Beyond auditing: What we have learned from phantom credentialing for clinical trials

Stephen F. Kry PhD

Radiological Physics Center, The University of Texas MD Anderson Cancer Center, Houston, TX

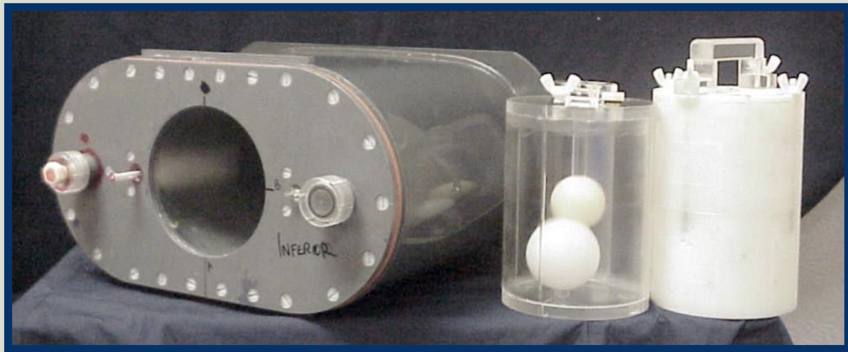
Spring Clinical AAPM Meeting
2014

RPC

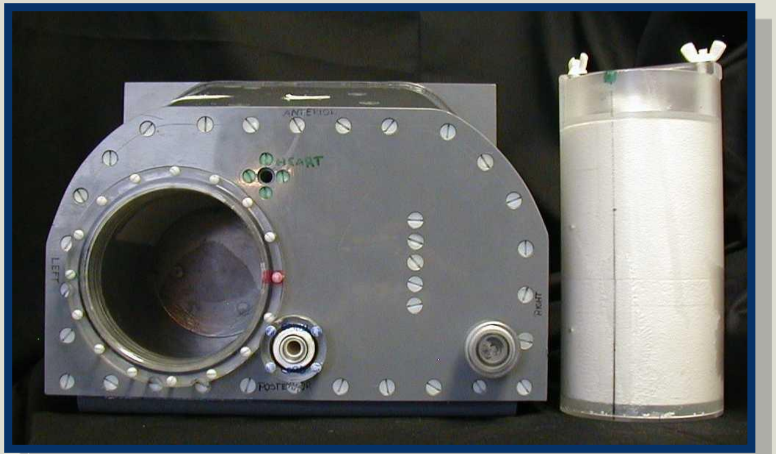
- Radiological Physics Center is scientific agency that supports NCI-run clinical trials
  – Verify that institutions are delivering the dose they believe they are delivering
- Have been doing this since 1968
- Monitor >1800 RT facilities
  – Many tools – mailable output checks, site visits, phantoms, patient chart dose recalculations

- Phantoms!

The RPC Phantom Family



10 prostate phantoms (IMRT)

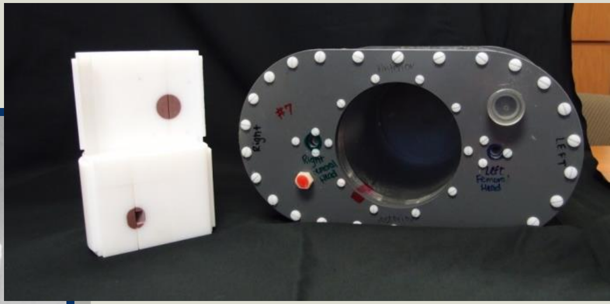


13 lung phantoms

25 H&N phantoms (IMRT)

8 Spine phantoms

16 SRS phantoms



8 liver insert

RPC activities - Phantoms

- Mail a phantom to an institution
  - Includes target(s) and dosimeters
- The institution treats it like a patient
  - Sim, plan, setup, treat
- The RPC analyses the results and compares the measured dose distribution to the institution’s TPS calculation
- Large history of irradiations
  - >3000 phantoms

Phantom Audits

- Can an institution deliver the dose they intended
- Pass ⇨ participation in clinical trials
- What else have we learned?
  – Lung phantom: Heterogeneous calculations
  – H&N phantom: phantom versus IMRT QA
  – Proton phantoms: material stopping power

Lessons from the Lung phantom

- Different algorithms show different levels of dose agreement in the RPC lung phantom.

Physics Contribution

Algorithms Used in Heterogeneous Dose Calculations Show Systematic Differences as Measured With the Radiological Physics Center's Anthropomorphic Thorax Phantom Used for RTOG Credentialing

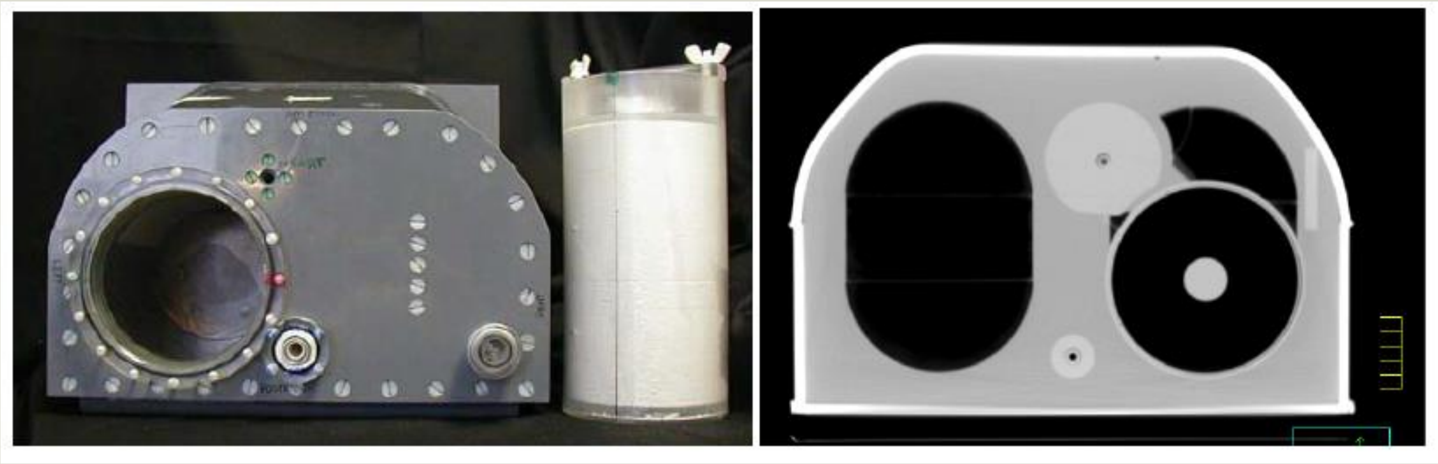
Stephen F. Kry, PhD,* Paola Alvarez, MS,* Andrea Molineu, MS,* Carrie Amador, BS,* James Galvin, DSc,† and David S. Followill, PhD*

