Louisiana State University
LSU Digital Commons

LSU Doctoral Dissertations | Graduate School

May 2020

Towards Optimizing Quality Assurance Outcomes of Knowledge-Based Radiation Therapy Treatment Plans Using Machine Learning

Phillip Douglas Hardenbergh Wall
Louisiana State University and Agricultural and Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

 Part of the Health and Medical Physics Commons

Recommended Citation

Wall, Phillip Douglas Hardenbergh, "Towards Optimizing Quality Assurance Outcomes of Knowledge-Based Radiation Therapy Treatment Plans Using Machine Learning" (2020). *LSU Doctoral Dissertations*. 5266.
https://digitalcommons.lsu.edu/gradschool_dissertations/5266

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contactgradetd@lsu.edu.

TOWARDS OPTIMIZING QUALITY ASSURANCE OUTCOMES OF KNOWLEDGE-BASED RADIATION THERAPY TREATMENT PLANS USING MACHINE LEARNING

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Physics and Astronomy

by
Phillip Douglas Hardenbergh Wall
B.S., Davidson College, 2014
M.S., Louisiana State University, 2017
August 2020

To my mother and father, who have given everything for me

ACKNOWLEDGMENTS

I first acknowledge my dissertation advisor, Dr. Jonas Fontenot, for his contributions to this work. I thank him for his guidance and expertise that he provided over the course of this project. I thank him for his scientific, professional, and personal mentorship and the opportunities he as afforded me over the course of our collaboration. I feel very fortunate to have undergone my graduate training under his direction. I also acknowledge my committee members Drs. Jianhua Chen, Juhan Frank, Wayne Newhauser, and Justin Sick for their time and effort in monitoring and steering the progression of this project. I thank Dr. Levent Dirikolu for serving as the Dean's Representative on my committee and Dr. Mark Wilde for his participation in the supervision of this project as well.

I acknowledge dosimetrists Frank Apollo, Eddie Singleton, Chad Dunn, and Hamlet Spears for sharing their insight and expertise regarding radiation therapy treatment planning. I acknowledge physicists Connel Chu, David Perrin, and Dan Neck for lending their knowledge of treatment planning systems, patient data management, and relevant clinical software, which was instrumental for collecting and anonymizing previous patient data needed for this project. I also acknowledge physics residents Desmond Fernandez, Brittany Moore, Addie Barron, Chris Schneider, and John Doiron for sacrificing their time to advise and instruct me in performing the quality assurance measurements and analysis for this study. I acknowledge Dr. David Solis for his interest and lending his machine learning expertise in support of this project.

I acknowledge RaySearch Laboratories Support for prompt responses to inquiries and for granting numerous requests for temporary research licenses to enable

DICOM export capabilities needed for this project. Specifically, I acknowledge Freddie Cardel, Khai Le, Khiem Le, Sam Painter, and Anthony Fong. I acknowledge Jason Stephens, director of IT, for his support and communication in facilitating the installation of said licenses. I acknowledge Adam Watts for insightful discussions regarding beam modeling and quality assurance within RayStation.

I acknowledge and thank Dr. Kip Matthews for his invaluable and selfless guidance, teachings, and personal mentorship over the course of this project and during my graduate studies. I also acknowledge Susan Hammond, Yao Zeng, Katelynn Fontenot, Katherine Pevey, and Megan Jarrell for their administrative and logistical support over the course of this project.

I acknowledge my peers and colleagues within the LSU Medical Physics & Health Physics Graduate Program for their scientific and moral support over the course of this project. Specifically, I thank Cameron Sprowls, Joe Steiner, Stephanie Wang, Elizabeth Hilliard, Krystal Kirby, Will Donahue, Lydia Wilson, Yibo Xie, Payton Bruckmeier, Andrew McGuffey, and Audrey Copeland.

I acknowledge and thank my close friends and family – namely my mother, brother, and grandfather – for their unwavering and unending love and support throughout this work and my life.

This work was supported in part by a grant through the Mary Bird Perkins Cancer Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

%DD	percent dose-difference
3DCRT	three-dimensional conformal radiation therapy
AA	aperture area
ANN	artificial neural network
AP	aperture perimeter
CART	classification and regression tree
CAS	cross-axis score
CI	conformity index
CLS	closed leaf score
COV	coefficient of variation
CP	control point
DICOM	Digital Imaging and Communications in Medicine
DTA	distance-to-agreement
DVH	dose volume histogram
EBRT	external beam radiation therapy
EM	edge metric
EQD_2	(biologically) equivalent physical dose in 2 Gy fractions
EUD	equivalent uniform dose
FAOC	fractional area outside of circle
FFF	flattening filter free
GPR	gamma passing rate
HI	homogeneity index

IMRT	intensity modulated radiation therapy
JM	jaw motion
JP	jaw position
JT	jaw travel
KBP	knowledge-based planning
LG	leaf gap
LINAC	linear accelerator
LM	leaf motion
LT	leaf travel
MAE	mean absolute error
MCS	modulation complexity score
MLC	multi-leaf collimator
MSE	mean squared error
MU	monitor unit
NTCP	normal tissue complication probability
OAR	organ at risk
OVH	overlap volume histogram
PI	plan irregularity
PM	plan modulation
PTV	planning target volume
QA	quality assurance
RBF	radial basis function
SAS	small aperture score

SVM	support vector machine
TCD_{50}	tumor control dose 50%
TCP	tumor control probability
TD_{50}	tolerance dose 50%
TG	task group
TPS	treatment planning system
VMAT	volumetric modulated arc therapy

ABSTRACT

Knowledge-based planning (KBP) techniques have been shown to provide improvements in plan quality, consistency, and efficiency for advanced radiation therapies such as volumetric modulated arc therapy (VMAT). While the potential clinical benefits of KBP methods are generally well known, comparatively less is understood regarding the impact of using these systems on resulting plan complexity and pre-treatment quality assurance (QA) measurements, especially for in-house KBP systems. Therefore, the overarching purpose of this work was to assess QA implications with using an in-house KBP system and explore data-driven methods for mitigating increased plan complexity and QA error rates without compromising dosimetric plan quality. Specifically, this study evaluated differences in dose, complexity, and QA outcomes between reference clinical plans and plans designed with a previously established in-house KBP system. Further, a machine learning model – trained and tested using a database of 500 previous VMAT treatment plans and QA measurements – was developed to predict VMAT QA measurements based on selected mechanical features of the plan. This model was deployed as a feedback mechanism within a heuristic optimization algorithm designed to modify plan parameters (identified by the machine learning model as important for accurately predicting QA outcomes) towards improving the predicted delivery accuracy of the plan. While KBP plans achieved average reductions of 6.4 Gy ($p < 0.001$) and 8.2 Gy ($p < 0.001$) in mean bladder and rectum dose compared to reference clinical plans across thirty-one prostate patients, significant ($p < 0.05$) increases in both complexity and QA measurement errors were observed. A support vector machine (SVM) was developed – using a database of 500

previous VMAT plans – to predict gamma passing rates (GPRs; 3%/3mm percent dose-difference/distance-to-agreement with local normalization) based on selected complexity features. A QA-based optimization algorithm was devised by utilizing the SVM model to iteratively modify mechanical treatment features most commonly associated with suboptimal GPRs. The feasibility was evaluated on 13 prostate VMAT plans designed with an in-house KBP method. Using a maximum random leaf gap displacement setting of 3 mm, predicted GPRs increased by an average of 1.14 ± 1.25% (p = 0.006) with minimal differences in dose and radiobiological metrics.

1. INTRODUCTION

1.1. BACKGROUND

1.1.1. Radiation Therapy Treatment Delivery and Planning

Radiation therapy or radiotherapy involves the treatment of disease with the use of high-energy radiation, which can take different forms such as x-rays, gamma rays, electrons, and protons. The primary disease treated with radiation therapy is cancer, with over half of all cancer patients receiving radiation therapy during the course of their care.[1] Typically, a cancer treatment team – comprising of physicians, physicists, dosimetrists, therapists, and other healthcare professionals – works to plan, simulate, test, and deliver a course of radiation therapy with the goal of simultaneously maximizing cancer cell death and minimizing cell damage in surrounding healthy tissues. This is an intricate process, marked by a plethora of treatment variables and parameters that the treatment team must define in order to provide the patient with high-quality radiotherapy.

One of the first choices that must be made is how to deliver the prescription dose of radiation to the targeted disease. For photon radiotherapy, this dose can be delivered either from outside of the patient or from directly within the patient. These two treatment modalities are called external beam radiotherapy (EBRT) and brachytherapy. While brachytherapy can present dosimetric and efficiency gains in specific treatment sites, EBRT is the far more commonly used type of radiation therapy and will be the focus of this work.[2,3] Today, most courses of EBRT are delivered via a medical electron linear accelerator (or LINAC), which generates a beam of Bremsstrahlung x-rays by bombarding a tungsten target with electrons accelerated through a large potential difference. These beams of x-rays are directed toward the patient and shaped

specifically to result in a dose distribution conforming to the target. Physical beam blocks – such as collimating jaws and thin multi-leaf collimators (MLCs) in the head of the LINAC or customized Cerrobend cutouts – can combine to shape the radiation beam as desired by the treatment team. These LINAC features, among others, have permitted the development of modern advanced radiotherapy delivery techniques.

Intensity modulated radiotherapy (IMRT) is a class of EBRT techniques defined by the utilization of radiation fields with spatially varying fluence patterns. While requiring more sophisticated software and hardware specifications, IMRT has been shown to produce dose distributions with significantly improved target conformality, healthy tissue sparing,[4-6] and overall patient outcomes compared to non-modulated delivery techniques (such as three dimensional conformal radiotherapy, or 3DCRT).[7-9] While there are several specific implementations of using intensity modulated radiation fields for treatment delivery, the two most common are fixed-gantry IMRT and volumetric modulated arc therapy (VMAT). Fixed-gantry IMRT deliveries typically employ five to seven fields of spatially varying fluence patterns of radiation spaced over discrete angles around the patient. On the other hand, VMAT treatments involve the continuous rotation of the LINAC gantry around the patient while delivering intensity-modulated segments of radiation at varying dose rates.[10] Although the literature is inconclusive regarding dosimetric superiority,[11-16] the VMAT technique has been shown to significantly increase treatment efficiency compared to fixed-gantry IMRT.[17-23] VMAT represents the latest technological advancement of rotational delivery techniques and has become routine in clinical practice. Its clinical prevalence has even led some to

debate whether its advantages will soon make conventional fixed-gantry IMRT obsolete.[24]

In order to facilitate VMAT's rise in clinical popularity, software innovations have been necessary to take advantage of the hardware advances that enable a radiotherapy delivery with modulated gantry rotation speeds, MLC positions, and dose rates. Treatment planning systems (TPSs) are computerized applications that provide a virtual environment for the treatment team to efficiently design and simulate a patient's treatment. A TPS models the physical treatment delivery devices (i.e. LINACs) and provides an algorithm for calculating the radiation dose upon computed tomography images of the patient's anatomy. The general process of designing and evaluating simulated treatments within such a system is referred to as treatment planning.

Treatment plans for VMAT deliveries must define hundreds of machine parameters to instruct the LINAC control systems how to deliver the radiation for a given patient. These parameters – notably the gantry speed, MLC leaf positions and speeds, and dose rate – must be chosen in a way that combine to result in the desired dose distribution. For simpler, non-modulated delivery techniques such as 3DCRT, a "forward" planning technique is utilized whereby planning parameters are manually defined and refined iteratively until the plan becomes clinically acceptable. Whereas for more sophisticated delivery modalities like IMRT and VMAT, an "inverse" planning approach – whereby the planner specifies the required clinical endpoints for each patient after which an optimization algorithm searches for a feasible solution – is required to efficiently determine clinically acceptable treatment plans.

Modern TPSs are equipped with these inverse optimization algorithms for addressing the impractical challenge of manually defining those mechanical parameters prior to dose calculation. As mentioned previously, the planner instead specifies arc angles, target dose goals, and dose goals for sparing surrounding organs at risk (OARs) so that an algorithm can heuristically search for a suitable set of intensity patterns and mechanical specifications that achieve the stated dosimetric goals.[25] If a planner wants to improve the plan or assess clinical trade-offs, they must adjust the initial dosimetric objectives and weights and run another optimization round followed by dose computation. Represented as a numerical cost function to be minimized internally by the TPS, deterministic and stochastic optimization methods are typically combined to maximize the likelihood of finding a global – as opposed to local – minimum. However, the existence of these sophisticated inverse planning algorithms still does not guarantee a perfect or even clinically acceptable solution.[25]

Chief among the many factors limiting a planner's ability to arrive at the best treatment plan every time is the unique anatomy of each patient. Traditional planning techniques do not provide planners any *a priori* information regarding what parameters would result in the optimal plan for each patient. This limitation, along with the trial-and-error nature of inverse planning, causes the quality of inversely optimized plans to be susceptible to planner bias, time constraints, and subjectivity. As a result, investigators have observed numerous instances of variation in inverse treatment plan quality.[26-31] Specifically, Nelms *et al*. found a wide variation in plan quality (defined in the study as the ability of planners and plans to meet specified goals) among different treatment planners and institutions.[30] Interestingly, this finding was not statistically dependent on

technologic parameters such as TPS and modality (i.e. fixed-gantry IMRT versus VMAT) nor on planner demographics such as years of experience, certification, and education. Although, Batumalai *et al*. found more experienced planners produced higher quality head-and-neck IMRT plans.[27] The authors instead suggest attributing the variation in plan quality to planner "skill," an undefined and potentially unquantifiable term that underscores how inverse treatment planning can sometimes be more art than science. To this end, researchers began investigating techniques for mitigating these known deficiencies in traditional inverse planning.

1.1.2. Knowledge-Based Planning

One such area of research is knowledge-based planning, or KBP. Although there are many different categories and specific implementations, KBP systems generally leverage retrospective anatomical and dosimetric patient data to guide the planning of new patients. These data-driven approaches have been shown to improve the quality, consistency, and efficiency of IMRT and VMAT planning compared to traditional planning methods.[32] Ge and Wu recently published a review article on KBP systems for IMRT where they classified the different types of KBP implementations into six categories based on the specific variables the models are designed to predict: (1) the entire dose-volume histogram (DVH),[33-64] (2) one or more dose metrics,[45,65-75] (3) voxel-level dose,[76-88] (4) objective function weights,[45,89] (5) beam-related parameters (e.g. number of beams, beam angles, and jaw settings),[90-94] and (6) quality assurance (QA) metrics.[95-101] These specific categories can further be stratified into two major classes according to their underlying mechanisms: (A) case or atlas-based methods and (B) statistical modeling and machine learning methods.

Whereas KBP methods in class (B) form predictive models (e.g. regression and machine learning models) from an established database of previous patients, methods in class (A) query the database for matches or "similar" cases and transfer selected data to the new case. One such case-based technique was one of the first KBP implementations investigated due primarily to its robustness and simple clinical implementation. A seminal method was originally developed by Wu *et al*., where an anatomical similarity metric is used to query a database of previous patients to predict achievable dose-volume histogram objectives in IMRT planning.[65] Specifically, the similarity metric used to quantify patient anatomy is the overlap-volume histogram (OVH).

The OVH, introduced by Kazhdan *et al*., is a shape relationship descriptor that defines the distance at which fractional volumes of OARs lie from the target's surface.[102] More specifically, it is defined for a target *T* and organ *O*, where the value of the OVH of *O* with respect to *T* at distance *r* is defined as the fractional organ volume a distance of *r* or less from the target:

$$OVH_{O,T}(r) = \frac{|\{p \in O | d(p,T) \leq r\}|}{|O|}$$

where *d(p,T)* is the signed distance of a point *p* from the target's boundary and |*O*| is the volume of the OAR. An in-field OVH can be defined similarly for a structure *O'*, or the portion of the organ *O* within the treatment fields. Such in-field volumes can be estimated only considering voxels lying between the transverse planes 6 mm superior and inferior to the most superior and inferior aspects of the planning target volume (PTV) respectively (approximating the beam penumbra at depth). The in-field OVH has

been found to produce improved bladder and rectum dose prediction accuracy for VMAT prostate plans.[103,104]

As mentioned previously, the clinical viability of OVH-driven KBP methods has been investigated due to the OVH's robustness and its simple implementation.[65,66,68,105] All of these methods assume that the dose received by a fractional OAR volume depends on its proximity to the PTV, which is quantified by the OVH. Therefore, each point of an OAR's OVH can be mapped to one point of the corresponding DVH, establishing a one-to-one relationship for each OAR of a given patient (Figure 1.1). This one-to-one distance-to-dose mapping can be formed by relating a distance r_v of an OVH for a fractional OAR volume v to a dose-volume D_v of a DVH. This is the fundamental principle for how the OVH is used to predict achievable DVH dose metrics.

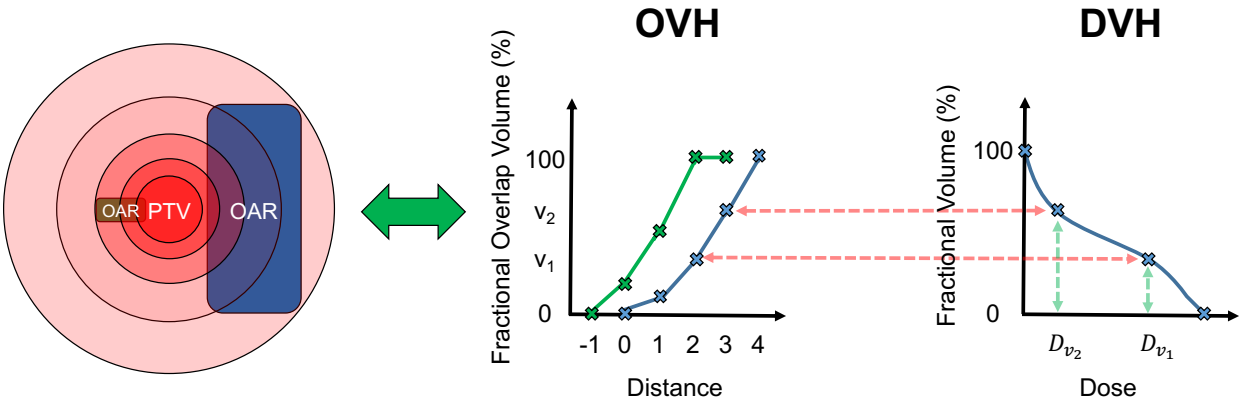


Figure 1.1. Graphic (adapted from Wu *et al*.) illustrating how the OVH is defined for two example OARs and one PTV (left). It also shows the one-to-one relationship between an OAR fractional volume's distance from the target (OVH) and the dose it receives (DVH), which is how OVHs are used to quantify patient anatomy for KBP dose-volume prediction.[105]

Regardless of specific implementation, KBP methods have been observed to contribute to overall improvements in OAR sparing and planning efficiency for several treatment sites. For example, in Ge and Wu's review of KBP approaches for prostate

cancer, they found a mean reduction in bladder and rectum dose of 2.0 and 2.6 Gy, respectively, across four different studies.[32,51,55,59,103] Nevertheless, while reports of these KBP methods leading to improved plan quality and planning efficiency are encouraging, it is important to consider and investigate whether any clinical tradeoffs arise incidental to these positive results.

1.1.3. Plan Complexity

One potential consequence of improved plan quality is an increase in plan complexity. As alluded to in Chapter 1.1.1, IMRT and VMAT have become the preferred treatment delivery techniques for EBRT due to improved dose conformity to diseased tissue and sparing of surrounding healthy tissue compared to traditional 3DCRT. However, these techniques often result in increased beam modulation (i.e. complexity), which has previously been described by the changes in MLC leaf positions, the number of monitor units (MUs), the dosimetric uncertainty owing to losses in charged particle equilibrium caused by smaller beam apertures, and the susceptibility to interplay between the motions of the linear accelerator and internal organs.[106-109]

IMRT and VMAT plans require combinations of many irregularly shaped and oftentimes small beam segments to obtain this increased dose conformity to target and sparing of OARs. Small beam segments have higher degrees of dosimetric uncertainty compared to traditional radiotherapies. This is primarily due to nonequilibrium conditions created by secondary electron track lengths and source sizes being comparable to small treatment field sizes, which results in increased beam penumbra.[110] This uncertainty places an emphasis on the TPS's ability to accurately model lateral electron scatter, MLC leaf ends, leaf transmission, and interleaf leakage for these plans. Among several sources of error, plan complexity has been linked to the deliverability of IMRT

and VMAT plans, with increased complexity often leading to decreases in quality assurance outcomes.[111-113] Therefore, quantifying and reducing IMRT/VMAT plan complexity is a reasonable strategy to improve deliverability.

Many different metrics have been established to quantify and describe plan complexity. Simple characteristic plan parameters such as total MUs give quick first-order indications about a plan's complexity. But other metrics are typically classified as fluence-based or aperture-based. More details on the complexity metrics used in this work can be found in Appendix D.

1.1.4. VMAT Quality Assurance

While VMAT is commonly available on modern commercial LINACs, the accurate delivery of this sophisticated treatment technique requires precise synchronization of MLC motion, gantry motion, and dose rate variations. Additionally, since continuous VMAT arcs are approximated by many discrete segments (or static beams) during the planning process, the delivery accuracy of VMAT treatments may depend on the discretization resolution and plan complexity.[114] This increase in plan complexity (relative to non-modulated delivery techniques) underscores the clinical importance of the QA process, which ensures IMRT treatment plans can be delivered as intended and verifies the accuracy of the TPS dose computation. The current standard for VMAT QA is to measure the planned radiation fields with a physical measurement device, and to compare the result with that computed by the TPS.

This is typically done by copying the approved plan from the patient's geometry onto a water- or tissue-equivalent phantom with known specifications, upon which the dose is then recomputed within the TPS. The plan is delivered to the QA phantom, which contains one or more radiation detectors (typically diodes or ionization chambers)

to measure the dose (Figure 1.2). The agreement between the measured and calculated dose is then evaluated, after which the plan is either approved for treatment or rejected depending on the institution's passing criteria.

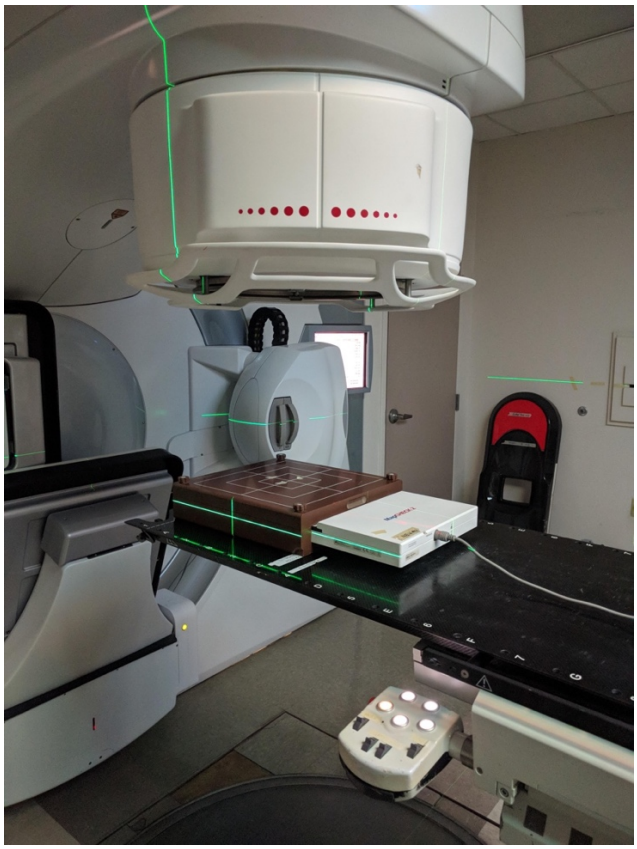


Figure 1.2. This shows the typical setup for VMAT QA at our institution, where the LINAC delivers the approved treatment plan to a known measurement device placed on the couch.

Gamma analysis is one of the most prevalent methods utilized for comparing computed and measured dose distributions in VMAT QA. Introduced by Low *et al.*, the gamma index is used to quantify both the percent dose-difference (%DD) and distance-to-agreement (DTA) between two dose distributions.[115] Specifically, the gamma index γ is defined as,

$$\Gamma(\vec{r_e}, \vec{r_r}) = \sqrt{\frac{r^2(\vec{r_e}, \vec{r_r})}{\Delta d^2} + \frac{\delta^2(\vec{r_e}, \vec{r_r})}{\Delta D^2}}$$

$$\gamma(\vec{r_r}) = \min\{\Gamma(\vec{r_e}, \vec{r_r})\} \forall \{\vec{r_e}\}$$

where $r(\vec{r_e}, \vec{r_r})$ is the distance between the reference and evaluated points, $\delta(\vec{r_e}, \vec{r_r})$ is the dose difference, and Δd and ΔD are the selected DTA and percent dose-difference criteria respectively. γ values equal to or less than one indicate that the comparison passed with respect to the selected %DD and DTA gamma criteria, whereas values greater than one indicate failure. The percent of points passing a given gamma criteria is typically referred to as the gamma passing rate (GPR) and is commonly used as a metric for quantifying the level of agreement between two dose distributions. Due to the complex nature of IMRT treatments, this QA assessment is needed to check for error-free data transfer, the accuracy of the TPS dose calculations, and the deliverability of the plan on the treatment machine.[114] Though there have been investigations into software-based QA protocols,[116-118] measurement-based techniques are still considered the standard for QA.

A consequence of increased plan complexity is the potential for reducing the accuracy of the delivered treatment, whereby the impact of uncertainties in relevant delivery parameters – such as MLC leaf positions – are exacerbated by small, irregularly shaped beam apertures and narrow leaf gap widths called for by the treatment plan. For instance, Masi *et al.* found that increased VMAT plan complexity (described by modulation complexity scores, or MCS, and leaf travel) was significantly correlated with lower quality assurance outcomes.[113] This reduced treatment delivery accuracy can have clinical implications.[113,119] These consequences have led

investigators to explore strategies for quantifying and reducing plan complexity without compromising plan quality.[112,120,121]

1.1.4.1. Machine Learning for Patient-Specific Quality Assurance

Several investigators have applied machine learning models for predicting QA outcomes based on treatment plan characteristics of fixed-gantry IMRT. Valdes *et al*. were able to predict GPRs within 3% accuracy for IMRT plans using a generalized Poisson regression model with Lasso regularization trained on a selection of 78 plan complexity features.[95,96] Interian *et al.* developed an ensemble of convolutional neural networks trained to predict IMRT gamma passing rates from fluence maps with results comparable to the Poisson regression model.[97] More recently, Lam *et al*. used tree-based machine learning algorithms to predict GPRs for portal dosimetry-based IMRT beams with a mean absolute error of less than 1%.[99] Such models can provide *a priori* information regarding the deliverability of plans during the optimization stage, which could provide many benefits that include minimizing wasted time in measuring and adjusting treatment plans that are likely to fail QA.

While machine learning techniques have been primarily applied to fixed-gantry IMRT QA, an evaluation of applying similar techniques for VMAT QA represents a logical next progression. Granville *et al.* showed the feasibility of using machine learning to classify results of VMAT QA measurements, where they trained a support vector machine classifier to predict whether median dose differences between measured and planned dose distributions were ‘hot’ (deviation more than 1%), ‘cold’ (deviation less than -1%), or ‘normal’ (deviation within ±1%).[122] Ono *et al*. used machine learning models to predict GPRs on a specific QA measurement device using plan complexity

features.[101] These previous works show the feasibility of using machine learning algorithms to accurately predict QA outcomes.

1.2. MOTIVATION FOR RESEARCH

While the advantages of KBP techniques in treatment planning have been widely described, the impact of their use on plan complexity and deliverability are less well understood. Comparatively few studies have even reported on simple complexity surrogates of KBP plans. Hussein *et al*. found no significant changes in MU and MCS values of prostate IMRT plans when using a commercial KBP system (RapidPlan, Varian Medical Systems, Palo Alto, CA).[49] Likewise, Tamura *et al.* found no significant changes in patient-specific quality assurance outcomes when using this commercial KBP system.[123] Conversely, Kubo *et al*. reported significantly increased MU values and higher plan complexity when using the same commercial KBP system as the studies noted previously.[124] These results suggest that further assessment of the complexity and deliverability of KBP-guided plans is needed. Therefore, the first purpose of this study was to examine differences in dose, complexity, and quality assurance outcomes between reference clinical plans and plans designed with an in-house KBP system.

Another objective of this work is to build upon existing literature for implementing machine learning models for predicting VMAT QA outcomes. As mentioned in Chapter 1.1.4.1, previous investigators have shown this application of machine learning algorithms to be feasible. However, each machine learning model is dependent on the characteristics of the available data. In this particular application of machine learning, each QA predictive model depends on the combination of technologies, the choice of machine learning model, and clinical protocols used for optimizing VMAT treatment plans, which can each vary across institutions. Therefore, the second purpose of this

work was to assess the feasibility of developing machine learning models for predicting VMAT GPRs at our institution, which utilizes a different combination of technologies (e.g. TPSs, delivery machines, and measurement devices) than previous works. Further, another aim of this work is to explore and assess different machine learning regression algorithms trained to predict gamma passing rates for VMAT QA from treatment planning parameters and metrics.

Current TPSs have simple penalties to globally reduce complexity and for controlling the likelihood of a plan failing VMAT QA. For example, the RayStation TPS (v4.5.1.14; RaySearch Laboratories, Stockholm, Sweden) has one option to constrain MLC leaf motion to a limited distance per degree of gantry rotation. Other works have explored integrating other aperture-based penalties during optimization, where Younge *et al*. were able to reduce complexity without degrading dose.[112] However, these penalties for reducing complexity do not guarantee a corresponding improvement in QA outcomes given their moderate correlation; any potential impact on QA outcomes would also be unknown until measurement. Investigators have consequently explored methods – like machine learning – for predicting delivery accuracy for purposes of identifying plans likely to fail QA prior to measurement. While these predictive models may save time by flagging those at-risk plans, they are limited to the post-planning stage. It would be ideal to actively optimize the treatment plans in terms of both dose- and QA-based metrics. To our knowledge, such an optimization workflow has not been investigated. Therefore, the third purpose of this work was to explore the feasibility of a planning QA tool that directly optimizes QA outcomes without compromising plan quality.

1.3. HYPOTHESES AND SPECIFIC AIMS

To this end, we hypothesized that KBP-guided plans result in significantly higher complexity and reduced gamma passing rates ($p < 0.05$) compared to reference clinical plans. Additionally, we hypothesize that a machine learning model designed to predict VMAT QA gamma passing rates can be used in plan optimization to increase predicted delivery accuracy without compromising KBP plan quality. In order to test these hypotheses, three specific aims were developed for this study:

Aim 1. Evaluate differences in plan complexity and delivery accuracy between KBP and reference clinical plans of prostate cancer patients treated with VMAT. Use four common metrics to describe overall plan complexity for the VMAT plans and perform QA measurements to quantify differences in delivery accuracy between the two sets of plans.

Aim 2. Develop, test, and compare different machine learning models for predicting gamma passing rates of VMAT QA measurements. Collect a large set of plan data to train and test the models for establishing performance.

Aim 3. Develop and establish the feasibility of a planning QA tool. Deploy the model from Aim 2 to interact with a commercial TPS to provide feedback regarding the predicted QA outcome of the treatment plan during plan optimization. Develop an in-house optimization algorithm to maximize predicted delivery accuracy of inversely optimized treatment plans. Demonstrate proof-of-concept on selected KBP plans.

