8-2016

Identifying Treatment Planning System Errors in IROC-Houston Head and Neck Phantom Irradiations

James Kerns

IDENTIFYING TREATMENT PLANNING SYSTEM ERRORS IN IROC-HOUSTON HEAD AND NECK PHANTOM IRRADIATIONS

by

James R. Kerns, M.S.

APPROVED:

_____
Stephen Kry, Ph.D.
Advisory Professor

_____
David Followill, Ph.D.

_____
Rebecca Howell, Ph.D.

_____
Adam Melancon, Ph.D.

_____
Francesco Stingo, Ph.D.

APPROVED:

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

IDENTIFYING TREATMENT PLANNING SYSTEM ERRORS IN IROC-HOUSTON HEAD AND NECK PHANTOM IRRADIATIONS

A

DISSERTATION

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

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

James Kerns, M.S.

Houston, Texas

August, 2016

Acknowledgements

I'd like to first thank my advisor Stephen Kry. He encouraged me to publish my Masters work way back when and eventually this led to where we are today. I'm grateful for his constant light-heartedness and open door. Our conversations were helpful not just with my dissertation but with life, and our profession. It goes to show that atheists and theists can get along in real life.

My committee, David Followill, Adam Melancon, Rebecca Howell, and Francesco Stingo were supportive of my work and helpful in all their suggestions. I'm grateful to have had the opportunity to pursue my Ph.D. and I wouldn't have been able to do it without them.

IROC-Houston staff and students have all been generous in their support and sharing of knowledge. The physicists and dosimetrists were very kind to share their time and wisdom.

My wife deserves the credit for getting me through most of my doctorate and her support of me as a person is overwhelming. I couldn't be the person I am today without her. As we look toward the future together, including a baby girl and moving across the country, I am so excited about being with her through all of it. I love you so much Love.

Finally, I want to give credit to my church community who have befriended me through my entire time in Houston and made life so much better in this terribly hot and stuffy city. They wouldn't have done any of it however without our common ground, Jesus Christ, by whose grace we can be hopeful of the future and real about our brokenness. If you make it this far, email me and I'll buy you a six pack of your favorite local beer.

IDENTIFYING TREATMENT PLANNING SYSTEM ERRORS IN IROC-HOUSTON HEAD AND NECK PHANTOM IRRADIATIONS

James R. Kerns, M.S.

Advisory Professor: Stephen Kry, Ph.D.

Abstract

Treatment Planning System (TPS) errors can affect large numbers of cancer patients receiving radiation therapy. Using an independent recalculation system, the Imaging and Radiation Oncology Core-Houston (IROC-H) can identify institutions that have not sufficiently modelled their linear accelerators in their TPS model. Linear accelerator point measurement data from IROC-H's site visits was aggregated and analyzed from over 30 linear accelerator models. Dosimetrically similar models were combined to create "classes". The class data was used to construct customized beam models in an independent treatment dose verification system (TVS). Approximately 200 head and neck phantom plans from 2012 to 2015 were recalculated using this TVS. Comparison of plan accuracy was evaluated by comparing the measured dose to the institution's TPS dose as well as the TVS dose. In cases where the TVS was more accurate than the institution by an average of >2%, the institution was identified as having a non-negligible TPS error. Of the ~200 recalculated plans, the average improvement using the TVS was ~0.1%; i.e. the recalculation, on average, slightly outperformed the institution's TPS. Of all the recalculated phantoms, 20% were identified as having a non-negligible TPS error. Fourteen plans failed current IROC-H criteria; the average TVS improvement of the failing plans was ~3% and 57% were found to have non-negligible TPS errors. Conclusion: IROC-H has developed an independent

recalculation system to identify institutions that have considerable TPS errors. A large number of institutions were found to have non-negligible TPS errors. Even institutions that passed IROC-H criteria could be identified as having a TPS error. Resolution of such errors would improve dose delivery for a large number of IROC-H phantoms and ultimately, patients.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

In the field of radiation therapy, dose is delivered to a patient with the intention of eliminating the cancerous cells while sparing normal healthy tissue as much as possible. Such delivery requires accurate knowledge of the patient anatomy, radiation-producing machines, patient setup at the time of treatment, and software planning systems that model dose delivery. Insufficient or inaccurate knowledge of any of these links may seriously compromise both the tumor control probability as well deliver more dose to healthy tissues.

It is the task of the Imaging and Radiation Oncology Core in Houston (IROC-Houston) to ensure that institutions that treat cancer patients with radiation therapy and that are participating in clinical trials do so accurately and safely. IROC-Houston has been performing this task since 1968. Since its inception, several programs have been developed to ensure this goal, two of which are relevant for this project. The first one is on-site dosimetry reviews. An on-site review is done by sending an IROC-Houston physicist to the institution with their own calibrated equipment. The IROC-Houston physicist measures basic dosimetric parameters of the institution's linear accelerators with a water phantom and ion chamber. These measurements are conducted in the presence of the institution physicist and are then compared to the institution's treatment planning system (TPS). The second relevant program uses anthropomorphic phantoms. Plastic phantoms that resemble human anatomy are sent to the institution and told to deliver a treatment plan to the phantom. The institution must go through the entire delivery process which includes scanning the phantom to identify anatomy, developing a treatment plan, setting up the phantom for radiation delivery, and actually delivering the dose. The phantom contains dosimeters that track the delivered dose and once the phantom and corresponding DICOM data are sent back to IROC-Houston they are read out. The delivered dose is then compared to the dose the

institution had planned to give via the DICOM files, based on the dose delivery simulation software. If the planned dose and delivered dose agree within a tolerance, the institution is allowed to treat clinical trial patients and the institution is generally said to be delivering dose accurately.

Unfortunately, a large percentage of institutions do not correctly deliver dose to these phantoms. Specifically, for the head and neck phantom, which is the most common phantom used, the failure rate of institutions in 2003 was ~35%.[1] This rate dropped to ~10% in 2012, signifying an improvement on the part of institutions and treatment planning software to correctly simulate and deliver dose to the phantom. Yet, this rate is still considered large. Additionally, the IROC-Houston tolerance for an acceptable delivery is wider than what is typically used in the clinic.

Determining the reason for the failure is a complex and difficult problem. Multiple steps are performed before the final delivery of dose to a phantom or real patient. Because IROC-Houston only knows the end result, i.e. whether the dose was delivered accurately, it is difficult to isolate where in the process an error may lie. For example, an institution may correctly scan the phantom and develop a dose plan, but at the time of delivery the phantom is not correctly positioned for the actual dose delivery. The dose delivered then is not like that of the simulation plan and will likely fail the IROC-Houston criteria. Usually, however, the problem is subtler. The phantom will proceed correctly through the planning and delivery process without any identifiable problem. If the results of the irradiation are not within criteria the institution may be at a loss for where the error lies. Many multiple small errors that are each within their own individual criteria may compound to an end result that is outside the criteria. Because the IROC-Houston phantom program is an end-to-end test, categorically speaking, these errors will never be identified. Experience on the part of IROC-Houston physicists may help in recognizing patterns in the dose delivery, but is suggestive at best.

There thus exists a serious problem in identifying where errors occur in the dose delivery process as well as in assisting institutions to rectify those problems.

One way of addressing this problem is to perform an independent audit or measure of a single feature that can be isolated. One feature that is increasing in popularity is a TPS independent second check. Such a system will independently calculate dose to the patient using the same geometric conditions as the clinical TPS. The point is to check that dosimetry calculations are accurate and add a level of safety to treatment planning. The second check system uses independent dosimetry data, algorithms, or assumptions to calculate the expected dose. This dose calculation is compared to the original TPS calculation; if there are differences between the two systems the physicist can then investigate discrepancies for a patient or phantom plan.

The goal of this study was to be able to identify one type of dose delivery error in an anthropomorphic phantom irradiation. Specifically, errors arising from inaccurate treatment beam modelling, either from erroneous input or physical modeling limitations. This goal assumes that a system can be developed that can correctly identify when an institution has a TPS calculation error. It also assumes that TPS errors contribute significantly to the number of phantom irradiations that fail to meet IROC-Houston criteria.

The hypothesis of this study was the following: ***By using an independent plan recalculation, IROC-Houston will be able to identify institutional treatment planning system calculation problems in 20% of head & neck phantom irradiation cases that fail credentialing.***

To test this hypothesis, the following 3 specific aims were developed and tested:

Aim 1: Acquire and develop reference data that accurately represent common linear accelerators. The goal of this aim is to create reference data that can be used by the

independent dose recalculation algorithm to accurately model linear accelerators that are currently in use. This is where IROC-Houston's site visit program comes in. Because data has already been acquired using calibrated equipment and using strict protocols it is accurate and comparable. This acquired data will serve as the basis for this aim. The working hypothesis for this aim was this: measurement data from linear accelerators is consistent between models and representative classes of multiple models can thus be formed.

Aim 2: Commission an accurate, independent dose recalculation system. This purpose behind this aim is to have an accurate dose recalculation system that can then be compared to the institution's TPS. The reference data from the above aim will be used as input to model the recalculation system. The system calculations will be compared to the reference data to determine the agreement. The working hypothesis is as follows: Beam models can be made in a treatment recalculation system that has the same or better agreement with input dosimetry data than a typical institutional TPS.

Aim 3: Recalculate dose to head and neck phantom irradiations and compare to institutional calculated dose. Once the recalculation system is accurately modeled, it can be used to recalculate dose to phantom irradiations. The accuracy of the institution TPS, i.e. the agreement between calculation and delivered dose, will be compared to the accuracy of the recalculation system. Irradiations where the recalculation system is considerably more accurate than the institution will be considered as having a TPS calculation error. All head and neck phantom irradiations from 2012 onward will be given to the recalculation system to generate a comparison. The hypothesis of this aim is that of the project.

Chapter 2: Reference photon dosimetry data for Varian accelerators

This chapter is based upon "Technical Report: Reference photon dosimetry data for Varian accelerators based on IROC-Houston Site Visit Data", by J. Kerns, D. Followill, J. Lowenstein, A. Molineu, P. Alvarez, P. Taylor, F. Stingo, and S. Kry, *Medical Physics* 43, 2374-2386 (2016). The journal allows a student's publication to be included in their dissertation.

2.1. Introduction

Using accurate dosimetry data is an essential part of providing high-quality radiation therapy treatments. This includes both acquiring accurate dosimetry data, and constructing an accurate beam model in the treatment planning system. The challenge in both of these steps has increased as new technologies like intensity-modulated radiation therapy (IMRT) and stereotactic body radiation therapy (SBRT) have become more common because these techniques have increased the necessary dosimetry data and data accuracy required.

One viable solution to help ensure accurate dose delivery is the creation of reference dosimetry data for linear accelerator (linac) beam data that can be used as a redundant dose verification tool. This dataset can, for example, be compared to commissioning measurements when important reference values are being established. Although dosimetry data for certain models of linacs have been published, including for multiple machines of the same type, no consistently collected large scale data source is yet available.[2-9] This study aims to evaluate and classify Varian linac models using statistical and clinical metrics. Furthermore, because of the large number of measurements, we provide not just reference dosimetry values for linacs but distribution characteristics so that physicists can evaluate their dosimetry data in the context of the distribution of similar linacs.

The Imaging and Radiation Oncology Core-Houston Quality Assurance Center (IROC-H), formerly named the Radiological Physics Center, was established to ensure that radiation therapy for institutions participating in the National Cancer Institute's clinical trials is delivered in a comparable, consistent and accurate manner. IROC-H has examined the dosimetric properties of linear accelerators since its inception. One way this is accomplished is through on-site dosimetry review visits by an IROC-H physicist to participating institutions. One component of the site visit is to acquire linac characteristics for basic dosimetry parameters.

In this work we present the measured dosimetry data from site visits for more than 500 Varian accelerators (Varian Medical Systems, Palo Alto, CA). Similarly-performing linac models have been grouped to form representative class datasets. These reference datasets can be used by physicists who might be commissioning a new treatment machine, considering matching different types of Varian machines or as a redundancy check of current baseline values. This work is a substantial expansion of previously published IROC-H photon data[10-12]; electron data exists as well but is not addressed here[13]. Because a large number of linacs have been measured, we can provide statistical metrics for each dataset so that a physicist can evaluate their machine's measurements not against a single value but against a distribution. IROC-H collects data from all types of linacs, but given the vast amount of data the analysis in this study was limited to one vendor, Varian.

2.2. Materials and Methods

2.2.1. Data Collection

All dosimetry data were acquired during IROC-H site visits using a 30 x 30 x 30 cm water phantom placed at a 100 cm source-to-surface distance. Point measurements were made with an Accredited Dosimetry Calibration Laboratory (ADCL) calibrated Farmer-type

chamber, typically a Standard Imaging Exradin A12 (Standard Imaging, Madison, WI), except for the IMRT and SBRT output fields (defined below), which used an Exradin A16 micro-chamber. Site visits and resulting measurements were performed by all physicists on staff at IROC-H in an approximately equal distribution following a consistent established standard operating procedure that included a detailed review by a second physicist. All measurements were conducted at the effective measurement location of the ion chamber, $0.6r_{cav}$ upstream of the physical center of the chamber.

Data from more than 500 Varian machines were collected during the period of 2000-2014 and are presented here. The number of measurements at a given point varied slightly as sometimes not every point was measured or recorded for a given parameter, and some parameters, like SBRT-style output factors, have only relatively recently started being collected.

The following dosimetric data point locations were measured. The percentage depth dose (PDD) was measured in a 6x6, 10x10, and 20x20 cm^2 field at effective depths of 5, 10, 15, and 20 cm; for 10x10 cm^2 fields a d_{max} measurement was also made. Field-size dependent output factors were measured at 10 cm depth. Values were converted to a d_{max} value based on the ratio of the institution's PDD values at 10 cm and d_{max} for 6x6, 10x10, 15x15, 20x20, and 30x30 cm^2 fields. Off-axis factors were measured at d_{max} at distances of 5, 10, and 15 cm away from the central axis in a 40x40 cm^2 field; at 10 cm off-axis, 4 measurements were made in the 4 cardinal directions of the field and averaged. Two sets of small field output factors were measured, both measured at 10 cm depth for the following field sizes: 2x2, 3x3, 4x4, and 6x6 cm^2. All measurements were normalized to a 10x10 cm^2 field. The first set of output factors, referred to hereafter as "IMRT-style output factors", had the jaws fixed at 10x10cm^2 and the MLCs moved to the mentioned field sizes; these are so called because they represent approximate segment sizes in an IMRT field. The second set

of output factors, referred to hereafter as “SBRT-style output factors”, moved both the jaws and MLCs to the given field size; these represent approximate positions during an SBRT treatment. A representative figure can be seen in Followill *et al*[10]. Wedge factors were measured at a depth of 10 cm in a 10x10 cm^2 field for 45° and 60° physical and enhanced dynamic wedges (EDW) when applicable; in addition, the 45° wedge was also measured at a depth of 15 cm in a 15x15 cm^2 field to verify depth and field size dependence of the wedge factor.

Although data have been collected for other energies, 6, 10, 15, and 18 MV are by far the most widely used energies and are thus presented here. Linac models not currently in widespread use were omitted from the analysis. Data were also reviewed for transcription and transfer errors to ensure integrity.

2.2.2. Data Analysis

All data analysis and visualization was done using the general programming language Python. The open-source “pandas” Python package was used for data munging and plotting.[14] Statistical testing used the “statsmodels” package.[15]

Varian linacs have been shown to have comparable dosimetric characteristics for machines of the same model and energy; beyond that, many different models have similar dosimetric properties. [1, 5, 8, 14, 15] This is understandable because different model names do not necessarily relate to differences in dosimetry – for example the EX and iX models differ only in the inclusion of an OBI system. IROC-H has measurements from over twenty Varian linac models. Each model may have multiple energies and produce specialized beams, e.g. flattening-filter free (FFF). If each energy and specialized beam is considered independent, there are over 50 measurement sets. Given the consistent dosimetric values and large

number of models, there was a desire to consolidate the different models into dosimetrically distinct groups, or “classes” of accelerator. Thus, models that fall into the same class can be considered dosimetrically equivalent at our criteria levels and our measurement points.

To categorize the different linac models into classes, two criterion were used to analyze comparability: statistical and clinical. The statistical criterion tested if a model’s mean parameter value (e.g. PDD(6x6cm^2, 5cm)) was significantly different from the comparison model’s mean value using analysis of variance and Tukey’s honest significant difference post-hoc analysis (α=0.05). The clinical criterion tested if the median value of a model’s dataset and the median value of the comparison model dataset had a local difference of less than 0.5%. This value was chosen because it is approximately equal to the overall standard deviation of the IROC-H measurements and these stricter criteria were deemed preferable to a looser one. If both criteria were not met, the dataset under consideration was rejected from that classification. The clinical criterion was added because statistical differences were occasionally achieved with very small differences in mean values (<0.5%).

Each model and energy combination was considered independent. Thus, using this classification it could be possible for two models to be dosimetrically equivalent at one energy but not at another energy. Specialized beams like Trilogy SRS and TrueBeam FFF were also independently evaluated. At each energy, the classification that represented the largest number of linacs was designated the “base” class. The base class was formed by starting with the most populous model dataset for that energy. This method was the most conservative approach since the most populous model had the narrowest confidence interval. The next most populous dataset was compared to the first. If it was within the criteria, that model was also said to be represented by the base class. Each subsequent model dataset was compared to the first. This process was then repeated for model datasets that were not within criteria of the base class, with the most populous remaining

dataset forming the start of the next class. This was repeated for each energy until no model datasets remained. Other classifications at the same energy were named as appropriate for the model(s) they represented. It should be noted that alternating the starting model had a negligible impact on the resulting classes. After all the models were assigned a class, the model datasets for a given class were assimilated into one dataset. Statistical metrics were derived from these combined datasets. Discussion of the dataset distributions can be found in the Appendix.

Finally, a comparison of classes was done for each dosimetric parameter at each energy. The 6 MV data are displayed via figures to fully describe the data distribution, while 10, 15 and 18 MV have been described in tables to save space. Quantitative data for all energies can be accessed through the online content which includes the number of measurements, median, standard deviation, and the 5^{th} and 95^{th} percentile values.[16] Because of the large number of data points the 6 MV figures are plotted in boxplot fashion. The central line within the box represents the median which is robust to outlier influences. The top and bottom of the box represent the 75^{th} and 25^{th} percentiles, respectively, and the whiskers above and below the box represent the 95^{th} and 5^{th} percentile values, respectively. Data are shown here graphically to quickly convey qualitative differences between machine classes and show the entire distribution, but the median value of each class is also given at the top of the plot.

2.3. Results

2.3.1. Model Comparison

For all energies, the Clinac 21EX model dataset was the most populous. At 6 MV, 17 models were evaluated. Six models were within comparability criteria and were assimilated

to form the base class. The remaining classes were generated using the same comparison process, resulting in a total of 8 classes. There were 11 models at 10 MV and 12 at 15 MV, which consolidated to 3 and 2 classes respectively. Eleven models were all consolidated to one class at 18 MV. The model classification results are shown in Table 2-1. This table is how a physicist can identify what class their linac is in. These classes are dosimetric representatives of the listed models. E.g. using Table 2-1, a 21iX 10 MV beam is said to be represented by the 10 MV base class, and evaluations of the individual machine should be performed against the results of that class.

	Class	Represented Models/Beams
6 MV	Base	21EX (D), 23EX, 21iX, 23iX, Trilogy
	TB	TrueBeam
	TB-FFF	TrueBeam FFF
	Trilogy SRS	Trilogy SRS
	2300	2300 (C) (CD)
	2100	2100 (C) (CD)
	600	600 (C) (CD)
	6EX	6EX
10 MV	Base	21EX (D), 23EX, 21iX, 23iX, Trilogy, 2100 (C) (CD), 2300
	TB	TrueBeam
	TB-FFF	TrueBeam FFF
15 MV	Base	21EX, 23EX, 21iX, 23iX, Trilogy, 2100 (C) (CD), 2300 (C) (CD)
	TB	TrueBeam
18 MV	Base	21EX (D), 23EX, 21iX, 23iX, Trilogy, 2100 (C) (CD), 2300 (CD)

Table 2-1. Derived classes with the machine models and/or beams they represent.

Overall differences between classes of machines were evaluated in Figure 2-1. This is clinically important when trying to match machines of different classes, or deciding how many TPS beam models to create. A comparison between the different classes is shown via a heatmap in Figure 2-1 relative to the base class. The color of the squares represents the maximum median difference that class had for that parameter in comparison to the base class. Darker squares indicate parameters that have a greater maximum difference than the dosimetric characteristics of the base class while lighter colors indicate smaller maximum differences. An asterisk indicates that the mean value of at least one measurement location is statistically different from the base class' mean value ($\alpha=0.05$). For example, for the PDD, the Clinac 2100 class had at least 1 PDD value that had a clinically and significantly different value as compared the base class. Although differences and significance are plotted relative to the base class, this does not imply that the base class is a benchmark; it is only meant as a guide in understanding class differences. Since only the maximum differences are plotted, it should be understood that a class may perform similarly to another class except at a single measurement point. For example, two classes' 5, 10, and 15 cm PDD measurements may agree well but if the 20 cm measurement value is significantly and clinically different, that is the value plotted in Figure 2-1.

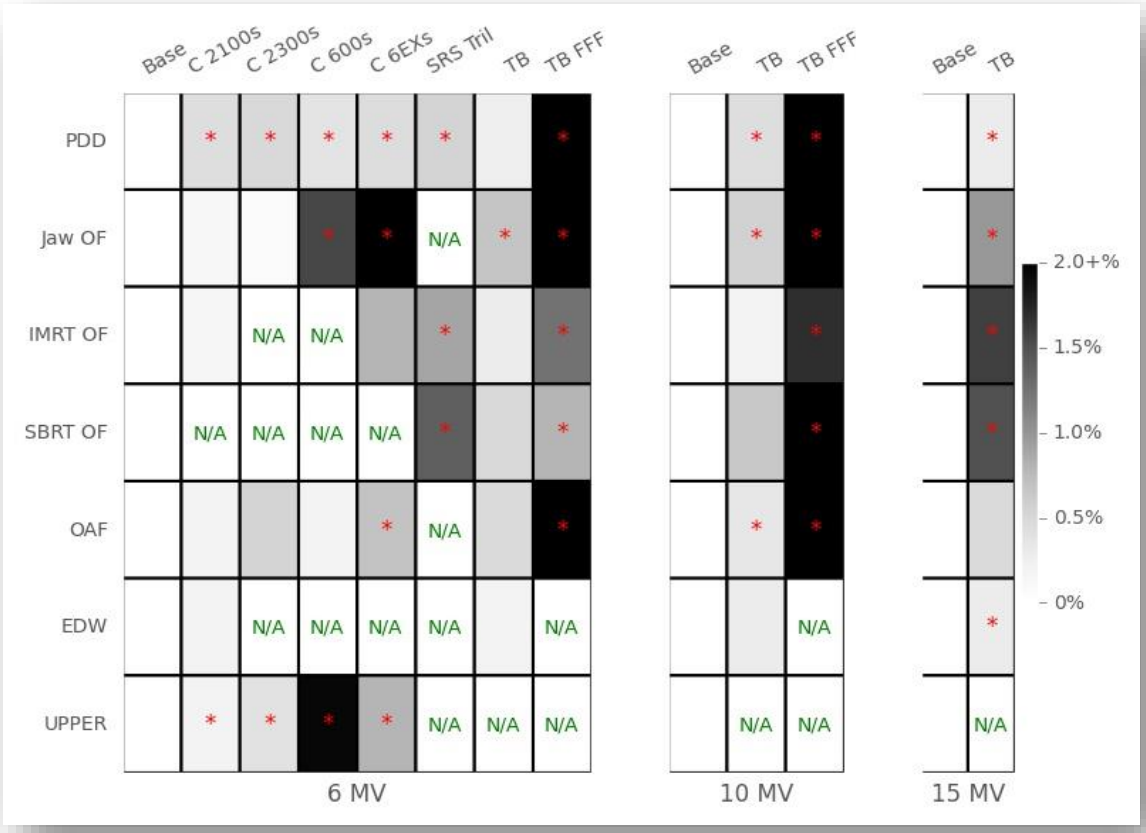

Figure 2-1. Heatmap comparing the maximum difference between parameters for the Base class compared to other classes. Parameters include percent depth dose (PDD), jaw-based output factor (Jaw OF), IMRT-style small field output factors (IMRT OF), SBRT-style small field output factors (SBRT OF), off-axis factors (OAF), and wedge factors for enhanced dynamic (EDW) and "upper" physical wedges (UPPER). Darker color indicates larger maximum differences from the base class. 18 MV is not shown because it only had one class. "N/A" means that measurements were not available for that parameter. An asterisk indicates a statistically different mean value from the base class.

2.3.2. 6 MV

For 6 MV, and all other energies, data are routinely described in two ways. The difference between classes, called the interclass difference or variability, represents the local difference of the median values. The average interclass difference is the mean of the differences at all the field sizes or depths. Difference can also be described within the class, which we labeled intraclass variability, which is synonymous with the coefficient of variation.

For 6 MV, data are given in figures for visual comparison, but the online content also contains tabular data.[16] The measured PDDs of the 10x10 cm^2 field are shown in Figure 2-2 and were normalized to the d_{max} measurement. Notably, all classes performed consistently with depth; i.e. if a class had a higher PDD at 10 cm, it was also almost always higher at 5, 15, and 20 cm. On average, the base class had 0.5% intraclass variability. The 2100 and 2300s classes had consistently harder spectra than the base class, while the 600, 6EX, and Trilogy SRS had softer. The TrueBeam class had a very similar spectrum to the base class, with the largest interclass difference between the two classes being -0.5% at 20 cm. Most previously published PDD data values were similar to our values although deviations are apparent, notably at 5cm depth. However, the previously published data also have the largest spread in values at 5cm.

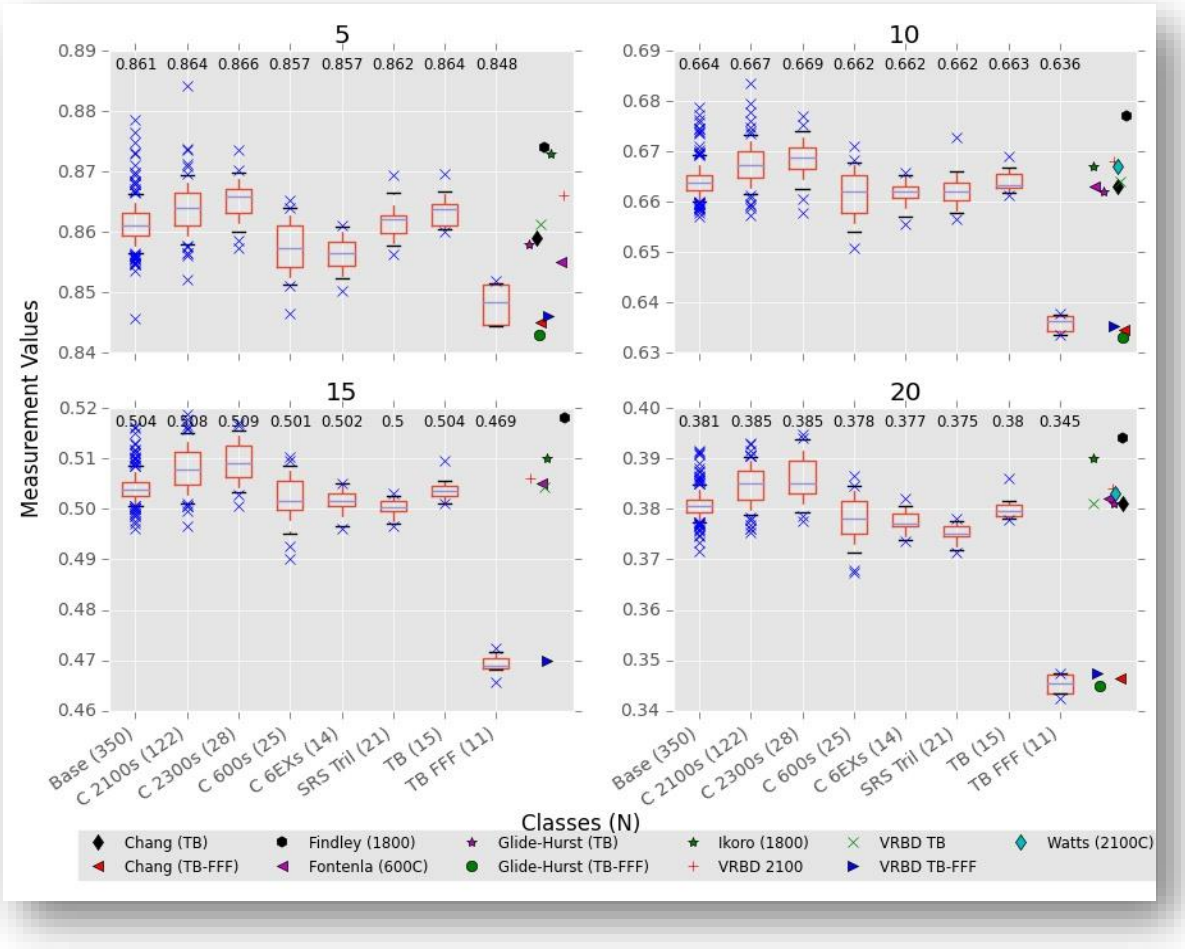

Figure 2-2. 6 MV 10x10 cm depth dose measurements at 5, 10, 15, and 20 cm. Comparisons to other data measurements are from Refs [3-8]. Class medians are posted at the top of the respective boxplot and are the central lines in the boxes. *N* is the number of measurements in a class. The top and bottom of the box represent the 75th and 25th percentiles and the whiskers represent the 95th and 5th percentiles.

