Addressing Misconceptions on IMRT Quality Assurance

This paper addresses statements made by Stephen Kry, Ph.D., regarding IMRT QA. The statements originated in a 2019 publication by Kry et al[1], and have subsequently been shared in a series of presentations, including a Point/Counterpoint session[2] during the 2020 Joint AAPM I COMP Virtual Meeting.

Sun Nuclear welcomes a fair and transparent conversation on the effectiveness, advantages, and disadvantages of any radiation therapy (RT) treatment or QA method. We submit these perspectives as part of this important conversation.

SUN NUCLEAR
corporation

First Things First

The 2020 Joint AAPM I COMP Virtual Meeting included a session titled, *A Point/Counterpoint on Current and Future Directions for Patient Specific QA.*[2] Dr. Andrea McNiven, Ph.D., presented the point position: current patient-specific QA will remain an essential part of practice. Dr. Kry presented the counterpoint position: calculation-based QA should be the future. In addition, Dr. Kry provided a PDF handout to support his perspective.[2] There is much in his handout we agree with, including the following assertions:

1. IMRT QA plays an important role, verifying deliverability of the plan, and verifying intended dose is delivered as planned.
2. There are errors to be caught and IMRT QA represents a detection opportunity.
3. Strong IMRT QA and in-vivo QA measurement systems have the potential to detect multiple failure modes (calculation, delivery, anatomical, patient setup, etc.).
4. IMRT QA is well-established, with a long history, ample available guidance, and deep experience to draw from.
5. There are clear cases of value from IMRT QA, including catching errors and enabling interventions, as well as highlighting opportunities to improve treatment planning.
6. Time invested in IMRT QA is acceptable if it's time well-spent.

Similar to the 2019 publication, there are also assertions presented that we believe misrepresent the full body of data. In the handout provided by Dr. Kry for the Point/Counterpoint session, these assertions are generally framed as "Cons" for current-state IMRT QA. With the following, we share our thoughts on these assertions and provide the factual evidence which was omitted.

Please direct questions to:
Jennifer Hamilton, M.E., DABR - jenniferhamilton@sunnuclear.com
Jeff Kapatoes, Ph.D. - jeffkapatoes@sunnuclear.com
Sun Nuclear Corporation

Implied Assertions

There are four general assertions made by Dr. Kry which we fill focus on throughout this white paper. They are:

1. **Beam modeling errors are more common and impactful than the physics community has realized due to poor QA methods**
   - "Traditional measurement-based methods of IMRT QA are suspect. These approaches have come under increasing scrutiny for their inability to detect major and substantial errors in the dose being delivered to the patient. Numerous standard measurement-based IMRT QA methods have been found to have poor sensitivity in the identification of low quality or unacceptable IMRT plans."[1]
2. **Measurement-based QA has poor sensitivity to errors**
   - "The traditional IMRT methods also performed consistently poorly regardless of whether a 3%/3mm criteria was used or a 2%/2mm criteria."[1]
   - "Traditional IMRT QA methods, as implemented clinically, struggle to detect low quality radiotherapy plans."[1]
   - "Arrays can't work."[2]
   - "Traditional QA devices are just 'green check mark generators.' These devices never do a good job of separating acceptable and unacceptable plans – there's no criteria you can use; there's no threshold you can use to make these devices work."[3]
3. **In-vivo QA has poor sensitivity to patient-related errors**
   - "In-vivo QA is unable to detect 2cm patient setup errors." - Point/Conterpoint AAPM 2020[2] – referencing Hsieh, et al, publication[4]
4. **3D Calculation has superior sensitivity to Measurement-based QA**
   - "Independent recalculation outperforms traditional measurement-based IMRT QA methods in detecting unacceptable plans" (title of Kry et al paper[1])
   - "Independent recalculation overwhelmingly outperformed the current measurement-based IMRT QA methods"[1]
   - "Compared to current, clinically implemented IMRT QA methods (in aggregate), and using common clinical criteria, Mobius3D-based recalculations were 12 times more sensitive at identifying failing phantom results. In particular, recalculation was significantly and dramatically superior to IMRT QA using an EPID, an ArcCHECK®, or a MapCHECK® device."[1]

Kry et al provide troubling examples from various studies for each of these assertions that may seem compelling. In the following pages, we deconstruct these examples and examine counter-examples omitted from Dr. Kry's publication and talks.

Addressing These Assertions via the Following Topics

- Focus on Beam Modeling, pages 5-6
- Independent Measurements are Necessary, page 7
- 3D QA is More Clinically Relevant than 2D QA, page 8
- Arrays Do Work, pages 8-10
- Automated In-Vivo QA is the Most Effective Use of Physics QA Time, pages 11-12
- Not All 3D QA is the Same, page 12

Types of Errors

It is important to begin by differentiating QA for IMRT/VMAT into three distinct categories. The nuances of these unique categories matter because different tools are effective at detecting different types of errors – there is no single QA tool that can universally detect all errors, though some are certainly more comprehensive than others.

Three distinct IMRT/VMAT QA categories:

- **3D Secondary Calculations** – historically used to verify MUs, but in modern radiotherapy 3D calculations can be used to verify that the treatment planning system (TPS) algorithm and beam models are accurate. 3D calculations can detect beam modeling errors and algorithm discrepancies. 3D secondary calculations can never detect changes in the beam output, beam quality, flatness, symmetry, plan transfer errors, deliverability errors, patient setup errors, patient anatomy changes, or daily deliverability errors.
- **Pre-Treatment QA** – used to verify that the plan has transferred correctly from the TPS to the linac, and can be delivered accurately by the linac. Pre-treatment QA can detect changes in the beam output, beam quality, flatness, symmetry, transfer errors, deliverability errors and MLC errors. Pre-treatment QA, especially through 3D reconstruction in patient anatomy, can also detect beam modeling errors and show the clinical significance of detected errors. Pre-treatment QA can never detect patient setup errors, patient anatomy changes, or daily deliverability errors.
- **In-Vivo Monitoring** – used to verify that the plan is delivered accurately with the patient on the treatment couch. In-vivo monitoring, especially through transit dosimetry measurement, can detect patient setup errors, some anatomical changes, couch insertion errors, localization accessory errors, intra-fraction motion, and daily deliverability errors.