*Radiological Physics Center, Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas; †American College of Radiology/Radiation Therapy Oncology Group, Philadelphia, Pennsylvania

Received Jan 10, 2012, and in revised form Aug 24, 2012. Accepted for publication Aug 29, 2012

Heterogeneous conditions:

- Lung phantom
  - 2 TLD in center of lung target (3 cm x 5 cm)
  - Film in 3 planes
- Homogeneous results within 1% (Ibbott)



Initial work: low quality algorithms

- Cyberknife pencil beam algorithm

TLD	Calculated		Measured		% Difference
	Dose (cGy)	% SD	Avg. Dose (cGy)	% SD	
PTV Sup.	610.3	1.1%	533.6	0.5%	-13.4%
PTV Inf.	592.9	2.4%	517.3	0.8%	-13.6%
Avg. PTV	**601.6**	**2.1%**	**525.4**	**1.8%**	**-13.5%**
Cord	42.1	2.2%	27.8	1.3%	-40.9%
Heart	63.6	5.6%	45.7	2.3%	-32.8%

Thoracic dose calculations

- Homogeneous and low quality heterogeneous dose calculations (e.g., Batho-corrected pencil beam) are highly inaccurate and inconsistent.
- These algorithms are not allowed in NCI-sponsored clinical trials involving the lung
- AAPM minimum practice statement
  - TG-244: Commissioning and QA of TPS in EBRT

b. Recommendations

i. To produce acceptable dosimetric accuracy in highly heterogeneous media (particularly in lung), an algorithm comparable to C/S, CC, MC, or GBBS-based dose calculation algorithm must be used.

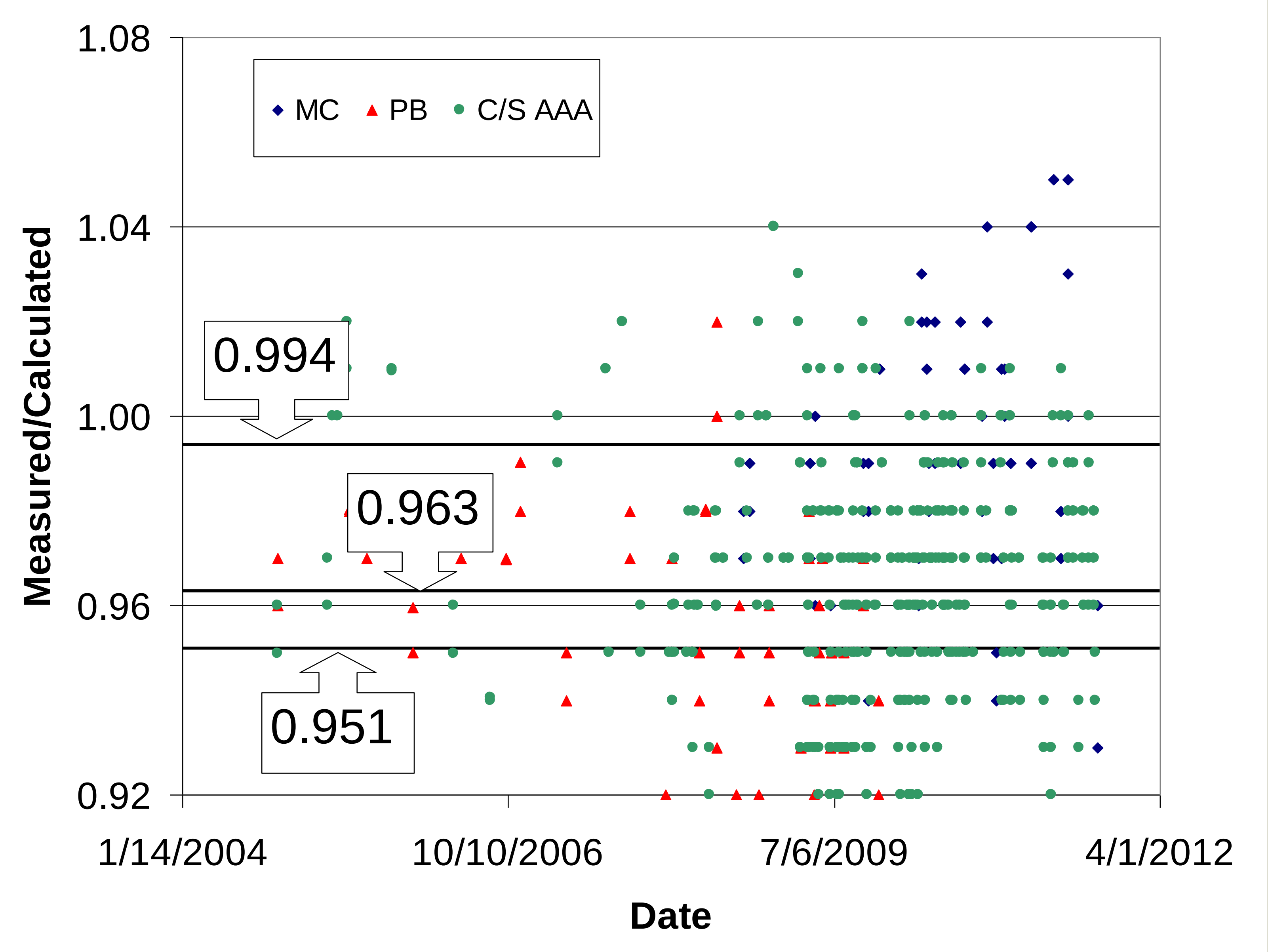
- Convolution-Superposition/AAA algorithms are generally considered accurate

Irradiations

- In this study:
  - 304 irradiations
  - 6 MV irradiations
  - IMRT or 3D CRT
  - Moving or static
  - Various algorithms
  - All used heterogeneity corrections
- Evaluate
  - TLD dose (vs TPS)
  - Planar agreement
    - DTA or gamma

