Developing a Quality Index for Dose-Volume Histograms Based on Physician Preference

by

Alex Tyler Price

Graduate Program in Medical Physics
Duke University

Date:_______________________
Approved:

___________________________
Joseph Y. Lo, Supervisor

___________________________
Shiva K. Das

___________________________
Robert Reiman Jr.

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Graduate Program in Medical Physics in the Graduate School of Duke University

2015
ABSTRACT

Developing a Quality Index for Dose-Volume Histograms Based on Physician Preference

by

Alex Tyler Price

Graduate Program in Medical Physics
Duke University

Date:_______________________

Approved:

___________________________
Joseph Y. Lo, Supervisor

___________________________
Shiva K. Das

___________________________
Robert Reiman Jr.

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Graduate Program in Medical Physics in the Graduate School of Duke University

2015
Abstract

The purpose of this study was to create a prostate IMRT plan quality index that may be used to quantitatively compare competing plans using a methodology that mimics physician preference. This methodology allows planners to choose between plans with equivocal characteristics, prior to physician scrutiny.

A training observer preference study was conducted to gather data from 3 radiation oncology physicians who ranked a set of 20 patients (each with 5 plans). The rankings were used to optimize a plan quality index that combined weighted portions of the rectum, bladder, and planning target volume DVHs, such that the relative plan quality index mimicked physician rankings as best as possible. Once optimized, a validation study assessed the plan quality index by comparison to physician rankings in a new set of 25 patients (each with 4 plans). The physician group in the validation study included 6 physicians, 5 of whom were not included in the training study. Plan quality indexes were evaluated against the physicians’ rank list using Spearman rank correlation. To validate the algorithm’s ability to rank plans across the whole scoring spectrum, a set of 15 patients was presented to the same 6 physicians to be ranked in order of quality.

The optimized plan quality index combined the following DVH features: high dose regions above 40Gy/60Gy (rectum/bladder), organ weightings, and PTV shoulder
coverage. Mean correlation of the plan quality index vs. physicians’ rankings in the training study was 0.507 (range: 0.345-0.706). By comparison, the mean correlation between physicians was 0.301 (range: 0.242-0.334). The mean correlation of the plan quality index vs. physician rankings in validation study was 0.726 (range: 0.416-0.936), indicating robustness of the plan quality index by virtue of producing similar results to the training study. Intra-physician correlation was 0.564 (range: 0.398-0.689). For the inter-patient study, the mean correlation of the plan quality index vs. the physicians was 0.821 (range: 0.657-0.964). By comparison, the mean correlation between the physicians was 0.690 (range: 0.570-0.784).

The correlation coefficients of the plan quality index vs. physicians were similar to and higher than the correlation coefficients of the physicians with each other, implying that the plan quality index developed in this work shows promise in reflecting physician clinical preference when selecting between competing, dosimetrically equivocal plans.
Dedication

I dedicate this work to my mother Tammi Price. I would not be where I am today without your parenting. Thank you for all that you have done and you deserve eternal happiness.
## Contents

Abstract ......................................................................................................................................... iv

List of Tables ................................................................................................................................. ix

List of Figures ................................................................................................................................ x

Acknowledgements .................................................................................................................... xii

1. Introduction ............................................................................................................................... 1

2. Methods and Materials: ............................................................................................................ 7

  2.1 Study Development.......................................................................................................... 7

  2.2 Training Intra-Patient Study ......................................................................................... 11

  2.3 Formulation of First Algorithm, PPA.1 ....................................................................... 11

  2.4 Intra-Patient Validation Study .................................................................................... 17

  2.5 Formulation of Second Algorithm, PPA.2 .................................................................. 18

  2.6 Inter-patient Study ......................................................................................................... 21

3. Results ....................................................................................................................................... 26

  3.1 Training Intra-patient Study ......................................................................................... 26

  3.2 Validation Intra-patient Study ...................................................................................... 27

  3.2.1 Analysis of Intra-Patient Results ............................................................................. 28

  3.3 Inter-patient Study ......................................................................................................... 29

  3.3.1 Analysis of Inter-Patient Results ............................................................................. 31

4. Discussion ................................................................................................................................ 34

  4.1 Future Work .................................................................................................................... 37
List of Tables

Table 1: Scores used to validate the scoring spectrum from the smallest score (2.62) to the highest score (7.797) ........................................................................................................................................ 22

Table 2: (Left) Represents the mean Spearman rank correlation of PPA.1 with each of the oncologists in the training study. (Right) Represents the mean Spearman rank correlation of each oncologist with the other two oncologists in the training study. .... 26

Table 3: (Left) Represents the Spearman rank correlation of PPA.2 with each Oncologist. (Right) Represents the mean Spearman rank correlation of an oncologist with the five other oncologists. ........................................................................................................................ 27

Table 4: This table shows the rank lists for each oncologist as compared to the rank list calculated by PPA.2 ........................................................................................................................................ 29

Table 5: (Left) Correlation of the PPA.2 with each oncologist for the inter-patient study. (Right) Mean correlation of an oncologist with every other oncologist for the inter-patient study. ................................................................................................................................. 31

Table 6: (Left) Correlation coefficients between the PPA.2 and the oncologists for the inter-patient study after eliminating the two inconsistent data points. (Right) Mean Correlation coefficients between the oncologists and the other oncologists for the inter-patient study after eliminating the two inconsistent data points ........................................ 32
List of Figures

Figure 1: 3D rendering of a prostate and organs at risk (OAR). The rectum is posterior to the prostate which is a storage site for feces. The bladder is the storage site for urine before disposal. The femoral heads is the top portion of the femur. ............................. 1

Figure 2: Example of a match case’s fluence map overlaid on a query case’s planning target volume before deformation (left) and after deformation (right) ............................. 4

Figure 3: Observer preference study graphical user interface ......................................................... 9

Figure 4: These DVHs demonstrate the similarities between each plan for the intra-patient study analysis ................................................................. 10

