IMRT commissioning: Multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119

Gary A. Ezzell
Department of Radiation Oncology, Mayo Clinic Scottsdale, 5777 East Mayo Boulevard, MCSB Concourse, Phoenix, Arizona 89054

Jay W. Burmeister
Wayne State University School of Medicine, Karmanos Cancer Center, 4100 John R Street, Detroit, Michigan 48201

Nesrin Dogan
Department of Radiation Oncology, Virginia Commonwealth University, 401 College Street B-129, Richmond, Virginia 23298

Thomas J. LoSasso and James G. Mechalakos
Department of Medical Physics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10065

Dimitris Mihailidis
Department of Radiation Oncology and Medical Physics, Charleston Radiation Therapy Cons, 3100 MacCorkle Avenue Southeast, Charleston, West Virginia 25304

Andrea Molineu
RPC, UT MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030

Jatinder R. Palta
Department of Radiation Oncology, University of Florida Health Science Center, 2000 Archer Road, Gainesville, Florida 32610–0385

Chester R. Ramsey
Thompson Cancer Survival Center, 1915 White Avenue, Knoxville, Tennessee 37916

Bill J. Salter
Department of Radiation Oncology, University of Utah, 1950 Circle of Hope Drive, Salt Lake City, Utah 84112

Jie Shi
Sun Nuclear Corp., 425–A Pineda Court, Melbourne, Florida 32940

Ping Xia
Department of Radiation Oncology, University of California, San Francisco, California 94143–1708

Ning J. Yue
Department of Radiation Oncology, The Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, New Jersey 08901

Ying Xiao[a)]
Department of Radiation Oncology, Thomas Jefferson University Hospital, 111 South 11th Street, Philadelphia, Pennsylvania 19107

(Received 15 May 2009; revised 4 September 2009; accepted for publication 5 September 2009; published 27 October 2009)

AAPM Task Group 119 has produced quantitative confidence limits as baseline expectation values for IMRT commissioning. A set of test cases was developed to assess the overall accuracy of planning and delivery of IMRT treatments. Each test uses contours of targets and avoidance structures drawn within rectangular phantoms. These tests were planned, delivered, measured, and analyzed by nine facilities using a variety of IMRT planning and delivery systems. Each facility had passed the Radiological Physics Center credentialing tests for IMRT. The agreement between the planned and measured doses was determined using ion chamber dosimetry in high and low dose regions, film dosimetry on coronal planes in the phantom with all fields delivered, and planar dosimetry for each field measured perpendicular to the central axis. The planar dose distributions were assessed using gamma criteria of 3%/3 mm. The mean values and standard deviations were used to develop confidence limits for the test results using the concept confidence limit$=|\text{mean}|+1.96\sigma$. Other facilities can use the test protocol and results as a basis for comparison to this group.

Locally derived confidence limits that substantially exceed these baseline values may indicate the need for improved IMRT commissioning.
[DOI: 10.1118/1.3238104]

Key words: IMRT, commissioning, quality assurance

TABLE OF CONTENTS

I. INTRODUCTION

The 2003 "Guidance Document" on IMRT[1] noted that "This complex but promising treatment modality is rapidly proliferating in both academic and community practice settings." The intervening years have seen the use of IMRT become commonplace. It is reported that approximately 30%–60% of cancer patients in the United States are currently being treated with IMRT.[2] However, there is evidence that IMRT treatments may not always be as accurate as practitioners believe. In 2008, the Radiological Physics Center (RPC) reported that of the 250 irradiations of a head and neck phantom as part of an IMRT credentialing process, 71 (28%) had failed to meet accuracy criteria of 7% for dose in a low gradient region and/or 4 mm distance to agreement in a high gradient.[3] This is a sobering statistic, especially considering that this is a sample of those institutions that felt confident enough in their IMRT planning and delivery process to apply for credentialing and presumably expected to pass.

This experience strongly suggests that some clinics have not adequately commissioned their planning and delivery systems for IMRT. By "commissioning," we mean the initial verification by phantom studies that treatments can be planned, prepared, and delivered with sufficient accuracy. Commissioning is different from per-patient phantom measurements for quality assurance purposes. In the latter case, the doses in the phantom are not the same as the doses predicted for the patient, and so are not complete tests of the total planning and delivery chain. Commissioning studies are best done by defining target and normal structure shapes on CT images of the dosimetry phantom, planning the treatment, and then comparing the measured dose in the phantom to the planned dose from the computer system. Commissioning studies should mimic the types of target and structure geometries along with the target doses and dose constraints that are likely to be encountered in the clinic. Commissioning studies should also be performed with particular care to minimize measurement uncertainties, which should be quantified. Differences between calculations and measurements can only be meaningfully evaluated if the uncertainties are understood.

The commissioning process was discussed in general terms in the 2003 Guidance Document.[1] Task Group 119 of the American Association of Physicists in Medicine (AAPM) was charged with expanding that guidance document. In this work, TG119 has focused on the problem of quantifying the overall performance of an IMRT system and determining reasonable confidence limits (CLs) for assessing the adequacy of the dosimetric commissioning. This report does not deal with many other important aspects of IMRT quality assurance, such as additional periodic QA of multileaf collimators, which are left for future work. The report from Task Group 142 (Working Group on Recommendations for Radiotherapy External Beam Quality Assurance),[23] will address some of these issues. The report from Task Group 120 (Writing Group on IMRT Metrology), also in preparation, will address specific issues related to measurement tools and analysis methods for IMRT.

The task group first developed a specific set of tests for IMRT commissioning that are representative of common clinical treatments. While not exhaustive, these tests pose a range of optimization problems requiring simple to complex modulation patterns. These represent total system checks of different types and levels of complexity. Differences between measurement and prediction may be caused by measurement

TABLE I. List of participating institutions and the systems utilized. Manufacturer's identifications are listed below the table. "DMLC" refers to dynamic MLC, sometimes called "sliding window." "SMLC" refers to static MLC, sometimes called "step and shoot" (Varian, ECLIPSE: Varian Medical Systems, Milpitas, CA; Siemens: Siemens AG, Healthcare Sector, Erlangen, Germany; Elekta, CMS: Elekta Inc., Norcross, GA; PINNACLE: Philips Healthcare, Andover, MA; TOMOTHERAPY: TomoTherapy Inc., Madison, WI).

Institution	Accelerator	Delivery technique	Planning system
Mayo Clinic Arizona	Varian 21EX	DMLC	ECLIPSE V7.5
Thomas Jefferson University Hospital	Elekta Synergy S	SMLC	CMS XIO V3.1
Robert Wood Johnson University Hospital	Varian 21EX	DMLC	ECLIPSE V7.5
Memorial Sloan Kettering Cancer Center	Varian Trilogy	DMLC	In-house
Karmanos Cancer Center/Wayne State University	Varian 23EX	DMLC	ECLIPSE V7.5
Karmanos Cancer Center/Wayne State University	Tomotherapy Hi-Art	BinaryMLC	TOMOTHERAPY V3.0
University of California at San Francisco	Siemens Oncor C	SMLC	PINNACLE V8.0d
University of Florida	Elekta Synergy	SMLC	PINNACLE V8.0d
Virginia Commonwealth University	Varian Trilogy	DMLC	PINNACLE V8.0d
Charleston Radiation Therapy Consultants	Siemens Primus	SMLC	PINNACLE V7.4f

uncertainty, limitations in the accuracy of dose calculations, and limitations in the dose delivery mechanisms. These tests do not serve to distinguish between these sources but test the overall accuracy of the IMRT system.

Each test includes target and normal structure shapes that a physicist can create on a simple slab phantom. Each test includes a specification of dose goals for the IMRT planning and the beam arrangement to be used. Each test also specifies the measurements to be taken to test the accuracy of the dose delivery and what is to be reported.

Members of the group have planned and delivered the treatments using the local planning and delivery systems and then assessed the resulting doses using broadly available dosimetry tools. The goal was to produce quantitative examples of the degree of agreement that should be expected for such tests, and thus provide the medical physics community with a useful set of benchmark data. Institutions that do similar tests and achieve similar results could then have more confidence that their system's performance is clinically acceptable, at least for the types of treatments modeled by the commissioning tests. Conversely, and we hope, helpfully, institutions with worse results can use these tests to refine their planning and delivery systems.

