Neuro-Oncology Advances

4(1), 1–9, 2022 | https://doi.org/10.1093/noajnl/vdac058 | Advance Access date 20 April 2022

The current status and shortcomings of stereotactic radiosurgery

Hunter Mehrens, Trang Nguyen, Sharbacha Edward, Shannon Hartzell, Mallory Glenn, Daniela Branco, Nadia Hernandez, Paola Alvarez, Andrea Molineu, Paige Taylor, and Stephen Kry

Department of Outreach Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA (H.M., T.N., S.E., S.H., M.G., D.B., N.H., P.A., A.M., P.T., S.K.); Imaging and Radiation Oncology Core, Houston, TX, USA (H.M., T.N., S.E., S.H., M.G., D.B., N.H., P.A., A.M., P.T., S.K.)

Corresponding Author: Stephen Kry, PhD, Imaging and Radiation Oncology Core, MD Anderson Cancer Center, 8060 El Rio, Houston, TX 77054, USA (sfkry@mdanderson.org).

Abstract

Background. Stereotactic radiosurgery (SRS) is a common treatment for intracranial lesions. This work explores the state of SRS treatment delivery to characterize current treatment accuracy based on treatment parameters.

Methods. NCI clinical trials involving SRS rely on an end-to-end treatment delivery on a patient surrogate (credentialing phantom) from the Imaging and Radiation Oncology Core (IROC) to test their treatment accuracy. The results of 1072 SRS phantom irradiations between 2012 and 2020 were retrospectively analyzed. Univariate analysis and random forest models were used to associate irradiation conditions with phantom performance. The following categories were evaluated in terms of how they predicted outcomes: year of irradiation, TPS algorithm, machine model, energy, and delivered field size.

Results. Overall, only 84.6% of irradiations have met the IROC/NCI acceptability criteria. Pass rate has remained constant over time, while dose calculation accuracy has slightly improved. Dose calculation algorithm ($P < .001$), collimator ($P = .024$), and field size ($P < .001$) were statistically significant predictors of pass/fail. Specifically, pencil beam algorithms and cone collimators were more likely to be associated with failing phantom results. Random forest modeling identified the size of the field as the most important factor for passing or failing followed by algorithm.

Conclusion. Constant throughout this retrospective study, approximately 15% of institutions fail to meet IROC/NCI standards for SRS treatment. In current clinical practice, this is particularly associated with smaller fields that yielded less accurate results. There is ongoing need to improve small field dosimetry, beam modeling, and QA to ensure high treatment quality, patient safety, and optimal clinical trials.

Key Points

- A substantial number of institutions fail to meet IROC/NCI standards for SRS treatment delivery.
- Smaller fields in SRS treatment yield less accurate results in phantoms.
- Pencil beam algorithms were more associated with failing phantom results.

Stereotactic radiosurgery (SRS) is a standard-of-care treatment for indications such as metastatic brain lesions that is supported by extensive clinical trial data.[1–5] SRS treatments have many technical differences from traditional radiotherapy because of the small tumor volumes targeted and the need for precise target localization. These special requirements have resulted in the development of unique treatment systems for SRS and, as a result, it is delivered with a wide range of different treatment delivery platforms and collimators, and the treatment dose is calculated using a wide range of treatment planning systems and dose calculation algorithms.

Importance of the Study

With the increased prevalence of stereotactic radiosurgery (SRS) as standard of care for brain metastasis, understanding the state of this technique and particularly the impact of treatment options that influence delivery accuracy is vital. While challenges to SRS treatments are known, eg, small field dosimetry and patient setup, the impact of these challenges, and prevalence of associated treatment shortcomings are not well studied or understood. This study highlights the state of accuracy of SRS across the radiotherapy community and identifies important treatment parameters that predict treatment accuracy. A substantial portion (15%) of the radiotherapy community fails to meet treatment delivery standards; while the field has progressed in both general knowledge and technology, challenges continue to limit the improvement of this treatment modality. Several key features are highlighted in predicting inaccurate treatment deliveries, specifically pencil beam algorithms and small fields, which should be approached with care when used for SRS treatments.

Regardless of the delivery platform and dose calculation method, SRS involves many technical challenges, including difficulties with small field dosimetry, dose calculation accuracy, and precise patient setup. Ensuring that the correct dose is delivered to the right location in SRS is of particular importantance[6–9] (and many tools have been developed to aid in this[13–17,19]) not only because of the unique challenges of this treatment, but also because of the inability to correct any error later in treatment. Despite this, radiotherapy institutions have been found to be routinely lagging behind current quality assurance recommendations,[10] both in the criteria used for assessing SRS plans and in dealing with identified plan inaccuracies.[11,12]

Particularly robust verification of plan accuracy is achieved with an independent end-to-end evaluation.[18,20–22] The largest program that supports such a test is the Imaging and Radiation Oncology Core (IROC), which supports clinical trials involving radiotherapy from the National Cancer Institute (NCI). IROC provides a standardized end-to-end SRS patient surrogate (phantom; Figure 1) as a credentialing test that serves as a prerequisite for institutions to participate in NCI-sponsored clinical trials using SRS. A phantom is sent to an institution where it is treated like a patient, undergoing radiotherapy planning and delivery according to the clinic's technique. After irradiation, it is returned to IROC and evaluated to see how the delivered and planning dose compare.