Output factors at d_{max} for field sizes of 6x6, 15x15, 20x20, and 30x30 cm^2, normalized to the measurement at 10x10 cm^2, are shown in Figure 2-3. The total range in output factor values across field sizes was largest for the base, 2100s and 2300s classes, which all performed comparably. The 600s, 6EXs, and TrueBeams all had output factors closer to unity at all field sizes (i.e. a flatter slope) although each of them displayed distinctive characteristics. The Trilogy SRS class also showed a similar effect as the latter

group for the applicable field sizes (15x15 cm^2 and smaller; data not shown but is contained in online content).

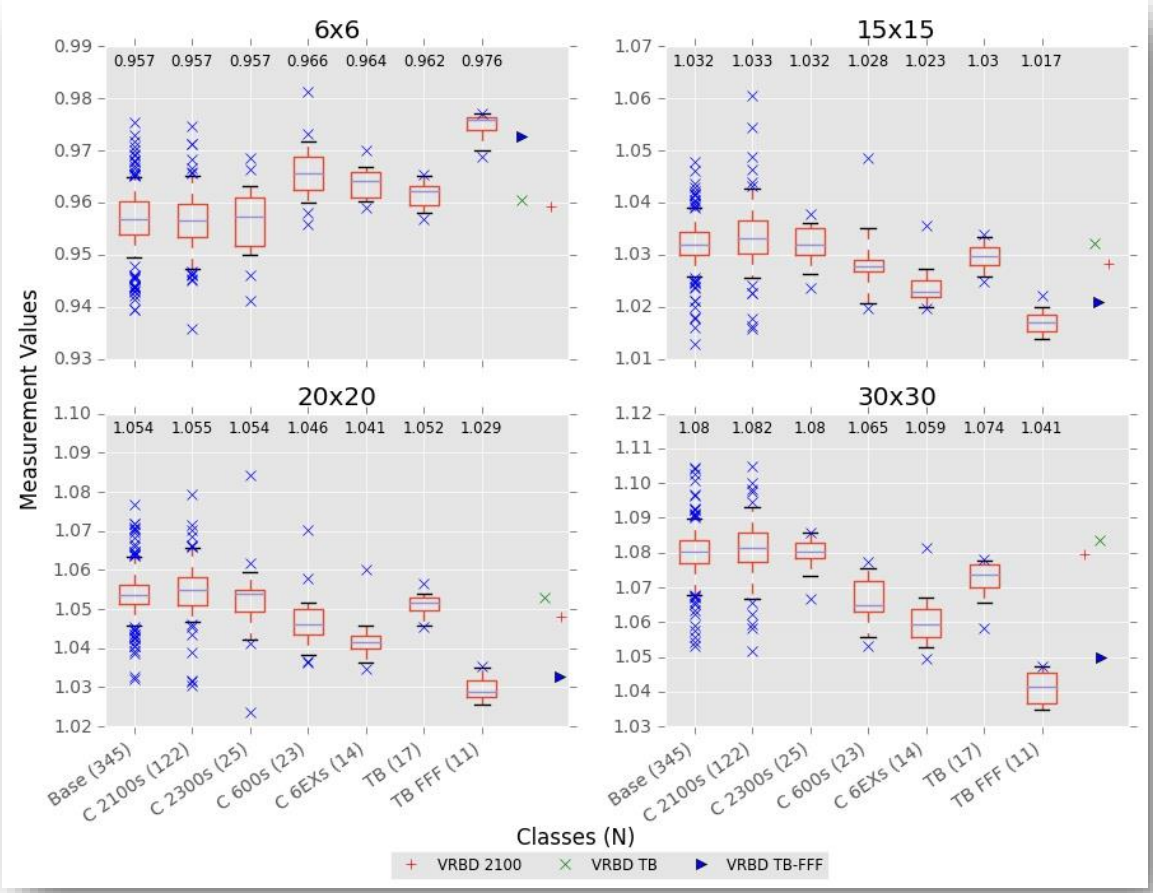

Figure 2-3. Jaw output factors for 6 MV classes at d_{max}, normalized to the 10x10cm^2 field. Field sizes are above each panel and in cm^2. Median class values are also included at the top of each panel. *N* is the number of measurements in a class.

Off-axis factors were measured at 5, 10, and 15 cm from the central axis (CAX) (Figure 2-4). Agreement of other classes to the base class was closest at 5 cm, but grew apart at further distances from the CAX. The base, TrueBeam, and TrueBeam FFF classes had the smallest interclass variability. The largest interclass variability was from the Clinac 600s, having 1.5% at 15 cm compared to the base

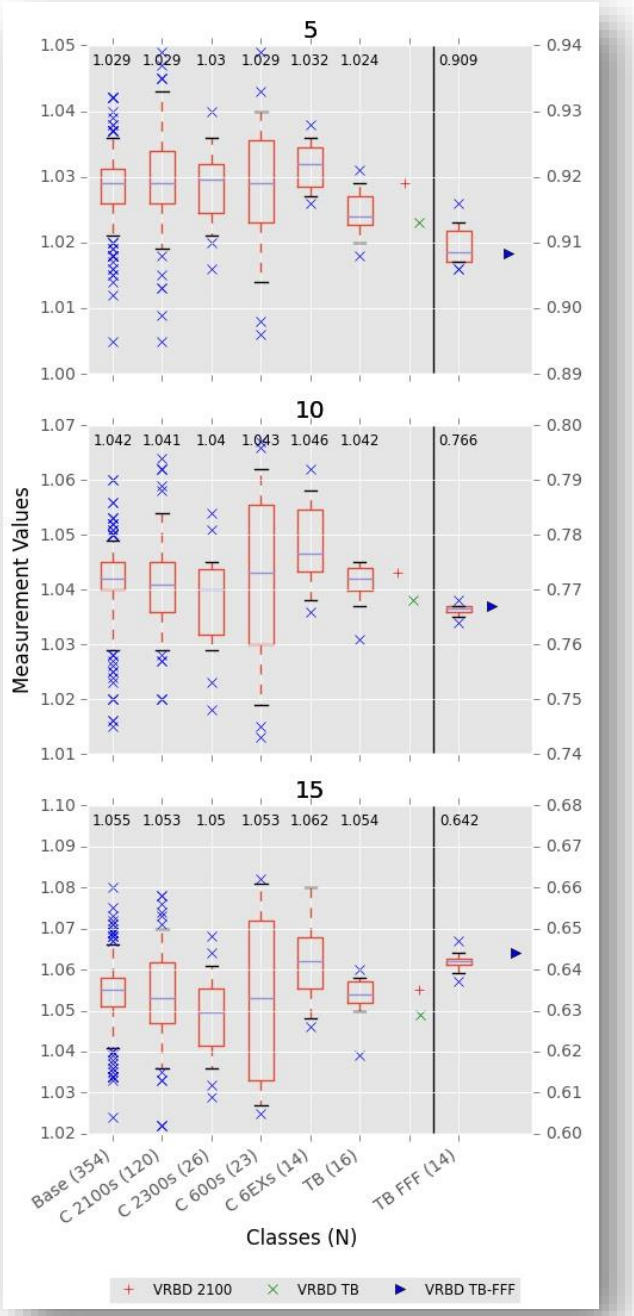

Figure 2-4. 6 MV Off-Axis Factors at the distances indicated (in cm) away from the CAX. TrueBeam FFF data and FFF reference data are aligned to the right axis and are visually separated by the vertical line. *N* is the number of measurements in a class.

class' 0.6%. On average, the off-axis parameter saw the greatest intraclass variation of all parameters. Interestingly, the inline waveguide machine models (600s, 6EXs) saw the greatest variability.

IMRT output factors, defined as having fixed jaws at 10x10 cm^2 and various MLC field sizes, are shown in Figure 2-5. No 600, 6EX, 2100, or 2300 class data were available. Collecting small field data is notoriously challenging as results are very sensitive to relatively small setup errors. However, intraclass variability of our measured data was comparable to the other parameters, having an average intraclass variation of 0.5% and a maximum of 0.8%, belonging to the base class. At 6x6cm^2, the interclass difference is relatively small, but the differences increase with smaller field sizes. At 2x2 cm^2, the 6EX and Trilogy SRS classes had at least 1.0% interclass difference compared to the base class.

SBRT output factors, defined as having both the jaws and MLCs at the given field size, are shown in Figure 2-6. SBRT output factors have only recently started being collected by IROC-H, thus the number of measurements compared to other parameters is fewer. As would be expected, these output factors are smaller than the corresponding IMRT output factors shown above. The base class and TrueBeam performed comparably, having an average interclass difference of 0.25%. The Trilogy SRS class had markedly different values from the base class with an average interclass difference of 1.4%, a difference even greater than the TrueBeam FFF class.

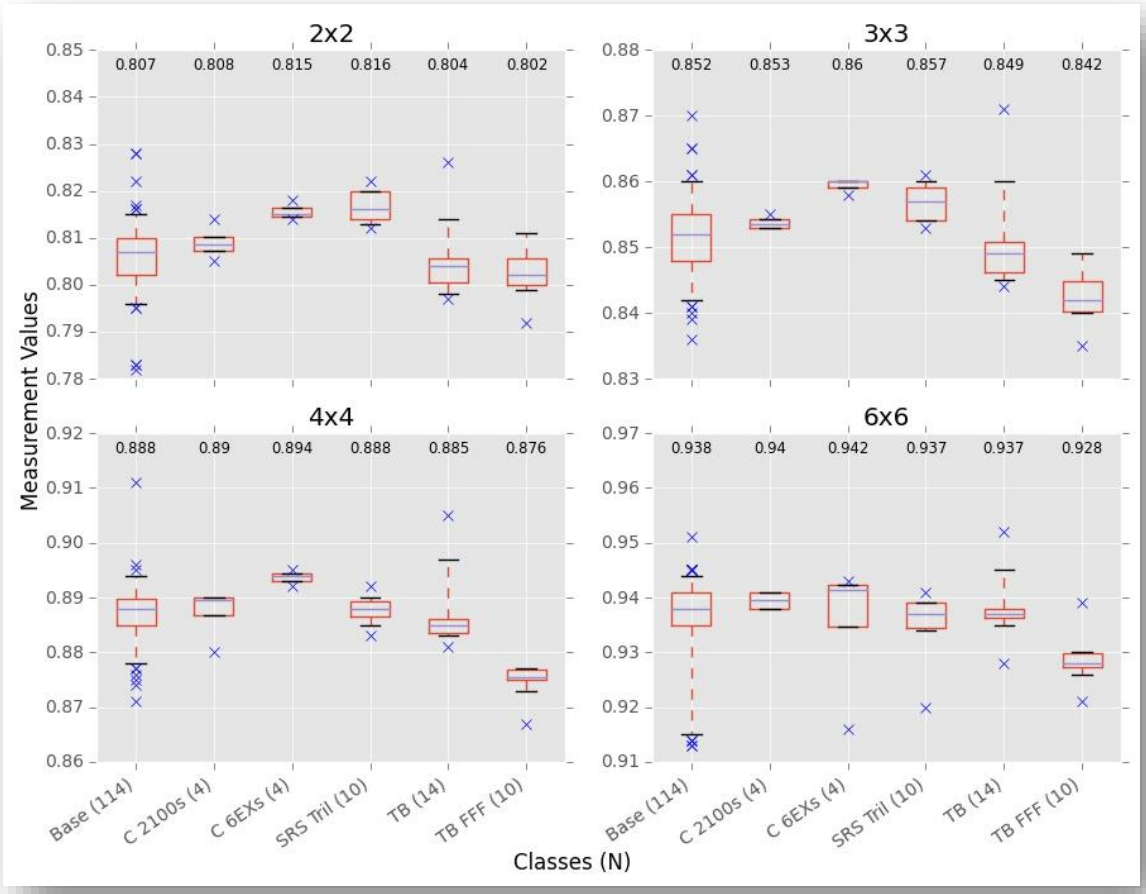

Figure 2-5. 6 MV IMRT-style output factors. Jaws were at 10x10 cm^2 for all measurements while the MLCs defined the field. Readings were normalized to a field where both the jaws and MLCs were at 10x10cm^2. Field sizes are in cm^2 and indicate the MLC field. *N* is the number of measurements in a class.

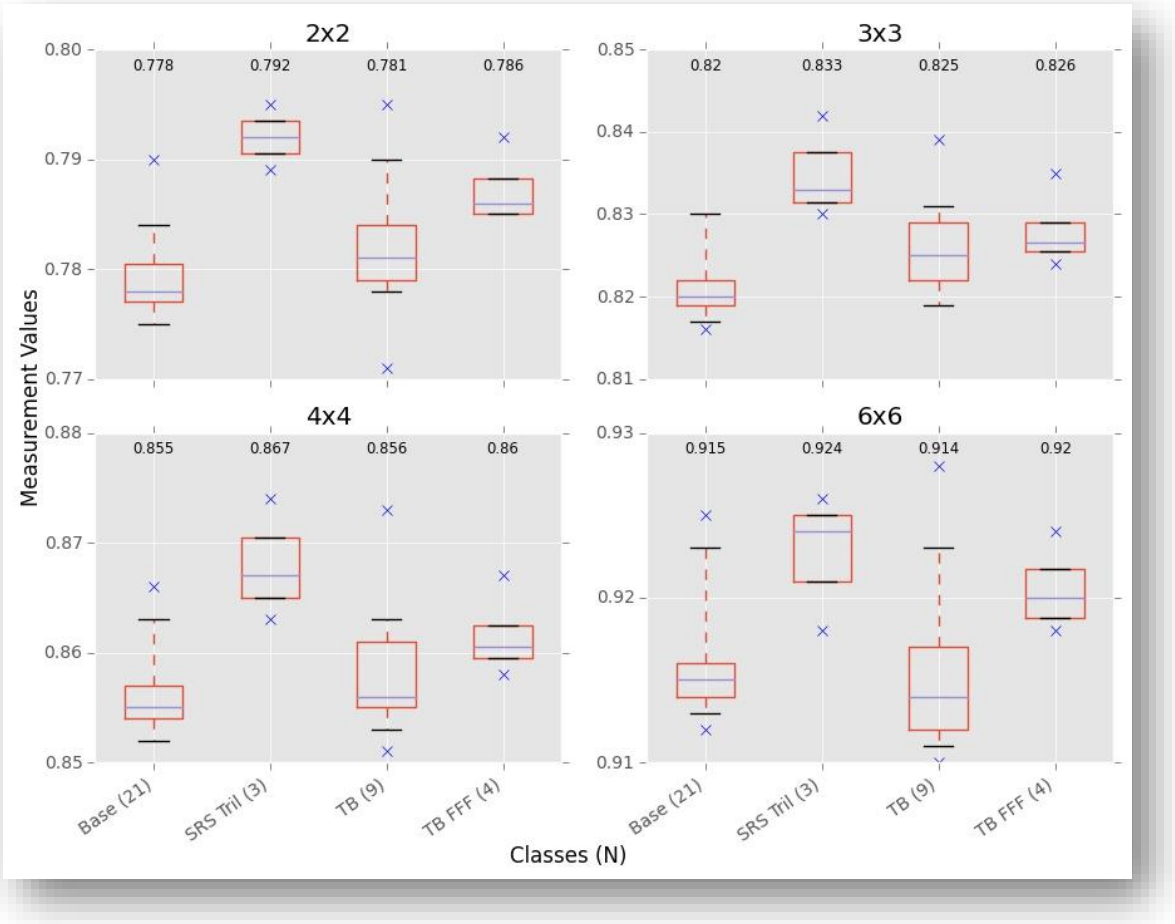

Figure 2-6. 6 MV SBRT output factors. Jaws and MLCs were both at the indicated field size above the panels. Readings were normalized to a field where both the jaws and MLCs were at 10x10cm^2. Fields are in cm^2. *N* is the number of measurements in a class.

Wedge factors for Varian include both “upper” physical wedges and enhanced dynamic wedges. The EDW results are shown in Figure 2-7 while the physical wedge results are shown in Figure 2-8. While only 3 classes had EDW measurements, both the interclass and intraclass variability is small with all classes performing similarly. The physical upper wedges however showed larger interclass variability. The base, 2100, and 2300 classes all had relatively low interclass variability, but the 600 and 6EX classes showed large differences, both interclass and intraclass.

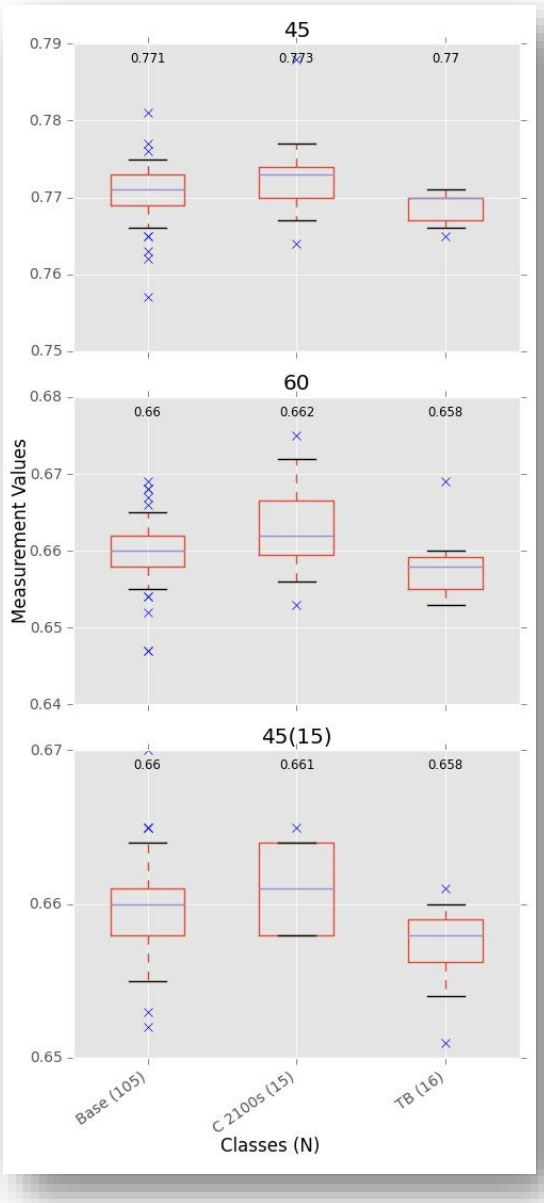

Figure 7. 6 MV EDW Factors. All measurements were at 10x10cm^2 and 10 cm depth except the 45(15) measurement, which was at 15 cm depth and 15x15 cm^2 field size. *N* is the number of measurements in a class.

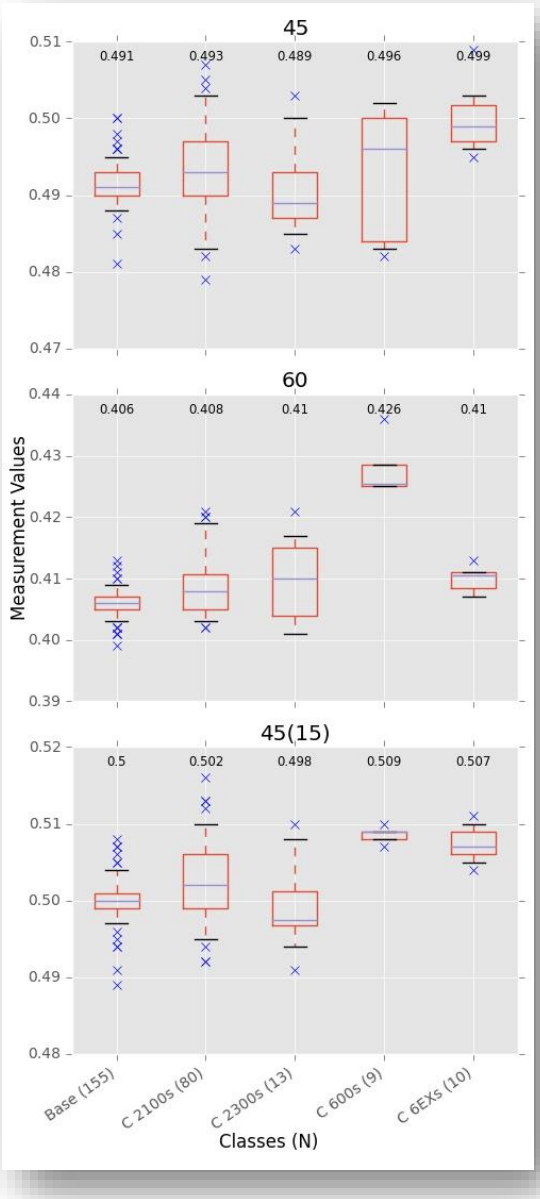

Figure 8. 6 MV "Upper" wedge factors. All measurements were at 10x10cm^2 and 10 cm depth except the 45(15) measurement, which was at 15 cm depth and 15x15cm^2 field size. *N* is the number of measurements in a class.

Parameter/Class		Base Class	TB-Flat	TB-FFF
	N	74	10	10
PDD	5 cm	0.913 (0.003)	0.918 (0.002)	0.908 (0.002)
	10 cm	0.733 (0.003)	0.737 (0.002)	0.712 (0.001)
	15 cm	0.582 (0.003)	0.586 (0.002)	0.554 (0.001)
	20 cm	0.460 (0.002)	0.463 (0.003)	0.430 (0.001)
	N	74	10	10
Output Factors	6x6 cm	0.953 (0.004)	0.956 (0.003)	0.980 (0.001)
	15 x 15 cm	1.033 (0.004)	1.032 (0.002)	1.015 (0.002)
	20 x 20 cm	1.054 (0.005)	1.053 (0.003)	1.026 (0.002)
	30 x 30 cm	1.083 (0.007)	1.077 (0.006)	1.034 (0.004)
	N	74	10	10
Off-Axis Factors	5 cm left	1.029 (0.006)	1.029 (0.004)	0.823 (0.003)
	10 cm avg	1.044 (0.006)	1.047 (0.003)	0.632 (0.002)
	15 cm left	1.053 (0.009)	1.056 (0.005)	0.497 (0.003)
	N	21	9	8
IMRT Output Factors	2x2 cm	0.825 (0.007)	0.823 (0.003)	0.842 (0.005)
	3x3 cm	0.881 (0.006)	0.880 (0.006)	0.892 (0.004)
	4x4 cm	0.918 (0.006)	0.916 (0.006)	0.918 (0.003)
	6x6 cm	0.959 (0.005)	0.958 (0.005)	0.956 (0.002)
	N	4	6	4
SBRT Output Factors	2x2 cm	0.794 (0.004)	0.790 (0.005)	0.825 (0.006)
	3x3 cm	0.846 (0.001)	0.849 (0.005)	0.881 (0.003)
	4x4 cm	0.877 (0.005)	0.884 (0.006)	0.907 (0.003)
	6x6 cm	0.929 (0.006)	0.933 (0.006)	0.948 (0.004)
	N	17	10	
EDW Factors	45°	0.803 (0.004)	0.800 (0.003)	N/A
	60°	0.701 (0.005)	0.698 (0.004)	
	45°(15x15,15)	0.703 (0.004)	0.702 (0.003)	
	N	40		
Upper Wedge Factors	45°	0.525 (0.004)	N/A	N/A
	60°	0.437 (0.003)		
	45°(15x15,15)	0.531 (0.003)		

Table 2-2. 10 MV Varian collected data for the three identified classes. Median values are given with the standard deviation in parentheses. *N* is the number of measurements.

2.3.3. 10 MV

The results of the 10 MV parameter measurements are shown in Table 2-2 for the three identified classes: Base, TrueBeam, and TrueBeam FFF. Overall, the base class commonly had a larger intraclass variability than did the TrueBeam and TrueBeam FFF classes. Of interest, deviations between the base and TrueBeam classes were seen that were not present at 6 MV. At 6 MV, the TrueBeam and base class had very similar PDDs. However, at 10 MV the TrueBeam had a harder beam (an average of +0.6% interclass difference at all depths). The intraclass variability was however comparable to 6 MV, with the base class having the largest average intraclass variability of 0.4%.

Jaw output factors showed similar results to 6 MV in that the TrueBeam class showed less range across the field sizes, i.e. a flatter slope, than the base class. The average intraclass variability of the base class was 0.5% compared to 0.3% and 0.2% for the TrueBeam and TrueBeam FFF classes, respectively. Off axis factors were similar between the base class and TrueBeam class, with the TrueBeam-FFF having larger differences, as expected.

IMRT output factors had small interclass variability, with the base class and TrueBeam having an average of 0.2%. Notably, the TrueBeam FFF class agreed with the other classes at 4x4 and 6x6 cm^2, but had a larger output factor at smaller field sizes.

SBRT factors appeared similar to IMRT output factors, but had less interclass agreement, and the TrueBeam FFF class was consistently higher than the other classes at all field sizes. Whereas the output factors were very close at 6 MV, at 10 MV the TrueBeam

class had an average deviation of +0.3%. The base class had an average intraclass variation of 0.3% compared to the TrueBeam at 0.7%.

EDW factors were very similar for the base and TrueBeam classes, having only 0.3% interclass variation and intraclass variations of 0.6% and 0.4% for the base and TrueBeam classes respectively. No measurements of upper physical wedges have yet been done for TrueBeam 10 MV.

2.3.4. 15 MV

Measurements for the 15 MV beams are shown in Table 2-3; the classes determined were the base class and TrueBeam. Some similarities to other energies were seen, but many differences were noted at 15 MV. The interclass differences also had the largest magnitudes at 15 MV.

Between the two classes, the TrueBeam beam had a consistently harder PDD than that of the base class, as was also seen at 10 MV. The TrueBeam had an average interclass difference of +0.4%.

At 6 and 10 MV, the TrueBeam output factors had a smaller slope across field sizes than the base class; at 15 MV the opposite was true: the slope was steeper. The TrueBeam had a median difference of -0.6% at 6x6 cm^2 and +0.9% at 30x30 cm^2 compared to the base class.

Parameter/Class		Base Class	TB-Flat
	N	100	14
PDD	5 cm	0.943 (0.003)	0.946 (0.002)
	10 cm	0.767 (0.003)	0.769 (0.001)
	15 cm	0.617 (0.003)	0.620 (0.001)
	20 cm	0.496 (0.002)	0.499 (0.001)
	N	97	14
Output Factors	6x6 cm	0.959 (0.005)	0.953 (0.003)
	15 x 15 cm	1.030 (0.004)	1.035 (0.002)

	20 x 20 cm	1.050 (0.006)	1.060 (0.003)
	30 x 30 cm	1.075 (0.007)	1.085 (0.005)
	N	100	13
Off-Axis Factors	5 cm left	1.038 (0.008)	1.036 (0.004)
	10 cm avg	1.049 (0.008)	1.045 (0.004)
	15 cm left	1.061 (0.010)	1.056 (0.004)
	N	18	13
IMRT Output Factors	2x2 cm	0.831 (0.006)	0.815 (0.006)
	3x3 cm	0.893 (0.008)	0.885 (0.004)
	4x4 cm	0.929 (0.008)	0.924 (0.002)
	6x6 cm	0.965 (0.007)	0.964 (0.002)
	N	7	7
SBRT Output Factors	2x2 cm	0.799 (0.004)	0.784 (0.006)
	3x3 cm	0.865 (0.002)	0.855 (0.004)
	4x4 cm	0.902 (0.003)	0.892 (0.004)
	6x6 cm	0.949 (0.003)	0.938 (0.003)
	N	27	11
EDW Factors	45°	0.814 (0.003)	0.811 (0.002)
	60°	0.716 (0.003)	0.713 (0.003)
	45°(15x15,15)	0.720 (0.001)	0.718 (0.003)
	N	44	
Upper Wedge Factors	45°	0.523 (0.003)	N/A
	60°	0.434 (0.003)	
	45°(15x15,15)	0.529 (0.003)	

Table 2-3. 15MV Varian collected reference data. Median values are given with the standard deviation in parentheses. N is the number of measurements.

Parameter/Class		Base Class
	N	243
PDD	5 cm	0.963 (0.003)
	10 cm	0.793 (0.003)
	15 cm	0.647 (0.002)
	20 cm	0.527 (0.002)
	N	243
Output Factors	6x6 cm	0.943 (0.006)
	15 x 15 cm	1.041 (0.005)
	20 x 20 cm	1.066 (0.007)
	30 x 30 cm	1.094 (0.010)
	N	243

Off-Axis Factors	5 cm left	1.029 (0.006)
	10 cm avg	1.044 (0.006)
	15 cm left	1.054 (0.009)
	N	37
IMRT Output Factors	2x2 cm	0.806 (0.005)
	3x3 cm	0.884 (0.004)
	4x4 cm	0.929 (0.004)
	6x6 cm	0.970 (0.003)
	N	6
SBRT Output Factors	2x2 cm	0.767 (0.003)
	3x3 cm	0.847 (0.001)
	4x4 cm	0.891 (0.000)
	6x6 cm	0.942 (0.001)
	N	53
EDW Factors	45°	0.824 (0.002)
	60°	0.729 (0.003)
	45°(15x15, 15)	0.734 (0.003)
	N	112
Upper Wedge Factors	45°	0.516 (0.003)
	60°	0.427 (0.002)
	45°(15x15, 15)	0.522 (0.003)

Table 2-4. 18MV Varian collected reference data. Median values are given with the standard deviation in parentheses. N is the number of measurements.

Off-axis factors, similar to the output factors, showed differences between the TrueBeam and base class that were not seen at 6 or 10 MV. At the other energies, the off-axis factors had little interclass variation; at 15 MV the TrueBeam had -0.4% and -0.6% median difference at 10 and 15 cm off axis, respectively. The base class had an average intraclass variation of 0.8% compared to 0.3% for TrueBeam.

IMRT output factors also showed marked differences in the two classes compared to other energies. Class medians are nearly the same at 6x6 cm^2, but the interclass variability increases to nearly 2% at 2x2cm^2.

SBRT output factors once again showed differences from the other energies. The TrueBeam consistently had lower median values than the base class for all field sizes. The average interclass difference was -1.3% for the TrueBeam, while the average intraclass variation was 0.4% for both classes.

EDW factors also showed the TrueBeam class as having a consistently lower median than the base class, although the difference was much less than for the IMRT and SBRT output factors and was nearly the same as the average intraclass variation (0.4%). No physical wedges have been measured for the TrueBeam at 15 MV.