This study has quantified the "degree of agreement that should be expected" using the concept of "confidence limit" as proposed by Venselaar *et al.*[4] and refined by Palta *et al.*[5] Whenever a measurement is made and compared to a calculation, one can expect some difference to be seen. If the difference is within a reasonable confidence limit, then the result can be considered acceptable. This task group has established confidence limits for different types of measurements by combining data from the participating institutions. Each of the institutions that have participated in this study has passed the RPC IMRT credentialing test using the RPC's head and neck dosimetry phantom.

The confidence limit is based on the average difference between measured and expected values for a number of measurements of comparable situations (systematic difference) summed with the standard deviation of the differences multiplied by some factor (random difference). In the formulation of Palta *et al.*, the CL is the sum of the absolute value of the average difference and the standard deviation of the differences multiplied by a factor of 1.96 [CL $= |\text{mean deviation}| + 1.96$ SD (Palta *et al.* used the symbol Δ for CL)]. In this formulation that is based on the statistics of a normal distribution, it is to be expected that 95% of the measured points will fall within the confidence limit. In this TG-119 study, the set of measurements for the group has been combined and analyzed in this fashion to provide a confidence limit for IMRT commissioning measurements. In order to use these benchmark data, a facility would perform a similar set of measurements, determine the local systematic and random variation from the expected values, calculate the local confidence limit using the same formulation, and see if it is similar to that from this task group. Note that the confidence limit will likely be dominated by the standard deviation term with its multiplier of nearly 2. However, should the facility find a tight distribution around a large mean difference, then the reason for that difference can very likely be found and the result improved.

II. METHODS AND MATERIALS

Table I lists the institutions participating in the study, along with the planning system and the delivery system used by each.

II.A. Phantoms

Institutions were instructed to choose a phantom in which to do the planning and measurements, following these specifications. The phantom should permit point measurements (e.g., ion chamber) and planar dose measurements (e.g., film) to be done on coronal planes. The phantom should consist of slabs of water-equivalent plastic, typically squares or rectangles 20–30 cm on a side, with a total thickness of about 15–20 cm, so that a chamber at its center is 7.5–10 cm below the anterior surface. (Note that the phantom shown in Figs. 1–5 has a different type of "water-equivalent" plastic used for the central section that is apparent because of the narrow CT imaging window used when the images were captured.)

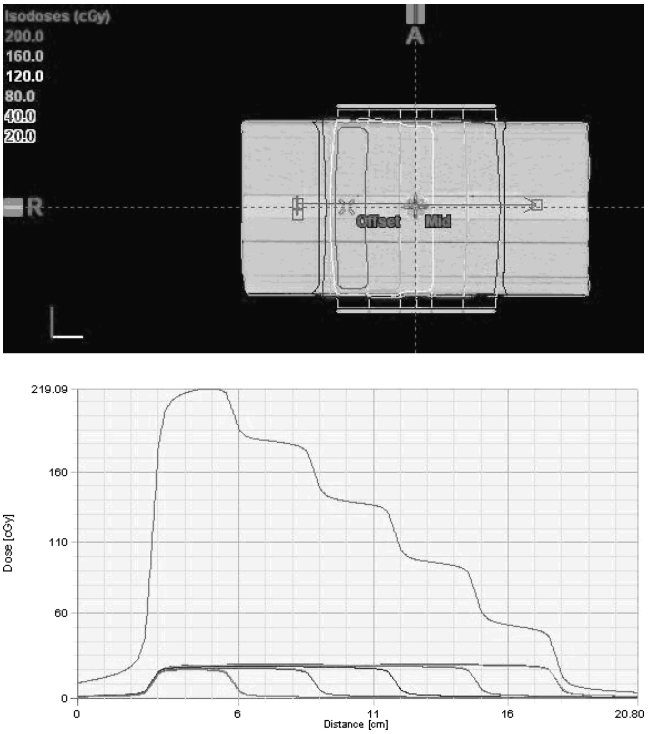


FIG. 1. Dose profile through central plane for bands. The lower curves are the individual contributions from each subfield (band); the upper curve is the summation.

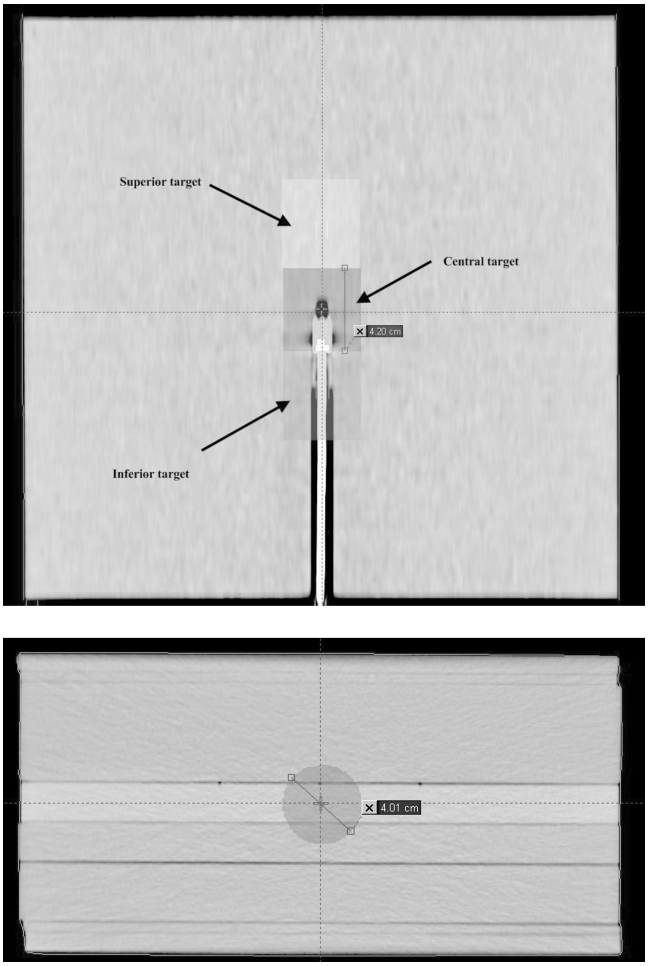


FIG. 2. Multitarget structures: Central target, superior target, and inferior target. These three cylindrical targets are stacked along the axis of rotation. Each has a diameter of approximately 4 cm and length of 4 cm. Coronal and transverse views are shown.

It should be possible to have either film or chamber on the central measurement plane, so that the film response can be normalized to the chamber. Each institution scanned its own phantom for planning and measurements. The plans were done either on that phantom with the structures outlined on it or on a downloaded CT study and then transferred to the local phantom for measurement, in a manner similar to performing patient quality assurance measurements.

II.B. Chamber measurements

Institutions were instructed to choose an ionization chamber suitable for IMRT commissioning and QA studies in the department. This typically would be smaller than a Farmer-type chamber, such as a 0.125 cm^3 scanning chamber. The chamber measurements were to be made with all fields irradiating the phantom using the planned gantry and collimator angles. For most of the tests, measurements were to be made in at least two locations, one in the target and one in a low dose avoidance structure. The doses were expected to be at least 30 cGy, so issues with very low dose measurements would not arise.

Conversion of chamber reading to dose was to be done by first irradiating the phantom with parallel-opposed 10×10 fields arranged isocentrically and establishing the ratio of reading to planned dose in that geometry. This was done in order to reduce the effects of daily linac output variations and differences between the phantom and liquid water. The institution with the tomotherapy device measured absolute doses for each delivered plan using $N_{D,w}^{\text{Co-60}}$ and k_Q values for chambers calibrated at an accredited dosimetry calibration laboratory.

