Advances in Radiation Oncology (2020) 5, 1296-1304

advances
in radiation oncology

www.advancesradonc.org

Scientific Article

Combination of a Big Data Analytics Resource System With an Artificial Intelligence Algorithm to Identify Clinically Actionable Radiation Dose Thresholds for Dysphagia in Head and Neck Patients

Charles S. Mayo, PhD,[a,*] Michelle Mierzwa, MD,[a] Jean M. Moran, PhD,[a] Martha M. Matuszak, PhD,[a] Joel Wilkie, MD,[a] Grace Sun, BS,[a] John Yao, PhD,[a] Grant Weyburn, BS,[a] Carlos J. Anderson, PhD,[a] Dawn Owen, MD,[a] and Arvind Rao, PhD[b]

Departments of [a]Radiation Oncology and [b]Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan

Received 3 April 2019; revised 23 December 2019; accepted 30 December 2019

Abstract

Purpose: We combined clinical practice changes, standardizations, and technology to automate aggregation, integration, and harmonization of comprehensive patient data from the multiple source systems used in clinical practice into a big data analytics resource system (BDARS). We then developed novel artificial intelligence algorithms, coupled with the BDARS, to identify structure dose volume histograms (DVH) metrics associated with dysphagia.

Methods and Materials: From the BDARS harmonized data of ≥22,000 patients, we identified 132 patients recently treated for head and neck cancer who also demonstrated dysphagia scores that worsened from base line to a maximum grade ≥2. We developed a method that used both physical and biologically corrected ($\alpha/\beta = 2.5$) DVH curves to test both absolute and percentage volume based DVH metrics. Combining a statistical categorization algorithm with machine learning (SCA-ML) provided more extensive detailing of response threshold evidence than either approach alone. A sensitivity guided, minimum input, machine learning (ML) model was iteratively constructed to identify the key structure DVH metric thresholds.

Results: Seven swallowing structures producing 738 candidate DVH metrics were ranked for association with dysphagia using SCA-ML scoring. Structures included superior pharyngeal constrictor (SPC), inferior pharyngeal constrictor (IPC), larynx, and esophagus. Bilateral parotid and submandibular gland (SG) structures were categorized by relative mean dose (eg, SG_high, SG_low) as a dose versus tumor centric analog to contra and ipsilateral designations. Structure DVH metrics with high SCA-ML scores included the following: SPC: D20% (equivalent dose [EQD2] Gy) ≥47.7; SPC: D25% (Gy) ≥50.4; IPC: D35% (Gy) ≥61.7; parotid_low: D60% (Gy) ≥13.2; and SG_high: D35% (Gy) ≥61.7. Larynx: D25% (Gy) ≥21.2 and SG_low: D45% ≥28.2 had high SCA-ML scores but

Sources of support: This work was supported in part by a grant from Varian Medical Systems to Dr. Mayo. Dr. Moran has grants from Varian Medical Systems and from Blue Cross Blue Shield of Michigan.

Disclosures: Grant support from Varian Medical Systems.

Data sharing: Research data are not available at this time.

* Corresponding author: Charles S. Mayo, PhD; E-mail: cmayo@med.umich.edu

https://doi.org/10.1016/j.adro.2019.12.007

were segmented on less than 90% of plans. A model based on SPC: D20% (EQD2 Gy) alone had sensitivity and area under the curve of 0.88 ± 0.13 and 0.74 ± 0.17, respectively.

Conclusions: This study provides practical demonstration of combining big data with artificial intelligence to increase volume of evidence in clinical learning paradigms.

Introduction

Dysphagia is a significant acute and late toxicity for patients undergoing radiation therapy for head and neck cancers, increasing the probability of an aspiration pneumonia posttreatment, with modern multi-institutional trials demonstrating 10% to 20% long-term dysphagia.[1] Organ sparing of the superior constrictor muscles has been demonstrated as an advantageous use of intensity modulated radiation therapy early in application of that technology.[2-6] Owing to the extensive manual effort required, most single institution studies tend to be modest in size, examining a limited set of manually selected dose volume histograms (DVHs) metrics.

Reliance on manual aggregation methods decreases the likelihood of follow-up studies as findings are implemented, and treatment planning approaches are subsequently modified. In addition, the manual effort required to collect DVH metrics constrains the range of metrics examined, introducing potential biases in selection of metrics for testing.

Recently, we have constructed a big data analytics resource system (BDARS) that automates aggregation, integration, and harmonization of key data elements and relationships for all treated patients in a standardized framework.[7,8] Aggregated elements include dose volume histograms (DVHs) for all treated plans and the course cumulative as treated plan sum in both physical (Gy) and bio-corrected (equivalent dose [EQD2] Gy with α/β = 2.5, 5, 10) doses.[8,9] Common Terminology Criteria for Adverse Events toxicity grades were entered in our electronic health record (Epic, Verona, WI) using standardized smart list objects we developed to enable accurate, automated extraction from encounter notes with aggregation into our BDARS.[10]

Our objective in this study was to develop an automatable, systematic approach that enabled consideration of both physical and biologically corrected doses to both percentage and absolute volumes of organs at risk, detailing levels of evidence for each candidate metric. We developed a novel algorithmic approach that combined a statistical categorization algorithm (SCA) with a machine learning (ML) algorithm to identify the DVH metrics with the strongest associations for each structure. From these, a multistructure predictive ML model, extending the SCA, then was iteratively constructed to identify a minimal set of predictive cofactors. In this approach the end product is not the model. Instead, the end product is a minimal set of clinically actionable DVH metric inputs and thresholds, identified through use of the model, with the strongest levels of evidence for association with worsening dysphagia.