QA Methods and Detection Capabilities

	3D Secondary Calculations	Pre-Treatment QA	In-Vivo Monitoring
Beam Modeling Errors	X	X[5]	*Unlikely**
Algorithm Discrepancies	X	X[5]	*Unlikely**
Beam Output		X	X
Beam Quality		X	X
Flatness/Symmetry		X	X
Transfer Errors		X	X
Deliverability/Complexity Errors		X	X
Patient Setup Errors			X
Patient Anatomy Changes			X

* Unless the error is large, beam modeling and algorithm discrepancies will be hidden by small daily patient variations during in-vivo monitoring.

Focus on Beam Modeling

Addressing Assertions 1, 3, 4

The paper and presentations in question are solely focused on **beam modeling errors.** The Kry et al[1] work does not seriously consider other sources of error that could only be detected by an independent measurement (for example: drifts in output, beam quality, or flatness/symmetry; MLC motor errors, data transfer errors, or deliverability errors). A 3D Secondary Calculation, if fully independent, may be excellent at catching modeling errors, but it can never detect errors related to data transfer and the linac's delivery of the plan, nor can it detect in-vivo errors such as patient setup or anatomy changes. The Kry et al publication addresses this in the Discussion section, about Mobius3D™ from Varian Medical Systems®: *"This recalculation approach (Mobius3D) would not be expected to have comprehensive sensitivity because it evaluates only one component of the radiotherapy process: the calculation. Errors in delivery or machine output could not be detected with the process implemented herein, limiting its sensitivity."*[1]

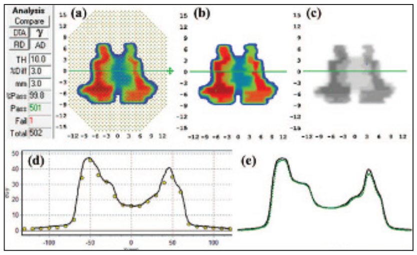
We concur beam modeling errors can be a pervasive and systemic source of error that is often overlooked due to poor QA methods. Some may assume small beam modeling errors are unlikely to create errors large enough for discrepancies to be clinically important. This is untrue. In modern radiotherapy, with highly modulated and complex treatments, very small beamlets can comprise a majority of treatment fields. Small, irregularly-shaped beamlets are highly dependent on accurate penumbra modeling and accurate MLC modeling – including tongue and groove effects, dosimetric leaf gap, leaf transmission, leaf thickness, etc.

One of many publications[5,6,7,8,9] to illustrate beam modeling impacts is Nelms et al.[5]

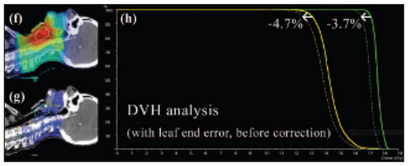
Nelms' publication on IMRT and VMAT errors lists several clinical examples of modeling errors that were missed by loose gamma criteria, but easily detected by measurement-based 3D QA using 3DVH™. Among the cases were dosimetric leaf gap errors, tongue-and-groove errors, volume averaging due to use of an overly large chamber for scanning, algorithm errors, and underestimation of small, narrow fields (overmodulation).

Case 1: Incorrect leaf-end modeling

Despite a high 3%/3mm average passing rate (99.2%), there was a systemic error in the TPS model of the rounded MLC leaf ends. The calculated penumbra at the MLC edges was too wide compared to measurements, resulting in TPS dose being overestimated.

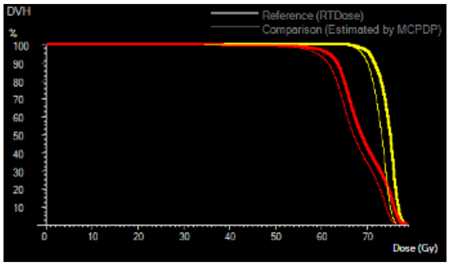


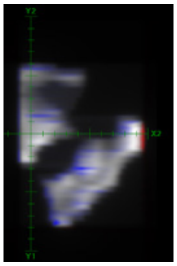
Analysis
Compare
DTA γ
RD AD
TH 10.0
%Diff 3.0
mm 3.0
%Pass 99.8
Pass 501
Fail 1
Total 502

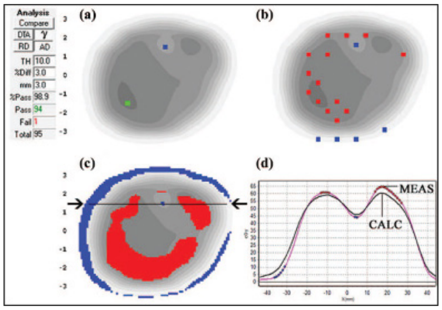


Case 2: TPS setting caused tongue-and-groove effects

Despite a high 3%/3mm average passing rate (99.4%), there was a systemic error in the TPS: the tongue-and-groove correction was turned off.





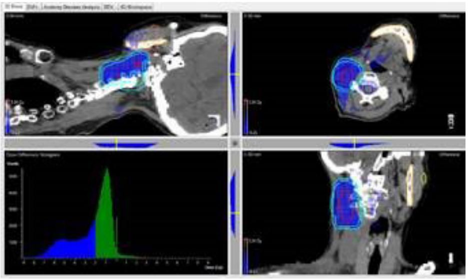


Case 3: Inaccurate (dose-averaged) profiles entered into beam model

Despite a high 3%/3mm average passing rate (99.3%), there was a systemic error due to the profiles for this model being acquired with a Farmer chamber

Case 4: Underestimation of dose for narrow MLC segments – Over-modulation

Despite a 3%/3mm passing rate of 94.7% with Delta4 and 93.9% with ArcCHECK®, a large number of very narrow fields (several mm in width) produced a ~5.5% cold region across the target areas.



