Physics in Medicine & Biology

IPEM
Institute of Physics and
Engineering in Medicine

PAPER • OPEN ACCESS

Treatment plan complexity does not predict IROC Houston anthropomorphic head and neck phantom performance

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

CrossMark

RECEIVED
1 May 2018

REVISED
23 July 2018

ACCEPTED FOR PUBLICATION
19 September 2018

PUBLISHED
17 October 2018

PAPER

Treatment plan complexity does not predict IROC Houston anthropomorphic head and neck phantom performance

Mallory C Glenn[1,2], Victor Hernandez[3], Jordi Saez[4], David S Followill[1,2], Rebecca M Howell[1,2], Julianne M Pollard-Larkin[1,2], Shouhao Zhou[1,5] and Stephen F Kry[1,2,6]

[1] The University of Texas MD Anderson Cancer Center, UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, United States of America
[2] Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, United States of America
[3] Department of Medical Physics, Hospital Universitari Sant Joan de Reus, Tarragona, Spain
[4] Department of Radiation Oncology, Hospital Clínic de Barcelona, Barcelona, Spain
[5] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, United States of America
[6] Author to whom any correspondence should be addressed.

E-mail: sfkry@mdanderson.org

Keywords: anthropomorphic phantom, IROC, credentialing, intensity-modulated radiotherapy, quality assurance, plan complexity

Abstract

Previous works indicate that intensity-modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) plans that are highly complex may produce more errors in dose calculation and treatment delivery. Multiple complexity metrics have been proposed and associated with IMRT QA results, but their relationships with plan performance using *in situ* dose measurements have not been thoroughly investigated. This study aimed to evaluate the relationships between IMRT treatment plan complexity and anthropomorphic phantom performance in order to assess the extent to which plan complexity is related to dosimetric performance in the IROC phantom credentialing program. Sixteen complexity metrics, including the modulation complexity score (MCS), several modulation indices, and total monitor units (MU) delivered, were evaluated for 343 head and neck phantom irradiations, comprising both IMRT (step-and-shoot and sliding window techniques) and VMAT. Spearman's correlations were used to explore the relationship between complexity and plan performance, as measured by the dosimetric differences between the treatment planning system (TPS) and thermoluminescent dosimeter (TLD) measurement, as well as film gamma analysis. Relationships were likewise determined for several combinations of subpopulations, based on the linear accelerator model, TPS used, and delivery modality. Evaluation of the complexity metrics presented here yielded no significant relationships ($p > 0.01$, Bonferroni-corrected) and all correlations were weak (less than ± 0.30). These results indicate that complexity metrics have limited predictive utility in assessing plan performance in multi-institutional comparisons of IMRT plans. Other factors affecting plan accuracy, such as dosimetric modeling or multileaf collimator (MLC) performance, should be investigated to determine a more probable cause for dose delivery errors.

1. Introduction

Intensity modulated radiation therapy (IMRT), including volumetric modulated arc therapy (VMAT), is currently a standard of care technique for many disease sites. This delivery technique allows for better dose conformity than traditional 3D conformal radiation therapy while simultaneously sparing normal tissues from extraneous radiation dose. However, this technique also requires variations in multileaf collimator (MLC) motion, as well as gantry rotation speed and dose rate in some cases. Such sources of variability increase the plan 'complexity', a term describing the frequency and amplitude of fluctuations in IMRT dose distributions (Mohan *et al* 2000). Thus, a simple IMRT treatment consists of large beam apertures of regular shapes, and complex IMRT beams tend to have small, narrow, or irregularly shaped apertures.

Many have previously reported that the degree of complexity (i.e. beam modulation) may be associated with greater uncertainties in radiation treatments (McNiven *et al* 2010, Younge *et al* 2012, Masi *et al* 2013, Crowe *et al*

2014, Du *et al* 2014, Park *et al* 2014, Götstedt *et al* 2015). This is a logical supposition as high-complexity treatment plans include more challenging dose calculations and increased sensitivity to mechanical delivery performance, especially when using very small fields. The potential for delivery errors associated with highly complex plans has ushered the need to characterize and mitigate complexity in IMRT. To do so, researchers have developed several metrics as indicators of plan complexity, consisting of both fluence map-based and aperture-based metrics (McNiven *et al* 2010, Younge *et al* 2012, Masi *et al* 2013, Crowe *et al* 2014, Du *et al* 2014, Park *et al* 2014, Götstedt *et al* 2015). Fluence map-based metrics, such as the modulation index proposed by Webb, measure the variations in photon fluence between adjacent pixels in a fluence map (Webb 2003). Aperture-based approaches measure complexity by directly measuring the irregularity of the treatment field, as defined by the MLC, although some metrics also evaluate other plan parameters, such as leaf speed and variations of the dose rate and gantry speed.