Methods and Materials

Patients

Records were examined for 439 patients treated for head and neck cancer from January 2014 to September 2018 using either intensity modulated radiation therapy or volumetric arc therapy treatment plans designed on a commercial system (Varian Medical System Eclipse, Palo Alto, CA). Toxicity and DVH curves for patients whose Common Terminology Criteria for Adverse Events dysphagia toxicity scores increased from baseline recorded during the first week of radiation therapy was used in the analysis. Patients were stratified for toxicity by maximum grade ≥2. Table 1 summarizes characteristics of 132 patients identified in this cohort. Three percent of patients were enrolled on clinical trials. Overall rates of toxicity that worsened from baseline were 17.8% ≥grade 2 and 5.5% ≥grade 3.

Contouring

Structures were contoured in a consistent fashion by a small number of physicians using agreed upon guidelines that have been in place for several years at our institution. The cervical esophagus was contoured as a tubular structure beginning at the bottom of inferior constrictor and extending to the thoracic inlet. The larynx was contoured extending from inferior border of hyoid to the inferior border of cricoid, and inferior constrictors were contoured from bottom of the hyoid to esophageal inlet, including anterior commissure and arytenoids. Superior constrictors were contoured from

Table 1 Characteristics of patients demonstrating worsening dysphagia

Characteristics of 132 out of 439 demonstrating worsening dysphagia	
Sex	
Male	35
Female	97
Age (median [25% quantile, 75% quantile])	62 [53, 67]
Count of patients by diagnosis site	
Pharynx	63
Oral cavity	22
Larynx	22
Nasopharynx	8
Other	17
Follow-up days (median [25% quantile, 75% quantile])	152 [52, 270]
Count of patients with dysphagia details	
Max dysphagia = 1	54
Max dysphagia = 2	54
Max dysphagia = 3	24
Max-Min dysphagia = 1	63
Max-Min dysphagia = 2	50
Max-Min dysphagia = 3	19

pterygoid plates to the inferior border of the hyoid. Inferior constrictors were contoured from inferior hyoid to cervical esophagus.

Statistical categorization algorithm and machine learning for algorithmic evidence-based identification of DVH metric predictors

We applied an approach combining a statistical categorization algorithm and machine learning (SCA-ML) to rank combined levels of evidence DVH metrics for ability to predict among patients demonstrating dysphagia scores that increased from start of treatment, which reached a maximum grade $\geq$2. Nine swallowing structures were examined (Table 2). DVH metrics were written using standardized TG-263 nomenclature.[11] Four as treated plan sum DVH curves were used for each structure to select from among physical and bio-corrected dose with respect to absolute and percent volume for each structure. Curves were rendered as sets of DVH metrics: Dx% (Gy), Dxcc (Gy), Dx% (α/β = 2.5) (EQD2 Gy), Dxcc (α/β = 2.5) (EQD2 Gy). Percentage volumes examined were.

$$x \in [100, 99.5, 99-96\, by\, 1, 95-5\, by\, 5, 4-1\, by\, 1, 0.5, 0.0].$$

For absolute volume $x \in [vq1 - 0.5\, by\, 0.5]$, where vq1 is the lower 1% quantile of volumes for structure in the sample.

For each DVH metric we calculated a statistical screening metrics set (SSMS) to identify an optimal threshold and detail statistical evidence for its predictive value. All calculations were carried out using R (Vienna,

Table 2 Summary statistics from statistical screening metrics set and combined statistical categorization algorithm and machine learning (SCA-ML) for the top physical and bio-corrected dose metrics for each swallowing structure examined

Structure	DVH metric	TV	N	AUC	PPV	NPV	SN	SP	OR	PETR	SCA-ML
SPC	D25% (Gy)	50.4	129	0.68	0.69	0.76	0.92	0.37	2.9	0.55	4.1
SPC	D20% (EQ2D Gy) (✓)	47.7	129	0.68	0.70	0.90	0.97	0.35	7.0	0.57	4.1
Parotid_low	D60% (Gy)	13.2	123	0.66	0.72	0.55	0.69	0.58	1.6	0.47	2.4
Parotid_low	D80% (EQD2 Gy) (✓)	6.0	123	0.65	0.75	0.52	0.6	0.69	1.6	0.44	2.9
SG_high	D35% (Gy) (✓)	61.7	124	0.68	0.74	0.58	0.66	0.67	1.7	0.47	2.60
SG_high	D30% (EQD2 Gy)	57.8	124	0.68	0.73	0.58	0.69	0.63	1.7	0.48	1.80
Oral_cavity	D95% (Gy) (✓)	15.3	129	0.68	0.78	0.53	0.55	0.77	1.7	0.45	2.5
Oral_cavity	D96% (EQD2 Gy)	9.8	129	0.67	0.78	0.53	0.55	0.77	1.7	0.45	2.1
Parotid_high	D28.5cc (Gy) (✓)	13.9	129	0.66	0.80	0.68	0.78	0.70	2.5	0.52	2.4
Parotid_high	D28.5cc (EQD2 Gy)	8.9	129	0.66	0.8	0.68	0.78	0.70	2.5	0.52	2.4
Esophagus	D2cc (Gy) (✓)	22.6	124	0.61	0.69	0.59	0.82	0.42	1.7	0.45	1.5
Esophagus	D3cc (EQD2 Gy)	24.3	121	0.58	0.79	0.45	0.36	0.85	1.4	0.25	1.5
IPC	D90% (Gy) (✓)	12.8	124	0.66	0.73	0.59	0.73	0.59	1.8	0.48	1.4
IPC	D95% (EQD2 Gy)	7.5	124	0.66	0.72	0.63	0.80	0.53	2.0	0.50	1.2
Larynx	D25% (Gy) (☒)	21.2	110	0.60	0.67	0.88	0.97	0.31	5.4	0.49	4.5
Larynx	D25% (EQD2 Gy)	15	110	0.59	0.66	0.81	0.95	0.29	3.5	0.46	3.7
SG_low	D45% (Gy) (☒)	28.2	95	0.71	0.73	0.85	0.95	0.46	4.9	0.60	5.4
SG_low	D35% (EQD2 Gy)	23.5	95	0.69	0.70	0.93	0.98	0.35	9.9	0.58	4.2

Columns correspond to the threshold value (TV), number of plans with the structure drawn (N), area under the curve (AUC) from the receiver operator characteristic analysis, positive predictive value (PPV), negative predictive value (NPV), sensitivity (SN), specificity (SP), and risk ratio determined using TV to construct a 2 × 2 contingency table. Structures not contoured on at least 90% of treatment plans (☒) are noted. For each structure, dose volume histograms (DVH) metric with the higher statistical categorization algorithm with machine learning (SCA-M) score is checked (✓).

