Reference dosimetry data and modeling challenges for Elekta accelerators based on IROC-Houston site visit data

James R. Kerns and David S. Followill

Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Imaging and Radiation Oncology Core-Houston, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Graduate School of Biomedical Sciences, The University of Texas Health Science Center-Houston, Houston, TX 77030, USA

Jessica Lowenstein, Andrea Molineu, Paola Alvarez, and Paige A. Taylor

Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Imaging and Radiation Oncology Core-Houston, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Stephen F. Kry^{a)}

Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Imaging and Radiation Oncology Core-Houston, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Graduate School of Biomedical Sciences, The University of Texas Health Science Center-Houston, Houston, TX 77030, USA

(Received 2 October 2017; revised 5 January 2018; accepted for publication 2 February 2018; published 1 April 2018)

Purpose: Reference dosimetry data can provide an independent second check of acquired values when commissioning or validating a treatment planning system (TPS). The Imaging and Radiation Oncology Core at Houston (IROC-Houston) has measured numerous linear accelerators throughout its existence. The results of those measurements are given here, comparing accelerators and the agreement of measurement versus institutional TPS calculations.

Methods: Data from IROC-Houston on-site reviews from 2000 through 2014 were analyzed for all Elekta accelerators, approximately 50. For each, consistent point dose measurements were conducted for several basic parameters in a water phantom, including percentage depth dose, output factors, small-field output factors, off-axis factors, and wedge factors. The results were compared by accelerator type independently for 6, 10, 15, and 18 MV. Distributions of the measurements for each parameter are given, providing the mean and standard deviation. Each accelerator's measurements were also compared to its corresponding TPS calculation from the institution to determine the level of agreement, as well as determining which dosimetric parameters were most often in error.

Results: Accelerators were grouped by head type and reference dosimetric values were compiled. No class of linac had better overall agreement with its TPS, but percentage depth dose and output factors commonly agreed well, while small-field output factors, off-axis factors, and wedge factors often disagreed substantially from their TPS calculations.

Conclusion: Reference data has been collected and analyzed for numerous Elekta linacs, which provide an independent way for a physicist to double-check their own measurements to prevent gross treatment errors. In addition, treatment planning parameters more often in error have been highlighted, providing practical caution for physicists commissioning treatment planning systems for Elekta linacs. © 2018 American Association of Physicists in Medicine [<https://doi.org/10.1002/mp.12865>]

Key words: IROC-Houston, linac, percentage depth dose, reference data, small field

1. INTRODUCTION

Accurate dose determination is the foundation of good radiotherapy. Dose calculations for radiation therapy patients are done through a treatment planning system (TPS) model, thus there is high priority on ensuring that the TPS model is as accurate as possible as detailed in several IAEA and AAPM reports.^{1–3} 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 achieved and maintained between the TPS model and physical linac characteristics. To achieve this, it can be helpful to check basic dosimetric measurements against those from similar linacs. In this way, the measurements from a

single linac can be validated as being reasonable; that is, a “sanity check” against other similar machines. This has been a primary motivation behind the publication of several sets of linac commissioning data,^{4–10} although such datasets are based almost exclusively on a small sample of linacs or even just a single linac.

It is the goal of the Imaging and Radiation Oncology Core in Houston (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 the dosimetric characteristics of their linacs. 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). Results from IROC-H have, unfortunately, consistently indicated that such agreement is often poor.^{11–13}

Since IROC-H has collected data on numerous linacs over many years, it was thought that the dosimetric characteristics of linac photon beams should be evaluated and presented to the community. This would allow individual institutions to compare their measurement data to IROC-H's large dataset distribution. 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. We previously reported on the dosimetric characteristics of Varian linacs;¹⁴ in this study we analyzed Elekta linacs.

Our goal was to analyze the characteristics of each linac model compared to the others. Models that performed similarly would be grouped together into classes based on statistical and clinical criteria. This basic dosimetric data from the different linac classes was tabulated for each evaluation condition (including mean values and statistical descriptors) and is included as a supplemental excel file. In this work, we also evaluated the level of agreement seen in the radiotherapy community between the institution's linac and TPS model. This was done to understand areas of common agreement and disagreement achieved clinically. Last, this agreement was categorized by TPS to determine areas where individual TPSs excelled or underperformed.