Complexity metrics have also been suggested to be a time-efficient complement to current IMRT quality assurance (QA) methods, as they further inform the extent of beam modulation in the treatment and therefore may flag cases where modulation is higher than would normally be expected. This application is of particular interest to the Imaging and Radiation Oncology Core Houston (IROC) Quality Assurance Center. IROC seeks to confirm that institutions participating in National Cancer Institute sponsored clinical trials, including those utilizing IMRT, can calculate and deliver radiation doses consistently and accurately. For IMRT, this is done through the use of end-to-end anthropomorphic phantom irradiations whereby institutions irradiate an IROC phantom containing thermoluminescent dosimeters (TLD) and radiochromic film (Molineu *et al* 2005). The measured dose distribution is then compared to the institution's calculated dose distribution. Yet, even with improvements in IMRT planning and delivery over time, and relatively lax dosimetric agreement criteria for the phantom (7%), a sizeable percentage of institutions continue to fail the phantom test; only 85%–90% of institutions have passed in recent years (Molineu *et al* 2013). Of concern, dose calculation inaccuracies have been shown to be a leading cause of treatment delivery error (Carson *et al* 2016, Kerns *et al* 2017). If complexity could be used to predict treatment accuracy, such analysis would aid in identifying the cause of phantom failures.

Therefore, the purpose of this study was to investigate the relationship between treatment plan complexity and treatment accuracy, with the aim of identifying which complexity metrics best predict planning and/or delivery errors and how much complexity contributes to dosimetric errors in IMRT delivery. To date, a comprehensive evaluation of a broad range of complexity metrics has not been done, particularly using a single, controlled patient geometry. This evaluation, as performed using IROC phantoms, has the potential to identify metrics related to the agreement between dose calculations and measurements. In addition, the information produced in this work may be used to better inform the treatment planning process or guide QA testing in order to mitigate potential errors.

2. Methods

2.1. Phantom plans

A total of 343 IMRT and VMAT irradiations of IROC's head and neck (H&N) phantom (including 11 repeat irradiations) were performed by 312 different institutions between September 2011 and December 2016 as part of IROC's phantom credentialing program. The H&N phantom was chosen for evaluation because it is the most frequently irradiated phantom and is the default credentialing phantom for IMRT. The phantom contains two PTV targets and an organ at risk, and the dose was assessed with six double-loaded TLD and two sheets of film. Phantom performance was evaluated by comparing the dose calculated by the TPS with the dose actually delivered. Additional details on the phantom and analysis program are available in the literature (Molineu *et al* 2005).

Despite the uniform geometry and planning objectives, the phantom irradiations were done with a broad cohort of delivery methods, and thereby a variety of different complexities. The demographics of these are detailed in table 1. This cohort was limited to 6 MV photon treatments administered by Varian and Elekta linear accelerators, which account for the vast majority of H&N phantom irradiations. For all of these irradiations, institutions irradiated identical phantoms and were instructed to follow the same IROC protocol for phantom irradiation, thus achieving very similar dose distributions (Molineu *et al* 2005).

2.2. Complexity metrics

In this study, sixteen identified measures of complexity were computed for each of the 343 phantom plans in order to provide a comprehensive view of complexity definitions, including both aperture-based and fluence map-based metrics. Here we considered both established measures of IMRT complexity from the literature, as well as several additional metrics describing variations within the MLC position, gantry position, and dose rate, thus allowing for a more well-rounded assessment of IMRT treatment delivery. For each of the metrics described herein, complexity was calculated for each beam or arc in a treatment plan, and subsequently averaged for all beams or arcs to yield the plan's average complexity. The following indices were evaluated:

Table 1. Demographics of IMRT technique, treatment planning system (TPS), linear accelerator manufacturer, and linac-TPS combination for the sample of this study.

	N	%
IMRT technique		
Dynamic MLC	93	27.1
Static MLC	43	12.5
VMAT	207	60.3
Linear accelerator manufacturer		
Elekta	39	11.4
Varian	304	88.6
Treatment planning system (TPS)		
Eclipse	249	72.6
Pinnacle	69	20.1
RayStation	9	2.6
Other[a]	16	4.7
Linac-TPS combination		
Elekta-Eclipse	1	0.3
Elekta-Pinnacle	24	7.0
Elekta-RayStation	4	1.2
Varian-Eclipse	248	72.3
Varian-Pinnacle	45	13.1
Varian-RayStation	5	1.5

[a] Other TPS include XiO, iPlan, Monaco, and Oncentra.

(a) Total MU delivered. In general, a high degree of complexity is typically associated with a large number of MU; this has been used as a surrogate for treatment plan complexity previously, though correlations have not been definitive (Masi *et al* 2013, Agnew *et al* 2014, Kry *et al* 2014, Crowe *et al* 2015).

(b) Modulation complexity score (MCS) (McNiven *et al* 2010). The MCS aims to characterize beam complexity in terms of the aperture shapes and area present throughout treatment. This metric was originally conceptualized for step-and-shoot delivery but was later adapted for sliding window and VMAT techniques (Masi *et al* 2013). A smaller MCS indicates more complex apertures.

(c) Edge metric (EM) (Younge *et al* 2012). This metric defines complexity as a ratio of MLC side length edge to aperture area. In this study the original recommendations for the input parameters ($C1 = 0$ and $C2 = 1$) were used. A larger EM index signifies larger positional differences between adjacent leaves.

(d) Plan irregularity (PI) and plan modulation (PM) (Du *et al* 2014). PI describes the non-circularity of the MLC apertures, whereas PM indicates to what extent the beam delivery is delivered into smaller apertures.

(e) Modulation indices (MI_{speed}, MI_{accel}, MI_{total}) (Park *et al* 2014). MI_{speed} and MI_{accel} evaluate the extent of variation within the speed and acceleration of the MLC, respectively. In addition to these variations, MI_{total} also considers variations in gantry speed and dose rate to quantify the total delivery complexity.