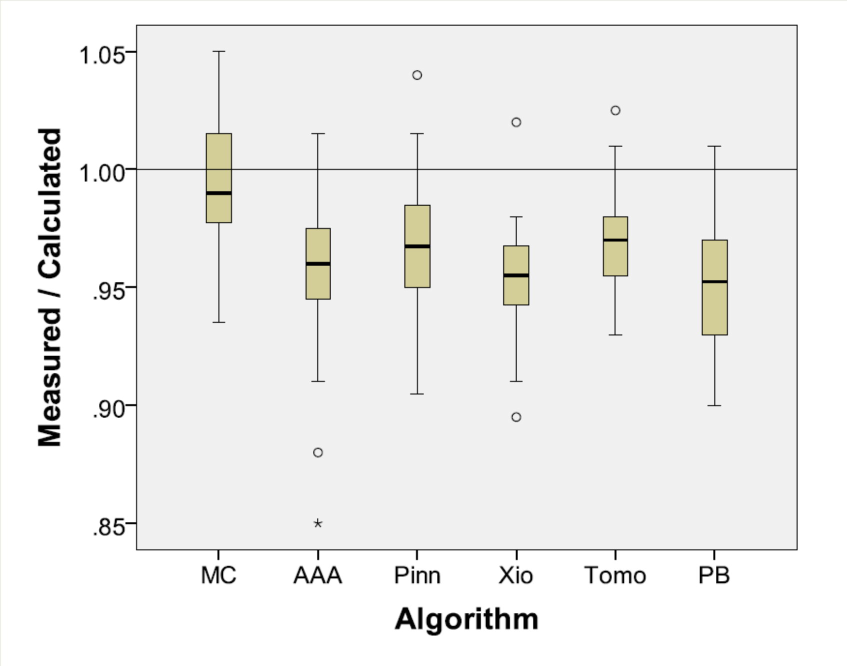
Algorithm Class	Commercial Product	N	
Monte Carlo		**32**	
	MultiPlan		25
	BrainLab		3
	CMS Monaco		2
	In-house		1
CS/AAA		**236**	
	Eclipse AAA		98
	Pinnacle CS		90
	CMS Xio CS		23
	Tomotherapy CS		25
Pencil Beam		**36**	
	Eclipse PBA		28
	Elekta PrecisePLAN		2
	BrainLab		2
	CMS Xio		1
	In-house		3

TLD Measurement vs TPS calculation



TLD Dose Findings

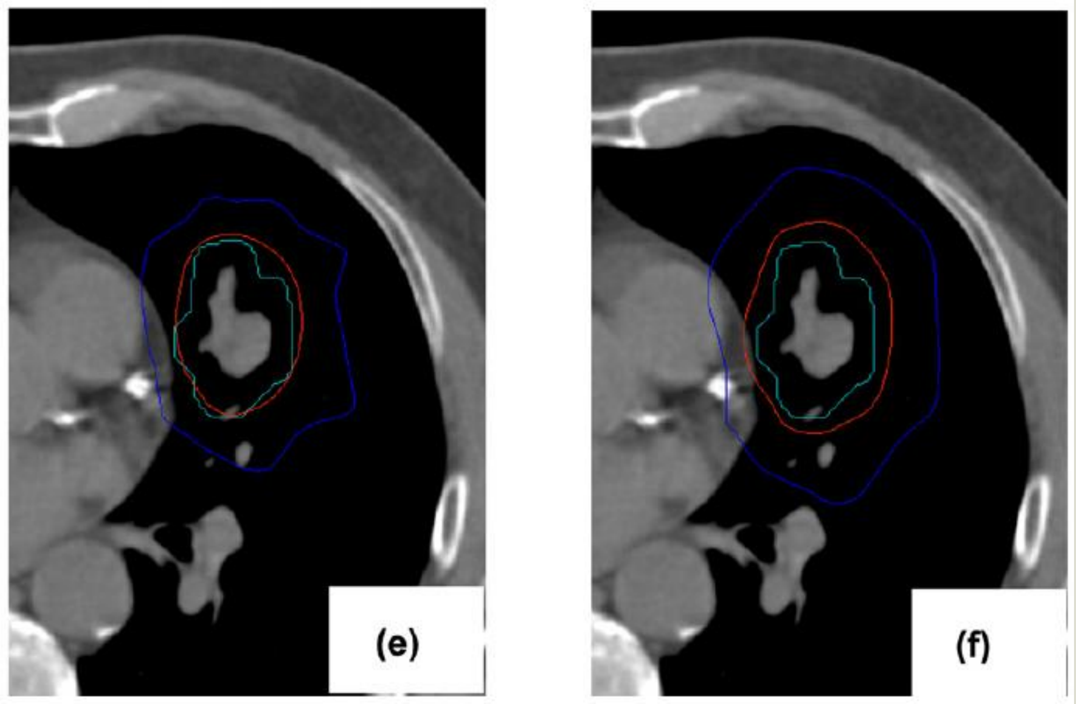
- Measured doses systematically lower than calculated doses for C/S AAA algorithms ($p<0.0001$)
- No significant difference between C/S AAA algorithms



- For C/S AAA algorithms:
- No significant difference between IMRT (mean=0.963) and 3D CRT (mean=0.964) irradiations ($p=0.7$)
- No significant difference between moving (mean=0.961) and static (mean=0.964) irradiations ($p=0.5$)
- No significant trend versus irradiation date ($p=0.2$)

Systematic calculation discrepancy

- Overestimation of dose with C/S AAA (3.7%)
- Dose to center of target
- Other studies showing similar results
  - Monte Carlo lung plans hotter than C/S
  - Larger 100% isodose volume



Physics Contribution

Dosimetric Verification Using Monte Carlo Calculations for Tissue Heterogeneity-Corrected Conformal Treatment Plans Following RTOG 0813 Dosimetric Criteria for Lung Cancer Stereotactic Body Radiotherapy

Jun Li, Ph.D.,* James Galvin, D.Sc.,* Amy Harrison, M.Sc.,* Robert Timmerman, M.D.,† Yan Yu, Ph.D.,* and Ying Xiao, Ph.D.*

Int J Radiation Oncol Biol Phys, Vol. 84, No. 2, pp. 508–513, 2012

What does this mean?