2. MATERIALS AND METHODS

2.A. 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. During an on-site audit, IROC-H physicists come to institutions with their own equipment to make point measurements in a water phantom for simple geometric conditions. The institution is asked to provide TPS calculated 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 $30 \times 30 \times 30$ cm³ water phantom at 100 cm source-to-surface distance. All data collected were point measurements. Percentage depth doses (PDD) were measured for 6×6 cm², 10×10 cm², and 20×20 cm² field sizes, each at 5, 10, 15, and 20 cm depths; at 10×10 cm² a d_{max} measurement was also taken. Off-axis factor (OAF) measurements were done with a 40×40 cm² field size at 5, 10, and 15 cm off-axis at d_{max} . Universal wedge factors were measured for the 60 degree wedge in a 10×10 cm² field at 10 cm depth and in a 15×15 cm² field at 15 cm depth. Two sets of output factors (OF) were measured. The first set, standard jaw-defined output factors, was taken at d_{max} for 30×30 , 20×20 , 15×15 , 10×10 , and 6×6 cm² field sizes. The second set, referred to as "IMRT output factors", was taken at 10 cm depth for 6×6 , 4×4 , 3×3 , and

2×2 cm² field sizes. All measurements except the small field output factors were measured with a Standard Imaging Exradin A12 (Standard Imaging, Madison, WI, USA) Farmer-type 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 the most widely used energies and are thus presented here.

2.B. Data analysis

All data analysis and visualization was done using the general programming language Python and the open-source "pandas" Python package.¹⁵

While it is common to differentiate linacs by model name, Elekta's 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 nominal linac model name.

To categorize the linacs into dosimetrically similar classes, linacs were first grouped by head configuration: MLCi, MLCi2, Beam Modulator (BMod), and Agility. Each energy was analyzed separately and considered independent. Two criteria were then applied to evaluate dosimetric comparability between head models at the given energy: a clinical and a statistical criterion. The clinical criterion was added because small but statistically significant differences were not uncommon, but were also deemed to not represent a meaningfully different dosimetric value. Each linac head's mean value for each measurement location was compared to the others' (e.g., PDD(10×10 cm², at 5 cm depth)) 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. In addition, the clinical criterion tested if the median values had at least a 0.5% difference in the local value (e.g., was outside the range of 49.75–50.25 on a 50% isodose value). This value was chosen since it is approximately equal to the standard deviation seen in IROC-H measurements. If a linac head model had both a statistically different and clinically different value from the other head models, it was put into its own class.

The dosimetric characteristics (PDD, OF, OAF, etc.) from each of the resulting linac classes were summarized and evaluated. The data were considered, and 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.

In addition to basic dosimetric measurements, we also evaluated how well institutions modelled their linacs. The dosimetric characteristics of each institution's linac and their specific TPS calculation were compared for the site visit

conditions described above. For each accelerator and each measured parameter, the institution's calculated value from their clinical TPS was assessed. The agreement across the different dosimetric parameters was investigated to determine areas of common disagreement. Each linac measurement set was compared to its own TPS calculations, and then binned according to the linac class and binned according to the TPS vendor to examine possible trends with class or TPS. Differences in agreement between TPS and measurement were flagged based on one of two criteria: a large standard deviation in the agreement ($>1\%$ on average); or a systematic difference in the median TPS-to-measurement ratio ($>1\%$ from unity).

3. RESULTS

3.A. Model comparison

There were four Elekta linac head configurations analyzed at each energy: MLCi, MLCi2, BMod, and Agility. Because the BMod MLC dimensions do not allow for regular field size definition, measurements on the BMod head were not directly 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 statistically or clinically equivalent at all locations, and were therefore combined into an "MLCi" class. The manufacturer claimed that the MLCi2 head was expected to be dosimetrically equivalent to the MLCi head for the conditions examined in this study, giving support to the dosimetric merging of these two models.