(f) Leaf travel (LT) (Masi *et al* 2013). LT indicates the average distance traveled by the MLC leaves. Because LT was originally designed for single full arc treatments, this metric is divided by the treatment's corresponding arc length to allow for comparisons with treatments with multiple arcs or partial arcs. Here the metric is denoted 'LT/AL' to establish this modification.

(g) Mean dose rate variation. This metric is defined as the sum of dose rate variations from all control points, divided by arc length (to allow comparisons of plans with different numbers of control points).

(h) Mean gantry speed variation. Like mean dose rate variation, this index is the sum of variations in gantry speed, divided by arc length.

(i) Percentage of MLC gaps >10 mm. This metric describes the cumulative window width for MLC leaves. Plans with a small cumulative metric indicate the use of many small MLC gaps. Here we choose 10 mm as an appropriate threshold in order to delineate the difference between large and small leaf gaps.

(j) Mean tongue and groove index. This index is calculated for each pair of leaves and for each control point as a fraction of the MLC gap adjacent to a consecutive leave. The mean index is then obtained by averaging all the pairs of leaves inside the beam and all the control points.

(k) MLC interdigitation. This index characterizes the overlap between consecutive leaves from opposing banks with respect to the maximum interdigitation, taking into account the complete irradiated area outline of the MLC.

(l) Mean MLC speed variation. The mean variation in MLC speed is computed as the sum of MLC speed variations (i.e. MLC accelerations), divided by the total leaf travel.

(m) Mean Gap speed variation. The mean variation of gap sizes is computed as the sum of gap size variations, divided by the total leaf travel.

2.3. Data analysis

To quantify the previously described complexity indices, a MATLAB-based software called PlanAnalyzer was used to read the DICOM plans submitted by the institutions undergoing phantom credentialing (Hernandez *et al* 2018). These measures of plan complexity were compared against the dosimetric error found for each delivered plan. The average TLD error was defined as the average magnitude percentage difference between the TPS-calculated doses and the corresponding measured doses for the TLD in the H&N phantom (six TLD per phantom). Because point dosimetry may not fully characterize the irradiation conditions, plan error was also measured by the percentage pixels passing from radiochromic film gamma analysis, following IROC's protocol for analysis with the criteria of 7% dose agreement and 4 mm distance to agreement (Molineu *et al* 2013).

Correlations between complexity metrics and phantom plan error were determined using Spearman's rank-order correlation coefficients (with Bonferroni corrections applied for multiple comparisons). For the purposes of this work, the strength of the association in absolute value was defined as follows: 0–0.19 was regarded as 'no correlation', 0.20–0.39 as 'weak', 0.40–0.59 as 'moderate', and 0.60–1 as 'strong'. Correlations were evaluated for the entire sample, as well as according to TPS (Pinnacle and Eclipse), machine type, and delivery technique, as delineated in table 1. Similarly, poor phantom results, those with at least one TLD measuring $>$5% error, were segregated, and the same analyses were applied to visualize whether such clinically underperforming plans had distinguishing features.

3. Results

3.1. The relationship between TLD-based plan error and complexity metrics

Despite the uniformity of the phantom and dose objectives, the plans in this study had a comprehensive assortment of treatment complexities; for example, MU used in delivery ranged from 458 to 3358 with a mean of 1883. Figure 1 shows the distributions of the MCS and corresponding plan error for the total sample and multiple subsamples examined in this work. Visually, these distributions represent poor ability of complexity metrics to be utilized as a means of distinguishing irradiations prone to error, at least under the circumstances examined herein. Other complexity metrics appeared similarly and yielded indistinguishable relationships.

Relationships between complexity metrics and plan error are shown in table 2. The only index to achieve significance was MLC interdigitation ($p < 0.003$); no other complexity metric significantly predicted dosimetric inaccuracies in the calculation or delivery of the radiation dose. However, even this single significant relationship, found for the total sample and also only Varian machines, had a very low correlation strength ($|r_s| < 0.18$), which indicated no clinically meaningful relationship by our criteria. The highest correlation was found between LT/AL and Elekta machines supported by Pinnacle ($r_s = -0.395, p = 0.116$), but this correlation coefficient was still classified as 'weak' and was not found to be a significant relationship, partly due to the small subsample size. From this data, it is evident that complexity metrics are not related to the TLD error observed in IROC's H&N phantom practice, regardless of delivery technique, TPS, or machine manufacturer.

3.2. The relationship between gamma-based plan error and complexity metrics

Figure 2 shows the distributions of treatment complexity and corresponding average film gamma percentage pixels passing for two prominent metrics, MCS and MU. Much like the results of section 3.1, no significant relationships were evident, regardless of how the sample was broken down ($r_s < 0.206, p > 0.05$). Upon further inspection, this result is expected because the average absolute TLD error is correlated with the average gamma pass rate ($r_s = -0.464$, $p < 0.001$), meaning similar information is provided by both methods of plan error measurement.

3.3. The relationship between poor performing phantom irradiations and complexity metrics

Of the 343 phantom irradiations initially analyzed, 96 cases were identified as 'poor performers' based on a threshold of 5% TLD dose error for any given TLD within the phantom. Figure 3 depicts two distributions of treatment complexity and corresponding plan error. Like previous cases, no relationships were found to be significant, meaning that trends could not be distinguished in even the most concerning of irradiations.