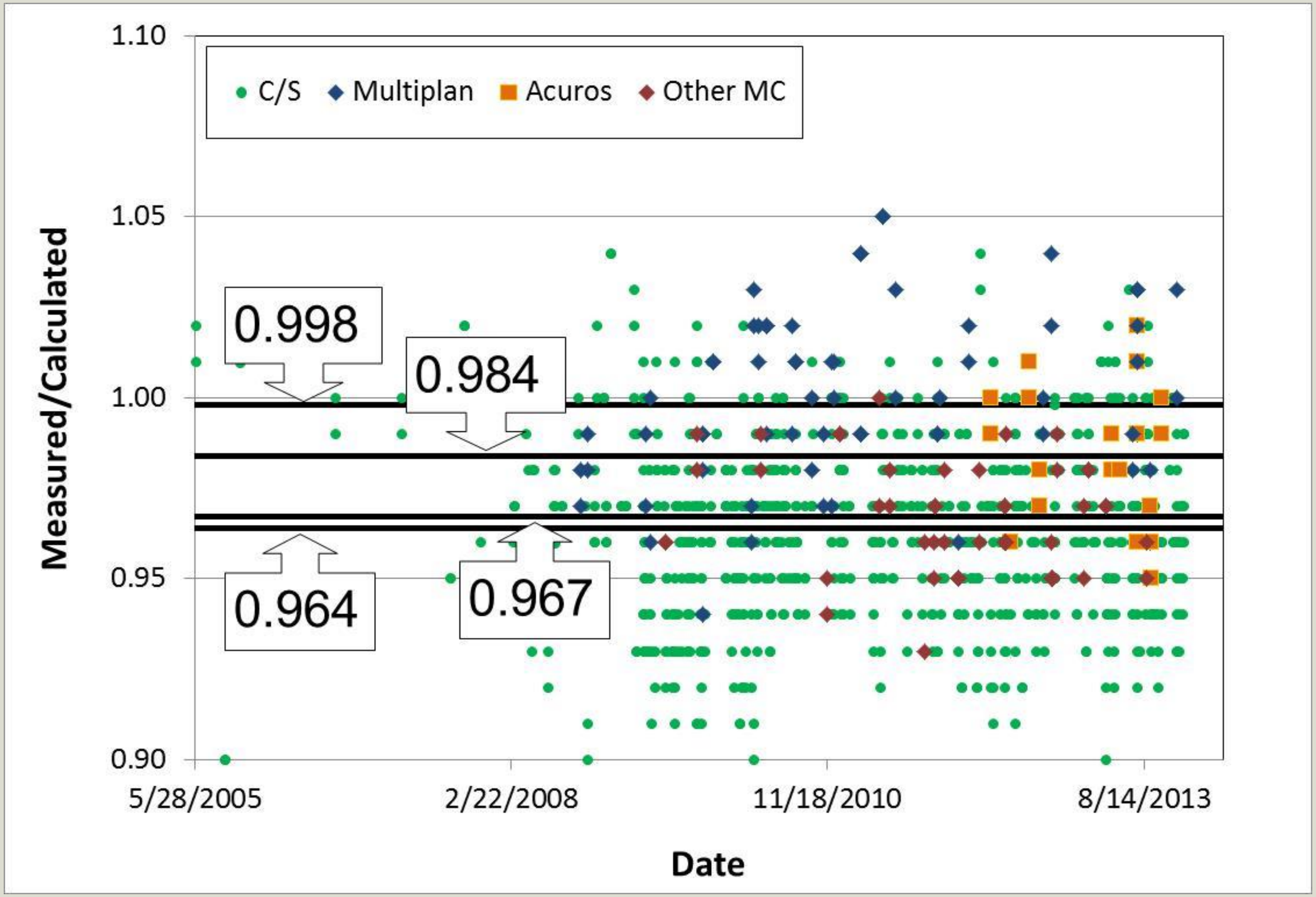
- Issue for dose calculation accuracy
  (AAPM TG-65 goal: 1-2%)

- Potentially issue for dose reporting/prescribing

Update to these results

- Another 1.5 years of phantom results
- Acuros (n=13)
  - Unique radiation transport algorithm
- More Monte Carlo (n=57)
  - Multiplan (n=34)
  - BrainLab (n=12)
  - Monaco (n=10)
- More C/S (n=457)

Updated results



Monte Carlo results are not consistent……

Update summary

- More variability than expected between different algorithms
- Acuros different than MC or C/S
- Monte Carlo results not uniformly consistent
- Why so much difference???

What to do?

- Note that we see some inconsistencies
- Understand where this arises in clinical practice, and how much difference there is

- Pressure manufacturers to improve dose calculation accuracy

Lessons from the H&N phantom

- How do phantom results compare to IMRT QA results?

- Does IMRT QA predict RPC phantom results?

IMRT QA

- IMRT QA comes in many flavours
  - Detectors, detector geometries, delivery geometry, tolerances, analysis techniques, ROI selected, analysis software and on and on…..
- All flavours are used. None are repeated
- At the end of the day, they should evaluate a treatment plan
  - Are you delivering what you think you are?

IMRT QA

- We collected institutional IMRT QA results for H&N phantom plans
  - Compare them with phantom results
- Abstracted 1005 H&N phantom results and corresponding IMRT QA results
- Excluded
  - No/unintelligible IMRT QA results
  - Adjusted MU between IMRT QA and phantom

Methods

- 855 records
  - 122 failed phantom irradiation
- First sorting:
  - Considered to pass IMRT QA unless stated otherwise
- Truth tables to calculate sensitivity and specificity of IMRT QA relative to RPC phantom

Results

- Institution declared “failed” IMRT QA

		RPC	
		Fail	Pass
Inst QA	Fail	2	3
	Pass	120	730

- Sensitivity: 2 (±1)% (Failing plan identified as failing)
- Specificity: 99.6 (±0.2)% (Passing plan identified as passing)

Results

- Re-evaluate institution IMRT QA
  - >3% absolute dose disagreement
  - <90% of pixels passing at least 3%/3mm