1.4. OVERVIEW OF DISSERTATION

This document follows a manuscript-style dissertation format, where the main chapters contain content derived from papers already published in or in preparation for submission to peer-reviewed scientific journals. Below is a summary of these works and their main contributions to the literature.

1.4.1. Chapter 2: Evaluation of complexity and deliverability of prostate cancer treatment plans designed with a knowledge-based VMAT planning technique

This chapter concerns the assessment of dosimetric, mechanical, and delivery properties of plans designed with a common KBP method for prostate cases treated via VMAT. Thirty-one prostate patients previously treated with VMAT were re-planned with an in-house KBP method based on the overlap-volume histogram. VMAT plan complexities of the KBP plans and the reference clinical plans were quantified via monitor units, modulation complexity scores, the edge metric, and average leaf motion per degree of gantry rotation. Each set of plans was delivered to the same diode-array and agreement between computed and measured dose distributions was evaluated using the gamma index. Varying percent dose-difference (1% to 3%) and distance-to-agreement (1 mm to 3 mm) thresholds were assessed for gamma analyses. KBP plans achieved average reductions of 6.4 Gy ($p < 0.001$) and 8.2 Gy ($p < 0.001$) in mean bladder and rectum dose compared to reference plans, while maintaining clinically acceptable target dose. However, KBP plans were significantly more complex than reference plans in each evaluated metric ($p < 0.001$). KBP plans also showed significant reductions ($p < 0.05$) in gamma passing rates at each evaluated criterion compared to reference plans. While KBP plans had significantly reduced bladder and rectum dose, they were significantly more complex and had significantly worse quality assurance

outcomes than reference plans. These results suggest caution should be taken when implementing an in-house KBP technique.

1.4.2. Chapter 3: Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning

This chapter describes the development and evaluation of machine learning models for predicting QA outcomes of VMAT treatment plans. A dataset of 500 VMAT treatment plans and diode-array QA measurements were collected for this study. GPRs were computed using a 3%/3mm percent dose-difference and distance-to-agreement gamma criterion with local normalization. 241 complexity metrics and plan parameters were extracted from each treatment plan and their relative importance for accurately predicting GPRs was assessed and compared using feature selection methods via forests of Extra-Trees, mutual information, and linear regression. Hyperparameters of different machine learning models – which included linear models, support vector machines (SVMs), tree-based models, and neural networks – were tuned using cross-validation on the training data (80%/20% training/testing split). Features were weakly correlated with GPRs, with the small aperture score (SAS) at 50 mm having the largest absolute Pearson correlation coefficient (0.38; $p < 0.001$). The SVM model, trained using the 100 most important features selected using the linear regression method, gave the lowest cross-validation testing mean absolute error of 3.75%. This represents a significant 41.1% improvement ($p < 0.001$) over “random guessing” error as simulated by randomly sampling a fitted normal distribution to the testing data. These predictive models can help guide the plan optimization process to avoid solutions which are likely to result in lower GPRs during QA.

1.4.3. Chapter 4: Use of machine learning algorithm during optimization to improve patient-specific quality assurance in volumetric modulated arc therapy plans

This chapter assesses the feasibility of optimizing plan deliverability of VMAT treatment plans using a machine learning model to predict QA outcomes. Current inverse planning algorithms incorporate specific mechanical restrictions – such as constraining leaf motion – that are designed to reduce the complexity of the treatment plan. However, mechanical constraints do not guarantee a corresponding improvement in measured QA outcomes. Therefore, this work explored the feasibility of an optimization framework for directly maximizing predicted QA outcomes of plans without compromising dosimetric quality. VMAT plans were retrospectively designed for 13 prostate patients using a previously established in-house KBP method. An SVM was developed – using a database of 500 previous VMAT plans – to predict GPRs (3%/3mm percent dose-difference/distance-to-agreement with local normalization) based on selected complexity features. An optimization algorithm was devised by utilizing the SVM model to iteratively modify mechanical treatment features most commonly associated with suboptimal GPRs. Specifically, leaf gaps (LGs) less than 5 cm were widened by random amounts, which impacts all aperture-based complexity features such as small aperture scores and aperture area uniformity. The original 13 VMAT plans were optimized with this QA-based algorithm using maximum LG displacements of 1, 3, and 5 mm before corresponding changes in predicted GPRs and dose were assessed. Predicted GPRs increased by an average of 0.30 ± 1.22% (p = 0.42), 1.14 ± 1.25% (p = 0.006), and 1.52 ± 1.27% (p =0.003) after QA-based optimization for 1, 3, and 5 mm maximum random LG displacements, respectively. Differences in dose were minimal, resulting in negligible changes in tumor control probability (maximum increase

= 0.05%) and normal tissue complication probability (maximum decrease = 0.22% among bladder, rectum, and femoral heads). A novel framework for optimizing predicted GPRs was developed and shown to increase predicted QA outcomes without degrading dosimetric quality of given plans. This method for integrating QA outcomes directly into planning optimization could help improve the probability and efficiency of arriving at a truly optimal treatment in terms of both dosimetric quality and QA outcomes.

2. EVALUATION OF COMPLEXITY AND DELIVERABILITY OF PROSTATE CANCER TREATMENT PLANS DESIGNED WITH A KNOWLEDGE-BASED VMAT PLANNING TECHNIQUE

2.1. PURPOSE

The purpose of this study was to evaluate differences in complexity and QA outcomes between reference clinical VMAT plans for prostate cancer and those designed with an in-house KBP method.

2.2. MATERIALS AND METHODS

2.2.1. Treatment Plans

A total of 31 prostate cancer patients previously treated at our institution were used for the treatment planning in this study. Selected patients were originally prescribed a dose to a single PTV and treated using two coplanar, 6 MV VMAT arcs. The clinical plans were originally created using the current TPS at our institution (Pinnacle^3 v9.10, Philips Medical Systems, Fitchburg, WI, USA). For research purposes, reference clinical plans were transferred or reconstructed in a research TPS (RayStation v4.5.1.14, RaySearch Laboratories, Stockholm, Sweden), where our in-house KBP method was developed. These reference clinical plans were re-computed or re-optimized to approximate the original clinical plans (see Figure C.1 in Appendix C.1).

In addition to these reference clinical plans, a KBP-guided plan was generated for each of the 31 patients using an in-house KBP technique. The KBP method, based on OVHs incorporating fractional OAR volumes only within the treatment fields,was

---

Contents of this chapter were previously published as Wall PDH, Fontenot JD. Evaluation of complexity and deliverability of prostate cancer treatment plans designed with a knowledge-based VMAT planning technique. *J Appl Clin Med Phys.* 2020;21(1):69-77. Reprinted by permission of Wiley Periodicals, Inc. (Appendix B.1)

used to generate patient-specific bladder and rectum dose-volume predictions at the 10%, 30%, 50%, 65%, and 80% relative volume levels. These dose-volume predictions were then input to the TPS as planning objectives and a KBP-guided plan was optimized for each patient. Additional details of our KBP method are described elsewhere[103,125], but this OVH-guided KBP method was used in this study because it can be easily implemented clinically and it has been previously shown to predict achievable OAR dose-volumes.[65,68,126] Moreover, it is useful to investigate these types of in-house KBP systems for clinics that may not have the resources or ability to acquire commercially available KBP systems.

A unique feature of this KBP method is that in addition to the manually-constructed clinical plans, the dose database was also populated with standardized Pareto-optimal plans that equally weighted sparing of each OAR. When a dose-volume was queried for patients with similar in-field OVHs to the new patient, the lowest dose value among both the clinical and Pareto plans was selected as the new patient's predicted dose-volume. Our previous work showed the knowledge from the Pareto plans often resulted in better achievable dose predictions.[125] So it is important to emphasize that the predicted dose-volume objectives from this KBP technique are selected from the lowest dose values among the clinical and Pareto plans available in the database.

It is also important to note that the KBP system is separate from the TPS optimization engine, whereby the KBP algorithm predicts dose-volume objectives to input into the TPS optimizer. For the KBP-guided plans, the planner strove to achieve the bladder and rectum dose predictions along with originally prescribed physician goals

for the target and remaining OARs. Once clinically acceptable target coverage was achieved, OAR sparing was optimized until either the KBP dose predictions were achieved or target coverage became clinically unacceptable.

Each set of 31 reference clinical plans and KBP plans were planned on the same commercial TPS under the same planning conditions, which included the same planner, machine, maximum leaf motion, dose grid resolution, and control point spacing. The primary reason for reconstructing the reference clinical plans was to account for the variations in these parameters that were used to design the original plans. Keeping these parameters and other complexity mitigation tools constant for the reference plans and KBP plans was desired in order to make the fairest comparison between these two sets of plans. All plans were designed to be delivered on a commercial linear accelerator equipped with a 160-leaf MLC (Infinity, Elekta AB, Stockholm, Sweden) and were optimized with a maximum leaf motion of 7 mm per degree of gantry rotation (mm/deg), dose grid resolution of 4 mm, and control point spacing of 4 degrees. Both reference and KBP plans were optimized under the same conditions with the same dosimetric endpoints for the target and OARs not including the bladder and rectum. The only optimization differences between reference and KBP plans were the bladder and rectum planning objectives, where reference plans utilized the original clinical goals and KBP plans used the predicted dose-volumes from the KBP method as described previously.

The plan quality of the reference plans and the KBP plans were compared qualitatively with DVHs and quantitatively with Wilcoxon signed-rank tests on an array of dose metrics at a significance level of $p = 0.05$.

2.2.2. Plan Complexity

The complexities of reference plans and KBP plans were quantified using the total planned MUs, MCS[113,120], the edge metric (EM)[112], and the average MLC leaf motion per degree of gantry rotation (LM). Planned MU values were normalized by the fractional prescription dose to enable the comparison between plans of differing prescriptions. The MCS was originally introduced by McNiven *et al*. to assess fixed-gantry IMRT modulation complexity and was later adapted for VMAT by Masi *et al*.[113,120] Briefly, the MCS is a metric ranging from zero (most modulated) to one (least modulated) that incorporates leaf sequence variability and aperture area variability components weighted by their segment contributions. The EM is computed as the segment weighted ratio of in-field MLC side length and aperture area, which was introduced by Younge *et al*. to characterize the amount of "edge" in apertures.[112] The values for the scaling factors used in this work were $C_1 = 0$ and $C_2 = 1$. LM was determined by averaging the change in leaf position per degree of gantry rotation calculated at each control point in the plan over each MLC leaf within the jaws.

These metrics were computed directly from the DICOM (Digital Imaging and Communications in Medicine) RT Plan files using in-house software.[127,128] MCS, EM, and LM values of the reference and KBP plans were compared using two-sided paired t-tests at a significance level of $p = 0.05$. Since the distribution of differences between reference and KBP plan MUs did not meet the normality assumption of the t-test, a two-sided paired Wilcoxon signed-rank test was used for comparing MUs.

2.2.3. Delivery Accuracy

Each of the 31 reference and KBP plans were delivered on the commercial linear accelerator platform for which it was planned (Infinity, Elekta AB, Stockholm, Sweden).

Dosimetric measurements were performed using a commercial diode-array housed in a water-equivalent phantom (MapCHECK2 and MapPHAN; Sun Nuclear Corporation, Melbourne, FL, USA). Each set of plans was delivered on three separate occasions in order to reduce effects of measurement noise and fluctuations. The diode array was calibrated prior to each measurement session to eliminate the influence of daily variations in machine output and detector response. While there are several other dosimeters with their own advantages and disadvantages, such as film or EPID panels, the MapCHECK2 was used in this study primarily to mimic the clinical protocol used at our institution. Additionally, while film and EPID panels provide high spatial resolution for relative measurements, they are not ideal absolute dosimeters. Detector arrays also measure dose at detector locations more accurately than film due to processing and densitometry uncertainties and are easier to use compared to film.[114]

Calculated dose distributions were generated for the measurement geometry and plane by the TPS and were compared to measured data using gamma analysis in this study as described in Chapter 1.1.4. Within gamma analysis, it is important to note the role dose normalization plays in the %DD gamma criterion. Two possible normalizations are global and local, where the former normalizes dose to the maximum dose in either dose distribution and the latter normalizes dose to the dose at the local point being evaluated. Typical tolerance limits are set as a percentage of points passing the given gamma criteria (i.e. GPR). Task Group No. 218 (TG-218) recently recommended tolerance and action limits for GPRs to be greater than or equal to 95% and 90%, respectively, using a %DD/DTA gamma criterion of 3%/2mm with global normalization.[114]

Percent dose-difference and distance-to-agreement criteria of 3%/3mm, 2%/2mm, and 1%/1mm with both global and local normalization were used to evaluate the agreement between the dose distributions. The 3%/2mm global criterion was additionally examined to align with the recent recommendations for universal tolerance and action limits from TG-218.[114] Gamma passing rates were computed for each plan using commercial quality assurance software (SNC Patient, Sun Nuclear Corporation, Melbourne, FL, USA), where only points with dose above 10% of the maximum dose were included in the analysis. The built-in calculated shift software feature was used to account for setup uncertainties. Passing rates of the reference and KBP plans were averaged over the three deliveries and statistically compared using two-sided paired t-tests at a significance level of $p = 0.05$.

2.3. RESULTS

2.3.1. Plan Quality

Table 2.1 shows how reference plans compare statistically with KBP plans for an array of DVH points and dose metrics. While KBP plans showed significant differences in some PTV dose metrics compared to reference plans, it should be noted KBP PTV doses were clinically acceptable and statistically equivalent to the original clinical PTV doses. KBP plans showed significant ($p < 0.001$) decreases in bladder and rectum doses compared to the reference plans. On average, D_{mean} for the bladder and rectum was 6.4 Gy and 8.2 Gy lower for KBP plans compared to reference plans, respectively. Average DVHs and the standard errors of the means are shown in Figure 2.1 for the reference and KBP plans.

Table 2.1. Statistical summary of dose values between reference and KBP plans.Note that all doses were normalized so that 95% of the PTV received 76 Gy.

Dose Metric	Means ± Standard Deviations		Wilcoxon *p*-value
	Reference	KBP	Reference vs. KBP
PTV			
D_2 (Gy)	78.7 ± 0.9	79.2 ± 1.1	0.001*
D_{50} (Gy)	77.4 ± 0.5	77.7 ± 0.8	0.02*
D_{98} (Gy)	75.2 ± 0.9	74.8 ± 0.9	< 0.001*
D_{min} (Gy)	67.1 ± 6.6	63.3 ± 7.0	< 0.001*
D_{mean} (Gy)	77.3 ± 0.5	77.6 ± 0.7	0.013*
D_{max} (Gy)	79.6 ± 1.3	80.6 ± 1.5	< 0.001*
V_{95} (%)	99.7 ± 0.7	99.5 ± 0.8	< 0.001*
V_{98} (%)	99.0 ± 1.2	98.4 ± 1.2	0.77
V_{100} (%)	94.6 ± 2.3	94.8 ± 1.4	0.98
V_{107} (%)	0.02 ± 0.1	0.2 ± 0.5	0.011*
HI†	0.05 ± 0.02	0.06 ± 0.02	< 0.001*
CI†	1.4 ± 0.1	1.4 ± 0.06	0.002*
Bladder			
D_{10} (Gy)	73.6 ± 6.1	68.4 ± 13.8	< 0.001*
D_{30} (Gy)	48.5 ± 18.7	38.2 ± 22.7	< 0.001*
D_{50} (Gy)	27.2 ± 18.5	19.5 ± 16.2	< 0.001*
D_{65} (Gy)	17.9 ± 15.7	11.6 ± 9.7	< 0.001*
D_{80} (Gy)	12.5 ± 13.0	7.7 ± 6.5	< 0.001*
D_{mean} (Gy)	35.1 ± 12.8	28.6 ± 12.0	< 0.001*
Rectum			
D_{10} (Gy)	72.8 ± 4.6	69.1 ± 7.9	< 0.001*
D_{30} (Gy)	51.4 ± 12.1	39.8 ± 17.0	< 0.001*
D_{50} (Gy)	35.4 ± 12.8	21.9 ± 12.9	< 0.001*
D_{65} (Gy)	24.6 ± 12.9	13.8 ± 9.0	< 0.001*
D_{80} (Gy)	15.0± 11.9	8.9 ± 6.5	< 0.001*
D_{mean} (Gy)	38.0 ± 8.9	29.8 ± 9.3	< 0.001*
Left Femoral Head			
D_2 (Gy)	40.1 ± 6.7	40.4 ± 7.1	0.019*
D_{max} (Gy)	45.9 ± 10.4	47.8 ± 10.2	< 0.001*
D_{mean} (Gy)	26.6 ±5.1	26.1 ± 5.6	0.95
Right Femoral Head			
D_2 (Gy)	39.5 ± 6.8	40.5 ± 7.7	< 0.001*
D_{max} (Gy)	44.8 ± 9.8	47.0 ± 9.3	< 0.001*
D_{mean} (Gy)	26.8 ± 5.1	26.6 ± 6.0	0.65

(table cont'd)

Dose Metric	Means ± Standard Deviations		Wilcoxon *p*-value
	Reference	KBP	Reference vs. KBP
Penile Bulb D_{mean} (Gy)	35.5 ± 18.4	34.9 ± 18.9	0.18

*Indicates a statistically significant result of $p < 0.05$
†Homogeneity and conformity indices were calculated according to their International Commission on Radiation Units & Measurements definitions.

2.3.2. Plan Complexity

KBP plans were significantly more complex than reference plans in every evaluated metric. On average, KBP plans required 143 more MUs ($p < 0.001$), had reduced MCS values of 18% ($p < 0.001$; indicating increased complexity), had 40% higher EM values ($p < 0.001$), and 47% higher LM ($p < 0.001$) compared to reference plans (Table 2.2).

Table 2.2. Statistical summary of the differences in complexity metrics between the reference and KBP plans.

Complexity Metric	Reference Plans ($\mu \pm \sigma$)	KBP Re-plans ($\mu \pm \sigma$)	t-test *p*-value
MU	450 ± 83	593 ± 113	< 0.001*†
MCS	0.5 ± 0.1	0.4 ± 0.1	< 0.001*
EM	0.06 ± 0.02	0.08 ± 0.01	< 0.001*
LM (mm/deg)	1.0 ± 0.6	1.5 ± 0.5	< 0.001*

*Indicates a statistically significant result of $p < 0.05$
†Result from two-sided Wilcoxon signed-rank test because the distribution of differences in MUs between reference and KBP plans was determined to break the t-test assumption of normality.

Complexity metrics were shown to be strongly correlated with each other. An increase in MUs correlated with more complex MCS, EM, and LM values with Pearson correlation coefficients (R) of -0.85 ($p < 0.001$), 0.91 ($p < 0.001$), and 0.84 ($p < 0.001$) respectively. Similarly, an increase in complexity in terms of the MCS score correlated

strongly with an increase in EM values ($R = -0.94$; $p < 0.001$) and LM ($R = -0.88$; $p < 0.001$). Lastly, an increase in EM values corresponded strongly with more LM with $R = 0.85$ ($p < 0.001$).

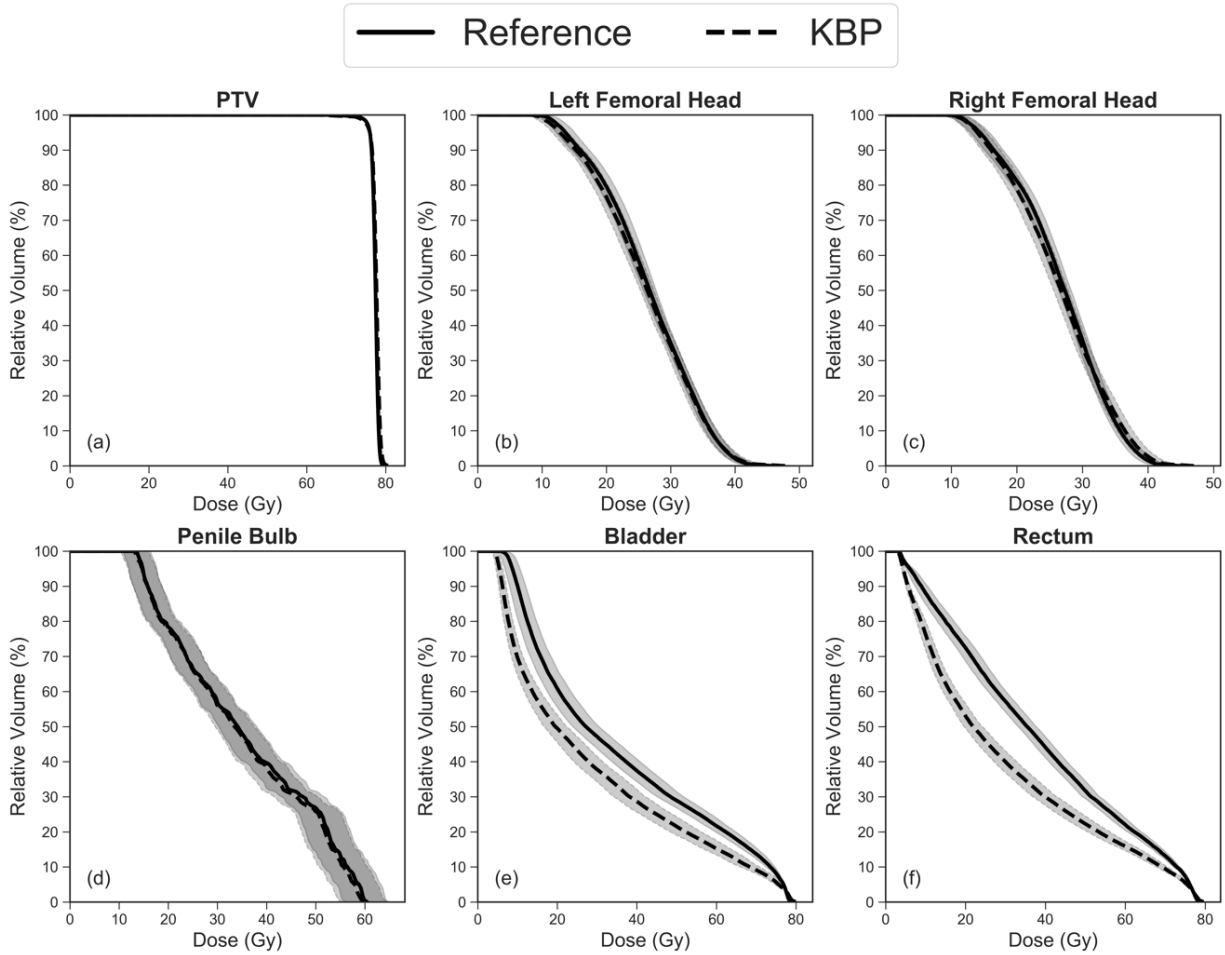


Figure 2.1. Average DVHs comparing reference clinical plans (solid) and KBP plans (dashed) for the 31 patients of each labelled planning structure (a-f). The standard error of the means is also included as filled bands with solid (reference) or dashed (KBP) edge lines. Note that doses were normalized so that 95% of the PTV received 76 Gy.

2.3.3. Delivery Accuracy

KBP plans showed significant reductions in quality assurance outcomes compared to reference plans as described by gamma passing rates. For criteria with global normalization, KBP plans on average had gamma passing rates that were 1.1,

1.6, 3.8, and 7.8 percentage points lower than reference plans at the 3%/3mm (p = 0.009), 3%/2mm (p = 0.003), 2%/2mm (p = 0.002), and 1%/1mm (p < 0.001) criteria respectively. Significant reductions in KBP plan gamma passing rates compared to the reference plans were also observed at each evaluated gamma criteria using local normalization (Table 2.3). Additionally, it is notable that KBP plans showed significantly greater inter-delivery variations (p < 0.05) in gamma passing rates than reference plans at each gamma criteria for both global and normalization methods (Table C.1).

Table 2.3. Statistical summary of the differences in gamma passing rates between the reference and KBP plans at different gamma criteria.

	Gamma Criteria	Reference Plans Gamma Pass Rates ($\mu \pm \sigma$)	KBP Plans Gamma Pass Rates ($\mu \pm \sigma$)	t-test p-value
Global	3%/3mm	98.8 ± 1.3	97.7 ± 2.5	0.009*
	3%/2mm	98.3 ± 1.7	96.6 ± 3.3	0.003*
	2%/2mm	93.8 ± 4.2	90.0 ± 6.8	0.002*
	1%/1mm	69.7 ± 8.7	61.9 ± 11.7	< 0.001*
Local	3%/3mm	91.8 ± 4.4	88.9 ± 6.6	0.03*
	2%/2mm	87.4 ± 6.0	82.0 ± 9.5	0.003*
	1%/1mm	75.4 ± 8.7	66.5 ± 11.9	< 0.001*

*Indicates a statistically significant result of p < 0.05

Patient-specific QA outcomes at the 3%/3mm global gamma criterion for all reference plans were greater than 95%. As for the KBP plans, only two plans had passing rates of less than 95% but greater than 90%. One KBP plan had a gamma passing rate of lower than 90% (87.7%).

Gamma passing rates were also found to be weakly to moderately correlated with the evaluated plan complexity metrics (Table 2.4). For instance, gamma passing rates at the 2%/2mm local criterion moderately correlated with MUs (R = -0.47; p <

0.001), MCS values (R = 0.42; $p < 0.001$), EM values (R = -0.40; $p = 0.001$), and LM (R = -0.37; $p = 0.003$).

Table 2.4. Pearson correlation coefficients between complexity metrics and gamma passing rates.

	Gamma Criteria	Pearson Correlation Coefficients (*p*-value)			
		MU	MCS	Edge Metric	Leaf Motion
Global	3%/3mm	-0.37 (0.003*)	0.36 (0.004*)	-0.36 (0.004*)	-0.33 (0.009*)
	3%/2mm	-0.43 (< 0.001*)	0.40 (0.001*)	-0.40 (0.001*)	-0.38 (0.002*)
	2%/2mm	-0.45 (< 0.001*)	0.39 (0.002*)	-0.39 (0.002*)	-0.36 (0.005*)
	1%/1mm	-0.50 (< 0.001*)	0.45 (< 0.001*)	-0.46 (< 0.001*)	-0.40 (0.001*)
Local	3%/3mm	-0.35 (0.006*)	0.31 (0.02*)	-0.29 (0.02*)	-0.21 (0.10)
	2%/2mm	-0.47 (< 0.001*)	0.42 (< 0.001*)	-0.40 (0.001*)	-0.37 (0.003*)
	1%/1mm	-0.56 (< 0.001*)	0.50 (< 0.001*)	-0.50 (< 0.001*)	-0.49 (< 0.001*)

*Indicates a statistically significant result of $p < 0.05$

2.4. DISCUSSION

In this work, VMAT plans for prostate cancer patients designed with an OVH-guided KBP method were significantly more complex and had significantly lower patient-specific quality assurance outcomes compared to manually-constructed reference plans. While KBP-guided plans led to significant improvements in OAR sparing, the values of MU, MCS, EM, and LM were all significantly more complex. In addition, a weak to moderate correlation was observed between the analyzed complexity metrics and quality assurance outcomes.

To our knowledge, this work is the first evaluation of plan complexity and deliverability of plans derived from an *OVH-guided* KBP method, whereas other studies have reported results from the commercial KBP product*, RapidPlan*. The observed

improvements in KBP plan quality are consistent with previous studies investigating KBP methods for prostate cancer.[67,76] The OAR dose-volume predictions generated from the OVH-guided KBP model are designed to output the lowest achievable dose levels based on previous data. The results from this study indicate that the achievability of these predictions seem to come at the cost of significant increases in plan complexity, which is consistent with the work of Kubo *et al*.[124] On the other hand, Tamura *et al*. reported KBP plans to have similar complexity to reference plans overall. They also observed significantly reduced ($p < 0.05$) leaf travel in KBP plans. Both Tamura *et al.* and Kubo *et al*. each evaluated 30 prostate patients using the same commercial KBP system. It is worth highlighting the differences between the KBP method used in the present study and the commercial system (RapidPlan) used in these previous studies. Whereas the RapidPlan training algorithm uses model-based principal component regression[129], the KBP system in this work follows an established library lookup algorithm to find the lowest dose achieved among a database of previously treated patients with similar in-field OVHs to the new patient.[126] Also, as mentioned previously in Chapter 2.2.1, standardized Pareto-optimal plans were added to the dose database and were found to further improve OAR sparing overall compared to using data from manually-constructed clinical plans.[125] This distinguishes this OVH-driven KBP system from RapidPlan's regression-model technique. Therefore, one possible explanation for this discrepancy among previous works could be the differences in dose objectives and resulting distributions*, i.e.*, the extent to which one study more aggressively pursued a better plan result than the other. Further, the differences between previous findings and our results may be explained by the differing

KBP techniques and also differences in the quality of the underlying dose databases these KBP systems used to generate their dose objectives. While KBP plans in both previous studies achieved similar dose to clinical plans overall, KBP plans reported by Kubo *et al*. showed significantly lower mean bladder dose along with significantly higher MUs and more complex MCS values. Our study achieved similar bladder dose reductions as Kubo *et al*. These results therefore suggest the possibility that increased complexity may be required in order to meet the "ideal" dose objectives, in which case efforts to mitigate complexity may reduce the quality of the KBP-guided dose distributions.

We did not observe a strong correlation between improved bladder mean dose and increased complexity and only a moderate ($R \geq 0.48$) correlation between improved rectum mean dose and increased complexity (Figure C.4). Plan complexity metrics were also not strongly correlated with gamma passing rates. This observation is consistent with previous works[113,130] and could indicate the selected plan quality metrics cannot fully describe plan complexity, even though available evidence suggests a relationship does exist. While there have been studies showing increased monitor units are necessary for achieving desired dose distributions for certain IMRT cases with complex geometry[131], other studies have observed instances of unnecessary VMAT plan complexity and were able to reduce complexity without substantially impacting plan quality using complexity penalties.[112] As Mohan *et al*. noted, the amount of possible complexity reduction is dependent on the difficulty of the underlying treatment plan.[131,132] However, it remains uncertain whether the increased complexity observed in the KBP plans of this study was required for the improved OAR dose. Further

investigation into what extent the complexity of these KBP plans could be mitigated by exploring different TPS optimization settings is warranted.

The clinical implications of increased plan complexity and reduced delivery accuracy have been studied extensively, which served as a primary motivation for this study. Investigators such as Younge *et al*. have implemented aperture complexity penalties into the plan optimization stage to limit plan complexity without degrading plan quality.[112] Others have examined how an array of metrics that quantify beam complexity (such as leaf travel in addition to plan irregularity and modulation) correlate with delivery accuracy and pre-treatment verification results.[119,121] Valdes *et al*. recently showed the feasibility of using machine learning techniques to accurately predict gamma passing rates of IMRT plans using many complexity features.[95,96] It is possible that further accounting for plan complexity using these similar methods during the optimization stage could reduce KBP's observed impact on complexity on a plan-specific level, thereby providing a more accurate delivery. This is an avenue of research we plan on investigating in future work.

