Advances in knowledge-based planning for radiotherapy with an open framework

by

Aaron Babier

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Mechanical and Industrial Engineering
University of Toronto

Abstract

Advances in knowledge-based planning for radiotherapy with an open framework

Aaron Babier

Doctor of Philosophy

Graduate Department of Mechanical and Industrial Engineering

University of Toronto

2022

Automated radiotherapy planning is on the verge of transforming personalized cancer treatment, and it will likely help address the growing global cancer burden. The most common type of automated planning is knowledge-based planning (KBP), which can generate treatment plans for radiotherapy without human intervention. In general, KBP is implemented as a two-stage pipeline that first predicts the dose that should be delivered to a patient, and then an optimization model converts that prediction into a treatment plan. Although KBP research is flourishing, it is largely limited to institution-specific datasets and evaluation metrics, which makes comparing competing approaches difficult.

The purpose of this thesis is to develop state-of-the-art KBP methods and an open framework for the KBP research community to benchmark new contributions. In this thesis, we developed the first generative adversarial network for KBP dose prediction, which outperformed several baselines. Next, we launched the OpenKBP Grand Challenge, which was the first platform that enabled researchers to compare KBP prediction methods fairly, and helped democratize KBP research by making it accessible to everyone. Next, we found that there were interaction effects between the two stages in KBP and that the choice of both stages can contribute to considerable variations in plan quality. To explore this further, we developed new plan optimization methods using open data and the dose predictions from the OpenKBP Grand Challenge. Overall, we improved KBP methods and promoted more collaboration within the KBP research community.

Dedicated to Mom and Dad.

Acknowledgements

First and foremost, thank you Tim. I am incredibly grateful for your mentorship and support. Thank you for giving me the freedom to pursue a variety of research projects. Your dedication to research is inspiring, and I could not have asked for a better supervisor.

I am also grateful for all the opportunities that I had to collaborate with others. Andrea, thank you for all of your guidance and for helping me better understand the clinical context of my research. Adam, thank you for always showing up with a contagious enthusiasm for our research. Rafid, thank you for pushing me to explore topics outside of my comfort zone; my fondest memory in research is of the time we spent collaborating on our first paper.

Thank you to my colleagues in the Applied Optimization Lab: Philip, Justin, Chris, Neal, Iman, Islay, Rafid, Ian, Minha, Ben P., Ben L., Jonathan, Clara, Bing, Nasrin, Matthew, Yusuf, Frances, Simon, Imran, Craig, Rachel W., Bo, Nathan, Albert, Coco, Rachel S., Jamal, Christina, Jesse, and Jangwon. You all made my time in graduate school memorable. A special thank you to Ian for always being open to a coffee run and, maybe more importantly, for your friendship.

I am also thankful to my good friends outside of academia: Alex, Jenson, and Zander. Thank you for being there when I needed respite from my studies. I always look forward to spending time with you.

I will forever be grateful to my parents. I would not be where I am today without their unconditional love and support. Thank you for always encouraging me to pursue my own interests.

Finally, thank you Minha. Meeting you was the highlight of my time in graduate school. Thank you for teaching me to be more spontaneous and bringing Chestnut into our family. I am excited to start the next chapter of our lives together.

Contents

Chapter 1

Introduction

Cancer is the leading cause of premature death in 57 countries[24] and that number will likely increase based on estimates that the global cancer burden will grow by 47% between 2020 and 2040.[97] Oncologists generally recommend treating cancer via surgery, systemic therapy, radiotherapy, or a combination of those three modalities. However, each of those modalities requires significant resources and trained personnel, which limits access to treatment in resource constrained healthcare systems around the world.[6] To reduce the strain on existing resources in other areas of healthcare, artificial intelligence (AI) has been implemented to augment human decision making and improve clinical efficiencies,[1] and recent clinical trials have demonstrated that AI can be implemented to augment decision making in radiotherapy.[77] The goal of this thesis is to develop AI tools for radiotherapy and promote more collaboration within the research community to improve the rate of innovation.

It is estimated that external beam radiotherapy is a recommended treatment for managing the disease in about half of all patients diagnosed with cancer.[16;21;40] This implies that improving processes related to external beam radiotherapy will have a significant impact on a large population of patients. External beam radiotherapy is delivered by a wide range of mechanisms that include radioactive sources (e.g. Cobalt-60), high-energy particle beams (e.g., electron, proton), and high-energy photon beams (e.g., three-dimensional conformal radiotherapy, volumetric modulated arc therapy, intensity modulated radiation therapy (IMRT)). During IMRT, which is the focus of this thesis, a linear accelerator (LINAC) projects high-energy photons from multiple angles around a patient (see Figure 1.1) to destroy cancerous tissue while minimizing damage to the healthy tissue. Treating a patient with IMRT requires a patient-specific *treatment plan*, which includes the instructions that a LINAC follows to deliver a patient-specific cancer treatment. The treatment plan is generated by a complex design process involving multiple medical professionals and a treatment planning system.

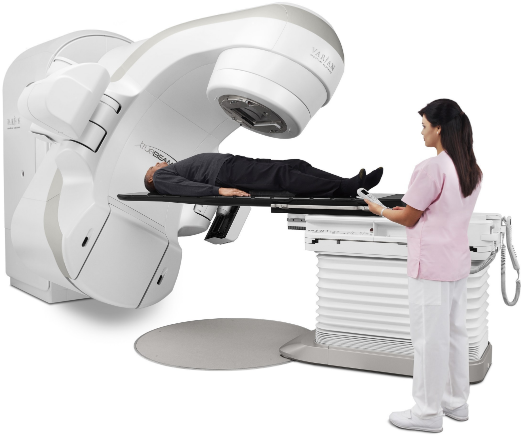


Figure 1.1: A patient lies on a bed as a LINAC rotates around him.

The time required to generate an acceptable patient-specific treatment plan varies depending on the type of cancer. For example, the median time to generate a prostate and head-and-neck treatment plan is 7.6 hours and 12.9 hours, respectively.[59] Developing head-and-neck cancer treatment plans are particularly challenging because important healthy tissue is often adjacent to or consumed by cancerous tissue. Additionally, the range in ways that head-and-neck cancer presents makes the patient cohort heterogenous[58] and contributes to considerable variation between treatment plans.

In this thesis, we develop models with open frameworks to generate treatment plans using AI. We focus exclusively on a type of head-and-neck cancer called oropharyngeal cancer, which occurs in the oropharynx (see Figure 1.2). Our motivation for focusing on oropharyngeal cancer stems from its complexity. Specifically, an AI tool that performs well on a complicated site like oropharynx should generalize to simpler sites where treatment plans have less variation between patients (e.g., prostate). We elaborate more on the details of the treatment planning process in the next section to provide more context for the contributions of this thesis.

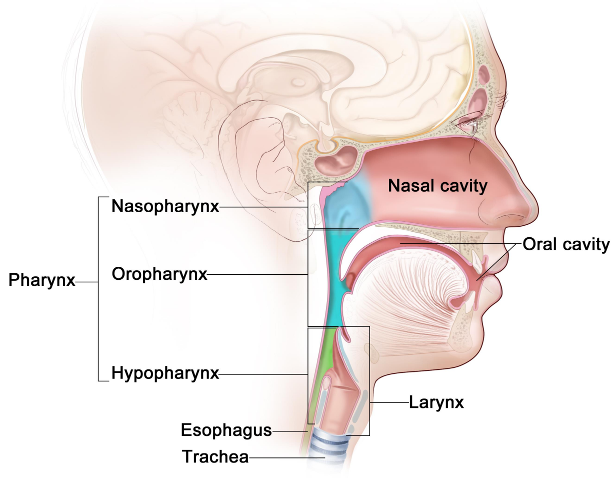


Figure 1.2: An overview of head-and-neck site.

1.1 Intensity modulated radiation therapy

Figure 1.3 summarizes the steps between an oncologist choosing to proceed with an IMRT treatment and treating the patient. The treatment plan will follow a protocol that defines the goals for the treatment. To provide context, we will summarize the three primary components that lead up to treating the patient (i.e., imaging the patient, segmenting the images, and generating the treatment plan). However, thereafter this thesis will focus on the process of generating the treatment plan.

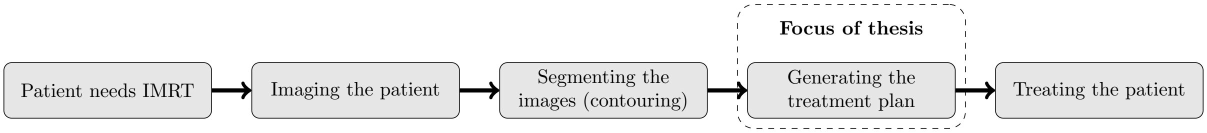


Figure 1.3: Overview of the steps involved in preparing a patient for IMRT.

1.1.1 Patient imaging

A computed tomography (CT) scan is performed to provide clinicians with the images of a patient's anatomy. The anatomy images are acquired as a series of two-dimensional (2D) cross-section images called *slices*, which are separated by small distances that equate to the *thickness* of the slices. Those slices are then stitched together to create a three-dimensional (3D) representation of the corresponding patient's anatomy. In that 3D representation, the anatomy of the patient is discretized into a grid of small volumes called *voxels*, which each encase a single CT image pixel within the thickness of its slice. CT images are essential for radiotherapy because they provide tissue density information, which is used to quantify how radiation travels through the patient and calculate the dose that will be deposited in various tissues from a radiation beam of known intensity. Additionally, CT images provide anatomical information that physicians use to distinguish between cancerous and non-cancerous tissue. Depending of the type of cancer, other imaging modalities (e.g., magnetic resonance imaging, positron emission tomography) may also be used to help the oncologist identify cancerous tissue that is difficult to ascertain on CT images.

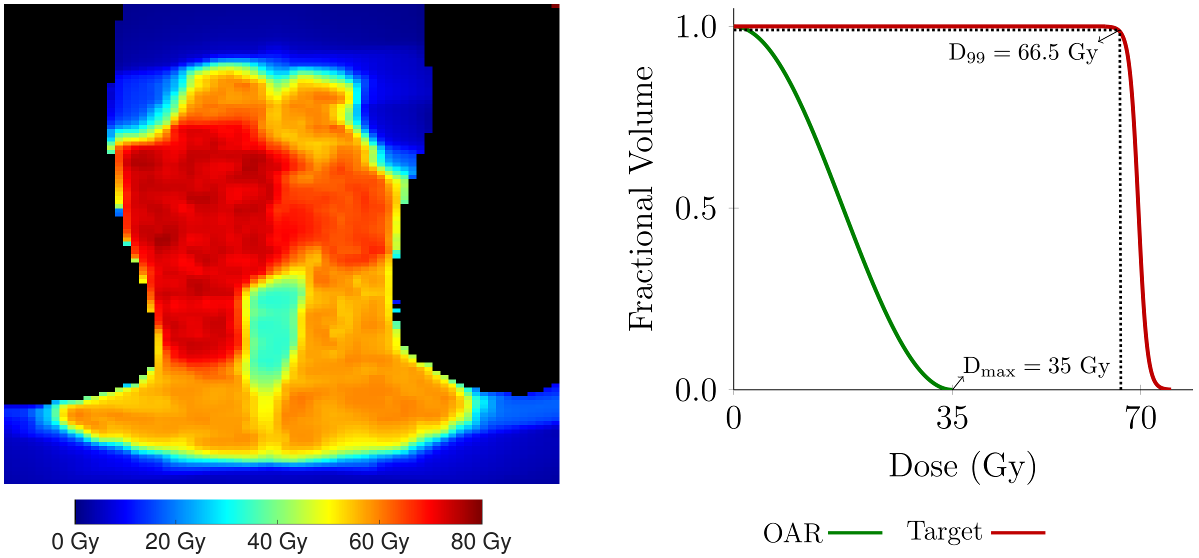
1.1.2 Image segmentation

Next, clinicians draw contours on the CT images to create *regions-of-interest* (ROIs). These ROIs are classified as either *organs-at-risk* (OARs), which are important healthy structures, or *targets*, which are cancerous or potentially cancerous tissues. One of the goals in radiotherapy is to minimize the dose delivered to OARs, which helps to limit treatment side effects and toxicity to the patient. However, it is inevitable that OARs will receive some dose during a treatment, and the acceptable dose level delivered to each OAR varies based on the *radiosensitivity* of the structure. The second goal in radiotherapy is to deliver the dose level prescribed to each target by the oncologist. In head-and-neck cancer, multiple dose levels are often prescribed to different targets. For example, the gross disease is generally prescribed the highest dose level, and a much larger volume that includes lymph nodes at risk for microscopic disease, which can not be visualized on the CT scan, is prescribed a lower dose level.

