2017 QADS Head & Neck Plan Study

A Study of Plan Quality and QA over a Population of Planners, Planning Systems, and Modalities

Ben Nelms, Ph.D.[1] & Vladimir Feygelman, Ph.D.[2]

[1] Canis Lupus LLC

[2] Moffitt Cancer Center

Outline

- Introduction
  - Goals
  - Brief history of the plan study program
  - Medical Physics Practice Guidelines 5.a (MPPG 5.a)
- Methods
  - Experimental design and discussion of variables
  - Plan scoring, QA metrics, and method of data collection
- Results
  - Analyses, comparisons, and studies of variation
  - Best practices (planning & physics)
- Conclusions

Introduction

Goals, History, etc.

Goals

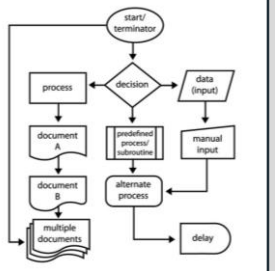
Assess Performance of:
TPS, Modality, & Planner vs. Dynamic Population Data

Assess Performance of:
TPS Dose Calculation Accuracy & Delivery

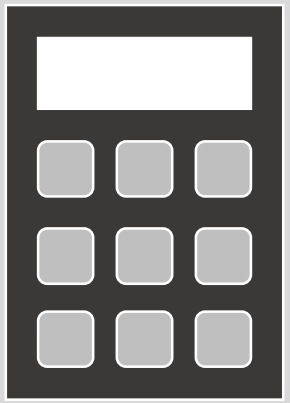
Download & Import Test Data → Create Treatment Plan ⇄ Score Plan Quality → High Quality Test Plan → True Composite Dose QA

INDUSTRY STANDARD: Process, Test Data, Scoring Methods and Benchmarks

Processes



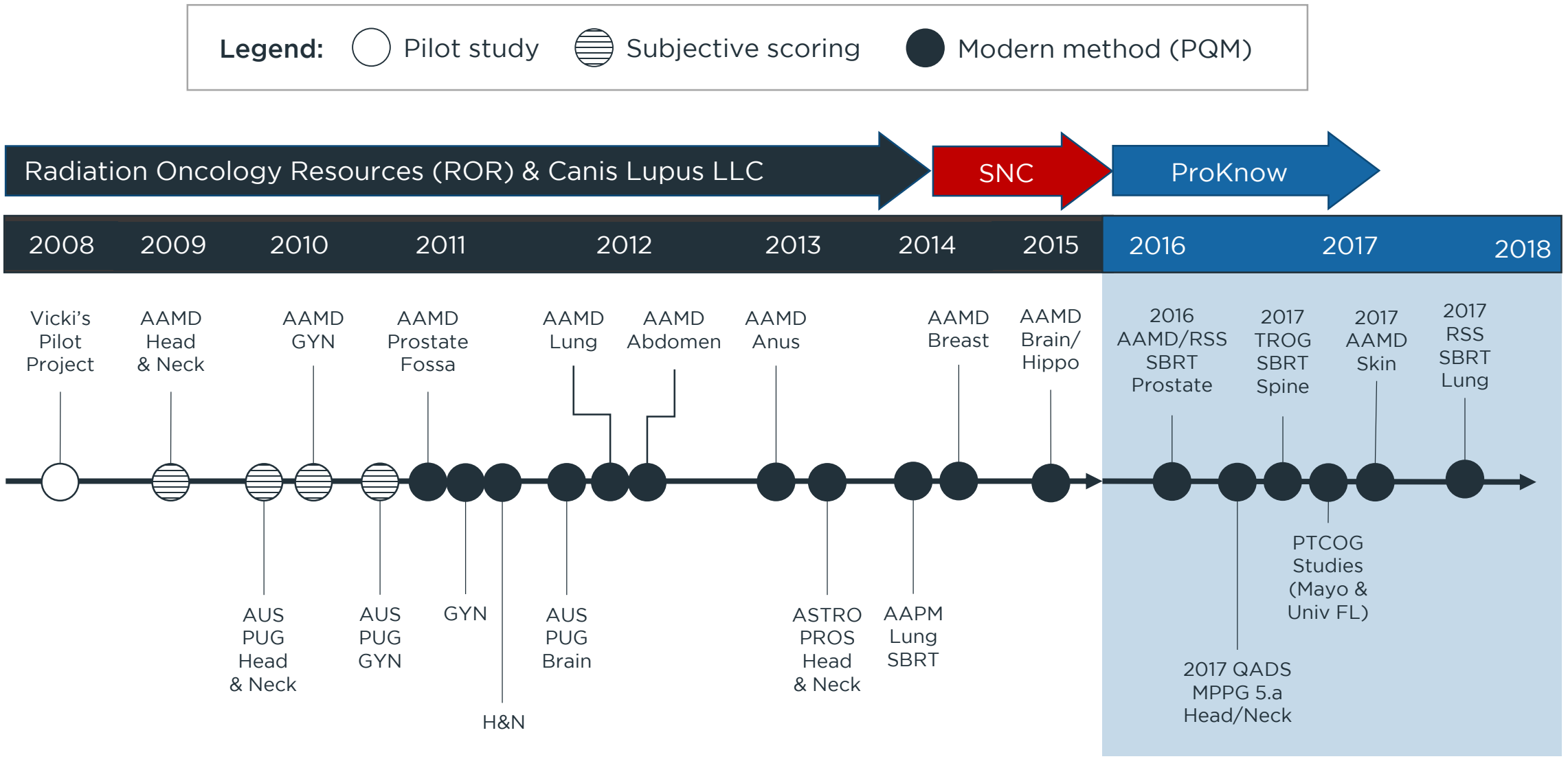
Standard
Clinical Datasets
(Images + Structures)



Goals

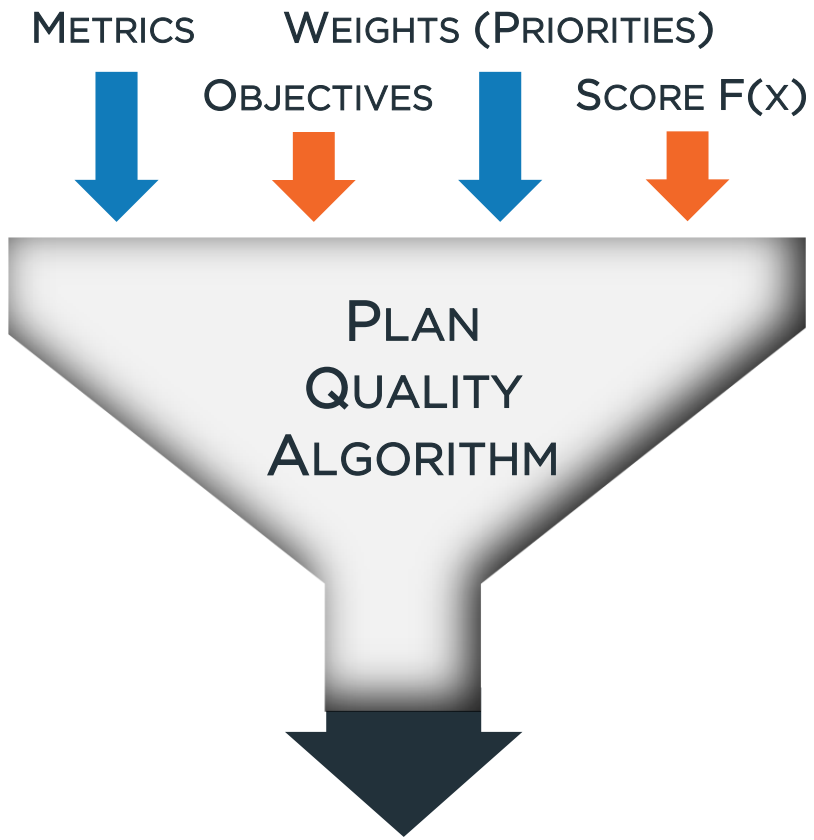
- **Benchmarking**
  - Achievable levels of quality for treatment plan quality
  - Achievable levels of quality for dose accuracy
- **Objective, Comparative Analyses**
  - Modality vs. Modality, TPS vs. TPS, etc.
  - Study of delivery efficiency and monitor unit usage
- **Study of Variation**
  - Potential causes and ways to improve
- **Sharing of Best Practices**
  - Identify and interview “high performers”
  - Share their successful tips and techniques (planning & physics)

History of Modern Plan Studies



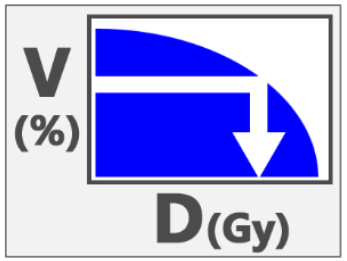
Modern Scoring Method

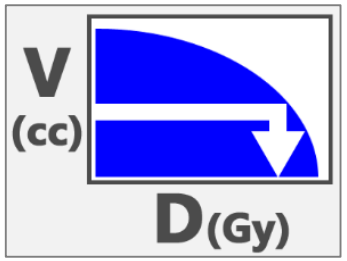
- **Technology**. Developed in 2011 as "Quality Reports," became Sun Nuclear's product called PlanIQ™. General method licensed for use for ProKnow Plan Studies
- **Identify critical metrics**. Dose, DVH, and formulaic metrics selected from a large library of options.
- **Define each metric's objective score f(x)**. For each metric, capture what defines success, i.e. specify priority along with: 1) minimally required result, 2) ideal result, and 3) variable scoring in between.
- **Define each metric's weight**. Assign point value (i.e. weight) for each metric, which scales the score f(x) ordinate values.

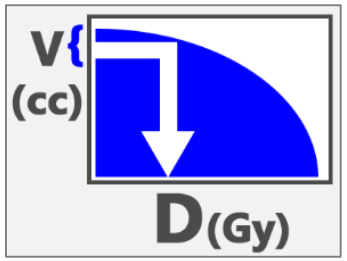


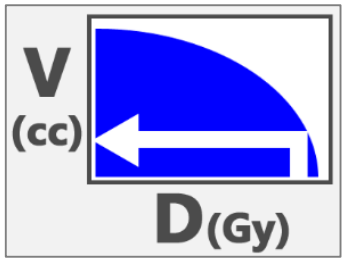
Σ PER METRIC SCORES = COMPOSITE PLAN SCORE

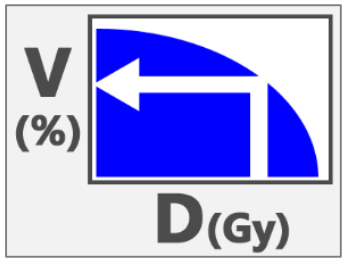
Library of Available Metrics

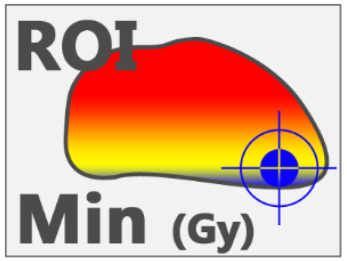


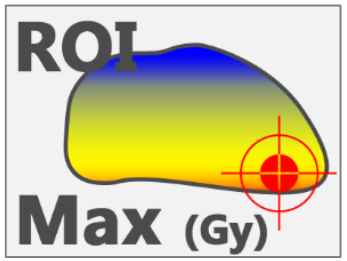


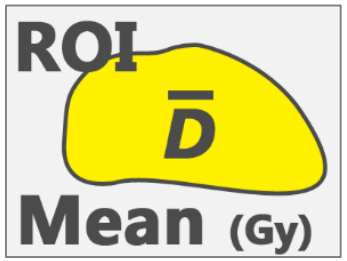


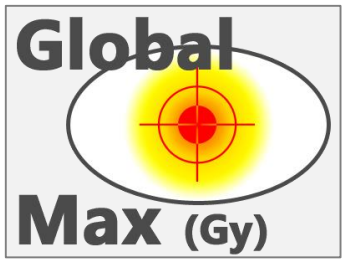


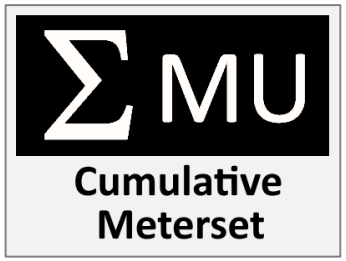


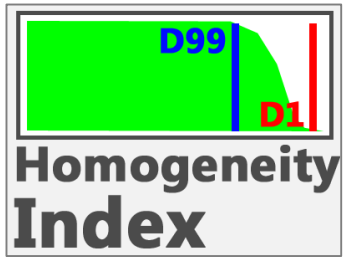


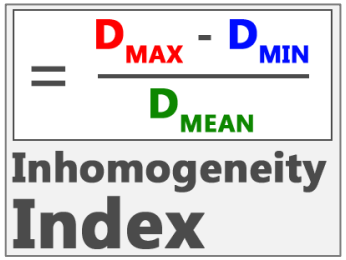










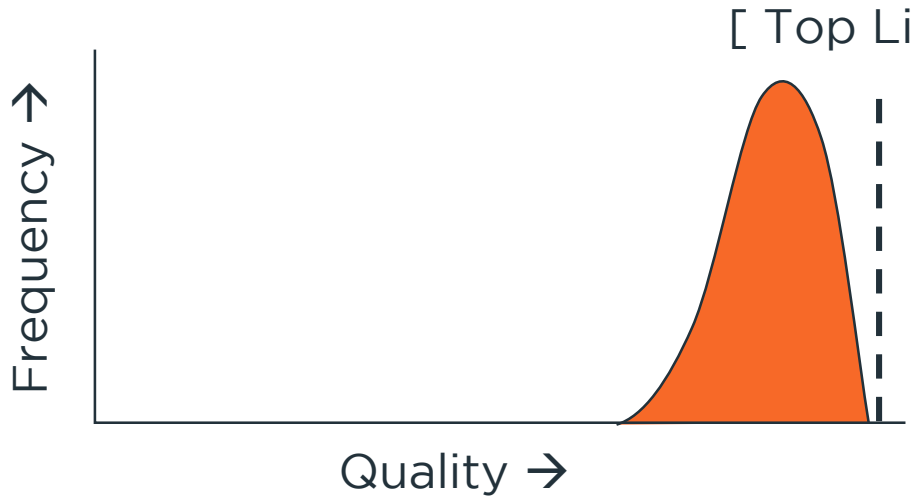


Modern Scoring Method

- 2012 Publication was the first study using the modern scoring method (prostate + nodes, in conjunction with the 2011 AAMD meeting)
- This has become a key paper cited by many new studies, in particular for studies of software auto-planning vs. manual planning

Statistical Process Control & Study of Variation

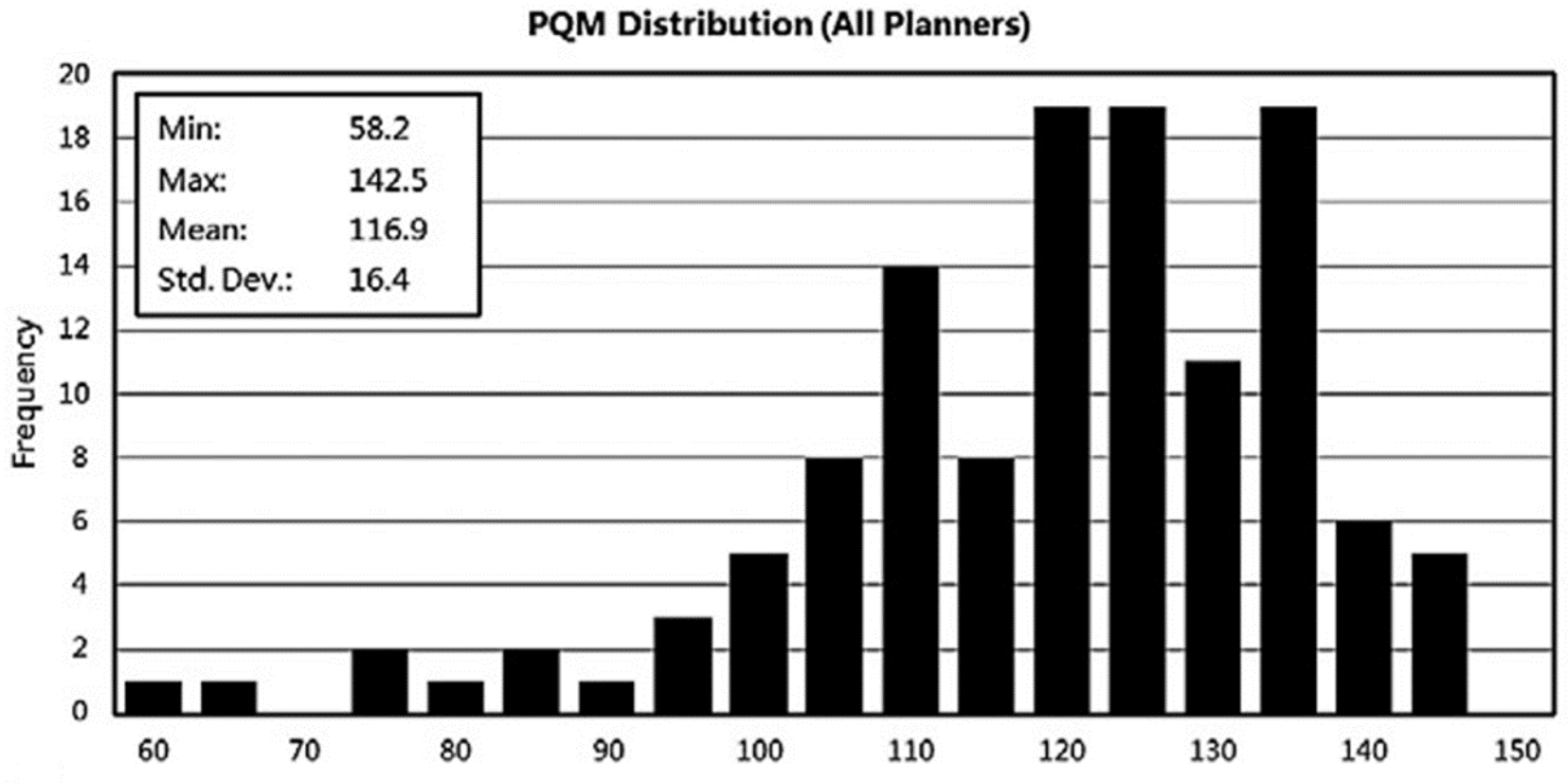
- High variation
- Average quality is low
- Lots of low quality items
- Few high quality items



- Lower variation
- Average quality is higher
- Fewer low quality items
- More high quality items

Key Findings from the 2012 Paper

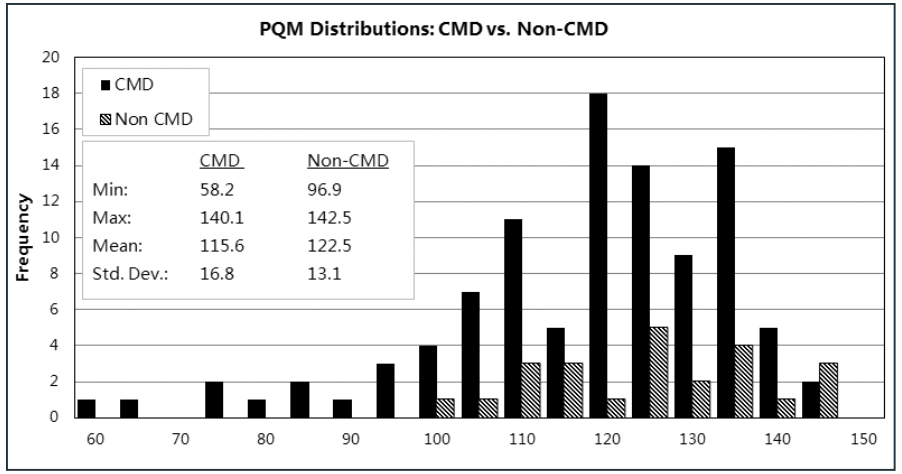
- Despite controlled inputs (CT and structures) and well-defined objectives (Plan Quality Algorithm), there was **very high variability in plan quality**.

