Machine learning for knowledge-based dose-volume histogram prediction in prostate cancer

Victor Immanuel Johannes Strijbis

A thesis presented for the degree of
Master of Science

Delft University of Technology
Faculty of Mechanical, Maritime and
Materials Engineering
Reactor Institute Delft

Antoni van Leeuwenhoek Hospital
Netherlands Cancer Institute
Radiotherapy Department

December 2018

Machine learning for knowledge-based dose-volume histogram prediction in prostate cancer

Victor I.J. Strijbis[1,*]

Supervision: dr. Zoltán Perkó (RID), dr. Tomas Janssen (NKI)

[1] MSc. BME student, TU Delft, The Netherlands
[*] victorstrijbis@gmail.com

Abstract

Introduction
Despite the vast amount of optimization algorithms, radiotherapy treatment planning remains a manual, time-consuming and iterative process. To increase plan standardization, we clinically use Pinnacle's autoplanner for several disease sites. However, this introduces new challenges: first, the autoplanner is not perfect and still requires substantial interaction from the radiotherapy technician (RTT). Second, it is difficult to judge whether a plan has indeed the most optimal trade-off between cure and toxicity, since the RTT has not worked the plan. Knowledge-based planning (KBP) could serve as a quality assurance tool to resolve these problems. It uses historical data (anatomical and dosimetric) from previous plans, to predict the likely dose distribution for the current patient. In this study, we construct an initial, simplistic KBP model that serves as the clinical practice. We then investigate of a variety of KBP modelling approaches to predict rectum dose-volume histograms (DVHs), in order to complement the current clinical practice in prostate cancer.
Methods
For model evaluation, we formulate a clinical tolerance criterion (TC) bandwidth based on a ground-truth set of existing radiotherapy plans. We evaluate on the overall prediction accuracy (RMS), the fraction of correctly predicted DVH bins (TC_α), and on the fraction of patients that have $\geq 90\%$ of their DVH correctly predicted (TC_β). We use the overlap volume histogram (OVH) to encode for organ geometrical information, and use reduced order modelling (ROM) to capture the most important characteristics of the DVH and OVH. Optimization methods we use are Principal Component Analysis (PCA) eigenvalue RMS minimization, direct DVH RMS minimization, and TC_α and TC_β maximization.
Results
Analyses of the KBP clinical practice yielded training and testing errors of 81.4% and 80.8% for TC_α and 53.3% and 51.1% for TC_β, with an RMS of 4.80 and 4.94 volume percentage [%]. Eigenvalue-optimization resulted TC_α of 86.5% and 82.4%, and TC_β of 68.8% and 59.1%, with respective RMS of 2.82 % and 3.22 %. Direct DVH-optimization yielded TC_α of 86.7% and 81.9%, and TC_β of 69.4% and 61.4%, with similar RMS. TC_α and TC_β maximizers resulted TC_α training and testing errors of 92.1% and 78.5%, and TC_β training and testing errors of and 84.3% and 53.4% respectively.
Discussion
The investigated models yielded significant improvements for direct eigenvalue- and DVH-optimization methods. We have also been able to perform optimizations for the clinical goal metrics, showing promising results in training data. Because TC_α- and TC_β-maximizers were unable generalize to perform well for unseen data, it is believed these metrics are too sensitive to be trained reliably, and more consistent data may be required for these optimizers to produce reliable test errors. Based on our findings, we advice the clinical practice to extend KBP-approaches to optimize for DVH-least squares.

Declaration

I declare that this thesis, presented for the degree of *Biomedical Engineering MSc.*, was composed by myself, that the work contained herein is my own unless explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Preface

In front of you lies the MSc. thesis *"Machine learning for knowledge-based dose-volume histogram prediction in prostate cancer"*, which has been written to fulfill the graduation requirements of the Biomedical Engineering programme at Delft University of Technology. My thesis work involved a joint collaboration between the Reactor Institute Delft (RID) and the radiotherapy department at the Netherlands Cancer Institute (NKI) of the Antoni van Leeuwenhoek hospital.

There are several acknowledgements that I wish to address. First, I would like to thank both dr. Zoltán Perkó, Assistant Professor from the Reactor Institute Delft at Delft University of Technology, and dr. Tomas Janssen, Medical Physicist at the Radiotherapy Department of the Antoni van Leeuwenhoek Hospital. The doors to their offices were always open whenever I had a question about my research or writing. They consistently allowed this study to be my own work, but steered me in the right the direction whenever they thought I needed it. In addition, they ensured an optimal environment for me to learn and conduct my work in.

Also, I would like to thank dr. Erik van der Bijl, Medical Physicist in training. His work on knowledge-based planning formed the basis for my thesis work. My conversations with him about his work have played an important role in my understanding of the topics at hand, and have greatly inspired me during the early stages of my research.

Finally, I would like to thank dr. Anne-Lisa Wolf, dr. Duncan den Boer, dr. Milena Smolic and dr. Geert Wortel from the NKI. They were always there to help whenever I had any kind of conceptual problem.

CONTENTS

CHAPTER 1

INTRODUCTION

Cancer is a generic term for diseases that are characterized by out-of-control cell-growth, and is the second leading cause of death with an estimated mortality rate of 9.6 million worldwide in 2017, World Health Organization reports show [WHO18]. The health care costs that accompany these numbers continue to burden society as politics, insurance companies, engineers and caregivers struggle to keep health care sustainable. These are problems that have become especially apparent with the increase in world population, and are expected to be ever more so in the future. Such societal challenges call for efficient engineering solutions.

During the last few decades, improved computational power and insights in the fields of data science and artificial intelligence have brought about an increased demand for machine learning. Machine learning is an application of artificial intelligence where computers are instructed to learn for themselves by using data to find a likely outcome. In order to combat cancer, and therefore to reduce its share in health care costs, machine learning is widely dispatched to take over tasks from radiotherapy physicians. The goal of this thesis is to propose a machine learning approach to treatment planning in prostate cancer, complementary to the current clinical practice. More specifically, this thesis will cover the knowledge-based planning approach to radiotherapy and investigates how different modelling choices can predict the dose to the rectum, as a result of prostate cancer treatment.

1.1 Introduction to radiotherapy

One of the main modalities to treat cancer is radiotherapy or radiation therapy (radiotherapy, RT), and is typically used in combination with chemotherapy and surgery. Radiotherapy is a field of cancer treatment in which ionizing radiation is used to battle the proliferation of tumour cells, by damaging their DNA and thereby inducing cell death.

1.1.1 Types of radiotherapy

Generally, there are two ways to deliver dose to the tumour: through internal sources and by external beams, each of which differs in the source of the radiation.

1. *Internal radiotherapy*
   Internal radiotherapy includes brachytherapy and radioligand therapy. In brachytherapy, radiation comes from radioactive sources that are placed inside the body. These sources are typically sealed vessels or seeds containing radioactive material that are implanted in or near the tumour site. In radioligand therapy, the radioactive source is brought to the tumour site either passively or actively. Passive radioligand therapy makes use of the enhanced permeability and retention effect, which is based on the retention of radioactive substances depending on the defective vascular architecture of tumour tissue. This can be exploited by attaching radionuclides to molecules or nanocarriers, which accumulate at the tumour site and consequently deliver their dose selectively. Active targeting uses vectors such as peptides or antigens to target specific tumour receptors. Typically, a high specific radionuclide activity is required to prevent cold (non-radioactive) nuclides from occupying tumour receptor sites.
2. *External beam radiotherapy*
   In external radiotherapy, rays are administered from outside the body, conformed to tumour cells. Ionizing electromagnetic waves (photons), charged particles or heavy ions (hadrons), or electrons can be used for dose delivery.
   (a) Photon therapy
   Photon therapy is the most widely used modality for external radiotherapy in the Netherlands, because it is relatively simple compared to hadron therapy, making it more cost-effective. However, the long (theoretically infinite) range of photons, together with the scattering physics of photon-matter interactions, poses a threat for all tissues in (the vicinity of) the beam line. This limits healthy tissue sparing.
   (b) Hadron therapy
   As opposed to photons, charged particles, such as protons, have the advantage that they completely stop inside the tissue after the maximum deposition peak, known as the Bragg peak. This potentially allows for a better dose distribution due to a more localized dose delivery. On the other hand, uncertainties are typically more difficult to handle than in photon therapy.
   (c) Electron therapy
   Similar to hadrons, electrons exhibit a Bragg peak in which they deposit their dose, however due to their smaller mass, they have a much smaller range, making them suitable for superficial tumour irradiation. In addition, electrons tend to scatter when interacting with tissue, causing a rather dispersed dose deposition.

1.1.2 External beam radiotherapy modalities

There are several photon radiotherapy modalities that have been used throughout history. The goal of these modalities is conforming the applied radiation dose as much as possible to the tumour. Technological advances have continued to increase dose conformity by means of new imaging modalities, computational methods and different dose delivery systems. This section briefly summarizes three main photon RT modalaties, 3D conformal radiotherapy (3DCRT), intensity modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT).

1. 3D conformal radiation therapy
   3DCRT is a type of conformal radiotherapy that uses 3D images to allow more localised dose administration than conventional 2D methods. In current 3DCRT, multi-leaf collimators (MLCs) are used to conform a beam of constant intensity to different projections of the tumour. The advantage with respect to its 2D predecessor is that it enables irradiation over multiple angles, spreading healthy tissue dose over a larger volume. The main drawback of 3DCRT is that the intensity within a beam cannot be varied.
2. Intensity modulated radiation therapy
   IMRT allows not only for better tumour conformity with the help of MLCs, but also allows for better modulation of the beam in accordance with the tumour shape and position by allowing controlling intensity within the beam. A drawback of IMRT is that it is still rather time-consuming.
3. Volumetric modulated arc therapy
   VMAT was developed to provide further flexibility with respect to IMRT by allowing the continuous movement of the gantry and MLCs, as well as changing the dose rate [Ott08], making treatment treatments more time-efficient. Currently, IMRT and VMAT are predominantly used.

1.1.3 Challenges in radiotherapy

Core challenges in external beam radiotherapy modalities may involve precise tumour or critical structure localization, image guidance, palliative treatments, dealing with uncertainties such as patient movement or organ displacement due to breathing, pulsation, or the filling and emptying of the bladder and bowels. What's more, a prominent challenge that this study is involved with is balancing trade-offs between tumour cure and healthy tissue toxicity, which is the main focus of radiotherapy treatment planning.

1.2 Treatment planning

Treatment planning in radiotherapy is a process with the goal of identifying a personalized treatment plan with an optimal trade-off between tumour cure and healthy tissue toxicity. In recent times, several tools have emerged in the market to automate treatment planning, but in spite of this, treatment planning remains a complex, manual, time-consuming and iterative process. A representation of the treatment planning process in terms of input and output is shown in Figure 1.1, which we will now discuss in brief.

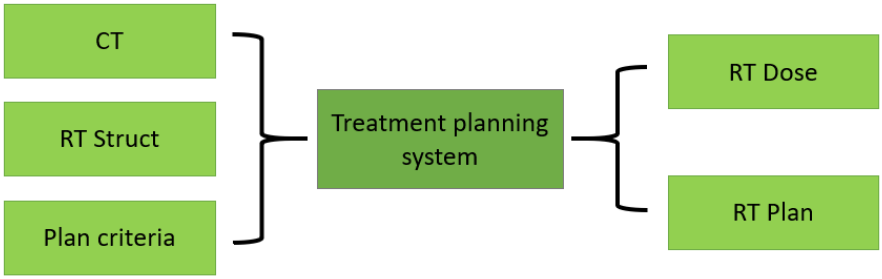


Figure 1.1: A simplified black box system representation of the treatment planning process

Here, the first input denotes the planning computed tomography (CT) scan, which contains the anatomical information of the patient at hand. In the clinic, *RT Struct* contains the information required by the treatment planning system (TPS) to tell apart important structures. *RT Dose* is one of the outputs of the TPS, and describes the desired dose distribution of the patient at hand. The process of finding this distribution is referred to as fluence map optimization. *RT Plan* contains a description of the machine parameters required to actually deliver this plan, which is find by an optimization referred to as machine parameter optimization. Finally, a treatment plan complies with a set of plan criteria, typically set by the radiation oncologist. We will now go over the treatment planning process in some more detail.

1.2.1 Structure delineation

Every radiotherapy patient is treated with an individualized treatment plan. At the basis of each treatment plan stands a planning computed tomography (CT) scan. This information is used for dose calculation. For additional information on dose calculation, the reader is referred to Appendix A.1. The CT is also used for delineation of the tumour, as well as critical surrounding structures, or organs at risk (OARs). The tumour delineated area considered in radiotherapy treatment planning is the planning target volume (PTV). For more information on how the PTV is determined, the reader is referred to Appendix A.2. Once the important structures have been delineated, the TPS requires the criteria we wish our treatment plans to comply with.

1.2.2 Plan criteria

Depending on the clinical indications of the patient, a personalized radiotherapy treatment plan should comply with a set of planning criteria. These criteria may be involved in soft constraints or hard constraints.

Soft constraints

In basic terms, soft constraints involve objectives (which will be discussed in greater detail in Section 1.2.3)) that hold true unless contradicted by another constraint that has a higher priority. Some soft constraints involve the concepts of conformity and homogeneity and the ALARA (As Low As Reasonably Achievable) principle. The conformity describes how well the delivered dose distribution is shaped to the tumour and can be expressed by the conformity index. This index is defined as the ratio between the PTV and the irradiated volume at a specified reference prescription isodose D_{ref}, typically being 95% isodose [Sal+17]. The homogeneity is a measure that assesses the uniformity of a dose distribution and can be expressed by the homogeneity index, defined as the maximum dose in the target volume and D_{ref} [Sal+17]. The ALARA principle represents a practice mandate adhering to the principle of minimizing radiation doses to patients (both healthy tissue and the PTV) as low as reasonably achievable [SK06].

Hard constraints

Hard constraints involve planning indications that absolutely cannot be violated. These indications, as prescribed by the radiation oncologist, are defined such that the tumour is expected to be cured, whilst limiting biological complications, and are typically represented by dose-volume metrics. These are metrics that prescribe a certain amount of dose to a fractional volume of an OAR or the PTV. In terms of prostate planning, for example, plans are hard-constrained to deliver 95% of the prescribed dose to at least 99% of the PTV volume, and 64 Gy to at most 35% of the rectum volume. Such dose-volume metrics can be displayed by means of a dose-volume histogram (DVH).

Dose-volume histograms

The DVH is a way of displaying dose to a structure, however it has lost all spatial information about the dose distribution [Drz+91]. DVHs serve as simple tools to compare treatment plans by presenting dose in the irradiated target and in critical adjacent structures (OARs). Two example DVHs of the rectum OAR and of the PTV of a prostate patient are shown in Figure 1.2. The cumulative DVH is a one-dimensional function that

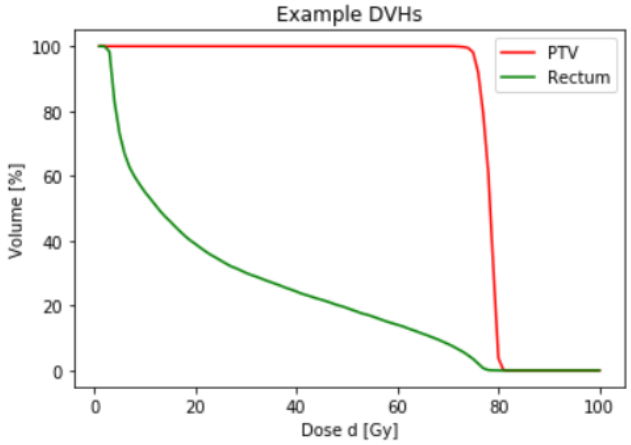


Figure 1.2: Typical example DVHs of a prostate primary PTV (red) and rectum (green)

displays a structure's fractional volume f_d that has received at least a dose d is described by Equation 1.1. Throughout this thesis, there will be a main focus on predicting the dose-volume histogram for prostate cancer patients.

$$f_d = DVH(d) \tag{1.1}$$

Fluence maps and machine parameters

Treatment planning relies on mathematical optimization, such that the tumour receives the prescribed therapeutic dose, while keeping dose to surrounding healthy and OARs, to a minimum [Bal17]. This is achieved by balancing trade-offs involving tumour conformity and homogeneity, and structure dose-volume metrics, leading to an optimal desired fluence map. There are many ways to find a desired dose distribution, and even when one has been found, the challenge of finding the optimal way to deliver this distribution remains. Finding the optimal machine parameter settings that best delivers the desired dose distribution is referred to as machine parameter optimization. However, in radiotherapy treatment planning practice, the balancing of trade-offs refers not only to plan quality in terms of tumour coverage and the sparing of organs at risk, but also to efficiency in terms of plan optimization time and dose delivery time, as well as planning robustness in handling uncertainties. Due to the vast amount of parameters involved, RT treatment planning is a complex problem.

1.2.3 Multi-criteria optimization

A treatment plan contains a description of radiation source locations, beam intensities, duration of dose delivery and beam collimation, and how much dose is prescribed for delivery to the tumour. These quantities are used to calculate the intended dose profiles. Obtaining the desired dose distributions is referred to as fluence map optimization. Fluence maps are two-dimensional maps of beamlet intensities that are found by the optimization algorithm, based on its objective function and constraints. Such optimization problems are typically approached by optimization of a set of conflicting objectives. Two general approaches to planning are forward planning and inverse planning. Forward planning approaches revolve around expert supervision to decide on treatment parameters before computing and evaluating the resulting dose distribution. In inverse planning, a solution on the Pareto frontier (i.e. the set of all solutions that cannot be improved for any criteria without deteriorating other criteria) is found by optimization of one or more objective functions. An objective function contains a mathematical description that maps values of the included variables onto a number, representing a "loss" that is associated with these variables. Optimality is found by minimization of this loss. Contemporary IMRT and VMAT systems use inverse planning approaches with multiple objective functions. When more than one objective function is optimized simultaneously, we refer to it as multi-criteria optimization (MCO). By means of example, objective functions may include the balancing conformity vs. homogeneity, or tumour dose vs. healthy tissue toxicity, each of which can be managed by controlling the values of the variables that are accepted by the optimization. A type of optimization where variables are only allowed strictly within a range is a constrained optimization. There are two main methods for identifying optimal plans with MCO. First, the epsilon-constraint method generally constrains all but one of the objectives to achievable levels, and then minimizes the remaining objective [Cra16]. Second, the weighted-sum method, being more common [BSH09], will be discussed in more detail in the following paragraph.

Weighted-sum optimization model

In a weighted-sum model, all associated objective functions, called objectives, are arbitrarily weighted and combined to form a general objective function. Depending on the mathematical formulation, this can effectively turn a multi-criteria problem into single-criterion objective, meaning that optimization is performed on one common objective function as a whole, rather than on all objectives simultaneously, which is a more complex procedure. Objectives can typically involve dose or dose-volume metrics for different organs, such as for example minimum dose (in case of the PTV), average dose or V_{95} (the volume that receives at least 95% of the prescribed dose). Simultaneous objective optimization is commonly done by using a scalarization approach, where objectives are weighted such that all (non-negative) weights add up to one. An advantage of such an approach is that the relative inter-criterion importance can be understood more easily [Bal17]. With the help of Balvert et al. [Bal17] [Bal+15], let us provide a mathematical description for the basic MCO model.

Basic MCO model

Let us denote a set of OARs by S_{OAR}, and the complete set of relevant tissue structures by $S = S_{OAR} \cup PTV$. All structures are discretized into voxels, and the set of voxels in structure s ∈ S is denoted by I_s, where s can be any delineated structure. Let us define the set Y of all beamlets (intensities that account for the dose deposition along its line of irradiation). Dose rates from each beamlet b to each voxel i are contained in matrix $\dot{\boldsymbol{d}}$ of dimensions, where $b \in [1, 2, \cdots, |Y|]$ and $i \in [1, 2, \cdots, |I_s|]$, where the absolute value brackets denote set cardinality. Vector $\boldsymbol{t}$ contains the beam-on time for each beamlet, such that t_b denotes the beam-on time of beamlet b. While keeping in mind that $\dot{\boldsymbol{d}}_b$ is a vector that contains dose rate from beamlet b to voxel i, it can be seen that the total dose (from all beamlets) to voxel i results from the multiplication $\dot{\boldsymbol{d}}_i^T \boldsymbol{t}$. The prescribed dose to the target is denoted by D_P. The variable u_i describes the difference between the delivered and prescribed dose in voxel i, $D_P - \dot{\boldsymbol{d}}_i^T \boldsymbol{t}$, if the delivered dose is less than D_P and 0 otherwise. Let w_s be the weight assigned to the constraints corresponding to structure $s \in \boldsymbol{S}$, which satisfies $\sum_{s \in \boldsymbol{OAR}} w_s = 1$. w_{PTV} is a value describing the trade-off between PTV dose and OAR tissue sparing. Now, the general optimization model can be formulated:

$$min_{u,t} \quad w_{PTV} \frac{1}{|I_{PTV}|} \sum_{i \in I_{PTV}} u_i + (1 - w_{PTV}) \sum_{s \in S_{OAR}} w_s \frac{1}{|I_s|} \sum_{i \in I_s} \dot{\boldsymbol{d}}_i^T \boldsymbol{t} \tag{1.2}$$

$$\begin{aligned} s.t. \quad u_i &\geq D_P - \dot{\boldsymbol{d}}_i^T \boldsymbol{t} && \forall i \in I_{PTV} \\ u_i &\geq 0 && \forall i \in I_{PTV} \\ t_b &\geq 0 && \forall b \in Y \end{aligned}$$

It should be noted that this is only a general model that aims to minimize per-voxel-average under-dose to the PTV and per-voxel-average over-dose to an arbitrarily weighted set of OAR criteria. The model can be extended to include, for example, restrictions on PTV over-dose or on the minimum, mean and maximum dose D_{min}, D_{mean} and D_{max}. The optimization problem described here is a convex optimization problem [Bal17] [Bal+15], so the Pareto frontier is guaranteed to be found and can be navigated by adjusting w_{PTV} and w_s.

1.2.4 Treatment plan segmentation

Inverse planning approaches yield fluence maps for each tumour projection. However, even when fluence map optimization has yielded an ideal dose distribution, the optimal machine parameter settings to deliver this distribution remain unknown. These settings are found by machine parameter optimization, or plan segmentation. Machine parameters include for example MLC positions, dose rates, gantry angles or (for VMAT) rotation speed. There are two main problems that are accompanied by machine parameter optimizations: non-convexity (hardware derived) of the optimization, and deterioration of time-efficiency. Hardware derived non-convexities arise from the practical desire to use a small number of MLC segments when delivering treatment plans [Cra16]. In addition, the MLC leaf thickness limits the number of deliverable segments. Two reasons for the deterioration of time-efficiency when delivering treatment plans are: first, although advancements in MLC technology have sped up inter-segment MLC-adjustments, the adjustment speed can still be an issue. As illustration, if the MLC leaf positions of two subsequent segments require the MLC leafs to travel large distances, the time required to deliver the plan will grow longer, and time-efficiency becomes an issue. Second (only for VMAT), subsequent MLC states are not independent, because not every MLC configuration can be attained from a previous MLC configuration, for time-efficiency reasons, making it impossible to always achieve ideal dose distributions for every patient.

Mathematical optimization techniques

Non-convexity of plan segmentation optimization discussed in Section 1.2.4 is problem that is always inherent to RT treatment planning. Although there are reasons to believe that resulting local minima are good approximations of global minima [Web03], extensive efforts have been made to ensure that optimizers converge to the best plan possible.

Sequential quadratic programming One approach used in this study is sequential quadratic progamming (SQP). SQP relies on a quadratic approximation of the objective function and the constraints [GMW81]. For a more detailed description of SQP, the reader is referred to [Bar+09].

Simulated annealing Another conventional approach is simulated annealing. Simulated annealing is based on exploring the area of a local minimum to converge to an even better spolution, before accepting the yielded result as the best one globally [Web89].

Lexicographic ordering Lexicographic optimization is a type of hierarchical prioritized optimization. For instance, lexicographic ordering provides a way to deal with a large number of competing clinical trade-offs by attempting to compromise less important criteria before more important ones such that a more optimal cure versus toxicity trade-off can be obtained. This makes complex multi-criteria problems more manageable and potentially makes the planning process more efficient [JMF07].

1.2.5 Biological models

Models that provide quantitative biophysical measures can also be utilized for cost functions [Nie98] [Nie97]. For additional information on radiobiology and an overview of some of the commonly used biological models, the reader is referred to appendix A.3. In this study, there is one biological model that we used, which is the generalized equivalent uniform dose (gEUD).

Generalized equivalent uniform dose

The gEUD is defined as the uniform dose that, when irradiated homogeneously and given over the same number of fractions, yields the same radiobiological effect as the actual non-homogeneous absorbed dose distribution does and can be written as [Wan+16]:

$$gEUD = \big(\sum_{i=1}^{V} f_i D_i^a\big)^{\frac{1}{a}} \tag{1.3}$$

where V is the total number of voxels in the structure considered, f_i is the fractional volume receiving a dose D_i, and a is a tissue-specific EUD-parameter. The practical use of the gEUD is that by changing parameter a, we can control the importance of a certain dose such that it counts more heavily towards the weighted average dose, which can in turn be incorporated in a cost function. For example, low doses are more important when an a-parameter closer to 0 is used, and vice versa for higher doses. It can be noted that in the limit $\lim_{a\to\infty} gEUD$, the gEUD converges to the maximum dose. Similarly, in the limit $\lim_{a\to 0^+} gEUD$, the gEUD converges to the smallest dose. If $a = 1$, the gEUD equals D_{mean}. In the case of the PTV, negative a is used [Nie97]. This essentially reverses the effect of the weighted average.

1.2.6 Prostate planning

The prostate is an organ that is part of the male's preproductive system and is located anterior to the rectum in the lower abdomen region. The prostate receives mature sperm cells from the testes, and secretes nutrients and buffers that protect the sperm against the acidic vaginal secretions [WRS11a]. The seminal vesicles (SVs), two glands that are attached posterior to the prostate, secrete chemicals that increase sperm motility and prostaglandins [WRS11a]. Together, the mature sperm cells and the secretions from the prostate and seminal vesicles make up the fluid referred to as semen [WRS11a]. A schematic drawing of the prostate and seminal vesicles is shown in Figure 1.3. The prostate is planned clinically in three separate groups that each have different planning objectives. Patients are distinguished by the radiation oncologist, based on their individual planning indications. All groups are planned with a primary PTV and a secondary PTV. The primary is

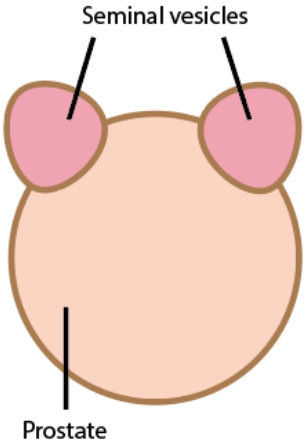


Figure 1.3: Schematic depiction of the prostate (beige) and seminal vesicles (pink).

(a) Patient group 1 (b) Patient group 2 (c) Patient group 3

Figure 1.4: Schematic depictions of the three distinguishable patient groups. The inner red area represents the surdose (primary) PTV, and the outer yellow area represents the secondary PTV.

also referred to as the surdose or boost PTV. The primary PTV consists of the CTV plus a 4 mm margin and is prescribed 77 Gy, whereas the secondary PTV consists of the primary PTV plus a 3 mm margin and is prescribed 70 Gy. In the first group, the seminal vesicles (SVs) were excluded from delineation in either of the PTVs. In the second group, the SVs were included in only the secondary PTV. As for the third group, the SVs were included in the primary PTV and thus also the second PTV. A visual interpretation of the patient group division can be seen in Figure 1.4.

1.2.7 Clinical situation: problem posing

Finding a clinically acceptable treatment plan is not trivial, since RT treatment planning is an intrinsically difficult problem due to the vast amount of parameters that can be controlled, and the conflicting objectives of achieving a uniform dose to the tumour while limiting dose in healthy tissues [JDK17]. Therefore, the expertise and experience of the treatment planner is heavily relied on, introducing subjectivities and increasing patient risk. Furthermore, the present treatment planning framework is laborious, time-consuming, and it is difficult to assess the quality of a certain plan, or if there is a plan with a more optimal trade-off between cure and toxicity. For these reasons, it becomes impossible to always achieve ideal, personalized treatments, leading to acceptable but potentially sub-optimal treatment plans. To resolve these problems, an automated approach to treatment planning is used. At the NKI, we clinically use Philips Pinnacle's autoplanner. However, autoplanning introduces new challenges. Namely, autoplanning is not perfect and still requires substantial interaction from the RT technician. In addition, it is next to impossible to judge whether a resulting plan is indeed the best personalized plan, since no RT technician has manually worked the plan. As a consequence, there is a need for standardized quality assurance tools for detecting outliers. Moreover, if optimal treatment plans could be used as feedback for new plans, we could potentially increase overall plan quality. Knowledge-based planning (KBP) could potentially provide a tool that resolves these issues. This thesis focuses on KBP for prostate cancer.

1.3 Knowledge-based planning

Knowledge-based planning approaches use both historical structural and dosimetric data from prior treatment plans to predict the likely dose distribution for the current patient. KBP does this by comparing the internal geometry of a given patient with patients treated in the past, and uses this knowledge to propose a plan for the current patient. In this way one can hope that, if patients treated in the past had an optimal trade-off, new patients will get this as well [JDK17]. Thus, KBP makes for an objective, automated and patient-specific approach that ensures a realistic and achievable treatment plan. Such methods are believed to enhance plan consistency, quality, standardization and planning efficiency. Throughout this thesis study we focus on KBP for prostate cancer planning.

1.3.1 Overlap volume histogram

Experience has shown that optimality of treatment plans is strongly influenced by the geometries of critical structures with respect to the target volume [Web03]. Specifically, an OARs' proximity to, or quite regularly even overlap with the PTV can be parameters of interest. For the purpose of studying the influence of the OARs' proximity to the PTV on its received dose, the overlap-volume histogram (OVH) was introduced [WRS11b] [Wu+13] [Pet+12]. The OVH is a one-dimensional function that describes the fraction of the OAR volume that is encompassed by a uniform expansion or contraction of the PTV by distance r [WRS11b] [Wu+13] [Pet+12]:

$$f_r = OVH(r) \tag{1.4}$$

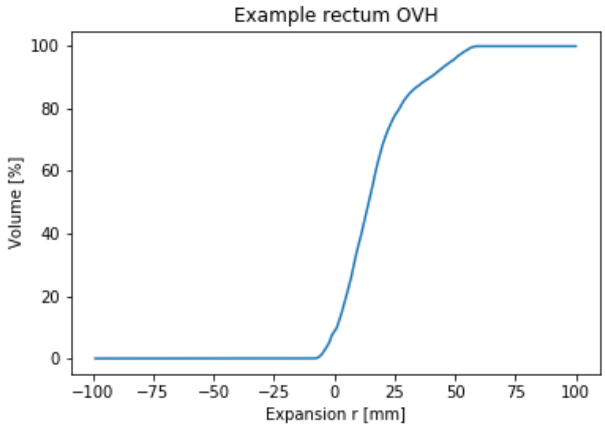


Figure 1.5: Example OVH for the rectum. The OVH was calculated from the secondary PTV. The OVH(r = 0) can be regarded as the PTV border

In its Euclidean form, the OVH value f_r at distance r is given by the fraction of the OAR voxels with its maximum distance to the PTV boundary less than r. Let us remember the discretized voxel representation introduced for discussing the basic MCO model in Section 1.2.3. If s denotes the delineated structure, the set of all voxels in structure s is denoted by I_s. Let voxel i in structure s within distance r from the PTV boundary be $v_i^s \in I_s$. Similarly, let voxel k in the PTV be $v_k \in I_{PTV}$. Let V be the voxels that define the surface

of the PTV. The distance function ρ from voxel v_i^s to the PTV surface is then formulated as:

$$\begin{aligned} \rho_s^{PTV} &= d(v_i^s, I_{PTV}) \\ &= min_k\{\parallel v_i^s - v_k \parallel \mid v_k \in V\} \end{aligned} \tag{1.5}$$

OAR voxels within the PTV volume are denoted by negative expansion. A more detailed description of the OVH has been given previously ([Kaz+09] [Wu+09]). It can be argued that in the case where a smaller expansion distance is needed to reach a certain OAR overlap, sparing it is more difficult [Yua+12]. An important assumption of the OVH model in KBP is that dose to OAR voxels decreases with increasing distance from the PTV.

1.3.2 KBP literature overview

For the sake of providing an outline of previous research, we will briefly discuss some studies that involve KBP. We discuss three general KBP-approaches: OVH-based methods, methods involving the projections of delineated structures, and methods involving mathematical frameworks that do not belong to the first two classes.

OVH-methods

There have been many studies that have combined historical data with OVH-methods for DVH prediction ([WRS11b] [Wu+13] [Moo+11] [Yua+12] [Pet+12] [Wan+16]). [WRS11b] used the OVH to generate achievable DVH objectives for head-and-neck plans as initial planning goals. [Wu+13] investigated the use of OVH for automated VMAT planning in head-and-neck patients. They did this by using DVH objectives estimated from historical IMRT plans as optimization parameters for VMAT plans. [Moo+11] et al. have used OVH-information to predict OAR dose metrics for head-and-nack and prostate IMRT. Yuan et al [Yua+12] used the OVH to explain inter-patient DVH variability in head-and-neck and prostate cancer OARs. They used the first three principal component modes to represent the OVH and DVH. Through this method, they identified three important factors that explain a significant amount of inter-patient DVH varability, being the *mean OAR-PTV distance*, *OVH metrics* and *out-of-field OAR volume*. [Pet+12] demonstrated that a prior lexicographic ordering model for head-and-neck patients could be used to predict the achievable dose to an abdomen OAR for a new pancreatic tumour patient. For this, the OVH was used. They assumed that the minimal achievable OAR dose depends mainly on its distance to and orientation with respect to the PTV, which holds for a typical prostate case. They put this assumption to the test in head-and-neck cases. Wang et al. [Wan+16] generated a ground-truth set of consistently planned Pareto-optimal treatment plans for prostate patients, using lexicographic MCO. They then proceeded to use an OVH-based KBP method on this ground-truth data set, improving planning standardization and preventing validation with possibly suboptimal benchmark plans.

Structure projections

Some studies use the best matching OAR and PTV projections vor DVH prediction. Chanyavanich et al. [Cha+11] demonstrated the use of a knowledge base of prior, clinically approved archived plans for the creation of new, also clinically acceptable plans, by finding the best matching reference case in the data base by matching OAR and PTV projections.

Subsequently, treatment parameters of this best matching plan are used as a starting point for planning. Similarly, [Goo+13] used KBP approaches that involved matching query plans to the most similar reference OAR and PTV projections. Then, the treatment parameters of the corresponding plan were taken and further individualized by applying a deformable registration. Matching plans were then compared for PTV homogenieties, which were found to be significantly lower in KBP plans.

Other mathematical frameworks

Lastly, there are studies that use other OAR-specific mathematical frameworks to do DVH prediction. For example, [App+12] divided OARs into several sub-volumes. Then, they determine the parameters of a skew-normal distribution by regression for each sub-volume. In turn, the results of these regression models combined yield the DVH prediction.

Connection to this research

Although there are endless approaches to knowledge-based planning, KBP approaches commonly involve the use of the OVH in some way to encode for patient anatomies. In addition, reduced order modeling (ROM) (will be discussed in Section 3.2) is widely dispatched to capture the most important characteristics of the DVH and OVH. Throughout this thesis study, we will follow these directives.

1.3.3 Aims and objectives

As a part of this thesis work, we have developed a relatively simplistic KBP tool, which will be discussed in detail in Section 4.3. This tool has has yielded sufficient results to be used in a pilot study at the NKI, and has been in use since October 2018. This model is what will be referred to as the current clinical practice throughout this study. The aim of this thesis project is to build a machine learning model that is able to predict the DVH for new prostate patients, from their anatomical information and thereby to better understand how to complement the current clinical practice. To do this, we will investigate a variety of different machine learning modelling approaches in a KBP context. To evaluate these models, we have formulated a quantitative, clinical goal (Section 4.2) to which our models should comply. In short, this formulation entails that our goal is achieved when we create a model that successfully predicts 90% of the DVH bins for 90% of patients. However, as we are most interested in testing accuracy, we will only use models that are believed to retain a sufficient degree of generalizability.