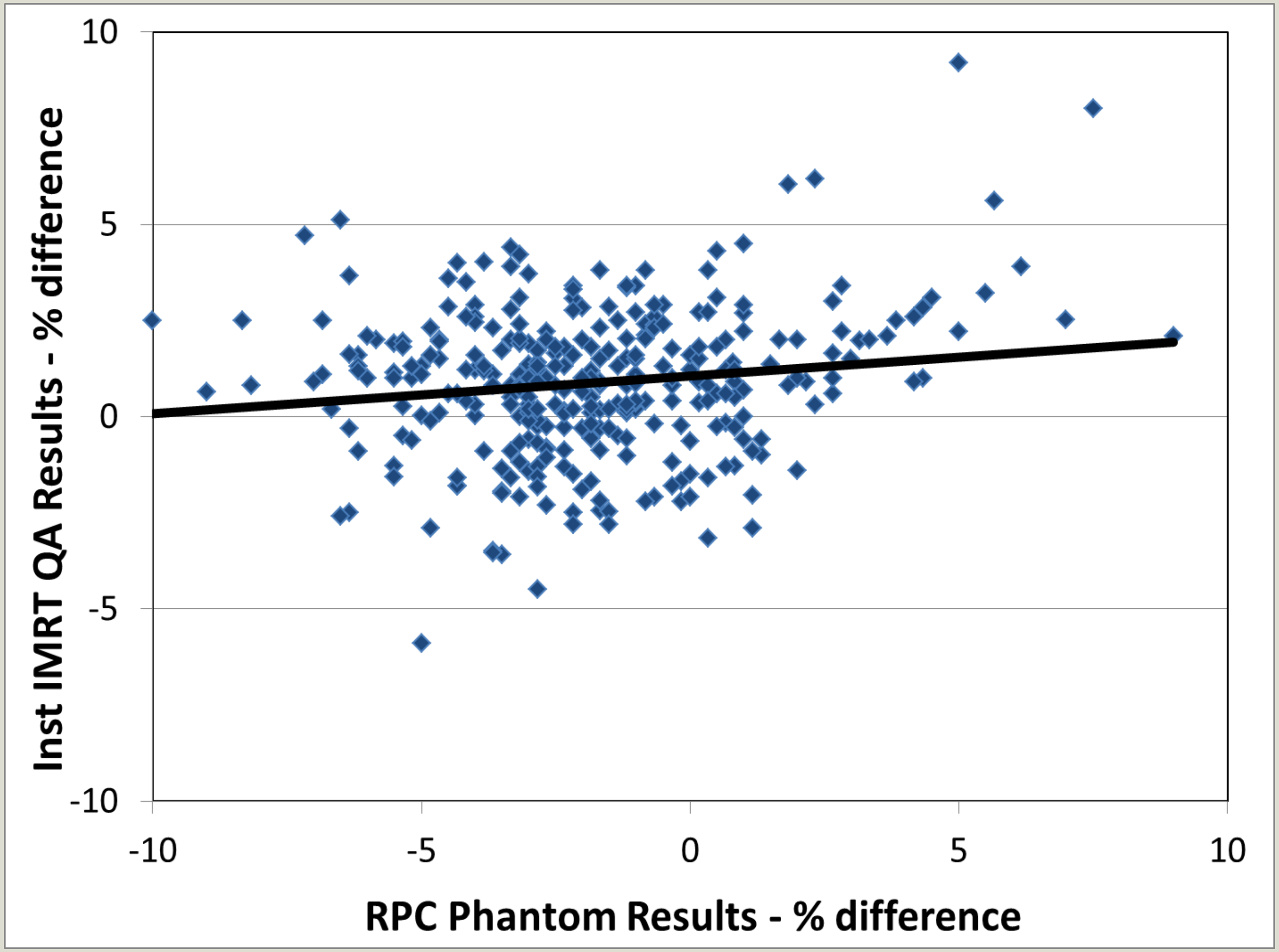
		RPC	
		Fail	Pass
Inst QA	Fail	19	57
	Pass	84	585

- Sensitivity: 18(±4)% (Failing plan identified as failing)
- Specificity: 91(±1)% (Passing plan identified as passing)

Summary of all Results

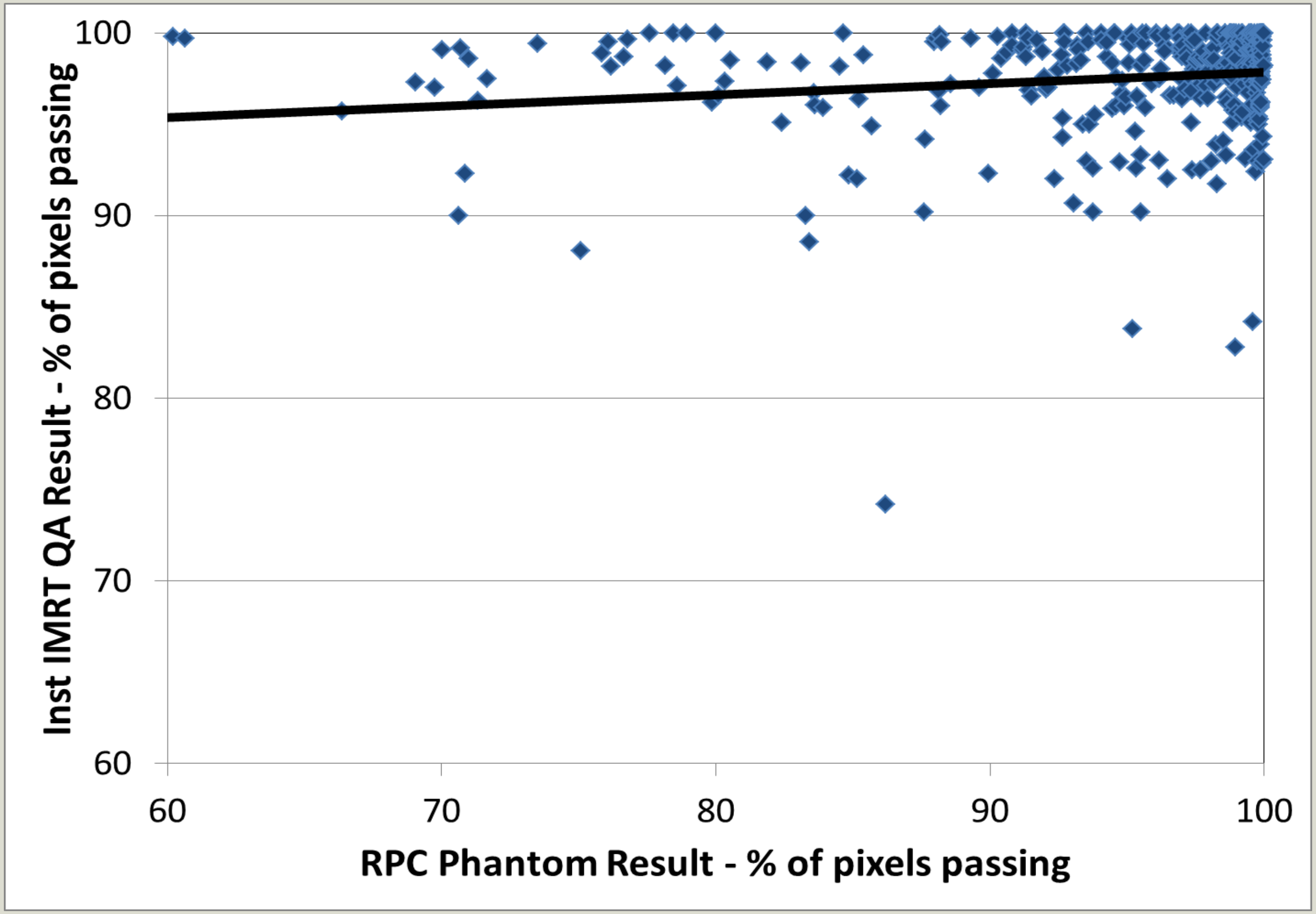
	Number	Sensitivity in % (st. dev.)	Specificity in % (st. dev.)
ALL RESULTS			
Institution claim	855	2 (1)	99.6 (0.2)
Re-evaluated	745	18 (4)	91 (1)
DEVICE			
Ion chamber + planar	91	54 (14)	79 (5)
Ion chamber	325	25 (6)	90 (2)
Film	71	33 (16)	82 (5)
MapCheck	322	14 (5)	94 (2)
MODE			
Absolute	295	3 (3)	94 (1)
Relative	97	21 (9)	91 (3)