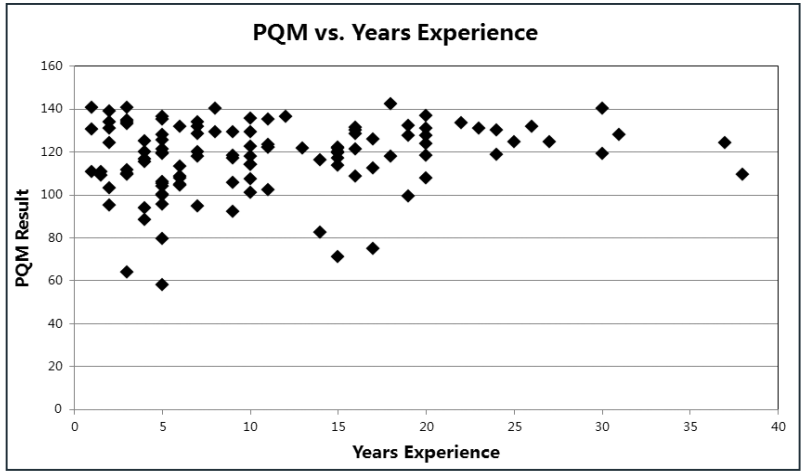


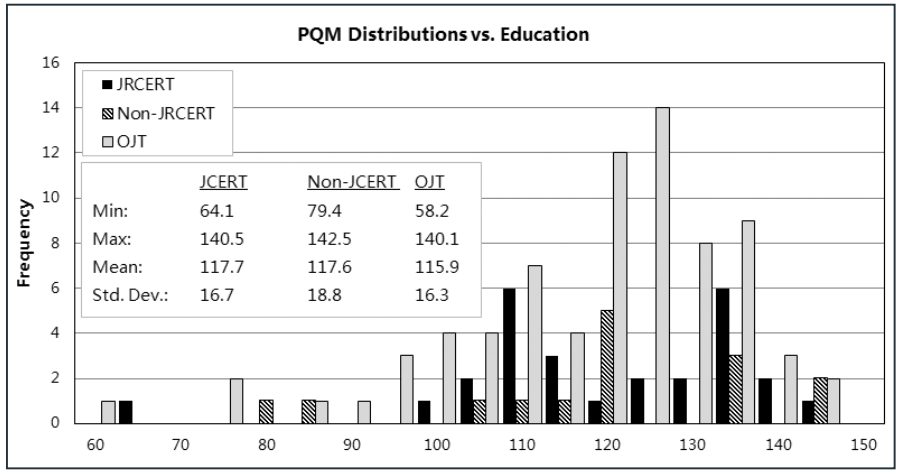
Key Findings from the 2012 Paper

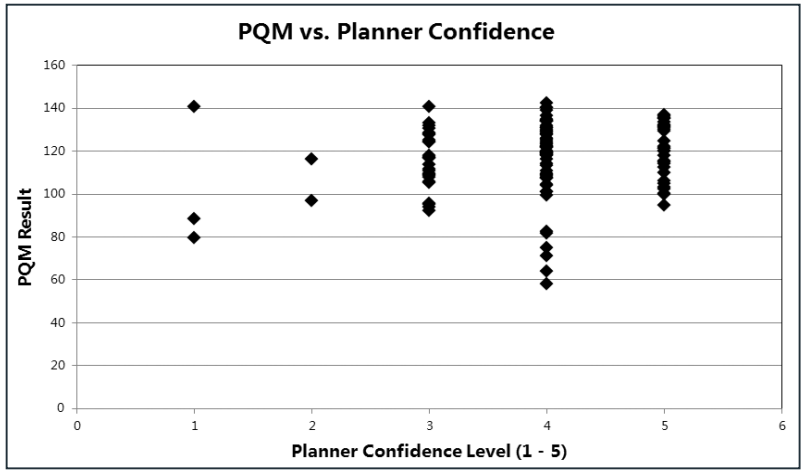
NO correlation with:

- Certification
- Education level
- Experience
- Confidence



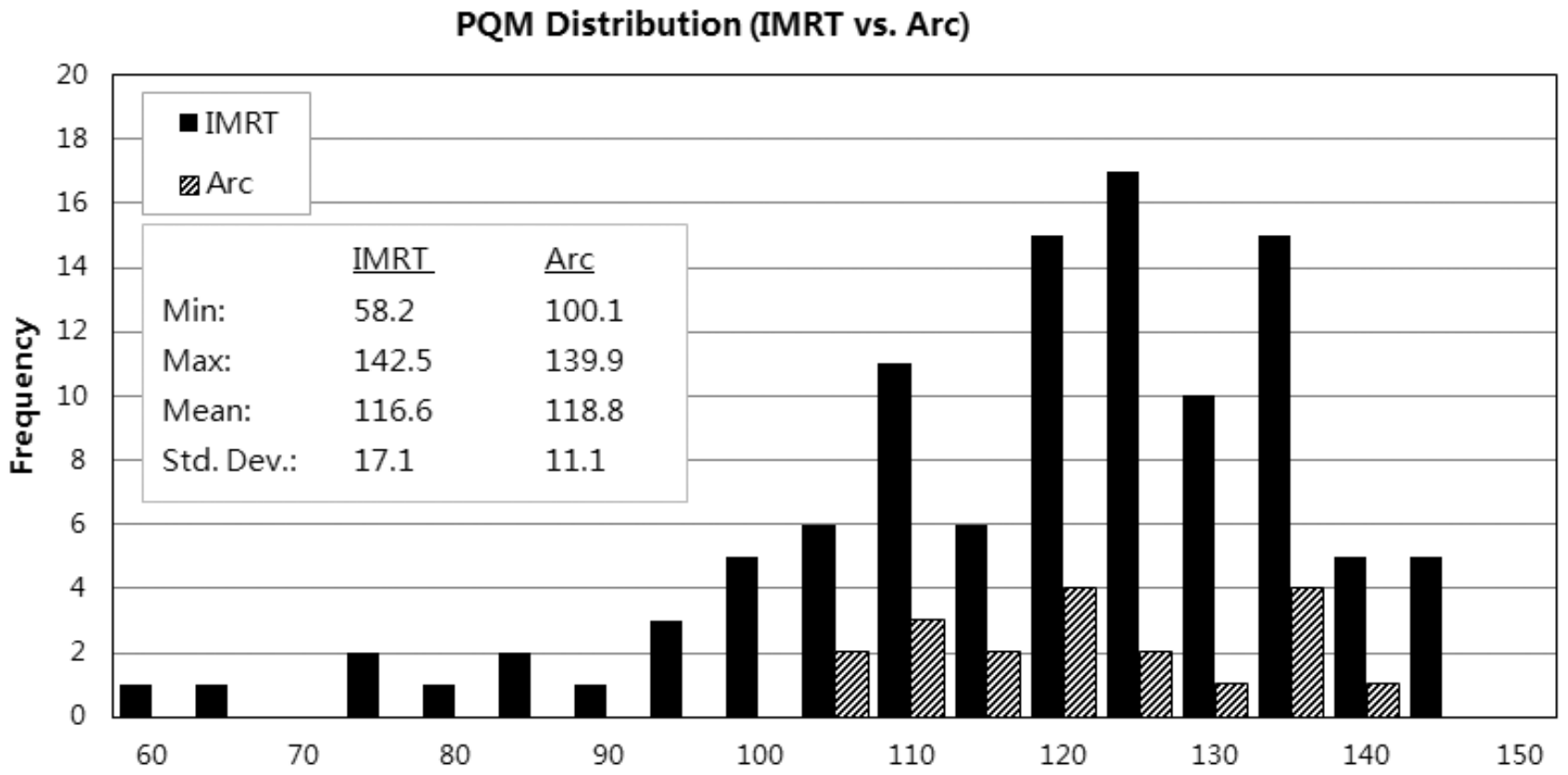






Key Findings from the 2012 Paper

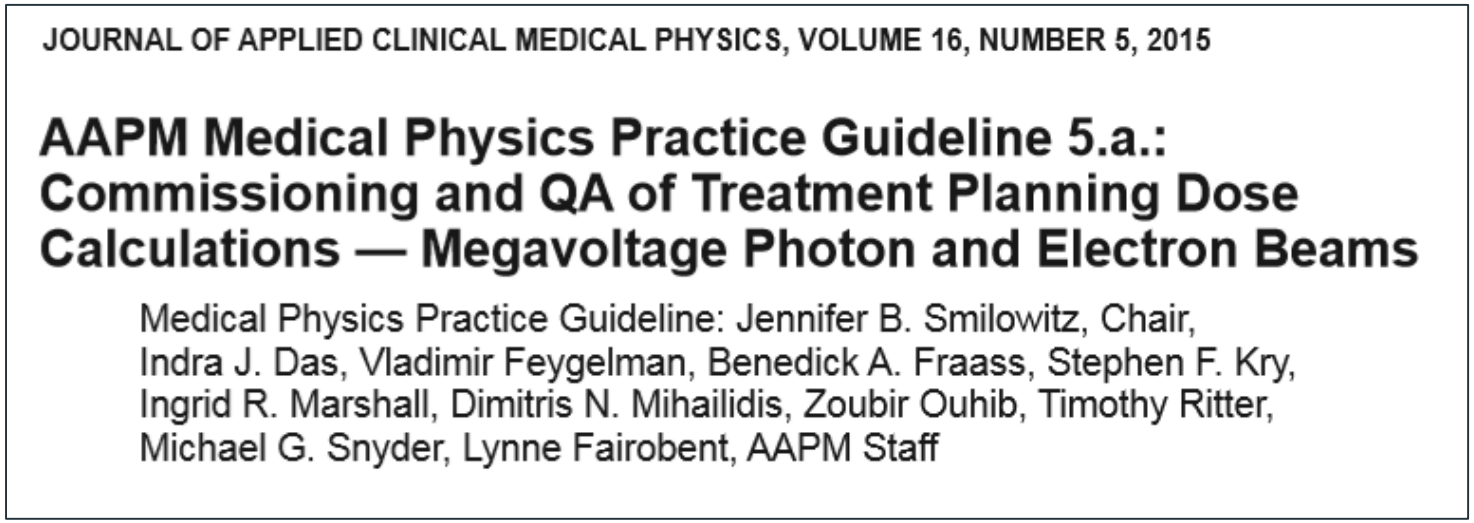
- VMAT was not statistically “better” than IMRT, though it was somewhat less variable.



Key Findings from the 2012 Paper

- **No significant dependence on TPS**
  - All TPS models show large variation in plan quality.
  - Over many studies over the years, there are reproducible trends in terms of benchmarks such as max scores or 90th percentile, i.e. top 10%.
- **No dependence on plan "complexity"**
  - As evidenced by total monitor units or treatment time
  - Some very efficient/low MU plans score very high while some very inefficient/high MU plans score low
- **Quality is determined by Planner Skill**
  - This first (and all subsequent studies so far) assert that plan quality depends less on modality or technology, and more on planner skill.

MPPG 5.a (TG244)



JOURNAL OF APPLIED CLINICAL MEDICAL PHYSICS, VOLUME 16, NUMBER 5, 2015

AAPM Medical Physics Practice Guideline 5.a.: Commissioning and QA of Treatment Planning Dose Calculations — Megavoltage Photon and Electron Beams

Medical Physics Practice Guideline: Jennifer B. Smilowitz, Chair, Indra J. Das, Vladimir Feygelman, Benedick A. Fraass, Stephen F. Kry, Ingrid R. Marshall, Dimitris N. Mihailidis, Zoubir Ouhib, Timothy Ritter, Michael G. Snyder, Lynne Fairobent, AAPM Staff

- **Goal**
  - To provide an overview of the minimum requirements for TPS dose calculation algorithm commissioning (data acquisition, modeling, and verification) and QA in a clinical setting.
- **Six standard patient datasets are provided**
  - Five for studying plan quality and dose accuracy, and one for studying heterogeneities

Methods

Experimental Design

Scoring Methods

Data Collection

Control Variables

- **Patient anatomy** (CT images)
- **Contoured targets and organs** (RT Structure Set)
- **Objective planning goals** (plan scoring algorithm)
- Common **scoring software**, to eliminate inter-TPS variation in DVH calculation methods
  - Nelms BE, Stambaugh C, Hunt D, Tonner B, Zhang G, and Feygelman V. "Methods, software and datasets to verify DVH calculations against analytical values: Twenty Years Late(r)," **Med Phys.** 2015 Aug; 42(8).

Control Variables (cont.)

- Modern **dose calculation algorithm** (superposition or better)
- Minimum requirements for **dose grid resolution** (≤ 3.0 mm) and **size** (covering all scored structures)
- Realistic and **practical delivery time**
  - Estimated “beam on” time is calculated and displayed on each plan quality scorecard

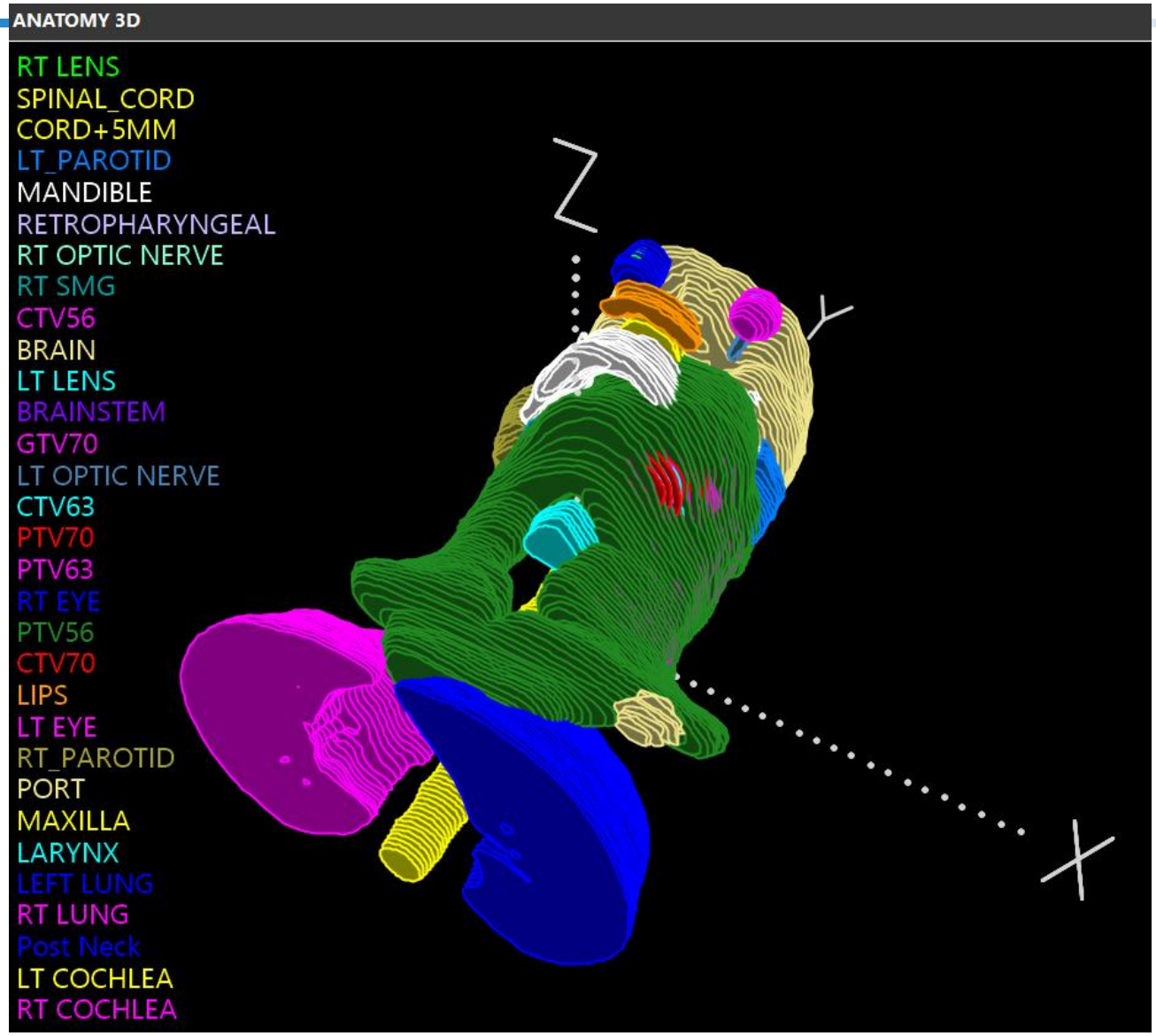
Independent Variables

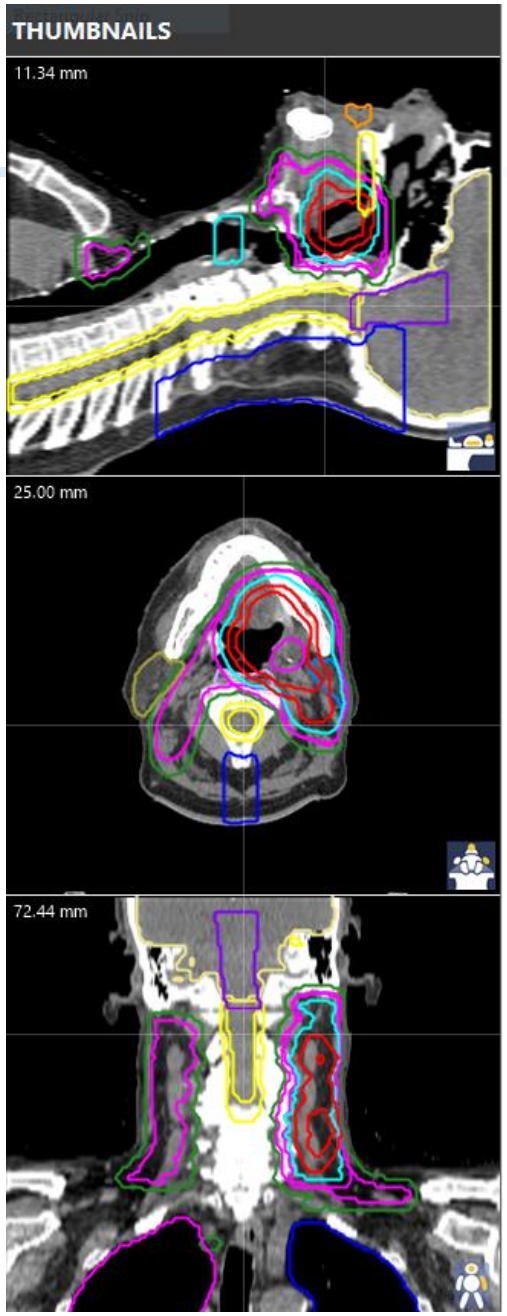
- **Treatment Planning System**
- **Treatment Modality** (VMAT, IMRT, protons, helical Tomotherapy, robotic, etc.)
- **Human Skill**
  - Relative talents of treatment planner(s) & physicist(s)
- **Planning Techniques**
- **Beam Energy**

Dependent Variables

- **Plan Quality Scores** (Composite, max 150)
- **Per Metric Scores**
- **Estimated Treatment Times**
- **Pretreatment Dose QA** (Calc vs. Meas)

Patient Dataset





Plan Quality Algorithm

- 21 Scored Metrics
  - 150 points total
  - Target coverage accounted for 94 of the 150 points
  - Conformation and sparing of organs-at-risk: 56 points
  - 14 of 21 metrics use advanced, non-linear scoring
  - 2 “hard constraint” OAR objectives
- 2 unscored metrics
  - Estimated Treatment Time (min)
  - Cumulative meterset (e.g. monitor units) over all treatment beams

Plan Scoring Scoresheet, At-a-Glance

Blank Scoresheet

METRIC	RESULT	MIN REQ	IDEAL	POINTS	WEIGHT
Volume (%) of the PTV70 covered by 70 (Gy)		90 (0p 90, 12p 93)	95		15.00
Volume (%) of the PTV70 covered by 73.5 (Gy)		10 (0p 10)	0		7.00
Dose (Gy) covering 0.03 (cc) of the PTV70		77 (0p 77, 5p 73.5)	71.3		7.00
Volume (%) of the CTV70 covered by 70 (Gy)		95 (0p 95, 5p 97)	99		7.00
Volume (%) of the PTV63-PTV70 covered by 63 (Gy)		90 (0p 90, 13p 93)	95		15.00
Volume (%) of the PTV63-PTV70 covered by 66.15 (Gy)		60 (0p 60)	20		7.00
Volume (%) of the CTV63-CTV70 covered by 63 (Gy)		95 (0p 95, 5p 97)	99		7.00
Volume (%) of the PTV56-PTV63 covered by 56 (Gy)		90 (0p 90, 13p 93)	95		15.00
Volume (%) of the PTV56-PTV63 covered by 58.8 (Gy)		60 (0p 60)	30		7.00
Volume (%) of the CTV56-CTV63 covered by 56 (Gy)		95 (0p 95, 5p 97)	99		7.00
Conformation Number [53.2 (Gy), PTV56]		0.5 (0p 0.5, 10p 0.8)	1		12.00
Dose (Gy) covering 0.03 (cc) of the SPINAL_CORD		< 48 (0p < 48)			0.00
Dose (Gy) covering 0.03 (cc) of the BRAINSTEM		< 52 (0p < 52)			0.00
Dose (Gy) covering 0.03 (cc) of the RT COCHLEA		45 (0p 45)	35		5.00
Dose (Gy) covering 0.03 (cc) of the LT COCHLEA		45 (0p 45)	35		5.00
Volume (%) of the LIPS covered by 30 (Gy)		40 (0p 40, 3.5p 35)	20		5.00
Mean dose (Gy) to the RT_PAROTID		30 (0p 30, 5p 26)	24		7.00
Volume (%) of the MANDIBLE covered by 70 (Gy)		25 (0p 25, 2.5p 20)	10		5.00
Mean dose (Gy) to the LARYNX		50 (0p 50, 5p 45)	40		7.00
Volume (%) of the POST NECK covered by 35 (Gy)		30 (0p 30, 2p 25)	10		5.00
Structure(s) containing the global max dose point		Elsewhere (0p ELSEWHERE, 3p PTV70)	CTV70		5.00
Estimated 'beam-on' time, all beams (minutes)		---	---	---	---
Cumulative meterset over all treatment beams		---	---	---	---
TOTALS		21 Goals	21 Goals		150.00

