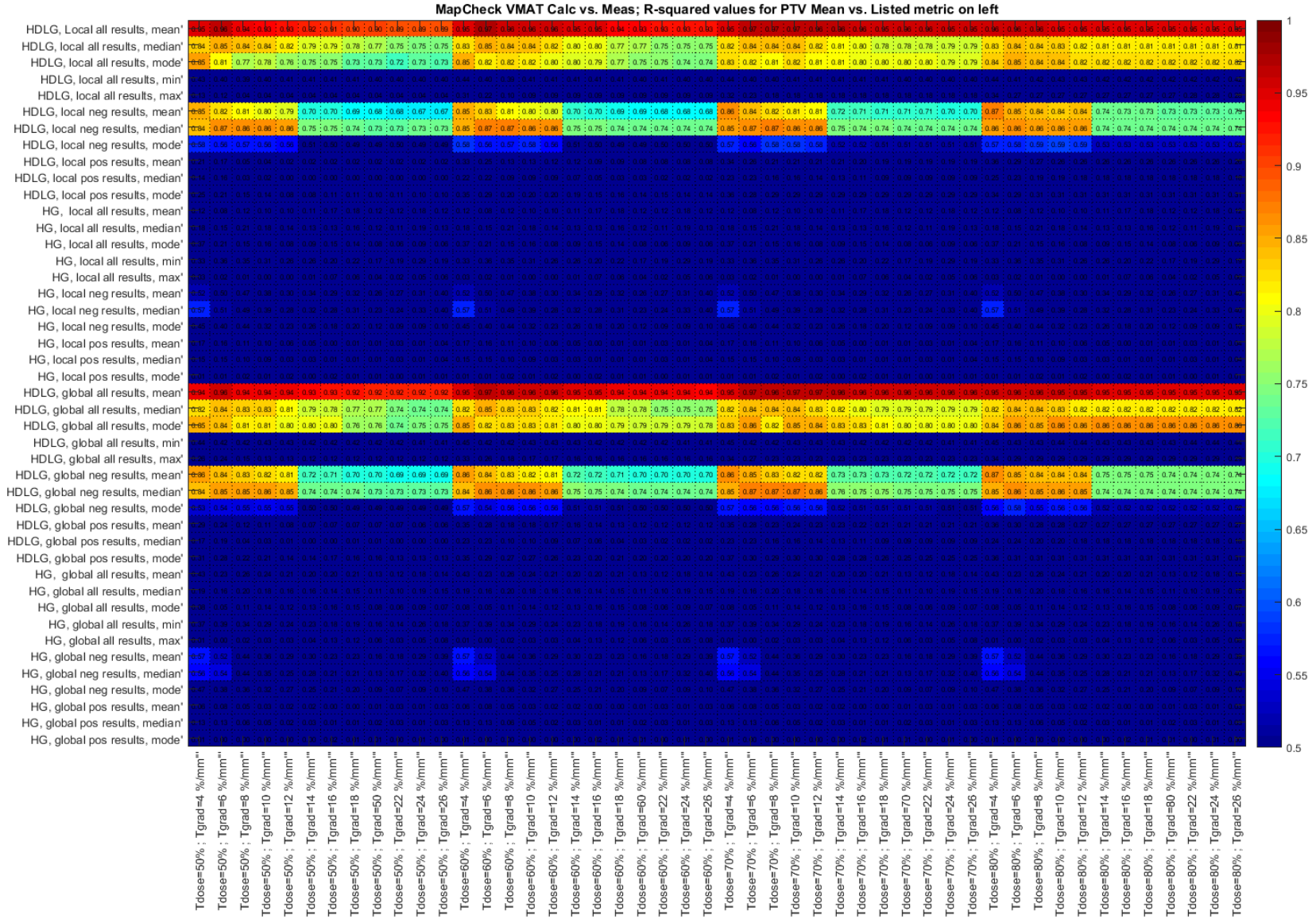

Figure A.5-71. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in PTV mean dose in the patient plan using histogram metrics obtained from high-dose low-gradient (HDLG) segmented dose differences and high-gradient (HG) dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

A.V.12 MapCHECK VMAT Ring Correlations

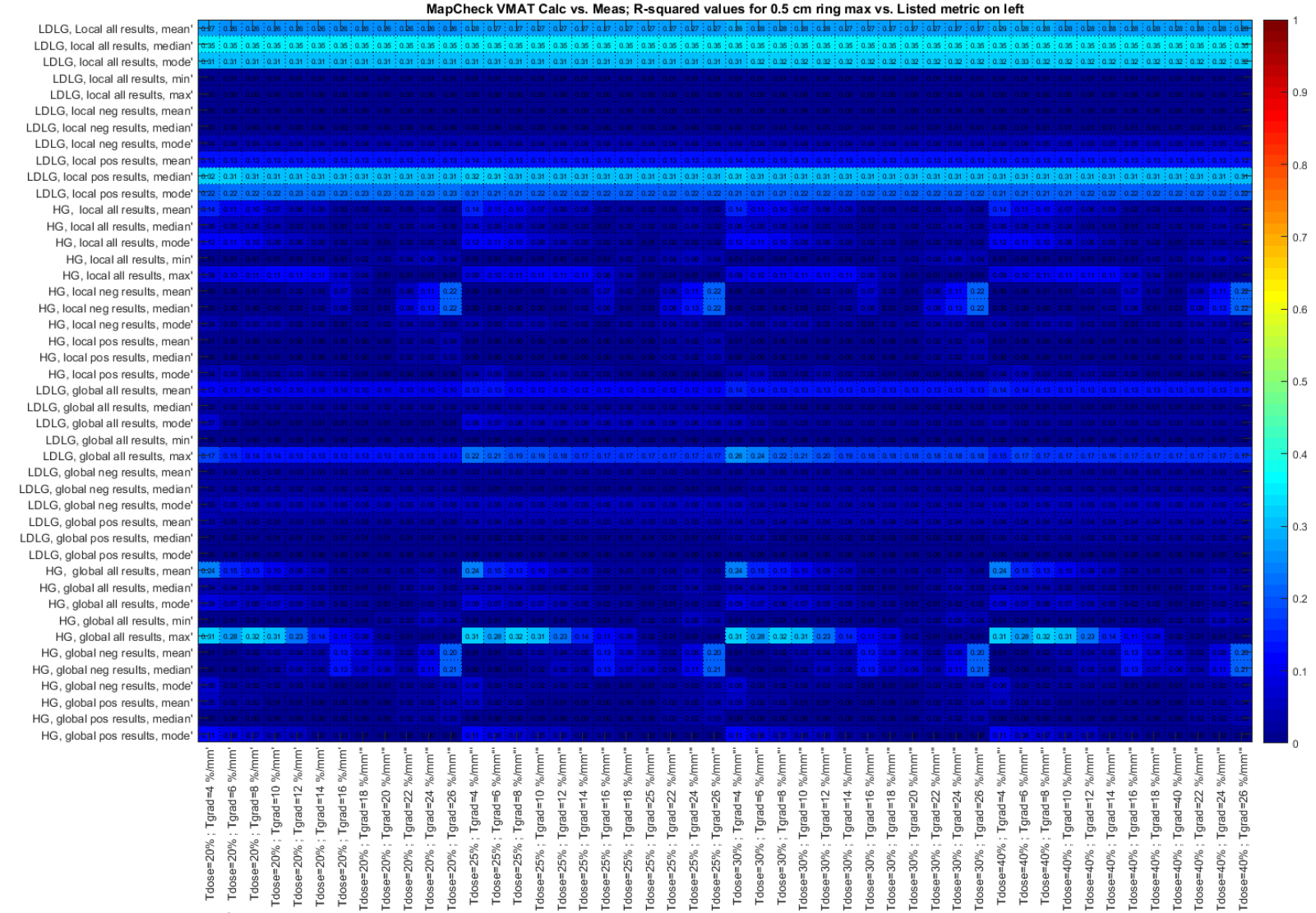

Figure A.5-72. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 0.5 cm ring maximum dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

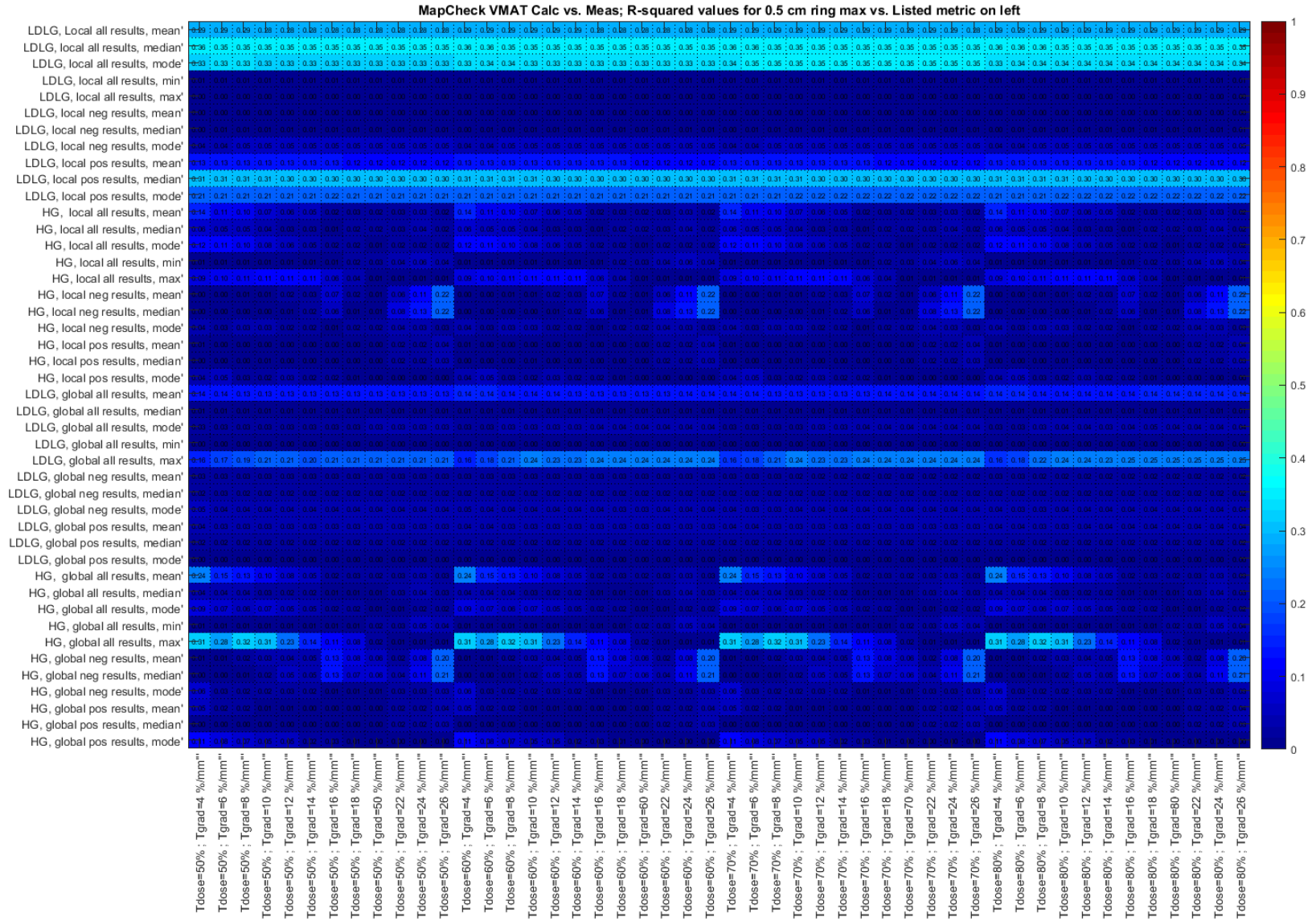

Figure A.5-73. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 0.5 cm ring maximum dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

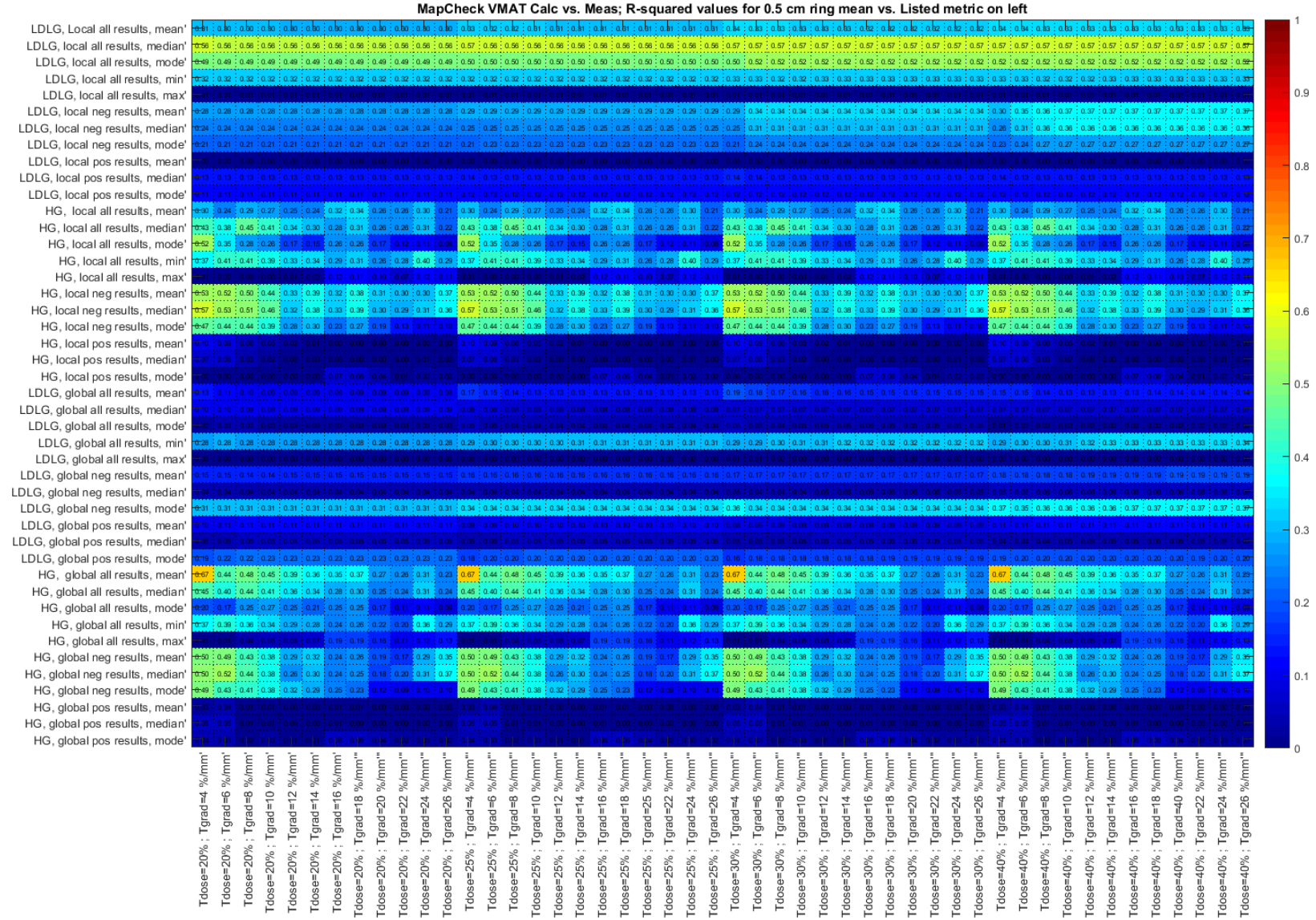

Figure A.5-74. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 0.5 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