The resulting classes were consistent across all energies with each head. That is, at each energy there were three classes, named for the head they represented: Agility, MLCi, and BMod, where the MLCi class represented the MLCi and MLCi2 head.

Figure 1 shows an overall comparison of dosimetric performance between classes. Because the measurement conditions (field sizes) of the BMod head were not the same as the others, a direct comparison is not possible with the BMod, so it is left out of the overall comparison. The figure is intended to quickly show differences between the MLCi class and the Agility class. The color of the squares represents the maximum difference that the MLCi class had for that parameter in comparison to the Agility class (selected as the arbitrary comparison class). For example, for the PDD parameter, the darkness of the square is the greatest difference between the different field size and depth conditions evaluated. Lighter colors indicate smaller differences from the Agility class, while darker colors 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.

3.B. 6 MV

The 6 MV summary data is given in Table I. Further statistical metrics (including the 5th and 95th percentile ranges) and complete PDD data (including other field sizes) are given in the supplemental online content. For the BMod class, the field sizes used are noted in the table legend.

FIG. 1. Heatmap comparing the Agility class to the MLCi class by parameter. Parameters include percent depth dose (PDD), jaw-defined output factor (Jaw OF), small-field output factors as used for IMRT (IMRT OF), off-axis factors (OAF), and wedge factors defined using the universal wedge (Universal). Darker color indicates larger maximum differences from the agility class. "N/A" means that measurements were not available for that parameter. [Color figure can be viewed at wileyonlinelibrary.com]

Between classes, the PDDs were all quite similar, having an average interclass variation of 0.7% and average intra-class variation of 0.8% (averaged over all field sizes and depths). Almost all other parameters also showed similar consistency between classes. The largest difference between the Agility and MLCi classes of accelerator were in the off-axis factors. The Agility class had less pronounced horns compared to the MLCi. 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, indicating relatively larger variability between different individual machines.

TABLE I. 6 MV measured Elekta data. Median values are given and the standard deviation is given in parentheses. N is the number of measurements.

N	Class		
	Agility 10	MLCi 40	Bmod ^a 6
PDD (10 × 10)			
5 cm	0.871 (0.004)	0.868 (0.004)	0.870 (0.003)
10 cm	0.677 (0.004)	0.675 (0.004)	0.676 (0.004)
15 cm	0.519 (0.004)	0.516 (0.004)	0.519 (0.003)
20 cm	0.396 (0.004)	0.393 (0.004)	0.395 (0.003)
N	10	33	3
Jaw OF			
6 × 6 cm ²	0.965 (0.008)	0.965 (0.004)	0.965 (0.006)
15 × 15 cm ²	1.026 (0.008)	1.030 (0.004)	1.028 (0.014)
20 × 20 cm ²	1.052 (0.008)	1.054 (0.005)	N/A
30 × 30 cm ²	1.072 (0.009)	1.075 (0.006)	N/A
N	10	39	0
OAF			
5 cm	1.021 (0.007)	1.025 (0.007)	N/A
10 cm	1.051 (0.009)	1.054 (0.013)	N/A
15 cm	1.052 (0.009)	1.061 (0.015)	N/A
N	5	11	6
IMRT OF			
2 × 2 cm ²	0.793 (0.006)	0.792 (0.005)	0.756 (0.006)
3 × 3 cm ²	0.837 (0.007)	0.836 (0.003)	0.839 (0.001)
4 × 4 cm ²	0.871 (0.002)	0.870 (0.003)	0.866 (0.004)
6 × 6 cm ²	0.925 (0.002)	0.924 (0.002)	0.928 (0.002)
N	3	23	4
Universal			
60°	0.270 (0.006)	0.268 (0.005)	0.271 (0.004)
60° (15 × 15)	0.277 (0.008)	0.277 (0.005)	N/A

^aThe BMod head cannot form the same field sizes as the other classes; PDD data was taken at 10.4 × 10.4 cm², jaw output factors at 6.4 × 6.4, 10.4 × 10.4, and 15.2 × 15.2 cm²; IMRT-style output factors at 6.4 × 6.4, 4 × 4, 3.2 × 3.2, and 1.6 × 1.6 cm².