This study had several limitations. Planning time was not explicitly recorded in this research since KBP has been extensively shown to improve planning efficiency.[44,46,52,53] However, average KBP planning time was qualitatively comparable to average reference planning time in the present study. Also, standard clinical values for control point spacing and dose grid resolution were used in this work. While it is possible increasing the resolution of these two parameters could mitigate the observed differences in calculated and measured KBP dose[113], the settings used in this study have been shown to provide an acceptable balance of calculation accuracy and

speed.[133,134] Also, the leaf motion was not constrained beyond the default limits of the modeled linear accelerator. It is possible that adjusting these specific optimization parameters may diminish KBP plan complexity and delivery accuracy deficiencies to an extent.[135-137] Another potential limitation is that only a limited number of metrics where chosen to quantify plan complexity, though the chosen metrics are commonly used in the literature.[113,121]

The use of a diode-array also presents potential limitations. A previous study have observed a slight temperature dependence for individual MapCHECK diodes ranging from 0.52% to 0.57%/°C.[138] The impact of any existing temperature dependence would likely be negligible in the present study as the measurements for KBP and reference plans were acquired consecutively and in temperature-controlled rooms. Also, other studies have shown an angular dependence to be the factor that most affects the accuracy of MapCHECK2 measurements – particularly at gantry angles of 90 and 270 degrees – which could potentially affect gamma passing rates.[139,140] While this study did not directly investigate the effects on these temperature and angular dependencies on delivery accuracy, the gamma passing rates at clinically-relevant criterion for the plans in this study were consistent with those observed at our clinic for prostate cases. Additionally, the commercial diode array used in this work has been shown to provide accurate VMAT QA measurements despite this angular dependence.[140-142] It is also important to note the results seen here with this specific combination of technologies (OVH-guided KBP method, RayStation TPS, Elekta treatment machine, and MapCHECK2 diode array) may not hold for different KBP methods and planning, delivery, and measurement technologies as evidenced by the

results from Tamura *et al*.[123] Regardless, the results of this study indicate that caution is needed regarding the effects of plan complexity and quality assurance outcomes when implementing any KBP system as they become more clinically prevalent. However, these results supplement the available literature showing KBP's potential in providing immediate and substantial clinical impact. In-house OVH-guided KBP systems similar to the one described in this work could be developed and implemented clinically without disrupting the existing inverse optimization workflow. The KBP system would provide patient-specific predicted bladder and rectum dose-volume planning objectives prior to planning, and the planner could then strive to meet these KBP goals as they would normally. The observed increase in complexity and reduction in QA outcomes from this study may warrant additional focus on the quality control of KBP plan delivery. The qualified medical physicist would be responsible for monitoring the deliverability of VMAT plans designed with any KBP system. Provided that any reduction in QA outcomes does not result in consistently unacceptable plans, the substantial potential improvement in plan quality provided by KBP systems should persuade the clinical physicist to investigate the feasibility of implementing a KBP system within his or her institution.

Our study did not investigate the source of the reduced quality assurance outcomes of KBP-guided plans. While it would certainly be important and desirable to characterize the specific causes of delivery accuracy discrepancies between KBP and reference plans we leave this for future research as it lies outside the scope and purpose of the current work. However, there are known categories of uncertainties in the IMRT planning and delivery process that include: limitations of the beam model (e.g.

MLC modeling, modeling output factors for small fields, etc.), mechanical and dosimetric uncertainties of the delivery system (e.g. MLC leaf position and speed errors, gantry rotation and table motion stability, beam stability, etc.), and measurement and analysis uncertainties (e.g. setup errors).[114] Given evidence available in the literature, the primary source of error in the discrepancies between KBP and reference GPRs is most likely inaccuracies in the TPS dose computation. For instance, Masi *et al*. observed increased GPRs with plans optimized with a finer control point spacing compared to plans of similar complexity optimized with a courser control point spacing.[113] Therefore, since the KBP and reference plans were optimized under the same TPS settings, the resulting differences in GPRs may primarily be caused by limitations in the TPS's ability to accurately model and compute dose of plans of higher complexity. Increasing the control point spacing resolution during KBP plan optimization could mitigate the observed delivery errors to some extent, but future work is needed to fully describe the specific sources of error. KBP effects on plan complexity and gamma passing rates for different treatment sites, such as the head and neck, would also be instructive to explore. But overall, this research gives reason to further validate and verify all aspects of the treatment workflow when implementing KBP systems, whether they be established in-house or commercially available methods.

2.5. CONCLUSION

While KBP methods have been shown to improve the quality and consistency of treatment plans across institutions, the results of this study suggest their use can significantly increase plan complexity and reduce patient-specific QA outcomes. An in-field OVH-guided KBP method was used to generate 31 VMAT plans for previous prostate cancer patients. KBP plans showed significantly reduced bladder and rectum

dose but were significantly more complex compared to reference plans. The KBP plans showed a significant reduction in delivery accuracy - as measured by patient-specific QA measurements. These results demonstrate that care should be taken when implementing KBP models to ensure resulting plans achieve acceptable quality and deliverability.

3. APPLICATION AND COMPARISON OF MACHINE LEARNING MODELS FOR PREDICTING QUALITY ASSURANCE OUTCOMES IN RADIATION THERAPY TREATMENT PLANNING

3.1. PURPOSE

The purpose of this study was to develop, validate, and compare different machine learning models for predicting quality assurance outcomes of VMAT treatment plans.

3.2. MATERIALS AND METHODS

3.2.1. Data

This study was approved by the Institutional Review Board at Louisiana State University under IRB# E11428 (see Appendix A). In total, 500 dataset samples were collected and anonymized from patients who were previously treated with VMAT at our institution. The specific data collected from each patient case consisted of the following: the DICOM RT-Plan file, containing the technical parameters of the clinically approved treatment plan; the QA measurement file, containing the measured dose acquired during QA; and the predicted QA dose output from the TPS. The inclusion criteria for each sampled case required each patient to have been previously treated with at least one VMAT arc and the subsequent QA measurement to have been performed with a MapCHECK2 diode-array housed in a MapPHAN water-equivalent phantom (Sun Nuclear Corporation, Melbourne, FL, USA). The diode-array was calibrated prior to each measurement to eliminate the influence of daily variations in machine output and

---

Contents of this chapter were previously published as Wall PDH, Fontenot JD. Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning. *Informatics in Medicine Unlocked.* 2020;18:100292. Reprinted by permission of Elsevier Ltd. (Appendix B.2)

detector response. All treatment plans were designed using Pinnacle^3 (v9.10, Philips Healthcare, Amsterdam, Netherlands) and delivered using one of four matched Elekta linear accelerators equipped with Agility MLC heads (Elekta AB, Stockholm, Sweden).

3.2.1.1. Features

Treatment plan parameters and characteristics were extracted from the DICOM RT-Plan file of each sample using in-house software. These parameters were used to formulate an array of features that would later be used to develop machine learning regression models for predicting QA outcomes. This consisted of 241 raw features, which can be categorized into the 23 groups of treatment plan parameters and complexity metrics listed in Table 3.1.

Table 3.1. Summary of 23 feature groups assembled for this study.

Feature Group	Reference(s)	Notes
Modulation Complexity Score (MCS)	McNiven *et al.*[120] and Masi *et al.*[113]	–
Edge Metric (EM)	Younge *et al.*[112]	–
Leaf Travel (LT)	–	Total leaf travel per leaf; includes 1st, 2nd, 3rd, 4th, and 5th moments unweighted and weighted by segment MU
Leaf Motion (LM)	–	Defined as average leaf travel per degree of gantry rotation; includes 1st, 2nd, 3rd, 4th, and 5th moments unweighted and weighted by segment MU
Arc Length	–	Total degrees of gantry rotation in plan
MU Factor	–	Total planned MUs normalized by fractional dose to specification point
Number of Arcs	–	
Average Jaw Position (JP)	–	Feature for each jaw and feature with average combined jaw position
Jaw Travel (JT)	–	Total travel for each jaw
Average Jaw Motion (JM)	–	Average jaw travel per degree of gantry rotation for each jaw; unweighted and weighted

(table cont'd)

Feature Group	Reference(s)	Notes
Average collimator angle	–	–
Aperture Area (AA)	Du *et al.*[143]	1st, 2nd, 3rd, 4th, and 5th moments unweighted and weighted by segment MU
Aperture Perimeter (AP)	Du *et al.*[143]	1st, 2nd, 3rd, 4th, and 5th moments unweighted and weighted by segment MU
Plan Irregularity (PI)	Du *et al.*[143]	1st, 2nd, 3rd, 4th, and 5th moments unweighted and weighted by segment MU
Leaf Gap (LG)	–	Distance between opposing in-field leaf pairs; 1st, 2nd, 3rd, 4th, and 5th moments unweighted and weighted by segment MU
Plan Modulation (PM)	Du *et al.*[143]	–
Small Aperture Score (SAS)	Crowe *et al*.[144]	Additional feature included maximum SAS among all control points
Cross-axis Score (CAS)	Crowe *et al*.[144]	–
Fractional Area Outside of Circle (FAOC)	Valdes *et al.*[95]	1st, 2nd, 3rd, 4th, and 5th moments unweighted and weighted by segment MU
Nominal Dose Rate	–	–
Treatment Machine	–	One of four dosimetrically matched machines
Treatment Site	–	–
Flattening Filter Free (FFF)	–	–

Three of these feature groups include the following categorical features: (1) the specific treatment machine on which the plan was delivered, (2) the treated anatomical site, and (3) the use of a specialized high-dose delivery mode. In sum, 29 abdomen, 13 breast, 36 chest, 13 chest wall, 148 head and neck, 127 lung, 61 prostate, 30 prostatic fossa, 32 pelvis, and 11 miscellaneous (knee, spine, and shoulder) treatment plans were collected for this study. The remaining 20 feature groups were numerical and derived from complexity features found in the literature. For groups characterized by distributions, features such as mean, standard deviation, and up to the 5th moment

about the mean were extracted. Features were also implemented with and without accounting for segment MU weightings where applicable.

3.2.1.2. Target Values

GPRs were computed using in-house software from each sample's predicted QA dose file from the TPS and the measured dose from performing the QA. The software was modified from existing open-source code for computing the gamma index to interface with the input data and to include a feature accounting for setup errors . Setup errors were mitigated by shifting the planned dose distribution in 1 mm steps along the cardinal axes within a 10 mm radius to search for better agreement. The foundation of the computation algorithm and validation procedure was based on previous work.[145] The gamma index for each point was computed with a percent dose-difference criterion of 3%, a distance-to-agreement criterion of 3 mm, local normalization, and a 10% dose threshold.

3.2.1.3. Data Processing

In order to avoid data snooping bias in the development and evaluation of machine learning models, the overall dataset was split into a training set and a testing set with 400 and 100 samples respectively (i.e. 80%/20% split). Data was split using a stratified technique based on the distribution of GPRs to guarantee the testing set be representative of the overall population of GPRs. Figure 3.1 shows the differences in random and stratified sampling for splitting the data, where the standard randomized technique resulted in a test set distribution that was less representative of the total distribution compared to the stratified technique. This is particularly pronounced when comparing the standard deviation of the two testing set distributions against that of the overall distribution in the Figure 3.1 example. The overall GPR distribution of the

dataset had a standard deviation of 6.01%, compared to 6.13% and 5.49% for the testing sets resulting from a stratified and random sampling, respectively. Therefore, the stratified sampling avoids possible sampling bias in the target variable by generating training and testing sets with target distributions representative of the overall dataset distribution.

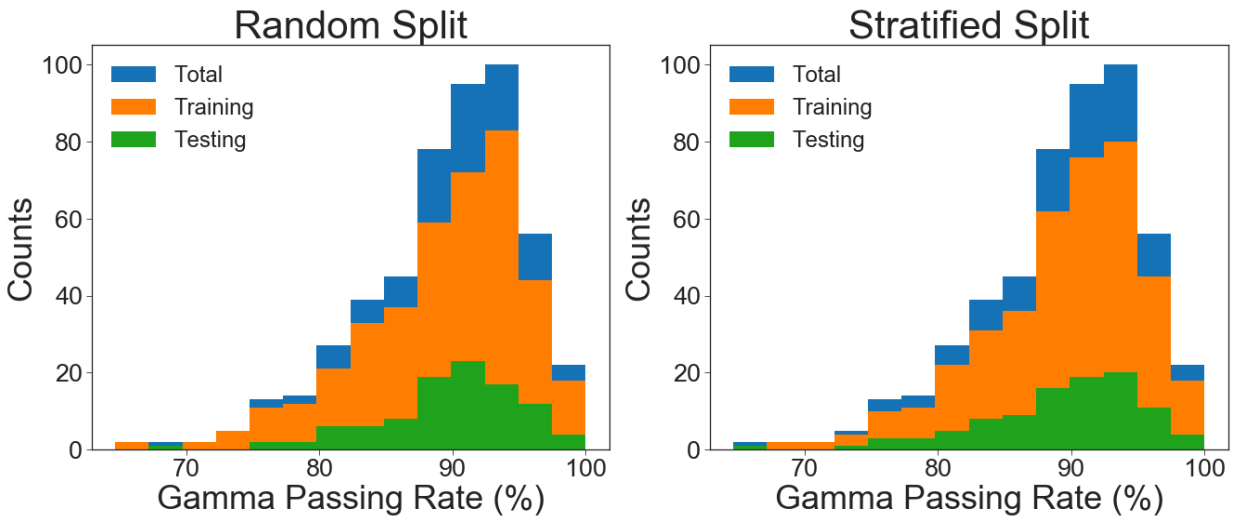


Figure 3.1. Histograms of GPR distributions for the entire dataset (blue), training set (orange), and testing set (green) when utilizing a random (left) and stratified (right) sampling technique. The stratified technique provides training and testing sets with GPR distributions sampled proportionally from the entire dataset, which avoids bias from over- or under-sampling certain ranges of GPRs as seen with the purely randomized technique.

After splitting the data, categorical features were one-hot encoded to represent every category in each feature group as a binary attribute. Feature standardization was applied to numerical features, where the mean value was subtracted and the variance normalized to one. This transformation was fit to the training data and was then applied to the testing set i.e. the testing set data was not used in the initial standardization process.

3.2.2. Feature Analysis and Selection

Feature selection is a process used in developing machine learning models that can have several benefits. Selecting a subset of the most important features can simplify interpretation and visualization of the data, reduce training and utilization times, or improve the overall performance and robustness of the predictive model.[146] In this study, each feature's relative importance in contributing to the accurate prediction of GPRs was quantified and ranked using three different methods: forests of extremely randomized decision trees (Extra-Trees), mutual information, and linear regression.

Decision Trees are a powerful machine learning algorithm that work to search for the feature and the decision threshold within that feature that best splits the training data into similar response categories. More specifically, the feature and decision cut-point for each node of a Decision Tree is determined such that the cost function – e.g. mean squared error (MSE) between the predicted output and the true value – is minimized. Decision Trees are formed deterministically and typically employ a greedy training algorithm, where the optimum split is searched for at each decision node. While classical Decision Trees are versatile machine learning algorithms and capable of fitting complex datasets, they can suffer from overfitting the training data.

Random Forests are another class of tree-based algorithms developed in an effort to reduce model variance found with regular decision trees.[147] A Random Forest is an ensemble of Decision Trees that grows its trees by searching for the best feature among a random subset of features. Random Forests are a bagging technique, where a collection of estimators is trained on different random subsets of the training set with replacement. Once all of the estimators are trained, the Random Forest ensemble aggregates all estimator predictions by averaging them. Generally, the random

sampling of the training set results in greater tree diversity, which trades higher bias for lower variance for an improved overall model.

The Extra-Trees algorithm furthers the randomization of Random Forests by also using random thresholds for each feature when growing each random Decision Tree, rather than searching for the best possible thresholds as in regular Decision Trees and Random Forests. This explicit randomization of the cut-points and features combined with ensemble averaging have been shown to reduce variance compared to other weaker randomization schemes found in other algorithms like Random Forests.[148] A useful property of Extra-Trees is that the relative importance of each feature can be measured incidentally. Relative feature importance is computed by averaging the amount each feature contributes to reducing the prediction error over all trees in the forest.

Relative feature importance was calculated in this study by averaging the results from 50 different Extra-Trees, where each Extra-Tree forest consisted of 500 trees. Standard values were used for other parameters within the Extra-Trees regression estimator provided by scikit-learn, which is the open-source machine learning software package utilized for feature selection and model development in this study unless otherwise specified.[149]

Univariate statistical methods using mutual information and linear regression were used to quantify relative feature importance to compare with the Extra-Trees method. Mutual information is a statistic for measuring the dependency between variables. In contrast to linear regression, which only quantify linear relationships between data sets, mutual information can detect either linear or nonlinear

relationships. Specifically, mutual information is defined as the reduction in uncertainty, or entropy, of one random variable due to the knowledge of another random variable.[150] Given two continuous variables X and Y with a joint probability density function $p(x,y)$ and marginal probability density functions $p(x)$ and $p(y)$, the mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$, or

$$I(X;Y) = \iint_{x\in X\ y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy$$

Mutual information values were computed for each feature to measure their dependence relative to the target values.

Linear regression was used to quantify the linear relationship between the feature and target variable space. First, the correlation between each feature set X and the target set Y was computed as

$$\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}$$

where μ and σ represent the mean and standard deviation respectively. The strengths of the linear correlation between each feature and the set of target values were computed and used to measure the relative feature importance in this study.

After relative feature importance was assessed with these three different selection methods, different numbers of features were selected for each set of relative importance rankings. Specifically, subsets of the 5, 10, 25, 50, 100, 175, and 241 (i.e. all features) most important features according to each of the three feature importance methods were selected for training a given machine learning model, resulting in 21 models (three feature selection methods times seven dataset subsets of varying size)

for each class of machine learning algorithm. The impact of the number of selected features and type of feature selection method on model performance was then assessed.

3.2.3. Training Machine Learning Regression Algorithms

This study surveyed and evaluated the performance of four different categories of machine learning algorithms for this regression problem. Below are brief descriptions of each specific algorithm along with the model-specific, tuned hyperparameters determined via cross-validated searches. The interested reader can find more details in the supplementary material regarding the hyperparameter tuning methods used for each model. The tuned hyperparameters of the models discussed in ensuing results are also summarized in the supplementary material for convenience (Table C.2). For each of the 21 combinations of number of features selected and feature selection methods detailed in Chapter 3.2.2, the best predictor was selected from these cross-validated searches and were then refit to the entire training set. This optimized fit was evaluated on the testing set and was used to compare the performance of different learning algorithms and to assess the impact of number of features selected and type of feature selection method on model performance.

3.2.3.1. Linear Regression Models

One of the simplest machine learning models is linear regression. A linear model makes a prediction by computing a weighted sum of the input features, plus a constant bias term. This equation for linear regression model prediction can be expressed as

$$\hat{y} = h_\theta(\boldsymbol{x}) = \theta^T \cdot \boldsymbol{x}$$

where $\hat{y}$ is the predicted value, θ is the model's parameter vector containing the bias term θ_0 and the feature weights θ_1 to θ_n, $\boldsymbol{x}$ is the instance's feature vector (containing x_0

to x_n, with x_0 always equal to 1), and h_θ is the hypothesis (i.e. prediction) function, using the model parameters θ. Training a linear regression model means finding the set of parameters such that the model best fits the training set. This can be done by finding the parameter vector θ that minimizes the MSE, defined as

$$\mathrm{MSE}(\boldsymbol{X},\, h_\theta) = \left(\frac{1}{m}\right)\sum_{i=1}^{m}\left(h_\theta\left(\boldsymbol{x}^{(i)}\right) - \boldsymbol{y}^{(i)}\right)^2$$

where m is the number of instances in the dataset, $\boldsymbol{x}^{(i)}$ is a vector of all the feature values of the i^{th} instance in the dataset with $\boldsymbol{y}^{(i)}$ being its associated target value (the desired output value for that instance), $\boldsymbol{X}$ is a matrix containing all the feature values of all instances in the dataset, and h_θ is the system's prediction or hypothesis function parametrized by θ.

A strategy to reduce the likelihood of overfitting in linear regression and other machine learning models is to regularize, or constrain, the weights. In addition to unregularized linear regression, three types of regularized linear models were explored in this study: Ridge Regression, Lasso Regression, and Elastic Net. Ridge Regression adds an ℓ_2-norm regularization term to the cost function, equal to $\frac{\alpha}{2}\sum_{i=1}^{n}\theta_i^2$, where α is a hyperparameter to control the amount of regularization.[151] Similarly, Lasso Regression instead adds an ℓ_1-norm regularization term to the cost function, equal to $\alpha\sum_{i=1}^{n}|\theta_i|$.[152] Both algorithms include the hyperparameter α to control the amount of regularization. Elastic Net utilizes both regularization terms found in Ridge Regression and Lasso regression and allows control of each type of regularization through a mixing parameter r; Elastic Net is equivalent to Ridge Regression when $r = 0$ and equivalent to Lasso Regression when $r = 1$. Therefore, since the hyperparameter optimization of the Elastic

Net model included both Lasso and Ridge Regression models, only Elastic Net results are shown here for simplicity. The Elastic Net hyperparameters α and r were tuned to 0.594 and 1, respectively, via a 10-fold cross-validated grid search.

3.2.3.2. Support Vector Machine

SVMs are powerful and versatile models capable of performing linear or nonlinear classification and regression. SVMs implement kernels to map input features to higher dimensional spaces to facilitate nonlinear predictive modeling. The SVM regression algorithm originally proposed by Drucker *et al*. depends on only a subset of the training set such that if the predicted value is within a certain tolerance ε, the loss is zero, while if the predicted point is outside this ε-tube, the loss is the magnitude of the difference between the predicted value and the radius ε of the tube.[153] Therefore, SVMs are ε-insensitive, where adding training instances within the ε margin does not affect model predictions.

The linear, polynomial, Gaussian Radial Basis Function (RBF), and sigmoid kernels were tested in this study. Each are defined as

$$\begin{aligned}
&\text{Linear:} && K(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^T \cdot \boldsymbol{b} \\
&\text{Polynomial:} && K(\boldsymbol{a}, \boldsymbol{b}) = (\gamma \boldsymbol{a}^{\boldsymbol{T}} \cdot \boldsymbol{b} + r)^d \\
&\text{Gaussian RBF:} && K(\boldsymbol{a}, \boldsymbol{b}) = \exp\left(-\gamma \|\boldsymbol{a} - \boldsymbol{b}\|^2\right) \\
&\text{Sigmoid:} && K(\boldsymbol{a}, \boldsymbol{b}) = \tanh(\gamma \boldsymbol{a}^{\boldsymbol{T}} \cdot \boldsymbol{b} + r)
\end{aligned}$$

where γ and r are kernel hyperparameters. Tuned hyperparameters for SVMs were optimized via a 5-fold cross-validated randomized search with 250 iterations (Table 3.2).

Table 3.2. Selected hyperparameter values for the SVM model.

SVM Parameter	Selected Value
Kernel	Gaussian RBF
C	6.407
ε	0.094

(table cont'd)

SVM Parameter	Selected Value
γ (for Polynomial, Gaussian RBF, and Sigmoid kernels)	$\frac{1}{n_{features} \times \sigma^2}$

Note: σ^2 is the variance of the given feature distribution

3.2.3.3. Tree-Based Regression

3.2.3.3.1 Decision Trees

As mentioned previously in Chapter 3.2.2, Decision Trees are flexible algorithms capable of classification or regression tasks and can even predict multiple outputs. While they can fit complex datasets, Decision Trees are nonparametric models that have a strong tendency to overfit the training data if left unconstrained or not regularized appropriately. Regularization parameters for the tuned Decision Tree model in this study were determined to be a maximum tree depth of 3, a minimum number of samples required at each node of 17, with the mean absolute error (MAE) loss function used to measure the quality of a split. Standard values were used for remaining model parameters during training. The Classification and Regression Tree (CART) algorithm introduced by Breiman *et al.* was used to train the decision tree models in this work.[154]

3.2.3.3.2 Random Forests

Random Forests are a class of machine learning algorithm consisting of an ensemble of Decision Trees which are grown by searching for the best feature among a random subset of the feature space, instead of searching for the best among all features as in normal Decision Trees. As indicated in Chapter 3.2.2, this increases bias and decreases variance to generally yield a better model overall. The same tree-growing hyperparameters and their associated ranges of values given in Chapter 3.2.3.3.1 were similarly optimized for Random Forests. The optimized Random Forest

model had a maximum tree depth of 12, a minimum number of samples required at each node of 4, and an MSE loss function. Additionally, the number of trees in the optimized forest was determined to be 124.

3.2.3.3.3 AdaBoost

Boosting algorithms are those incorporating and combining an ensemble of weak models into a stronger composite model. This process is generally sequential, where subsequent models are trained based on the errors of the preceding model. AdaBoost is a popular boosting learning algorithm that begins by fitting a base regressor to the training data and then iteratively fits copies of the regressor on the same dataset while adjusting the relative weights of training instances associated with the largest errors.[155]

A standard Decision Tree with a maximum depth of 5 levels was the base regression model used for optimizing and training the AdaBoost models in this study. The optimized AdaBoost-specific hyperparameters of describing the maximum number of estimators at which boosting was terminated and learning rate, which scales the contribution of each regressor, were determined to be 91 and 1.311 respectively.

3.2.3.3.4 Gradient Boosting

Gradient Boosting is another popular boosting algorithm that differs from AdaBoost, which adjust instance weights at every iteration, by fitting new predictors to the residual errors made by the previous predictor.[156] This allows for optimization of arbitrary differentiable loss functions, where each predictor is fit on the negative gradient of the given loss function at each iteration.

Like with the AdaBoost models, Decision Trees were used as the base regression model for training and optimizing Gradient Boosting models. The determined hyperparameter values for the optimized Gradient Boosting model were a maximum

number of estimators of 616, a learning rate of 0.007, a fraction of training samples to be used for fitting the individual base predictors of 0.444, a maximum tree depth of 14, and a minimum number of samples required to split an internal node of 6.

3.2.3.4. Artificial Neural Network

Artificial Neural Networks (ANNs) are popular machine learning algorithms that take inspiration from the biological architecture found in the brain. ANNs are commonly used models due to their robustness and scalability, which makes them useful for large and complex tasks like mastering the game of Go.[157]

Hyperparameter tuning in neural networks is a challenging problem given the potentially large parameter space. Although the number of tunable parameters gives ANNs their flexibility, the topology of even a simple network can be altered by parameters such as the number of hidden layers, the number of neurons per layer, the type of activation function used in each layer, among others.

Neural networks in this study were developed using Keras.[158] The set of hyperparameters yielding the best validation MAE (20% of training set) was chosen as optimal and were used for final training and testing. Table 3.3 lists the optimized hyperparameter values for the ANN model.

Table 3.3. Selected hyperparameters for the ANN model.

Talos-specific hyperparameter	Tuned Parameters
Learning Rate	0.280
Number of neurons in first layer	27
Batch Size	94
Number of hidden layers	3
Topology Shape†	'funnel'
Epochs	275
Dropout	0

(table cont'd)

Talos-specific hyperparameter	Tuned Parameters
Optimizer	Nadam
Losses	MAE
Hidden layers activation function	Linear
Output layer activation function	relu

†Topology shapes are package-specific names where 'brick' assigns the same number of neurons in each layer, 'triangle' decreases the number of neurons by a constant number with each layer so that the shape resembles a triangle, and 'funnel' decreases the number of neurons by floor of the difference between the specified number of neurons in the first layer and last layer divided by the number of desired hidden layers, resulting in a funnel shape.

3.2.4. Overall Model Comparison

Once optimal hyperparameters were selected via cross-validated searches as described in Chapter 3.2.3, models for each type of algorithm, feature selection method, and number of selected features were evaluated on the separate testing set that was unseen during training. Models with the best MAE within each class of learning algorithm were used to compare overall performance. It is important to note MAE is reported as a percentage in this work because the variable being evaluated (GPR) is expressed as a percentage. Specifically, MAE is defined in this study as:

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$$

where $\hat{y}_i$ is the predicted GPR and y_i is the true GPR. Since GPRs are the percentages of points passing the given gamma criteria, the individual absolute errors $|\hat{y}_i - y_i|$ and therefore the MAE of predicted GPRs are expressed as percentages. Further, values of MAE in this work should not be confused with mean absolute percentage errors.

Top performing models were selected to further assess the impact of suspected outliers among the labeled data (i.e. GPRs), which was performed by iteratively removing samples with GPRs outside 1.5 times the interquartile range (25%-75% distribution quartiles). Performance for these top performing models was also

statistically compared using Wilcoxon signed-rank tests (with significance level set at $p = 0.05$) to "random guessing" by fitting the training and testing datasets to Gaussian distributions and then randomly sampling those distributions to obtain 'random' predictions.

3.3. RESULTS

3.3.1. Feature Analysis

Relative feature importance was measured using three feature selection methods: forests of Extra-Trees, mutual information, and linear regression. Table 3.4 shows the relative importance of each feature category – according to the sum of feature importances of each individual feature within each category – for each of these three selection methods. The SAS feature category was most important for each selection method. Generally, feature groups related to aperture size and shape or MLC leaf gaps (e.g. FAOC, LG, etc.) were also important for accurately predicting GPRs.

Regarding the raw features within each feature category, the five most important features using forests of Extra-Trees were (in order of decreasing importance) the lung treatment site, the prostatic fossa treatment site, the mean weighted LG, the mean collimator angle, and maximum SAS at 40 mm. The five most important features with the mutual information feature selection method were the mean weighted FAOC at 25 mm, the SAS at 75 mm, the mean weighted LG, the SAS at 80 mm, and the SAS at 45 mm. Lastly, the five most important features using the linear regression method were the SAS at 50 mm, the SAS at 45 mm, the SAS at 55 mm, the maximum SAS at 45 mm, and the SAS at 60 mm. It is interesting to note the differences in feature importance among the three selection methods. The two most important features using forests of Extra-Trees were lung and prostatic fossa, both of which are treatment sites.

Whereas for the mutual information and univariate linear regression methods, the top features were related to aperture shape and the size of the gaps between opposing MLC leaves. Features were weakly correlated with GPRs, with SAS 50 mm having the Pearson correlation coefficient with the largest magnitude of 0.38 ($p < 0.001$; Figure 3.2). Overall, 113 out of the 241 raw features having significant *p*-values of less than 0.05 and 32 features had correlation coefficient magnitudes of at least 0.3.

Table 3.4. Rankings of relative importance of feature categories according to the sum of all raw features within each classification for each feature analysis method.

Summed Feature Rank	Extra-Trees	Mutual Information	Linear Regression
1	SAS	SAS	SAS
2	FAOC	FAOC	LG
3	Site	PI	FAOC
4	LT	LG	AA
5	AP	LM	Site
6	LM	LT	JT
7	PI	AA	EM
8	LG	AP	FFF
9	JM	Site	Dose Rate
10	Machine	JP	MU Factor
11	AA	JM	AP
12	JT	JT	JM
13	Collimator Angle	MU Factor	PI
14	JP	EM	LM
15	Arc Length	Machine	Machine
16	MU Factor	CAS	LT
17	FFF	FFF	JP
18	Dose Rate	Dose Rate	Collimator Angle
19	CAS	Collimator Angle	Arc Length
20	MCS	MCS	Number of Arcs
21	EM	PM	MCS
22	PM	Arc Length	CAS
23	Number of Arcs	Number of Arcs	PM

Figure 3.3 shows the distribution of dataset GPRs as a function of treatment site. Lung had the highest mean GPR overall of 91.05% with a standard deviation of 6.67%,

while prostatic fossa had the lowest mean GPR of 83.35% with a standard deviation of 6.29%. Additionally, lung had the most samples with GPRs less than or equal to 80% (n = 10) while prostatic fossa had the highest proportion of samples that had GPRs less than or equal to 80% (30%).