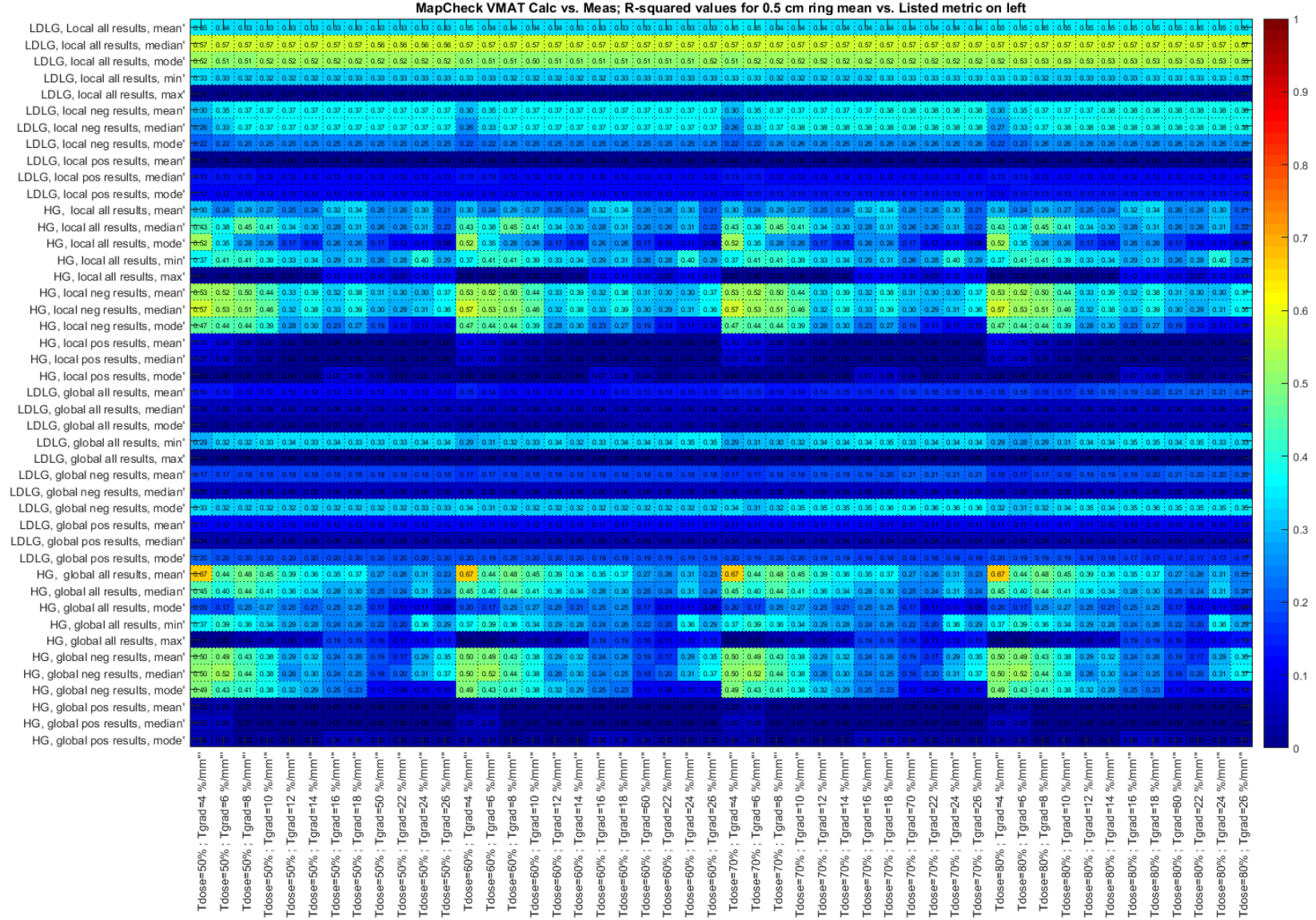

Figure A.5-75. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 0.5 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

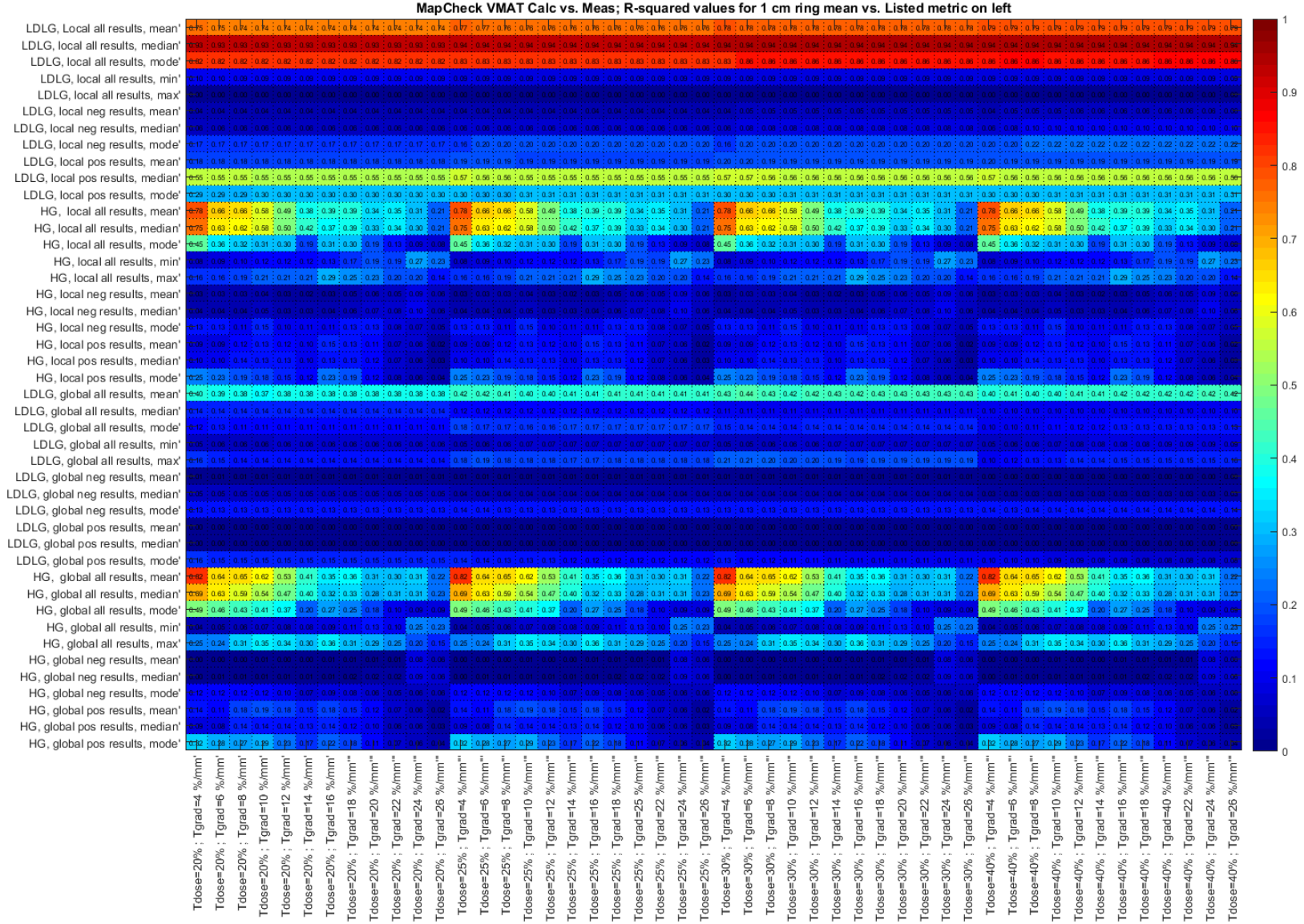

Figure A.5-76. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 1 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

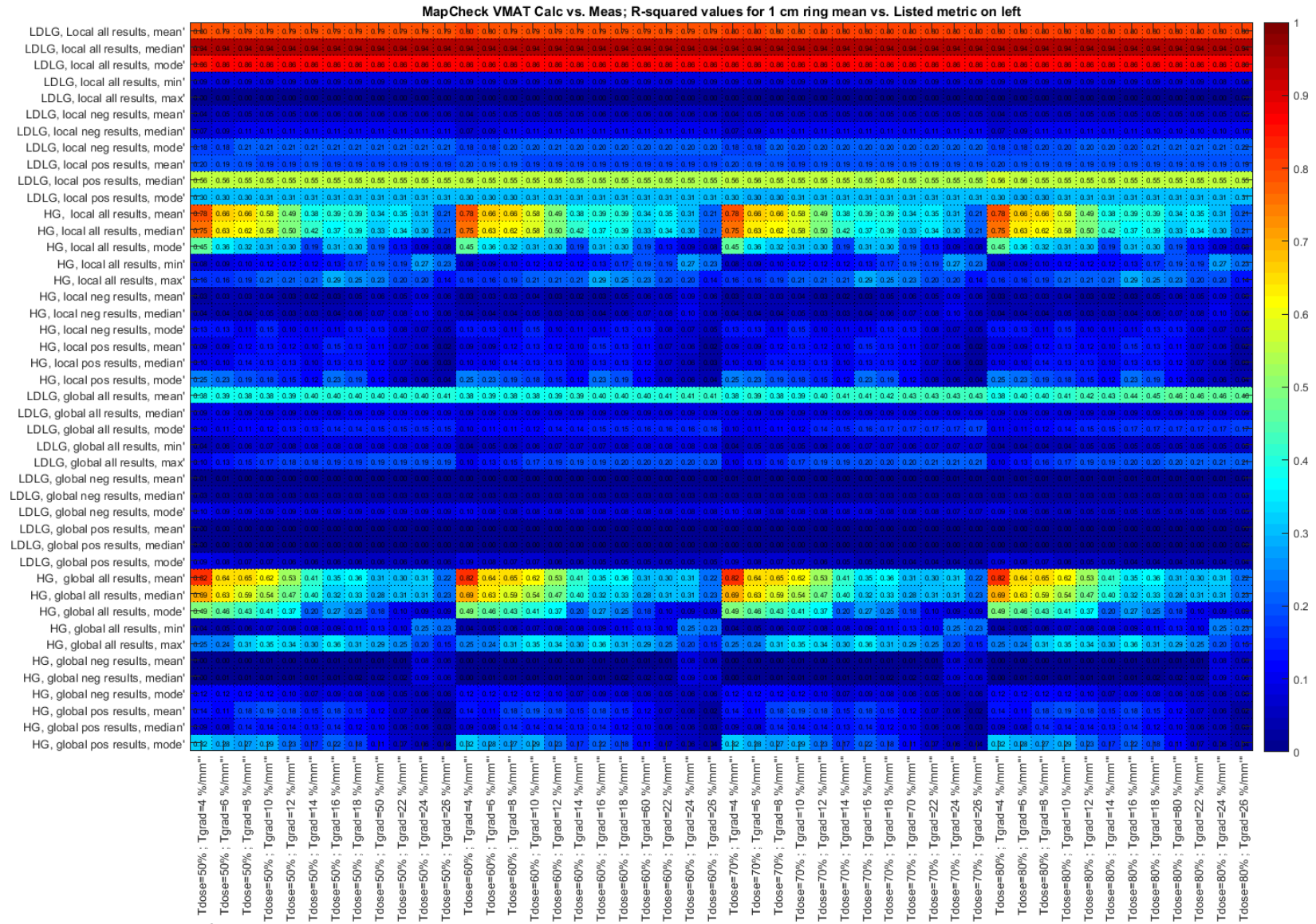

Figure A.5-77. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 1 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

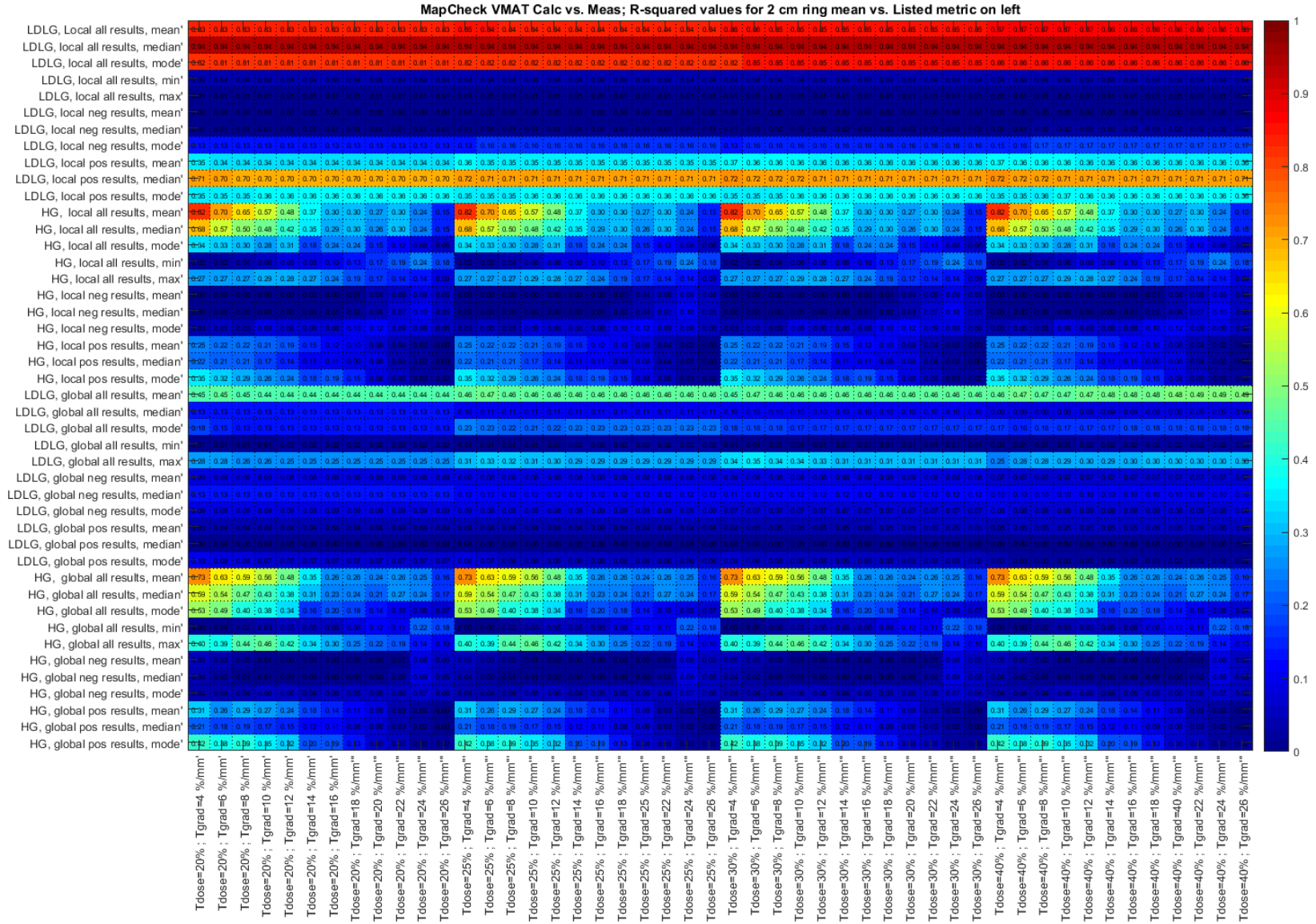

Figure A.5-78. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 2 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

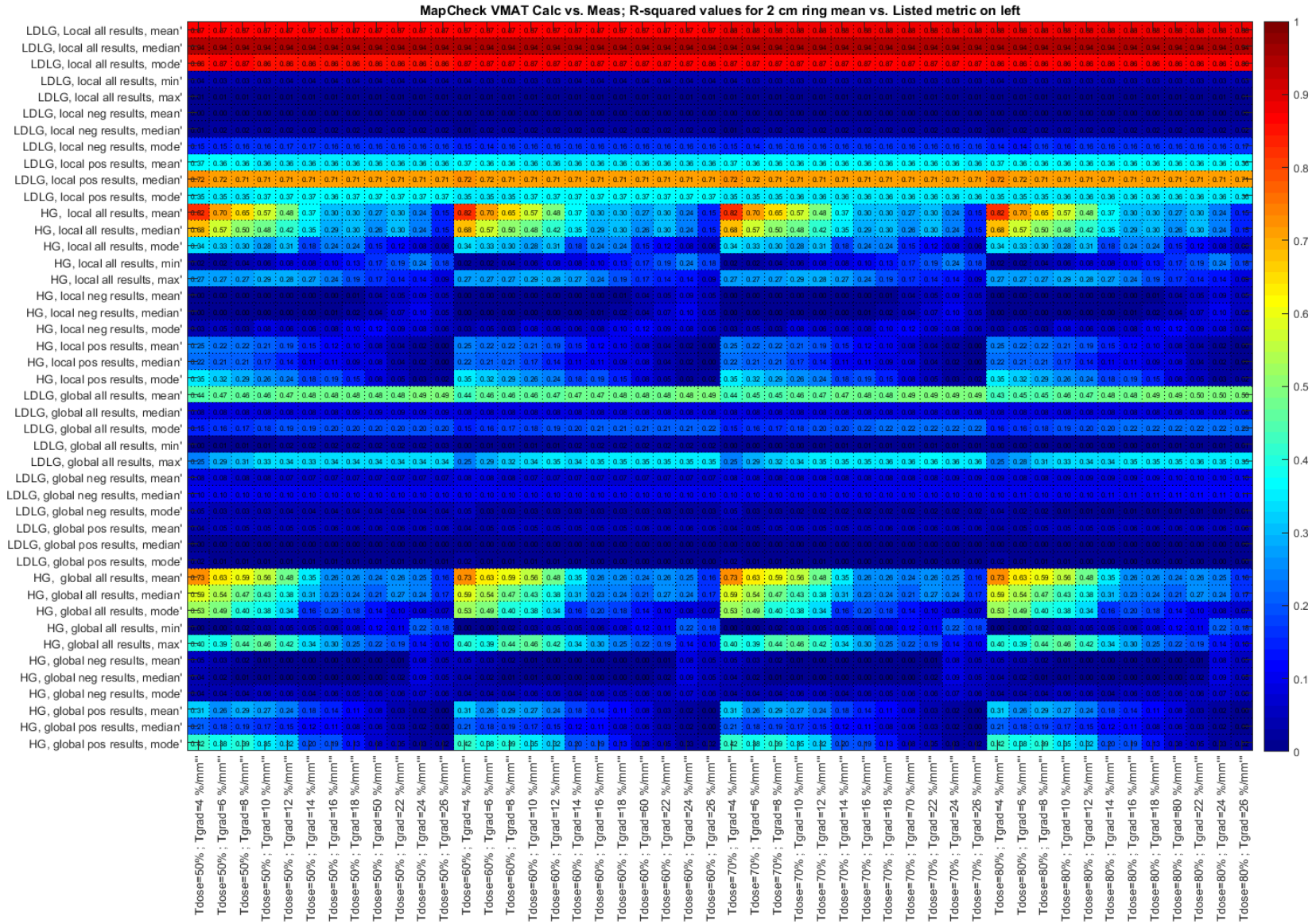

Figure A.5-79. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 2 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