CHAPTER 2

DATA

This chapter will cover the characteristics of the data that were used throughout this study. This includes by which means the data were acquired, before being ready for use in model development. This chapter also discusses how the data were selected, and describes the characteristics of the data sets.

2.1 Data collection

2.1.1 Acquisition

Retrospective analyses were performed based on clinical VMAT plans and planning CTs of prostate cancer patients treated at the NKI. Philips Pinnacle^3 was used for creating VMAT plans, where optimization was done according to RaySearch White Paper guidelines [Eri+09]. DVHs and OVHs were calculated from the available DICOM (Digital Imaging and Communications in Medicine) files stored in PACS (Picture Archiving and Communication System). All geometrical information (OVHs and organ sizes) and DVHs were matched by their Medical Record Number and Unique Plan Identifier.

2.1.2 Selection

The data set used for this study consisted of the combined data of three patient groups as discussed in Section 1.2.6. The sample size of group 1 alone was considered too small for reliable model development and was therefore omitted from this study. Also, despite the differences in inter-group planning indications, intermediary analyses have shown that groups 2 and 3 did not exhibit significant inherent differences in their secondary OVH and rectum DVHs. For these two reasons, we focused on model development for the unification of groups 2 and 3, and omitted group 1. Structures of interest were the rectum and anal sphincter. The bladder was not believed to be of importance for planning purposes for three reasons: first, the majority of the dose ends up in the urine, which in turn is excreted from the body. Second, the complications caused by damage to the bladder wall are not believed to be of primary concern, compared to other OARs. Third, sparing the bladder wall often proves to be difficult in practice and typically compromises other planning objectives. However, there are exceptional cases in which it was not deemed possible by the radiotherapy technician to ignore the bladder as an OAR. These cases were omitted from model development. Furthermore, patients with a bowel-loop or hip- or femoral prosthesis are excluded from investigations. The reason for this is to ensure that model development is done with data that best represents the the average cases. This way, model development starts off simple, but exceptional cases, such as those including a bowel-loop or femoral prosthesis, can be included when more complex models are constructed.

2.1.3 Data set specification

After selection of our data, a total of 92 "good" patients remained for model development. Of this, 51 and 41 patients belonged to group 2 and group 3 respectively. For the sake of consistency, the plans within this data set contains plans created with Pinnacle's auto-planner. DVHs and OVHs were calculated with a dose resolution of 1 Gy and a spatial resolution of 1 mm respectively.

CHAPTER 3

THEORY

With machine learning being one of the cornerstones of this study, there are several topics involved that require some mathematical background to understand their working principle. Some of the general topics relevant to machine learning are regression analysis, feature extraction, feature selection and model validation methods. In this chapter, we will go over the mathematical bases of some regression methods, Principal Component Analysis (PCA), and we will discuss the rationale of validation methods used to validate our results.

3.1 Regression

3.1.1 Basis function regression

Regression is used to find a general trend that exists between a reponse (dependent) variable y, and one or more explanatory (independent) variables $\boldsymbol{x}$. All regression models have the same basic form:

$$y = f(\boldsymbol{x}) \tag{3.1}$$

where f($\boldsymbol{x}$) involves a set of regression parameters or coefficients that need to be fit. How such a fit is found will be discussed in Section 3.1.4. Let us imagine a problem with a single explanatory input variable x. We can write Equation 3.1 in terms of the regression coefficients, for the one-dimensional case [HFB15]:

$$y = f(x) = \sum_{i=0}^{I} w_i \phi_i(x) \tag{3.2}$$

where the functions $\phi_i(x)$ are called basis functions, w_i are the regression coefficients, and I is an arbitrary maximum number of basis functions. For convenience, let us express $\phi_i(x)$ and w_i in vector notation: the basis function vector $\Phi(x) = [\phi_0(x),\phi_1(x),\cdots,\phi_I(x)]^T$, and weight vector $\boldsymbol{w} = [w_0, w_1, \cdots, w_I]^T$. Equation 3.2 can be rewritten as the multiplication of the weight vector transpose and basis function vector:

$$f(x) = \boldsymbol{w}^T \Phi(x) \tag{3.3}$$

Basis functions should be chosen such that they fit the regression problem at hand. As illustration, in an experiment with a known linear dependence between x and y, it would make sense to use basis functions linear in x to predict y. However, especially in machine learning, dependencies are not always known. So, basis functions are typically decided upon by investigating the correlations that exist in the data. Alternatively, when no a priori knowledge about the data's dependency is available, one may choose to use basis functions that allow for more flexibility. These funcions may involve radial basis functions, support vector machines or artificial neural networks. Regressors using such models are typically referred to as being non-parametric, because the free parameters that need to be determined have no real relation to the regression problem, as opposed to parametric basis functions [Orr96].

3.1.2 Linear regression

One of the most basic forms of regression is simple linear regression. It is called simple when there is only a single input variable involved. In this case, it suffices to assume a basis function with the monomial bases: $\Phi(x) = [1, x]^T$. The result of Equation 3.3 becomes the standard simple linear model:

$$y = w_0 + w_1 x \tag{3.4}$$

where w_0 and w_1 are the intercept and slope regression coefficient. If we wish to expand this to the M-dimensional case, we have to choose $\phi_m(\boldsymbol{x}) = x_m$, for m = [1, 2, $\cdots$, M]. When substituting this into Equation 3.2, the regression model becomes that of multiple linear regression:

$$y = \sum_{m=0}^{M} w_m \phi_m(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_M x_M \tag{3.5}$$

Let the reader be informed that $\phi_0(x) = 1$ will be used in all regression models from here.

3.1.3 Polynomial regression

One of the most common basis functions used in regression is a polynomial. The basis function used for Equation 3.4 is a first degree polynomial. In order to extend this model to the Q-th order polynomial case, we would have to use the bases $\phi_q(x) = x^q$, for q = [0, 1, 2, $\cdots$, Q]. Substituting this into Equation 3.2 will yield the one-dimensional Q^{th} order polynomial regression model:

$$y = \sum_{q=0}^{Q} w_q \phi_q(x) = w_0 + w_1 x + w_2 x^2 + ... + w_Q x^Q \tag{3.6}$$

The M-dimensional polynomial is more complex, because we cannot describe either the basis functions or the regression coefficients with a single index. Instead, the indexer in 3.6 has to become multi-index: $\hat{q} = [q_1, q_2, \cdots, q_M]$, and the sum over q becomes a sum over all indices in $\hat{q}$ with $q_1 + q_2 + \cdots q_m \leq Q$ [Kon04]. The resulting general M-dimensional, Q^{th} order polynomial regression model and an example with M = 3 and Q = 2 are shown in Equations 3.7 and 3.8 respectively:

$$y = \sum_{(q_1,q_2,\cdots,q_M)}^{Q} w_{(q_1,q_2,\cdots,q_M)} \phi_{(q_1,q_2,\cdots,q_M)}(\boldsymbol{x}) \tag{3.7}$$

$$\begin{aligned} y = & \sum_{(q_1,q_2,q_3)}^{2} w_{(q_1,q_2,q_3)} \phi_{(q_1,q_2,q_3)}(\boldsymbol{x}) = \\ & w_0 + w_{(1,0,0)} x_1 + w_{(0,1,0)} x_2 + w_{(0,0,1)} x_3 + \\ & w_{(1,1,0)} x_1 x_2 + w_{(1,0,1)} x_1 x_3 + w_{(0,1,1)} x_2 x_3 + \\ & w_{(2,0,0)} x_1^2 + w_{(0,2,0)} x_2^2 + w_{(0,0,2)} x_3^2 \end{aligned} \tag{3.8}$$

3.1.4 Cost function minimization

The linear and polynomial regression methods discussed in sections 3.1.2 and 3.1.3, are examples of parametric regression methods. In parametric regression, models are fitted by minimizing a cost function. Most regularly, least squares minimizers (or ordinary least squares (OLS)) are used as a cost function, but absolute deviations may also be used. OLS refers to finding a fit such that the total squared vertical discrepancies between this fit and all given sample points are minimized. OLS-based methods are convenient for a two reasons. First, it ensures that outliers that are very far off weigh more heavily in the cost function. Second, when used in optimization, squared objective functions typically have the convenient property that they are convex. This ensures the optimizer to reach a global optimum. Another convenient property of OLS-solutions, is that they can be found relatively easily with the help of the Moore-Penrose pseudoinverse [Weixxa] [Pen56]. Let us explore how OLS methods work mathematically.

Least squares methods

Let us assume a two-variable linear regression model. If $\hat{y}_n$ is the n^{th} observed data point $\hat{y}$, the cost function is the total sum of squared residuals, RSS:

$$\begin{aligned} RSS &= \sum_{n=1}^{N} (\hat{y}_n - y_n)^2 \\ &= \sum_{n=1}^{N} \left[\hat{y}_n - \left(w_0 + w_1[x_1]_n + w_2[x_2]_n\right)\right]^2 \end{aligned} \tag{3.9}$$

The set of regression coefficients $\boldsymbol{w}$ that result in residuals minimization are found by setting their partial derivatives to 0 and solving the resulting system of linear equations:

$$\frac{\partial(RSS)}{\partial w_0} = -2 \sum_{n=1}^{N} \left[\hat{y}_n - \left(w_0 + w_1[x_1]_n + w_2[x_2]_n\right)\right] = 0 \tag{3.10}$$

$$\frac{\partial(RSS)}{\partial w_1} = -2 \sum_{n=1}^{N} \left[\hat{y}_n - \left(w_0 + w_1[x_1]_n + w_2[x_2]_n\right)\right][x_1]_n = 0 \tag{3.11}$$

$$\frac{\partial(RSS)}{\partial w_2} = -2 \sum_{n=1}^{N} \left[\hat{y}_n - \left(w_0 + w_1[x_1]_n + w_2[x_2]_n\right)\right][x_2]_n = 0 \tag{3.12}$$

As long as there are as many equations as there are free variables, this is a solveable system of linear equations. It can be shown that by using the Moore-Penrose pseudoinverse, the solution to such a system of linear equations (3.10 - 3.12) can be automatically found [Weixxa] [Weixxb] [Pen56]. We will now explain why this works mathematically, with the help of [TK09]. Let $[\boldsymbol{x}]_n$ be the n-sample vector containing all M variables. Minimizing Equation 3.9 with respect to $\boldsymbol{w}$ results in the following vector notation:

$$\begin{gathered} \sum_{n=1}^{N} (\hat{y}_n - [\boldsymbol{x}]_n^T \boldsymbol{w})[\boldsymbol{x}]_n = 0 \longrightarrow \\ \left(\sum_{n=1}^{N} [\boldsymbol{x}]_n [\boldsymbol{x}]_n^T\right) \boldsymbol{w} = \sum_{n=1}^{N} ([\boldsymbol{x}]_n \hat{y}_n) \end{gathered} \tag{3.13}$$

We can use the matrix notation:

$$X = \begin{bmatrix} [\boldsymbol{x}]_1^T \\ [\boldsymbol{x}]_2^T \\ \vdots \\ [\boldsymbol{x}]_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{bmatrix}, \quad \hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} \tag{3.14}$$

Here, X is an N x M matrix which rows are the sample-specific feature vectors, and $\boldsymbol{y}$ is a vector that contains the corresponding observed responses. From Equation 3.14 it can be seen that:

$$\sum_{n=1}^{N} [\boldsymbol{x}]_n [\boldsymbol{x}]_n^T = X^T X \tag{3.15}$$

$$\sum_{n=1}^{N} [\boldsymbol{x}]_n \hat{y}_n = X^T \hat{\boldsymbol{y}} \tag{3.16}$$

where $\Sigma = X^T X$ is the covariance matrix. Hence, Equation 3.13 can now be rewritten to yield the Moore-Penrose pseudoinverse matrix notation of the OLS approach to linear regression.

$$(X^T X)\hat{\boldsymbol{w}} = X^T \hat{\boldsymbol{y}} \longrightarrow \hat{\boldsymbol{w}} = (X^T X)^{-1} X^T \hat{\boldsymbol{y}} \tag{3.17}$$

Here, the pseudoinverse $X^+ = (X^T X)^{-1}$ is meaningful only if Σ is invertible, that is, of rank M [TK09].

3.2 Dimensionality reduction

As a result of an increased demand for machine learning approaches, as well as the computational resources to handle large data sets, comes a surge of data that is used in data analysis. Although having more data is typically desirable, the implications of having too many descriptive features at ones disposal is three-fold. First, features may hold no additional descriptive value over other features: they can be redundant. Second, additional features may correlate in no way to the response we are interested in: they are irrelevant. Third, features are more likely to cause overfitting as their number grows, and thereby reduce model generalization (referring to Section 3.3). Additionally, data require storage and large data sets increase processing time as well as complexity, hindering straightforward interpretation. For these reasons, dimensionality reduction methods are used to decrease the number of variables. Fundamentally, there are two steps that can be distinguished in dimensionality reduction: feature extraction and feature selection [RS00].

3.2.1 Feature extraction

In feature extraction we start from an initial set of features, and subsequently redefine them with the intention for them to be more informative. This can be referred to as reduced order modeling (ROM). A widely exploited ROM method is Principal Component Analysis (PCA), sometimes also referred to as the Karhunen-Loève (KL) transformation. These are terms that are often used interchangeably, but are not equivalent. In order to understand how the KL transformation and PCA are distinguished, let us imagine a binary classification problem where we aim to classify some object $\boldsymbol{O}$ as either of the two classes. Suppose that $\boldsymbol{O}$ can be represented by a number of m characteristics (features) in an m-dimensional vector called the feature vector. If m is large, it means that we are dealing with a high dimensional classification problem, whereas some characteristics may be more important than others for correct classification of $\boldsymbol{O}$. This makes the problem unnecessarily complex and hence it would be desirable to have a method that reduces dimensionality without compromising the classification accuracy of $\boldsymbol{O}$. The KL transformation and PCA provide such a method. The primary purpose of the KL transformation is to reduce the dimensionality of a data set that contains interrelated variables into a smaller set of mutually uncorrelated variables. PCA is then used to identify the features that retain most of the variation among the data [Owe14].

Karhunen-Loève transform

The KL transformation essentially does a re-mapping of the original coordinates in which the data are expressed into a more "meaningful" basis of coordinates, such that separability among these coordinates can be maximized. This transformation generates mutually

uncorrelated features (but not necessarily independent) [TK09]. The new coordinate bases are referred to as the principal components (PCs). Let us now explain why the KL transformation works mathematically with the help of Theodoritis et al. [TK09]. Initially, let us assume zero sample means. Let us define an input sample matrix (m measurements x n samples) of random variables $\boldsymbol{x}$, that is to be expressed in a new (m x n) basis of coordinates $\boldsymbol{c}$:

$$\boldsymbol{c} = A^T \boldsymbol{x} \tag{3.18}$$

with A being a square (m x m) matrix which rows form the new basis for $\boldsymbol{x}$. Let the reader be reminded that a correlation and covariance matrix describe the degree to which two random variables can deviate from their respective means. From linear algebra we know that the correlation matrix of $\boldsymbol{x}$ can be expressed as the expectation of the outer product of $\boldsymbol{x}$ with its transpose, and similarly for $\boldsymbol{c}$:

$$R_x \equiv E[\boldsymbol{x}\boldsymbol{x}^T] \tag{3.19}$$

$$R_c \equiv E[\boldsymbol{c}\boldsymbol{c}^T] \tag{3.20}$$

Then, by substituting Equation 3.18 in the definition of the correlation matrix, we get:

$$R_c \equiv E[\boldsymbol{c}\boldsymbol{c}^T] = E[A^T \boldsymbol{x}\boldsymbol{x}^T A] = A^T R_x A \tag{3.21}$$

and it can be seen that, if A is chosen such that its columns contain N orthonormal eigenvectors $\boldsymbol{a}_i$ of R_x (with i = 0, 1, ..., N-1) of length m, then R_y is the diagonal eigenvalue matrix $\boldsymbol{\Lambda}$:

$$R_c = A^T R_x A = \boldsymbol{\Lambda} \tag{3.22}$$

where $\boldsymbol{\Lambda} = \boldsymbol{I}\boldsymbol{\lambda}$, where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{\lambda}$ is an eigenvalue vector that contains the eigenvalue λ_i of each respective eigenvector $\boldsymbol{a}_i$. This resulting transformation is what is known as the Karhunen-Loève transformation. In the typical case when the zero mean assumption is not valid, the sample means needs to be subtracted. In summary, with the KL transformation, we achieve a new, orthonogal basis of correlation matrix of $\boldsymbol{x}$, by means of an eigenvalue decomposition. However, we have not yet explained how we can identify the characteristics that explain the largest variation in the data.

Principal component analysis

Once the principal components of R_x have been found, they need to be structured, such that we can identify the fraction of the total variance in the data that they account for. This fraction is what is referred to as the explained variance ratio (EVR). In terms of the problem in Section 3.2.1, PCA is a linear transformation that retains most of the total variance associated with an original random variable vector $\boldsymbol{x}$, by approximating it with a smaller subset of vectors, thereby finding the explained variance ratios of each PC. The vectors in this subset turn out to be the eigenvectors, or principal component modes, of R_x, such that a minimal amount of variance is lost. Let us select P eigenvectors, such that we approximate $\boldsymbol{x}$, from its projection $\widehat{\boldsymbol{x}}$ spanned by the P orthonormal eigenvectors involved:

$$\widehat{\boldsymbol{x}} = \sum_{i=1}^{P} \dot{y}_i \boldsymbol{a}_i \tag{3.23}$$

where, $\dot{y}_i = \boldsymbol{a}_i^T \boldsymbol{x}$. Generally, the mean square error (MSE) estimate $\tilde{\boldsymbol{z}}$ of a random variable $\boldsymbol{z}$ with mean $\overline{\boldsymbol{z}}$ is given by:

$$\tilde{\boldsymbol{z}} = \arg\min_{\overline{\boldsymbol{z}}} E\big[\parallel \boldsymbol{z} - \overline{\boldsymbol{z}} \parallel^2 \big] \tag{3.24}$$

So, if we try to approximate $\boldsymbol{x}$ by its projection $\widehat{\boldsymbol{x}}$, the resulting MSE is given by:

$$\tilde{\boldsymbol{x}} \approx E\big[\, \| \boldsymbol{x} - \widehat{\boldsymbol{x}} \|^2 \,\big] = E\left[\left\| \sum_{i=P+1}^{N} \dot{y}_i \boldsymbol{a}_i \right\|^2\right] \tag{3.25}$$

The goal is then to identify the $p \leq P$ eigenvectors that best approximate $\boldsymbol{x}$ (i.e. minimizes the MSE). From Equation 3.25 and by remembering the orthonormality property of eigenvectors $\boldsymbol{a}_i$, it can be showed that:

$$\begin{aligned} E\left[\left\| \sum_{i=P+1}^{N} \dot{y}_i \boldsymbol{a}_i \right\|^2\right] &= E\left[\sum_{i=P+1}^{N} \sum_{j=P+1}^{N} (\dot{y}_i \boldsymbol{a}_i^T)(\dot{y}_j \boldsymbol{a}_j) \right] \\ &= \sum_{i=P+1}^{N} E[\dot{y}_i^2] = \sum_{i=P+1}^{N} \boldsymbol{a}_i^T E[\boldsymbol{x}\boldsymbol{x}^T] \boldsymbol{a}_i \end{aligned} \tag{3.26}$$

If we combine this result with the MSE in Equation 3.25 and remember the eigenvector definition, it can be seen that:

$$E\big[\, \| \boldsymbol{x} - \widehat{\boldsymbol{x}} \|^2 \,\big] = \sum_{i=P+1}^{N} \boldsymbol{a}_i^T \lambda_i \boldsymbol{a}_i = \sum_{i=P+1}^{N} \lambda_i \tag{3.27}$$

Thus, if we select the eigenvectors in Equation 3.23 corresponding to the P largest eigenvalues of the correlation matrix, we are left with the minimized MSE from Equation 3.27, being the sum of the $N - P$ smallest eigenvalues. The PCs that account for the largest EVR can simply be found by selecting the PCs with the largest corresponding eigenvalues. However, the derivation in this section has the advantage over the derivation in Section 3.2.1 that the *first* P components are chosen, because they are ensured to describe the highest variance out of all PCs.

PCA additional remarks

In conclusion, the advantage of PCA is that it re-maps a data set into mutually uncorrelated variables while retaining most of the variation in the data in the first "few" PC modes. One difficulty in practice is how to choose the number of PCs to include. In addition, there are some limitations to PCA which may make it less attractive to use for certain applications. First, it should remembered that PCA generally realizes purely linear output mappings, whereas some applications (e.g. some neural networks) require non-linearity from its input features [Kar94]. Second, PCA relies purely on covariances or correlations, which can only describe completely Gaussian and stationary processing options [Kar94], although this does not mean Gaussianity is a prerequisite for PCA [TK09], [Kar94]. Third, PCA outputs are mutually uncorrelated but not independent, which is a stronger condition than uncorrelatedness [TK09].

3.2.2 Feature selection

As opposed to feature extraction, in feature selection we make do with the features currently at our disposal, and select the most meaningful ones. This helps to simplify the data, shorten the time required for model training and increase model generalization (decrease overfitting). Three approaches to feature selection are filter methods, wrapper methods and embedded methods. In this study, mainly a combination between filter and wrapper methods were used.

Filter methods

The filter method is an information gain approach. Let us explain the basis for filter methods with the help of Scheaffer et al. [SMM11], and let the reader be reminded of the two-variable regresion model discussed in Section 3.1.4. Filter approaches to feature selection use correlations between the independent and response variables to quantify the importance of features. The metric used for this is the coefficient of determination R^2:

$$\begin{aligned} R^2 =& 1 - \frac{\sum_{n=1}^{N} (\hat{y}_n - y_n)^2}{\sum_{n=1}^{N} (\hat{y}_n - \overline{y}_n)^2} \\ =& 1 - \frac{SSE}{SS_{yy}} = \frac{RSS}{SS_{yy}} \end{aligned} \tag{3.28}$$

Here, the SSE being the sum of squared errors between the regressor and observed values, SS_{yy} being the sample variation of y (total sum of squares), and RSS being the regressor summed squares [SMM11]. $R^2 = 0$ implies a complete lack of fit of the model to the data, whereas $R^2 = 1$ implies a perfect fit.

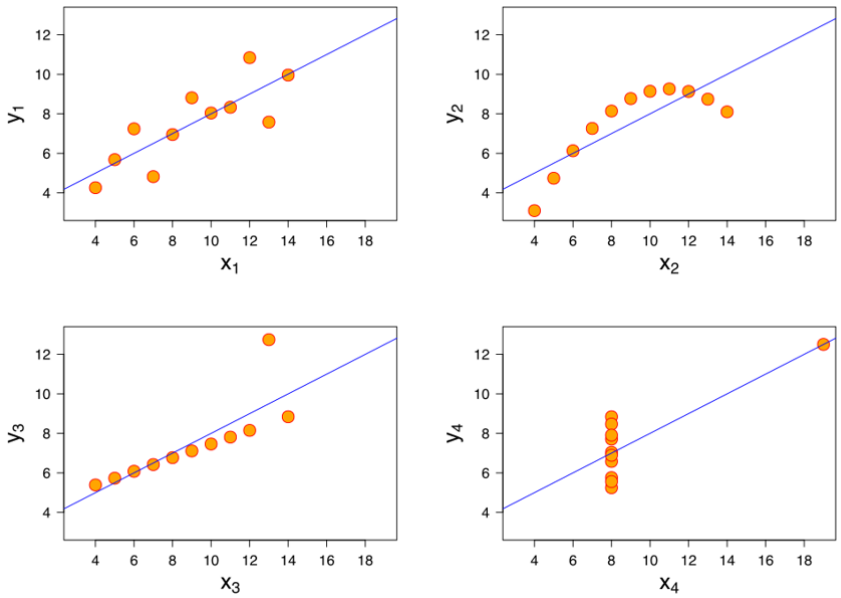


Figure 3.1: Four typical linear regression programs, named Anscombes quartet [Ans73], each yielding the same standard outputs (e.g. means, regression coefficients and R^2). Image adapted from [Asa18])

The main danger of this approach is that correlation coefficients can be rather misleading, as is illustrated from Anscombe's quartet where clearly different relations result in the same standard outputs [Ans73], as displayed in Figure 3.1. This can lead to the erroneously accepting a feature as the "best" feature in terms of R^2 improvement, even though there is no clear relation between the response and dependent variables considered.

Wrapper methods

Wrapper feature selection methods search for the best set of features by assessing model performance. The difference with filter methods is that wrapper methods use combinations

of features to test whether a certain model performance improves. This performance can be anything, but for the sake of simplicity let us once more assume R^2 as a result from a regression model. Let us imagine having K features to select for model use, from which we wish to choose the L best features. When feature sets are large, it becomes impossible to assess model performance exhaustively. Therefore, heuristic approaches are typically relied on. One such approach is hill climbing, in which features are iteratively added, until no further model improvement can be achieved [RN09]. The forward sequential hill climbing framework is as follows: in the first iteration, K models are trained with each individual feature and the best predictive feature is selected before progressing to the second iteration. Then, $K-1$ models are trained with each feature and the previously selected feature and the best predictive second feature is added. This is repeated until a subset of L features are selected. An example of such a feature selection scheme is displayed in Figure 3.2. The main drawbacks of such methods are two-fold:

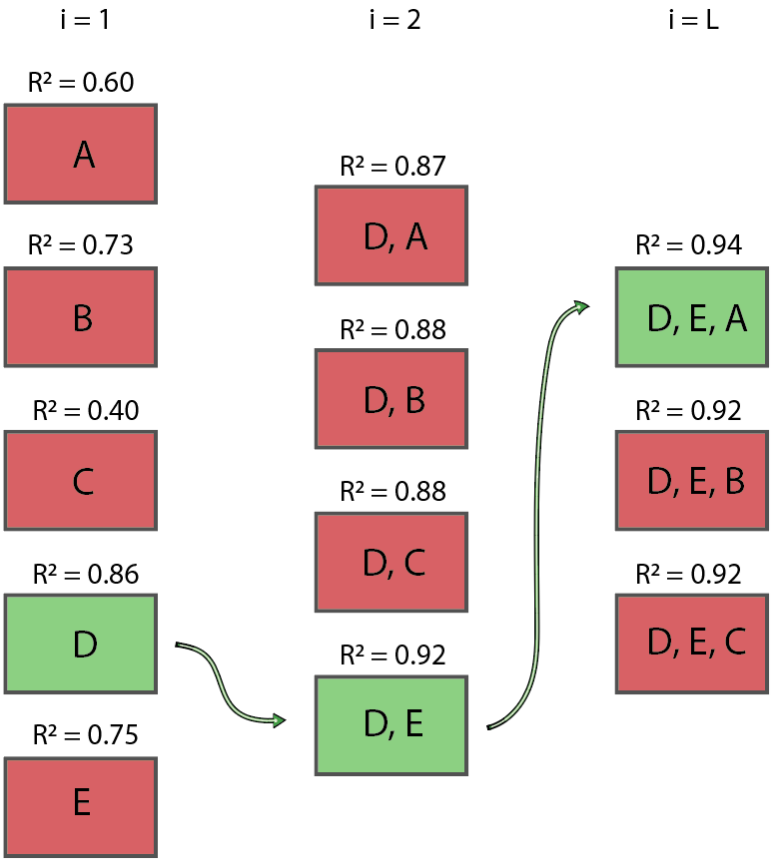


Figure 3.2: Wrapper method example framework, for selecting the L = 3 best features out of K = 5 features A, B, C, D and E. By means of example, model performance is assessed by R^2-value, where the best L features yield D, E, A.

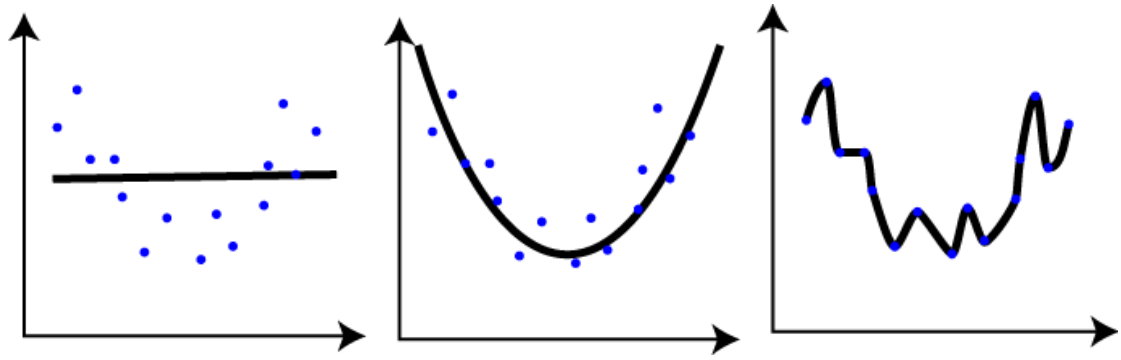
1. Computationally expensive
   Since these methods rely on model performance with respect to feature inclusion, formally, models have to be trained and cross-validated for each feature that is added. This becomes an increasingly time- and resource expensive problem when models are complex (e.g. time-consuming optimizations are involved rather than simple linear fits) or when the number of variables is large [Ras18].

2. Risk of overfitting

In greedy methods where one continues to add features until model performance is maximized with a minimal set of features easily allows for overtraining, especially when the number of samples is insufficient.

3.3 Validation

Proper validation of results requires an assessment of whether results really reflect what they seem to reflect. To illustrate this, let us imagine a scenario where a specific approach yields great results for one data set, whereas the same approach may completely fail to perform well on other data. In such case, it would be desirable to make an adaptation to the model based on the validated results, such that it maintains more of its ability to generalize. In other words, proper results validation may affect model development. The provided scenario is an example of overfitting and imposes a trade-off that is always inherent to machine learning model design: overfitting versus underfitting. Overfitting occurs when a model is trained to the point where it recognizes the variablilties inherent to that specific dataset and it begins to fit noise. Consequently, the model loses its ability to generalize (i.e. to perform well on unseen data drawn from the same distribution of the data it has been trained on). On the other hand, underfitting occurs when a model is inaccurate in identifying the general characteristics of a dataset, and typically happens when training sets are small or when the selected features are insufficient in describing a response variable. So, we need a way to know the degree of overfitting and underfitting. An example of the effect of overfitting can be seen in Figure 3.3.



(a) Underfitting (b) Just right (c) Overfitting

Figure 3.3: The danger of overfitting. The scattered data show a clearly quadratic trend. Using a linear fit (a) would lead to underfitting, whereas a higher-order fit would cause overfitting to the noise in the data set. Image adapted from [Joh13].

3.3.1 Cross-validation

A good way to test for overfitting is to examine the models' training and testing errors. These are found by evaluating model performance on training and on testing data respectively. If a model shows great results on training data and does not perform well on testing data, it is likely that overfitting occurs. In order to quantify training and testing errors, a data set is typically split into a training set and a testing set. The model is trained on the training set, before evaluating the model performance to obtain training errors. Evaluating the performance of the trained model on the unseen data of the testing set then yields the testing errors. However, an obvious consequence of splitting the data, is that the size

of the data set available for training is compromised. This can be especially troublesome when little data are available. Cross-validation (CV) is a technique that allows models to be tested using the full training set by means of repeated resampling [RF08], before averaging the final result. This maximizes the total number of points available for testing, and simultaneously allows for results validation. Let K be the number of folds chosen to use for cross-validation. We then split the available data samples into K sets, which leaves K ways to choose K - 1 training folds, while always leaving one fold for testing. This is repeated for every way there is to reclassify training testing folds and the result is normalized over K afterwards. This technique is known as K-fold cross-validation and an example where K = 4 is schematically displayed in Figure 3.4. When K equals the total sample size N, K-fold CV becomes leave-one-out cross-validation (LOOCV), where N-1 samples are used for training, and testing is done on the remaining sample. This can be particularly useful when data are scarce and training data must be maximized.

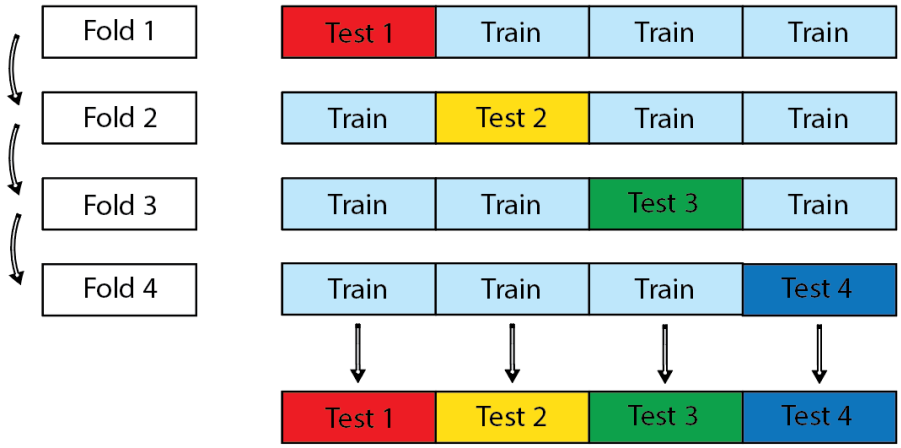


Figure 3.4: K-fold cross-validation in the case where K = 4. Each fold uses the same data set but differ in the way they were split. The test statistics of the entire data set are made up of the results from all CV folds.

CHAPTER 4

METHODS & MATERIALS

In this chapter we will first cover the PCA DVH-reconstruction, around which all models are built. We will continue to cover the methodologies of the current clinical practice, and then proceed our methodologies in developing our other, more complex, optimization-based models. The latter involves feature selection, objective function formulations, as well as the used validation methods.

4.1 Principal Components DVH reconstruction framework

At the basis of all of our models stands a simplification of both the OVH and DVH by means of PCA eigenvalues and eigenvectors. Instead of predicting few points on the DVH, we predict only the first few PCA eigenvalues, and use the eigenvectors to reconstruct the entirety of the DVH. The full DVHs follow from a linear combination of these eigenvalues and their corresponding PCA eigenvectors. It can be said that any DVH in the dataset can be perfectly reconstructed from the eigenvectors and their corresponding eigenvalues that result from the DVH dataset PCA decomposition, as follows:

$$\widetilde{DVH}_n = \overline{DVH} + \sum_{i=1}^{D} \tilde{\lambda}_{i,n} \boldsymbol{V}_i \tag{4.1}$$

where the tilde denotes parameters that are identical to the data set, and D is the total number of available dose bins. This works, because the PC set resulting from the PCA is as large as the number of dose bins. $\tilde{\lambda}_{i,n}$ is the i^{th} PCA eigenvalue of patient n resulting from the PCA on the DVH data, and $\boldsymbol{V}_i$ is the i^{th} PCA eigenvector which is the same for all samples. $\tilde{\lambda}_{i,n}$ and $\boldsymbol{V}_i$ are both ranked from large to small according to their explained variance ratio (EVR). The proper PCA DVH, DVH_p, which is the DVH as approximated by the first C PC modes, can then be written as follows:

$$DVH_{p,n} = \overline{DVH} + \sum_{i=1}^{C} \tilde{\lambda}_{i,n} \boldsymbol{V}_i \tag{4.2}$$

where C is the number of PCs chosen for DVH prediction. C should be picked such that a large amount of the dataset variation is captured by the PCA, while remaining generally descriptive for unseen data (i.e. the PCA does not "overfit" to our data). If we now predict C eigenvalues ($\lambda_{i,n}$), we can use the result to reconstruct a new DVH prediction from Equation 4.2.

$$DVH_n = \overline{DVH} + \sum_{i=1}^{C} \lambda_{i,n} \boldsymbol{V}_i \tag{4.3}$$

In words, the general framework of the optimization-based methods is always to predict C eigenvalues, such that a certain objective function is minimized. Such methods essentially boil down to finding an estimation function that describes these eigenvalues from a set of input features $[\boldsymbol{\xi}]_n$ and regression coefficients $\boldsymbol{A}$. Mathematically, we can write:

$$\lambda_{i,n} = \left[f_i(\boldsymbol{A}_i, [\boldsymbol{\xi}]_n) \right]_n \tag{4.4}$$

where $\boldsymbol{A}$ denotes the regression coefficient matrix. The optimal result is found by identifying the regression coefficients, such that a certain objective function is minimized. The models we used differed in the formulation of these objective functions, for which decisions were based on intermediate analyses of DVH prediction accuracies on testing and training data, and on the TC scores. We will first discuss the current KBP clinical practice.