The results of these credentialing tests can elucidate the pervasiveness and degree to which the challenges in SRS, particularly how the various delivery and dose calculation platforms, affect current clinical practice. In this study, we reviewed the most recent 1072 SRS phantom irradiations using univariate analysis and random forest models[23–26] to explore the state of SRS treatment delivery as well as determine underlying trends and features that lead to suboptimal delivery.

Materials and Methods

IROC Phantom

IROC's SRS head phantom contains a 1.9 cm spherical target made of solid water. The accuracy of the delivered dose in the target is measured with two single-loaded thermoluminescent dosimeters (TLDs) and two planes (coronal and sagittal) of GAFchromic film. The TLD provide a highly precise[27] assessment of dose accuracy at two points in the target; the film provides planar assessment of the dose distribution.

Institutions are instructed to deliver a treatment consistent with their clinical practice with a maximum target dose of ~30 Gy. The process includes immobilization, treatment planning, localization, and delivery. To pass, the measured point dose (via TLD) must agree with the treatment planning system (TPS) calculation within 5%, and ≥85% of pixels must pass a 5%/3 mm gamma analysis on each of the two film planes. Figure 1 shows the two types of SRS head phantoms available: the original SRS head phantom is water-filled with a dosimetric and imaging insert that must be interchanged between simulation and treatment, while the newer design (available post-2017) has a single insert for both imaging and dosimetry that is constructed primarily from high-impact poly. Both phantom types are used interchangeably and have identical targets and dosimetry.

Dataset

SRS phantom data were collected from 2012 to October 31, 2020. Nine values were extracted for each irradiation: overall phantom performance (pass/fail) status, average TLD ratio (measured:calculated), average percentage of pixels passing gamma for both planes (gamma analysis was implemented after 2013), as well as irradiation date, TPS algorithm,[28] treatment machine, beam energy, collimator, and the delivered field size (while the phantom has a single target size, different institutions treated larger or smaller volumes; this was acceptable as long as the delivered dose matched the intended dose).

Irradiation date was defined only based on the year and considered as a categorical variable. Treatment machine was defined based on machine class from the dosimetric characteristics in IROC's reference data.[29,30] Collimators were divided into three categories: cone of any size (including variable), high-definition MLC (leaf width <0.5 cm; HD120, micro MLC models, Incise2), and low-definition MLC (leaf width ≥0.5 cm; Millennium 120, Elekta Agility, Elekta 80, Tomotherapy binary MLC). The delivered field size was calculated from the average extent of the 50% isodose line based on the film measurements and averaged over all three-profile directions (superior-inferior, right-left, and anterior-posterior). Specifically, the 50% isodose line was calculated based on the average dose across 80% of the tumor extent. Demographic data are shown in Table 1.

Analysis

We first evaluated the relationship between treatment conditions and the performance (pass/fail rate) of the phantom using univariate analysis (Pearson chi-squared test with Tukey post hoc analysis; IBM SPSS 24). Follow-up analysis was done using random forest methods to predict pass/fail

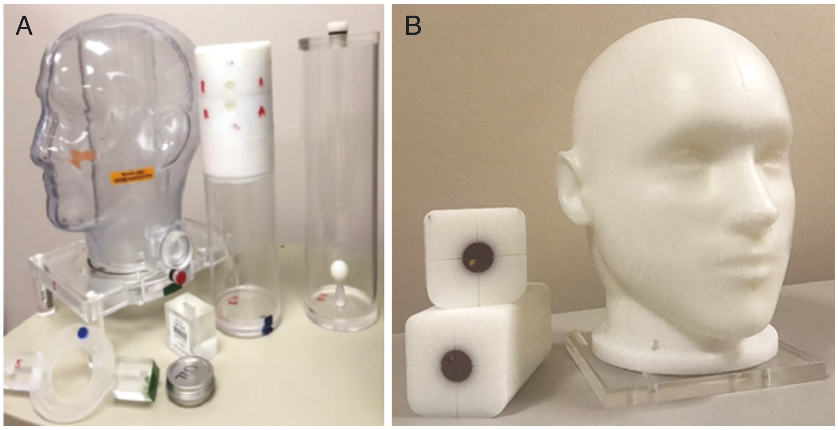

Figure 1. IROC Houston's two SRS head phantoms: (a) water-filled with a dosimetric and imaging insert and (b) single insert that is constructed with high-impact poly. The phantom contains a 1.9 cm spherical target made of solid water where dose is measured with two single-loaded TLDs and two planes of GAFchromic film. Abbreviations: IROC, Imaging and Radiation Oncology Core; SRS, stereotactic radiosurgery; TLD, thermoluminescent dosimeters.