Ion chamber versus average TLD



- P = 0.006, R^2 = 0.02

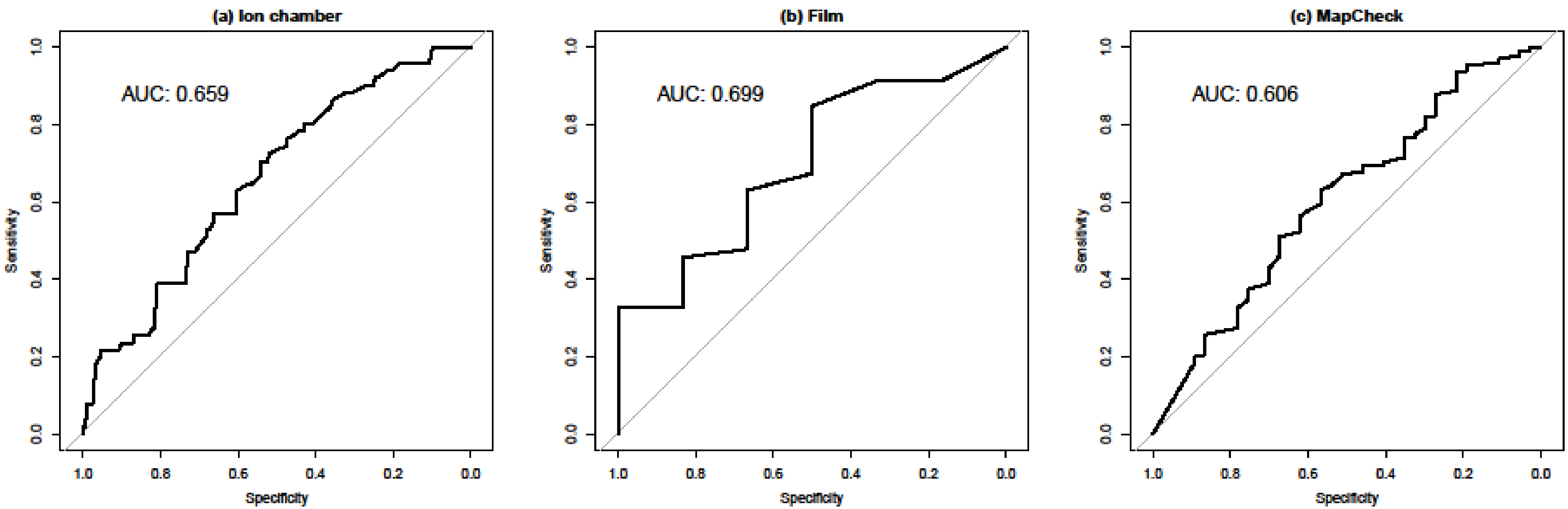
Planar detector versus average film



- Just 3%/3mm for any planar device
- $P = 0.002$, $R^2 = 0.05$

Is this a criteria problem?

- AUC – all devices equal (poor)
- No good criteria that has good sensitivity and specificity
- 50% sensitivity
  - 2% ion chamber, 97% of pixels passing (3%/3mm)



Conclusions

- In-house IMRT QA does not well predict external phantom audit results
  - Phantom failure rate ~20%
  - In house IMRT QA failure rate ~3%
    - Dong IJROBP 2003, Fenoglietto Radiat Oncol 2011
- True for all devices and criteria
  - Some criteria better than others
- We need to better understand our QA processes
  - Why don’t these two tests for QA agree better?
  - What QA device/techniques are superior

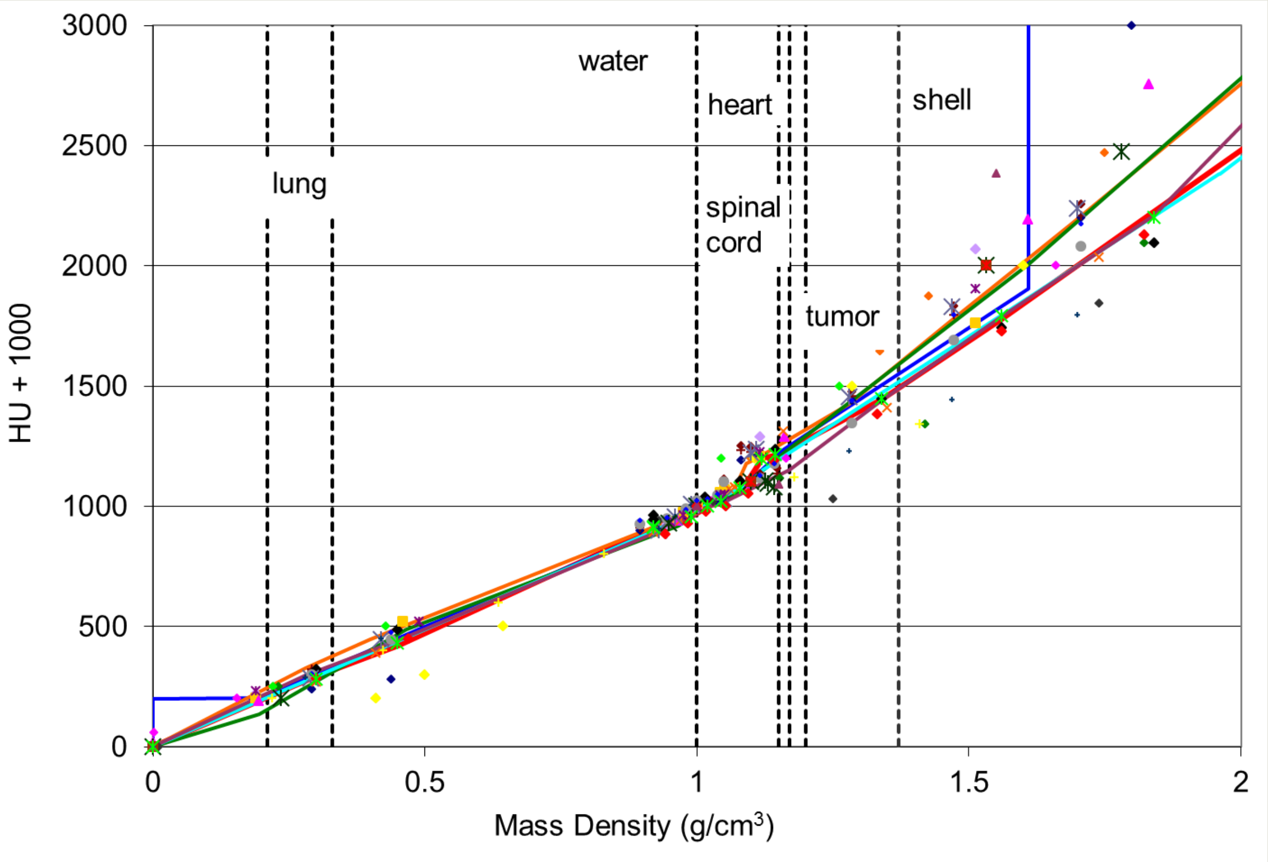
Lessons from Proton phantoms

- What do you mean:

“Proton equivalent”?