Throughout this thesis we will only consider manually drawn contours, however, we acknowledge that segmenting these images manually is another time consuming step in the process of preparing a patient for IMRT.[59] Currently, there are several groups developing tools that do segmentation without human intervention using a processes called auto-segmentation.[27] Some of those tools have also been adopted into clinical practice.[93]

1.1.3 Generating treatment plans

As mentioned previously, the goals of radiotherapy are to minimize the dose delivered to healthy tissues and deliver a prescribed dose of radiation to the targets. These goals are often evaluated by clinicians who inspect the corresponding *dose distribution*, which shows how much dose a treatment plan will deliver to every voxel in the patient (see Figure 1.4(a)). Dose distributions are often consolidated into summary statistics like *dose-volume histograms* (DVHs), which show how much dose is delivered to fractional volumes of each ROI (see Figure 1.4(b)). There are generally specific points along a DVH that are especially important to consider during the evaluation process. Most notably there are *clinical criteria* that put limits on specific DVH points for each ROI (e.g., $D_{max} \leq 35$ Gy is an upper bound of 35 Gy to an OAR, $D_{99} \geq 66.5$ Gy is a lower bound of 66.5 Gy on 99% of the voxels in a target). These clinical criteria are effectively institutional guidelines that are based on existing literature and population statistics.



(a) Sample dose distribution

(b) Sample DVHs

Figure 1.4: A sample treatment plan that is summarized by (a) a slice of its dose distribution and (b) a DVH for an OAR and target where two important DVH points are also indicated.

A treatment plan is usually generated according to an institutional planning protocol that is also informed by population statistics. However, since all patients are different they need personalized treatment plans, which may need to compromise on some guidelines in the protocol (e.g., OAR clinical criteria). As a result, a treatment plan is usually generated by solving a multi-criteria optimization model, which is known as an *inverse planning model*, that balances tradeoffs with the guidelines.[26] Typically, the tradeoffs involve increasing the dose to OARs in an effort to achieve acceptable levels of dose delivered to the targets. Those goals are quantified using various constraints and a cost function that sums a series of objective functions that each have an objective function weight, which is chosen to reflect the relative importance of the corresponding objective function.

The decision variables in the inverse planning model are related to the intensity modulation of the beam. At each angle, the LINAC projects a beam with constant intensity that is shaped into a series of irregular shapes called *apertures*. The cumulative dose that is delivered by each aperture corresponds to the dose distribution delivered by a treatment plan. Selecting the optimal apertures is a non-convex optimization problem, and to simplify it practitioners often use *fluence map optimization* (FMO).[26] In FMO, the beam at each angle is divided into a grid of *beamlets*, which comprise the fluence map of the treatment plan. Solving an FMO model returns beamlets with intensities that make the specified cost function optimal. A second optimization model, called *leaf sequencing*, is then used to convert the fluence map into a set of deliverable apertures.[98] In this thesis, we only generate fluence-based plans via FMO and do not do leaf sequencing. However, to ensure that our plans represent realistic dose distributions we use a *sum-of-positive gradients* constraint that makes the fluence-based plans closely resemble plans that are deliverable by apertures.[35]

In practice, an inverse optimization model is developed via a process called *inverse planning* where a dosimetrist specifies patient-specific goals for the treatment. Dosimetrists perform inverse planning using specialized optimization software to iteratively tune several parameters and solve the corresponding inverse planning model, which generates a treatment plan that is subsequently evaluated by an oncologist (see Figure 1.5).[19] The oncologist usually proposes modifications to the plan that require the dosimetrist to adjust the plan via inverse planning. The total process is labor intensive, time-consuming, and costly, as the back-and-forth between the planner and oncologist is often repeated multiple times until the plan is finally approved.[59] In this thesis, we use AI to automate this treatment planning process.

1.2 Knowledge-based planning

The significant manual effort associated with the current treatment planning paradigm, along with the fact that IMRT plans are generally quite similar for patients with similar geometries, has motivated researchers to investigate how automation can be used in the

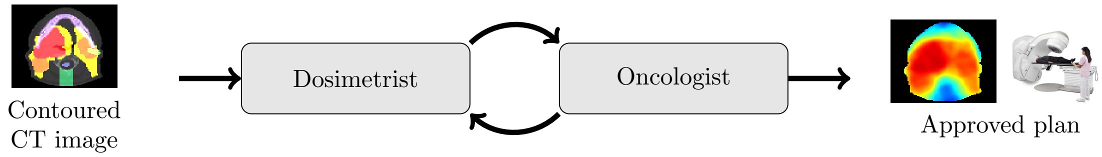


Figure 1.5: An overview of the iterative clinical treatment planning process called inverse planning. A medical physicist is also often involved in this iterative process but is excluded in this overview in order to show the most simplified workflow.

planning process.[94] A key enabler of automation is known as knowledge-based planning (KBP), which leverages historically delivered treatments to generate new plans for similar patients. Figure 1.6 depicts the two main components of a KBP-driven automated planning pipeline: (i) a dose prediction model that uses CT-derived patient geometric features to predict a clinically acceptable dose distribution or its summary statistics (e.g., DVH points);[5;96;111;112] and (ii) a plan optimization model that converts the prediction into a treatment plan.[7;75;109] The second step is needed because the dose prediction from the first component does not include delivery instructions (e.g., fluence to deliver dose).

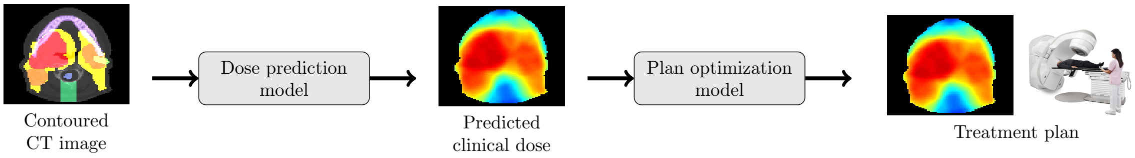


Figure 1.6: An overview a knowledge-based planning pipeline.

1.2.1 Dose prediction models

Dose prediction models predict the amount and location of dose that an acceptable treatment plan should deliver to a patient. All dose prediction models use machine learning to "learn" relationships from previously generated high-quality treatment plans, but the models have changed a lot over the last decade.[79] Many of the early models used machine learning methods like linear regression,[111] principal component analysis,[7;113;116] and random forests.[74] Those models relied heavily on feature engineering to condense 3D patient images into lower dimensional features (e.g., overlap-volume histogram).[107] More recently, the field of dose prediction has been dominated by computer vision models,[79] and feature engineering is now used to supplement, not replace, 3D patient images.[85] The first prediction model that used computer vision was developed in 2017,[79;82] and the work in this thesis was the first to incorporate computer vision models into a full KBP pipeline to generate treatment plans.[71]

The motivation for developing dose prediction models has also evolved in the last decade. Early models were largely developed to help dosimetrists identify achievable dosimetric objectives (i.e., DVH metrics), which would enable clinics to decrease variability in planning and increase the overall quality of treatments.[46] Dosimetrists could use these predictive models to estimate when and to what degree it was appropriate to compromise on some treatment goals or objectives.[113] More recently, there is a move towards incorporating dose prediction models into KBP pipelines as an intermediate step that provide patient-specific parameters for the optimization model.[76] A KBP pipeline recently produced promising results in clinical trials where clinicians selected the KBP generated plan over a manually generated plan to treat patients in 72% of cases.[77]

1.2.2 KBP plan optimization models

Optimization models are used to translate dose predictions into treatment plans. The dominant type of KBP optimization model is *dose mimicking*, which attempts to recreate plans that have similar objective values to the dose prediction.[76] In general, a practitioner can choose to optimize over a set of structure-based objective functions, which quantify a measure of the dose delivered to a single ROI, or voxel-based objective functions, which quantify a measure of the dose delivered to a single voxel.[11] These dose mimicking models are generally designed as *fully-automated* processes that generate plans without human intervention. As a result, there are no intuitive processes for fixing problems with a KBP generated plan.

Another type of KBP optimization model is *inverse optimization* (IO).[7] An IO model estimates the objective function weights that would make the prediction optimal in an inverse planning model.[8] Note that we use the term IO to be consistent with the operations research community,[31] however, it is effectively the reverse of inverse planning in the radiotherapy community; specifically, inverse planning takes objective function weights as input to generate a treatment plan, and inverse optimization takes a dose distribution as input to generate objective function weights. One benefit of IO is that it can be used as a *semi-automated* process that generates plans without human intervention and then enables a human to improve the plan via an intuitive process. Specifically, IO generated weights can be used to generate a plan via inverse planning, and if the dosimetrist needs to improve the plan then those weights can be adjusted manually before re-solving the inverse planning model. In this thesis, we focus mostly on IO models for plan optimization, and we also demonstrate that the plans generated by dose mimicking and an IO process are equivalent under certain conditions.

1.2.3 Evaluating KBP generated plans

No single metric can effectively quantify the quality of a treatment plan.[80] The gold standard for KBP is to have a radiation oncologist evaluate several attributes of the KBP generated treatment plan before approving it for treatment.[77] However, reviewing a treatment plan is time consuming, and in most research studies it is impractical to subject each iteration of KBP generated plans to a rigorous review by a radiation oncologist. Instead, it is common for researchers to define several summary statistics that compare dose predictions and KBP generated plans to the corresponding clinical treatment plan (i.e., ground truth or reference plan).[44;64] Common types of evaluation metrics involve clinical criteria and DVHs that radiation oncologists use to evaluate plan quality during their plan review process.[7;61]

Different institutions adopt a wide range of planning protocols.[54] This heterogeneity among institutions has crept into KBP research because most research groups tailor their methods and evaluation metrics to their institution.[7;59] As a result, most models are evaluated on institutional-specific metrics that are not always widely adopted, which makes it challenging to compare competing models. Adopting a small set of standardized metrics to augment institutional-specific evaluation metrics would enable us to better measure progress in the field. Prior to our work, there were no recommended standardized metrics to evaluate KBP predictions or plans.[13]

1.2.4 Datasets

Access to data is a major barrier to knowledge-based planning research, which is dominated by models developed on private datasets.[79] Developing models on private datasets limits the number of KBP researchers to those with access to data, thereby stifling innovation. Additionally, it is difficult to rigorously benchmark methods that are developed on drastically different datasets (e.g., prostate versus head-and-neck). Open datasets (e.g., CIFAR)[63] are a staple in thriving AI-driven fields that democratize research efforts and enable researchers to evaluate their methods using a common dataset. Prior to our work, there was no open datasets to develop KBP prediction or plan optimization methods.[13]

1.3 Contributions and outline

Incorporating AI tools into radiotherapy is an effective means for improving cancer care, however, development of these tools is impeded by a lack of standardized metrics and datasets. The purpose of this thesis is to improve knowledge-based planning techniques and promote more collaboration within the KBP research community. To achieve this purpose we accomplished the following:

1. Developed the first dose prediction models that use generative adversarial networks
2. Implemented the first KBP pipelines that use computer vision and optimization
3. Garnered widespread adoption of standardized metrics for KBP research
4. Published the first open datasets for KBP research
5. Identified interaction effects between the stages in KBP that affect performance

The remainder of this thesis is organized into five self-contained chapters. The first three chapters focus on dose prediction models and the last two focus more on KBP optimization models. Each of these chapters is summarized in more detail below.

1.3.1 Dose prediction with 2D computer vision

This chapter was published as "Automated treatment planning in radiation therapy using generative adversarial networks" in *Proceedings of Machine Learning Research*, Vol. 85 (*Machine Learning for Healthcare*), pp. 484-499, 2018.[7] This chapter was developed with significant contributions from Rafid Mahmood. As a result, this chapter closely resembles Chapter 4 from his thesis.[70] My primary contributions were formatting the patient data for computer vision models, implementing three of the four baseline models, constructing the optimization models, and evaluating the clinically relevant metrics. Rafid's primary contributions were training the neural networks and tailoring this work for a machine learning audience.