2.3.5. 18 MV

The 18 MV measurement results are shown in Table 2-4. There was only one resultant class, so no comparison between classes could be done. Intraclass variability of the base class was similar to other energies however, with an average intraclass variability of 0.45% across all parameters.

2.4. Discussion

The IROC-H site visit data for Varian linacs and have been compared, and dosimetrically similar linac models have been grouped into representative “classes” at each energy. A base class which represented the most popular models of linacs was developed for each energy. A total of 8 classes were developed for 6 MV and 3, 2, and 1 class for 10, 15, and 18 MV respectively. The data presented here are the first to show how Varian linac models perform relative to each other using a systematic approach. Differences between models and classes have been quantified so that physicists can understand how their specific

machine performs relative to the community, and also how different linac models compare dosimetrically to one another.

Previous datasets have been published that describe Varian linac dosimetric properties. These publications are typically based on a limited number of machines. Data that were taken under the same conditions are presented throughout Figures 2-2 through 2-4 and in the online content.[16] Although other reference datasets exist, the large dataset provided by IROC-H, drawn from hundreds of machines and institutions, allows for statistically robust reference values and metrics, over against a handful of measurements. Most of the values agree however, so these data are not in opposition—they are well within the range given by our large collection of data. We have added to the robustness of the "average" machine values as well as described the range of values seen from the machines. As well, we have described observed differences between the linac types.

The parameter with the most numerous comparisons with previously published values is the PDD. As well, the PDD measured by IROC-H occasionally disagreed with previous data in the literature. For example, for the 6 MV base class, IROC-H measured a PDD(10) of 66.4% for the 10x10 field. Most data in the literature, including the Varian Reference Beam Data, describe a value between 66.7% and 67.7%.[3, 5-7] Although small, this is a notable difference in a core physics parameter. This difference likely arises from the method of measurement. IROC-H, since 2000, has measured PDDs accounting for the effective point of measurement of the ion chamber. Measurements taken before TG-51 or not accounting for the $0.6r_{cav}$ shift, such as Ikoro *et al*[5] and Findley *et al*[7] were systematically higher than our values because the effective point of measurement was actually upstream of the measurement point. This theory is supported in that PDD(10) data published post-TG-51[2, 3, 6, 9] agree well with our value of 66.4%.

Output factors also proved noteworthy. At 6 MV, the jaw output factors showed the most interclass variability of all the classes. For 6EXs and TrueBeams, this is the parameter with the greatest mean difference from the base class. Those classes showed less sensitivity to field size, giving a flatter slope of output factors. However, at 15 MV the differences between the base and TrueBeam class are the opposite. The TrueBeam class shows greater change with field size; at 30x30 cm^2 the differences are statistically and clinically different from the base class. Beyer[2] also concluded that there were output factor differences between the TrueBeam and Clinac 2100, but that the TrueBeams values varied less with field size. Our study confirms this result for 6 and 10 MV but not 15 MV. This difference may lie in our use of many linac measurements compared to one machine. We have also shown whether these differences are statistically and clinically significant at the given energies.

Two uses of this data are noteworthy. First, the dosimetric properties of an individual linac can be quantitatively compared directly to these reference data. The goal of quantifying differences is not to match or try to match the class median, but rather to identify the magnitude of those differences and where they lie. Most linacs should have dosimetric properties that are consistent with the reference data presented here. However, if a machine value is obtained for a particular machine that is different than one presented here for that machine's class, this should not necessarily be interpreted as an error as it could be a non-standard machine. However, such a difference should raise awareness and warrant an investigation to ensure that the difference is justified. Furthermore, differences should be evaluated in context. One value with a large difference may represent a one-off collection or transcription error; many measurements with a systematic difference may represent a setup or collection error. Additionally, machines may truly be different, either through manufacturing or from physicist customization. In any case, awareness of differences can be raised. This can be especially helpful when commissioning a new machine when no prior

reference values can be used for comparison. The most important agreement is that of the machine to the treatment planning system. Differences between dosimetry values do not inherently carry clinical impact until modeled. Thus, reference datasets like this study can add a check that the acquired values used in TPS modelling are sufficiently accurate.

A second important use of these data arises when considering the purchase of another machine or trying to match existing machines and the impact it may have. Clinics with multiple linacs can use one beam model for their TPS only if all the machines have similar dosimetric properties. If machines of different classes are being matched, the clinically acceptable window and therapy type needs to account for the underlying differences in the machine parameters, not including the general uncertainty in the beam model. A useful case for comparison is institutions that have, or will transition from older Varian models to the newer TrueBeam platform. An important clinical question is “how similar are these machines?” and therefore “Can I use the same beam model for both?” These differences, including the parameter(s) that are different and the magnitude of difference, are described in this study and an example of comparison to the base class is shown in Figure 2-1. In general, the differences between the base and TrueBeam classes were fewer at 6 and 10 MV, but had numerous differences at 15 MV. As described above, the output factors were typically the biggest difference between classes. Such data show that the TrueBeam is not dosimetrically similar to prior linac models at all points. Institutions that have a mixture of linac classes should be careful in their data acquisition and beam modelling to avoid systematic bias toward one class of machine.

Finally, it should be emphasized that this dataset is not meant to replace machine measurements at a clinic. Conformity to median values is not the goal, which is why the distributions have also been included. This dataset can be consulted to compare a machine

or clinic to that of the community at large, understanding that each machine may be slightly unique but that differences should be identified as such.

2.5. Conclusion

Data from the IROC-Houston's site visits have been analyzed, comparing and classifying Varian linacs based on dosimetric characteristics. Linac models with comparable dosimetric properties were grouped into classes, and the reference class data, including the underlying distribution, is presented here. Dosimetric characteristics included point measurements of the PDD, jaw output factors, two types of MLC output factors, off-axis factors, and wedge factors. The data can be used as a secondary check of acquired values of a new machine, to understand how a machine performs relative to the community. These reference data can also be used as a guide of how much variability a physicist should expect between different models of linear accelerators.

Many thanks go to Paul Holguin for querying and compiling the site visit data. This work was supported by Public Health Service Grant CA180803 awarded by the National Cancer Institute, United States Department of Health and Human Services.

2.6. Appendix: Data distribution & statistical metrics

In this appendix we examine the data distribution of the measurements and reasons for using certain statistical metrics. Measurement distributions with a bimodal distribution would suggest systematic error in the collection of data. Distributions with a relatively large standard deviation may suggest random error, perhaps describing aspects of the linac that were not as uniform during the manufacturing or operation processes or setup uncertainty.

The data for the 30x30 cm^2 jaw output factor of the Clinac 21EX model, the most populous model, is shown in Figure 2-7. The data presented are representative of nearly all

the model measurement distributions. A fitted normal distribution and fitted student's t distribution are also plotted. The fitted normal distribution did not described the distribution well because of the relatively narrow peak and heavy tails, but the fitted student's t distribution described the data much better.

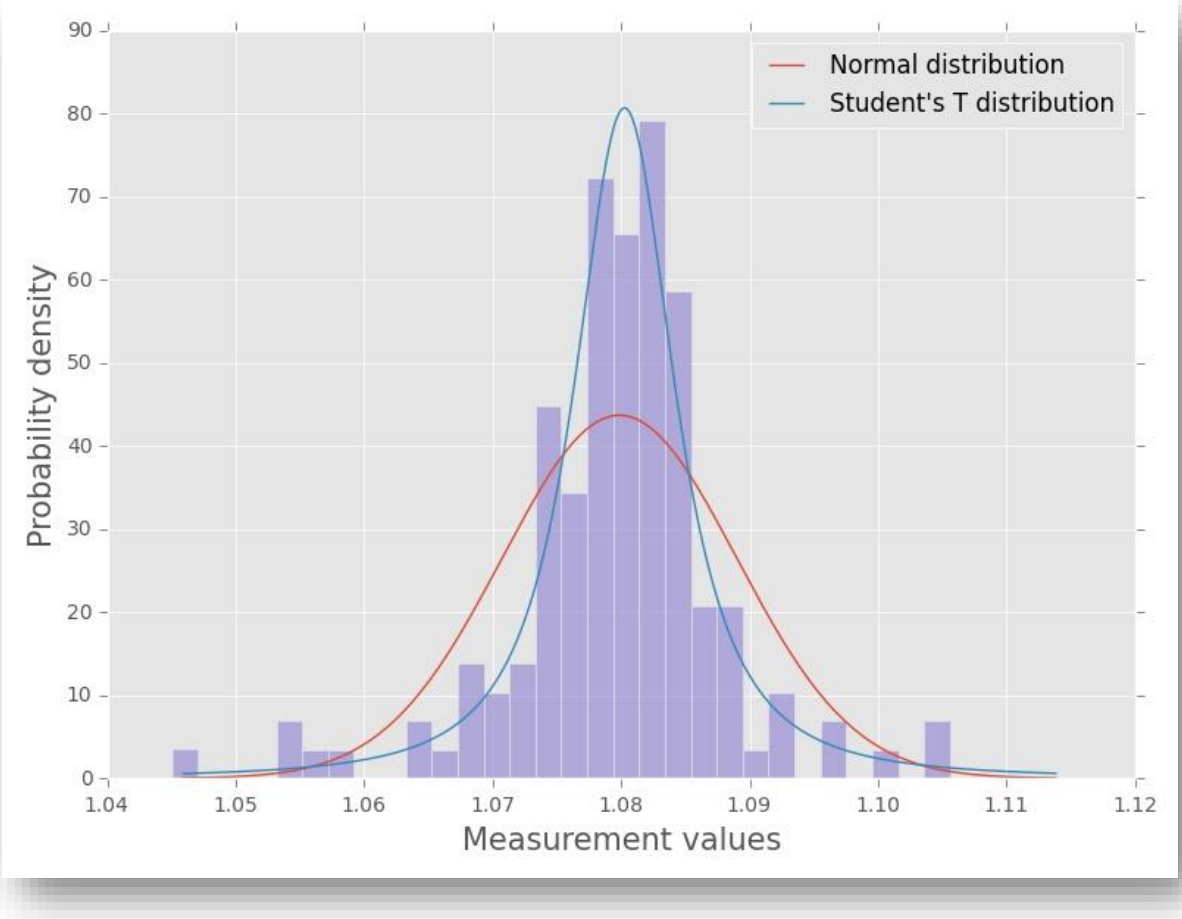

Figure 2-7. Measurement histogram for the 6 MV Clinac 21EX 30x30cm^2 jaw output factors along with fitted distributions.

A Kolmogorov-Smirnov (K-S) goodness of fit statistical test was also done to test whether the distribution could be said to be derived from either a normal or student's t distribution. The K-S test quantifies the maximum distance between the measured

distribution function and the reference distribution function. The distribution was statistically different than a normal distribution but not a student's t distribution (α=0.05). Since the data can reasonably be described by a student's t distribution, this suggests that data metrics like the mean, median, and standard deviation are valid statistics to describe our observed data.

The use of the median value over the mean was chosen due to the median's robustness to outliers for small datasets. As Figure 2-7 shows, there are a number of outliers in the data. Because the number of measurements per dataset varies, the mean describes large datasets well, but can be influenced by outliers in the smaller datasets. The median however is robust to these influences. For example, for the Clinac 21EX 30x30 cm^2 jaw output factor shown in Figure 2-7, the median and mean, with an *N* of 141, are both 1.080. However, the TrueBeam SBRT-style 4x4 cm^2 output factor, with an *N* of 9, had a median and mean of 0.856 and 0.859, respectively, which is a relative difference of 0.35%.

The spread in the data distributions are likely from three inextricable sources: machine delivery uncertainty, setup uncertainty, and measurement uncertainty. IROC-H maintains strict procedures for equipment setup and measurement, which minimizes as much as possible the contributions from these error sources, while delivery uncertainty is not easily controlled. Given that the distribution is symmetric and unimodal it is reasonable to assume that systematic error was minimized and the three sources contribute to random error.

Chapter 3: Reference photon dosimetry data for Elekta accelerators

3.1. Introduction

Accurate measurement of linear accelerator (linac) dosimetric characteristics is the foundation of good radiotherapy dose calculation and delivery. Dose calculations for radiation therapy patients are done through a TPS model, thus there is high priority on

ensuring that the TPS model is as accurate as possible. The TPS model should be based on the dosimetric performance of the machine(s) that it represents. It is the duty of the medical physicist to ensure that a high level of agreement is maintained between the TPS model calculations and measured physical linac characteristics. It is also helpful to check measurements against similar linacs. In this way the measurements from a single linac can be validated as being reasonable; i.e. a “sanity check” against other similar machines.

It is the goal of IROC-H to ensure consistency in the delivery of radiation for clinics involved in clinical trials. One way this is done is by visiting institutions and independently measuring their linac dosimetric characteristics. The data are then compared to the institution’s TPS, providing an independent dosimetry audit. The level of agreement between the measurements and TPS model indicate how well the institution has modelled their linac(s).

Since IROC-H has collected data on numerous linacs over many years, it was thought the distribution of linac photon beam characteristics should be studied and presented to the community which would allow individual radiotherapy institutions to compare their measurement data to IROC-H’s large dataset of dosimetry data. In this way, institutions can compare themselves not to a single value, but a distribution so that differences can be understood in the context of the entire community. Our previous study looked at Varian linacs (see Chapter 2); in this study we analyzed Elekta linacs.

Our goal was to analyze the characteristics of each Elekta linac model and compare them to the other models. Similarly performing models would be grouped together into classes based on statistical and clinical criteria as outlined in chapter 2. Furthermore, the level of agreement between the linac and institution TPS model was studied to understand areas of common agreement and disagreement. Lastly, this agreement was categorized by TPS to determine areas where individual TPSs excelled and underperformed.

3.2. Materials and Methods

3.2.1. Data Collection

The collection of data involved two sets of values: the independently-acquired linac measurements and the institution's TPS calculation values for the same conditions. IROC-H physicists went to institutions with their own equipment to make point measurements in a water phantom for simple geometric conditions. The institution was then asked to provide TPS calculation values for the same geometric conditions, which gave a direct comparison of measurement to calculation.

To collect the measurement data, IROC-H physicists used a 30x30x30cm^3 water phantom at 100cm source-to-surface distance. All data collected were point measurements. Percentage depth doses (PDD) were measured for 6x6 cm^2, 10x10 cm^2, and 20x20 cm^2 field sizes, each at 5, 10, 15, and 20 cm depths; at 10x10 cm^2 a d_{max} measurement as also taken. Off-axis measurements were done with a 40x40 cm^2 field size at 5, 10, and 15 cm off-axis at d_{max}. Universal wedge factors were measured for the 60 wedge in a 10x10 cm^2 field at 10cm depth and in a 15x15 cm^2 field at 15cm depth. Two sets of output factors were measured on Elekta linacs. The first set, labeled "midsize", was taken at d_{max} for 30x30, 20x20, 15x15, 10x10, and 6x6 cm^2 field sizes. The second set, labeled "small", was taken at 10 cm depth for 10x10, 6x6, 4x4, 3x3, and 2x2 cm^2 field sizes. IROC-H's output measurements must be consistent across vendors and linac head configurations, which explains the two sets of measurements despite both being simple output factors. All measurements except the small output factors were taken with a Standard Imaging Exradin A12 (Standard Imaging, Madison, WI) ion chamber; small field output factors were measured with an Exradin A16 microchamber.

Although data have been collected for other energies, 6, 10, 15, and 18 MV are by far the most widely used energies and are thus presented here.

3.2.2. Data Analysis

All data analysis and visualization was done using the general programming language Python and the open-source “pandas” Python package.[14]

While it is common to differentiate linacs by model name, Elekta’s linac construction configuration allows linacs of the same model name to have different head configurations. Since the head components determine the dosimetric characteristics, linac aggregation was performed using the head configuration rather than the linac nominal model name.

To categorize the linacs into dosimetrically similar classes, linacs were first grouped together by head configurations. Each energy was analyzed separately and considered independent. Two criteria were then applied to compare dosimetric comparability at the given energy: clinical and statistical. A clinical criterion was added because small but statistically significant differences were not deemed to have an effectively different dosimetry value. Each group’s mean value for each measurement location was compared to the others’ (e.g. PDD(10x10cm^2, 5cm)). Using ANOVA and Tukey’s honest significant difference test, groups were tested to see if there was statistical significance between the means of any measurement location. Additionally, the clinical criterion tested if the median values had at least a 0.5% local difference. This value was chosen since it is approximately equal to the standard deviation seen in IROC-H measurements. If a linac group had a statistically different and clinically different value from the other groups, it was put into its own class.

The resulting linac classes were compared to each other to highlight where differences and agreements occurred. Linacs were also compared to their institution TPS calculations to determine areas of common disagreement. Each linac measurement set was compared to

its own TPS, but the results were binned according the linac class. Finally, the measurement-to-TPS results were binned according to TPS vendor to sift out differences.

3.3. Results

3.3.1. Model Comparison

There were 4 Elekta linac head configurations analyzed at each energy: BMod, MLCi, MLCi2, and Agility. Due to physical construction limitations, measurements of the BMod head were not perfectly comparable to the other heads and thus the BMod linacs were placed immediately into their own class at each energy. The MLCi and MLCi2 heads were within the comparability criteria and were combined into an “MLCi” class. The manufacturer claimed that the MLCi2 head was equivalent to the MLCi head and the results agreed within our stringent criteria.

The resulting classes were the same at all energies with each head. Thus, there were three classes, named for the head they represented: Agility, MLCi, and BMod, where the MLCi class represented the MLCi and MLCi2 head and the others represented their own head configuration.

Figure 3-1 shows an overall comparison of dosimetric performance between the Agility and MLCi classes. Because the measurement conditions of the BMod head were not the same as the others, it is left out of the overall comparison. The figure is intended to quickly show differences between two of the three classes. The shade of the squares represents the maximum difference that class had for that parameter in comparison to the base class. Lighter shades indicate smaller differences from the base class, while darker shades indicate more pronounced differences. Differences between classes of machines should be noted, especially when trying to match machines of different classes, or deciding how many TPS beam models to create.

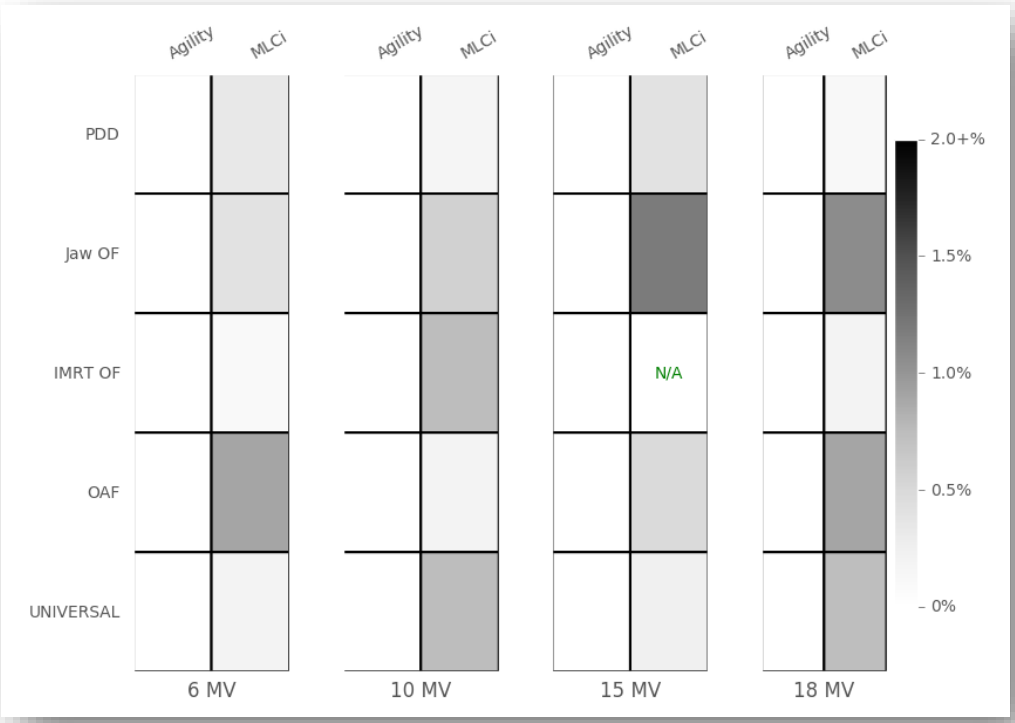

Figure 3-1. Heatmap comparing the Base class to other classes by parameter. Parameters include percent depth dose (PDD), jaw-based output factor (Jaw OF), IMRT-style small field output factors (IMRT OF), SBRT-style small field output factors (SBRT OF), off-axis factors (OAF), and wedge factors for enhanced dynamic and "upper" physical wedges. Darker color indicates larger maximum differences from the base class. "N/A" means that measurements were not available for that parameter.

3.3.2. 6 MV

For 6 MV, and all other energies, data are described in two ways. The difference between classes, called the interclass difference or variability, represents the local difference of the median values. The average interclass difference is the mean of the differences at all the field sizes or depths. Difference can also be described within the class, which we labeled intraclass variability, which is synonymous with the coefficient of variation.

The 6 MV summary data is given in Table 3-1. Further statistical metrics are given in the online content. The BMod class, with a different head design, could form similar, but not the exact same open field shapes, and thus should be taken into consideration when directly comparing classes. The most common class was MLCi, with 40 sets of measurements. The PDDs were all quite similar, having an average interclass variation of 0.7%. The MLCi class had a consistently softer spectrum than the other two classes although the difference was small. The average PDD intraclass variation was 0.8% for all the classes. Jaw output factors also showed close agreement. The MLCi class however had a much larger intraclass variation than did the Agility. Off-axis factors was where the classes deviated from each other, with the Agility class having less increase with distance off-axis compared to the MLCi, especially at 15cm. Off-axis factors were also where the measurements had the largest intraclass variation, having 0.8% and 1.1% for the Agility and MLCi class respectively. No off-axis measurements were taken for the BMod class. The IMRT-style output factors attained very close values for the Agility and MLCi classes, having no more than 0.1% interclass variation. The BMod measurements are somewhat similar but show differences from the other classes. Since these measurements were not at the same geometry, direct comparison should be limited. Universal wedge factors also showed good agreement between Agility and MLCi, having an average of 0.7% interclass variation but a relatively large 2.5% and 1.8% intraclass variation for Agility and MLCi, respectively.

		Class		
		Agility	MLCi	Bmod*
PDD	N	10	40	6
	5 cm	0.871 (0.004)	0.868 (0.004)	0.872 (0.006)
	10 cm	0.677 (0.004)	0.675 (0.004)	0.678 (0.004)
	15 cm	0.519 (0.004)	0.516 (0.004)	0.5205 (0.004)
	20 cm	0.396 (0.004)	0.393 (0.004)	0.3965 (0.004)
	N	8	33	6

Jaw OF	6x6 cm	0.965 (0.004)	0.965 (0.008)	0.968 (0.006)
	15 x 15 cm	1.030 (0.004)	1.026 (0.008)	1.025 (0.012)
	20 x 20 cm	1.054 (0.005)	1.052 (0.008)	N/A
	30 x 30 cm	1.075 (0.006)	1.072 (0.009)	N/A
	N	10	39	0
OAF	5 cm left	1.021 (0.007)	1.025 (0.007)	N/A
	10 cm avg	1.051 (0.009)	1.054 (0.013)	N/A
	15 cm left	1.052 (0.009)	1.061 (0.015)	N/A
	N	5	11	6
IMRT OF	2x2 cm	0.793 (0.006)	0.792 (0.005)	0.756 (0.006)
	3x3 cm	0.837 (0.007)	0.836 (0.003)	0.839 (0.001)
	4x4 cm	0.871 (0.002)	0.870 (0.003)	0.866 (0.004)
	6x6 cm	0.925 (0.002)	0.924 (0.002)	0.928 (0.002)
	N	3	18	4
Universal	60 deg	0.270 (0.006)	0.268 (0.005)	0.271 (0.004)
	60 deg (15x15)	0.277 (0.008)	0.277 (0.005)	N/A

Table 3-1. 6 MV measured Elekta data. Median values are given and the standard deviation is given in parentheses. N is the number of measurements. *The BMod head cannot form the same exact field sizes as the other classes; PDD data was taken at 10.4x10.4cm^2, jaw output factors at 6.4x6.4, 10.4x10.4, and 15.2x15.2cm^2; IMRT-style output factors at 6.4x6.4, 4x4, 3.2x3.2 and 1.6x1.6cm^2.

3.3.3. 10 MV

The 10 MV results are shown in Table 3-2. PDD results were very similar across classes with an average intraclass variation of 0.4% and 0.7% for the Agility and MLCi class, respectively. The Agility class had larger output factors above 10x10cm than the MLCi class, although the larger field sizes had a much larger intraclass variation. Off-axis factors also had large intraclass deviation although the medians were quite similar. For the IMRT-style output factors, the Agility class had consistently lower values; the values were not statistically different although the 2x2cm^2 value was clinically different. The universal wedge factors proved to also be clinically significantly different.

	Class		
	Agility	MLCi	Bmod*
N	8	25	6

PDD	5 cm	0.910 (0.004)	0.908 (0.005)	0.911 (0.009)
	10 cm	0.727 (0.002)	0.727 (0.004)	0.730 (0.006)
	15 cm	0.576 (0.002)	0.575 (0.005)	0.581 (0.006)
	20 cm	0.455 (0.002)	0.454 (0.004)	0.459 (0.004)
	N	6	20	4
Jaw OF	6x6 cm	0.964 (0.003)	0.964 (0.006)	0.973 (0.005)
	15 x 15 cm	1.035 (0.002)	1.029 (0.005)	1.019 (0.005)
	20 x 20 cm	1.055 (0.014)	1.052 (0.007)	N/A
	30 x 30 cm	1.074 (0.018)	1.070 (0.005)	N/A
	N	6	19	0
OAF	5 cm left	1.030 (0.014)	1.030 (0.010)	N/A
	10 cm avg	1.044 (0.017)	1.043 (0.013)	N/A
	15 cm left	1.053 (0.023)	1.051 (0.017)	N/A
	N	3	9	5
IMRT OF	2x2 cm	0.793 (0.011)	0.800 (0.008)	0.750 (0.003)
	3x3 cm	0.857 (0.003)	0.858 (0.004)	0.863 (0.002)
	4x4 cm	0.891 (0.004)	0.894 (0.006)	0.887 (0.002)
	6x6 cm	0.936 (0.004)	0.939 (0.003)	0.940 (0.001)
	N	5	18	4
Universal	60 deg	0.289 (0.001)	0.283 (0.006)	0.287 (0.006)
	60 deg (15x15)	0.290 (0.008)	0.282 (0.005)	N/A

Table 3-2. 10 MV measured Elekta data. Median values are given and the standard deviation is given in parentheses. N is the number of measurements. *The BMod head cannot form the same exact field sizes as the other classes; PDD data was taken at 10.4x10.4cm2, jaw output factors at 6.4x6.4, 10.4x10.4, and 15.2x15.2cm^2; IMRT-style output factors at 6.4x6.4, 4x4, 3.2x3.2 and 1.6x1.6cm^2.

3.3.4. 15 MV

The 15 MV measurement data are shown in Table 3-3; no BMod data were available. The Agility class PDD, unlike 6 and 10 MV, was slightly harder than the MLCi class. The data were not statistically significant but the 10 cm measurement pair was clinically significant. The off-axis factors were mostly similar, except at 10cm where the medians had a local difference of 0.5%. No IMRT-style output factor data was available for either class, nor were they available for the Agility universal wedge factors.

	Class	
	Agility	MLCi
N	3	10

PDD	5 cm	0.931 (0.003)	0.934 (0.003)
	10 cm	0.755 (0.003)	0.759 (0.003)
	15 cm	0.608 (0.002)	0.610 (0.003)
	20 cm	0.489 (0.002)	0.490 (0.003)
	N	3	8
Jaw OF	6x6 cm	0.957 (0.001)	0.955 (0.004)
	15 x 15 cm		1.033 (0.001)
	20 x 20 cm	1.052 (0.004)	1.057 (0.003)
	30 x 30 cm		1.072 (0.008)
	N	3	9
OAF	5 cm left	1.033 (0.006)	1.033 (0.009)
	10 cm avg	1.061 (0.002)	1.056 (0.010)
	15 cm left	1.081 (0.015)	1.080 (0.012)
	N		
IMRT OF	2x2 cm		
	3x3 cm	N/A	N/A
	4x4 cm		
	6x6 cm		
	N		4
Universal	60 deg	N/A	0.271 (0.001)
	60 deg (15x15)		0.277 (0.001)

Table 3-3. 15 MV measured Elekta data. Median values are given and the standard deviation is given in parentheses. N is the number of measurements.

3.3.5. 18 MV

The 18 MV results are shown in Table 3-4. The PDDs of the two classes are very similar, with an average interclass variation of 0.2%. Jaw output factors had larger interclass variability and the Agility class had large intraclass variability with an average of 1.8%. The MLCi off-axis factors had less variability with increasing distance from the central axis than did the Agility class although the intraclass variation was similar at 1.1% and 1.2% for Agility and MLCi, respectively. Notably, the IMRT-style output factors had a much lower intraclass variability for the MLCi class at 0.6%, which is much closer to the other parameters.

	Class	
	Agility	MLCi
N	5	24
5 cm	0.955 (0.006)	0.956 (0.006)

PDD	10 cm	0.782 (0.005)	0.783 (0.005)
	15 cm	0.635 (0.005)	0.635 (0.005)
	20 cm	0.513 (0.005)	0.515 (0.004)
	N	5	20
Jaw OF	6x6 cm	0.956 (0.009)	0.957 (0.008)
	15 x 15 cm	1.023 (0.024)	1.034 (0.005)
	20 x 20 cm	1.055 (0.013)	1.058 (0.005)
	30 x 30 cm	1.066 (0.027)	1.077 (0.005)
	N	5	21
OAF	5 cm left	1.036 (0.009)	1.039 (0.011)
	10 cm avg	1.059 (0.010)	1.055 (0.012)
	15 cm left	1.044 (0.014)	1.053 (0.014)
	N		5
IMRT OF	2x2 cm	N/A	0.786 (0.010)
	3x3 cm		0.865 (0.005)
	4x4 cm		0.903 (0.003)
	6x6 cm		0.944 (0.001)
	N		10
Universal	60 deg	N/A	0.272 (0.005)
	60 deg (15x15)		0.275 (0.005)

Table 3-4. 18 MV measured Elekta data. Median values are given and the standard deviation is given in parentheses. N is the number of measurements.