The fact that 3%/3mm gamma criteria are insensitive to many errors, and that 3D QA is preferable for the detection of small, pervasive modeling errors is certainly not new. Over the past decade, numerous publications[4-9] confirm global gamma using 3%/3mm criteria are insufficient to detect clinically impactful errors.

AAPM Task Group 218[10] likewise rejected 3%/3mm criteria as too insensitive and recommends 3%/2mm for IMRT/VMAT, with tighter criteria used for SBRT/SRS or any plan with a margin <2mm for the Target or critical OARs.

Despite the precedent for appropriate gamma criteria for sensitive error detection, the Kry et al work uses 90%/3%/3mm criteria (which many agree is outdated) as a basis for conclusions that imply arrays aren't sensitive to beam modeling errors.[1] Even when the data is parsed into tighter criteria, all QA methods (many of them poor) are grouped into one batch, grouping single ion chamber readings with 3D arrays. Poor passing rates data were also excluded from the dataset.* This presentation of data has the effect of masking the sensitivity of the 3D arrays.

* The IMRT QA result was declared to have passed if at least one point dose assessment agreed within 3% or if >90% of pixels passed a composite gamma criteria of 3%/3 mm (or tighter). Field-by-field gamma results could have at most one field with <90% of pixels passing and still be declared as passing. The IMRT QA result was declared to have failed if all point dose assessments showed a disagreement of >3% or if <90% of pixels passed a gamma criteria of 3%/3 mm (or looser). Results that could not be categorized according to this system (e.g., >90% of pixels passing very loose gamma criteria or <90% of pixels passing very stringent gamma criteria) were excluded from this evaluation.

Independent Measurements are Necessary

Addressing Assertions 2, 3, 4

In the AAPM Point/Counterpoint talk[2], Dr. Kry correctly noted measurement-based QA has a place in detecting transfer and delivery errors. Transfer and delivery errors can be clinically significant and adversely affect patient outcomes, and both types of errors have been seen frequently in the clinic and in publications.[9, 11]

Transfer errors are a serious and semi-frequent cause of catastrophic plan failure. An infamous example is a head and neck plan referenced in the 2010 New York Times article on misadministrations. In this case, open beams were delivered to a patient because the MLC failed to transfer. The error resulted in the patient's death. Another example comes from Mans et al[11] where an SBRT plan was corrupted during transfer without displaying any errors. The first fraction of the plan was consequently delivered in error, with the MLCs and Jaws mis-synced for every beamlet. In-vivo QA detected the error, which enabled correction before the 2nd fraction. This publication found 17 serious errors out of 4,337 Fractions – 4 of which were file transfer errors.

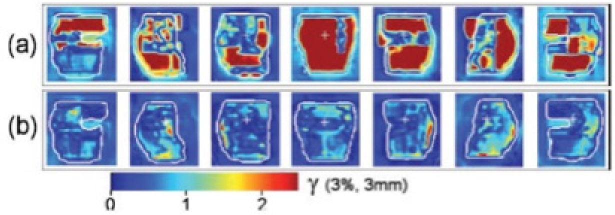
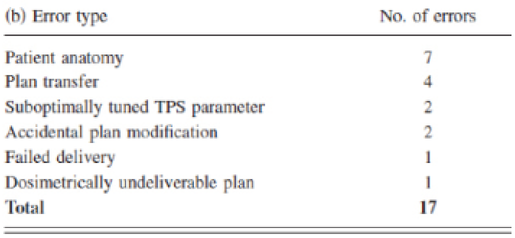


FIG. 2. γ-evaluations of (a) the first (malformed plan) and (b) the second (corrected plan) fractions in a plane parallel to the EPID, intersecting the isocenter. The white "+" indicates the isocenter.



(b) Error type	No. of errors
Patient anatomy	7
Plan transfer	4
Suboptimally tuned TPS parameter	2
Accidental plan modification	2
Failed delivery	1
Dosimetrically undeliverable plan	1
Total	**17**

Deliverability errors that reach clinical significance are not common, but when plans are highly complex, they can adversely affect clinical outcomes. The example published in Nelms et al[5] showed a 5.5% pervasive cold spot in a head and neck tumor. The error was undetectable with insensitive criteria of 3%/3mm and global normalization; however, with the available 3DVH algorithm applied to the same measurement to obtain full-density 3D results, the failure was obvious with only 19% of voxels passing in the PTV volume.

It is Sun Nuclear's assertion, based on the above and the collective experience of many clinical physicists, that measurement-based QA is required to detect Transfer and Deliverability errors.

Measurement-based QA is also required per ACR/ASTRO guidelines.[12] ASTRO 2016 Users Guide states: "There are a number of products that support calculation based IMRT validations ('software' calculation measurement); however, these do not satisfy the current requirements."

3D QA is More Clinically Relevant than 2D QA

Addressing Assertion 4

One other area of agreement is that 3D high resolution QA is by far the most sensitive and clinically useful QA method. Importantly, a 3D measurement is also the most efficient approach, in that all errors can be detected with one QA event.

Given the focus from Kry et al[1] on 3D volumetric QA, it is again curious that all 3D methods of QA other than Mobius3D™ from Varian Medical Systems® are excluded from consideration. Numerous other 3D products were developed prior to Mobius3D, and after, including Sun Nuclear's 3DVH™, SunCHECK™ Patient – PerFRACTION™ and DoseCHECK™, PTW's Octavius™, ScandiDos' Delta4 Anatomy, and IBA Dosimetry's COMPASS.

Well before Mobius3D and before the paper and presentations in question, volumetric QA had clearly been shown to be much more sensitive and specific in finding clinically relevant errors.[5,6,7,9] Dr. Kry's observations with Mobius3D are not treading new ground. **However, the work seems to imply a false choice – either use flawed, insensitive gamma criteria on an array or use 3D calculation-based QA from Mobius3D**. Rather than focusing on modeling errors through secondary calculations, a clinician can easily use a 3D array with stringent criteria and/or any of the existing **measurement-based 3D QA tools** that are widely available to assess both beam modeling and data transfer and deliverability errors.