Table 1. Pass Rate of Treatment Parameters

Category	Constituents	N (% of Total Samples)	Pass Rate (%)
Treatment machine class	Truebeam	419 (39.1)	84.5
	Trilogy	162 (15.1)	82.7
	Gammaknife	130 (12.1)	89.2
	Varian Base	121 (11.3)	81.0
	Cyberknife	98 (9.1)	87.8
	Elekta Agility	60 (5.6)	83.3
Beam energy	6	526 (49.1)	85.0
	6 SRS	60 (5.6)	81.7
	6 FFF	248 (23.1)	84.3
	Co-60	129 (12.0)	89.2
	10 FFF	61 (5.7)	78.7
TPS algorithm	Pencil Beam	231 (21.6)	75.8
	Monte Carlo	76 (7.1)	90.8
	AAA	285 (26.6)	86.0
	S/C	112 (10.5)	85.7
	GBBS	66 (6.2)	87.9
	Measured	172 (16.0)	90.7
Collimator group	Cone	231 (21.6)	78.8
	High definition	205 (19.1)	87.3
	Low definition	374 (34.9)	85.8

Abbreviations: SRS, stereotactic radiosurgery; TPS, treatment planning system; FFF, flattening filter-free; GBBS, grid-based Boltzmann solver.

(classification), as well as average TLD ratio and average percentage of pixels passing gamma (regression), through the above categories. For all random forest analyses, missing data were imputed using 100 trees for all random forest models and then upsampled to ensure a balanced dataset between pass and fail. In addition, four hyperparameters (number of trees, number of variables split to randomly sample for each descending node, percentage of samples to train on, and minimum number of samples within the terminal nodes) were tuned to minimize out-of-bag (OOB) error. These hyperparameters were tuned over the following hyperspace: number of trees (200-1000), number of variables split to randomly sample for each descending node (2-6), percentage of samples to train on (55%, 63.2%, 70%, 80%), and minimum number of samples within the terminal nodes (1-20). Since the random forest algorithm is self-validating through bootstrap aggregating, we split our dataset into a training (70%) and a testing set (30%). The implementation of this random forest classification was run three times to assess fluctuations within the data and results. Random forest was performed using R (4.0.2) via the ranger and missRanger packages.[31–33]

Results

For the 1072 irradiations of the IROC SRS phantom between 2012 and 2020, the average pass rate was 84.6%, indicating that more than 15% of institutions failed to deliver their intended dose accurately. Of the 164 failures, 56.4% failed due to point dose assessment only (TLD; typically indicating the wrong dose was delivered to the right location), 30.3% failed due to planar assessment only (typically indicating that the right dose was delivered to the wrong location), and 13.1% failed due to both point and planar assessment (indicating multiple or substantial issues). The average dose ratio in the target (measured/predicted) and planar agreement (percentage of pixels passing gamma) were 0.982 ± 0.035 and 95.6 ± 6.8%, respectively, indicating that the dose was, on average, underestimated by 1.8%. Figure 2 shows the evolution of the pass rate, average TLD ratio, and number of irradiations per year. The passing rate and percentage of pixels passing gamma have not significantly changed (regression analysis, $P = .975$ and $P = .415$, respectively), while the average TLD ratio (point dose) has steadily improved, moving closer to 1.000, at a rate of 0.002 per year (significantly positive sloping trend line, $P = .005$). The relationship between each of the different treatment parameters and TLD ratio and gamma results is shown in Figure 3.

Univariate analysis showed that the institutions' TPS algorithm ($P < .001$), collimator ($P = .024$), and field size ($P < .001$) were significantly associated with phantom performance, while other variables, including energy ($P = .531$), machine ($P = .832$), and year of irradiation ($P = .118$) were not statistically significant (Table 2). In particular, calculations based on a pencil beam algorithm had a 10%-15% lower pass rates than other algorithms (AAA: $P = .01$, GBBS: $P = .015$, Monte Carlo: $P = .001$, Measured: $P < .001$, and Superposition/Convolution: $P = .015$); no other differences existed between TPS algorithms. Irradiations delivered with cone collimators had an 8.5% lower pass rates than high definition MLC collimators ($p = 0.035$), and 7% lower pass rates than low-definition collimators (although not significant; $p = 0.055$); low definition and high definition MLCs showed no difference. Finally, treatments that used field sizes <3 cm were significantly less likely to pass (15.3% reduction in pass rate) than irradiations done with larger fields.

Our random forest results reinforced the findings of the univariate analysis and provided the additional depth of information. The variables evaluated were found to account for the vast majority of the variations seen by different institutions: particularly for classification, which had an accuracy of 90.9 ± 0.3% (regression accuracy: 83.1 ± 2.5%). For overall pass/fail, average TLD ratio, and average percentage of pixels passing, the field size was the most important predictor. The next three most important variables were year of irradiation, machine, and algorithm; however, the specific order of importance varied depending on which outcome was of interest. For example, for average TLD ratio regression, the TPS algorithm was the second most important variable, while for percentage of pixels passing gamma, TPS algorithm was the fourth most important. Interestingly, and different from the univariate analysis, collimator was found to be among the least important parameters in predicting phantom performance. Figure 4 shows the ranking of the variables of importance for classification and regression random forest modeling.