II.C. Composite film measurements

Each test called for a film to be placed in at least one coronal plane and to be exposed to all fields irradiating the phantom with the planned gantry and collimator angles. Institutions were expected to use their most accurate protocols for film dosimetry. Dose distributions were analyzed using gamma criteria[6] of 3% dose and 3 mm distance to agreement. The planar dose distributions obtained with the film could be normalized to the dose measured with the chamber at a suitable point in a high dose, low gradient region. The film analysis was done with the software tools available at each institution. The gamma analysis was to be restricted to regions to avoid those of very low dose; this was done in one of two ways. If the software defined the region of interest using a threshold dose, then that was set to 10% of the maximum dose. If the software required a rectangular region of interest to be defined, then that was taken to be the jaw

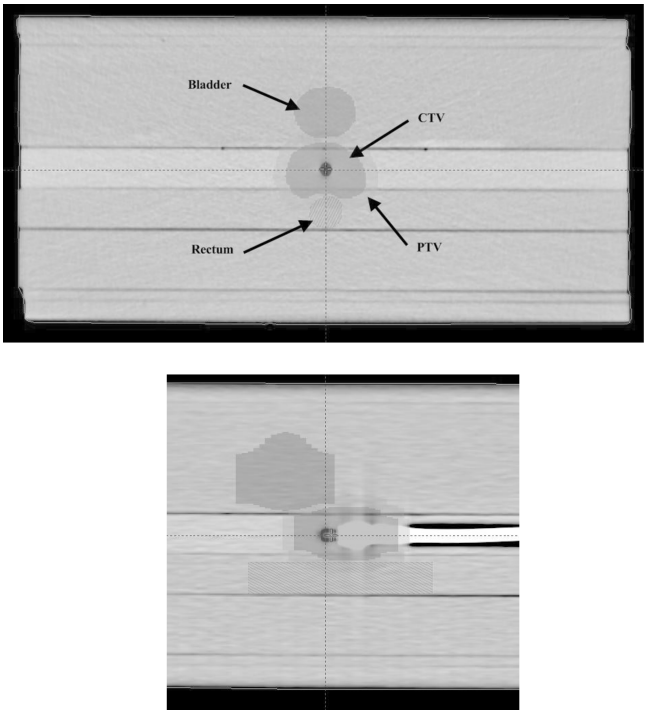


FIG. 3. Mock prostate structures: The prostate CTV, PTV, rectum, and bladder. The prostate CTV is roughly ellipsoidal with RL, AP, and SI dimensions of 4.0, 2.6, and 6.5 cm, respectively. The prostate PTV is expanded 0.6 cm around the CTV. The rectum is a cylinder with diameter of 1.5 cm that abuts the indented posterior aspect of the prostate. The PTV includes about 1/3 of the rectal volume on the widest PTV slice. The bladder is roughly ellipsoidal with RL, AP, and SI dimensions of 5.0, 4.0, and 5.0 cm, respectively, and is centered on the superior aspect of the prostate. Transverse and coronal views are shown.

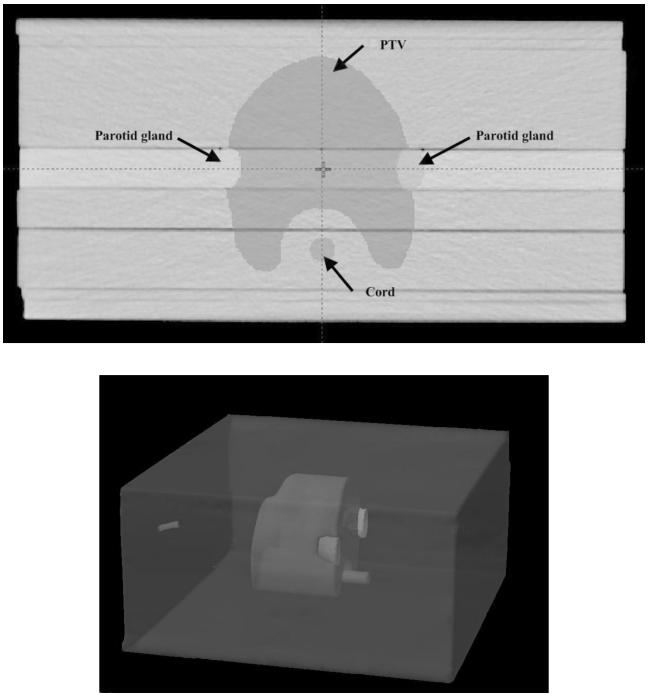


FIG. 4. Mock head/neck structures: HN PTV, cord, and parotid glands. The PTV is retracted from the skin by 0.6 cm. There is a gap of about 1.5 cm between the cord and the PTV. The parotid glands are to be avoided and are at the superior aspect of the PTV. Transverse and 3D views are shown.

settings for the field at gantry 0° or 180°. This restriction was done because the percentage of points that pass the gamma criteria can depend on the region chosen and the details of how low dose points are handled in the algorithm implemented in the particular software used.

II.D. Per-field measurements

Each institution was asked to evaluate the dose distribution produced by each field individually using the dosimetry system available, which was either film, detector array, or EPID. Gamma criteria of 3% dose and 3 mm distance to agreement were used and the region of interest was specified as above: Either 10% dose threshold or a region of interest determined by the jaw settings.

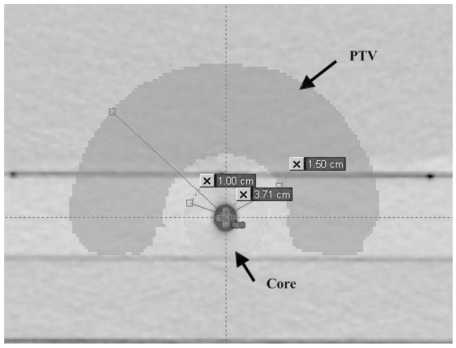
Five of the institutions performed these measurements using the MAPCHECK diode array device (Sun Nuclear Corporation, Melbourne, FL).[7] They agreed on a common set of user preferences in order to standardize the analysis to the extent possible. These choices (with brief explanation) were absolute dose (measured doses were not scaled to some normalization value), 10% threshold (the region of interest was defined by the isodose line representing 10% of maximum dose),[11] Van Dyk percentage difference[12] (the percentage difference in dose was with respect to the maximum point in the region, not the local point), and applied measurement uncertainty (a presumed measurement error of about 1% is included in the analysis, so that a nominal 3% dose difference can be 4%). The plan and measurement data from these institutions were sent to one location for analysis using version MAPCHECK 3.04. This selection does not imply endorsement of either this particular device or this particular set of parameter options for use in clinical evaluations.

II.E. Test suite: Planning conditions and measurement specifications

Two preliminary tests with simple fields irradiating the phantom were requested to demonstrate the reliability of the assessment system for non-IMRT dose delivery, followed by five tests of IMRT plans with increasing complexity. The dose goals for the IMRT plans were expressed in total doses with the daily dose to be 180–200 cGy. The volumes for the IMRT plans could be either drawn *de novo* by the institution or downloaded as DICOM-RT data from a central server and transferred to the scans of the institution's phantom. These tests were all performed at 6 MV, which was an energy available to all the participating institutions.

II.E.1. Test P1: AP:PA

- Calculate a simple parallel-opposed irradiation of the phantom using AP:PA 10×10 fields to a dose of 200 cGy to the isocenter, placed at the phantom midline.
- Measure the central dose with the chamber and the dose distribution on the central plane with film.



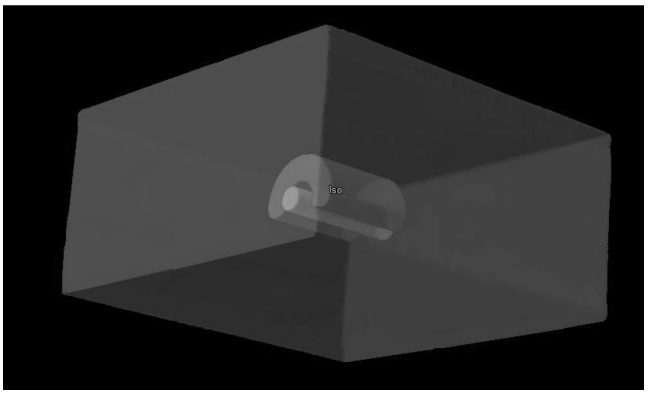


FIG. 5. CShape structures: CShape PTV and core. The center core is a cylinder 1 cm in radius. The gap between the core and the PTV is 0.5 cm, so the inner arc of the PTV is 1.5 cm in radius. The outer arc of the PTV is 3.7 cm in radius. The PTV is 8 cm long and the core is 10 cm long. Transverse and 3D views are shown.

- Use this chamber measurement to set the dose/chamber reading ratio for subsequent tests.
- Analyze the film dosimetry and report the fraction of points passing the gamma criteria.

II.E.2. Test P2: Bands

- Calculate a parallel-opposed irradiation of the phantom using a series of AP:PA fields to create a set of five bands, 3 cm wide, receiving doses from roughly 40 to 200 cGy (Fig. 1). This could be done using asymmetric jaws or static MLC fields.
- Measure the central dose with the chamber and the dose distribution on the central plane with the film. Analyze the film dosimetry and report the fraction of points passing the gamma criteria.