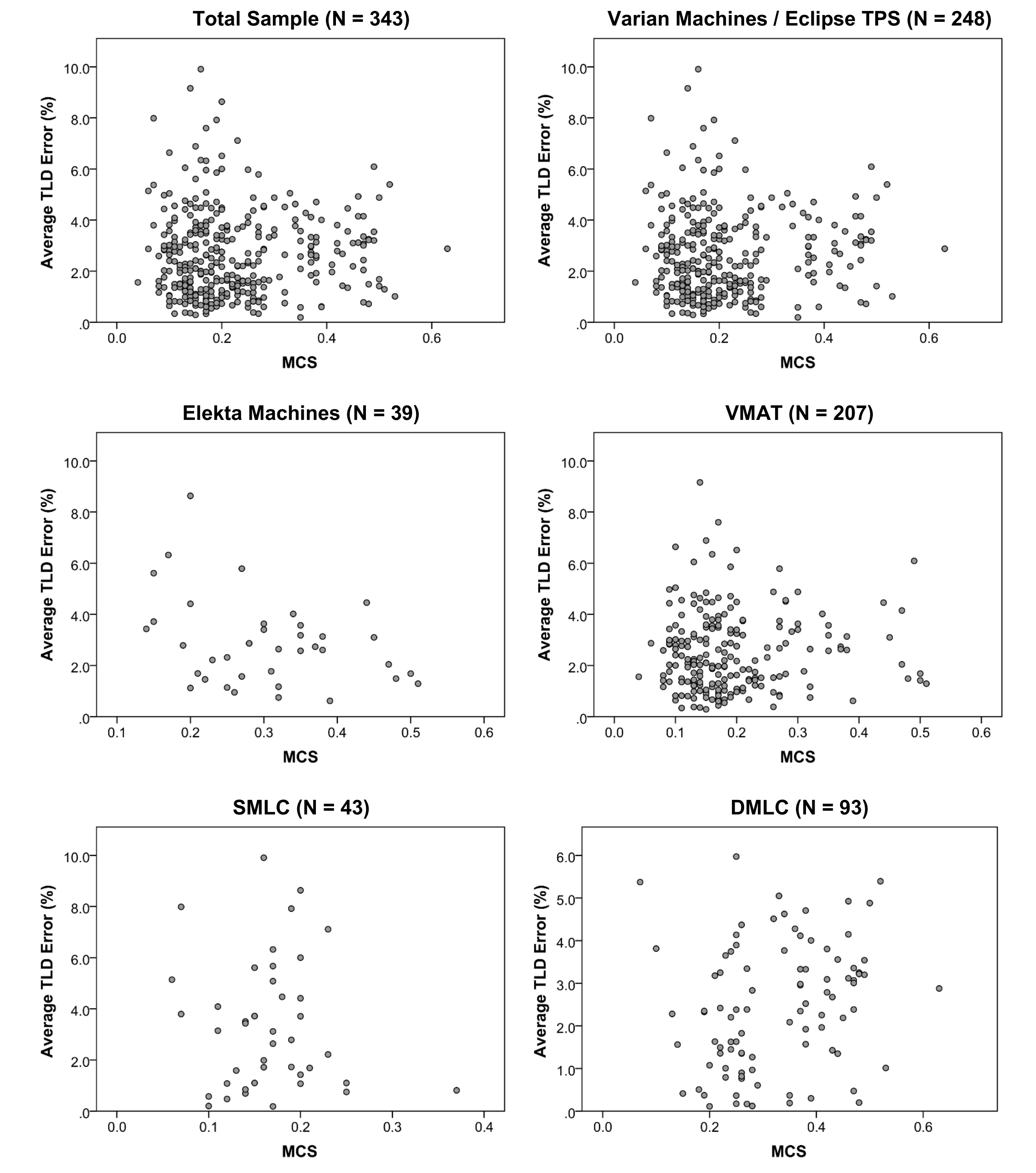

Figure 1. Scatter plots of average absolute TLD error versus complexity metrics for the whole sample and several subsamples in this work. The distributions showed no correlation between the measured TLD error and the assessed complexity metric, here shown as the MCS (smaller values of MCS correspond to more complex plans). SMLC = static MLC, DMLC = dynamic MLC.

4. Discussion

In this study, we examined several known measures of complexity, as well as additional plan metrics. These metrics generally have clear physical meanings that describe the beam aperture or fluence. The rationale for examining sixteen metrics was because others have suggested that a single measure may not be able to reveal all the details of the complexity in IMRT plans, nor may an individual metric be suitable for all TPSs (Du *et al* 2014, Hernandez *et al* 2018). By evaluating several metrics, we quantify different aspects of complexity and generate a much clearer, more comprehensive picture of a treatment plan and its potential challenges. Here the use of complexity metrics allowed for the potential identification of specific relationships that can influence plan performance in IROC's uniform phantom program.

Unfortunately, the results of this work show that there are no observable correlations between complexity metrics and the observed plan error in recent IROC H&N phantom performance. These results are interesting because they corroborate well with preliminary work that evaluated the utility of the MCS with a small cohort of H&N phantom plans on a single machine (Tonigan 2011). It was expected that certain metrics would not pro-

Table 2. Summary of Spearman correlations (r_s) comparing average absolute TLD error and complexity metric value for the subsamples described in this study (i.e. machine manufacturer, TPS, or delivery method).