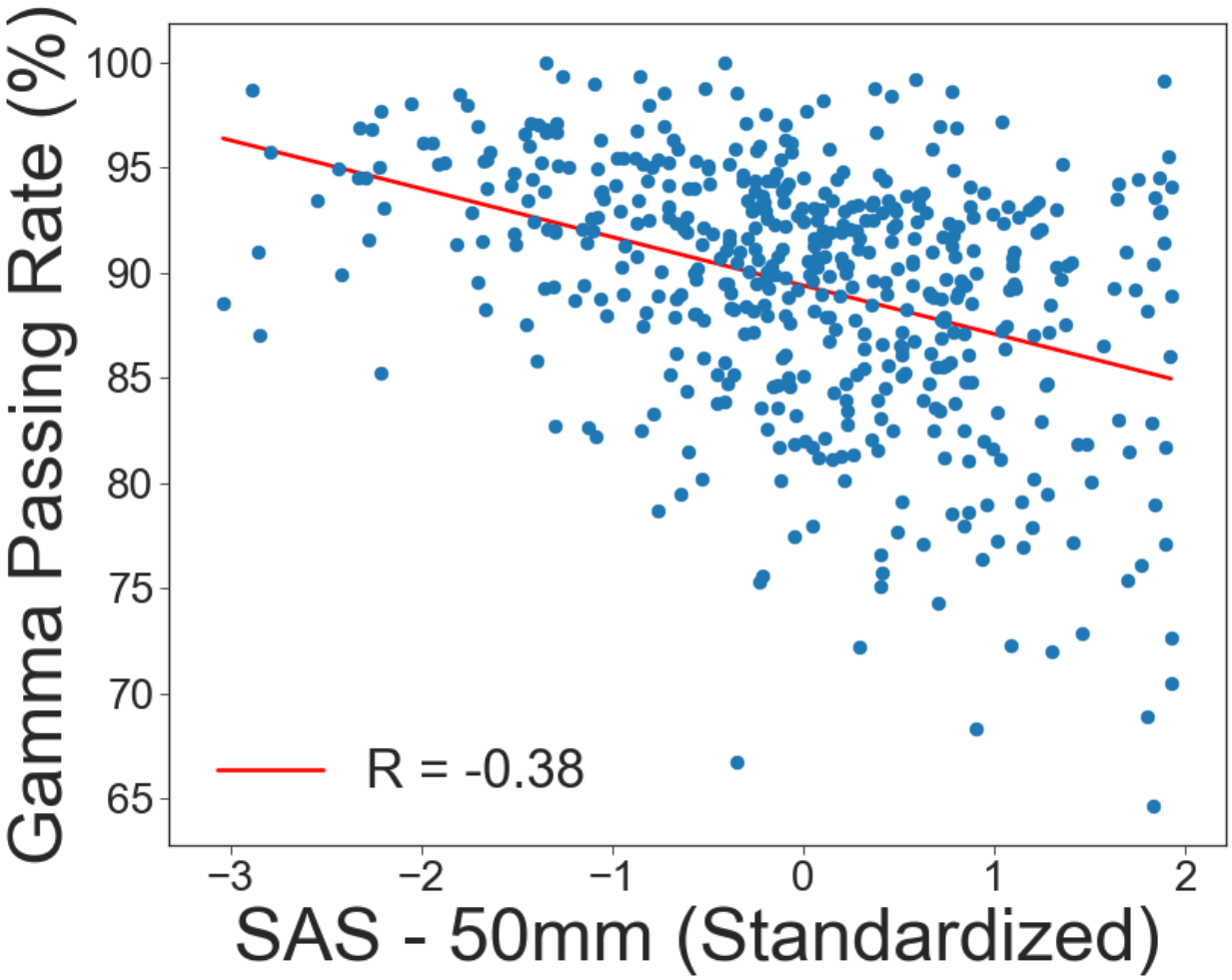


Figure 3.2. Scatter plot of GPR vs. SAS – 50 mm over the dataset along with the Pearson correlation coefficient (R). Note the feature axis has been standardized.

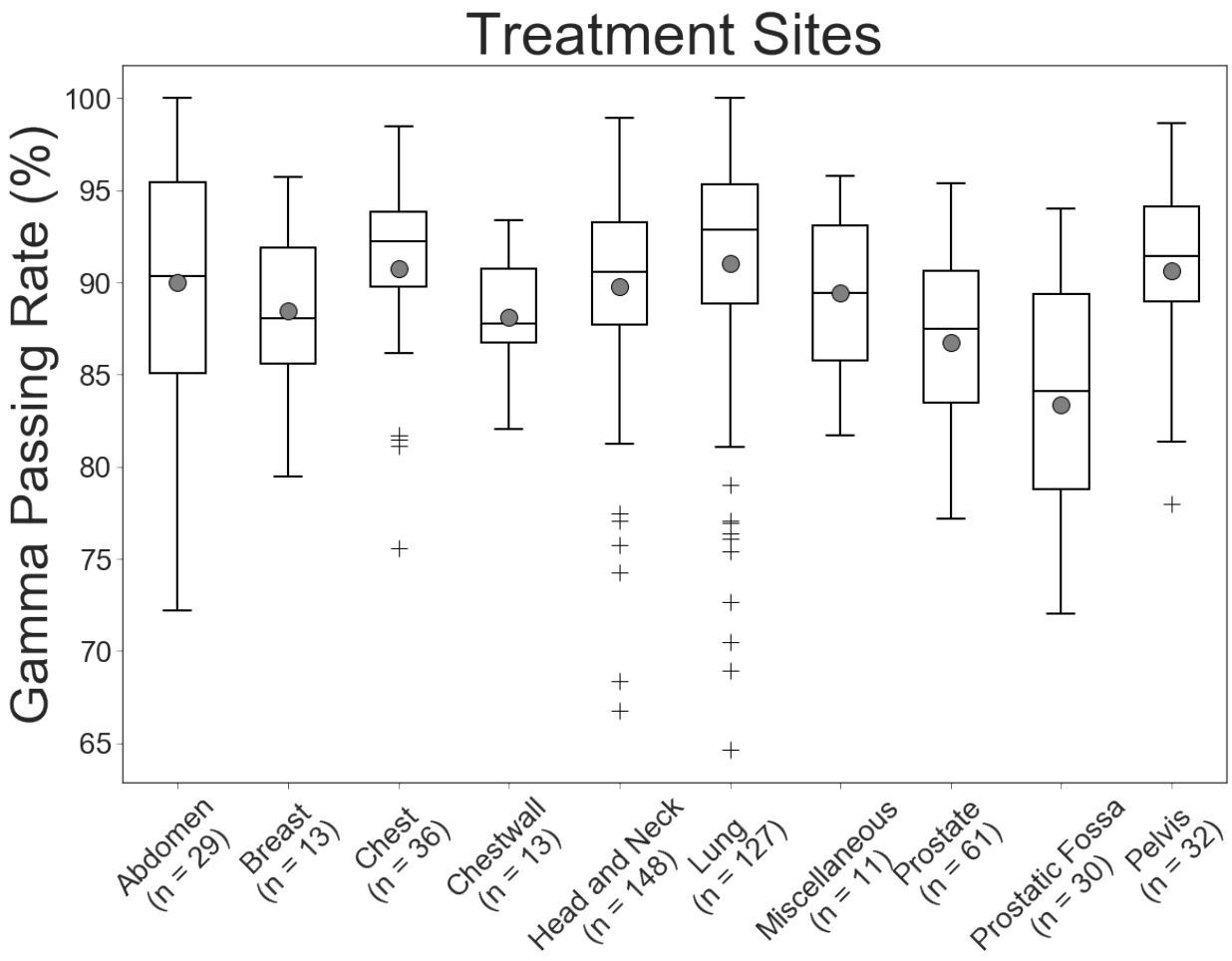


Figure 3.3. Distribution breakdown of dataset with respect to treatment site. Here, the grey circles indicate the distribution mean and crosshairs indicate values lying outside 1.5 times the interquartile (25%-75%) range.

3.3.2. Optimized Model Performance

After relative feature importance was quantified using the three selection methods, different machine learning algorithms were trained using subsets of varying numbered selected features as determined by each method. Hyperparameters were optimized for each learning algorithm and feature set pair using a cross-validation searching method as described in Chapter 3.2.3. Each model was trained and subsequently evaluated on the testing set, after which the models within each class of algorithm with the lowest test MAE was selected for comparison (Table 3.5). Optimized

hyperparameters for each model listed in Table 3.5 can be found in the supplemental material.

Table 3.5. Testing error for best performing model within each class of learning algorithm with associated feature selection method and number of features. Mean and standard deviations of testing error from cross-validation is also included. Note that all models underwent 10-fold cross-validation, except for SVM, for which 5 folds were used.

Algorithm	Cross-Validation MAE ($\mu \pm \sigma$; %)	Test MAE (%)	Feature Selection Method	Number Of Features
Linear Regression	4.20 ± 0.57	4.29	ET	5
Elastic Net	4.34 ± 0.69	4.17	LR	50
SVM	3.96 ± 0.42	3.85	LR	100
Decision Tree	4.37 ± 0.72	4.14	LR	10
Random Forest	3.91 ± 0.60	3.98	ET	100
AdaBoost	3.99 ± 0.51	3.98	ET	50
Gradient Boosting	4.06 ± 0.52	3.94	ET	50
ANN	4.24 ± 0.69	4.01	ET	50

Abbreviations: LR = Linear Regression; ET = Extra-Trees

Mean cross-validation MAE was consistent with test MAE for each learning algorithm, indicating the models generalized as expected from training to testing. The best (i.e. lowest) MAE on the testing set was the SVM model (3.85%) followed by the Gradient Boosting model (3.94%). Random Forest and AdaBoost models also achieved test MAEs of less than 4%. In addition, the optimal number of features for each model seemed to be between 50 and 100 of the most important features according to ranking with Extra-Trees and linear regression feature selection methods.

3.3.3. Top-Performing Models

The top two performing models, SVM and Gradient Boosting, were selected for further inspection and analysis. Figure 3.4 shows the behavior of cross-validation error as a function of feature selection method and number of features selected for each model. Generally, the cross-validation error decreased as the number of selected

features increased for both models but plateaued after 100 selected features. SVM models using the Extra-Trees and linear regression feature selection methods resulted in lower cross-validation error than the mutual information method when using higher numbers of features. Also, cross-validation error for both models using the Extra-Trees feature selection varied less with the number of selected features compared to mutual information and linear regression selection methods.

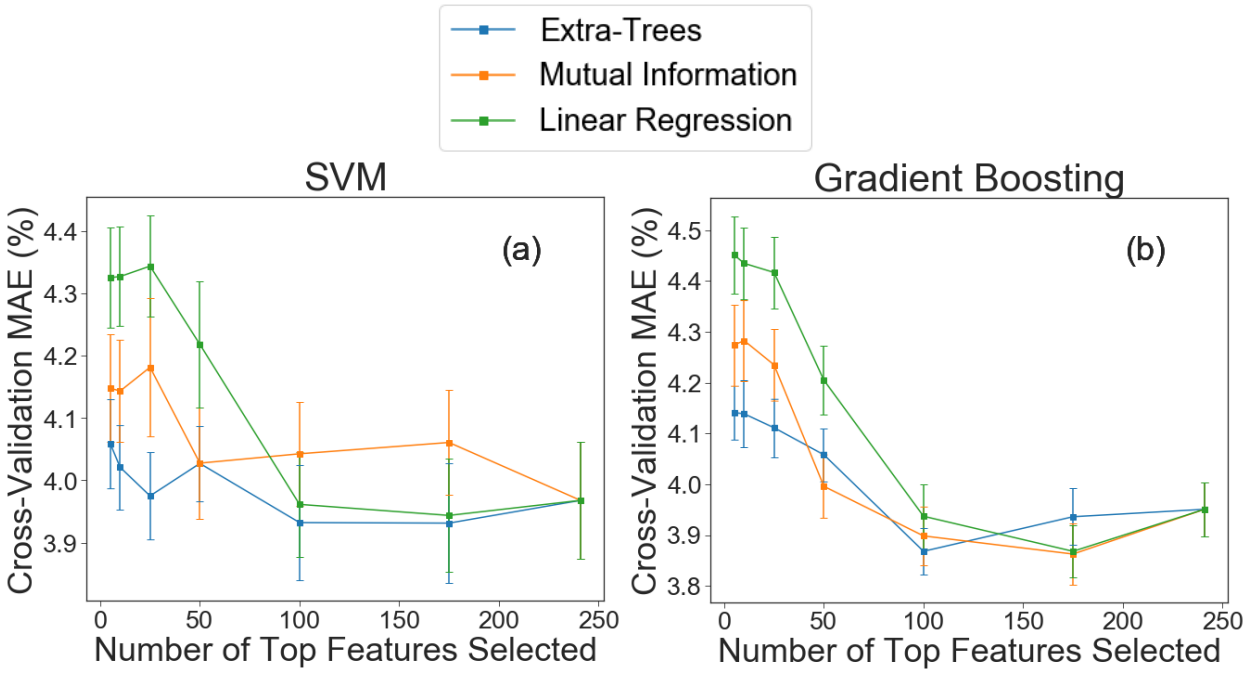


Figure 3.4. Cross-validation MAE of SVM (a) and Gradient Boosting (b) models as a function of feature selection method and number of selected features. The mean cross-validation error and standard deviation is plotted. Note that 5 folds were used for SVM models and 10 folds were used for Gradient Boosting models.

The impact of the number of training samples on training and cross-validation error was also assessed for these models. Figure 3.5 shows these learning curves (which plot training and cross-validation error against number of training samples) for the optimized models. The SVM learning curve shows the validation error gradually decreasing with increasing training samples, while the training error stabilizes after

about 75 training samples. The Gradient Boosting learning curve shows training and validation error to also be decreasing with the size of the training set. However, the difference in the training and validation errors for the Gradient Boosting model highlight the tendency of Decision Trees to overfit the training data compared to the SVM model.

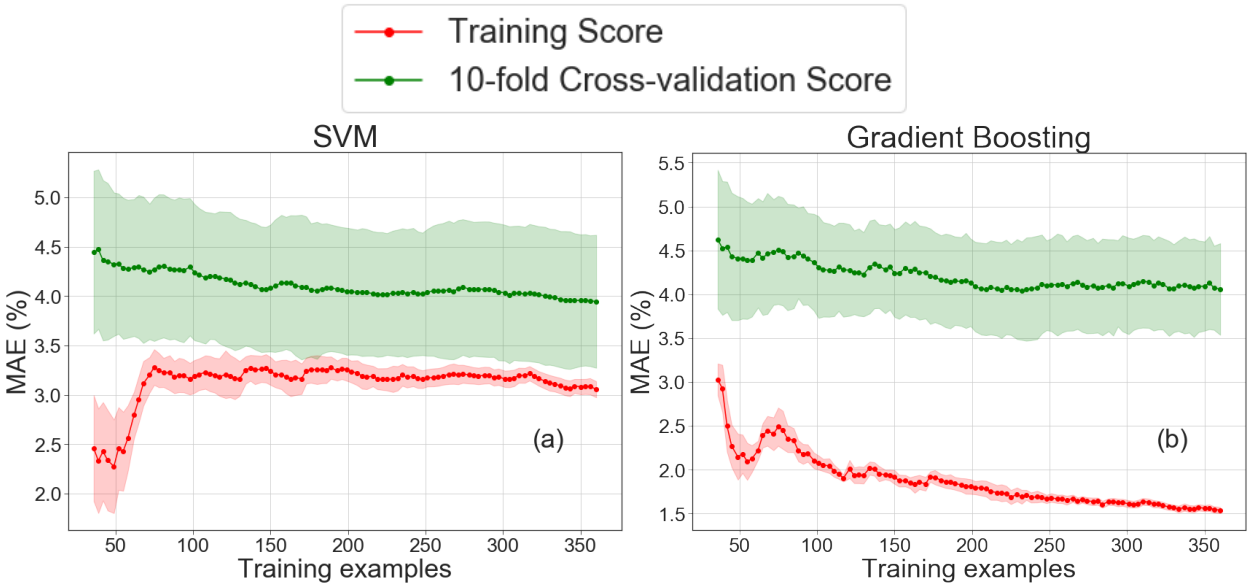


Figure 3.5. Learning curves, which plot error as a function of the number of training samples, for the optimized SVM (a) and Gradient Boosting (b) models. Note that 10-fold cross-validation was used for both models to construct these curves. Average values resulting from cross-validation are shown via markers, with the associated standard deviation given via the shaded regions.

Composite testing error results from 5-fold cross-validation over the entire dataset for the SVM and Gradient Boosting models are shown in Figure 3.6 and Figure 3.7 respectively. The SVM model resulted in 3.75 ± 0.29% (mean ± standard deviation) testing MAE while the Gradient Boosting model resulted in 3.81 ± 0.22% MAE. This represents a significant 41.1% ($p < 0.001$) and 40.6% ($p = 0.02$) average improvement respectively over "random guessing," which had an MAE of 6.41 ± 4.98%. The SVM model predicted 51.2% and 74.4% of the testing samples to within 3% and 5% error, respectively. Whereas the Gradient Boosting model predicted 47.2% and 71.2% of the

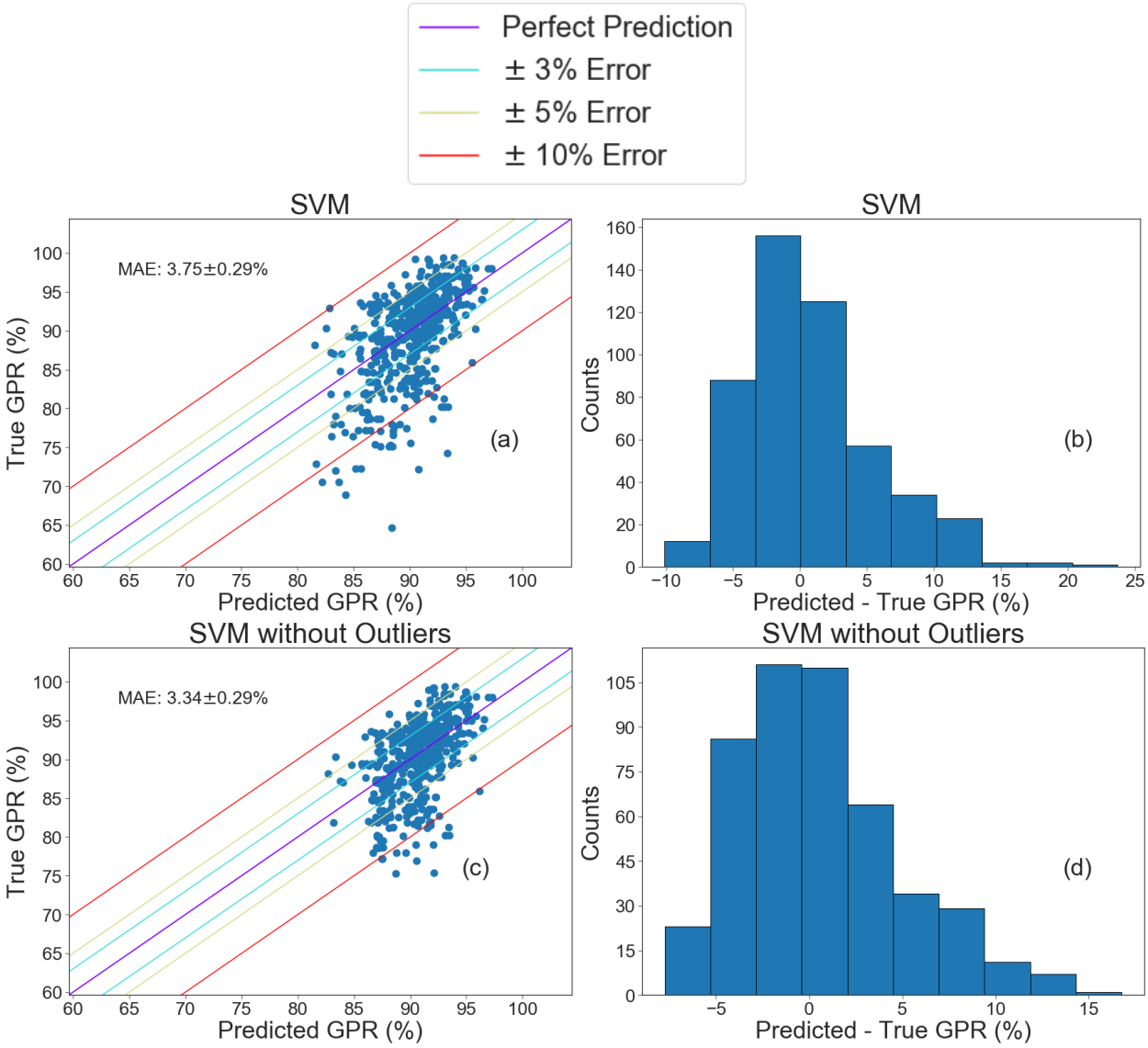


Figure 3.6. 5-fold cross-validation testing performance for the optimized SVM model. (a) and (c) plot the true GPR values against the predicted GPR values from the model when trained and tested with the full dataset and when trained and tested with suspected outliers removed from the dataset, respectively. (b) and (d) are associated histograms of the differences between the true and predicted values.

testing samples to within 3% and 5% error, respectively. The maximum errors were 23.7% and 17.8% for the SVM and Gradient Boosting models respectively. When suspected outliers were removed from the dataset, the average testing MAE improved by 0.41% and 0.47% for the SVM and Gradient Boosting models respectively. Most noticeably, the percentages of instances with errors less than 10% were improved from

94.2% to 96.64% and from 95.8% to 98.32% for the SVM and Gradient Boosting model respectively when removing these suspected outliers.

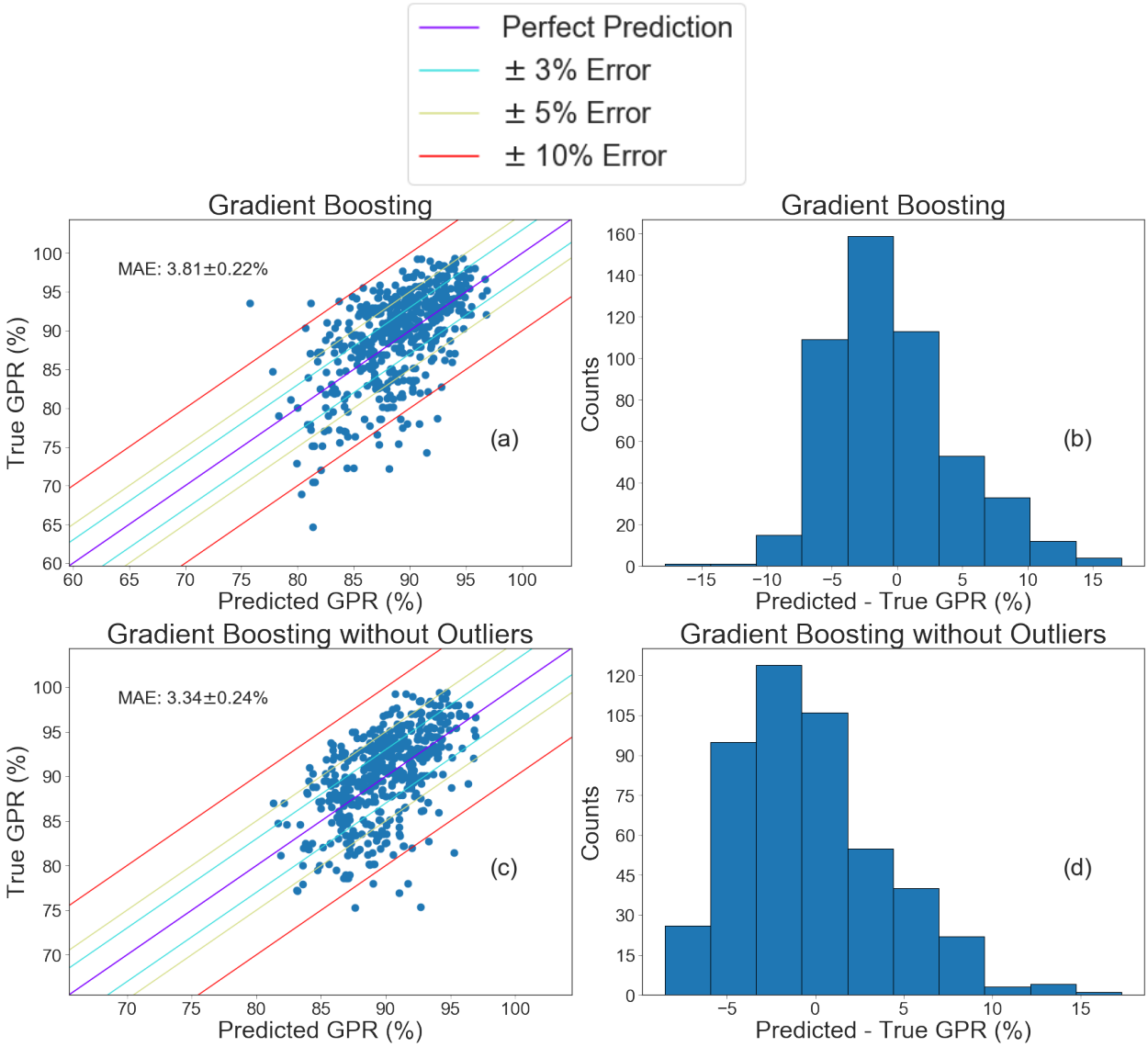


Figure 3.7. 5-fold cross-validation testing performance for the optimized Gradient Boosting model. (a) and (c) plot the true GPR values against the predicted GPR values from the model when trained and tested with the full dataset and when trained and tested with suspected outliers removed from the dataset, respectively. (b) and (d) are associated histograms of the differences between the true and predicted values.

3.4. DISCUSSION

This study collected a large retrospective dataset of VMAT plan and QA data in order to investigate the feasibility of developing machine learning models for predicting GPRs from an array of treatment plan parameters and complexity metrics. The selected features were found to be weakly correlated with the target variable, which resulted in the best-performing machine learning model (SVM) yielding an average cross-validated test MAE of 3.75%. These results may reveal potential limitations within the present dataset when compared to model performance in existing literature such as in Valdes *et al*., which reported being able to predict GPRs for fixed-gantry IMRT plans to within 3%.[95] However, each machine learning model investigated in this study achieved improved testing MAE compared to "random guessing" and was able to capture the overall relationship between the independent and dependent variables present in the available dataset.

Several previous studies have showed promising results for predicting GPRs for "fixed-gantry" IMRT plans using machine learning and deep learning.[95-99] Fewer studies have been published concerning VMAT QA. Granville *et al*. recently applied an SVM model for classifying results of VMAT QA measurements performed with biplanar diode arrays into 'hot', 'cold', or 'normal' median dose difference categories.[122] Ono *et al*. used regression tree analysis, multiple regression analysis, and ANNs to predict GPR measurements using a helical diode array based on plan complexity metrics.[101] Novel aspects of this study compared to these previous works are the comprehensive comparison of different standard machine learning models, the investigation of feature importance using three different selection methods, and the assessment of the relationship between both the number of features and the type of selection method used

during training and the resulting model performance. The SVM model was found to be the best-performing model and the machine learning algorithms generally performed better with fewer number of features selected with either forests of Extra-Trees or linear regression. Additional strengths of this study are the collection of a large dataset along with inspecting a large array of plan complexity metrics and features. Although the features were found to be weakly correlated with the target variable overall, more than 100 features were significantly ($p < 0.05$) correlated. The feature selection methods used were able to improve model performance compared to models trained with all available features as shown in Table 3.5, which indicates reducing the number of selected features provided the best performance across all models, in addition to reducing model dimensionality. Also, the specific combination of technologies (Pinnacle TPS, Elekta LINACs, and MapCHECK2 diode-array) used in this study is unique relative to previous studies using machine learning to predict GPR measurements to our knowledge.

An interesting result from analyzing the relative feature importance in accurately predicting GPRs was that the lung and prostatic fossa treatment sites were the two most important features according to the forests of Extra-Trees selection method. Unlike previous studies, which have not included treatment site as an input feature, this result could indicate the need for site-specific models. This is similar to the results Valdes *et al*. observed, where one treatment machine was found to have different profile characteristics affecting model performance that suggested machine-specific models may lead to more robust predictions. Further research and a larger number of samples

within each treatment site are warranted in the future to evaluate differences in site-specific models.

The present work shares similarities with previous studies such as the number of samples and the plan complexity metrics and parameters included in the feature space. The features selected for this study (which were taken or adapted from the literature listed in Table 3.1) showed weaker correlations with the target GPR values than the correlations found in previous studies for those same features.[112,113,120] However, features based on the aperture size (e.g. SAS) were found to have the strongest linear correlation with GPRs (Figure 3.2), which is consistent with Crowe *et al.*[144] The weaker correlations found in this study could be a result of differences in the underlying data in each work, such as the institution-specific combination of technologies and clinical protocols utilized for treatment planning QA.

The differences in testing errors between this study and previous results may be due to the spread in the distribution of GPRs in the present dataset. The majority of GPR distributions in these previously mentioned studies were heavily concentrated towards the 100% GPR boundary, which potentially clouds the true relationship between features and dose differences. In contrast, the target distribution used in this study had an average value of 89.39% and a standard deviation of 6.01%. The percent dose-difference/distance-to-agreement gamma index criteria of 3% and 3 mm with local normalization was selected for this study to enable comparison with previous studies and to obtain a target distribution with meaningful information about the underlying differences between planned and measured dose distributions. It is also important to note each sample used in the study passed the clinical QA protocol at our institution.

The GPRs reported here were computed without incorporating measurement uncertainty (which is typically turned on for clinical evaluation) associated with the device due to factors such as, temperature change, accelerator output fluctuation, array calibration accuracy, and electronic measurement precision.

Training time was not considered for the purposes of this study but is still important to contemplate prior to clinical implementation. Training times qualitatively varied depending on the type of machine learning algorithm, number of features used, and the number of iterations executed for hyperparameter optimization searches. However, for each of the machine learning models in this study, once the model has been trained and has had its hyperparameters tuned, real-time predictions can be given. Therefore, predictions can be quickly provided during a clinical scenario given the necessary input features for the model.

It is possible the correlation between plan complexity and GRPs in this study was also limited by the 1 cm minimum leaf gap constraint used at our institution. Additionally, all but four samples in the dataset utilized four degrees for control point spacing during VMAT plan optimization. Using a finer gantry spacing resolution could result in more accurate dose computation, although four degrees has been shown to provide an optimal balance between plan quality and complexity.[113,135] Further, the specific combination of technologies and clinical techniques – such as the radiation-delivery machine, the TPS software and optimization settings used to design the clinical radiation treatment plans, the measurement device and analysis technique used for performing QA, etc. – used in this study could also have led to a dataset with weaker relationships between plan features and delivery errors. For example, the 1 cm

minimum leaf gap constraint likely restricted the possible range of complexity for the plans in this study's dataset, which could have led to weaker correlations with QA outcomes when compared to studies with plans with a looser leaf gap constraint. Differences in these types of underlying factors that characterize the feature-target relationship possibly prevented the machine learning algorithms surveyed in this work to achieve error rates as low as previously reported in IMRT and VMAT QA studies. However, each machine learning model was able to achieve a minimum of 37% improvement over "random guessing," with the top-performing SVM model improving by 41.1%. Therefore, it is important to note that results from the present and previous studies are specific to the particular dataset used, which warrants future research investigating how the relationship between complexity features and QA outcomes behaves as a function of varying delivery, planning, and QA parameters.