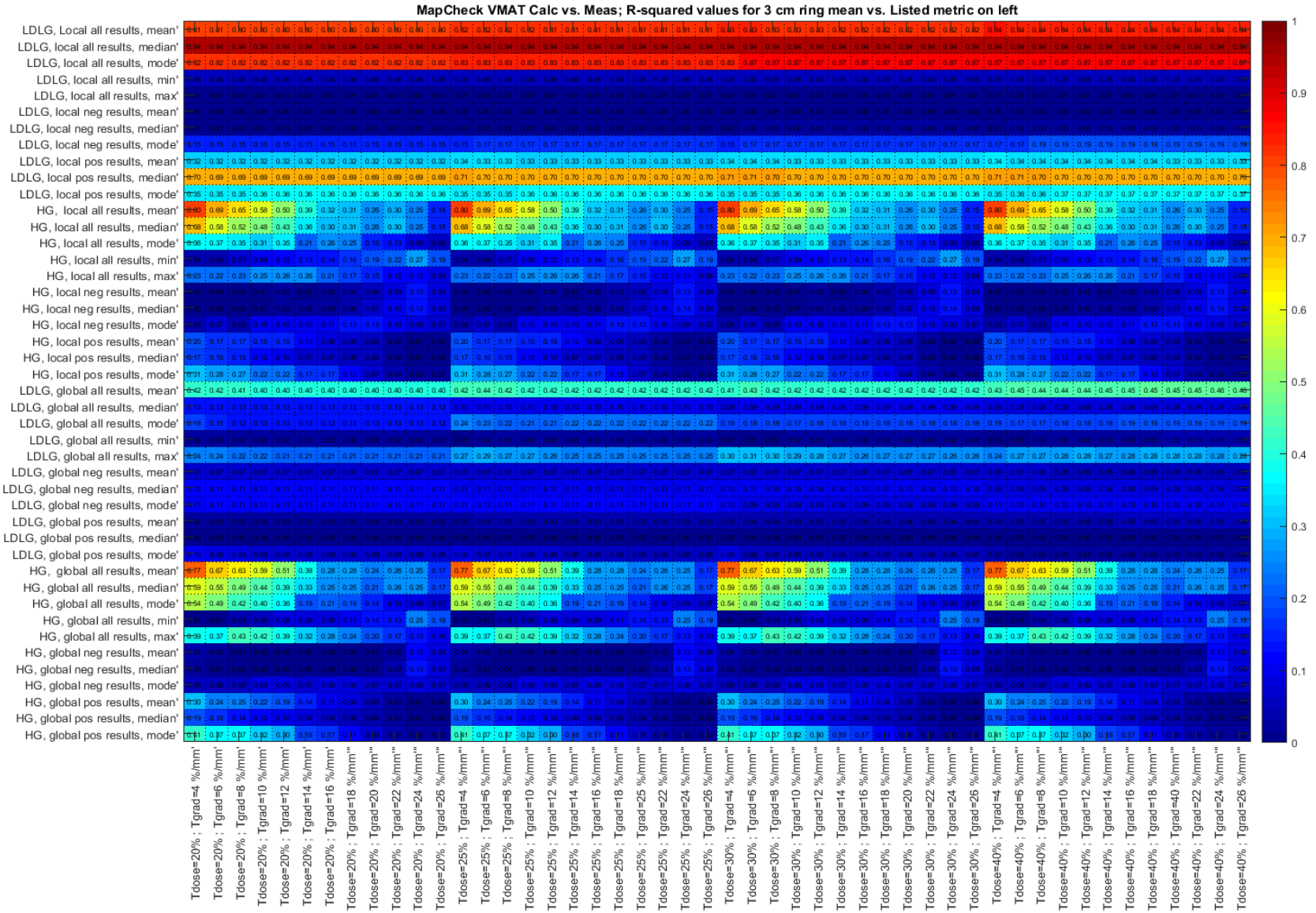

Figure A.5-80. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 3 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 20%-40% for all gradient thresholds used for segmentation are shown on the x-axis.

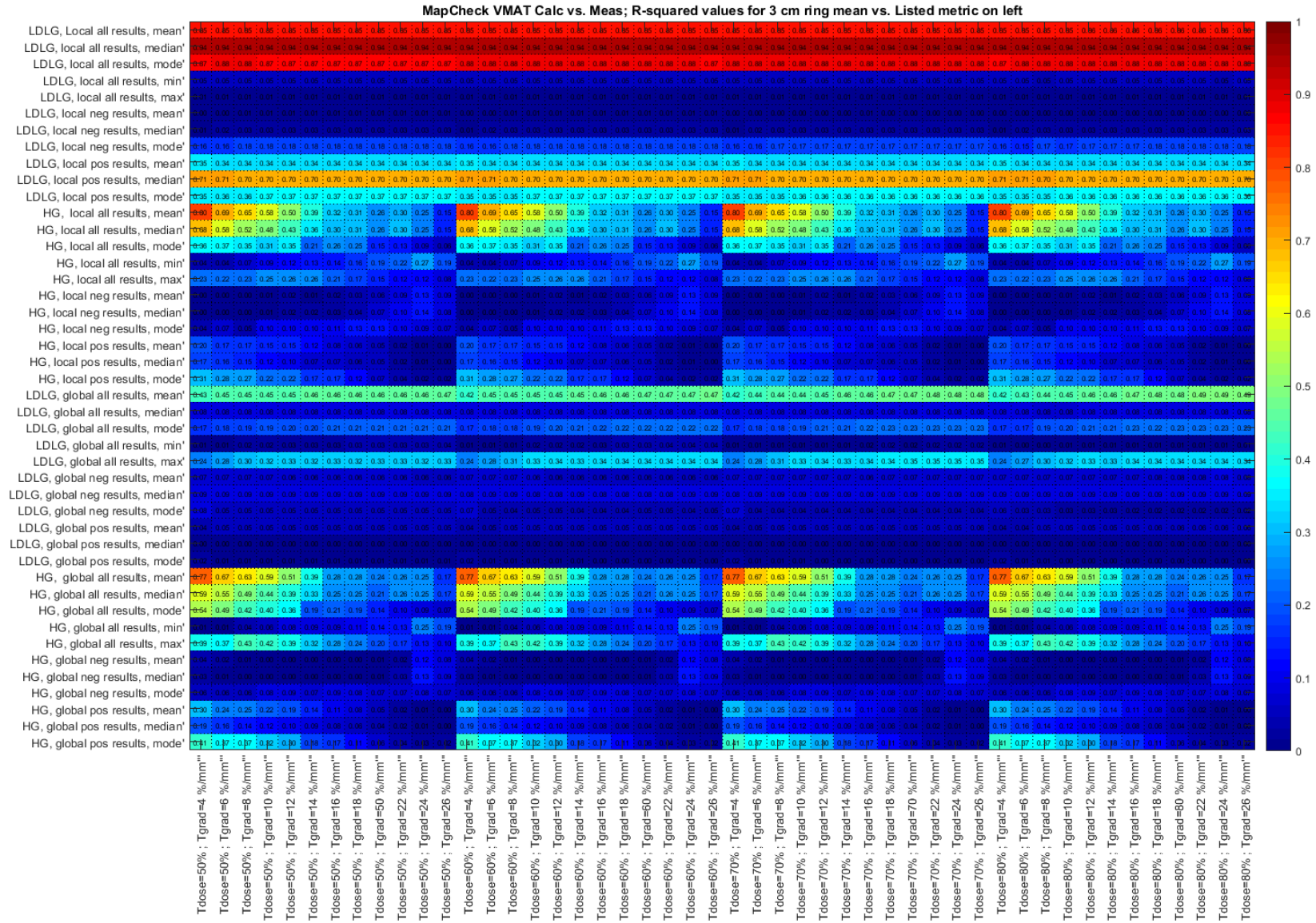

Figure A.5-81. Pearson r^2 heat map for VMAT cases on the MapCHECK for predicting the change in 3 cm ring mean dose in the patient plan using histogram metrics obtained from high-gradient (HG) segmented dose differences and low-dose low-gradient (LDLG) segmented dose differences (shown on the y-axis) on the phantom geometry. Dose and gradient thresholds ranging from 50%-80% for all gradient thresholds used for segmentation are shown on the x-axis.

B.V. Appendix B for Chapter V

The predicted change in PTV mean dose (utilizing local mean dose differences in high-dose, low-gradient comparison points) is shown for IMRT and VMAT cohorts for each detector geometry with results separated by delivery error type.

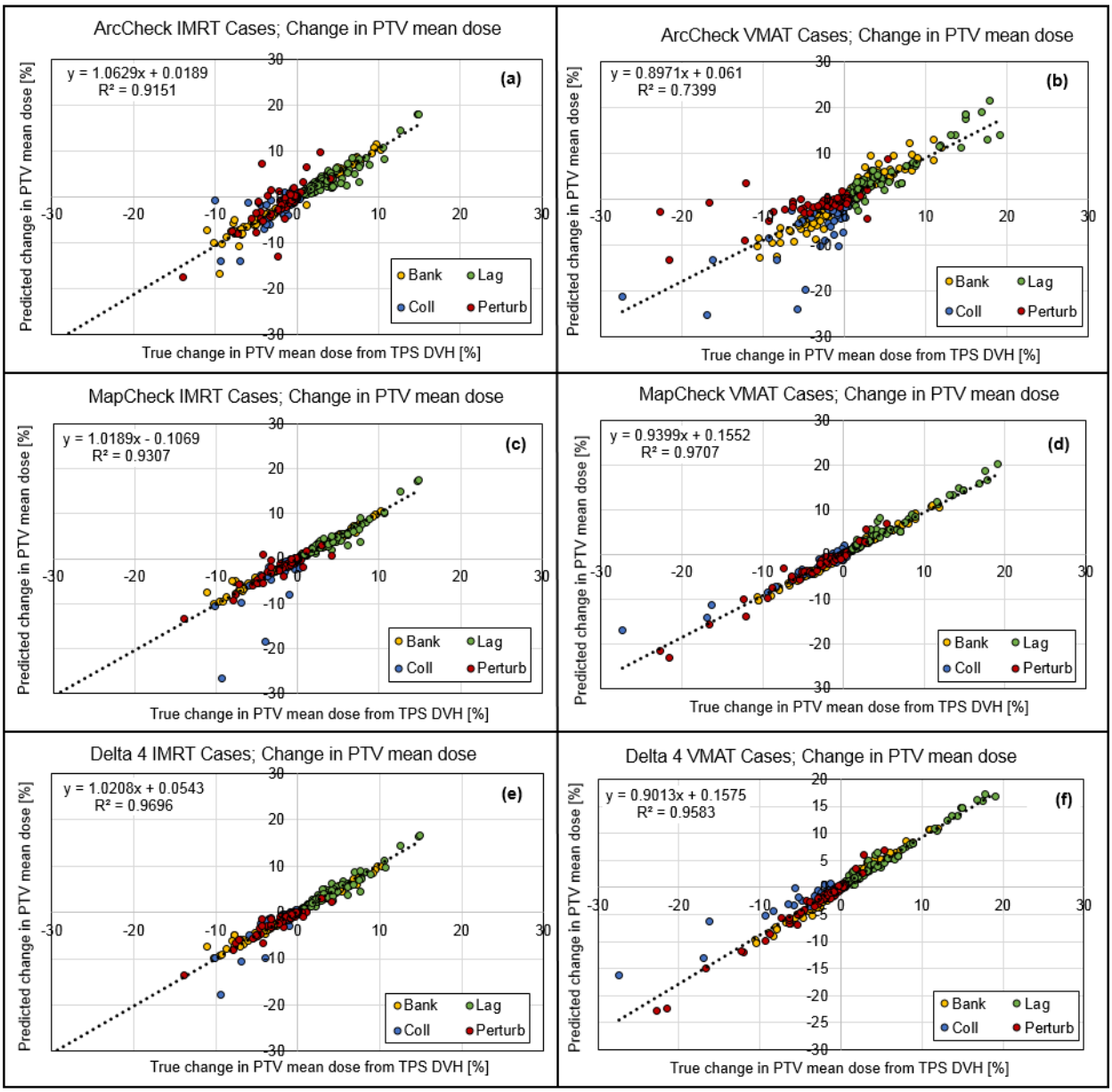

Figure B.5-1. Correlations between predicted change in PTV mean dose from the high-dose low-gradient points compared to the true change in the PTV mean in the patient geometry from the TPS DVHs split by error type for both IMRT cases (left) and VMAT cases (right) for each studied device geometry – ArcCHECK (a-b), MapCHECK (c-d), and Delta 4 (e-f).

C.V. Appendix C for Chapter V

Figure C.5-1, Figure C.5-2, and Figure C.5-3 show the true changes in PTV $D_{95\%}$ and PTV mean dose against the gamma passing rates in our calculation-only scenario for VMAT cases on the ArcCHECK, MapCHECK, and Delta 4, respectively. Some plans with large changes to the PTV $D_{95\%}$ and PTV mean dose can pass QA using the gamma criteria shown. Additionally, these figures further illustrate the point that when a plan has passing rates even above 99%, there can still be appreciable dose differences in the plan (here dose differences in the PTV $D_{95\%}$ and PTV mean exceeding 2-4%).

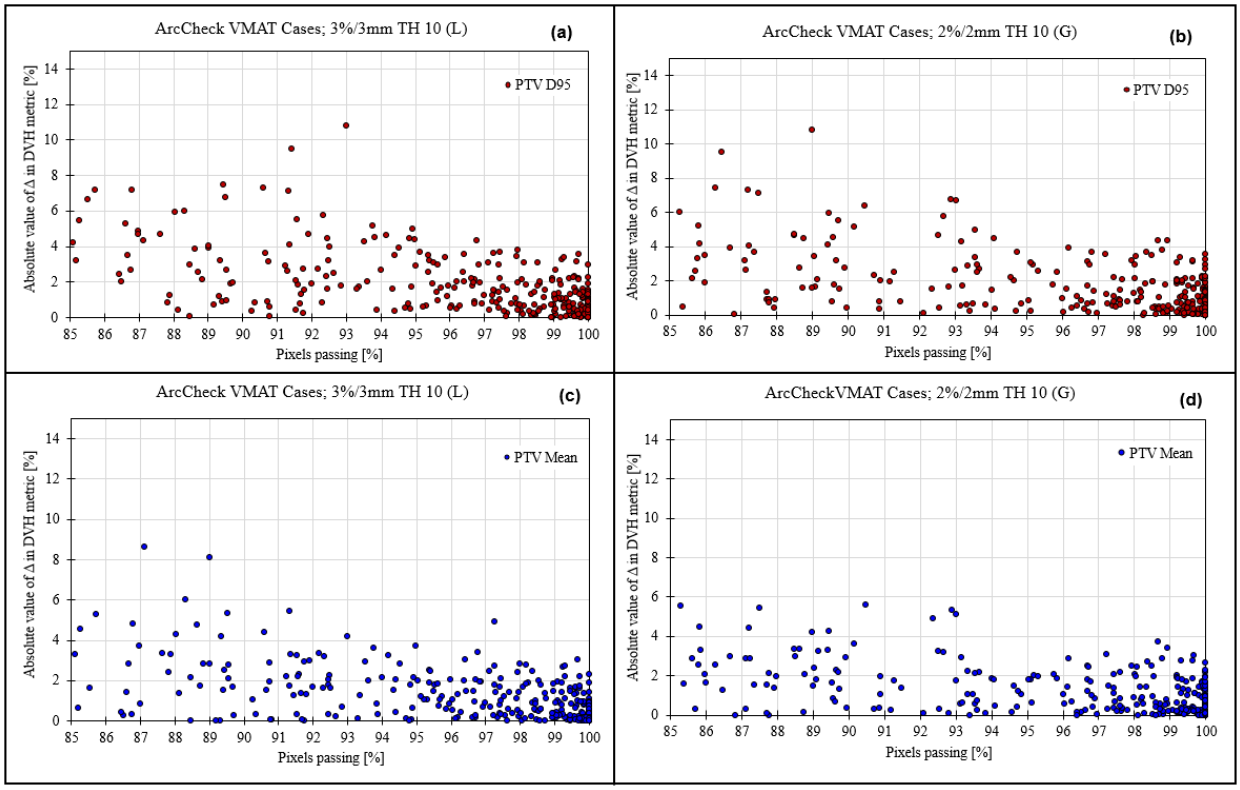

Figure C.5-1. The true change in PTV $D_{95\%}$ for (a) 3%/3mm TH10 (local), (b) 2%/2mm TH10 (global) and the true change in the PTV mean dose for (c) 3%/3mm TH10 (local) and (d) 2%/2mm TH10 (global) for IVMAT cases on the ArcCHECK geometry.