Figure 5: Graphical representation of the scalar map 2D integration. The lighter the color of black it is, the higher the integer value at that point is ................................................. 12

Figure 6: Graphical representation of PPA.1’s weighting scheme for the rectum ......................... 14

Figure 7: Graphical representation of PPA.1’s weighting scheme for the bladder ................. 15

Figure 8: Graphical Representation of PPA.1’s area differences. The darker red represents the penalization score where the lighter red represents the area under the curve ................................................................. 16

Figure 9: Graphical representation of PPA.2’s weighting scheme for the rectum ................. 19

Figure 10: Graphical Representation of PPA.2’s weighting scheme for the bladder .......... 19

Figure 11: Graphical representation of PPA.2’s area differences to describe shoulder coverage .............................................................................................................. 20

Figure 12: Inter-patient study screen that allowed the oncologists to switch the order of graphs on the left slide bar and quickly toggle back and forth between patients using the arrow keys ................................................................. 23

Figure 13: The multiple slide view in PowerPoint allowed oncologists to compare multiple DVHs at once and reorder the DVHs at the same time ........................................ 24

Figure 14: The graph represents different DVHs with varying plan quality to represent the spectrum of quality that the oncologists were ranking for the inter-patient study ... 25
Figure 15: The plot represents the distribution of correlations as compared to the mean correlation for PPA.2. .......................................................... 28

Figure 16: A heat map representation of Table 8 to represent the trends of the oncologists’ rankings in the inter-patient study. .......................................................... 30
Acknowledgements

I want to thank Dr. Shiva Das and Dr. Joseph Lo for advising me on this project and allowing me to explore different ideas and options during this process which has made realize that I enjoy researching. The time spent researching under your tutelage will be impactful on my career and I greatly appreciate the opportunity.

I also want to thank Dr. Catherine Chang, Dr. Bridget Koontz, Dr. Jason Lee, Dr. Robert Lee, Dr. Manisha Palta, Dr. Brad Perez, Dr. Joseph Salama, and Dr. Kolby Sidhu for taking the time out of their busy schedule to participate in my study.

Thank you Dr. Robert Reiman for serving on my thesis defense committee. Many thanks to Dr. Kingshuk Roy Choudhury for answering all my statistics based questions.
1. Introduction

Prostate cancer forms in the glandular tissues of the prostate in the male reproductive system\(^1\). Most prostate cancers are characterized by its slow growth occurrence in older man causing difficulty or abnormalities in urination, erectile dysfunction, general pain in the pelvic and lower body area, or no symptoms at all\(^2,3\). According to the SEER Stat Fact Sheet from Cancer.gov, there was an estimated 233,000 new cases in 2014, making it the most prevalent form of diagnosed cancer accounting for 14\% of all new cancer cases with about 30,000 deaths from prostate cancer in that same year\(^4\). As those statistics are alarming, even more men live their daily lives unbeknownst to the prostate cancer which grows inside of them\(^5\).

Figure 1: 3D rendering of a prostate and organs at risk (OAR). The rectum is posterior to the prostate which is a storage site for feces. The bladder is the storage site for urine before disposal. The femoral heads is the top portion of the femur.
As a result of the staggering number of men affected by prostate cancer, additional research and improvements in treatment are sought after by the medical community. The current standard of care for prostate cancer is radiation therapy, surgery, chemotherapy, immunotherapy, and vaccine therapy\(^1\). Radiation therapy, whether it be external beam radiation therapy or interstitial radiation therapy, is used in stages I-IV of prostate cancer in combination with the previously stated methods of treatment and is used on almost half of all prostate patients\(^1,6\). Intensity modulated radiation therapy (IMRT) is the most common and optimal method of external beam radiation therapy for prostate cancer\(^7,8\).

External beam radiation therapy (EBRT) involves the use of high energy megavoltage photons delivered into the body from a linear accelerator. IMRT is an advanced form of EBRT that allows one to produce highly conformal dose distributions and increased dose to the planning target volume (PTV) to improve local control while also decreasing the dose to the surrounding normal tissues or organs at risk (OAR) leading to reductions in normal tissue complications\(^9\). IMRT achieves these characteristics by modulating the fluence, or number of photons per unit area, throughout the treatment with the use of a multi-leaf collimator (MLC) to paint the dose distribution on the PTV. For the prostate site, IMRT lowers patients’ risk of rectal bleeding, incontinence, or bowel irritation as compared to conventional 3D conformal radiation therapy\(^10,11,12\). However, IMRT requires more stringent immobilization
techniques due to the high conformity of the distribution and is a more labor intensive process.

IMRT plans are created using inverse planning. With inverse planning, the treatment planner inputs dose constraints for the OARs and the dose prescription for the PTV into the treatment planning system (TPS), resulting in the TPS using an optimizing algorithm to best satisfy the input parameters\(^{13}\). To best reach the desired optimized parameters, multiple iterations of the inverse planning are required which takes time and effort from the treatment planners. In addition, plan quality is dependent on treatment planner skill and the amount of time he or she has to complete the plan. This has led to a recent focus in knowledge-based radiation therapy (KBRT) to optimize the IMRT planning process through dose limiting optimization and decreasing the duration of the treatment planning process.

In previous work from our group, the goal was to create high-quality plans using a database of previously treated patients as the baseline of the knowledge base process. A “query” case, or a case to be planned, was paired with a similar “match” case, such that the dosimetric parameters from the match case could be applied to the query case. These cases were matched using 2D projections from the beam’s eye view of the 3D rendering of the patient anatomy. Mutual information comparisons yielded the match case with closest anatomical features. Since no two patient anatomies are completely identical, the match case’s PTV projections were deformed to the query case’s PTV
projections. These deformation transformations were applied to the fluence map of the match case to match of the query case as demonstrated in the following figure.