Discussion

The state of clinical SRS practice was elucidated in this study. The pass rate of the SRS phantom has remained steady over the past 8 years at approximately 85%, despite an increase in the average TLD ratio. This implies that clinical practice still struggles to accurately align and deliver dose to targets in SRS treatments, despite increases in utilization and advancements in this treatment modality. Specifically, our data show that the interplay of three categories contributed to the majority of failing phantom results: TPS algorithm (pencil beam), collimator (cone), and small treatment fields.

Pencil beam algorithms have been well known to poorly model the true dose distribution, particularly in heterogeneous anatomy, and including in other IROC phantoms (non-SRS).[34,35] However, the SRS head phantom, which has homogenous anatomy, indicates that the failure of the pencil beam algorithm may extend beyond poor modeling in heterogeneous media. Interestingly, the pencil beam algorithm not only underperformed compared to more advanced algorithms but also simpler measurement-based algorithms, highlighting a pronounced deficiency.

Collimator and field size need to be examined together to understand their effect on SRS delivery performance. First, we found that cone collimators were used for the majority of small treatment fields (field sizes ≤3 cm). Second, all random forest models indicate that field size was the most important variable while collimator was the least important variable. Combining these two analyses indicates that field size is the driving force behind failures, and cone

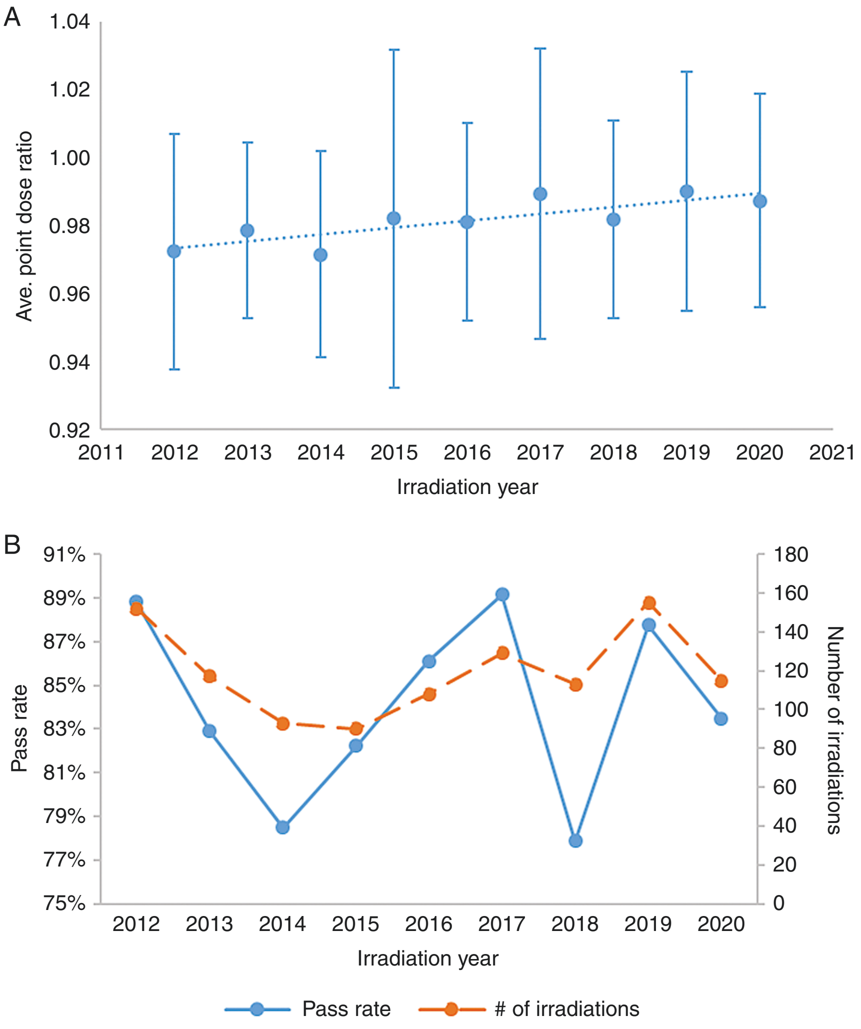

Figure 2. a) The average measured to delivered point dose (average TLD ratio) value for the SRS phantom through the years b) Pass rates for the SRS head phantom through the years with an overall average of 84.6% (solid line) combined with the number of irradiations performed per year (dashed line). The error bars presented in graph a represent ± one standard deviation. There has been an overall increase in point dose performance relative to a site's dose calculation but passing rates have remained constant over the span of our data set.

collimators are simply correlated with field size. There are two reasons why small fields are likely to predict phantom failure. From an institution's perspective, smaller fields are harder to characterize dosimetrically and contain larger dose calculation uncertainties compared to conventional radiotherapy.[18] On IROC's side, the dose measurement is also more challenging: in particular, volume averaging effects over the TLD are more pronounced for smaller fields, especially as dose uniformity across the target in SRS treatments is often not a priority (while the measured TLD dose is compared to a corresponding volume in the treatment planning system, this is nevertheless more sensitive to any positioning uncertainty).