		MU	MCS	EM	PI	PM	MI speed	MI accel.	MI total	LT/AL	Mean DR Var.	Mean GS Var.	Gap > 10 mm	Mean TG	MLC interdigitation	Mean MLC speed var.	Mean gap speed var.
All machines	r_s	−0.025	0.031	−0.119	−0.082	−0.018	−0.024	−0.028	−0.007	−0.006	0.025	−0.006	−0.018	−0.080	−0.176	−0.132	−0.134
$N = 343$	p	0.645	0.570	0.027	0.130	0.742	0.655	0.599	0.899	0.931	0.650	0.917	0.740	0.141	0.001*	0.015	0.013
Varian machines	r_s	−0.022	0.040	−0.107	−0.065	−0.011	−0.010	−0.011	0.008	0.002	0.039	0.008	0.015	−0.065	−0.174	−0.120	−0.124
$N = 304$	p	0.699	0.484	0.063	0.256	0.853	0.864	0.853	0.896	0.976	0.500	0.890	0.798	0.256	0.002*	0.036	0.030
Elekta machines	r_s	−0.016	−0.303	−0.150	−0.071	0.016	−0.241	−0.223	−0.283	−0.095	−0.237	−0.227	−0.044	−0.075	0.002	−0.130	−0.174
$N = 39$	p	0.924	0.061	0.363	0.668	0.922	0.140	0.173	0.081	0.636	0.147	0.165	0.791	0.649	0.991	0.429	0.289
Pinnacle TPS	r_s	0.208	−0.103	0.233	0.188	0.158	0.018	0.022	0.020	−0.173	0.002	0.049	0.200	0.171	0.089	0.140	0.159
$N = 69$	p	0.087	0.397	0.054	0.122	0.196	0.883	0.855	0.872	0.274	0.986	0.688	0.100	0.159	0.466	0.253	0.193
Eclipse TPS	r_s	0.025	0.017	−0.093	−0.058	0.055	−0.080	−0.066	−0.070	−0.059	−0.063	−0.075	0.029	−0.022	−0.110	−0.131	−0.104
$N = 249$	p	0.697	0.785	0.142	0.364	0.390	0.208	0.298	0.269	0.467	0.322	0.236	0.646	0.730	0.082	0.039	0.102
Varian + Eclipse	r_s	0.024	0.019	−0.094	−0.058	0.056	−0.078	−0.065	−0.070	−0.054	−0.065	−0.076	0.028	−0.022	−0.114	−0.131	−0.106
$N = 248$	p	0.708	0.768	0.141	0.362	0.382	0.221	0.306	0.274	0.509	0.309	0.236	0.659	0.725	0.073	0.039	0.096
Varian + Pinnacle	r_s	0.143	0.116	0.235	0.200	0.080	0.149	0.124	0.127	0.327	0.103	0.192	0.144	0.124	−0.042	0.231	0.242
$N = 45$	p	0.349	0.447	0.121	0.187	0.603	0.327	0.417	0.405	0.110	0.499	0.207	0.344	0.417	0.782	0.127	0.109
Elekta + Pinnacle	r_s	0.296	−0.446	0.131	0.120	0.225	−0.295	−0.282	−0.314	−0.395	−0.311	−0.187	0.244	0.195	0.203	−0.144	−0.111
$N = 24$	p	0.160	0.029	0.541	0.576	0.291	0.161	0.182	0.135	0.116	0.139	0.381	0.250	0.361	0.341	0.502	0.607
VMAT	r_s	0.005	0.021	−0.077	−0.008	0.040	0.100	0.089	0.148	−0.006	0.122	0.119	0.176	−0.008	−0.140	−0.159	−0.177
$N = 207$	p	0.945	0.765	0.272	0.903	0.571	0.151	0.200	0.033	0.931	0.079	0.086	0.011	0.904	0.044	0.022	0.011
DMLC	r_s	−0.218	0.240	−0.188	−0.079	−0.088							−0.052	−0.179	−0.220	0.065	0.160
$N = 93$	p	0.036	0.021	0.071	0.450	0.402							0.623	0.086	0.034	0.534	0.127
SMLC	r_s	0.212	0.057	−0.003	0.119	0.097							0.067	0.111	−0.031		
$N = 43$	p	0.172	0.719	0.987	0.446	0.535							0.670	0.480	0.846		

Note. DMLC = dynamic MLC, SMLC = static MLC, MU = monitor units, MCS = modulation complexity score, EM = edge metric, PI = plan irregularity, PM = plan modulation, MI = modulation index, LT/AL = leaf travel per arc length, DR = dose rate, GS = gantry speed, TG = tongue and groove.

* Correlation is significant at the 0.3% level (required for Bonferroni correction).