![Figure 2: Example of a match case’s fluence map overlaid on a query case’s planning target volume before deformation (left) and after deformation (right)\textsuperscript{14}.](image)

The deformed matched case provides dosimetric information that is entered into the IMRT optimization algorithm, thus providing a high quality starting point for treatment planning\textsuperscript{14}.

With the emergence of these varying KBRT techniques such as RapidPlan (Varian Medical Systems, Palo Alto CA)\textsuperscript{15}, our group\textsuperscript{14,16}, and other research groups\textsuperscript{17,18,19,20}, there needs to be a clinically relevant plan quality metric that allows for the comparison of competing equivocal plans. In the long term, KBRT has the potential to not only expedite the radiation therapy planning process, but also create plans of
similar or even better quality than current plans created by dosimetrists. It may also allow a clinic to tailor its treatment planning towards physician preference. It introduces another layer of quality control within the clinic, by limiting the delivery of sub-optimal plans and making planners more aware of the quality of plans prior to physician approval. In the current state of radiation therapy, plan quality metrics used to compare separate plans or determine quality are dose distributions in axial CT slices, dose-volume histograms (DVH), and biological models.

In dose distributions, the dose is distributed within the body in three dimensions impinging onto healthy organs and the target. Dose distributions allow one to see if there are any hotspots in a particular organ or if dose is spilling outside the target into an organ that is not desired. For two-dimensional evaluations, a cumulative dose-volume histogram allows one to see how much of a particular organ is receiving a certain amount of dose. Finally, in a one-dimensional sense, tumor control probability (TCP) and normal tissue complication probability (NTCP) describe how the dose impacts the target or OARs based on the biological factors and dose delivered to each structure. TCP and NTCP allows one to calculate the therapeutic ratio to maximize tumor kill while minimizing complications to the normal tissues.

While the previous mentioned metrics are useful and effective, the purpose of this research is to create a plan quality index based on dose-volume histograms that can simplify the gamut of plan quality metrics used in the clinic. Such a simplified metric
may be used to create a more robust quality control program. As Drzymala et al has stated in regards to DVHs, “because of the loss of positional information in the volume(s) under consideration, it should not be the sole criterion for plan evaluation.” Our group’s index is not an index to determine overall plan quality, but rather to describe the quality of a DVH which is a determining factor in overall plan quality. Previous research looked to quantify this dose-volume relationship with the use of biological relationships or volume overlaps but are still debated within radiation oncology.

Currently in the clinic, radiation oncologists drive plan quality and are the decision makers when it comes to which plan is used to treat a patient. Because of this clinic dynamic, physician preference was considered as the basis of dose-volume quantification rather than biological models. However, differences in opinions among the oncologists will arise as the determination of a “good” DVH is subjective. With sufficient data, it may be possible to customize a plan quality metric to suit preferences for any specific oncologist. The goal of this DVH plan quality index is to represent a “general” physician in radiation oncology and not any one particular physician.
2. Methods and Materials:

2.1 Study Development

To study the trends and preferences of different oncologists evaluating DVHs, an observer preference study was created based on previous observer preference studies of radiologist preference of image quality\textsuperscript{32,33}. Not only did we want to gather the qualitative data to shape our quality index algorithm, but also quantitative data to validate the quality index against oncologists’ preferences. To achieve the purpose of the study, an intra-patient training study, an intra-patient validation study, and an inter-patient validation study were performed to develop the algorithm, to validate the algorithm for scoring equivocal plans for the same patient, and to validate the scoring spectrum across all patients respectively.

Because of its relative simplicity in treatment planning, as well as the large clinical volume, prostate cases were chosen to be the site to develop the DVH quality index. In prostate cases, the OARs (rectum, bladder, femoral heads, and penile bulb) are consistent from patient to patient with well-defined constraints for each OAR.

To gauge specifically what oncologists’ preferences are when determining a DVH quality index, multiple plans of varying plan quality were developed in Eclipse (Varian Medical Systems, Palo Alto, CA). Since for a given treatment plan there could be infinite possibilities of the resultant DVH, our goal was to isolate changes in either a specific organ, dose region, relationships between two organs, or PTV shoulder coverage
on a case by case basis. This allowed our group to teach the algorithm to fit the data, drawing its basis from oncologists’ preferences. To create a robust sampling pool, 45 subjects were selected from patients previously treated at our institution. For each subject, four or five different treatment plans were developed, including the original plan used to treat the patient as well as alternative plans that were varied as mentioned above. All plans were simultaneously presented to the oncologist who would rank them in order of plan quality. To include a diverse array of IMRT techniques, postop, intact low risk, and intact intermediate risk treatments were included in the pool of 45 subjects.

To facilitate physician viewing and ranking of plans, a simple the graphical user interface (GUI) was developed to be visually similar to the clinical treatment planning system. As the oncologist began the observer preference study, the GUI that was presented is shown in the following figure:
Figure 3: Observer preference study graphical user interface.
As one can see, the observer was allowed to see all the DVHs at once, zoom in on any particular DVH, compare any number of DVHs, and select which structures to be displayed or not. For example, if one were wanted to compare all four graphs with every structure, he or she would select those structures and comparison boxes to produce the following figure:

![Figure 4: These DVHs demonstrate the similarities between each plan for the intra-patient study analysis.](image)

Once the oncologist ranked each of plans from 1 to 4, 1 being the best and 4 being the worst, a dialog box was presented that allowed the oncologist to provide free text describing the rationale for their preferences. This information was used to guide development of an automated plan quality index. At the end of the study, another
explanation was required that requested the oncologists to discuss the features of each OAR that he or she found pertinent in their decision making processes.

2.2 Training Intra-Patient Study

To gather initial data for the training of the algorithm, prostate specialist Dr. Robert Lee, and two senior residents, Dr. Jason Lee and Dr. Robert Perez, were chosen to perform the observer preference study. Each oncologist was presented with 20 patients, each with 5 corresponding treatment plans. Once both the qualitative and quantitative data were gathered from the observers, the first version of the physician preference algorithm (PPA.1) was developed. The qualitative data was first used to set the general weighting factors needed for the algorithm followed by adjustments to those same weighting factors to best fit the quantitative correlative values.