Previous investigators have noted the numerous clinical benefits of models for predicting VMAT QA measurements. Most notably, a machine learning model could identify a treatment plan that is predicted to present unacceptable dose delivery errors before measurement, allowing the plan to be modified beforehand to save time. Alternatively, a machine learning model for predicting GPRs for VMAT plans could be inserted into the planning stage to provide QA-based information to the optimizer and planner in real-time. Predicted GPR output from a machine learning model could be added to the optimization cost function to penalize search solutions that reduce the predicted GPR, which would result in a plan guided by both plan quality (e.g. dose-based) and delivery accuracy (e.g. QA-based) endpoints. Our group is currently pursuing this avenue of research. Previous works have also investigated the feasibility

of developing convolutional neural networks to predict GPRs based on fluence maps for fixed-gantry IMRT and were able to achieve comparable performance to machine learning models based on complexity metrics.[97] While the purpose of this work was specifically to perform a systematic evaluation of different machine learning models for predicting VMAT QA outcomes based on complexity features, a logical next progression would be to develop a convolutional neural network for predicting VMAT QA outcomes to mirror the progression of previous literature on predictive models for fixed-gantry IMRT QA.

3.5. CONCLUSION

Models predicting VMAT QA outcomes can help improve clinical efficiency by highlighting treatment plans likely to fail QA prior to measurement. This work is among the first to investigate and compare several machine learning algorithms for predicting VMAT QA measurements using the specific planning and measurement technologies at our institution. Model features were based on treatment plan complexity metrics and parameters and their relative importance in accurately predicting GPRs was assessed via forests of Extra-Trees, mutual information, and linear regression. Features were found to be weakly correlated with GPRs, resulting in a test MAE of 3.75% for the best performing model (SVM). While previous studies have shown the ability of machine learning models to predict QA outcomes with a high degree of accuracy, the results of this study show model performance may be limited by characteristics of the underlying data, particularly the unique and specific combination of technologies and clinical parameters used to generate treatment plans and perform quality assurance. Further, machine learning models were developed and shown to be significantly better than "random guessing," but the results from this study indicate feature analysis and

selection should be performed when establishing a machine learning model for predicting QA measurements.

4. USE OF MACHINE LEARNING ALGORITHM DURING OPTIMIZATION TO IMPROVE PATIENT-SPECIFIC QUALITY ASSURANCE IN VOLUMETRIC MODULATED ARC THERAPY PLANS

4.1. PURPOSE

The purpose of this study was to evaluate the feasibility of a QA-based planning tool, whereby QA outcomes for VMAT treatment plans are directly optimized using machine learning without substantially degrading the dosimetric quality of the original plan.

4.2. MATERIALS AND METHODS

4.2.1. Description of Algorithm

An in-house algorithm was developed to optimize predicted QA endpoints by modifying existing mechanical parameters determined from Chapter 3 to be important for predicting the deliverability of a radiotherapy treatment plan. This was accomplished through utilization of the machine learning algorithm described in Chapter 3 to predict the QA outcomes based on complexity features of VMAT treatment plans.

The proposed algorithm – existing independently from a commercial TPS – takes an existing VMAT treatment plan file as input with an initial predicted GPR, and returns a modified version of the original VMAT treatment plan with mechanical parameters adjusted to result in an increased predicted GPR (see Figure 4.1). Before this QA optimization, the initial predicted GPR is computed for the original plan file by the previously mentioned machine learning model. This initial prediction serves as a reference point for the performance of subsequent plan modifications relative to changes in predicted QA outcome. After this initial QA outcome prediction, the algorithm selects specific features of the plan that the machine learning model deems to be the

most important or influential to plan deliverability. These selected plan features are then modified, using logic informed by the feature analysis results detailed in Chapter 3.3.1, before the machine learning model assesses the new predicted QA outcome as a result of the modifications to the mechanical parameters. This process is iterated over a given number of optimization iterations, after which the algorithm returns the plan file with the best predicted QA outcome.

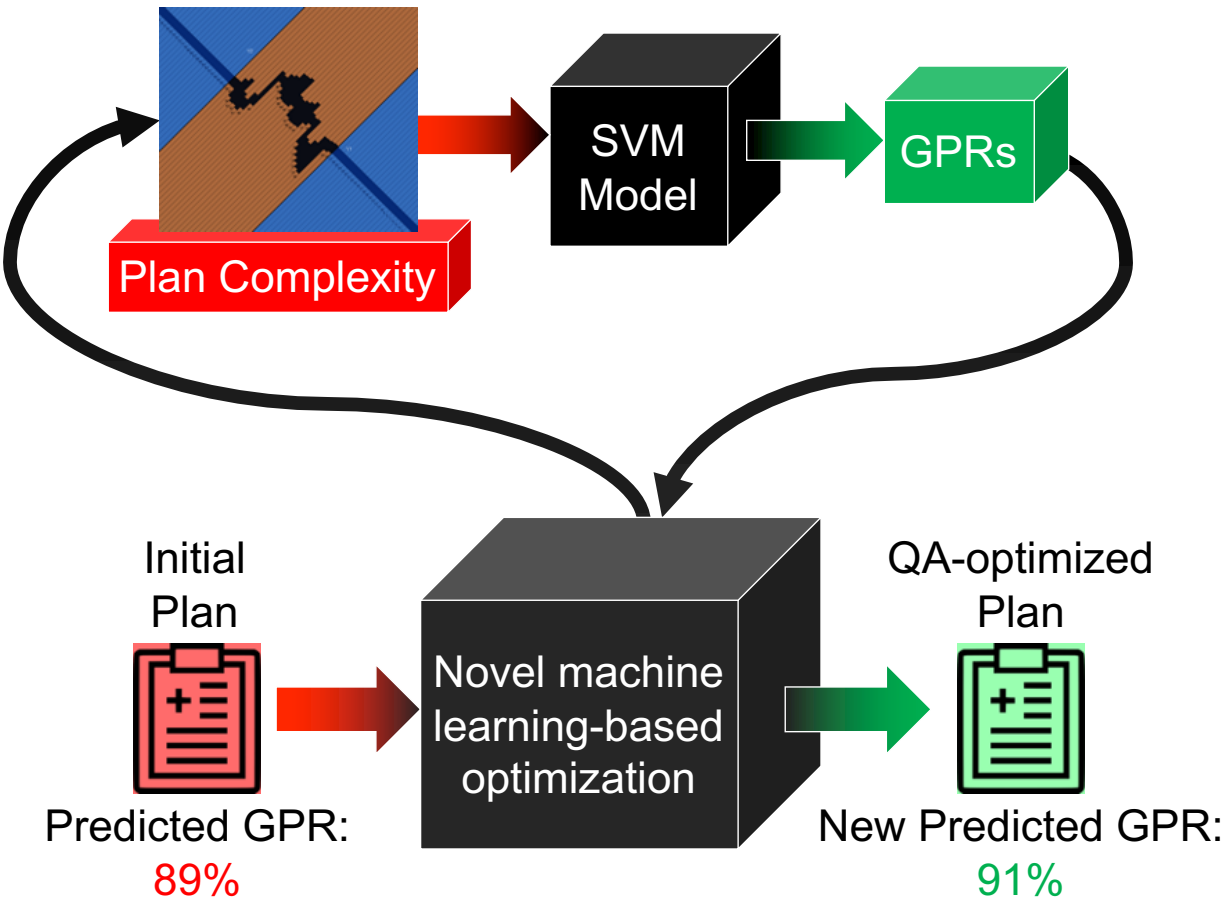


Figure 4.1. Conceptual overview of the proposed QA-based treatment planning optimization technique. First, an inversely optimized plan is taken as input with an initial predicted GPR. Then mechanical features identified by our machine learning model are randomly displaced in order to produce a new GPR prediction. This process is iterated to produce the largest positive GPR change.

More specifically, the algorithm was designed using the Python programming language. Given an initial DICOM RT-Plan file of a VMAT treatment plan, the algorithm begins by extracting the necessary complexity features (see Chapter 3.2.1.1 and Table 3.1) as required by the machine learning algorithm to predict the QA outcome of the original plan. The machine learning model implemented in Chapter 3 was a support vector machine (SVM), developed on a training and testing set of 400 and 100 previous VMAT treatment plans, respectively. The SVM model was designed to use an array of plan complexity features to predict GPRs of a given VMAT plan using 3%/3mm percent dose-difference and distance-to-agreement gamma criteria with local normalization. The 100 most important complexity features were selected using linear regression tests out of a total of 241 raw plan features, belonging to 23 general plan parameter categories (Table 3.1). As mentioned in Chapter 3.3.1, the feature category selected as the most important feature group for accurately predicting GPRs was the small aperture score, or SAS, for all three selection methods. Further, the five most important features selected according to the linear regression method were all based on the SAS, defined as the fraction of open MLC leaf gaps less than a given distance. As a result, although features were found to be weakly correlated with GPRs, the SAS feature at 50 mm had the strongest correlation with GPR (Figure 3.2). The model hyperparameters – including the kernel function, associated kernel hyperparameters, and ε – were tuned via a 5-fold cross-validated randomized search over 250 iterations (Table 3.2). This SVM model, which was the best performing model out of those surveyed in Chapter 3, achieved a 5-fold cross-validated testing mean absolute accuracy of 3.75 ± 0.29% (mean ± standard deviation).

After the GPR of the initial VMAT plan was predicted using the SVM model, the algorithm selected every open LG less than 50 mm over all control points in every beam within the plan file, as these features were previously determined to be most important for predicting GPRs. Then, each LG was widened by a random amount sampled from a uniform distribution between 0 and a maximum distance specified by the user. Randomized MLC leaf displacements were selected for investigation because their impact on the resulting calculated dose distribution is smaller compared to systematic leaf displacements.[159] In order to avoid having the resulting changes in MLC leaf positions violate the minimum MU per leaf travel distance machine constraint, the algorithm also increased the segment weight so that the minimum MU per leaf travel distance never fell below 0.25 MU/cm. The modified MLC leaf positions and segment MUs were then saved, with which new complexity features were computed and input to the SVM model for evaluating the modified plan's new predicted GPR. The modified MLC positions, segment MUs, and predicted GPRs were recorded after each optimization round. This process would be repeated for a given number of iterations set by the user, after which the algorithm would return a DICOM RT-Plan file with the MLC positions and segment MUs that yielded the highest predicted GPR. Since this QA-based optimization occurred external to a commercial TPS, changes in the resulting dose distribution of these QA-optimized plans were unable to be assessed until re-importing the plan and computing the dose. As mentioned previously however, small random MLC leaf displacements were utilized specifically to minimize meaningful changes to the subsequent dose distribution.

4.2.2. Feasibility Assessment

The feasibility of this in-house, QA-based optimization algorithm was evaluated on a set of 13 treatment plans of previously treated prostate patients. Each plan was designed with two co-planar 6 MV VMAT arcs with the guidance from a previously established in-house KBP system.[103,125,160] This KBP implementation predicts the lowest achievable bladder and rectum dose-volumes for new patients among anatomically similar previous patients available in a database, which has been shown to lead to significant reductions in the mean dose to the bladder and rectum compared to reference clinical plans while maintaining clinically acceptable dose to the target.[103,125] However, significant increases in complexity and diminished QA outcomes were also observed for these plans.[160] Therefore, plans designed with this KBP method were selected for this study to assess the extent to which their predicted deliverability could be improved via this algorithm without compromising their dosimetric quality.

The DICOM RT-Plan file for each of the 13 KBP plans were exported from a commercially available TPS. Each plan was modified with the previously described QA-based optimization algorithm (see Chapter 4.2.1) with varying maximum LG displacement and number of iteration settings. Specifically, maximum LG displacements of 1, 3, and 5 mm were used with both 25 and 1000 optimization iterations. This resulted in a total of 6 QA-optimized DICOM RT-Plan files for each patient. In order to assess the dosimetric changes resulting from these mechanical parameter modifications, each DICOM file was then imported back into the same TPS in which the original KBP plan was designed. Then, the same dose grid settings (i.e. 4 mm/voxel resolution and grid coordinates) from the original KBP plan were applied to each QA-

optimized plan before the dose was recomputed. Dose distributions of all plans for each patient were scaled so that 95% of the PTV received the full prescribed dose.

Differences in predicted GPRs, plan complexity metrics, dose, and radiobiological metrics were assessed as a function of maximum LG displacement and number of optimization iterations. The specific plan complexity metrics that were investigated in this study were: MU factor, defined as the total planned monitor units normalized by the fractional prescription dose; modulation complexity score (MCS), a metric introduced by McNiven *et al*. for fixed-gantry IMRT and later adapted for VMAT by Masi *et al*. quantifies aperture area variability and leaf sequence variability into a composite value;[113,120] edge metric (EM), a parameter introduced by Younge *et al*. that measures the "edge" in a plan through the ratio of MLC leaf side lengths over aperture areas;[112] mean leaf motion (LM), defined as the average distance an MLC leaf travels per degree of gantry rotation in mm/deg; mean LG in mm; mean aperture area in mm^2; and the small aperture score at 50 mm.

The radiobiological metrics used in this study were equivalent uniform dose (EUD), tumor control probability (TCP) and normal tissue complication probability (NTCP).[161] EUD has been described by Wu *et al.* as the biologically equivalent dose from a uniform distribution that would result in the same cell kill in the volume as the given non-uniform dose distribution.[162] Differing from this linear-quadratic cell survival model, Niemierko introduced a phenomenological model of EUD defined as

$$EUD = \left(\sum_{i=1}\left(v_i {EQD_2}_i^{\,a}\right)\right)^{\frac{1}{a}}$$

that can be used for both tumors and normal tissues, where a is a unitless model parameter that is specific to the normal structure or tumor of interest, and v_i is unitless and represents the i'th partial volume receiving the biologically equivalent physical dose of 2 Gy, EQD_{2_i}.[161,163] Specifically, EQD_{2_i} is defined as

$$EQD_{2_i} = D_i \times \frac{\left({}^{\alpha}/_{\beta} + {}^{D_i}/_{n_f}\right)}{\left({}^{\alpha}/_{\beta} + 2\right)}$$

where n_f and $d_f = {}^{D_i}/_{n_f}$ are the number of fractions and dose per fraction size of the treatment course that the i'th partial volume receives, respectively. ${}^{\alpha}/_{\beta}$ is the tissue-specific linear-quadratic parameter of the organ being exposed.[164] A parametrization of the dose-response characteristics of tissues was proposed to calculate EUD-based TCP and NTCP, where TCP is calculated according to

$$TCP = \frac{1}{1 + \left(\frac{TCD_{50}}{EUD}\right)^{4\gamma_{50}}}$$

The TCD_{50} is the tumor dose needed to control 50% of the tumor assuming a homogenous irradiation, and γ_{50} is a unitless model parameter that is specific to the normal structure or tumor of interest and characterizes the dose-response curve. Similarly, NTCP is computed according to

$$NTCP = \frac{1}{1 + \left(\frac{TD_{50}}{EUD}\right)^{4\gamma_{50}}}$$

where TD_{50} is the tolerance dose for a 50% complication rate at a specific time interval (typically taken to be 5 years in normal tissue tolerance studies) when the whole organ is irradiated homogenously.

TCP for the prostate and NTCPs for the bladder, rectum, and femoral heads were computed for all plans for each patient. Model parameters used in this study were taken from previous literature and are given in Table 4.1.[164]

Table 4.1. Tissue-specific model parameters used to compute EUD-based TCP and NTCP.[164]

Tissue	a	γ_{50}	TCD_{50} (Gy)	TD_{50} (Gy)	α/β (Gy)
Prostate	-10	1	28.34	NA	1.2
Bladder	2	4	NA	80	8.0
Rectum	8.33	4	NA	80	3.9
Femur	4	4	NA	65	0.85

Differences in these deliverability (i.e. predicted GPRs) and dosimetric characteristics of the QA-optimized plans relative to the original KBP plans were statistically compared using a two-sided Wilcoxon signed-rank test with a significance level set to 0.05.

4.3. RESULTS

4.3.1. Changes in Predicted QA Outcomes

The predicted GPRs of plans optimized with 25 iterations increased by an average of 0.30 ± 1.22% (p = 0.42), 1.14 ± 1.25% (p = 0.006), and 1.52 ± 1.27% (p = 0.003) compared to the original plan when using maximum random LG displacements of 1, 3, and 5 mm respectively. Using 1000 optimization iterations, the predicted GPRs of the optimized plans increased by an average of 0.32 ± 1.17% (p = 0.31), 1.18 ± 0.99% (p = 0.004), and 1.57 ± 1.08% (p = 0.002) compared to the original plans when using maximum random LG displacements of 1, 3, and 5 mm, respectively. There were no significant differences in predicted GPRs of QA-optimized plans when using 25 iterations versus using 1000 iterations for each of the three displacement settings (p =

0.92, p = 0.35, and p = 0.92 for 1, 3, and 5 mm respectively). Figure 4.2 and Figure 4.3 show changes in predicted GPRs for each patient using different maximum LG displacements at 25 and 1000 optimization iterations, respectively. The optimization algorithm runtime was about two minutes using 25 iterations for a given plan and optimization settings, whereas 1000 iterations resulted in a runtime of about 52 minutes.

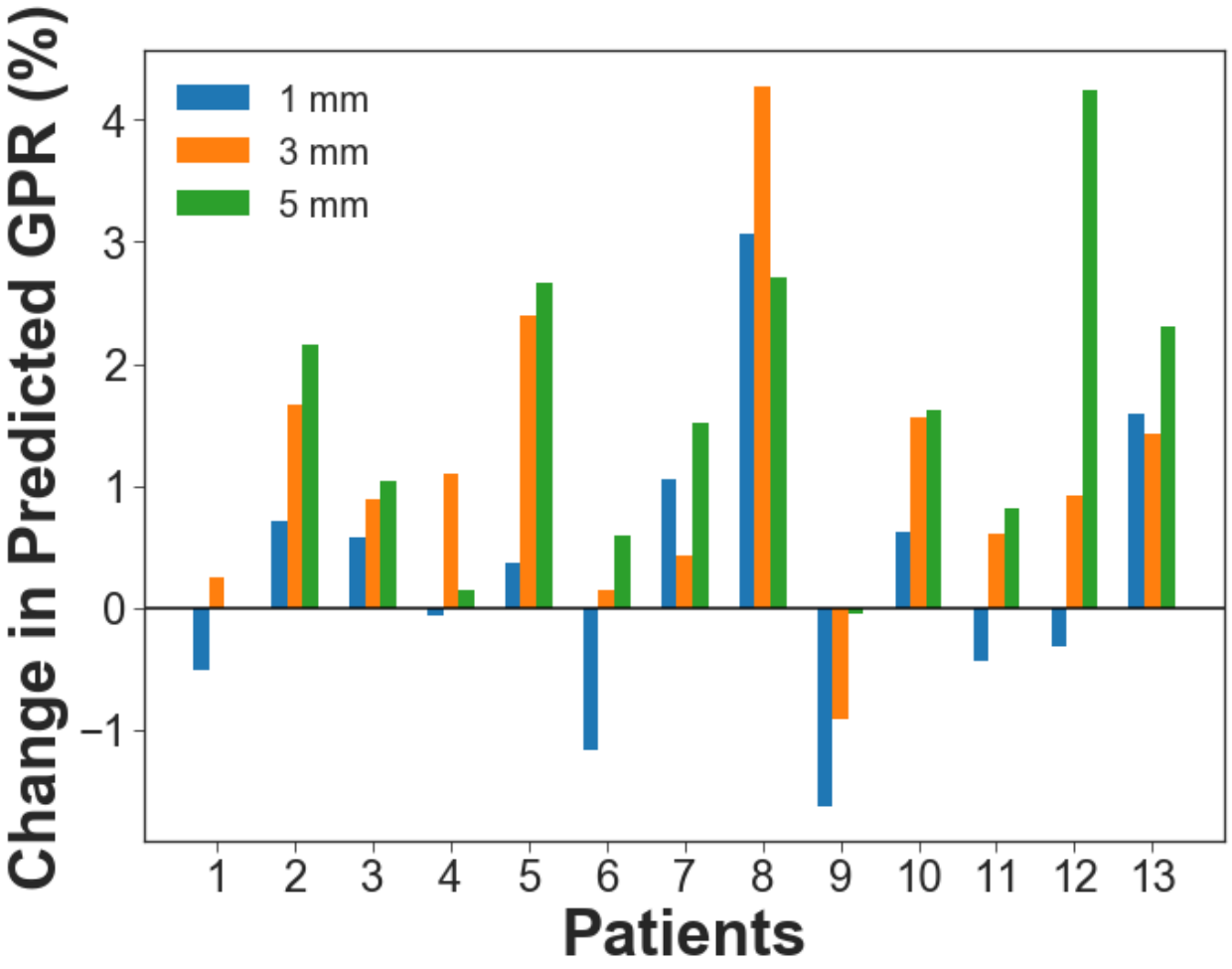


Figure 4.2. Changes in predicted GPRs of QA-optimized plans for each patient relative to the original KBP plans using maximum random LG displacements of 1 (blue), 3 (orange), and 5 (green) mm with 25 optimization iterations. A positive value indicates the QA-optimized plan has the higher predicted GPR. Note these GPRs were calculated with a 3%/3mm percent dose-difference/distance-to-agreement gamma criterion using local normalization.

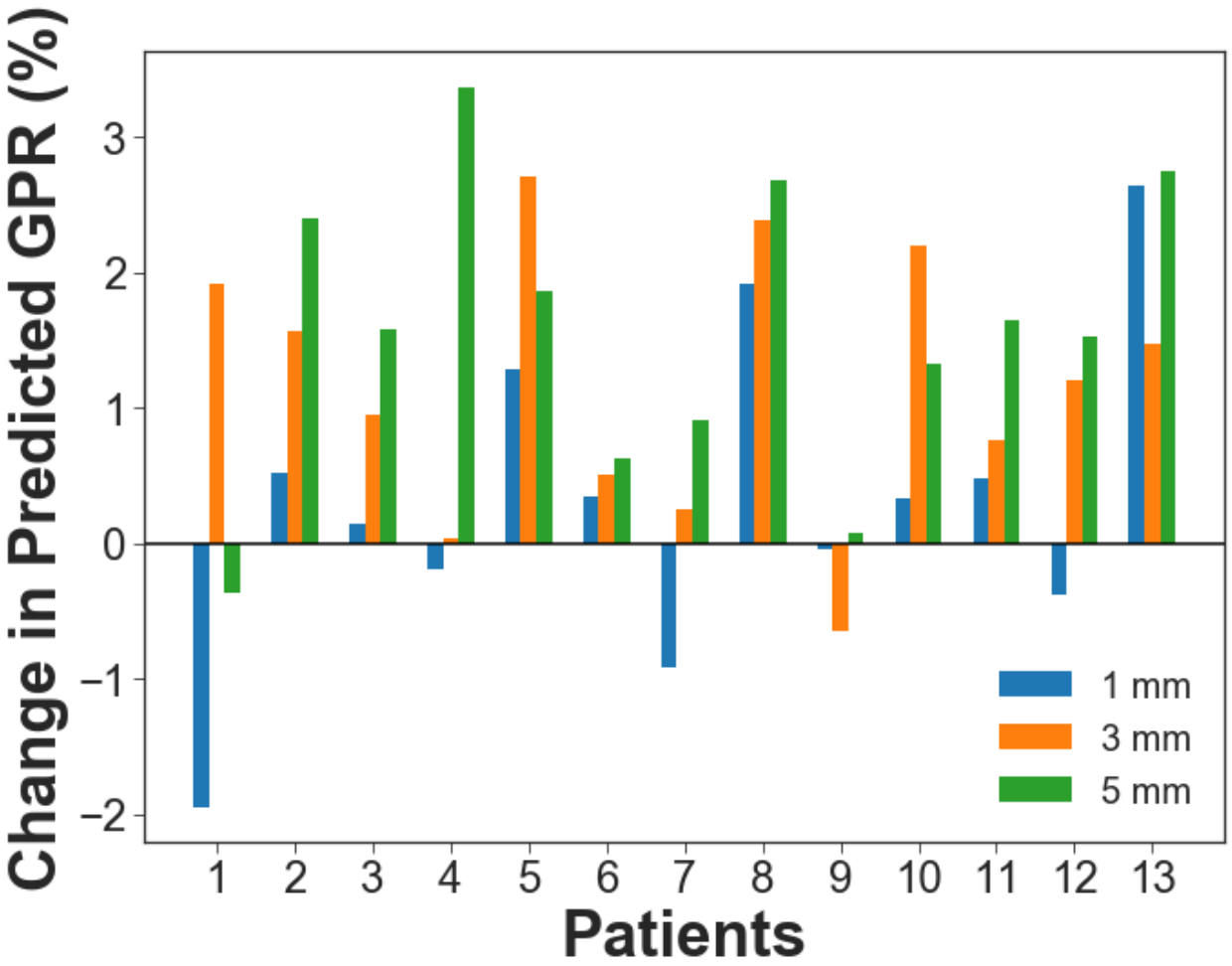


Figure 4.3. Changes in predicted GPRs of QA-optimized plans for each patient relative to the original KBP plans using maximum random LG displacements of 1 (blue), 3 (orange), and 5 (green) mm with 1000 optimization iterations. A positive value indicates the QA-optimized plan has the higher predicted GPR. Note these GPRs were calculated with a 3%/3mm percent dose-difference/distance-to-agreement gamma criterion using local normalization.

4.3.1.1. Changes in Plan Complexity

Table 4.2 and Table 4.3 show changes in complexity metrics between the QA-optimized plans and the original KBP plans overall all patients when using 25 and 1000 optimization iterations respectively. Overall, the QA-optimized plans showed significantly ($p < 0.05$) reduced levels of plan complexity compared to the original plans, with the magnitude increasing with maximum random LG displacement.

Table 4.2. Mean ± standard deviations (μ ± σ) of the differences in complexity metrics between QA-optimized KBP plans and the corresponding original KBP plan for 1, 3, and 5 mm maximum random LG displacements. The QA-optimized plans were generated after 25 iterations.

Complexity Metric	KBP-QA – KBP (μ ± σ)		
	1 mm	3 mm	5 mm
MU Factor	-2.92 ± 1.04*	-7.93 ± 3.27*	-13.01 ± 5.28*
MCS	0.0034 ± 0.0004*	0.010 ± 0.002*	0.015 ± 0.003*
EM	-0.001 ± 0.0001*	-0.002 ± 0.0004*	-0.004 ± 0.001*
Mean LM (mm/deg)	-0.001 ± 0.002	0.008 ± 0.007*	0.025 ± 0.013*
Mean LG (mm)	0.39 ± 0.02*	1.18 ± 0.07*	1.97 ± 0.12*
Mean Aperture Area (mm^2)	41.80 ± 11.09*	125.61 ± 33.87*	209.30 ± 54.12*
SAS – 50 mm	-0.005 ± 0.002*	-0.018 ±0.006*	-0.029 ± 0.005*

*indicates a statistically significant result with $p < 0.05$

The optimization algorithm reduced MU factors by 2.92 ± 1.04 (μ ± σ; $p = 0.001$) and increased the mean LG by 0.39 ± 0.02 mm ($p = 0.001$) using a maximum LG displacement of 1 mm. Aperture areas of the QA-optimized plans also increased significantly, where a 1 mm maximum LG displacement setting resulted in increasing mean aperture areas by 41.80 ± 11.09 mm^2 ($p = 0.001$) compared to the original plan. Differences in complexity metrics were negligible when comparing plans using 25 and 1000 optimization iterations.

Table 4.3. Mean ± standard deviations (μ ± σ) of the differences in complexity metrics between QA-optimized KBP plans and the corresponding original KBP plan for 1, 3, and 5 mm maximum random LG displacements. The QA-optimized plans were generated after 1000 iterations.

Complexity Metric	KBP-QA – KBP (μ ± σ)		
	1 mm	3 mm	5 mm
MU Factor	-2.93 ± 1.00*	-8.07 ± 3.29*	-12.92 ± 5.66*
MCS	0.0034 ± 0.0004*	0.010 ± 0.002*	0.015 ± 0.003*
EM	-0.001 ± 0.0001*	-0.003 ± 0.0004*	-0.004 ± 0.001*
Mean LM (mm/deg)	-0.001 ± 0.002*	0.008 ± 0.006*	0.026 ± 0.014*
Mean LG (mm)	0.40 ± 0.02*	1.18 ± 0.07*	1.97 ± 0.13*

(table cont'd)

Complexity Metric	KBP-QA – KBP ($\mu \pm \sigma$)		
	1 mm	3 mm	5 mm
Mean Aperture Area (mm^2)	42.12 ± 10.82*	125.42 ± 32.82*	208.99 ± 55.42*
SAS – 50 mm	-0.005 ± 0.002*	-0.017 ± 0.004*	-0.028 ± 0.007*

*indicates a statistically significant result with $p < 0.05$

4.3.2. Changes in Plan Quality

Figure 4.4 shows the average of the 13 DVHs across the original KBP plan and the QA-optimized plans with the three different maximum random LG displacements using 25 iterations. Table 4.4 shows the resulting differences among selected dose metrics between the QA-optimized plans and the original KBP plan. Deviations in dose from the original plan generally increased with maximum LG displacement. While statistical significances were observed for most of the dose metrics for the target and each of the OARs, the magnitude of the differences were relatively small. For the PTV dose coverage, significant ($p < 0.05$) increases were observed for D_{min} and D_{mean} for each maximum LG displacement. However, the largest average increase that was observed was minimal, on the order of 1.09 Gy and 0.27 Gy for D_{min} and D_{mean} respectively when the maximum LG displacement was set at 5 mm. The target dose distributions also became slightly less conformal and less homogenous as the maximum LG displacement increased compared to the original KBP plan, where the conformity index and homogeneity index significantly increased by an average of 0.044 and 0.009, respectively, with the maximum LG displacement set to 5 mm.