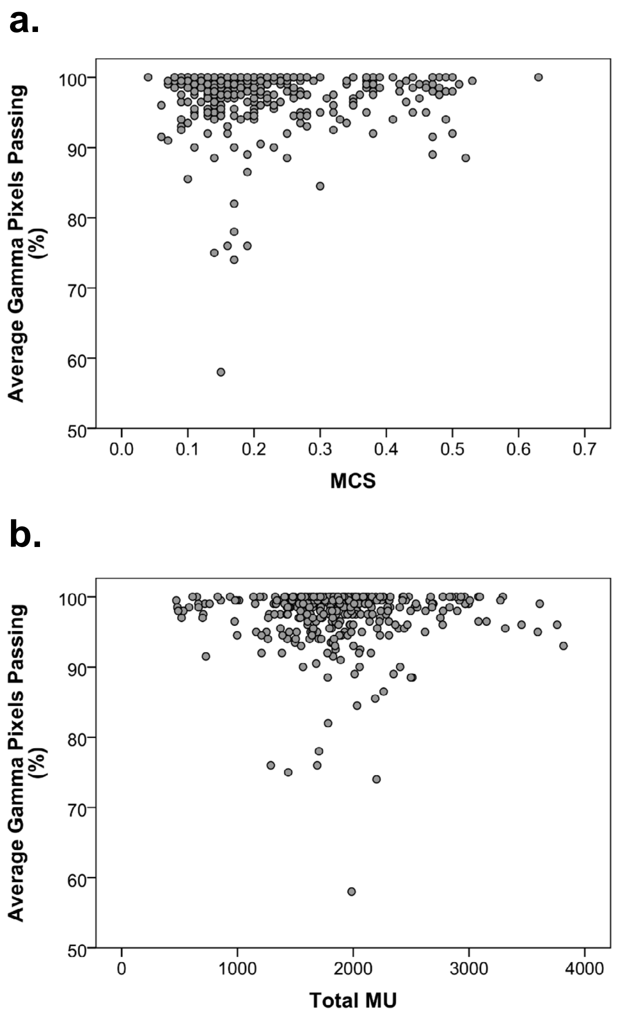

Figure 2. Scatter plots of average gamma pass rate versus complexity metrics for all irradiations examined in this work: (a) MCS, (b) total MU.

duce strong correlations based on typical clinical practices: for example, variations of the dose rate are generally well-controlled, and many IMRT plans do not have any dose rate variation. Other studies have also described how some metrics, such as the MCS, do not have a large effect on IMRT QA performance, which may then translate to a lack of relationship in our work (McNiven *et al* 2010, Rajasekaran *et al* 2015). Additionally, certain metrics provide similar information, as was determined by Hernandez *et al* in their comparisons of MCS, PI, and EM, meaning these indices should consequently produce similar results (Hernandez *et al* 2018). However, what is somewhat surprising from this work is that none of the sixteen complexity metrics even remotely produced viable relationships with the IROC H&N phantom results under the range of conditions evaluated. Our analyses show that complexity metrics were poor tools for predicting phantom performance and may have limited utility in determining the accuracy of treatment delivery.

Previous works examining IMRT complexity have shown mixed success in determining relationships with plan performance. Götstedt *et al* observed strong correlations between several metrics and gamma pass rates for both EBT3 film and portal imaging using a variety of MLC aperture shapes (Götstedt *et al* 2015). Likewise, both Masi *et al* and Agnew *et al* determined that the MCS correlates with gamma analyses for patient-specific QA (Masi *et al* 2013, Agnew *et al* 2014). Building on these ideas, Crowe *et al* suggested that there exist threshold complexity values that can defined to identify plans that are likely to fail QA (Crowe *et al* 2014). However, others, such as Du *et al* and McNiven *et al*, conceded that their proposed complexity indices did not yield correlations with plan quality metrics (including IMRT QA) but could still have utility in limiting the uncertainty in IMRT performance (McNiven *et al* 2010, Du *et al* 2014). Of particular interest, the work of McGarry *et al*, which observed QA phantom irradiations from multiple institutions, discovered weak but significant relationships between MU and plan quality, as well as MCS and plan quality, for all linear accelerators or Varian accelerators considered in their work (McGarry *et al* 2016). The work presented here clearly shows no indication of significant relationships between complexity and plan performance based on a relatively large sample of irradiations performed using multiple TPS, linear accelerator models, and delivery techniques.

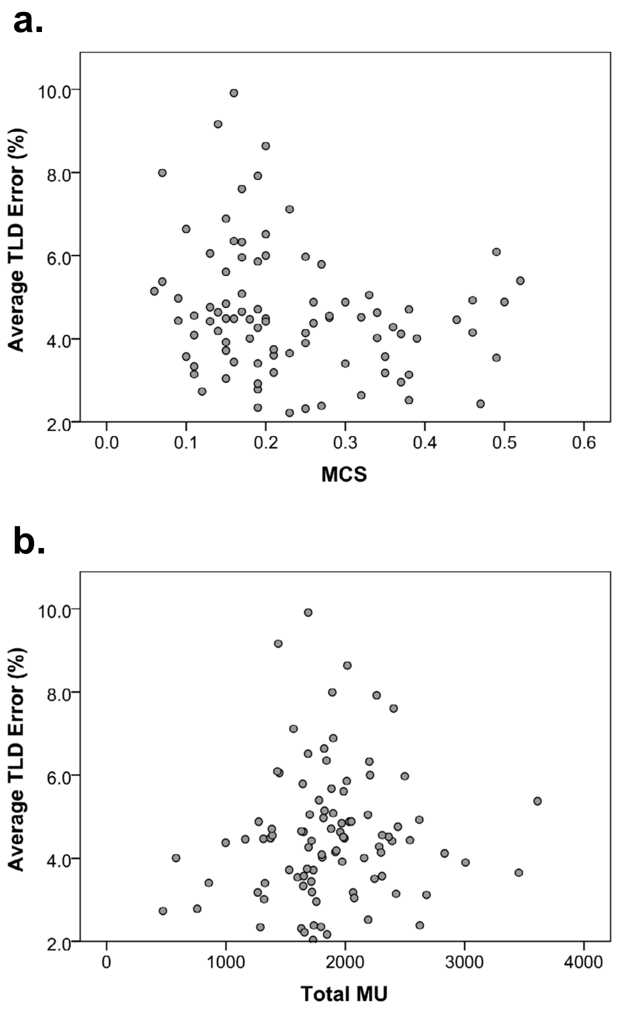

Figure 3. Scatter plots of average TLD error versus complexity metrics for poor performing phantom irradiations: (a) MCS, (b) total MU.