3.C. 10 MV

The main 10 MV results are shown in Table II. Additional statistical measures and PDDs for other field sizes are included in the supplemental online content. Similar to the 6 MV data, 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 (averaged over all field sizes and depths). At 10 MV the median off-axis factors were similar between accelerator classes, but the standard deviations were notably higher than for other parameters, indicating (as with 6 MV) that there is increased variability between machines for off-axis factors

TABLE II. 10 MV measured Elekta data. Median values are given and the standard deviation is given in parentheses. N is the number of measurements.

N	Class		
	Agility 5	MLCi 19	Bmod ^a 5
PDD (10 × 10)			
5 cm	0.910 (0.004)	0.909 (0.005)	0.907 (0.010)
10 cm	0.727 (0.002)	0.727 (0.004)	0.727 (0.008)
15 cm	0.576 (0.002)	0.575 (0.005)	0.578 (0.009)
20 cm	0.455 (0.002)	0.454 (0.004)	0.457 (0.007)
N	5	16	2
Jaw OF			
6 × 6 cm ²	0.964 (0.003)	0.964 (0.006)	0.973 (0.005)
15 × 15 cm ²	1.035 (0.002)	1.029 (0.005)	1.021 (0.005)
20 × 20 cm ²	1.055 (0.014)	1.052 (0.007)	N/A
30 × 30 cm ²	1.074 (0.018)	1.070 (0.006)	N/A
N	5	19	0
OAF			
5 cm	1.030 (0.014)	1.030 (0.009)	N/A
10 cm	1.044 (0.017)	1.043 (0.013)	N/A
15 cm	1.053 (0.023)	1.051 (0.017)	N/A
N	2	7	4
IMRT OF			
2 × 2 cm ²	0.793 (0.011)	0.800 (0.008)	0.750 (0.003)
3 × 3 cm ²	0.857 (0.003)	0.858 (0.004)	0.863 (0.002)
4 × 4 cm ²	0.891 (0.004)	0.894 (0.006)	0.887 (0.002)
6 × 6 cm ²	0.936 (0.004)	0.939 (0.003)	0.940 (0.001)
N	2	12	4
Universal			
60°	0.289 (0.001)	0.283 (0.006)	0.287 (0.006)
60° (15 × 15)	0.290 (0.008)	0.282 (0.005)	N/A

^aThe BMod head cannot form the same field sizes as the other classes; PDD data was taken at 10.4 × 10.4 cm², jaw output factors at 6.4 × 6.4, 10.4 × 10.4, and 15.2 × 15.2 cm²; IMRT-style output factors at 6.4 × 6.4, 4 × 4, 3.2 × 3.2, and 1.6 × 1.6 cm².

as compared to other parameters. Differences between classes in terms of mean values were seen for output factors and wedge factors.

3.D. 15 MV

The main 15 MV measurement data are shown in Table III, although some specific values were not collected. In addition, no small-field output factors (IMRT OF) were measured at this energy as this measurement suite was not implemented at the time of these measurements. Additional statistical measures and PDDs for other field sizes are included in the supplemental online content.

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

N	Class		
	Agility 3	MLCi 10	Bmod ^a 2
PDD (10 × 10)			
5 cm	0.931 (0.003)	0.934 (0.003)	0.924 (0.018)
10 cm	0.755 (0.003)	0.759 (0.003)	0.756 (0.010)
15 cm	0.608 (0.002)	0.610 (0.003)	0.608 (0.008)
20 cm	0.489 (0.003)	0.490 (0.003)	0.489 (0.010)
N	3	8	0
Jaw OF			
6 × 6 cm ²	0.957 (0.001)	0.955 (0.004)	
15 × 15 cm ²	N/A	1.033 (0.001)	
20 × 20 cm ²	1.052 (0.004)	1.057 (0.003)	N/A
30 × 30 cm ²	N/A	1.072 (0.008)	
N	3	9	0
OAF			
5 cm	1.033 (0.006)	1.033 (0.009)	
10 cm	1.061 (0.002)	1.056 (0.010)	N/A
15 cm	1.081 (0.015)	1.080 (0.012)	
N	0	0	0
IMRT OF			
2 × 2 cm ²			
3 × 3 cm ²	N/A	N/A	N/A
4 × 4 cm ²			
6 × 6 cm ²			
N	1	4	2
Universal			
60°	0.272	0.271 (0.001)	0.279 (0.001)
60° (15 × 15)	0.279	0.277 (0.001)	N/A