II.E.3. Test I1: Multitarget

II.E.3.a. Structures. Three cylindrical targets are stacked along the axis of rotation. Each has a diameter of approximately 4 cm and length of 4 cm (Fig. 2). They are to receive different doses, with the central target to receive the largest dose per fraction. The superior target was to receive 50% of that and the inferior 25%.

II.E.3.b. Dose goals used for planning. The dose goals used for planning were expressed in terms of dose to 99% of the volume (D99) and dose to 10% of the volume (D10) for the three targets. Table II includes the specific numerical goals.

TABLE II. Treatment plan statistics for multitarget.

Planning parameter	Plan goal (cGy)	Mean (cGy)	Standard deviation (cGy)	Coefficient of variation
Central target D99	>5000	4955	162	0.033
Central target D10	<5300	5455	173	0.032
Superior target D99	>2500	2516	85	0.034
Superior target D10	<3500	3412	304	0.089
Inferior target D99	>1250	1407	185	0.132
Inferior target D10	<2500	2418	272	0.112

II.E.3.c. Beam arrangement.

- 6 MV, 7 fields at 50° intervals from the vertical (e.g., 0°, 50°, 100°, 150°, 310°, 260°, 210°).

II.E.3.d. Chamber measurement points.

- Isocenter, middle of the central target.
- Center of the other two targets.

II.E.3.e. Film measurement.

- Midphantom.

II.E.4. Test I2: Mock prostate

II.E.4.a. Structures. The prostate CTV is roughly ellipsoidal, with posterior concavity, with RL, AP, and SI dimensions of 4.0, 2.6, and 6.5 cm, respectively. The prostate PTV is expanded 0.6 cm around the CTV.

The rectum is a cylinder with diameter of 1.5 cm that abuts the indented posterior aspect of the prostate. The PTV includes about 1/3 of the rectal volume on the widest PTV slice. The bladder is roughly ellipsoidal with RL, AP, and SI dimensions of 5.0, 4.0, and 5.0 cm, respectively, and is centered on the superior aspect of the prostate (Fig. 3).

II.E.4.b. Dose goals used for planning. For the prostate PTV, dose goals were specified as D95 and D5. For rectum and bladder, D30 and D10 were used. Table III includes the specific numerical goals.

II.E.4.c. Beam arrangement.

- 6 MV, seven fields at 50° intervals from the vertical.

II.E.4.d. Chamber measurement points.

- Isocenter, in the mid-PTV.

TABLE III. Treatment plan statistics for mock prostate.

Planning parameter	Plan goal (cGy)	Mean (cGy)	Standard deviation (cGy)	Coefficient of variation
Prostate D95	>7560	7566	21	0.003
Prostate D5	<8300	8143	156	0.019
Rectum D30	<7000	6536	297	0.045
Rectum D10	<7500	7303	150	0.020
Bladder D30	<7000	4394	878	0.200
Bladder D10	<7500	6269	815	0.130

TABLE IV. Treatment plan statistics for mock head and neck.

Planning parameter	Plan goal (cGy)	Mean (cGy)	Standard deviation (cGy)	Coefficient of variation
PTV D90	5000	5028	58	0.013
PTV D99	>4650	4704	52	0.011
PTV D20	<5500	5299	93	0.018
Cord maximum	<4000	3741	250	0.067
Parotid D50	<2000	1798	184	0.102

- 2.5 cm posterior, midrectum.

II.E.4.e. Film measurement.

- Midphantom.

II.E.5. Test I3: Mock head/neck

II.E.5.a. Structures. The volumes for the head/neck (HN) case were first drawn on a scan of an anthropomorphic phantom and then transferred to the rectangular phantom. The HN PTV includes all anterior volume from the base of the skull to the upper neck, including the posterior neck nodes. The PTV is retracted from the skin by 0.6 cm. There is a gap of about 1.5 cm between the cord and the PTV. The parotid glands are to be avoided and are at the superior aspect of the PTV (Fig. 4).

II.E.5.b. Dose goals used for planning. For the head and neck PTV, dose goals were specified as D99, D90, and D20. For normal structures, D50 was used for parotid and maximum dose was used for cord. Table IV includes the specific numerical goals.

II.E.5.c. Beam arrangement.

- 6 MV, 9 fields at 40° intervals from the vertical.

II.E.5.d. Chamber measurement points.

- Isocenter, in the mid-PTV.
- 4.0 cm posterior, midspinal cord.

II.E.5.e. Film measurements.

- Midphantom, includes parotids.
- 4.0 cm posterior, through cord.

II.E.6. Tests I4 and I5: Cshape

II.E.6.a. Structures. The target is a CShape that surrounds a central avoidance structure. The center core is a cylinder 1 cm in radius. The gap between the core and the PTV is 0.5 cm, so the inner arc of the PTV is 1.5 cm in radius. The outer arc of the PTV is 3.7 cm in radius. The PTV is 8 cm long and the core is 10 cm long (Fig. 5).

Two versions of the problem are given. In the easier, the central core is to be kept to 50% of the target dose. In the harder, the central core is to be kept to 20% of the target dose. This latter goal is probably not achievable and tests a system that is being pushed very hard.

II.E.6.b. Dose goals for planning (easier version and harder version). For the CShape PTV, dose goals were specified as D95 and D10. For the core normal structure, D10 was used. Table V includes the specific numerical goals for the easier version and Table VI includes those for the harder version.

TABLE V. Treatment plan statistics for CShape (easier).

Planning parameter	Plan goal (cGy)	Mean (cGy)	Standard deviation (cGy)	Coefficient of variation
PTV D95	5000	5010	17	0.003
PTV D10	<5500	5440	52	0.010
Core D10	<2500	2200	314	0.141

II.E.6.c. Beam arrangement.

- 6 MV, 9 fields at 40° intervals from the vertical.

II.E.6.d. Chamber measurement points.

- Central core.
- Mid-PTV, 2.5 cm anterior to isocenter.

II.E.6.e. Film measurements.

- Midphantom.
- Mid-PTV, 2.5 cm anterior to isocenter.

III. RESULTS

III.A. Planning results

The statistics for the plans from institutions for test I1 (multitarget), test I2 (mock prostate), test I3 (mock head and neck), test I4 (CShape easier), and test I5 (CShape harder) are listed in Tables II–VI, respectively. In these tables, the notation "D99" means the dose covering 99% of the volume.

The planning instructions did not specify a minimum calculation grid size. Participants reported using grid intervals ranging from 0.1 to 0.4 cm.

III.B. Measurement results

III.B.1. Results for preliminary test P2: Bands

Six of the institutions reported ion chamber results for the band test. These ranged from 1.3% more than predicted to 0.9% less with a mean of 0.3% more. Four of the institutions reported gamma results from film for the bands test with gamma pass rates ranging from 98.3% to 99.4% with a mean of 99.1%.

III.B.2. Ion chamber results

The results of the ion chamber measurements are shown in Tables VII–X. (In subsequent tables, the facilities are identified by letter only, not corresponding to the order by which

TABLE VI. Treatment plan statistics for CShape (harder).

Planning Parameter	Plan goal (cGy)	Mean (cGy)	Standard deviation (cGy)	Coefficient of variation
PTV D95	5000	5011	16.5	0.003
PTV D10	<5500	5702	220	0.039
Core D10	<1000	1630	307	0.188

TABLE VII. High dose point in the PTV measured with ion chamber: [(measured dose) −(plan dose)]/prescription dose, averaged over the institutions, with associated confidence limits.