Abbreviations: IPC = inferior pharyngeal constrictor; PETR = positive evidence of a threshold response; SG = submandibular gland; SPC = superior pharyngeal constrictor.

Austria, version 4.3.3).[2-15] For each SSMS, a receiver operator characteristic curve was constructed, and the area under the curve (AUC) was calculated for each set of toxicity and DVH metric dose records. A DVH metric value threshold was determined with the Youden index and used to construct a 2 × 2 contingency table. Values for the 95% confidence interval for the AUC, sensitivity (SN), specificity (SP), positive predictive value (PPV), and negative predicted value were calculated. The Fisher exact test was used to calculate the P value of the 2 × 2 contingency table. Relative risk and odds ratio were calculated. Standard and scaled values for the number of true positive, false positive, true negative, and false negative values were calculated with the square root of the number of samples ($\sqrt{N}$) as the scaling factor. A single-tailed Kolmogorov-Smirnov (ks) test was used to determine the P value that the distribution of doses for those without toxicities was stochastically less than the distribution of doses for those with toxicities. A single-tailed Welch T test was used to determine P for the probability that the mean of the distribution of values without toxicities is less than that with toxicities. The 15% and 25% quantiles for the distribution of doses with toxicities and the 75% and 85% quantiles for the distribution of doses without toxicities were used to demark dose-response regions.

Using the SSMS for each structure-DVH metric, we introduced a ranking metric combining elements for positive evidence of a threshold response (PETR). PETR was based on the AUC, with weighting factors (1-0) for sTP, ks, PPV, and SN.

$$PETR = AUC\, x\, LF_{sTP}(sTP, sTP_0, k_{sTP}) \; x\, LF_{ks}\,(ks, ks_0, k_{ks})\, x\, \frac{(PPV + SN)}{2} \quad (1)$$

We noted that AUC can be high when TP is small. Small values could be due to random events. To screen for the possibility of high AUC due to "noisy" data, we used a logistic function (LF_{sTP}) with coefficients selected so that LF_{sTP} = (0.5, 1.0) for sTP = (0.5, >1)

$$LF_{sTP}(sTP, sTP_0, k_{sTP}) = \frac{1}{1 + e^{k_{sTP}(sTP - sTP_0)}} \quad (2)$$

with $sTP_0 = 0.5$ *and* $k_{sTP} = -6/0.5$.

We noted that AUC can be high when the distribution of DVH metric values associated with the toxicity is not separated from, and higher than, the distribution of values without toxicity (ie, single sided ks is large). To screen distributions not demonstrating a transition to increased likelihood of toxicity with increasing dose (ie, a response-threshold) we used ks in a logistic function (LFks) with coefficients selected so that LFks = (0.5, 1) for ks = (0.1, < 0.01).

$$LF_{ks}(ks, ks_0, k_{ks}) = \frac{1}{1 + e^{ks_{ks}(ks - ks_0)}} \quad (3)$$

with $ks_0 = 0.1$, $k_{ks} = {}^{6}\!/_{0.09}$

Next, a machine learning model was used, like PETR, to rank each structure-DVH metric. Machine learning models are nondeterministic, vary substantially in selection of ranking metric (MLRM) used to score relative importance of input values, and frequently differ in which input variables are selected in models as most relevant for predicting outcomes.[16] For this study, random forest was selected using percent incremental increase in mean square error to rank the relative relevance of input variables (ie, MLRM = percent incremental increase in mean square error).

The product of PETR and MLRM was used for relative ranking of structure-DVH metrics for predictive ability, based on combined evidence from machine learning and more conventional statistical methods.

$$\text{SCA-ML} = \text{PETR} \times \text{MLRM} \quad (4)$$

Peak SCA-ML was used to cull the large number of candidate DVH metrics, selecting one physical and one bio-corrected DVH metric for each structure. These were categorized as primary and secondary according to their relative SCA-ML score. Absolute volume statistics (Dxcc [Gy], Dxcc [EQD2 Gy]) were dropped from consideration if x was greater than the 5% quantile of the structure volumes.

Minimum input set for multistructure predictive model

The minimal set of SCA-ML based metrics needed to predict dysphagia within the data set was identified through iterative construction of a machine learning model. Structures that were not drawn on at least 90% of the plans were excluded. For each remaining structure in the culled data set, the physical or biological dose metric with the largest SCA-ML was selected for the modeling data set (MDS). Plans with incomplete sets of structure-DVH metrics were excluded. At each iteration, 10-fold cross validation was used to calculate the average and standard deviation of the SP, SN, PPV, and negative predicted value across the folds.

A baseline model was first constructed using the full MDS as inputs. The next iterative construction of a minimal input model began with constructing single input models for each element of the MDS. The element with the largest sensitivity was selected as the first input element. Elements were incrementally added to the model and ranked according to sensitivity. Model iterations were stopped when the average SN was not significantly ($P <$ .05) different from the baseline value according to a Student's t test.

In routine clinical practice, physical doses are more readily available in commercial treatment planning systems than bio-corrected doses. Therefore, if the resulting model contained bio-corrected dose metrics, then the

process was repeated using the physical dose metric identified in the culled data set. The sensitivity of initial iterative model to the physical dose model was compared.