4.2 Tolerance criterion

The purpose of the tolerance criterion is to define a boundary (the tolerance criterion boundary, TCB), wherein DVH-predictions should lie to be considered a successful prediction. This is not a statistical measure that is based on the spread of a population, hence it is not the same as a confidence interval (CI). This boundary is patient-specific, and is determined from the true DVHs in the data set, $\widetilde{DVH}_n$, where n denotes the patient. Let us imagine a point H on a hypothetical DVH, DVH^H. Based on this point, four TC boundary points can be defined, each a distance ϵ away from H, $P^{dwn}_{volume,n}$, $P^{up}_{volume,n}$, $P^{dwn}_{dose,n}$ and $P^{up}_{dose,n}$. The TC boundary, denoted by TCB, is then constructed from the outer rim of the four sub-boundaries that result from these lower and upper vertical (volume axis) and horizontal (dose-axis) DVH points. These points can more conveniently be named:

$$
\begin{aligned}
&1.\ P^{dwn}_{volume,n} = P^{south}_{n} \quad ; \quad TC^{dwn}_{volume,n} = TC^{south}_{n} \\
&2.\ P^{up}_{volume,n} = P^{north}_{n} \quad ; \quad TC^{up}_{volume,n} = TC^{north}_{n} \\
&3.\ P^{dwn}_{dose,n} = P^{west}_{n} \quad ; \quad TC^{dwn}_{dose,n} = TC^{west}_{n} \\
&4.\ P^{up}_{dose,n} = P^{east}_{n} \quad ; \quad TC^{up}_{dose,n} = TC^{east}_{n}
\end{aligned}
$$

where the subscript n denotes patient-dependence, which we drop for the rest of this paragraph for convenience. The naming of each cardinal direction is intuitively explained by the fact that each point dictates the furthest lower, upper, left and right point a predicted DVH dose bin respective to point H may be to be counted as a successully predicted point. In order to visualize this, let us imagine a hypothetical dose-volume

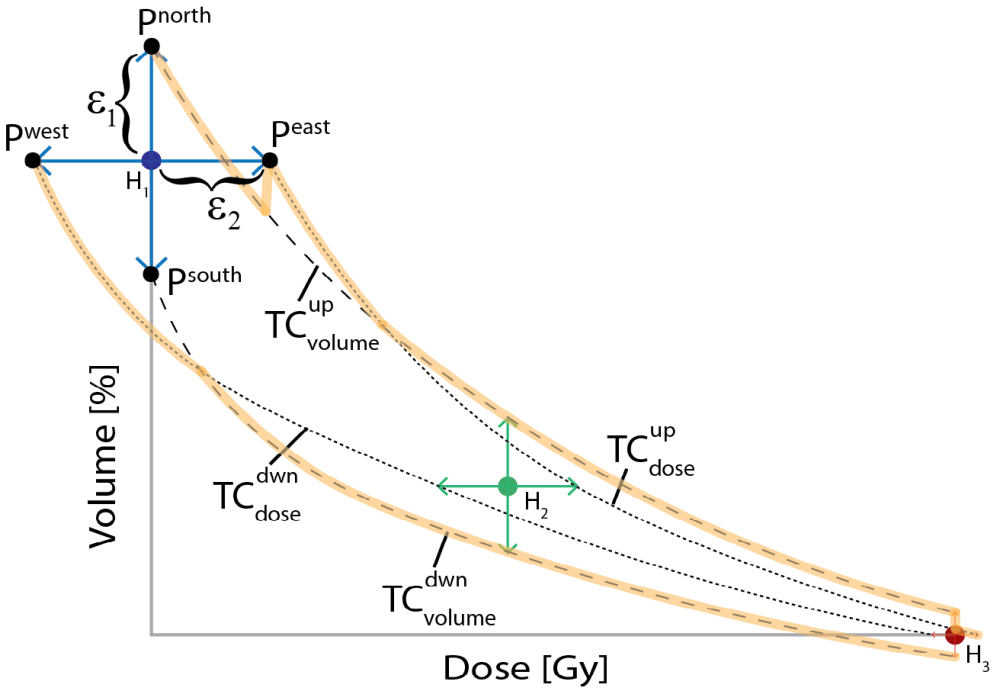


Figure 4.1: Three DVH points, H_1, H_2, and H_3, sampled from DVH^H, are indicated by the blue, green and red dots. The outer boundaries are indicated by the semi-opaque yellow line.

histogram, DVH^H, referring to Figure 4.1. Let us select three points on DVH^H, denoted by the blue (H_1), green (H_2) and red (H_3) dots. If each point is assigned a lower, upper, left and right point a distance ϵ away from $DVH^H(j)$, where $j \in [1, 2, 3]$, the four TC-boundaries are constructed from the lines that run through each of its corresponding boundary points. The remainder of the line is determined by cubic interpolation. It should be noted that ϵ, which spans the separation between DVH^H and each of the above-mentioned sub-boundaries, decreases linearly with dose. Formally, ϵ should be distinguished by ϵ_1 and ϵ_2, describing vertical and horizontal separations respectively, meaning they have different units (volume percentage and Gy). Throughout this study, we have chosen $\epsilon_1(d)$ and $\epsilon_2(d)$ such that the initial separations Δ_I^V and Δ_I^D are 5 % and 5 Gy at $\epsilon(d = 0 \text{ Gy})$ and the final separations Δ_F^V and Δ_F^D amount 1 % and 1 Gy at $\epsilon(d = 80 \text{ Gy})$. Remembering that D denotes the prescribed dose of the primary PTV, 77 Gy. Mathematically, $\epsilon(d)$ then translates to:

$$\epsilon_1(d) = \Delta_I^V - \frac{(\Delta_I^V - \Delta_F^V)}{D} d \tag{4.5}$$

$$\epsilon_2(d) = \Delta_I^D - \frac{(\Delta_I^D - \Delta_F^D)}{D} d \tag{4.6}$$

Each of the four sub-boundaries are calculated from:

$$TC^{south}(d) = DVH^H(d) - \epsilon_1(d) \tag{4.7}$$

$$TC^{north}(d) = DVH^H(d) + \epsilon_1(d) \tag{4.8}$$

$$TC^{west}(d - \epsilon_2(d)) = DVH^H(d) \tag{4.9}$$

$$TC^{east}(d + \epsilon_2(d)) = DVH^H(d) \tag{4.10}$$

The yellow boundary indicated in Figure 4.1 results from the outer rim of all sub-boundaries:

$$TC^{dwn'}(d) = min(TC^{west}(d), TC^{south}(d)) \tag{4.11}$$

$$TC^{up'}(d) = max(TC^{north}(d), TC^{east}(d)) \tag{4.12}$$

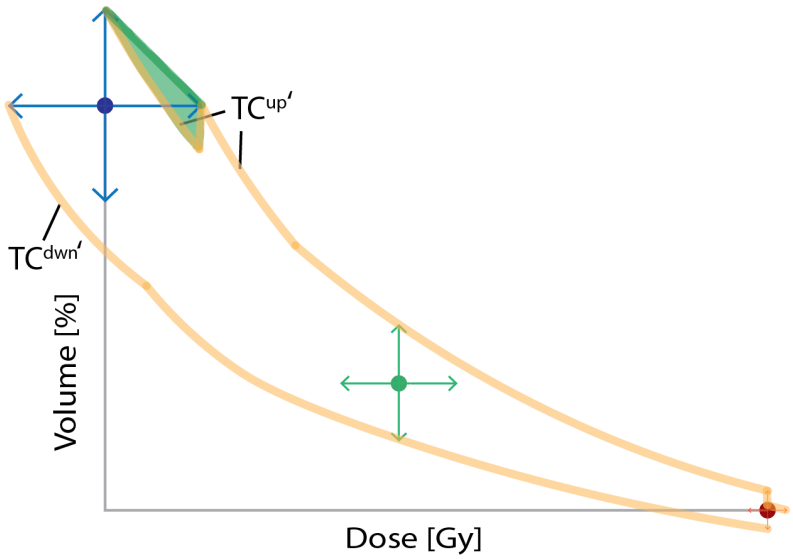


Figure 4.2: The resulting lower and upper boundaries $TC^{dwn'}$ and $TC^{up'}$ are indicated by the yellow line. The artefacts visible at the beginning and at the tail are due to sampling differences between the four sub-boundaries.

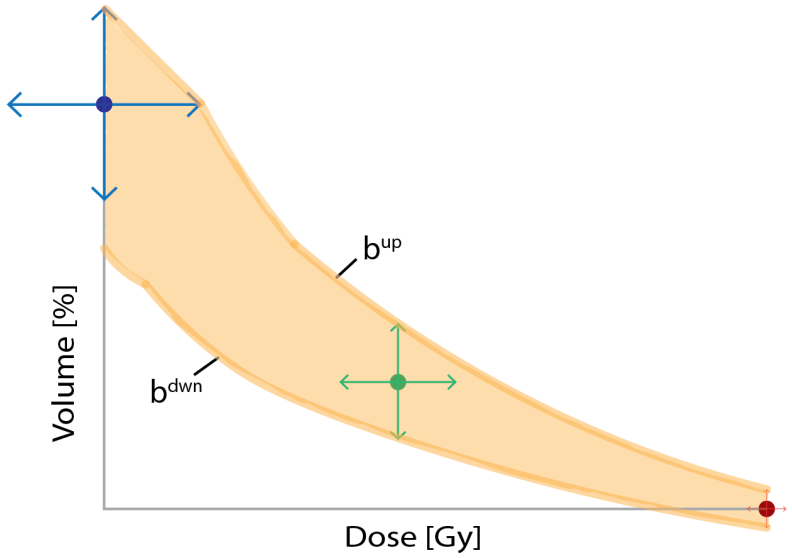


Figure 4.3: The resulting TCB of DVH^H is spanned by the area between b^{dwn} and b^{up}, and is filled in yellow.

The resulting lower and upper TC boundaries, $TC^{dwn'}$ and $TC^{up'}$ (note the apostrophe) are yet incomplete; as can be seen in Figures 4.1 and 4.2, the beginning and the tail need to be processed. For the tail, simply cutting off at the final dose bin suffices, and the same goes for $TC^{dwn'}(d < 0)$. However, for $d < \Delta_I^D$ where TC^{east} is undefined, we simply linearly interpolate with the the maximum of TC^{north}. The restored area by linear interpolation is indicated by green in Figure 4.2. This finally yields our two TC boundaries: b^{dwn} and b^{up}.

4.2.1 Tolerance criterion evaluation metric 1: TC_α

Now that we have discussed how the tolerance criterion is defined, we can continue to define our final two evaluation metrics. If we imagine a population for which DVH-predictions were done, one of the first TC-based metrics that comes to mind is the number of predicted DVH-points that are within this boundary. The fraction of these points is what determines TC_α. This can be modelled with a step function, that is 1 inside the TCB, and 0 outside. Naturally, as the boundary varies per patient and dose bin, so does the step function. Given that we again include patient dependence, this step function can be written for patient n and dose bin d:

$$\begin{aligned} H_{\alpha,n}(d) = \frac{1}{2}\big((sgn(DVH_n(d) - b_n^{dwn}(d)) + 1) \\ - sgn(DVH_n(d) - b_n^{up}(d)) + 1)\big) \end{aligned} \tag{4.13}$$

$TC_{\alpha,n}$, being patient-specific, can then be found by summing over all patients and dose bins, before normalizing:

$$TC_{\alpha,n} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{d=1}^{D} H_{\alpha,n}(d) \tag{4.14}$$

It should be noted that $H_{\alpha,n}$ is a series of functions, that is defined differently, depending on n and d.

4.2.2 Tolerance criterion evaluation metric 2: TC_β

The second and final TC-based metric is one to serve as a pass-fail criterion for a patient DVH prediction. The criterion is that we aim for patients to have at least 90% of its dose points within the TCB. If it does, it counts as a pass. The fraction of the patients who pass is what defines TC_β. This can once more be modeled with a step function that is 0 for $TC_{\alpha,n} \leq \nu = 90\%$, and 1 above it:

$$H_\beta(TC_{\alpha,n}) = \frac{1}{2}(sgn(TC_{\alpha,n} - \nu) + 1) \tag{4.15}$$

$$TC_\beta = \frac{1}{N}\sum_{n=1}^{N} H_\beta(TC_{\alpha,n}) \tag{4.16}$$

With this criterion in mind, the aim of our model to reach a TC_α and TC_β accuracy of at least $\nu = 90\%$ in training data. From here, symbols with a subscript α and β denote the their connection to TC_α and TC_β respectively.

4.3 Two-point predictor

One of the first approaches we believed to be fruitful was by predicting only a few points through which a predicted DVH runs, and using the PCA eigenvectors to reconstruct the remainder of the DVH. The goal was to investigate how well we can predict DVHs by modelling with a clear focus on simplicity. To do this, we investigated dose-volume metrics that correlated highly with OVH metrics, and used the DVH PCA modes to reconstruct the rest of the DVH. Two points that we have found to work well are the V_{95} and V_{mean}.

4.3.1 DVH to OVH correlations

In order to select two DVH points to use for modelling, we investigated correlations in the data. The data confirmed the use for the V_{95} and V_{mean} metrics for DVH prediction, as they were found to correlate highly with the OVH at r = 0 and r = 10 mm respectively ($R^2 = 0.946$ and 0.866 respectively). For the first point, we determined the R^2 of each DVH point with several OVH points. The correlations can be seen in Figure 4.4. For the second point, we investigated how the gEUD (Equation 1.3) with different EUD-parameter values correlated to different OVH points. Correlations can be found in Figure 4.5. Scatter plots of the V_95 and V_{mean} for the two-point predictor model are shown in Figure 4.6.

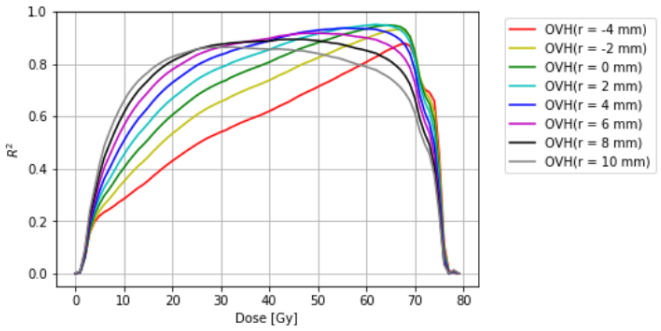


Figure 4.4: The R^2 values were collected for scatter plots of all DVH-dose bins with OVH values at several distances r. This yields a number of highly correlating dose-volume metrics in the high-dose region. The V_{95} was confirmed to correlate well with $OVH(0)$ ($R^2 = 0.94596212$), as indicated by the peak in the green line at d = 66 Gy ($\approx 0.95 \cdot D_P$, D_P being the prescribed dose of the secondary PTV, 70 Gy)

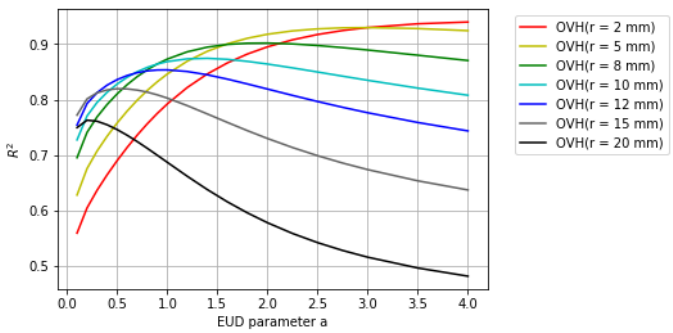
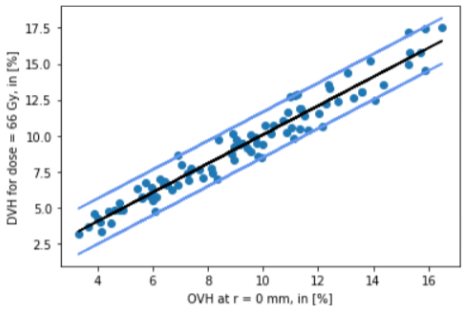
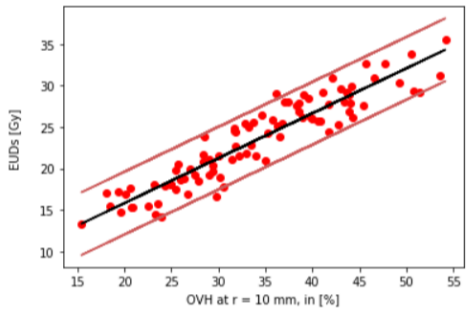


Figure 4.5: The R^2 values were collected for the scatter plots of DVH gEUD-dose for different a, with respect to OVH values at several distances r. It can be seen that for the average dose (a = 1), we find an R^2-optimum for OVH(r = 10 mm) ($R^2 = 0.86594754$).



(a) V_{95} scatter plot with regression function $y = 0.208 + 0.994x$, ($R^2 = 0.946$)



(b) V_{mean} scatter plot with regression function $y = 3.95 + 0.553x$, ($R^2 = 0.866$)

Figure 4.6: V_{95} and V_{mean} linear regression models and 95% CIs

There were two reasons why the gEUD for a = 1 was chosen, instead of any of the higher-correlating gEUD metrics. First, the better correlating gEUDs result from scatter plots with OVH metrics that lie closer to the OVH we are already using for the first point: OVH(r = 0 mm). Therefore choosing points closer to this OVH point are more likely to describe redundant information. Second, the OVH(r = 10 mm) gEUD peaks close to a = 1. Since a = 1 corresponds to the average dose, this was chosen for convenience. In addition, the OVH(r = 10 mm) seemed like a good trade-off between a decent R^2 and "new" information.

4.3.2 DVH reconstruction

Once we know the linear relationships, we can easily estimate the two DVH points for a new patient, given his OVH. Let $\boldsymbol{V}_i$ be the i^{th} DVH PC mode, for $i = [0, 1, 2]$, and let D_S be the maximum dose of the secondary PTV, 70 Gy. The predicted DVH for patient n, DVH_n, follows from a linear combination of the PC modes and two to-be-determined coefficients α and β:

$$DVH_n = \boldsymbol{V}_0 + [\lambda_1]_n \boldsymbol{V}_1 + [\lambda_2]_n \boldsymbol{V}_2 \tag{4.17}$$

where the subscript n indicates patient-specific parameters and $[\lambda_1]_n$ and $[\lambda_2]_n$ are scalars. This results in a solveable system of two linear equations:

$$\begin{aligned} \overline{DVH}_n &= \overline{\boldsymbol{V}}_0 & +[\lambda_1]_n \overline{\boldsymbol{V}}_1 & & +[\lambda_2]_n \overline{\boldsymbol{V}}_2 \\ DVH_n(0.95\ D_S) &= \boldsymbol{V}_0(0.95\ D_S) & +[\lambda_1]_n \boldsymbol{V}_1(0.95\ D_S) & & +[\lambda_2]_n \boldsymbol{V}_2(0.95\ D_S) \end{aligned} \tag{4.18}$$

where the overline denotes averages. This may be written in matrix form:

$$\begin{bmatrix} \overline{DVH}_n - \overline{\boldsymbol{V}}_0 \\ DVH_n(0.95\ D_S) - \boldsymbol{V}_0(0.95\ D_S) \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{V}}_1 & \overline{\boldsymbol{V}}_2 \\ \boldsymbol{V}_1(0.95\ D_S) & \boldsymbol{V}_2(0.95\ D_S) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}_n \tag{4.19}$$

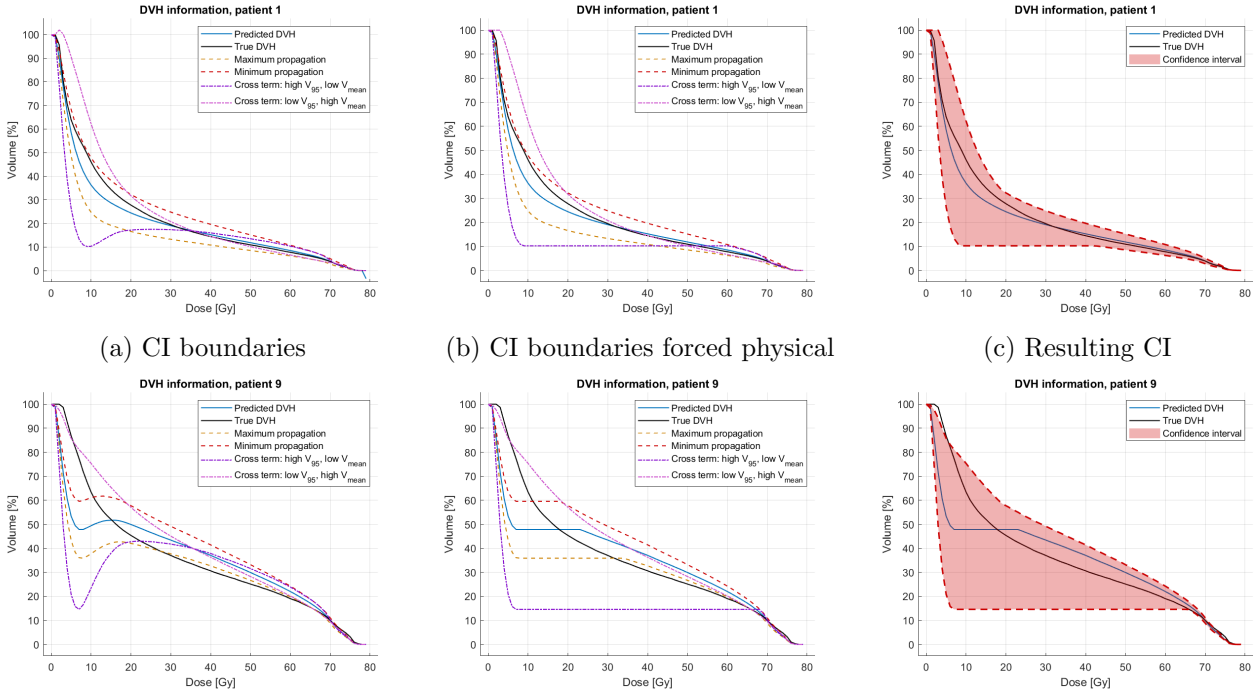
The solutions to λ_1 and λ_2 are found from:

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}_n = \begin{bmatrix} \overline{\boldsymbol{V}}_1 & \overline{\boldsymbol{V}}_2 \\ \boldsymbol{V}_1(0.95\ D_S) & \boldsymbol{V}_2(0.95\ D_S) \end{bmatrix}^{-1} \begin{bmatrix} \overline{DVH}_n - \overline{\boldsymbol{V}}_0 \\ DVH_n(0.95\ D_S) - \boldsymbol{V}_0(0.95\ D_S) \end{bmatrix} \tag{4.20}$$

The resulting scalar coefficients are put into Equation 4.17 to obtain the DVH prediction. 8-fold CV as well as LOOCV were used for result validation.

4.3.3 Confidence intervals

In addition, under the assumption of their Gaussian distribution, we determine a 95% confidence interval CI based on the V_{95} and V_{mean} spreads that resulted from the scatter plots in Figure 4.6. This confidence interval was determined by using adding and subtracting two standard deviations of the V_{95} and V_{mean} averages. By propagating the minima and maxima of these spreads through the DVH reconstruction method as described by 4.3.2, the CI of the entire resulting DVH is determined. However, it should be noted that $V_{95} - V_{mean}$ correlations are unknown. Therefore, it is difficult to precisely determine the 95% DVH confidence interval. Namely, let us consider the CI boundaries that result from propagation of the maximum V_{95} and V_{mean}, and the minimum V_{95} and V_{mean}, this would be most conservative estimation of the 95% CI (i.e. it is likely to not be broad enough over the entire DVH). The other side of the spectrum is where, in addition, we take into account both cross-terms ((minimum V_{95} & and maximum V_{mean} and vice versa), and take the outer boundary together with the non-cross terms to estimate the 95% CI. This only occurs if the V_{95} and V_{mean} are completely uncorrelated, which is not expected. Moreover, because it yields too many unphysical results as exemplified in Figures 4.7, we propose two alternatives, which are shown in Figures 4.9 and 4.8. Note that forcing the DVH to be physical is a post-processing step that will be discussed in Section 4.5.5.



(a) CI boundaries (b) CI boundaries forced physical (c) Resulting CI

(d) CI boundaries (e) CI boundaries forced physical (f) Resulting CI

Figure 4.7: A Typical (a-c) and an extreme (d-f) DVH prediction where the 95% CI resulting from the progressive cross-term propagation does not yield articulate or physical results

As for the proposed solutions, we use two alternatives to determine the cross terms that result in the 95% CI. An overview of how each model is found is shown in Table 4.1.

Alternative 1

The first alternative involves finding four additional cross-terms, by selecting: the maximum V_{95} and average V_{mean} for cross term 1, the minimum V_{95} and average V_{mean} for cross term 2, the average V_{95} and maximum V_{mean} for cross term 3 and the average V_{95} and minimum V_{mean} for cross term 4. The CI is found from taking the encompassing of all six (including the already found upper and lower boundaries) boundaries. Examples using this methodology are shown in Figure 4.9.

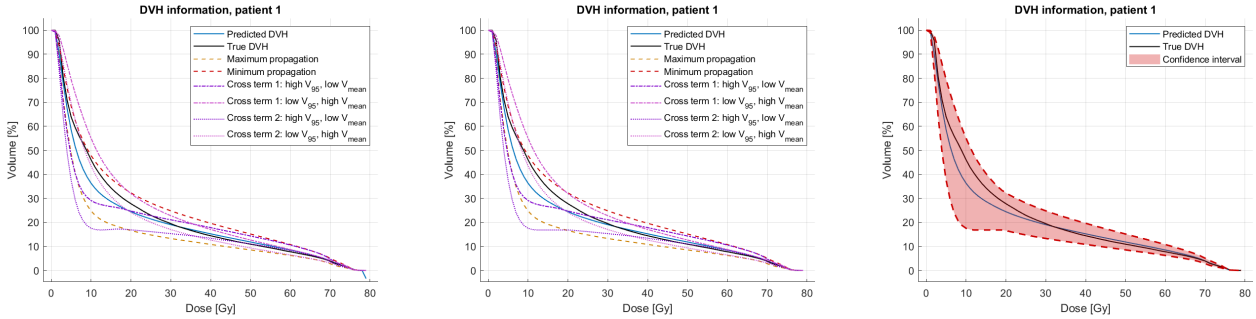
Alternative 2

The second alternative involves finding four cross-terms, by selecting: the maximum V_{95} and average V_{mean} for cross term 1,the average V_{95} and maximum V_{mean} for cross term 2, the minimum V_{95} and average V_{mean} for cross term 3 and the average V_{95} and minimum V_{mean} for cross term 4. Then, cross term 1 is averaged with cross term 3 and cross term 2 is averaged with cross term 4. The resulting four boundaries define the CI by their encompassing. Examples using this methodology are shown in Figure 4.9.

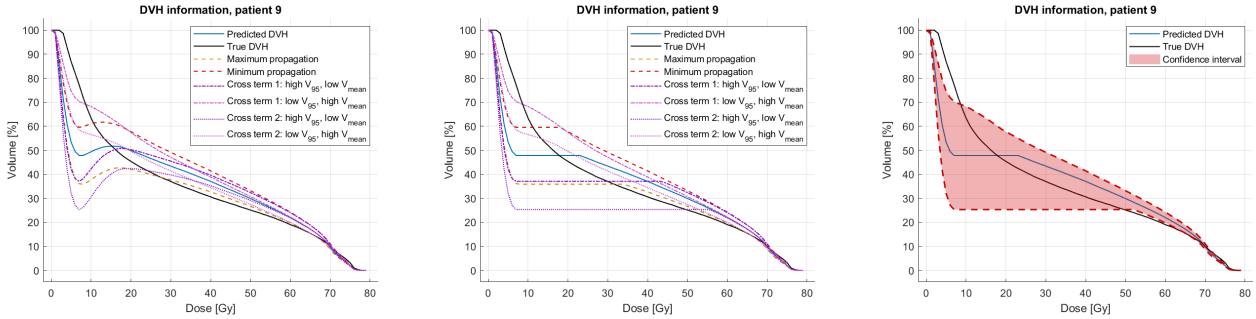
	Upper boundary		Lower boundary		Cross term 1	
	V_{95}	V_{mean}	V_{95}	V_{mean}	V_{95}	V_{mean}
Conservative	Max	Max	Min	Min	x	x
Progressive	Max	Max	Min	Min	Max	Min
Alternative 1	Max	Max	Min	Min	Max	Med
Alternative 2	Max	Max	Min	Min	Max	Med

	Cross term 2		Cross term 3		Cross term 4	
	V_{95}	V_{mean}	V_{95}	V_{mean}	V_{95}	V_{mean}
Conservative	x	x	x	x	x	x
Progressive	Min	Max	x	x	x	x
Alternative 1	Min	Med	Med	Max	Med	Min
Alternative 2	Med	Max	Min	Med	Med	Min

Table 4.1: Overview of the conservative, progressive and the two alternative approaches for determining the boundaries that approximate of the 95% CI. This table should be read as follows. For every model, the boundaries in each colum are obtained from taking either the maximum, minimum or average V_{95} and V_{mean} values that result from their Gaussianity. Med denotes the average, and is simply the determined linear regression model value as can be seen in Figure 4.6. The 'x' indicates that that particular boundary does not apply to the corresponding model.

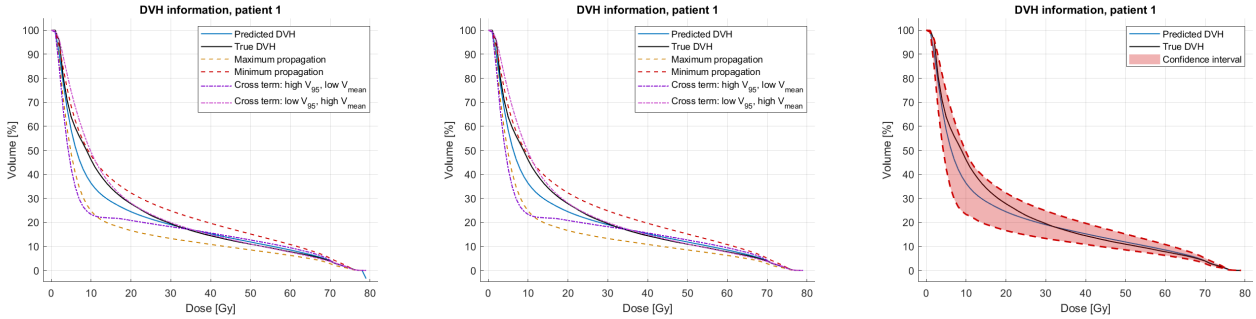


(a) CI boundaries (b) CI boundaries forced physical (c) Resulting CI

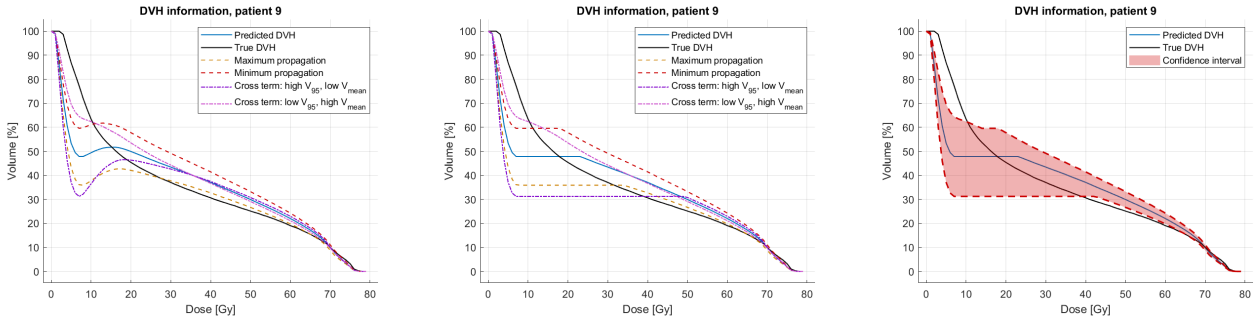


(d) CI boundaries (e) CI boundaries forced physical (f) Resulting CI

Figure 4.8: These figures show the same two example DVH two-point predictions as in Figure 4.7 and 4.9, but these are calculated with the first alternative approach



(a) CI boundaries (b) CI boundaries forced physical (c) Resulting CI



(d) CI boundaries (e) CI boundaries forced physical (f) Resulting CI

Figure 4.9: These figures show the same two example DVH two-point predictions as in Figure 4.7 and 4.8, but now calculated with the second alternative approach.

4.4 Feature selection

4.4.1 Feature specification

In the context of this research, we predict C PCA eigenvalues $\lambda_{i,n}$ for $i \in [1, 2, \cdots C]$, on the basis of geometrical anatomical features (Equation 4.3). OVH-based features we use are the OVH metrics OVH(d = 0mm) and OVH(d = 10mm) and the first three OVH PCA eigenvalues of both the rectum and anal sphincter (AS). In addition, we use the volumes of the PTV, V_{PTV} and the rectum, V_{rect}, and we use three additional features that are derived from CT slice numbers. These features are: first, the PTV length, denoted as O_1. Second, the PTV/rectum length ratio, denoted by $\frac{O_1}{O_2}$. Third, the fraction of the anal sphincter within the PTV + 5 mm, denoted by $\frac{O_3}{O_4}$. All O_1, O_2, O_3 and O_4 are expressed in millimeters. Figure 4.10 schematically shows a representation of the PTV and OARs in question. In addition, the meaning of $O_1 - O_4$ is visualized. Let us define our feature vector $\boldsymbol{\xi}$ of length M. The features $\boldsymbol{\xi}$ contains are enumerated as follows:

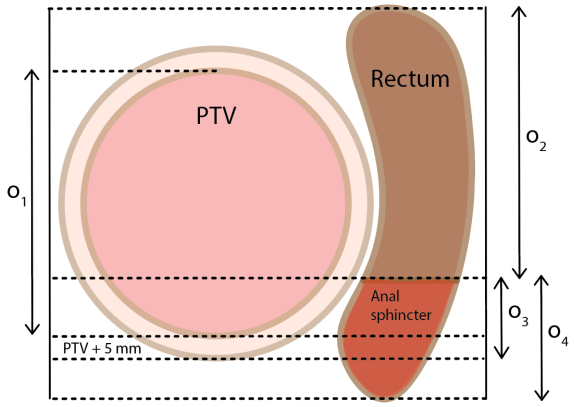


Figure 4.10: Schematic craniocaudal view of the coronal plane, showing the prostate PTV, the PTV expanded by 5 mm, the rectum and anal sphincter. As illustration, CT slice values are based on height coordinates in this image.

ξ_1:	Rectum OVH(0)	ξ_8:	Anal sphincter OVH PCA λ_3
ξ_2:	Rectum OVH(10)	ξ_9:	V_{PTV}
ξ_3:	Rectum OVH PCA λ_1	ξ_{10}:	V_{rect}
ξ_4:	Rectum OVH PCA λ_2	ξ_{11}:	o_1
ξ_5:	Rectum OVH PCA λ_3	ξ_{12}:	$\frac{o_1}{o_2}$
ξ_6:	Anal sphincter OVH PCA λ_1	ξ_{13}:	$\frac{o_3}{o_4}$
ξ_7:	Anal sphincter OVH PCA λ_2		

4.4.2 Logarithmic regression

We have investigated the use for logarithmic regression basis functions for eigenvalue prediction. For the analyses where we include logarithmic regression, $\boldsymbol{\xi}$ is appended with the natural logarithm of ξ_i for $i \in [1, 2, ..., 12]$. ξ_{13} is omitted, because its logarithm reaches minus infinity in the case where there is no AS-(PTV+5) overlap. Thus, in the cases where

logarithmic regression is used, $||\boldsymbol{\xi}|| = 25$. However, logarithmic regressions only proved useful for training data and deteriorated testing errors (thus causing overfitting), and was therefore omitted from analyses in this thesis.