^aThe BMod head cannot form the same field sizes as the other classes; PDD data was taken at 10.4 × 10.4 cm², jaw output factors at 6.4 × 6.4, 10.4 × 10.4, and 15.2 × 15.2 cm²; IMRT-style output factors at 6.4 × 6.4, 4 × 4, 3.2 × 3.2, and 1.6 × 1.6 cm².

3.E. 18 MV

The main 18 MV measurement data are shown in Table IV, although little data existed for the BMod class at this energy. Additional statistical measures and PDDs for other field sizes are included in the supplemental online content. The PDDs of the Agility and MLCi classes are very similar, with an average interclass variation of 0.2% (averaged over all field sizes and depths). In contrast, relative large differences between classes were seen with the output factors, although the Agility class had notably large intraclass variability with an average of 1.9%. The off-axis factors also showed differences between classes.

TABLE IV. 18 MV measured Elekta data. Median values are given and the standard deviation is given in parentheses. N is the number of measurements.

N	Class		
	Agility 5	MLCi 26	Bmod ^a 1
PDD (10 × 10)			
5 cm	0.955 (0.006)	0.956 (0.006)	0.959
10 cm	0.782 (0.005)	0.783 (0.005)	0.788
15 cm	0.635 (0.005)	0.635 (0.005)	0.641
20 cm	0.514 (0.005)	0.515 (0.004)	0.521
N	5	20	0
Jaw OF			
6 × 6 cm ²	0.956 (0.009)	0.957 (0.008)	
15 × 15 cm ²	1.024 (0.024)	1.034 (0.005)	
20 × 20 cm ²	1.055 (0.013)	1.058 (0.005)	N/A
30 × 30 cm ²	1.066 (0.028)	1.077 (0.005)	
N	5	21	0
OAF			
5 cm	1.036 (0.009)	1.039 (0.011)	
10 cm	1.059 (0.010)	1.055 (0.012)	N/A
15 cm	1.044 (0.014)	1.053 (0.014)	
N	1	5	0
IMRT OF			
2 × 2 cm ²	0.784	0.786 (0.010)	
3 × 3 cm ²	0.862	0.861 (0.005)	N/A
4 × 4 cm ²	0.902	0.900 (0.003)	
6 × 6 cm ²	0.946	0.944 (0.001)	
N	2	16	0
Universal			
60°	0.277 (0.002)	0.272 (0.005)	N/A
60° (15 × 15)	N/A	0.275 (0.005)	

^aThe BMod head cannot form the same field sizes as the other classes; PDD data was taken at 10.4 × 10.4 cm², jaw output factors at 6.4 × 6.4, 10.4 × 10.4, and 15.2 × 15.2 cm²; IMRT-style output factors at 6.4 × 6.4, 4 × 4, 3.2 × 3.2, and 1.6 × 1.6 cm².