3.3.6. Measurement and TPS agreement

The results of the comparison between an institution's dosimetric characteristics and their TPS calculations for the Elekta site visit conditions are shown in Figure 3-2. Although each institution was compared to its own TPS, the results were binned according to the linear accelerator's resultant class. The shades of grey in Figure 3-2 denote the level of agreement, with darker shades indicating worse agreement. Notably, the parameters with good agreement at one energy may have poor agreement at another. The Agility class had good agreement for PDDs and the larger field output factors across all energies as did the MLCi class for energies 10MV and greater. However, the IMRT-style output factors had a wide range of agreement. The 6 MV Agility class' values disagreed sharply while all other energies' values were well within agreement criteria. Nearly all off-axis factors had a large

standard deviation although the median agreement was always within tolerance. The universal wedges, as the IMRT-style output factors, had a large range of agreement. For the MLCi class' wedge factors, the agreement was generally worse than the Agility class.

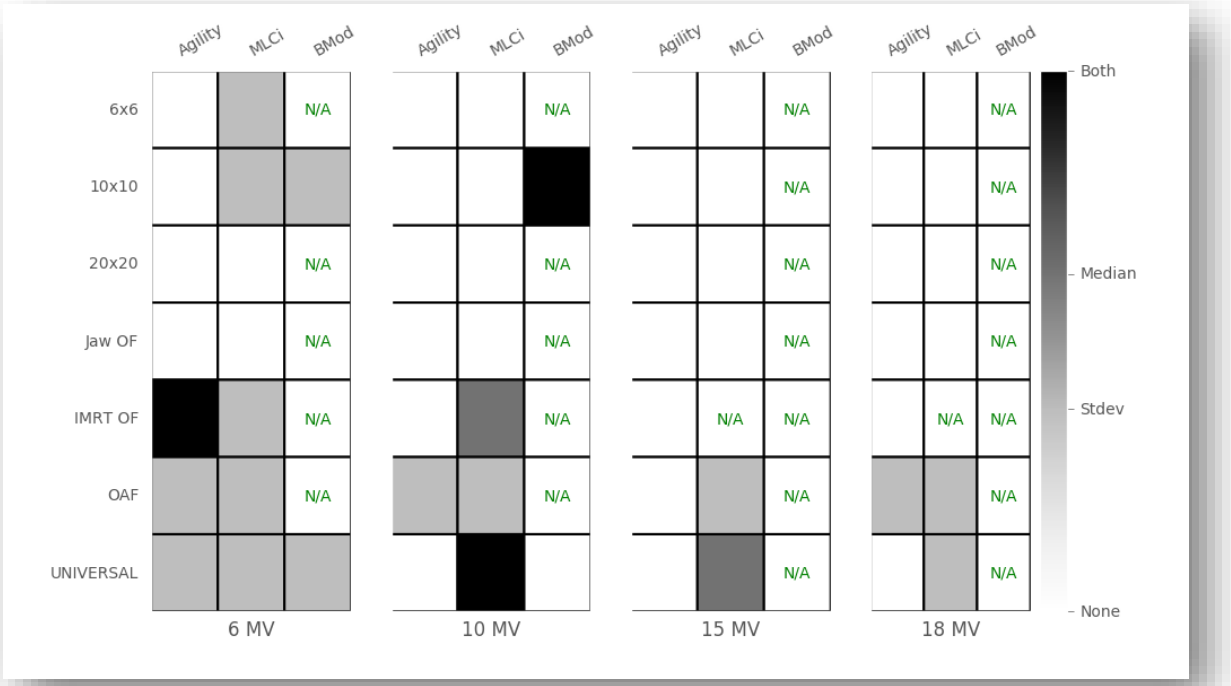

Figure 3-2. Agreement between linear accelerator dosimetry characteristics and the institution's TPS, sorted according to the accelerator class. The shade of gray denotes the level of agreement with the TPS calculations, with darker shades indicating greater disagreement. White indicates that the both the standard deviation and median agreement were good. Lighter grey indicates that the standard deviation of agreement across institutions was more than 1%; dark gray indicates the median TPS-to-measurement ratio across institutions was more than 1% from unity. Black indicates both the standard deviation and median values were above their respective thresholds. N/A indicates that not enough comparisons were available for that energy/parameter.

3.3.7. Agreement by TPS

The agreement of an institution's accelerator measurement and TPS calculation was studied to see if there was a difference in agreement between the two most common TPSs and the

results are shown in Figure 3-3. Results were binned regardless of linac class by the institution's TPS. Agreement results were mixed between the TPSs. At 6 MV, Eclipse had both poor median and standard deviation agreement for the off-axis and universal wedge factors while Pinnacle had only a large standard deviation. Eclipse however had good agreement for the 6x6cm^2 PDD and jaw output factors. At 10 and 15 MV, Eclipse had good agreement for all parameters while Pinnacle had a few areas of disagreement.

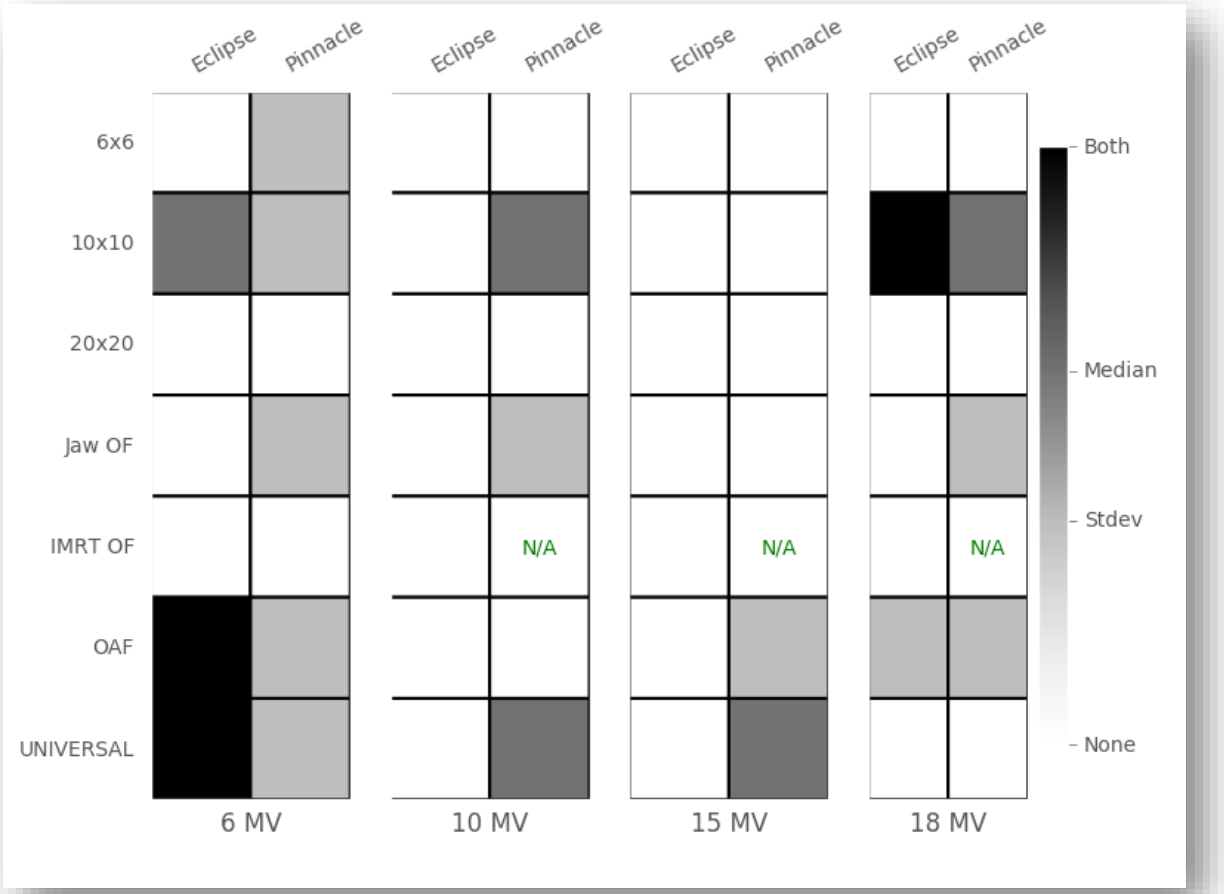

Figure 3-3. Differences between the linear accelerator dosimetric characteristics and the institution's TPS, according to each TPS and grouped by energy. The shade of gray denotes the level of agreement with the TPS calculations, with darker shades indicating greater disagreement. White indicates that the both the standard deviation and median agreement were good. Lighter grey indicates that the standard deviation of agreement across institutions

was more than 1%; dark gray indicates the median TPS-to-measurement ratio across institutions was more than 1% from unity. Black indicates both the standard deviation and median values were above their respective thresholds. N/A indicates that not enough comparisons were available for that energy/parameter.

3.4. Discussion

The goal of this study, as with our prior study of Varian machines, was to combine linacs into dosimetrically equivalent classes. The idea being that several different models could be aggregated into a distinct few. Additionally, the dosimetric characteristics of these classes relative to one another was of interest. Note that comparing characteristics does not imply one class is superior, only that there may or may not be significant differences between them. Finally, reference data was compiled so that physicists can compare their own measurement values to these reference distributions.

An important question for the reader to consider is how to use the data provided. Reference datasets like these show what the average linac dosimetric characteristics are, and thus can only give so much information about an individual machine; individual measurements may deviate from the median values given in this study. Inevitably, when an individual machine performs slightly differently than a reference dataset the question of how significant that difference is becomes important. To provide a better understanding of the entire distribution, statistical metrics of the distribution such as median and standard deviation are given so that physicists are aware of parameters that may expect a wide or narrow range of values. In this way, the physicist can flag measurements that are outside a reasonable range. As well, the goal of the dataset is not to set an expectation that machines should be close to the median. Institutions may have specialized linacs. In such cases, the physicist should simply be aware that this is the case and that their datasets should not be expected to be similar to other linacs. However, there should be careful analysis of where

those differences are and why they exist. For non-specialized linacs any significant differences should be investigated. Measurement error can be found that is either systematic or one-off. Finally, the ultimate test of performance should be the agreement between the dosimetric measurement and TPS calculation. Reference datasets exist to validate those dosimetric measurements.

The data of Figure 3-2 showing TPS and measurement agreement and Figure 3-3 showing agreement by TPS are important ones for any physicist commissioning or validating their beam models. Assuming any differences between the reference datasets and an individual machine are well-understood, the agreement of the TPS is the next area of investigation. Figure 3-2 shows that each linac class has its own areas of agreement and disagreement. Specifically, the Agility class had good agreement for PDD and jaw output factors across all energies, while the other parameters had varying levels of agreement across energies. The MLCi class had disagreement of several parameters at each energy. Three of the 4 comparable parameters for the BMod class had disagreement of some kind. There is thus no linac class that is without areas of disagreement and care should be taken when measuring and validating those parameters. Figure 3-3 shows the agreement split by the institution's TPS for the Eclipse and Pinnacle TPSs. Notably, the IMRT-style output factors are all in good agreement where applicable. This contrasts with some of the results shown in Figure 3-2. The discrepancy arises due to not all the institutions reporting their TPS. So, Figure 3-3 is only a subset of the available data where the institution explicitly reported their TPS. This means that parameters in Figure 3-3 that are in good agreement might not actually be good in every case, and should only be taken as suggestive. Those parameters that are in poor agreement should however be noted and investigated since even a subset of the data suggests it is a problematic parameter.

3.5. Conclusion

In this study, Elekta linac measurements from all over the United States were analyzed and compiled to create reference datasets. Statistical metrics were given for each dataset so that physicists could understand not just the median value but the distribution and so recognize the expected variance of each measured parameter. Agreement between the treatment planning system and the dosimetric measurements was also studied by linac class and planning system.

The datasets generated in this study can be used as a second check against an institution's measurements to detect any anomalies. Any significant difference should be investigated and the reasons be well-understood or identified as measurement errors. No class of accelerator had better agreement with its respective TPS than another across all energies although energy-specific exceptions existed.

Many thanks go to Paul Holguin for querying and compiling the site visit data. This work was supported by Public Health Service Grant CA180803 awarded by the National Cancer Institute, United States Department of Health and Human Services.

Chapter 4: Agreement of institutional measurements and treatment planning systems

This chapter is based upon “Agreement between institutional measurements and treatment planning system calculations for basic dosimetric parameters as measured by IROC-Houston”, by J. Kerns, D. Followill, J. Lowenstein, A. Molineu, P. Alvarez, P. Taylor, and S. Kry, *International Journal of Radiation Oncology Biology Physics* (in press) (2016). The journal allows a student’s publication to be included in their dissertation.

4.1. Introduction

Obtaining accurate dosimetry data has always been a major goal in the field of medical physics. Although delivery methods have improved and evolved, it is still a challenge to match even basic dosimetry data between the radiation treatment machine and the treatment planning system (TPS). The percentage of institutions that pass an IROC-H phantom irradiation, as determined by head and neck anthropomorphic phantoms, has improved over time, but even with relaxed criteria, a relatively large number of institutions still fail to meet the minimum standards.[1] Reasons for failure vary; often several TPS factors may be involved, leading to an additive effect.[17] Although machine measurement data have been analyzed in numerous studies[3, 4, 6, 8], no large-scale, systematic comparison of machine data with TPS data has been done.

In an effort to ensure high-quality radiation therapy for patients in clinical trials, IROC-H has developed several ways to measure and confirm various aspects of radiation delivery accuracy. One of these ways is through on-site dosimetry reviews visits. During an on-site visit, an IROC-H physicist goes to the institution and, among other things, takes independent dosimetry measurements of the linear accelerators. These measured values

are compared with those calculated by the institution's TPS to assess how well the institution has modeled certain basic dosimetry parameters.

IROC-H measurements correspond with several tests recommended by the AAPM Multidisciplinary Program Planning Group 5 (MPPG-5) for basic photon validation in TPSs.[18] Owing to limitations in the beam modeling and dose calculation algorithm, TPS-calculated doses do not always perfectly agree with measured values. However, for basic photon parameters, the TPS calculated dose and the measured dose should agree to within 2% in the high-dose regions.[7] Given that these are calculations of basic photon dosimetry parameters, any disagreement discovered may have an impact on all radiotherapy patients. It is thus of the utmost importance that these basic parameters are modeled well in the TPS. Raising an awareness of TPS dosimetry parameters that have been found to disagree with measurements can help physicists focus their time and energy on verifying those parameters.

The goal of the current study is to compare acquired measurement dosimetry data with the institution's TPS calculation data to determine how institutions are actually faring as they work toward meeting MPPG-5's dosimetric agreement goal (i.e., 2%). Examination of these comparisons can identify common problem areas. Armed with this information, physicists can be more prepared when commissioning a TPS or a new linear accelerator.

4.2. Materials and Methods

4.2.1. Data Collection

Data collection involved two steps. First, measurement values were acquired during an IROC-H on-site dosimetry review visit, in which an IROC-H physicist used their own equipment to make point measurements in a water phantom for simple irradiation geometries. The institution's physicist was always present for data collection. Second, TPS-

calculated values were determined and provided by the institution's physicist for the same geometric conditions and points as were measured. In this way, a direct comparison of institution TPS-calculated values with independent machine measurements could be performed. Although institution linac measurement data was not required, the institution physicist was free to compare their results at the time of acquisition. Any large discrepancies in acquired values were investigated for validity. In the vast majority of cases when institution measurements were comparable, IROC-H's acquired values were similar.

The collection process and geometries of the point measurement data were discussed fully in our prior study.[19] In summary, all measurements were taken in a 30x30x30 cm water phantom at a source-to-surface distance of 100 cm. A Standard Imaging Exradin A12 (Standard Imaging, Madison, WI) ion chamber was used for all measurements except small fields that used the multileaf collimator (MLC). For such measurements, an Exradin A16 microchamber was used. The A16 has been shown to have minimal influence from spectrum changes for the field sizes measured under similar conditions.[20] Percentage depth dose (PDD) was measured for 3 field sizes: 6x6 cm^2, 10x10 cm^2, and 20x20 cm^2. For each field, a measurement was taken at 5, 10, 15, and 20 cm depth; at 10x10 cm^2 a d_{max} measurement was also taken. Output factors were sampled at 6x6, 10x10, 15x15, 20x20, and 30x30 cm^2 field sizes, all at 10 cm depth and corrected to d_{max} using the institution's own clinical PDD data. Off-axis measurements were taken at 5, 10, and 15 cm off-axis at d_{max} in a 40x40 cm^2 field. Wedge output factors were measured for the 45° and 60° enhanced dynamic wedge (EDW) for a 10x10 cm^2 field at 10 cm depth; additionally, a 45° EDW measurement was taken in a 15x15 cm^2 field at 15 cm depth. Two sets of small field MLC output factors were measured, representing fields that may be seen in both intensity-modulated radiation therapy (IMRT) and stereotactic body radiation therapy (SBRT), called "IMRT-style" and "SBRT-style" output factors respectively. IMRT-style fields were measured by fixing the jaws at 10x10 cm and varying the MLC field size to 6x6, 4x4, 3x3, and 2x2 cm^2,

representing various possible segment sizes. Measurements were normalized to an open 10x10 cm^2 field. SBRT-style measurements were taken using the same field sizes as for IMRT, but both jaws and MLCs were moved to the same position for each given field size.

Measurements were taken at all points described at all photon energies commissioned by the institution. Although more photon energies exist, the most common energies of 6, 10, 15, and 18 MV are presented.

4.2.2. Data Analysis

The goal of our analysis was to determine where institutional TPS calculated dosimetry data commonly agreed and commonly disagreed with the measured data and where agreement varied widely. In a prior study, analysis of IROC-H data collected between 2000 and 2014 for Varian machines resulted in the establishment of a number of machine classes. These classes were a result of clinical and statistical criterion to determine which machine models were dosimetrically equivalent. The resulting classes consolidated the number of datasets necessary to describe the Varian linear accelerators currently in service.[19] At each energy, the class that represented the most machine models was called the “base class”; e.g. at 6 MV this class represented the 21/23EX, 21/23iX and Trilogy platforms. Although each institution’s machine measurement point was compared to the institution TPS calculation point, the resulting ratios were binned according to the machine class; i.e. binned with machines shown to be dosimetrically equivalent.

Machine data were compared with TPS values by dividing IROC-H measurement values for that institution’s machine by institutional TPS calculation values at a given point, thus providing a ratio. This was done for every measurement point, machine, energy, and institution; more than 250 institutions and 500 machines were measured and compared. Two additional comparisons were done by separating results by TPS and by agreement

over time. For the TPS comparison, measurements of the base class, the most populous class, were separated by TPS. Sufficient data existed only to compare Pinnacle and Eclipse TPSs. To examine the agreement of parameters over time, we binned data from the base class into three time periods according to the site visit date: 2000-2005, 2006-2010, and 2011-2014.

Two sets of criteria were used to identify troublesome parameters. First, for each energy and class dataset, median values for a given parameter were tested for statistically and clinically significant differences from unity. That is, we tested to see which parameters (if any) had a systematic bias between the measured and calculated values. Statistical significance was measured using a Wilcoxon rank-sum test against the null hypothesis of unity ($\alpha = 0.05$). For clinical significance, a median value greater than 1% different from unity was deemed significant. Because of the large number of measurements, statistical significance was extremely easy to achieve and nearly all parameters reached significance, even for very small distances from unity. Thus, clinical significance became the dominant watershed for median comparison. Distribution differences that were statistically and clinically significant were thought to represent parameters that TPSs systematically did not model well.

The second criterion indicating a troublesome parameter was a ratio distribution with a standard deviation greater than 1%. Distributions with a large standard deviation, even when the median was close to unity, were thought to represent parameters that had a wide range of modeling discrepancies and no common agreement amongst institutions; as such, these parameters were considered poorly modeled or challenging to model, either by the vendor or physicist.

4.3. Results

4.3.1. Class Comparison

Figure 4-1 presents the fitted distribution density fits of dosimetric parameters (ratio of measurement to TPS value) for the 6 MV base class accelerator. The top

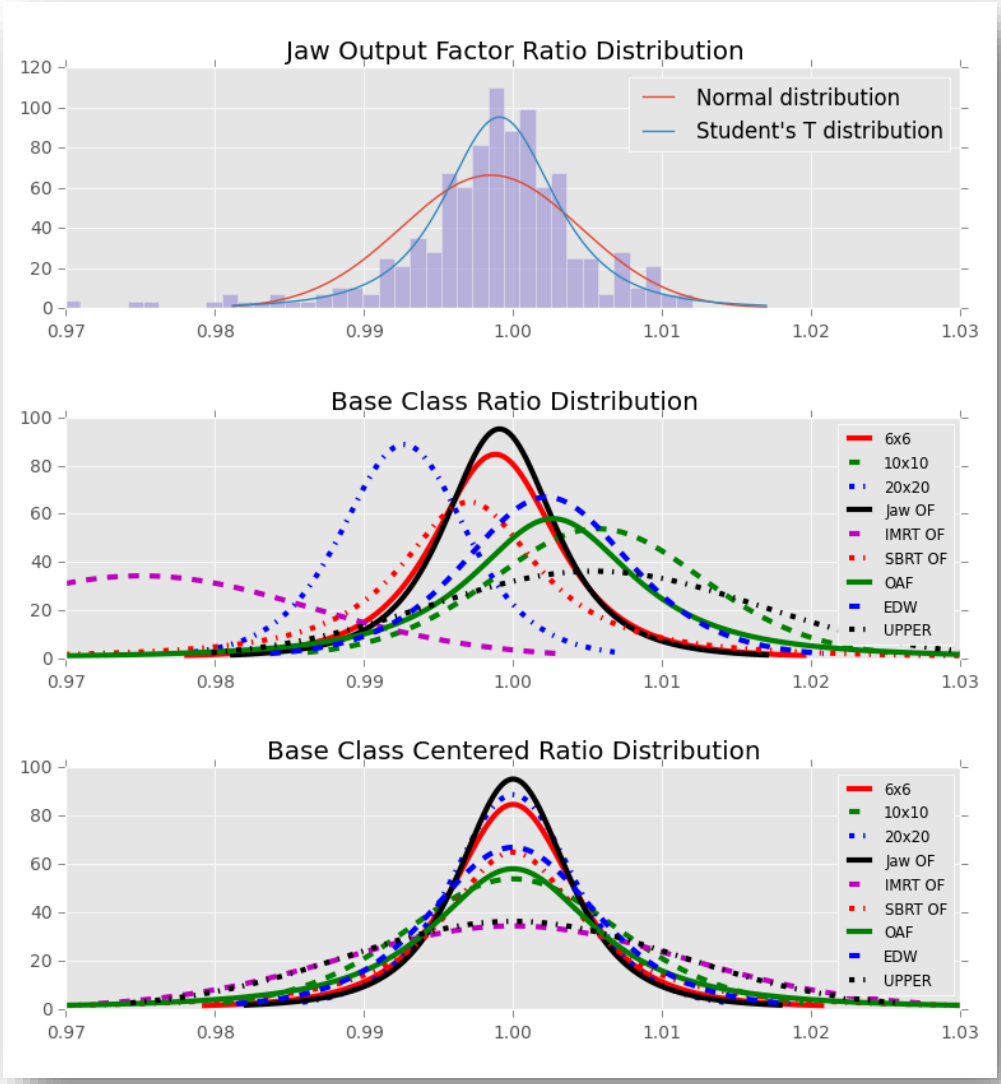

Figure 4-1. Density distributions of the ratio of machine measurement to TPS-calculated values. The top plot is a histogram of the base class jaw output factor ratios along with a fitted normal and student's t distribution. The lower two plots show fitted student's t distributions of all the parameters of the base class. Distributions in the middle plot are centered about the median measurement value while those in the bottom plot are centered about unity for visual comparison of the distribution spread. The 6x6, 10x10, and 20x20 cm^2 lines represent the field size for PDD measurements; "OF" indicates output factor; "OAF", off-axis factor.

plot shows a histogram of the base class jaw output factor ratios, along with a fitted normal and student's t distribution. To test which type of distribution best described the data, a Kolmogorov-Smirnov goodness-of-fit test was run against a normal and student's t distribution (α=0.05). The test rejected the null hypothesis that the data could be described by a normal distribution, but could not reject the student's t distribution. The student's t distribution was then used to represent a parameter's data for Figure 4-1. The middle plot shows the distributions centered at the median measurement value, and the bottom plot shows the same distributions centered about unity to visualize distribution width. Although there are several distributions for each parameter (e.g., for a given PDD there is a distribution at each of the evaluation depths, a.k.a. subparameter: 5, 10, 15, and 20 cm), only the distribution from the worst-performing subparameter is shown. Thus if the 5 cm depth distribution was the worst performing subparameter for 6 × 6 cm^2, it was the distribution plotted. This approach was more conservative than grouping all subparameter measurements together, which may wash out differences, and was more consistent with the MPPG-5 criteria of individual point comparison.[18] Systematic offsets in the measurement to TPS ratio can be seen in the middle panel, particularly for the small-field IMRT output factors, in which the TPS systematically overestimated the output compared with the measurement. Although the upper physical wedge output factors were also notably offset from unity, the median fell just within the 1% criteria. Other parameters typically had measurement to TPS ratios that were centered close to unity. The bottom plot shows that the IMRT-style output factors, as well as the upper physical wedge output factors, had the widest distributions, with >1% standard deviation. The off-axis factors also showed a relatively wide distribution, although these fell just within the 1% criteria. The jaw output, EDW, and PDD distributions were relatively tight.

The analysis of Figure 4-1 was generalized for all classes to produce a heat map, shown in Figure 4-2. Shaded boxes represent parameters that were identified as problematic,

either because of a median difference (dark shading) or a standard deviation greater than the specified criteria (light shading). Black boxes indicate that both the median and standard deviation were too high. As in Figure 4-1, each parameter's worst-performing subparameter distribution was chosen for analysis. The results from Figure 4-1 can be seen in the base class column in Figure 4-2: the upper physical wedge output factors and SBRT-style output factors had high standard deviations (gray boxes), and the IMRT-style output factors had high standard deviations and a systematic offset (black boxes).

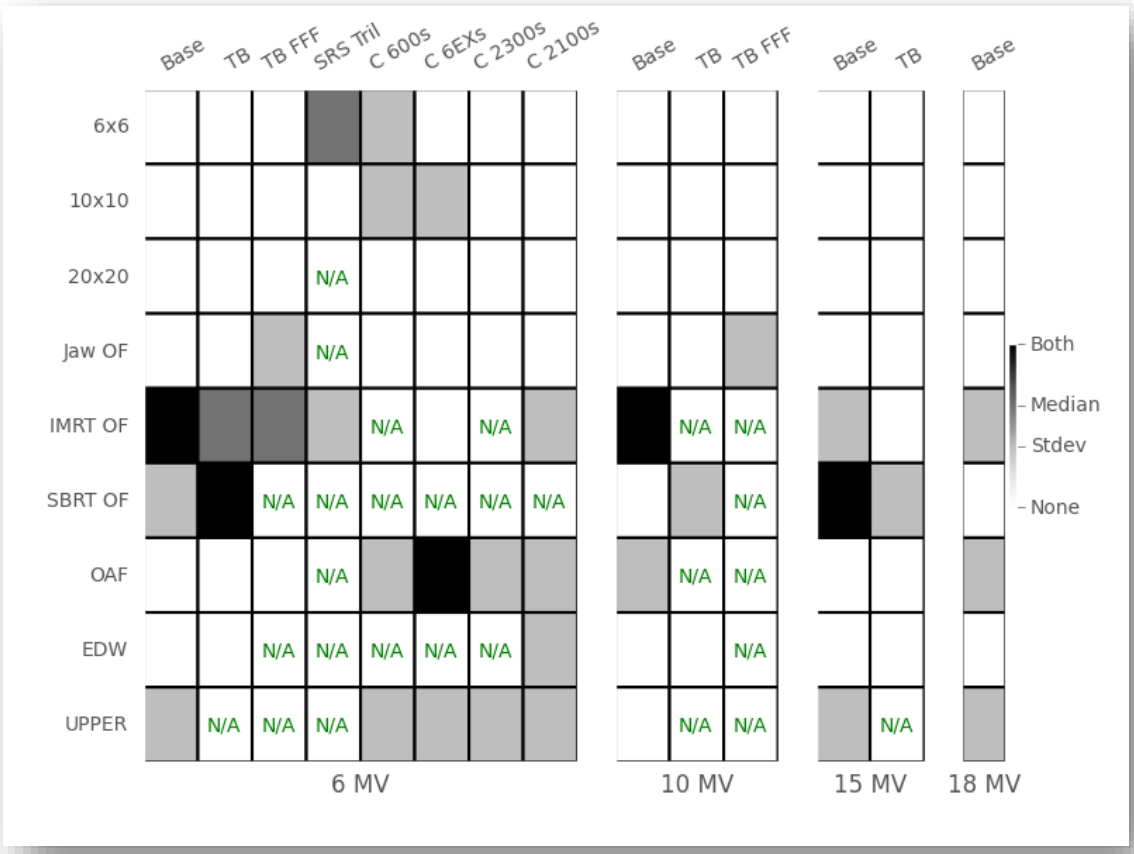

Figure 4-2. A heat map of differences between treatment planning system values and machine measurements, broken down by machine class. Shaded boxes represent distributions that had a median or standard deviation (or both) greater than the criteria described in the text. Median differences are shaded darker than high standard deviations only for visualization purposes. N/A indicates that not enough data were available for comparison. 6x6, 10x10, and 20x20 cm^2 represent the field size for PDD measurements; "OF" indicates output factor; "OAF", off-axis factor.