The suitability of the criteria set forth by IROC in conjunction with the NCI can be evaluated, and separated into two categories: (1) 5% dose difference for TLD and gamma analysis and (2) 3 mm distance to agreement for gamma analysis. Through IROC's head and neck phantom, Carson et al justified the 5% dose difference as a reasonable threshold to evaluate phantom performance with a probability of measurement-noise induced failure being less than 0.5%.[21] However, due to the ablative nature of SRS therapy, such dose errors may be less clinically important. Nevertheless, medical physics practice standards are clear that 5% should be obtainable in terms of dosimetric accuracy, despite the additional challenges associated with small field dosimetry.[18,36] The registration error associated with IROC's distance to agreement is approximately 1 mm, which is consistent with medical physics tolerances standards for SRS.[9,37] Given the nature and goal of SRS treatment, 3 mm is a generous spatial margin to assess performance within the phantom.

It is of note that the two analysis techniques utilized in this paper differ in their goals and interpretation. Univariate analysis directly compared two variables, while the random forest models analyzed the larger

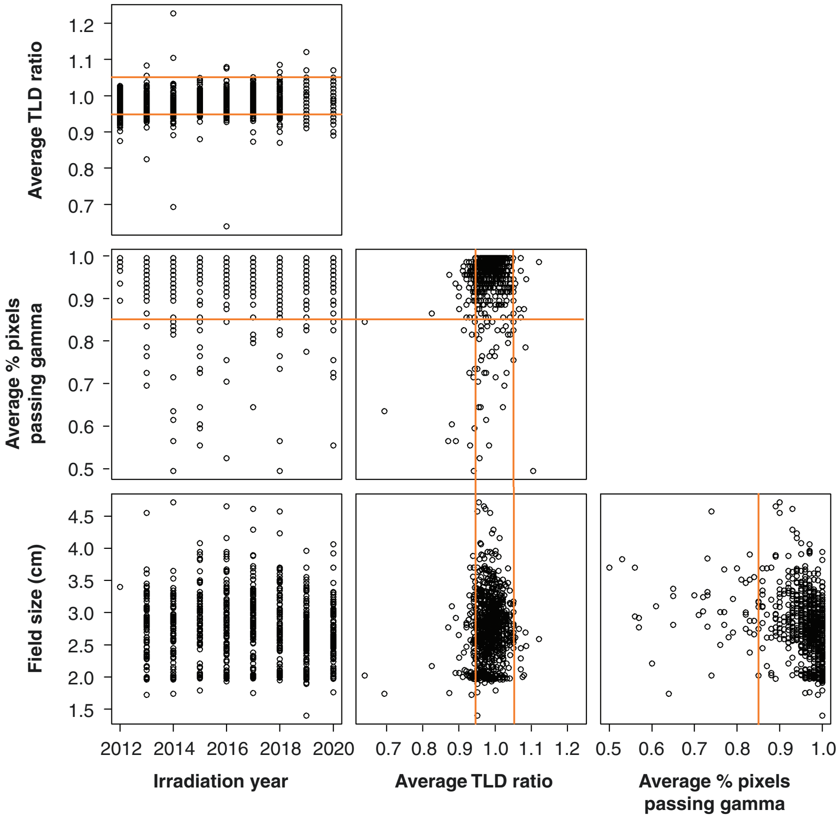

Figure 3. Inter-comparison of irradiation year, average TLD ratio, average percentage of pixels passing gamma, and field size. The lines represent the passing criteria for the SRS head phantom. Noteworthy are that as the irradiation year progresses, average TLD ratio begins to tighten around 1.0 (panel a). In addition, failures in the average TLD ratio tend to be more prevalent for smaller field sizes (≤ 3.0 cm; mid-bottom panel).

Table 2. Detailed Output Parameters of Algorithm and Collimator

TPS Algorithm	Pencil Beam[a]	Monte Carlo	AAA	S/C	GBBS	Measurement-based
N	231	76	285	112	66	172
% (N)	21.6%	7.1%	26.6%	10.5%	6.16%	16.0%
Average TLD ratio	0.967	0.974	0.992	0.981	0.999	0.987
Average % pixels passing	94.8%	96.6%	95.1%	94.0%	95.8%	97.8%
Pass rate	75.8%	90.8%	86.0%	85.7%	87.9%	90.7%
Collimator	**Cone[a]**		**HD**		**LD**	
N	231		205		374	
% (N)	21.6%		19.1%		34.9%	
Average TLD ratio	0.973		0.985		0.987	
Average % pixels passing	96.8%		95.2%		94.5%	
Pass rate	78.8%		87.3%		85.8%	

Abbreviations: AAA, Anisotropic Analytical Algorithm; HD, high definition; LD, GBBS, grid-based Boltzmann solver; low definition; TLD, thermoluminescent dosimeters; TPS, treatment planning system.