Scored Plan

METRIC	RESULT	MIN REQ	IDEAL	POINTS	WEIGHT
Volume (%) of the PTV70 covered by 70 (Gy)	95.64	90 ✓ (0p 90, 12p 93)	95 ✓	15.00	15.00
Volume (%) of the PTV70 covered by 73.5 (Gy)	1.37	10 ✓ (0p 10)	0 ☐	6.04	7.00
Dose (Gy) covering 0.03 (cc) of the PTV70	73.82	77 ✓ (0p 77, 5p 73.5)	71.3 ☐	4.54	7.00
Volume (%) of the CTV70 covered by 70 (Gy)	99.49	95 ✓ (0p 95, 5p 97)	99 ✓	7.00	7.00
Volume (%) of the PTV63-PTV70 covered by 63 (Gy)	95.85	90 ✓ (0p 90, 13p 93)	95 ✓	15.00	15.00
Volume (%) of the PTV63-PTV70 covered by 66.15 (Gy)	55.86	60 ✓ (0p 60)	20 ☐	0.72	7.00
Volume (%) of the CTV63-CTV70 covered by 63 (Gy)	100.00	95 ✓ (0p 95, 5p 97)	99 ✓	7.00	7.00
Volume (%) of the PTV56-PTV63 covered by 56 (Gy)	93.90	90 ✓ (0p 90, 13p 93)	95 ☐	13.90	15.00
Volume (%) of the PTV56-PTV63 covered by 58.8 (Gy)	41.63	60 ✓ (0p 60)	30 ☐	4.29	7.00
Volume (%) of the CTV56-CTV63 covered by 56 (Gy)	99.34	95 ✓ (0p 95, 5p 97)	99 ✓	7.00	7.00
Conformation Number [53.2 (Gy), PTV56]	0.74	0.5 ✓ (0p 0.5, 10p 0.8)	1 ☐	8.16	12.00
(Gy) covering 0.03 (cc) of the SPINAL_CORD	46.74	< 48 ✓ (0p < 48)	✓	0.00	0.00
Dose (Gy) covering 0.03 (cc) of the BRAINSTEM	51.30	< 52 ✓ (0p < 52)	✓	0.00	0.00
Dose (Gy) covering 0.03 (cc) of the RT COCHLEA	24.66	45 ✓ (0p 45)	35 ✓	5.00	5.00
Dose (Gy) covering 0.03 (cc) of the LT COCHLEA	33.15	45 ✓ (0p 45)	35 ✓	5.00	5.00
Volume (%) of the LIPS covered by 30 (Gy)	24.39	40 ✓ (0p 40, 3.5p 35)	20 ☐	4.56	5.00
Mean dose (Gy) to the RT_PAROTID	23.07	30 ✓ (0p 30, 5p 26)	24 ✓	7.00	7.00
Volume (%) of the MANDIBLE covered by 70 (Gy)	12.23	25 ✓ (0p 25, 2.5p 20)	10 ☐	4.44	5.00
Mean dose (Gy) to the LARYNX	40.66	50 ✓ (0p 50, 5p 45)	40 ☐	6.73	7.00
Volume (%) of the POST NECK covered by 35 (Gy)	9.66	30 ✓ (0p 30, 2p 25)	10 ✓	5.00	5.00
Structure(s) containing the global max dose point	(7 values)	Elsewhere ✓ (0p ELSEWHERE, 3p PTV70)	CTV70 ✓	5.00	5.00
Estimated 'beam-on' time, all beams (minutes)	7.60	---	---	---	---
Cumulative meterset over all treatment beams	1826.42	---	---	---	---
TOTALS		21 (of 21)	12 (of 21)	131.40	150.00

Dose QA: Metrics Collected

- **QA Device**
- **QA Method**
  - True Composite, Single Angle Composite, EPID-based planar, independent dose recalc, etc.
- **Gamma Passing Rates**
  - 3% (global) / 3 mm / 10% lower threshold
  - 3% (global) / 2 mm / 10% lower threshold
  - 2% (local) / 2 mm / 20% lower threshold

Dose QA: Special Rules

- For the dose QA method field, generic and/or industry-standard terms were required, such as:
  - True composite
  - Single angle composite
  - Portal dosimetry
  - Log-file based re-calculation
  - etc.
- Users not allowed to use “measurement uncertainty”
  - Bailey et al., “Measurement Uncertainty function and its effect on planar dose pass rates,” JACMP 17(2), 2016.

Refresher: The X's and O's of 3D Dosimeters

JOURNAL OF APPLIED CLINICAL MEDICAL PHYSICS, VOLUME 12, NUMBER 2, SPRING 2011

Evaluation of a new VMAT QA device, or the "X" and "O" array geometries

Vladimir Feygelman,[1a] Geoffrey Zhang,[1] Craig Stevens,[1]
Benjamin E. Nelms[2]
Division of Radiation Oncology,[1] H. Lee Moffitt Cancer Center, Tampa, Florida 33612, USA; Canis Lupus LLC,[2] Sauk County, Wisconsin 53561, USA.
vladimir.feygelman@moffitt.org

Received 19 April, 2009; accepted 8 November, 2010

- All else equal, QA metrics (e.g. passing rates) measured by different detector geometries will **NOT** be the same. The more sensitive the metric, the more varied the results.
- Sensitivities vary based on error type and where the errors manifest in 3D.
- This presents a real problem for any group trying to publish general guidelines that apply across all QA devices.

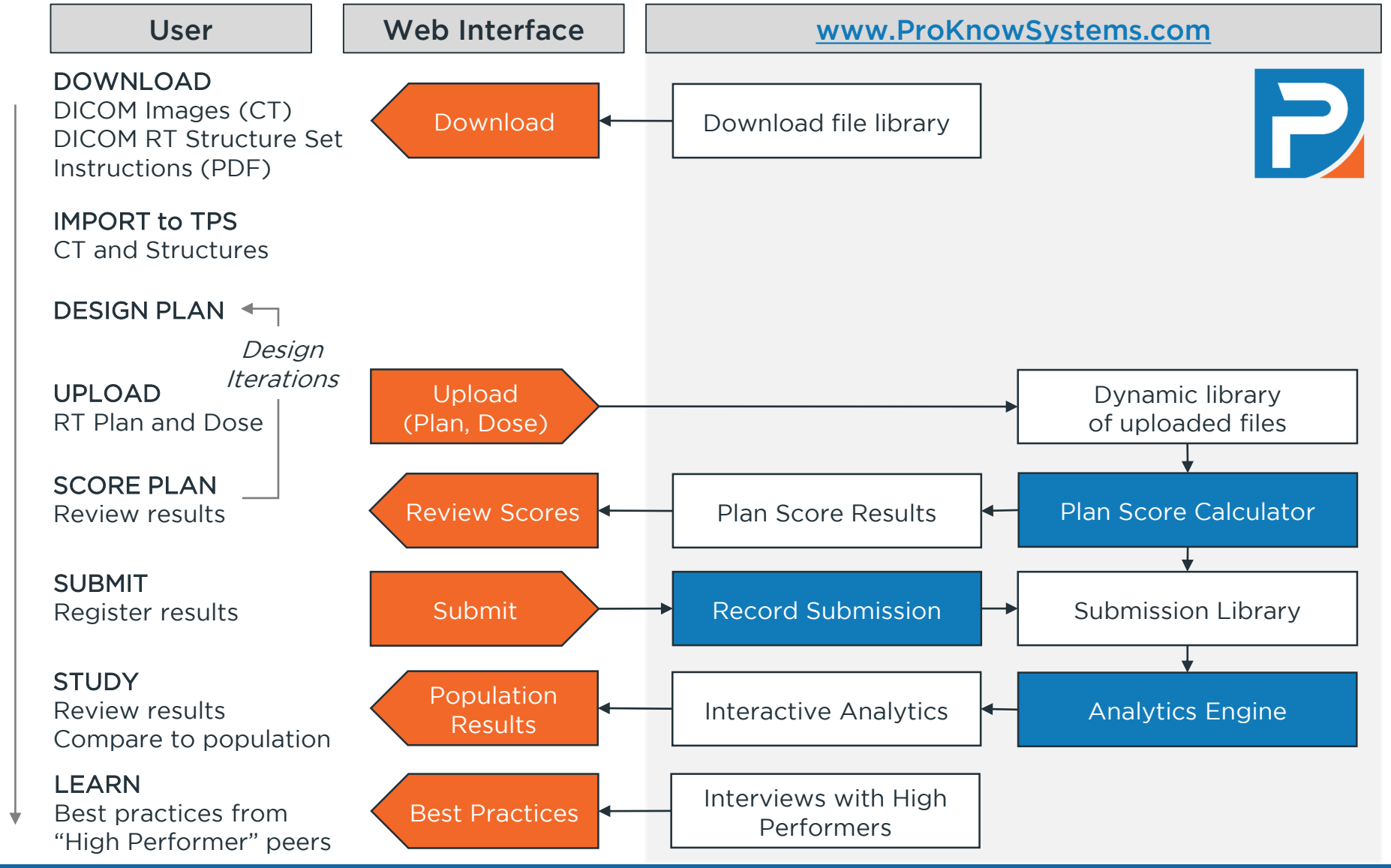
Refresher: Some Inconvenient Truths

- Jin X, Yan H, Han C, Zhou Y, Yi J, Xie C. Correlation between gamma index passing rate and clinical dosimetric difference for pre-treatment 2D and 3D volumetric modulated arc therapy dosimetric verification. Br J Radiol 2015;88:20140577.
- Nelms BE, Chan MF, Jarry G, Lemire M, Lowden J, Hampton C, and Feygelman V. 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. 2013 Nov; 40(11).
- Nelms BE, Opp D, Robinson J, Wolf TK, Zhang G, Moros E, Feygelman V. VMAT QA: measurement-guided 4D dose reconstruction on a patient. Med Phys. 2012 Jul; 39(7).
- Feygelman V, Stambaugh C, Opp D, Zhang G, Moros E, and Nelms, BE. Cross-validation of two commercial methods for volumetric high-resolution dose reconstruction on a phantom for non-coplanar VMAT beams. Radiother Oncol. 2014 Mar; 110(3).
- Opp D, Nelms BE, Zhang G, Stevens C, Feygelman V. Validation of measurement-guided 3D VMAT dose reconstruction on a heterogeneous anthropomorphic phantom. J Appl Clin Med Phys. 2013 Jul; 14(4).
- Chan MF, Li J, Schupak K, Burman C. 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. Technol Cancer Res Treat. 2013 Jun; 13(1).
- Stasi et al. Pretreatment patient-specific IMRT quality assurance: A correlation study between gamma index and patient clinical dose volume histogram. Med Phys. 2012 Dec; 39(12).
- Carrasco et al. 3D DVH-based metric analysis versus per-beam planar analysis in IMRT pretreatment verification. Med Phys. 2012 Aug; 39(8).
- Zhen H, Nelms BE, Tome WA. Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA. Med Phys. 2011 Oct; 38(10).
- Nelms BE, Zhen H, Tomé WA. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. Med Phys. 2011 Feb;38(2).

Refresher: Some Inconvenient Truths

- No Gamma passing rate at any setting of % difference or distance has been proven to be sensitive and specific in terms of detecting relevant clinical errors in the TPS dose calc (or dose delivery).
- In fact, all common permutations of gamma have been proven insensitive and non-specific.
  - Among those with low specificity and sensitivity are: 3% (global)/ 3 mm (TG119), 3% (global) / 2 mm (TG218), and even 2% (global) / 2 mm (TG244).
  - Higher sensitivity observed with 2% (local) / 2 mm, but with some concern if it's too sensitive, i.e. false positives. Needs benchmarking to see what is achievable.
- Accurate, measurement-guided 3D dose reconstruction and DVH analysis is promising, and by definition is both sensitive and specific.

The ProKnow Plan Study System



SAM Questions

Ready, Set, Go!

SAM Question #1

1. Which of the following is currently the biggest driver of radiation treatment plan quality?
   A. Advanced modalities such as VMAT & protons
   B. TPS model
   C. Allowing more monitor units for modulation
   D. Planner experience level
   E. None of the above

SAM Question #1: Answer

1. Which of the following is currently the biggest driver of radiation treatment plan quality?
   A. Advanced modalities such as VMAT & protons
   B. TPS model
   C. Allowing more monitor units for modulation
   D. Planner experience level
   E. None of the above

REFERENCE: Nelms BE, Robinson G, Markham J, Velasco K, Boyd S, Narayan S, Wheeler J, Sobczak M. Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems. Practical Radiation Oncology 2012 Oct;2(4):296-305.

SAM Question #2

2. Which of the following summarizes the goal of the MPPG 5.a / TG244 guidelines?
   A. To summarize the standard methods and performance benchmarks for dose QA.
   B. To summarize minimum requirements for TPS dose calculation algorithm commissioning and QA.
   C. To provide methods and minimum requirements for linear accelerator commissioning.
   D. To provide standard datasets to verify inter-connectivity of medical devices.

SAM Question #2: Answer

2. Which of the following summarizes the goal of the MPPG 5.a / TG244 guidelines?

   A. To summarize the standard methods and performance benchmarks for dose QA.

   B. To summarize minimum requirements for TPS dose calculation algorithm commissioning and QA.

   C. To provide methods and minimum requirements for linear accelerator commissioning.

   D. To provide standard datasets to verify inter-connectivity of medical devices.

REFERENCE: AAPM TG244 subcommittee. AAPM Medical Physics Practice Guideline 5.a.: Commissioning and QA of Treatment Planning Dose Calculations — Megavoltage Photon and Electron Beams. J Appl Clin Med Phys. 2015;16(5).

SAM Question #3

3. Which of the following is important for a study aimed to quantify and compare the abilities of different TPS and treatment planners to produce a high quality plan?
   A. Each planner must use the same patient imageset and contoured critical volumes.
   B. Dose calculation grids must be of sufficient resolution and size.
   C. Each submitted plan and corresponding dose must be scored exactly the same way, with 100% transparency and objectivity.
   D. All of the above.

SAM Question #3: Answer

3. Which of the following is important for a study aimed to quantify and compare the abilities of different TPS and treatment planners to produce a high quality plan?
   A. Each planner must use the same patient imageset and contoured critical volumes.
   B. Dose calculation grids must be of sufficient resolution and size.
   C. Each submitted plan and corresponding dose must be scored exactly the same way, with 100% transparency and objectivity.
   D. All of the above.

REFERENCE: Nelms BE, Robinson G, Markham J, Velasco K, Boyd S, Narayan S, Wheeler J, Sobczak M. Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems. Practical Radiation Oncology 2012 Oct;2(4):296-305.

SAM Question #4

4. Which of the following is true about pretreatment dose QA using different dosimetry methods:
   A. If an identical fraction is analyzed with multiple commercial QA systems, a physicist should expect identical, or at least very similar, metric results.
   B. A gamma passing rate as measured by a Delta 4 “X” geometry will produce the same or higher passing rate if measured by an ArcCHECK “O” geometry.
   C. Passing rates of 100% as measured by methods using an EPID have been proven to ensure a passing rate of 90% or higher for the same criteria in a true 3D dosimeter.
   D. None of the above.

SAM Question #4: Answer

4. Which of the following is true about pretreatment dose QA using different dosimetry methods:
   A. If an identical fraction is analyzed with multiple commercial QA systems, a physicist should expect identical, or at least very similar, metric results.
   B. A gamma passing rate as measured by a Delta 4 “O” geometry will produce the same or higher passing rate if measured by an ArcCHECK “X” geometry.
   C. Passing rates of 100% as measured by methods using an EPID have been proven to ensure a passing rate of 90% or higher for the same criteria in a true 3D dosimeter.
   D. None of the above.

REFERENCE: Feygelman V, Zhang G, Stevens C, Nelms BE. Evaluation of a new VMAT QA device, or the “X” and “O” array geometries. J Appl Clin Med Phys. 2011 Jan 31;12(2):3346.

REFERENCE: Zhen H, Nelms BE, Tome WA. Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA. Med Phys. 2011 Oct;38(10):5477-89.

SAM Question #5

5. Which of the following is true regarding the use of gamma passing rates for dose QA:
   A. Clinically relevant errors can still occur even for passing rates > 95% for conventional criteria such as 3% / 3 mm.
   B. Local percent dose normalization is more sensitive than global normalization.
   C. All else equal, different dosimetry methods can produce different gamma results.
   D. No gamma method has ever been proven to be sensitive and specific relative to detection of clinically relevant errors.
   E. All of the above.

SAM Question #5: Answer

5. Which of the following is true regarding the use of gamma passing rates for dose QA:
   A. Clinically relevant errors can still occur even for passing rates > 95% for conventional criteria such as 3% / 3 mm.
   B. Local percent dose normalization is more sensitive than global normalization.
   C. All else equal, different dosimetry methods can produce different gamma results.
   D. No gamma method has ever been proven to be sensitive and specific relative to detection of clinically relevant errors.
   E. All of the above.

REFERENCES: *Many* (see earlier slide), including: Nelms BE, Chan MF, Jarry G, Lemire M, Lowden J, Hampton C, and Feygelman V. 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. 2013 Nov; 40(11).