Since initial data from the training study produced favorable results, a more widespread observer preference study was initiated to validate PPA.1 and eventually lead to PPA.2.

2.3 Formulation of First Algorithm, PPA.1

Our plan quality index is based upon DVH curves, which summarize the 3D dose distribution in a manner that is simple but effective. Within our group, a simple area under the curve integration had been used to model OAR dose volume relationships in the past\textsuperscript{35}. However, many different dose-volume curves could provide similar or equal areas under curve despite the possibilities of being diverse in their
clinical implications. In a DVH, ideally one would want 100% of an OAR receiving no
dose. As a higher amount of the volume is irradiated by a higher amount of dose, the
more at risk the organ is to radiation-induced side effects.

Because of this, we modified the DVH by an x*y scalar map such that the regions
with a higher percentage of volume receiving a higher amount of dose will produce a
larger integral sum than a simple area under the curve model. For our plan quality
index, such a scalar map would preferentially increase the contribution of DVH curve
regions with higher dose (x) or volume (y).

Figure 5: Graphical representation of the scalar map 2D integration. The
lighter the color of black it is, the higher the integer value at that point is.
Based upon the data gathered from the oncologists, as long as the femoral heads and penile bulb were in proximity or below their constraints, those OARs were not factored into the evaluation of plan quality.

The rectum and bladder were the two OARs that the oncologists considered and were two of three components that led to the determination of the DVH quality index. For each OAR, portions of the DVH curve were scalar mapped and integrated, then those areas were linearly weighted to create a “penalty” metric for that organ. The more dose there was to higher dose or volume regions, the greater was the penalty. The dose regions were defined in the following manner. Along each dose-volume curve, the dose region below 40Gy was rarely considered by oncologists, whereas the dose region above 60Gy, the highest dose region, was given the highest consideration when determining plan quality. These dose region considerations led to separate weighting factors for the dose regions between 40-60Gy and greater than 60Gy. Moreover, oncologists indicated higher priority to lowering the dose to the rectum than to the bladder. We therefore used a larger weighting factor for the highest dose regions of the rectum than the bladder, leading to the following equations:

\[
P_{\text{rect}} = \left[ w_{r_{40Gy-60Gy}} \int_{40Gy}^{60Gy} f(x) x \ast y \, dy \, dx + w_{r_{60Gy-\text{Max}}} \int_{60Gy}^{\text{Max}} f(x) x \ast y \, dy \, dx \right]
\]
Equation 2

\[ P_{\text{blad}} = \left[ w_{b_{40Gy-60Gy}} \int_{40Gy}^{60Gy} g(x) x \, dy \, dx + w_{b_{60Gy}} \int_{60Gy}^{\text{Max}} g(x) x \, dy \, dx \right] \]

The weighting factors, \( w \), determined for each feature were initially estimated based on the oncologists’ qualitative preferences, then adjusted to provide the best correlation to their rankings. \( f(x) \) and \( g(x) \) represent the dose-volume curve for the rectum and bladder respectively. Finally, the \( P \) represents the penalty associated with each organ that takes away from the plan quality index. These equations can be represented in the following two figures:

Figure 6: Graphical representation of PPA.1’s weighting scheme for the rectum.
The perfect DVH plan quality index was arbitrarily picked to be a score of 10. The penalty factors for each OAR were scaled to single digit values and subtracted away from 10. When performing the scalar map integration, the resultant values would be on the scale of $1E09$, resulting in that integral value being divided by $1E09$ to be relative to the number 10. In this regard, as the value of the integration becomes higher, the more detrimental the dose is to that organ which results in a lower DVH quality index.

For the PTV, shoulder coverage penalization took on a simpler form than the OARs. Weaker shoulder coverage was much less preferred than hotter shoulder coverage, so this was characterized without a scalar map but rather by simple differences in areas.
Figure 8: Graphical Representation of PPA.1’s area differences. The darker red represents the penalization score where the lighter red represents the area under the curve.

As shown in the figure, we considered only the region of the PTV (yellow curve) that was bounded within the box defined by dose of $D_{<100\%}$ to $D_{<90\%}$ and the volume from 0 to 100%. Within this regions, the area under the PTV, (pink), was subtracted away from the box, resulting in the volume shown in red. This difference (representing missed PTV coverage) was scaled down to single digit values and had an overall weighting factor applied to the equation:

Equation 3

$$P_{PTV} = w_{PTV} \left[ 100 \ast (D_{90\%} - D_{<100\%}) - \int_{D_{<100\%}}^{D_{90\%}} h(x)dx \right]$$

The $P_{PTV}$ is also subtracted away from the perfect score of 10. For the PTV penalization, the differences in areas would result in values on the scale of 1000 which would then be divided by 1000 to be relative to 10. The three penalty values all contribute to the give and take relationship that is associated with the OARs and PTV
coverage when trying to determine plan quality. The plan quality index was ready for testing on a larger group of cases and physicians, which is described in the following section.

2.4 Intra-Patient Validation Study

This validation study involved 6 radiation oncologists within the Duke network: Dr. Catherine Chang, Dr. Bridget Koontz, Dr. Robert Lee, Dr. Manisha Palta, Dr. Joseph Salama, and Dr. Kolby Sidhu. The intention was to not only have prostate specialists, but to gather information from a larger and more diverse background of radiation oncologists from multiple hospitals in Durham and Raleigh, NC. The hospitals included Duke University Hospital, Duke Regional Hospital, Durham V.A. Medical Center, and Duke Raleigh Hospital. For this study, 25 previously treated patients were chosen each with 3 alternative and 1 original treatment plans. Therefore, each oncologist interpreted 100 plans, yielding 600 total data points. The study went from 5 treatment plans at a time to 4 treatment plans at a time to allow for more contrasting differences between each treatment plan since some of the treatment plans were too similar when 5 treatment plans were used.