As can be seen in Figure 4-2, no class of accelerator was free from challenging parameters. Most of these challenging parameters were identified by the standard deviation criterion, and a handful had problematic median differences or both problematic standard deviations and problematic median differences. The 10-, 15-, and 18-MV energies performed similarly; most troublesome parameters were consistent across energies. However, this was not universally true. For the base class of accelerators, SBRT-style output factors ranged from thorough agreement at 10 and 18 MV to thorough disagreement at 15 MV.

In general, the worst-performing parameters were IMRT-style output factors, SBRT-style output factors, and upper physical wedge output factors, and the best-performing parameters were PDD, EDW, and jaw output factors.

4.3.2. TPS Comparison

To determine the effect of the TPS used on measurement to TPS model agreement, the machines of the base class of accelerators were split according to the institution's reported TPS. Figure 4-3 shows the results of the analysis for the Eclipse and Pinnacle TPSs. Although other TPSs have been recorded, these TPSs account for the vast majority used clinically. These results show similar but not identical problems between the TPSs. Eclipse data showed larger standard deviations than Pinnacle data for several 6 MV parameters, whereas Pinnacle had more troublesome parameters than Eclipse data at 10 and 15 MV. Both TPSs accurately modeled PDD, EDW, and jaw output factors and had trouble modeling the IMRT-style output factors at 6, 10, and 18 MV.

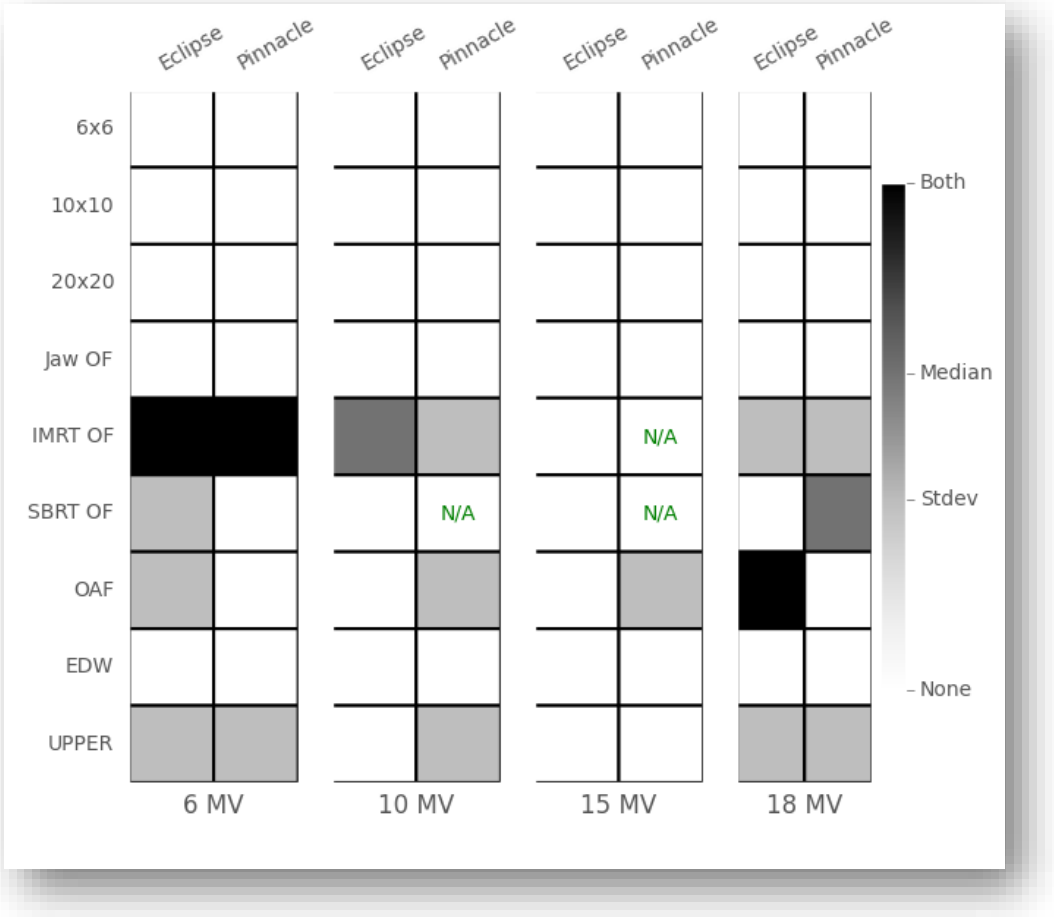

Figure 4-3. Ratios of machine measurement and treatment planning system-calculated values broken down by treatment planning system and energy. 6x6, 10x10, and 20x20 cm^2 represent the field size for PDD measurements; "OF" indicates output factor; "OAF", off-axis factor.

4.3.3. Time Period Comparison

Figure 4-4 shows the measurement to TPS ratios for the base class of accelerators according to the time period of the site visit. The data clearly show that the parameters with the worst agreement have always had the worst agreement, and agreement has not improved with time; only agreement for the 10 MV 10 × 10 cm^2 PDD distribution has changed since 2000, and it got worse.

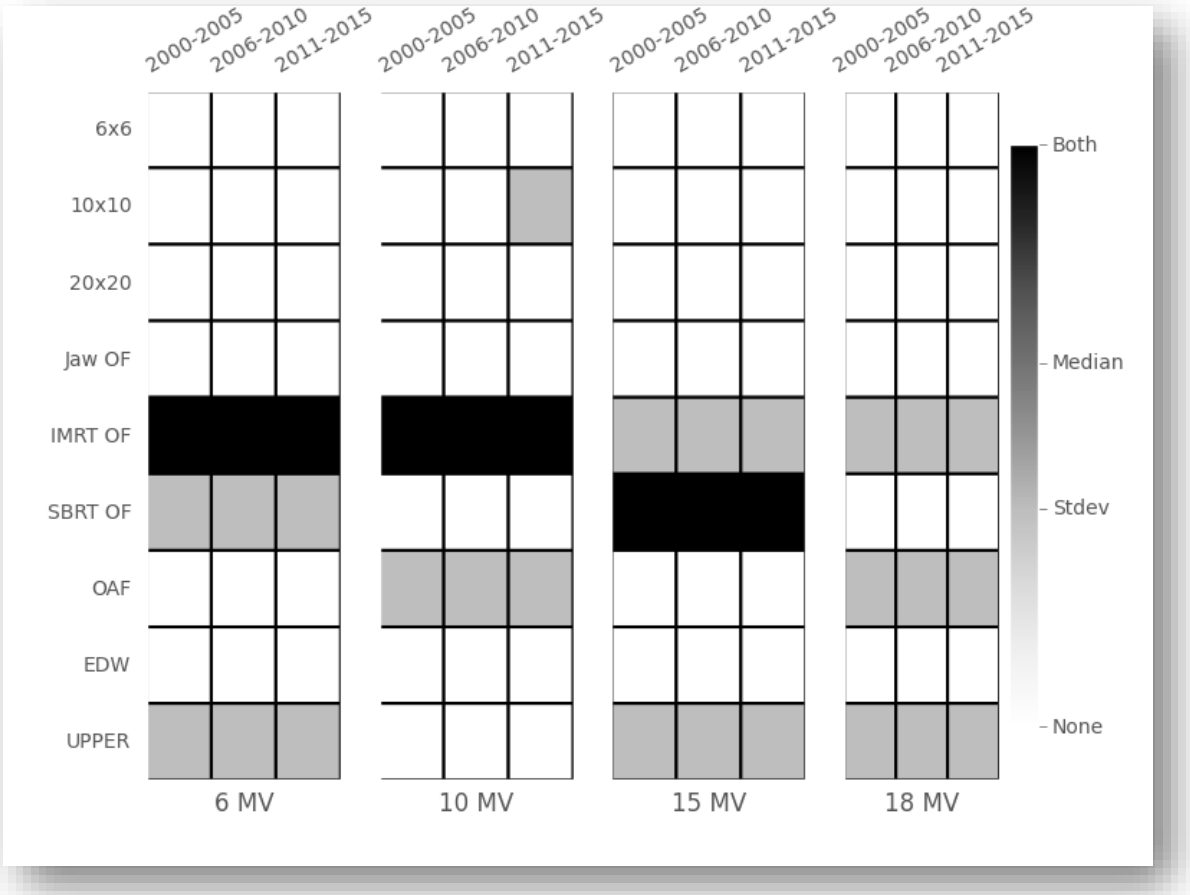

Figure 4-4. Ratios of machine measurement and treatment planning system-calculated values broken down by energy and time period of the site visit. 6x6, 10x10, and 20x20 cm^2 represent the field size for PDD measurements; "OF" indicates output factor; "OAF", off-axis factor.

4.4. Discussion

Our study highlights areas of common agreement and disagreement between linear accelerator measurements and TPS calculated values. PDD and jaw output factors nearly always showed good agreement, but IMRT- and SBRT-style output factors and upper physical wedge output factors generally did not show good agreement. Although some of these results may not be surprising, given that institutions have long reported various disagreements between measurements and TPS values[10, 11, 21], our findings more

specifically characterize the disagreements, i.e., whether the disagreement is systematic (large median difference from unity) or represents a wide range of disagreement (large standard deviation).

We found the most pronounced disagreements for the IMRT-style and SBRT-style small field output factors. The measured 6 MV IMRT output factor values in particular were consistently lower than the TPS values across a large number of commissioned TPSs, having an average discrepancy of 1.6% for all field sizes, and 64% of measurements having a discrepancy of over 1%. For SBRT-style output factors the results were slightly better with an average 6 MV discrepancy of 0.5% and 38% of measurements with a >1% discrepancy. However, the fact that nearly all class distributions also had a large standard deviation highlights the wide range of output factors physicists use in TPS models.

Upper physical wedge distributions nearly always had a large standard deviation across all energies, whereas EDW distributions nearly always had good agreement. Because EDW output factors are based on open field measurements, the agreement is not surprising. Physical wedge output factors require more input from the physicist; additionally, because physical wedge output factors are less commonly used in the era of IMRT, the physicist may not require the same level of accuracy as for open fields or may not spend as much time modeling. Of note, IROC-H evaluations are performed along the central axis only; off-axis wedge values may disagree even more. Implementing EDWs in place of physical wedges would reduce the chance of dosimetric error.

Although we observed some differences between the Eclipse and Pinnacle TPSs (Figure 4-3), neither TPS outperformed the other across all energies. Our analysis did not take into account the TPS version number, and it is possible that stronger differences are present for specific TPS versions.

Perhaps most notable of our findings is the consistency of distributions across time. Parameters that were problematic a decade ago are still problematic. Note that data were

binned by site visit date, and some sites were visited more than once. Furthermore, the data may be influenced by institutions that initially commissioned their TPS and never adjusted it for new machines or TPS versions. Still, physicists continue to struggle to accurately model their machines despite advances in accelerator manufacturing technology and TPS modeling. The lack of improvement in TPS agreement is most concerning because new radiation therapy techniques such as stereotactic radiosurgery and volumetric modulated arc therapy have become more common. These techniques generally require higher levels of TPS accuracy, especially for small fields. Therefore, physicists commissioning or adjusting a TPS model should seriously investigate the differences between their TPS and machine.

Given the tolerances of the AAPM MPPG-5 report[18], most institutions are in compliance for most basic dosimetric parameters. However, the tolerances given in the AAPM report are intended to be the maximum allowable difference between measurements and TPS values. A few parameters approach or exceed these tolerances, even on average, and physicists should carefully review these parameters. The systematic disagreement may be due completely or in large part to TPS physics modeling limitations. Improperly measured model input data may also be a factor, although when the institution had comparable acquisition data it was usually similar to IROC-H data. The results presented here can be used as a guide to identify parameters that should be given more time and attention so that conformance to MPPG-5 is assured.

Ultimately, we cannot make sweeping conclusions about why a measured parameter has poor agreement with the TPS model because there could be numerous reasons, including data collection, beam modeling, and TPS limitations. Our data suggest that physicists should spend additional time examining the problem parameters of their machine, according to its machine class. However, no matter which machine an institution has, IMRT- and SBRT-style output factors and upper physical wedge output factors should be carefully

modeled. Future research would include determining non-dosimetric TPS settings that may influence model agreement as well as whether institutions have improved over multiple visits.

4.5. Conclusion

This study examined the agreement between radiation machine measurement and TPS values for basic dosimetric parameters. Parameters that disagreed between measurement and TPS value were highlighted by machine class. Small differences were found between TPSs, but neither TPS examined uniformly outperformed the other. Agreement was also found not to change with time; problem parameters have always, and continue to be, problem parameters.

Chapter 5: TPS calculation errors are the leading cause of IROC-Houston phantom failures

5.1. Introduction

Accuracy in treatment planning of radiation therapy is extremely important.[22] Errors, shortcomings, and limitations in the beam models of the treatment planning system cause differences between what was planned and what was actually delivered to the patient. New technologies like intensity-modulated therapy (IMRT) and volumetric modulated arc therapy (VMAT) allow unprecedented plan conformality and dose distributions. It is imperative that the delivered dose distribution actually matches the distribution planned.

One widely used test to ensure that planned and delivered doses agree is the anthropomorphic phantom program of the Houston branch of the Imaging and Radiation Oncology Core (IROC-Houston). IROC-Houston has been charged with ensuring that institutions participating in clinical trials delivery radiation dose safely and accurately. In the anthropomorphic phantom program, the institution irradiates an IROC-Houston phantom containing thermoluminescent dosimeters (TLD) and radiochromic film.[23-25] The institution-calculated dose distribution is compared to the measured dose distribution and passes or fails the irradiation based on the agreement.

Despite the advances in delivery, localization, and imaging, irradiation pass rates have only risen modestly, reaching ~90% in recent years.[1] This rate is concerning because IROC-Houston's current criteria are looser than most institutional criteria (7% TLD agreement and 7%/4mm gamma criteria) There are many reasons why an institution may fail the phantom test, including setup or positioning errors, linac delivery performance, or beam modeling errors in the treatment planning system. One limitation of the phantom program is that it is an end-to-end test, so understanding the underlying causes of disagreement between measured and calculated doses are very difficult to identify. To determine the cause of the discrepancy, the institution physicist has relatively little

information to start from. Although setup errors are easy to spot, they are relatively rare; most often, the dose is systematically different from the calculation.

To better inform the institution of where problems may lie, IROC-Houston is developing new tools to diagnose specific issues. Through an independent recalculation, IROC-Houston could identify errors in the institution's TPS model. Although other causes or multiple causes exist, this would be the first step toward evidence-based error diagnosis.

5.2. Materials and Methods

We studied the head & neck (H&N) anthropomorphic phantom because it is the most frequently irradiated phantom. Irradiations from 2012-2016 were studied so that results were up-to-date. This phantom is made of a hollow plastic shell (filled with water during irradiation) with a solid insert containing 6 TLDs in the target (4 within a primary target, 2 within a secondary target). Two EBT films are also positioned axially and sagittally at the center of the primary target volume. A full description of phantom design and construction can be found elsewhere.[23] Upon receiving a phantom, the institution treats it as a patient, CT scanning it and then designing and delivering a therapy plan. After delivery, the phantom and associated DICOM data, containing CT scan images, treatment plan, and TPS calculated dose, is sent back to IROC-Houston. The dose delivered to the TLDs is then read out. The film is read and normalized to the dose of the adjacent TLDs. The measured TLD dose is compared to the TPS calculated dose over the volume of each TLD. Film dose is compared to the TPS calculation for the same region and plane of the film and a gamma analysis is run. If all TLD doses are within ±7% of the calculations and the percent of pixels passing gamma analysis for each of the two films at 7%/4mm is above 85, the phantom is said to pass credentialing requirements.

In this work, we independently recalculated the institution's treatment plan to the H&N phantom. To independently recalculate the institution's dose distribution, a treatment

verification system (TVS) was utilized. Mobius3D (v1.5.3, Mobius Medical Systems, Houston, TX) was chosen as the TVS due to minimal additional workflow requirements, IROC-Houston's high volume of phantom irradiations, the 3D dose recalculation, and the customizable beam models.[26, 27] Mobius3D has several default beam models, but models are also customizable to acquired dosimetry data. IROC-Houston has acquired data from hundreds of linacs and grouped them into representative classes.[28] We created three 6 MV recalculation beam models intended to match the 3 most common classes of linear accelerator: the Varian Base class (including iX, EX, and trilogy machines), Varian TrueBeam class (flattened beams), and the Elekta Agility class. Excluding Cyberknife and Tomotherapy, these three classes represent over 90% of the linacs that have irradiated a phantom. Each beam model started with the default model but was iteratively tuned to match its respective IROC-Houston reference beam dataset. A fitness metric was used that calculated the absolute sum of local differences between the reference beam data point dose values and the model's calculation of the point dose values under the same geometric conditions. Each model was iteratively tuned until the model could no longer achieve a lower fitness metric value.

After the TVS beam models were customized to best match the reference data values, H&N phantom DICOM datasets were given to the TVS for recalculation. All available irradiations from 2012-2016 were given recalculation attempts. In order to be recalculated, irradiations had to include the full DICOM dataset from the institution and be delivered by a linac within the 3 classes. This resulted in 259 irradiation datasets being recalculated. Because the vast majority of phantom failures result from TLD disagreement, we focused on TLD data results.[1, 29] We examined the entire cohort of irradiations as well as the subset of irradiations that failed the IROC-Houston criteria. The entire cohort was also divided by TPS for the two most common planning systems, Eclipse and Pinnacle, to determine relative TPS performance. It was also divided by linac class and by linac-TPS combination. And finally,

the cohort was divided by delivery type: segmental IMRT, dynamic IMRT, and VMAT. All these subset analyses tested if the mean was statistically significant (α=0.05).

The difference in accuracy between IROC-Houston's recalculation and the institution calculation was defined as follows:

$$\mathrm{D}_n = \left(\left|1 - \frac{TPS_n}{TLD_n}\right| - \left|1 - \frac{TVS_n}{TLD_n}\right|\right) * 100$$

Where D_n represents the difference in accuracy between IROC-Houston's recalculation (TVS_n) and the institution's original calculation (TPS_n) for the measured dose (TLD_n) of a given TLD (n). Positive D_n values indicate that IROC-H's recalculation was closer to the measured dose (i.e. more accurate) than was the institution, whereas negative values indicate that the institution's calculation was more accurate. Of particular interest were irradiations where the TVS recalculation matched the TLDs "considerably better" than the institution TPS, which would indicate TPS modelling errors. The threshold for the TVS being "considerably better" used a clinical and a statistical criterion. The clinical criterion specified that the TVS must have an average positive value of D >2% over the 6 PTV TLD locations or a single TLD D value >3%. In other words, the recalculation must be more accurate by an average of 2% or 3% at a single location. The statistical criterion required that the mean value of the distribution of the 6 D_n values must be statistically different from zero (α=0.05). This was done using a 2-sided t-test with failure detection rate correction applied to the p-values. The statistical criterion removed irradiations where the D distribution may have shown a large positive improvement but the individual TLD results were varying substantially.

5.3. Results

Modeling results from the Varian base class TVS beam models are shown in Table 5-1. The values describe the local difference between the acquired reference data point from IROC-Houston's standard dataset and the model recalculation of the same point. The default recalculation model had a fitness value of 11.8. After tuning the model using the built-in tools, the fitness value was lowered to 5.1. All three beam models were tuned from the default and had similar final fitness values. The tuned TVS beam models had mean dose differences from all evaluation points from the reference data of 0.27%, 0.27%, and 0.36% for the Base, TrueBeam, and Agility models, respectively. These differences are the same or smaller than the average institution measurement-to-TPS difference of 0.36% (see Chapter 4).

PDD			Jaw Output Factor		
	Default	Tuned		Default	Tuned
5cm	-0.1%	-0.1%	6x6cm	0.9%	0.2%
10cm	-0.2%	-0.2%	15x15cm	-0.3%	0.0%
15cm	0.6%	0.2%	20x20cm	-0.2%	0.0%
20cm	0.3%	-0.8%	30x30cm	0.3%	-0.1%

IMRT-style Output Factors			SBRT-style Output Factors		
	Default	Tuned		Default	Tuned
6x6cm	0.4%	0.1%	6x6cm	1.0%	0.0%
4x4cm	-0.3%	-0.8%	4x4cm	1.3%	-0.1%
3x3cm	-0.2%	-0.8%	3x3cm	1.7%	0.0%
2x2cm	-0.7%	-1.2%	2x2cm	2.1%	-0.4%

Off-Axis Factors		
	Default	Tuned
5cm	-0.6%	-0.1%
10cm	-0.2%	0.0%
15cm	-0.4%	0.0%

Table 5-1. TVS model discrepancies between the reference data and calculation for the default beam model and the final customized model for the Varian base class.

Two case studies are given here to detail irradiation and recalculation results and to better understand the following results. The agreement of the original TPS calculation and IROC-Houston recalculation of the case studies are given in Table 5-2. The first institution failed the credentialing requirements with two TLD dose/calculation discrepancies beyond criteria and had a 6% average dose discrepancy. Upon recalculation by the TVS, the maximum dose/calculation discrepancy was 4% and an average of 3%. Notably, the institution would have easily passed the irradiation requirements if this single issue were addressed. The second irradiation had very good accuracy from the original institution's calculation, with the measured TLD dose and TPS calculation having no more than 1% discrepancy. The TVS calculation had poorer accuracy, however, with the average discrepancy rising to 7%.

	Inst #1		Inst #2	
TLD #	TPS/TLD	TVS/TLD	TPS/TLD	TVS/TLD
1	0.96	0.98	1.00	1.04
2	0.94	0.97	0.99	1.04
3	0.92	0.97	0.99	1.06
4	0.92	0.98	1.00	1.07
5	0.96	0.99	1.00	1.10
6	0.94	0.96	1.01	1.11
Avg Ratio	0.94	0.97	1.00	1.07
Avg *D*	+3.87%		-5.9%	

Table 5-2. Two irradiations comparing the institution's original dose agreement (TPS/TLD) and IROC-Houston recalculation agreement (TVS/TLD).

The 259 irradiation datasets were recalculated utilizing all 3 beam models to determine difference values, D_n, averaged over all 6 TLD locations in each phantom. Results of the average D_n value are shown via a waterfall plot in the top and middle panel of Figure 5-1. The top and middle panel show the same data but with different color overlays. Recalculations with large negative difference values indicate that the recalculation had poorer accuracy than did the institution's original calculation. Middle values show that the recalculation had comparable accuracy with the original calculation. Irradiations with high positive difference values are those where the recalculation system obtained much better accuracy than the original calculation. The color of the difference value in the top panel denotes how accurate the original calculation was. Cyan indicates that the original calculation had less than a 2% maximum discrepancy with the TLD measured dose across the 6 TLDs. Light green indicates between 2 and 5% discrepancy; orange indicates >5% and red indicates that the irradiation failed IROC-Houston criteria.

The median TVS recalculation difference value was +0.2%, meaning that on average the recalculation was closer to the measured TLD dose than the institution's calculation; the mean was not statistically significant (p=0.9). Of the 259 recalculations, 45 (17%) had differences above the clinical and statistical thresholds, meaning that the TVS recalculation was considerably better and that the institution has serious calculation differences between its linear accelerator and TPS model. These data are shown in pink in the middle panel, which is the same data as the top panel, but color-coded by whether the irradiation had a considerable calculation error. Irradiations without considerable calculation differences are shown in slate gray.

The recalculations of the irradiations that failed to meet current IROC-Houston criteria (the red-colored values of the top panel) are shown in the bottom panel of Figure 5-1. Nineteen phantoms were in this subset, and of those 13 (68%) had considerable TPS calculation differences, with a median TVS difference of +3.1%; the mean was statistically significant ($p<<0.01$).

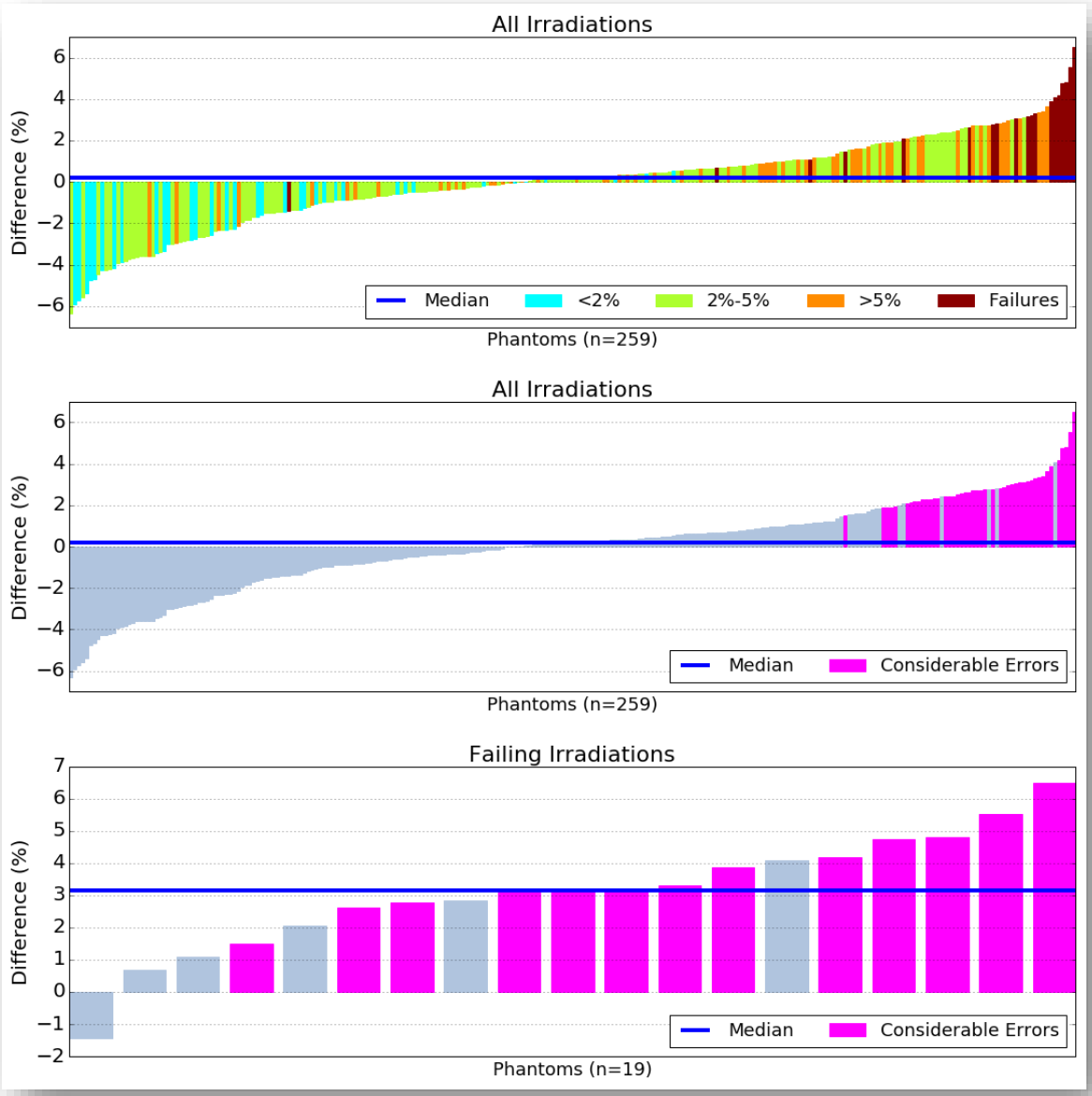

Figure 5-1. Difference values in accuracy between the institution TPS calculation and IROC-Houston's recalculation. Positive values indicate the recalculation was more accurate. The top and middle panel show the same data with different color overlays. The top overlay indicates the institution's original agreement with the TLDs. Pink values in the middle and bottom panel indicate a considerable TPS calculation error on the part of the institution.

The recalculation cohort was analyzed according to several parameters: the delivery technique, the linac class that delivered the irradiation, institution TPS, and linac-TPS combination. The results are shown in Table 5-3. For the delivery techniques, VMAT was the most common delivery technique and had a negative mean D value of -0.39% (p=0.01) indicating that, in general, institutions had more accurate results than the recalculation when utilizing VMAT. In contrast, both the segmental and dynamic IMRT delivery types had D values of 1.18% (p<0.01) and 0.32% (p=0.28), meaning IROC-H's TVS achieved more accurate results on average. Regarding the linac class, the TrueBeam performed the best with a median D value of -0.51% (p=0.04), while the Varian base and Elekta Agility class had values of 0.32% (p=0.06) and 0.68% (p=0.20) respectively. While the Eclipse TPS had a negative but non-significant difference of -0.22% (p=0.16), the Pinnacle TPS had a positive and significant difference of 0.90% (p<0.01), meaning the average recalculation was more accurate than Pinnacle.

	N	Mean D (%)	%CE
Technique			
Segmental IMRT	32	+1.18*	30
Dynamic IMRT	72	+0.32	26
VMAT	153	-0.39*	11
Linac Class			
Varian Base	140	+0.32	21
Varian TrueBeam	74	-0.51*	12
Elekta Agility	23	+0.73	26
TPS			
Eclipse	181	-0.22	16
Pinnacle	53	+0.90*	22
Linac-TPS			
Varian & Eclipse	181	-0.22	16
Varian & Pinnacle	36	+0.87*	19
Elekta & Pinnacle	17	+0.94	29

Table 5-3. Recalculation data broken down by delivery technique, linac class, TPS, and linac-TPS combination. *N* is the number of recalculations. *D* is the difference value of the recalculation. %CE is the percent of irradiations with a considerable TPS error. An asterisk indicates statistical significance.