4.4.3 Polynomial feature vector

All of our optimization-based models use polynomial regression (i.e. regression usin polynomial basis functions, resulting in the use of higher-order features). To obtain a Q^{th}-order polynomial feature vector, we simply redefine our basis feature vector $\boldsymbol{\xi}$ such that it contains all Q^{th}-order features. This gives us the polynomial feature vector $\boldsymbol{\zeta}$. For example, if M = 2 and Q = 2, $\boldsymbol{\zeta}$ would be of the form: $[1, \xi_1, \xi_2, \xi_1\xi_2, \xi_1^2, \xi_2^2,]$, there being 1 0^{th} order term, 2 1^{st} order terms and 3 2^{nd} order terms, totaling 6 elements. Similarly, if M = 2 and Q = 3, $\boldsymbol{\zeta} = [1, \xi_1, \xi_2, \xi_1\xi_2, \xi_1^2, \xi_2^2, \xi_1^2\xi_2, \xi_1\xi_2^2, \xi_1^3, \xi_2^3]$, there being 1 0^{th} order term, 2 1^{st} order terms, 3 2^{nd} order terms and 4 3^{rd} order terms, totaling 10 elements. For the general case, the length of $\boldsymbol{\zeta}$, Z (i.e. the number of all unique combinations), is found from an experiment of unordered sampling with replacement of the terms in ξ. Mathematically, this means that Z results from summing the number of possible q^{th}-order combinations from q = 0 to q = Q, without using the same combination more than once:

$$Z = \sum_{q=0}^{Q} \binom{M+q-1}{M-1} = \sum_{q=0}^{Q} \frac{(M+q-1)!}{(M-1)!q!} \tag{4.21}$$

The number of possible individual q^{th} order combinations for varying M were calculated and shown in Table 4.2

Table 4.2: The number of possible q^{th} degree combinations (features) that can be made for different numbers of features M, up to a 3^{rd} order. The length of $\boldsymbol{\zeta}$ is found by summing over corresponding column M up to q = Q.

q\M	1	2	3	4	5	6	7	8	9	10	11	12	13
0	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	2	3	4	5	6	7	8	9	10	11	12	13
2	1	3	6	10	15	21	28	36	45	55	66	78	91
3	1	4	10	20	35	56	84	120	165	220	286	364	455

q\M	14	15	16	17	18	19	20	21	22	23	24	25	26
0	1	1	1	1	1	1	1	1	1	1	1	1	1
1	14	15	16	17	18	19	20	21	22	23	24	25	26
2	105	120	136	153	171	190	210	231	253	276	300	325	351
3	560	680	816	969	1140	1330	1540	1771	2024	2300	2600	2925	3276

4.4.4 Feature selection

We use the "two subjects per variable" rule of thumb as proposed by Austin and Steyerberg [AS15] as a limit for the maximum number of features that can be included, L. This is a rule of thumb that is used to estimate the amount of regression coefficients that can be predicted reliably, based on the sample size available. So: the feature vector limit $L = N/2$. Since we typically end up with way more features (see also Table 4.2), we

need a feature selection algorithm to find the most descriptive ones. For this, we use a combination of a filter and a wrapper method. First, we do a simple, linear fit for every individual feature with λ_1 to see how well it correlates. This is the filter method part. The feature corresponding to the highest R^2 score is chosen. Next, we iteratively make a new fit for each feature added to the previously chosen feature. The feature that brings about the greatest R^2 increase is the second feature chosen. This follows a heuristic, forward sequential hill climbing framework as proposed by [RN09] (see the wrapper methods paragraph of Section 3.2.2). The process is repeated until either all features are used or when L features have been selected. Then, this process is repeated C times, resulting in a specific optimal feature set for each eigenvalue. M is the minimal value of L and the maximum feature vector length:

$$M = min(L, Z) \tag{4.22}$$

4.5 Optimization-based models

The following methods have all involved optimization, but differed in the way their objective functions were formulated. Optimizations were done by sequential quadratic programming in MATLAB's nonlinear programming solver *fmincon*.

4.5.1 EV-optimization

The first optimization-based approach was to use an objective function that minimizes the PCA eigenvalues. This approach essentially is the same as C multiple linear regression models with polynomial (and logarithmic) features, because the objective function is simply optimized by OLS. Equation 4.4 becomes for EV-optimization:

$$\lambda_{i,n} = \big[f(\boldsymbol{A}_i, \boldsymbol{\zeta}_n)\big]_n = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_C \end{bmatrix}_n = \begin{bmatrix} a_{10} & a_{11} & \dots & a_{1M} \\ a_{20} & a_{21} & \dots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{C0} & a_{C1} & \dots & a_{CM} \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_M \end{bmatrix}_n \tag{4.23}$$

in which the estimation function may also be written as:

$$\big[f(\boldsymbol{A}_i, \boldsymbol{\zeta}_n)\big]_n = \sum_{m=0}^{M} a_{i,m}[\zeta_m]_n \tag{4.24}$$

The objective function is the squared differences between the true and predicted EVs:

$$E_1^i = \frac{1}{N}\sum_{n=1}^{N}\Big(\tilde{\lambda}_{i,n} - \lambda_{i,n}\Big)^2 = \frac{1}{N}\sum_{n=1}^{N}\Big(\tilde{\lambda}_{i,n} - \big[f(\boldsymbol{A}_i, \boldsymbol{\zeta}_n)\big]_n\Big)^2 \tag{4.25}$$

In which we want to find $\boldsymbol{A}_i$ for which E_i is minimized for $i \in [1, 2, \cdots, C]$.

$$\min_{a_{i,m}} E_1^i = \min_{a_{i,m}} \frac{1}{N}\sum_{n=1}^{N}\Big(\tilde{\lambda}_{i,n} - \sum_{m=0}^{M} a_{i,m}[\zeta_m]_n\Big)^2 \tag{4.26}$$

This results in the OLS-optimal regresion coefficients for predictions at the PCA eigenvalue level. A drawback of this approach is this model optimizes for something that we are not directly interested in; it would make more sense to use an optimizer that ensures finding a fit for the eigenvalues, such that the DVH that results from it is predicted optimally. A second drawback of this approach is that it is not trivial to constrain the EV optimizer on the resulting DVH, because the relation between eigenvalues and DVH-imposed constraints is not intuitive.

4.5.2 DVH-optimization

In order to have the optimization better fit our goal of predicting accurate DVHs, the objective function was formulated such that quadratic differences between predicted and true DVHs are minimized. This means that we still predict eigenvalues, however the result is optimized for the DVH root mean square. Referring to Equation 4.25, we use the DVH dose bins least squares instead of EVs least squares:

$$E_2 = \frac{1}{ND}\sum_{d=1}^{D}\sum_{n=1}^{N}\Big(\widetilde{DVH}_n(d) - \sum_{i=1}^{C}\big[f(\boldsymbol{A}_i, \boldsymbol{\zeta}_n)\big]_n \boldsymbol{V}_i(d)\Big)^2 \tag{4.27}$$

and for the optimal regression coefficients:

$$\min_{a_{i,m}} E_2 = \min_{a_{i,m}} \frac{1}{ND} \sum_{d=1}^{D} \sum_{n=1}^{N} \left(\widetilde{DVH}_n(d) - \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m} [\zeta_m]_n \boldsymbol{V}_i(d) \right)^2 \tag{4.28}$$

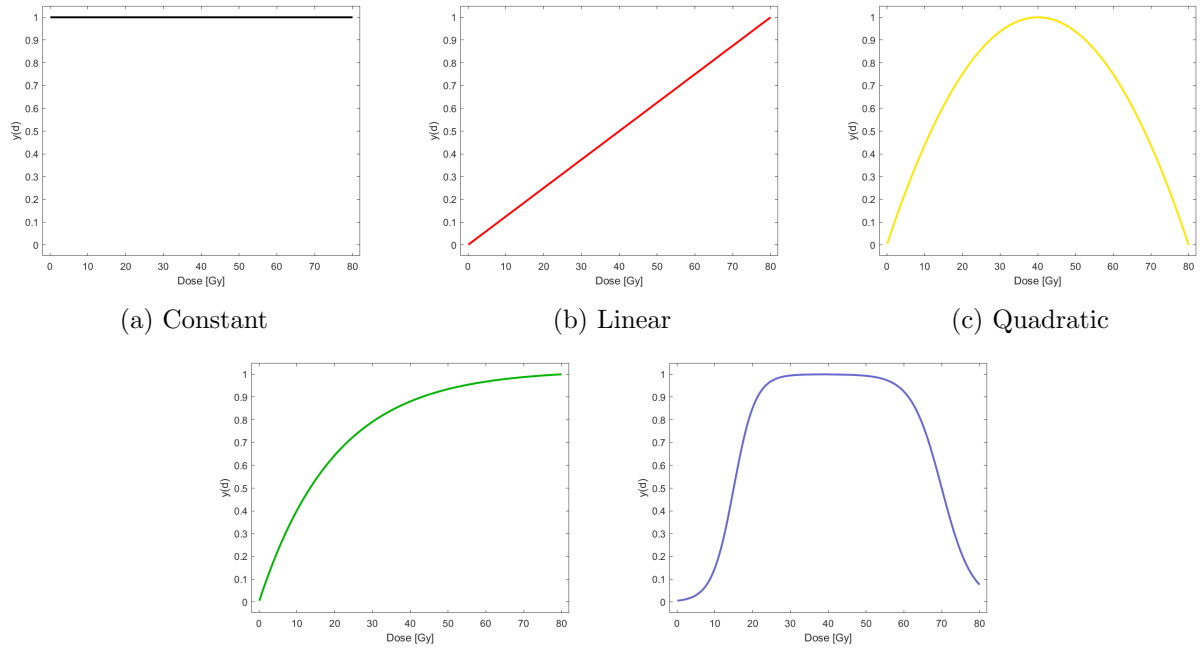
As a result, we can find regression coefficients by means of DVH OLS minimization (i.e. as close to the true DVH along the entire dose grid for the whole population as can get).

4.5.3 Penalized DVH-optimizations

We further investigated ways to improve DVH optimizers to yield better results with respect to our set tolerance criterion boundary (TCB). In addition to minimizing overall DVH squared differences, we use additional weighted penalty terms for points that are outside of our tolerance criterion. This means that we sacrifice a bit of overall prediction accuracy, but we trade that for pushing the DVH prediction closer to or inside the tolerance criterion. There are two types of weighted DVH optimizations that we have looked into. They first one assigns weighted penalties purely based on the TCB. The second one involves a second boundary, the halfway-boundary (HWB), which is the boundary that runs halfway between the true DVH and the TCB. Both methods handle weights in the same way. Let us define $\psi_l(d)$, the weight basis function of each weight l, that regulates the penalty weights between 0 and 1, based on the dose bin:

$$W_{k,l}(d) = J_k \cdot \psi_l(d) \tag{4.29}$$

where J_k is an arbitrary base weight factor, $W_k(d)$ is the resulting weight at dose bin d, and k denotes the region for which the weight penalizes, the significance of which becomes apparent in the next paragraph, and $l \in [1, \ldots, 5]$. We investigated weight basis functions of different parametrizations, which are summarized in Figure 4.11.



(a) Constant (b) Linear (c) Quadratic

(d) Exponential (e) Double sigmoidal

Figure 4.11: Figures (a) through (e) display the forms of the used weight basis functions.

The reasoning behind the shape of these weight basis functions is as follows. First, the shapes of Figures 4.11a and 4.11c were decided upon to assign more importance to the higher dose region. Although this basically means that mispredicted points in the high-dose region are penalized double (once by the TC boundary, once by the increased penalty weight), it was believed to further improve the model's predictive power. This was confirmed for high-dimensional training data. Second, the shapes of the quadratic and double sigmoidal weight basis functions (Figures 4.11c and 4.11e) were decided upon, based on both training and testing errors in the best performing model that showed for a large number of models the erroneously predicted DVH points were distributed rather uniformly along the dose grid. This can be seen in Figure 4.12. Supposedly, using penalty functions with these shapes would improve predictions roughly between 25 Gy and 65 Gy, at the cost of accuracy outside of these dose regions. The rationale behind the stationary weight to penalize points outside the TC boundary along the whole dose grid, and thereby to aid the optimizer in pushing points closer to the TCB.

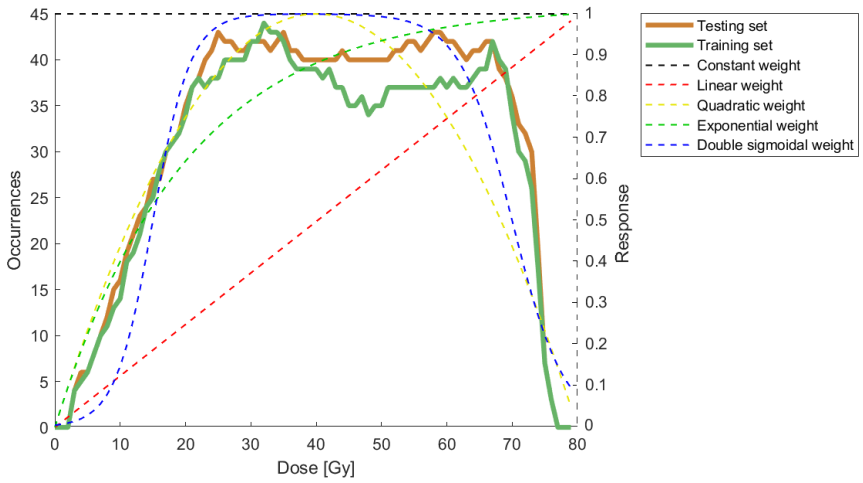


Figure 4.12: Wrongly predicted points (i.e. outside the TCB) by the weightless DVH-optimized predictor . The weight basis function dashed lines are included to visualize how each corresponding weight penalty relates to model inaccuracy, and are represented by the right axis.

TC-penalized DVH-optimizations

We included a penalty term for points outside of the TCB. That way, two regions that can be distinguished, separated by the TC boundary. Regions are denoted by $k \in [1, 2]$, referring to Equation 4.29, where J_1 is the union of the two regions and J_2 denotes the region outside of the TC boundary. Intermediate analyses have shown that a base weight $(J_1 : J_2)$ ratio of $(1 : 10)$ for points inside vs outside the TCB yielded the best predictions (much lower penalty weights generally didn't improve TC scores, whereas higher penalty weights confounded the overall DVH prediction too much to improve the result in any way). In line with Equation 4.29, the resulting weights are:

$$\boldsymbol{W} = \begin{bmatrix} W_{1,l}(d) \\ W_{2,l}(d) \end{bmatrix} = \psi_l(d) \begin{bmatrix} J_1 \\ J_2 \end{bmatrix} \tag{4.30}$$

From here on, let us simplify our notations by naming $x_{d,n}(d) = \sum_{i=1}^{C} \left[f(\boldsymbol{A}_i, \boldsymbol{\zeta}_n)\right]_n \boldsymbol{V}_i(d)$, being the predicted DVH resulting from Equation 4.3. In the objective function used for the penalized DVH optimizer, OLS is still optimized for. The penalties alter the objective function and are incorporated as follows:

$$\begin{aligned} E_{3,l} = \frac{1}{ND} \sum_{d=1}^{D} \sum_{n=1}^{N} \Bigg[& W_{2,l}(d) \cdot \Big(\widetilde{DVH}_n(d) - x_{d,n}(d)\Big)^2 \quad + \\ & H^{dwn}_{\alpha,n}(d) \cdot W_{2,l}(d) \cdot \Big(b^{dwn}_n(d) \quad - x_{d,n}(d)\Big)^2 \quad + \\ & H^{up}_{\alpha,n}(d) \cdot W_{2,l}(d) \cdot \Big(b^{up}_n(d) \quad - x_{d,n}(d)\Big)^2 \Bigg] \end{aligned} \tag{4.31}$$

where $H^{dwn}_{\alpha,n}(d)$ and $H^{up}_{\alpha,n}(d)$ are the lower and upper boundary terms of the TC_α step function (Equation 4.13):

$$H^{dwn}_{\alpha,n}(d) = \frac{1}{2}(sgn\Big(x_{d,n}(d) - b^{dwn}_n(d)\Big) + 1) \tag{4.32}$$

$$H^{up}_{\alpha,n}(d) = \frac{1}{2}(sgn\Big(b^{up}_n(d) - x_{d,n}(d)\Big) + 1) \tag{4.33}$$

With $\left[f(\boldsymbol{A}_i, \boldsymbol{\zeta}_n)\right]_n$ being expressed as $\sum_{m=0}^{M} a_{i,m}[\zeta_m]_n \boldsymbol{V}_i(d)$, written explicitly, the optimal regression coefficients result from:

$$\begin{aligned} \min_{a_{i,m}} E_{3,l} = \min_{a_{i,m}} \frac{1}{ND} \sum_{d=1}^{D} \sum_{n=1}^{N} \Bigg[& W_1(d) \cdot \Big(\widetilde{DVH}_n(d) \quad - \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m}[\zeta_m]_n \boldsymbol{V}_i(d)\Big)^2 \quad + \\ & H^{dwn}_{\alpha,n}(d) \cdot W_2(d) \Big(b^{dwn}_n(d) \quad - \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m}[\zeta_m]_n \boldsymbol{V}_i(d)\Big)^2 \quad + \\ & H^{up}_{\alpha,n}(d) \cdot W_2(d) \Big(b^{up}_n(d) \quad - \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m}[\zeta_m]_n \boldsymbol{V}_i(d)\Big)^2 \Bigg] \end{aligned} \tag{4.34}$$

It should be noted that the normalization factor $\frac{1}{ND}$ normalizes simply for the patients and DVH bins without penalty terms. Because the normalization factor does not influence the optimization, the normalization factor was left as it is for the non-penalized DVH case, and does not corect for penalty terms. Also, similar to Equation 4.13, it should be noted that $H^{dwn}_{\alpha,n}$ and $H^{up}_{\alpha,n}$ are a series of functions that is defined for all n and d.

HWB-penalized DVH-optimizations

To complement the penalized DVH-optimization model, we included an additional boundary, halfway between the true DVH and the TC boundary, the halfway boundary (HWB). We did this in order to further enable the optimizer to improve the TC score. The rationale was that the regular DVH weighted optimizer may help to push DVH points towards the outer boundary of the TCB, but does not necessarily optimize points to be within the TCB. Moreover, such an approach does not prevent points that are initially inside the TCB to be pulled outside in order to push another point closer to but not within the TCB. This may result in an overall deterioration of the TC scores. Having an additional boundary at the mid-point was believed to further facilitate the algorithm in pushing these points inside the TCB, and to be less likely to lose accuracy in other points. We included

an additional penalty term for points outside of the TCB. That way, three regions can now be distinguished are: within the halfway boundaries, within the TCB and outside the TCB. Based on these regions, referring to Equation 4.29, we can re-define our weight bases for $k \in [1, 2, 3]$. Here, J_1 denotes the union of all three regions, J_2 denotes the region outside the halfway boundary, and J_3 denotes the region outside the TC boundary. To judge the effect of the halfway boundary model, we initially used weights of equal magnitude. Let us denote the weight base factors for equal weights as J_k^0, where for this initial model $(J_1^0 : J_2^0 : J_3^0)$ is $(1 : 1 : 1)$. The objective function is set up in the same way (only J_k are different) as the other models to be discussed now, and is denoted by E_4^0. Continuing, based on intermediate analyses, we chose weight ratios of $(1 : 5 : 10)$ for $(J_1 : J_2 : J_3)$, respectively referring to points within the HWB vs. points in between the HWB and the TCB vs. points outside the TCB.

$$\boldsymbol{W} = \begin{bmatrix} W_{1,l}(d) \\ W_{2,l}(d) \\ W_{3,l}(d) \end{bmatrix} = \psi_l(d) \begin{bmatrix} J_1 \\ J_2 \\ J_3 \end{bmatrix} \tag{4.35}$$

Note that J_2 is re-defined with respect to the penalized regular DVH model. Similar to penalized DVH-optimization, DVH OLS is still optimized for. However, the halfway boundaries are now involved, which are defined as the midpoint between the true DVH and TC boundaries:

$$c_n^{dwn}(d) = \frac{1}{2}\big(\widetilde{DVH}_n(d) + b_n^{dwn}(d)\big), \qquad \forall d \in [1, 2, \ldots, D] \tag{4.36}$$

$$c_n^{up}(d) = \frac{1}{2}\big(\widetilde{DVH}_n(d) + b_n^{up}(d)\big), \qquad \forall d \in [1, 2, \ldots, D] \tag{4.37}$$

such that the objective function can be written:

$$\begin{aligned} E_4 = \frac{1}{ND}\sum_{d=1}^{D}\sum_{n=1}^{N}\Bigg[& W_{1,l}(d)\cdot\Big(\widetilde{DVH}_n(d) - x_{d,n}(d)\Big)^2 + \\ & G_{\alpha,n}^{dwn}(d)\cdot W_{2,l}(d)\cdot\Big(c_n^{dwn}(d) - x_{d,n}(d)\Big)^2 + \\ & G_{\alpha,n}^{up}(d)\cdot W_{2,l}(d)\cdot\Big(c_n^{up}(d) - x_{d,n}(d)\Big)^2 + \\ & H_{\alpha,n}^{dwn}(d)\cdot W_{3,l}(d)\cdot\Big(b_n^{dwn}(d) - x_{d,n}(d)\Big)^2 + \\ & H_{\alpha,n}^{up}(d)\cdot W_{3,l}(d)\cdot\Big(b_n^{up}(d) - x_{d,n}(d)\Big)^2\Bigg] \end{aligned} \tag{4.38}$$

where $G_{\alpha,n}^{dwn}(d)$ and $G_{\alpha,n}^{up}(d)$ are the lower and upper halfway boundary step functions:

$$G_{\alpha,n}^{dwn}(d) = \frac{1}{2}\Big(sgn\big(x_{d,n}(d) - c_n^{dwn}(d)\big) + 1\Big) \tag{4.39}$$

$$G_{\alpha,n}^{up}(d) = \frac{1}{2}\Big(sgn\big(c_n^{up}(d) - x_{d,n}(d)\big) + 1\Big) \tag{4.40}$$

With $\big[f(\boldsymbol{A}_i, \boldsymbol{\zeta}_n)\big]_n$ being expressed as $\sum_{m=0}^{M} a_{i,m}[\zeta_m]_n V_i(d)$, written explicitly, the optimal regression coefficients result from Equation 4.41:

$$\begin{aligned}\min_{a_{i,m}} E_{4,l} = \min_{a_{i,m}} \frac{1}{ND} \sum_{d=1}^{D} \sum_{n=1}^{N} \Bigg[& W_{1,l}(d) \cdot \Big(\widetilde{DVH}_n(d) - \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m} [\zeta_m]_n \boldsymbol{V}_i(d) \Big)^2 + \\ & G^{dwn}_{\alpha,n}(d) \cdot W_{2,l}(d) \cdot \Big(c^{dwn}_n(d) - \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m} [\zeta_m]_n \boldsymbol{V}_i(d) \Big)^2 + \\ & G^{up}_{\alpha,n}(d) \cdot W_{2,l}(d) \cdot \Big(c^{up}_n(d) - \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m} [\zeta_m]_n \boldsymbol{V}_i(d) \Big)^2 + \\ & H^{dwn}_{\alpha,n}(d) \cdot W_{3,l}(d) \cdot \Big(b^{dwn}_n(d) - \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m} [\zeta_m]_n \boldsymbol{V}_i(d) \Big)^2 + \\ & H^{up}_{\alpha,n}(d) \cdot W_{3,l}(d) \cdot \Big(b^{up}_n(d) - \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m} [\zeta_m]_n \boldsymbol{V}_i(d) \Big)^2 \Bigg] \end{aligned} \tag{4.41}$$

Similar to the regular penalized DVH objective function, it should be noted that there is no intuitive interpretation of the physical meaning of the objective function value, because the normalization factor does not account for penalized terms. Similar to Equation 4.13, $G^{dwn}_{\alpha,n}$ and $G^{up}_{\alpha,n}$ are separately defined for all n and d.

4.5.4 TC-optimizations

TC optimizers are different from its predecessors in the sense that they do not use any form OLS optimization in order to determine the optimal linear regression coefficients. Instead, cost functions are based on the TC evaluation metrics discussed in Section **??**. We have investigated TC-based regressors for both the TC_α and TC_β. As discussed, TC boundaries can be perfectly modeled with step functions (Equations 4.13 and 4.15). However, because the finite-differencing methods used by *fmincon* to estimate the gradient field fall short in this situation (gradients are 0 at every infinitesimally small point in coefficient hyperspace), it fails to converge to a global optimum. To deal with this, we used sigmoid-approximations of the point and patient step functions, with the goal of better enabling the optimizer to navigate through the objective function hyperspace. The TC_β sigmoid, an example TC_α sigmoid function and their approximations are shown in Figure 4.13. Providing the gradient field directly to the optimizer for more accurate results has been looked into, but the gradients were not used to obtain the results for these models. We provide the method to demonstrate how it can be done for future reference.

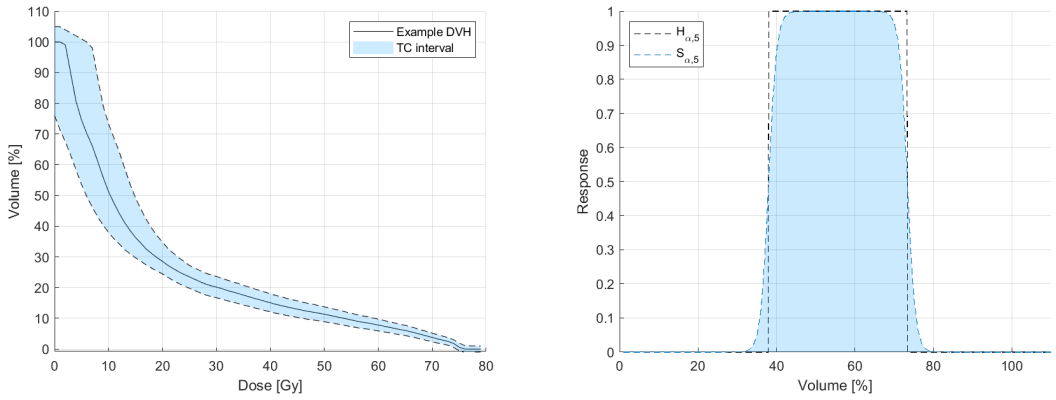
TC_α-optimization

The first TC-based model optimizes for the total number of points within the TCB. We do this by approximating the step function in Equation 4.13, as this proved to be sufficient for the optimizer to converge. This is a double-sided function that must be determined for each patient, at every dose bin. Its sigmoid approximation can be described as:

$$S_{\alpha,n}(d) = \left[\left[1 + e^{-\eta \left(x_{d,n}(d) - b^{dwn}_n(d) \right)} \right]^{-1} - \left[1 + e^{-\eta \left(x_{d,n}(d) - b^{up}_n(d) \right)} \right]^{-1} \right] \tag{4.42}$$

where η is the sigmoid steepness and its value was arbitrarily chosen to be 100. $b^{dwn}_n(d)$ and $b^{up}_n(d)$ define the centerpoints of the sigmoid, which are simply the lower and upper

TCB values at dose bin d. The TC_α optimizer incorporates Equation 4.42 in the objective function to find a solution that maximizes the TC_α.



(a) (b)

Figure 4.13: (a) Example DVH, n = 5 (b) Example representation of the step function and its double-sigmoid approximation that models the TC boundary at dose d = 10 Gy.

For each patient n separately, every point in DVH_n is fed into $S_{\alpha,n}(d)$ and summed over the number of dose bins to find the amount of points inside the TC boundary:

$$\begin{aligned} TC_{\alpha,n} &= \sum_{d=1}^{D} S_{\alpha,n}\big(x_{d,n}(d)\big) \\ &= \sum_{d=1}^{D} \left[\left[1 + e^{-\eta(x_{d,n}(d) - b_n^{dwn}(d))}\right]^{-1} - \left[1 + e^{-\eta(x_{d,n}(d) - b_n^{up}(d))}\right]^{-1} \right] \end{aligned} \tag{4.43}$$

Since we want to maximize TC_α, the objective function can be written as the sum of $-1 \cdot TC_{\alpha,n}$ averaged over all patients and dose bins:

$$E_{\alpha,5} = \frac{-1}{ND} \sum_{n=1}^{N} \sum_{d=1}^{D} \left[\left[1 + e^{-\eta(x_{d,n}(d) - b_n^{dwn}(d))}\right]^{-1} - \left[1 + e^{-\eta(x_{d,n}(d) - b_n^{up}(d))}\right]^{-1} \right] \tag{4.44}$$

where $DVH_n(d) = \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m}[\zeta_m]_n \boldsymbol{V}_i(d)$, and the optimal regression coefficients are found by minimizing E_5:

$$\begin{aligned} \min_{a_{i,m}} E_{\alpha,5} = \min_{a_{i,m}} \frac{-1}{ND} \sum_{n=1}^{N} \sum_{d=1}^{D} \Bigg[&\left[1 + e^{-\eta(x_{d,n}(d) - b_n^{dwn}(d))}\right]^{-1} - \\ &\left[1 + e^{-\eta(x_{d,n}(d) - b_n^{up}(d))}\right]^{-1} \Bigg] \end{aligned} \tag{4.45}$$

TC_β-optimization

The second TC-based optimizer is one that optimizes for the number of patients that pass the criterion of having more than ν % of its dose bins inside the TC boundary. A straightforward way to do this is by approximating the TC_β step function (Equation 4.15) with a sigmoid as we did for TC_α. This approximation is shown by $S_{\beta,n}$:

$$S_{\beta,n}\big(TC_{\alpha,n}\big) = \left[1 + e^{-\theta(TC_{\alpha,n} - \nu \cdot D)}\right]^{-1} \tag{4.46}$$

where ν denotes the threshold required for the number of correctly predicted points to be considered a passing DVH prediction, and θ is the steepness of the sigmoid. However, in the TC_β optimization, the objective function becomes significantly more complex, and convexity issues proved problematic. Either the optimizer was unable to identify the gradient field of the objective function, or converged immediately to a local optimum. It would require excessive smoothing with the θ term, such that *fmincon*s optimization becomes unreliable. For that reason, a more complicated approximation was chosen, in order to better enable *fmincon* to identify the gradient field. Instead of Equation 4.46, we used an alternative of the following general form:

$$I_\beta(TC_{\alpha,n}) = \frac{I'_\beta(TC_{\alpha,n})}{I'_\beta(D)} \tag{4.47}$$

$$I'_\beta(TC_{\alpha,n}) = U_1(TC_{\alpha,n})\Big(1 - U_2(TC_{\alpha,n})\Big) + U_2(TC_{\alpha,n}) \tag{4.48}$$

Where U_1 and U_2 are:

$$U_1(TC_{\alpha,n}) = \left(2^{\frac{TC_{\alpha,n}}{\nu \cdot D}} - 1\right)^R \tag{4.49}$$

$$U_2(TC_{\alpha,n}) = \left(\frac{1}{\pi}\left(tan^{-1}(TC_{\alpha,n} - \nu \cdot D) + \frac{\pi}{2}\right)\right) \tag{4.50}$$

Here, R = 9 and is chosen such that $I_\beta\big(TC_{\alpha,n} = \nu \cdot D)\big) = 90\%$. The I_β and $S_{\beta,n}$ approximations of H_2 are shown in Figure 4.14.

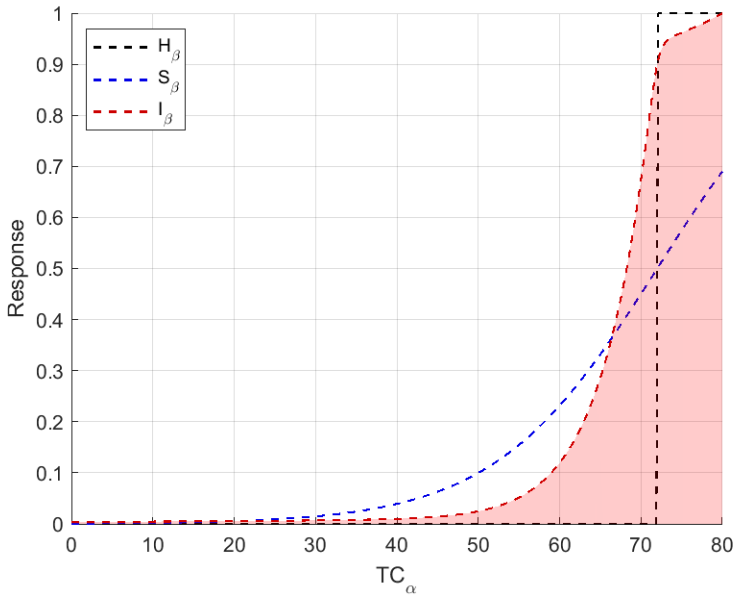


Figure 4.14: TC_β step function, H_β, and its approximations, I_β and S_β

This function shape was chosen to help the objective function push patients with $TC_{\alpha,n}$ close to but below $\nu \cdot D$ within the TC boundary. This is expected to happen, because there is much to win in terms of cost minimization in the $60 - 72$ TC_α region. In the region $TC_\alpha > 72$, I_β increases more slowly, and allows for further optimization of $TC_{\alpha,n}$, even when the objective of reaching $\geq 90\%$ of the DVH has already been reached. The red area under I_β is used to emphasize the are under the curve used for the objective function. As can be seen, S_β requires too much smoothing with the θ term to be used as a reliable approximation of H_β.