Sun Nuclear pioneered the use of 3D volumetric patient QA in 2010 with the release of 3DVH which provided a revolutionary method to see, with 1mm resolution, the full volumetric impact of dose delivery on a patient's anatomy based on real measurement.

In 2015, Sun Nuclear further innovated in response to calls for an easier method to achieve 3D results. Sun Nuclear released PerFRACTION, which uses EPID-based measurements, with fully automated image retrieval and calculations, to allow automated and comprehensive 3D patient QA for both pre-treatment and in-vivo QA.

Sun Nuclear pioneered these technologies because we believed strongly that 3D QA was superior to planar QA, and that in-vivo QA could offer a solution to the persistent problem of patient setup and anatomy change errors.

Arrays Do Work

Addressing Assertions 2, 4

In the AAPM Point/Counterpoint talk[2], Dr. Kry makes the astonishing statement, "Arrays can't work." He implies that – regardless of criteria, QA methodology, or the types of errors looked for – arrays are not useful for QA. Interestingly, in the publication[1], he contradicts this statement by noting, *"This study did not evaluate or demonstrate that measurement based approaches cannot work, nor is it implied that measurements are not a critical component of beam model validation and the radiotherapy evaluation process. However, IMRT QA measurement methods, evaluated in aggregate based on current clinical practice, did not produce meaningful results in interpreting the suitability of a treatment plan."*

Dr. Kry singles out the ArcCHECK for criticism. Below is a subset of sensitivity studies performed on the ArcCHECK, which Dr. Kry excludes.

- Hussein et al[13], performed a sensitivity study on several measurement arrays and film. They introduced various errors and then predicted the gamma pass rate that the devices should produce. Each array's results were compared to the predicted pass rate. Errors introduced were: MLC positional errors of 1mm, 2mm, 5mm, Collimator rotation errors of 1°, 2°, 5°, and Hot/Cold spots (+/- 10% dose range with 0.5cm – 2cm dimensions). They concluded: *"Out of all the systems, ArcCHECK measurements exhibited the closest statistical agreement with the predicted gamma index."*[13]

- Leif et al[14], found that the ArcCHECK discovered a MLC mis-alignment that their MapCHECK had missed. *“ArcCheck was instrumental in uncovering those [MLC] discrepancies because the measured dose was spread onto a larger surface with less overlap, and cumulative discrepancies in low-dose regions added up to unacceptable level. Conclusion: Highly modulated large IMRT fields with sliding window tend to deliver a large number of monitor units with a potential of excessive dose to the patient of about 3%, due to MLC misalignment. This discrepancy can be better measured by ArcCheck.”*

- Yu et al[15], evaluated DLG adjustments for the HDMLC - *“Conclusion: ArcCHECK proved to be sensitive for detecting variations in dose distribution calculated with different DLG values. Based on the QA results, the original measured DLG value was adjusted to an optimal DLG value and IMRT QA results were improved, especially for highly modulated plans.”*

- Templeton et al[16], studied the sensitivity of ArcCHECK on Tomotherapy plans. The study unfortunately used insensitive criteria of 90%/3%/3mm/global normalization, but was still able to detect large errors. This proves once again both that the ArcCHECK can detect errors and that tighter criteria are necessary to find smaller errors. *“Errors were introduced in each of the couch speed, leaf open time, and gantry starting position in increasing magnitude while the resulting gamma passing rates were tabulated. The error size required to degrade the gamma passing rate to 90% or below was on average a 3% change in couch speed, 5° in gantry synchronization, or a 5 ms in leaf closing speed for a 3%/3 mm Van Dyk gamma analysis.”*

- Wang et al[17], tested the sensitivity of early ArcCHECK devices by introducing MLC errors, noting: *“For the intentionally introduced systematic leaf positioning errors of −0.5 and +1 mm, the detected leaf positioning errors was −0.46 ± 0.14 and 1.02 ± 0.26 mm, respectively. This demonstrated the submillimeter sensitivity of the proposed method.”*

Given the above studies[13-17], it is perplexing that the data reported in the presentations from Dr. Kry have a plan with an 8% dose difference that ArcCHECK failed to detect. Concerns of such a dose difference can be addressed by the publications above, a subset of more than 1,200 publications on ArcCHECK. However, it’s worth contextualizing the error. This error was caused by an inaccurate Dosimetric Leaf Gap (DLG) setting[3] – because DLG errors are each individually very small, appropriate criteria are crucial. This institution chose a 3%/3mm criteria with a 20% threshold; the distance to agreement (DTA) of 3mm easily masked this type of modeling error since each small error can be hidden with a small shift. This systematic modeling error would have been easily detected during the modeling process if appropriate criteria or a 3D QA method were used.[5] As in the Nelms et al[5] example previously discussed (Case 1), the institution would have seen this error by observing the nested profiles, by using a 3D method, or by using % Dose criteria only (as Hsieh et al[4] would suggest). The insensitive criteria caused the false negative result and inability to detect the error.

A review of AAPM Task Group 218[10,] a guideline outlining Patient Safety standards for pre-treatment measurements, addresses this issue directly. The TG-218 report promotes error detection by recommending:

- 3D measurements – noting that *“detector devices designed to measure VMAT beams such as ArcCHECK or Delta4 generally sample the entire beam area”*

- True Composite measurements with arrays that include robust angular correction methods
  - *“IMRT QA measurements should be performed using a TC (true composite) delivery method provided that the QA device has negligible angular dependence, or the angular dependence is accurately accounted for in the vendor software.”*

- Finally, that 2D Perpendicular Composite measurements should not be used – noting they don’t sample the entire volume and may mask clinically impactful errors (e.g., EPID composites or 2D array composites when rotated with the gantry).
  - *“IMRT QA measurements should not be performed using the PC (Perpendicular Composite) delivery method which is prone to masking delivery errors.”*
  - *“The PC method has the distinct disadvantage of potentially masking errors due to the summation.”*
  - *“Using the EPID to obtain an integrated image (2D composite image) is considered Perpendicular Composite.”*