5.4. Discussion

The positive overall median value of D for the phantom recalculation are surprising given that the treatment verification system (TVS) beam models were meant to represent an "average" machine. If every institution modelled their TPS to perfectly match their linac, the recalculation error values would always be negative; i.e. IROC-Houston's recalculation would always less accurate. However, in roughly half of all cases, the recalculation was closer to the TLD measured dose than the TPS. Of the irradiations that failed IROC-Houston criteria, nearly all recalculations had improved accuracy.

Figure 5-1 demonstrates several important findings. First, irradiations that had good agreement between the institution's calculation and measured dose (colored in cyan) generally had negative values. These irradiations could only improve by 2% or less and

show that these institutions have accurately modeled their linac characteristics. Additionally, since the institutions have customized their TPS, generic models like the ones developed in this study will almost certainly have a lower accuracy. This is demonstrated in the first case study results. Second, recalculations of irradiations that failed IROC-Houston criteria (colored in red) almost always improved accuracy. This was clearly shown in the second case study. The single failing irradiation with a negative difference value was determined to be a setup error, and thus improvement would not be expected. Finally, the question of using “stock” data or models is raised.

The recalculation models used in this study are based on representative beam data from the community. The models perform comparably to the community, with roughly half the recalculations being more accurate and half being less accurate. For the irradiations with negative difference values, using stock data would reduce accuracy, in some cases by a large amount. This alone proves that stock data is not suitable for all scenarios. Yet, for irradiations that failed IROC-Houston criteria, using a model based on generic data was always more accurate. In these cases, stock data would have been a superior choice. These data underscore the need of the physicist to validate the match of their TPS calculations to the linac characteristics as well as comparing their data to community data. The goal is not to match stock data, but to identify where differences are and whether those differences are justified.

Using the statistical and clinical criteria, the recalculations where IROC-Houston’s independent recalculation was considerably closer to the TLD measured dose than the institution’s TPS were the cases of particular interest in this study and to IROC-Houston because these are cases where dramatic improvements can clearly be made in the institution’s dose calculation accuracy. Given that such a TPS is systematically miscalculating dose to every patient, addressing the discrepancies would make a large impact on patient care.

There are several surprising conclusions that can be drawn from the recalculation results. A relatively large percentage of irradiations (17%) were identified as having a considerable TPS calculation error. Furthermore, of the irradiations that failed IROC-H's criteria, two-thirds were shown to have a TPS calculation error and thus is the leading error contributing to failures. Given the boom of advanced therapy delivery techniques and accuracy needed for such treatments, these values are alarmingly high.

Because IROC-H anthropomorphic phantoms are end-to-end tests it can be difficult to determine causes of error in the irradiation workflow. The recalculation tool of IROC-Houston now adds one more layer of problem-solving. Although not every error is due to TPS calculation errors and not every calculation error will be caught by the TVS, those that are identified will have a much clearer picture of where differences between their planned dose and IROC-H's measured dose lie. With this information, the institution physicist can immediately start diagnosing TPS errors rather than spending time trying to identify sources of error that ultimately don't contribute significantly to the problem.

5.5. Conclusion

IROC-Houston has utilized an independent dose recalculation tool, modified after community reference data, to identify institutions that have considerable treatment planning system errors via the anthropomorphic phantom program. 259 head and neck phantom irradiations were recalculated. Of all the irradiations, 17% were found to have considerable TPS errors. Of the irradiations that failed current IROC-H criteria, 68% had this error, making it the leading cause of irradiations failing IROC-Houston criteria. IROC-Houston now has the ability to flag when an institution has a TPS error and can pass that information along to the institution.

Chapter 6: Conclusion

This study has examined the use of an independent recalculation system to identify treatment planning system (TPS) modelling errors. The recalculation system was commissioned from hundreds of measurements of linear accelerators throughout the country. This reference data was systematically collected and thus accurate and representative of the community. From these data, the recalculation system was able to match the reference data better than the average institution. Thus, IROC-Houston's model is able to be trusted. This model was used to recalculate hundreds of head & neck phantom irradiations. Based on the difference in accuracy between the recalculation system and the institution TPS, 17% of all irradiations were found to have a non-trivial modelling error. Further, 68% of irradiations that did not pass IROC-Houston credentialing requirements were found to have this non-trivial TPS error. This shows that despite the advancement in radiation therapy technology and education, an alarming number of institutions have not modeled their TPS accurately. The study was based on the following hypothesis and specific aims. The strategies and results of each aim are explained with a final evaluation of the hypothesis.

Hypothesis: ***By using an independent plan recalculation, IROC-Houston will be able to identify institutional treatment planning system calculation problems in 20% of head & neck phantom irradiation cases that fail credentialing.***

Specific Aim #1: *Acquire and develop reference data that accurately represent common linear accelerators.* This aim was the foundation for the independent recalculation system. Inaccurate data collection leads to inaccurate modeling, thus underscoring the need for reliable reference data. This data was collected consistently using IROC-Houston internal protocols to ensure collection integrity. Data measurements from 2000 through 2015

were queried and analyzed. What made this aim more than simply data collection was studying the underlying distributions of the various linear accelerator models to identify those that could be considered dosimetrically equivalent. Using statistical and clinical criteria, >30 Varian nominal models were condensed to a handful of representative classes as described in chapter 2. Elekta did not have many models to begin with and thus did not experience much reduction in the resulting classes as described in chapter 3. The resulting classes were also compared to other reference data where applicable and largely agreed. The benefit of IROC-Houston's data is that so many linear accelerators were measured that instead of simply a single value for a measurement point, an entire distribution was given. This allows a physicist to understand whether the difference between their own data and the reference data is significant.

Specific Aim #2: *Commission an accurate, independent dose recalculation system.* Once accurate reference data was available, the recalculation system could be compared to them. Mobius3D, the recalculation system used, comes with a default beam model, but the models are also somewhat customizable. A fitness metric was utilized that evaluated overall how close the model calculations were to the reference data. Starting with the default model, the fitness metric was evaluated. Based on the results of the evaluation, new customization values were derived and input into the beam modelling tool. The evaluations continued iteratively until the fitness metric could not be improved. This customization was performed independently for 3 beam models that represented the 3 most common linear accelerator classes. Customization was able to improve the fitness of each model by approximately double from the default. Additionally, the results of the final, tuned beam model were compared to that of the average institution and found to be on par or better as described in chapter 4.

Specific Aim #3: *Recalculate dose to head and neck phantom irradiations and compare to institutional calculated dose*. Given that the recalculation system was verified to be accurate, it was now able to be used to recalculate phantom irradiations retrospectively. Head and neck irradiation data from 2012-2016 were queried and sent to the recalculation system. Over 250 irradiations were able to be recalculated and compared to the institutional calculation. Using a difference equation, the accuracy of each recalculation was compared to the accuracy of the institution's calculation. If the IROC-Houston recalculation was considerably more accurate than the institution's, the institution was flagged as having a TPS modelling error, as described in chapter 5. This equation and metric were used to evaluate the hypothesis.

Using the results from specific aim #3, the percent of irradiations that had considerable error were computed. Of the 19 irradiations that failed IROC-Houston credentialing requirements, 13 (68%) were found to have a considerable TPS error. Given that the definition of the error includes a statistical component, there is no error associated with this result; each irradiation is evaluated independently. Thus, we can say with full confidence that the project's hypothesis is true. Beyond this, there were several other findings unrelated to the hypothesis that are explained in the Appendix.

There are several areas where this study can be continued on in future work. First, only the 3 most common linear accelerator classes were modeled. Although this contained roughly 90% of the available irradiation data, there are still irradiations that could be recalculated if a model were commissioned. Potential models include older Varian machines (2100, 2300), older Elekta machines (MLCi, BMod), and a Varian flattening-filter free model.

Second, more phantom types could be evaluated. This study focused solely on the head & neck phantom, but IROC-Houston offers several varieties of phantoms. The

incorporation of these other phantom types into the recalculation system would allow IROC-Houston to identify more institutions that may not have irradiated a head & neck phantom.

Chapter 7: Appendix

Modified IROC-Houston workflow

IROC-Houston's current workflow and new potential workflow based on this project would look similar to Figure 7-1. The top row explains IROC-Houston's current anthropomorphic phantom workflow. In the first step, the institution receives and irradiates the phantom with TLD dosimeters measuring the delivered dose. In the next step, the institution returns the irradiated phantom and the associated DICOM data including the CT image data, RT plan, RT dose, and RT structures files. These files contain the expected delivery information and metadata (e.g. gantry angles and MLC positions). Next, IROC-Houston removes and measures the TLD dosimeters. The institution's DICOM plan contains the contours of the TLDs; this dose calculation is the expected dose. Next, these two values, the TLD measured dose and TPS expected dose, are compared to one another for all 6 TLDs within the target volumes. Finally, based on the agreement of these values, the institution passes or fails the credentialing requirements.

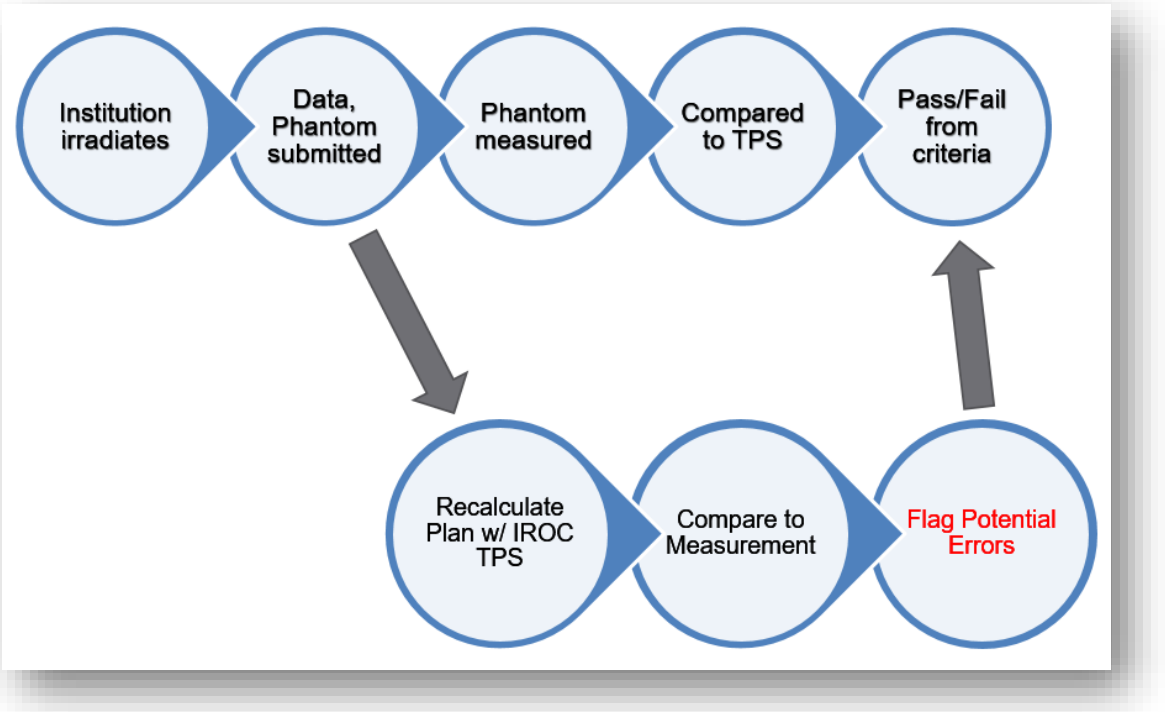

Figure 7-1. The current and proposed workflow for IROC-Houston phantom irradiations.

The new proposed workflow includes all of the current workflow steps and adds 3 additional steps. At the time that the institution submits their phantom and DICOM data, this data is passed to the recalculation system. The recalculation system will calculate its own values for the expected dose to the TLDs. These values can then be compared to both the measurement and institution TPS calculation. Finally, in the case that the recalculation system is much more accurate than the institution's calculation, the institution is flagged as having a TPS calculation error and the institution is passed the relevant information.

Sending an IROC-Houston phantom irradiation dataset to the treatment verification system

To send an institution dataset to the treatment verification system, specifically to Mobius3D (M3D), a handful of steps must be performed

1. **Ensure that the irradiation is valid and contains all relevant DICOM data**. Currently, Tomotherapy and Cyberknife irradiations cannot be given to M3D, thus eliminating those from possible recalculation. Further, very old Varian and old to middle-aged Elekta machines are not yet modeled in M3D. Assuming the irradiation does not fall into one of the above categories, it can be recalculated. Ensure that all DICOM data is present in the IROC-Houston shared drive or wherever this DICOM data is archived. This includes the CT image data, RTPlan file, RTDose file, and RTStructure file. Make a copy of the data so as not to alter the original dataset.
2. **Convert the DICOM data to be associated with the appropriate linac class.** Each institution has their own name for their linac. This data is stored in a DICOM tag and read by M3D. If M3D does not have a listed linac that exactly matches the tag, it will not proceed with recalculation. To remedy this, the DICOM tag must be changed to one of the linacs listed in M3D. First, identify which class the linac derives from. For example, an institution may have a machine name of “TB_1”. This is almost certainly a TrueBeam linac, which derives from the Varian TrueBeam class of reference machines. Using the program created for this project or any DICOM tag-editing software, change the name to the linac class. Continuing from above, the new DICOM machine name should be “TrueBeam”.
3. **Compress and upload to Mobius3D**. Compress the newly-edited set of DICOM files to a ZIP archive. Then, log into M3D and go to settings, then DICOM files. There is an “upload files” file browser button. Find the ZIP archive and then click

"Upload Files". M3D will give a confirmation message when uploading is complete and validated.

Process of tuning Mobius3D beam models and results of tuned models

Mobius3D comes with a default beam model out of the box, but IROC-Houston desired a set of models that corresponded to the reference datasets from the site visits as much as possible. Mobius3D fortunately has tools that allow for beam customization. This includes in-field dosimetry tuning and model meta-parameters such as dosimetric leaf gap. To determine how well a beam model configuration agrees with the reference data, a fitness metric was used. The fitness metric was defined as follows:

$$FM = \sum_{p=1}^{15} |\, TVS_p - \, RD_p|$$

Where *FM* is the fitness metric, *TVS* is the treatment verification system (Mobius3D), *RD* is the reference data, and *p* is the specific parameter that is being considered (e.g. PDD10x10cm^2(10cm)) and there are 15 total parameters (excludes wedge factors). Thus, the fitness metric is the absolute sum of the agreement between the TVS calculation and the reference data point for all the site visit parameters. The minimization of this fitness metric is the goal. A calculation of the fitness metric was done for the untuned beam models. Based on the individual differences of the parameters, new tuning parameters were estimated and input into Mobius3D's beam modeling tools in an attempt to attain a lower fitness metric. This process was repeated iteratively until the fitness metric could not be lowered further. The individual parameter differences of the default beam model and final, tuned beam model for the 3 models created for this project are shown from Figure 7-2 through Figure 7-4 and Table 7-1 through Table 7-3.

Default Varian Model

cm/cm^2/cm^2/cm^2/cm	PDD	Jaw OF	IMRT OF	SBRT OF	Off-Axis
5/6x6/2x2/2x2/5	-0.1%	0.9%	-0.7%	2.1%	-0.6%
10/15x15/3x3/3x3/10	-0.2%	-0.3%	-0.2%	1.7%	-0.2%
15/20x20/4x4/4x4/15	0.6%	-0.2%	-0.3%	1.3%	-0.4%
20/30x30/6x6/6x6/--	0.3%	0.3%	0.4%	1.0%	

Tuned Varian Base Model

5/6x6/2x2/2x2/5	-0.1%	0.2%	-1.2%	-0.4%	-0.1%
10/15x15/3x3/3x3/10	-0.2%	0.0%	-0.8%	0.0%	0.0%
15/20x20/4x4/4x4/15	0.2%	0.0%	-0.8%	-0.1%	0.0%
20/30x30/6x6/6x6/--	-0.8%	-0.1%	0.1%	0.0%	

Table 7-1. The local difference between the IROC-Houston standard reference dataset for the 6 MV Varian Base class for the default Mobius3D model and the final, tuned model.

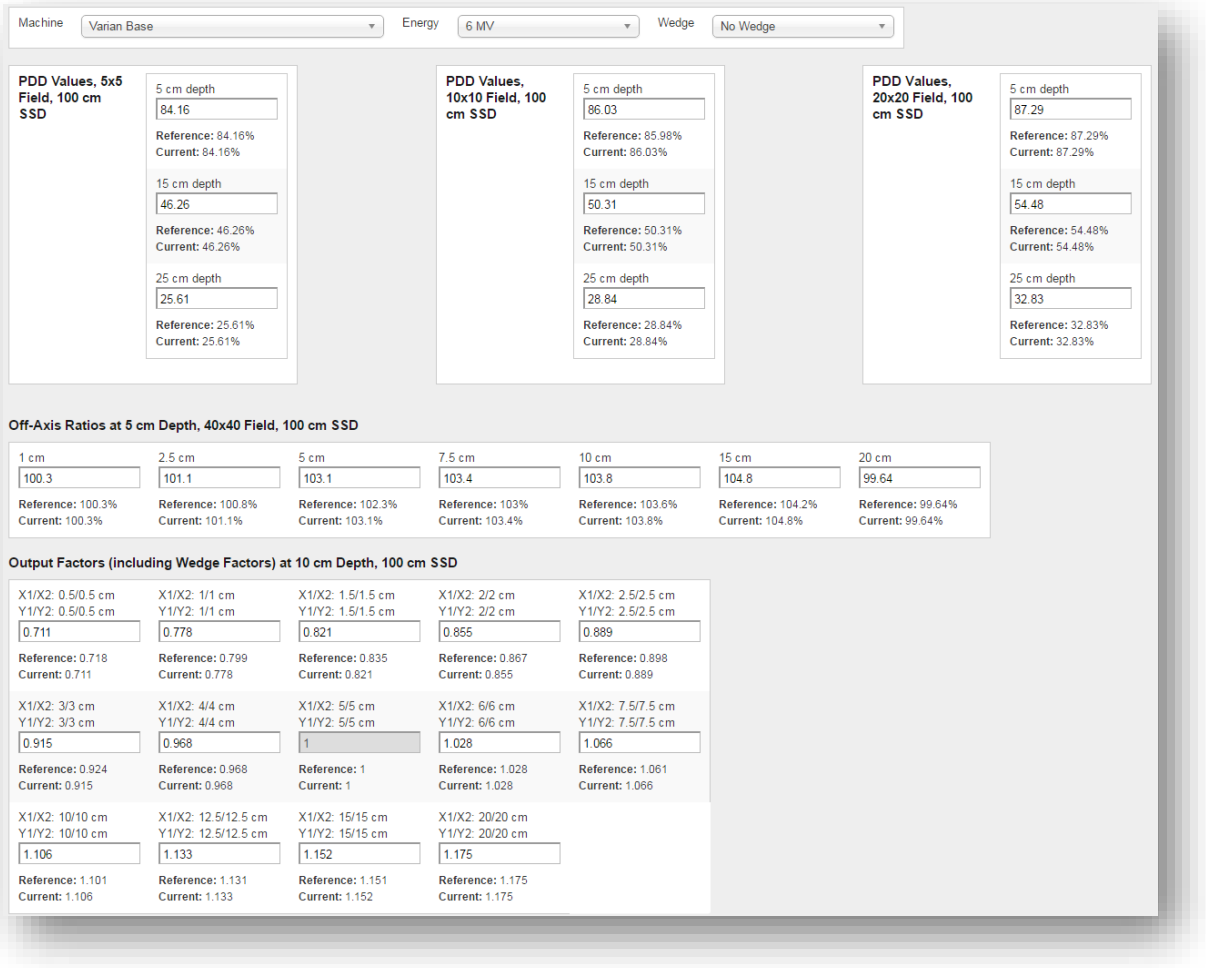

Figure 7-2. Screenshot of final tuning parameters for the Varian Base class in Mobius3D.

Default Varian Model

cm/cm^2/cm^2/cm^2/cm	PDD	Jaw OF	IMRT OF	SBRT OF	Off-Axis
5/6x6/2x2/2x2/5	-0.1%	0.9%	-0.7%	2.1%	-0.6%
10/15x15/3x3/3x3/10	-0.2%	-0.3%	-0.2%	1.7%	-0.2%
15/20x20/4x4/4x4/15	0.6%	-0.2%	-0.3%	1.3%	-0.4%
20/30x30/6x6/6x6/--	0.3%	0.3%	0.4%	1.0%	

	Tuned Varian TrueBeam Model				
5/6x6/2x2/2x2/5	-0.5%	-0.4%	-0.9%	-0.4%	0.2%
10/15x15/3x3/3x3/10	0.0%	-0.1%	-0.5%	-0.4%	0.0%
15/20x20/4x4/4x4/15	0.2%	0.0%	-0.5%	-0.2%	0.0%
20/30x30/6x6/6x6/--	-0.5%	-0.2%	0.2%	0.0%	

Table 7-2. The local difference between the IROC-Houston standard reference dataset for the 6 MV Varian TrueBeam class for the default Mobius3D model and the final, tuned model.

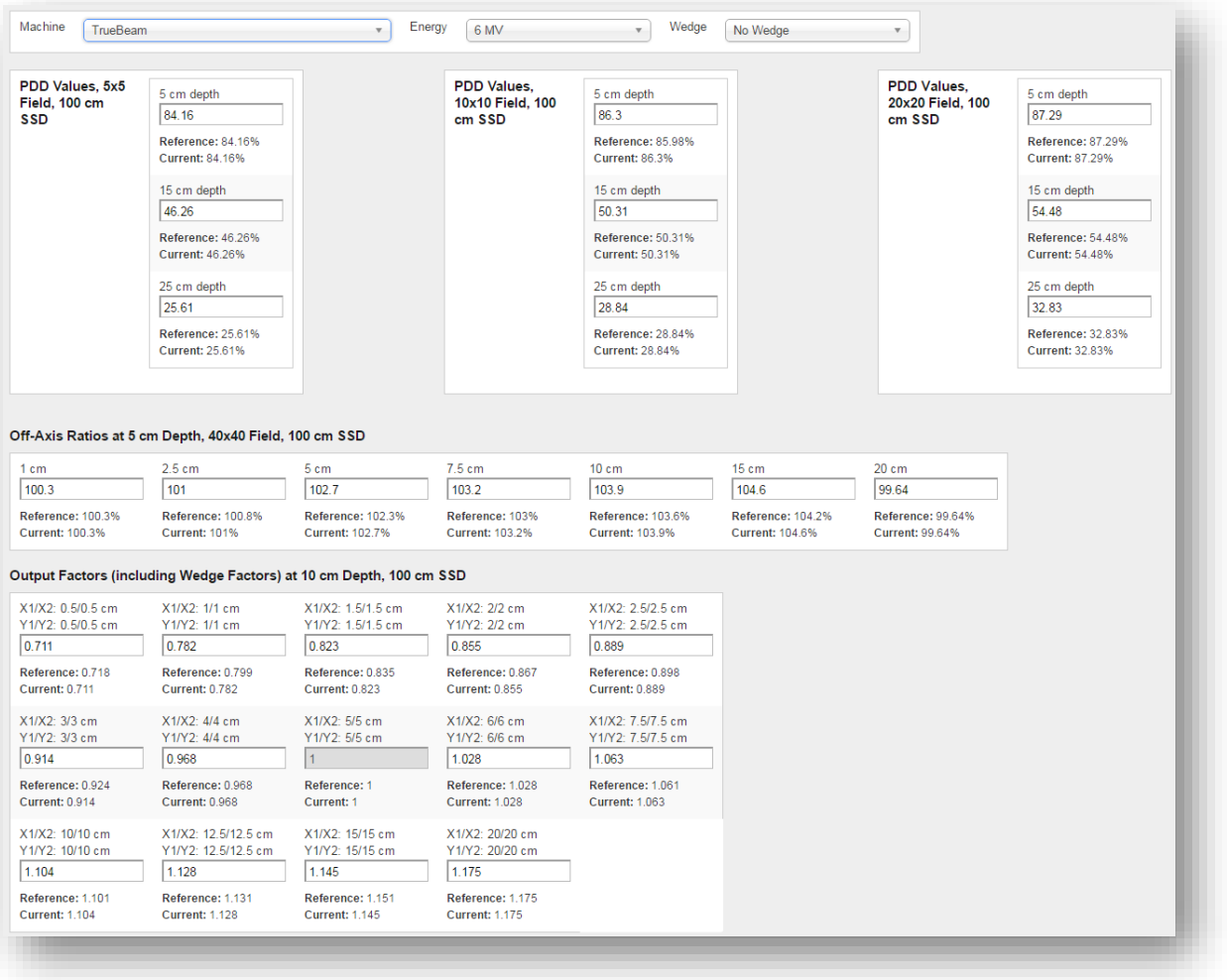

Figure 7-3. Screenshot of final tuning parameters for the Varian TrueBeam class in Mobius3D.

Default Elekta Model

cm/cm^2/cm^2/cm	PDD	Jaw OF	IMRT OF	Off-Axis
5/6x6/2x2/5	1.7%	0.1%	-4.5%	-0.1%
10/15x15/3x3/10	1.6%	-0.1%	-0.1%	-1.1%
15/20x20/4x4/15	2.1%	-0.2%	0.1%	-1.3%
20/30x30/6x6/--	2.0%	-0.5%	0.6%	

Tuned Elekta Agility Model

5/6x6/2x2/5	-0.1%	-0.2%	-2.6%	-0.1%
10/15x15/3x3/10	-0.3%	0.2%	-0.4%	-0.1%
15/20x20/4x4/15	0.0%	0.0%	-0.5%	-0.2%
20/30x30/6x6/--	-0.5%	0.0%	0.2%	

Table 7-3. The local difference between the IROC-Houston standard reference dataset for the 6 MV Elekta Agility class for the default Mobius3D model and the final, tuned model.

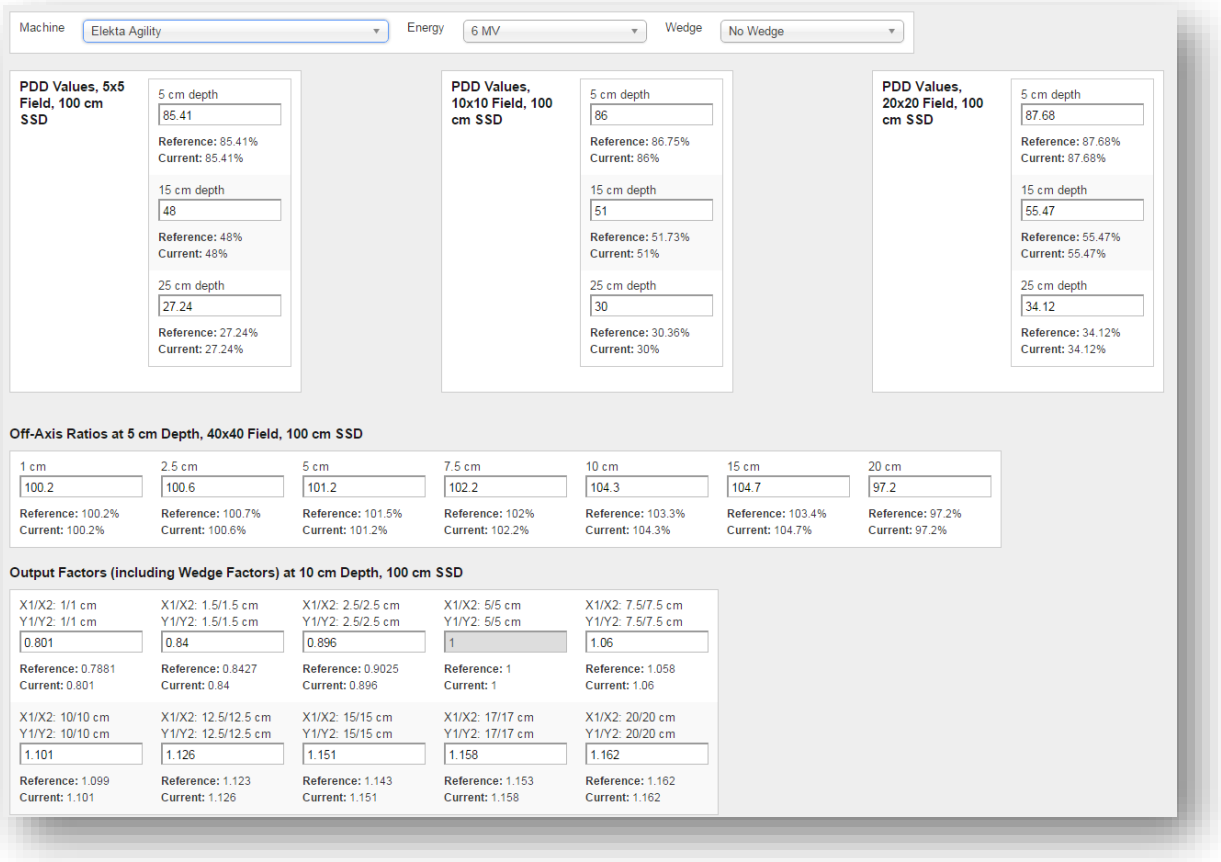

Figure 7-4. Screenshot of final tuning parameters for the Elekta Agility class in Mobius3D.

Monte Carlo dose comparisons to the TVS

During the commissioning of the TVS beam models it was thought that verification of the model's accuracy for points not along the central axis, where the reference data was, could be done using an independent and accurate Monte Carlo (MC) dose calculation. IROC-Houston has developed MC simulations and phase space files that accurately represent several common linear accelerators. By calculating dose for the same fields as used during a site visit and then comparing that to the TVS' calculation of the same fields the off-axis and penumbra doses could be validated. Open fields of 6x6, 10x10, and 20x20cm^2 were created along with IMRT-style fields of 2x2, 3x3, 4x4 and 6x6cm^2 fields. A

total of 10^9 particles were run, being split into 10 batches, and the results were averaged. Profiles and percent depth dose curves were sampled for several field sizes and are shown below, along with a table of field widths and penumbras for the complete dataset.