When the observer preference study was performed and all the data was collected from each oncologist, PPA.1 was validated and analyzed for its ability to effectively represent a general oncologist within the Duke network. Since the validation study collected a wealth of information regarding physician preference, PPA.1 was
reevaluated and modified to produce a more robust metric designated as PPA.2 which will be discussed further.

2.5 Formulation of Second Algorithm, PPA.2

PPA.2 was formulated after the completion of the validation intra-patient study. Its goal was to be more representative of the physicians’ preference than PPA.1 by increasing the correlation coefficients between the plan quality index and the physicians. During the validation study, it was noted that the oncologists were also considering general overall OAR weightings along with specific dose regions, i.e. physicians occasionally would preferentially select rectum sparing over bladder sparing with little consideration for certain dose regions. For that reason, PPA.2 was modified to include weighting factors that were applied to the penalty factor as a whole for each structure to account for the overall OAR curves resulting in the following equations.

Equation 4

\[ P_{\text{rect}} = w_r \left[ w_{R_{40 Gy-60 Gy}} \int_{40 Gy}^{60 Gy} f(x) x \cdot y \, dx + w_{R_{60 Gy-Max}} \int_{60 Gy}^{Max} f(x) x \cdot y \, dx \right] \]

Equation 5

\[ P_{\text{blad}} = w_b \left[ w_{B_{40 Gy-60 Gy}} \int_{40 Gy}^{60 Gy} g(x) x \cdot y \, dx + w_{B_{60 Gy-Max}} \int_{60 Gy}^{Max} g(x) x \cdot y \, dx \right] \]

The study proved the tendency of the oncologists to weight the rectum higher than the bladder causing the weighting factor on the rectum to be higher than that of the bladder resulting in the following graphical representations.
Figure 9: Graphical representation of PPA.2’s weighting scheme for the rectum.

Figure 10: Graphical Representation of PPA.2’s weighting scheme for the bladder.
The previous PPA.1 algorithm lacked information of the prescribed dose to be delivered to the target, which is necessary to describe the tumor coverage. Thus, the prescribed dose became a key factor in developing the improved PPA.2 algorithm as represented in the following figure:

![Graphical representation of PPA.2’s area differences to describe shoulder coverage.](image)

To do this, the area under the PTV curve from 5Gy less than the prescribed dose to $D_{95\%}$ was calculated. That area was subtracted away from a rectangular area between 5Gy less than the prescribed dose to 105% of the prescribed dose. Thus if the shoulder coverage was weaker, the rectangular area set by prescribed dose would result in a higher penalization compared to a sharper PTV shoulder. The difference was normalized by the rectangular area with a weighting factor applied as shown in the following equation.
The $P_{PTV}$ is again subtracted away from the perfect score of 10 where the normalization factor reduces $P_{PTV}$ to be on the scale of the scoring spectrum. PPA.2 relies on these equations to produce a DVH quality index that are used to rank plans, allowing for Spearman’s rank correlation analysis between the PPA.2 rank order and the oncologists’ rank order. The DVH information was exported from Eclipse and then processed within MATLAB (Mathworks Inc., Natick MA).

2.6 Inter-patient Study

For any one patient plan in the previous intra-patient study, the alternative plans were intentionally chosen to be equivocal and thus challenging to physicians. As a result, the plan quality indexes did not vary over the overall spectrum of scores. The intra-patient study only validated that the ranking algorithm’s ability to differentiate amongst similar plans and not the trend of scoring across multiple patients who would have more substantial differences in quality.

The inter-patient study was then introduced to validate the algorithms ability to rank the plans across the whole scoring spectrum. To create this study, 15 patients were chosen to represent the broad spectrum of scores. It was important that these 15 patients all received the same intermediate risk treatment with a prescription dose of 7600cGy.
However, two of the patients had a dose of 7560cGy which was deemed acceptable to be used in the study. The following scores were used to validate the PPA.2 scoring spectrum:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Score</th>
<th>Patient</th>
<th>Score</th>
<th>Patient</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient A</td>
<td>7.797</td>
<td>Patient F</td>
<td>6.573</td>
<td>Patient K</td>
<td>4.792</td>
</tr>
<tr>
<td>Patient B</td>
<td>7.512</td>
<td>Patient G</td>
<td>6.126</td>
<td>Patient L</td>
<td>3.865</td>
</tr>
<tr>
<td>Patient C</td>
<td>7.263</td>
<td>Patient H</td>
<td>5.706</td>
<td>Patient M</td>
<td>3.730</td>
</tr>
<tr>
<td>Patient D</td>
<td>7.180</td>
<td>Patient I</td>
<td>5.553</td>
<td>Patient N</td>
<td>2.620</td>
</tr>
<tr>
<td>Patient E</td>
<td>6.885</td>
<td>Patient J</td>
<td>5.315</td>
<td>Patient O</td>
<td>2.094</td>
</tr>
</tbody>
</table>

Oncologists were presented with all 15 cases simultaneously and asked to rank them in order. To provide a simple but familiar interface, PowerPoint (Microsoft Corporation, Redmond WA) was used. Each case was presented on a separate slide with the same OAR color scheme, and all plans were positioned and scaled to overlap exactly. Oncologists were able to quickly toggle through plans, rank them in the order he or she deemed correct by reordering the slides, and display the graphs simultaneously. The following figure represents a certain technique that was used by the oncologists to sort the plans.
Figure 12: Inter-patient study screen that allowed the oncologists to switch the order of graphs on the left slide bar and quickly toggle back and forth between patients using the arrow keys.

As one can see from the arrows, the oncologist could drag and drop the graphs in the order desired after toggling between each graph which would eventually result in a sorted list.