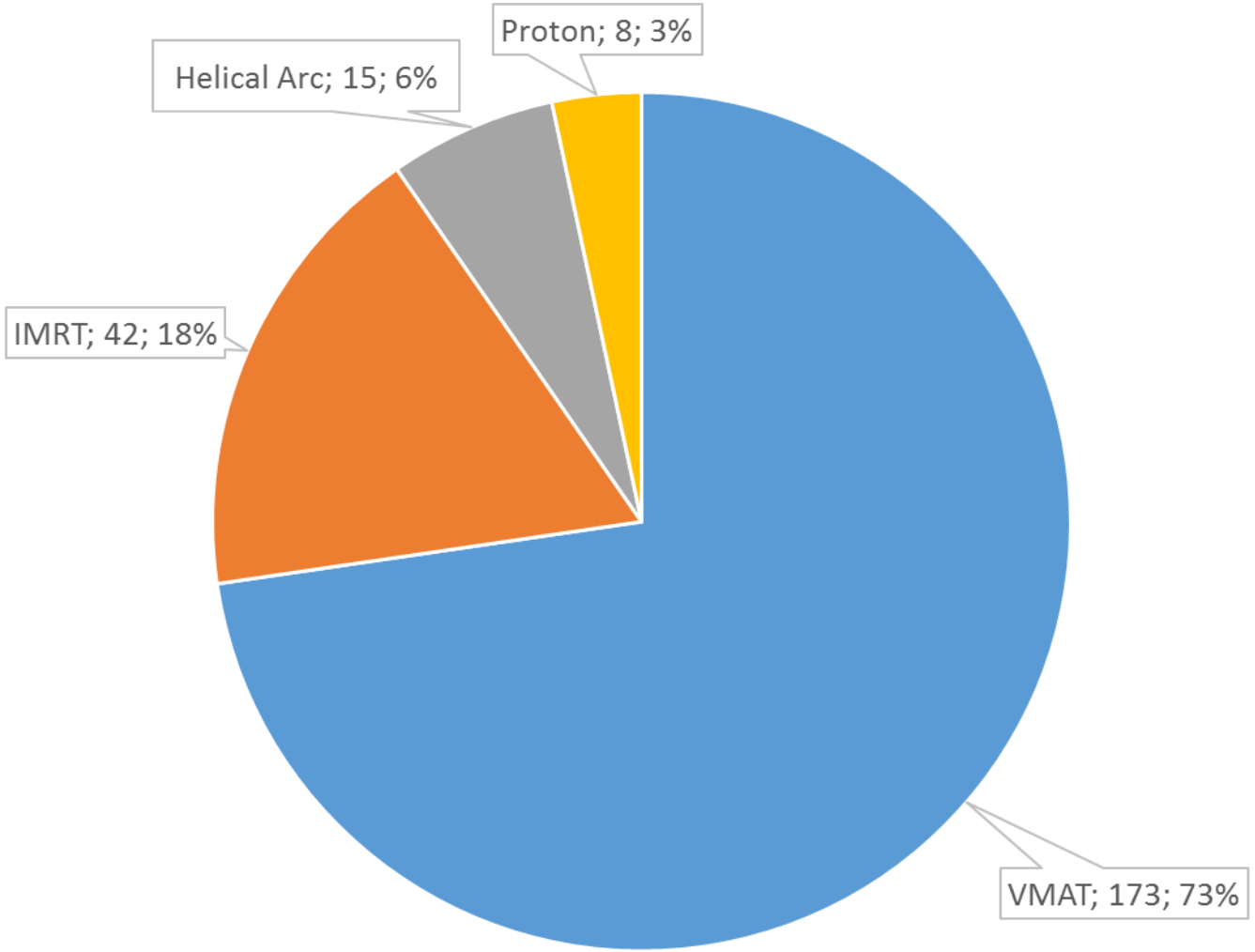
Results

Data Analyses

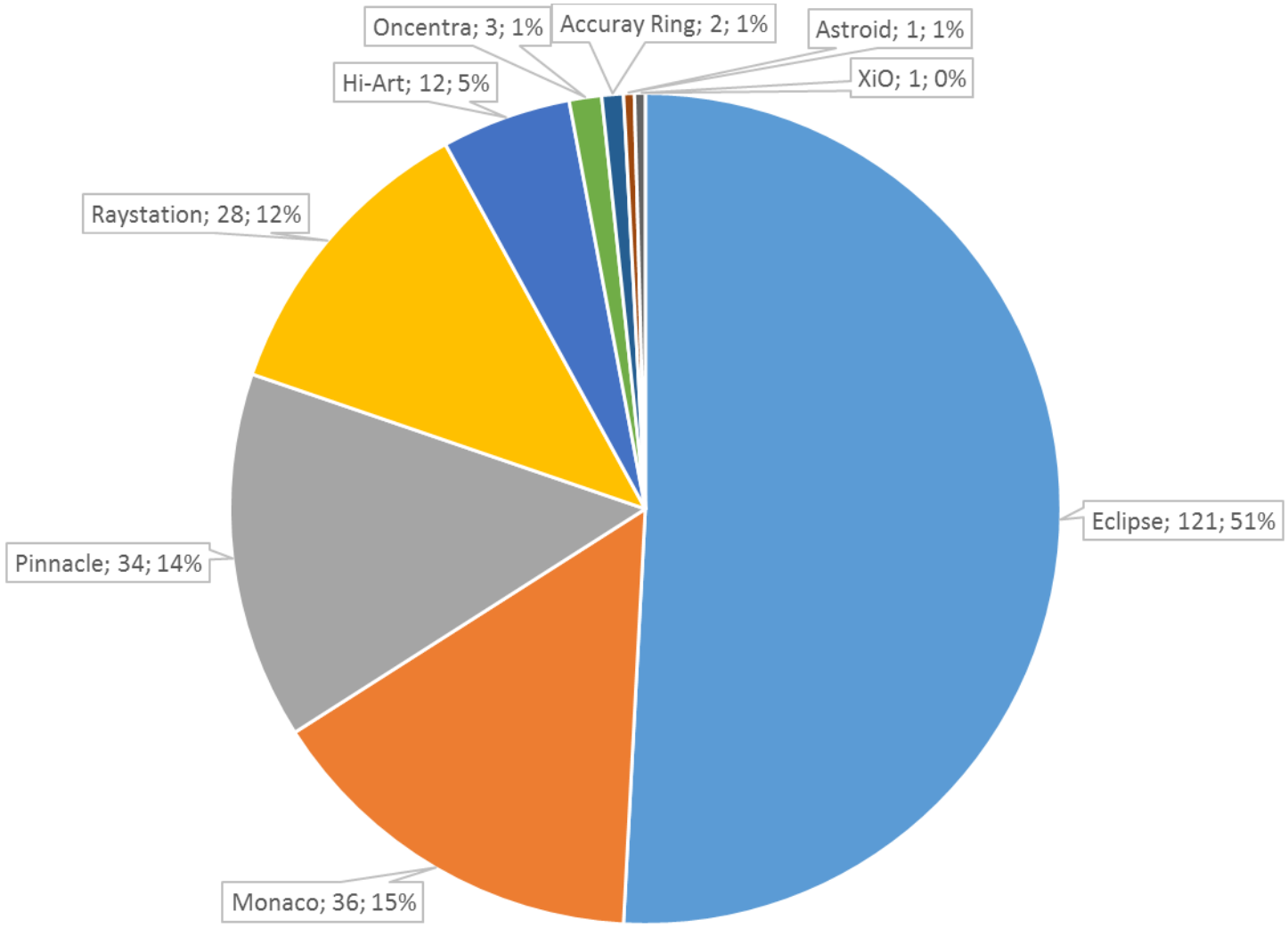
Comparisons

Studies of Variation

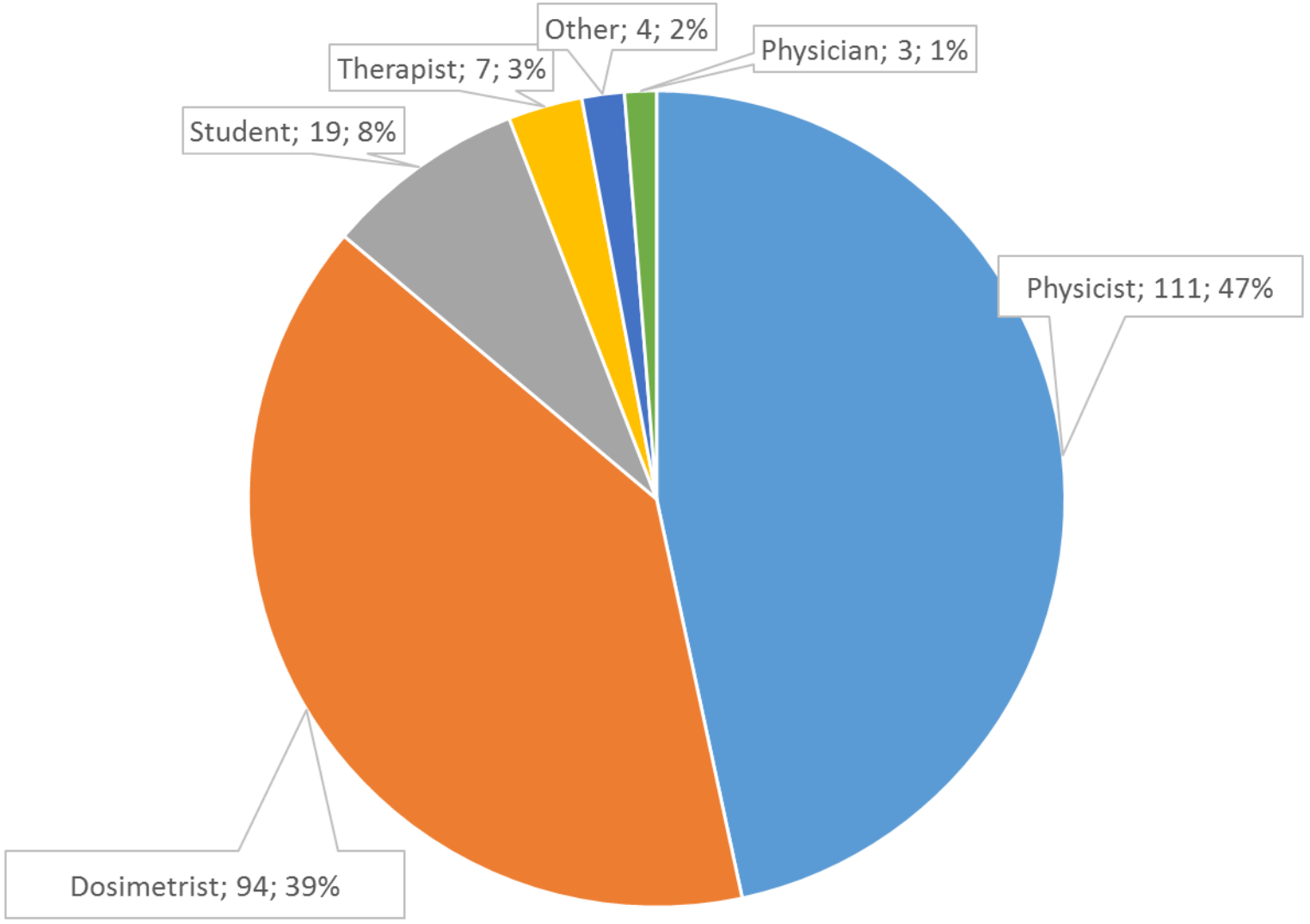
Participation by Modality (N = 238)



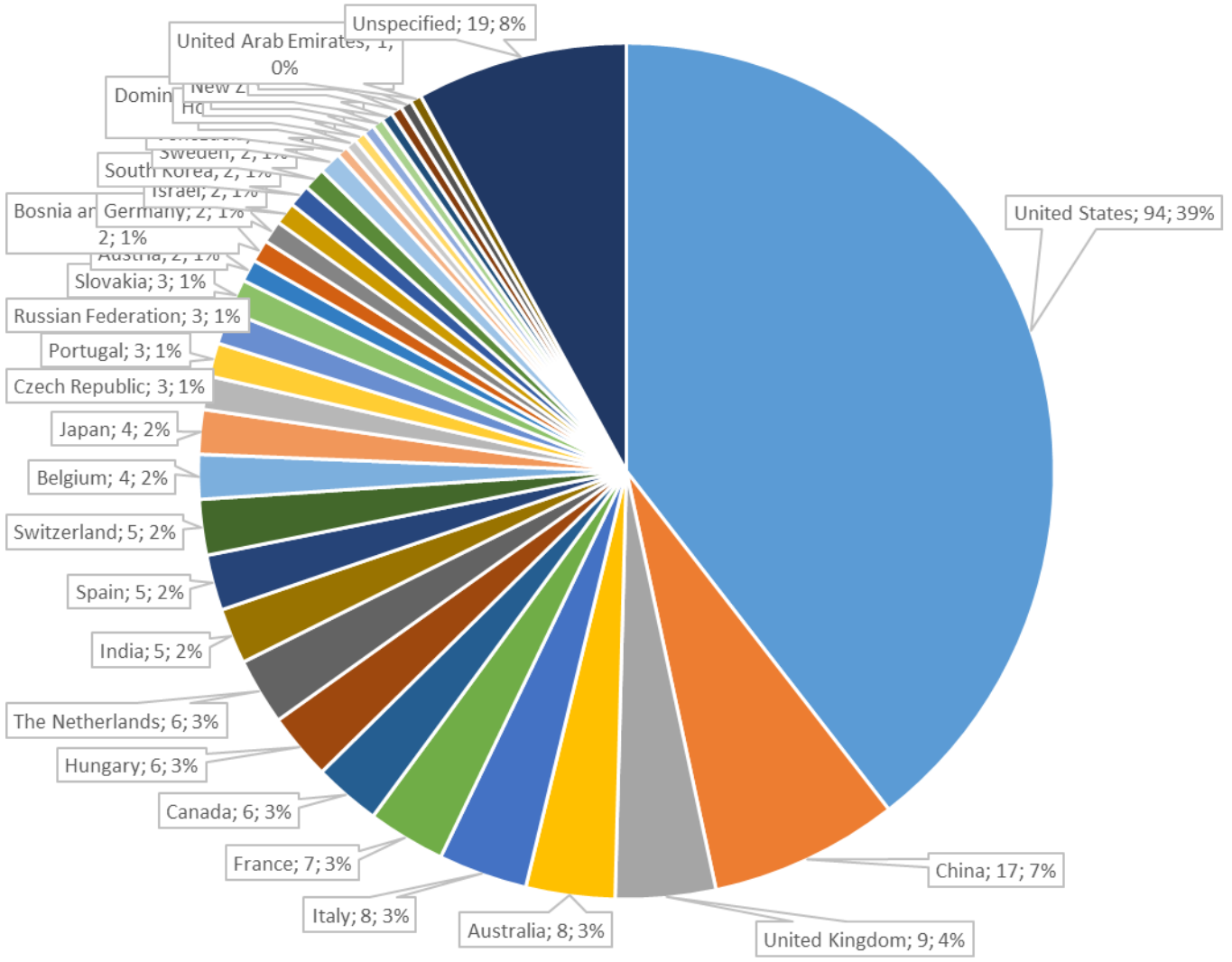
Participation by TPS (N = 238)



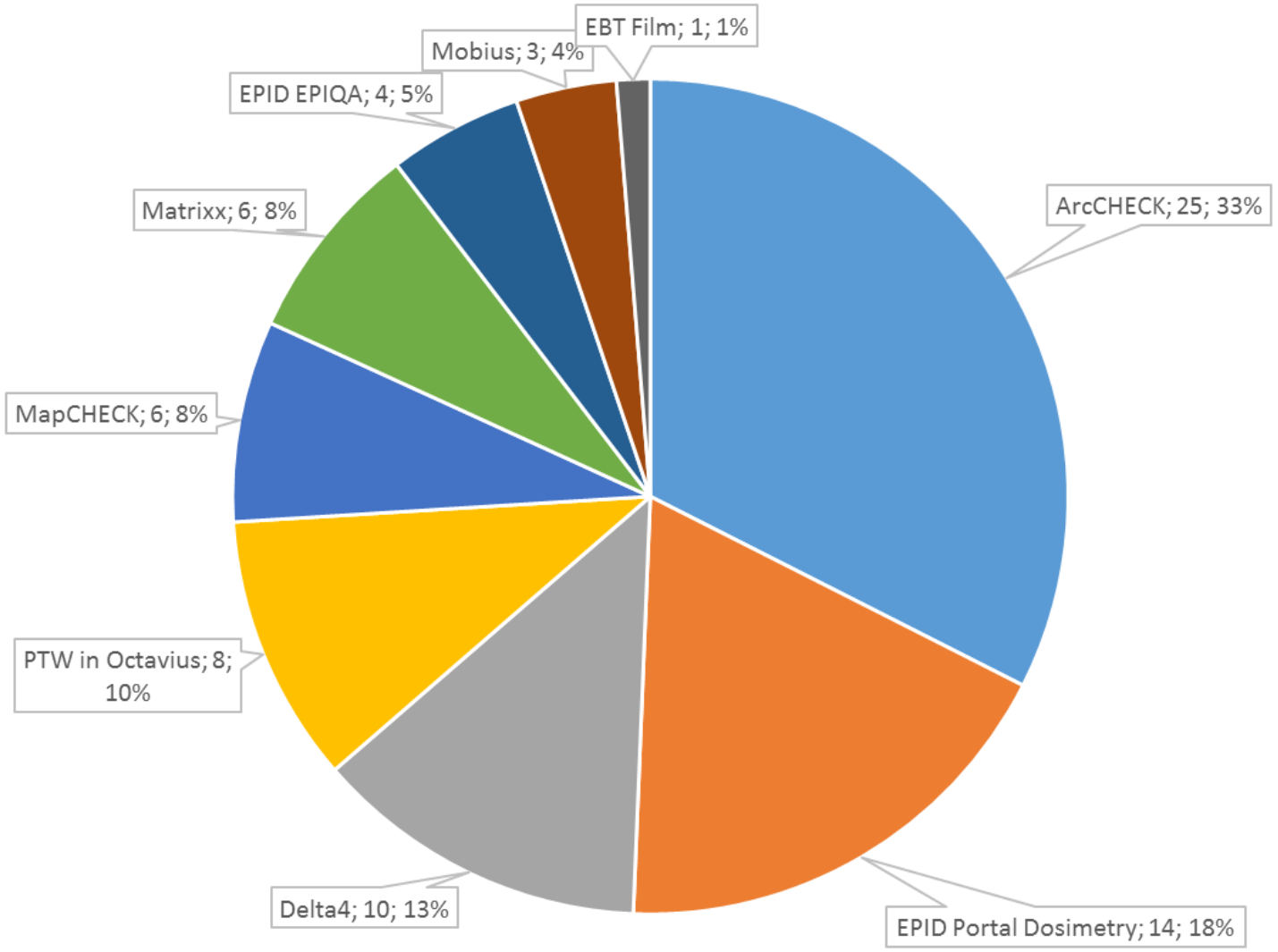
Participation by Role (N = 238)



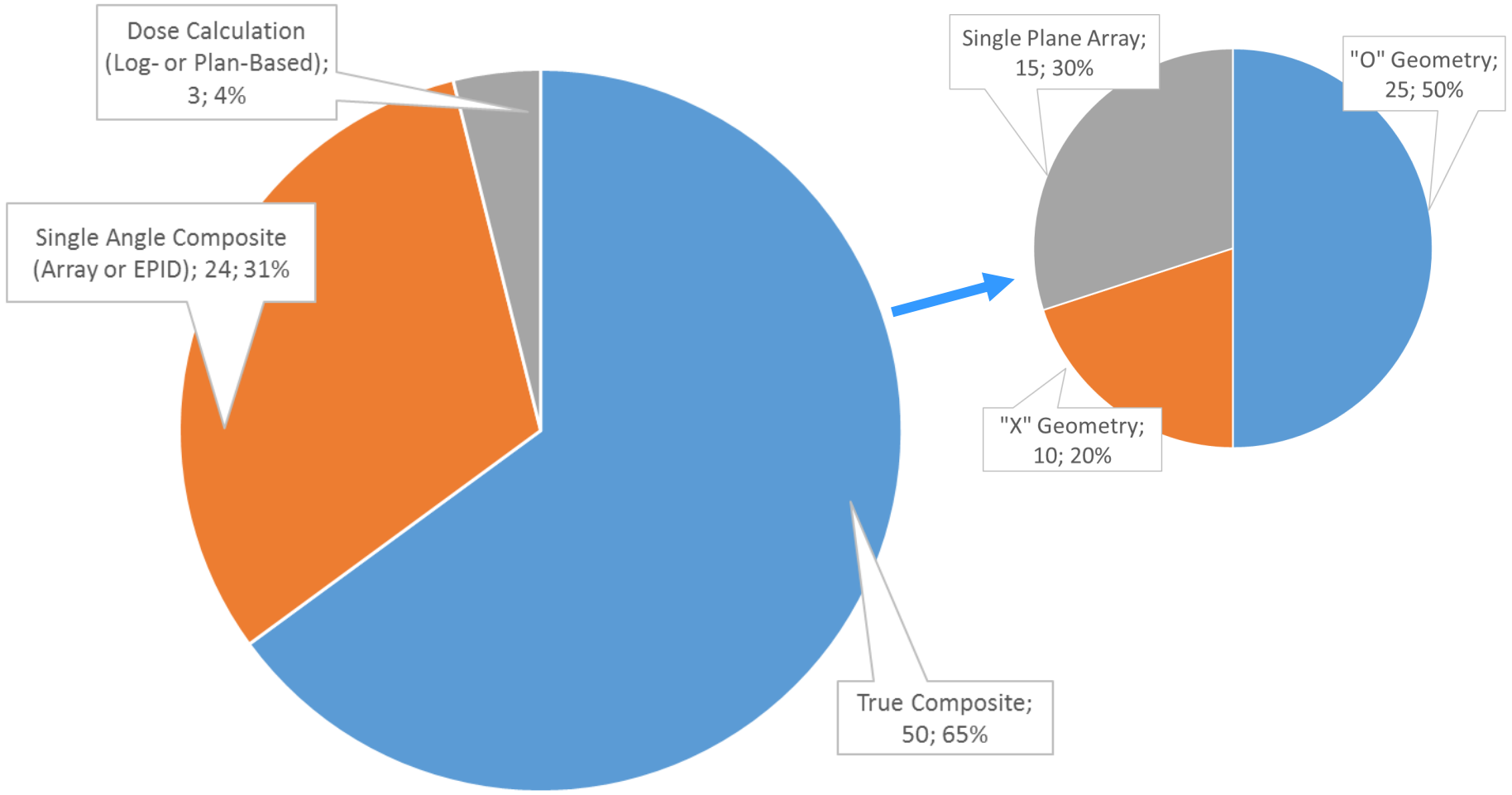
Participation by Country (34 Countries)