In photons

- Lots of plastics behave well
  - Fall on the HU:ED curve

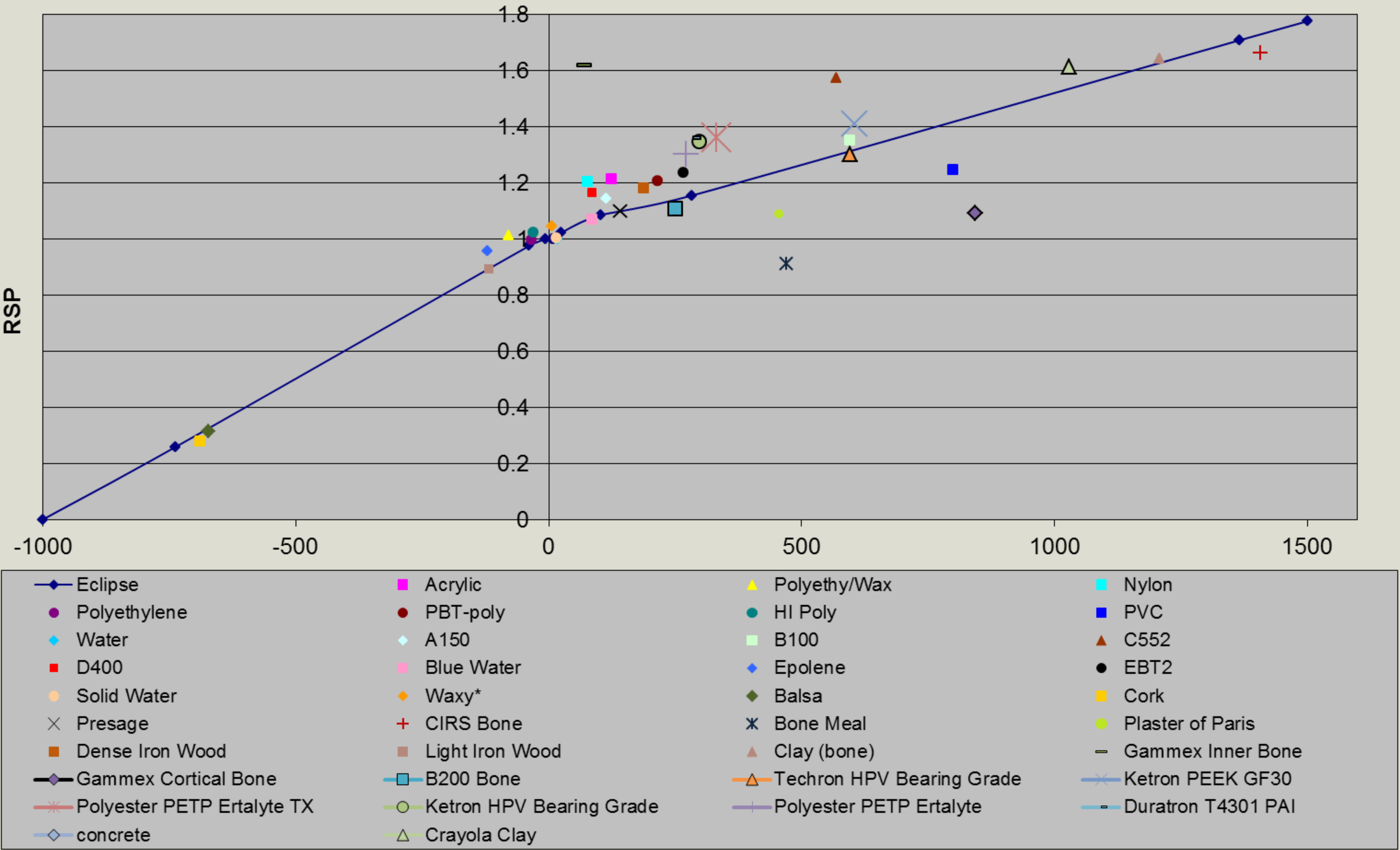


HU/ RSP Data Collection

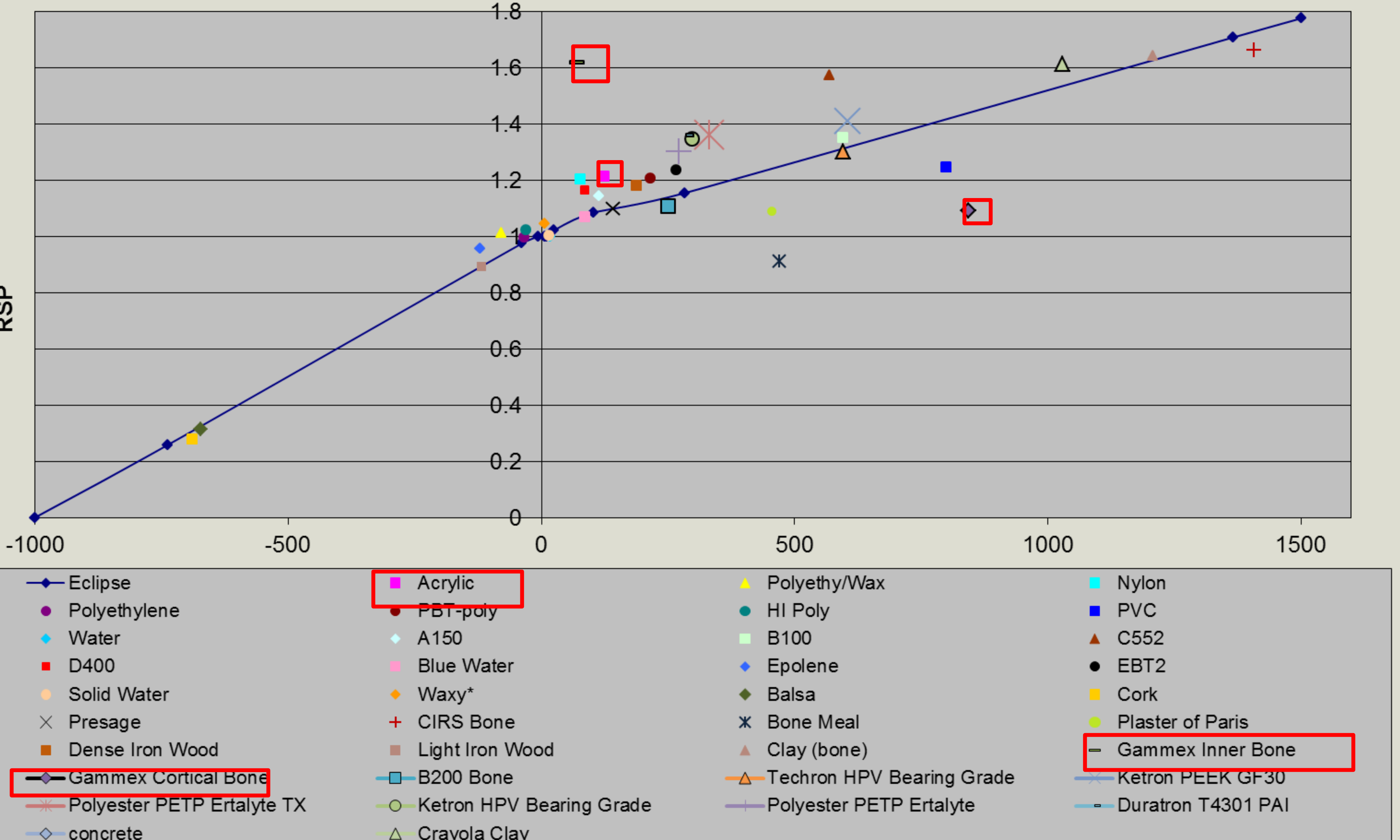
- Based on Moyers et. Al, “Ion Stopping Powers and CT Numbers”
- CT imaging of materials at 120 kVp, 120 mAs, 48cm diameter FOV, slice thickness of 5mm
- HU measurement using Eclipse
- Proton RSP measured at 160 MeV and 250 MeV