Additionally, the oncologist used the PowerPoint function where all the slides can be displayed congruently in the same window as shown in the following figure:
Figure 13: The multiple slide view in PowerPoint allowed oncologists to compare multiple DVHs at once and reorder the DVHs at the same time.

With many graphs being displayed at once, it allowed the oncologists to see the spectrum of quality among many plans at once leading to more informed decisions. The oncologists used a mix of both techniques to rank the plans. Once the plans were set in the desired order, the ranks were compared against PPA.2 using a Spearman’s rank correlation test. Figure 15 demonstrates the differences in scores using PPA.2.
Figure 14: The graph represents different DVHs with varying plan quality to represent the spectrum of quality that the oncologists were ranking for the inter-patient study.

Each plan is normalized to 100% of the dose covering 95% of the volume. The dose prescribed for these three patients was 7600cGy. Notice the differences in PTV coverage and OAR sparing, resulting in a give and take relationship between OAR sparing and PTV coverage.

With the intra-patient and inter-patient study now explained, one can look at the results of the studies for further analysis and discussion.
3. Results

3.1 Training Intra-patient Study

The training intra-patient study allowed our group to develop an algorithm and test the ease of our observer preference that could be presented to oncologists within the Duke Network.

This data was the basis of the development of our algorithm. The weighting factors and shoulder coverage calculation went through many iterations until we arrived at the optimized PPA.1. The algorithm’s rank order list was compared with the oncologists’ rank list using the Spearman’s rank correlation coefficient. The mean correlation coefficient was calculated among all patients. The following tables represent correlations between PPA.1 and each physician, as well as each physician against the other two physicians:

Table 2: (Left) Represents the mean Spearman rank correlation of PPA.1 with each of the oncologists in the training study. (Right) Represents the mean Spearman rank correlation of each oncologist with the other two oncologists in the training study.

<table>
<thead>
<tr>
<th>Oncologist Correlation with PPA.1</th>
<th>Oncologist Correlation with Oncologists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician A</td>
<td>0.470</td>
</tr>
<tr>
<td>Physician B</td>
<td>0.345</td>
</tr>
<tr>
<td>Physician C</td>
<td>0.706</td>
</tr>
<tr>
<td>Physician A</td>
<td>0.328</td>
</tr>
<tr>
<td>Physician B</td>
<td>0.242</td>
</tr>
<tr>
<td>Physician C</td>
<td>0.334</td>
</tr>
</tbody>
</table>

The mean correlation for PPA.1 correlation with the oncologists is 0.507 and the mean correlation for the oncologists against each other is 0.301. Despite the mean correlations
being fairly weak, the correlations associated with PPA.1 were still higher than the mean correlation of the oncologists.

### 3.2 Validation Intra-patient Study

For each patient, the rank list from both PPA.2 and the oncologist were compared with the mean Spearman rank correlation test across all cases. The correlation was calculated between PPA.2 and each physician, and the mean of each physician against all other physicians. That data is shown in the following tables.

<table>
<thead>
<tr>
<th>Oncologist Correlation with PPA.2</th>
<th>Oncologist Correlation with Oncologists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician A</td>
<td>0.920</td>
</tr>
<tr>
<td>Physician B</td>
<td>0.540</td>
</tr>
<tr>
<td>Physician C</td>
<td>0.936</td>
</tr>
<tr>
<td>Physician D</td>
<td>0.712</td>
</tr>
<tr>
<td>Physician E</td>
<td>0.832</td>
</tr>
<tr>
<td>Physician F</td>
<td>0.416</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician A</td>
<td>0.689</td>
</tr>
<tr>
<td>Physician B</td>
<td>0.470</td>
</tr>
<tr>
<td>Physician C</td>
<td>0.677</td>
</tr>
<tr>
<td>Physician D</td>
<td>0.551</td>
</tr>
<tr>
<td>Physician E</td>
<td>0.601</td>
</tr>
<tr>
<td>Physician F</td>
<td>0.398</td>
</tr>
</tbody>
</table>

The mean correlation between PPA.2 and the physicians was 0.726, which was higher than the mean correlation between each physician and all other physicians of 0.564.

Another representation of this data is shown in the following figure:
Figure 15: The plot represents the distribution of correlations as compared to the mean correlation for PPA.2.

3.2.1 Analysis of Intra-Patient Results

The inherent complexities and factors in an oncologist’s rank order that are not explicitly stated or measurable within the observer preference study lead to the correlation coefficients being lower than what is conventionally expected. When looking at Figure 4, an observer can notice the varying shoulder coverages that could influence the oncologist’s opinion on target coverage. Additionally, the rectum and bladder have differing weighted integrals. Graph B has the lowest integral dose for the bladder, however, it also highest integral dose for the rectum in the 40Gy-60Gy range and the worst shoulder coverage. With Graph C, it has the lowest integral dose in the rectum but the highest integral dose for the bladder. From this example, one can fully see the complexities and varying features within the DVH that can influence an oncologist’s
decision making. For this reason, inconsistencies from patient to patient can arise when ranking DVH quality since many factors are considered and difficult to isolate since evaluating a DVH is a give-and-take process.

**3.3 Inter-patient Study**

Once the inter-patient study was completed, our data contained rank lists of the numbers 1 through 15 for each physician. The following results of each rank list are shown in the table below:

**Table 4: This table shows the rank lists for each oncologist as compared to the rank list calculated by PPA.2.**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient A</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Patient B</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Patient C</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>13</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Patient D</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Patient E</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Patient F</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>11</td>
<td>4</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Patient G</td>
<td>7</td>
<td>8</td>
<td>12</td>
<td>5</td>
<td>11</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Patient H</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Patient I</td>
<td>9</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Patient J</td>
<td>10</td>
<td>6</td>
<td>9</td>
<td>14</td>
<td>7</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Patient K</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>9</td>
<td>3</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>Patient L</td>
<td>12</td>
<td>13</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>Patient M</td>
<td>13</td>
<td>12</td>
<td>13</td>
<td>10</td>
<td>13</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>Patient N</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>12</td>
<td>15</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Patient O</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>15</td>
<td>14</td>
<td>15</td>
<td>13</td>
</tr>
</tbody>
</table>

When looking at the above table, it may be difficult to discern the trend of each rank list compared to PPA.2 or with an oncologist. For this reason, a heat map was produced of the above table to visually show the trend of the separate rank lists.
Figure 16: A heat map representation of Table 8 to represent the trends of the oncologists’ rankings in the inter-patient study.