3.F. Measurement and TPS agreement

While the above data present the measurement data for different classes of accelerator, here we present how well each institution was able to model their linac. While each institution was compared to its own TPS, the results were binned in

two different ways. In Fig. 2 the agreement has been binned according to the linear accelerator class. Notably, there were differences in which parameters were problematic to model based on the class of accelerator. There were also differences based on the energy. Regardless, most of the disagreement was because of the standard deviation criterion — that is,

FIG. 2. Agreement between linear accelerator dosimetry characteristics (including PDDs for $6 \times 6 \text{ cm}^2$, $10 \times 10 \text{ cm}^2$, and $20 \times 20 \text{ cm}^2$) and the institution's TPS, grouped according to the accelerator class. The shade of gray denotes the type of disagreement with the TPS calculations. White indicates that the both the standard deviation and median agreement were good. Light gray 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, which was defined as three linacs. [Color figure can be viewed at [wileyonlinelibrary.com](#)]

FIG. 3. Differences between the institution's linear accelerator dosimetric characteristics (including PDDs for $6 \times 6 \text{ cm}^2$, $10 \times 10 \text{ cm}^2$, and $20 \times 20 \text{ cm}^2$) and the institution's TPS, grouped by TPS. The shade of gray denotes the type of disagreement with the TPS calculations. White indicates that the both the standard deviation and median agreement were good. Lighter gray 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, which was defined as 3. [Color figure can be viewed at [wileyonlinelibrary.com](#)]

there was a lot of variability in how well the community modelled the parameter, although on average the modelling was correct. This was particularly apparent for the universal wedge factors and off-axis factors. Some differences were also systematic though, meaning that, consistently, institutions tended to systematically model their linacs incorrectly. This was seen most often for small-field IMRT output factors and universal wedge factors.

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 (regardless of linac class) and the results are shown in Fig. 3. 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 large standard deviations. Eclipse however had good agreement for the 6×6 cm² PDD and jaw output factors where Pinnacle performed less well on these metrics. At 10 and 15 MV, Eclipse had good agreement for all parameters while Pinnacle had a few areas of disagreement.

4. DISCUSSION

The goal of this study, as with our prior study of Varian machines,^{12,14} was to present dosimetric characteristics (median values and descriptions of the distributions) for the different linacs currently available. To facilitate such data presentation, linac models were first combined into dosimetrically equivalent classes. The idea being that several different models could be aggregated into a distinct few. In addition, the dosimetric characteristics of these classes relative to one another were of interest. Note that comparing characteristics does not imply one class is superior, only that there may or may not be differences between them. Reference data was compiled so that physicists can compare their own measurement values to these reference distributions.

An important question for the reader 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 shown here. When an individual machine performs slightly differently than a reference dataset, it is of interest to know how if that difference is reasonable. To provide this perspective, statistical metrics of the distribution such as standard deviation are given so that physicists are aware of which parameters should be expected to show a wide or narrow range of values. While a specific machine may show differences from the median linac data shown here, physicists can flag measurements that are outside a reasonable range for additional study and verification. There should be careful analysis of these differences to rule out such things as measurement error. Ultimately, and critically, the true test of performance is the agreement between the dosimetric measurement and TPS calculation.

The data of Figs. 2 and 3 show where physicists struggle most often when commissioning and validating their beam

models. These are cases where, quite simply, the beam model does not accurately describe the true machine output. There was a lot of variations seen with different beam energies, linac classes, and TPSs. Importantly, there was no class or TPS without areas of disagreement, and particular care should be taken when measuring, modeling, and validating those parameters. All steps of the commissioning and modeling process can introduce errors, and therefore all steps must be undertaken with care. Measurement errors in the commissioning process can certainly lead to disagreement between the TPS and linac output. Most of the conditions evaluated in this work are in highly uniform regions of dose and therefore not particularly prone to commissioning measurement errors, although the small field output factors require additional attention on this subject. Generation of the beam model based on the commissioning data, however, is a process identified as prone to error.^{13,16} These sorts of errors in beam modeling have been implicated in the majority of IROC-Houston anthropomorphic phantom failures.¹⁷ Beam modeling is therefore a task that requires increased attention by the medical physics community.

5. CONCLUSION

In this study, Elekta linac measurements from across 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 value and variability in a host of dosimetric parameters. 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 anomalies that would warrant further attention. Any sizeable difference should be investigated and confirmed. Physicists should be particularly cautious in measuring, modeling, and validating dosimetric parameters identified as problematic in this study. No class of accelerator had better agreement with its respective TPS than another across all energies, although energy-specific exceptions existed.