In previous works relating complexity and plan quality, patient-specific IMRT QA was typically used as a measure for plan performance accuracy. More recently, IMRT QA has come under scrutiny because of it is inability to discriminate unacceptable plans (Kruse 2010, Nelms *et al* 2011, Kry *et al* 2014, Stojadinovic *et al* 2015). Between different institutions, IMRT QA is completed using a plethora of devices, delivery techniques, and criteria for acceptability, which also limit the reproducibility and applicability of results derived from IMRT QA analysis. This work differs from previous studies of IMRT plan complexity in that it is among the first to examine complexity on a single patient geometry, the H&N phantom, using *in situ* dose measurement to characterize plan error for a multitude of institutions. Although both IMRT QA and IROC phantoms are designed to verify the accurate delivery of IMRT, this distinction is important because the phantom provides a direct comparison between the dose that was planned and that which was delivered, whereas IMRT QA measurements serve as a proxy for treatment accuracy. The H&N phantom is advantageous because all the plans observed had similar treatment objectives, thus eliminating the variability found between patient plans. This phantom also has a conceptual advantage over IMRT QA because its analysis was designed with consistent dose delivery in mind: all its irradiations are processed, analyzed, and evaluated in a consistent manner, and the uncertainties of this process are documented and well controlled (Molineu *et al* 2005). Such standardized treatment limits variability and allows for better understanding of overall performance trends in the radiotherapy community.

However, this approach is not without its own limitations. While the phantom test can control for many factors that other studies could not, such as patient geometry, this process also introduces other forms of variance, which can arise from the multitude of beam models used to calculate the dose distributions. Additionally, because the phantom is an end-to-end assessment of the treatment delivery process, it is possible that some of the plan errors observed here do not have a causal link with treatment delivery, but rather other external factors, such as phantom setup. Fortunately, based on IROC experience, incorrect setup does not contribute near as much to phantom errors as do systematic dosimetric inaccuracies (Carson *et al* 2016). There may also exist cases for which our methods are not sufficiently sensitive to characterize dose errors caused by excessively complexity plans, yet these would not be of clinical concern, as the measurement uncertainty for each double-loaded TLD is approximately 1.6% (Kirby *et al* 1992). Lastly, another factor that was not examined, but would pose a valid concern for patient treatment, is the potential effect of motion. Longer treatment times, as is common with high complexity treatments, may increase the sensitivity of dose accuracy to patient/target motion, but this could not be tested with a static phantom.

Though limiting the complexity of a plan may be good practice to limit some planning and delivery uncertainties, other factors may contribute to the degradation of plan accuracy. First and foremost of possibilities is the TPS calculation, which includes the beam modeling and inputs for beam characterization. The use of MLC-shaped beam segments, as is standard in IMRT, requires accurate modeling of several factors, including the leaf end, leaf transmission, and inter-leaf leakage (Ezzell *et al* 2009). If modeled improperly, the dose distributions delivered through MLC-defined apertures will have introduced error; systematic dosimetric errors have been documented for small fields (Followill *et al* 2012). Second, errors could be related to phantom or QA device positioning, which is user-dependent. Third, errors could be caused by inaccurate machine delivery characteristics, especially concerning the MLC positioning and dose rate accuracies. Because complexity measurement cannot encompass all potential failure modes, it is essential that these and other treatment delivery factors also be considered when assessing the potential for poor plan performance.

5. Conclusions

This study evaluated IMRT treatment plan complexity metrics with the purpose of identifying those which best predicted irradiation errors. Surprisingly, existing complexity metrics were universally not predictive of dosimetric errors in the IROC H&N phantom irradiations. That is, all metrics evaluated in this study failed to show a statistically significant relationship between phantom performance and the degree of complexity of the treatment plan, regardless of delivery technique, machine model, or TPS. This is interesting, because unlike previous experiments evaluating complexity metrics, the irradiated geometry is constant and without the heterogeneities or uncertainties found in real patient cases. These findings indicate that variations in beam complexity could not explain the disparities in phantom plan performance and that other factors affecting treatment delivery, such as beam modeling inaccuracies, dictate the accuracy of phantom treatment plans.

Acknowledgments

This work was supported by Public Health Service Grants CA180803 and CA214526 awarded by the National Cancer Institute, United States Department of Health and Human Services. Mallory C Glenn is a recipient of the Rosalie B Hite Fellowship in Cancer Research, awarded by the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.

Disclosure of conflicts of interest

The authors have no conflicts of interest to disclose.