It's important to note that in the paper[1] and presentations[2,3] in question, TG-218's recommendations are unheeded in all of the data tables (Tables II, III, and IV). In addition, there is very little detail on the criteria Dr. Kry has used to arrive at his results. For example, the Kry et al[1] Table II shows each device listed separately, but uses gamma criteria that is too insensitive (90%/3%/3mm) with no detail on critical analysis settings such as measurement uncertainty, threshold, normalization methods, or any separation of static-IMRT vs. VMAT. Table IV groups together all QA methods into one dataset, with ~80% of the methods being ones explicitly "not recommended" by TG-218 for VMAT (either single ion chamber, Perpendicular Composite, or arrays without angular dependence corrections). and 13% of the methods being a single ion chamber measurement. The gamma criteria of the group as a whole is then considered at various levels, but again without thresholding, normalization, or measurement uncertainty being addressed. The paper notes in the Methods section that QA results falling below 90% passing rates with tighter criteria than 3%/3mm were excluded from the study.[1] This appears to skew the research by removing IMRT QA that revealed errors.

Nowhere in the study can a reader look up sensitivity and specificity on a given device (such as the ArcCHECK) using the TG-218 criteria of 3%/2mm (or tighter for SRS/SBRT).

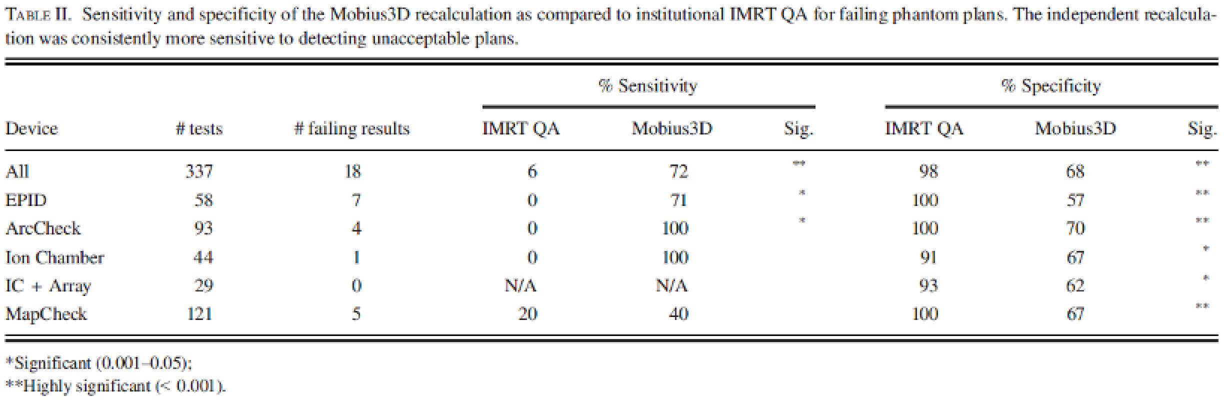


TABLE II. Sensitivity and specificity of the Mobius3D recalculation as compared to institutional IMRT QA for failing phantom plans. The independent recalculation was consistently more sensitive to detecting unacceptable plans.

			% Sensitivity			% Specificity		
Device	# tests	# failing results	IMRT QA	Mobius3D	Sig.	IMRT QA	Mobius3D	Sig.
All	337	18	6	72	**	98	68	**
EPID	58	7	0	71	*	100	57	**
ArcCheck	93	4	0	100	*	100	70	**
Ion Chamber	44	1	0	100		91	67	*
IC + Array	29	0	N/A	N/A		93	62	*
MapCheck	121	5	20	40		100	67	**

*Significant (0.001–0.05);
**Highly significant (< 0.001).

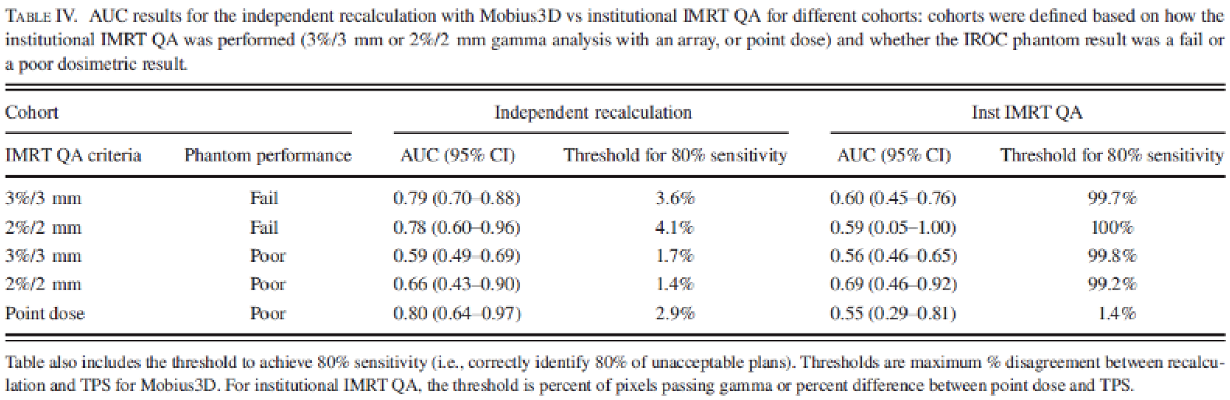


TABLE IV. AUC results for the independent recalculation with Mobius3D vs institutional IMRT QA for different cohorts: cohorts were defined based on how the institutional IMRT QA was performed (3%/3 mm or 2%/2 mm gamma analysis with an array, or point dose) and whether the IROC phantom result was a fail or a poor dosimetric result.