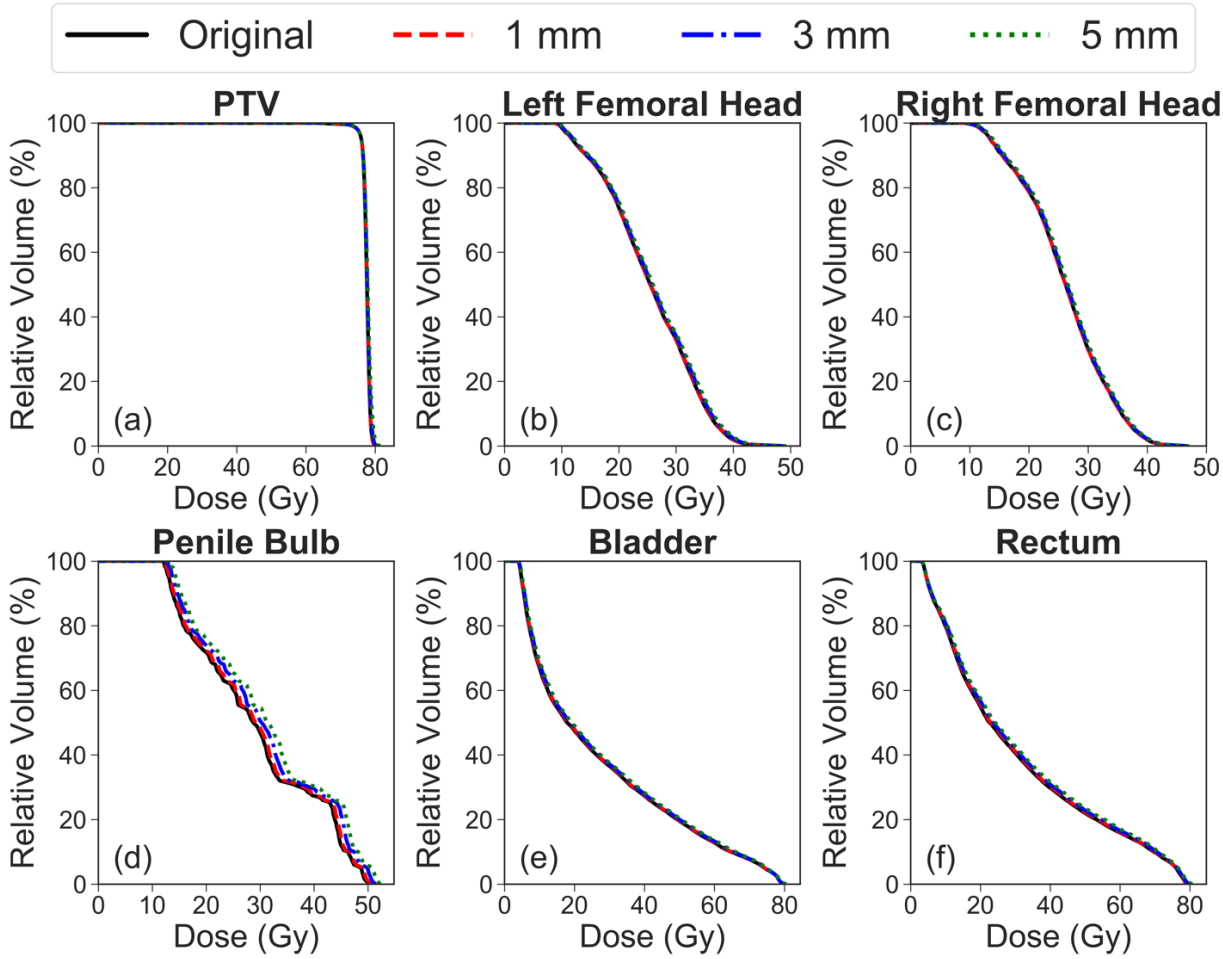


Figure 4.4. Average dose volume histograms over the 13 patients comparing the original plan (solid black) with the QA-optimized plans using maximum random LG displacements of 1 (dashed red), 3 (dashdot blue), and 5 (dotted green) mm. These QA-optimized plans were generated after 25 iterations.

Statistically significant increases in bladder and rectum doses were observed for each of the dose metrics assessed via Table 4.4. However, the largest average magnitude of increased D_{mean} to the bladder and the rectum across all possible maximum LG displacements was 0.83 Gy and 1.22 Gy, respectively, which correspond to increases of 3.06% and 4.08% of D_{mean}. Significant increases in D_{mean} were also observed for the left and right femoral heads and penile bulb, with the largest observed increases being 0.53 Gy, 0.44 Gy and 2.29 Gy on average (representing 2.04%, 1.69%, and 7.79% of average D_{mean} values), respectively.

Table 4.4. Average differences in dose metrics between QA-optimized plans (KBP-QA) and the original KBP plans for maximum random LG displacements of 1, 3, and 5 mm. The QA-optimized dose values resulted from plans were generated after 25 iterations.

Dose Metric	Mean (KBP-QA – KBP)		
	1mm	3mm	5mm
PTV			
D_2 (cGy)	3.50	32.16*	79.80*
D_{50} (cGy)	4.77*	8.94	21.93*
D_{98} (cGy)	0.21	3.90	6.36
D_{min} (cGy)	22.89*	65.80*	109.63*
D_{mean} (cGy)	4.02*	10.97*	26.97*
D_{max} (cGy)	-9.12	19.34	87.06*
V_{95} (%)	0.008	0.04*	0.08*
V_{98} (%)	-0.01	0.02	0.06
V_{100} (%)	-0.12	-0.12	-0.12
V_{107} (%)	0.008	0.12	0.59
HI†	0.0004	0.004*	0.009*
CI†	0.008*	0.025*	0.044*
Bladder			
D_{10} (cGy)	23.72*	74.63*	129.33*
D_{30} (cGy)	32.68*	85.22*	147.22*
D_{50} (cGy)	26.89*	75.44*	123.43*
D_{65} (cGy)	19.74*	45.56*	76.12*
D_{80} (cGy)	10.92*	25.78*	41.83*
D_{mean} (cGy)	18.94*	49.51*	82.78*
Rectum			
D_{10} (cGy)	27.77*	84.26*	140.87*
D_{30} (cGy)	46.57*	137.14*	226.39*
D_{50} (cGy)	35.31*	98.69*	158.54*
D_{65} (cGy)	25.26*	61.61*	102.77*
D_{80} (cGy)	15.99*	38.91*	62.22*
D_{mean} (cGy)	25.78*	72.86*	122.35*
Left Femoral Head			
D_2 (cGy)	7.26	46.19*	84.64*
D_{max} (cGy)	7.82	38.70*	80.30*
D_{mean} (cGy)	6.22*	30.25*	53.10*
Right Femoral Head			
D_2 (cGy)	-1.05	23.53*	47.14*
D_{max} (cGy)	-1.79	27.52*	52.74*
D_{mean} (cGy)	0.82	23.20*	43.89*

(table cont'd)

Dose Metric	Mean (KBP-QA – KBP)		
	1mm	3mm	5mm
Penile Bulb			
D_{mean} (cGy)	41.36*	133.93*	229.14*

*indicates a statistically significant result with $p < 0.05$
†Homogeneity and conformity indices were calculated according to their International Commission on Radiation Units & Measurements definitions.

The changes in dose observed in Figure 4.4 and Table 4.4 for QA-optimized plans with 25 iterations were similar to the dose differences observed for QA-optimized plans resulting from 1000 iterations, as can be seen in Figure 4.5 and Table 4.5.

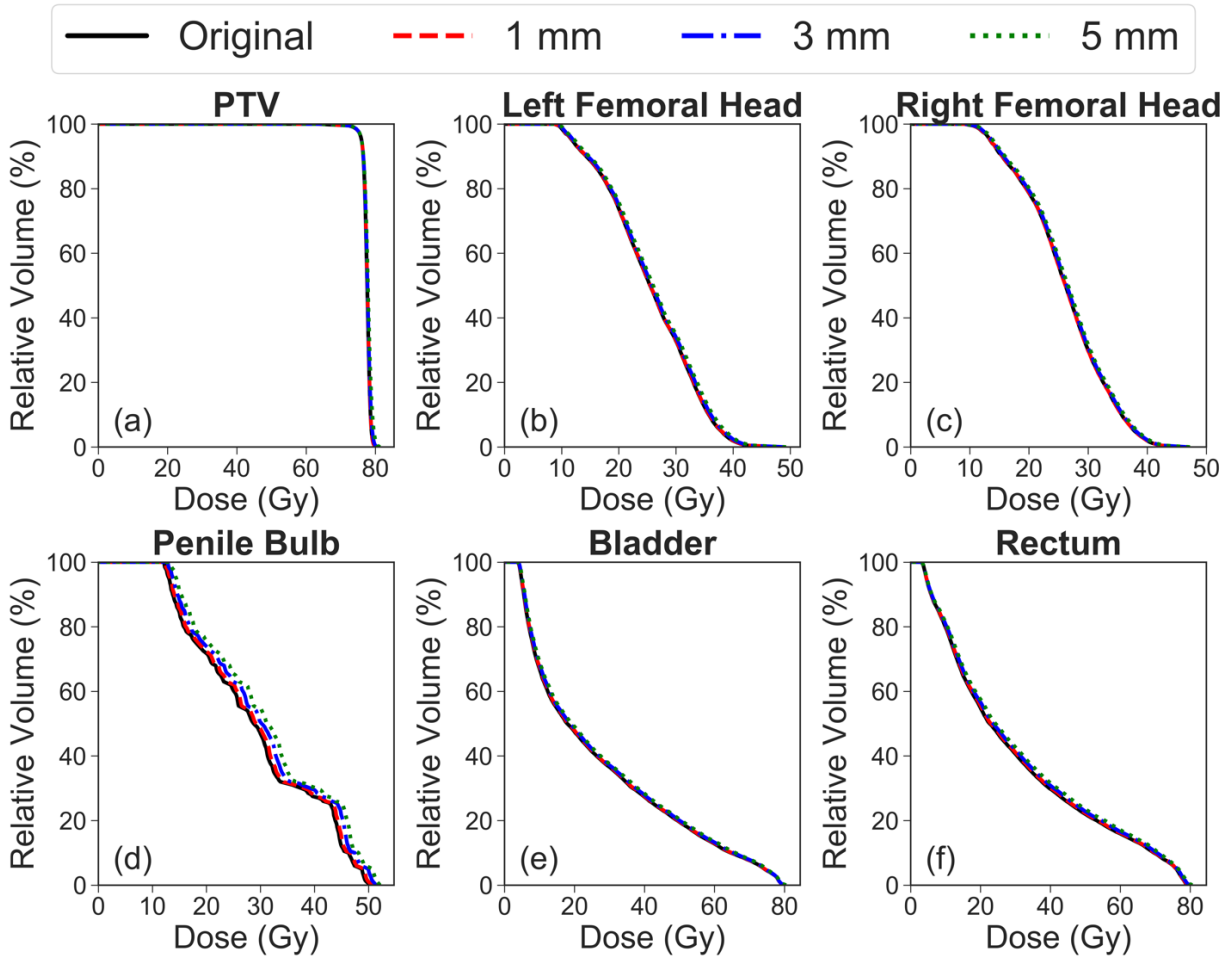


Figure 4.5. Average dose volume histograms over the 13 patients comparing the original plan (solid black) with the QA-optimized plans using maximum random LG displacements of 1 (dashed red), 3 (dashdot blue), and 5 (dotted green) mm. These QA-optimized plans were generated after 1000 iterations.

Again, small, but significant dose differences were observed with increased maximum LG displacement for these QA-optimized plans resulting from 1000 iterations. These changes were not noticeably different from those exhibited from the QA-optimized plans resulting from 25 iterations.

Table 4.5. Average differences in dose metrics between QA-optimized plans (KBP-QA) and the original KBP plans for maximum random LG displacements of 1, 3, and 5 mm. The QA-optimized dose values resulted from plans were generated after 1000 iterations.

Dose Metric	Mean (KBP-QA – KBP)		
	1mm	3mm	5mm
PTV			
D_2 (cGy)	3.21	28.91*	81.46*
D_{50} (cGy)	4.10*	8.33	22.03*
D_{98} (cGy)	-0.11	1.72	3.78
D_{min} (cGy)	21.94*	67.53*	115.99*
D_{mean} (cGy)	3.47*	9.79	26.82*
D_{max} (cGy)	-11.14	21.46	79.98*
V_{95} (%)	0.008	0.04*	0.07*
V_{98} (%)	-0.015	-0.006	0.07
V_{100} (%)	-0.12	-0.12	-0.12
V_{107} (%)	0.003	0.09	0.70
HI†	0.0004	0.003*	0.010*
CI†	0.007*	0.025*	0.045*
Bladder			
D_{10} (cGy)	24.41*	72.70*	131.35*
D_{30} (cGy)	32.81*	87.39*	142.09*
D_{50} (cGy)	25.45*	72.33*	129.82*
D_{65} (cGy)	19.90*	46.69*	75.37*
D_{80} (cGy)	11.00*	25.05*	41.95*
D_{mean} (cGy)	18.80*	49.14*	82.92*
Rectum			
D_{10} (cGy)	25.83*	84.41*	138.40*
D_{30} (cGy)	48.10*	132.83*	231.93*
D_{50} (cGy)	35.26*	98.53*	166.56*
D_{65} (cGy)	24.50*	61.43*	101.76*
D_{80} (cGy)	15.41*	37.45*	63.32*
D_{mean} (cGy)	25.91*	72.03*	123.69*

(table cont'd)

Dose Metric	Mean (KBP-QA – KBP)		
	1mm	3mm	5mm
Left Femoral Head			
D_2 (cGy)	8.73*	44.04*	83.00*
D_{max} (cGy)	7.33	51.57*	84.15*
D_{mean} (cGy)	6.30*	30.53*	57.11*
Right Femoral Head			
D_2 (cGy)	-1.44	22.26*	49.77*
D_{max} (cGy)	0.12	23.49	64.87*
D_{mean} (cGy)	0.53	22.10*	45.31*
Penile Bulb			
D_{mean} (cGy)	43.38*	132.55*	224.14*

*indicates a statistically significant result with $p < 0.05$
†Homogeneity and conformity indices were calculated according to their International Commission on Radiation Units & Measurements definitions.

Changes in the EUD-based radiobiological metrics TCP and NTCP can be found in Table 4.6 and Table 4.7 for QA-optimized plans generated after 25 iterations and 1000 iterations, respectively. Significant increases in TCP were observed across each maximum LG displacement setting. Statistically significant increases in NTCP were observed for the bladder, rectum, and femoral heads, although the magnitudes of each were less than 0.22%, on average.

Table 4.6. Summary of mean differences (∆) in the radiobiological metrics based on equivalent uniform doses: tumor control probability (TCP) and normal tissue complication probability (NTCP). Here, ∆ reflects the difference between the QA-optimized values and the original values. Specifically, a positive ∆TCP (or ∆NTCP) indicates the QA-optimized plan had a higher TCP (or NTCP) value than the original plan. These QA-optimized plans were generated after 25 iterations.

Max LG Displacement (mm)	∆TCP (%) (10^{-2})	∆NTCP (%)			
		Bladder (10^{-4})	Rectum (10^{-1})	Left Femur (10^{-7})	Right Femur (10^{-8})
1	0.6*	2.2*	0.2	0.1	-0.5
3	2.4*	4.2*	1.1*	0.7*	2.1*
5	4.7*	8.0*	2.2*	1.3*	5.6*

*indicates a statistically significant result with $p < 0.05$

The magnitudes in the changes observed for these radiobiological metrics were similar regardless of the number of iterations used to generate the QA-optimized plans.

Table 4.7. Summary of differences (∆) in the radiobiological metrics based on equivalent uniform doses: tumor control probability (TCP) and normal tissue complication probability (NTCP). Here, ∆ reflects the difference between the QA-optimized values and the original values. Specifically, a positive ∆TCP (or ∆NTCP) indicates the QA-optimized plan had a higher TCP (or NTCP) value than the original plan. These QA-optimized plans were generated after 1000 iterations.

Max LG Displacement (mm)	ΔTCP (%) (10^{-2})	ΔNTCP (%)			
		Bladder (10^{-4})	Rectum (10^{-1})	Left Femur (10^{-7})	Right Femur (10^{-8})
1	0.6*	2.0*	0.2	0.1	0.5
3	2.1*	4.6*	1.1*	0.7*	2.3*
5	5.1*	7.9*	2.2*	1.5*	5.9*

*indicates a statistically significant result with $p < 0.05$

4.4. DISCUSSION

This study explored the idea of integrating the prediction of VMAT QA outcomes into the plan design stage by using a machine learning model as a feedback mechanism within a heuristic algorithm for adjusting mechanical parameters of treatment plans towards increasing their predicted deliverability without degrading their dosimetric quality. Specifically, an SVM model based on QA measurements of 500 previous VMAT treatment plans was implemented into the plan optimization workflow. The algorithm selects plan features (deemed to be important for accurately predicting QA outcomes using the SVM model) and randomly modifies them in searching for a solution that maximally increases predicted deliverability. Using 13 previous KBP-guided VMAT plans for purposes of demonstrating feasibility, the QA-optimization algorithm was found to yield statistically significant increases in predicted GPRs and reductions in plan complexity, while minimally altering the plan quality in terms of dosimetric and radiobiological metrics. Although the magnitude of each of these

changes increased with maximum random LG displacement, a 3 mm maximum displacement and 25 iterations yielded QA-optimized plans with a 1.14 ± 1.25% (μ ± σ; p = 0.006) increase in predicted GPRs and negligible increases (largest average increase of 0.1% among bladder, rectum, and femoral heads) in OAR NTCPs, although these increases were statistically significant (p < 0.05).

To our knowledge, this is the first study exploring the feasibility of a planning and quality assurance framework that directly integrates predicted QA outcomes into the plan design process. Previous investigators have researched the efficacy of using individual complexity penalties to improve the deliverability of treatment plans. For instance, Younge *et al*. found it was possible to improve dose delivery accuracy by penalizing plans with complex aperture shapes via their "edge" metric without resulting in substantial changes to the dose distribution quality.[112] While many different complexity metrics and mechanical plan features have been studied for quantifying and predicting pre-treatment QA results, few significant relationships have been found between them for VMAT QA.[144,165] In other words, penalizing plan complexity without direct knowledge of how the associated changes in mechanical plan parameters impact resulting QA measurements is not guaranteed to result in improved dose delivery accuracy. This has led to assessing the capability of machine learning models of accurately predicting QA outcomes from a large array of different plan complexity metrics.[95-97,99,101,122,166] The present work advances and builds upon these previous works by employing one such machine learning model for purposes of developing a planning QA tool for improving predicted delivery accuracy as defined by GPRs from QA measurements.

The novel and innovative component of this study is the implementation of such a machine learning model directly into the planning optimization process, in effect including predicted QA outcomes as an additional, plan-specific objective function. QA-based optimization has the potential to improve overall plan quality without compromising the deliverability of the plan, in addition to producing efficiency gains in the treatment planning optimization and QA process. As opposed to previous studies that apply penalties based on single complexity or plan features, QA-based optimization modifies a group of aperture-based complexity metrics selected specifically to improve predicted GPRs. This inserts QA-specific endpoints into the plan optimization process, thereby providing *a priori* information to the planner regarding a plan's likely QA outcome. Also, given the recent recommendations by TG-218 of stricter tolerance and action levels in evaluating patient-specific QA, QA-based optimization could help maintain the dosimetric quality of treatment plans while still meeting these QA guidelines.[114] In a more general sense, QA-based optimization represents a step towards creating the best possible plan and delivery for each patient.

It is important to note the limitations of utilizing radiobiological metrics for purposes of predicting clinical outcomes. While the goal for every radiation therapy treatment is to simultaneously maximize tumor cell death and minimize risk of normal tissue complications, treatment plans are typically evaluated by physical dose-volume metrics, which are merely a surrogate for these biological endpoints. This has led to an effort to integrate biologically relevant metrics into the treatment planning design and evaluation process.[167,168] Even though several commercial TPSs have begun to incorporate biologically based models, predictive NTCP and TCP models have had

limited clinical presence to date due to questions of reliability and uncertainties. Specifically, variations in model parameter estimates (e.g. a, α/β, etc.) have been shown to yield different radiobiological predictions.[169-171] These uncertainties are largely due to a lack of satisfactory datasets for confidently establishing the correlation between these predictive biological metrics and realized clinical outcomes.[172] Acknowledging these deficiencies in using NTCP and TCP models for accurately predicting absolute biological metrics for plan evaluation, these models' ability to correctly capture volume effects and general radiobiological trends can still prove useful for a relative comparison in plan quality. This serves as the rationale for implementing the phenomenological, EUD-based models of NTCP and TCP in this study for evaluating differences in plan quality between QA-optimized plans and their original counterparts. Further, while statistically significant differences were observed in TCP and several OAR NTCP values among QA-optimized plans, the magnitude of these differences (maximum difference of 0.22%; Table 4.6 and Table 4.7) can be considered small within the context of the uncertainties associated with these radiobiological models in general.

This study had other limitations. One practical limitation was the inability to integrate QA-based optimization directly into an existing TPS optimization and dose calculation framework. Several options were explored and considered before deciding to evaluate feasibility via the present workflow. Integration of QA-based optimization within an existing TPS would facilitate a more efficient and comprehensive investigation into the tradeoff between increased levels of predicted QA outcomes and corresponding changes in dosimetric plan quality. This would enable a future study with a larger cohort

of initial testing plans as well. Nonetheless, the present results demonstrate the viability of implementing a QA-based optimization into an existing inverse planning workflow.

The results from this work show that in general, predicted GPRs increased with larger maximum LG displacement settings. This validates the previous feature analysis study indicating LGs and small apertures as reasonable surrogates for adjusting the aperture-based complexity features used to predict GPRs by the SVM model. Further, larger maximum LG displacements tended to result in decreased plan complexity, but increased changes in dose distributions when compared to the original plans. Among the different magnitudes in adjustments made to the complexity features used by the SVM, the 3 mm LG setting seems to be a suitable selection for future integration and testing within a commercially available TPS as it provided the best trade-off in maximizing predicted GPRs and minimizing changes in dosimetric quality of the plans. While further investigation may reveal a more optimal maximum LG displacement setting, the present results suggest that the improvements in the aforementioned trade-off would likely be clinically negligible.

Another limitation of this study was the inherent accuracy of the machine learning model used to inform the heuristic search. The SVM model implemented in this work yielded a testing mean absolute error of 3.75% for predicting VMAT GPRs, which is on the order of the improvements in predicted GPRs achieved by incorporating the SVM model into the optimization process. However, a significant reduction in overall plan complexity was observed in the QA-optimized plans relative their original KBP plans (Table 4.2 and Table 4.3). Additionally, preliminary QA measurements of these optimized plans seem to agree with the differences in predicted QA yielded from the

SVM model, although a more thorough and complete measurement assessment is needed for verification.

While this study focused on solely prostate treatment plans, it would be interesting to investigate the potential impact of this optimization workflow for other, more complex treatment sites, such as head and neck cancers. It is possible the differences in predicted QA outcomes would be more pronounced for treatment sites requiring more complex MLC leaf sequences and larger, more irregularly shaped apertures than prostate cases. However, given a significant increase in predicted QA was observed for these prostate plans without substantially degrading dosimetric quality provided by KBP-guidance is an indication of the potential clinical utility and applicability for any treatment site and inverse optimization planning environment.

The proposed QA-based optimization framework for improving plan deliverability, when paired with a data-driven KBP method, represents a step forward towards a planning workflow integrating both dose- and QA-based objectives for designing a treatment plan with the best composite dosimetric quality and delivery accuracy. The more immediate clinical impact of this framework would be an increase in efficiency, as the number and likelihood of plans failing QA would decrease and could be predicted prior to measurement. Also, an important and novel consequence of this work is the ability to establish the direct connection between global plan complexity mitigation tools within existing TPSs and corresponding QA outcomes during the planning process.

4.5. CONCLUSION

While current TPSs have simple global penalties aiming to reduce complexity and control the likelihood of a plan failing QA measurements, these penalties are surrogates for QA outcomes; any potential impact on QA outcomes would be unknown

until measurement. This study is the first to investigate the feasibility of a planning tool that incorporates a machine learning model into the optimization of VMAT plans. An SVM model, previously trained and tested to predict VMAT GPRs based on an array of plan complexity features, was used to inform a heuristic algorithm for modifying plan parameters with the aim of increasing the predicted GPR. Significant increases in predicted GRPs were observed over 13 QA-optimized plans compared to original reference plans without substantially compromising the dosimetric quality of the plan. Therefore, this study has shown the feasibility of a QA-based optimization routine – with a maximum LG displacement of 3 mm and just 25 iterations – for increasing predicted plan deliverability without impacting dosimetric nor radiobiological plan quality. This novel QA-based optimization could be a useful addition to inverse planning workflow to improve both the overall quality and deliverability of clinical treatment plans.

5. CONCLUSIONS

5.1. SUMMARY OF FINDINGS

The overall goals of this dissertation were threefold: to evaluate differences in dose, complexity, and QA outcomes between reference clinical plans and those designed with an in-house KBP technique (Chapter 2); to develop and compare machine learning models for predicting VMAT QA outcomes based on an array of plan complexity features (Chapter 3); and to assess the feasibility of optimizing plan deliverability of VMAT treatment plans using a machine learning model to predict QA outcomes (Chapter 4).

Thirty-one prostate patients previously treated with VMAT were re-planned with an in-house KBP method based on the overlap volume histogram. In addition to evaluating differences in dose, differences in VMAT plan complexity were quantified via normalized MUs, modulation complexity scores, the edge metric, and average leaf travel per degree of gantry rotation (i.e. leaf motion) for both the reference clinical plans and KBP plans. Each set of plans for each patient was delivered to the same diode-array and GPRs were utilized to quantify the level of agreement between the computed and measured dose distributions. While KBP plans achieved noticeable gains in bladder and rectum dose – with average reductions of 6.4 Gy ($p < 0.001$) and 8.2 Gy ($p < 0.001$) in mean bladder and rectum dose compared to reference plans – they were found to be significantly more complex than reference plans. On average, KBP plans required 143 ± 93 more MUs ($p < 0.001$), had reduced MCS values of 18% ($p < 0.001$; indicating increased complexity), had 40% higher EM values ($p < 0.001$), and 47% higher LM ($p < 0.001$) compared to reference plans. Further, KBP plans were also more susceptible to QA measurement errors. For gamma criteria with global normalization, KBP plans on

average had gamma passing rates that were 1.1, 1.6, 3.8, and 7.8 percentage points lower than reference plans at the 3%/3mm (p = 0.009), 3%/2mm (p = 0.003), 2%/2mm (p = 0.002), and 1%/1mm (p < 0.001) criteria respectively.

These observed differences in plan complexity and deliverability between KBP and reference clinical plans served as motivation for further investigating the relationship between complexity features and QA outcomes through the use of machine learning models. This was established using a dataset of 500 VMAT treatment plans and diode-array QA measurements, upon which an array of machine learning models was trained. GPRs were computed using a 3%/3mm percent dose-difference and distance-to-agreement gamma criterion with local normalization. 241 complexity metrics and plan parameters were extracted from each treatment plan and their relative importance for accurately predicting GPRs was assessed and compared using feature selection methods via forests of Extra-Trees, mutual information, and linear regression. Hyperparameters of different machine learning models – which included linear models, support vector machines (SVMs), tree-based models, and neural networks – were tuned using cross-validation on the training data (80%/20% training/testing split). While features were weakly correlated with GPRs in general, with the small aperture score (SAS) at 50 mm having the largest absolute Pearson correlation coefficient (0.38; p < 0.001), the SVM model, trained using the 100 most important features selected using the linear regression method, yielded the lowest cross-validation testing mean absolute error of 3.75%. While not as accurate as previously published models designed to predict GPRs for fixed-gantry IMRT plans (e.g. Valdes *et al*. reported being able to predict GPRs within 3%), the SVM model in this study provided a significant

improvement in performance compared to “random guessing” for predicting VMAT QA. Furthermore, its ability to capture the general relationship between plan complexity and resulting QA outcomes was determined to be sufficient for utilization in the subsequent objective of this work: to explore the feasibility of developing a QA-based optimization framework for inverse planning workflows.

A heuristic optimization framework was proposed for directly maximizing predicted QA outcomes of plans without degrading the quality of the plan’s dose distribution using the aforementioned SVM model. Thirteen of the prostate VMAT plans designed with an in-house KBP system from Chapter 2 were used to assess the feasibility of this framework. An algorithm was devised by utilizing the SVM model to guide iterative modification of mechanical treatment features most commonly associated with suboptimal GPRs. Specifically, leaf gaps (LGs) less than 5 cm were widened by random amounts, which impacts several complexity features such as small aperture scores and aperture area uniformity. The original 13 plans were optimized with this QA-based algorithm using maximum LG displacements of 1, 3, and 5 mm before corresponding changes in predicted GPRs and dose were assessed. Predicted GPRs increased by an average of 0.30 ± 1.22% (p = 0.42), 1.14 ± 1.25% (p = 0.006), and 1.52 ± 1.27% (p =0.003) after QA-based optimization for 1, 3, and 5 mm maximum random LG displacements, respectively. Differences in dose were minimal, resulting in clinically negligible changes in tumor control probability (maximum increase = 0.05%) and normal tissue complication probability (maximum decrease = 0.2% among bladder, rectum, and femoral heads).

The hypotheses of this study were (1) that KBP-guided plans would result in significantly higher complexity and reduced gamma passing rates ($p < 0.05$) compared to reference clinical plans and (2) that a machine learning model designed to predict VMAT QA gamma passing rates can be used within an in-house optimization workflow to increase predicted delivery accuracy without compromising KBP plan quality. To this end, the results of from this study support the first hypothesis as KBP plans were observed to have significantly increased levels of complexity and QA errors. As for the second hypothesis, the feasibility of using a QA-based optimization framework to improve predicted plan deliverability was demonstrated in this study, with improved levels of predicted GPRs without sacrificing the dose gains provided by the KBP technique.

5.2. LIMITATIONS

There were several limitations of this work. The use of a diode-array, while shown to be a viable measurement device for performing VMAT QA,[140-142] lacks the spatial resolution advantages that can be found in other dosimeters such as film. Planar diode-arrays have also been shown to be overresponsive to lateral beam angles, which could have impacted measured GPRs in this study.[139,140] A more complete characterization of the sources of error would be a useful future avenue of research.

Another limitation of this study was the inability to integrate and test the QA-based optimization concept directly within an existing TPS. This restricted the extent and scope of possible testing, as performing plan modifications external to a TPS before importing the modified plans to compute dose and evaluate changes in the resulting dose distribution is inefficient for testing many different patients and combinations of optimization settings. Still, the methods described in this work demonstrate feasibility of

such a method and the potential clinical advantages that warrant further TPS integration.

As mentioned in Chapter 4.4, the accuracy of the SVM model used within the QA-based optimization algorithm was another limitation of this study. The testing MAE of 3.75% for predicting VMAT GPRs was larger than any of the average increases in predicted GPRs across the 13 patient plans that were tested via the QA-based optimization algorithm. However, significant reductions in overall plan complexities were observed for the modified plans, which aligns with what would be expected. Also, preliminary QA measurements of a sample of QA-optimized plans seem to support the predicted differences in QA outcomes from the SVM model, although a more complete validation study is needed.

5.3. FUTURE WORK

A logical next progression for this project would be to explore how these results generalize to other, generally more complicated treatment sites such as the head-and-neck and to other combinations of technologies (e.g. KBP implementation, TPS, treatment machine, and measurement device). An important note for this study is that these results were observed for the specific dataset and clinical tools used in the experimental methods. These results, combined with those from previous works, indicate there may be some variation present when considering different treatment sites and technologies.

It would also be worthwhile to investigate and characterize the specific sources of error that led to the observed decreases in KBP QA outcomes relative to the reference clinical plans. Although sources of error in IMRT planning and delivery have been studied extensively, it would be interesting to assess the primary sources of error within

the context of the in-house KBP method used in this study. Along the same lines, it would be useful to investigate a machine learning model that can predict multiple classes of delivery accuracy metrics that may be more clinically relevant than the GPR. For instance, being able to predict measured dose differences along the central axis or the mean gamma value, when combined with GPRs, could yield more useful indications for clinically relevant patient dose errors.[173]

In preparation for these tools to be implemented clinically, future work would need to be performed to further validate the potential clinical benefits of an in-house KBP method and QA-based optimization framework. One or several radiation oncologists would need to provide feedback on the quality of the KBP-guided plans to ensure their clinical acceptability. Further testing of the QA-based optimization algorithm is also needed to verify its clinical utility, which would involve a larger number of patients with varying treatment sites. This testing would ideally be performed directly within an existing TPS. This integration with commercial TPSs represents an important and necessary next step towards realizing a clinically useful tool. Discussions were held with an interested vendor in the development of this work regarding the possibility of implementing and testing this QA-based optimization within their existing TPS framework, which did not materialize due to timeline restrictions of the project. However, it should be possible to integrate the present QA-based optimization framework within any existing TPS inverse planning process by adding a predicted GPRs objective to the composite objective function. This TPS integration would also facilitate further fruitful investigations, such as the assessment of other optimization algorithms for modifying the mechanical parameters that may be more efficient and lead to even larger

improvements in predicted GPRs. Lastly, verification measurements should be performed to confirm the QA predictions made by the machine learning algorithm.

