Plan Scoring Metric Using Only DVH Information for Evaluation and Comparison of IMRT Treatment Plans

James Giltz, William Weaver, Mark Deweese
Alyzen Medical Physics, University of Kentucky

Purpose

To develop a meaningful metric which can be used to evaluate the quality of IMRT treatment plans (both absolutely and comparatively) using only information from a dose volume histogram. This can provide quick evaluation to give the reviewer an initial indication as to plan quality.

As automation in treatment planning continues to develop, tools are necessary for modeling what a "good" plan looks like and how several plans compare. Up until very recently, this has been entirely a data-driven but subjective judgment of clinicians in practice. If there is a mathematical model that can simulate the kinds of judgements made by these clinicians, it may facilitate modeling for plan scores that would closely match actual physicists and physicians in practice.

Objectives & Tools

Objectives:

- Design a mathematical model that approximates clinician evaluations of treatment plans
- Design and define any structures needed to refine the ranking system
- Determine a single set of constraints and action levels that can be applied to plans for evaluation
- Evaluate clinically approved plans of varying sites/doses to determine how well the model agrees with clinicians
- Create and evaluate plans of varying quality and determine how well the system can compare plans on the same structure set.
- Evaluate utility for routine clinical use, and determine if model will need fine tuning between clinics/clinicians or if there is a stable set of values that will apply for all.

Tools:

- All treatment plans created on Varian Eclipse 13.7, using: AAA 11.0.3.1, DVO 11.0.3.1
- All treatment plans created using Static Gantry IMRT on a Varian IX machine
- All treatment plans created with 6X only.
- Analysis initially completed with Microsoft Excel, later tools have been created for rapid analysis in a proprietary program.
- The dose constraint form used is a combination of the constraints provided by Mobius and Quantec Clinically used at a site covered by Alyzen Medical Physics; sample shown below:

Organs	Dose Limits	% Vol
Brachial Plexus[1]	6600	MAX
Brachial Plexus[1]	6000	5%
Brainstem[2]	5400	MAX
Brainstem (PTV)[2]	6000	1%
Brain[17]	6000	MAX
Cochlea Lt[17]	4500	MEAN
Cochlea Lt[3]	5500	5%
Cochlea Rt[17]	4500	MEAN
Cochlea Rt[3]	5500	5%
Ear Lt (Inner/Middle)[4]	5000	MEAN
Ear Rt (Inner/Middle)[4]	5000	MEAN
Eye Lt[3]	5000	MAX
Eye Lt[4]	3500	MEAN
Eye Rt[3]	5000	MAX
Eye Rt[4]	3500	MEAN
Glotic Larynx[4]	4500	MEAN
Lens Lt[5]	2500	MAX
Lens Rt[5]	2500	MAX
Mandible[2]	7000	MAX
Mandible (PTV)[2]	7500	1cc
Optic Nerve Lt[2]	5400	MAX
Optic Nerve Lt (PTV) [2]	6000	1%
Optic Nerve Rt[2]	5400	MAX
Optic Nerve Rt (PTV) [2]	6000	1%
Optic Chiasm [17]	5500	MAX
Oral Cavity - PTV[3]	4000	MEAN
Parotid Gland Lt [6]	2600	Mean
Parotid Gland Lt [6]	3000	50%
Parotid Gland Rt [6]	2600	Mean

Organs	Dose Limits	% Vol
Parotid Gland Rt [6]	3000	50%
Parotid Gland (both)[6]	2000	20cc
Temporal Lobes[2]	6000	MAX
Temporal Lobes (PTV)[2]	6500	1%
T-M Joint Lt[2]	7000	MAX
T-M Joint Lt (PTV)[2]	7500	1cc
T-M Joint Rt[2]	7000	MAX
T-M Joint Rt (PTV)[2]	7500	1cc
Tongue [2]	5500	MAX
Tongue (PTV)[2]	6500	1%
Larynx[17]	5000	27%
Larynx[17]	4400	MEAN
Pharangeal Constrictor[10]	5400	MEAN
Pharangeal Constrictor[10]	5000	51%
Pharangeal Constrictor[10]	5200	60%
Esophagus[17]	3400	MEAN
Esophagus[7]	5400	15%
Esophagus[7]	4500	33%
Heart[8]	6000	33%
Heart[8]	4500	67%
Heart[8]	4000	100%
Lung (Right and Left)[9]	2000	MEAN
Lung (Right and Left)[17]	2000	30%
Duodenum[11]	6000	MAX
Duodenum[11]	4500	33%
Rt Kidney[12]	5000	33%
Rt Kidney[12]	3000	67%
Rt Kidney[12]	2300	100%