Results

Of the 439 patients examined, 132 (27%) had dysphagia that worsened from beginning of treatment. Of those with worsening dysphagia, 78 (16%) had a maximum grade ≥2. The median (25% quantile, 75% quantile) number of days from beginning of treatment to the highest recorded toxicity greater than or equal to grade ≥2, was 37 (22, 80) days. Figure 1 illustrates the time to maximum dysphagia score.

Seven swallowing structures were evaluated: esophagus, larynx, superior pharyngeal constrictor (SPC), inferior pharyngeal constrictor (IPC), parotids, and submandibular glands (SG). Parotids and submandibular glands were subcategorized according to their relative mean doses (parotid_high, parotid_low, SG_high, SG_low).

In the analysis, 738 structure-DVH metrics were calculated and ranked for evidence for predicting dysphagia using SCA-ML. The top 18 are presented in Table 2. Primary (checked ✔) and secondary structure−DVH metrics identified with the SCA-ML are listed in Table 2. In order of decreasing SCA-ML, the top 3 primary structure-DVH identified in the MDS were SPC D20% (EQD2 Gy) ≥47.7, parotid_low: D80% (EQD2 Gy) ≥6, SG_high D35% ≥61.7. The top secondary structure-DVH metric was SPC D25% (Gy) ≥50.4.

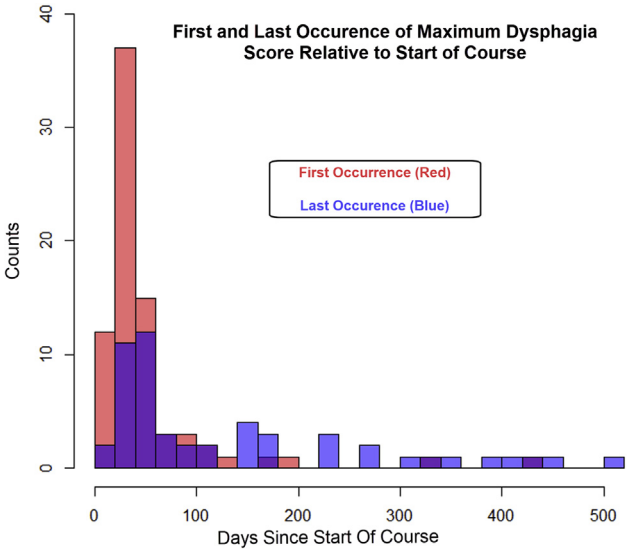

Figure 1 For patients demonstrating dysphagia scores that worsened from start of treatment, the median time to the first maximum toxicity record was 37 days. Median time to the last occurrence of the maximum score was 48 days.

Both SG_low D45% (Gy) ≥28.8 and larynx D25% (Gy) ≥21.2 Gy had high SCA-ML scores. They were not present on at least 90% of the treatment plans. Reasons include involvement in the target volume (eg, cancer of the larynx), laryngectomy, or removal as part of neck dissection.

Figure 2 illustrates statistical DVH curves for the physical and bio-corrected doses to the SPC, and for physical doses to SG_high, SG_low, larynx.[9] Curves are color coded for patient subsets with and without worsening dysphagia scores. Statistical DVHs show the median Dx% (Gy or EQD2 Gy) values (dotted line) layered with a shaded area encompassing the central 70% of Dx% values to highlight where subsets separate.

Figure 3 illustrates application of the method for physical and bio-corrected doses to the SPC and for physical doses to SG_high, SG_low, larynx. In Fig 3b, SPC Dx% (EQD2 Gy) AUCs did not vary greatly with volume or highlight specific narrow regions with evidence for response thresholds (ks). Fractional volumes of 15% to 35% demonstrated the region with the strongest evidence based on PETR scores. Note in the figure the low predictive strength near median (Gy). Also note that although AUC was elevated near to Max (Gy) (ie, D0% [Gy]), SCA-ML scoring indicated low combined evidence for dose-response threshold.

Figure 4 shows the toxicities along with the SCAL-ML identified thresholds. A logistic regression of the data was used to characterize the overall probability of toxicity for each structure independent of the others. Comparing distributions for physical and bio-corrected SCP doses, D20% (EQD2 Gy) and D25% (Gy) graphically demonstrated dose-response thresholds with similar SCA-ML (4.092 vs 4.067) and PETR (5.4 vs 4.3) scores.

SG_low and the larynx had high scores but were excluded from the multistructure model because they had only been contoured on 95 out of 132 of the treatment plans. In the multistructure iterative model construction, there were 108 complete data sets in the MDS for the 5 candidate structures (SPC, IPC, esophagus, SG_high, parotid_high, and parotid_low) that had been contoured on at least 90% of treatment plans. The baseline sensitivity of the model constructing using the 5 primary structure-DVH metrics was 0.79 ± 0.21. Only one structure-DVH metric input, D20% (EQD2 Gy), was needed in the iterative model to achieve sensitivity comparable to the baseline. Although SPC D20% (EQD2 Gy) ≥47.7 had a higher relative risk than D25% (Gy) ≥50.4 (20.7 vs 7.1) in the SSMS, the overall sensitivity (0.78 ± 0.18 vs 0.76 ± 0.26) and AUC (0.70 ± 0.16 vs 0.70 ± 0.15) of the iteratively constructed, cross validated random forest models was comparable.