APPENDIX A. IRB APPROVAL FORM

ACTION ON EXEMPTION APPROVAL REQUEST

Institutional Review Board
Dr. Dennis Landin, Chair
130 David Boyd Hall
Baton Rouge, LA 70803
P: 225.578.8692
F: 225.578.5983
irb@lsu.edu
lsu.edu/research

TO: Phillip Wall
Physics & Astronomy

FROM: Dennis Landin
Chair, Institutional Review Board

DATE: January 8, 2019

RE: **IRB#** E11428

TITLE: QUALITY ASSURANCE OF KNOWLEDGE-BASED RADIATION THERAPY TREATMENT PLANNING

New Protocol/Modification/Continuation: New Protocol

Review Date: 1/8/2019

Approved X **Disapproved**

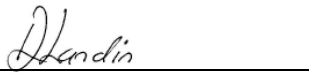
Approval Date: 1/8/2019 **Approval Expiration Date:** 1/7/2022

Exemption Category/Paragraph: 4a

Signed Consent Waived?: N/A

Re-review frequency: (three years unless otherwise stated)

LSU Proposal Number (if applicable):

By: Dennis Landin, Chairman 

PRINCIPAL INVESTIGATOR: PLEASE READ THE FOLLOWING –
Continuing approval is CONDITIONAL on:

1. Adherence to the approved protocol, familiarity with, and adherence to the ethical standards of the Belmont Report, and LSU's Assurance of Compliance with DHHS regulations for the protection of human subjects*
2. Prior approval of a change in protocol, including revision of the consent documents or an increase in the number of subjects over that approved.
3. Obtaining renewed approval (or submittal of a termination report), prior to the approval expiration date, upon request by the IRB office (irrespective of when the project actually begins); notification of project termination.
4. Retention of documentation of informed consent and study records for at least 3 years after the study ends.
5. Continuing attention to the physical and psychological well-being and informed consent of the individual participants, including notification of new information that might affect consent.
6. A prompt report to the IRB of any adverse event affecting a participant potentially arising from the study.
7. Notification of the IRB of a serious compliance failure.
8. **SPECIAL NOTE: When emailing more than one recipient, make sure you use bcc. Approvals will automatically be closed by the IRB on the expiration date unless the PI requests a continuation.**

* *All investigators and support staff have access to copies of the Belmont Report, LSU's Assurance with DHHS, DHHS (45 CFR 46) and FDA regulations governing use of human subjects, and other relevant documents in print in this office or on our World Wide Web site at http://www.lsu.edu/irb*

APPENDIX B. COPYRIGHT INFORMATION

B.1. CHAPTER 2

Copyright Clearance Center RightsLink®

Home | Help | Email Support | Sign in | Create Account

WILEY

Evaluation of complexity and deliverability of prostate cancer treatment plans designed with a knowledge-based VMAT planning technique

Author: Jonas D. Fontenot, Phillip D. H. Wall

Publication: Journal of Applied Clinical Medical Physics

Publisher: John Wiley and Sons

Date: Dec 9, 2019

© 2019 The Authors. Journal of Applied Clinical Medical Physics published by Wiley Periodicals, Inc. on behalf of American Association of Physicists in Medicine.

Welcome to RightsLink

This article is available under the terms of the Creative Commons Attribution License (CC BY) (which may be updated from time to time) and permits use, distribution and reproduction in any medium, provided that the Contribution is properly cited.

For an understanding of what is meant by the terms of the Creative Commons License, please refer to Wiley's Open Access Terms and Conditions.

Permission is not required for this type of reuse.

Wiley offers a professional reprint service for high quality reproduction of articles from over 1400 scientific and medical journals. Wiley's reprint service offers:

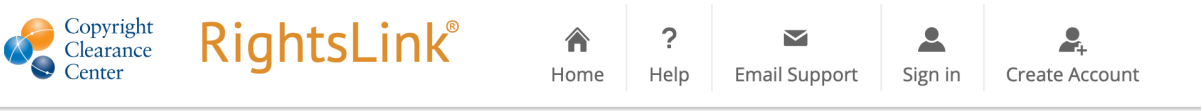
- Peer reviewed research or reviews
- Tailored collections of articles
- A professional high quality finish
- Glossy journal style color covers
- Company or brand customisation
- Language translations
- Prompt turnaround times and delivery directly to your office, warehouse or congress.

Please contact our Reprints department for a quotation. Email corporatesaleseurope@wiley.com or corporatesalesusa@wiley.com or corporatesalesDE@wiley.com .

CLOSE WINDOW

© 2020 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

B.2. CHAPTER 3



Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning

Author: Phillip D.H. Wall,Jonas D. Fontenot

Publication: Informatics in Medicine Unlocked

Publisher: Elsevier

Date: 2020

© 2020 The Authors. Published by Elsevier Ltd.

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

BACK CLOSE WINDOW

© 2020 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

APPENDIX C. SUPPLEMENTARY MATERIAL

C.1. CHAPTER 2

Table C.1. Statistical summary of the differences in coefficients of variation (COV) of inter-delivery measurements at each gamma criteria between reference and KBP plans over the three separate measurements.

	Gamma Criteria	Reference Plans COV ($\mu \pm \sigma$ x 10^{-2})	KBP Plans COV ($\mu \pm \sigma$ x 10^{-2})	t-test *p*-value
Global	3%/3mm	0.3 ± 0.3	0.5 ± 0.6	0.005*
	3%/2mm	0.4 ± 0.4	0.8 ± 0.7	0.001*
	2%/2mm	0.8 ± 0.5	1.8 ± 1.4	< 0.001*
	1%/1mm	3.1 ± 2.1	4.8 ± 3.4	0.005*
Local	3%/3mm	0.9 ± 0.6	1.3 ± 1.0	0.02*
	2%/2mm	1.3 ± 1.0	2.1 ± 1.6	0.004*
	1%/1mm	2.8 ± 1.9	4.2 ± 2.4	0.003*

*Indicates a statistically significant result of $p < 0.05$

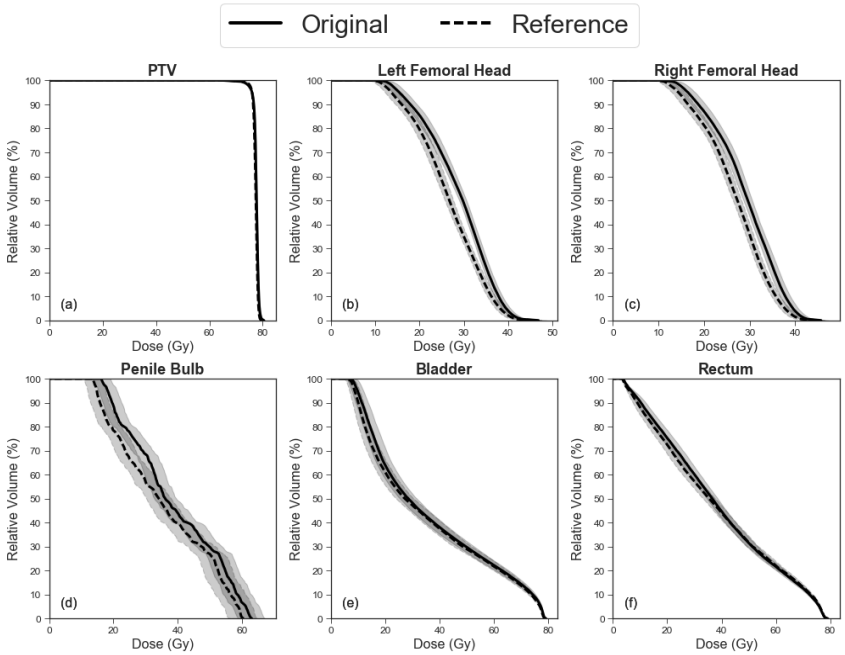


Figure C.1. Average DVHs comparing original clinical plans (solid) to the reconstructed reference clinical plans (dashed) for the 31 patients of each labelled planning structure (a-f). The standard error of the means is also included as filled bands with solid (original) or dashed (reference) edge lines. Note that doses were normalized so that 95% of the PTV received 76 Gy.

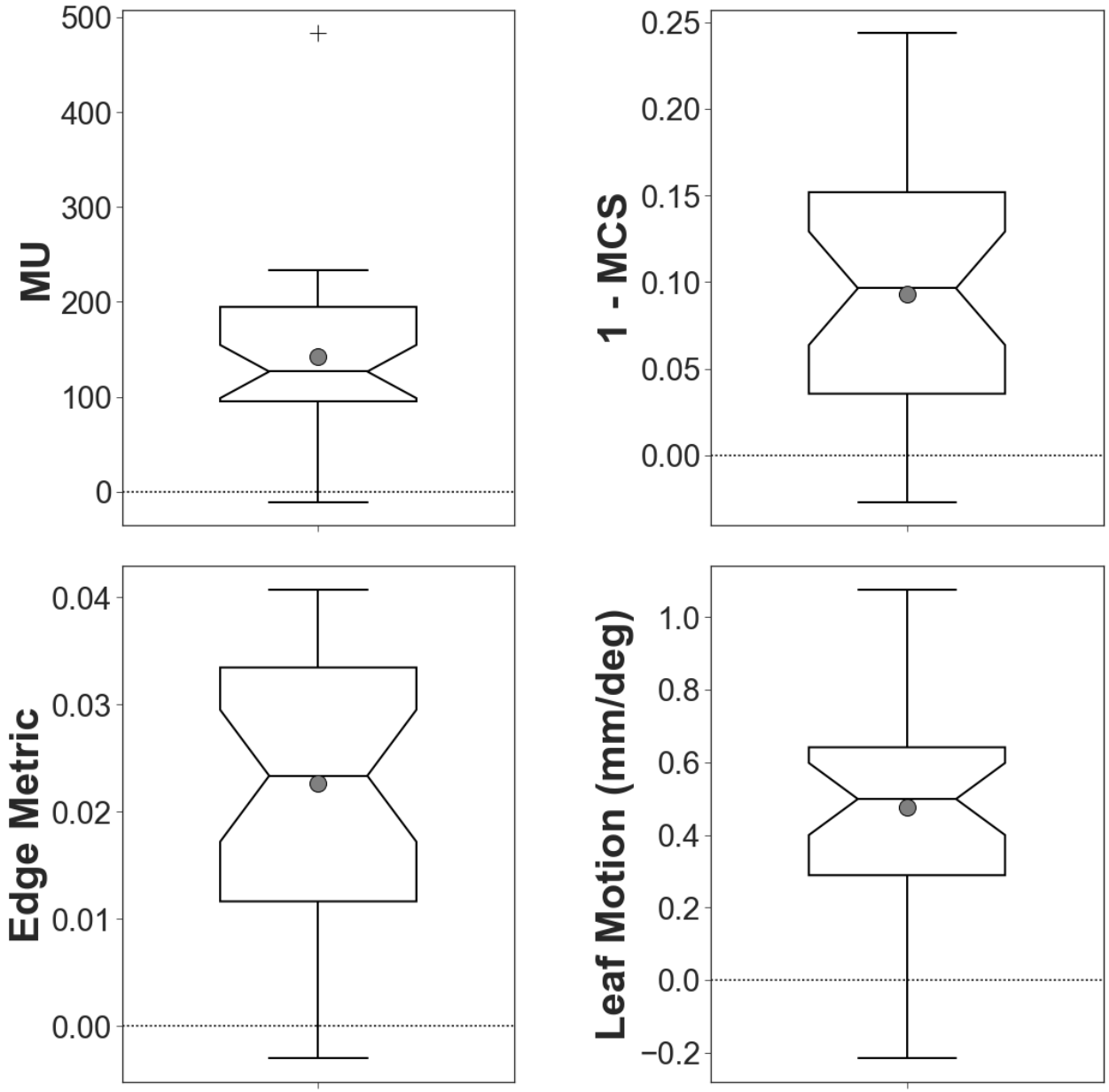


Figure C.2. Distributions of the 31 paired differences between KBP and reference plans for planned MUs (a), MCS values (b), EM values (c), and LM (d). Positive values in each complexity metric plot indicate the KBP value was larger (i.e. more complex) than the corresponding reference plan value. Note in (b), 1 – MCS values were plotted so that higher values indicate higher complexity in each plot. Horizontal black lines within each box indicate distribution medians; notches indicate the 95% confidence intervals around each median, grey circles indicate the distribution means; whiskers indicate the range of data points lying within the 1.5 times the interquartile range and crosses indicate points outside this range.

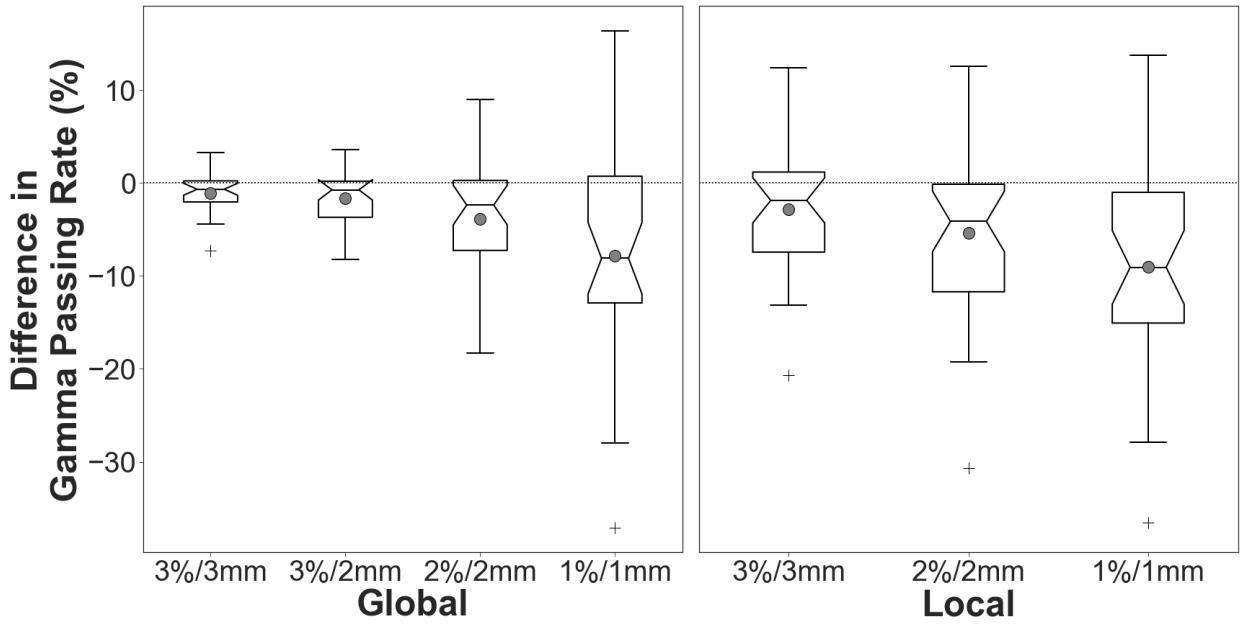


Figure C.3. Distributions of differences in gamma passing rates between reference plans and KBP plans at each gamma index criteria calculated with both global (left) and local (right) normalization. Negative values indicate the KBP plan had a lower gamma passing rate. Same boxplot characteristics from the Figure C.2 caption apply here.

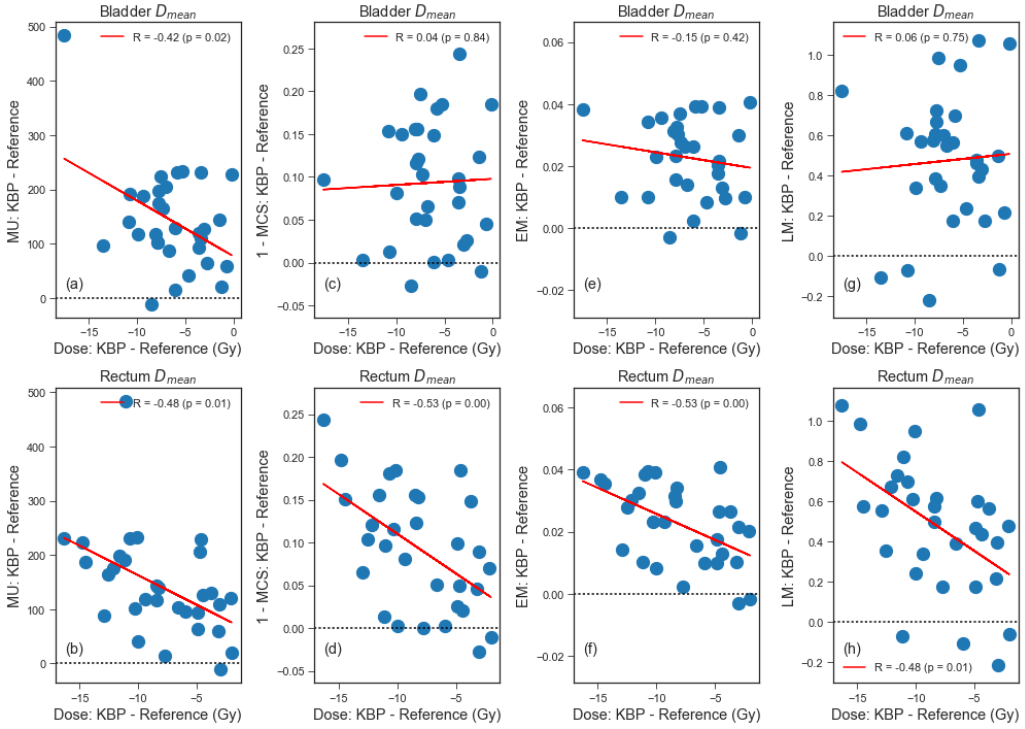


Figure C.4. Correlation between increased plan complexity and improvement in plan quality. Differences between KBP and reference plans are shown, where positive y values indicate increased KBP plan complexity and negative x values indicate improved or lower bladder (a, c, e, g) or rectum dose (b, d, f, h).

C.2. CHAPTER 3

Table C.2. Hyperparameters of the models listed in Table 3.5 that were tuned with cross-validated searches.

Algorithm	α	r	Kernel	C	ε	γ	Max Depth	Min Samples per Leaf	Loss	Number of Estimators	Learning Rate	Subsample
Elastic Net	0.594	1	-	-	-	-	-	-	-	-	-	-
SVM	-	-	RBF	6.407	0.094	$\frac{1}{n_{features} \times \sigma^2}$	-	-	-	-	-	-
Decision Tree	-	-	-	-	-	-	3	17	MAE	-	-	-
Random Forest	-	-	-	-	-	-	12	4	MSE	124	-	-
AdaBoost	-	-	-	-	-	-	-	-	-	91	1.311	-
Gradient Boosting	-	-	-	-	-	-	14	6		616	0.007	0.444

C.2.1. Hyperparameter Tuning

The following describes the specifics of the cross-validated searches carried out for each model to determine the optimal hyperparameters reported in the main text.

C.2.1.1. Elastic Net

The Elastic Net hyperparameters α and r were optimized by a 10-fold cross validated grid-search over specified parameter spaces. The optimal α was determined from 100 uniformly sampled values ranging from 0.005 to 0.25, while the optimal r was determined from 10 uniformly sampled values ranging from 0 to 1.

C.2.1.2. Support Vector Machine

Hyperparameters for SVMs were optimized with a 5-fold cross-validated randomized search with 250 iterations. Table C.3 lists the defined parameter space over which the randomized search was conducted.

Table C.3. SVM parameter space ranges over which optimal parameter values were randomly searched with 5-fold cross-validation.

SVM Parameter	Defined Search Domain
Kernel	Linear, Polynomial, Gaussian RBF, Sigmoid
Degree (Polynomial kernel only)	1-11
C	1000 equally spaced points between 0.001 and 100
ε	1000 equally spaced points between 0.001 and 1.5
γ (for Polynomial, Gaussian RBF, and Sigmoid kernels)	$\frac{1}{n_{features}}, \frac{1}{n_{features} \times \sigma^2}$

C.2.1.3. Decision Tree

The Decision Tree hyperparameters optimized in this study were the maximum tree depth, the minimum number of samples required at each node, and the loss function to measure the quality of a split. These hyperparameters were tuned with a 10-

fold cross-validated randomized search of 100 iterations over the specified parameter space. The range of searched values were 1 to 20 levels for the maximum tree depth parameter, 1 to 20 samples for the minimum number of samples required at each node parameter, and either MSE or mean absolute error (MAE) for the loss function.

C.2.1.4. Random Forest

The same tree-growing hyperparameters and their associated ranges of values given in C.2.1.3 were similarly optimized for Random Forests using a 10-fold cross-validated randomized search over 100 iterations. The number of trees in each forest was also optimized in this randomized search, where models could have anywhere between 10 and 1000 estimators.

C.2.1.5. AdaBoost

The AdaBoost-specific hyperparameters were optimized via a 10-fold cross-validated randomized search with 100 iterations. The parameters tuned in this study were the maximum number of estimators at which boosting was terminated (values ranged from 10 to 1000) and learning rate (values ranged from 0 to 2), which scales the contribution of each regressor.

C.2.1.6. Gradient Boosting

The optimal Gradient Boosting parameters were selected via a 10-fold cross-validated randomized search with 100 iterations over the defined parameter space. The parameters tuned in this study were the maximum number of estimators (ranging from 10 to 1000), learning rate (ranging from 10^{-5} to 1), the fraction of training samples to be used for fitting the individual base predictors (ranging from 10^{-5} to 1), the maximum tree depth (ranging from 1 to 20), and the minimum number of samples required to split an internal node (ranging from 1 to 20).

C.2.1.7. Artificial Neural Network

For ANN hyperparameter optimization, a randomized search of 194 iterations over the given parameter space (Table C.4) was performed for each of the 21 combinations of feature sets using the open-source package Talos (http://github.com/autonomio/talos). Each of the 21 models was trained using 10-fold cross-validation.

Table C.4. ANN hyperparameter space defined for optimization using the Talos package.

Talos-specific hyperparameter	Parameter space
Learning Rate	0.001 to 1
Number of neurons in first layer	5 to 50
Batch Size	32 to 125
Number of hidden layers	1 to 10
Topology Shapes†	'brick', 'triangle', 'funnel'
Epochs	50 to 500
Dropout	0 to 0.5
Optimizer	Adam, SGD, Nadam
Losses	MSE, MAE
Hidden layers activation function	Sigmoid, tanh, relu, linear
Output layer activation function	Sigmoid, tanh, relu, linear

†Topology shapes are package-specific names where 'brick' assigns the same number of neurons in each layer, 'triangle' decreases the number of neurons by a constant number with each layer so that the shape resembles a triangle, and 'funnel' decreases the number of neurons by floor of the difference between the specified number of neurons in the first layer and last layer divided by the number of desired hidden layers, resulting in a funnel shape.

APPENDIX D. COMPLEXITY METRICS

Below is the list of complexity features used in this work along with their definitions and works from which they were derived, if applicable.

Number	Name	Definition	Reference(s)
1.	AA moments, weighted†	Moments of AA weighted by CP MU distribution over all CPs in plan	143
2.	AA moments†	Moments of AA distribution over all CPs in plan	143
3.	AA, average	Average AA over all CPs in plan	143
4.	AA, average weighted	Average AA weighted by CP MU over all CPs in plan	143
5.	AA, weighted	Sum of AAs weighted by CP MU over all CPs in plan	143
6.	AP moments, weighted†	Moments of AP weighted by CP MU distribution over all CPs in plan	143
7.	AP moments†	Moments of AP distribution over all CPs in plan	143
8.	AP, average	Average AP over all CPs in plan	143
9.	AP, average weighted	Average AP weighted by CP MU over all CPs in plan	143
10.	AP, weighted	Sum of APs weighted by CP MU over all CPs in plan	143
11.	Aperture Area (AA)	Sum of AAs over all CPs in plan	143
12.	Aperture Perimeter (AP)	Sum of APs over all CPs in plan	143
13.	Arc length	Sum of degrees of gantry rotation in plan	
14.	Closed Leaf Score (CLS)	Ratio of closed leaf pairs to all leaf pairs weighted by CP MU	144
15.	Collimator angle, average	Collimator angle averaged over each beam	95
16.	Cross-Axis Score (CAS)	Ratio of number of leaf pairs where one leaf crosses central axis over total number of in-field leaf pairs weighted by CP MU	144
17.	Edge Metric (EM)	Ratio of MLC side lengths and aperture perimeter	112
18.	FAOC moments, weighted†*	Moments of FAOC weighted by CP MU distribution over all CPs in plan	95
19.	FAOC moments†*	Moments of FAOC distribution over all CPs in plan	95

Number	Name	Definition	Reference(s)
20.	FAOC, average weighted*	Average FAOC weighted by CP MU over all CPs	[95]
21.	Fractional Area Outside of Circle (FAOC), average*	Average fraction of AA outside a circle of radius *r* centered at isocenter over all CPs	[95]
22.	Leaf Gap (LG)	Average LG of each leaf pair over all CPs in plan	[95]
23.	Leaf Motion (LM)	Average leaf travel per degree of gantry rotation per leaf, averaged over all CPs in plan	[121]
24.	Leaf Travel (LT)	Average total leaf travel per leaf	[113, 121]
25.	LG moments, weighted†	Moments of LG weighted by CP MU distribution over all leaf pairs in each CP in plan	[95]
26.	LG moments†	Moments of LG distribution over all leaf pairs in each CP in plan	[95]
27.	LG, average weighted	Average LG of each leaf pair weighted by CP MU over all CPs in plan	[95]
28.	LM moments, weighted†	Moments of average LM distribution of leaves weighted by CP MU	[121]
29.	LM moments†	Moments of average LM distribution of leaves	[121]
30.	LM, weighted	LM weighted by CP MU	[121]
31.	LT moments, weighted†	Moments of total leaf travel distribution of leaves weighted by CP MU	[113, 121]
32.	LT moments†	Moments of total leaf travel distribution of leaves	[113, 121]
33.	LT, average weighted	LT weighted by CP MU	[113, 121]
34.	Machine	Name of treatment machine	[95]
35.	MLC – agility	Binary, yes or no	[95]
36.	MLC – i2	Binary, yes or no	[95]
37.	Modulation Complexity Score (MCS)	Product of aperture area variability, leaf sequence variability, and control point (CP) weight	[120, 113]
38.	MU factor	Total planned MUs divided by total fractional dose to the specification point i.e. iso	[143]
39.	Number of arcs		[95]
40.	PI moments, weighted†	Moments of AI weighted by CP MU distribution over all CPs in plan	[143]
41.	PI moments†	Moments of AI distribution over all CPs in plan	[143]
42.	PI, average	Average AI over all CPs in plan	[143]

Number	Name	Definition	Reference(s)
43.	PI, average weighted	Average AI weighted by plan MU over all CPs in plan	[143]
44.	PI, weighted	Sum of AI values weighted by CP MU over all CPs in plan	[143]
45.	Plan irregularity (PI)	Computed aperture irregularity (AI) as non-circularity of each aperture: $\frac{AP^2}{4\pi \cdot AA}$; sum of AI values over all CPs in plan	[143]
46.	Plan modulation (PM)	Computed as the weighted sum of each beam's modulation value: $BM_i = 1 - \frac{\sum_j MU_{ij} \cdot AA_{ij}}{MU_i \cdot U(AA_{ij})}$; where $U(AA_{ij})$ is the union area of all apertures of beam *i*	[143]
47.	Small Aperture Scores (SAS)‡	Ratio of leaf gaps smaller than given distance *r* weighted by CP MU	[144]
48.	SAS, maximum‡	Maximum of SAS scores over all CPs in plan	[144]
49.	Y jaw position, average	Average of Y1 and Y2 positions	[95]
50.	Y1/Y2 jaw motion, average weighted	Average Y1/Y2 jaw motion weighted by CP MU	[95]
51.	Y1/Y2 jaw motion, average†	Average jaw travel per degree of gantry rotation of each Y1/Y2 jaw	[95]
52.	Y1/Y2 jaw position, average	Averaged over all CPs	[95]
53.	Y1/Y2 jaw travel	Total Y1/Y2 jaw travel over plan	[95]

†the k = 2nd, 3rd, 4th, and 5th moments were used in this study
‡distance criteria of r = 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 mm were used in this study
*circles of radii = 25, 50, 75, 100, 125, 150, 175, and 200 mm were used in this study

REFERENCES

1. How Radiation Therapy Is Used to Treat Cancer. American Cancer Society. https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/radiation/basics.html. Published 2019. Accessed 2020.

2. Skowronek J. Current status of brachytherapy in cancer treatment - short overview. *J Contemp Brachytherapy.* 2017;9(6):581-589.

3. Getting External Beam Radiation Therapy. American Cancer Society. https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/radiation/external-beam-radiation-therapy.html. Published 2019. Accessed 2020.

4. Gupta T, Agarwal J, Jain S, et al. Three-dimensional conformal radiotherapy (3D-CRT) versus intensity modulated radiation therapy (IMRT) in squamous cell carcinoma of the head and neck: a randomized controlled trial. *Radiother Oncol.* 2012;104(3):343-348.

5. Staffurth J, Radiotherapy Development B. A review of the clinical evidence for intensity-modulated radiotherapy. *Clin Oncol (R Coll Radiol).* 2010;22(8):643-657.

6. Veldeman L, Madani I, Hulstaert F, De Meerleer G, Mareel M, De Neve W. Evidence behind use of intensity-modulated radiotherapy: a systematic review of comparative clinical studies. *Lancet Oncol.* 2008;9(4):367-375.

7. Abel E, Silander E, Nyman J, et al. Impact on quality of life of IMRT versus 3-D conformal radiation therapy in head and neck cancer patients: A case control study. *Advances in radiation oncology.* 2017;2(3):346-353.

8. Yu T, Zhang Q, Zheng T, et al. The Effectiveness of Intensity Modulated Radiation Therapy versus Three-Dimensional Radiation Therapy in Prostate Cancer: A Meta-Analysis of the Literatures. *PLoS One.* 2016;11(5):e0154499-e0154499.

9. Shimizuguchi T, Nihei K, Okano T, Machitori Y, Ito K, Karasawa K. A comparison of clinical outcomes between three-dimensional conformal radiotherapy and intensity-modulated radiotherapy for prostate cancer. *Int J Clin Oncol.* 2017;22(2):373.

10. Otto K. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Med Phys.* 2008;35(1):310-317.

11. Tsai CL, Wu JK, Chao HL, Tsai YC, Cheng JC. Treatment and dosimetric advantages between VMAT, IMRT, and helical tomotherapy in prostate cancer. *Med Dosim.* 2011;36(3):264-271.

12. Hardcastle N, Tome WA, Foo K, Miller A, Carolan M, Metcalfe P. Comparison of prostate IMRT and VMAT biologically optimised treatment plans. *Med Dosim.* 2011;36(3):292-298.

13. Kopp RW, Duff M, Catalfamo F, Shah D, Rajecki M, Ahmad K. VMAT vs. 7-field-IMRT: assessing the dosimetric parameters of prostate cancer treatment with a 292-patient sample. *Med Dosim.* 2011;36(4):365-372.