[a] The categories that showed further statistical significance utilizing the Tukey post hoc test. Specifically, pencil beam compared to all other algorithms and cone collimators compared to HD were statistically significant.

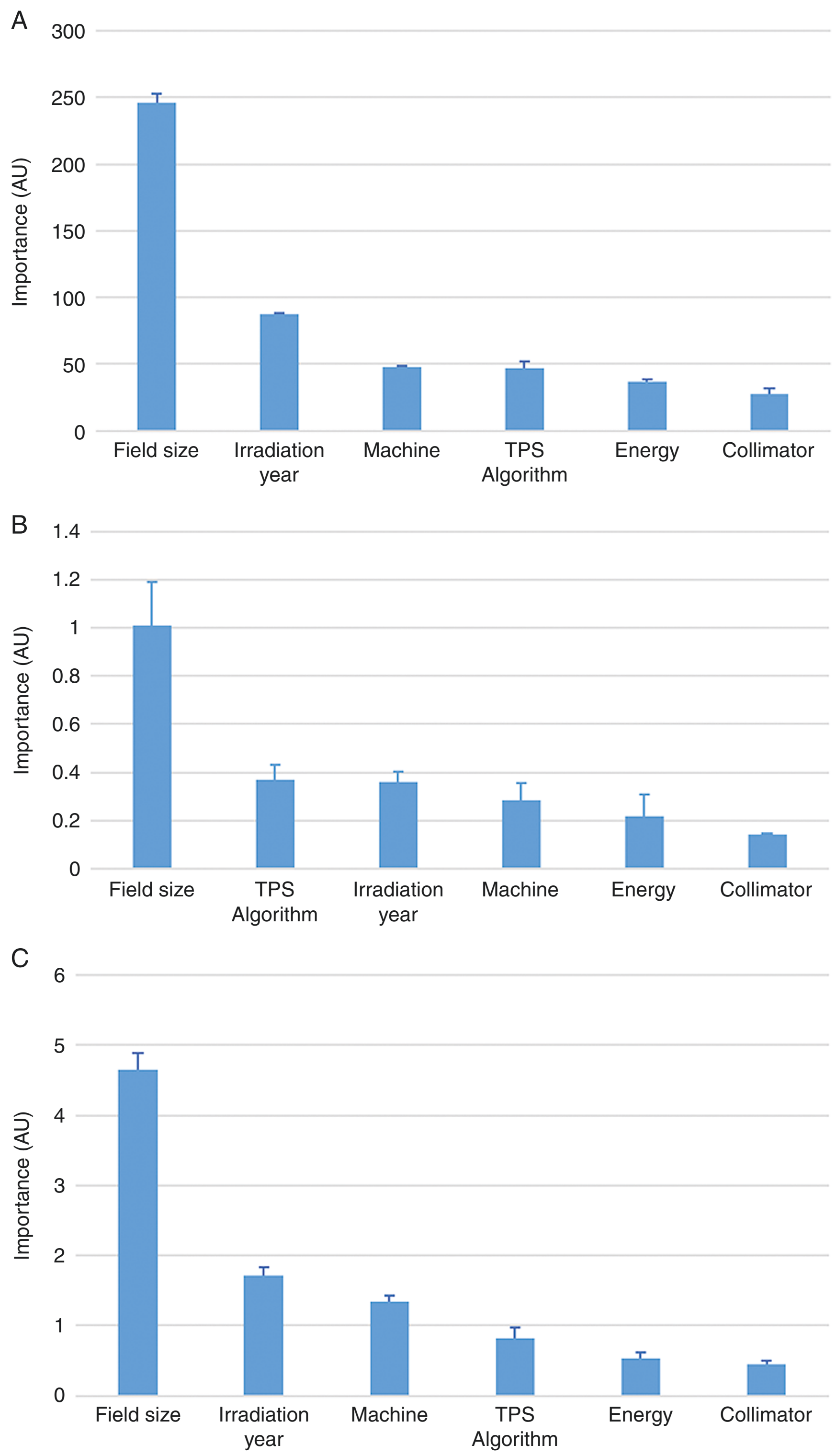

Figure 4. The variables of importance from our (a) performance classification, (b) regression average TLD ratio, and (c) regression percentage of pixels passing gamma random forest models. Error bars represent 1 SD. The level of importance is in arbitrary units (AU) and can only be compared to other variables on a single graph. Abbreviation: TLD, thermoluminescent dosimeters.

picture, modeling the interplay of all the variables. Both techniques allow for different viewpoints of the data, and through their combination, we can understand the complex nature of our data and the true interplay of treatment parameters on performance. In addition to elucidating the interplay between collimator and field size as described above, a deeper understanding is also available in terms of the TPS algorithm's impact. Figure 4 shows TPS algorithm's ranking in importance at predicting outcome. While it is the second most important factor for the TLD dose accuracy, it is only the fourth most important variable for gamma agreement or overall pass/fail. This information, combined with our univariate analysis, allows for the reasonable conclusion that pencil beam algorithms struggle to accurately predict dose compared to other algorithms (affecting the TLD ratio) but does not affect alignment (ie, the gamma pass rate).