In Chapter 2, we develop the first automated treatment planning pipeline for oropharyngeal cancer that uses a conditional generative adversarial network (GAN)[57] to predict slices of 3D dose distributions as a colored red-green-blue (RGB) heatmap. In contrast to previous machine learning methods before 2017,[46] our approach does not require the pre-specification of an extensive set of feature variables for prediction. Instead, our model learns what features are important to predict clinical quality dose distributions from contoured patient images. We compare our approach to several other techniques from the literature: a query-based method,[7] linear regression,[7] random forest,[75] and a U-Net.[82] We demonstrate that our approach outperforms all other models across several clinical metrics when it is used as the intermediate step in a full KBP pipeline.

1.3.2 Dose prediction with 3D computer vision

This chapter was published as "Knowledge-based automated planning with three-dimensional generative adversarial networks" in *Medical Physics* Vol. 47, pp. 297-306, 2020.[10]

In Chapter 3, we build on our previous work by adjusting our GAN architecture to predict dose in standard units of gray and as a full 3D dose distribution (i.e., not a single slice of the distribution). Once again, we use these predictions as input into an optimization model to produce plans.[8] The plans are compared to plans generated using predictions from two baseline models: 1) 2D-RGB,[71] which is the model developed in Chapter 2 and 2) DoseNet,[61] which is based on a model in the literature with a U-Net style architecture. Additionally, we investigate the impact of multiplicatively scaling the predictions before optimization, such that the predicted dose distributions achieve all target clinical criteria. We find that the best performing plans are generated using predictions from the 3D model that are multiplicatively scaled.

1.3.3 OpenKBP: The open dose prediction challenge

This chapter was published as "OpenKBP: The open-access knowledge-based planning grand challenge and dataset" in *Medical Physics*, Vol. 48, pp. 5549-5561, 2021.[13]

In Chapter 4, we report on an international machine learning competition for dose prediction that we organized called the Open Knowledge-Based Planning Grand Challenge (OpenKBP) that was sponsored by the American Association of Physicists in Medicine.

OpenKBP advances KBP research by providing a platform to enable fair and consistent comparisons of dose prediction methods. To facilitate the Challenge, we publish the first open dataset and standardized metrics that cater to KBP dose prediction methods. Participants in the Challenge use a large dataset to train, test, and compare their prediction methods, using our standardized metrics, with those of other participants. The Challenge proceeds in two phases. In the first (validation) phase, teams develop their models and compare their performance in real time to other teams via a public leaderboard. In the second (testing) phase, teams submit their dose predictions for an unseen data set and a final set of winners is determined. The Challenge also provides the first platform that enables researchers to compare KBP prediction methods fairly, and it helps democratize KBP research by making it accessible to everyone. We also collect equity, diversity, and inclusion (EDI) data on the Challenge participants to publish the first set of EDI data for the KBP research community.

1.3.4 Evaluating complete automated planning pipelines

This chapter was published as "The importance of evaluating the complete automated knowledge-based planning pipeline" in *Physica Medica*, Vol. 72, pp. 73-79, 2020, and it was also selected for the *Rising Stars Completion* at the *International Conference on the Use of Computer in Radiotherapy.*[11]

In Chapter 5, we evaluate how KBP prediction methods combine with optimization methods in a two-stage KBP pipeline. Although there have been significant advances in KBP research, improvements have typically been measured by modifying one stage while holding the other constant.[7;11;76] Whether specific combinations of KBP and optimization models produce superior plans has not been considered in the extant literature. Thus, we compare the plans generated by four different KBP pipelines that were assembled from all possible permutations of each component from two state-of-the-art KBP pipelines.[11;76] We show that the way in which these two stages combine alters the quality of the output treatment plan. We find that state-of-the-art prediction methods when paired with different optimization algorithms, produce treatment plans with considerable variation in quality.

1.3.5 OpenKBP: An open framework for plan optimization

This chapter is published as "OpenKBP-Opt: An international and reproducible evaluation of 76 knowledge-based planning pipelines" as a preprint on arXiv.[14]

In Chapter 6, we compare the performance of four novel KBP plan optimization models using 19 different prediction models that were developed in the OpenKBP Grand Challenge. Our plan optimization methods all perform dose mimicking to generate plans with dose objective values that "mimic" (i.e., closely match) the input dose objective values, and we demonstrate a clear link between dose mimicking and inverse optimization methods, which could be used to enable human planners to improve the plans generated by KBP. We also publish our code and data to make this the first open dataset that caters to researching KBP optimization methods. We find that many dose prediction methods can achieve low error, however, optimization often improves upon the predictions and often minimizes clinically relevant differences between prediction methods. Thus, it is critical that we improve the optimization stage in KBP to get better utility out of the existing high-quality dose prediction methods.

Chapter 2

Dose prediction with 2D computer vision

As outlined in Chapter 1, knowledge-based planning (KBP) is an automated approach to radiotherapy treatment planning that involves a dose prediction model and an optimization model. In this chapter, we develop a new dose prediction model and compare it to other recent models. Our new model eschews the previous paradigms of site-specific feature engineering and predicting low-dimensional representations of the plan, which are common in earlier dose prediction models. This is the first study that uses a generative adversarial network (GAN) to predict dose, and it also the first study to compare the performance of computer vision models in a full KBP pipeline. We compare the performance through a series of experiments on a private dataset of 217 oropharyngeal cancer treatment plans.

2.1 Introduction

External beam radiotherapy is recommended for about half of all patients diagnosed with cancer.[40] During external beam radiotherapy, a linear accelerator (LINAC) outputs high-energy x-ray beams from multiple angles around a patient to deliver a prescribed dose of radiation to a set of targets while minimizing dose to healthy tissue. The LINAC delivers radiotherapy by following the instructions contained in a patient-specific treatment plan, which is the result of a design process that involves multiple medical professionals and a treatment planning system. This includes specialized optimization software that determines the beam characteristics (e.g., aperture shapes for each beam angle, dose delivered from each aperture) required to deliver the final dose distribution. The optimization model takes as input a set of various dosimetric objectives, constraints, and other parameters that guide the optimization process. The model outputs a treatment plan that is subsequently evaluated by an oncologist. The oncologist usually suggests revisions to the plan, which then requires the dosimetrist (i.e., the practitioner who generates treatment plans) to re-solve the optimization model using updated parameters. The total process is labor intensive, time-consuming, and costly, as the back-and-forth between the dosimetrist and oncologist is often repeated multiple times until the plan is finally approved.

Contrasting the iterative clinical procedure, knowledge-based planning is a data-driven approach that learns from historical plans to generate new plans for future patients. Figure 2.1 depicts the two main stages of a KBP pipeline: (a) a machine learning model that predicts a clinically satisfactory dose;[5;96;111;112] and (b) an optimization model that generates a deliverable treatment plan.[7;75;109] The second step is needed to ensure the treatment plan produced by the machine learning model satisfies the physical delivery constraints imposed by the LINAC.

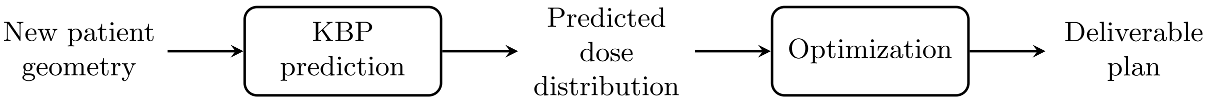


Figure 2.1: Overview of KBP-driven automated treatment planning pipeline.

A major limitation of most existing KBP prediction methods is their reliance on low-dimensional hand-tailored features derived from patient geometry to predict new dose distributions. In contrast, we propose a new paradigm for generating KBP predictions that automatically learns to predict a 3D dose distribution directly from a CT image. More specifically, we recast the dose prediction problem as an image colorization problem, which we solve using a GAN.[48] GANs, which have produced impressive results in other image colorization applications,[57;115] involve a pair of neural networks: a generator that performs a task and a discriminator that evaluates how well the task is performed. In our application, the generator serves as a dosimetrist that designs a treatment, while the discriminator plays the role of the oncologist who critiques the generated dose distribution by comparing it to the real treatment plan. Both neural networks train simultaneously on historical data, effectively replicating and aggregating the combined knowledge gained during the iterative manual process used to design clinically acceptable treatments.

In this paper, we develop a novel automated treatment planning pipeline for oropharyngeal cancer that uses a GAN to predict 3D dose distributions. In contrast to previous machine learning methods, our approach does not require the pre-specification of an extensive set of feature variables for prediction. Instead, our model learns what features are important to produce clinically acceptable treatment plans. We apply our KBP methodology to a dataset consisting of 26,279 CT images from 217 patients with oropharyngeal cancer that have undergone radiation therapy. Approximately 60% of these images are used to train the GAN, which is used to predict high quality dose distributions for the remaining out-of-sample patients. These predictions are used as input into an optimization model to produce deliverable plans. We compare our approach to several other techniques, including three feature-based machine learning models and a standard convolutional neural network (CNN). We demonstrate that our approach outperforms all other models in achieving several clinically relevant criteria and in matching the clinical (benchmark) plans.

Technical Significance We demonstrate the first use of GANs for generating radiation treatment plans in cancer. We recast KBP prediction as an image colorization problem for which GANs are known to perform well. Moreover, we provide the first full pipeline comparison between different KBP prediction methods by optimizing the predicted dose distribution and comparing the final result to deliverable plans. We find that, in this setting, our GAN approach outperforms all other methods, including the latest in machine learning-based KBP approaches, in meeting clinical criteria.

Clinical Relevance Oropharyngeal cancer is one of the most difficult cancers to plan a treatment for, and as a result, generating deliverable treatment plans is particularly time consuming.[37] Our GAN approach automates the planning approach producing, on average, plans that are superior to clinical ones in several key metrics. Our site-independent method suggests similar performance for simpler sites, such as prostate and stomach cancers, while showing that high-quality oropharynx treatment plans can be automatically generated.

2.2 Related work

2.2.1 Knowledge-based planning

Many different approaches have been tested for the machine learning component of a KBP-driven automated planning pipeline. Query-based methods identify previously treated patients who are sufficiently similar to the new patient, and use the historically achieved dose metrics as predictions for the new patient.[107;108] Another common approach uses principal component analysis (PCA), in conjunction with linear regression, to predict dose metrics for new patients.[113;116] However, these well-established techniques only predict two-dimensional dose metrics. Recent research has shown that 3D dose distribution predictions can also be generated using random forest or neural network-based models.[76;82;95] Nevertheless, for many of these approaches to work effectively, significant effort must be spent in feature engineering, i.e., introducing features specific to the cancer site. Furthermore, some of these approaches compare the predicted dose distributions, rather than deliverable plans post-optimization, to the clinical plans.

For the optimization phase of KBP, there are two main approaches for turning predictions into treatments: dose mimicking[87] and inverse optimization.[30] The dose mimicking model minimizes the L_2 loss between the predicted dose distribution and one that satisfies all physical constraints. Alternatively, inverse optimization (IO) is a methodology that estimates parameters of an optimization problem from its observed solutions.[2] In the RT context, IO finds parameters, e.g., objective function weights, that allow a deliverable treatment plan to re-create the predicted dose distribution as closely as possible.[30] A key advantage of inverse optimization is that it better replicates the trade-offs implicit in clinical treatment plans.[29]

2.2.2 Generative adversarial networks

GANs are a well-studied class of deep learning algorithms used in *generative* modeling, i.e., in the creation of new data.[48] Although initially used to artificially generate 2D images, and later 3D models,[110] their success has garnered increasing interest for healthcare applications. GANs have been used for medical drug discovery,[60] generating artificial patient records,[32;43] the detection of brain lesions,[3] and image augmentation for improved liver lesion classification.[45]

A GAN consists of two neural networks, a generator and a discriminator, working in tandem. The generator $G(\cdot)$ takes an initial random input $\mathbf{z} \sim p_{\mathbf{z}}$ and attempts to generate an artificial data sample $\mathbf{x} = G(\mathbf{z})$ (i.e., the 3D dose distribution). The discriminator $D(\cdot)$ is a classifier that takes generated and real data samples, and tries to identify which is which, i.e., $D(\mathbf{x}) \in [0, 1]$ where 1 suggests the generated sample is satisfactory. The interaction between the networks can be formalized mathematically as a minimax game. If $\mathbf{x} \sim p_{\text{data}}$ is the probability distribution over the real data samples, then the game is defined as

$$\min_G \max_D \left\{ V(G, D) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [\log(1 - D(G(\mathbf{z})))] \right\}.$$

GANs have been proven effective in style transfer problems, where the generator input $\mathbf{z}$ is a data sample corresponding to one style (or characteristic) and the output $\mathbf{x}$ is a mapping to a different style.[57;115] For example, style transfer can be used to transform grayscale images to colored photos,[92] in facial recognition for surveillance-based law enforcement,[104] and in 3D reconstruction of damaged artifacts.[52] Here, the generator $G(\mathbf{z})$ learns the mapping between styles that generates samples resembling the ground truth. Since key structures in the output may be entangled with noise from the generator, the desired output is often achieved by modifying the original minimax game with a penalty term on large deviations between the real and generated samples:

$$\min_G \max_D \left\{ V(G, D) + \lambda \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}, \mathbf{z} \sim p_{\mathbf{z}}} [\|\mathbf{x} - G(\mathbf{z})\|_1] \right\}, \tag{2.1}$$

where λ is a regularizer that balances the trade-off between learning style and the real data.