Red represents the best ranked plan and white represents the worst ranked plan.

Looking at the above heat map, one can notice the more noticeable outliers for physician C at position 3 and the two separate outliers for physician D and physician F at position 11. Other outliers are present in the heat map but not as impactful as the outliers stated above.

The Spearman rank correlations are shown in the tables below for correlation of PPA.2 vs. each physician, and each physician vs. other physicians.
Table 5: (Left) Correlation of the PPA.2 with each oncologist for the inter-patient study. (Right) Mean correlation of an oncologist with every other oncologist for the inter-patient study.

<table>
<thead>
<tr>
<th>Oncologist Correlation with PPA.2</th>
<th>Oncologist Correlation with Oncologists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician A</td>
<td>0.930</td>
</tr>
<tr>
<td>Physician B</td>
<td>0.850</td>
</tr>
<tr>
<td>Physician C</td>
<td>0.657</td>
</tr>
<tr>
<td>Physician D</td>
<td>0.761</td>
</tr>
<tr>
<td>Physician E</td>
<td>0.964</td>
</tr>
<tr>
<td>Physician F</td>
<td>0.768</td>
</tr>
<tr>
<td>Physician A</td>
<td>0.718</td>
</tr>
<tr>
<td>Physician B</td>
<td>0.759</td>
</tr>
<tr>
<td>Physician C</td>
<td>0.570</td>
</tr>
<tr>
<td>Physician D</td>
<td>0.645</td>
</tr>
<tr>
<td>Physician E</td>
<td>0.784</td>
</tr>
<tr>
<td>Physician F</td>
<td>0.715</td>
</tr>
</tbody>
</table>

As shown previously in the intra-patient study, the PPA.2 used in the inter-patient is also able to have a higher mean correlation with the oncologists (0.821) than the mean correlation of the oncologists with each other (0.690).

3.3.1 Analysis of Inter-Patient Results

When comparing Table 4, the intra-patient results, and Table 9, the inter-patient results, one can see that Physicians A and E have similar correlation coefficients. Physicians B and F are significantly more correlated with the PPA.2’s rank order for the inter-patient study as compared to the intra-patient study. To explain this, consider Figure 14 which represents three separate patients’ DVH plots. The scores associated with each graph in Figure 14 is more intuitive as compared to Figure 4. Because of this observation, one can predict that oncologists ranking DVHs across separate patients is more correlative to the PPA.2 rankings.

To explain the significant drop in correlation for Physician C from the intra-patient study to the inter-patient study, one should analyze Table 8 and Figure 16. For
Physician C’s rank list, PPA.2’s 3rd ranked plan (Patient C) is Physician C’s 13th worst ranked plan whereas every other physician has ranks of 3rd, 4th, 2nd, 5th and 7th. Similarly, for Patient K (PPA.2’s 11th ranked plan), Physicians D and F ranked that patient the 3rd best while the other physicians ranked it 11th, 10th, 9th, and 12th. If one were to throw out Patient C and Patient K due to some possible error in the ranking by the oncologists, a new set of correlation coefficients can be generated which produced higher correlation coefficients which only makes the data stronger and reveals the physicians to be more consistent against the PPA.2.

Table 6: (Left) Correlation coefficients between the PPA.2 and the oncologists for the inter-patient study after eliminating the two inconsistent data points. (Right) Mean Correlation coefficients between the oncologists and the other oncologists for the inter-patient study after eliminating the two inconsistent data points

<table>
<thead>
<tr>
<th>Oncologist Correlation with PPA.2</th>
<th>Oncologist Correlation with Oncologists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician A</td>
<td>0.912</td>
</tr>
<tr>
<td>Physician B</td>
<td>0.846</td>
</tr>
<tr>
<td>Physician C</td>
<td>0.863</td>
</tr>
<tr>
<td>Physician D</td>
<td>0.852</td>
</tr>
<tr>
<td>Physician E</td>
<td>0.951</td>
</tr>
<tr>
<td>Physician F</td>
<td>0.896</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician A</td>
<td>0.782</td>
</tr>
<tr>
<td>Physician B</td>
<td>0.811</td>
</tr>
<tr>
<td>Physician C</td>
<td>0.727</td>
</tr>
<tr>
<td>Physician D</td>
<td>0.773</td>
</tr>
<tr>
<td>Physician E</td>
<td>0.843</td>
</tr>
<tr>
<td>Physician F</td>
<td>0.809</td>
</tr>
</tbody>
</table>

The mean correlation coefficients, after correcting for the two outlying data points, are 0.887 for the PPA.2 with the oncologists and 0.791 for the oncologists with each other which are higher than previous mean correlation coefficients. Additionally, the correlation coefficient for Physician C after removing the outlier is 27.1% higher as compared to the full data set. The corrected correlation coefficient for the inter-patient is
also closer in proximity to the intra-patient mean correlation coefficient, providing a more reasonable inter-observer preference consistency from Physician C. What is important to note is that our DVH quality index is converging to a higher mean correlation coefficient from the training to validation study. If we are able to raise our mean correlation with the physicians then the PPA.2 is becoming more representative of a general oncologist within the Duke Network.
4. Discussion

The purpose of the study was to create a prostate IMRT plan quality index based on physician preference to quantitatively compare competing treatment plans. Thus allowing planners to choose between plans with equivocal characteristics, prior to physician scrutiny. To achieve this purpose, the goal was to present an observer preference study to formulate and validate a plan quality index. The correlation between the plan quality index and each physician needed to be similar to the physicians’ correlation with each other, thus demonstrating that our algorithm can reflect typical physician preference for plan quality. However, the ability to accurately quantify physician preference for DVHs was in question as it had not been explored in literature. Regarding the varying opinions of physicians of what makes a “good” DVH, some physicians have described this as an “art form.” There was also a concern in the observer consistency or reliability as there are many different factors that are taken into consideration when determining plan quality that extend past DVHs such as dose distributions or structure contouring.