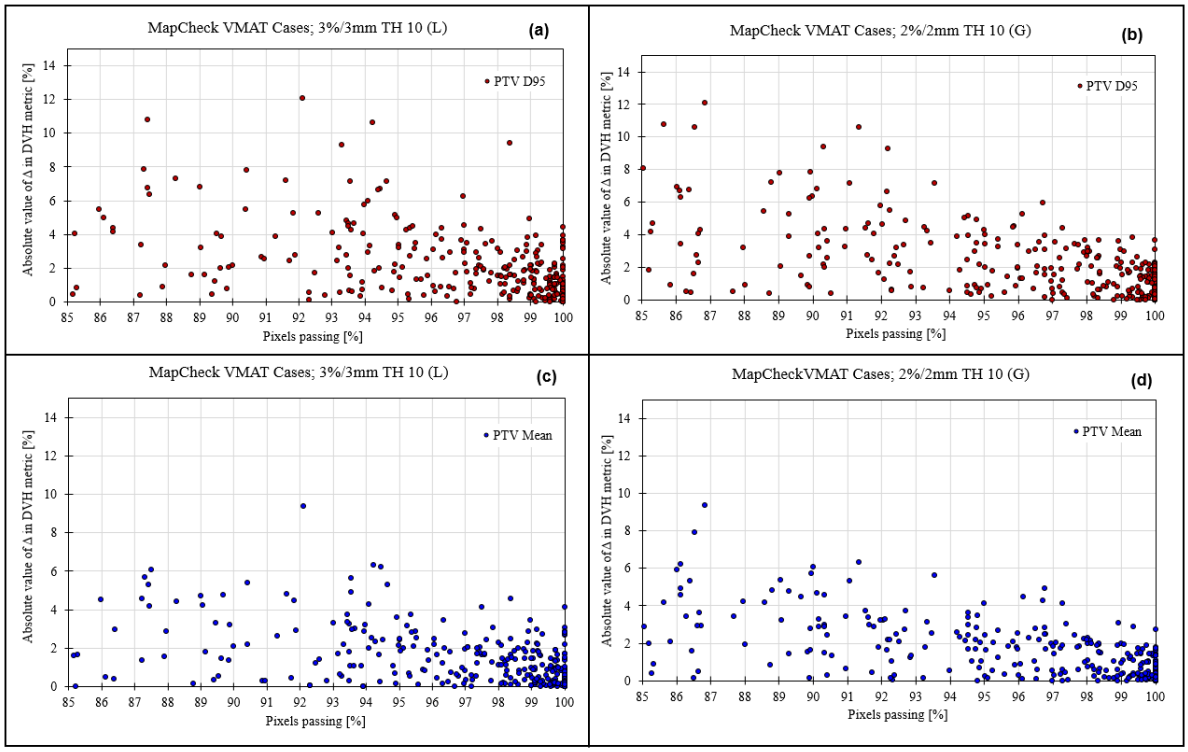

Figure C.5-2. The true change in PTV $D_{95\%}$ for (a) 3%/3mm TH10 (local), (b) 2%/2mm TH10 (global) and the true change in the PTV mean dose for (c) 3%/3mm TH10 (local) and (d) 2%/2mm TH10 (global) for IVMAT cases on the MapCHECK geometry.

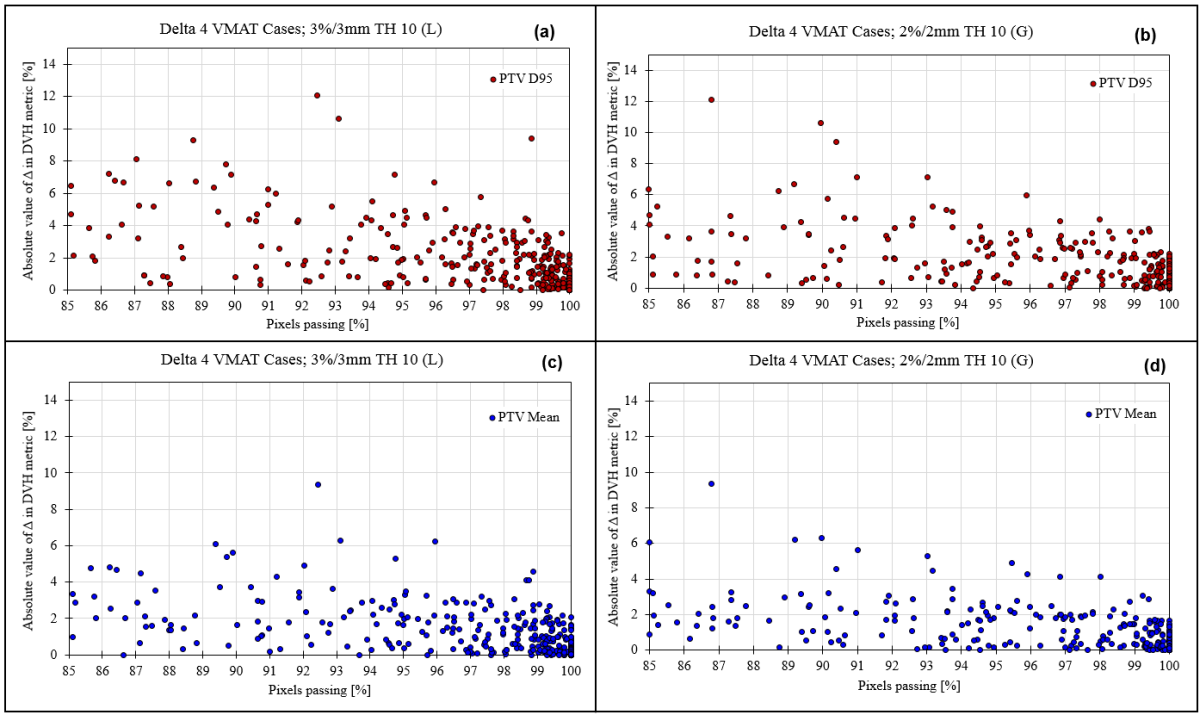

Figure C.5-3. The true change in PTV $D_{95\%}$ for (a) 3%/3mm TH10 (local), (b) 2%/2mm TH10 (global) and the true change in the PTV mean dose for (c) 3%/3mm TH10 (local) and (d) 2%/2mm TH10 (global) for IVMAT cases on the Delta 4 geometry.

References

1. McKenzie EM, Balter P a, Stingo FC, Jones J, Followill DS, Kry SF. Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans. *Med Phys*. 2014;41(12):121702. doi:10.1118/1.4899177.
2. Kry SF, Molineu A, Kerns JR, et al. Institutional patient-specific IMRT QA does not predict unacceptable plan delivery. *Int J Radiat Oncol Biol Phys*. 2014;90(5):1195-1201. doi:10.1016/j.ijrobp.2014.08.334.
3. Steers JM, Fraass B a. IMRT QA: Selecting gamma criteria based on error detection sensitivity. *Med Phys*. 2016;43(4):1982-1994. doi:10.1118/1.4943953.
4. Kruse JJ. On the insensitivity of single field planar dosimetry to IMRT inaccuracies. *Med Phys*. 2010;37(6):2516. doi:10.1118/1.3425781.
5. Nelms B, Jarry G, Chan M, Hampton C, Watanabe Y, Feygelman V. Real-world examples of sensitivity failures of the 3%/3mm pass rate metric and published action levels when used in IMRT/VMAT system commissioning. *J Phys Conf Ser*. 2013;444:012086. doi:10.1088/1742-6596/444/1/012086.
6. Nelms BE, Chan MF, Jarry G, et al. Evaluating IMRT and VMAT dose accuracy: practical examples of failure to detect systematic errors when applying a commonly used metric and action levels. *Med Phys*. 2013;40(11):111722. doi:10.1118/1.4826166.
7. Heilemann G, Poppe B, Laub W. On the sensitivity of common gamma-index evaluation methods to MLC misalignments in Rapidarc quality assurance. *Med Phys*. 2013;40(3):031702. doi:10.1118/1.4789580.
8. Nelms BE, Zhen H, Tomé W a. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. *Med Phys*. 2011;38(2):1037-1044. doi:10.1118/1.3544657.
9. Kruse JJ. On the insensitivity of single field planar dosimetry to IMRT inaccuracies. *Med Phys*. 2010;37(6):2516-2524. doi:10.1118/1.3514143.
10. Stasi M, Bresciani S, Miranti A, Maggio A, Sapino V, Gabriele P. Pretreatment patient-specific IMRT quality assurance : A correlation study. *Med Phys*. 2012;39(12):7626-7634. doi:10.1118/1.4767763.
11. Templeton AK, Chu JCH, Turian J V. The sensitivity of ArcCHECK-based gamma analysis to manufactured errors in helical tomotherapy radiation delivery. *J Appl Clin Med Phys*. 2015;16(1):4814. doi:10.1120/jacmp.v16i1.4814.
12. Jin X, Yan H, Han C, Zhou Y, Yi J, Xie C. Correlation between gamma index passing rate and clinical dosimetric difference for pre-treatment 2D and 3D volumetric modulated arc therapy dosimetric verification. *Br J Radiol*. 2015;88(1047). doi:10.1259/bjr.20140577.
13. Renner WD, Norton K, Holmes T. A method for deconvolution of integrated electronic portal images to obtain incident fluence for dose reconstruction. *J Appl Clin Med Phys*. 2005;6(4):22-39. doi:10.1120/jacmp.v6i4.2104.
14. Boggula R, Lorenz F, Mueller L, et al. Experimental validation of a commercial 3D dose verification system for intensity-modulated arc therapies. *Phys Med Biol*. 2010;55(19):5619-5633. doi:10.1088/0031-9155/55/19/001.
15. Olch AJ. Evaluation of the accuracy of 3DVH software estimates of dose to virtual ion chamber and film in composite IMRT QA. *Med Phys*. 2011;39(1):81-86. doi:10.1118/1.3666771.
16. Nakaguchi Y, Araki F, Maruyama M, Saiga S. Dose verification of IMRT by use of a COMPASS transmission detector. *Radiol Phys Technol*. 2012;5(1):63-70. doi:10.1007/s12194-011-0137-y.
17. Nelms BE, Opp D, Robinson J, et al. VMAT QA: Measurement-guided 4D dose

reconstruction on a patient. *Med Phys*. 2012;39(7):4228-4238. doi:10.1118/1.4729709.

18. Opp D, Nelms BE, Zhang G, Stevens C, Feygelman V. Validation of measurement-guided 3D VMAT dose reconstruction on a heterogeneous anthropomorphic phantom. *J Appl Clin Med Phys*. 2013;14(4):70-84. doi:10.1120/jacmp.v14i4.4154.
19. Chan. Using a Novel Dose QA Tool to Quantify the Impact of Systematic Errors Otherwise Undetected by Conventional QA Methods: Clinical Head and Neck Case Studies. *Technol Cancer Res Treat*. 2013;13(1):1-11. doi:10.7785/tcrt.2012.500353.
20. Ezzell G a, Galvin JM, Low D, et al. Guidance document on delivery, treatment planning, and clinical implementation of IMRT: report of the IMRT Subcommittee of the AAPM Radiation Therapy Committee. *Med Phys*. 2003;30(8):2089-2115. doi:10.1118/1.1591194.
21. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. *Med Phys*. 2018;45(4):e53-e83. doi:10.1002/mp.12810.
22. Wootton L, Chaovalitwongse A, Li N, Nyflot M, Ford E. MO-FG-202-06: Improving the Performance of Gamma Analysis QA with Radiomics- Based Image Analysis. *Med Phys*. 2016;43(6Part31):3713. doi:10.1118/1.4957309.
23. Valdes G, Chan MF, Lim SB, Scheuermann R, Deasy JO, Solberg TD. IMRT QA using machine learning: A multi-institutional validation. *J Appl Clin Med Phys*. 2017;18(5):279-284. doi:10.1002/acm2.12161.
24. Childress NL, Rosen II. The design and testing of novel clinical parameters for dose comparison. *Int J Radiat Oncol Biol Phys*. 2003;56(5):1464-1479. doi:10.1016/S0360-3016(03)00430-9.
25. Hodapp N. [The ICRU Report 83: prescribing, recording and reporting photon-beam intensity-modulated radiation therapy (IMRT)]. *Strahlentherapie und Onkol Organ der Dtsch Rontgengesellschaft . [et al]*. 2012;188(1):97-99. doi:10.1007/s00066-011-0015-x.
26. Das IJ, Andersen A, Chen ZJ, et al. State of dose prescription and compliance to international standard (ICRU-83) in intensity modulated radiation therapy among academic institutions. *Pract Radiat Oncol*. 2017;7(2):e145-e155. doi:10.1016/j.prro.2016.11.003.
27. Shaw E, Kline R, Gillin M, et al. Radiation therapy oncology group: Radiosurgery quality assurance guidelines. *Int J Radiat Oncol*. 1993;27(5):1231-1239. doi:10.1016/0360-3016(93)90548-A.
28. Feuvret L, Noël G, Mazeron J-J, Bey P. Conformity index: A review. *Int J Radiat Oncol*. 2006;64(2):333-342. doi:10.1016/J.IJROBP.2005.09.028.
29. Yan G, Liu C, Simon TA, Peng LC, Fox C, Li JG. On the sensitivity of patient-specific IMRT QA to MLC positioning errors. *J Appl Clin Med Phys*. 2009;10(1):120-128. doi:10.1118/1.2961902.
30. Zhen H, Nelms BE, Tomé W a. Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA. *Med Phys*. 2011;38(10):5477. doi:10.1118/1.3633904.

Chapter VI.

Validating the gradient-dose segmented analysis (GDSA) method

VI.A. Motivation

In the previous chapter the gradient-dose segmented analysis (GDSA) method was introduced for comparing measurements and calculations in patient-specific IMRT QA. GDSA was developed as an additional/alternative test to the gamma comparison, as gamma comparisons can be insensitive to clinically meaningful errors, and results lack a relationship to clinical dose metrics[1–14]. GDSA separates the analysis into regions of high-gradient, high-dose low-gradients, and low-dose low-gradients. The calculation-only comparisons in Chapter V showed that the mean local dose difference in the high-dose low-gradient regions of the phantom comparisons predicts changes in PTV mean dose in the patient DVH for three studied detector array geometries – ArcCHECK, MapCHECK, and Delta 4. The relationship between predicted change and true change was nearly 1:1, with strong linear correlations. Additionally, used as a binary metric with a threshold of 3% change in PTV mean for passing/failing a QA comparison, GDSA has improved sensitivity and specificity when compared to five different gamma criteria – 3%/3mm TH 10 (G), 3%/3mm TH 10 (L), 2%2mm TH 10 (G), 3%/3mm TH 50 (G), and 3%/3mm TH 50 (L).

Since the development of GDSA in Chapter V relied on 20 IMRT and 20 VMAT cases for calculation-only comparisons, real MapCHECK measurements were compared to calculations with both the GDSA method and the gamma comparison to evaluate if results of GDSA with real measurements would cause a large increase in plan failures in the clinic compared to the gamma comparison. Additionally, errors were induced in real plan deliveries to determine if GDSA could accurately predict the relative change in PTV mean between the error-free measurement and error-induced measurement. These results were compared against the known change in PTV mean from the patient TPS DVH. In this chapter we also validate the GDSA method in calculation-

only comparisons for a smaller validation data set consisting of randomly selected patient cases with induced errors for the ArcCHECK, MapCHECK, and Delta 4 geometries.

VI.B. Methods

VI.B.1. GDSA vs. gamma comparisons for clinical MapCHECK measurements

Clinical MapCHECK measurements and plan dose calculations were compared for 20 patient cases using both the gamma comparison and GDSA. Cases were chosen by chronologically selecting the plans that were previously measured in the clinic after the most recent dose and array calibrations on the MapCHECK 2 device.

The MATLAB code from Chapter V for the GDSA method was extended to perform GDSA comparisons with real measurements. The measurement and dose calculation files were opened using the MapCHECK software (Sun Nuclear Corporation, Melbourne FL) and dose matrices were copied manually from the software into MATLAB arrays to allow the calculation of 3D gradient maps in MATLAB. The new method was then evaluated on the MapCHECK device and evaluated against gamma comparison results obtained from the vendor software. Gamma comparisons were performed for 2%/2mm TH 10 (global) and 3%/3mm TH 10 (global) criteria (uncertainty off). All cases were 6 MV VMAT plans, with a large number being prostate or pelvis treatment sites. Unlike the work in Chapter V which made use of calculation-only comparisons, it is important to note that comparisons with real measurements lack a ground truth.

VI.B.2. GDSA vs. gamma comparisons for delivered plans with intentional errors

A second set of comparisons was made with real MapCHECK measurements both with and without intentional errors introduced in the plan delivery. Error-induced plans were created by modifying the error-free DICOM RT plans in MATLAB, re-importing the plans into Eclipse for dose calculation, and subsequently pushing both the clinical plan without induced errors (labeled the "error-free" plan) and the plan with intentional errors to the treatment machine. Intentional errors

included collimator rotation, MU scaling, bank shift, lagging leaf, and random MLC perturbational errors to induce a range of dose differences in PTV mean among the nine studied plans.