Cohort		Independent recalculation		Inst IMRT QA	
IMRT QA criteria	Phantom performance	AUC (95% CI)	Threshold for 80% sensitivity	AUC (95% CI)	Threshold for 80% sensitivity
3%/3 mm	Fail	0.79 (0.70–0.88)	3.6%	0.60 (0.45–0.76)	99.7%
2%/2 mm	Fail	0.78 (0.60–0.96)	4.1%	0.59 (0.05–1.00)	100%
3%/3 mm	Poor	0.59 (0.49–0.69)	1.7%	0.56 (0.46–0.65)	99.8%
2%/2 mm	Poor	0.66 (0.43–0.90)	1.4%	0.69 (0.46–0.92)	99.2%
Point dose	Poor	0.80 (0.64–0.97)	2.9%	0.55 (0.29–0.81)	1.4%

Table also includes the threshold to achieve 80% sensitivity (i.e., correctly identify 80% of unacceptable plans). Thresholds are maximum % disagreement between recalculation and TPS for Mobius3D. For institutional IMRT QA, the threshold is percent of pixels passing gamma or percent difference between point dose and TPS.

The audience is left with the impression that ArcCHECK would not detect errors even with 2%/2mm criteria, but this is never proven with data. Furthermore, sensitive methods such as a full 3D measurement-based QA solution (e.g., 3DVH or PerFRACTION) were excluded from the analysis.

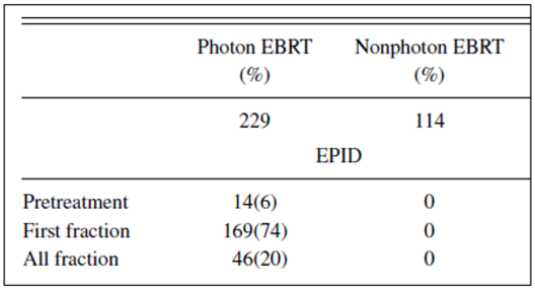
Automated In-Vivo QA is the Most Effective Use of Physics QA Time

While pre-treatment QA is important, and does occasionally catch catastrophic errors, the proportion of errors detected through pre-treatment QA is small.[18] Conversely, patient setup errors and anatomy changes are a frequent and (historically) difficult to detect source of treatment errors. The advent of IMRT/VMAT and SBRT only intensified the problem since these precise modalities depend on accurate and reproducible patient alignment.

Over the past decade there has been increasing interest in EPID-based in-vivo QA. Some interest was a result of catastrophic radiation errors leading some countries, such as France, to mandate daily in-vivo measurements and prompting others, such as the United Kingdom's NHS, to provide funding for in-vivo measurements to all radiation clinics. Advances in automation, EPID stability and positioning, and in CBCT image quality have also enabled adoption of in-vivo QA.

There have been many studies[18,19,20,21,22] demonstrating the effectiveness of in-vivo QA. Minjheer[23] provides a very effective and recent summary of this topic. In a seminal study[18], Bojechko et al reviewed clinical errors and "near misses" for two and a half years. There were 343 "potentially severe" and "critical" errors and near misses during this period. Lastly, the study asked an important question – "Where could these errors be detected in the QA process?" The answer: 74% of the errors and "near misses" could be detected by performing one additional QA task – a first fraction in-vivo QA check.

It is remarkable that detectability of serious errors improved from 6% to 80% with one automated first fraction in-vivo QA. In conclusion, the publication states that, *"The most effective EPID-based dosimetry verification is in vivo measurements during the first fraction."*



	Photon EBRT (%)	Nonphoton EBRT (%)
	229	114
	EPID	
Pretreatment	14(6)	0
First fraction	169(74)	0
All fraction	46(20)	0

An overly broad assertion with respect to in-vivo QA is made during the AAPM Point/Counterpoint session. Dr. Kry mentions in-vivo QA theoretically could be useful in detecting patient setup errors and anatomy errors, but that PerFRACTION has been proven to not even detect a 2cm shift in a patient.[2]

The example used is from the publication by Hsieh et al[4] from UC Davis. Dr. Kry mistakenly states that the authors did not detect a 2cm shift. In fact, the study stated, "a 5mm left shift was undetected by gamma analysis, and up to a 2cm shift had to be introduced for the average gamma pass rate of 3%/3mm to fall below a 95% pass rate criteria.... Because gamma proved insensitive to the small shifts to be studied in this work, the subsequent analysis was focused on the more sensitive DTA metric." The authors concluded PerFRACTION was able to detect 5mm, 3mm, and 1mm errors for SBRT/SRS when appropriate and sensitive criteria were used.

Given the intent of the Hsieh et al publication to identify appropriate gamma criteria and passing rates, the authors concluded by stating: **"PerFRACTION 2D mode successfully detected setup errors outside the systematic error tolerance for SRS, IMRT and 3D when an appropriate analysis metric and pass/fail criteria was implemented. Our data confirms that percent difference may be more sensitive in detecting plan failure than gamma analysis."[4]**

Another compelling finding from the Hsieh et al study[4] is: "PerFRACTION 2D mode successfully detected setup errors outside our systematic error tolerance for IMRT (3mm shift) and SRS (1mm) when an appropriate analysis metric and pass/fail criteria was implemented", providing recommended tolerance criteria for both non-SRS and SRS plans.

Desired Shift Detection Level	1 mm	3 mm	5 mm	5° yaw
3% Difference	Not advised	97%	96%	73%
1% Difference	89%	63%	62%	37%

Table 3: Matrix listing recommended clinical parameter settings for detecting shifts using the 2D EPID dosimetry function in PerFRACTION. The columns indicate the desired shift detection level, while the rows list the % Difference setting in PerFRACTION. Each cell indicates which pass rate tolerance setting would be required to flag at least on field for each of the five cadaver heads and the solid water phantom as failing.[4]

The Hsieh et al publication[4], which implemented an early 2D relative dose version of PerFRACTION, was used by Dr. Kry to dismiss all in-vivo QA as insensitive to patient errors.[2] In reality, this publication and several others[19,20,21,22] show PerFRACTION is sensitive to the detection of all manner of patient-related errors.