Organs	Dose Limits	% Vol
Lt Kidney[12]	5000	33%
Lt Kidney[12]	3000	67%
Lt Kidney[12]	2300	100%
Liver[13]	3500	50%
Liver[13]	3000	100%
Small Intestine[14]	5000	MAX
Small Intestine[14]	4500	65cc
Small Intestine[14]	4000	100cc
Small Intestine[14]	3500	180cc
Stomach[11]	5400	MAX
Stomach[11]	5000	2%
Stomach[11]	4500	25%
Bladder[15]	8000	15%
Bladder[15]	7500	25%
Bladder[15]	7000	35%
Bladder[15]	6500	50%
Rectum[17]	7500	15%
Rectum[17]	6500	25%
Rectum[17]	6000	35%
Rectum[17]	5000	50%
Lt Femoral Head[14]	5000	MAX
Lt Femoral Head[14]	4500	25%
Lt Femoral Head[14]	4000	40%
Rt Femoral Head[14]	5000	MAX
Rt Femoral Head[14]	4500	25%
Rt Femoral Head[14]	4000	40%
Penile Bulb[15]	5250	MEAN
Spinal Cord[16]	4500	MAX

Materials & Methods

Beginning with a simple linear/non linear piecewise function, equations were iteratively developed that were able to significantly differ in value based on a clinical priority, which are manually set to be between 1 and 5 based upon the subjective "value" of that constraint goal. This created a structure in which minor penalty is associated under a threshold, and then a non-linear section which approaches a power function with further deviation from the goal. It was determined additional evaluation criteria were needed to give the DVH sufficient information for evaluation, as there are other factors besides organ at risk doses to consider. For this iteration of the metric and for easy analysis, the only additional criteria used was the absolute max dose of the plan. The constants were chosen and applied in such a way that a score of more than 100 would represent a failing plan.

Priorities had to be set for each constraint evaluated and these were determined in consultation of the authors based on their combined clinical experience of how both physicists and physicians tended to view the relative importance of critical structures. It was agreed that specific clinics may wish to alter these depending on their clinical priorities.

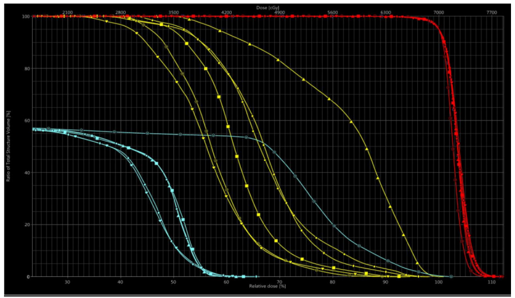
For initial evaluation, 30 patients' plans were evaluated directly with the following approximate distribution, representing the last year of IMRT treatments (non-Stereotactic) at one of the clinics covered by Alyzen Medical Physics.

- 10 Prostate plans
- 9 Head and Neck Plans
- 4 Pelvis (Anus/Rectum) plans
- 5 Lung plans
- 2 GYN plans

For all plans, we grouped scores as follows: 100+ - Failure, 20-100 – potentially clinically acceptable, 2-20 – good, 0-2 – passing well.

In order to evaluate the metric's ability to compare varied plans on the same structure set, one clinically accepted plan was reoptimized to create several variants. Some of these were potentially clinically acceptable and others were intentionally made to be unacceptable. Each plan was scored using the metric. The scores were then evaluated to determine how well the score matched with our clinical impression of the plan variations.

All plans are normalized such that coverage on PTVs meets physician intent, then analyzed.



Results

Prostate:

The average penalty applied to prostate plans was 87.76, with a maximum score of 290.88 and a minimum score of 1.47. Of these 10 plans, 3 were noted as failures, 5 were potentially clinically acceptable, and 2 were considered passing well.

While investigating the plans that scored higher it was noted that the institution from which the patients were selected used significantly different constraints on the small bowel and femur/hip objectives. Correcting for this difference the average changed to 51.86, the maximum to 220, and minimum to 1.47. With adjusted scoring, 2 were noted as failures, 2 were noted as potentially clinically acceptable, 4 were noted as good, and 2 were considered passing well.