14. Deng Z, Shen L, Zheng X, et al. Dosimetric advantage of volumetric modulated arc therapy in the treatment of intraocular cancer. *Radiat Oncol.* 2017;12(1):83.

15. Vanetti E, Clivio A, Nicolini G, et al. Volumetric modulated arc radiotherapy for carcinomas of the oro-pharynx, hypo-pharynx and larynx: a treatment planning comparison with fixed field IMRT. *Radiother Oncol.* 2009;92(1):111-117.

16. Bertelsen A, Hansen CR, Johansen J, Brink C. Single Arc Volumetric Modulated Arc Therapy of head and neck cancer. *Radiother Oncol.* 2010;95(2):142-148.

17. Davidson MT, Blake SJ, Batchelar DL, Cheung P, Mah K. Assessing the role of volumetric modulated arc therapy (VMAT) relative to IMRT and helical tomotherapy in the management of localized, locally advanced, and post-operative prostate cancer. *Int J Radiat Oncol Biol Phys.* 2011;80(5):1550-1558.

18. Holt A, van Vliet-Vroegindeweij C, Mans A, Belderbos JS, Damen EM. Volumetric-modulated arc therapy for stereotactic body radiotherapy of lung tumors: a comparison with intensity-modulated radiotherapy techniques. *Int J Radiat Oncol Biol Phys.* 2011;81(5):1560-1567.

19. Lee TF, Chao PJ, Ting HM, et al. Comparative analysis of SmartArc-based dual arc volumetric-modulated arc radiotherapy (VMAT) versus intensity-modulated radiotherapy (IMRT) for nasopharyngeal carcinoma. *J Appl Clin Med Phys.* 2011;12(4):3587.

20. Nguyen K, Cummings D, Lanza VC, et al. A dosimetric comparative study: volumetric modulated arc therapy vs intensity-modulated radiation therapy in the treatment of nasal cavity carcinomas. *Med Dosim.* 2013;38(3):225-232.

21. Quan EM, Li X, Li Y, et al. A comprehensive comparison of IMRT and VMAT plan quality for prostate cancer treatment. *Int J Radiat Oncol Biol Phys.* 2012;83(4):1169-1178.

22. Studenski MT, Bar-Ad V, Siglin J, et al. Clinical experience transitioning from IMRT to VMAT for head and neck cancer. *Med Dosim.* 2013;38(2):171-175.

23. Teoh M, Clark CH, Wood K, Whitaker S, Nisbet A. Volumetric modulated arc therapy: a review of current literature and clinical use in practice. *Br J Radiol.* 2011;84(1007):967-996.

24. Popple RA, Balter PA, Orton CG. Point/Counterpoint. Because of the advantages of rotational techniques, conventional IMRT will soon become obsolete. *Med Phys.* 2014;41(10):100601.

25. Ezzell GA, Galvin JM, Low D, et al. Guidance document on delivery, treatment planning, and clinical implementation of IMRT: report of the IMRT Subcommittee of the AAPM Radiation Therapy Committee. *Med Phys.* 2003;30(8):2089-2115.

26. Matsuo Y, Takayama K, Nagata Y, et al. Interinstitutional variations in planning for stereotactic body radiation therapy for lung cancer. *Int J Radiat Oncol Biol Phys.* 2007;68(2):416-425.

27. Batumalai V, Jameson MG, Forstner DF, Vial P, Holloway LC. How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case. *Pract Radiat Oncol.* 2013;3(3):e99-e106.

28. Berry SL, Boczkowski A, Ma R, Mechalakos J, Hunt M. Interobserver variability in radiation therapy plan output: Results of a single-institution study. *Pract Radiat Oncol.* 2016;6(6):442-449.

29. Das IJ, Cheng CW, Chopra KL, Mitra RK, Srivastava SP, Glatstein E. Intensity-modulated radiation therapy dose prescription, recording, and delivery: patterns of variability among institutions and treatment planning systems. *J Natl Cancer Inst.* 2008;100(5):300-307.

30. Nelms BE, Robinson G, Markham J, et al. Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems. *Pract Radiat Oncol.* 2012;2(4):296-305.

31. Moore KL, Schmidt R, Moiseenko V, et al. Quantifying Unnecessary Normal Tissue Complication Risks due to Suboptimal Planning: A Secondary Study of RTOG 0126. *Int J Radiat Oncol Biol Phys.* 2015;92(2):228-235.

32. Ge Y, Wu QJ. Knowledge-based planning for intensity-modulated radiation therapy: A review of data-driven approaches. *Med Phys.* 2019;46(6):2760-2775.

33. Zhu X, Ge Y, Li T, Thongphiew D, Yin FF, Wu QJ. A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Med Phys.* 2011;38(2):719-726.

34. Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore KL. Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Med Phys.* 2012;39(12):7446-7461.

35. Yuan L, Ge Y, Lee WR, Yin FF, Kirkpatrick JP, Wu QJ. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. *Med Phys.* 2012;39(11):6868-6878.

36. Lian J, Yuan L, Ge Y, et al. Modeling the dosimetry of organ-at-risk in head and neck IMRT planning: an intertechnique and interinstitutional study. *Med Phys.* 2013;40(12):121704.

37. Fogliata A, Wang PM, Belosi F, et al. Assessment of a model based optimization engine for volumetric modulated arc therapy for patients with advanced hepatocellular cancer. *Radiat Oncol.* 2014;9(1):236.

38. Zarepisheh M, Long T, Li N, et al. A DVH-guided IMRT optimization algorithm for automatic treatment planning and adaptive radiotherapy replanning. *Med Phys.* 2014;41(6):061711.

39. Fogliata A, Nicolini G, Clivio A, et al. A broad scope knowledge based model for optimization of VMAT in esophageal cancer: validation and assessment of plan quality among different treatment centers. *Radiat Oncol.* 2015;10:220.

40. Stanhope C, Wu QJ, Yuan L, et al. Utilizing knowledge from prior plans in the evaluation of quality assurance. *Phys Med Biol.* 2015;60(12):4873-4891.

41. Tol JP, Dahele M, Delaney AR, Slotman BJ, Verbakel WF. Can knowledge-based DVH predictions be used for automated, individualized quality assurance of radiotherapy treatment plans? *Radiat Oncol.* 2015;10:234.

42. Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WF. Evaluation of a knowledge-based planning solution for head and neck cancer. *Int J Radiat Oncol Biol Phys.* 2015;91(3):612-620.

43. Wang J, Jin X, Zhao K, et al. Patient feature based dosimetric Pareto front prediction in esophageal cancer radiotherapy. *Med Phys.* 2015;42(2):1005-1011.

44. Yang Y, Li T, Yuan L, et al. Quantitative comparison of automatic and manual IMRT optimization for prostate cancer: the benefits of DVH prediction. *J Appl Clin Med Phys.* 2015;16(2):5204.

45. Boutilier JJ, Craig T, Sharpe MB, Chan TC. Sample size requirements for knowledge-based treatment planning. *Med Phys.* 2016;43(3):1212-1221.

46. Chang ATY, Hung AWM, Cheung FWK, et al. Comparison of Planning Quality and Efficiency Between Conventional and Knowledge-based Algorithms in Nasopharyngeal Cancer Patients Using Intensity Modulated Radiation Therapy. *Int J Radiat Oncol Biol Phys.* 2016;95(3):981-990.

47. Chin Snyder K, Kim J, Reding A, et al. Development and evaluation of a clinical model for lung cancer patients using stereotactic body radiotherapy (SBRT) within a knowledge-based algorithm for treatment planning. *J Appl Clin Med Phys.* 2016;17(6):263-275.

48. Deshpande RR, DeMarco J, Sayre JW, Liu BJ. Knowledge-driven decision support for assessing dose distributions in radiation therapy of head and neck cancer. *Int J Comput Assist Radiol Surg.* 2016;11(11):2071-2083.

49. Hussein M, South CP, Barry MA, et al. Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy. *Radiother Oncol.* 2016;120(3):473-479.

50. Wu H, Jiang F, Yue H, Li S, Zhang Y. A dosimetric evaluation of knowledge-based VMAT planning with simultaneous integrated boosting for rectal cancer patients. *J Appl Clin Med Phys.* 2016;17(6):78-85.

51. Cagni E, Botti A, Micera R, et al. Knowledge-based treatment planning: An inter-technique and inter-system feasibility study for prostate cancer. *Phys Med.* 2017;36:38-45.

52. Chatterjee A, Serban M, Abdulkarim B, et al. Performance of Knowledge-Based Radiation Therapy Planning for the Glioblastoma Disease Site. *Int J Radiat Oncol Biol Phys.* 2017;99(4):1021-1028.

53. Foy JJ, Marsh R, Ten Haken RK, et al. An analysis of knowledge-based planning for stereotactic body radiation therapy of the spine. *Pract Radiat Oncol.* 2017;7(5):e355-e360.

54. Li N, Carmona R, Sirak I, et al. Highly Efficient Training, Refinement, and Validation of a Knowledge-based Planning Quality-Control System for Radiation Therapy Clinical Trials. *Int J Radiat Oncol Biol Phys.* 2017;97(1):164-172.

55. Schubert C, Waletzko O, Weiss C, et al. Intercenter validation of a knowledge based model for automated planning of volumetric modulated arc therapy for prostate cancer. The experience of the German RapidPlan Consortium. *PLoS One.* 2017;12(5):e0178034.

56. Sheng Y, Ge Y, Yuan L, Li T, Yin FF, Wu QJ. Outlier identification in radiation therapy knowledge-based planning: A study of pelvic cases. *Med Phys.* 2017;44(11):5617-5626.

57. Wang J, Hu W, Yang Z, et al. Is it possible for knowledge-based planning to improve intensity modulated radiation therapy plan quality for planners with different planning experiences in left-sided breast cancer patients? *Radiation Oncology.* 2017;12(1):85.

58. Faught AM, Olsen L, Schubert L, et al. Functional-guided radiotherapy using knowledge-based planning. *Radiother Oncol.* 2018;129(3):494-498.

59. Masi K, Archer P, Jackson W, et al. Knowledge-based treatment planning and its potential role in the transition between treatment planning systems. *Med Dosim.* 2018;43(3):251-257.

60. Ueda Y, Fukunaga JI, Kamima T, Adachi Y, Nakamatsu K, Monzen H. Evaluation of multiple institutions' models for knowledge-based planning of volumetric modulated arc therapy (VMAT) for prostate cancer. *Radiat Oncol.* 2018;13(1):46.

61. Younge KC, Marsh RB, Owen D, et al. Improving Quality and Consistency in NRG Oncology Radiation Therapy Oncology Group 0631 for Spine Radiosurgery via Knowledge-Based Planning. *Int J Radiat Oncol Biol Phys.* 2018;100(4):1067-1074.

62. Yu G, Li Y, Feng Z, et al. Knowledge-based IMRT planning for individual liver cancer patients using a novel specific model. *Radiat Oncol.* 2018;13(1):52.

63. Zhang J, Wu QJ, Xie T, Sheng Y, Yin FF, Ge Y. An Ensemble Approach to Knowledge-Based Intensity-Modulated Radiation Therapy Planning. *Front Oncol.* 2018;8:57.

64. Cornell M, Kaderka R, Hild SJ, et al. Noninferiority Study of Automated Knowledge-Based Planning Versus Human-Driven Optimization Across Multiple Disease Sites. *Int J Radiat Oncol Biol Phys.* 2020;106(2):430-439.

65. Wu B, Ricchetti F, Sanguineti G, et al. Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys.* 2011;79(4):1241-1247.

66. Wu B, McNutt T, Zahurak M, et al. Fully automated simultaneous integrated boosted-intensity modulated radiation therapy treatment planning is feasible for head-and-neck cancer: a prospective clinical study. *Int J Radiat Oncol Biol Phys.* 2012;84(5):e647-653.

67. Wang Y, Zolnay A, Incrocci L, et al. A quality control model that uses PTV-rectal distances to predict the lowest achievable rectum dose, improves IMRT planning for patients with prostate cancer. *Radiother Oncol.* 2013;107(3):352-357.

68. Wu B, Pang D, Simari P, Taylor R, Sanguineti G, McNutt T. Using overlap volume histogram and IMRT plan data to guide and automate VMAT planning: a head-and-neck case study. *Med Phys.* 2013;40(2):021714.

69. Yang Y, Ford EC, Wu B, et al. An overlap-volume-histogram based method for rectal dose prediction and automated treatment planning in the external beam prostate radiotherapy following hydrogel injection. *Med Phys.* 2013;40(1):011709.

70. Zhou Z, Chen Y, Yu Z, et al. A study of quality control method for IMRT planning based on prior knowledge and novel measures derived from both OVHs and DVHs. *Biomed Mater Eng.* 2014;24(6):3479-3485.

71. Cooper BT, Li X, Shin SM, et al. Preplanning prediction of the left anterior descending artery maximum dose based on patient, dosimetric, and treatment planning parameters. *Adv Radiat Oncol.* 2016;1(4):373-381.

72. Kuo L, Yorke ED, Dumane VA, et al. Geometric dose prediction model for hemithoracic intensity-modulated radiation therapy in mesothelioma patients with two intact lungs. *J Appl Clin Med Phys.* 2016;17(3):371-379.

73. Powis R, Bird A, Brennan M, et al. Clinical implementation of a knowledge based planning tool for prostate VMAT. *Radiat Oncol.* 2017;12(1):81.

74. Valdes G, Simone CB, 2nd, Chen J, et al. Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiother Oncol.* 2017;125(3):392-397.

75. Millunchick CH, Zhen H, Redler G, Liao Y, Turian JV. A model for predicting the dose to the parotid glands based on their relative overlapping with planning target volumes during helical radiotherapy. *J Appl Clin Med Phys.* 2018;19(2):48-53.

76. Chanyavanich V, Das SK, Lee WR, Lo JY. Knowledge-based IMRT treatment planning for prostate cancer. *Med Phys.* 2011;38(5):2515-2522.

77. Good D, Lo J, Lee WR, Wu QJ, Yin FF, Das SK. A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: an example application to prostate cancer planning. *Int J Radiat Oncol Biol Phys.* 2013;87(1):176-181.

78. Nwankwo O, Sihono DS, Schneider F, Wenz F. A global quality assurance system for personalized radiation therapy treatment planning for the prostate (or other sites). *Phys Med Biol.* 2014;59(18):5575-5591.

79. Liu J, Wu QJ, Kirkpatrick JP, Yin FF, Yuan L, Ge Y. From active shape model to active optical flow model: a shape-based approach to predicting voxel-level dose distributions in spine SBRT. *Phys Med Biol.* 2015;60(5):N83-92.

80. Nwankwo O, Mekdash H, Sihono DS, Wenz F, Glatting G. Knowledge-based radiation therapy (KBRT) treatment planning versus planning by experts: validation of a KBRT algorithm for prostate cancer treatment planning. *Radiat Oncol.* 2015;10:111.

81. Schmidt M, Lo JY, Grzetic S, Lutzky C, Brizel DM, Das SK. Semiautomated head-and-neck IMRT planning using dose warping and scaling to robustly adapt plans in a knowledge database containing potentially suboptimal plans. *Med Phys.* 2015;42(8):4428-4434.

82. Schreibmann E, Fox T, Curran W, Shu HK, Crocker I. Automated population-based planning for whole brain radiation therapy. *J Appl Clin Med Phys.* 2015;16(5):76-86.

83. Sheng Y, Li T, Zhang Y, et al. Atlas-guided prostate intensity modulated radiation therapy (IMRT) planning. *Phys Med Biol.* 2015;60(18):7277-7291.

84. McIntosh C, Purdie TG. Contextual Atlas Regression Forests: Multiple-Atlas-Based Automated Dose Prediction in Radiation Therapy. *IEEE Trans Med Imaging.* 2016;35(4):1000-1012.

85. Shiraishi S, Moore KL. Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy. *Med Phys.* 2016;43(1):378.

86. Campbell WG, Miften M, Olsen L, et al. Neural network dose models for knowledge-based planning in pancreatic SBRT. *Med Phys.* 2017;44(12):6148-6158.

87. McIntosh C, Purdie TG. Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Phys Med Biol.* 2017;62(2):415-431.

88. McIntosh C, Welch M, McNiven A, Jaffray DA, Purdie TG. Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. *Phys Med Biol.* 2017;62(15):5926-5944.

89. Lee T, Hammad M, Chan TC, Craig T, Sharpe MB. Predicting objective function weights from patient anatomy in prostate IMRT treatment planning. *Med Phys.* 2013;40(12):121706.

90. Schreibmann E, Fox T. Prior-knowledge treatment planning for volumetric arc therapy using feature-based database mining. *J Appl Clin Med Phys.* 2014;15(2):19-27.

91. Amit G, Purdie TG, Levinshtein A, et al. Automatic learning-based beam angle selection for thoracic IMRT. *Med Phys.* 2015;42(4):1992-2005.

92. Yuan L, Wu QJ, Yin F, et al. Standardized beam bouquets for lung IMRT planning. *Phys Med Biol.* 2015;60(5):1831-1843.

93. Petrovic S, Khussainova G, Jagannathan R. Knowledge-light adaptation approaches in case-based reasoning for radiotherapy treatment planning. *Artif Intell Med.* 2016;68:17-28.

94. Huang Y, Yue H, Wang M, et al. Fully automated searching for the optimal VMAT jaw settings based on Eclipse Scripting Application Programming Interface (ESAPI) and RapidPlan knowledge-based planning. *J Appl Clin Med Phys.* 2018;19(3):177-182.

95. Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning. *Med Phys.* 2016;43(7):4323.

96. Valdes G, Chan MF, Lim SB, Scheuermann R, Deasy JO, Solberg TD. IMRT QA using machine learning: A multi-institutional validation. *J Appl Clin Med Phys.* 2017;18(5):279-284.

97. Interian Y, Rideout V, Kearney VP, et al. Deep nets vs expert designed features in medical physics: An IMRT QA case study. *Med Phys.* 2018;45(6):2672-2680.

98. Tomori S, Kadoya N, Takayama Y, et al. A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance. *Med Phys.* 2018;45:4055-4065.

99. Lam D, Zhang X, Li H, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. *Med Phys.* 2019;46(10):4666-4675.

100. Li J, Wang L, Zhang X, et al. Machine Learning for Patient-Specific Quality Assurance of VMAT: Prediction and Classification Accuracy. *Int J Radiat Oncol Biol Phys.* 2019;105(4):893-902.

101. Ono T, Hirashima H, Iramina H, et al. Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning. *Med Phys.* 2019;46(9):3823-3832.

102. Kazhdan M, Simari P, McNutt T, et al. A shape relationship descriptor for radiation therapy planning. *Med Image Comput Comput Assist Interv.* 2009;12:100-108.

103. Wall PDH, Carver RL, Fontenot JD. An improved distance-to-dose correlation for predicting bladder and rectum dose-volumes in knowledge-based VMAT planning for prostate cancer. *Phys Med Biol.* 2018;63(1):015035.

104. Petit SF, Wu B, Kazhdan M, et al. Increased organ sparing using shape-based treatment plan optimization for intensity modulated radiation therapy of pancreatic adenocarcinoma. *Radiother Oncol.* 2012;102:38-44.

105. Wu B, Pang D, Lei S, et al. Improved robotic stereotactic body radiation therapy plan quality and planning efficacy for organ-confined prostate cancer utilizing overlap-volume histogram-driven planning methodology. *Radiother Oncol.* 2014;112(2):221-226.

106. Ezzell GA, Burmeister JW, Dogan N, et al. IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119. *Med Phys.* 2009;36(11):5359-5373.

107. Yu CX, Jaffray DA, Wong JW. The effects of intra-fraction organ motion on the delivery of dynamic intensity modulation. *Phys Med Biol.* 1998;43(1):91-104.

108. Galvin JM, Ezzell G, Eisbrauch A, et al. Implementing IMRT in clinical practice: a joint document of the American Society for Therapeutic Radiology and Oncology and the American Association of Physicists in Medicine. *Int J Radiat Oncol Biol Phys.* 2004;58(5):1616-1634.

109. Court L, Wagar M, Berbeco R, et al. Evaluation of the interplay effect when using RapidArc to treat targets moving in the craniocaudal or right-left direction. *Med Phys.* 2010;37(1):4-11.

110. Das IJ, Ding GX, Ahnesjo A. Small fields: nonequilibrium radiation dosimetry. *Med Phys.* 2008;35(1):206-215.

111. Nauta M, Villarreal-Barajas JE, Tambasco M. Fractal analysis for assessing the level of modulation of IMRT fields. *Med Phys.* 2011;38:5385-5393.

112. Younge KC, Matuszak MM, Moran JM, McShan DL, Fraass BA, Roberts DA. Penalization of aperture complexity in inversely planned volumetric modulated arc therapy. *Med Phys.* 2012;39(11):7160-7170.

113. Masi L, Doro R, Favuzza V, Cipressi S, Livi L. Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy. *Med Phys.* 2013;40(7):071718.

114. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. *Med Phys.* 2018;45(4):e53-e83.

115. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys.* 1998;25(5):656-661.

116. Pawlicki T, Yoo S, Court LE, et al. Moving from IMRT QA measurements toward independent computer calculations using control charts. *Radiother Oncol.* 2008;89(3):330-337.

117. Siochi R, Huang Y, Bayouth J. WE-D-AUD B-03: Assessment of An In-House Independent Phantom Dose Calculation Algorithm for IMRT QA. *Med Phys.* 2008;35(6Part24):2944-2944.

118. Siochi RA, Molineu A, Orton CG. Point/Counterpoint. Patient-specific QA for IMRT should be performed using software rather than hardware methods. *Med Phys.* 2013;40(7):070601.

119. Park JM, Park SY, Kim H, Kim JH, Carlson J, Ye SJ. Modulation indices for volumetric modulated arc therapy. *Phys Med Biol.* 2014;59(23):7315-7340.

120. McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. *Med Phys.* 2010;37(2):505-515.

121. Hernandez V, Saez J, Pasler M, Jurado-Bruggeman D, Jornet N. Comparison of complexity metrics for multi-institutional evaluations of treatment plans in radiotherapy. *Phys Imag Radiat Oncol.* 2018;5:37-43.

122. Granville DA, Sutherland JG, Belec JG, La Russa DJ. Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics. *Phys Med Biol.* 2019;64(9):095017.

123. Tamura M, Monzen H, Matsumoto K, et al. Mechanical performance of a commercial knowledge-based VMAT planning for prostate cancer. *Radiat Oncol.* 2018;13(1):163.

124. Kubo K, Monzen H, Ishii K, et al. Dosimetric comparison of RapidPlan and manually optimized plans in volumetric modulated arc therapy for prostate cancer. *Phys Med.* 2017;44:199-204.

125. Wall PDH, Carver RL, Fontenot JD. Impact of database quality in knowledge-based treatment planning for prostate cancer. *Pract Radiat Oncol.* 2018;8(6):437-444.

126. Wu B, Ricchetti F, Sanguineti G, et al. Patient geometry-driven information retrieval for IMRT treatment plan quality control. *Med Phys.* 2009;36(12):5497-5505.

127. Van Rossum G, Drake Jr FL. Python tutorial. 1995.

128. Mason D. SU-E-T-33: Pydicom: An Open Source DICOM Library. *Med Phys.* 2011;38(6Part10):3493-3493.

129. Fogliata A, Belosi F, Clivio A, et al. On the pre-clinical validation of a commercial model-based optimisation engine: application to volumetric modulated arc therapy for patients with lung or prostate cancer. *Radiother Oncol.* 2014;113(3):385-391.

130. Agnew CE, Irvine DM, McGarry CK. Correlation of phantom-based and log file patient-specific QA with complexity scores for VMAT. *J Appl Clin Med Phys.* 2014;15(6):4994.

131. Craft D, Suss P, Bortfeld T. The tradeoff between treatment plan quality and required number of monitor units in intensity-modulated radiotherapy. *Int J Radiat Oncol Biol Phys.* 2007;67(5):1596-1605.

132. Mohan R, Arnfield M, Tong S, Wu Q, Siebers J. The impact of fluctuations in intensity patterns on the number of monitor units and the quality and accuracy of intensity modulated radiotherapy. *Med Phys.* 2000;27(6):1226-1237.

133. Feygelman V, Zhang G, Stevens C. Initial dosimetric evaluation of SmartArc - a novel VMAT treatment planning module implemented in a multi-vendor delivery chain. *J Appl Clin Med Phys.* 2010;11(1):3169.

134. Murtaza G, Cora S, Khan EU. Validation of the relative insensitivity of volumetric-modulated arc therapy (VMAT) plan quality to gantry space resolution. *J Radiat Res.* 2017;58(4):579-590.

135. Mihaylov IB, Curran B, Sternick E. The effect of gantry spacing resolution on plan quality in a single modulated arc optimization. *J Appl Clin Med Phys.* 2011;12(4):3603.

136. Snyder Karen C, Liu M, Zhao B, et al. Investigating the dosimetric effects of grid size on dose calculation accuracy using volumetric modulated arc therapy in spine stereotactic radiosurgery. *Journal of Radiosurgery and SBRT.* 2017;4:303-313.

137. Chen F, Rao M, Ye JS, Shepard DM, Cao D. Impact of leaf motion constraints on IMAT plan quality, deliver accuracy, and efficiency. *Med Phys.* 2011;38(11):6106-6118.

138. Letourneau D, Gulam M, Yan D, Oldham M, Wong JW. Evaluation of a 2D diode array for IMRT quality assurance. *Radiother Oncol.* 2004;70(2):199-206.

139. Keeling VP, Ahmad S, Jin H. A comprehensive comparison study of three different planar IMRT QA techniques using MapCHECK 2. *J Appl Clin Med Phys.* 2013;14:222-233.

140. Rinaldin G, Perna L, Agnello G, et al. Quality assurance of rapid arc treatments: performances and pre-clinical verifications of a planar detector (MapCHECK2). *Phys Med.* 2014;30(2):184-190.

141. Iftimia I, Cirino ET, Xiong L, Mower HW. Quality assurance methodology for Varian RapidArc treatment plans. *J Appl Clin Med Phys.* 2010;11(4):3164.

142. Gloi AM, Buchanan RE, Zuge CL, Goettler AM. RapidArc quality assurance through MapCHECK. *J Appl Clin Med Phys.* 2011;12:39-47.

143. Du W, Cho SH, Zhang X, Hoffman KE, Kudchadker RJ. Quantification of beam complexity in intensity-modulated radiation therapy treatment plans. *Med Phys.* 2014;41(2):021716.

144. Crowe SB, Kairn T, Kenny J, et al. Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results. *Australas Phys Eng Sci Med.* 2014;37(3):475-482.

145. Wendling M, Zijp LJ, McDermott LN, et al. A fast algorithm for gamma evaluation in 3D. *Med Phys.* 2007;34(5):1647-1654.

146. Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of machine learning research.* 2003;3(Mar):1157-1182.

147. Ho TK. Random decision forests. Paper presented at: Proceedings of 3rd international conference on document analysis and recognition, 1995.

148. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Machine learning.* 2006;63(1):3-42.

149. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python *Journal of Machine Learning Research.* 2011;12:2825-2830.

150. Cover TM, Thomas JA. *Elements of information theory.* John Wiley & Sons; 2012.

151. Tikhonov AN, Arsenin VY. Solutions of ill-posed problems. 1977. *WH Winston, Washington, DC.* 1977;330.

152. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological).* 1996;58(1):267-288.

153. Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V. Support vector regression machines. Paper presented at: Advances in neural information processing systems, 1997.

154. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, CA: Wadsworth. *International Group.* 1984;432:151-166.

155. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences.* 1997;55(1):119-139.

156. Breiman L. *Arcing the edge.* Technical Report 486, Statistics Department, University of California at Berkeley;1997.

157. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature.* 2016;529(7587):484-489.

158. *Keras* [computer program]. 2015.

159. Bojechko C, Ford EC. Quantifying the performance of in vivo portal dosimetry in detecting four types of treatment parameter variations. *Med Phys.* 2015;42(12):6912-6918.

160. Wall PDH, Fontenot JD. Evaluation of complexity and deliverability of prostate cancer treatment plans designed with a knowledge-based VMAT planning technique. *J Appl Clin Med Phys.* 2020;21(1):69-77.

161. Gay HA, Niemierko A. A free program for calculating EUD-based NTCP and TCP in external beam radiotherapy. *Phys Med.* 2007;23(3):115-125.

162. Wu Q, Mohan R, Niemierko A, Schmidt-Ullrich R. Optimization of intensity-modulated radiotherapy plans based on the equivalent uniform dose. *International Journal of Radiation Oncology • Biology • Physics.* 2002;52(1):224-235.

163. Niemierko A. Reporting and analyzing dose distributions: a concept of equivalent uniform dose. *Med Phys.* 1997;24(1):103-110.

164. Rana S, Cheng C. Radiobiological impact of planning techniques for prostate cancer in terms of tumor control probability and normal tissue complication probability. *Ann Med Health Sci Res.* 2014;4(2):167-172.

165. Crowe SB, Kairn T, Middlebrook N, et al. Examination of the properties of IMRT and VMAT beams and evaluation against pre-treatment quality assurance results. *Phys Med Biol.* 2015;60:2587.

166. Wall PDH, Fontenot JD. Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning. *Informatics in Medicine Unlocked.* 2020;18:100292.

167. Allen Li X, Alber M, Deasy JO, et al. The use and QA of biologically related models for treatment planning: short report of the TG-166 of the therapy physics committee of the AAPM. *Med Phys.* 2012;39(3):1386-1409.

168. Deasy JO, Mayo CS, Orton CG. Treatment planning evaluation and optimization should be biologically and not dose/volume based. *Med Phys.* 2015;42(6):2753-2756.

169. Moiseenko V, Battista J, Van Dyk J. Normal tissue complication probabilities: dependence on choice of biological model and dose-volume histogram reduction scheme. *Int J Radiat Oncol Biol Phys.* 2000;46(4):983-993.

170. Muren LP, Jebsen N, Gustafsson A, Dahl O. Can dose–response models predict reliable normal tissue complication probabilities in radical radiotherapy of urinary bladder cancer? The impact of alternative radiation tolerance models and parameters. *International Journal of Radiation Oncology*Biology*Physics.* 2001;50(3):627-637.

171. Wang H, Cooper BT, Schiff P, et al. Dosimetric assessment of tumor control probability in intensity and volumetric modulated radiotherapy plans. *Br J Radiol.* 2019;92(1094):20180471.

172. Deasy JO, Bentzen SM, Jackson A, et al. Improving normal tissue complication probability models: the need to adopt a "data-pooling" culture. *Int J Radiat Oncol Biol Phys.* 2010;76(3 Suppl):S151-154.

173. Nelms BE, Zhen H, Tomé WA. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. *Med Phys.* 2011;38:1037-1044.

VITA

Phillip Wall was born in Birmingham, Alabama in 1992, where he was raised before graduating from The Altamont School in 2010. Moving to Davidson, North Carolina, he earned a Bachelor of Science degree in Mathematics and Physics from Davidson College in the spring of 2014. Phillip entered the Medical Physics & Health Physics Graduate Program at Louisiana State University (LSU) in the fall of 2014. He earned a Master of Science degree in medical physics from LSU in the summer of 2017, after which he continued at LSU for doctoral studies. Phillip expects to graduate from LSU with a Doctor of Philosophy degree in physics with a concentration in medical physics in 2020. After completing his degree requirements, Phillip will pursue clinical residency training in therapeutic radiation oncology physics at the University of California, San Francisco.