2.3 Methods

We used contoured CT images and clinically acceptable dose distributions from the treatment plans of past oropharyngeal cancer patients to train a style transfer GAN. We then passed out-of-sample predicted dose distributions through an IO pipeline[7] to generate the final treatment plans. For baseline comparisons, we also implemented several methods from the literature using the complete pipeline. Figure 2.2 shows a high-level overview of this automated planning pipeline.

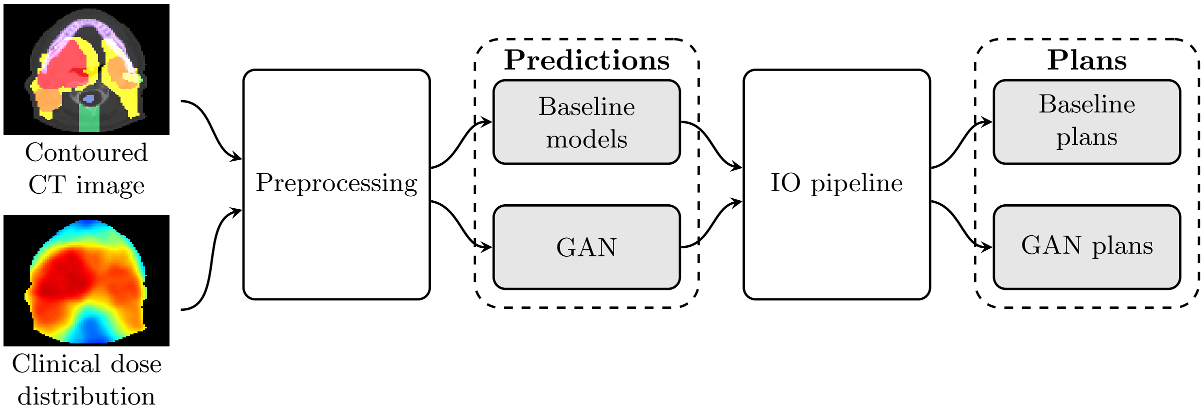


Figure 2.2: An schematic of our KBP-based automated planning pipeline.

2.3.1 Data

We obtained treatment plans from 217 oropharyngeal cancer patients treated at a single institution with 6 MV, step-and-shoot, intensity-modulated radiation therapy machine. All plans were for a prescription of 70 Gy, 63 Gy, and 56 Gy in 35 fractions to the gross disease, intermediate risk, and elective target volumes, respectively.

For each patient, we identified a set of targets and healthy organs-at-risk (OARs). Targets were denoted as planning target volumes (PTVs) along with the oncologist-prescribed dose (e.g., PTV70 corresponds the target with the highest dose prescription). OARs included the brainstem, spinal cord, right and left parotids, larynx, esophagus, and mandible. Every voxel (a 3D pixel of size 4 mm × 4 mm × 2 mm) of a CT image was classified by their clinically drawn contours. All voxels were assigned a structure-specific color, and in cases where the voxel was classified as both target and OAR, we reverted to target. All unclassified tissue was left as the original CT image grayscale.

2.3.2 GAN model

We first divided each 3D CT image into 2D slices of 128×128 pixels. The generator used a single CT image slice to predict the dose distribution along that same plane without considering the vertical relationship between different slices. This process was repeated for every slice until a full 3D dose distribution was produced. Our training set consisted of all 2D slices from the 3D CT images for 130 patients, totaling 15,657 images. The CT images from the remaining 87 patients were used for out-of-sample evaluation.

Our GAN learning model was built on the `pix2pix` style transfer architecture of.[57] We used a U-Net generator that passed a 2D contoured CT image slice through consecutive convolution layers, a bottleneck layer, and then through several deconvolution layers. The U-Net also employed skip connections, i.e., the output of each convolution layer was concatenated to the input of a corresponding deconvolution layer. This allowed the generator to easily pass "high dimensional" information (e.g., structural outlines) between the inputted CT image slice and the outputted dose slice. The discriminator passed a 2D slice of the dose distribution along several consecutive convolution layers, outputting a single scalar value. In the training phase, the discriminator received one real and one generated dose distribution before backpropagation. We disconnected the discriminator after training, at which point the generator only received a contoured CT slice. We refer the reader to Appendix A.1 for additional details regarding the network architectures.

We used the loss function given by Equation (2.1) with $\lambda = 90$, and trained using Adam,[62] with learning rate 0.0002 and $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for 25 epochs. We used the default Adam settings from,[57] as they were proven to be good for a variety of different style transfer problems. While we swept through various values for λ and the number of epochs, we found these default settings to be sufficient, with minimal subsequent improvement. We found it useful to stop training when the loss functions were roughly equal; if the loss from the l_1 penalty fell too low, the GAN began to simply memorize the dataset. The code for all experiments, along with the parameter settings is provided at https://github.com/ababier/gancer.

2.3.3 Plan generation

Predicted dose distributions were inputted into an IO pipeline to generate optimized plans. The IO model determined the weights of a parametric "forward" optimization model given a predicted dose distribution. The objective of the forward model was to minimize the sum of 65 objective functions: seven per OAR and three per target. Terms for the OARs included the mean dose, max dose, and the percentile (0.25, 0.50, 0.75, 0.90, and 0.975) above the maximum predicted dose to the OAR. Similarly, terms for the target included the maximum dose, average dose below prescription, and average dose above prescription. The complexity of the KBP-generated treatment plan was constrained to match the clinical treatment[35] where complexity represents a (convex) surrogate measure for the physical deliverability of a plan. We note that in reality, there are additional constraints in the IO pipeline that we omit for tractability. Thus, our notion of a deliverable plan does not include all physical constraints.

Physical parameters for the optimization model were derived from `A Computational Environment for Radiotherapy Research`.[38] To replicate the clinical plans, all KBP-generated plans were delivered from nine equidistant coplanar beams at angles 0°, 40°, . . . , 320°. We used `Gurobi 7.5` to solve the inverse and forward optimization problems associated with the IO pipeline. Additional details of the IO model can be found in Babier et al.[8]

2.3.4 Baseline approaches

We compared our GAN approach to generating predicted dose distributions with several state-of-the-art techniques. We briefly describe the baseline approaches here.

- **Bagging query (BQ):** A look-up method identifies patients with similar geometries who have undergone radiation therapy and outputs their doses as predictions. This approach predicts dose volume histograms (DVHs), i.e., 2D summaries of the

3D dose delivered to specific targets and OARs.[7]

- **Generalized PCA (gPCA):** A method combining PCA with linear regression using patient geometry features. Similar to BQ, this method also predicts DVHs.[7].

- **Random forest (RF):** Predicts dose to each voxel (3D dose prediction) using ten customized features based on patient geometry (based on McIntosh et al.[76]). Additional details can be found in Appendix A.2.

- **U-Net (CNN):** Predicts dose to each voxel in 2D slices from a CT image using a U-Net convolution neural network architecture (based on Nguyen et al.[82]).

All baseline predictions were fed into the same IO pipeline as the GAN approach to ensure a fair comparison between deliverable plans.

2.4 Results

2.4.1 Sample generated dose distributions

We observed that the style transfer function mapping the CT image to the predicted dose distribution appeared easy to learn. This is because the GAN generated dose distributions had the hallmarks of a deliverable plan, like the sharp dose gradients that are generated by individual beams. However, there were subtle deliverability characteristics that the GAN could not always identify. The optimization step enforced these physical deliverability constraints to correct for these idiosyncracies. This result can be observed in Figure 2.3, where five sample slices of a clinical, predicted, and optimized plan are presented.

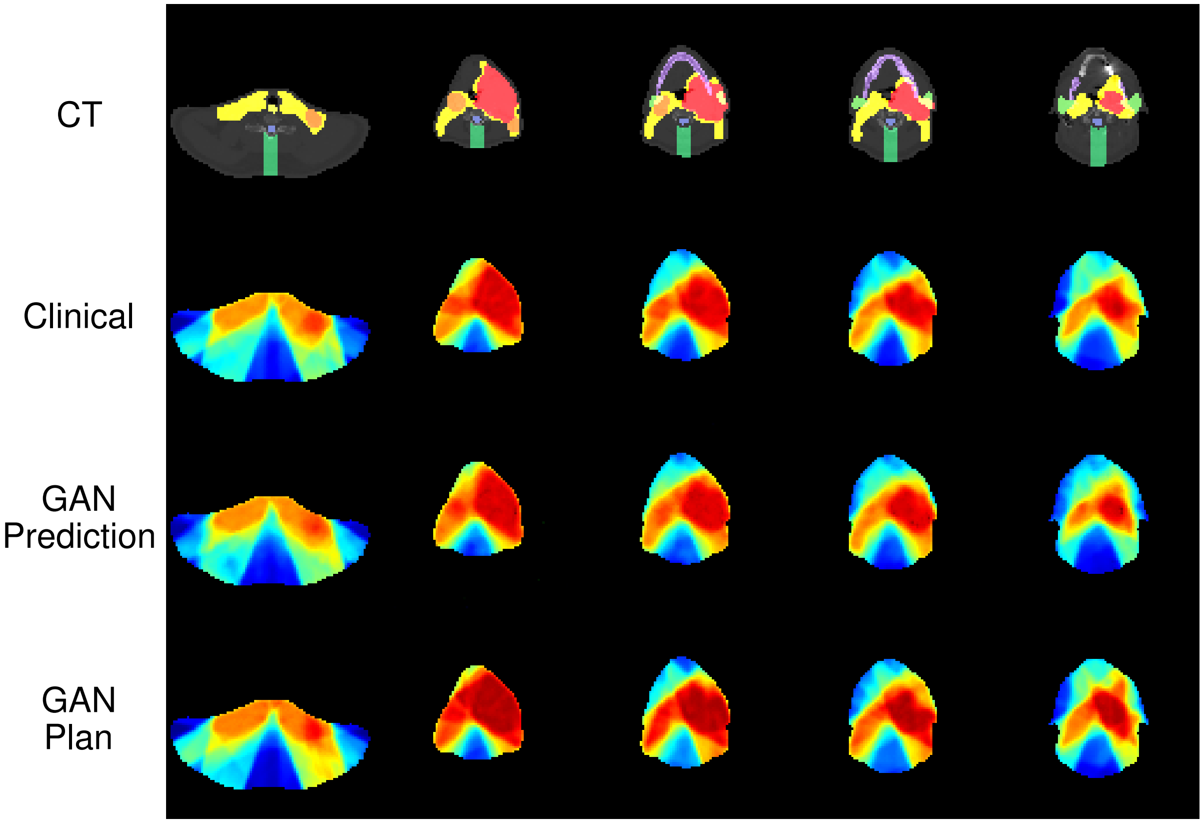


Figure 2.3: Sample of slices from a test patient. From top to bottom: contoured CT image (generator input), clinical plan (ground truth), GAN prediction, and GAN plan (post optimization).

2.4.2 Clinical criteria satisfaction

We measured plan quality by evaluating how frequently they satisfied the standard clinical criteria for oropharyngeal cancer treatment plans; see Table 2.1. Clinicians commonly use criteria satisfaction as a metric to evaluate plan quality and approve a treatment plan after it satisfies a sufficient number of the criteria. Thus, each criterion (one per OAR and target) was measured on a pass-fail basis depending on whether the mean dose $\mathcal{D}_{mean}$, maximum dose $\mathcal{D}_{max}$, or the dose to 99% of the volume of that structure $\mathcal{D}_{99}$, was above or below a given threshold. To facilitate the comparisons, we scaled the GAN and baseline treatment plans so that their PTV $\mathcal{D}_{99}$ was equal to the PTV $\mathcal{D}_{99}$ of the corresponding clinical plan.