Next, TC_β is found by averaging the optimized result of I_β, to obtain our final objective function value:

$$E_{\beta,5} = \frac{1}{N}\left[\sum_{n=1}^{N} I_\beta(TC_{\alpha,n})\right] \tag{4.51}$$

This leaves us with our objective function value that approximates the fraction of patients that pass the TC_β condition. Remembering that, in order to maximize TC_β, we need can simply multiply the objective function by -1 and minimize. For the resulting regression coefficients that maximize TC_β, we can write:

$$\min_{a_{i,m}} E_{\beta,5} = \min_{a_{i,m}} \frac{-1}{N}\left[\sum_{n=1}^{N} I_\beta(TC_{\alpha,n})\right] \tag{4.52}$$

TC_β objective function gradients

Alternatively, one could provide the objective function gradient field to *fmincon*. This has the advantage that the optimizer can better navigate through coefficient gradient space, allowing *fmincon* to converge to points that better represents the true outcome of TC_β. In order to provide the gradient field to *fmincon*, we have to provide the analytical expressions of:

$$\frac{\partial TC_\beta(\boldsymbol{a}_i)}{\partial a_{i,m}} \;\; \forall\, a_i, i \in [1,\ldots,C],\; \forall\, a_m, m \in [0,\ldots,M] \tag{4.53}$$

Generally, the C^{th} M coefficients contribute to the C^{th} eigenvalue. Or, one could say that we use the same regression model C times to predict each corresponding eigenvalue, where $\frac{\partial[\lambda_i]_n}{\partial a_{i,m}} = [\zeta_{i,m}]_n$. The partial derivatives are found from:

$$\frac{\partial[\lambda_i]_n}{\partial a_{i,m}} = [\zeta_{i,m}]_n \tag{4.54}$$

Then, from Equation 4.3 we find:

$$\frac{\partial DVH_n}{\partial a_{i,m}} = [\zeta_{i,m}]_n \boldsymbol{V}_i \tag{4.55}$$

Remembering that the derivative of a sigmoid has the general form:

$$\frac{d}{dx} sigm(x) = \frac{d}{dx}[1 + e^{-\gamma(x-c)}]^{-1} = \frac{\gamma e^{-\gamma(x-c)}}{(1 + e^{-\gamma(x-c)})^2} \tag{4.56}$$

It can then be seen from Equation 4.43 that $TC_{\alpha,n}$ for patient n results:

$$\begin{aligned}\frac{\partial TC_{\alpha,n}}{\partial a_{i,m}} = \sum_{d=1}^{D}\Big[&\frac{\eta\, e^{-\eta(DVH_n(d) - b_n^{dwn}(d))}}{(1 + e^{-\eta(DVH_n(d) - b_n^{dwn}(d))})^2} \\ &- \frac{\eta\, e^{-\eta(DVH_n(d) - b_n^{up}(d))}}{(1 + e^{-\eta(DVH_n(d) - b_n^{up}(d))})^2}\Big] \cdot \frac{\partial DVH_n(d)}{\partial a_{i,m}}\end{aligned} \tag{4.57}$$

Since we analytically provide the gradient, there is no need for a rough approximation of H_β in order for *fmincon* to estimate the gradient. In this case, using $S_{\beta,n}$ (Equation 4.46), with a large steepness suffices. We can then find the partial derivative of the final objective function TC_β:

$$\frac{\partial TC_\beta}{\partial a_{i,m}} = \frac{1}{N}\,\frac{\partial}{\partial a_{i,m}} \sum_{n=1}^{N} S_{\beta,n}(TC_{\alpha,n}) \;\cdot \frac{\theta e^{-\theta(x-c)}}{\left(1 + e^{-\theta(x-c)}\right)^2} \cdot \frac{\partial TC_{\alpha,n}}{a_{i,m}} \tag{4.58}$$

4.5.5 Constrained DVH-optimizations

All of the previously mentioned prediction approaches are unconstrained. So, these models are allowed to produce any DVH that best fits the objective function. Consequently, predicted DVHs are allowed to be unrealistic. This means that they can show DVH points above 100% or below 0%, or that the cumulative DVH has a non-negative derivative. For this reason, we investigated the possibilities to constrain optimizations to counteract these problems and ensure physical DVH predictions. Imposing constraints on the DVH becomes straightforward as a direct result of having the non-weighted DVH optimizer. This is because the DVH-optimizer objective function is directly expressed by the regression coefficients, as opposed to the EV-optimizer, that requires an intermediary DVH-reconstruction step. The optimization that was solved for is the same as shown in Equation 4.28, however now subject to the following constraints:

(i) $\forall j, \forall n, \qquad DVH_n(d_j) \leq 100\%$

(ii) $\forall j, \forall n, \qquad DVH_n(d_j) \geq 0\%$

(iii) $\forall j, \forall n, \qquad \frac{dDVH_n}{dd_j} \leq 0\%$

We have investigated strict and flexible hard constrained optimizations. In strict optimizations, we allow absolutely no violations of the requirements to DVH physicality. For the flexible case, we hard constrain on the requirements to DVH physicality, allowing a larger margin. Written explicity in terms of the regression coefficients, optimizations are constrained to:

$$\forall j, \forall n, \qquad \overline{DVH}(d_j) + \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m}[\zeta_{i,m}]_n \cdot \boldsymbol{V}_i(d_j) - 100\% \leq \Theta_1 \tag{4.59}$$

$$\forall j, \forall n, \qquad \overline{DVH}(d_j) + \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m}[\zeta_{i,m}]_n \cdot \boldsymbol{V}_i(d_j) \leq \Theta_2 \tag{4.60}$$

$$\forall j, \forall n, \qquad \frac{d(\overline{DVH}(d_j))}{dd_j} + \sum_{i=1}^{C} \sum_{m=0}^{M} a_{i,m}[\zeta_{i,m}]_n \cdot \frac{d(\boldsymbol{V}_i(d_j))}{dd_j} \leq \Theta_3 \tag{4.61}$$

For strictly constrained optimizations, Θ_1, Θ_2 and Θ_3 are 100%, 0% and 0% respectively. For the flexibly constrained optimizations, Θ_1, Θ_2 and Θ_3 are chosen 102%, -0.5% and 1% respectively. The model used for constrained optimizations was the non-weighted DVH optimizer.

4.5.6 Post-processing

A potential pitfall for strictly constrained optimizations could be that this may heavily influence DVH prediction accuracy of our models. For that reason, the most pragmatic approach would be to simply force the constraints mentioned in Section 4.5.5 as a post-processing step to all DVHs. This step is performed on all non-constrained methods.

4.6 Validation

4.6.1 K-fold cross-validation

In order to choose the appropriate fold sizes for K-fold CV, we performed multiple K-fold (from K = 2 to K = 10) cross-validations to investigate the sample size needed to sufficiently train our models. This was done for the EV-optimizer, the direct DVH-optimizer, and the penalized direct DVH optimizer models. In addition, we investigated for all evaluation metrics the training and testing errors with respect to the amount of features included, thereby getting an idea at which point overfitting occurs. The optimal number of features for all evaluation metrics were estimated 10. For this reason, L = 10 was chosen (Equation 4.22). The graphs on which these choices were based can be viewed in the appendix (Figures B.2 - B.4).

4.6.2 Training and testing metrics

All prediction methods resulted in a model that takes geometrical features as input, and outputs a set of DVHs. The resulting DVHs are evaluated by comparing the prediction to the true DVH in terms of the DVH bins RMS, $TC_{\alpha,n}$ and TC_{β}. These are also the error statistics that we use for both training and testing. We have chosen to use 8-fold CV in all of our optimization-based methods. However, as the only exception, also LOOCV was done for the two-point method. For the optimization-based methods, the 8-fold CV resulted in 8 differently trained models. All 8 CVs combined ensure that each patient is tested once. Since patients end up 7 times in a training set, there are 7 models that result in a training prediction for every patient. These predictions were averaged to find the average training prediction. The cross-validation spread is determined by the range $(maximum - minimum)$ of cross-validation errors.

4.7 Model overview

For the sake of clarity, this section contains an overview of all the used models. This overview is displayed below, in Table 4.3

General model objective	Model	Optimized quantity	Sub-model	Objective function
Residuals minimization	Two-point predictor	V_{95} V_{mean}	-	-
	EV	DVH PCA EVs	-	E_1^i
	DVH	DVH dose bins	-	E_2
	$DVH-TC$		ϕ_1	$E_{3,1}$
			ϕ_2	$E_{3,2}$
			ϕ_3	$E_{3,3}$
			ϕ_4	$E_{3,4}$
			ϕ_5	$E_{3,5}$
	$DVH-HWB$	DVH dose bins	J_k^0	E_4^0
			ϕ_1	$E_{4,1}$
			ϕ_2	$E_{4,2}$
			ϕ_3	$E_{4,3}$
			ϕ_4	$E_{4,4}$
			ϕ_5	$E_{4,5}$
	Constrained DVH	DVH dose bins	Strict Flexible	E_2 E_2
TC-metrics maximization	TC_α	S_α	-	$E_{5,\alpha}$
	TC_β	I_β	-	$E_{5,\beta}$

Table 4.3: An overview of the used models. Models can be roughly distinguished by the general shape of their objective function. Two classes of objective functions are those that minimize residuals between the true and optimized quantity, and those that maximize the proposed $TC-$metrics, as indicated by the blue and red rows respectively. The models are named based on the model structure. The models $DVH-TC$ and $DVH-HWB$ represent the respective models that were penalized based on the TCB and the HWB. For some models, we have investigated similar models that differed slightly in one aspect, but the general structure of the objective function is the same. The variable or manners by which sub-models differ is indicated by the sub-model column. The objective function column displays the symbols used to indicate each objective function as proposed throughout Section 4.5

CHAPTER 5

RESULTS

In this chapter we will start by displaying the PCA results of our data sets, followed by results from our feature selection scheme. From there, we will provide the results obtained with our used models, starting off with the two-point predictor. Each model section covers global results that say something about overall performances with respect to the entire population. Following, we will show a number of DVH predictions that are specific to that model. With that being said, it should be noted that for the two-point predictor, 95% CIs are included for their predictions, whereas for the optimization-based models there is not. They are all, however, evaluated in the same way. A second difference is that the two-point predictor is validated using LOOCV, in addition of 8-fold CV, whereas for all optimization-based methods we used only 8-fold CV. For all optimization methods, we train the models with 10 features. We decided to omit analyses done with logarithmic features from model evaluation sections, as none yielded better results for testing data, and ensued a greater degree of overfitting.

5.1 Preliminary analyses

5.1.1 Principal Component Analysis

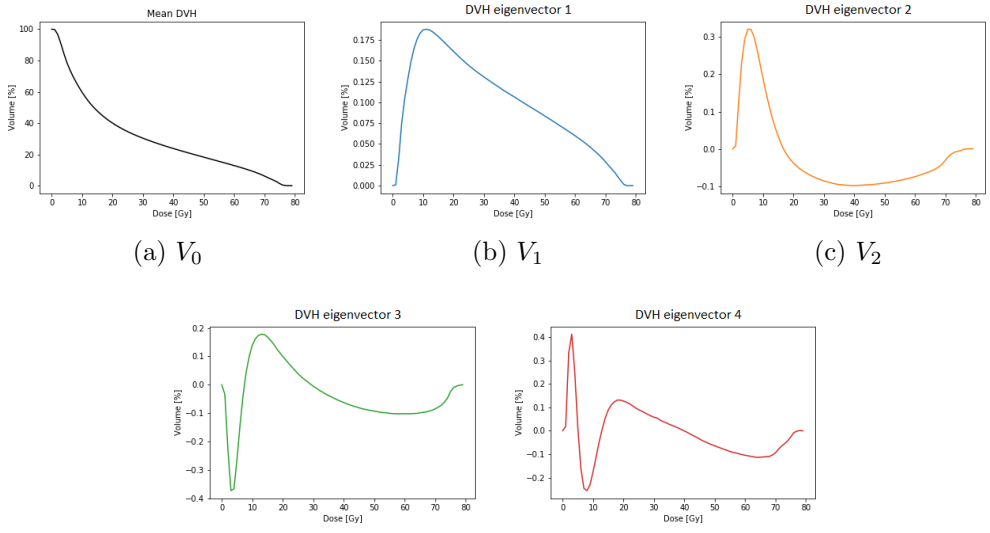
The first analysis involved doing the PCA, and finding the resulting variance ratios explained (EVRs) by each PC mode. We calculated the EVRs for the rectum DVH, for the rectum OVH and for the AS OVH. The results are displayed in Table 5.1. In addition, the PCA-reconstructed set of DVHs TC_β was calculated for each subsequent PC mode addition. These results are shown in Table 5.2. It was based on the information in these tables that we have decided to include 4 rectum PCA components in our models for prediction. These PC modes are shown in Figure 5.1

EVR:	1	2	3	4	5	6	7	8
Rectum DVH	87.79	95.66	98.73	99.42	99.66	99.83	99.89	99.93
Rectum OVH	84.46	95.69	98.41	99.37	99.69	99.85	99.91	99.94
AS OVH	88.82	98.25	99.37	99.66	99.81	99.89	99.92	99.94

Table 5.1: Cumulative explained variance ratios for the first 10 principal components of the rectum DVH and OVH, and the anal sphincter OVH

PC modes:	0	1	2	3	4	5	6	7	8
RMS	8.11	2.83	1.69	0.915	0.617	0.473	0.334	0.271	0.220
CF_α	42.70	85.84	93.37	98.61	99.48	99.63	99.76	99.95	99.95
CF_β	11.96	57.61	76.09	94.56	100	100	100	100	100

Table 5.2: The evaluation metrics of the DVH dataset reconstructed with an increasing number of PC modes. The 0^{th} PC mode designates the average DVH. These values serve as an upper boundary for model performance.



(a) V_0 (b) V_1 (c) V_2

(d) V_3 (e) V_4

Figure 5.1: Eigenvectors resulting from the rectum data set PCA

5.1.2 Feature analysis

The best 5 polynomial features resulting from the methods described in Section 4.4.4, and their corresponding cumulative correlation coefficients with each of the DVH eigenvalues are shown Table 5.3. For the remaining features, as well as the best linear features, and the best features including logarithmic features, the reader can examine Tables B.1 and B.2, as appended. In addition, a comparison between the cumulative R^2 for different feature selection methods are shown in Figure 5.2. This can be seen for linear features and for polynomial features in appendix Figure B.1

DVH	EVs	λ_1	λ_2	λ_3	λ_4
1	Feature	$OVH(10)$	$\lambda_3^{rect} \cdot V_{PTV}$	$\lambda_3^{rect} \cdot V_{PTV}$	$\lambda_1^{rect} \cdot \lambda_2^{rect}$
	R^2	0.8458	0.3696	0.2743	0.1453
2	Feature	$OVH(10) \cdot V_{rect}$	λ_2^{rect}	$\lambda_2^{rect} \cdot V_{PTV}$	$OVH(0) \cdot \frac{O_3}{O_4}$
	R^2	0.8794	0.6606	0.4878	0.2523
3	Feature	$\lambda_1^{rect} \cdot V_{PTV}$	$V_{PTV} \cdot \frac{O_1}{O_2}$	$(OVH(0))^2$	$OVH(10) \cdot \frac{O_3}{O_4}$
	R^2	0.8972	0.7590	0.6322	0.3081
4	Feature	$\lambda_3^{rect} \cdot \lambda_2^{AS}$	$(\lambda_3^{AS})^2$	$OVH(10) \cdot V_{rect}$	$(\lambda_3^{AS})^2$
	R^2	0.9075	0.7778	0.6855	0.3699
5	Feature	$(\lambda_2^{AS})^2$	$\lambda_2^{AS} \cdot V_{rect}$	$(\lambda_1^{AS})^2$	$\lambda_3^{rect} \cdot \lambda_3^{AS}$
	R^2	0.9120	0.7909	0.7113	0.4164

Table 5.3: The first 5 best correlating polynomial features for each of the four DVH eigenvalues, as indicated by the columns. The eigenvalues written in the table indicate rectum and anal sphincter OVH PCA eigenvalues

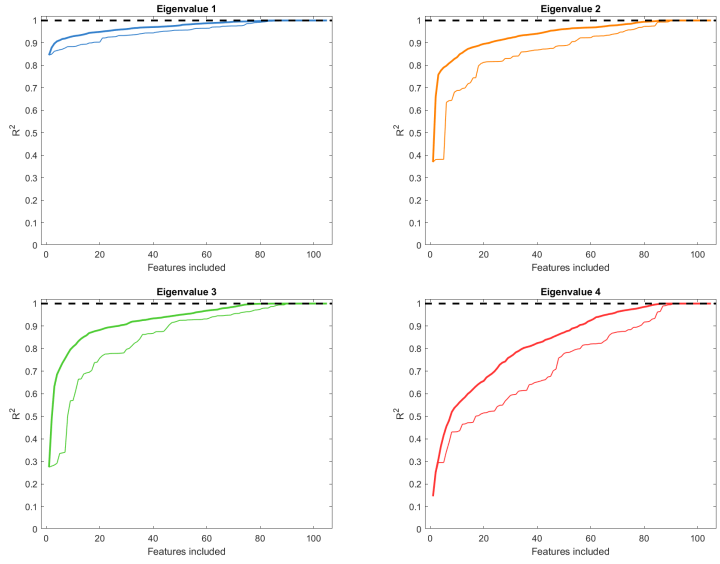
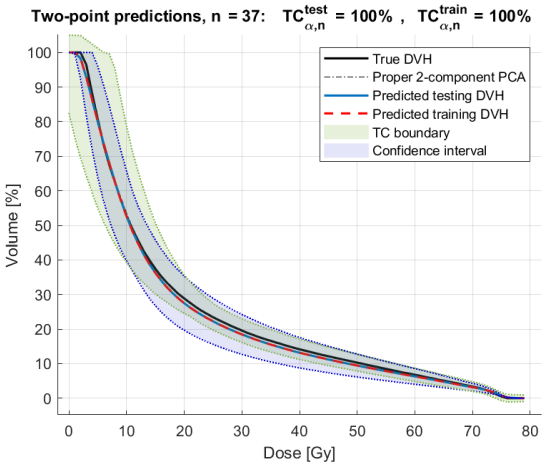


Figure 5.2: R^2 relations for all EVs when fitted with an increasing number of polynomial (Q = 2) features, using only non-logarithmic features. Thin and thick lines denote filter and filter-wrapper methods results respectively.

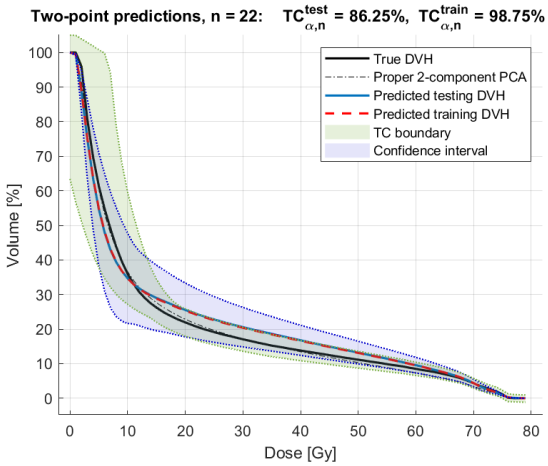
5.1.3 Clinical practice analysis

	RMS	TC_α	TC_β
8-fold CV train errors	4.8042	81.40%	53.26%
8-fold CV test errors	4.9902	80.79%	51.09%
LOOCV test errors	4.9385	80.45%	50.00%

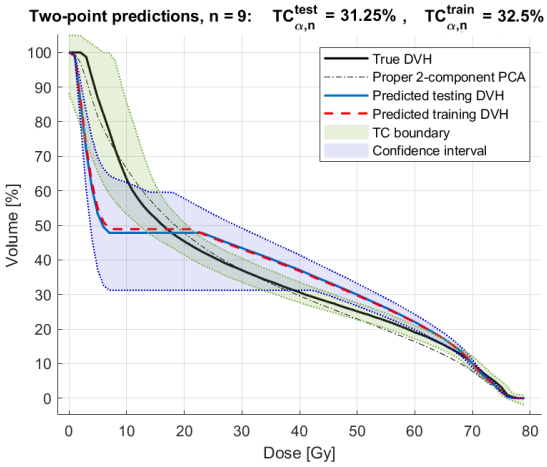
Table 5.4: Training and testing errors resulting from the two-point predictor model



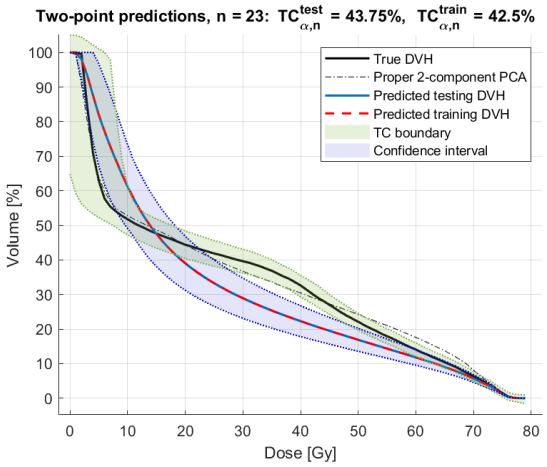
(a) A good prediction



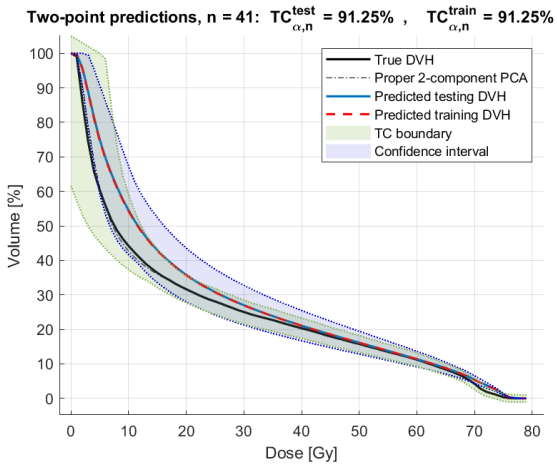
(b) An average prediction



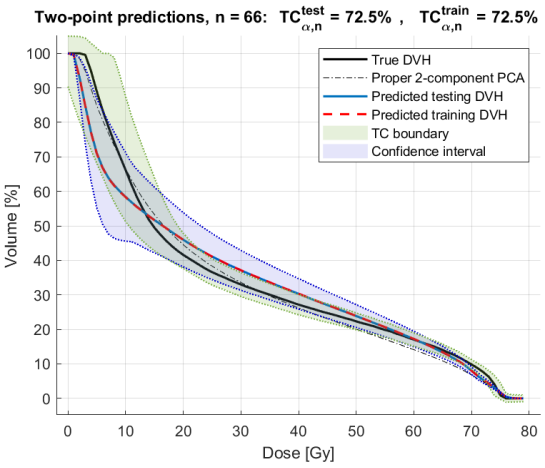
(c) A bad prediction



(d) The PCA falls short



(e) Predictions barely fit the TCB



(f) Predictions barely miss the TCB

Figure 5.3: Overview of typical DVH-predictions acquired with the two-point predictor

The prediction of patient n = 9 in Figure 5.3c is an example of a bad prediction, where the two-point predictor is insufficient in correctly predicting most of the DVH, and the confidence interval even more so. Figure 5.3d shows an example where the (two-component) PCA-reconstructed DVH falls for a large part outside of the TC boundary. Figures 5.3e and 5.3f show examples of decent (but typical) DVH predictions that run closely along the boundary of the TCB.

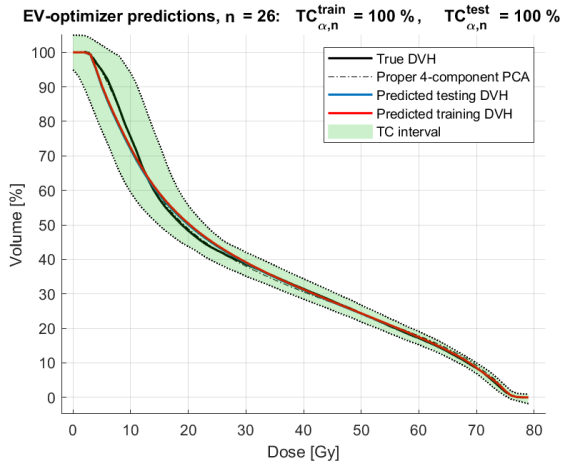
5.2 Optimization-based predictions

5.2.1 EV-optimization predictions

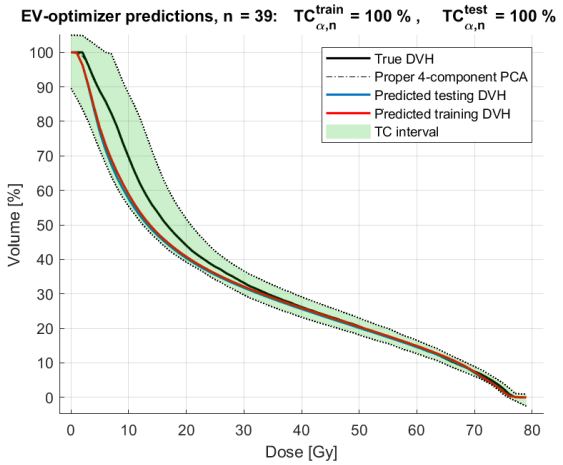
The CV average and spread train and test errors resulting from the EV-optimization method are shown in Tables 5.5. Figure 5.4 shows some representative DVHs predicted with this model.

	Training errors			Testing errors		
	RMS	TC_α	TC_β	RMS	TC_α	TC_β
Average	2.8213	86.46%	68.83%	3.2242	82.40%	59.09%
Spread	0.1445	2.47%	6.17%	1.1891	20.34%	36.37%

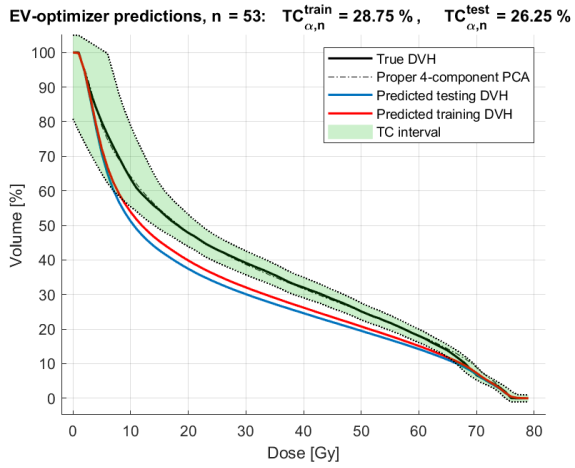
Table 5.5: EV-optimization 8-fold CV train and test errors



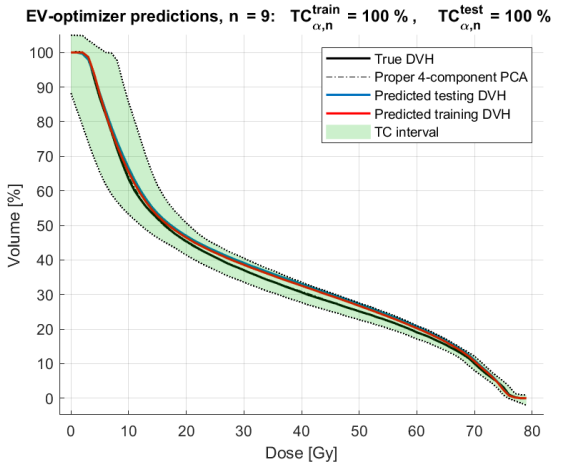
(a) A good prediction



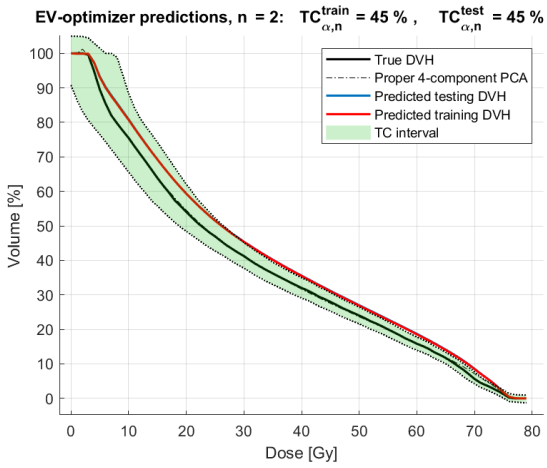
(b) An average prediction



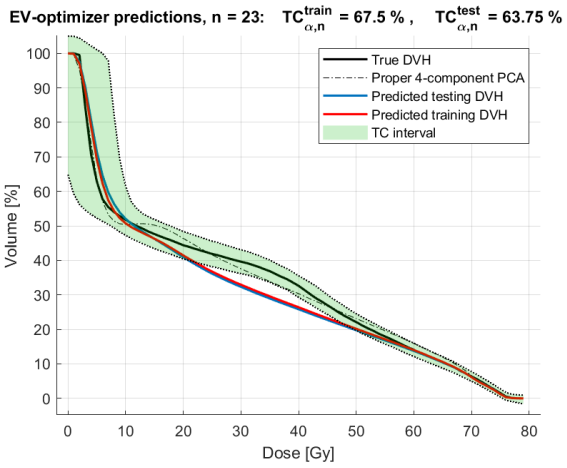
(c) A bad prediction



(d) An improved prediction (Figure 5.3a)



(e) Otherwise decent prediction failing the TC



(f) The PCA falls short

Figure 5.4: Some typical DVH-predictions acquired with the EV-optimized model

In Figure 5.4d we see an improved prediction for specifically patient n = 9 (Figure 5.4d, compared to the two-point predicted DVH in Figure 5.3a. Also, we can see in Figure 5.4f an example where the DVH reconstructed from the true DVH eigenvalues fails to be in the tolerance criterion along the entirety of the DVH. What's more, this proper reconstruction exhibits unphysical behaviour. In addition, the DVH prediction fails to capture the behaviour of the PCA.

5.2.2 DVH-optimization predictions

This subsection covers the results of direct DVH-optimization predictions of the regular DVH optimizer, as well as the penalized DVH optimizers, as well as the halfway-boundary-assisted models. This includes both the non-weighted and weighted least squares models.

Regular DVH optimization

The CV average and spread training and testing errors resulting from the non-weighted and weighted regular DVH-optimization method are summarized in Table 5.6. Figure 5.5 shows some representative DVHs predicted with the non-weighted DVH optimizer. Predictions made with the penalized DVH predictions were omitted, because based on Table 5.6, it was believed that this optimizer performed about equally well.

Halfway boundary optimization

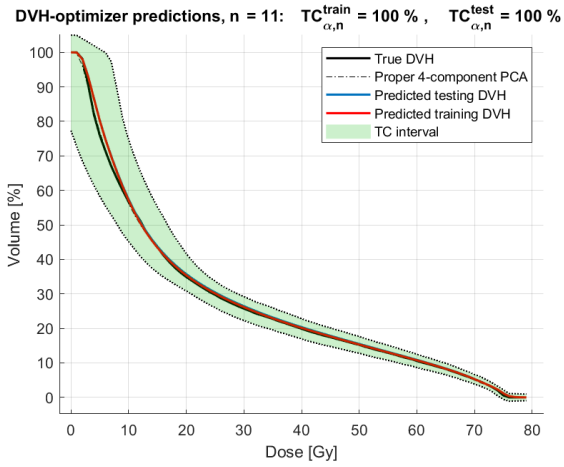
The CV average and spread training and testing errors resulting from the non-weighted and weighted DVH halfway-boundary-optimization method are summarized in Table 5.7. Figure 5.6 shows some representative DVHs predicted with this model. For the same reason as for the regular DVH-optimizer, we omitted figures of predictions by the penalized halfway-boundary-assisted models.

		Training errors			Testing errors		
Sub-model		RMS	TC_α	TC_β	RMS	TC_α	TC_β
Weightless	Average	2.8271	86.66%	69.44%	3.2294	81.90%	61.36%
	Spread	0.1064	1.18 %	1.23 %	0.7869	13.97%	36.37%
Stationary	Average	2.9028	85.46%	65.12%	3.3390	81.53%	63.64%
	Spread	0.1059	2.11 %	6.18 %	0.7532	8.29 %	27.27%
Linear	Average	3.0911	87.20%	66.36%	3.5946	82.02%	57.95%
	Spread	0.1503	2.87 %	11.11%	0.8962	9.55 %	9.09 %
Quadratic	Average	3.0010	85.81%	64.20%	3.4923	81.18%	57.95%
	Spread	0.1371	1.75 %	4.94 %	0.8011	8.87 %	27.27%
Exponential	Average	2.9973	86.46%	65.12%	3.4803	82.02%	62.50%
	Spread	0.1320	2.02 %	7.41 %	0.8078	8.30 %	18.18%
Double	Average	3.2867	86.28%	62.04%	3.8295	81.36%	56.82%
sigmoid	Spread	0.1920	2.36 %	9.88 %	0.9331	15.22%	27.27%

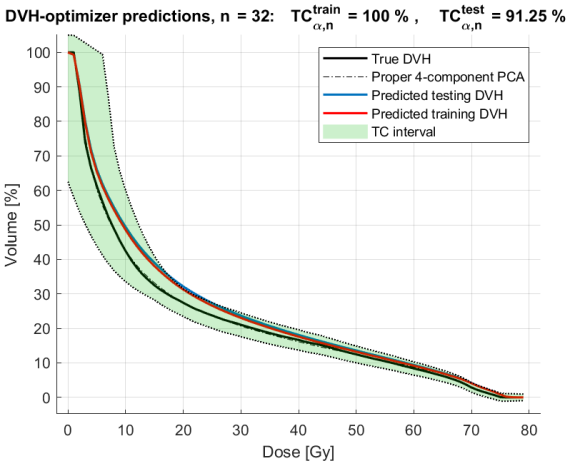
Table 5.6: Summary for all regular non-weighted and weighted DVH optimizations. Each color represents the weight as shown in Figure 4.11

		Training errors			Testing errors		
Sub-model		RMS	TC_α	TC_β	RMS	TC_α	TC_β
Weightless	Average	2.8362	86.57%	68.21%	3.2274	82.37%	63.64%
	Spread	0.1889	2.88 %	9.87%	1.4093	17.04%	27.27%
Stationary	Average	2.9541	87.19%	68.98%	3.4120	82.76%	60.23%
	Spread	0.1343	3.29%	7.41%	1.4257	18.86%	27.27%
Linear	Average	3.1463	88.10%	67.44%	3.5105	81.76%	56.82%
	Spread	0.2197	3.89%	12.34%	1.6298	13.64%	36.37%
Quadratic	Average	3.0826	87.63%	67.75%	3.4203	82.20%	57.95%
	Spread	0.2023	3.67%	12.34%	1.4983	13.18%	27.27%
Exponential	Average	3.0765	88.08%	68.21%	3.4114	82.22%	56.82%
	Spread	0.2053	4.95%	13.58%	1.5110	12.61%	36.37%
Double	Average	3.3657	88.17%	66.36%	3.7512	81.51%	55.68%
sigmoid	Spread	0.2824	3.16%	11.11%	1.8101	16.70%	45.46%

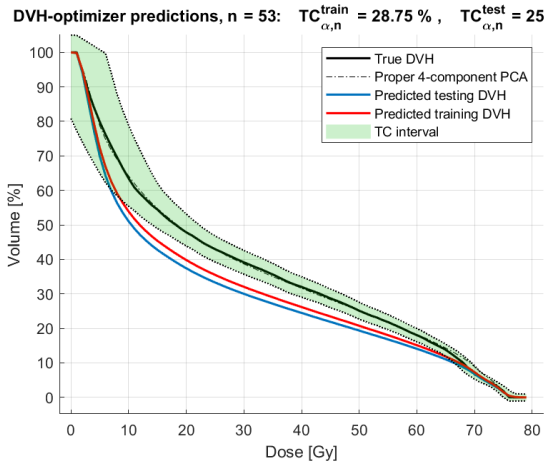
Table 5.7: Summary for all halfway boundary non-weighted and weighted DVH optimizations. Each color represents the weight as shown in Figure 4.11