Test	Location	Mean	Standard deviation (σ)	Maximum	Minimum
Multitarget	Isocenter	0.001	0.017	0.030	−0.020
Prostate	Isocenter	−0.001	0.016	0.022	−0.026
Head and neck	Isocenter	−0.010	0.013	0.011	−0.036
CShape (easier)	2.5 cm anterior to isocenter	−0.001	0.028	0.038	−0.059
CShape (harder)	2.5 cm anterior to isocenter	−0.001	0.036	0.054	−0.061
Overall combined		−0.002	0.022		
Confidence limit=(|mean|+1.96σ)			0.045		

they are listed in Table I.) Ion chamber predictions were obtained with averaged values over a number of points within the chamber volume for institutions A–C and G–I. Institution D used a single point at the chamber center. For E and F (one institution with two planning/delivery systems), one point prediction was used, but the variation within the chamber volume was inspected and found to be 1% within the PTV region and 2% within the OAR. For institution J, the chamber volume was so small that a single point prediction was deemed sufficiently accurate. The difference between the measured and planned doses are expressed as a ratio of the prescription dose instead of the predicted local dose. This choice was deemed more clinically relevant, especially for low dose regions, for which reporting the difference from the local dose can overstate the clinical importance of the deviation. For the high dose low gradient regions in the target, the average difference between the measured and planned doses, expressed as a ratio to the prescribed dose and averaged over all tests and institutions, was -0.002 ± 0.022, corresponding to a confidence limit (mean $+1.96\sigma$) of 0.045. 94% of the results fell within the confidence limit. The average of the absolute value of the ratio was 0.009.

For the low dose avoidance structures, the average difference between the measured and planned doses, expressed as a ratio to the prescription dose and averaged over all tests and institutions, was 0.006 ± 0.030, corresponding to a confidence limit of 0.064. However, this result is skewed by a single number coming from institution J, which had much larger variations that were attributed to the presence of high dose gradients. For the low dose region in the prostate case, J reported a difference ratio of 0.142, more than twice the difference in any other cases. Repeat measurements with the chamber shifted by 1–2 mm produced better agreement. Discarding that single result changed the overall average difference between the measured and planned doses to 0.003 ± 0.022, corresponding to a confidence limit of 0.047, similar to the result for the high dose regions. With that change, 91% of the results fell with the confidence limit. Before the change, 98% of the points fell within the larger confidence limit. The average of the absolute value of the ratio was 0.011.

III.B.3. Composite film measurements

Seven of the nine facilities analyzed films exposed within the phantom, although not all seven did each of the suggested planes. These institutions all had their film dosimetry normalized to a point or to an area that corresponds to ion chamber measurement. The results are presented in Tables XI and XII. For the high dose planes, the percentage of points passing the gamma criteria, averaged over all tests and institutions, was 96.6 ± 4.1. For the low dose planes, the percentage of points passing the gamma criteria, averaged over all tests and institutions, was 96.1 ± 4.8. Combining all the film planes gives an average of 96.3 ± 4.4. Using the same approach to establishing a confidence limit but recognizing that it is the reduction from 100% of points passing that is important leads to a somewhat different formulation: (100 −mean)+1.96σ is the percentage less than 100 that constitutes the limit. This gives a value of 12.4, or 87.6%. 93% of the film results reported gamma pass rates of 88% or higher. Note that this formulation may not correspond to the 95% confidence level associated with a two-tailed Gaussian distribution but is nevertheless used here as a reasonable method to compare results.

TABLE VIII. High dose point in the PTV measured with ion chamber: [(measured dose)−(plan dose)]/prescription dose, averaged over all the test plans measured at each institution, with associated confidence limits.

	Institution									
	A	B	C	D	E	F	G	H	I	J
Mean	−0.004	−0.012	−0.006	−0.007	0.017	0.002	−0.013	−0.014	−0.009	0.008
Standard deviation (σ)	0.023	0.021	0.011	0.004	0.014	0.012	0.044	0.004	0.030	0.019
Local confidence limit (|mean|+1.96σ)	0.049	0.053	0.028	0.015	0.044	0.026	0.098	0.022	0.068	0.044
Number of measurements	6	6	5	6	5	3	5	6	6	5

TABLE IX. Low dose point in the avoidance structure measured with ion chamber: [(measured dose) −(plan dose)]/prescription dose, averaged over the institutions, with associated confidence limits.

Test	Location	Mean	Standard deviation (σ)	Maximum	Minimum
Multitarget	4 cm inferior to isocenter	−0.008	0.019	0.014	−0.050
Prostate	2.5 cm posterior to isocenter	0.000	0.018	0.030	−0.025
Head and neck	4 cm posterior to isocenter	0.004	0.024	0.061	−0.017
CShape (easier)	Isocenter	0.010	0.024	0.050	−0.037
CShape (harder)	Isocenter	0.009	0.025	0.055	−0.021
Overall combined		0.003	0.022		
Confidence limit ($\|\text{mean}\|+1.96\sigma$)			0.047		

III.B.4. Per-field measurements

Seven facilities did field-by-field measurements. Five used a diode array (MAPCHECK, Sun Nuclear Corporation, Melbourne, FL), one used film, and one used EPID. All used gamma criteria of 3%/3 mm. Tables XIII and XIV present the average percentage of points passing the gamma criteria for the different institutions and test cases. As was done for the composite film measurements, the confidence limit here is expressed as the reduction from 100%. The overall results are 97.9 ± 2.5, leading to a confidence limit of 7.0 or 93.0%. 94% of the per-field results reported gamma pass rates of 93% or higher.

IV. DISCUSSION

IV.A. Test suite

The test suite is a useful starting point but it is neither comprehensive nor necessarily representative of a particular clinic's practice. The suite uses only 6 MV, for example. The head and neck case has a PTV volume that is relatively large, such as for a postoperative treatment, while clinical cases often have multiple targets prescribed to different doses. None of the test cases represent the broad targets found in pelvic cases in which lymph node chains are targeted and bowel is to be spared. Facilities should create mock clinical cases that reasonably represent the types of cases that they see in clinical practice, including tests of other energies if used.

IV.B. Planning results

The planning results demonstrate that the various institutions were able to produce comparable plans. The purpose of the study was not to compare planning results but to test how well the measured doses matched those planned. The planning results needed to be comparable so that the degree of beam modulation would likely be similar. It would be desirable to have measures of beam modulation to confirm that the plans were comparable in that regard, since the level of complexity of individual plans is related to the delivery accuracy and associated quality assurance metrics. As an example, one participating institution generated multiple head and neck and prostate plans meeting the TG-119 planning goals with varying complexity to evaluate the effect of plan complexity on delivery accuracy for these standardized test cases.[8] Plans were done with the ECLIPSE planning system. Complexity was varied using smoothing parameters available in ECLIPSE and quantified using the number of monitor units for delivery. Results revealed a decrease in gamma pass rate with increasing plan complexity. While this decrease was less than 1% for the prostate cases using both film and MAPCHECK, measurements for the more complex head and neck case revealed differences in gamma pass rate of approximately 3% from composite film analysis and almost 9% from individual field measurements using MAPCHECK. Unfortunately, surrogates such as the total monitor units are not readily useful when comparing different delivery techniques, such as sliding window, step and shoot, or tomotherapy. Thus, selected dose-volume values were used to assess that the plans were reasonably similar to each other.

The variation in the target dose-volume parameters was typically less than 1.5%, except for the harder CShape test which stipulated unachievable goals. The high dose in that PTV exceeded the D10 (i.e., dose to 10% of the volume) limit with a variation of just over 4%.

The variation in the specified dose-volume parameters to the normal structures ranged from 2% to 20%. Doses varied

TABLE X. Low dose point in the avoidance structure measured with ion chamber: [(measured dose)−(plan dose)]/prescription dose, averaged over all the test plans measured at each institution, with associated confidence limits.

	Institution									
	A	B	C	D	E	F	G	H	I	J
Mean	−0.006	−0.010	0.006	0.013	−0.005	n/a	−0.005	0.008	−0.008	0.045
Standard deviation (σ)	0.007	0.018	0.034	0.006	0.013	n/a	0.005	0.024	0.014	0.021
Local confidence limit ($\|\text{mean}\|+1.96\sigma$)	0.020	0.045	0.072	0.024	0.030	n/a	0.014	0.056	0.036	0.086
Number of measurements	5	5	5	5	5	1	5	5	5	4

TABLE XI. Composite film: Percentage of points passing gamma criteria of 3%/3 mm, averaged over the institutions, with associated confidence limits.