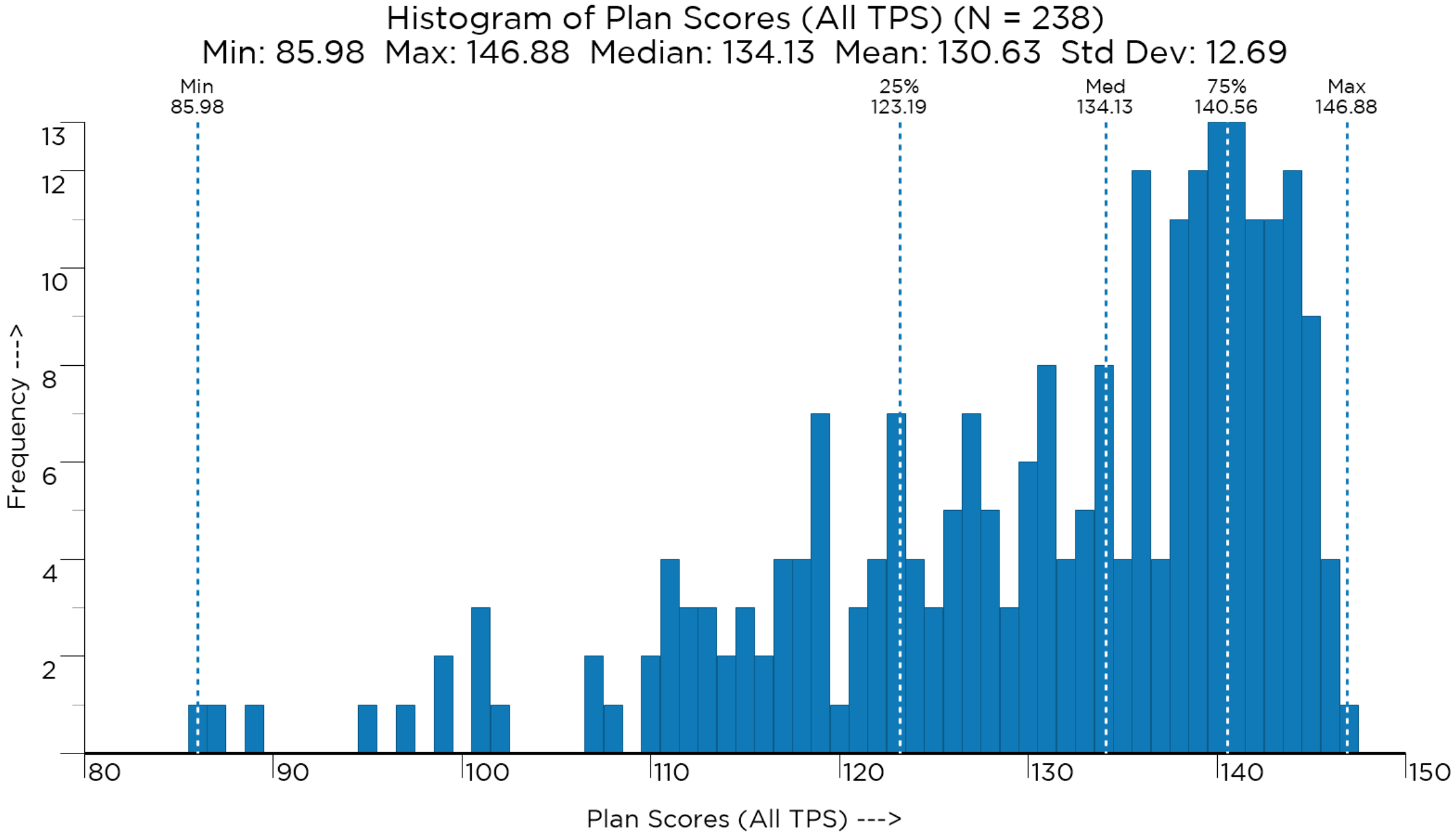


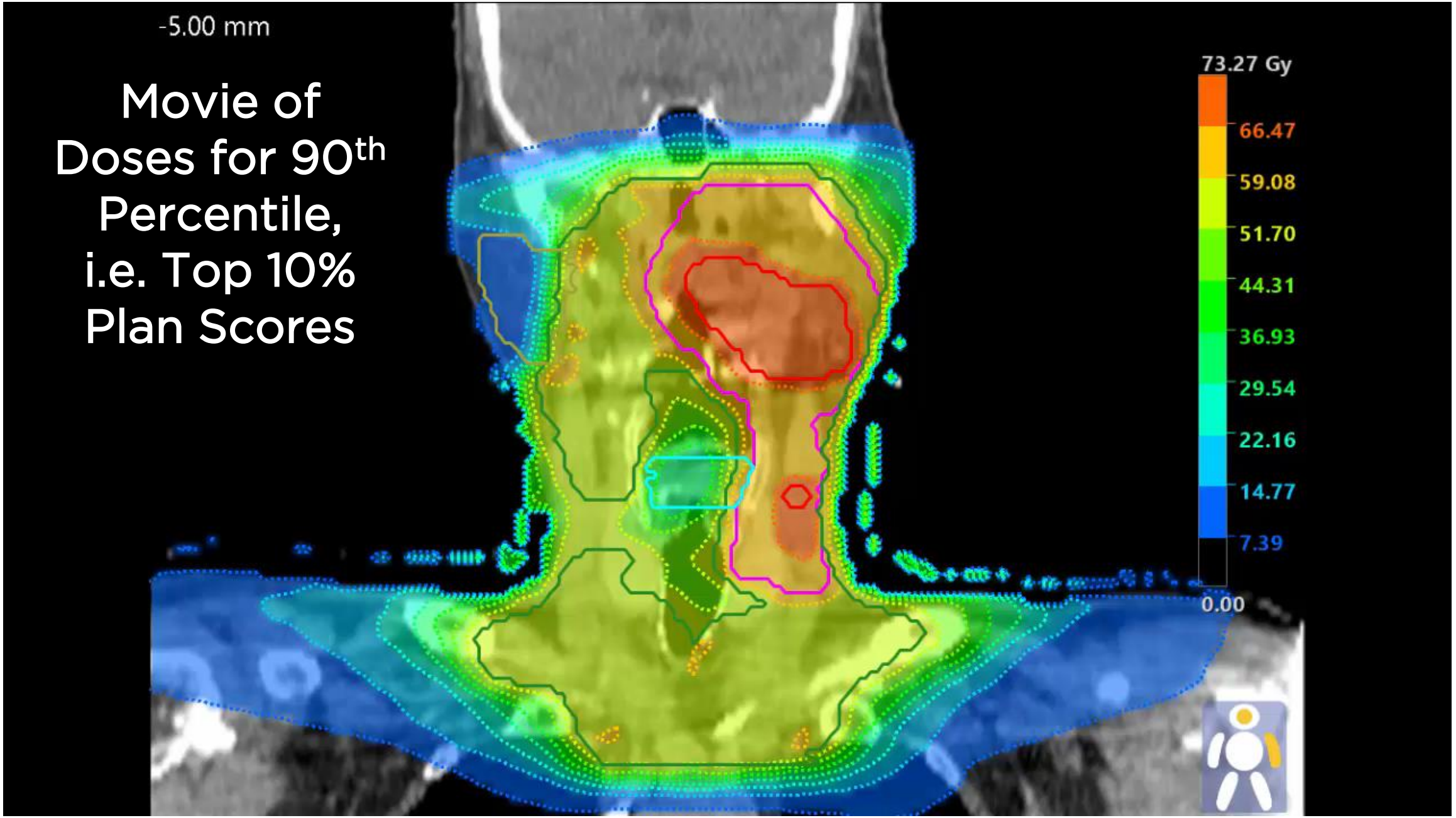
Participation by QA Device (N = 77)

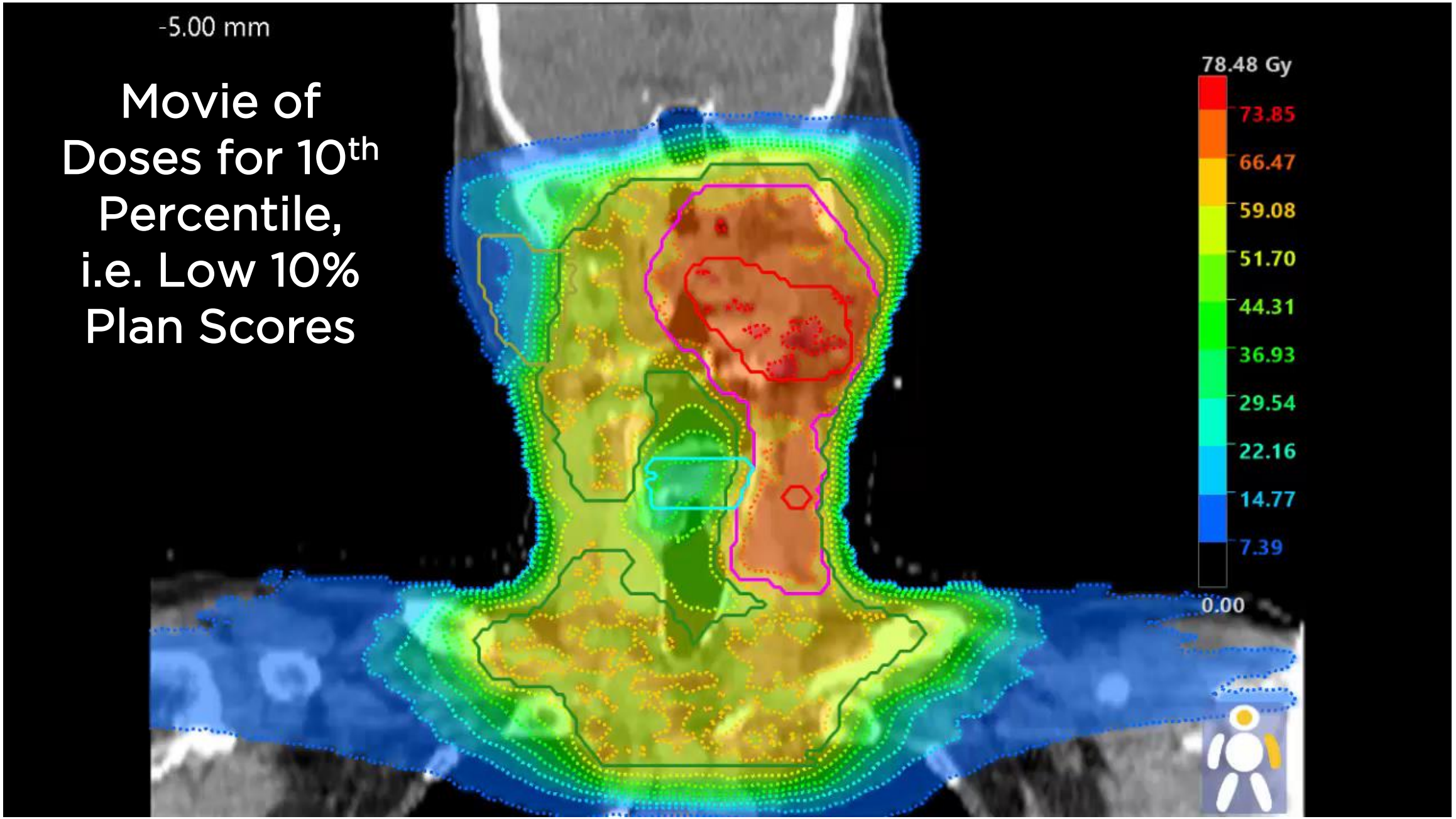


Participation by QA Method (N = 77)



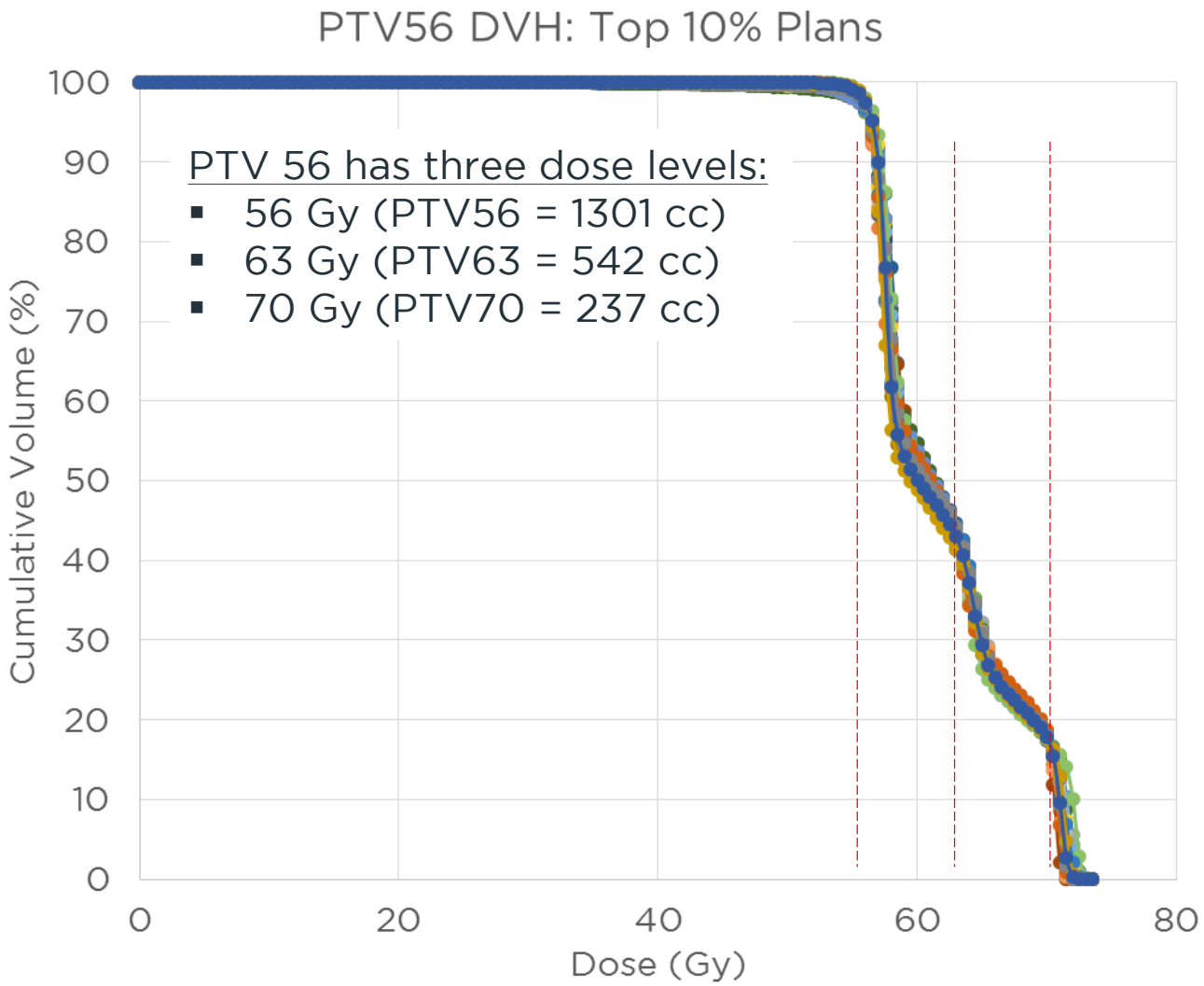

Histogram of Plan Scores (All TPS) (N = 238)
Min: 85.98 Max: 146.88 Median: 134.13 Mean: 130.63 Std Dev: 12.69
Min
85.98
25%
123.19
Med
134.13
75%
140.56
Max
146.88
Frequency --->
13
12
10
8
6
4
2
80
90
100
110
120
130
140
150
Plan Scores (All TPS) --->

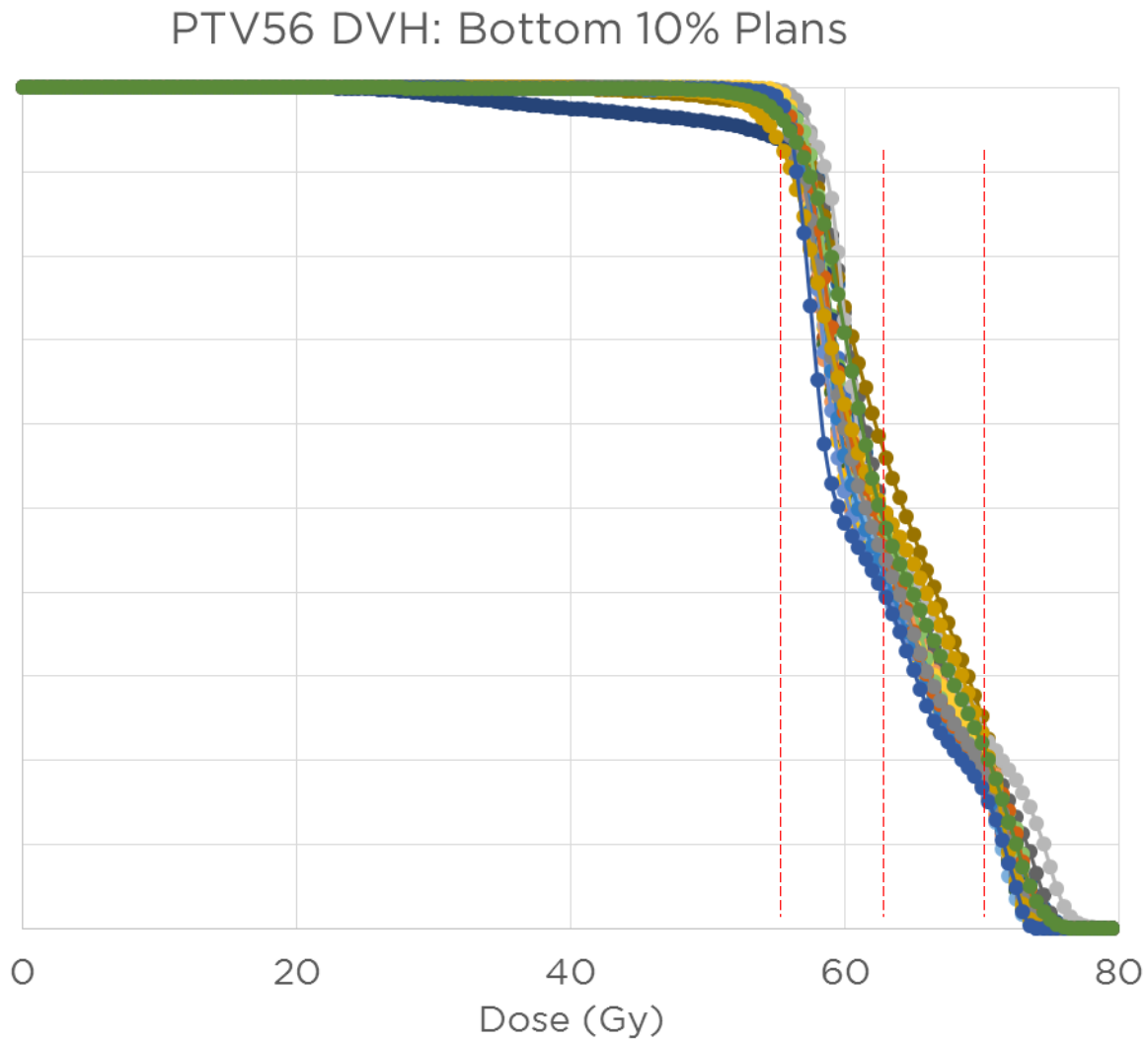

-5.00 mm
Movie of Doses for 90th Percentile, i.e. Top 10% Plan Scores
73.27 Gy
66.47
59.08
51.70
44.31
36.93
29.54
22.16
14.77
7.39
0.00