Through the use of random forest and univariate analysis, three main categories of failure were found that continued to suppress the pass rate for SRS treatment. To avoid continuing this trend into the future, additional steps can be taken by institutions. First, a careful understanding of the difficulty and methods for accurate characterization of small fields is essential. The opportunity for improvement has been aided greatly by recent guidance from the IAEA,[36] and it is essential that such guidance be followed by any SRS program. Second, independent evaluations of an SRS program (such as using IROC's end-to-end QA phantom) should be used. Such evaluations provide an opportunity for review and testing of the program and are strongest when based on an independent test using independent dosimeters and analysis methods. Any unexpected or undesirable results should be thoroughly reviewed, for example with help from IROC's team, to determine likely reasons for the problems. Once properly investigated, diagnosed, and rectified, the independent testing should be repeated.

Conclusions

A collection of 1072 SRS head phantoms irradiated from 2012 to 2020 were analyzed to understand the performance and state of the radiotherapy community and shortcomings in different treatment techniques. Despite an increase in the dosimetric accuracy, the pass rate has remained steady at approximately 85%. Field size had the largest impact in determining performance (pass/fail) and, specifically, small fields (≤3 cm) were associated with more failures across the radiation oncology community. In addition, pencil beam algorithms lagged in performance compared to other dose calculation algorithms in terms of accuracy.

These results broadly indicate a need for improvement in SRS treatments across the community. Treatment parameters that need particular attention to ensure accurate treatment delivery include, smaller field sizes and pencil beam algorithms (which should be used with caution).

Keywords

IROC | phantoms | QA | random forest | SRS

Funding

This work was supported by the National Institute of Health (NIH)/National Cancer Institute (NCI) [grant #CA180803, #CA214526 to S.K.].

Conflict of interest statement. None of the above authors has any conflicts of interest to declare.

Authorship statement. Random forest and statistical analysis: H.M., S.K. Phantom analysis and data collection: H.M., T.N., S.E., S.H., M.G., D.B., N.H., P.A., A.M., P.T., and S.K. Writing and revision of the manuscript: H.M., T.N., S.E., S.H., M.G., D.B., N.H., P.A., A.M., P.T., and SK.

References

1. Brown PD, Ballman KV, Cerhan JH, et al. Postoperative stereotactic radiosurgery compared with whole brain radiotherapy for resected metastatic brain disease (NCCTG N107C/CEC·3): a multicentre, randomised, controlled, phase 3 trial. *Lancet Oncol.* 2017;18(8):1049–1060.
2. Churilla TM, Chowdhury IH, Handorf E, et al. Comparison of local control of brain metastases with stereotactic radiosurgery vs surgical resection: a secondary analysis of a randomized clinical trial. *JAMA Oncol.* 2019;5(2):243–247.
3. Mitrasinovic S, Zhang M, Appelboom G, et al. Milestones in stereotactic radiosurgery for the central nervous system. *J Clin Neurosci.* 2019;59(1):12–19.
4. Sahgal A, Ruschin M, Ma L, et al. Stereotactic radiosurgery alone for multiple brain metastases? A review of clinical and technical issues. *Neuro Oncol.* 2017;19(suppl_2):ii2–ii15.
5. Hartgerink D, Swinnen A, Roberge D, et al. LINAC based stereotactic radiosurgery for multiple brain metastases: guidance for clinical implementation. *Acta Oncol.* 2019;58(9):1275–1282.
6. Potters L, Kavanagh B, Galvin JM. American Society for Therapeutic Radiology and Oncology (ASTRO) and American College of Radiology (ACR) practice guideline for the performance of stereotactic body radiation therapy. *Int J Radiat Oncol Biol Phys.* 2010;76(2):326–332.
7. Benedict SH, Yenice KM, Followill D, et al. Stereotactic body radiation therapy: the report of AAPM Task Group 101. *Med Phys.* 2010;37(8):4078–4101. Erratum in: Med Phys. 2012;39(1):563. Dosage error in article text. PMID: 20879569.
8. Schell M, Bova F, Larson D, et al. *AAPM Report No. 54: Stereotactic Radiosurgery.* Boston, MA: American Association of Physicists in Medicine; 1995.
9. Klein EE, Hanley J, Bayouth J, et al. Task Group 142 report: quality assurance of medical accelerators. *Med Phys.* 2009;36(9):4197–4212.
10. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM Task Group No. 218. *Med Phys.* 2018;45(4):e53–e83.
11. Mehrens H, Taylor P, Followill D, Kry S. Survey results of 3D-CRT and IMRT quality assurance practice. *J Appl Clin Med Phys.* 2020;21(7):70–76.