Test	Location	Mean	Standard deviation (σ)	Maximum	Minimum	Number of submissions
Multitarget	Isocenter	99.1	0.9	100	97.5	8
Prostate	Isocenter	98.0	2.24	99.8	94.2	7
	2.5 cm posterior	93.2	7.6	99.9	85	3
Head and neck	Isocenter	96.2	3.0	100	92.4	8
	4 cm posterior	97.6	1.5	98.9	95.6	4
CShape (easier)	Isocenter	97.6	3.9	100	88.9	7
	2.5 cm anterior to isocenter	93.9	5.0	99.6	87.9	5
CShape (harder)	Isocenter	94.4	6.0	99.4	86.2	5
	2.5 cm anterior to isocenter	93.0	7.2	99.9	81.3	5
Overall combined		96.3	4.4			
Confidence limit = (100 − mean) + 1.96σ				12.4 (i.e., 87.6% passing)		

more for structures with goals that were either very easy or very difficult to meet. In some cases, such as for the Bladder D30 for the prostate plan, the dose limit was easy to satisfy and so the actual dose could vary without penalty. Some planners forced the dose as low as it could go without compromising other goals, while others did not. At the other extreme, the harder CShape, dose goal to the core critical structure could not be met and the actual dose achieved depended on the choices made by the planner and the capabilities of the planning and delivery system. In order to reduce the variability in the planning results, additional plan goals and indications of priority would need to be specified.

IV.C. Ion chamber results

For the target regions, each institution's average ion chamber measurements were within 2% of the planned dose. Four of the nine institutions had at least one measurement that differed from planned by more than 3%. Facility G's results for the two CShape cases were 6% less than planned and that for the prostate and multitarget cases 2.2% and 3.0% more than planned, respectively, for a mean of −1.3% but a standard deviation of 4.4%. On the other extreme, facility D was more consistent with a mean of −0.7% and standard deviation of 0.4%. The confidence limit for the combined group for these measurements in the high dose, low gradient region was 4.5%.

For the lower dose measurements in the avoidance structure regions, eight of the nine facilities reported average dose within 2% of planned (where the percentage is of the prescription dose, not the local dose) with standard deviations of the same magnitude. The confidence limit for the combined group for these measurements in the low dose, avoidance structure was 4.7%.

Based on these collective results, it seems reasonable to expect that an institution's average agreement between predicted and measured doses measured with an ion chamber should certainly be at least within 3% (of prescription dose). Most of the participants in this study reported averages within 1.5% of expected from the treatment plan. Some outliers were seen but few outside the confidence limits determined by this group. To be quantitative, an institution can calculate its own confidence limit with this methodology, and the result should be comparable to this group's. The confidence limit for the group was obtained by combining many measurements. A single institution performing only the tests in this test suite will have weaker statistics that could be improved with more repetitions, either of the same tests or similar ones derived from clinical plans. However, if the confidence limit derived from the test suite is much larger than the group's (as for facility G, for example), then it is likely that the IMRT system can be improved before clinical treatments commence.

IV.D. Composite film measurements

The first point of interest regarding the composite film results is that two of the nine facilities did not report any. The increasing prevalence of digital imaging and decreasing availability of well-maintained film processors is making it more difficult to accomplish planar dose measurements in phantom. This is a concern, because it is important to know

TABLE XII. Composite film: Percentage of points passing gamma criteria of 3%/3 mm, averaged over the test plans, with associated confidence limits.

	Institution							
	A	B	D	E	F	G	I	J
Number of film planes	9	9	4	7	4	9	5	5
Mean	99.5	92.6	99.9	97.6	98.0	93.0	95.8	97.5
Standard deviation (σ)	0.4	4.3	0.3	2.3	1.1	6.5	3.6	2.9
Local confidence limit (100 − mean) + 1.96σ	1.2(98.8%)	15.7(84.3%)	0.6(99.4%)	6.9(93.1%)	4.5(95.5%)	19.7(80.3%)	11.2(88.8%)	8.2(91.8%)

TABLE XIII. Per-field measurements: Average percentage of points passing the gamma criteria of 3%/3 mm, averaged over the institutions, with associated confidence limits.

Test	Mean	Standard deviation (σ)	Maximum	Minimum
Multitarget	97.8	3.5	99.8	90.8
Prostate	98.6	2.4	100	93.3
Head and neck	98.1	2.0	100	94.2
CShape (easier)	97.4	2.8	99.8	93.0
CShape (harder)	97.5	2.6	99.9	94.0
Overall combined	97.9	2.5		
Confidence limit=(100−mean)+1.96σ		7.0 (i.e., 93.0% passing)		

how well the different IMRT fields combine to produce a predicted distribution within the geometric shape of the phantom. It is not possible to assess the accuracy of the cumulative doses by only analyzing the dose distribution for each field in a geometry radically different from the phantom/patient. The commissioning process needs to test all the components of the planning and delivery system, as components and as an integrated system. Certainly, if the gantry is maintained in a vertical direction for the individual field measurements, then problems with delivery with different orientations with respect to gravity will not be found. Issues with transmission through couch support assemblies would also not be identified without doing composite measurements, and this might be relevant if gantry angles are used for IMRT that were not used for 3D conformal plans.

If a facility cannot perform reliable planar dosimetry in phantom, then a larger set of individual point doses needs to be measured, but that is not the recommended solution. Facilities that are losing or have lost the ability to do film dosimetry with radiographic film should be moving to alternatives such as radiochromic film,[9] computed radiography plates,[10] and detector arrays with attendant scanning and analysis tools.

For each of the six facilities performing film dosimetry, the average percentage of points passing the gamma criteria exceeded 90%, where the average is over all the analyzed planes. Combining all the results gives an overall average of 96.4% with a standard deviation of 4.3%. Facilities B and G reported more variation than did the others. Facility B reported their test of the reference band case as having 99.9% points passing, so its results for the IMRT test cases are not likely to be heavily influenced by film dosimetry problems. Facility G did not report results for the band case, so one cannot assess the film dosimetry accuracy. The confidence limit for these collective results was 12%, which indicates that the percentage of points passing the gamma criteria should be more than 88% approximately 95% of the time. For our collective results, 93% of the tests fell within the confidence limit.

The reported percentage of points passing gamma criteria depends heavily on the details of the implementation of the data analysis. Examples include using a region of interest or a threshold to exclude some points from assessment, normalizing the measurements to some reference point, and defining the percentage agreement in terms of local dose or prescription dose. In practice, physicists use commercial tools that have different available options, and so it is difficult to offer definitive guidance regarding acceptance levels for gamma analysis results. It seems reasonable, however, to expect that if one normalizes the film results to ion chamber measurements in the high dose region on the same plane, then on average about 95% of the points on the plane within the region of interest should pass gamma criteria of 3%/3 mm with a confidence limit that ranges down to 88%.

IV.E. Per-field measurements

Five of the seven facilities that performed this test used the same model of dosimeter and software and so the analysis could be standardized. Doing so is important in order to compare results, because the percentage of points passing the gamma criteria can change dramatically depending on the details of the analysis.

TABLE XIV. Per-field measurements: Average percentage of points passing the gamma criteria of 3%/3 mm, averaged over the test plans, with associated confidence limits.

	Institution						
	A	B	C	D	E	F	H
Measurement device	Diode array	Diode array	EPID	Diode array	Diode array	Film	Diode array
Mean	98.9	93.3	99.4	99.2	98.6	99.6	96.8
Standard deviation(σ)	1.5	1.5	0.4	1.3	1.5	0.3	2.5
Local confidence limit (100−mean)+1.96σ	3.9 (96.1%)	9.5 (90.5%)	1.3 (98.7%)	3.4 (96.6%)	4.3 (95.7%)	1.0 (99.0%)	8.1 (91.9%)
Number of studies	5	5	5	5	4	4	5

Two of the institutions used film or EPID as the device for assessing per-field quality. Such devices have greater spatial resolution than an array of diodes or ion chambers. These two institutions reported average gamma pass rates that exceeded 99%, which was generally larger than those from the diode array. This study does not provide enough data to independently derive confidence limits for film or EPID per-field measurements, but it is reasonable to assume that these should not be worse than the combined results reported here.

With the gamma analysis parameters used in this study, the average percentage of points passing the criteria was quite high: the overall average was 97.9% with a standard deviation (of the average) of 2.5%. This corresponds to a confidence limit of 7.0%, which means that the percentage of points passing the criteria should exceed 93% approximately 95% of the time. For our collective results, 94% of the tests fell within the confidence limit.

IV.F. Overall comments

This test protocol asks for both composite planar dosimetry in phantom and per-field measurements. The task group recommends that both be done whenever possible at the time of commissioning, because the information provided is complementary. Composite delivery checks that the doses add together as planned, but it is possible that the magnitude of deviations with some beam angles could be suppressed when combined with the other fields. Checking each field individually on a plane perpendicular to the beam permits that beam's delivery to be analyzed in detail but does not assure that the beams combine appropriately. Multiple ion chamber measurements may substitute for composite planar dosimetry if necessary.