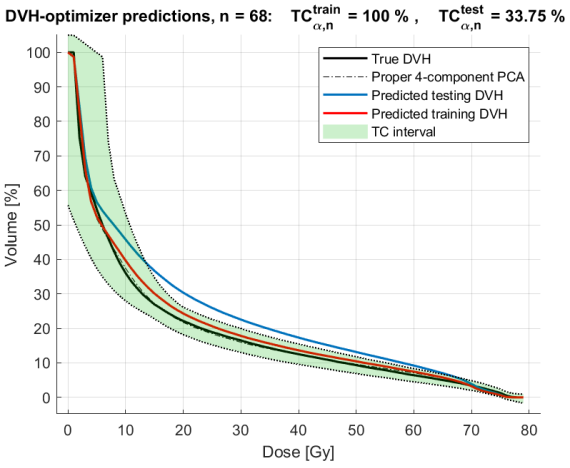
(a) A good prediction



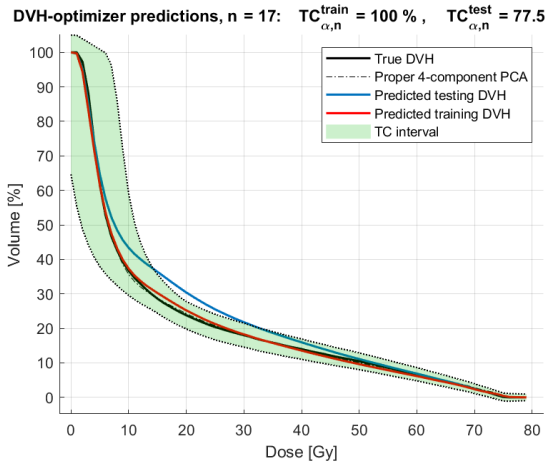
(b) An average prediction



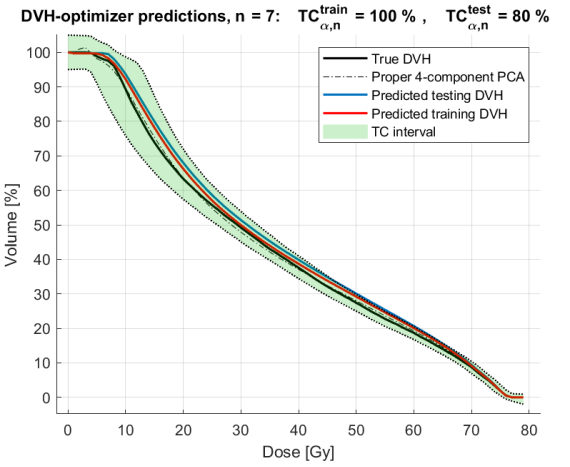
(c) A bad prediction



(d) The test case barely fails

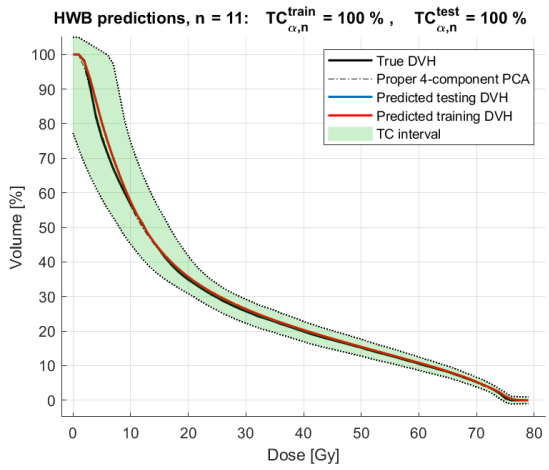


(e) The test prediction fails

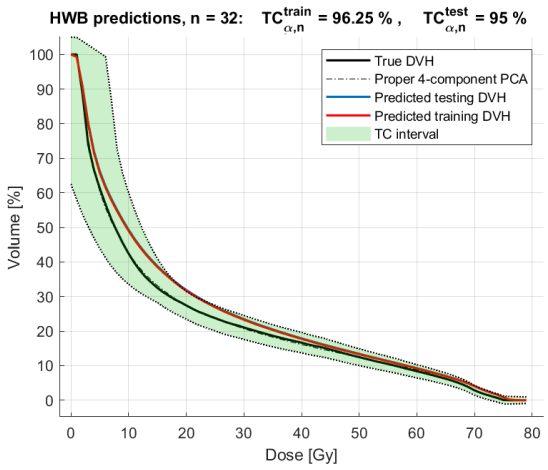


(f) The predictions are barely outside of the TCB

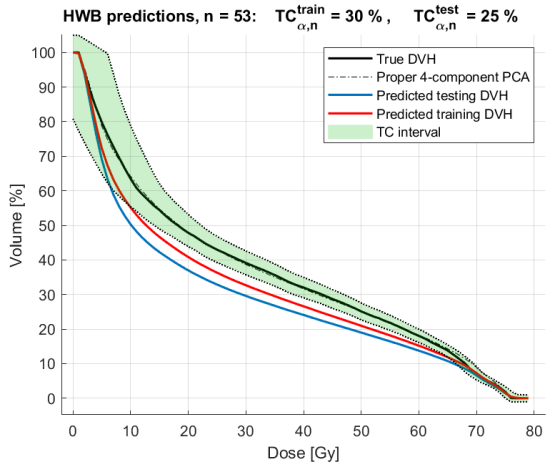
Figure 5.5: Overview of some typical DVH-predictions acquired with the DVH-optimized model



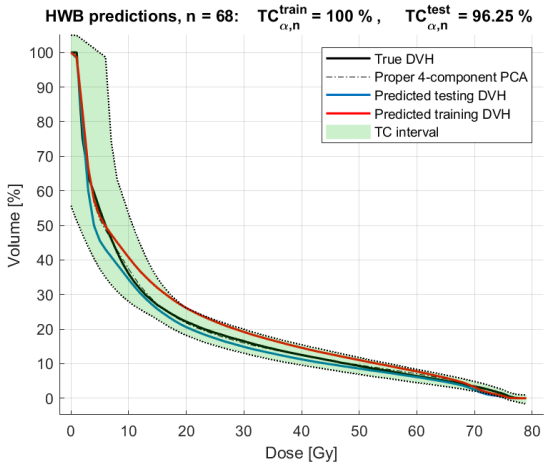
(a) A good prediction



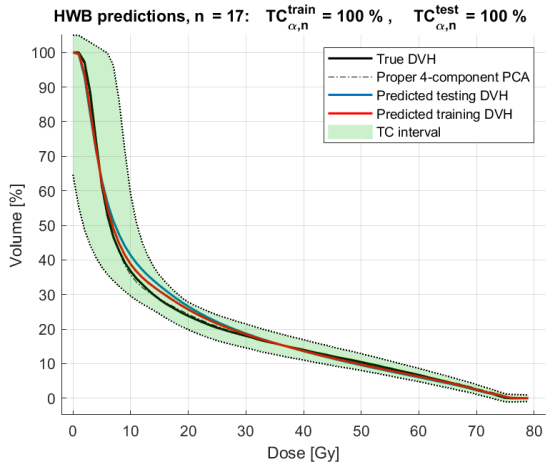
(b) An average prediction



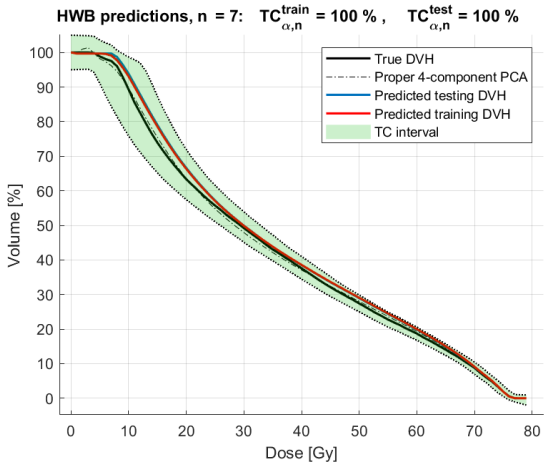
(c) A bad prediction



(d) Improved example compared to 5.5d



(e) Another improved example compared to 5.5e



(f) A third improved example compared to 5.5f

Figure 5.6: Overview of some typical DVH-predictions acquired with the DVH halfway-boundary-assisted model

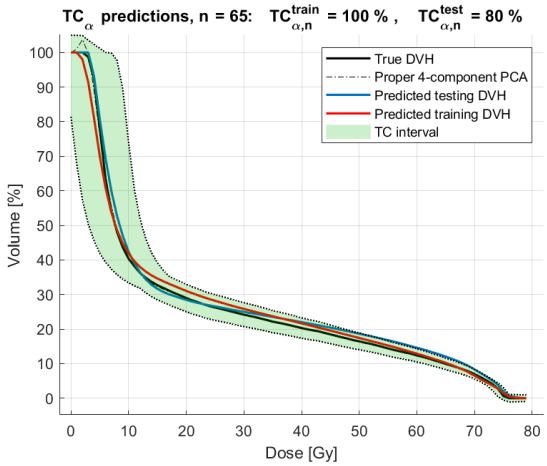
5.2.3 TC-optimization predictions

Maximizing TC_α

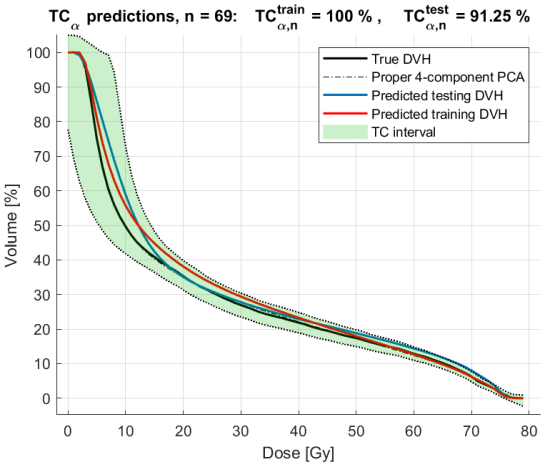
The CV average and spread train and test errors resulting from the TC_{pts}-optimization method are summarized in Table 5.8. Figure 5.7 shows some representative DVHs predicted with this model.

	Training errors			Testing errors		
	RMS	TC_α	TC_β	RMS	TC_α	TC_β
Average	3.9304	92.05%	80.09%	4.4255	78.51%	46.59%
Spread	0.5867	2.24%	8.64%	2.7218	22.50%	45.46%

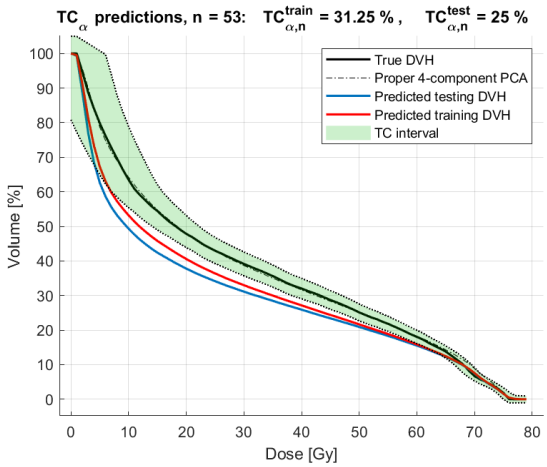
Table 5.8: TC_α score optimization 8-fold CV training and testing errors



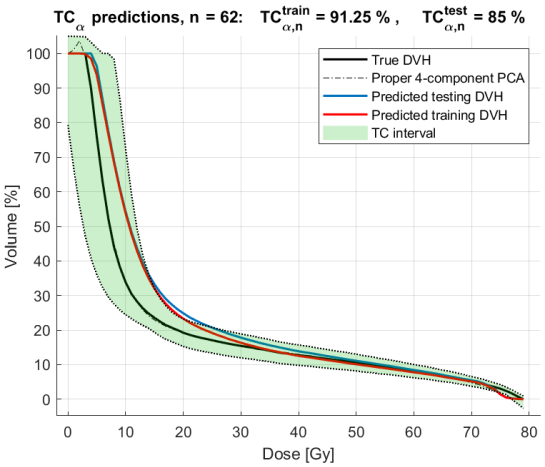
(a) A good prediction



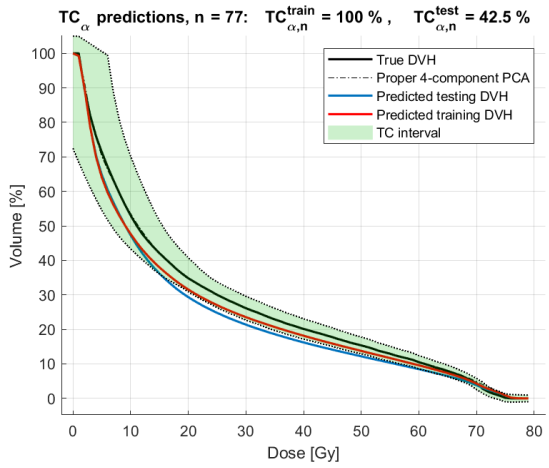
(b) An average prediction



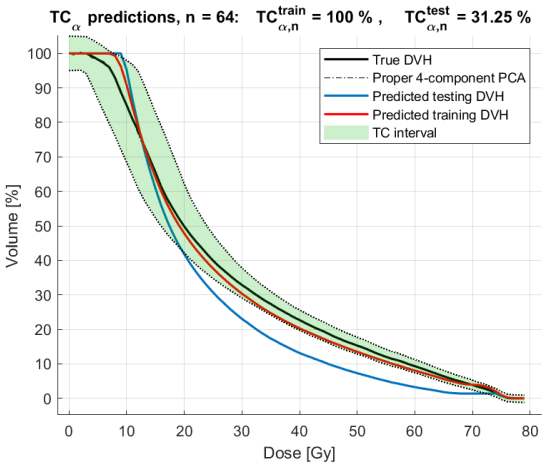
(c) A bad prediction



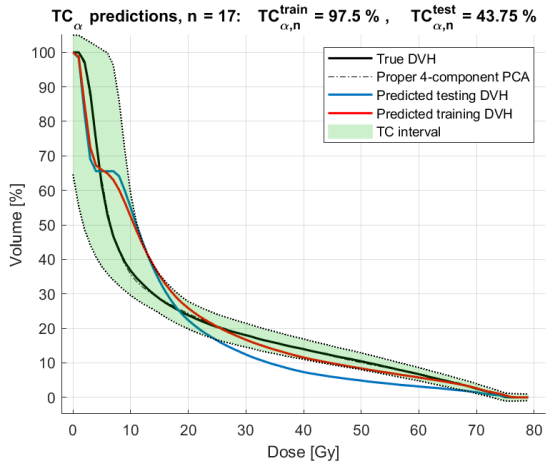
(d) The test case fails



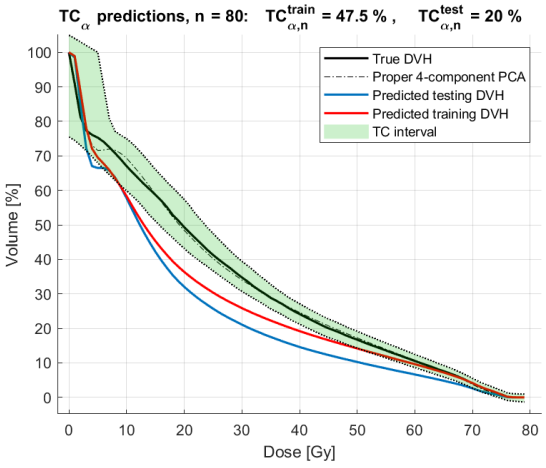
(e) The train DVH barely passes for all points



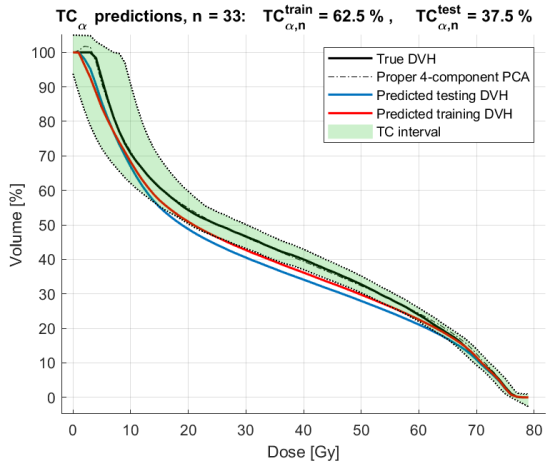
(f) Similar to 5.7e, but the test case fails



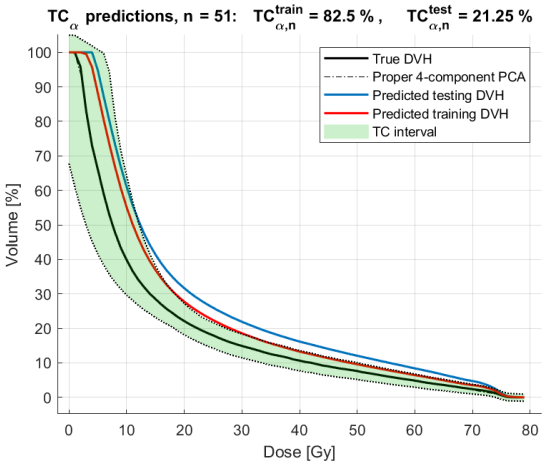
(g) The test case fails



(h) Both the regression and the PCA are insufficient



(i) The train case barely misses the TC_β criterion



(j) Both predictions miss the criterion

Figure 5.7: Some typical DVH-predictions acquired with the TC_α-optimized model

In general, the trend that we see from the TC_α-optimizer is that it pushes as many DVH points within the TC boundary, at the expense of overall prediction accuracy. In training

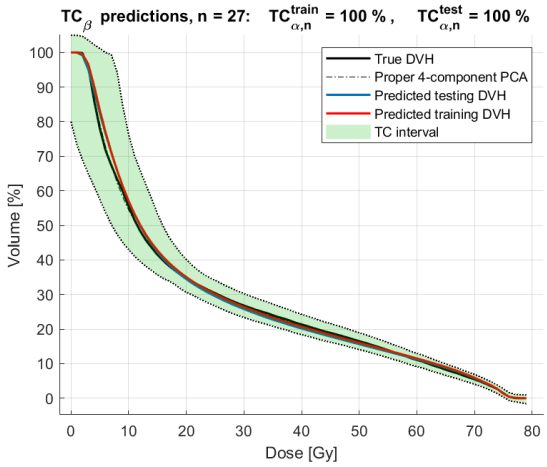
data, we see this confirmed in Table 5.8. We see that TC_α has improved compared to all other models, showing a relatively small CV spread (2.24%). The increase in TC_α brings about an increase in TC_β as well (80.1%). On the other hand, we see that the RMS deteriorates compared to all previous models. The same observations are not seen in testing data, where typically all evaluation metrics are seen to have deteriorated.

Maximizing TC_β

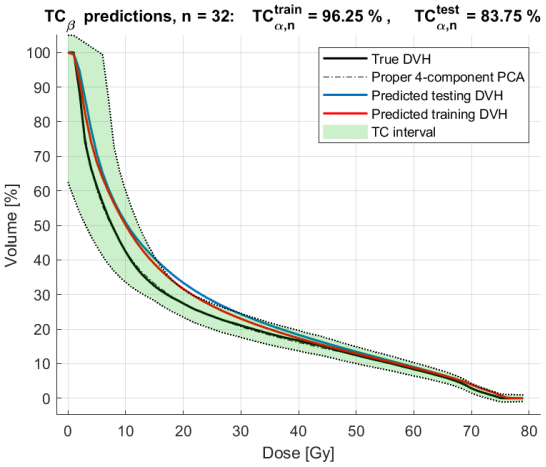
The CV average and spread training and testing errors resulting from the TC_β-optimization method are summarized in Table 5.9. Figure 5.8 shows some representative DVHs predicted with this model.

	Training errors			Testing errors		
	RMS	TC_α	TC_β	RMS	TC_α	TC_β
Average	4.4607	88.73%	84.26%	4.7371	79.32%	53.41%
Spread	1.0915	5.32%	6.17%	1.5435	9.32%	27.27%

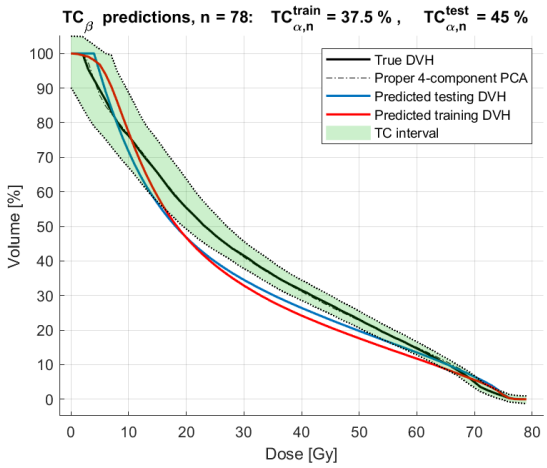
Table 5.9: TC_β score optimization 8-fold CV training and testing errors



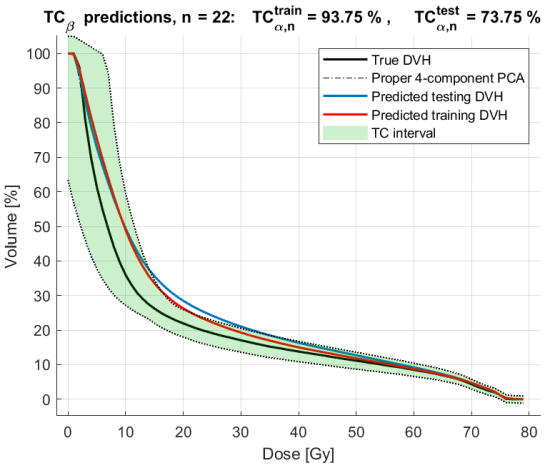
(a) A good prediction



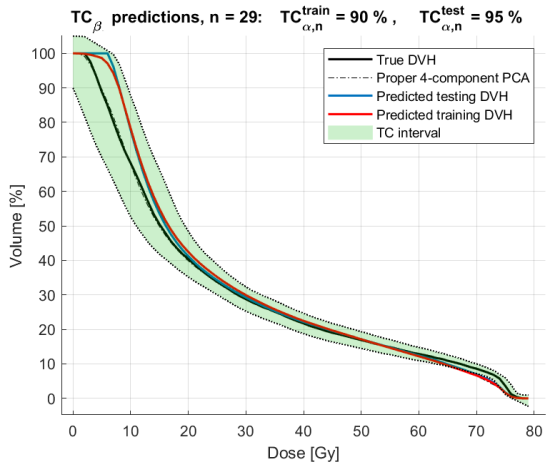
(b) An average prediction



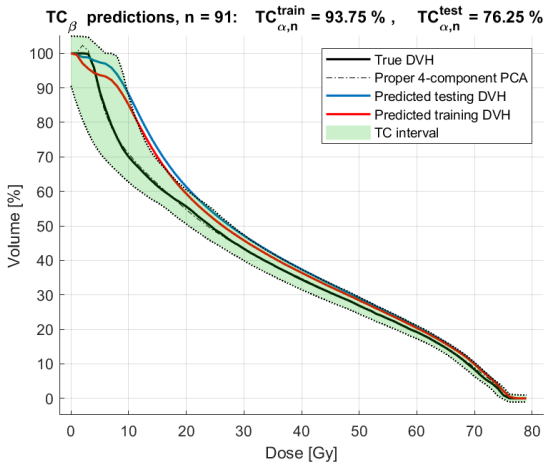
(c) A bad prediction



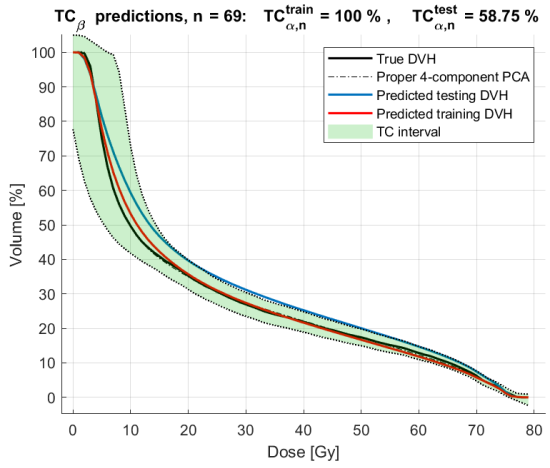
(d) The test case fails



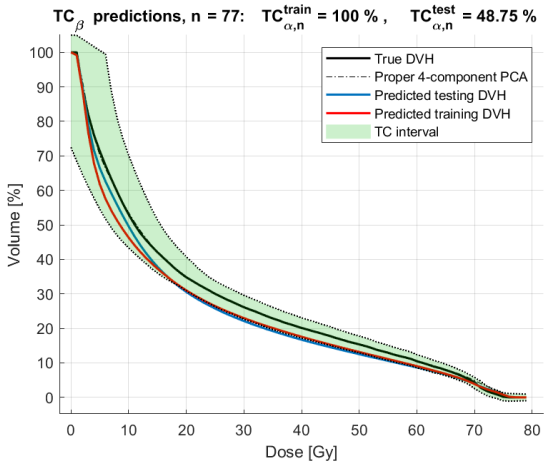
(e) The training DVH barely passes the TC_β criterion



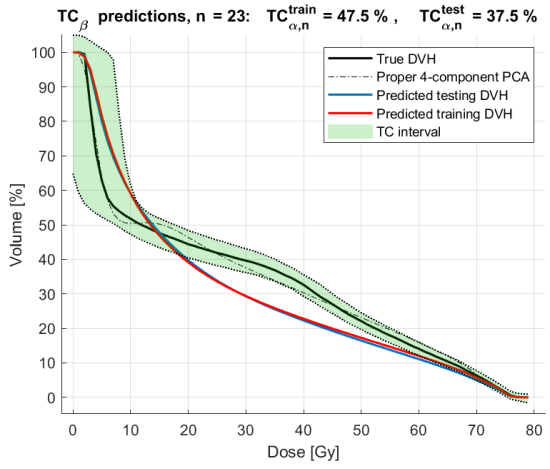
(f) The testing DVH barely fails the TC_β criterion



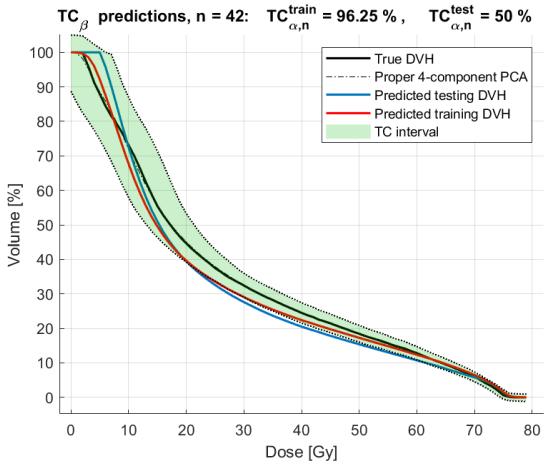
(g) The testing DVH barely fails the TC_β criterion



(h) The testing DVH barely fails the TC_β criterion



(i) Both the regression and the PCA are insufficient



(j) The testing DVH barely fails the TC_β criterion

Figure 5.8: Some typical DVH-predictions acquired with the TC_β-optimized model

The training predictions in Figure 5.8 are often seen to run close to the inner side of the TC boundary border. In addition, compared to the TC_α-model, we observe more predictions with $\geq 90\%$ of their dose bins inside the TCB, passing the TC_β condition. Examples of

this are shown in Figures 5.8b, 5.8d, 5.8f, 5.8h and 5.8j. This is also confirmed by Table 5.9, where the TC_β value scores the highest of all of our models, even with a smaller CV spread than in the TC_α-model. Even in atypical situations such as shown by Figure 5.8i (where even the (4-component) PCA reconstruction of the DVH does not suffice), it appears that the optimizer tries to fit the prediction in the TCB. Another thing that stands out is that training and testing predictions typically do not show much variation. Finally, similar to the TC_α model, the optimizer does this at the cost of overall accuracy (and TC_α in this case), and this model also does not score very well in terms of testing errors.

5.2.4 Constrained DVH-optimization predictions

	Strictly constrained		Flexibly constrained	
	Training	Testing	Training	Testing
Max DVH	100.0000	101.8634	101.2585	102.4015
Min DVH	-0.0026	-3.2598	-0.04980	-0.03790
Max derivative	0.0029	1.7271	1.0000	2.3317

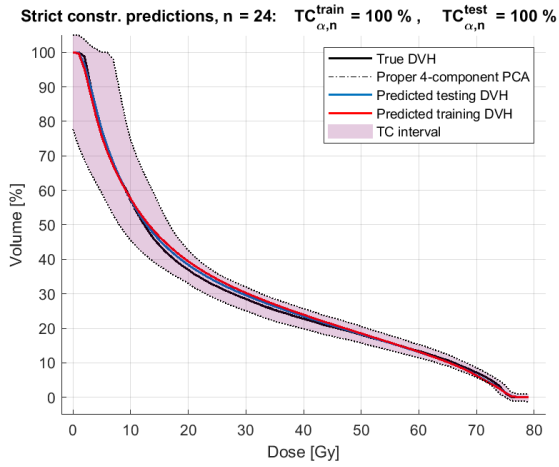
Table 5.10: An overview of the requirements to DVH physicality in the resulting training and testing DVHs for the strict and flexible models.

Two sub-models that were investigated, were constrained optimization with strict constraints and with more flexible constraints. In the strictly constrained case, no violations of the requirements of realistic DVHs as described in 4.5.5 were allowed. In the flexibly constrained case, we allowed DVH values up between 102 and -0.5, and a DVH derivative of at maximum 1. The CV average and spread training and testing errors resulting from constrained direct DVH-optimization are summarized in Tables 5.11 and 5.12. Figures 5.9 and 5.10 show some representative DVHs predicted with strict and flexible models respectively. Lastly, it was checked if predictions complied with the set constraints. An assessment of the requirements to DVH physicality is shown in Table 5.10. Unexpectedly, there were cases where the strictly constrained training DVHs still showed very minor violations of the physicality requirements. It was uncertain why this happened, but since these violations were of such a small magnitute ($\leq$ 0.01 volume %), this was deemed acceptable even for the strict optimization.

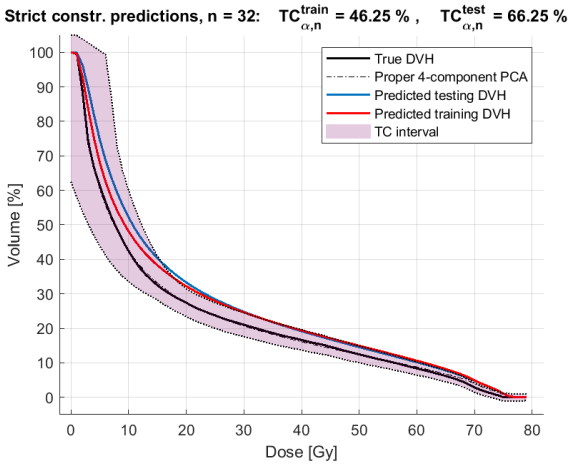
Strict constraints

	Training errors			Testing errors		
	RMS	TC_α	TC_β	RMS	TC_α	TC_β
Average	4.7998	64.80%	31.64%	4.9166	61.95%	28.41%
Spread	1.1830	11.76%	14.81%	1.5407	25.46%	27.27%

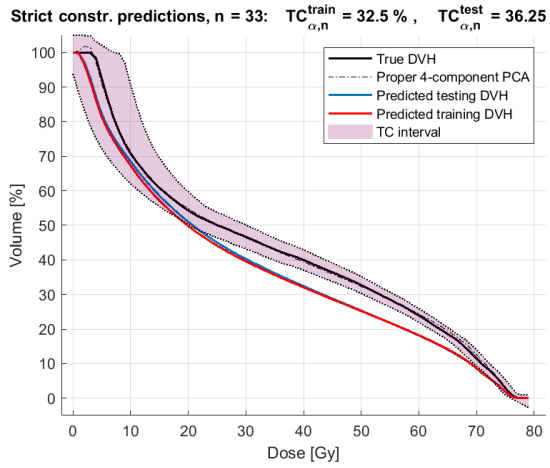
Table 5.11: Strict hard-constrained optimization directly for the DVH dose bins, 8-fold CV training and testing errors.



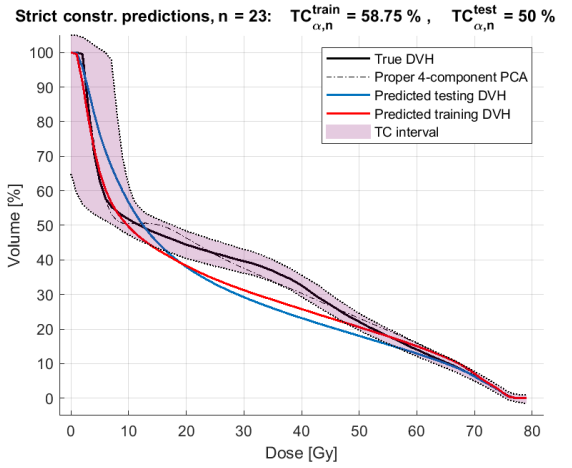
(a) A good prediction



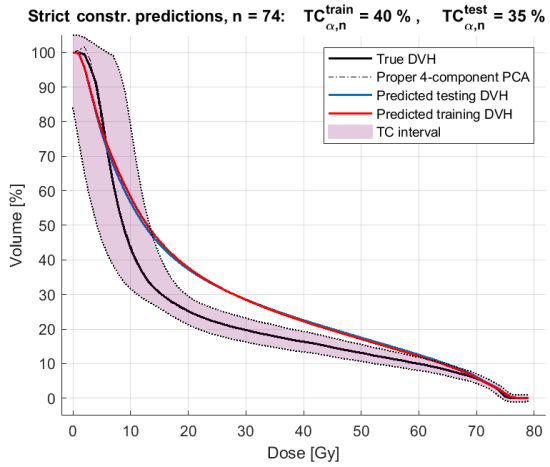
(b) An average prediction



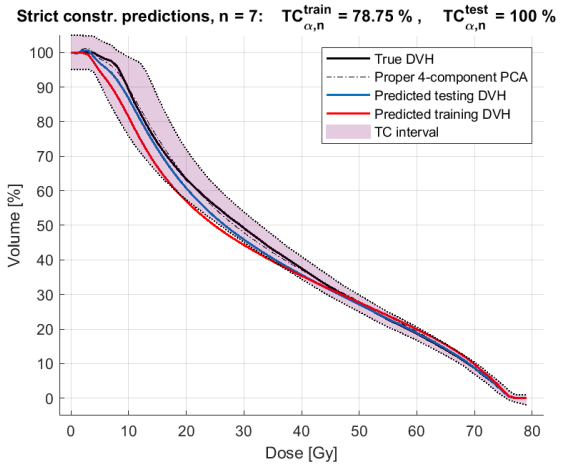
(c) A bad prediction



(d) The PCA-reconstruction falls short



(e) Another average prediction



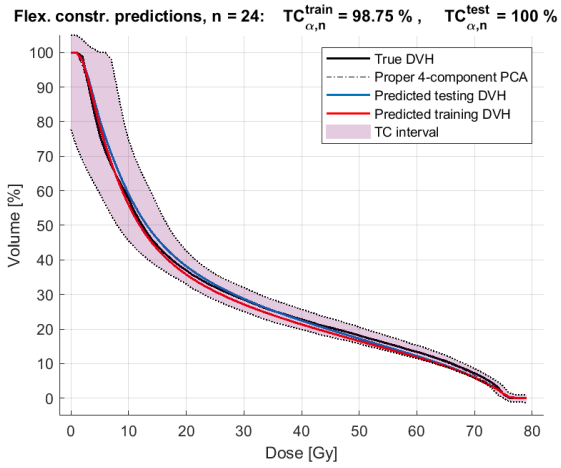
(f) Relatively good, realistic predictions

Figure 5.9: Overview of some typical DVH-predictions acquired with a strictly hard-constrained DVH-optimized model.

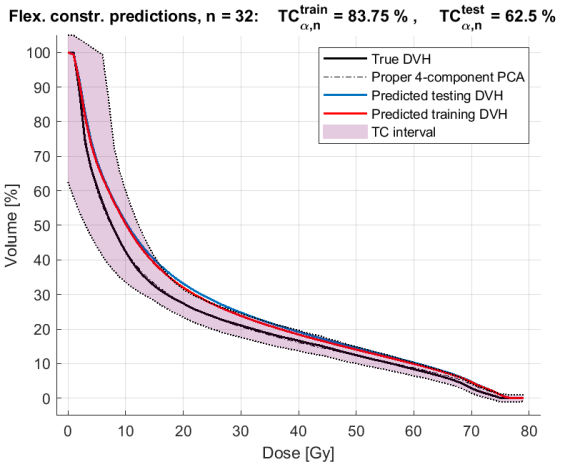
Flexible constraints

	Training errors			Testing errors		
	RMS	TC_α	TC_β	RMS	TC_α	TC_β
Average	3.5161	79.87%	55.09%	3.9947	74.13%	39.77%
Spread	1.9274	22.49%	34.56%	2.4553	20.11%	45.46%

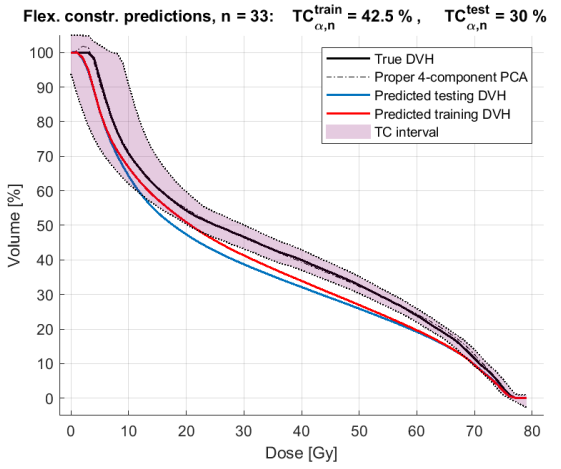
Table 5.12: Flexible hard-constrained optimization directly for the DVH dose bins, 8-fold CV training and testing errors.