12. Xia Y, Adamson J, Zlateva Y, Giles W. Application of TG-218 action limits to SRS and SBRT pre-treatment patient specific QA. *J Radiosurg SBRT.* 2020;7(2):135–147.
13. Sarkar V, Huang L, Huang YJ, et al. Head to head comparison of two commercial phantoms used for SRS QA. *J Radiosurg SBRT.* 2016;4(3):213–223.
14. Capaldi DPI, Skinner LB, Dubrowski P, et al. An integrated quality assurance phantom for frameless single-isocenter multitarget stereotactic radiosurgery. *Phys Med Biol.* 2020;65(11):115006.
15. Brezovich IA, Popple RA, Duan J, et al. Phantom and procedure providing submillimeter accuracy in daily QA tests of accelerators used for stereotactic radiosurgery. *J Appl Clin Med Phys.* 2016;17(4):246–253.
16. Rose MS, Tirpak L, Van Casteren K, et al. Multi-institution validation of a new high spatial resolution diode array for SRS and SBRT plan pretreatment quality assurance. *Med Phys.* 2020;47(7):3153–3164.
17. Brezovich IA, Wu X, Popple RA, et al. Stereotactic radiosurgery with MLC-defined arcs: verification of dosimetry, spatial accuracy, and end-to-end tests. *J Appl Clin Med Phys.* 2019;20(5):84–98.
18. Smilowitz JB, Das IJ, Feygelman V, AAPM Medical Physics Practice Guideline Task Group, et al. 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):14–34.
19. Dimitriadis A, Tsang Y, Thomas R, et al. Multi-institutional dosimetric delivery assessment of intracranial stereotactic radiosurgery on different treatment platforms. *Radiother Oncol.* 2020;147(1):153–161.
20. Molineu A, Hernandez N, Nguyen T, Ibbott G, Followill D. Credentialing results from IMRT irradiations of an anthropomorphic head and neck phantom. *Med Phys.* 2013;40(2):022101.
21. Carson ME, Molineu A, Taylor, PA, et al. Examining credentialing criteria and poor performance indicators for IROC Houston's anthropomorphic head and neck phantom. *Med Phys.* 2016;43(12):6491–6496.
22. Kerns JR, Stingo F, Followill DS, et al. Treatment planning system calculation errors are present in most Imaging and Radiation Oncology Core-Houston phantom failures. *Int J Radiat Oncol Biol Phys.* 2017;98(5):1197–1203.
23. Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics.* 2018;19(1):270.
24. Ho TK. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Quebec, Canada, August 14–15, 1995, Vol. 1, pp. 278–282.
25. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach Learn.* 2000;40(1):139–157.
26. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32.
27. Kirby TH, Hanson WF, Johnston DA. Uncertainty analysis of absorbed dose calculations from thermoluminescence dosimeters. *Med Phys.* 1992;19(6):1427–1433.
28. Kry SF, Feygelman V, Balter P, et al. AAPM Task Group 329: reference dose specification for dose calculations: dose-to-water or dose-to-muscle? *Med Phys.* 2020;47(3):e52–e64.
29. Kerns JR, Followill DS, Lowenstein J, et al. Technical Report: Reference photon dosimetry data for Varian accelerators based on IROC-Houston site visit data. *Med Phys.* 2016;43(5):2374.
30. Kerns JR, Followill DS, Lowenstein J, et al. Reference dosimetry data and modeling challenges for Elekta accelerators based on IROC-Houston site visit data. *Med Phys.* 2018;45(5):2337–2344.
31. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.R-project.org/.
32. Marvin NW, Andreas Z. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw.* 2017;77(1):1–17.
33. Mayer M. missRanger: Fast Imputation of Missing Values R package version 2.1.0. 2019. https://CRAN.R-project.org/package=missRanger
34. Kry SF, Alvarez P, Molineu A, et al. Algorithms used in heterogeneous dose calculations show systematic differences as measured with the Radiological Physics Center's anthropomorphic thorax phantom used for RTOG credentialing. *Int J Radiat Oncol Biol Phys.* 2013;85(1):e95–e100.
35. Taylor PA, Kry SF, Followill DS. Pencil beam algorithms are unsuitable for proton dose calculations in lung. *Int J Radiat Oncol Biol Phys.* 2017;99(3):750–756.
36. Palmans H, Andreo P, Huq MS, et al. Dosimetry of small static fields used in external photon beam radiotherapy: summary of TRS-483, the IAEA-AAPM International Code of Practice for reference and relative dose determination. *Med Phys.* 2018;45(11):e1123–e1145.
37. Halvorsen PH, Cirino E, Das IJ, et al. AAPM-RSS Medical Physics Practice Guideline 9.a. for SRS-SBRT. *J Appl Clin Med Phys.* 2017;18(5):10–21.