Error-free plans and error-induced plans were delivered using a TrueBeam linear accelerator and measured using a MapCHECK 2 device on the same day to avoid variations in output from the machine and differences in set-up. Measurements were compared to calculations without any intentional errors from the Eclipse TPS using AAA dose calculation and a 1mm grid size. Gamma comparisons were performed using the MapCHECK software (Sun Nuclear Corporation, Melbourne, FL) and GDSA predictions were performed using the in-house developed MATLAB software. For the GDSA analysis, the necessary dose planes for analysis were manually copied from the MapCHECK software into MATLAB. Changes in PTV mean predicted by GDSA between the error-free and error-induced deliveries were compared to the true changes in PTV mean from the patient DVHs calculated in the TPS. The reduction in gamma passing rates between error-free and error-induced plans for 3%/3mm TH 10 (G) And 2%/2mm TH 10 (G) were also recorded and compared. It should be noted that while the plan labeled “error-free” did not have any intentional errors, these measurements may still show differences when compared to the calculations due to measurement limitations on the MapCHECK device, MapCHECK calibration uncertainties, as well as beam modeling limitations, and plan delivery limitations.

VI.B.3. GDSA vs. gamma for calculation-only validation dataset

Validation cases were chosen by selecting the most recently treated cases at our institution. A set of 25 VMAT and 25 IMRT cases were selected. Two different error magnitudes for a selected error type were induced in each of the 25 VMAT and IMRT cases, resulting in 50 error-induced calculations for both IMRT and VMAT groups. As the training dataset from Chapter V consisted of 480 error induced plans for each delivery type, the validation cohort here was chosen to be approximately 10% the size of the training data set. This follows the general rule of thumb that validation datasets typically be 10-20% the size of the training dataset.

Instead of inducing all the errors described in Chapter III on each patient case, an in-house MATLAB script was created to randomly select from one of the following error types for each case:

- Whole plan MU errors
- One beam/arc MU error
- MLC perturbational error
- MLC lagging leaf error
- MLC bank shift error
- Collimator rotation error
- Gantry rotation error (for IMRT cases only)
- Random combination of two of any of the above errors

The MATLAB code assigned an index to each error type, and then used the random integer value function in MATLAB to select an error type for a given case. If the random combination error was chosen, the selection was repeated twice and selected one error type and removed this error type from the second iteration when selecting the second induced error. After the error type was selected, three error magnitudes were also randomly selected. MU errors ranged in magnitude from -15% to 15%, MLC perturbational errors from 0 to ±2 cm, lagging leaf from 0 to 4 cm, bank shift errors from -0.3 to +0.3 cm., collimator errors from 0 to 25 degrees, and gantry angle errors from 0 to 10 degrees. Three error magnitudes were selected with the expectation of obtaining a range of different error magnitudes, and ultimately choosing the smallest and largest of these three errors to induce in each patient plan on each detector geometry. The purpose of this was to have both a large magnitude and a small magnitude error for each case so as to span a range of error magnitudes.

Errors were induced using the same MATLAB codes in Chapter III and error-induced DICOM RT plan files were re-imported back into Eclipse for dose calculation on both the patient geometry and the detector phantom geometries. Dose calculations were performed using an Eclipse ESAPI script to run batch dose calculations and dose matrices were extracted for analysis in MATLAB using CERR. 1mm error-free calculations were compared to error-induced calculations at the real spatial sampling of each detector device using gamma criteria of 3%/3mm TH 10 (global) and 2%/2mm TH 10 (global). These gamma comparisons were performed using the in-house MATLAB code previously described in Chapter III. The GDSA method was also performed between error-free and error-induced calculations at the real spatial sampling for each device. The mean local percent dose difference in the high-dose low-gradient regions of the field, using the specified thresholds in Table 6-1, was calculated for each case. As in Chapter V, this result was evaluated for its ability to predict change in the PTV mean dose from the patient DVH by plotting the predicted change from the GDSA method versus the actual change in PTV mean from the TPS DVH.

Table 6-1. Optimal dose and gradient thresholds for each detector geometry and delivery technique.

	Delta 4		**MapCHECK**		**ArcCHECK**	
	IMRT	VMAT	IMRT	VMAT	IMRT	VMAT
Dose Threshold [%]	70%	80%	60%	70%	40%	70%
Gradient Threshold [%/mm]	4%/mm	8%/mm	4%/mm	6%/mm	4%/mm	6%/mm

Additionally, ROC curves and AUC values were calculated for both the gamma comparison and the GDSA method to compare the sensitivity and specificity between GDSA and the gamma comparison. The ability of the GDSA to properly identify plans with a PTV mean dose difference greater than ±3% was evaluated and compared to the ability of 3%/3mm TH 10 (G) and 2%/2mm TH 10 (G) to fail a plan with a PTV mean dose difference greater than ±3%. ROC curves and AUC values were calculated in *R*.

True positives and negatives, and false positives and negatives were defined as

- True negative (TN): An error-induced plan has <3% change in PTV mean dose and gamma/GDSA correctly labels the plan as passing
- True positive (TP): An error-induced plan has ≥3% change in PTV mean dose and gamma/GDSA correctly labels the plan as failing
- False negative (FN): An error-induced plan has ≥3% change in PTV mean dose, but gamma/GDSA incorrectly labels the plan as passing
- False positive (FP): An error-induced plan has <3% change in PTV mean dose, but gamma/GDSA method incorrectly labels the plan as failing

VI.C. Results

VI.C.1. Comparisons with clinical MapCHECK measurements

Comparisons were made using real MapCHECK measurements from 20 consecutive cases measured on the same TrueBeam linear accelerator over the course of several weeks. These measurements were made on different days by different users. While this introduces more uncertainties into the measurements, it also is representative of how QA comparisons are carried out in a clinical setting. For several cases it was apparent that small geometrical shifts in the x- and y-directions were appropriate to correct for set-up uncertainties. Automatic calculation shifts were not used, as this feature in the MapCHECK software only looks for shifts based on improving the gamma passing rate. Instead, we evaluated dose difference maps for structure that hinted at set-up issues. This typically exhibits itself in a planar measurement as a band or line of hot and cold points adjacent to one another such as that seen Figure 6-1 (a) and (c). This was also further verified by comparing measurements taken on the same day with the same setup. An example of this is shown in Figure 6-1. MapCHECK cases 5 and 6 were taken on the same day, and both show similar banding of hot and cold in the same direction. After applying small shifts of 1mm in both the x- and y-directions, the dose difference maps look slightly better.

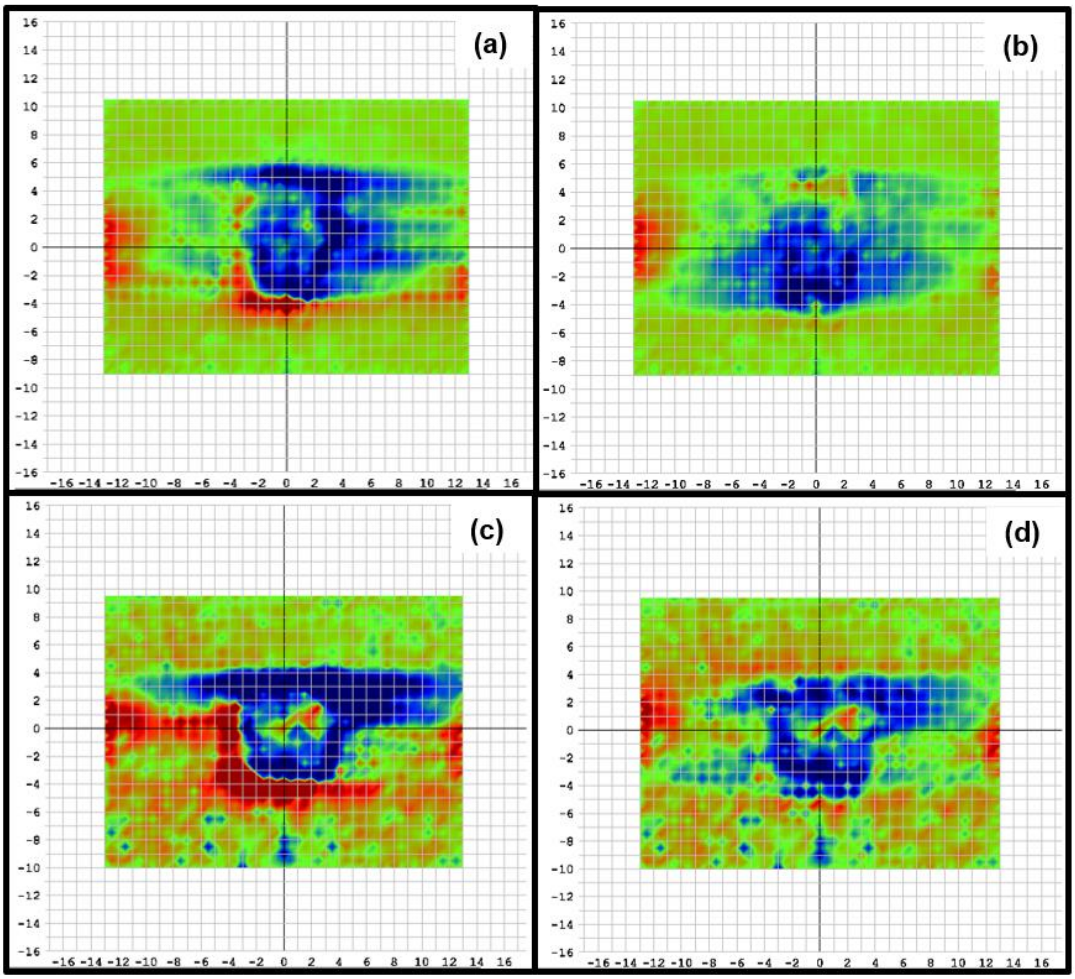

Figure 6-1. Dose difference maps shown for *(a)* MapCHECK Case 5 without any shifts, *(b)* MapCHECK Case 5 with x=1mm and y=1mm shifts applied, *(c)* MapCHECK Case 6 without any shifts, and *(d)* MapCHECK Case 6 with x=1mm and y=1mm shifts applied. In both case measurements before shifting, lines of hot and cold dose differences are shown that are remedied by this small shift. Since this is present in both measurements, it is likely a set-up error and not a delivery error.

Results for GDSA and the gamma comparison in Table 6-2 and Figure 6-3 show that no apparent relationship exists between GDSA and gamma passing rates for the two gamma criteria studied (2%/2mm TH 10 global, and 3%/3mm TH 10 global). For this case cohort, the results for GDSA predict that the change in PTV mean for all the studied cases is less than 3% in magnitude. However, if a 2% dose threshold were imposed as the GDSA pass/fail criterion for the cases shown in Table 6-2, nearly half of the cases would fail IMRT QA. Interestingly, the results all showed that the measurements were always slightly lower than the calculations, and it appears that there may be a systematic dose difference of approximately 1.5%. This may be due to several factors such as Eclipse dose calculation limitations for the phantom geometry or MapCHECK

calibration limitations for VMAT plans. However, this illustrates that the increased sensitivity of GDSA may also allow easier identification of systematic issues in the QA process that could be improved.

Table 6-2. Results for 20 real MapCHECK measurements compared to calculations from Eclipse TPS. Positive differences for change in PTV mean dose represent where the calculation is hotter than the measurement, whereas negative PTV mean dose differences indicate that the measurement is hotter than the calculation.

Case Name	Treatment Site	2%/2mm TH 10 (G)	3%/3mm TH 10 (G)	GDSA Predicted Δ in PTV Mean dose [%]
MpCk 1	Pelvic LNs	88 %	97.8 %	1.93 %
MpCk 2	Prostate	92.6 %	96.2 %	2.33 %
MpCk 3	Prostate	85.5 %	93.4 %	1.96 %
MpCk 4	Thigh	86.6 %	94 %	0.91 %
MpCk 5	Prostate	90.2 %	95.9 %	2.20 %
MpCk 6	Prostate	92.1 %	97.8 %	1.48 %
MpCk 7	Prostate + LNs	89.7 %	96.3 %	2.17 %
MpCk 8	Pelvis	88.6 %	96.6 %	1.45 %
MpCk 9	Pelvis	89.6 %	96.9 %	2.05 %
MpCk 10	Prostate	92.6 %	98.7 %	1.47 %
MpCk 11	Prostate + LNs	92.2 %	98.2 %	0.53 %
MpCk 12	Breast	73.4 %	90.4 %	1.25 %
MpCk 13	Prostate	83.8 %	95.5 %	2.32 %
MpCk 14	Anus/LN	88.7 %	96.8 %	0.22 %
MpCk 15	Prostate SBRT	87.9 %	97.2 %	1.04 %
MpCk 16	Pelvis	88.1 %	97.1 %	1.06 %
MpCk 17	Prostate	90.2 %	96 %	1.50 %
MpCk 18	Prostate	90.1 %	97.5 %	2.12 %
MpCk 19	Pelvis	89.7 %	97.5 %	0.56 %
MpCk 20	Rectum	94 %	98.7 %	2.04 %

In Table 6-2 the case with the lowest gamma passing rate, MpCk 12, did not show large dose differences with the GDSA metric. When investigating where the points failed the gamma comparison, shown in Figure 6-2, we see that the gamma comparison fails many points both inside and outside the high dose regions of the measurement. The points failing gamma where the measurement is colder than the calculation, shown in blue, mostly fall within the PTV volume,

suggesting that the dose in the center of the plan is low, which is similar to GDSA results predicting the measurement has a 1.25% lower PTV mean dose than the calculation.

Conversely, points failing the gamma comparison outside the high dose region all appear hotter in the measurement. This can likely be explained by two factors – when a high percentage of fluence enters through lateral beam angles on the MapCHECK device, the diodes are known to overrespond[15–17]. Additionally, for cases of larger field size, poor agreement on the MapCHECK has also been reported, particularly for 6 MV VMAT plans[17]. Here, the gamma comparison would result in a case barely passing IMRT QA since the gamma comparison does not distinguish between comparison points that fall in/outside the PTV volume. Thus, gamma comparisons for cases like these could result in near failures or increased failures in the clinic. Conversely, the results for GDSA for this case predict that changes in PTV mean dose between measurement and calculation were less than 2% and it is likely that this measurement suffers from multiple MapCHECK limitations added together. This suggests that GDSA has the potential to reduce false positives in the clinical setting. Indeed, reduced rates of false positives were also observed for GDSA compared to gamma comparisons in calculation-only comparisons performed in Chapter V. These results further support the argument that the percent pixels passing metric alone is not robust enough for passing or failing an IMRT QA result, which is also supported by current TG-218 recommendations[18].

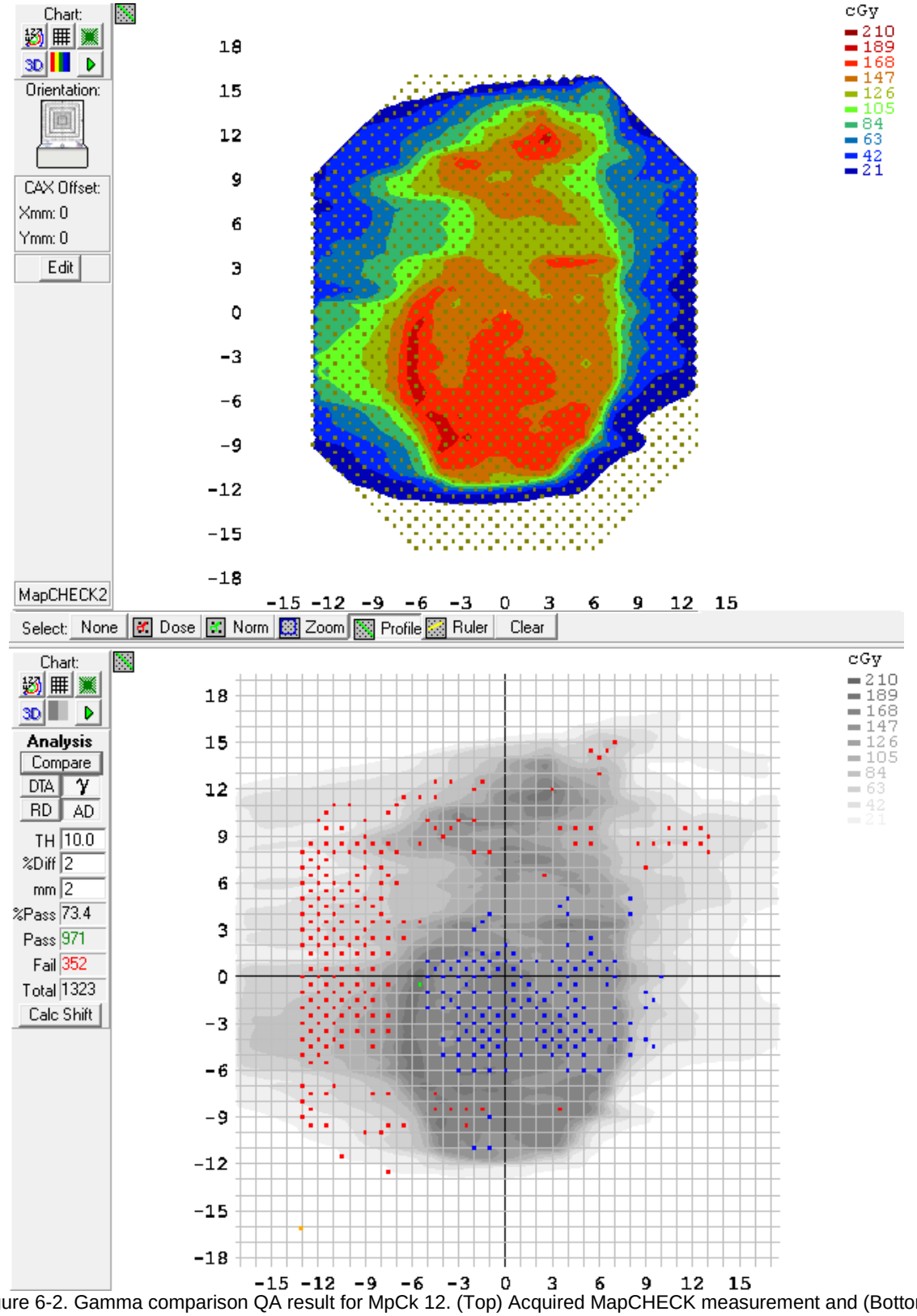

Figure 6-2. Gamma comparison QA result for MpCk 12. (Top) Acquired MapCHECK measurement and (Bottom) points failing the gamma comparison at 2%/2mm TH 10 (global, uncertainty off).