PPA.1 is the first generation of the plan quality index algorithm that was derived from the training intra-patient study. The mean correlations for PPA.1 in the training intra-patient study are higher than the correlations between the oncologists which suggests that an observer preference study can be used to quantify physician preference. It can also demonstrate that PPA.1 can be representative of physician preference but the
lower correlation values raised questions of the algorithm's ability to accurately represent physician preference. However, the validation intra-patient study demonstrated that the mean correlation coefficient is higher between PPA.2 and the physicians than the physicians with each other. Since the correlation coefficients were higher than the training study, we can claim that PPA.2 can more accurately represent physician preference for prostate IMRT plans.

The inter-patient studies also proved to have stronger correlations of PPA.2 with the oncologists than the oncologists with each other. This study not only provided validation of the scoring spectrum calculated by that of PPA.2, it also demonstrated observer reliability and consistency from the intra-patient study to the inter-patient study. Since the PPA.2 was unchanged from the intra-patient study to the inter-patient study, one could analyze how consistent an oncologist was when ranking plans against the algorithm's rankings. Finally, the mean correlation coefficient between the PPA.2 and the oncologists is of a higher correlation which proves that the inter-patient study provides a very strong correlation with all of the oncologists when ranking plans across a wide range of patients.

The trend of the automated plan quality index correlating better to physicians than the physicians with the other physicians may seem surprising. To explain this further, if there were spectrum of physician preference, PPA.2 has the ability to be somewhere in the middle of that spectrum whereas certain oncologists are on opposite
sides of the spectrum with highly contrasting preferences. By focusing only on a few simple, quantitative components, the plan quality index is likely to be more robust and consistent than human experts.

The DVH plan quality index based on physician preference provides a quality index based on the individuals who are the final decision makers on whether or not a treatment is to be delivered. This tool integrates an oncologist’s opinion on a treatment plan throughout its optimization instead of at the completion of optimization. Because of that quality control, potential re-optimizations of a treatment plan could be avoided, decreasing patient planning time to allow for more patients to be treated within a clinic. Since this type of plan quality index has not been considered in literature, the results are untested and without comparison to other studies. The expectations of this study were exceeded because correlations consistently higher than the physicians was not expected, supporting the need for more of these types of studies to be performed.

This study had several limitations. Discrepancies in plan quality rankings will arise from oncologist to oncologists and even with one oncologist over time, and these may add uncertainty to their ranking data. In addition, such simple quantitative metrics cannot integrate all of the qualitative data that comes from physicians’ decisions; in fact much of that data goes unsaid. Additionally, physicians are not accustomed to viewing more than two plans at once or ranking multiple plans at a time, which may have
resulted in training bias during the time it takes for the physician to adjust to the observer study. Additional training may help to improve their consistency and accuracy.

Although, this study already involved up to eight physicians with different levels of expertise from many hospitals, all physicians were drawn from the Duke University Health System network. The study could benefit from more physicians from different networks participating in the study with a larger patient pool to add more variability in the quality of plans analyzed.

4.1 Future Work

The next step in plan quality index development is to test other weighting factors to bring the algorithm more towards the central point of the DVH quality spectrum that the oncologists have dictated in his or her decision making process. To do this, one could analyze the data on a case by case basis to find the percentages of when each structure and each dose region was considered to develop percentages for the weighting factors. Another possible method to pursue the convergence of the DVH quality index spectrum is to fit the equations used in the algorithm to the correlation data. Instead of using an inferential-based approach to creating the algorithm, one can use numerical or statistical approaches to fit the equations to the best possible correlation results. However, since the oncologists’ decision making process is inferential at its core, the use of a data fitting tool could prove to be no better or worse than the inferential fitting
based approach due to the inherent marginalized random nature and complexity of the data that was gathered in this work.

The plan quality index algorithm can also benefit in its robustness if volume overlap normalization is pursued. Since the contouring of structures and patient anatomy can greatly impact the outcomes of DVHs, it would be useful to account for these situations. To perform this volume overlap normalization, one would look at the volume contours in Eclipse and find the percentage of the OARs overlapped by the PTV to create a normalization factor. Thus, if a patient were to have large PTV OAR overlaps, the score from the algorithm can be increased as to not penalize the DVH quality index due a non-optimal patient anatomy. This addition to the DVH quality algorithm would shift the algorithm in the direction of patient-specific medicine which has become a major push in the field of medicine. Finally, observer preference studies may be performed at separate hospitals to use the DVH plan quality index to compare and improve quality across multiple sites.
5. Conclusion

The goal of our research was to create a DVH quality index that would be useful as a comparison tool between separate DVHs and provide a multi-functional plan quality tool to improve patient treatment. Our plan quality metric has proven to be a viable tool that correlates well with clinical DVH plan quality. In fact, the algorithm consistently correlated better to oncologists than they did with each other. For this reason, we can state that our algorithm quantitatively scores and ranks DVH quality in a way that is representative of a general expert oncologist here in the Duke Network. Because the algorithm can at least in part represent the opinions of an oncologist, it can be used as a comparison tool for plan quality and as a clinical workflow tool in dosimetry workrooms. This research can provide useful and exciting advances in knowledge-based radiation therapy and patient specific plan quality.
References