Head and Neck:

The average penalty applied to head and neck plans was 26.73, with a maximum score of 49.43, and a minimum score of 5.95. Of these plans, 6 were considered potentially clinically acceptable, and 3 were considered good. None were considered failures or to be passing well.

Lung:

The average score for lung plans was 131.31, with a maximum score of 304.71, and a minimum score of 4.45. Of the plans, 3 were considered failures and 2 were considered good. None were considered passing well, or potentially clinically acceptable.

Pelvis:

The average score for lung plans was 13.08 with a maximum score of 45.2, and a minimum score of 0.26. Of the plans, 1 was considered potentially clinically acceptable, 1 was considered good, and 2 were considered to be passing well. None were considered failures

GYN:

The scores for the 2 GYN plans were 5.97 and 21.28 resulting in a good and potentially clinically acceptable plan respectively.

Plan Comparison:

Starting with a head and neck plan that had been considered good, 5 plans were designed with varying constraint failures to test the model's ability to differentiate between plans on a structure set, results can be found in the figures below/on the left.

Tonsil Head and Neck - 70Gy		
Plan Name	Score	Comments/Evaluation
Good_Test	10.75	Minor failures on both parotid glands and the hot spot was slightly high, otherwise everything meets well and plan looks good on evaluation.
LarynxMajor	190.95	No constraint on Larynx, received RX dose - bad plan
LarynxMinor	26.88	Larynx fails, though not horribly. Parotids and hot spot virtually unchanged, may be accepted
Cord_min	23.11	Spinal cord barely fails, structure represents the entire spinal canal - may be accepted
Cord_maj	1040	Removed constraint from spinal cord, received RX dose - bad plan
cord&laryn	38.02	Both spinal cord and Larynx barely fail to meet constraints, still potentially clinically acceptable.

In the DVH to the left, the PTV is in red, the Larynx is in yellow, and the Spinal Cord is in Cyan.

Discussion

The two failing prostate plans on review were plans that had taken a significant amount of time to plan, a large amount of physician involvement, and were not clinically considered stellar – rather treated out of necessity. In this case, the model accurately shows this.The difference in the constraints on the prostate plans highlights a need for different models built between facilities/physicians, as demonstrated by the significant difference in the prostate evaluations between the original and modified model.

During the analysis of the plans it became clear that a single model would be insufficient to analyze all types of treatment plans. It is also evident that different treatment sites will need more differentiated models. For example, lung plans were very binary – either passing well or failing horribly, entirely related by proximity to the esophagus. Additionally, with the Lung, Pelvis, and GYN plans, another issue that arose was a lack of constraints – the relative low number resulted in a lack of specificity provided by the model. In a clinical setting, more constraints may be added to these models to better represent their differences.

It was also determined during evaluation that another structure that would be of benefit is an "Outside PTV" structure (Body-(PTV+1cm)) in which a cc volume based penalty could be applied (as opposed to dose). In addition this could be used for further homogeneity limits within target structures, or volumetric evaluation of critical structures. We feel the same principle could be used as in the dosimetric evaluation, however different constants would be needed within that model.

Conclusion

This data suggests that the model created is a valid system for plan evaluation. It is also functional for both the goal of absolute evaluation and comparison between plans on the same structure set

It is recommended that different scoring parameters be applied between different types of plans (Prostate/H&N/Lung, etc) so that more meaningful determinations can be made compared to a single metric applied for all treatment locations. There are two primary methods to achieve this score differentiation:

- Apply Different constants throughout the equation to change the form of the cost curve, thereby letting all plans fall on the same 0-100 scale.
- Use the same constants, and set different score ranges for different treatment areas.

It has also been determined that models will need to be created by each facility so that the constraints and valuations can be made relative to the clinicians at the particular location.

Only two additional structures have been deemed necessary for conventional IMRT treatments – Outside PTV, and overall hot spot as defined in methods and discussion. More complex treatments such as Stereotactic plans would likely require more structures be created.

Contact

James Giltz, M.S. – jgiltz@alyzenmed.com
William Weaver, M.S., DABR – wweaver@alyzenmed.com
Mark Deweese, M.S., DABR – mdeweese@alyzenmed.com