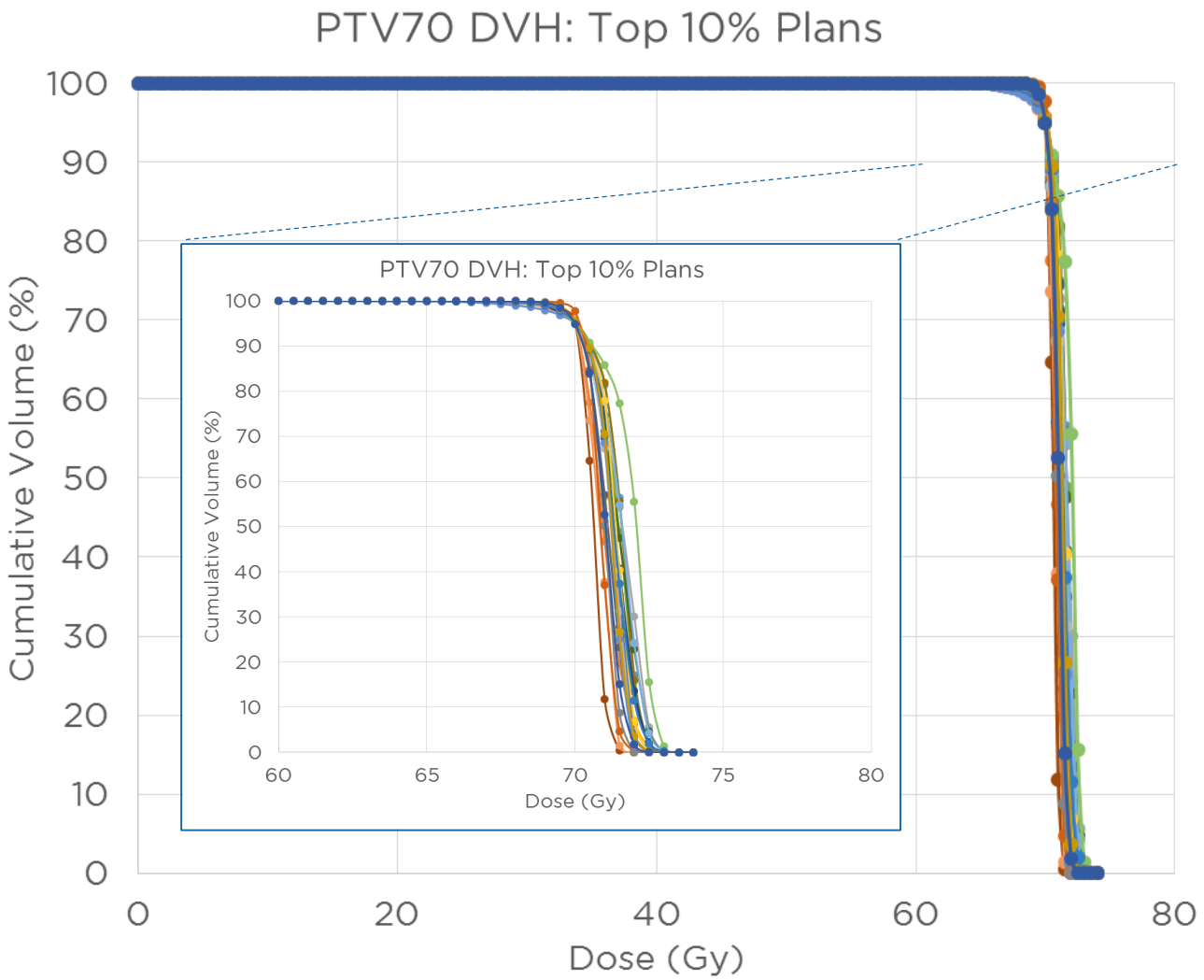
Movie of Doses for 10th Percentile, i.e. Low 10% Plan Scores

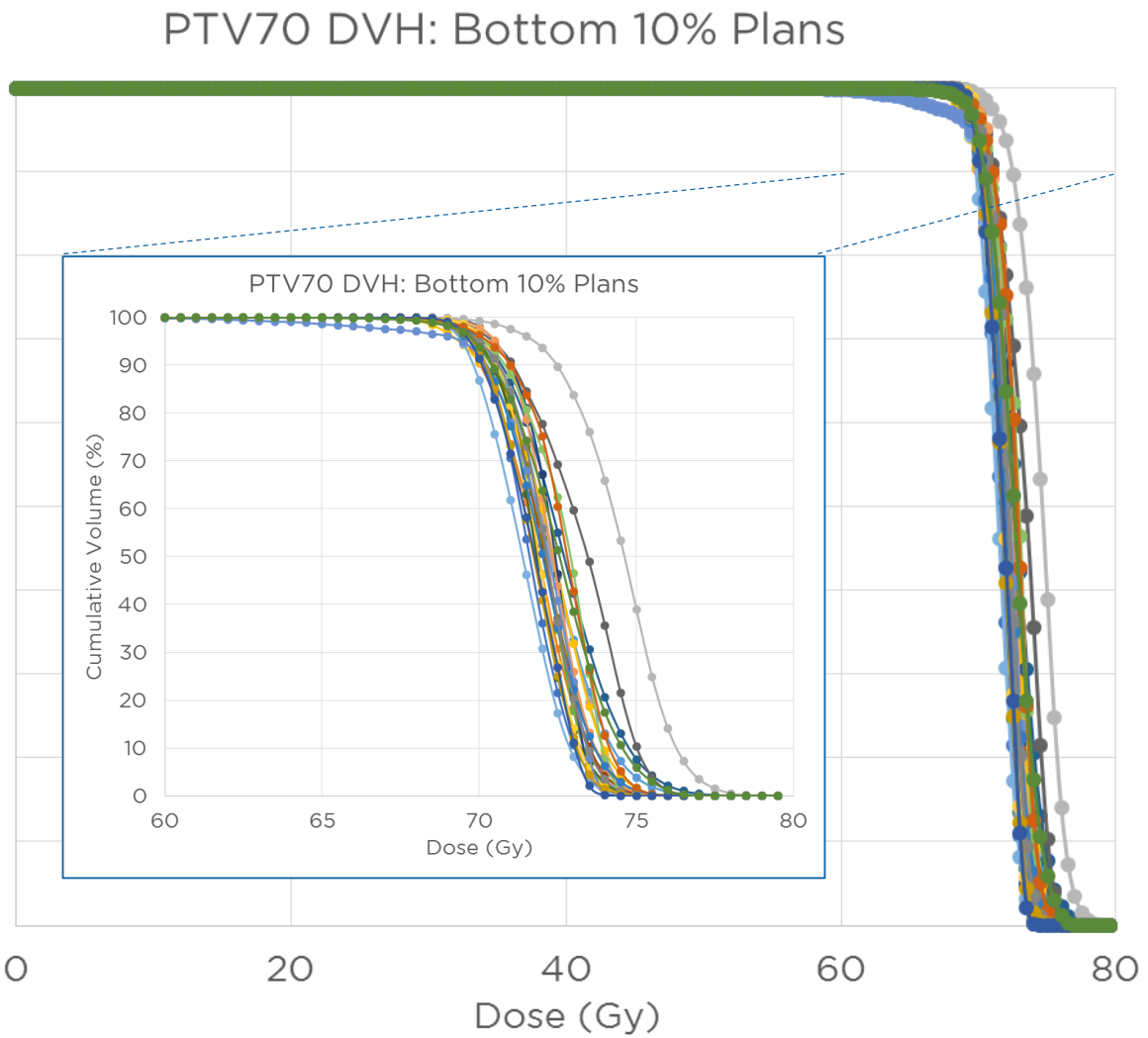
Visual Examples of Variation



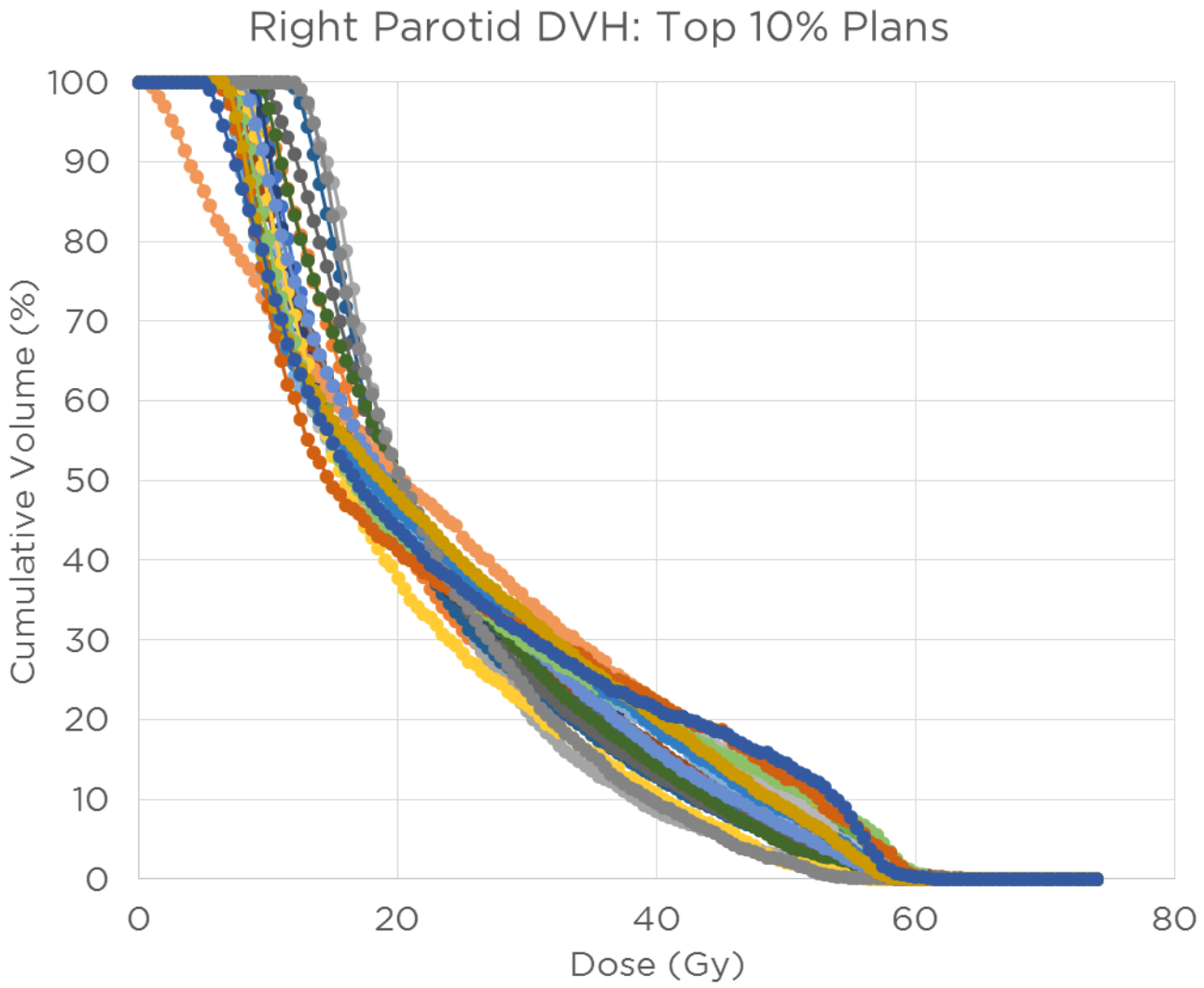


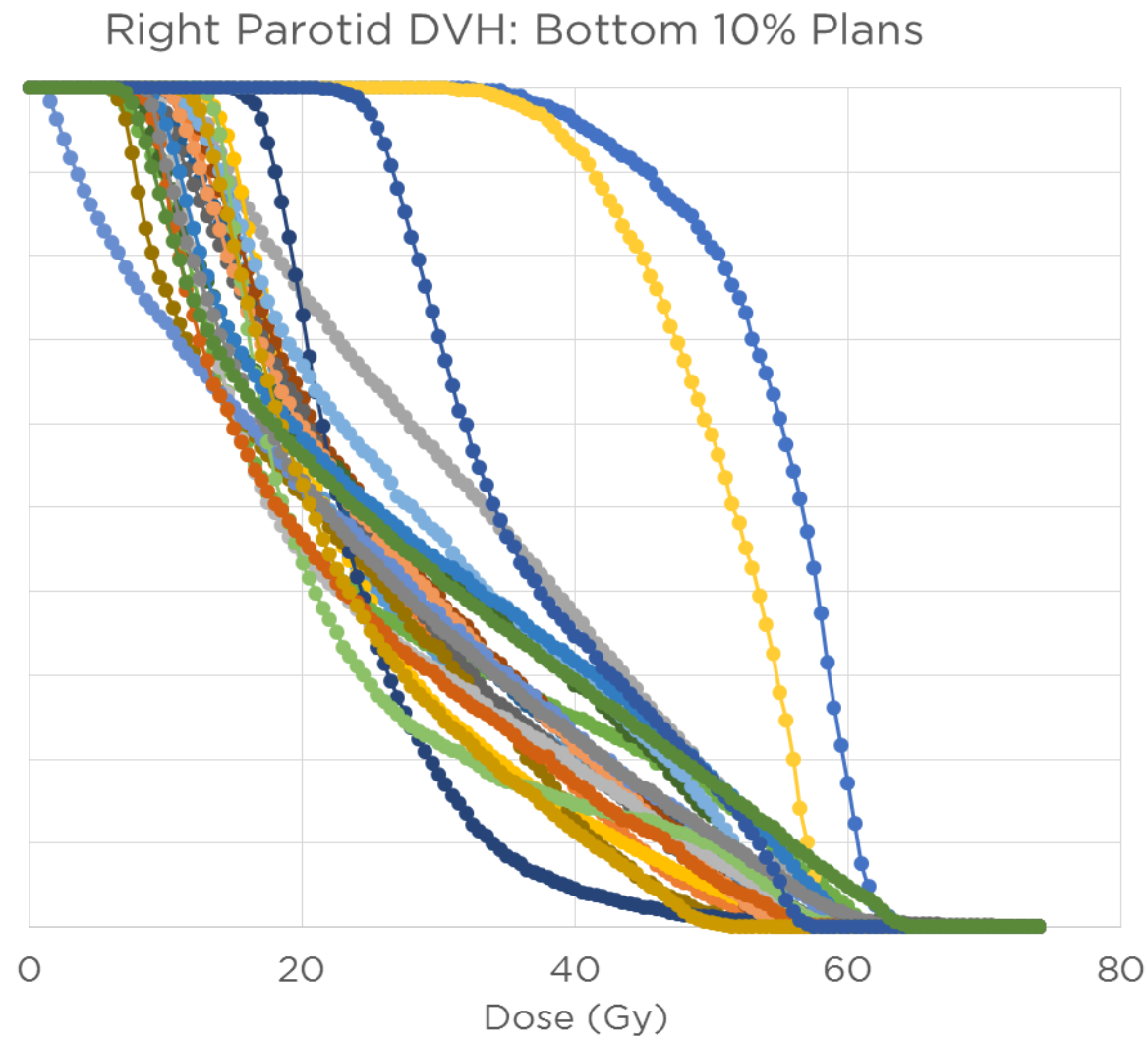
Visual Examples of Variation

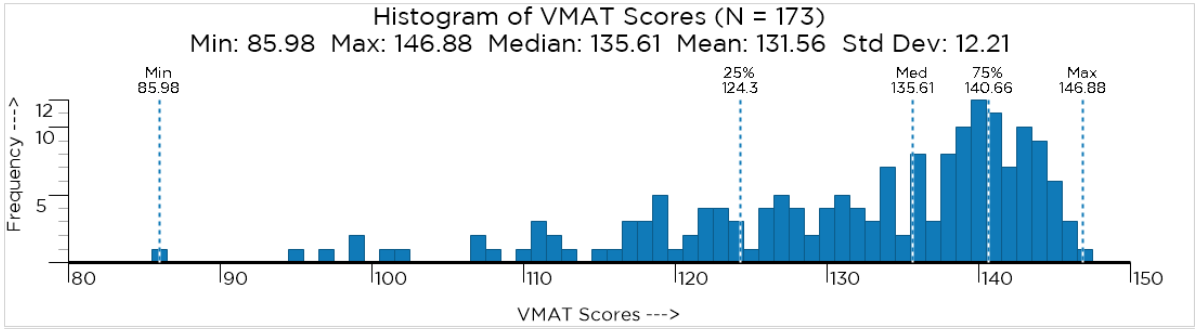


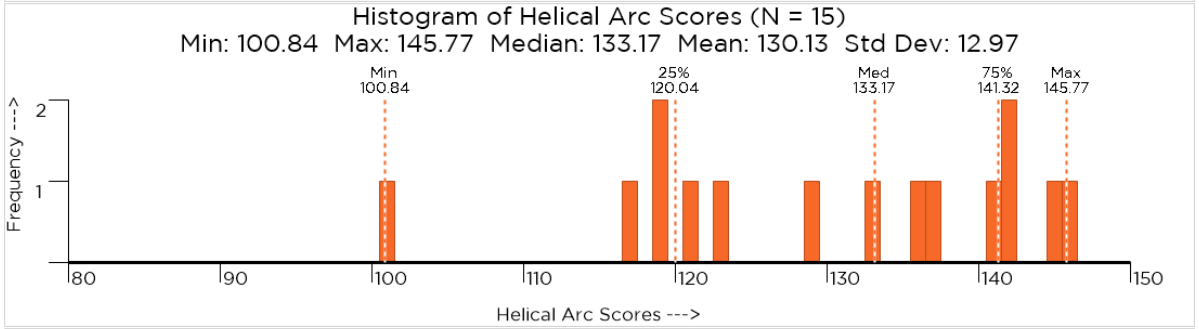


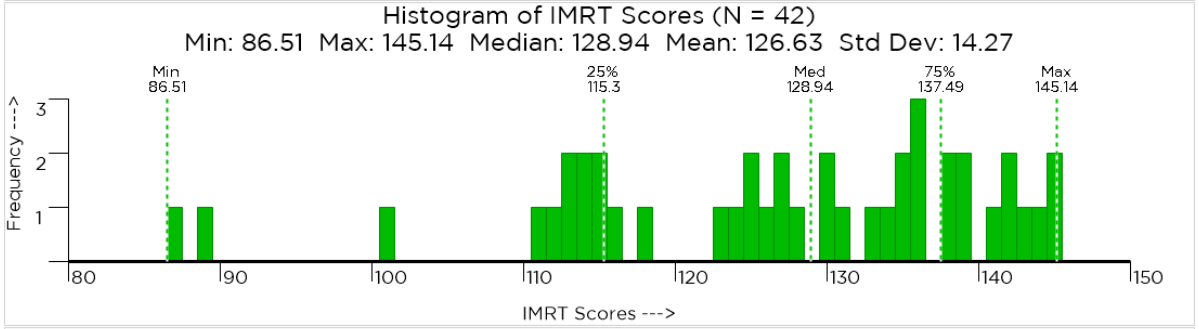
Visual Examples of Variation

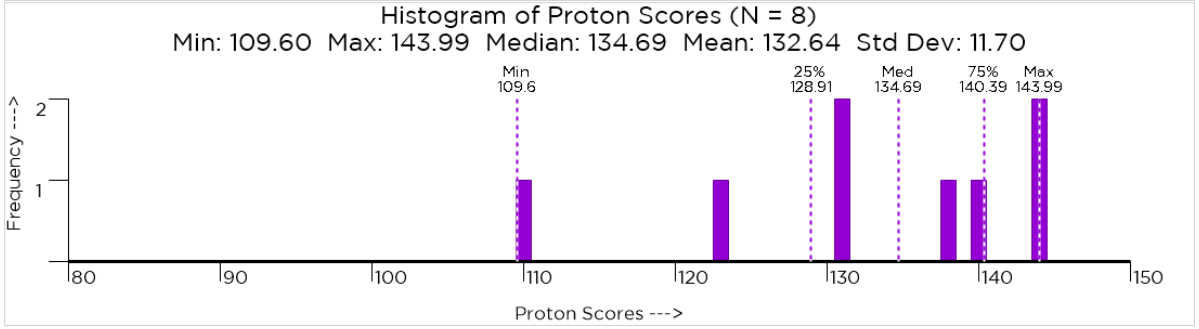


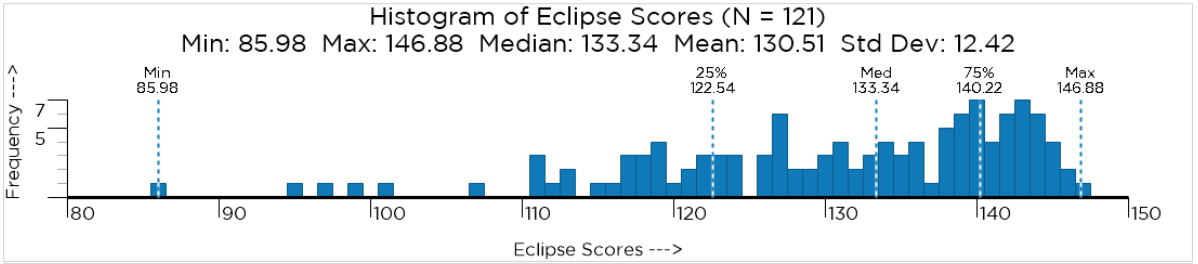


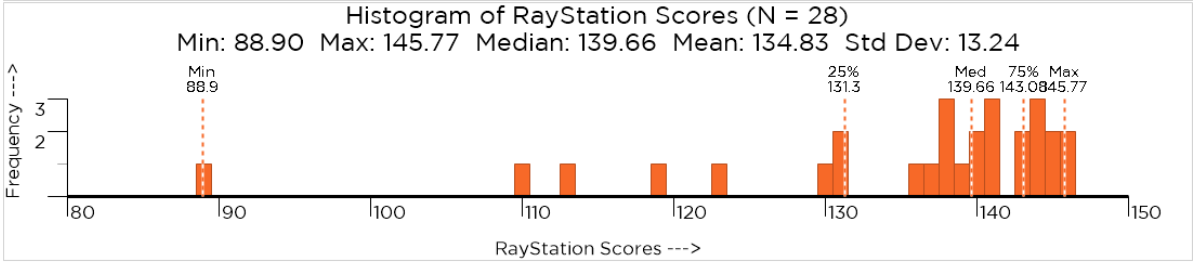

Histogram of VMAT Scores (N = 173)
Min: 85.98 Max: 146.88 Median: 135.61 Mean: 131.56 Std Dev: 12.21
Min 85.98
25% 124.3
Med 135.61
75% 140.66
Max 146.88
Frequency --->
VMAT Scores --->