Table 2.1: Clinical criteria used to evaluate all plans. $\mathcal{D}_{mean}$ refers to the mean dose, $\mathcal{D}_{max}$ the maximum dose, and $\mathcal{D}_{99}$ dose to 99% of the structure.

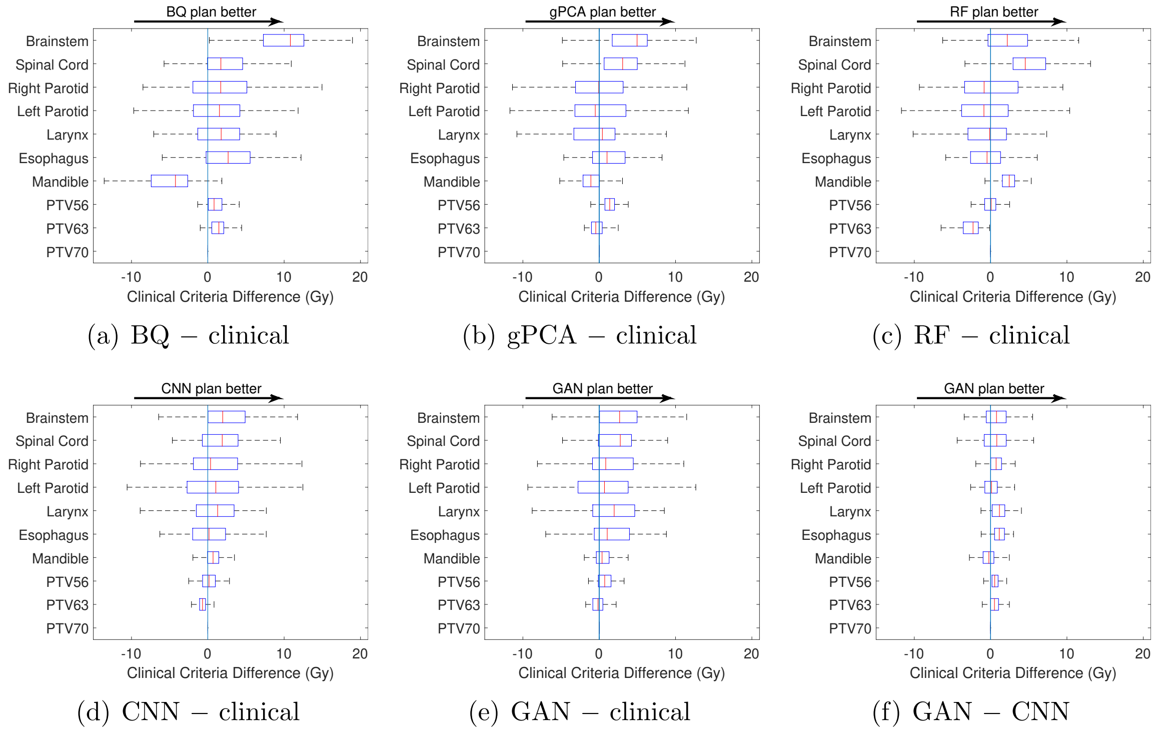
Structure	Criteria
Brainstem	$\mathcal{D}_{max} \leq 54$ Gy
Spinal Cord	$\mathcal{D}_{max} \leq 48$ Gy
Right Parotid	$\mathcal{D}_{mean} \leq 26$ Gy
Left Parotid	$\mathcal{D}_{mean} \leq 26$ Gy
Larynx	$\mathcal{D}_{mean} \leq 45$ Gy
Esophagus	$\mathcal{D}_{mean} \leq 45$ Gy
Mandible	$\mathcal{D}_{max} \leq 73.5$ Gy
PTV56	$\mathcal{D}_{99} \geq 53.2$ Gy
PTV63	$\mathcal{D}_{99} \geq 59.9$ Gy
PTV70	$\mathcal{D}_{99} \geq 66.5$ Gy

Table 2.2 presents the percentage of the GAN and baseline treatment plans that satisfied the clinical criteria. The clinically acceptable plans typically violated some criteria because of the proximity of the targets to the OARs and the complexity of the head-and-neck site in general. We observed that the BQ and gPCA plans tended to satisfy PTV criteria more frequently, which suggested that they may recommend delivering a higher dose to the target relative to the clinical plan. However, they failed to achieve mean and maximum dose criteria to the OARs (note: there are more than triple the number of OAR criteria as PTV criteria once all plans are normalized to $\mathcal{D}_{99}$ of the PTV70). On the other hand, the RF plans appeared to satisfy fewer clinical criteria associated with the target as compared to the clinical plans. The CNN plans achieved the closest level of performance to the clinical plans. However, the GAN plans had the best overall performance among all approaches. They offered a balanced trade-off between the OARs and targets, and even outperformed the clinical plans on clinical criteria satisfaction.

Table 2.2: Frequency of clinical criteria satisfaction.

	BQ	gPCA	RF	CNN	GAN	Clinical
OAR criteria	61.6%	65.8%	71.5%	72.5%	72.8%	72.0%
PTV criteria	83.5%	85.7%	68.0%	76.3%	81.3%	76.8%
All criteria	67.6%	71.2%	70.7%	73.6%	75.2%	73.3%

The previous results focused on pass-fail performance with respect to the clinical criteria. We also examined the magnitude of passing or failing via head-to-head comparisons of the GAN/baseline plans to the clinical plans, and between the GAN and CNN plans (see Figure 2.4). The x-axis in each figure is the difference in Gray (Gy) between the KBP and the clinical plans for the criterion on the corresponding y-axis. We found that for each criterion, the majority of GAN plans outperformed their clinical counterparts by several Gy (Figure 2.4 (e)). This is a significant result given that the clinical plans were heavily optimized and delivered to actual patients. The BQ, gPCA, and RF plans displayed substantial variability in performance when compared to the clinical plan. Consistent with Table 2.2, performance of the CNN plans were closest to the GAN plans although, as shown in Figure 2.4 (f), the GAN plans were slightly better.



(a) BQ − clinical

(b) gPCA − clinical

(c) RF − clinical

(d) CNN − clinical

(e) GAN − clinical

(f) GAN − CNN

Figure 2.4: Head-to-head comparisons: (a)–(e) the plans generated from each KBP model versus their clinical counterparts where positive difference implies the KBP plans were better; (f) the plans from the GAN versus the CNN. Upper and lower boundaries of each box represent the 75th and 25th percentiles respectively, and the vertical line in the box depicts the median. Whiskers extend to 1.5 times the interquartile range.

Finally, we compared the KBP plans against the clinical plans using the gamma passing rate (GPR) metric. GPR measures the similarity between two dose distributions on a voxel-by-voxel basis, computing for each voxel, a pass-fail test. We considered the standard choice of GPR, i.e., a 3%/3 mm tolerance,[68] which roughly means a voxel in the evaluated dose distribution (KBP) "passes" if there is at least one voxel in the reference dose distribution (clinical) within 3 mm that receives a dose that is within ±3% of the reference dose. Table 2.3 summarizes the average GPR achieved over all KBP-generated plans. A score of 1.0 means that every voxel has passed the criteria; in other words, the two dose distributions were considered identical (within the tolerance). Overall, we observed that the GAN plans generated dose distributions that most closely resembled the clinical dose distributions, followed by the CNN, and then the gPCA plans. Notably, the GAN dose distributions best resembled the clinical dose distribution around the target, which is of primary importance. The GAN plans performed less well on the OARs, but this result was expected given the results from Table 2.2, which indicated that the GAN plans achieved more OAR clinical criteria than the clinical plan (i.e., the GAN was able to deliver a lower dose to the OARs as compared to the clinical dose distribution).

Table 2.3: Average GPR for each population of KBP plans compared to clinical plans.

	BQ	gPCA	RF	CNN	GAN
All OARs	0.548	0.584	0.535	0.566	0.549
All PTVs	0.533	0.728	0.503	0.741	0.761
All Structures	0.536	0.669	0.518	0.670	0.675

2.5 Discussion and Future Work

In this paper, we proposed the first GAN-based KBP method to generate radiation therapy treatment plans. We trained our complete pipeline on 130 patients, tested on 87 out-of-sample patients diagnosed with oropharyngeal cancer, and compared our technique with several state-of-the-art planning methods including a query-based approach, a PCA-based method, a random forest, and a CNN. All methods were evaluated on standard clinical criteria for plan evaluation (i.e., OARs sparing and target coverage), showing that the GAN plans outperformed all baseline KBP methods. We also demonstrated that the GAN plans outperformed the clinical plans by satisfying additional criteria on OAR dose sparing and target dose coverage. Finally, we used the gamma passing rate, a standard metric in the radiation therapy literature, to evaluate the similarity of the full 3D dose distribution between the KBP and clinical plans demonstrating that the GAN plans were the most similar to clinical plans on average. Note that the performance of automated planning methods should be measured based on their ability to re-create clinical quality plans with minimal manual effort. Of course, if the auto-generated plans manage to improve upon the clinical plans, that would be even better.

Our approach eschews the classical paradigm of predicting low-dimensional representations, or engineering features, by training a generic neural network to learn desirable dose distributions. Specifically, the GAN recasts KBP prediction as an image colorization problem. Moreover, the GAN is trained by mimicking the iterative process between the dosimetrist and oncologist; the generator network acts as the dosimetrist by designing dose distributions while the discriminator acts as the oncologist by determining whether the plans are good or bad. The implication is that selecting the appropriate neural network architecture may be sufficient when creating an automated KBP pipeline that generates deliverable plans. Further, our approach does not add site-specific feature variables which suggests that the good performance we observe may not be limited to patients with oropharyngeal cancer. Finally, since the GAN plans improve upon the clinical plans, it may be useful to analyze the results to generate useful insights for practitioners.

This work has four major limitations. First, the GAN and U-Net dose prediction models only learn about 2D relationships in dose distributions. As a result, they are unable to learn that the dose delivered to adjacent slices in a patient image is often similar. A second limitation is that this analysis is limited to a single cancer site (i.e., oropharynx). Further studies are needed to explore how GANs perform on different cancer sites. By adding site labels, we expect that a GAN can learn from the augmented training set of different cancer sites to better develop plans for specific sites. Third, this approach requires contoured CT images, which are time consuming to generate via conventional processes. Future studies could address this limitation by incorporating an automated image segmentation model (i.e., models that generated contours without human intervention) into a preprocessing stage that contours unlabelled CT images. Lastly, these approaches have only been evaluated using summary statistics. Future studies are required to understand how well these statistics translate into true quality, which can only be evaluated via physician review.

Chapter 3

Dose prediction with 3D computer vision

In the previous chapter, we conducted a series of experiments on a dataset of oropharyngeal cancer patients to show that our generative adversarial network (GAN) dose prediction model outperforms previous models on several clinical metrics. In this chapter, we improve the GAN model that was developed in the previous chapter. Out improvements lead to the first knowledge-based planning (KBP) pipeline uses a 3D GAN to predict a complete 3D dose distribution (i.e., not a single slice of the distribution that are then stitched together). Additionally, we investigate the impact of multiplicatively scaling the predictions before optimization, such that the predicted dose distributions achieve all target clinical criteria before they are input into the optimization model. We evaluate the performance on our contributions using a large private data set of 217 oropharyngeal cancer treatment plans.

3.1 Introduction

The conventional radiation therapy treatment planning process consists of an iterative, back-and-forth procedure between a treatment planner and an oncologist. The duration of a single iteration, compounded by the number of iterations that may take place, means that it can take several days for a treatment plan to be completed.[37] Automated treatment planning systems have the potential to replace this conventional approach with a more efficient operational paradigm that reduces plan generation lead time.[88] Hospitals that adopt these techniques should also be better equipped to efficiently produce high-quality treatment plans for complicated sites[117] and serve the growing demand for radiation therapy.[6]

Knowledge-based planning is a two-step approach to automated radiation therapy treatment planning that first predicts a clinically acceptable dose before using optimization to convert the prediction into a deliverable plan.[8;44;75;109] The prediction component of the pipeline, referred to as a dose prediction model, typically uses a library of historical treatment plans to learn the dose characteristics of previously delivered plans. Accordingly, it is essential that this prediction model be accurate as the quality of the final plans strongly correlate with the quality of the predictions.[8]

Many KBP prediction approaches have been introduced to either predict a dose distribution or a dose volume histogram (DVH).[5;7;44;95;96;111–113;116] While the majority of these methods use hand-tailored or low dimensional features for prediction, recent advances in machine learning have spurred the development of KBP methods that predict full dose distributions using automatically generated high-dimensional features.[44;71;74] The most recent work in this space has focused on neural network-based KBP methods, which are trained on libraries of historical plans to predict dose for each axial slice separately (i.e., 2D KBP methods)[44;71;84] or all slices concurrently (i.e., 3D KBP methods).[61;83] Among the 2D methods, generative adversarial networks have been shown to perform the best[71] while among the 3D methods, DoseNet is considered state-of-the-art.[61] It remains an open question as to whether a combination of the two approaches, a 3D GAN, will achieve even better results.