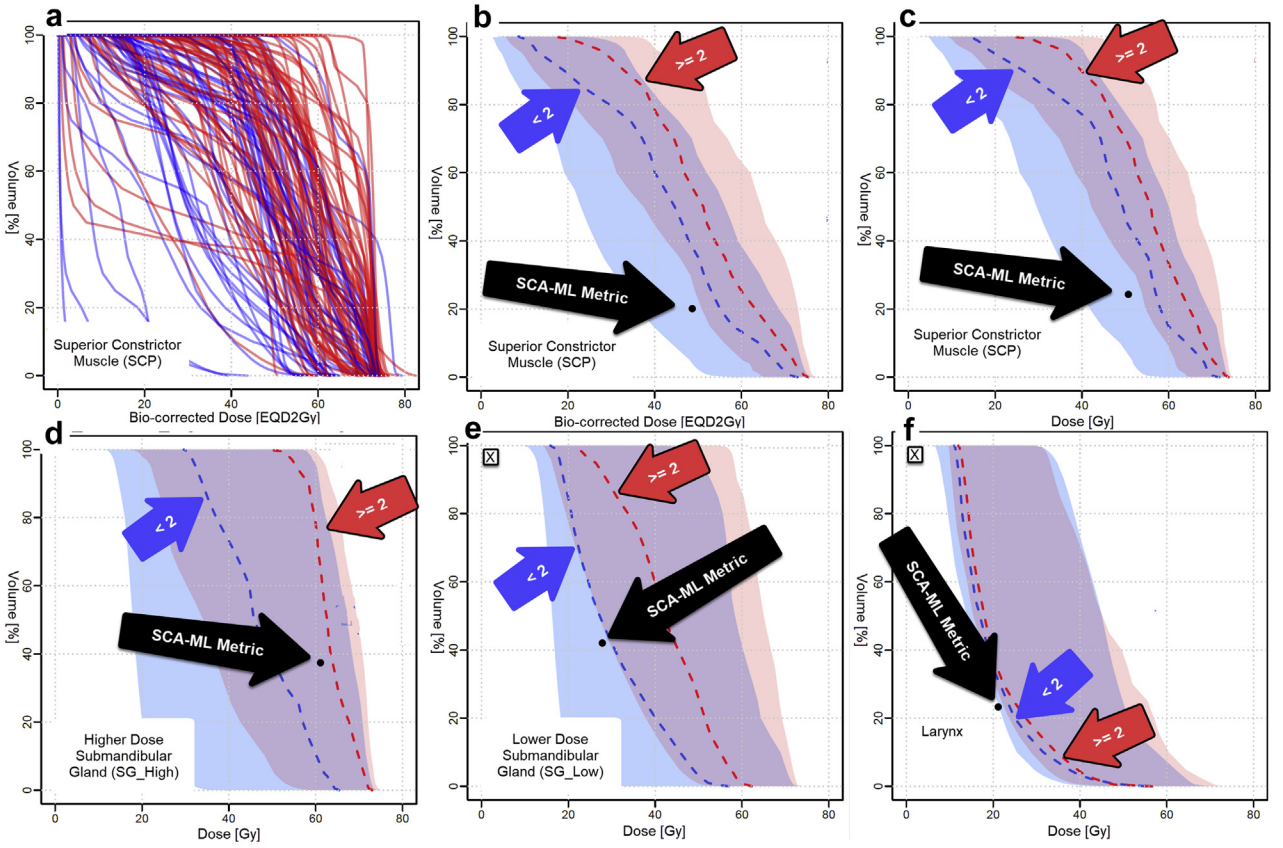

Figure 2 (a) Plots of statistical dose volume histograms (DVH) curves. Superior constrictor muscle (SCP) bio-corrected DVH curves are shown for patients with (red) and without (blue) worsening dysphagia. To clarify visualization and provide more quantitative detail, statistical DVH curves show median (dashed line) and 70% confidence intervals of DVH curves for (b) SCP Dx% (EQD2 Gy), (c) SCP Dx% (Gy), (d) the submandibular gland receiving the higher relative mean dose (SG_high) Dx [Gy], (e) the submandibular gland receiving the lower relative mean dose (SG_low) Dx [Gy], and (f) larynx Dx [Gy]. SG_low and larynx were not included in multi-structure model due to lack of contouring on at least 90% of plans (☒). The DVH metric and threshold with the highest combined statistical categorization algorithm and machine learning (SCA-ML) score is shown for each (black dot).

Discussion

Combining the big data analytics resource system with artificial intelligence enabled systematic investigation of a much larger range of structure-DVH metrics than used by other studies, using historic evidence to identify a minimal set of clinically actionable metrics and thresholds. This provides a means to incrementally improve the set of constraints used.

Although AUC is useful, we did not find it necessarily sufficient as a sole metric for identification of dose-response thresholds. To add levels of evidence, we introduced PETR as an algorithmic method for layering combined information from conventional statistical measures that have well understood interpretability (ks, sensitivity, positive predictive value) onto AUC. We further extended the approach, by layering on "importance" metrics used by machine learning algorithms, such as random forest by introduction of SCA-ML. This layered approach enabled illustrating where combined evidence of different types of measures agree.

The purpose for use of ML in the method was not to generate a specific model for predicting toxicity. Instead, the approach combined evidence from statistical categorization, ML and iterative construction of parsimonious model to winnow a large number of candidate inputs down to a minimal set of DVH metric inputs and thresholds with the strongest clinical evidence for increasing dose contributing to increasing toxicity. This method provides a means to follow observational data accumulated in the BDARS to identify inputs that are also clinically actionable. By objectively comparing both physical and biologically corrected doses with absolute and percentage volume cut points, it avoids a-priori judgment, of which is most relevant. In this case 738 candidate model metrics were winnowed down to the one with the strongest combined levels of evidence that was also actionable in a routine clinical setting.