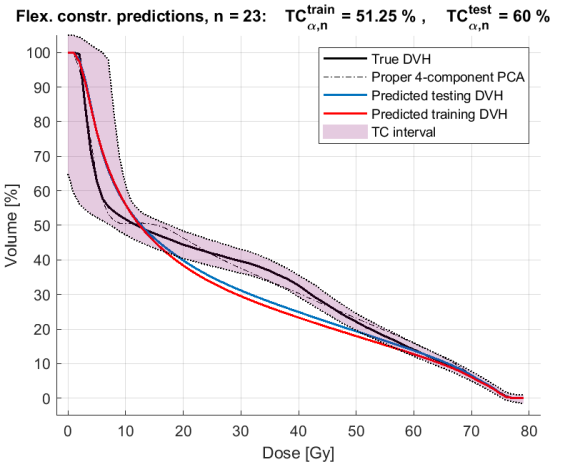
(a) A good prediction



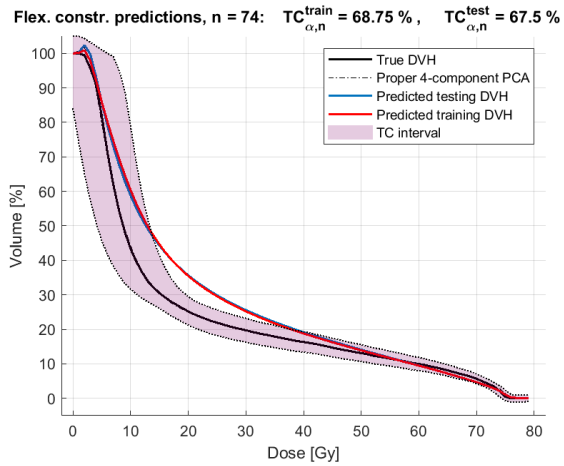
(b) An average prediction



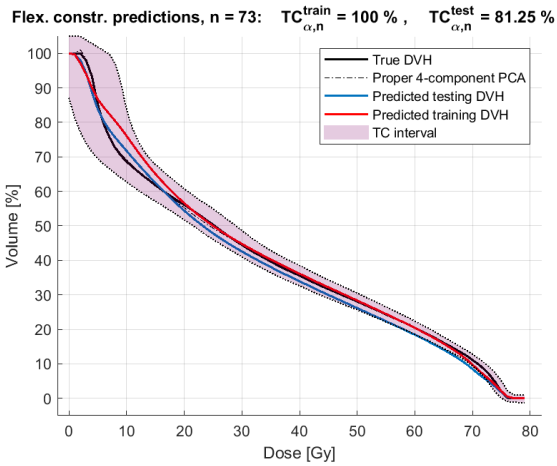
(c) A bad prediction



(d) The PCA-reconstruction falls short



(e) Another average prediction



(f) A correct training prediction

Figure 5.10: Overview of some typical DVH-predictions acquired with a flexibly hard-constrained DVH-optimized model.

One of the first things that comes forward from Table 5.10 is that testing DVHs resulting from constrained models do not have to be purely physical, or even have imposed the same constraints. From Table 5.11 we see that using strict constraints largely deteriorates DVH prediction, at least in the direct DVH-optimization model, as we see a decrease in all evaluation metrics. In addition, the DVH graphs shown for all 6 examples in Figure 5.9 show similar (but obviously not equal) training and testing DVHs, even when the true DVH clearly differs. When we look at the flexible constraint results in Table 5.12, we see that these effects are still present (accuracies are still substantially worse than for the unconstrained DVH predictor), but are already much less expressive. In the DVH prediction graphs in Figure 5.10, we now see increased DVH variability, as well as improved predictions when compared to the strictly constrained case.

CHAPTER 6

DISCUSSION

Through a variety of computational approaches, we have been able to make DVH prediction models, each showing different characteristic DVH predictions, and have shown how these models perform on the used clinical data set. By constructing a clinical criterion, we were able to evaluate these models in terms of metrics that were believed better serve clinical interests, complementary to global prediction accuracy metrics. The two-point predictor method has been assumed as the current clinical practice, and with this study we aim to assess how to further improve this current clinical practice. An additional aim of our study was to construct a model with a TC_β score of $\geq 90\%$. We will now discuss how the results coming from each model should be interpreted, cover some of their strengths and drawbacks, and address the validity of the results of our models.

6.1 Results interpretation

6.1.1 Clinical practice

The two-point predictor is the simplest of our methods. Its implementation stems from a clinical wish for methods to be no more complicated than they need to be. But this simplicity comes at a price; since there are two PC modes and geometrical descriptors involved, its predictive power is limited. Nevertheless, judging from Table 5.4, the two-point predictor generally shows a decent performance, where approximately 80% of dose bins are correctly predicted. One of the assets is its ability to generalize, as training and testing error differences are minimal. However, there is still much room for improvement in terms of our evaluation metrics. Something that stands out about the individual prediction grahps (Figure 5.3), is the production of unphysical DVHs by this model. They show some very clear violations of the three requirements to DVH physicality. This is an effect that is inherent to the model, for a large part because this model utilizes (and is limited to) 2 principal components. Because these violations propagate multiplicatively into the CI, these effects are expressed even more in the confidence intervals. It should be noted that the way this CI is determined, the CI on display is not a true CI because the true correlations between V_{95} and V_{mean} have not been properly investigated. In addition, as can be seen from Table 5.2, 2 PCs simply do not explain enough of the DVH data's variance to perform well.

6.1.2 EV-optimizer

By optimization on the PCA eigenvalues, we can bring about an increase in overall accuracy in the resulting DVH predictions, and thereby increase the correctly predicted points, and the number of patients that pass our TC_β criterion. Compared to the current clinical practice, this method shows a clear improvement in all of our evaluations. Although this model already loses some of its ability to generalize, training and testing errors in 5.5 suggest that this model still reatins a good degree of generalizability. Another peculiar result is the large spread that is found in the testing errors. As this is something we see in all our cross-validations, this will be discussed in a later section (Section 6.2.5). The DVH-prediction graphs resulting from this exhibit lesser degrees of DVH unphysicality than those from the clinical practice, Figures 5.3a and 5.4d. There are also cases where the model is insufficient (5.4c) and cases where the PCA is insufficient (5.4f). In the former case, the regression model is not able to sufficiently predict the eigenvalues, whereas in the latter case is a rather unique case that contains DVH variantion that is not captured by the 4 PC modes.

6.1.3 DVH-optimizer

By optimization directly for precision (in terms of quadratic differences) at the level of the DVH, we have observed this model to allow increased prediction accuracy. Though only a minor improvement in both training and testing errors, different cross-validations yielded consistently better average CV-errors. The DVH-prediction graphs (Figure 5.5) confirm that we see subtle improvements with respect to the EV-optimizer. Finally, the TC_α and TC_β training spreads of 1.18% and 1.23% suggest that this model is more robust variations due to the random patient sampling of cross-validation.

6.1.4 Penalized DVH-optimizer

By imposing penalty terms on erroneously predicted points, we have not been able to make a model that improves traning or testing errors. Although Figure 5.6 shows a minor testing improvement for the stationary-weighted model (63.64% vs. 61.36%), the deterioration of training errors suggest that this is a coincidental cross-validation result. None of the other sub-models yielded better results. However, during earlier stages of this work, we have seen clear improvements with these models when more features were included, as well as logarithmic features.

6.1.5 Halfway-boundary DVH-optimizer

Similar to the regular penalized DVH optimizer, from the halfway boundary-assisted model (Figure 5.7) we see a minor increase in testing error for the weightless sub-model. However, due to the deterioration of training errors, as well as the increase of training CV spreads, this model was not directly accepted as superior. Namely, it is uncertain whether this is an effect that is inherent to the model, or if it is due to the randomness that in CV sampling. Regardless, three examples of improved generalizability are shown in Figures 5.5d - 5.5f and 5.6d - 5.6f. Although the difference is subtle, the DVH shown in Figure 5.6f shows a testing results that are pushed within the TCB.

6.1.6 TC_α-optimizer

By constructing a model that maximizes the amount of correctly predicted points for the entire patient population in the data set, we have been able to clearly improve TC_α and TC_β scores for training data. With $TC_\alpha = 92.1\%$ and $TC_\beta = 80.1\%$, as we can see from Table 5.8, that the model is effective at reaching this goal. This is confirmed by the reported CV training spreads, which are of the same order of magnitude most of the other models penalized models, as well as the EV-optimizer. The TC-metric improvements come at the cost of overall prediction accuracy, as is confirmed by the deteriorated RMS. Based purely on the data in the table, this model does not generalize well, as testing errors show substantially decreased RMS, TC_α and TC_β averages. However, representing the accuracy of a model purely judging from these metrics seems inappropriate. The reason why becomes clear when examining the DVH prediction graphs (Figure 5.7). A general trend that arises is that predictions end up close to the TC boundary, such that it can fit the other DVHs within their respective boundary as good as possible. The problem of the TC-metrics is that when DVHs are predicted near or on the boundary on such a large scale, small anatomical deviations from training cases can cause the resulting testing DVH to be rejected. This makes TC-metrics very sensitive to anatomical changes with respect to the "typical" anatomy. As for the DVH prediction graphs (Figure 5.7a), we observe that the model is still able to predict typical cases well (5.7a). A potential pitfall of this model may be for the model to completely disregard a "difficult" patient, in order to increase the number of correctly predicted points for the majority of the patients. However, we have not observed this, and DVHs that models were unable to predict well end up to be not much worse (Figures 5.5c, 5.6c and 5.7c). We can see that the model predicts the course of the DVH just within the TCB, as exemplified by Figures 5.7c - 5.7e. However, the DVH shown in Figure 5.7e is a rather peculiar example, as it shows a trend in the 0 - 10 Gy dose region that is not representative of a typical DVH, compared to the "typical" results shown in Figures 5.5e and 5.6e. Another patient that appeared difficult for the

TC_α model (and all other models) to predict predict is shown in Figure 5.7h. However, this result has improved in terms of TC_{pnts}. The prediction shown in Figures 5.7i and 5.7j end up fairly well, but fail for the test case. These two predictions demonstrate the motivation to use a TC_β optimizer. Seeing TC_α result 62.5% and 82.5% in the respective figures, they do not reach the TC_β criterion to be counted a pass. What's more, since the entirety of the DVH training predictions are very close to the TC-boundary, improving their result in terms of TC_β should require only minimal cost from its cost function. The ability of this model to generalize will be discussed in the next section, as the TC_β model exhibits the same behaviour as TC_α.

6.1.7 TC_β-optimizer

By incorporating an approximation of the TC_β criterion in out TC_α objective function, we have been able to construct a model that maximizes the number of successfully predicted patients. Judging from Table **??**, this optimizer manages to further increase the average TC_β score (84.3%). Expectedly, this comes at the expense of TC_α (and overall accuracy). As for the DVH prediction graphs (Figure 5.8), we observe that there are fewer examples of patients that have a TC_α score just below 90%. An example of this is shown in Figure 5.8j. Similar to the TC_α model, the TC-metric improvements of this model are carried over to testing predictions. It is believed in the case for TC-based optimizers that the proposed TC-metrics allow for a misrepresentation of the testing errors, demonstrating the sensitivity of these TC-metrics when predictions are close to the TC boundary. This is reinforced by the fact that the resulting training and testing RMS are very similar, suggesting that this model does not show a good degree of generalizability.

6.1.8 Constrained DVH-optimizer

By imposing strict constraints on the requirements to DVH physicality, we have forced realistic DVH predictions from the direct DVH-optimization model. Tables 5.11 and 5.12 suggest that these constraints greatly deteriorate the optimizers ability to make accurate predictions, and the more flexible we choose our constraints, the better predictions get. In addition, the strictly constrained model (Figure 5.9) consistently predicts DVHs that are very much alike. The more flexible we set our constraints, the more variation is observed in the predicted DVHs (Figure 5.10). However, the desired result of this optimizer is to produce good DVH that are also realistic. An example that shows the optimizer is able to do this is from patient n = 7 in Figures 5.5f and 5.6f. These unconstrained models show good predictions by the DVH and HWB-optimizers, that have been forced physical as can be seen by the flattened area in the $d < 10$ Gy dose region. Figure 5.9e shows a decent and entirely realistic, prediction by the hard-constrained optimizer, whereas the flexibly constrained optimizer further improves this result (5.10e).

6.2 Assumptions

All results were acquired under certain assumptions. Some of the most important assumptions concern the number of PC-modes, the composition of our feature vector $\boldsymbol{\zeta}$, the way our TC-boundary was determined, the way we chose to validate our results and lastly our data selection. Let us now discuss each of these.

6.2.1 PCA

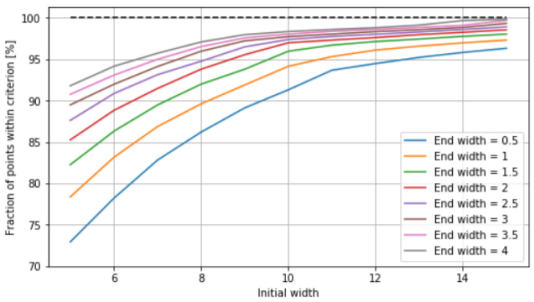
Throughout all of our analyses, we chose to use 4 components for model training and DVH reconstruction. This was based on the the explained variance ratios of each PC mode, and on TC-score performances of the proper PCA-reconstructions. It is likely that using 4 components is a good trade-off to include enough variance vs. endangering the model of over-fitting, but the use of a smaller or larger set of components has not been properly investigated.

6.2.2 Features

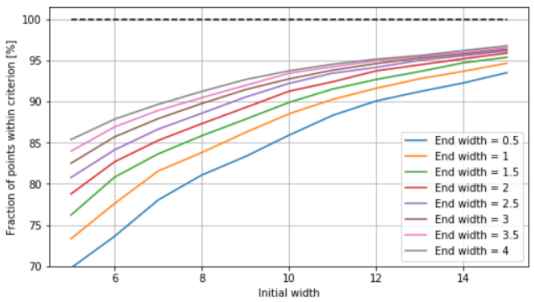
The feature selection method proposed was of a heuristic nature, where we included only the polynomial features that were most descriptive in terms of DVH PCA eigenvalues R^2. It is believed that this method should provide a set of features that should work generally well for all models, however it would make more sense to have our feature selection method custom fit the objective function of each optimization model. In addition, our feature selection method has not covered all of the feature combinations exhaustively for computational efficiency reasons. Finally, we have chosen to use the 10 best features as suggested by our selection method, to minimize over-fitting and to optimize our results in terms of testing errors. An overview of the least squares error, TC_α and TC_β with respect to different CV fold sizes and features included can be seen in figures B.2, B.3 and B.4. Based on these figures, it can be argued that the number over-fitting starts to occur from the inclusion of 10% of the maximum features and onward for least squares error, and even from 6-8% for TC_α and TC_β.

6.2.3 TC-sensitivity

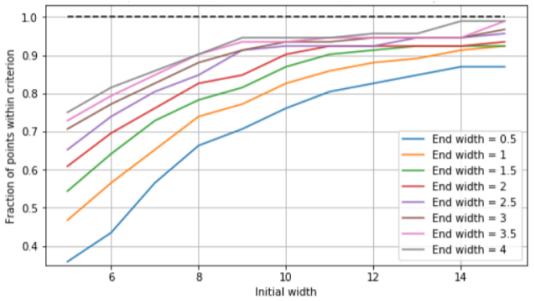
As observed, TC_α and TC_β scores exhibit a large range in cross-validation errors, especially in testing cases. Naturally, the way the TC-boundary was determined influences this. We performed analyses to make an assessment on the sensitivity of these criteria to its initial and end width, Δ_I and Δ_F respectively. The rationale behind this was to confirm if the TC-based metrics could fairly reflect the accuracy of a model, by investigating how these metrics differ for a varying TC boundary initial and end width. To illustrate this, let us imagine a model with a considerable amount of DVH-predictions that run barely out of the TCB. Based on its TC scores, it is considered a bad model, although the global predictions may be considered as quite good, especially if the chosen TCB width is narrow. The TC_α and TC_β scores with respect to different Δ_I and Δ_F are displayed for two typical (linear and polynomial regression) EV-optimization models in Figure 6.1.



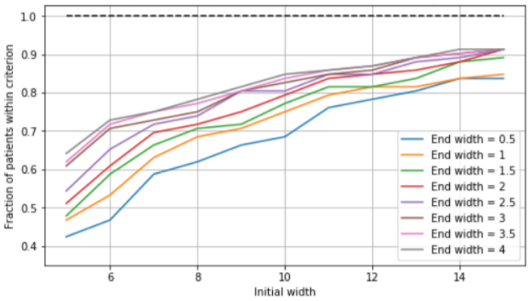
(a) Polynomial regression model TC_α sensitivity



(b) Linear regression model TC_α sensitivity



(c) Polynomial regression TC_β sensitivity



(d) Linear regression model TC_β sensitivity

Figure 6.1: An overview of the TC_α and TC_β scores for training DVHs predicted training data from two typical EV-regression models done with a manual selection of features.

Judging from Figure 6.1, we can see that TC_α and TC_β scores increase at the same rate when increasing the end-width vs. when increasing the initial width for both the polynomial model and linear model. This means there is no clear indication that prediction accuracies would improve mostly by changing either one of the parameters Δ_I, Δ_F. However, due to the steep increase in TC_α and TC_β accuracies in both models in the lower-width regions, one could argue that the proposed $\Delta_I = 5$ and $\Delta_F = 1$ are slightly harsh. Although the TC boundary dimensions were chosen as a mere initial assumption (i.e. it was not exactly known how strict or exactly how strict or mild the TC boundary width would be), they were chosen to reflect the clinical desire of KBP models.

6.2.4 Cross-validation

Our validation methods were mostly based on 8-fold CV, in which we use 87.5% of the available data for model training and 12.5% for testing. It is good practice to reserve a sample set outside the set that is used for CV, in order to gain an unbiased idea of how models are expected to perform on completely new data. This would be valuable in order to provide an unambiguous advice to the clinic on how to improve the clinical practice. However, because the data set was on the small side, we decided to include the complete data set for training and testing. In addition, the large spread exhibited by testing data may be an indication that more training data is required to provide reliable predictions. For that reason, it may be ideal to cross-validate the results in a leave-one-out (LOOCV)-manner.

6.2.5 Data selection

Cross-validations exhibit a large spread for testing data. This suggests that cross-validations depend highly on the random sampling method used to choose which patients are in the training and testing set. One of the assumptions we have made when selecting the data to use for model development was that patient groups 2 and 3 were sufficiently similar in order to unify them. This was done to maximize the amount of data available for model training. In hindsight, this may explain some of the spread seen in testing data. Ideally, models should be re-evaluated with patient groups separated.

CHAPTER 7

CONCLUSION

Radiotherapy treatment planning remains a complex, time-consuming and often manual process. Various tools have emerged in the market that automate the treatment planning process and thereby increase plan standardization. However, automated planning approaches are not perfect, and still require substantial interaction from the RT treatment planner. Additionally, it is difficult to judge whether a plan with a more ideal trade-off between tumour cure and healthy tissue toxicity exists, since the RT technician has not worked an automated plan. To resolve these issues, knowledge-based planning could be used. As a part of this thesis study, we have constructed an initial, simplistic KBP tool that serves as the current clinical practice. We continued to investigate several KBP modelling approaches, in order to improve this clinical practice. Finally, we aimed to build a model that could correctly predict $\geq 90\%$ of the dose-volume histogram for $\geq 90\%$ of patients.

7.1 Conclusion

7.1.1 Summary

In compliance with knowledge-based planning literature, we have shown that the overlap-volume histogram (OVH) can be used for KBP modelling to predict VMAT rectum dose-volume histograms in prostate cancer. Reduced order modelling of the OVH and DVH can be effective in simplifying the data before being used for KBP modelling. However, the two-point predictor has shown that OVH-metrics can be used to accurately predict specific DVH points. Regardless, our investigated models included the OVH PCA. Compared to the current clinical practice, the greatest improvement in overall DVH prediction accuracy was by means of direct DVH dose bin squares minimization. This model yielded a good degree of generalization. We then further improved predictions by using the clinically-driven tolerance criterion (TC), and were able to further improve the amount of correctly predicted DVH points ($TC_\alpha = 92.1\%$), as well as the amount of patients that have $\geq 90\%$ of their DVH correctly predicted ($TC_\beta = 84.3\%$). This means that we have not reached our initial goal of building a model that yielded a TC_β of $\geq 90\%$ in training data.

7.1.2 Study limitations

There are some limitations to this study that should be acknowledged. First, it is important to note that the criteria on which data selection was based involved only patient groupification and the prevalence of a femoral prosthesis or bowel loop. There is a possibility that improved guidelines on data selection may result in more consistent plans, resulting in better models. In addition, it should be remembered that the data involved only patients treated at the NKI. On the one hand this increases plan consistency, but on the other hand limits the available data, and may introduce some bias.

7.1.3 Future directions

Based on our findings, there are several proper recommendations we can make for future research. First, although, based on the population-wide accuracy evaluations, one may conclude that the TC_α and TC_β models are unable to generalize for unseen data, it is believed that additional research is required to investigate whether metrics that include a tolerance criterion are useful for DVH prediction. If the DVH prediction uncertainty could be better understood based on the anatomy, we could adapt TC-based models and make robust optimization models such that we can account for these prediction uncertainties. This would require a quantitative, statistical analysis of the DVH prediction uncertainties for unseen patients, with respect to their anatomies. Second, investigating how these models perform when analysed separately for patient groups 2 and 3 may allow models to be more consistently trained and better generalize for unseen data. To reliably train such models for the separate groups, additional data may be required. Third, it is believed that there is a use for non-parametric regression models. This has the advantage that fitting parameters have no real relation to the independent variables at hand. This way, the predictor is constructed according to information derived from the data. Previous KBP literature has shown the support vector regressor to be suitable for DVH prediction in head-and-neck and prostate cancer [Yua+12]. Additionally, the radial basis regressor and random forest regressor may be well-suited. To do non-parametric regression reliably, additional data may be required.

7.1.4 Proton therapy

Under the veil of personal interest, I will very briefly cover the potential of knowledge-based planning approaches for proton therapy. As discussed in Section 1.1.1, protons are used as an alternative external beam radiotherapy modality to photons. Because protons have the advantage that they completely come to rest inside the tissue after the Bragg peak, they potentially allow for a better dose distribution due to a more localized dose delivery. However, since protons are more sensitive to changes in the anatomy of the patient, robust alrogithms that deal with anatomical undertainties are needed. One way to do this is to use computational optimization algorithms that find treatment planning settings such that all anatomical uncertainties are accounted for. This way, for the majority of patients, a good plan can be found, whereas for the remaining patient additional measures are required. Although it is not yet much used throughout in state-of-the-art proton therapy clinics, knowledge-based planning could potentially provide an additional quality assurance tool also for proton therapy. As an extension of the photon KBP model proposed by Appenzoller et al. [App+12], one study has successfully incorporated an KBP tool that can predict bladder and rectum DVHs for proton therapy of prostate cancer [YMS17]. However, KBP for proton therapy is not commonly used, for which some reasons are enumerated. First, KBP modelling relies heavily on high-quality data (the model can only be as good as the data allows). With proton therapy being more scarcely used than photon therapy, having this data at ones disposal can prove problematic. Second, optimization methods may Second, dose distributions in proton therapy strongly depend on beam angles and intensities, unlike in IMRT [YMS17]. Therefore, in order to use KBP for proton therapy, first it needs to be understood how the physics of proton therapy translates into machine learning modelling.

APPENDIX A

BACKGROUND INFORMATION

A.1 Dose calculation

For accurate dose calculation, the interaction between radiation and tissue needs to be known. For photons, this is given by the electron densities, which are given by CT Hounsfield Units (HUs). The relation of HUs to relevant interaction properties of radiation with tissue have been published and are the basis of dose calculation [Cho+84] [PHC79]. With the known interactions, treatment planning becomes a matter of finding proper treatment parameters, such as angles, intensities and machine parameters.

A.2 Uncertainty handling in radiotherapy

In radiotherapy there are many different sources of error during treatment preparation and execution that limit the accuracy of dose delivery. As a consequence, several safety margins are required to ensure actually delivering the planned dose. This appendix gives a short description of only of the most important errors in radiotherapy and the margins that account for them. The major error sources regarded here are tumour delineation uncertainties, organ positional variation within the patient and patient setup variations, and errors are regarded as any deviation between planned and executed treatment.

Systematic and random errors

Error sources have systematic and random components. In fact, each of which has a different effect on the dose distribution. Random errors blur the dose distribution, which can in practice be described as a convolution of the dose distribution with the probability distribution function of the random error [Her04] [LLB99]. Systematic errors cause a shift of the cumulative dose distribution relative to the target [Her04]. A description of these errors can be used to create a margin that protects for under-dose of the tumour volume.

Target volume definition

Typically, the process of radiotherapy commences by delineation of relevant OAR structures and primary tumour volume, or gross tumour volume (GTV), in a planning computed tomography (CT) scan. The clinical target volume (CTV) incorporates the GTV, and microscopic disease extensions not shown by clinical examination and the CT. Definition of the CTV is based on radiation oncologist experience, as well as local recurrence patterns and histological examination of post mortem specimen assessment [Bar+09]. In order to ensure adequate dose to the CTV, the planning target volume (PTV)) is used in radiotherapy.

Setup accuracy

Setup errors arise from inaccuracies in the positioning of the patient with respect to the treatment field. In particular, motion of the skin with respect to internal anatomy limits reproducibility of the patient setup, introducing a systematic error [Her+00]. However, gross setup errors are typically prevented by anatomy matching software [Mur+08]. Depending on the RT indication, immobilization may also aid in minimizing setup errors and improving reproducibility of treatment. Immobilization is particularly helpful in head-and-neck treatment, because organ position variation relative to the rigid skull is relatively small. Studies have indicated that with present-day immobilization methods and well-designed setup protocols, a setup error standard deviation (SD) of 2 mm or better for each axis is achievable for prostate irradiation [Bel+96].

Organ motion

A third major source of error is organ motion. It includes periodic movement such as breathing or pulsation and non-periodic movement such as the filling and emptying of the bladder and bowels. Organ motion can cause both systematic and random errors. In prostate irradiation, organ motion errors were found to be more significant than setup errors, with a motion SD of 5.8 mm in anterior-posterior and 3.3 mm in superior-inferior direction [Ala+01].

Incorporation of uncertainties into treatment planning margins

Population-based studies have showed the relation between the PTV)-CTV margin and PTV) coverage. From these, a margin recipe for the CTV was derived, such that 90% of patients in the population are guaranteed to receive a minimum cumulative CTV dose of at least 95% of the prescribed dose [Her04] [Her+00]. This CTV margin is approximately $2.5(SD_{systematic}) + 0.7(SD_{random})$. This directive is followed by the NKI.

A.3 Biological modelling

Since the goal in radiotherapy is to bring about a biological effect in order to control a tumour, biological models can be useful for optimization purposes in treatment planning. Such models are more complex and more difficult to control because of the sheer amount of uncertainties introduced by the radio-biology.

A.3.1 Radiobiology

The biological target

DNA molecules are the carriers of all genetic information of the human cell (genome). The DNA is known to encode for the information of a large number of proteins, which in turn are responsible for many of the cells processes that are necessary for it to carry out its function and proliferate. The DNA is made up of two strands of nucleotides (the building blocks of the DNA), that are tied into one another to form a double helix. All DNA molecules are contained within chromosomes. These are macromolecules that are known to encode the information for a large number of proteins, which in turn are responsible for many of the cells processes that are necessary for it to carry out its function and proliferate. The fraction of the DNA containing important genetic information is also known as the biological target. To provide a quantitative feeling for the scale of this target, it can be stated that only about 1% of human cell volumes consists of chromosomes, and that about 10% of the chromosomes' content actually concerns relevant genetic information [BDO07]. Radiotherapy aims to disrupt cancer cells by effectuating irreparable damage to this biological target.

Types of damage

The most important forms of radiation-induced DNA-damage are nucleotide damage, DNA cross-links, single strand breaks (SSBs) and double-strand breaks (DSBs) [BDO07]. Whenever DNA damage has occured, it can typically be repaired completely from to the complementary genetic information in the non-damaged DNA strand, leaving no lasting effects. However, damage may also be repaired incompletely. In this case, the effect of damage may or may not become apparent at a later stage of the cell or after reproduction. If sufficient damage is caused, the cell may enter a state of cell cycle arrest (reproductive death), or apoptosis (self-induced, programmed cell-death). These effects may be caused through either the direct effect of radiation, or through the indirect effect.

1. Direct effect
   The minority ($\sim$ 30%) of DNA damage is caused by ionising rays directly hitting the DNA. This is referred to as the direct effect of radiation.

2. Indirect effect
   The majority ($\sim$ 70%) of the damage is caused indirectly by molecules that are radicalized due to the ionising radiation. This is referred to as the indirect effect of radiation. As a consequence, these radicals react with and cause damage to the DNA. Oxygen-containing molecules, such as the hydroperoxy- ($HO_2^{\bullet}$) and hydrogen-peroxide ($H_2O_2^{\bullet}$) radicals, are known to largely contribute to this effect. This is used for modeling, and will be discussed in paragraph A.3.2.

Fractionation

Tumour cells are known to be intrinsically less effective in repairing DNA damage than normal tissue cells. This difference is widely exploited in radiotherapy by fractionation of the total dose over a large number of daily fractions during treatment, allowing normal tissue to regenerate, ultimately destroying the tumour more effectively.

A.3.2 Biological models

NTCP and TCP

Basic models that consider biological responses in RT involve tumour control probability (TCP) and the normal tissue complication probability (NTCP), which are used to model tumour control and healthy tissue complications. TCP and NTCP models are believed to be effective cost functions for MCO to map dose distributions into toxicity or cure outcomes [Cra13].

Linear-quadratic model

In cell studies, survival curves are used to describe the survival fraction of a cell population when exposed to a certain condition. Such curves are typically assessed on a logarithmic scale with respect to dose and can be modeled by a linear-quadratic model of the form:

$$S = e^{-(\alpha D + \beta D^2)} \tag{A.1}$$

where S is the fraction of surviving cells, D is the dose delivered, and α and β are parameters capturing intrinsic sensitivity of the cells to ionizing radiation. The mechanistic interpretation of the model is that cell death is either caused by an SSB, which is characterized by α, or by a DSB, characterized by β [BDO07]. The ratio of α/β denotes the relative importance of the linear and quadratic dose terms, and controls the shape of the curve [Bar+09]. When α/β is large, the linear term dominates and the survival fraction shows a so-called shoulder in the low dose region. Conversely, when α/β is low, the quadratic term dominates and causes a steepening curve.

Radio-biological effectiveness

A metric that is used for quantifying biological damage is the relative- or radio-biological effectiveness (RBE). The RBE is defined as the ratio between a reference dose $D_{ref,b}$ (typically 250 keV photons) necessary to induce a certain biological effect b, and the dose $D_{a,b}$ of radiation type a of interest, required to achieve the same biological effect under the same circumstances:

$$RBE_{a,b} = \frac{D_{ref,b}}{D_{a,b}} \tag{A.2}$$

For the purpose of quantifying the contribution of oxygen radicals to DNA damage, the oxygen enhancement ratio (OER) was introduced. The OER can be used to quantify the enhancement of biological damage due to oxygen in the vicinity of the irradiation field. It is defined as the dose $D_{hypoxia,j}$ required to induce a certain biological effect j in the presence of oxygen, compared to the dose $D_{non-hypoxia,j}$ inducing the same biological effect in the absence of oxygen:

$$OER = \frac{D_{hypoxia,j}}{D_{non-hypoxia,j}} \tag{A.3}$$

This is a principle that can be exploited in tumour sites, because they typically have an altered vascular structure such that these sites receive sufficient nutrition required for consistent tumour growth [BDO07].