It is also of interest to note that a low gamma passing rate does not always mean that GDSA predicts a large change in PTV mean dose. Conversely, a high gamma passing rate does not correlate with smaller changes in PTV mean predicted by GDSA. This is illustrated in Figure 6-3 showing that no apparent correlation exists between the GDSA-predicted change in PTV mean dose and the gamma passing rate for 3%/3mm TH 10 (G) and 2%/2mm TH 10 (G).

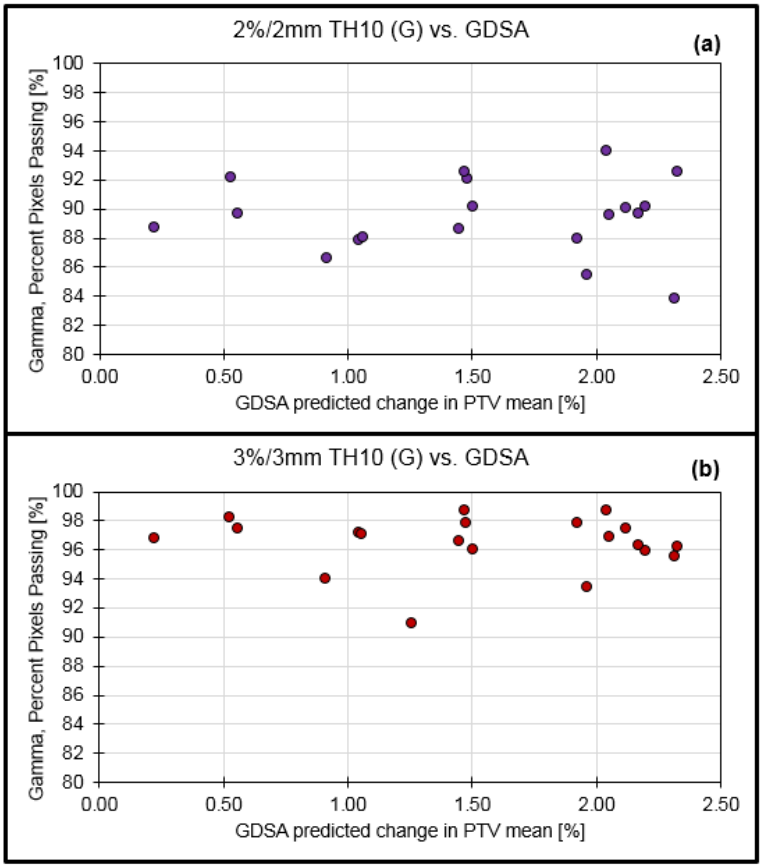

Figure 6-3. The results from the gamma comparison for (a) 2%/2mm TH 10 (global) and (b) 3%/3mm TH 10 (global) are plotted against the results from the GDSA method that predicts the change in PTV mean dose.

VI.C.2. Comparisons with error-induced plan deliveries

Errors were also introduced into the actual delivery for nine of the cases from Table 6-2. The purpose of this test is to evaluate if the GDSA predictions between the "error-free" and error-induced plans can accurately predict the change in PTV mean shown in the TPS DVH (ground truth). Additionally, changes in gamma passing rates between error-free and error-induced plans

are shown. While gamma passing rates for error-free plans did not reach 100%, any error-induced plans with a change in passing rate >10% would most certainly cause a failure in the clinic (i.e. a passing rate below 90%).

The results for gamma comparisons and GDSA comparisons are shown in Table 6-3 for plans without intentional errors and plan with intentional errors. The true change in PTV mean dose between the error-free plans and error-induced plans are also shown for reference in Table 6-3. For some cases, the error-induced plans resulted in slightly improved gamma passing rates or smaller GDSA predictions for the change in PTV mean dose, which is most noticeable for MpCk 12. This same case had a higher passing rate for the clinical measurements made a year prior, but the pattern of gamma failures was similar to that shown in Figure 6-2, with a large number of points falling outside the PTV that failed hot, but with simply more points in this outer region of the measurement failing the gamma comparison for the new measurement. While the gamma passing rate for this plan is low, the GDSA predicted PTV mean dose difference for MpCk 12 is only 1.5%, and thus a pass/fail threshold of even 2% would allow this plan to pass clinically. Additionally, if a clinic applied a pass/fail threshold of 2% change in PTV mean dose for the GDSA method, two of the plans without intentional errors in Table 6-3 would fail IMRT QA (MpCk 3 and MpCk 6) with GDSA but not with the 3%/3mm TH 10 (global) gamma criterion. Since these real measurements lack a ground truth, these cases could indeed have PTV mean dose differences greater than 2%, but this would not be flagged in clinical gamma comparisons.

The relative differences between the results from Table 6-3 are shown in Table 6-4 and are visualized in Figure 6-4. As a point of interest, while error-free plans never achieved a gamma passing rate of 100%, the change in gamma passing rate between error-free and error-induced plans was always >10% for those cases that had true changes in the PTV mean dose >3%. That is, the error-induced plans would cause a failure with the gamma comparison in the clinic (i.e. trigger a passing rate below 90%). Most importantly, Table 6-4 shows that that for all but one

plan (MpCk 1), the change in GDSA predictions between error-free and error-induced comparisons match very closely to the changes in PTV mean dose predicted by the TPS DVHs between error-free and error-induced plans, illustrating that GDSA works well not only in calculation-only scenarios, but for real MapCHECK measurements. For eight of the nine studied cases, the relative difference predicted by GDSA between the error-free and error-induced plans was less than 1% different from the ground truth.

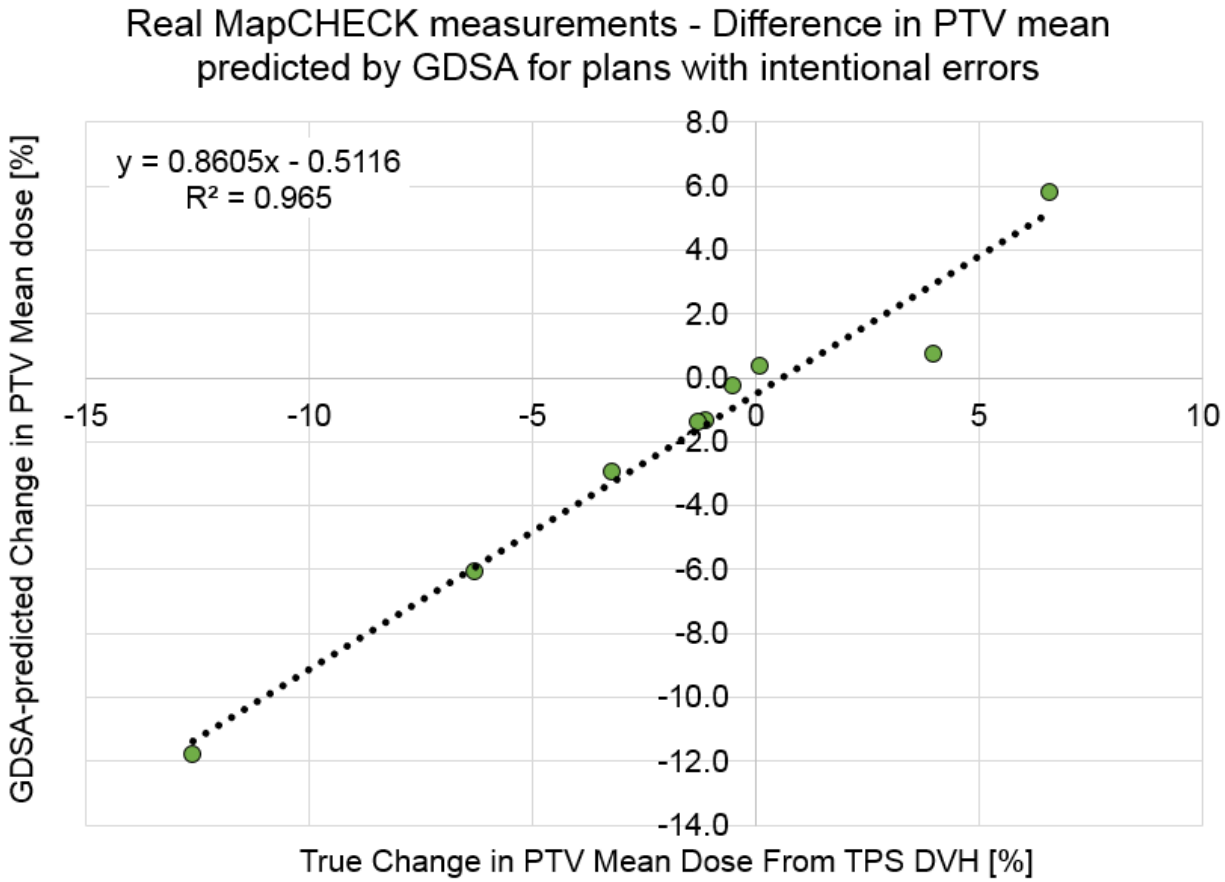

Figure 6-4. The GDSA-predicted change in PTV mean between the error-free and error-induced plans versus the true change in PTV mean dose from the TPS DVHs between the error-free and error-induced plans.

Table 6-3. The gamma passing rates and GDSA predictions are shown for nine cases that were delivered with and without intentional errors.

				Without intentional errors			With intentional errors		
	Induced error	Error magnitude	True change in PTV mean [%] between error-free and error-induced plan	2%/2mm TH 10 (G) passing rate	3%/3mm TH 10 (G) passing rate	GDSA predicted change in PTV mean dose [%]	2%/2mm TH 10 (G) passing rate	3%/3mm TH 10 (G) passing rate	GDSA predicted change in PTV mean dose [%]
MpCk 1	Collimator	10°	4 %	78.5 %	93.8 %	-1.3 %	38.7 %	54.6 %	-0.6 %
MpCk 2	MU whole plan	6%	-6.3 %	93.6 %	98.1 %	-0.8 %	71 %	81.5 %	-6.8 %
MpCk 3	Perturb	±2 cm	6.6 %	83.8 %	94.5 %	2.5 %	58.1 %	75.9 %	8.3 %
MpCk 4	Bank	+0.2 cm	-1.1 %	90 %	95.6 %	0.5 %	85.7 %	93.9 %	-0.8 %
MpCk 5	Lag	2 cm	-3.2 %	85.6 %	94.4 %	-0.6 %	71 %	85 %	-3.6 %
MpCk 6	Collimator	5°	0.1 %	88.9 %	97.6 %	2.0 %	80.7 %	93.7 %	2.4 %
MpCk 12	MU, one beam	3%	-1.3 %	61 %	78.8 %	1.5 %	61.1 %	79.9 %	0.1 %
MpCk 14	Bank	+1 cm	-12.6 %	74.6 %	90.9 %	0.7 %	5.1 %	14.6 %	-11.1 %
MpCk 16	Lag	1 cm	-0.5 %	82.8 %	94.6 %	1.2 %	82.5 %	95 %	1.0 %

Table 6-4. For nine real deliveries, plans without intentional errors were delivered to the MapCHECK. GDSA and gamma comparison results were calculated. Plans were then delivered with the listed induced error and GDSA and gamma comparisons were calculated. Differences between these results between error-free and plan deliveries with intentional errors are shown. The true change in PTV mean dose from the TPS DVH is also shown as the ground truth for the change in GDSA results.

	Induced error	**Error magnitude**	**True change in PTV mean [%] between error-free and error-induced plan**	**Change in GDSA result between error-free and error-induced plan**	**Change in 3%/3mm TH 10 (G) passing rate between error-free and error-induced plan**	**Change in 2%/2mm TH 10 (G) passing rate between error-free and error-induced plan**
MpCk 1	Collimator	10°	4 %	0.7 %	39.2 %	39.8 %
MpCk 2	MU whole plan	6%	-6.3 %	-6.0 %	16.6 %	22.6 %
MpCk 3	Perturb	±2 cm	6.6 %	5.8 %	18.6 %	25.7 %
MpCk 4	Bank	+0.2 cm	-1.1 %	-1.3 %	1.7 %	4.3 %
MpCk 5	Lag	2 cm	-3.2 %	-3.0 %	9.4 %	14.6 %
MpCk 6	Collimator	5°	0.1 %	0.4 %	3.9 %	8.2 %
MpCk 12	MU, one beam	3%	-1.3 %	-1.4 %	-1.1 %	-0.1 %
MpCk 14	Bank	+1 cm	-12.6 %	-11.8 %	76.3 %	69.5 %
MpCk 16	Lag	1 cm	-0.5 %	-0.2 %	-0.4 %	0.3 %

VI.C.3. Calculation-only validation cases

The breakdown of cases by treatment site that were randomly chosen for the IMRT and VMAT cohorts are shown in Figure 6-5 and the frequency of the different randomly chosen errors that were induced in the IMRT and VMAT case cohorts is shown in Figure 6-6. For these cases and induced errors, the gradient segmented method was performed, and the results from that method were compared to the true change in the PTV mean from the TPS DVH. Results in Figure 6-7 show the predicted change in PTV mean dose from the GDSA method compared to the true change in PTV mean dose from the TPS DVH for each detector geometry and delivery type. These results are markedly similar to those seen in Chapter V with the training dataset and suggests that the thresholds chosen in the training portion of this work appear to work well for an entirely different set of cases. As in Chapter V, the GDSA method is able to best predict changes in PTV mean dose for the MapCHECK and Delta 4 geometries. While GDSA does not work as well in the ArcCHECK geometry, correlations are still moderate to strong. Here, the ArcCHECK VMAT results show stronger Pearson correlations than seen in the training dataset, however, this may be simply due to the fewer number of evaluated cases in the validation dataset compared to the training dataset.

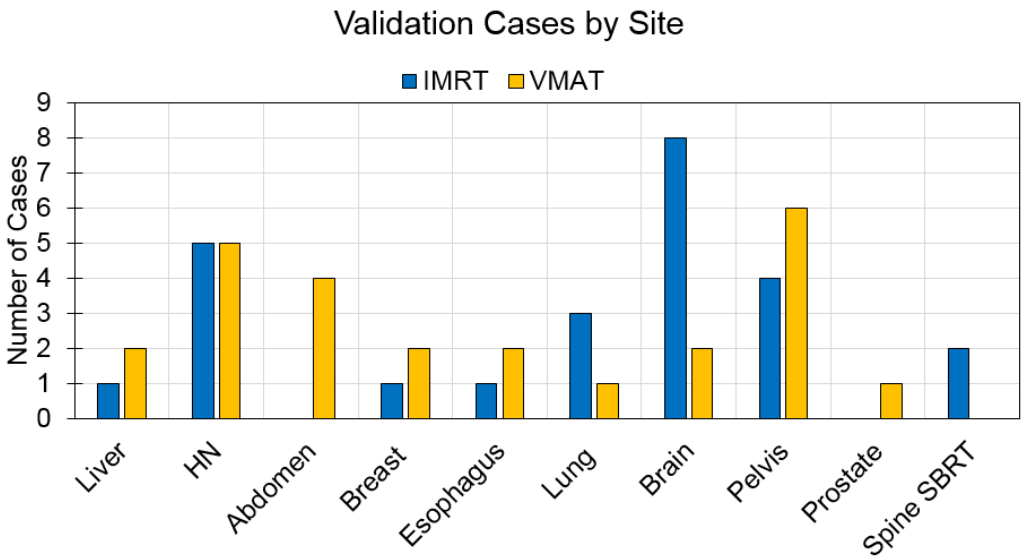

Figure 6-5. Treatment site breakdown for the validation cases for IMRT (n=25) and VMAT (n=25) cohorts.