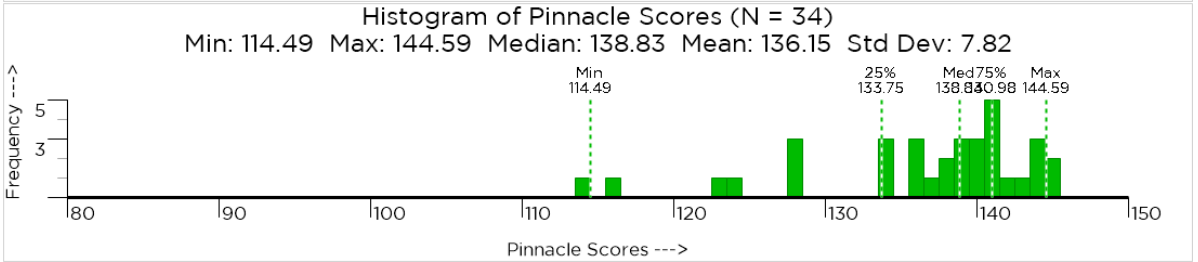

Histogram of Helical Arc Scores (N = 15)
Min: 100.84 Max: 145.77 Median: 133.17 Mean: 130.13 Std Dev: 12.97
Min 100.84
25% 120.04
Med 133.17
75% 141.32
Max 145.77
Frequency --->
Helical Arc Scores --->

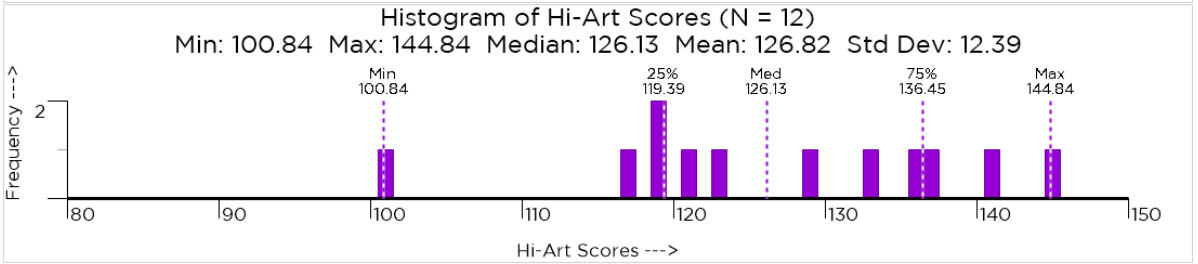

Histogram of IMRT Scores (N = 42)
Min: 86.51 Max: 145.14 Median: 128.94 Mean: 126.63 Std Dev: 14.27
Min 86.51
25% 115.3
Med 128.94
75% 137.49
Max 145.14
Frequency --->
IMRT Scores --->

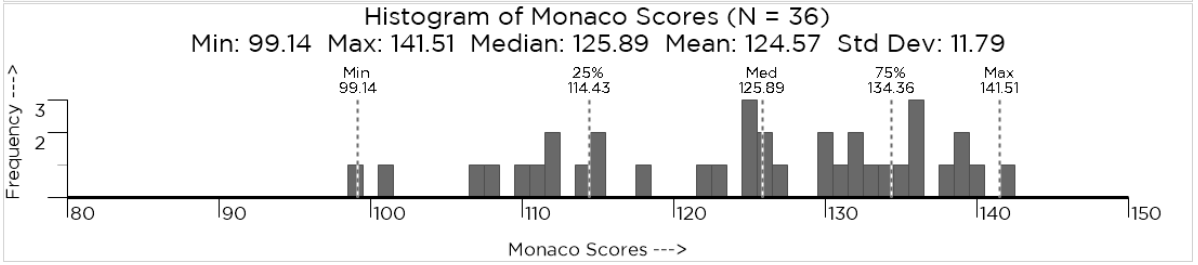

Histogram of Proton Scores (N = 8)
Min: 109.60 Max: 143.99 Median: 134.69 Mean: 132.64 Std Dev: 11.70
Min 109.6
25% 128.91
Med 134.69
75% 140.39
Max 143.99
Frequency --->
Proton Scores --->


Histogram of Eclipse Scores (N = 121)
Min: 85.98 Max: 146.88 Median: 133.34 Mean: 130.51 Std Dev: 12.42
Min 85.98
25% 122.54
Med 133.34
75% 140.22
Max 146.88
Frequency --->
Eclipse Scores --->

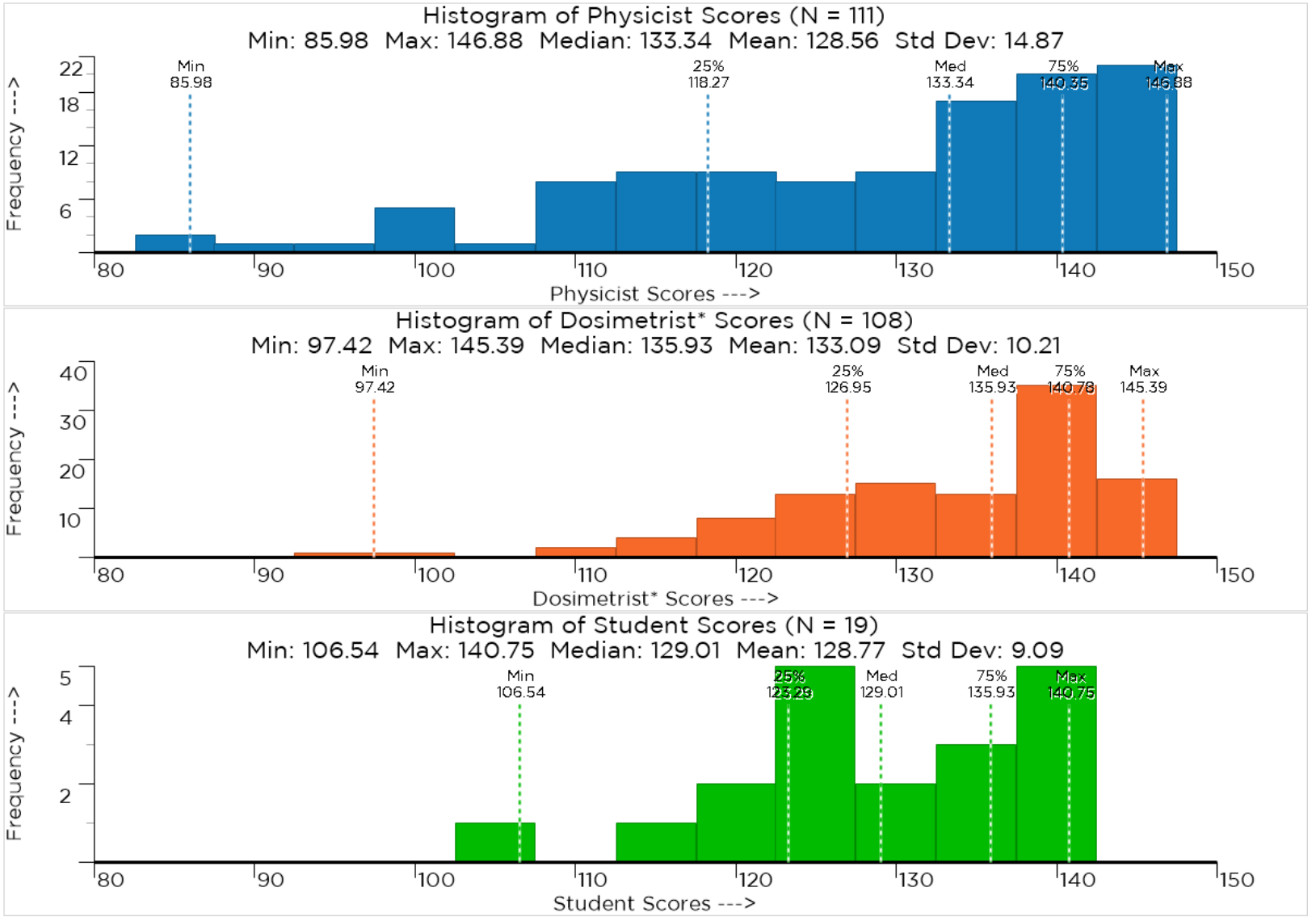

Histogram of RayStation Scores (N = 28)
Min: 88.90 Max: 145.77 Median: 139.66 Mean: 134.83 Std Dev: 13.24
Min 88.9
25% 131.3
Med 139.66
75% 143.08
Max 145.77
Frequency --->
RayStation Scores --->

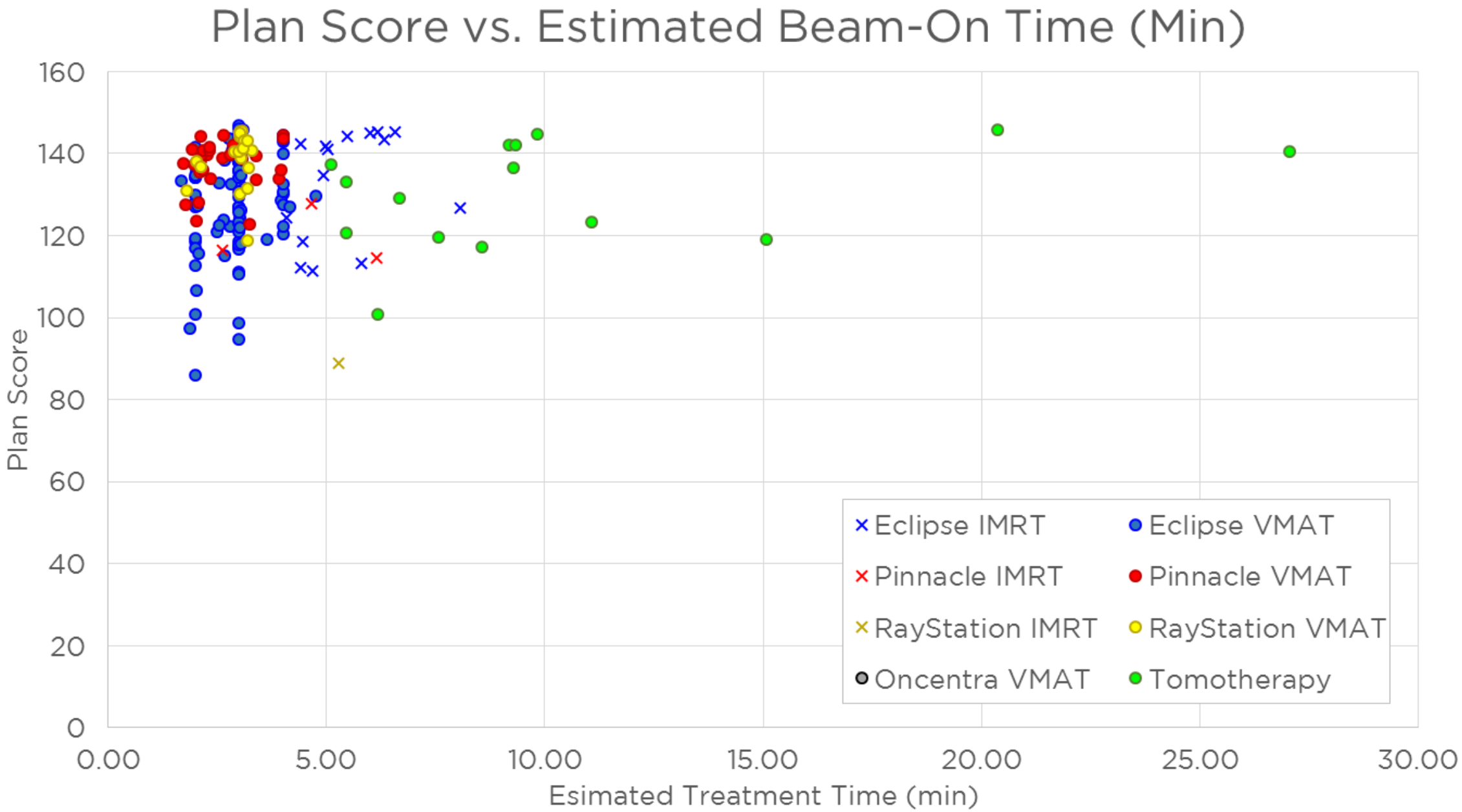

Histogram of Pinnacle Scores (N = 34)
Min: 114.49 Max: 144.59 Median: 138.83 Mean: 136.15 Std Dev: 7.82
Min 114.49
25% 133.75
Med 138.83
75% 140.98
Max 144.59
Frequency --->
Pinnacle Scores --->

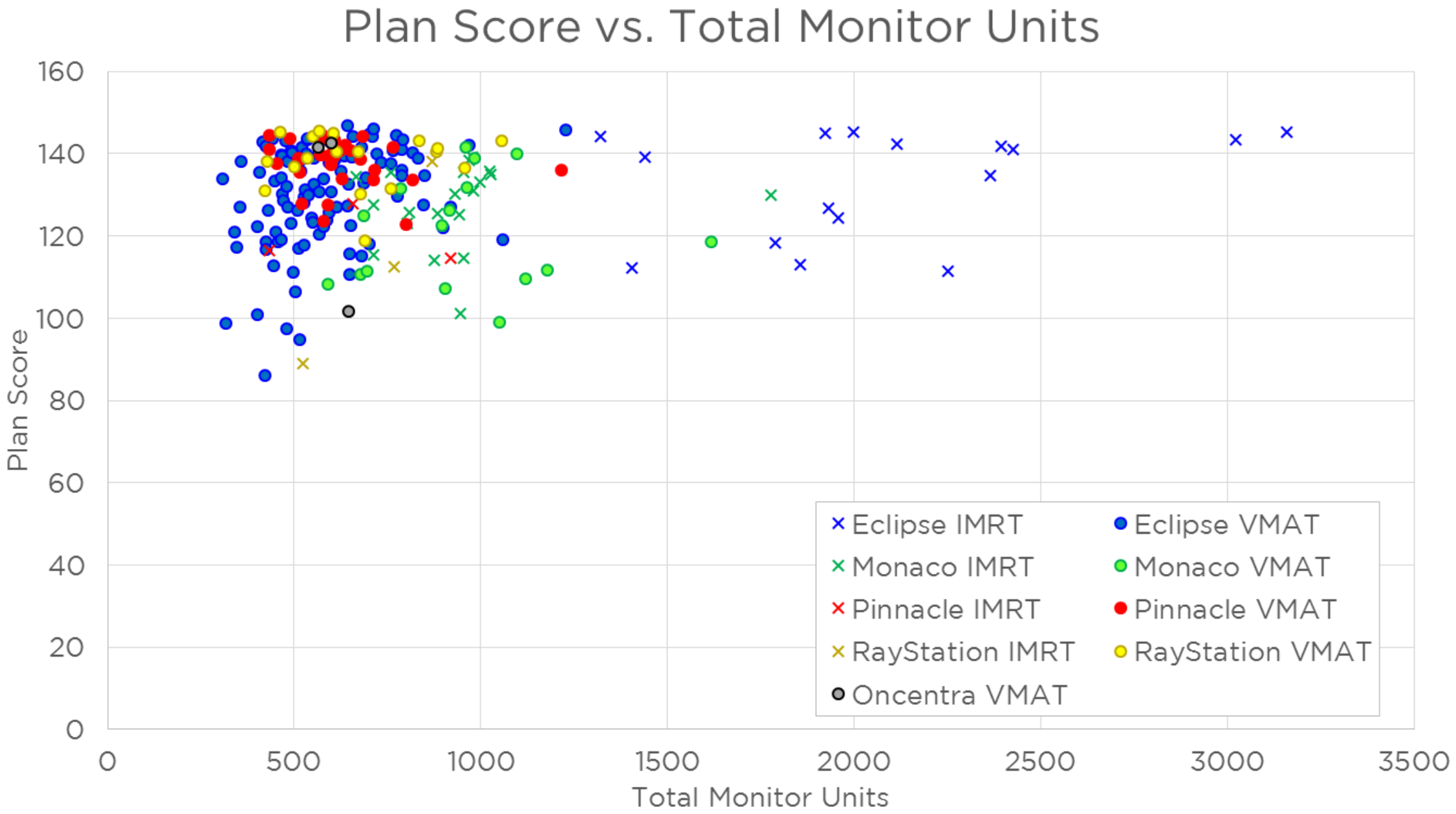

Histogram of Hi-Art Scores (N = 12)
Min: 100.84 Max: 144.84 Median: 126.13 Mean: 126.82 Std Dev: 12.39
Min 100.84
25% 119.39
Med 126.13
75% 136.45
Max 144.84
Frequency --->
Hi-Art Scores --->


Histogram of Monaco Scores (N = 36)
Min: 99.14 Max: 141.51 Median: 125.89 Mean: 124.57 Std Dev: 11.79
Min 99.14
25% 114.43
Med 125.89
75% 134.36
Max 141.51
Frequency --->
Monaco Scores --->