$$RSLP = \frac{R_{80,w} - R_{80,m}}{t_m}$$

Stopping Power vs. HU Curve

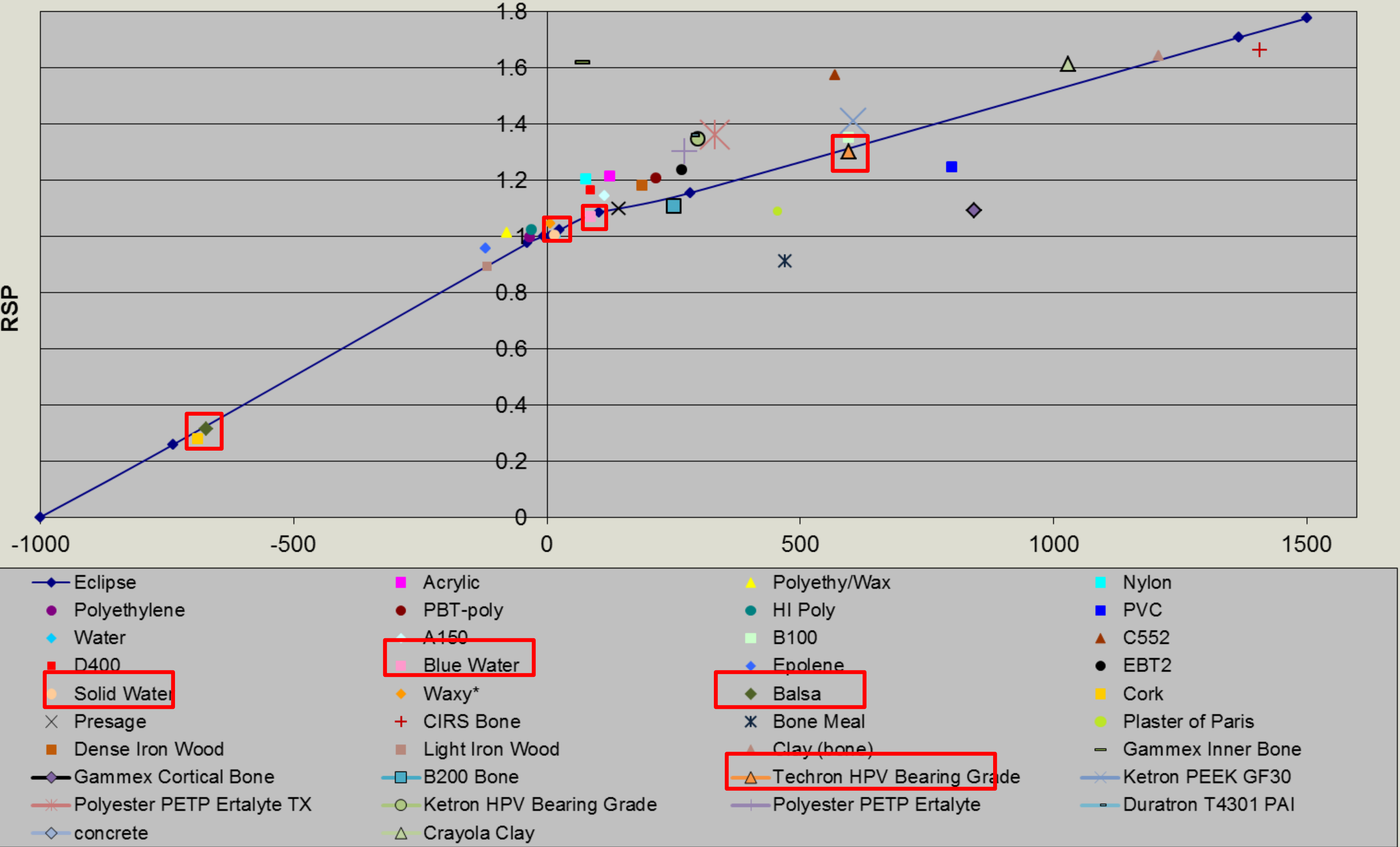


Not so good…..



RSP
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-1000
-500
0
500
1000
1500
Eclipse
Acrylic
Polyethy/Wax
Nylon
Polyethylene
PBT-poly
HI Poly
PVC
Water
A150
B100
C552
D400
Blue Water
Epolene
EBT2
Solid Water
Waxy*
Balsa
Cork
Presage
CIRS Bone
Bone Meal
Plaster of Paris
Dense Iron Wood
Light Iron Wood
Clay (bone)
Gammex Inner Bone
Gammex Cortical Bone
B200 Bone
Techron HPV Bearing Grade
Ketron PEEK GF30
Polyester PETP Ertalyte TX
Ketron HPV Bearing Grade
Polyester PETP Ertalyte
Duratron T4301 PAI
concrete
Crayola Clay

Stopping Power vs. HU Curve



Summary

- Be careful with proton beams!
- Good luck finding materials that behave like tissues!

Conclusions

- Phantoms are useful for credentialing
- Phantoms are also a unique tool to evaluate many different aspects of radiation therapy

Thank You!

- This work was supported by grants CA010953, CA081647