In this paper, we develop the first 3D GAN-based KBP method, which takes as input a 3D CT image and predicts the full 3D dose distribution all at once. We embed our prediction model in a KBP pipeline for oropharynx treatment planning.[7] Similar to other 3D approaches,[61;83] our approach uses a patient's entire 3D CT image as input and learns to construct spatial features without human intervention. In doing so, it learns to produce the entire 3D dose distribution rather than separate 2D dose distributions for each axial slice. Further, we also investigate the effect of multiplicatively scaling dose predictions such that they satisfy all target criteria before using them as input to a fluence map optimization method. We also improve upon a previously developed 2D GAN approach[71] ("2D-RGB") by specializing the GAN to the radiation therapy context so that it predicts dose as a scalar value rather than a color representation (i.e., heat map).

We apply our models to a dataset of clinical radiation therapy plans for 217 patients with oropharyngeal cancer. Approximately 60% of these clinical plans are used to train our models, which are then used to predict the dose distribution for the remaining out-of-sample patients. Next, those predictions are used as input into a fluence map optimization pipeline to generate treatment plans.[8] We compare our models to two notable deep learning methods, DoseNet[61] and 2D-RGB,[71] and demonstrate that: (i) 3D GANs are better suited for KBP than previous state-of-the-art prediction methods; (ii) adjusting predictions via multiplicative scaling - such that the predictions satisfy all target criteria prior to fluence map optimization - generally improves the quality of the resulting KBP plans; and (iii) dose should be represented as a scalar value rather than an RGB heat map. Finally, we observe that good KBP predictions, after fluence map optimization, do not necessarily result in the best treatment plans. Thus, we recommend that future KBP research should report fluence map-optimized performance metrics.

3.2 Methods and Materials

We use contoured CT images and dose distributions from clinically accepted treatment plans to train all models in the KBP pipeline. Each KBP model was trained to predict the dose distribution given a contoured CT image. For testing, we passed out-of-sample contoured CT images through each of the models to generate dose distributions. These predictions were then passed through an optimization pipeline[8] to generate the final fluence-based treatment plans. Figure 3.1 shows a high-level overview of this automated planning pipeline.

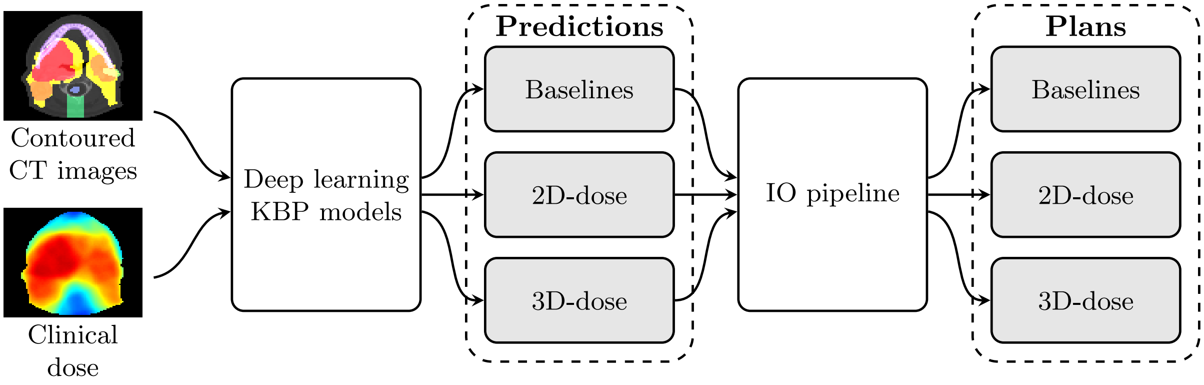


Figure 3.1: Overview of the knowledge-based automated treatment planning pipeline.

3.2.1 Prediction Using Generative Adversarial Networks

A GAN consists of two neural networks known as a *generator* and *discriminator*.[48] We focus on *conditional* GANs which, in addition to a Gaussian input, learn to generate different outputs conditioned on known problem-specific characteristics (e.g., CT images).[57] Specifically, let $\mathbf{z} \sim p_{\mathbf{z}}$ denote a sample from a Gaussian input. The generator network takes as input $\mathbf{z}$ and a CT image $\mathbf{c}$ and outputs a predicted dose distribution $\mathbf{x} = G(\mathbf{z}, \mathbf{c})$. The discriminator network then takes a CT image and the predicted dose distribution as input and outputs a "belief" regarding whether the dose distribution is the actual clinical dose (as opposed to artificially produced by the generator). That is, $D(\mathbf{x}, \mathbf{c}) \in [0, 1]$

where $D(\mathbf{x}, \mathbf{c}) = 1$ suggests the discriminator is confident the dose distribution is the clinically delivered dose. Both networks are trained iteratively with a single loss function $\mathcal{L}(D, G)$. Letting $\mathbf{x} \sim p_{\text{data}}$ denote the distribution of real delivered plans, we write the training problem as:

$$\min_{G} \max_{D} \ \mathcal{L}(D, G)$$

where

$$\mathcal{L}(D, G) = \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} \left[\log(1 - D(G(\mathbf{z}, \mathbf{c}), \mathbf{c})) \right] + \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \left[\log D(\mathbf{x}, \mathbf{c}) \right] + \lambda \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}, \mathbf{z} \sim p_{\mathbf{z}}} \left[\| \mathbf{x} - G(\mathbf{z}, \mathbf{c}) \|_1 \right].$$

The above formulation represents the objective for the most common class of conditional GANs used in problems associated with image generation.[57;115] By minimizing the first term in $\mathcal{L}(D, G)$, the generator learns to construct dose distributions such that $D(G(\mathbf{z}, \mathbf{c}), \mathbf{c}) = 1$. That is, the generator attempts to fool the discriminator into believing that the generated dose is a real clinical dose. The discriminator adversarially maximizes the second term in $\mathcal{L}(D, G)$ to output $D(G(\mathbf{z}, \mathbf{c}), \mathbf{c}) = 0$ for $\mathbf{z} \sim p_{\mathbf{z}}$ and $D(\mathbf{x}, \mathbf{c}) = 1$ for $\mathbf{x} \sim p_{\text{data}}$, i.e., the discriminator attempts to correctly distinguish between artificially generated versus clinically delivered plans. The final term in $\mathcal{L}(D, G)$ is an l_1 loss function which forces the generated samples to better resemble the ground truth dataset (i.e., the clinically delivered dose distribution). The hyperparameter λ balances the tradeoff between minimizing the GAN loss (first two terms) and having images resemble deliverable plans.

We constructed a generator and a discriminator network using the `pix2pix` architecture proposed in the canonical Style Transfer GAN.[57] The generator possesses a U-net architecture that passes a contoured CT image through consecutive convolution layers, a bottleneck layer, and then several deconvolution layers. The U-net employs skip connections, i.e., the output of each convolution layer is concatenated to the input of a corresponding deconvolution layer. This allows the generator to easily pass "high-dimensional" information (e.g., structural outlines) between the inputted CT image and the outputted dose. The discriminator takes as input a dose distribution and a CT image and passes them through several consecutive convolution layers until outputting a single scalar value between zero and one. We refer the reader to the original Style Transfer GAN work for full details on the number and size of the convolution and deconvolution layers in the `pix2pix` architecture.[57]

Two GAN models were created; 3D-dose represents our main contribution while 2D-dose represents a credible benchmark and allows us to ascertain the impact of separate versus simultaneous axial slice prediction. That is, the "2D" and "3D" designation refers to whether dose is being predicted for each 2D slice independently (i.e., 128×128 pixel images) or the full 3D distribution (i.e., $128 \times 128 \times 128$ voxel images). The generator and discriminator architectures for these GAN models are summarized in Tables VII and VIII in the Appendix B, respectively. Both models output dose predictions as a single scalar value rather than a color image. Additionally, they share the same general architecture (e.g., number of layers and filter sizes) except that 2D-dose uses two-dimensional convolution and deconvolution filters (i.e., 4×4 kernels) whereas 3D-dose uses three-dimensional filters (i.e., $4 \times 4 \times 4$ kernels).[52] In order to construct a 3D dose distribution using the 2D-dose GAN, we concatenated all outputted axial slices for each patient.

We compared the GAN models to two benchmarks. The first is DoseNet, which is a state-of-the-art convolutional neural network model with residual network blocks for predicting dose.[61] Like 3D-dose, it simultaneously predicts the entire dose distribution, but unlike 3D-dose, it is trained with only a generator network. Note that because code for DoseNet is not publicly available, our implementation recreates the original model after corresponding with the authors. The second benchmark is 2D-RGB GAN, which is the state-of-the-art in GAN-based dose prediction.[71] This model is identical to the original pix2pix network, and therefore, shares a similar architecture as 2D-dose. However, it outputs a 128×128 pixel image with three channels (RGB) for color. To convert this to a dose distribution, we simply mapped elements in the color vector to their corresponding scalar dose values.

The generator and discriminator networks in our models were trained iteratively using gradient descent. After training was complete, the discriminator was disconnected and the generator was used on out-of-sample CT images. Figure 3.2 summarizes the difference between how the GANs were trained and tested. In our experiments, we used the loss function given by $\mathcal{L}(D, G)$ with $\lambda = 90$, and trained the networks using the Adam optimizer[62] with learning rate $\alpha = 0.0002$ and $\beta_1 = 0.5$ and $\beta_2 = 0.999$. These hyperparameters are the default Adam settings and have been used in many style transfer problems.[57]

Finding the optimal hyperparameters for KBP is difficult because we primarily evaluate the quality of the final plans (which involves solving a computationally expensive fluence map optimization problem) rather than the KBP predictions. While we use a consistent set of hyperparameters for all models where possible (e.g., optimizer parameter settings and regularizer weight), we varied the batch size and number of epochs used to train each model. Training was performed on a single Nvidia 1080 Ti GPU with 12 GB RAM and we set the batch size of each model as high as possible to fill the memory. This ensured that all models were trained with the same computational resources. Moreover, we varied the number of epochs for each model in order to prevent overfitting. We stopped training each model when the regularized l_1 loss function, $\lambda \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}, \mathbf{z} \sim p_{\mathbf{z}}} \left[\|\mathbf{x} - G(\mathbf{z}, \mathbf{c})\|_1 \right]$, fell to 0.5 for the in-sample training data.[71] Intuitively, if the l_1 loss falls too low (e.g., below the adversarial loss), the GAN begins to overfit. This is particularly dangerous due to the relatively small (within the deep learning context) dataset used for training. To further validate our stopping rule, we plotted the training and testing loss up to 200 epochs for all models (see Figure 3.5). In all cases, a training loss of 0.5 was roughly the point at which the out-of-sample loss also reached a steady state. The code for all experiments with the parameter settings is provided at http://github.com/ababier/gancer.

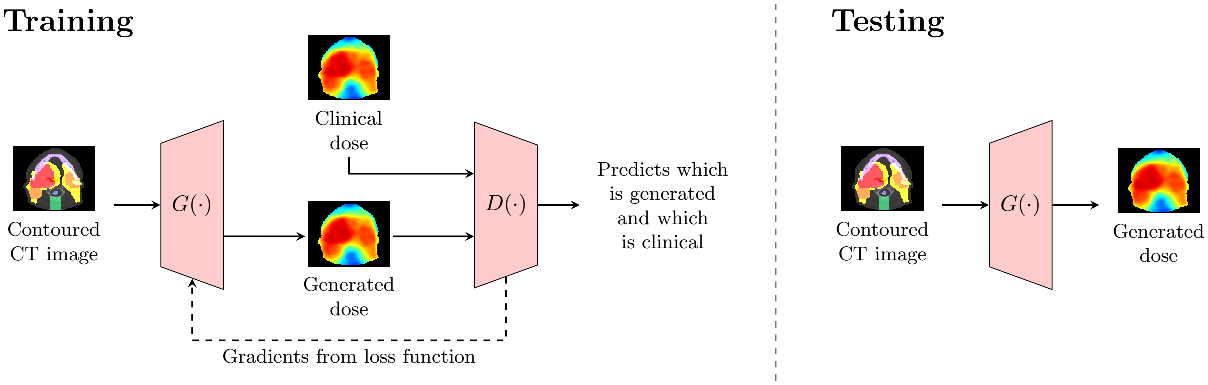


Figure 3.2: Overview of the GAN training and testing phases.