APPENDIX B

FIGURES AND TABLES

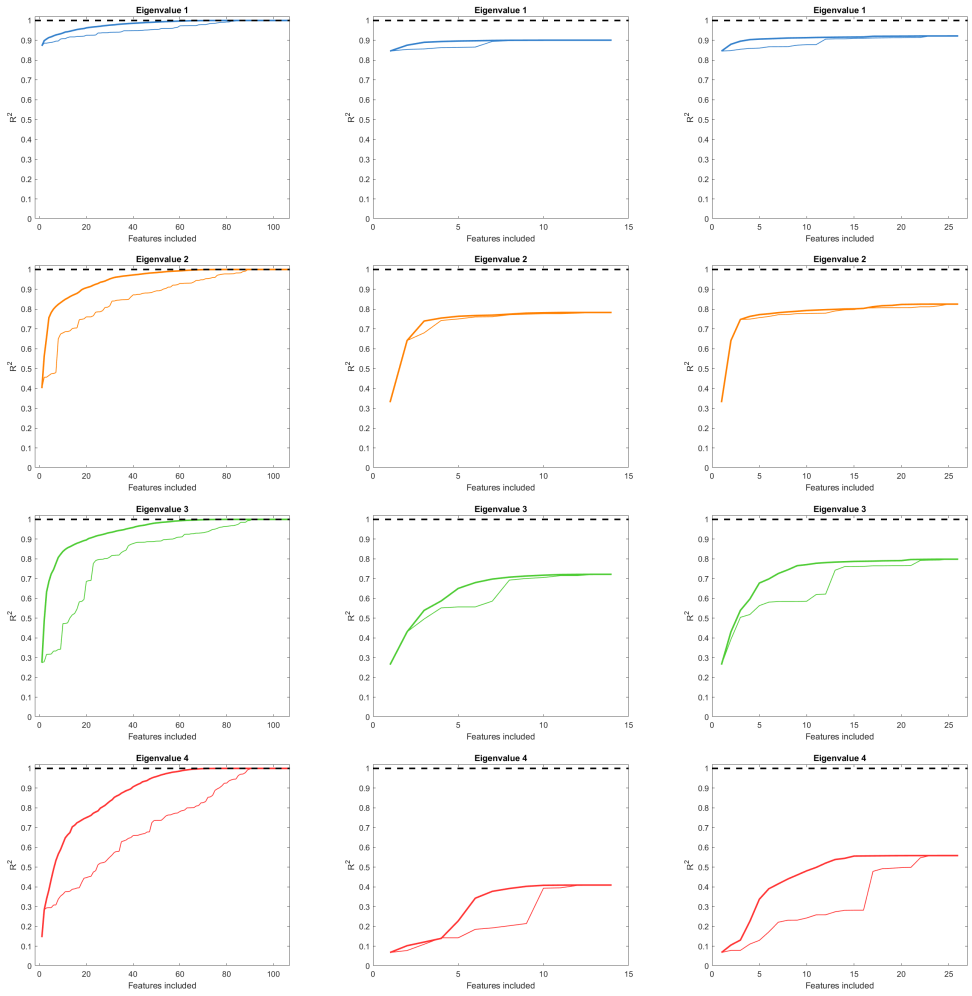


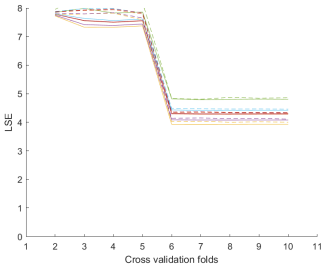
Figure B.1: R^2 relations for all EVs when fitted with an increasing number of features. The left column shows polynomial feature results with logarithmic features. The middle and right columns show linear feature results, without and with logarithmic features respectively. Thin and thick lines denote filter and filter-wrapper results respectively

	Non-log. features				Incl. log. features			
	λ_1	λ_2	λ_3	λ_4	λ_1	λ_2	λ_3	λ_4
1	ξ_2	ξ_4	ξ_5	ξ_4	ξ_2	ξ_4	ξ_5	ξ_4
2	ξ_{10}	ξ_5	ξ_4	ξ_{13}	ξ_{23}	ξ_5	ξ_4	ξ_{23}
3	ξ_3	ξ_{12}	ξ_1	ξ_{10}	ξ_3	ξ_{25}	ξ_1	ξ_1
4	ξ_6	ξ_{11}	ξ_3	ξ_1	ξ_{18}	ξ_{11}	ξ_{15}	ξ_2
5	ξ_7	ξ_9	ξ_{10}	ξ_2	ξ_6	ξ_{22}	ξ_{23}	ξ_5
6	ξ_1	ξ_8	ξ_{12}	ξ_5	ξ_7	ξ_{12}	ξ_{25}	ξ_{22}
7	ξ_5	ξ_{10}	ξ_{13}	ξ_9	ξ_{14}	ξ_7	ξ_{13}	ξ_9
8	ξ_9	ξ_2	ξ_7	ξ_{11}	ξ_5	ξ_{21}	ξ_{12}	ξ_8
9	ξ_{11}	ξ_1	ξ_8	ξ_8	ξ_1	ξ_8	ξ_7	ξ_{24}
10	ξ_{13}	ξ_{13}	ξ_2	ξ_{12}	ξ_9	ξ_{19}	ξ_3	ξ_{21}
11	ξ_8	ξ_7	ξ_{11}	ξ_6	ξ_{24}	ξ_{16}	ξ_2	ξ_{19}
12	ξ_{12}	ξ_6	ξ_9	ξ_7	ξ_{19}	ξ_{13}	ξ_8	ξ_{18}
13	ξ_4	ξ_3	ξ_6	ξ_3	ξ_{15}	ξ_{18}	ξ_{18}	ξ_{15}
14	1	1	1	1	ξ_{22}	ξ_9	ξ_{10}	ξ_{11}
⋮					⋮	⋮	⋮	⋮

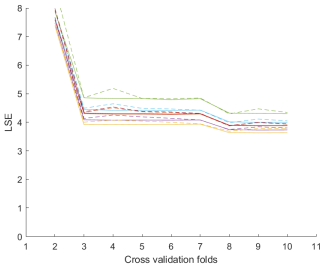
Table B.1: The 14 linear features most descriptive of each DVH eigenvalue, as selected by filter-wrapper feature selection. The ones at index 14 are included to account for the intercept coefficient. This table resembles the middle and right columns in Figure B.1

	Non-log features				Incl. log features			
	λ_1	λ_2	λ_3	λ_4	λ_1	λ_2	λ_3	λ_4
1	ξ_2	$\xi_5\xi_9$	$\xi_5\xi_9$	$\xi_3\xi_4$	$\xi_2\xi_{23}$	$\xi_5\xi_{16}$	$\xi_5\xi_9$	$\xi_3\xi_4$
2	$\xi_2\xi_{10}$	ξ_4	$\xi_4\xi_9$	$\xi_1\xi_{13}$	$\xi_3\xi_9$	$\xi_4\xi_{17}$	$\xi_4\xi_9$	$\xi_1\xi_{15}$
3	$\xi_3\xi_9$	$\xi_9\xi_{12}$	ξ_1^2	$\xi_2\xi_{13}$	$\xi_5\xi_7$	$\xi_{24}\xi_{25}$	ξ_1^2	$\xi_9\xi_{25}$
4	$\xi_5\xi_7$	ξ_8^2	$\xi_2\xi_{10}$	ξ_8^2	$\xi_3\xi_{16}$	$\xi_5\xi_9$	$\xi_2\xi_{10}$	$\xi_7\xi_{20}$
5	ξ_4^2	$\xi_7\xi_{10}$	ξ_6^2	$\xi_5\xi_8$	$\xi_5\xi_{21}$	$\xi_4\xi_{23}$	$\xi_2\xi_{25}$	$\xi_1\xi_8$
6	$\xi_4\xi_5$	$\xi_5\xi_{10}$	ξ_4^2	$\xi_4\xi_6$	$\xi_5\xi_{25}$	$\xi_9\xi_{25}$	$\xi_6\xi_{19}$	$\xi_5\xi_{18}$
7	ξ_{10}^2	$\xi_3\xi_5$	$\xi_1\xi_5$	$\xi_2\xi_7$	$\xi_9\xi_{18}$	ξ_8^2	$\xi_4\xi_{13}$	$\xi_2\xi_8$
8	ξ_2^2	$\xi_5\xi_8$	$\xi_4\xi_{10}$	$\xi_6\xi_{13}$	$\xi_4\xi_{16}$	ξ_{22}	$\xi_{16}\xi_{17}$	$\xi_{18}\xi_{25}$
9	$\xi_3\xi_8$	$\xi_4\xi_8$	$\xi_4\xi_7$	$\xi_3\xi_8$	$\xi_5\xi_6$	$\xi_7\xi_{10}$	$\xi_1\xi_5$	ξ_8^2
10	$\xi_4\xi_{13}$	$\xi_8\xi_9$	$\xi_8\xi_9$	$\xi_1\xi_7$	ξ_4^2	$\xi_5\xi_{10}$	$\xi_6\xi_{18}$	$\xi_5\xi_6$
11	ξ_8^2	$\xi_8\xi_{12}$	$\xi_2\xi_4$	$\xi_1\xi_5$	$\xi_{14}\xi_{17}$	$\xi_4\xi_{19}$	$\xi_1\xi_{16}$	$\xi_5\xi_{19}$
12	$\xi_3\xi_5$	$\xi_4\xi_5$	$\xi_3\xi_5$	ξ_5^2	$\xi_5\xi_{20}$	$\xi_2\xi_5$	$\xi_1\xi_{17}$	$\xi_3\xi_5$
13	$\xi_4\xi_6$	ξ_2	$\xi_3\xi_{13}$	$\xi_7\xi_{11}$	$\xi_5\xi_{10}$	$\xi_1\xi_8$	$\xi_{15}\xi_{17}$	ξ_9^2
14	$\xi_1\xi_4$	$\xi_3\xi_{12}$	$\xi_{10}\xi_{12}$	$\xi_7\xi_9$	$\xi_5\xi_{18}$	$\xi_6\xi_{20}$	$\xi_6\xi_{10}$	$\xi_{12}\xi_{18}$
15	$\xi_3\xi_{13}$	$\xi_1\xi_4$	ξ_8	ξ_3^2	$\xi_8\xi_{13}$	$\xi_5\xi_8$	$\xi_5\xi_7$	ξ_{14}^2
16	ξ_{10}	$\xi_1\xi_9$	$\xi_8\xi_{12}$	$\xi_4\xi_5$	$\xi_7\xi_{18}$	$\xi_4\xi_{13}$	$\xi_{12}\xi_{15}$	$\xi_9\xi_{15}$
17	$\xi_5\xi_8$	ξ_{11}	ξ_{13}^2	ξ_{13}	$\xi_8\xi_{16}$	$\xi_{13}\xi_{20}$	$\xi_1\xi_4$	$\xi_6\xi_{20}$
18	$\xi_8\xi_9$	$\xi_2\xi_{12}$	ξ_8^2	$\xi_9\xi_{13}$	$\xi_7\xi_{12}$	$\xi_4\xi_6$	$\xi_4\xi_{11}$	ξ_{24}
19	ξ_8	$\xi_9\xi_{11}$	$\xi_3\xi_7$	$\xi_3\xi_{13}$	$\xi_8\xi_{25}$	$\xi_{13}\xi_{16}$	$\xi_1\xi_{15}$	$\xi_9\xi_{18}$
20	$\xi_7\xi_8$	$\xi_2\xi_{11}$	$\xi_6\xi_7$	$\xi_4\xi_{12}$	$\xi_8\xi_9$	$\xi_4\xi_{15}$	$\xi_{13}\xi_{19}$	$\xi_8\xi_{25}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

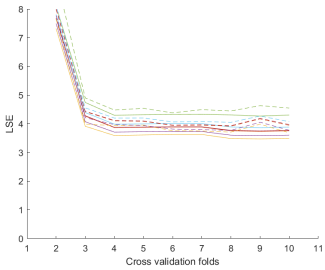
Table B.2: 20 polynomial features most descriptive of each DVH eigenvalue, as selected by filter-wrapper feature selection. This table resembles the left column of Figure B.1.



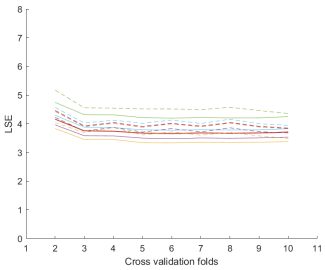
(a) 4% features



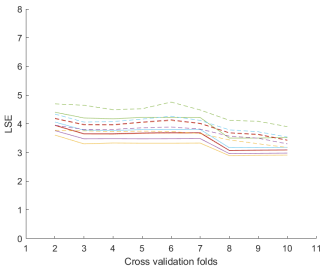
(b) 5% features



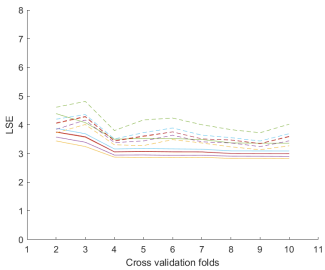
(c) 6% features



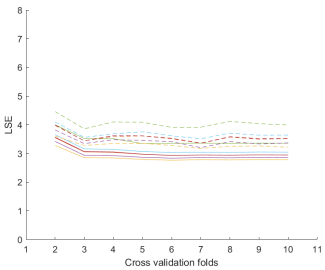
(d) 8% features



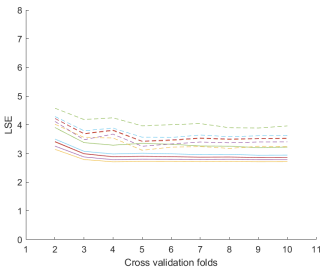
(e) 10% features



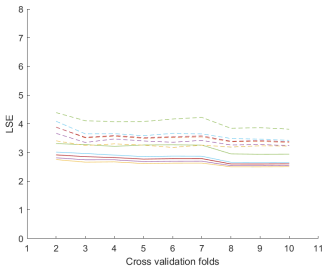
(f) 12% features



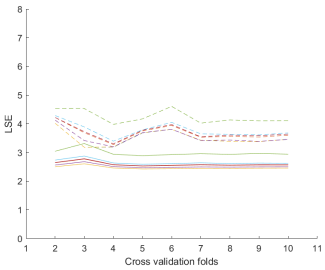
(g) 14% features



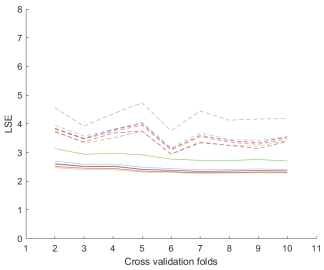
(h) 16% features



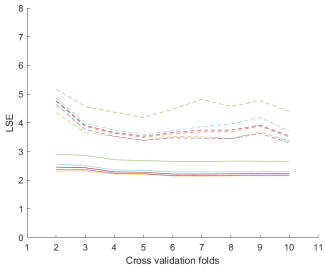
(i) 20% features



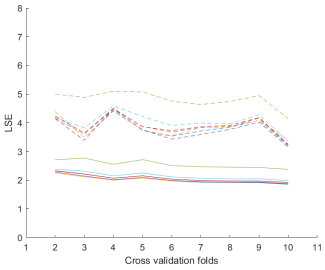
(j) 25% features



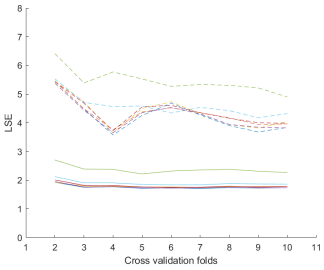
(k) 28% features



(l) 33% features

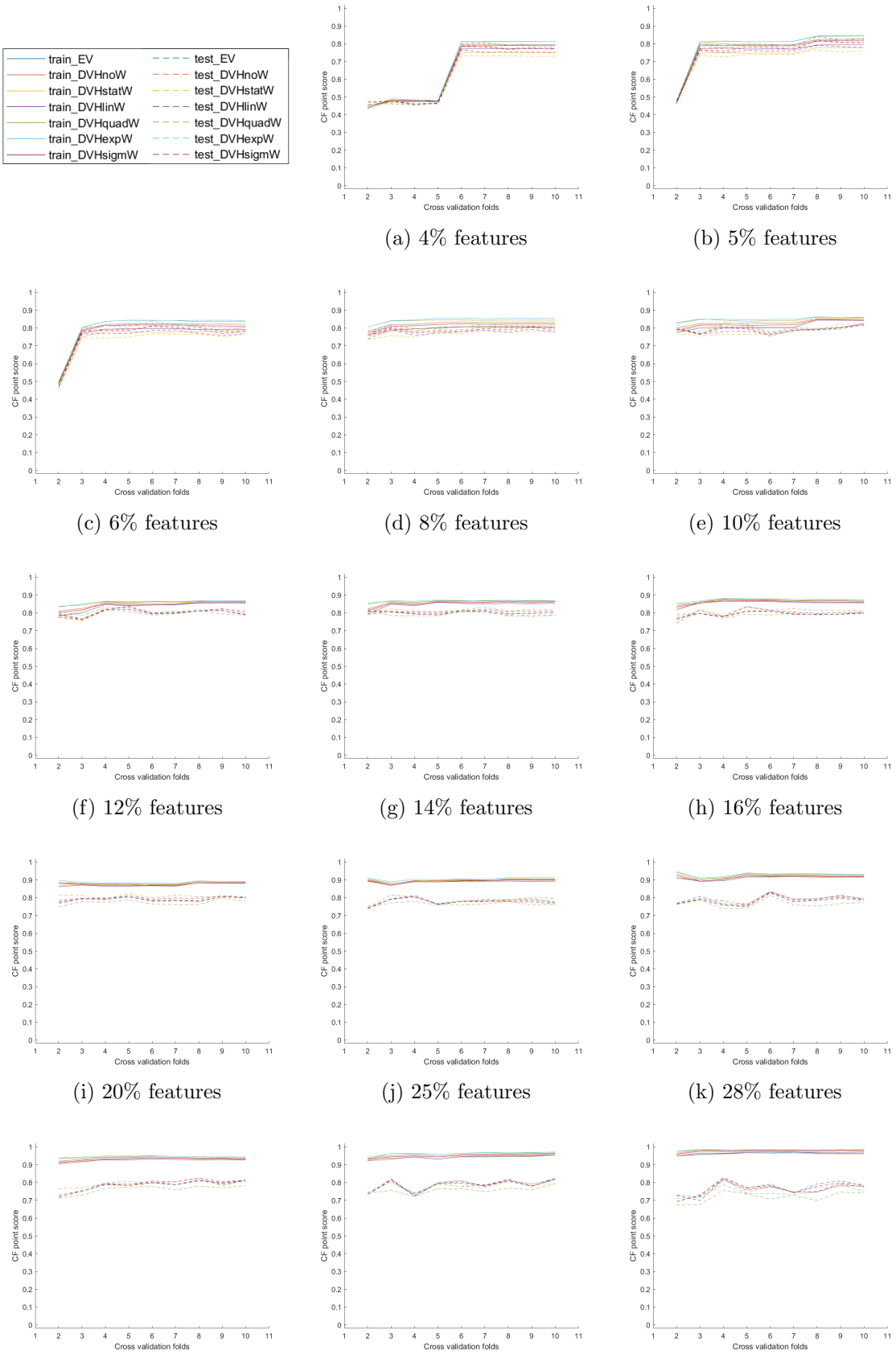


(m) 40% features



(n) 50% features

Figure B.2: Least square error train and test errors for several models, with respect to K folds and L maximum features. Only non-logarithmic features were included.



(a) 4% features (b) 5% features

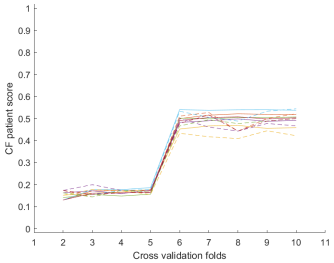
(c) 6% features (d) 8% features (e) 10% features

(f) 12% features (g) 14% features (h) 16% features

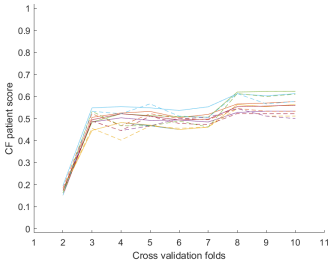
(i) 20% features (j) 25% features (k) 28% features

(l) 33% features (m) 40% features (n) 50% features

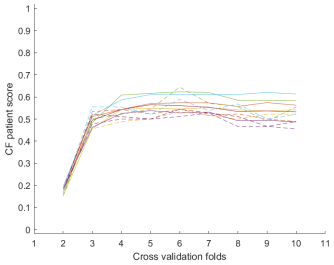
Figure B.3: TC_α train and test errors for several models, with respect to K folds and L maximum features. Only non-logarithmic features were included.



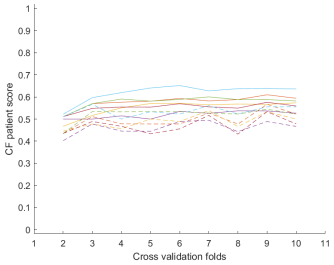
(a) 4% features



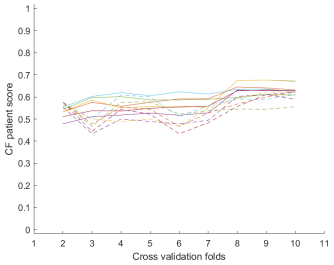
(b) 5% features



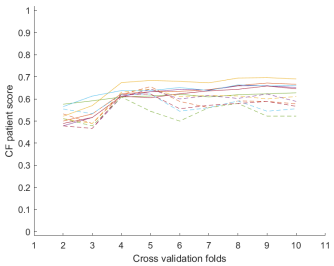
(c) 6% features



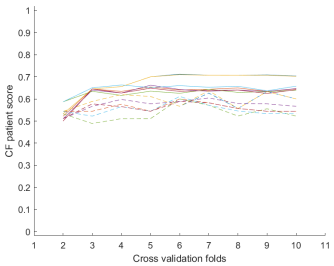
(d) 8% features



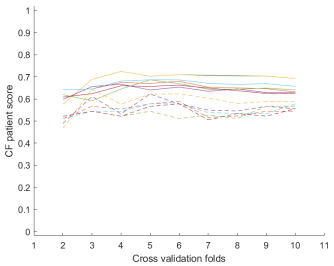
(e) 10% features



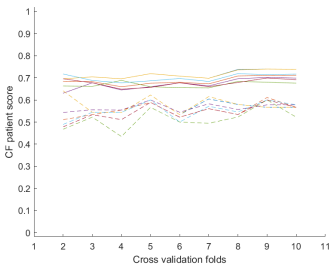
(f) 12% features



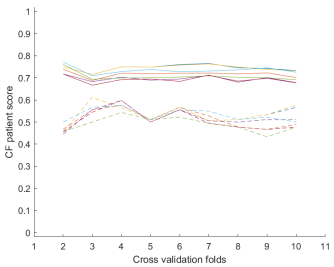
(g) 14% features



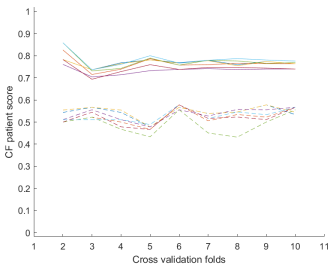
(h) 16% features



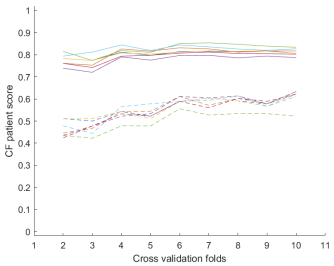
(i) 20% features



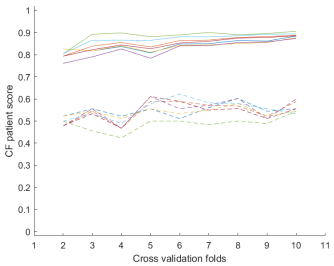
(j) 25% features



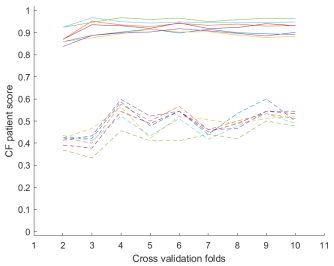
(k) 28% features



(l) 33% features



(m) 40% features



(n) 50% features

Figure B.4: TC_β train and test errors for several models, with respect to K folds and L maximum features. Only non-logarithmic features were included.

ABBREVIATIONS

3DCRT 3-dimensional conformal radiotherapy. 13

ALARA As Low As Reasonably Achievable. 15

AS Anal sphincter. 50, 66

CI Confidence interval. 41, 46–49, 65, 84

CT Computed Tomography. 14, 26, 50, 94, 95

CTV Clinical Target Volume. 20, 95

CV Cross-validation. 37, 46, 63, 65, 70, 71, 75, 77, 79, 81, 84, 85, 87, 88

DSB Double-Strand Break. 96, 97

DVH Dose Volume Histogram. 3, 7, 15, 22, 23, 26, 39–41, 43–47, 49, 53–60, 62, 63, 65–67, 69–88, 92, 93

EV Eigenvalue. 7, 53, 62, 63, 67, 70, 71, 84, 85, 87, 88, 98

EVR Explained Variance Ratio. 32, 33, 40, 66

gEUD Generalized equivalent uniform dose. 19, 45, 46

GTV Gross Target Volume. 95

HU Hounsfield Unit. 94

HWB Halfway boundary. 54, 56, 57, 64, 86

IMRT Intensity Modulated Radio-Therapy. 13, 16, 22, 93

KBP Knowledge-based planning. 3, 20–23, 40, 88, 91–93

KL Karhunen-Loève. 31, 32

LOOCV Leave-one-out cross-validations. 37, 46, 63, 65, 88

MCO Multi-criteria optimization. 16, 17, 21, 22, 97

MLC Multi-leaf collimator. 13, 18

MSE Mean Square Error. 32, 33

NKI Netherlands Cancer Institute. 3, 5, 20, 23, 26, 92, 95

NTCP Normal tissue complication probability. 97

OAR Orgat at Risk. 14–17, 21–23, 26, 50, 95

OER Oxygen-Enhancement Ratio. 97

OLS Ordinary least squares. 29, 31, 53, 54, 56–58

OVH Overlap Volume Histogram. 3, 7, 21–23, 26, 40, 45, 46, 50, 66, 67, 92

PC Principal Components. 32, 33, 40, 46, 66, 84, 87

PCA Principal Component Analysis. 3, 27, 31–33, 39, 40, 45, 50, 53, 65–69, 71, 76, 78–81, 84, 87, 92

PTV Planning Target Volume. 14, 15, 17, 19–23, 42, 45, 46, 50, 95

RBE Relative or Radiobiological Effectiveness. 97

RID Reactor Institute Delft. 3, 5

RMS Root mean square. 3, 63, 70, 72, 75, 77, 79, 81, 85, 86

ROM Reduced Order Modelling. 3, 23, 31

RT Radiotherapy. 12, 13, 16, 18, 20, 91, 97

RTT Radiotherapy Technician. 3

SD Standard deviation. 95

SQP Sequential quadratic programming. 18

SSB Single-Strand Break. 96, 97

SV Seminal Vesicles. 19, 20

TC Tolerance criterion. 3, 7, 40–44, 55–60, 69, 71, 76, 78, 85–88, 92

TCB Tolerance criterion boundary. 41, 43, 44, 54–59, 64, 69, 73, 78, 79, 85, 87

TCP Tumour control probability. 97

TPS Treatment Planning System. 14

VMAT Volumetric Modulated Arc Therapy. 13, 16, 18, 22, 26, 92

BIBLIOGRAPHY

[Ala+01] H. Alasti et al. "Portal imaging for evaluation of daily on-line setup errors and off-line organ motion during conformal irradiation of carcinoma of the prostate". In: *Int. Journ. Radiat. Oncol. Biol. Phys.* 49.3 (2001), pp. 869–884. DOI: 10.1016/S0360-3016(00)01446-2.

[Ans73] F.J. Anscombe. "Graphs in Statistical Analysis". In: *The American Statistician* 27.1 (1973), pp. 17–21. DOI: 10.2307/2682899.

[App+12] L.M. Appenzoller et al. "Predicting dose-volume histograms for organs-at-risk in IMRT planning". In: *Int. Journ. of Med. Phys.* 39.12 (2012), pp. 7446–7461. DOI: 10.1118/1.4761864.

[AS15] P.C. Austin and E.W. Steyerberg. "The number of subjects per variable required in linear regression analyses". In: *Journal of Clinical Epidemiology* 68.6 (2015), pp. 627–636. DOI: 10.1016/j.jclinepi.2014.12.014.

[Asa18] S. Asaithambi. *Why, How and When to apply Feature Selection.* 2018. URL: https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection-e9c69adfabf2.

[Bal+15] M. Balvert et al. "A framework for inverse planning of beam-on times for 3D small animal radiotherapy using interactive multi-objective optimisation". In: *Phys. Med. Biol.* 60.14 (2015), pp. 5681–1598. DOI: 10.1088/0031-9155/60/14/5681.

[Bal17] M. Balvert. "Improving the quality, efficiency and robustness of radiation therapy planning and delivery through mathematical optimization". PhD thesis. Tilburg University, 2017. ISBN: 978-90-5668-508-9.

[Bar+09] A. Barrett et al. *Practical radiotherapy planning.* Fouth edition. USA: CRC Press, 2009. ISBN: 978-034-09-2773-1.

[BDO07] A.J.J. Bos, F.S. Draaisma, and W.J.C. Okx. *Inleiding tot de stralingshygiëne.* Den Haag, the Netherlands: SDU publishers, 2007. ISBN: 978-90-12-11-905-4.

[Bel+96] A. Bel et al. "High-precision prostate cancer irradiation by clinical application of an offline patient setup verification procedure, using portal imaging." In: *Int. Journ. Radiat. Oncol. Biol. Phys.* 35.2 (1996), pp. 321–332. DOI: 10.1016/0360-3016(95)02395-X.

[BSH09] S. Breedveld, P.R.M. Storchi, and B.J.M. Heijmen. "The equivalence of multicriteria methods for radiotherapy plan optimization". In: *Phys. Med. Biol.* 54.23 (2009), pp. 7199–7206. DOI: 10.1088/0031-9155/54/23/011.

[Cha+11] V. Chanyavanich et al. "A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: an example application to prostate cancer planning". In: *Med. Phys.* 38.5 (2011), pp. 2515–1522. DOI: 10.1118/1.3574874.

[Cho+84] K.H. Cho et al. "The Effect of the CT Number for Each CT on Photon Dose Calculation". In: *Int. Fed. Med. Biol. Eng.* 14 (1984), pp. 1984–1986. DOI: 10.1007/978-3-540-36841-0_496.

[Cra13] D. Craft. "Multi-criteria optimization methods in radiation therapy planning: a review of technologies and directions". In: *Harvard Medical School* (2013).

[Cra16] D. Craft. *Multi-criteria optimization methods in radiation therapy planning: a review of technologies and directions*. 2016. URL: https://www.researchgate.net/publication/301542664_Multi-criteria_optimization_methods_in_radiation_therapy_planning_a_review_of_technologies_and_directions.

[Drz+91] R.E. Drzymala et al. "Dose-volume histograms". In: *Int. Journ. Radiat. Oncol. Biol. Phys.* 21.1 (1991), pp. 71–78. DOI: 10.1016/0360-3016(91)90168-4.

[Eri+09] K. Ericsson et al. "Volumetric Modulated Arc Therapy (VMAT) optimization with RayArc: RaySearch White Paper". In: *RaySearch Laboratories* (2009).

[GMW81] P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. New York, USA: Academic Press, 1981. ISBN: 978-0122839528.

[Goo+13] D. Good et al. "A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: an example application to prostate cancer planning". In: *Int. J. Radiat. Oncol. Biol. Phys.* 87.1 (2013), pp. 176–181. DOI: 10.1016/j.ijrobp.2013.03.015..

[Her+00] M. van Herk et al. "The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy". In: *Int. Journ. Radiat. Oncol. Biol. Phys.* 47.4 (2000), pp. 1121–1135. DOI: 10.1016/S0360-3016(00)00518-6.

[Her04] M. van Herk. "Errors and margins in radiotherapy". In: *Seminars in Radiation Oncology* 14.1 (2004), pp. 52–64. DOI: 10.1053/j.semradonc.2003.10.003.

[HFB15] A. Herzmann, D. Fleet, and M. Brubaker. *Machine Learning and Data Mining: Lecture Notes*. Department of Computer and Mathematical Sciences University of Toronto Scarborough. 2015.

[JDK17] T.M. Janssen, E. Damen, and M. Kusters. *Knowledge based Pareto front generation and automatic plan optimization in radiotherapy*. Open Technology Programme application, Netherlands Organization for Scientific Research. 2017.

[JMF07] K.W. Jee, D.L. McShan, and B.A. Fraass. "Lexicographic ordering: intuitive multicriteria optimization for IMRT". In: *Phys. Med. Biol.* 52.7 (2007), pp. 1845–1861. DOI: 10.1088/0031-9155/52/7/006.

[Joh13] J. Johnson. *General regression and over fitting*. 2013. URL: https://shapeof\\data.wordpress.com/2013/03/26/general-regression-and-over-fitting/.

[Kar94] J. Karhunen J. Joutensalo. "Representation and separation of signals using nonlinear PCA type learning". In: *Neural Networks* 7.1 (1994), pp. 113–127. DOI: doi.org/10.1016/0893-6080(94)90060-4.

[Kaz+09] M. Kazhdan et al. “A shape relationship descriptor for radiation therapy planning”. In: *Medical Image Computing and Computer-Assisted Intervention* (2009), pp. 100–108.

[Kon04] R. Kondor. *Regression by linear combination of basis functions*. Computer Science, Columbia University, New York. 2004. URL: http://www.cs.columbia.edu/~jebara/4771/tutorials/regression.pdf.

[LLB99] A.E. Lujan, E.W. Larsen, and J.M. Balter. “A method for incorporating organ motion due to breating into 3D dose calculations”. In: *Journ. Med. Phys.* 26.5 (1999), pp. 715–720. DOI: 10.1118/1.598577.

[Moo+11] K.L. Moore et al. “Experience-based quality control of clinical intensity-modulated radiotherapy planning”. In: *Int. Journ. Rad. Oncol. Biol. Phys.* 81.2 (2011), pp. 545–551. DOI: 10.1016/j.ijrobp.2010.11.030.

[Mur+08] K.K. Murthy et al. “Verification of setup errors in external beam radiation therapy using electronic portal imaging”. In: *Journ. Med. Phys.* 33.2 (2008), pp. 49–53. DOI: 10.4103/0971-6203.41192.

[Nie97] A. Niemierko. “Reporting and analyzing dose distributions: A concept of equivalent uniform dose”. In: *Med. Phys.* 24.1 (1997), pp. 103–110. DOI: 10.1118/1.598063.

[Nie98] A. Niemierko. “Radiobiological models of tissue response to radiation in treatment planning systems”. In: *Tumori* 84.2 (1998), pp. 140–143. DOI: 10.1177/030089169808400208.

[Orr96] M.J.L. Orr. *Introduction to Radial Basis Function Netoworks*. Center for Cognitive Science, University of Edinburgh, Edinburgh. 1996. URL: https://www.cc.gatech.edu/~isbell/tutorials/rbf-intro.pdf.

[Ott08] K. Otto. “Volumetric modulated arc therapy: IMRT in a single gantry arc.” In: *Med. Phys.* 35.1 (2008), pp. 310–317. DOI: 10.1118/1.2818738.

[Owe14] J.A. Owen. “Principal Component Analysis: Data Reduction and Simplification”. In: *McNair Scholars Research Journal* 1.2 (2014). URL: https://commons.erau.edu/mcnair/vol1/iss1/2.

[Pen56] R. Penrose. “On best approximate solution of linear matrix equations”. In: *Proceedings Cambr. Phil. Society* 52.2 (1956), pp. 17–19. DOI: 10.1017/S0305004100030929.

[Pet+12] S.F. Petit et al. “Increased organ sparing using shape-based treatment plan optimization for intensity modulated radiation therapy of pancreatic adenocarcinoma”. In: *Int. Journ. Radiother. Oncol.* 102.1 (2012), pp. 38–44. DOI: 10.1016/j.radonc.2011.05.025.

[PHC79] P. Parker, P.A. Hobday, and K.J. Cassell. “The direct use of CT numbers in radiotherapy dosage calculations for inhomogeneous media”. In: *Med. Phys. Biol.* 24.4 (1979), pp. 802–809. DOI: 10.1088/0031-9155/24/4/011.

[Ras18] S. Raschka. *Machine Learning FAQ: What is the difference between filter, wrapper, and embedded methods for feature selection?* 2018. URL: https://sebastianraschka.com/faq/docs/feature_sele_categories.html.

[RF08] R.B. Rao and G. Fung. “On the dangers of cross-validation: an experimental evaluation”. In: *Proceedings of the SIAM International Conference on Data Mining* (2008). DOI: 10.1137/1.9781611972788.54.

[RN09] S. Russell and P Norvig. *Artificial Intelligence: A Modern Approach*. Third edition. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. ISBN: 978-0-136-04259-4.

[RS00] S. T. Roweis and L. K. Saul. "Nonlinear Dimensionality Reduction by Locally Linear Embedding". In: *Science* 290.5500 (2000), 2323–2326. DOI: 10.1126/science.290.5500.2323.

[Sal+17] M. Salimi et al. "Assessment and Comparison of Homogeneity and Conformity Indexes in Step-and-Shoot and Compensator-Based Intensity Modulated Radiation Therapy (IMRT) and Three-Dimensional Conformal Radiation Therapy (3D CRT) in Prostate Cancer". In: *Journ. Med. Signals Sens.* 7.2 (2017), 102–107.

[SK06] K.J. Strauss and S.C. Kaste. "The ALARA (as low as reasonably achievable) concept in pediatric interventional and fluoroscopic imaging: striving to keep radiation doses as low as possible during fluoroscopy of pediatric patients—a white paper executive summary". In: *Pediatr Radiol.* 36.2 (2006), pp. 110–112. DOI: 10.1007/s00247-006-0184-4].

[SMM11] R.L. Scheaffer, M.S. Mulekar, and J.T. McClave. *Probability and Statistics for Engineers, International Edition*. Fifth edition. USA: Cengage Learning, Inc., 2011, pp. 607–622. ISBN: 978-0-538-73590-2.

[TK09] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Fourth edition. USA: Elsevier Inc., 2009. ISBN: 978-1-59749-272-0.

[Wan+16] Y. Wang et al. "Evaluation of plan quality assurance models for prostate cancer 1 patients based on fully automatically generated Pareto-optimal treatment plans". In: *Phys. Med. Biol.* 61.11 (2016), pp. 4268–4282. DOI: 10.1088/0031-9155/61/11/4268.

[Web03] S. Webb. "The physical basis of IMRT and inverse planning". In: *Br. J. Radiol.* 76.910 (2003), pp. 678–689. DOI: 10.1259/bjr/65676879.

[Web89] S. Webb. "Optimisation of conformal radiotherapy dose distributions by simulated annealing". In: *Phys. Med. Biol.* 34.10 (1989), pp. 1349–1370. DOI: 10.1088/0031-9155/36/9/005.

[Weixxa] E.W. Weisstein. *Least Squares Fitting*. 20xx. URL: http://mathworld.wolfram.com/LeastSquaresFitting.html.

[Weixxb] E.W. Weisstein. *Least Squares Fitting–Polynomial*. 20xx. URL: http://math\\world.wolfram.com/LeastSquaresFittingPolynomial.html.

[WHO18] WHO World Health Organization. *Cancer*. 2018. URL: http://www.who.int/news-room/fact-sheets/detail/cancer.

[WRS11a] E.P. Widmaier, H. Raff, and K.T. Strang. *Vander's Human Physiology*. Twelfth edition. New York, USA: McGraw-Hill, 2011. ISBN: 978-0-07-122215-0.

[WRS11b] B. Wu, F. Richcetti, and G. Sanguineti. "Data-driven approach to generating achievable dose-volume histogram objectives in intensity modulated radiotherapy planning". In: *Int. Journ. Radiat. Oncol. Biol. Phys.* 79.4 (2011), pp. 1241–1247. DOI: 10.1016/j.ijrobp.2010.05.026.

[Wu+09] B. Wu et al. "Patient geometry-driven information retrieval for IMRT treatment plan quality control". In: *Med. Phys.* (2009), pp. 5497–5505.

[Wu+13] B. Wu et al. "Using overlap volume histogram and IMRT plan data to guide and automate VMAT planning: A head-and-neck case study". In: *Am. Assoc. Phys. Med.* 40.2 (2013). DOI: 10.1118/1.4788671.

[YMS17] S. Yu H. Takao, T. Matsuura, and S. Shimizu. "Development of the DVH prediction method considering dose distribution in proton therapy". In: (2017).

[Yua+12] L. Yuan et al. "Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans". In: *Int. Journ. Med. Phys.* 39.11 (2012), pp. 6868–6878. DOI: 10.1118/1.4757927.