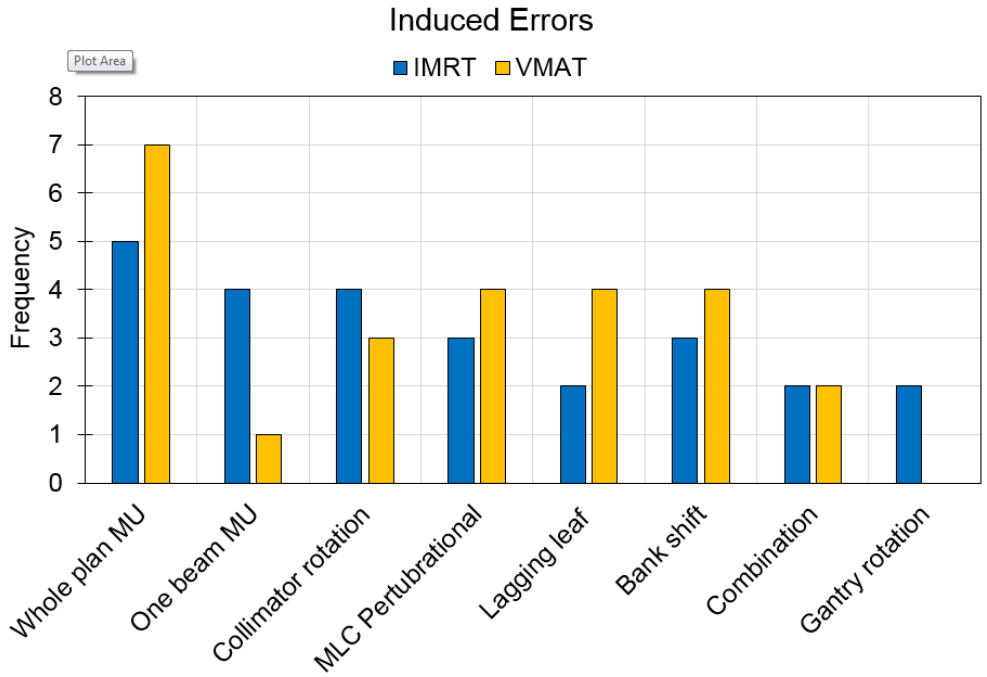

Figure 6-6. The frequency of different induced errors (chosen randomly) for the 25 IMRT and 25 VMAT cases

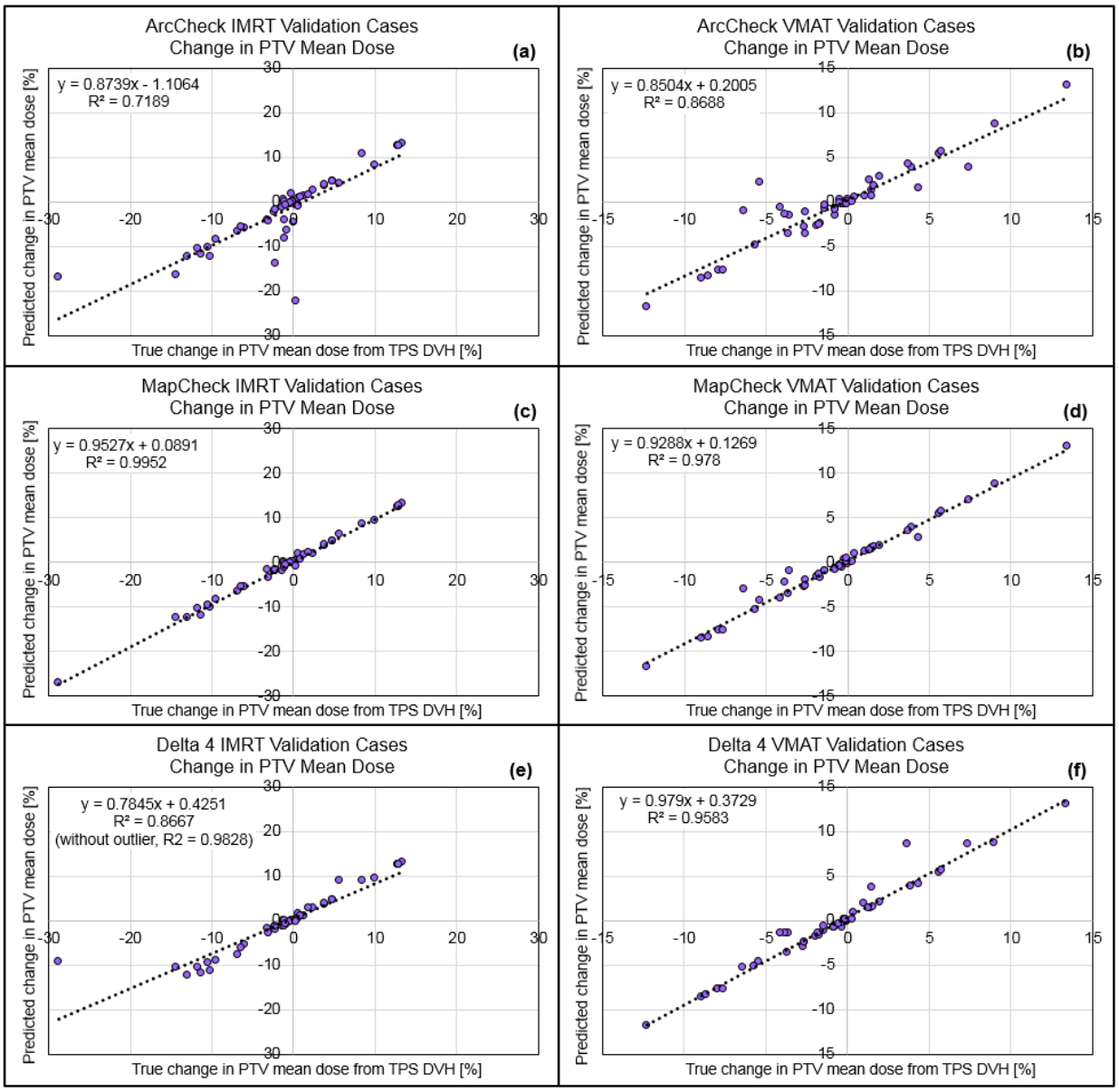

Figure 6-7. Correlations between predicted change in PTV mean dose from the high-dose low-gradient points compared to the true change in the PTV mean in the patient geometry from the TPS DVHs for both IMRT cases (left) and VMAT cases (right) for each studied device geometry – ArcCHECK (a-b), MapCHECK (c-d), and Delta 4 (e-f).

ROC analyses were carried out in *R* and the ground truth was considered a true change >3% in PTV mean dose as the "gold standard" in this case to illustrate the sensitivity and specificity differences between two commonly used gamma criteria and the ability of the high-dose low-gradient points to flag plans that have a PTV mean dose difference greater than ±3%. Evaluation of the gamma criteria here also includes comparison to a true change in PTV mean, meaning that

we are testing the ability of the listed gamma criteria to flag a plan as failing that has greater than a 3% change in PTV mean.

Due to the smaller number of induced errors for each case, the ROC curves and respective AUC values are calculated by combining the results from VMAT and IMRT cohorts for each device. ROC curves are shown in Figure 6-8 and the corresponding AUC values in Table 6-5. Similar to the data in Chapter V, the ROC curves show that GDSA has higher sensitivity and specificity for all three devices, with MapCHECK and Delta 4 achieving similar AUC values for GDSA. Of note, it also appears that the sensitivity and specificity of the gamma comparison is higher for Delta 4 than for MapCHECK, which is consistent with data from Chapter V. As before, while ArcCHECK AUC values for GDSA are noticeably lower compared to MapCHECK and Delta 4, GDSA is still more accurate as a binary test when compared to 3%/3mm TH 10 (G) and 2%/2mm TH 10 (G) gamma criteria.

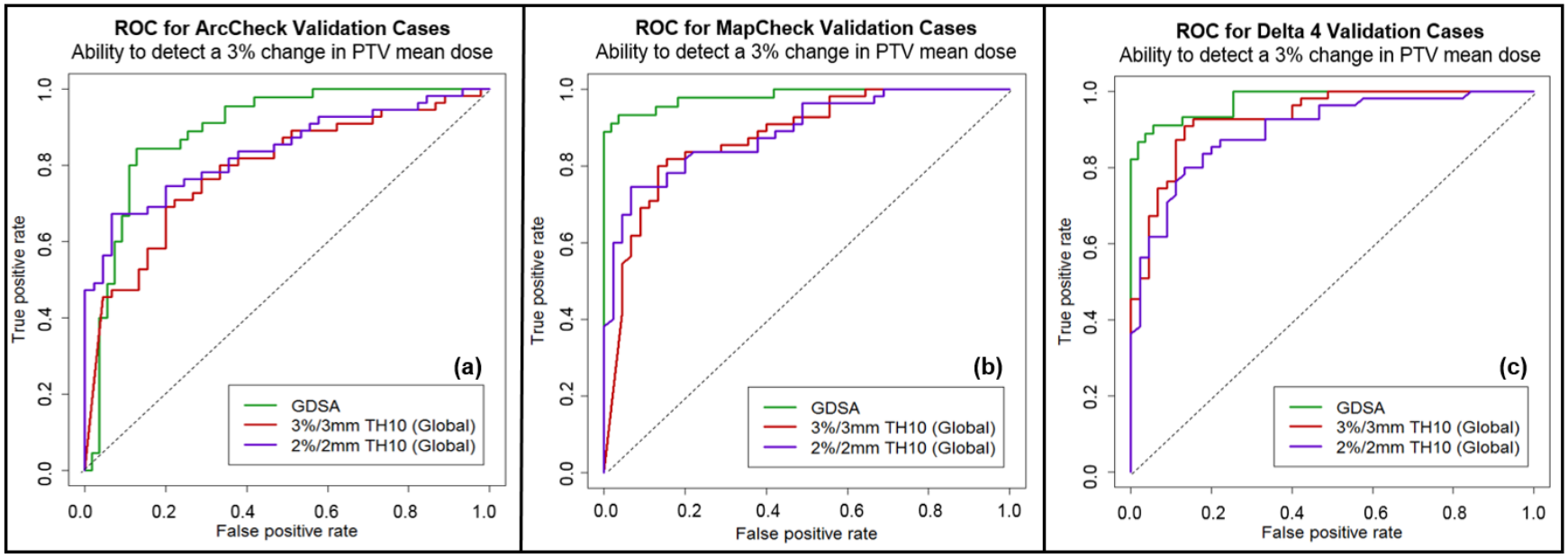

Figure 6-8. ROC curves showing the ability of the new method (i.e. using high-dose low gradient points to predict changes in PTV mean) compared to the listed gamma criteria to flag a plan as failing in the presence of a PTV mean dose difference greater than ±3%. Results shown for (a) ArcCHECK cases (b) MapCHECK cases, and (c) Delta 4 cases.

Table 6-5. Area under the ROC curve (AUC) values for the three devices are shown for the new method and the two listed gamma criteria for the validation cases. VMAT and IMRT results are combined for each device

	ArcCHECK	MapCHECK	Delta 4
New method	0.893	0.983	0.977
3%/3mm TH 10 (G)	0.790	0.877	0.934
2%/2mm TH 10 (G)	0.838	0.889	0.900

VI.D. Discussion

This chapter has validated the gradient segmented analysis (GDSA) method with a separate set of 25 IMRT and 25 VMAT cases recently treated at our institution. The results from the validation data set showed very similar results to those seen in Chapter V with the initial training data set, thus confirming that this method works well to predict true changes in PTV mean dose, especially for the MapCHECK and Delta 4 geometries. Both the data in Chapter V and in the validation data set made use of calculation-only comparisons to not only remove uncertainties present in real measurements, but to allow the calculation of DVH dose differences as the ground truth.

The GDSA method was also evaluated for 20 real measurements made on the MapCHECK device. Dose differences in PTV mean between measurement and calculation predicted by GDSA for the 20 MapCHECK cases were all less than 3%. All studied cases also passed clinical IMRT QA with >90% pixels passing at 3%/3mm TH 10 (G), uncertainty corrections off. It is important to note that any limitations in beam modeling, device calibration, machine output variations, and delivery limitations are folded into these comparisons with real measurements. No correlations appeared to exist between gamma passing rates and GDSA predicted changes in PTV mean dose.

Additionally, for these 20 measurements, the GDSA results showed that the measurements had systematically lower doses of approximately 1-2% compared to the calculations, which is not information directly reported in gamma comparison results. While issues such as this could also be flagged by inspecting dose difference maps and profile comparisons, there is no simple metric that can be used to track this for a large number of cases when using the gamma comparison. Instead, the differences predicted by the gradient-dose segmented analysis method could easily be stored for each case and reviewed regularly to watch for concerning trends.

Finally, the ability of GDSA to predict changes in PTV mean doses between delivered plans without intentional errors and delivered plans with intentional errors was evaluated. For eight of

the nine cases delivered to the MapCHECK, the relative change in PTV mean predicted by GDSA was within 1% of the ground truth (from the TPS DVH). This further confirms that GDSA works well not only in calculation-only comparisons, but in real measurement scenarios as well.

When planning to implement a new technique for IMRT QA comparisons in the clinic, one concern is if the new technique will increase the number of false positives in clinical IMRT QA comparisons, which could ultimately result in an increase in the physics workload. A method that fails too many plans when there is not actually a concerning dose discrepancy could lead to a large amount of time spent investigating, measuring, or even re-planning cases. However, it is equally important that the new comparison method does not pass too many plans that do indeed have considerable dose differences. For the 20 real MapCHECK measurements we found that the new GDSA method, using a ±3% change in PTV mean dose as a pass/fail metric, would not have failed any case that passed IMRT QA with the standard 3%/3mm TH 10 (global) 90% pixels passing criterion. We also evaluated the number of false negatives and false positives for the GDSA method using the calculation-only comparison data from Chapter V. For the MapCHECK and Delta 4 geometries, assuming a pass/fail threshold of ±3% change in PTV mean dose, the number of false positives and more importantly, false negatives are very low which is important to consider when implementing this method clinically. False positive rates and false negative rates from the training dataset in Chapter V (*n*=480) are shown for GDSA and three different gamma criteria for IMRT cases and VMAT cases in Table 6-6 and Table 6-7, respectively.

Table 6-6. False negative and false positive rates for IMRT cases on each device are shown for the gradient segmented method and three different gamma criteria given a threshold of ±3% for change in PTV mean dose.

IMRT Cases						
	False Negative Rate (FN / 480)			False Positive Rate (FP / 480)		
	ArcCHECK	MapCHECK	Delta 4	ArcCHECK	MapCHECK	Delta 4
GDSA	4.8 %	1.3 %	2.5 %	9.2 %	4.4 %	2.5 %
3%/3mm TH 10 (G) 95%	7.9 %	1.9 %	7.1 %	10.6 %	17.1 %	10.2 %
3%/3mm TH 10 (L) 95%	0.3 %	1.9 %	2.1 %	23.3 %	17.1 %	17.9 %
2%/2mm TH 10 (G) 95%	0.8 %	1.5 %	1.0 %	24.8 %	17.9 %	24.0 %

Table 6-7. False negative and false positive rates for VMAT cases on each device are shown for the gradient segmented method and three different gamma criteria given a threshold of ±3% for change in PTV mean dose.

VMAT Cases						
	False Negative Rate (FN / 480)			False Positive Rate (FP / 480)		
	ArcCHECK	MapCHECK	Delta 4	ArcCHECK	MapCHECK	Delta 4
GDSA	3.5 %	3.1 %	2.9 %	5.4 %	1.3 %	1.5 %
3%/3mm TH 10 (G) 95%	14.0 %	5.6 %	3. 8 %	11.0 %	12.5 %	15.2 %
3%/3mm TH 10 (L) 95%	1.5 %	1.7 %	0.4 %	19.4 %	20.6 %	27.1 %
2%/2mm TH 10 (G) 95%	3.5 %	0.2 %	0.0 %	21.3 %	23.8 %	28.3 %

Most importantly, GDSA has fewer false negatives and false positives compared to the most commonly used criterion of 3%/3mm TH 10 (global). As gamma criteria are tightened, the number of false negatives decreases, but this also causes a rise in false positives which often precludes the use of stricter gamma criteria in many clinics. This is where the GDSA method appears to have an advantage in that the number of false positives always appears lower than that of the three listed gamma criteria. In general, the rate of false positives and false negatives for GDSA are all under 5% for MapCHECK and Delta 4 geometries across both delivery types. This shows that if this method were to be implemented clinically, it is not likely to have a noticeable impact on the clinical physics workflow, which is highly desirable when introducing a new technique into the clinic. However, as clinics investigate the use of GDSA for IMRT QA comparisons, it is likely prudent to use both the gamma comparison technique and GDSA in tandem over an extended period of time.

In developing a new technique for evaluating IMRT QA comparisons, this new method would ideally,

i. Offer more clinically meaningful results than current methods
ii. Increase sensitivity and specificity to errors compared to current methods
iii. Not increase clinical physics workload, either from the time spent analyzing the data or by increasing the number of false positives that would require investigation/re-planning

iv. Be simple enough to easily implement in other clinics or in vendor software

While we have met these benchmarks, it is our hope that this work is only the beginning of moving toward more meaningful comparisons in patient-specific IMRT QA and has presented one new way of beginning to evaluate these complex dose distributions. Further validation of this method, particularly third-party and vendor validation of appropriate dose and gradient thresholds is appropriate moving forward, as well as extending this analysis to other measurement scenarios, such as portal dosimetry IMRT QA comparisons. However, this method is simple enough to be easily replicated by other clinics or in vendor software. This method also only requires one additional piece of information from the end-user, the dose per fraction, which is used for normalizing the gradient maps. The other information required for GDSA, the 3D calculated dose and the measured dose, is already required to perform gamma comparisons, and thus implementing this method should not require any foreseeable increase in time spent analyzing the data.