3.2.2 Training the GAN

We obtained plans for 217 oropharyngeal cancer treatments delivered at a single institution with 6 MV, step-and-shoot, intensity-modulated radiation therapy. All plans were prescribed 70 Gy and 56 Gy in 35 fractions to the gross disease (PTV70) and elective target volumes (PTV56), respectively; in 130 plans there was also a prescription of 63 Gy to the intermediate risk target volume (PTV63). The organs-at-risk (OARs) included the brainstem, spinal cord, right parotid, left parotid, larynx, esophagus, mandible, and the limPostNeck, which is an artificial structure used to spare the posterior neck. The geometry of each patient was discretized into voxels of size 4 mm × 4 mm × 2 mm.

The CT images and dose distributions for all 217 treatment plans were converted into a suitable format for use by the neural networks. The CT images were reconstructed so that each voxel had RGB channels, which were assigned values according to Table 3.1, and converted into 128 axial slices of 128×128 voxels. The images were separated into a training set of 130 random samples (a total of 16,640 pairs of CT image slices and dose distributions) and a testing set of the remaining 87 samples for out-of-sample evaluation.

Table 3.1: Colors assigned to each voxel. Voxels that were classified as both OAR and target were assigned nonzero green and red channel values, respectively.

Structure	Red Channel	Green Channel	Blue Channel
Brainstem	0	125	CT Grayscale
Spinal Cord	0	147	CT Grayscale
Right Parotid	0	190	CT Grayscale
Left Parotid	0	190	CT Grayscale
Larynx	0	233	CT Grayscale
Esophagus	0	212	CT Grayscale
Mandible	0	168	CT Grayscale
limPostNeck	0	255	CT Grayscale
PTV70	255	0	CT Grayscale
PTV63	205	0	CT Grayscale
PTV56	155	0	CT Grayscale
Unclassified	0	0	CT Grayscale
Empty Space	0	0	0

3.2.3 Creating Plans Using Inverse Planning

During out-of-sample testing, predictions produced by the generator were used as input into a fluence map optimization model with two stages. In the first stage, given a predicted dose distribution, the objective weights for a standard inverse planning model were estimated using a parameter estimation technique that has been previously validated in oropharynx.[8] In the second stage, the estimated weights were used in an inverse planning optimization model to generate treatment plans. The objective minimized the sum of 65 functions: seven per OAR and three per target. In our experiments, objectives for the OARs included the mean dose, max dose, and the average dose above 0.25, 0.50, 0.75, 0.90, and 0.975 of the maximum predicted dose to the OAR. Objectives for the target included the maximum dose, average dose below prescription, and average dose above prescription. The complexity of all generated treatment plans was constrained to a sum-of-positive-gradients (SPG) value of 55.[35] SPG was used since it is a convex surrogate for the physical deliverability of a plan and the parameter 55 was chosen as it is two standard deviations above the average SPG.[7] The dose influence matrices required for the optimization model were derived with `A Computational Environment for Radiotherapy Research`.[38] Each of the KBP-generated plans were delivered from nine equidistant coplanar beams at angles 0°, 40°, . . . , 320°. We used `Gurobi 7.5` to solve the optimization model.

We also generated fluence map-optimized treatment plans using a scaling procedure that multiplicatively increased the entire predicted dose distribution by the smallest amount to satisfy all target dose criteria. The scaled predictions were then input into the optimization model. Note that multiplicative scaling does not affect the fairness of the analysis because the final KBP plans must satisfy the same constraints (e.g., SPG) as plans generated from the unscaled predictions. To study the full effect of the scaling step, we generated four additional populations of *scaled* final plans alongside the initial four *unscaled* plans corresponding to 2D-RGB, DoseNet, 2D-dose, and 3D-dose. We refer to the four scaled KBP plans as 2D-RGB′, DoseNet′, 2D-dose′, and 3D-dose′ respectively. In these plans, the median scaling factor was 1.00 for the 2D-RGB' predictions, 1.02 for both the DoseNet′ and 2D-dose′ predictions, and 1.05 for the 3D-dose′ predictions.

3.2.4 Performance Analysis

We conducted two primary analyses. First, we evaluated the quality of the KBP plans by computing the fraction of clinical planning criteria that were satisfied. We compared these results against the performance of the clinical plans. Second, we evaluated the quality of the KBP predictions by comparing the predicted dose distributions to clinical dose distributions.

KBP Plan Quality: The quality of the final KBP plans was measured by evaluating how often they satisfied the clinical criteria presented in Table 3.2. For each class of clinical criteria, i.e., OARs, targets, and all regions-of-interest (ROIs), which includes both OARs and targets, we generated confusion matrices to compare the KBP plans with the clinical plans. We used a one-sided binomial test to determine whether 3D-dose′

plans passed the same (null hypothesis) or a greater proportion (alternative hypothesis) of criteria, which were also satisfied by the clinical plans, than the other automated plans; an analogous test was used over all the criteria that the clinical plans failed.

We then analyzed the organ-specific criteria that was achieved by each clinical plan to determine whether the corresponding KBP plan also passed that criterion. We again used a one-sided binomial test to determine whether the proportion of 3D-dose$'$ plans satisfying the same criteria as their clinical counterparts across all ROIs was the same (null hypothesis) or greater (alternative hypothesis) than the other automated plans. For these and all subsequent hypothesis tests, $p < 0.05$ was considered significant.

Table 3.2: The planning criteria used for evaluation: $\mathcal{D}_{99}$ is the dose to a fractional volume of 0.99, $\mathcal{D}_{mean}$ is the mean dose to a structure, and $\mathcal{D}_{max}$ is the max dose to a structure.

Structure	Criteria
Brainstem	$\mathcal{D}_{max} \leq 54$ Gy
Spinal Cord	$\mathcal{D}_{max} \leq 48$ Gy
Right Parotid	$\mathcal{D}_{mean} \leq 26$ Gy
Left Parotid	$\mathcal{D}_{mean} \leq 26$ Gy
Larynx	$\mathcal{D}_{mean} \leq 45$ Gy
Esophagus	$\mathcal{D}_{mean} \leq 45$ Gy
Mandible	$\mathcal{D}_{max} \leq 73.5$ Gy
PTV70	$\mathcal{D}_{99} \geq 66.5$ Gy
PTV63	$\mathcal{D}_{99} \geq 59.9$ Gy
PTV56	$\mathcal{D}_{99} \geq 53.2$ Gy

KBP Prediction Quality: Although KBP plan quality is the ideal metric for evaluating these models, we also measured KBP prediction quality to determine whether better predictions lead to better final treatment plans. Specifically, we evaluated how similar the KBP predictions were to their corresponding clinical plans. For every ROI and every patient, we calculated the average absolute error between the KBP predicted DVH and the clinical plan DVH, which were plotted in a box plot. We then used a one-sided

Mann-Whitney U test to determine whether the 3D-dose′ predictions had the same (null hypothesis) or greater absolute error (alternative hypothesis) than the predictions from the other KBP models. Finally, we plotted the predicted dose distributions to detect visual differences between the predictions of each KBP model. We also plotted the loss function of the training and testing sets to retrospectively validate our stopping rule for model training.

3.3 Results

KBP Plan Quality

In Table 3.3, we present confusion matrices comparing the quality of the final KBP plans with the clinical plans. The rows represent KBP performance using each of the eight KBP approaches while the columns indicate the clinical plans and the performance targets. Overall, 3D-dose′ plans best replicated the performance of the clinical plans since they agreed most on what criteria passed and failed (i.e., Pass/Pass and Fail/Fail). Where they differed, 3D-dose′ plans satisfied five times as many criteria (Pass/Fail) as the clinical plans (Fail/Pass). We also observed that scaling made a substantial difference as scaled plans outperformed their unscaled counterparts. For example, scaled 3D plans satisfied 99.5% of all target criteria whereas only 52.3% were satisfied for unscaled 3D plans. Finally, 2D-dose′ and 3D-dose′ performed the best satisfying 77.0% and 76.6% of all ROI criteria (i.e., the sum of the appropriate “Pass” rows), respectively.

Table 3.3: For each KBP approach, the percentage of clinical criteria that passed and failed as compared to the corresponding clinical plans. The highest percentage of KBP plan criteria that were also satisfied by the clinical plans are bolded in each column. p-values from the one-sided binomial test are presented in the last two columns, which compare the corresponding plans to 3D-dose′ plans.

		Clinical plans						p-value	
		OARs		Targets		All ROIs		All ROIs	
		Pass	Fail	Pass	Fail	Pass	Fail	Pass	Fail
2D-RGB	Pass	63.4	6.2	45.5	9.8	58.5	7.2	< 0.001	< 0.001
	Fail	3.2	27.2	23.2	21.4	8.7	25.6		
DoseNet	Pass	63.6	5.3	46.9	8.0	60.3	8.1	< 0.001	< 0.001
	Fail	1.4	29.6	21.9	23.2	6.9	24.7		
2D-dose	Pass	64.9	7.9	46.9	9.4	60.0	8.3	< 0.001	< 0.001
	Fail	1.7	25.5	21.9	21.9	7.2	24.5		
3D-dose	Pass	65.8	7.9	43.8	8.5	59.7	8.1	< 0.001	< 0.001
	Fail	0.8	25.5	25.0	22.8	7.5	24.7		
2D-RGB′	Pass	60.5	4.7	60.3	21.4	60.5	9.3	< 0.001	< 0.001
	Fail	6.1	28.7	8.5	9.8	6.7	23.5		
DoseNet′	Pass	62.5	4.3	65.6	25.4	63.9	12.1	0.115	0.708
	Fail	2.6	30.6	3.1	5.8	3.3	20.7		
2D-dose′	Pass	63.1	5.9	67.9	30.4	64.4	12.6	0.315	0.971
	Fail	3.5	27.5	0.9	0.9	2.8	20.2		
3D-dose′	Pass	63.4	4.6	68.3	31.2	64.7	11.9	—	—
	Fail	3.2	28.8	0.4	0.0	2.4	20.9		

Table 3.4 summarizes the performance of the KBP plans, focusing only on the criteria that the corresponding clinical plans also passed. The top ten rows mark satisfaction for each individual criteria. That is, if a clinical plan satisfied a certain criteria, whether the KBP plan also satisfied that same criteria. Both 3D-dose and 3D-dose′ performed the best across all OAR and target criteria. In particular, 3D-dose′ achieved the highest passing rate for every single target criteria. It also achieved the highest passing rate for every OAR criteria except the larynx and mandible. However, for some regions such as

the brainstem, esophagus, and PTV63, multiple KBP approaches also yielded the top result.

Table 3.4: The percentage of KBP plans that satisfied the same clinical criteria as the clinical plans. The rows under the heading "All" rows summarize the percentage of KBP plans that satisfied all clinical criteria that were satisfied by the clinical plans. p-values for the binomial tests using "All ROIs" are presented in the final row, which compare the corresponding plans to 3D-dose′ plans. The highest percentage of satisfied criteria is bolded in each row.