Visually, the open field profiles are very close to one another, while the IMRT-style fields have distinctive differences in the profile shapes, notably at the penumbra. The PDD curves match well, although a slight increase with depth in the ratio can be seen as the field size decreases. Looking at the tabular data, the field size widths (defined as the full-width half-max) are very similar, being within 1mm of each other. The penumbra widths were very different however, having a difference between 0.6 and 1.7mm, with Mobius3D always having a wider penumbra.

In the customization of the Mobius3D beam models, the only adjustable parameter that affected MLC leaves was the dosimetric leaf gap (DLG). The DLG however, controls field width for MLC-defined fields, not the penumbra width. Thus, no change could be made to the Mobius3D models to adjust the penumbra. It should also be stressed that the Monte Carlo has uncertainty in its calculations, and even conforming to the Monte Carlo values may not have been ideal based on the favorable results of this study.

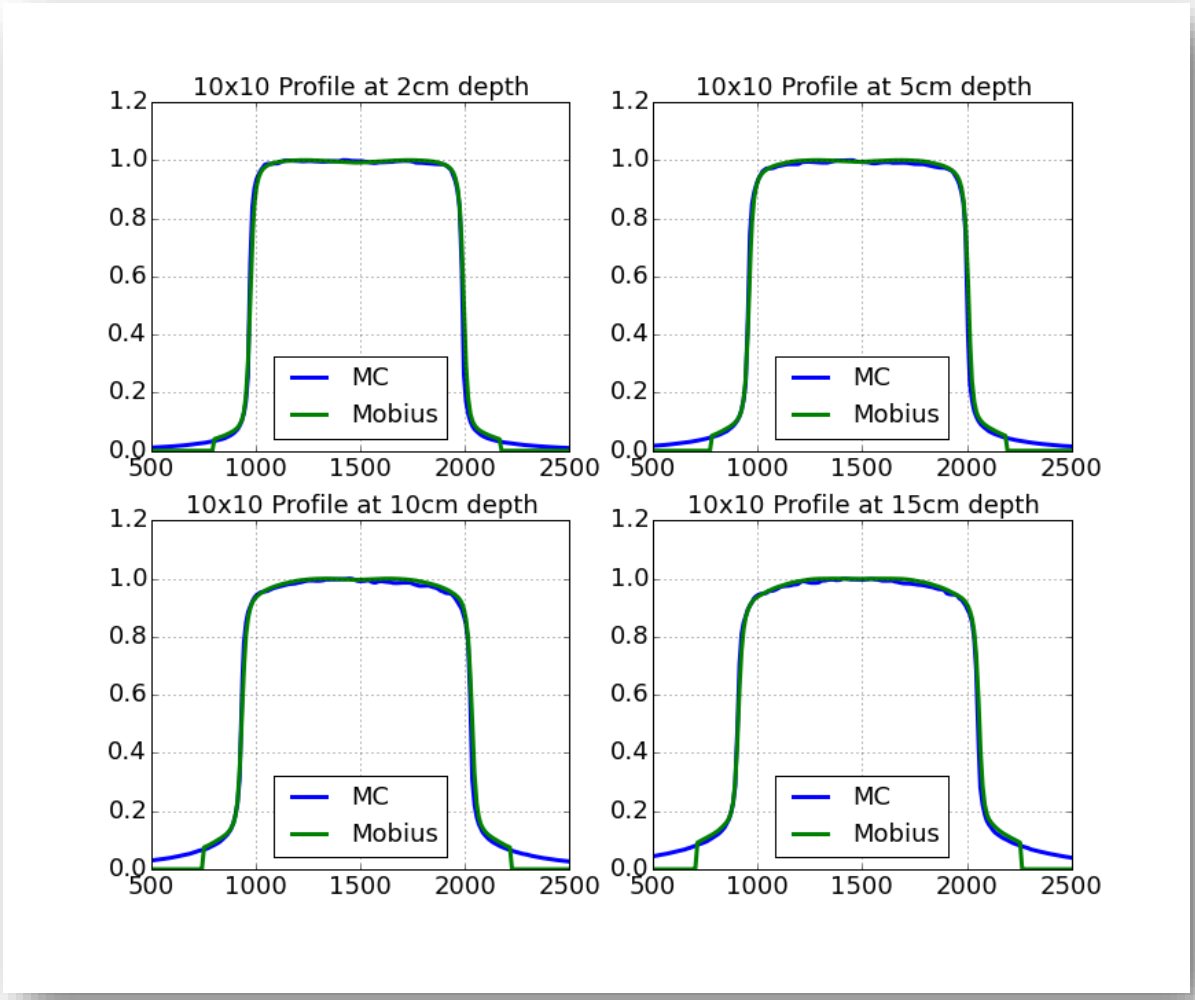

Figure 7-5. Profiles of a 10x10cm^2 open field at various depths for Monte Carlo (MC) and Mobius3D (M3D). A ratio of the profiles is also given.

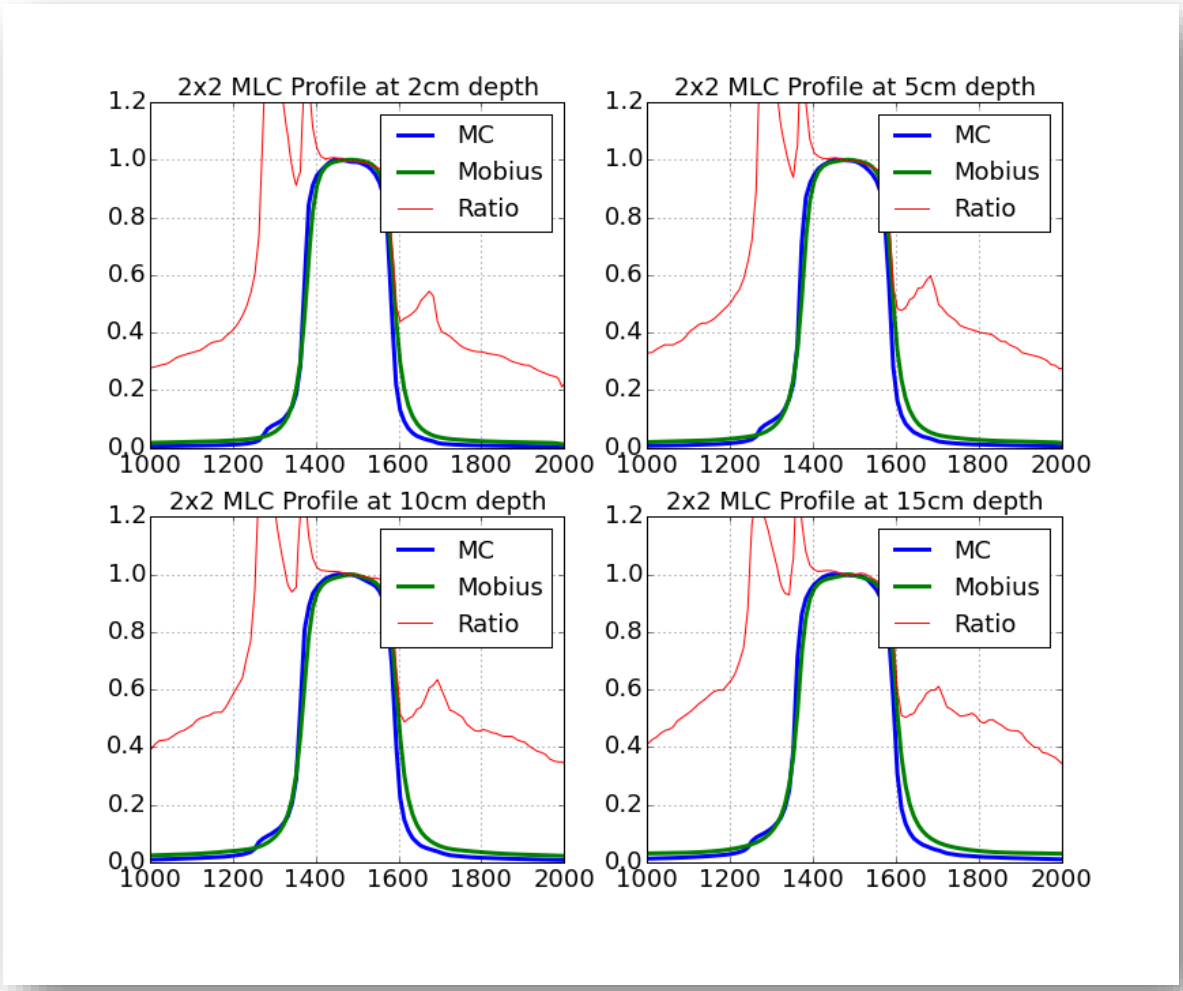

Figure 7-6. Profiles of a 2x2cm^2 MLC-defined field at various depths for Monte Carlo (MC) and Mobius3D (M3D). A ratio of the profiles is also given.

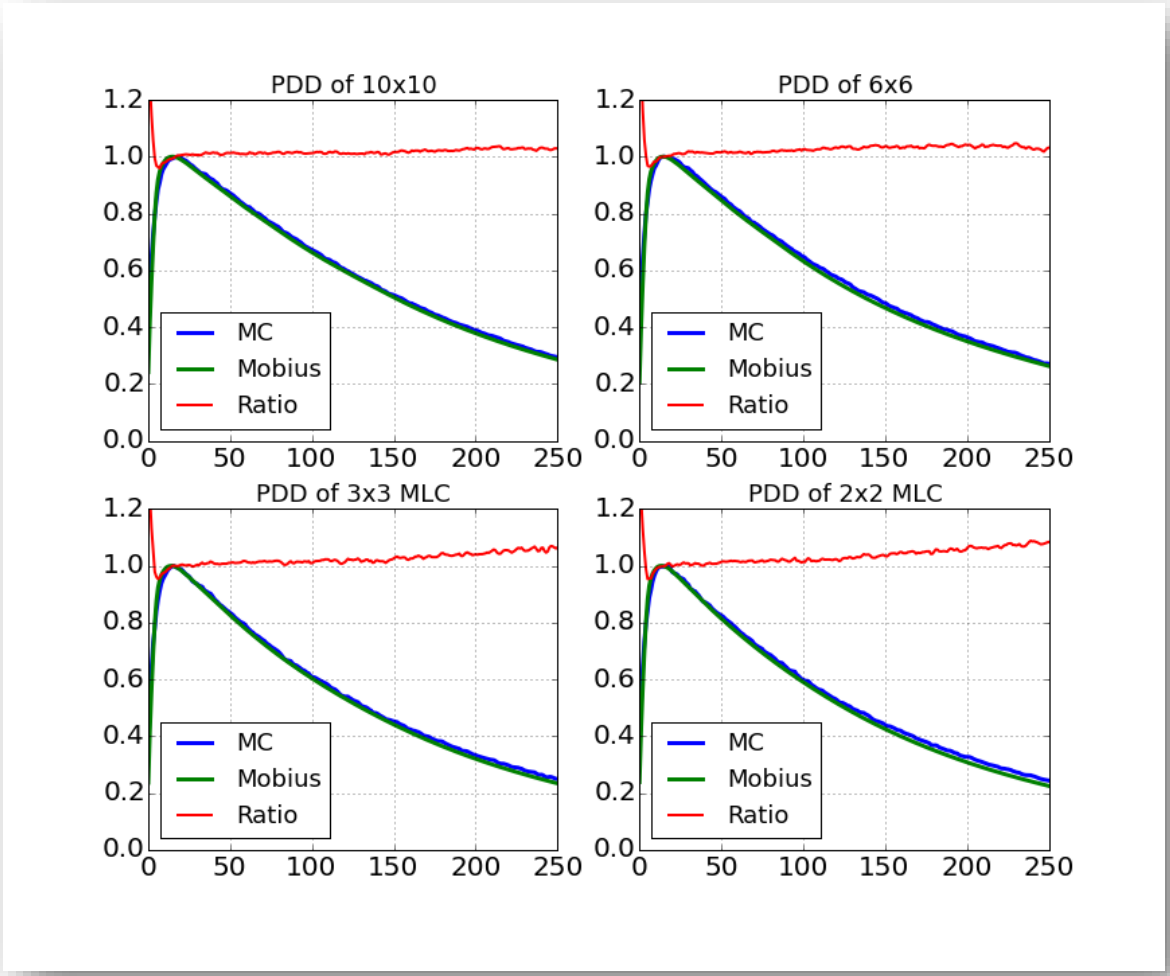

Figure 7-7. Plots of percent depth dose curves of both Monte Carlo and Mobius3D. A ratio of the curves is also given.

Field size	Field Widths @ 10cm (cm)		Penumbra @ 10cm (mm)	
	MC	M3D	MC	M3D
20x20	21.94	22.04	4.15	5.40
10x10	11.01	11.04	3.10	3.85
6x6	6.61	6.65	3.05	3.60
6x6 MLC	6.69	6.75	4.15	5.70
4x4 MLC	4.51	4.56	3.65	5.25
3x3 MLC	3.41	3.45	3.80	4.80
2x2 MLC	2.30	2.35	3.05	4.55

Table 7-4. Field widths and penumbra widths for a range of field sizes at 10cm depth comparing Monte Carlo and Mobius3D.

Comparison of accuracy of TVS models and the average institution

Even though the TVS beam models were tuned to reach a minimum fitness score, their accuracy and agreement should be evaluated against that of the community in order to understand the expectations of the TVS. The IROC-Houston site visit data carries two components: the direct dosimetric measurement and the institutions TPS calculation for that same point and geometry. The agreement of the institution and their own TPS calculations can then be quantified, and statistical metrics can be acquired from the entire site visit cohort. The cohort metrics can then be compared to the TVS beam model metrics. These metrics are given the table below.

	Avg diff. %	% points >1%
Varian Base	0.27	5.3
TrueBeam	0.27	15.8
Agility	0.36	15.8
Avg Inst	0.36	21

Table 7-5. Agreement between measured dosimetric data and calculation for the TVS beam models and average institution TPS. Average difference percent represents the average local difference between measured data and calculation for all measured parameters. Percent of points greater than 1% are the number of individual measurement points where the calculation had a greater local difference than 1%.

Based on the above table it can be seen than the TVS beam models either meet or beat the average local difference between measurement and calculation. Further, all the models have a lower percentage of points where the calculation is >1% different from the measurement. Based on this data the TVS beam models can be said to be comparable if not better in accuracy than the average institution.

Site Visit measurements compared to phantom irradiation agreement

Several institutions that irradiated a H&N phantom also had a site visit performed within the span of a few years, either before or after the irradiation. These overlaps allow for a more detailed comparison of how well the institution agrees with its own TPS and how well they fare in a phantom irradiation. To do the comparison the improvement of the TVS' recalculation over the TPS calculation (the difference D as defined in chapter 4) was plotted against the general disagreement of the site visit data. The disagreement value of the site visit was determined by summing the absolute differences of the measurements and TPS calculations for a subset of parameters relevant to a H&N irradiation: 6x6 and 10x10cm^2 open jaw output factors and 2x2, 3x3, 4x4, and 6x6cm^2 IMRT-style output factors. These values are plotted along with a fitted linear trendline in Figure 7-8.

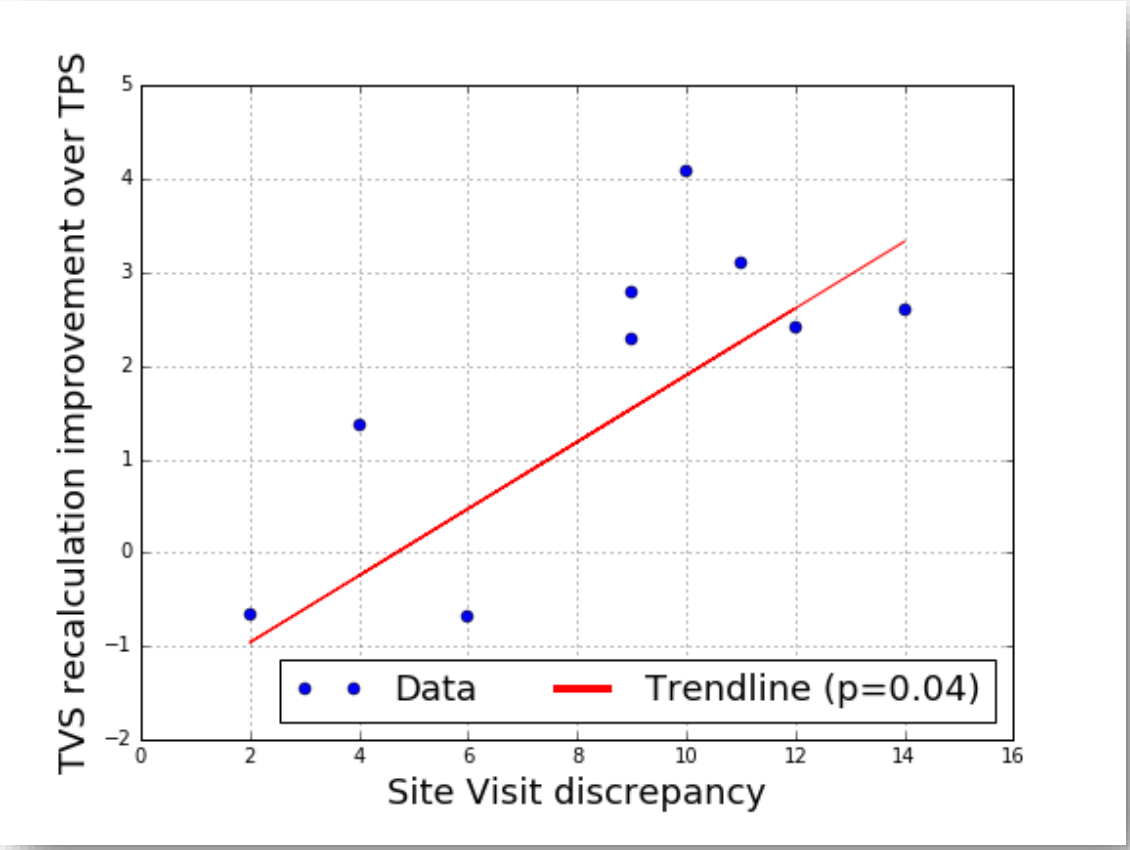

Figure 7-8. Phantom recalculation differences plotted against the overall discrepancy of a site visit done at the institution. The difference values (y-axis) are the difference values calculated in chapter 4 and the site visit discrepancy (y-axis) is the sum of absolute differences between the dosimetric characteristics and TPS calculation for relevant parameters.

A slope regression test was performed on the fitted trendline to determine if it was statistically significant from zero (α=0.05). The test showed that the slope was in fact significant (p=0.04). The positive slope demonstrates that IROC-Houston's recalculation improvement in accuracy rose as the institution had worse agreement between its dosimetric characteristics and TPS calculation. This correlation proves an important finding: the greater the disagreement between the institution measurement values and the institution TPS calculation, the lower the accuracy of the phantom irradiation. Such a finding may be intuitively obvious, but the fact that so many institutions fail a phantom irradiation shows that

despite this straightforward issue many institutions still do not model their TPSs very well. This result proves that increasing TPS accuracy for even basic parameters such as output factors would improve dose accuracy for all patient calculations.

There are limitations to these findings however. First, the calculation of site visit discrepancy was somewhat arbitrary, avoiding any weighting of the discrepancies. Weighting and/or including other parameters may change the results or prove the slope is not significantly positive. Also, there are several TPS model meta-parameters that can have a large effect on small field calculations (dosimetric leaf gap, MLC transmission and leakage, etc). None of these parameters were examined and may prove to have a significant correlation with phantom dose accuracy.

Graphical results of recalculation groups

Chapter 4 described several findings of phantom irradiations by recalculating dose via an independent TVS. Due to space limitations not all findings were shown and some findings were condensed to a table. The findings are graphically reproduced here in full.

The first two graphs show the recalculation difference D for 3 sets of phantom irradiations: the entire cohort, those that had at least one original TLD discrepancy >5%, and those that failed IROC-Houston credentialing (7% TLD, 85% pixels with γ<1 at 7%/4mm film DTA). The first graph shows the three sets colored according to the institution's original agreement. The second graph is the same underlying data, but highlights recalculations where the IROC-Houston TVS demonstrated considerable improvement over the institution TPS.

The second set of graphs show all the phantom irradiation recalculations together, but split according to the class of accelerator that delivered the dose. The third set of graphs

show the recalculation differences according to delivery techniques: segmental IMRT, dynamic IMRT, and VMAT. The fourth set of graphs show the recalculation differences according to linac vendor and TPS configurations.

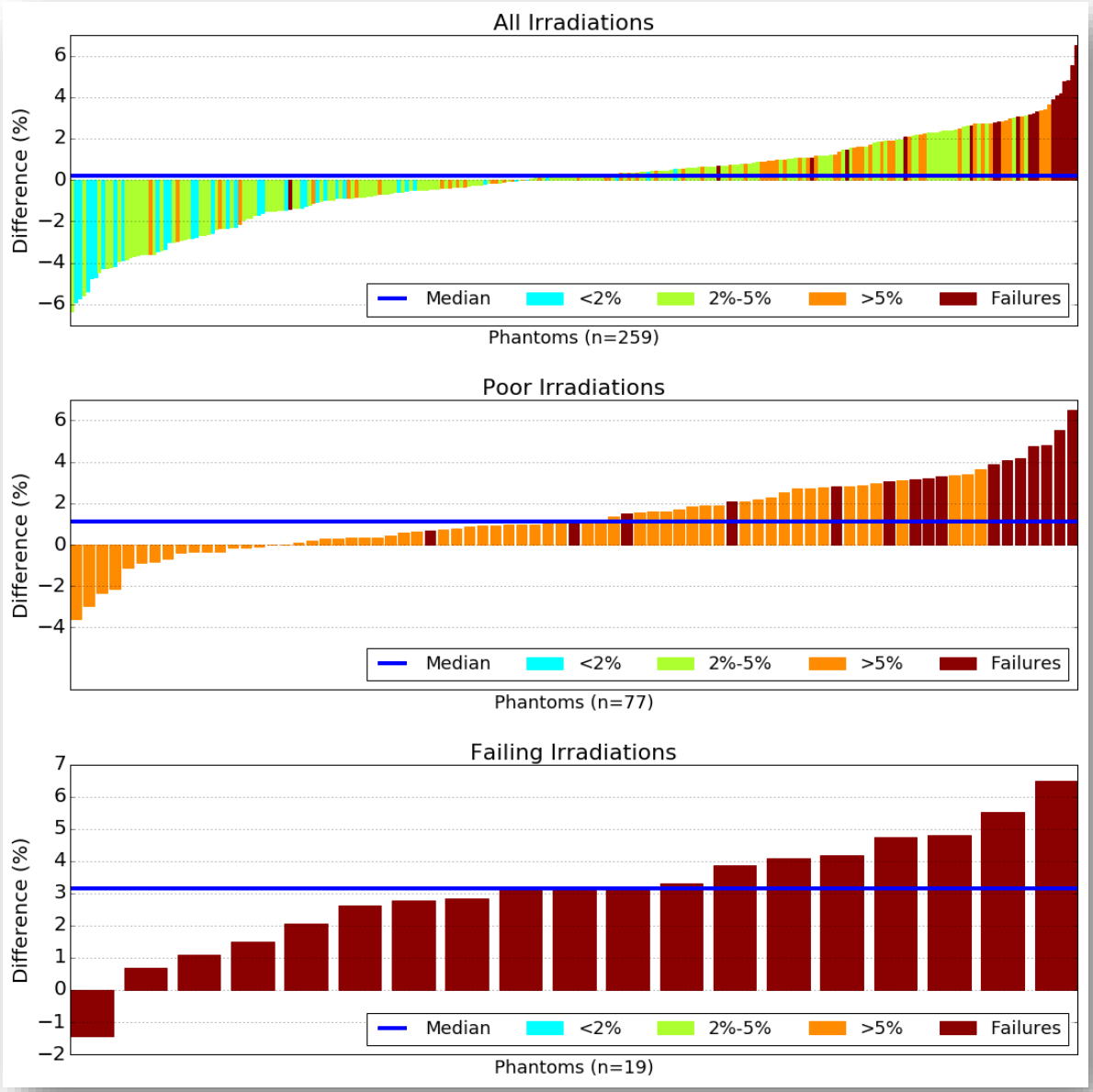

Figure 7-9. Phantom recalculation difference values plotted according to 3 subsets; each graph contains a subset. Colors indicate tiers of original agreement between the TPS and TLD doses.

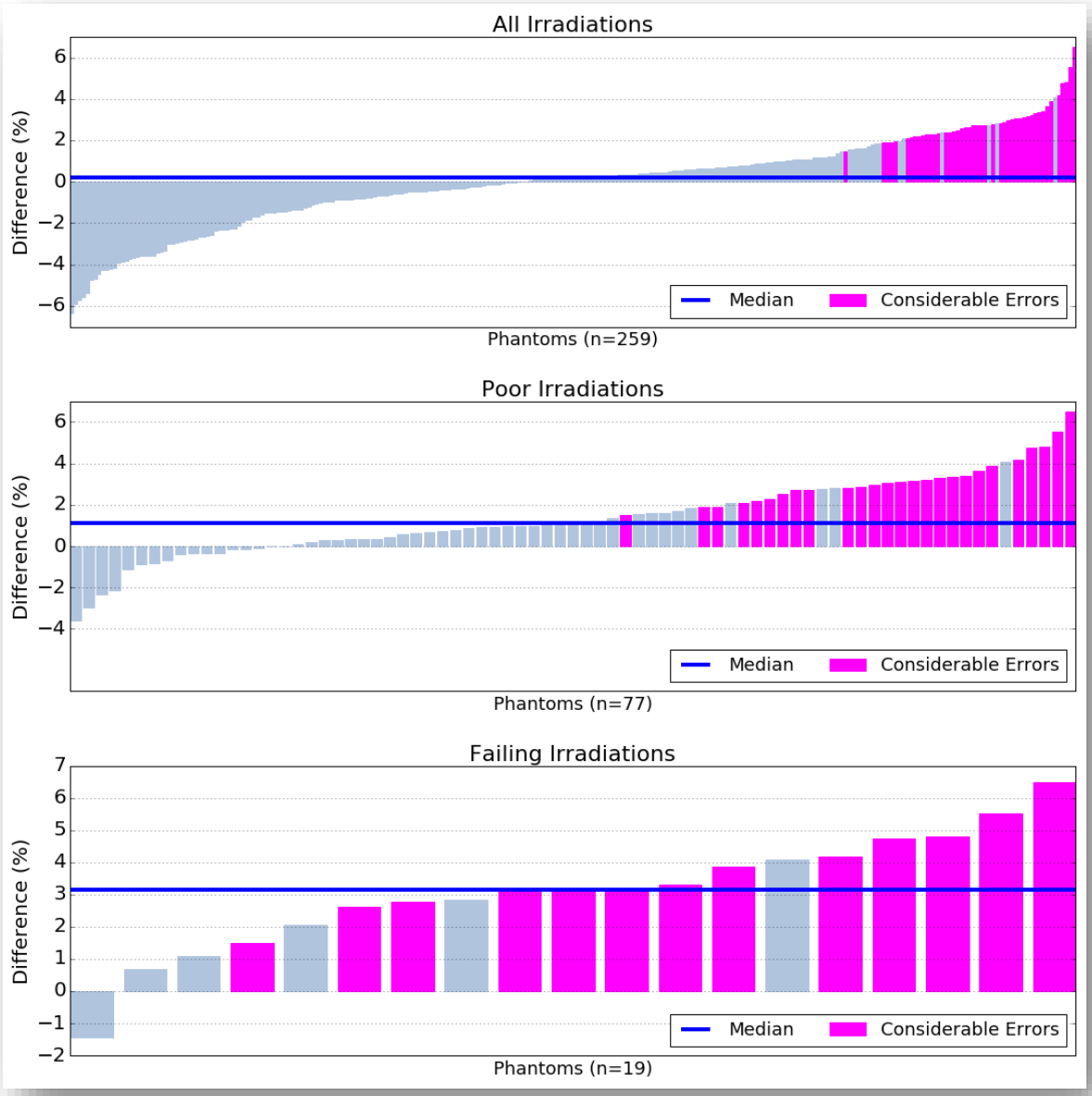

Figure 7-10. Phantom recalculation difference values plotted according to 3 subsets; each graph contains a subset. Colors indicate whether the institution TPS disagreed considerably with the TVS and thus had a considerable error.