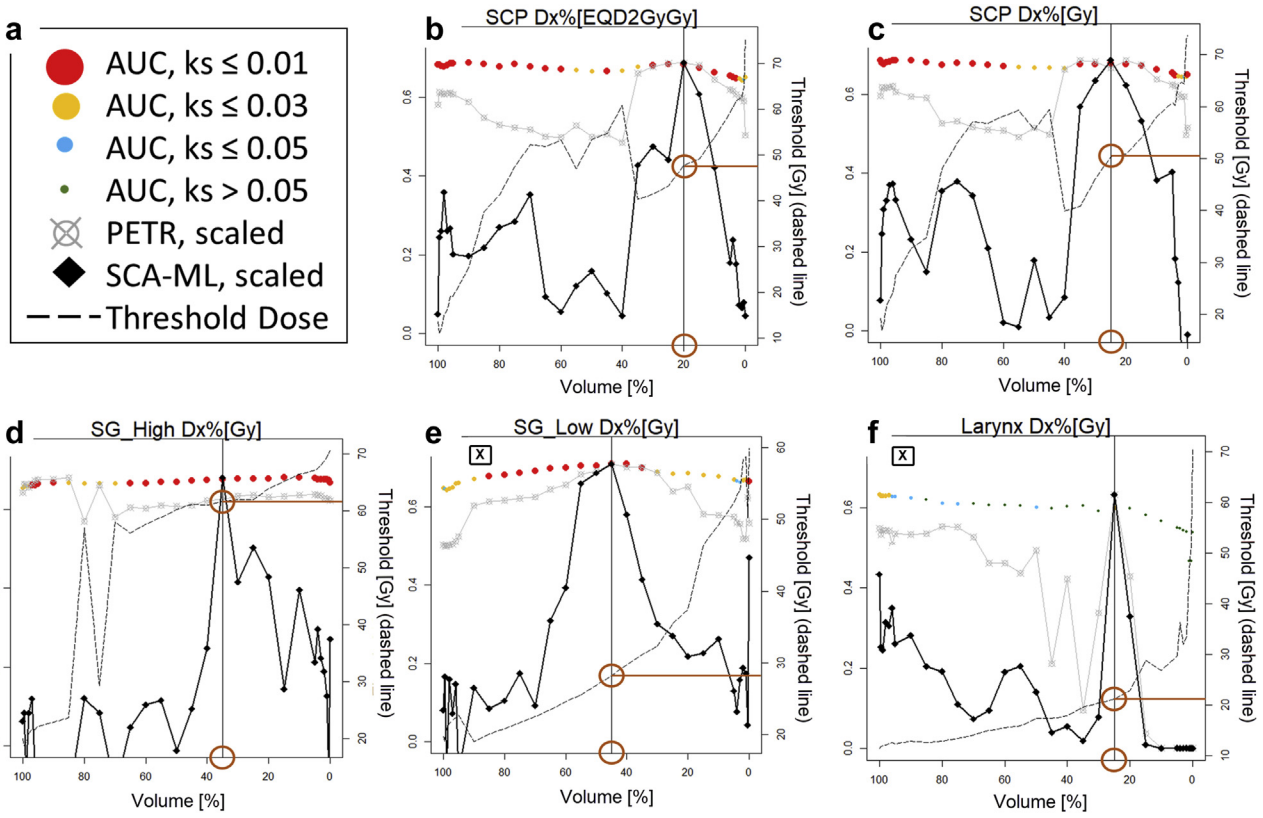

Figure 3 (a) Illustration of combined statistical categorization algorithm and machine learning (SCA-ML) plots for determining dose volume histograms (DVH) metric demonstrating strong evidence of dose-response threshold for (b) superior constrictor muscle (SCP) Dx% (EQD2 Gy), (c) SCP Dx% (Gy), (d) the submandibular gland receiving the higher relative mean dose (SG_high) Dx (Gy), (e) the submandibular gland receiving lower relative mean dose (SG_low) Dx (Gy), and (f) larynx Dx (Gy). Area under the curve (AUC) values are plotted for each metric with color coding and symbol size differentiating *P* values for Kolmogorov-Smirnov test. Positive evidence of threshold response (PETR) and SCA-ML scores are scaled using the highest relative value to select metric. The threshold dose determined for each metric is plotted (dashed line). Peak SCA-ML values and thresholds are circled on the graph.

Without the advantage of a BDARS, prior studies have used substantially smaller sets of patients and of metrics tested for predicting various endpoints related to dysphagia. In a 2007 study of 36 patients who examined a total of 15 physical-dose based DVH metrics for 3 swallowing structures, Feng et al found that total pharyngeal constrictor (PC) mean (Gy) >60, V65 Gy (%) >65, and supraglottic larynx mean V50 Gy (%) >50 values had strong correlations with videoflouroscopy based aspirations.[2] They found that only PC mean (Gy) was correlated with both patient- and provider-rated worsening of swallowing solids.

In a 2010 study of 83 evaluable patients, Caudell et al examined 16 physical dose DVH metrics for 2 swallowing structures.[5] They reported glottis and supraglotic larynx (GSL) V55 Gy (%) <32 and IPC V60 Gy (%) <11.8 were significant for stricture and risk of aspiration with odds ratios of 1.03 and 1.02, respectively. Larynx mean (Gy) ≥41 and V60 Gy (%) >24 in addition to IPC V60 Gy (%) >12 were significant for percutaneous endoscopy gastrostomy tube dependence and aspiration. SPC V65 Gy (%) ≥33 and IPC V65 Gy (%) ≥75 were associated with pharygoesophageal stricture that required dilation. Median time to diagnosis of stricture was 7 months. No aspiration was noted for larynx mean (Gy) ≤40.6.

In a 2011 study of 73 patients, Eisbruch et al[3] found that esophagus mean (Gy) ≥48 was significant for strictures. For increased video fluoroscopy-based aspiration, scoring of PC mean (Gy) >56 and GSL mean (Gy) >39 correlated with 25% toxicity incidence. They examined 5 physical dose DVH metrics for 6 structures: SPC, IPC, mid pharyngeal constrictors and PC, GSL, and esophagus.

In a 2017 study, Chera et al reported on 9 out of 45 patients studied with worsening dysphagia scores at 6 months.[4] Limiting their study to fractional volumes receiving physical doses, they found that for SPC V55 Gy (%) ≥78 and V60 Gy (%) ≥40 were associated with 20% risk of toxicity. They reported 6 patients evaluated at 12 months. They did not find dose associations with esophagus, IPC, or middle constrictor muscles.