The task group also cautions against relying solely on per-field gamma analysis. When a beam is highly modulated, a gamma analysis may fail to identify some types of problems because it is possible to find some point that matches the intended dose by searching up to 3 mm in all directions. Per-field 2D dose measurement differs from the measurement with ion chamber in that the ion chamber is normally placed on a high dose, low gradient region where a difference from the predicted dose may be more indicative of the change in delivered dose to the patient. The gamma passing rate with a 2D array may not directly reflect a dose scaling factor error since it compares not only the scale of the dose but also the distance between agreement points. The gamma values depend on the data analysis method and criteria used as well.[11] If, for example, per-field QA is based on the Van Dyk gamma criterion that is normalized to the maximum dose as the default,[12] some errors could be hidden such as those from the MLC transmission factor, tongue and groove effect, and dose calibration. Another viable option is to combine gamma analysis and the average percentage error of all the measured points with a predefined threshold.[13]

Each of the facilities participating in this comparison had passed the RPC credentialing tests with its IMRT head and neck phantom.[14] That phantom uses TLDs to assess dose and radiochromic film to assess the dose gradient between the target and the organ at risk. For this group of facilities, the average agreement of the dose measured by TLDs in the target regions with the planned dose was −0.4% with a standard deviation of 2.6%. For the TLDs in the organ at risk, the average agreement was −1.4% with a standard deviation of 18.8%. (This percentage is of the local dose, not the prescription dose. The predicted doses in the PTV region averaged 7.13 Gy and the predicted doses in the avoidance structure averaged 2.77 Gy.) The average displacement of measured isodose lines in the gradient region from the planned positions was 1.1 mm with a standard deviation of 1.3 mm. These data provide independent confirmation of the accuracy of the IMRT planning and delivery by these facilities.

The data in this report would be stronger if (1) more facilities had reported the results of the preliminary band test in a consistent fashion and (2) if there were repetitions of the measurements at each facility to assess the reproducibility of the results. Those facts, along with the uncertainty in the details of the various gamma analyses, make it difficult to put error bars on the results and therefore draw stronger conclusions about the agreement between plan and measurement that should be expected. Nevertheless, these data should be helpful for institutions assessing their own IMRT commissioning. Additionally, independent verification using IMRT phantoms available through the RPC is always prudent.

IV.G. An example of the practical utility of the tests

Institution B used these sets of tests as part of the commissioning evaluation of the beam modeling for one newly installed linear accelerator with multiple photon energies. The following paragraphs briefly describe key elements of the commissioning process and illustrate how using these tests identified that the commissioning needed to be improved for one beam energy and how the main source of error was identified.

The commissioning process for the synergy S system included collections of a complete set of scan and point measurement data for photons and electrons as specified by the CMS/XIO beam modeling guide for the beam modulator. Beam data collection and calibration were internally verified by at least two independent measurements and checked against standard data sets. Treatment planning system modeling followed the guidelines of TG53.[15] When compromises had to be made, the best fits were chosen for situations mimicking IMRT. Before actual clinical implementation, periodic QA baselines were established, and site specific IMRT plans and QA measurements were performed on phantoms. QA measurements of 3D conformal plans achieved the following agreement statistics: 3 mm DTA, 3% difference, and produced pass rate of 97.8% average (2.6% STD).[16]

In the initial testing of these sets of IMRT plans and measurements, however, a larger than expected discrepancy was observed for the prostate plan with one of the photon beams (10 MV). The field-by-field analysis of diode array measurements yielded an average gamma pass rate of 80.5% with 3 mm DTA and 3% dose difference.

When faced with such a finding, the clinical physicist must consider various reasons for the discrepancy. The reported sources of deviation between planned and measured doses fall into the following three major categories, treatment planning system (major source, close to 50%), delivery system, and measurement process.[3,17]

For the treatment planning systems, there could be inaccurate data input, inaccurate/insufficient modeling (one example is given in Sec. IV H), and software glitches. Complexity of the IMRT plans may exacerbate the above mentioned inadequacies. For the delivery system, there could be output inaccuracy, beam definition system error (e.g., MLC error), or error in patient positioning system. For the sources in measurement process, there could be suboptimal measurement techniques, limitation/inaccuracy in measurement devices, human error in execution process. The uncertainties due to some of the factors described above were quantified in the dose distributions.[18,19]

In this case, the parameters that constitute beam models for the superposition/convolution algorithm were closely evaluated, including energy spectrum, build-up electron contamination, Gaussian parameters for profile tail modeling, transmissions of beam modifiers, and penumbra modeling for beam modifiers. These factors were found to have significant influence upon the dosimetric outcome from treatment planning systems.[20,21] Perturbations were applied to the energy spectrum, Gaussian parameters, and transmission factors of multileaf collimators, without significantly affecting the fitting of the measurements during the modeling process. Corresponding dose distributions were created to be compared with measurements for IMRT plans. It was found that gamma passing rates given certain DTA and percentage dose deviation were most sensitive to MLC transmission factors. By decreasing the MLC transmission from 3% to 1.5% for the 10 MV beam, average gamma passing rate for the prostate IMRT plan QA changed from 80.5% to 93.3% (3 mm DTA, 3% dose difference). Adjustments were therefore made to corresponding beam modeling parameters to improve the agreements for IMRT QA measurements, keeping similar fitting performances for the modeling process. Subsequent dosimetric testing using this suite of IMRT tests and other tests of 3D conformal beams demonstrated improved correspondence between calculation and measurement for the IMRT cases and continued agreement for the 3D conformal cases.

IV.H. Comparison to other work

In 2005, Gillis *et al.* published the results of a similar study conducted in Europe.[22] Eight European institutions planned and delivered an IMRT treatment to a horseshoe-shaped PTV surrounding a central avoidance structure in an idealized pelvic phantom, a geometry similar to that used in the CShape tests in this study. A variety of planning and delivery systems were used. 95% of the PTV volume was to receive at least 99% of the prescription dose and no more than 1% of the avoidance structure was to receive 70% of the prescription dose. The avoidance structure was separated from the PTV by 1 cm. Ion chamber measurements were made in the PTV and the avoidance structure, and film measurements were made in seven axial planes. The chamber results were used to adjust the film calibration. The films were processed and analyzed at one facility to improve consistency and were analyzed using gamma criteria of 3 mm DTA and 4% dose agreement. The overall average difference between the measured and the planned doses to the PTV, expressed as a ratio to the planned dose, was -0.014 ± 0.017. The mean dose to the avoidance structure was 53% of the prescription dose, and the overall average difference between the measured and the planned mean doses to the avoidance structure, expressed as a ratio to the prescription dose, was 0.000 ± 0.020. Thus, the European group's dose results were similar to the ion chamber results from this study. The European group summarized their gamma index results in terms 95th percentiles. Overall, the gamma index representing the 95th percentile was 0.84 ± 0.28. For their tests, at least 95% of the points in the PTV and the avoidance structure passed a gamma test using 3 mm DTA and 4% dose. Thus, the European study provides additional corroboration for the degree of agreement to be expected for a properly commissioned IMRT system.

A recent ESTRO booklet on "Guidelines for the Verification of IMRT"[17] summarizes the experience of several European institutions. It also discusses the use of confidence limits as expressed here. They recommend tolerance limits of $\pm 3\%$ for ion chamber measurement in target areas and action limits of $\pm 5\%$ for point dose verification.

IV.I. Confidence limits and action levels

This study has focused on IMRT commissioning, not on per-patient quality assurance tests. In this context, the group data have been used to develop confidence limits to assist in judging the adequacy of IMRT commissioning. A confidence limit is a statistical term, and its application requires the acquisition of a reasonable number of data points. Thus, this task group report recommends that measurements of a suite of IMRT tests be performed, mimicking the range of cases that will be encountered in practice. The average and standard deviation of the results can be used to compare with those obtained by this group. The confidence limit obtained by the local facility can be compared to this group's as given at the bottom of Tables VII, IX, XI, and XIII. The local confidence limit should be on the same order or less than this group's. If it is much larger, then that is an indication that the local IMRT system is not commissioned as well as it could be. However, that conclusion presumes that the analysis has been performed in a comparable manner and that the number of tests is sufficient to warrant a statistical judgment. Even though the 1.96 multiplier used in the confidence limit calculation strictly applies when a very large number of samples is available, we have chosen to use it to be clinically conservative instead of using a numerically larger multiplier consistent with a smaller sample size. Repetition of these tests is suggested in order to enlarge the sample size and make the

statistical judgment more reliable. The number of test cases can also be enlarged by using contours and dose goals from the local practice.