ACKNOWLEDGMENTS

Many thanks go to Paul Hougin for querying and compiling the site visit data. This work was supported by Public Health Service Grant CA180803 and CA214526 awarded by the National Cancer Institute, United States Department of Health and Human Services.

CONFLICT OF INTEREST

The authors have no conflicts to disclose.

^{a)}Author to whom correspondence should be addressed. Electronic mail: sfkry@mdanderson.org

REFERENCES

1. Fraass B, Doppke K, Hunt M, et al. American association of physicists in medicine radiation therapy committee task group 53: quality assurance for clinical radiotherapy treatment planning. *Med Phys*. 1998;25:1773–1829.
2. International Atomic Energy Agency. *Commissioning and Quality Assurance of Computerized Planning Systems for Radiation Treatment of Cancer. TRS 430*. Vienna: International Atomic Energy Agency; 2004.
3. Smilowitz JB, Das IJ, Feygelman V, et al. AAPM, medical physics practice guideline 5.a.: commissioning and QA of Treatment planning dose calculations, megavoltage photon and electron beams. *J Appl Clin Med Phys*. 2015;15:14–34. PMID: 26894369.
4. Ikoro NC, Johnson DA, Antich PP. Characteristics of the 6-MV photon beam produced by a dual energy linear accelerator. *Med Phys*. 1987;14:93–97.
5. Findley DO, Forell BW, Wong PS. Dosimetry of a dual photon energy linear accelerator. *Med Phys*. 1987;14:270–273.
6. Palta JR, Komanduri A, Daftari I, Suntharalingam N. Characteristics of photon beams from Philips SL25 linac accelerators. *Med Phys*. 1989;17:106–116.
7. Fontenla DP, Napoli JJ, Chui CS. Beam characteristics of a new model of 6-MV linear accelerator. *Med Phys*. 1992;19:343–349.
8. Watts RJ. Comparative measurements on a series of accelerators by the same vendor. *Med Phys*. 1999;26:2581–2585.
9. Chang Z, Wu Q, Adamson J, et al. Commissioning and dosimetric characteristics of TrueBeam system: composite data of three TrueBeam machines. *Med Phys*. 2012;39:6981–7018.
10. Glide-Hurst C, Bellon M, Foster R, et al. Commissioning of the Varian TrueBeam linear accelerator: a multi-institutional study. *Med Phys*. 2013;40:031719.
11. Followill DS, Kry SF, Qin L, et al. The radiological physics center's standard dataset for small field size output factors. *J Appl Clin Med Phys*. 2012;13:3962. Erratum 2014;15(2):356–57.
12. Kerns JR, Followill DS, Lowenstein J, et al. Agreement between institutional measurements and treatment planning system calculations for basic dosimetric parameters as measured by the imaging and radiation oncology core-Houston. *Int J Radiat Oncol Biol Phys*. 2016;95:1527–1534. PMID: 27315667.
13. Kry SF, Dromgoole L, Alvarez P, et al. Radiotherapy deficiencies identified during on-site dosimetry visits by the IROC Houston QA center. *Int J Radiat Oncol Biol Phys*. 2017;99:1094–1100.
14. Kerns JR, Followill DS, Lowenstein J, et al. Technical report: reference photon dosimetry data for Varian accelerators based on IROC-Houston site visit data. *Med Phys*. 2016;43:2374.
15. McKinney W. Data structures for statistical computing in python. 2010.
16. Wexler A, Bruce G, Goddu S, et al. FMEA of manual and automated methods for commissioning a radiotherapy treatment planning system. *Med Phys*. 2017;44:4415–4425.
17. Kerns JR, Stingo F, Followill DS, Howell RM, Melancon A, Kry SF. Treatment planning system calculation errors are present in most imaging and radiation oncology Core-Houston phantom failures. *Int J Radiat Oncol Biol Phys*. 2017;98:1197–1203.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

Table SI. 2000–2014 Data to accompany Kerns et al, “Reference dosimetry data and modeling challenges for Elekta accelerators based on IROC-Houston Site Visit Data”.