In the most comprehensive study[20], Bossuyt et al note: “Errors were caught such as: weight loss at start of treatment, problem with bellyboard, errors in planning, problems at simulation with 4DCT artifacts or contrast agents in bowel, pleural effusions cleared up by the time of treatment, poor breathing for gated breast patients.” The study concludes: “Absolute verification for transit in-vivo dosimetry enhanced detectable errors.” Finally Bossuyt, et al, note: “The number of plan adjustments increased, showing the increased confidence in the system as a base for adaptive planning.”

In this study from Iridium Kankernetwerk (in press)[20], **over 56,000 fractions were examined and over 4,000 clinically meaningful errors were detected.** This publication lists the gamma criteria per body site and decision trees for in-vivo QA workflows. The contents of Iridium Kankernetwerk’s study are available in a 2020 ESTRO abstract, and on the Sun Nuclear website via an on-demand webinar. The full-length manuscript will be published soon. This will be extremely helpful to the radiation oncology community by offering specific guidance on how to implement in-vivo QA efficiently.

Finally, in Zhuang et al[19] it was noted that PerFRACTION was sensitive enough to detect the following errors: 0.2mm Jaw errors, 0.4mm MLC errors, 0.2% output, 0.5 collimator rotation, 0.2mm couch shift, and incorrect rail positions, if appropriate criteria were used.

These and many other in-vivo studies are readily available; one wonders why during the AAPM 2020 Point/Counterpoint session Dr. Kry chose to quote, without the full context, the findings of a single study[4] to dismiss in-vivo QA.

Not All 3D QA is the Same

Addressing Assertions 2, 3, 4

It is also worthwhile to examine the accuracy of the algorithm used for secondary checks and beam model QA. In order to detect the beam modeling errors Dr. Kry focuses his research on[1,2,3], the 3D algorithm and beam models must be equal or superior to the TPS algorithm. There are also concerns with log-file only 3D QA – these are addressed in multiple publications.[27,28,29]. Assuming that a 3D secondary calculation is the best method to detect beam modeling errors, any independent 3D calculation should detect the beam modeling errors referenced in Dr. Kry’s work. This leaves the physicist with a choice of several commercially available 3D dose calculators – Sun Nuclear's DoseCHECK, Varian Medical Systems® Mobius3D™, RadCalc® 3D from LAP, SciMoCa™ from IBA Dosimetry to name a few. For the purposes of this discussion, we’ll narrow the scope to reviewing a series of publications looking at Varian Medical Systems® Mobius3D™ and Sun Nuclear's DoseCHECK/PerFRACTION dose calculator (SDC).[24,25,26]

SDC has been shown to have superior beam models with respect to small fields, MLC-modeling, and heterogeneity in three separate studies. On heterogeneity, Nakaguchi et al[26], concluded: *“For the planning of the whole neck, the differences in the M3D and the TPS dose profiles led to the inability of the former to calculate a complex dose distribution for VMAT” and that “the M3D system appears to be unsuitable for highly accurate dose calculations in anatomical regions filled with air.” And finally, that “the M3D dose measurements differed by 5-10% in the lung and bone regions.”* Hillman et al[24] found that for small field outputs of 0.5cm^2 fields, Mobius3D showed a >17% output discrepancy, while DoseCHECK was within 1.5% of ion chamber measurements. Finally, Kim et al[25] concluded: *“It was demonstrated that Mobius3D has dose calculation uncertainties for small fields and MLC tongue-and-groove design is not adequately taken into consideration in Mobius3D.”* Kim, also noted that, *“Unfortunately, Varian Medical Systems® Mobius3D™ users are not allowed to customize parameters related to the fluence model, except for the DLG correction factor.”* Conversely, in DoseCHECK there are four different MLC-related factors that can be adjusted in the model process – Radius of curvature, tongue-and-groove thickness, transmission, and leaf gap/offset.

Conclusion

A flawed comparison has been made between Varian Medical Systems® Mobius3D™ and array-based and single ion chamber QA.[1,2,3] It is an argument seemingly only using handpicked examples to support a pre-determined answer, and failing to acknowledge well-established QA methods and standards documented in numerous publications.[4-10]

3D measurement-based QA (such as SunCHECK Patient - PerFRACTION or ArcCHECK and 3DVH) are superior to calculation or log-based QA by achieving truly independent, high resolution, and clinically useful QA. Array-based QA is sensitive to errors when used with appropriate and stringent criteria[13-17] and is likewise superior to calculation or log-based QA, in that it is independent of the linac and can detect all modes of pre-treatment failure. Importantly, a 3D measurement is the most efficient approach in that all errors can be detected with one Quality Assurance event. This is the most efficient way to catch errors as opposed to attempting to determine which errors are missed by an incomplete QA method.

Finally, if in-vivo QA such as SunCHECK Patient - PerFRACTION is used, it is superior to calculation-based QA by enabling the clinician to see, for the first time, the dosimetric impact of day to day changes in setup and anatomy.[19,20,21,22] The full automation of in-vivo QA also means little additional effort is needed to make this important gain in patient safety.

Sun Nuclear has pioneered patient QA, leading to numerous advances in radiation therapy patient safety. We take our role seriously as an independent monitor of quality and efficacy for the industry and the patients we serve.

Sun Nuclear employs over 50 physicists working throughout our organization to ensure our products meet the highest standards. Yet our ultimate measure of value is determined by our users, their feedback, and their decision to continue to use our solutions. Well over 60% of global cancer treatment centers choose Sun Nuclear solutions and a majority of those centers rely on our proven and effective 2D/3D arrays and 3D analysis software for their IMRT/VMAT QA.

Total MapCHECK, MapCHECK 2, MapCHECK 3, SRS MapCHECK publications:

3,390

Total ArcCHECK publications:

1,220

Varian Medical Systems® is a registered trademark of Varian Medical Systems, Inc. Sun Nuclear Corporation is not affiliated with Varian Medical Systems, Inc., PTW, IBA Dosimetry, LAP, or ScandiDos.