One limitation of this method is its relatively poorer performance in ArcCHECK measurements. While it still performed better than the gamma comparison, the increases in sensitivity and specificity, as well as the ability to properly predict the true changes in PTV mean dose, were not as straightforward as they were in the Delta 4 and MapCHECK geometries. Additionally, it appears the gamma comparison in the ArcCHECK is less sensitive and specific than in the planar-type geometries, suggesting that a separate type of analysis may be required for this unique measurement geometry. Another limitation here is the difficulty in validating these results in real patient measurements, due to the lack of a ground truth in all IMRT QA measurements.

VI.E. Conclusions

The work in this chapter has validated the gradient-dose segmented (GDSA) method for clinical use. In evaluating real MapCHECK measurements it was found that the GDSA method is unlikely to cause an increase in clinical IMRT QA failures when using a 2-3% threshold for change in PTV

mean as the pass/fail metric. Not surprisingly, gamma passing rates for 2%/2mm TH 10 (G) and 3%/3mm TH 10 (G) did not correlate with GDSA results. Errors were also introduced in a subset of real plan deliveries and measured with the MapCHECK to evaluate the ability of GDSA to predict the relative changes in PTV mean dose between error-induced plans and error-free plans. For all but one case, GDSA predicted the true change in PTV mean to within 1% of the true change in PTV mean predicted by the TPS DVHs.

Finally, the GDSA method using the dose and gradient thresholds selected in Chapter V was evaluated for a validation cohort of 25 IMRT and 25 VMAT cases, different from those utilized in Chapter V. Calculation-only comparisons for all three detector geometries (ArcCHECK, MapCHECK, and Delta 4) showed markedly similar results to those obtained in Chapter V, which further confirms the dose and gradient thresholds selected in Chapter V for each device and delivery technique to be appropriate. While GDSA performs relatively worse for the ArcCHECK geometry compared to planar-type arrays, GDSA offers slightly better accuracy as a binary pass/fail metric when compared to gamma for the ArcCHECK. For MapCHECK and Delta 4 geometries, the correlations between predicted change in PTV mean (from the GDSA method) compared to the true change in PTV mean doses (from the TPS DVHs) closely matched the results presented in Chapter V. As before, correlations for these geometries exhibited a clear 1:1 relationship between predicted and true change in PTV mean dose, with Pearson correlation coefficients exceeding 0.9 for Delta 4 and MapCHECK geometries. Additionally, ROC analyses showed similar AUC values between the training dataset in Chapter V and the validation dataset in this chapter when evaluating GDSA as a binary pass/fail metric

We have shown that the GDSA method works well for a large range of case types and range of induced errors of varying magnitude. This method offers more clinically meaningful results with improved sensitivity to errors compared to current IMRT QA comparisons performed with the gamma comparison.

References

1. Nelms B, Jarry G, Chan M, Hampton C, Watanabe Y, Feygelman V. Real-world examples of sensitivity failures of the 3%/3mm pass rate metric and published action levels when used in IMRT/VMAT system commissioning. *J Phys Conf Ser*. 2013;444:012086. doi:10.1088/1742-6596/444/1/012086.

2. Nelms BE, Zhen H, Tomé W a. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. *Med Phys*. 2011;38(2):1037-1044. doi:10.1118/1.3544657.

3. Heilemann G, Poppe B, Laub W. On the sensitivity of common gamma-index evaluation methods to MLC misalignments in Rapidarc quality assurance. *Med Phys*. 2013;40(3):031702. doi:10.1118/1.4789580.

4. Templeton AK, Chu JCH, Turian J V. The sensitivity of ArcCHECK-based gamma analysis to manufactured errors in helical tomotherapy radiation delivery. *J Appl Clin Med Phys*. 2015;16(1):4814. doi:10.1120/jacmp.v16i1.4814.

5. Jin X, Yan H, Han C, Zhou Y, Yi J, Xie C. Correlation between gamma index passing rate and clinical dosimetric difference for pre-treatment 2D and 3D volumetric modulated arc therapy dosimetric verification. *Br J Radiol*. 2015;88(1047). doi:10.1259/bjr.20140577.

6. Kry SF, Molineu A, Kerns JR, et al. Institutional patient-specific IMRT QA does not predict unacceptable plan delivery. *Int J Radiat Oncol Biol Phys*. 2014;90(5):1195-1201. doi:10.1016/j.ijrobp.2014.08.334.

7. McKenzie EM, Balter P a, Stingo FC, Jones J, Followill DS, Kry SF. Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans. *Med Phys*. 2014;41(12):121702. doi:10.1118/1.4899177.

8. Steers JM, Fraass B a. IMRT QA: Selecting gamma criteria based on error detection sensitivity. *Med Phys*. 2016;43(4):1982-1994. doi:10.1118/1.4943953.

9. Yan G, Liu C, Simon TA, Peng LC, Fox C, Li JG. On the sensitivity of patient-specific IMRT QA to MLC positioning errors. *J Appl Clin Med Phys*. 2009;10(1):120-128. doi:10.1118/1.2961902.

10. Zhen H, Nelms BE, Tomé W a. Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA. *Med Phys*. 2011;38(10):5477. doi:10.1118/1.3633904.

11. Bresciani S, Di Dia A, Maggio A, et al. Tomotherapy treatment plan quality assurance: the impact of applied criteria on passing rate in gamma index method. *Med Phys*. 2013;40(12):121711. doi:10.1118/1.4829515.

12. Chan. Using a Novel Dose QA Tool to Quantify the Impact of Systematic Errors Otherwise Undetected by Conventional QA Methods: Clinical Head and Neck Case Studies. *Technol Cancer Res Treat*. 2013;13(1):1-11. doi:10.7785/tcrt.2012.500353.

13. Sjölin M, Edmund JM. Incorrect dosimetric leaf separation in IMRT and VMAT treatment planning: Clinical impact and correlation with pretreatment quality assurance. *Phys Medica*. 2016;32(7):918-925. doi:10.1016/j.ejmp.2016.06.012.

14. Vieillevigne L, Molinier J, Brun T, Ferrand R. Gamma index comparison of three VMAT QA systems and evaluation of their sensitivity to delivery errors. *Phys Medica*. 2015;31(7):720-

725. doi:10.1016/j.ejmp.2015.05.016.

15. Jursinic PA, Sharma R, Reuter J. MapCHECK used for rotational IMRT measurements: Step-and-shoot, Tomotherapy, RapidArc. *Med Phys*. 2010;37(6):2837-2846. doi:10.1118/1.3431994.

16. Li QL, Deng XW, Chen LX, Huang SM. The angular dependence of a 2 - dimensional diode array and the feasibility of its application in verifying the composite dose distribution of intensity - modulated radiation therapy. *Chin J Cancer*. 2010:617-620.

17. Jin H, Keeling VP, Johnson DA, Ahmad S. Interplay effect of angular dependence and calibration field size of MapCHECK 2 on RapidArc quality assurance. *J Appl Clin Med Phys*. 2014;15(3):80-92. doi:10.1120/jacmp.v15i3.4638.

18. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. *Med Phys*. 2018;45(4):e53-e83. doi:10.1002/mp.12810.

Chapter VII.

Summary

Each year in the United States, millions of patients receive radiation therapy treatment for cancer. Radiation therapy utilizes focused beams of high-energy particles to induce cell death, with the goal of reducing or eliminating tumor burden. The delivery of this treatment requires a multitude of safety checks in the treatment planning and delivery process in order to ensure that patient treatments are delivered without deviations that could cause harm. Since plans are complex in nature, unexpected deviations can occur from a variety of sources ranging from limitations in calculation algorithms, machine delivery limitations, corrupted delivery files, to human error. In particular, when patients are treated with intensity modulated fields, one important pre-treatment safety check is to deliver the plan to an array of detectors and compare the measurement to a calculation on the same detector geometry in the treatment planning system (TPS). This process is called patient-specific quality assurance, or intensity modulated radiation therapy quality assurance (IMRT QA).

The most commonly used method for analyzing IMRT QA comparisons, the gamma comparison, has been called into question in recent years due to difficulty in interpreting the results and numerous publications illustrating that large errors can go undetected with commonly used gamma criteria. A wide variety of measurement devices with very distinct geometries and spatial sampling are in clinical use, which are likely to have unique effects on error sensitivity with the gamma comparison. Beyond this, different delivery techniques may also have differing error sensitivities for each measurement device. The independent effects of these factors on the sensitivity for current IMRT QA comparisons and how to develop more sensitive and clinically meaningful metrics were the focus of this work.

Chapter I introduced the range of detector geometries available for clinical use, as well as various analysis techniques for determining if a patient's plan is free of clinically significant errors. The most commonly used comparison technique, the gamma comparison, allows the user to quickly evaluate a QA comparison result based on the percentage of comparison points that pass or fail this gamma metric. While the gamma comparison remains the standard by which the majority of IMRT QA comparisons are evaluated, it is also well-known that this metric is 1) difficult to interpret 2) does not relate to clinically meaningful endpoints and 3) can be insensitive to large errors for certain gamma criteria. The end of Chapter I outlines current gaps in knowledge, how this work aims to elucidate the effects of various factors potentially affecting error sensitivity in IMRT QA, as well as methods to improve error sensitivity in IMRT QA.

In order to objectively quantify the sensitivity of many different gamma criteria, Chapter II presented the error curve method. The error curve method was utilized to analyze the error sensitivity of 36 different combinations of gamma criteria to induced errors of varying magnitude in ArcCHECK comparisons. This method allowed a quantitative understanding of gamma comparison sensitivity by reporting the range of errors passing a given gamma criterion. Results from this chapter showed that for ArcCHECK IMRT cases, higher dose thresholds increased error sensitivity. Additionally, these results showed that surprisingly large errors exceeding 10% in MU and ±1 cm random MLC errors could pass IMRT QA using the most common gamma criterion of 3%/3mm TH 10 (global).

Since a range of measurement devices are available for clinical IMRT QA, three devices were studied in all subsequent chapters to evaluate differences in sensitivity. The effects of detector geometry, detector spatial sampling, and delivery technique on gamma comparison error sensitivity have not been independently evaluated up to this point. Thus, the work in Chapter III separates these effects with the use of extensive in-house developed MATLAB gamma comparison code for three detector devices – ArcCHECK, MapCHECK, and Delta 4 – and a

carefully selected cohort of 20 IMRT and 20 VMAT cases, matched for similar field sizes and plan complexity. Gamma comparison error sensitivity for five different induced error types was evaluated with all devices at a spatial sampling of 1mm and also at the real spatial sampling of the device to separate the effects of detector geometry and detector spatial sampling. While particular detector geometries showed different sensitivities to different error types for 1mm vs. 1mm comparisons, decreasing the spatial sampling of each device did not noticeably reduce error sensitivity for any of the studied devices. Thus, it is likely that device geometry, and not differences in spatial sampling, is a more important factor in gamma comparison sensitivity. It was also found that locally normalized comparisons with a 10% low dose threshold offered similar error sensitivity between devices. However, for commonly used gamma criteria such as 3%/3mm TH 10 (global), the Delta 4 typically achieved the highest sensitivity to most induced errors. Finally, differences in VMAT and IMRT deliveries were evaluated, showing that for most of the induced errors studied, errors were more difficult to detect in VMAT deliveries than in IMRT deliveries. While current guidance documents have recommended the use of a universal gamma criterion for all IMRT QA comparisons, these results suggest that delivery- and detector-specific gamma criteria may be appropriate.

Chapter IV utilized a variety of manufactured test fields and patient plans to investigate patterns of gamma failures as increasingly larger errors were induced in these plans. Gamma failures appeared most readily in low-gradient regions of the field whereas regions of high gradients did not cause gamma failures for even very large magnitude errors. The results from this chapter suggested that comparison points falling along high gradients may be partly responsible for gamma comparison insensitivity. Additionally, the relationship between high gradients and low gamma values was observed regardless of gamma normalization technique, induced error type, detector geometry, and plan type. If many measurement points fall on these high dose gradients, the number of diodes in real IMRT QA measurements that may be sensitive to errors could be

alarmingly low for some case measurements. This leads us to believe that gamma comparisons are likely limited in part by the complexity of overlapping gradients in composite IMRT QA measurements.

Based on current limitations of gamma comparisons and the results from Chapter IV, Chapter V introduced a new method for comparing IMRT QA dose distributions. This new method, gradient-dose segmented analysis (GDSA), was developed considering the three detector geometries studied in previous chapters – ArcCHECK, MapCHECK, and Delta 4 – for both VMAT and IMRT delivery techniques. In developing this new method, comparisons were first segmented into regions of high-gradient, high-dose low-gradient, and low-dose low-gradient. Dose difference histogram statistics were evaluated for each of these regions and analyzed to determine if any dose difference histogram metrics could predict changes in 11 different DVH metrics. With the use of over 180,000 comparisons, device- and delivery-specific dose and gradient thresholds were selected for segmenting dose and gradient maps. Most notably, the mean local dose differences in high-dose low-gradient regions of the phantom comparisons predicted real changes in PTV mean in the patient geometry with a nearly 1:1 correlation. This new method, gradient-dose segmented analysis (GDSA), was not only able to yield more clinically meaningful results, but also showed increased accuracy as a binary metric when compared to five different gamma criteria. The predicted change in PTV mean dose versus the true change in PTV mean was predicted best in MapCHECK and Delta 4 geometries, with Pearson correlation coefficients typically exceeding 0.9. GDSA performed relatively poorer for the ArcCHECK geometry, but still offered an improvement in error sensitivity compared to gamma for this measurement geometry. These results, as well as results from previous chapters suggest that analysis of ArcCHECK data may require different considerations and perhaps modified methods.

Chapter VI evaluated GDSA results for real MapCHECK measurements and showed that GDSA would be unlikely to cause increased rates of IMRT QA plan failures in the clinic when using a 2%

or 3% change in PTV mean as a pass/fail threshold. A handful of cases would have failed QA if using a GDSA pass/fail metric of 2% change in PTV mean dose, whereas a looser threshold of 3% would not cause any failures of these plan cases. Likewise, the 3%/3mm TH 10 (global) gamma criterion would not have caused failure of any of these plans. Real MapCHECK plans were also delivered with intentional errors in Chapter VI, and the relative difference in PTV mean predicted by GDSA and the TPS DVHs (ground truth) was evaluated. For all but one case, the differences between predicted change in PTV mean between the error-free and error-induced real plan deliveries for GDSA compared to the TPS DVHs was less than 1%, showing the utility of GDSA for real measurements.

Since Chapter V utilized a cohort of 480 error-induced IMRT plans and 480 error-induced VMAT plans as a training dataset to select the best dose and gradient thresholds for segmenting comparisons, a smaller validation dataset was evaluated in Chapter VI consisting of 50 error-induced IMRT plans and 50 error-induced VMAT plans on the three studied devices. Results using the dose and gradient thresholds from Chapter V showed markedly similar results, suggesting that the GDSA method with the chosen dose and gradient thresholds works remarkably well for many different case types.

This dissertation has focused on several aspects of IMRT QA. First, we have presented a new method for objectively quantifying gamma comparison sensitivity, the error curve method, which is applicable for any combination of detector device and gamma criterion. The error curve method allows the evaluation of the magnitude of errors that can pass QA as opposed to simply reporting the percent pixels passing in the presence of an induced error. Secondly, we have elucidated sensitivity differences between three unique detector geometries and two very different delivery techniques in a way that removes differences in spatial sampling and uncertainties from real measurements. Additionally, a study of the effects of spatial sampling for each device and delivery technique illustrates that sparse spatial sampling is unlikely the main factor driving gamma

comparison insensitivity. Rather, comparisons between gamma value maps and gradient maps of the composite dose distributions on the phantom geometries suggest that the complexity of gradients in current comparisons limits gamma comparison error sensitivity. Finally, we have introduced a new method, gradient-dose segmented analysis (GDSA), for the analysis of IMRT QA comparisons. GDSA predicts the change in PTV mean dose using only information from the measurement and calculation in the phantom geometry. GDSA not only offers more clinically meaningful results than current methods, but also shows improved sensitivity and specificity as a binary test compared to the gamma comparison. Additionally, the use of GDSA in the clinic would not require an increase in time spent analyzing QA comparisons, which is an important consideration for implementation in the busy atmosphere of Radiation Oncology clinics.

The results from this work may aid in refining and improving error sensitivity of current IMRT QA comparisons, whether it be by selecting more appropriate device-specific gamma criteria, or with the clinical use of GDSA for IMRT QA comparisons. Future work may involve a more in-depth analysis of current gamma comparison limitations in analyzing ArcCHECK data, as well as modifications to the GDSA method that account for this unique detector geometry. The extension of GDSA to other measurement devices is straightforward and would simply require a study to determine the appropriate dose and gradient thresholds for devices other than those studied here. Further development of GDSA analysis may also allow the prediction of other relevant DVH metrics. The use of both the gamma comparison results from this work and the introduction of the GDSA will help to advance our understanding of current IMRT QA comparisons and allow more sensitive and clinically meaningful comparisons moving forward.