In a 2018 study, Kamal et al[17] reported on 30 out of 97 patients found with moderate to severe radiation induced dysphagia at 3 to 6 months after XRT, using the Dynamic Imaging Grade for Swallowing Toxicity ≥2.

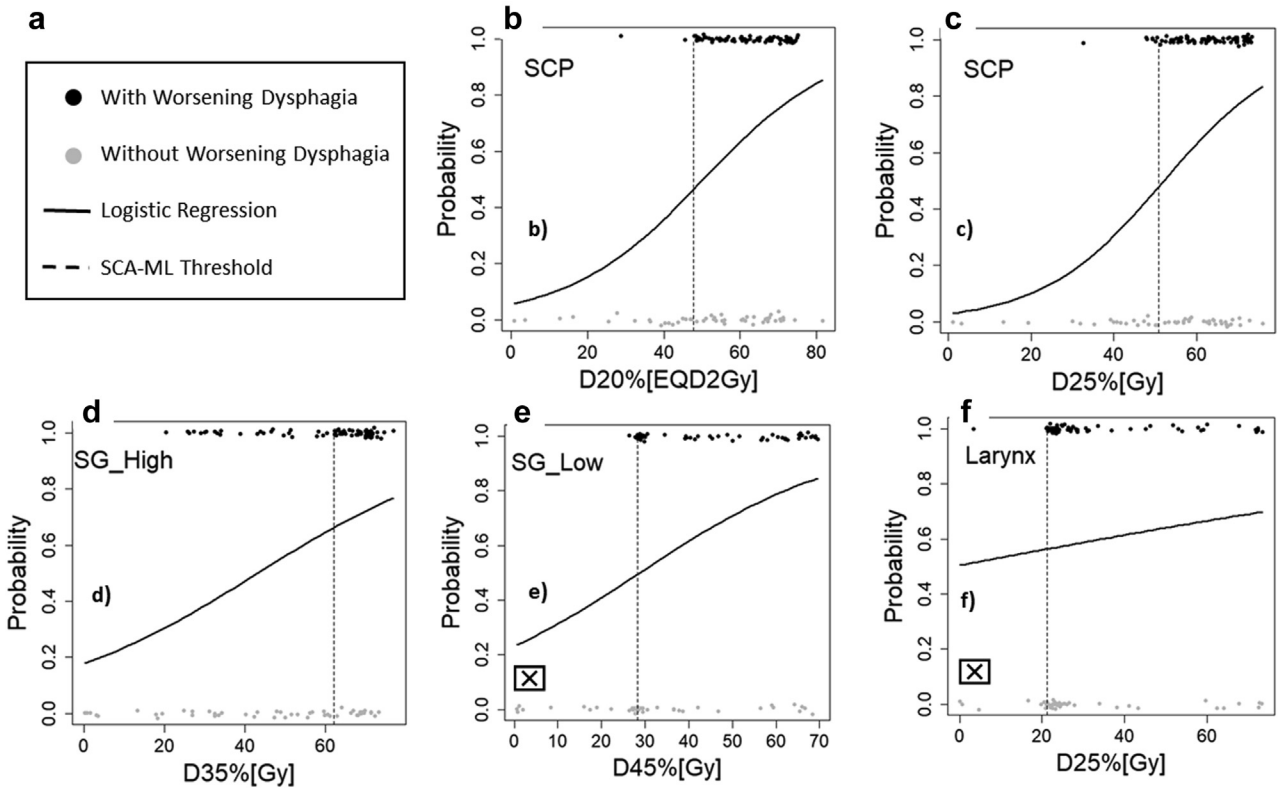

Figure 4 (a) Univariate plots of worsening dysphagia versus ranking metrics using combined statistical categorization algorithm and machine learning (SCA-ML) selected for (b) superior constrictor muscle (SCP) Dx% (EQD2 Gy), (c) SCP Dx% (Gy), (d) the submandibular gland receiving the higher relative mean dose (SG_high) Dx (Gy), (e) the submandibular gland receiving lower relative mean dose (SG_low) Dx (Gy), and (f) larynx Dx (Gy). Threshold corresponding to peak SCA-ML is plotted (dashed line) to highlight association with the distribution. A small amount of noise was added to the binary outcome, to reduce point overlap masking the density of points. A logistic regression is plotted to characterize probability of toxicity.

They identified geniohyoid muscle V61 Gy (%) ≥18.6 was the strongest predictor. SPC V55 Gy (%) ≥97.5 and supraglottic area V23 Gy (%) ≥92.5 were also identified as predictive.

Our specific findings that SPC D20% (EQD2 Gy) ≥47.7 and D25% (Gy) ≥50.4 are strongly associated with dysphagia are more specific, but consistent with the results of Chera et al and Caudell et al.[4,5] The finding that SG_high D35% ≥61.7 was predictive may be a surrogate for sensitivity of the proximal musculature. That interpretation is consistent is with the finding of finding of Kamal et al for the geniohyoid muscle. Sparing at least one salivary structure conveyed benefit for reducing odds for worsening dysphagia. Higher observed sensitivity of SG_low D45% (%) ≥28.2 compared with parotid_low D65% (Gy) ≥13.2 (0.95 vs 0.65) at minimum signals the importance of routine contouring of these structures and monitoring of their doses, which is consistent with the results of Jackson et al.[16]

The studies of Feng et al,[2] Eisbruch et al,[3] and Caudell et al[5] focused on mean dose to the larynx or GSL and identified differing thresholds. Drawing from these early results, the historic plans examined in this data set had used larynx:mean (Gy) ≤50 as a high priority constraint. The finding that D25% (Gy) ≥21.2 had a high sensitivity (SN = 0.97) suggests that controlling dose to small volumes may convey additional advantage.

Esophagus was noteworthy for identifying absolute versus a percentage volume D2cc [Gy] ≥22.6 as the strongest predictor. One interpretation is that the small volume of the esophagus proximal to the larynx could act as a surrogate measure for larynx dose. Additional inspection of the relative location of these sub volumes would be needed to confirm that interpretation.