References:

1. S.F. Kry, M.C. Glenn, C.B. Peterson, et al. ***Independent recalculation outperforms traditional measurement-based IMRT QA methods in detecting unacceptable plans***, Med. Phys. 46 (8), August 2019: 3700-3708
2. ***Session Title: A Point/Counterpoint on Current and Future Directions for Patient Specific QA***, and ***Handout Title: Current IMRT QA Pros and Cons*** (https://amos3.aapm.org/abstracts/pdf/155-50587-1531640-157243.pdf), 2020 Joint AAPM I COMP Virtual Meeting, MO-C-TRACK 3-C
3. Varian Medical Systems® Webinar: ***The Failures and Needs of IMRT QA,*** 2020/08/12
4. E.S. Hsieh, K.S. Hansen, M.S. Kent, et al. ***Can a commercially available EPID dosimetry system detect small daily patient setup errors for cranial IMRT/SRS?***, Practical Radiation Oncology (2016)
5. B. Nelms, G. Jarry, M. Lemire, J. Lowden, et al. ***Evaluating IMRT and VMAT dose accuracy: Practical examples of failure to detect systematic errors when applying a commonly used metric and action levels***, Med. Phys. 40(11), November 2013:111722
6. H. Zhen, B. Nelms, W. Tome. ***Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA,*** Med Phys, 2011 Oct;38(10):5477-89
7. M. Stasi, S Bresciani, A. Miranti, et al. ***Pretreatment patient-specific IMRT quality assurance: A correlation study between gamma index and patient clinical dose volume histogram***, Medical Physics 39(12), December 2012: 7626-34
8. B. Nelms, H. Zhen, W. Tome. ***Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors,*** Med Phys. February 2011;38(2):1037-44.
9. M. Chan, et al. ***Using a novel dose QA Tool to quantify the impact of systematic errors otherwise undetected by conventional QA methods:*** Clinical head and neck case studies, Technology in Cancer Research and Treatment, 13 (1), 57-67 (2014)
10. M. Miften, A. Olch, D. Mihailidis, J. Moran, et al. ***Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218,*** Med. Phys. 45 (4), April 2018
11. A. Mans, et al. ***Catching errors with in vivo EPID dosimetry,*** Med Phys. 2010 Jun; 37(6):2638-44
12. Source: https://www.ncbi.nlm.nih.gov/pubmed/29443390
13. M. Hussein, P. Rowshanfarzad, M. A. Ebert, et al. ***A comparison of the gamma index analysis in various commercial IMRT/VMAT QA systems,*** Radiotherapy and Oncology 109 (2013) 370-376
14. E. Lief, P. Jeffe, J. Restrepo, A. Cheuk, ***Excessive IMRT Patient Dose Due to the MLC Misalignment and Its Determination Using Archeck,*** AAPM 2017, SU-I-GPD-T-272
15. S. Yu, D. Rosenzweig, T. Barry, M. Pacella, ***Optimization of Dosimetric Leaf Gapvalue Using ArcCHECK and Highly Modulated VMAT Plans,*** AAPM 2017, WE-RAM2-GePD-TT-04
16. A. K. Templeton, J. C. H. Chu, J. V. Turian, ***The sensitivity of ArcCHECK-based gamma analysis to manufactured errors in helical tomotherapy radiation delivery,*** J Appl Clin Med Phys, 2015 Jan 8;16(1):4814
17. Q. Wang, et al. ***A novel method for routine quality assurance of volumetric-modulated arc therapy,*** Med. Phys. 40 (10) (2013)
18. C. Bojechko, et al. ***A quantification of the effectiveness of EPID dosimetry and software-based plan verification systems in detecting incidents in radiotherapy,*** Med Phys 42(9), Sept 2015
19. A. H. Zhuang, A. Olch. ***Sensitivity study of an automated system for daily patient QA using EPID exit dose images,*** J Appl Clin Med Phys. 2018 May;19(3):114-124
20. E. Bossuyt, R. Weytjens, D. Nevens, S. De Vos, D. Verellen. ***Results of 2 years of automated pre-treatment and absolute transit in vivo dosimetry,*** ESTRO 2020, PH-0050; Related manuscript in press: PHIRO-D-20-00026R3, ESTRO's phiRO journal
21. N. Jornet, I. Valverde, N. Espinosa, et al. ***EPID 2D transit In Vivo Dosimetry: Can relevant anatomy and positioning differences be detected?,*** ESTRO 2020, PO-1365
22. N. Kadoya, T. Matsumoto, K. Sato, et al. ***Evaluation of the feasibility of EPID-based in vivo dosimetry system for prostate cancer patients,*** ESTRO 2020 PO-1646
23. B Mijnheer. ***EPIDs and QA of Advanced Treatments,*** 2019 J. Phys.: Conf. Ser. 1305 012061
24. Y. Hillman, J. Kim, I. Chetty, N. Wen. ***Refinement of MLC modeling improves commercial QA dosimetry system for SRS and SBRT patient-specific QA,*** Med. Phys. 45 (4), April 2018
25. J. Kim, M. C. Han, E. Lee, K. Park, et al. ***Detailed evaluation of Mobius3D dose calculation accuracy for volumetric modulated arc therapy,*** Physica Medica 74 (2020) 125–132
26. Y. Nakaguchi, Y. Nakamura, Y. Yotsuji. ***Validation of secondary dose calculation system with manufacturer provided reference beam data using heterogeneous phantoms,*** Radiological Physics and Technology, Japanese Society of Radiological Physics and Technology [25 Jan 2019, 12(1):126-135]
27. M. Savvas, et al. ***Error detection between Log file and Measurement based VMAT QA,*** 2016 QA & Dosimetry Symposium presentation
28. B. Neal, M. Ahmed, K. Kathuria, T. Watkins, et al. ***A clinically observed discrepancy between image-based and log-based MLC positions,*** Med. Phys. 43, 2933 (2016)
29. A. Agnew, C. E. Agnew, M. W. D. Grattan, et al. ***Monitoring daily MLC positional errors using trajectory log files and EPID measurements for IMRT and VMAT deliveries***, Phys. Med. Biol. 59 (2014) N49–N63

All data used is best available at time of publication. Data is subject to change without notice. ©2021 Sun Nuclear Corporation. All Rights Reserved.