This approach to testing and analysis is not limited to 6MV, although the specific results from this study were for that energy. It would be reasonable to assume that similar confidence limits should hold for higher energies, but that remains to be proven.

The confidence limit does provide a mechanism for determining reasonable action levels for per-patient IMRT verification studies. If the confidence limit is established with enough points to provide good statistics, then using the value of 1.96σ suggests that variances in excess of the limit may occur about 5% of the time (one can decide on a higher or lower potential action triggering percentile by using a $x\sigma$ value, where x can be larger or smaller than 1.96). For this group, the confidence limit for ion chamber measurements in the target region was 4.5% and for the low dose region was 4.7%. Thus, the recommendations of Palta *et al.*[5] are consistent with this result: For point dose measurements they recommended an action level of $\pm 5\%$ in a high dose, low gradient region and $\pm 7\%$ in a low dose, low gradient region. This work provides additional support for action levels expressed in terms of percentage of points passing gamma criteria of 3%/3 mm: 90% for per-field measurements and 88%–90% for composite irradiations analyzed with radiographic film. (As noted above, however, the results of a gamma analysis depend heavily on the details of the implementation, so these recommendations must be considered in conjunction with the specific method used here.)

In practice, one would expect that simple cases (e.g., prostate) would rarely approach these limits, but highly modulated cases (e.g., perispinal) might exceed it if the system was pushed beyond its capabilities. Thus, this type of examination of the IMRT commissioning accuracy provides a baseline for the initial assessment of per-patient verification. A full discussion of per-patient quality assurance is beyond the scope of this report, but careful commissioning is a prerequisite for quality treatments.

V. CONCLUSION

Treatment planning and delivery in radiation therapy are never perfect, and so the practical question is "how good is good enough?" This study has not attempted to answer that question in a rigorous way but instead has studied the question "what is a reasonable and achievable standard for IMRT commissioning?" To provide a basis for that judgment, a group of institutions that have passed the RPC IMRT credentialing used a suite of standardized test cases to determine the degree of agreement achievable with their planning and delivery systems. These results, summarized at the bottom of Tables VII, IX, XI, and XIII, can be used as a practical baseline for comparison by other facilities as they evaluate their own IMRT commissioning. Facilities interested in using this test suite can download the DICOM-RT images and structure sets from http://www.aapm.org/pubs/tg119/default.asp along with a detailed description of the planning, measurement, and analysis process.

ACKNOWLEDGMENTS

The authors would like to thank all the physicists and dosimetrists from these institutions who performed the planning and measurements that made this cross comparison possible, Todd Bossenberger, Shalini Pandya, and Adrian Nalichowski from Wayne State University and Amy Harrison, Kevin Fallon, and Anthony Doemer from Thomas Jefferson University. The authors would like to acknowledge their industrial consultant, Bill Simon from Sun Nuclear. They also would like to thank their reviewers from the AAPM Therapy Physics Committee and those from Medical Physics for their helpful comments and suggestions.

[a]Author to whom correspondence should be addressed. Electronic mail: ying.xiao@jeffersonhospital.org

[1]G. A. Ezzell *et al.*, "AAPM REPORT: Guidance document on delivery, treatment planning, and clinical implementation of IMRT: Report of the IMRT subcommittee of the AAPM radiation therapy committee," Med. Phys. **30**, 2089–2115 (2003).
[2]I. J. Das, C. Cheng, K. L. Chopra, R. K. Mitra, S. P. Srivastava, and E. Glatstein, "Intensity-modulated radiation therapy dose prescription, recording, and delivery: Patterns of variability among institutions and treatment planning systems," J. Natl. Cancer Inst. **100**, 300–307 (2008).
[3]G. S. Ibbott, D. S. Followill, H. A. Molineu, J. R. Lowenstein, P. E. Alvarez, and J. E. Roll, "Challenges in credentialing institutions and participants in advanced technology multi-institutional clinical trials," Int. J. Radiat. Oncol., Biol., Phys. **71**, S71–S75 (2008).
[4]J. Venselaar, H. Welleweerd, and B. Mijnheer, "Tolerances for the accuracy of photon beam dose calculations of treatment planning systems," Radiother. Oncol. **60**, 191–201 (2001).
[5]J. Palta, S. Kim, J. Li, and C. Liu, in *Intensity-Modulated Radiation Therapy: The State of Art*, edited by J. R. Palta and T. R. Mackie (Medical Physics, Madison, WI, 2003), pp. 593–612.
[6]D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," Med. Phys. **25**, 656–661 (1998).
[7]D. Letourneau, M. Gulam, D. Yan, M. Oldham, and J. W. Wong, "Evaluation of a 2D diode array for IMRT quality assurance," Radiother. Oncol. **70**, 199–206 (2004).
[8]S. Pandya and J. Burmeister, "SU-GG-T-127: Effect of fluence smoothing on plan quality and delivery accuracy in intensity modulated radiotherapy," Med. Phys. **35**, 2755 (2008).
[9]A. Niroomand-Rad *et al.*, "Radiochromic film dosimetry: Recommendations of AAPM radiation therapy committee task group 55," Med. Phys. **25**, 2093–2115 (1998).
[10]A. J. Olch, "Evaluation of a computed radiography system for megavoltage photon beam dosimetry," Med. Phys. **32**, 2987–2999 (2005).
[11]MAPCHECK user manual, March 2006, 1175011 Rev H.
[12]J. Van Dyk, R. B. Barnett, J. E. Cygler, and P. C. Shragge, "Commissioning and quality assurance of treatment planning computers," Int. J. Radiat. Oncol., Biol., Phys. **26**, 261–273 (1993).
[13]S. Both *et al.*, "A study to establish reasonable action limits for patient specific quality assurance in intensity-modulated radiation therapy," J. Appl. Clin. Med. Phys. **8**(2), 1–8 (2007).
[14]A. Molineu *et al.*, "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**, 577–583 (2005).
[15]B. Fraass *et al.*, "American association of physicists in medicine radiation therapy committee task group 53: Quality assurance for clinical radiotherapy treatment planning," Med. Phys. **25**, 1773–1829 (1998).
[16]A. Harrison *et al.*, "SU-FF-T-382: Special dosimetric/measurement considerations in commissioning a novel integrated MiniMLC linear accelerator," Med. Phys. **34**, 2489–2490 (2007).

[17]M. Alber *et al.*., *Guidelines for the Verification of IMRT* (ESTRO, Brussels, Belgium, 2008).
[18]H. Jin, J. Palta, T. S. Suh, and S. Kim, "A generalized a priori dose uncertainty model of IMRT delivery," Med. Phys. **35**, 982–996 (2008).
[19]H. Jin, H. Chung, C. Liu, J. Palta, T. S. Suh, and S. Kim, "A novel dose uncertainty model and its application for dose verification," Med. Phys. **32**, 1747–1756 (2005).
[20]P. Cadman, R. Bassalow, N. P. Sidhu, G. Ibbott, and A. Nelson, "Dosimetric considerations for validation of a sequential IMRT process with a commercial treatment planning system," Phys. Med. Biol. **47**, 3001–3010 (2002).
[21]A. Rangel, N. Ploquin, I. Kay, and P. Dunscombe, "Towards an objective evaluation of tolerances for beam modeling in a treatment planning system," Phys. Med. Biol. **52**, 6011–6025 (2007).
[22]S. Gillis, C. De Wagter, J. Bohsung, B. Perrin, P. Williams, and B. J. Mijnheer, "An inter-centre quality assurance network for IMRT verification: Results of the ESTRO QUASIMODO project," Radiother. Oncol. **76**, 340–353 (2005).
[23]E. E. Klein, J. Hanley, J. Bayouth, F. Yin, W. Simon, S. Dresser, C. Serago, F. Aguirre, L. Ma, B. Arjomandy, C. Liu, C. Sandin, and T. Holmes, "Task Group 142 Report: Quality Assurance of Medical Accelerators," Med. Phys. **36**, 4197–4212 (2009).