ORCID iDs

Mallory C Glenn https://orcid.org/0000-0001-6780-4458
Victor Hernandez https://orcid.org/0000-0003-3770-8486
David S Followill https://orcid.org/0000-0001-6744-0439
Julianne M Pollard-Larkin https://orcid.org/0000-0002-1196-656X
Stephen F Kry https://orcid.org/0000-0001-6899-197X

References

Agnew C E, Irvine D M and McGarry C K 2014 Correlation of phantom-based and log file patient-specific QA with complexity scores for VMAT *J. Appl. Clin. Med. Phys.* **15** 204–16

Carson M E, Molineu A, Taylor P A, Followill D S, Stingo F C and Kry S F 2016 Examining credentialing criteria and poor performance indicators for IROC Houston's anthropomorphic head and neck phantom *Med. Phys.* **43** 6491–6

Crowe S B *et al* 2014 Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results *Australas. Phys. Eng. Sci. Med.* **37** 475–82

Crowe S B *et al* 2015 Examination of the properties of IMRT and VMAT beams and evaluation against pre-treatment quality assurance results *Phys. Med. Biol.* **60** 2587–601

Du W, Cho S H, Zhang X, Hoffman K E and Kudchadker R J 2014 Quantification of beam complexity in intensity-modulated radiation therapy treatment plans *Med. Phys.* **41** 21716

Ezzell G A *et al* 2009 IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119 *Med. Phys.* **36** 5359–73

Followill D S *et al* 2012 The Radiological Physics Center's standard dataset for small field size output factors *J. Appl. Clin. Med. Phys.* **13** 282–9

Götstedt J, Hauer A K and Bäck A 2015 Development and evaluation of aperture-based complexity metrics using film and EPID measurements of static MLC openings *Med. Phys.* **42** 3911–21

Hernandez V, Saez J, Pasler M, Jurado-Bruggeman D and Jornet N 2018 Comparison of complexity metrics for multi-institutional evaluations of treatment plans in radiotherapy *Phys. Imaging Radiat. Oncol.* **5** 37–43

Kerns J R, Stingo F, Followill D, Howell R, Melancon A and Kry S F 2017 Treatment planning system calculation errors are present in the majority of IROC-Houston phantom failures *Int. J. Radiat. Oncol. Biol. Phys.* **98** 1197–203

Kirby T H, Hanson W F and Johnston D A 1992 Uncertainty analysis of absorbed dose calculations from thermoluminescence dosimeters *Med. Phys.* **19** 1427–33

Kruse J J 2010 On the insensitivity of single field planar dosimetry to IMRT inaccuracies *Med. Phys.* **37** 2516–24

Kry S F *et al* 2014 Institutional patient-specific IMRT QA does not predict unacceptable plan delivery *Int. J. Radiat. Oncol. Biol. Phys.* **90** 1195–201

Masi L, Doro R, Favuzza V, Cipressi S and Livi L 2013 Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy *Med. Phys.* **40** 71718

McGarry C K *et al* 2016 The role of complexity metrics in a multi-institutional dosimetry audit of VMAT *Br. J. Radiol.* **89** 20150445

McNiven A L, Sharpe M B and Purdie T G 2010 A new metric for assessing IMRT modulation complexity and plan deliverability *Med. Phys.* **37** 505–15

Mohan R, Arnfield M, Tong S, Wu Q and Siebers J 2000 The impact of fluctuations in intensity patterns on the number of monitor units and the quality and accuracy of intensity modulated radiotherapy *Med. Phys.* **27** 1226–37

Molineu A *et al* 2005 Design and implementation of an anthropomorphic quality assurance phantom for intensity-modulated radiation therapy for the Radiation Therapy Oncology Group *Int. J. Radiat. Oncol. Biol. Phys.* **63** 577–83

Molineu A, Hernandez N, Nguyen T, Ibbott G and Followill D 2013 Credentialing results from IMRT irradiations of an anthropomorphic head and neck phantom *Med. Phys.* **40** 22101

Nelms B E, Zhen H and Tomé W A 2011 Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors *Med. Phys.* **38** 1037–44

Park J M, Park S-Y, Kim H, Kim J H, Carlson J and Ye S-J 2014 Modulation indices for volumetric modulated arc therapy *Phys. Med. Biol.* **59** 7315–40

Rajasekaran D, Jeevanandam P, Sukumar P, Ranganathan A, Johnjothi S and Nagarajan V 2015 A study on the correlation between plan complexity and gamma index analysis in patient specific quality assurance of volumetric modulated arc therapy *Rep. Pract. Oncol. Radiother.* **20** 57–65

Stojadinovic S, Ouyang L, Gu X, Pompoš A, Bao Q and Solberg T D 2015 Breaking bad IMRT QA practice *J. Appl. Clin. Med. Phys.* **16** 154–65

Tonigan J R 2011 Evaluation of intensity modulated radiation therapy (IMRT) delivery error due to IMRT treatment plan complexity and improperly matched dosimetry data *MSc Thesis* The University of Texas Graduate School of Biomedical Sciences at Houston

Webb S 2003 Use of a quantitative index of beam modulation to characterize dose conformality: illustration by a comparison of full beamlet IMRT, few-segment IMRT (fsIMRT) and conformal unmodulated radiotherapy *Phys. Med. Biol.* **48** 2051–62

Younge K C, Matuszak M M, Moran J M, McShan D L, Fraass B A and Roberts D A 2012 Penalization of aperture complexity in inversely planned volumetric modulated arc therapy *Med. Phys.* **39** 7160–70