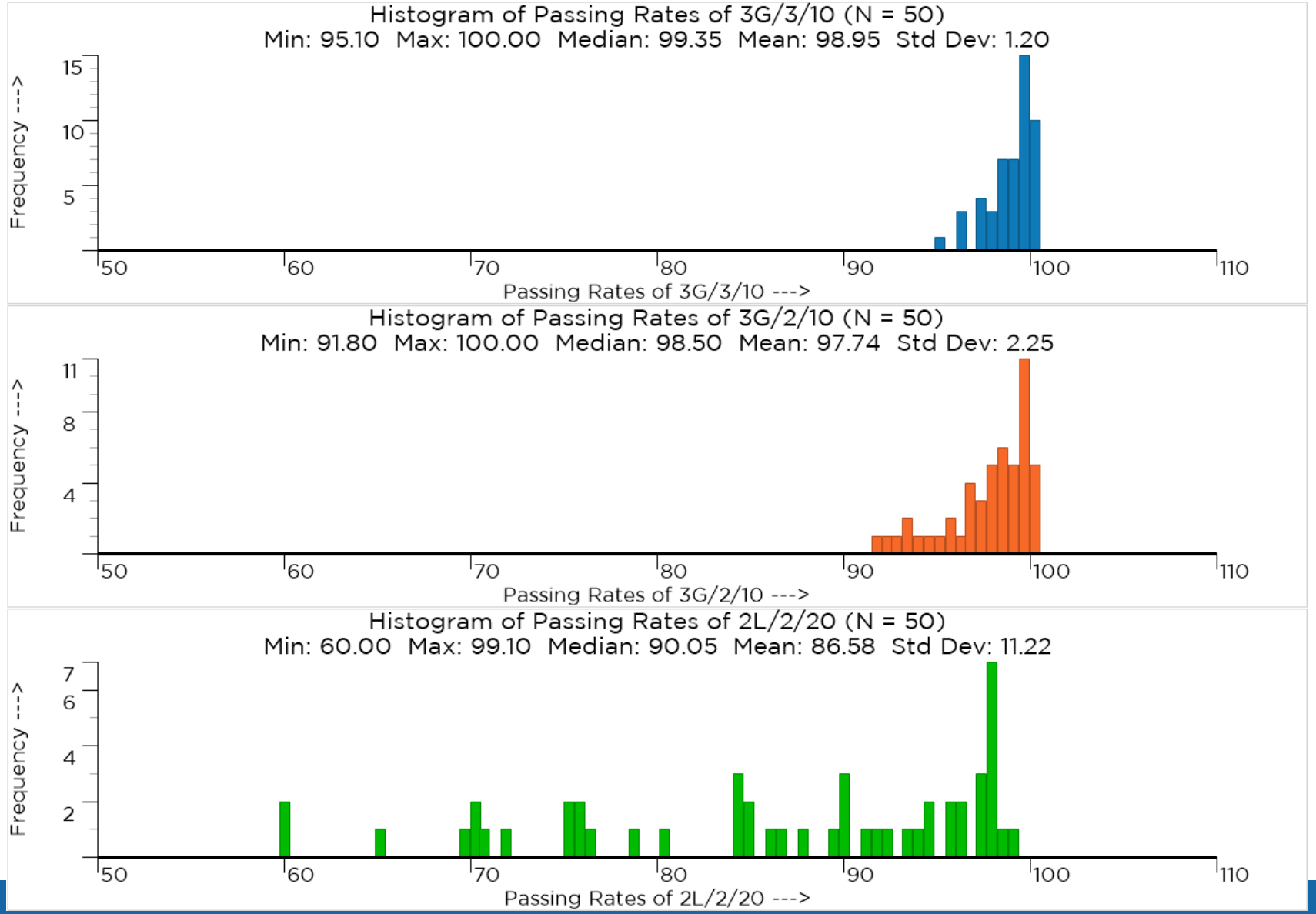
Plan Scores: By Planner Role



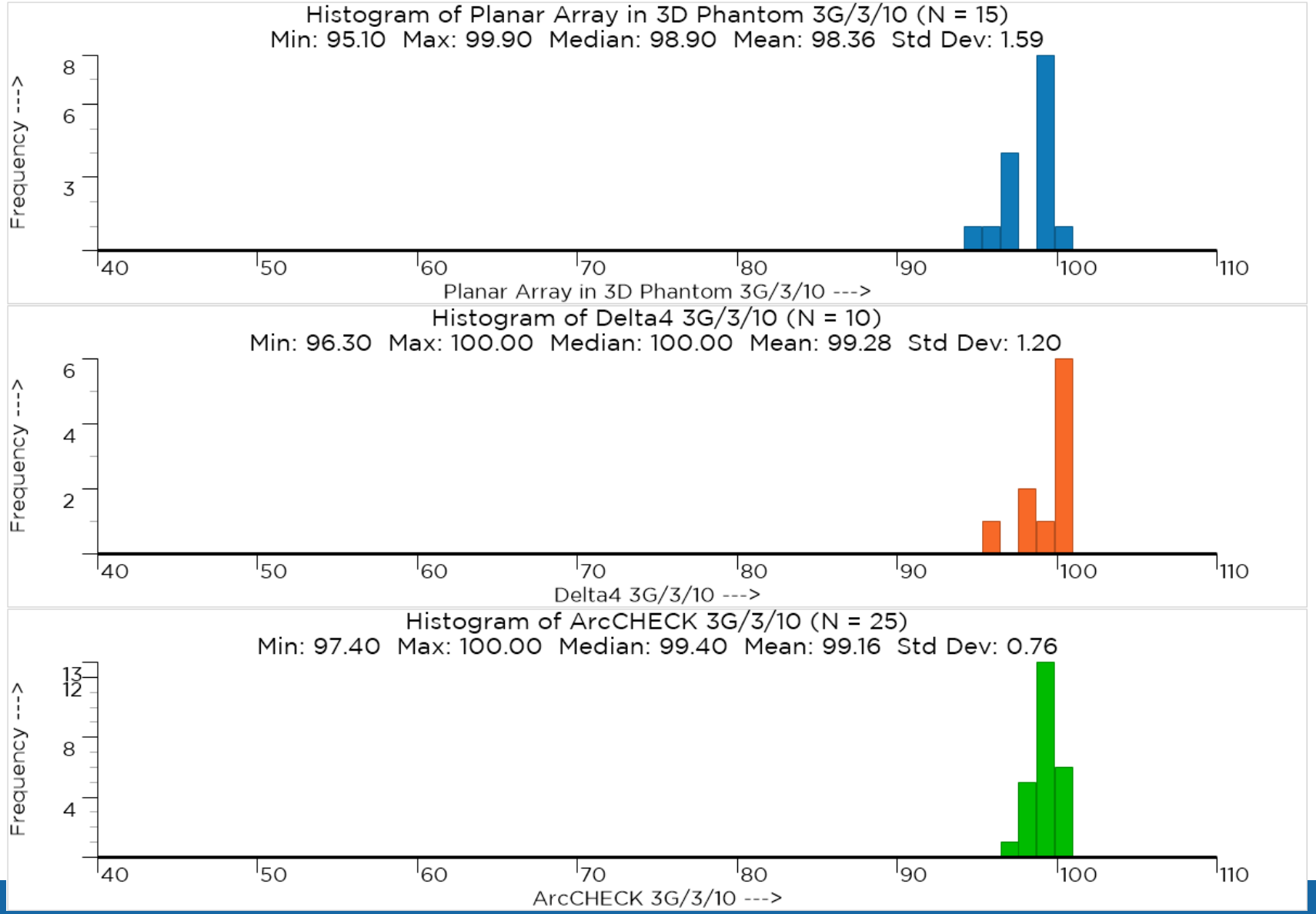

Plan Score vs. Estimated Beam-On Time (Min)
160
140
120
100
80
60
40
20
0
Plan Score
0.00
5.00
10.00
15.00
20.00
25.00
30.00
Esimated Treatment Time (min)
Eclipse IMRT
Eclipse VMAT
Pinnacle IMRT
Pinnacle VMAT
RayStation IMRT
RayStation VMAT
Oncentra VMAT
Tomotherapy


Plan Score vs. Total Monitor Units
Plan Score
Total Monitor Units
0
20
40
60
80
100
120
140
160
0
500
1000
1500
2000
2500
3000
3500
Eclipse IMRT
Eclipse VMAT
Monaco IMRT
Monaco VMAT
Pinnacle IMRT
Pinnacle VMAT
RayStation IMRT
RayStation VMAT
Oncentra VMAT

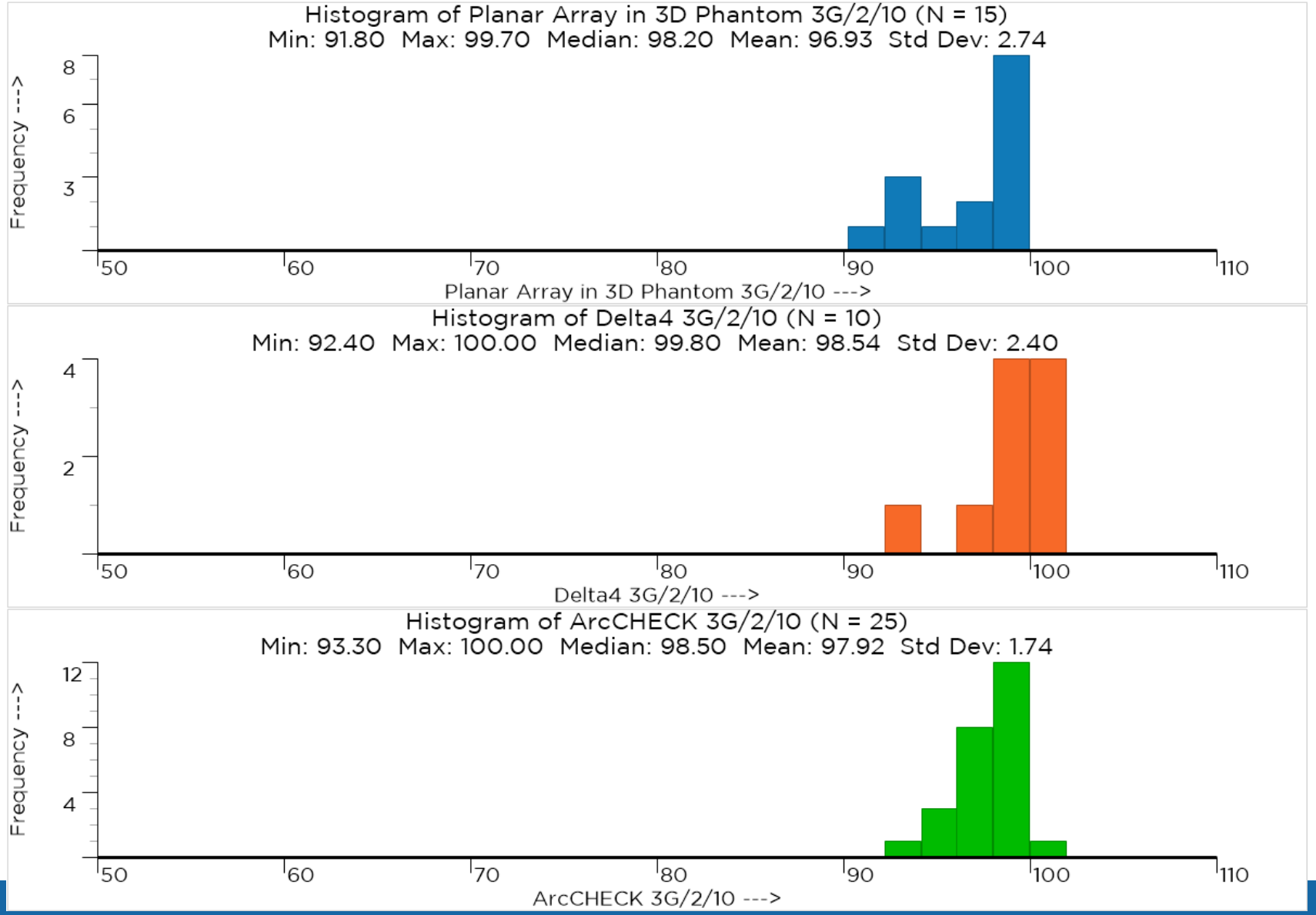
True Composite QA: 3G/3/10 vs. 3G/2/10 vs. 2L/2/20



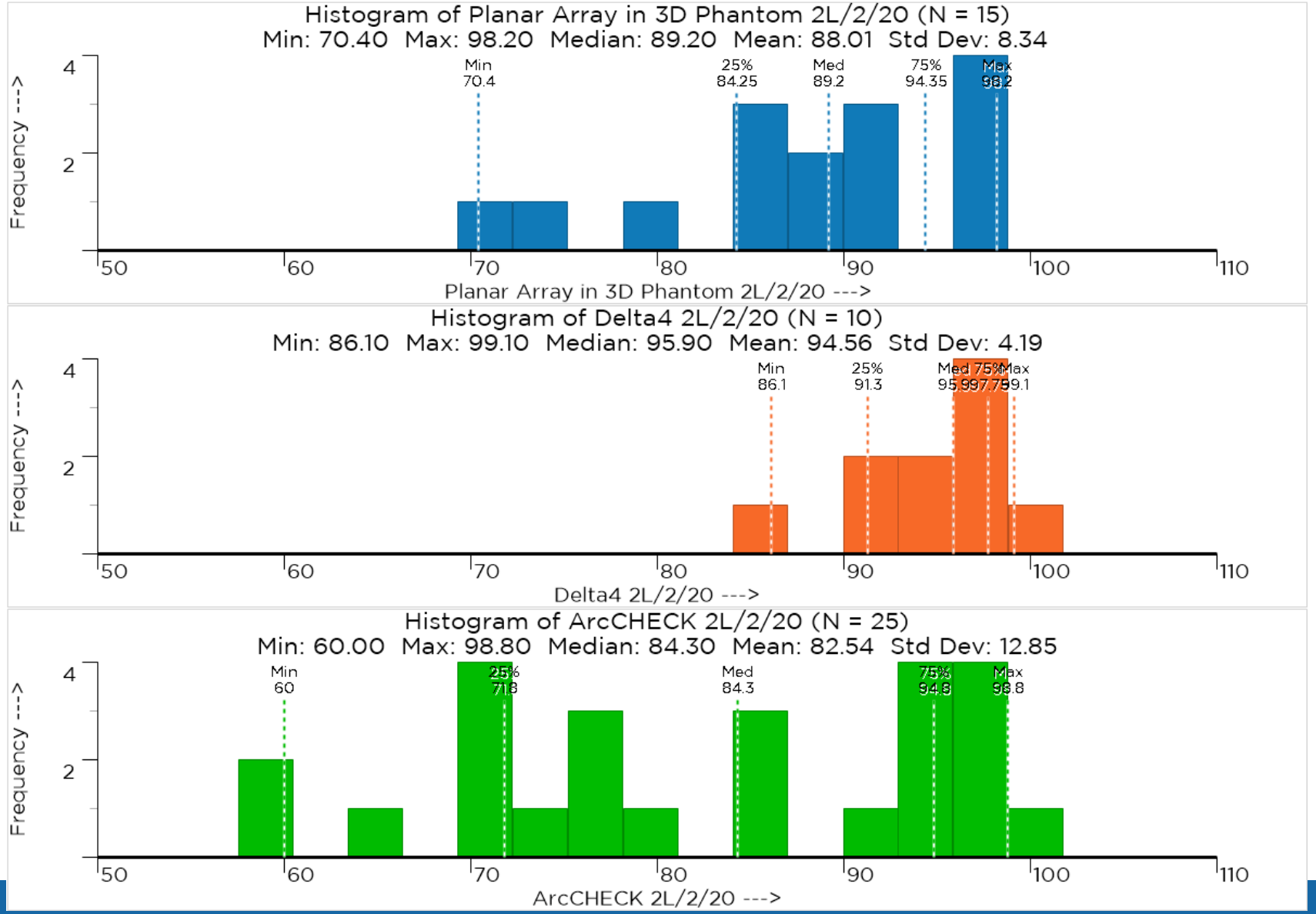
True Composite QA: 3% (Global) / 3 mm / 10%TH Passing Rates

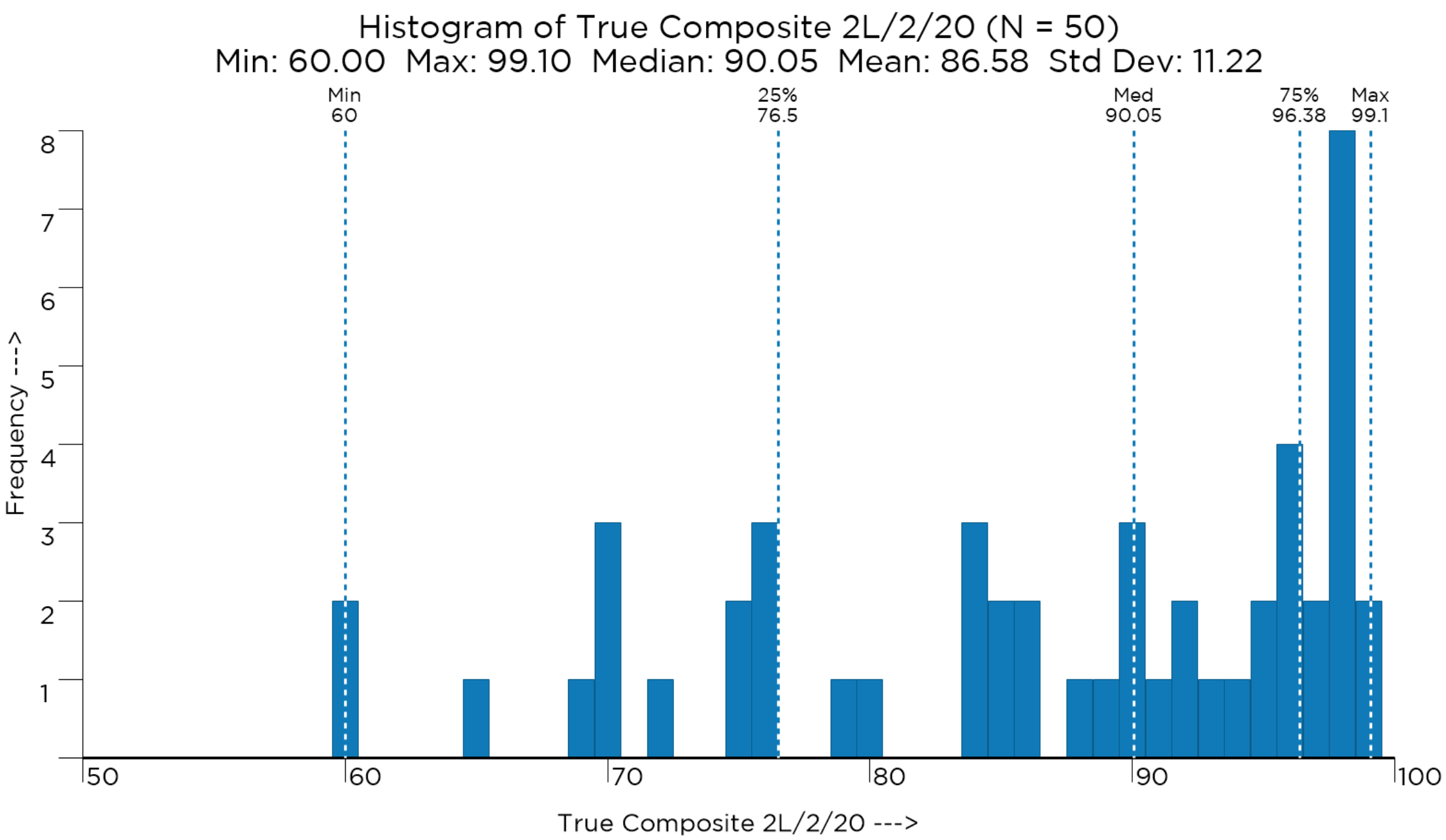


True Composite QA: 3% (Global) / 2 mm / 10%TH Passing Rates



True Composite QA: 2% (Local) / 2 mm / 20%TH Passing Rates




Histogram of True Composite 2L/2/20 (N = 50)
Min: 60.00 Max: 99.10 Median: 90.05 Mean: 86.58 Std Dev: 11.22
Min
60
25%
76.5
Med
90.05
75%
96.38
Max
99.1
Frequency --->
1
2
3
4
5
6
7
8
50
60
70
80
90
100
True Composite 2L/2/20 --->

Sharing of Best Practices

- What?
  - Planners and physicists from around the world who were recognized as “high performers” were contacted.
  - Many agreed to recorded interviews, and others to written interviews, sharing their methods.
  - Cross section of different TPS, modalities, and dose QA methods.
  - These interviews are shared worldwide through ProKnow.
- Where?
  - Sign in to ProKnow, go to the 2017 QADS Plan Study, and select “Learn”
  - Videos are embedded and documents are downloadable.
- For Who?
  - Everybody. Worldwide. For free.
  - You do not need to have participated in the plan study to access these learning materials.

Conclusions

Conclusions (TPS)

- All TPS studied were able to produce high quality plans.
  - 6 (of 6) TPS produced plans in the top 25% of plan quality.
  - 4 (of 6) TPS produced plans in the top 10% of plan quality.
- All TPS studied showed high variability in plan quality distribution.
- This suggests there would be high value in training, i.e. propagating best practices to help remove the low quality/low score tail.

Conclusions (Modality & Complexity)

- VMAT vs. IMRT vs. Tomotherapy
  - No statistical difference based on modality.
  - The VMAT plans were significantly more efficient in terms of time and monitor units than both IMRT and helical tomotherapy.
- Protons vs. Photons
  - Protons were a small sample size (8 out of 238 plans). Based on those, there was no measured advantage of protons over photons.
  - The max, 75th percentile, and median were:
    - 146.88, 140.66, and 135.61 (IMRT, VMAT, tomo)
    - 143.99, 140.39, and 134.69 (PROTONS)
- No correlation of plan quality to total MU (or time)
  - Many efficient (low MU) plans were very high quality.
  - Many inefficient (high MU) plans were lower quality.

Conclusions (True Composite Dose QA)

- Justification for aggressive benchmarks
  - Adoption of using more stringent criteria (2L/2/20) and tighter tolerances (> 95%) is **justified**, even for this sufficiently complex head & neck study.
    - Top quartile of dose QA performances showed passing rates > 95% for the stringent 2% (local normalization) / 2 mm criteria
    - Median 2 (local) / 2 mm passing rate was 90%
    - Was this sample biased towards high performers? Maybe, but when benchmarking that is fine, perhaps even preferred.
- Use of standard patient datasets for benchmarking
  - In terms of TPS commissioning, there is great value to the industry in using: 1) standard patient datasets, 2) object plan quality measures, and 3) consistent QA methods.
  - We should pursue similar benchmarking studies for the remaining MPPG 5.a datasets.