Historic plans had been created using IPC:mean (Gy) <20 as a high priority constraint. D90% (Gy) ≥12.8 reinforced use of the historic constraint to reduce doses to IPC. This highlights an important point to be noted in modeling dose responses. Results should be viewed in the context of intrinsic biases introduced by dose constraints used in creating treatment plans. In this instance not finding median (D50% [Gy]) dose more significant than D90% (Gy), could mean that the metric has already been sufficiently constrained by the default mean (Gy) <20 constraint and that significance of D90% (Gy) signals potential to augment, not replace, this default metric.

Ability to use historic data gathered from routine practice, by combining the BDARS with AI,

underscores the importance of consistency in contouring approaches within and among clinics. For example, we noted substantial differences in sensitivity of SPC versus IPC metrics for predicting worsening dysphagia. This highlights importance of contouring these structures separately. Other clinics may only contour a generalized PC structure as part of their practice guidelines. In that case, those clinics would miss the opportunity to detect differences for predicting toxicities or to use that information to reduce toxicities. Similarly, high SCA-ML scores for the parotid and submandibular gland structures underscore the value of consistently contouring both (if unresected) as part of routine treatment planning.

The potential for use of observational clinical data coupled with AI to improve hypothesis generation in design processes for randomized controlled trials has been discussed previously.[10] The method described here illustrates a potential example. Results provide strong levels of evidence for selection of specific DVH metrics and associations that could be tested in a subsequent multi-institutional trial. Evidence that larynx and SG_low DVH metrics may play a second order role to SC in predicting dysphagia underscore the need for consistent contouring of these structures to detail interactions in such a trial. Observation of the natural history occurrence of toxicity (Fig 1) could provide more specific guidance for selection of measurement time intervals.

Conclusions

By combining a big data analytics resource system with an AI algorithm, we were able to examine evidence for response thresholds for a much larger set of patients and DVH metrics than conventional approaches. Calculating both physical and biologically corrected doses and percentage and absolute volume DVH metrics, the approach was better able to follow the data and minimize metric selection bias. This presents a means that can in be automated to enable iterative learning from historic treatments to inform decision frameworks for future patients with clinically apprehensible metrics.

References

1. Gillison ML, Trotti AM, Harris J, et al. Radiotherapy plus cetuximab or cisplatin in human papillomavirus-positive oropharyngeal cancer (NRG Oncology RTOG 1016): A randomised, multicentre, non-inferiority trial. *Lancet*. 2019;393:40-50.
2. Feng FY1, Kim HM, Lyden TH, et al. Intensity modulated radiotherapy of head and neck cancer aiming to reduce dysphagia: Early dose–effect relationships for the swallowing structures. *Int J Radiat Oncol Biol Phys*. 2007;68:1289-1298.
3. Eisbruch A, Kim HM, Feng FY, et al. Chemo-IMRT of oropharyngeal cancer aiming to reduce dysphagia: Swallowing organs late complication probabilities and dosimetric correlates. *Int J Radiat Oncol Biol Phys*. 2011;81:e93-e99.
4. Chera BS, Fried D, Price A, et al. Dosimetric predictors of patient-reported xerostomia and dysphagia with deintensified chemoradiation therapy for HPV-associated oropharyngeal squamous cell carcinoma. *Int J Radiat Oncol Biol Phys*. 2017;98:1022-1027.
5. Caudell JJ, Schaner PE, Desmond RA, et al. Dosimetric factors associated with long-term dysphagia after definitive radiotherapy for squamous cell carcinoma of the head and neck. *Int J Radiat Oncol Biol Phys*. 2010;76:403-409.
6. Mazzola R, Ricchetti F, Fiorentino A, et al. Dose-volume-related dysphagia after constrictor muscles definition in head and neck cancer intensity modulated radiation treatment. *Br J Radiol*. 2014;87:20140543.
7. Mayo CS, Kessler ML, Eisbruch A, et al. The big data effort in radiation oncology: Data mining or data farming? *Adv Radiat Oncol*. 2016;1:260-271.
8. Mayo CS, Phillips M, McNutt T, et al. Treatment data and technical process challenges for practical big data efforts in radiation oncology medical physics. *Med Phys*. 2018;45:e793-e810.
9. Mayo CS, Yao J, Eisbruch A, et al. Incorporating big data into treatment plan evaluation -development of statistical DVH metrics and visualization dashboards. *Adv Radiat Oncol*. 2017;2:503-514.
10. Mayo CS, Matuszak MM, Schipper MJ, et al. Big data in designing clinical trials: Opportunities and challenges frontiers in oncology. *Front Oncol*. 2017;7:187.
11. Mayo CS, Moran JM, Bosch W, et al. American Association of Physicists in Medicine task group 263: Standardizing nomenclatures in radiation oncology. *Int Journal Radiat Oncol Biol Phys*. 2018;100:1057-1066.
12. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Available at: https://www.R-project.org/. Accessed January 16, 2020.
13. Andy Liaw A, Wiener M. Classification and regression by random forest. *R News*. 2002;2:18-22. Available at: https://CRAN.R-project.org/doc/Rnews/. Accessed January 16, 2020.
14. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
15. Kuhn M, Wing J, Weston S, et al. Caret: Classification and regression training, 2018 R package version 6.0-80. Available at: https://CRAN.R-project.org/package=caret. Accessed January 16, 2020.
16. Jackson WC, Hawkins PG, et al. Submandibular gland sparing when irradiating neck level IB in the treatment of oral squamous cell carcinoma. *Med Dosim*. 2019;44:144-149.
17. Kamal M, Mohamamed ASR, Volpe S, et al. Radiotherapy dose-volume parameters predict videofluoroscopy-detected dysphagia per DIGEST after IMRT for oropharyngeal cancer: Results of a prospective registry. *Radiother Oncol*. 2018;128:442-451.