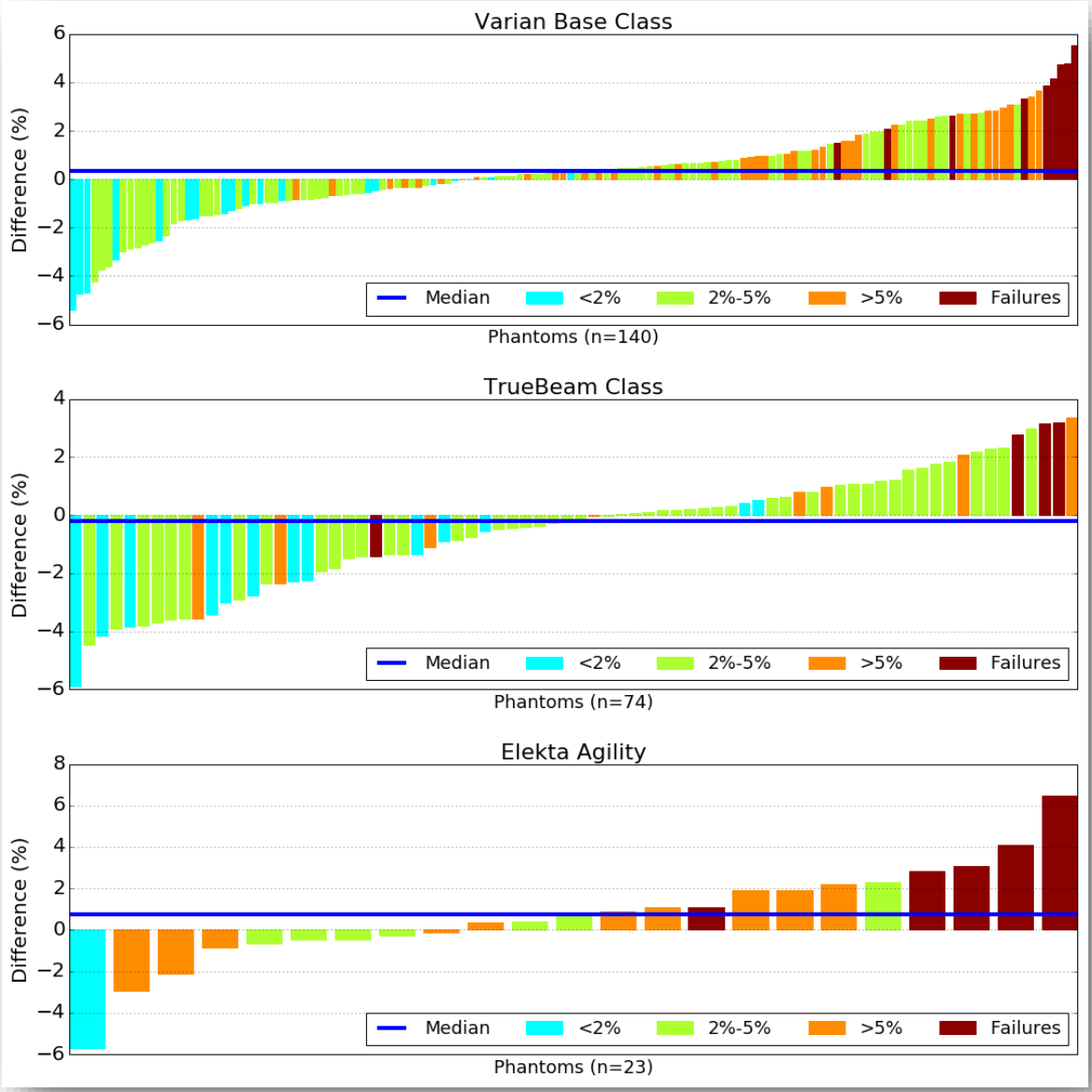

Figure 7-11. Phantom recalculation difference values plotted according to each linac class; each graph shows a class. Colors indicate tiers of original agreement between the TPS and TLD doses.

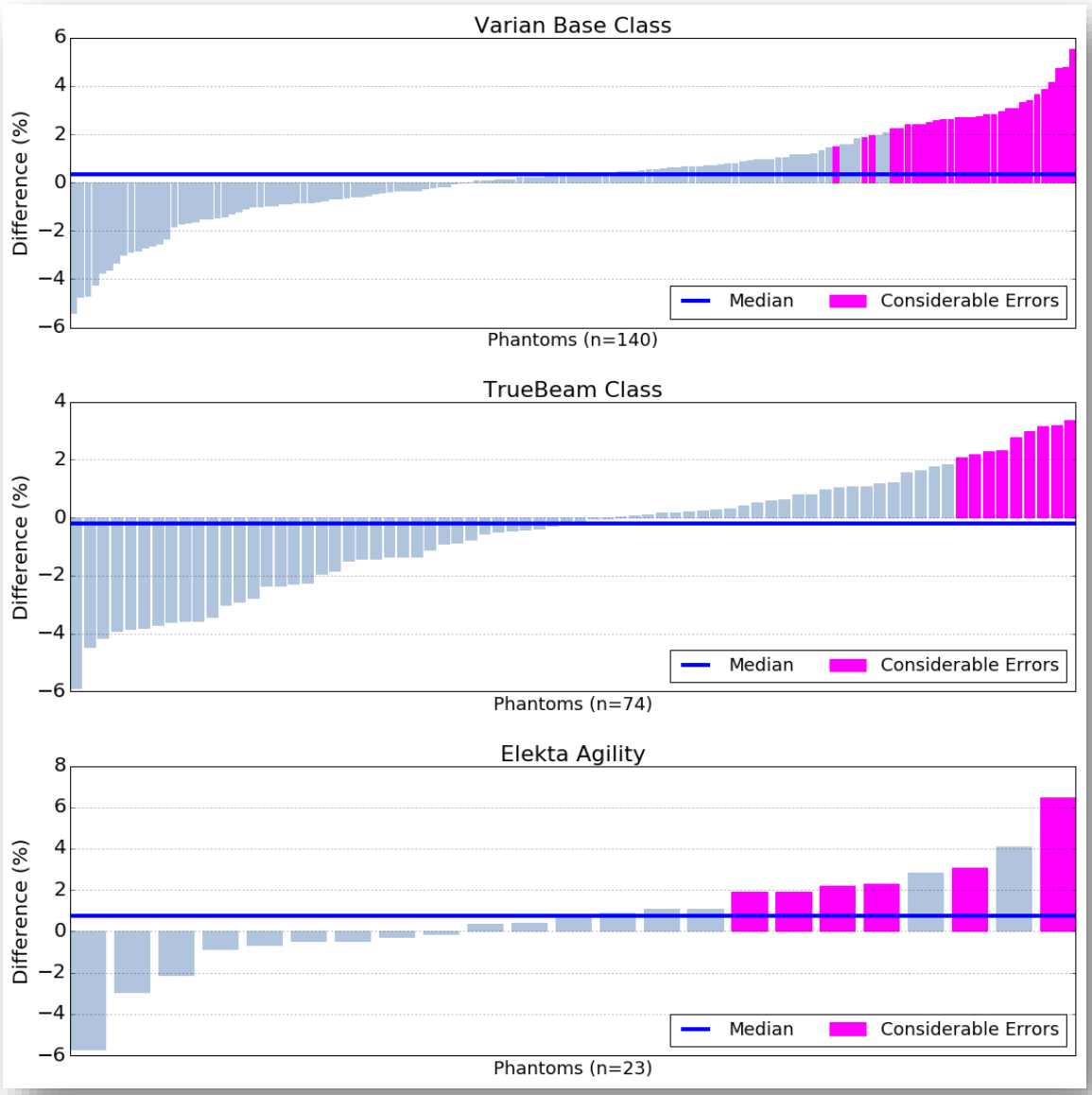

Figure 7-12. Phantom recalculation difference values plotted according to each linac class; each graph shows a class. Colors indicate whether the institution TPS disagreed considerably with the TVS and thus had a considerable error.

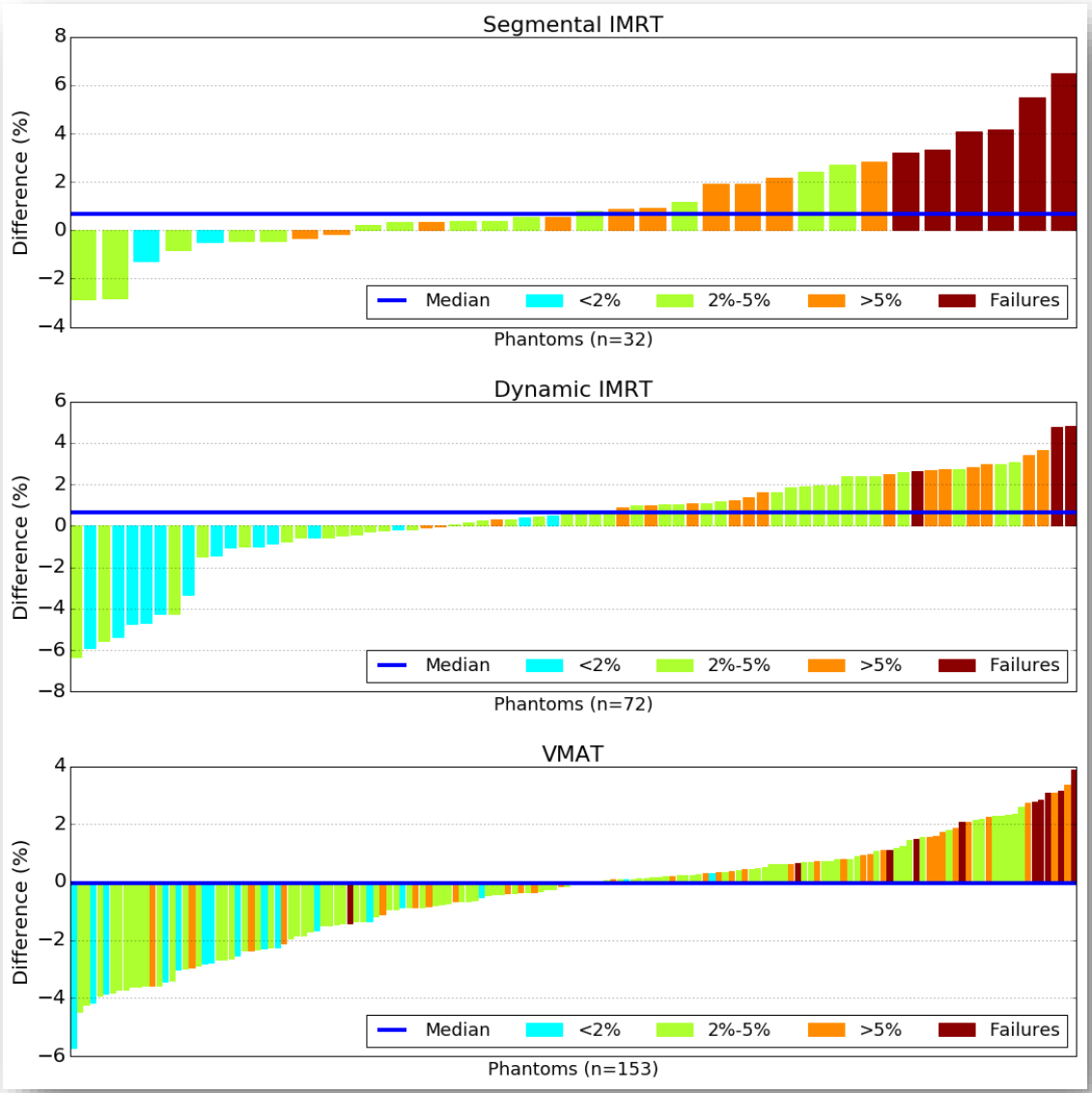

Figure 7-13. Phantom recalculation difference values plotted according to the 3 delivery techniques; each graph shows the results of that delivery technique. Colors indicate tiers of original agreement between the TPS and TLD doses.

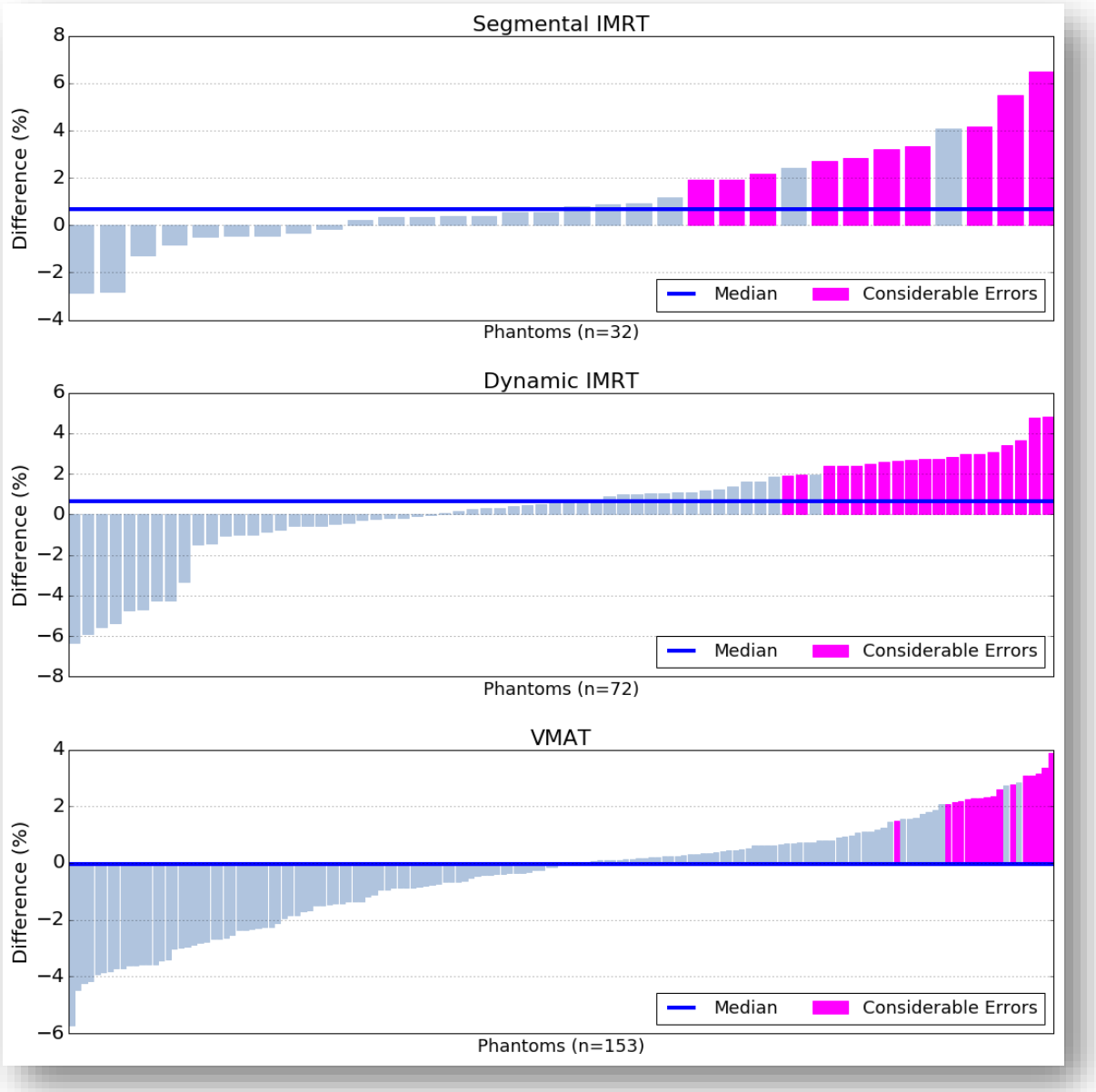

Figure 7-14. Phantom recalculation difference values plotted according to the 3 delivery techniques; each graph shows the results of that delivery technique. Colors indicate whether the institution TPS disagreed considerably with the TVS and thus had a considerable error.

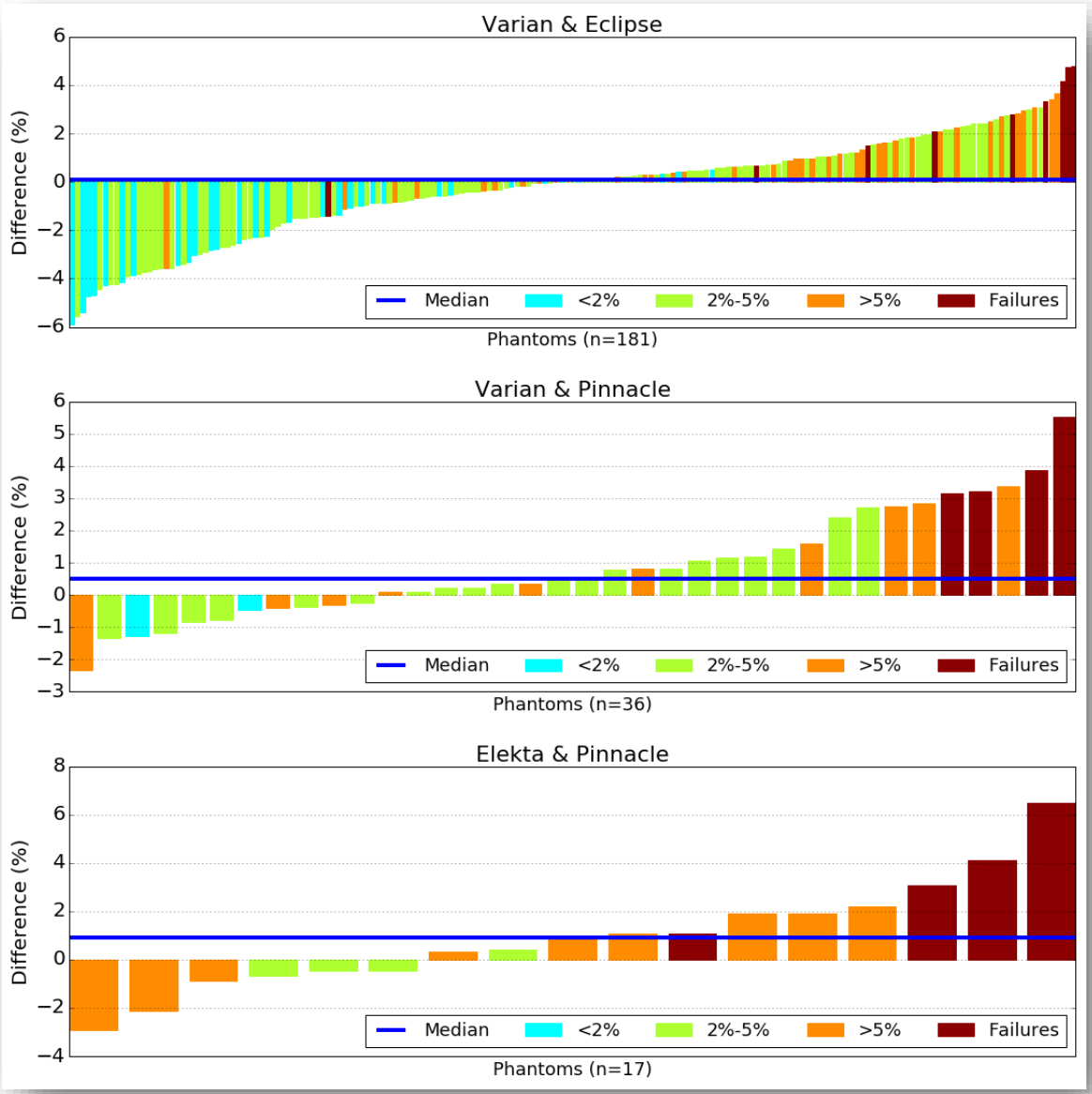

Figure 7-15. Phantom recalculation difference values plotted according to the linac/TPS configurations; each graph shows a linac/TPS configuration. Colors indicate tiers of original agreement between the TPS and TLD doses.

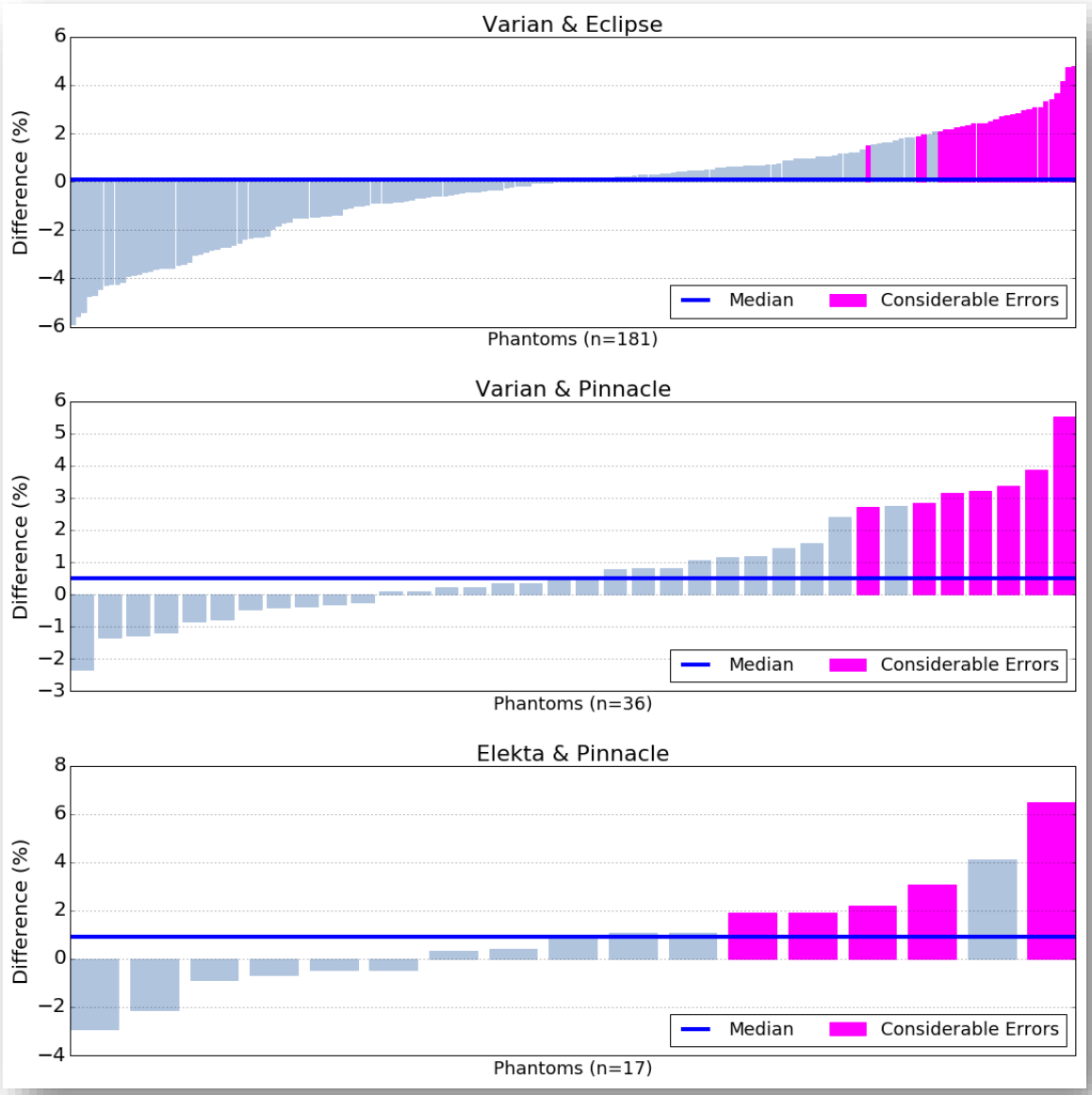

Figure 7-16. Phantom recalculation difference values plotted according to the linac/TPS configurations; each graph shows a linac/TPS configuration. Colors indicate whether the institution TPS disagreed considerably with the TVS and thus had a considerable error.

References

[1]A. Molineu, N. Hernandez, T. Nguyen, G. Ibbott and D. Followill, "Credentialing results from IMRT irradiations of an anthropomorphic head and neck phantom" Medical physics **40** (2), 022101 (2013).

[2]G. P. Beyer, "Commissioning measurements for photon beam data on three TrueBeam linear accelerators, and comparison with Trilogy and Clinac 2100 linear accelerators" Journal of applied clinical medical physics / American College of Medical Physics **14** (1), 4077 (2013).

[3]D. P. Fontenla, J. J. Napoli and C. S. Chui, "Beam characteristics of a new model of 6-MV linear accelerator" Medical physics **19** (2), 343-349 (1992).

[4]Z. Chang, Q. Wu, J. Adamson, L. Ren, J. Bowsher, H. Yan, A. Thomas and F. F. Yin, "Commissioning and dosimetric characteristics of TrueBeam system: composite data of three TrueBeam machines" Medical physics **39** (11), 6981-7018 (2012).

[5]N. C. Ikoro, D. A. Johnson and P. P. Antich, "Characteristics of the 6-MV photon beam produced by a dual energy linear accelerator" Medical physics **14** (1), 93-97 (1987).

[6]R. J. Watts, "Comparative measurements on a series of accelerators by the same vendor" Medical physics **26** (12), 2581-2585 (1999).

[7]D. O. Findley, B. W. Forell and P. S. Wong, "Dosimetry of a dual photon energy linear accelerator" Medical physics **14** (2), 270-273 (1987).

[8]C. Glide-Hurst, M. Bellon, R. Foster, C. Altunbas, M. Speiser, M. Altman, D. Westerly, N. Wen, B. Zhao, M. Miften, I. J. Chetty and T. Solberg, "Commissioning of the Varian TrueBeam linear accelerator: a multi-institutional study" Medical physics **40** (3), 031719 (2013).

[9]D. Sjostrom, U. Bjelkengren, W. Ottosson and C. F. Behrens, "A beam-matching concept for medical linear accelerators" Acta oncologica **48** (2), 192-200 (2009).

[10]D. S. Followill, S. F. Kry, L. Qin, J. Lowenstein, A. Molineu, P. Alvarez, J. F. Aguirre and G. S. Ibbott, "The Radiological Physics Center's standard dataset for small field size output factors" Journal of applied clinical medical physics / American College of Medical Physics **13** (5), 3962 (2012).

[11]D. S. Followill, S. Kry, L. Qin, J. Lowenstein, A. Molineu, P. Alvarez, J. F. Aguirre and G. Ibbott, "Erratum: “The Radiological Physics Center’s standard dataset for small field size output factors”" Journal of applied clinical medical physics / American College of Medical Physics **15** (2) (2014).

[12]S. H. Cho, O. N. Vassiliev, S. Lee, H. H. Liu, G. S. Ibbott and R. Mohan, "Reference photon dosimetry data and reference phase space data for the 6 MV photon beam from varian clinac 2100 series linear accelerators" Medical physics **32** (1), 137-148 (2005).

[13]D. S. Followill, D. S. Davis and G. S. Ibbott, "Comparison of electron beam characteristics from multiple accelerators" International journal of radiation oncology, biology, physics **59** (3), 905-910 (2004).

[14]W. Mckinney, *Data Structures for Statistical Computing in Python*. (2010).

[15]S. Seabold and J. Perktold, *Statsmodels: Economic and Statistical Modeling with Python*, presented at the Python in Science Conference, Austin, TX, 2010.

[16]See supplementary material at [URL] for quantitative data tables of all energies.

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

[18]J. B. Smilowitz, I. J. Das, V. Feygelman, B. A. Fraass, S. F. Kry, I. R. Marshall, D. N. Mihailidis, Z. Ouhib, T. Ritter, M. G. Snyder and L. Fairobent, "AAPM Medical Physics Practice Guideline 5.a.: Commissioning and QA of Treatment Planning Dose Calculations — Megavoltage Photon and Electron Beams" Journal of applied clinical medical physics / American College of Medical Physics **16** (5) (2015).

[19]J. Kerns, D. Followill, J. Lowenstein, A. Molineu, P. Alvarez, P. Taylor, F. Stingo and S. Kry, "Technical Report: Reference photon dosimetry data for Varian accelerators based on IROC-Houston Site Visit Data" Medical physics **(in press)** (2015).

[20]P. Francescon, S. Cora and N. Satariano, "Calculation of k(Q(clin),Q(msr)) (f(clin),f(msr)) for several small detectors and for two linear accelerators using Monte Carlo simulations" Medical physics **38** (12), 6513-6527 (2011).

[21]G. A. Ezzell, J. W. Burmeister, N. Dogan, T. J. LoSasso, J. G. Mechalakos, D. Mihailidis, A. Molineu, J. R. Palta, C. R. Ramsey, B. J. Salter, J. Shi, P. Xia, N. J. Yue and Y. Xiao, "IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119" Medical physics **36** (11), 5359-5373 (2009).

[22]B. E. Nelms, G. Robinson, J. Markham, K. Velasco, S. Boyd, S. Narayan, J. Wheeler and M. L. Sobczak, "Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems" Practical radiation oncology **2** (4), 296-305 (2012).

[23]A. Molineu, D. S. Followill, P. A. Balter, W. F. Hanson, M. T. Gillin, M. S. Huq, A. Eisbruch and G. S. Ibbott, "Design and implementation of an anthropomorphic quality assurance phantom for intensity-modulated radiation therapy for the Radiation Therapy Oncology Group" International journal of radiation oncology, biology, physics **63** (2), 577-583 (2005).

[24]D. S. Followill, D. R. Evans, C. Cherry, A. Molineu, G. Fisher, W. F. Hanson and G. S. Ibbott, "Design, development, and implementation of the radiological physics center's pelvis and thorax anthropomorphic quality assurance phantoms" Medical physics **34** (6), 2070-2076 (2007).

[25]G. S. Ibbott, A. Molineu and D. S. Followill, "Independent evaluations of IMRT through the use of an anthropomorphic phantom" Technology in cancer research & treatment **5** (5), 481-487 (2006).

[26]C. L. Nelson, B. E. Mason, R. C. Robinson, K. D. Kisling and S. M. Kirsner, "Commissioning results of an automated treatment planning verification system" Journal of applied clinical medical physics / American College of Medical Physics **15** (5), 4838 (2014).

[27]J. D. Fontenot, "Evaluation of a novel secondary check tool for intensity-modulated radiotherapy treatment planning" Journal of applied clinical medical physics / American College of Medical Physics **15** (5), 4990 (2014).

[28]J. R. Kerns, D. Followill, J. Lowenstein, A. Molineu, P. Alvarez, P. A. Taylor, F. Stingo and S. F. Kry, "Technical Report: Reference photon dosimetry data for Varian accelerators based on IROC-Houston site visit data" Medical physics **43** (2016).

[29]S. F. Kry, A. Molineu, J. R. Kerns, A. M. Faught, J. Y. Huang, K. B. Pulliam, J. Tonigan, P. Alvarez, F. Stingo and D. S. Followill, "Institutional patient-specific IMRT QA does not predict unacceptable plan delivery" International journal of radiation oncology, biology, physics **90** (5), 1195-1201 (2014).

Vita

James Russell Kerns was born in Phoenix, Arizona on October 22, 1986 to Brian and Christine Kerns. After completing high school as valedictorian at Christian Heritage School, he entered into Point Loma Nazarene University, San Diego, California. He graduated cum laude in May of 2008 with a Bachelor of Science degree in Engineering/Physics. In August of 2010, he graduated from the Medical Physics program at the University of Texas Health Science Center at Houston Graduate School of Biomedical Sciences. Following that, he entered the MD Anderson medical physics therapy residency program. In September of 2012, he returned to the Graduate School of Biomedical Sciences to pursue his doctorate.

Permanent Address:
21830 Crystal Ann Ct
Perris, CA 92570