	Unscaled				Scaled			
	2D-RGB	DoseNet	2D-dose	3D-dose	2D-RGB′	DoseNet′	2D-dose′	3D-dose′
OARs								
Brainstem	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Spinal Cord	**100.0**	**100.0**	**100.0**	**100.0**	98.9	**100.0**	98.9	**100.0**
Right Parotid	**94.1**	88.2	88.2	**94.1**	**94.1**	82.4	88.2	**94.1**
Left Parotid	63.6	**81.8**	**81.8**	**81.8**	54.5	72.7	**81.8**	**81.8**
Larynx	91.8	93.9	89.8	**98.8**	87.8	85.7	87.8	91.8
Esophagus	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Mandible	84.8	**100.0**	98.5	98.5	65.2	89.4	84.8	81.8
Targets								
PTV70	48.3	45.7	82.8	79.3	75.9	97.8	**100.0**	**100.0**
PTV63	**100.0**	96.0	**100.0**	94.0	**100.0**	**100.0**	**100.0**	**100.0**
PTV56	52.2	62.1	15.2	10.9	89.1	89.7	95.7	**97.8**
Totals								
All OARs	80.5	92.0	88.5	**94.3**	62.1	79.3	78.2	79.3
All Targets	51.7	54.0	52.9	47.1	78.2	92.0	97.7	**98.9**
All ROIs	42.5	49.4	47.1	43.7	48.3	72.4	75.9	**78.2**
p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.166	0.387	—

In Table 3.4, we also compare the KBP plans to clinical plans over groups of criteria. Here, we identified all OARs, Targets, and ROIs and recorded a pass for a given KBP plan only if it met all of the identified criteria, respectively. We found that plans generated from the scaled predictions performed better than their unscaled counterparts in terms of satisfying all ROI criteria; we observed the biggest improvement between 3D-dose and 3D-dose′ (34.5 percentage points). Like all scaled plans, the improvement of 3D-dose′, which satisfied the most target criteria (98.9%), was the result of much better target coverage at the expense of less OAR sparing as compared to the 3D-dose plans, which satisfied the most OAR criteria (94.3%). Dose encoded as a grayscale image led to an increase in plan quality best shown by the difference of 27.6% separating 2D-RGB′ and 2D-dose′. Finally, it was clear that the 3D GAN architecture resulted in large improvements in prediction quality. That is, 3D-dose′ satisfied criteria more frequently (by 2.3% points) than 2D-dose′ and outperformed DoseNet′ (by 5.8% points) as it delivered more dose to the target without sacrificing OARs. Overall, 3D-dose′ achieved the same criteria as the clinical plans in 78.2% of cases, which was more than any other approach. The last row in Table 3.4 demonstrates that the proportion of 3D-dose′ plans that satisfied the same criteria as the corresponding clinical plans was significantly greater ($p < 0.05$) than five of the seven alternative algorithms; we cannot reject the hypothesis that 3D-dose′ produced comparable plans to DoseNet′ and 2D-dose′.

KBP Prediction Quality

In Figure 3.3, we present the distribution of average absolute DVH differences between the predicted and clinical dose over the regions of interest, i.e., the absolute error between the KBP predictions and clinical plans. The DoseNet predictions had a lower median dose error (2.2 Gy) than any of the other predictions and dominated all prediction methods with the lowest median, 25^{th} percentile, and 75^{th} percentile error. Each of the alternative prediction models had significantly lower error ($p < 0.05$) than the 3D-dose′ predictions, which had the highest median error of 3.3 Gy.

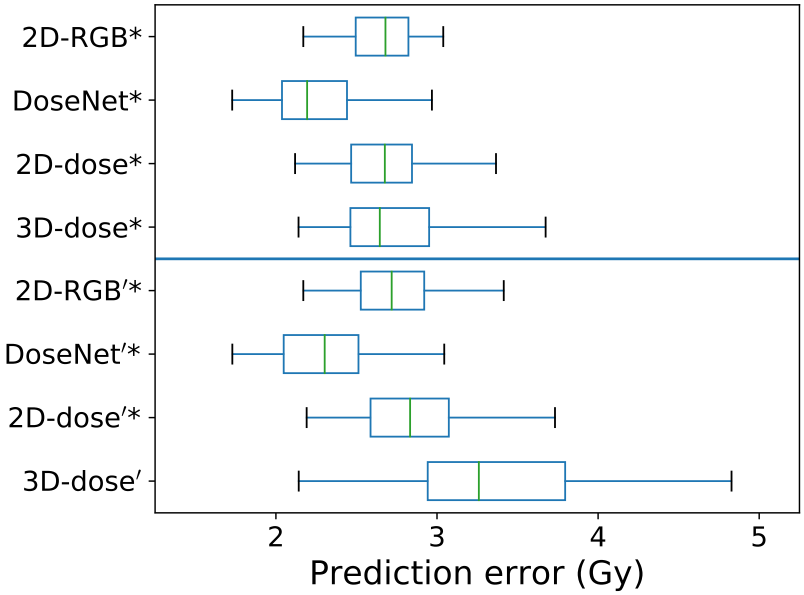


Figure 3.3: The distribution of average DVH differences between KBP prediction and clinical dose over all ROIs. The boxes indicate median and interquartile range (IQR). Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier. Asterisks denote predictions with significantly lower error ($p < 0.001$) than the 3D-dose′.

Figure 3.4 presents the predicted dose distributions corresponding to the 2D-RGB, DoseNet, 2D-dose, and 3D-dose KBP methods. Note that the 3D and 2D KBP models predicted distinct dose distributions that differed along the vertical (i.e., longitudinal) axis. In particular, 3D-dose and DoseNet better learned the vertical relationship between adjacent axial slices and thus, generated more natural dose predictions across the longitudinal axis. In contrast, the 2D KBP predictions included more "streaky" and unrealistic discontinuities in the dose distribution, particularly around axial slices that were adjacent to the target boundaries. For example, the dose falloff was impossibly steep in the plans generated using 2D predictions (Figure 3.4(e)), while plans generated using 3D predictions had more realistic dose gradients (Figure 3.4(f)). Lastly, we note that DoseNet predicted visually smoother dose distributions than the other models. This is because it uses an l_2 loss function (in contrast to an l_1 loss used by the other models) which in computer vision is known to generate blurry images[57].

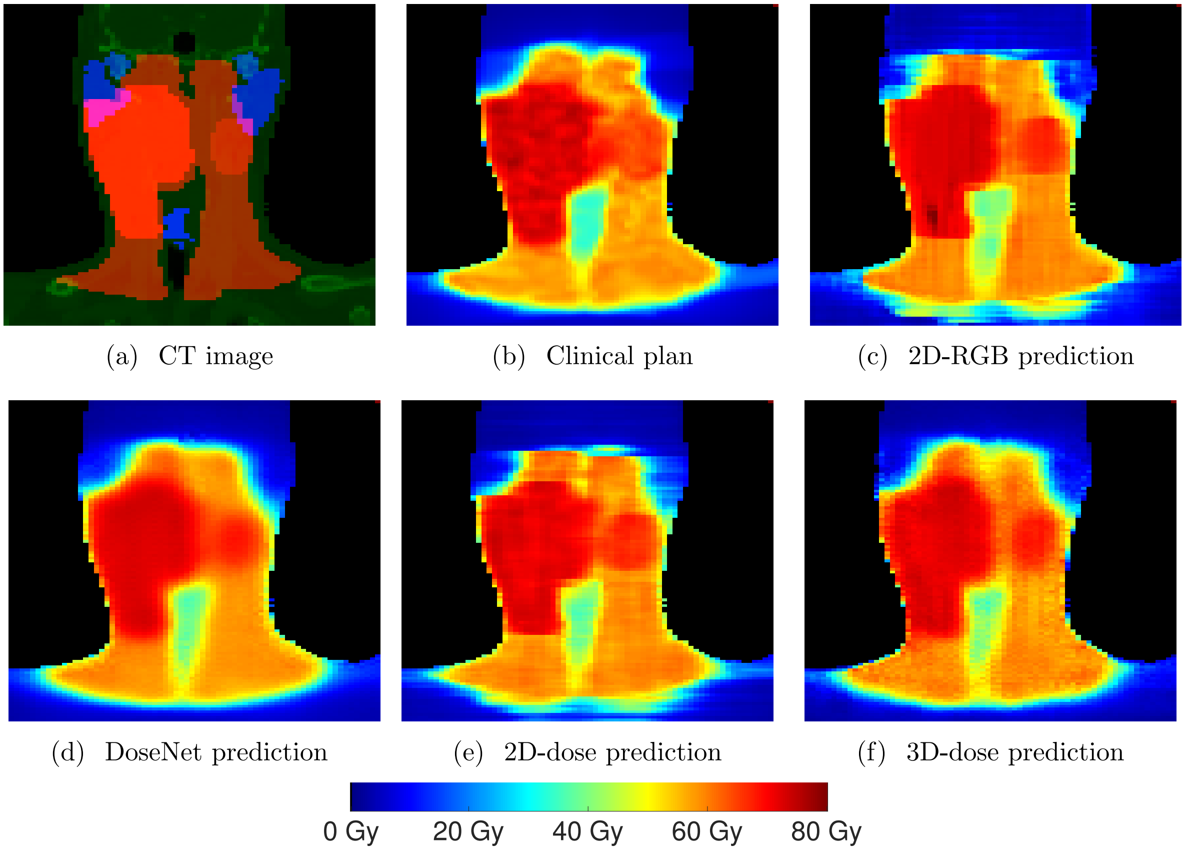
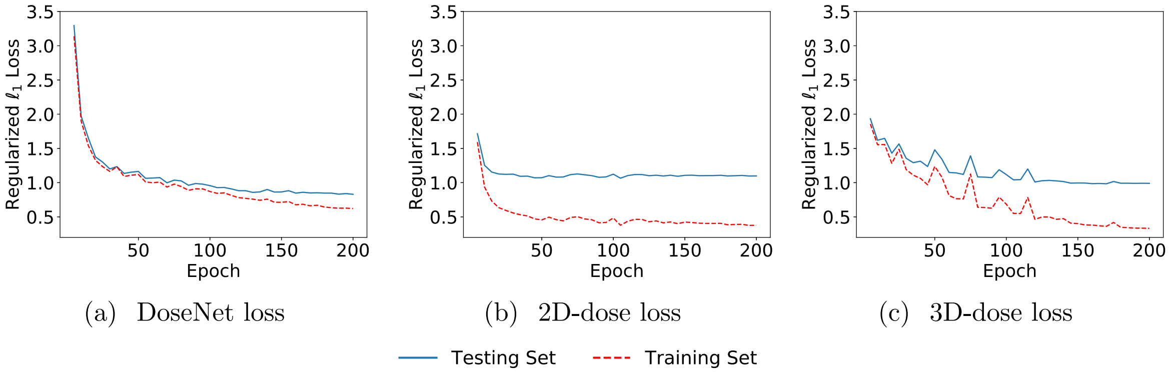

(a) CT image
(b) Clinical plan
(c) 2D-RGB prediction
(d) DoseNet prediction
(e) 2D-dose prediction
(f) 3D-dose prediction

Figure 3.4: The dose distributions for a sample patient over a single CT image (a) of their clinical plan (b), 2D-RGB prediction (c), DoseNet prediction (d), 2D-dose prediction (e), and 3D-dose prediction (f).

Finally, we reached a training loss of 0.5 at 50 epochs for 2D-dose and 2D-RGB, 120 epochs for 3D-dose, and 200 epochs for DoseNet. To validate these training duration decisions, which were based on when training loss fell to 0.5, we performed a retrospective analysis of the loss function on our training and testing sets at different epochs. Figure 3.5 shows that our stopping rule approximately identified the earliest point where the training loss continued to decrease but the testing set loss remained fixed, suggesting that beyond this point the model is simply overfitting to the training data. We emphasize that we did not base the training duration decision on out-of-sample performance.



(a) DoseNet loss (b) 2D-dose loss (c) 3D-dose loss

Figure 3.5: The regularized l_1 loss of the training and testing set.

3.4 Discussion

Although historically, KBP methods have predicted DVHs using hand-tailored features, there is widespread interest in automatically predicting dose distributions.[44;61;71;74] In this paper, we extend the literature by building a KBP pipeline that automatically generates treatment plans from CT images. The pipeline consists of two major components: prediction and optimization. The prediction stage uses a generative adversarial network to predict dose distributions from a CT image. The optimization stage consists of two optimization models, a parameter estimation model that learns objective function weights from the predicted dose distribution and an inverse planning model that produces the final fluence-based treatment plans. We demonstrate that our new GAN model produces treatment plans that are better than those produced by previous state-of-the-art KBP methods.[61;71]

Our framework includes two major enhancements that improved performance: 1) predicting the full 3D dose distribution from 3D CT images using GANs, and 2) multiplicatively scaling the KBP predictions prior to optimization.

Predicting the full 3D dose distribution from 3D CT images using GANs. Whereas the previous GAN KBP method predicted dose to each axial slice independently and then stitched the predictions together to form the full 3D dose distribution,[44;71;84;95] our 3D GAN was designed to take an entire contoured 3D CT image as input and generate the corresponding 3D dose distribution as output. Similar to previous 3D KBP methods,[61;83] the 3D GANs better learned the vertical relationship between adjacent axial slices than its 2D counterparts. Indeed, we observed that the 3D GAN predicted more realistic dose distributions than the 2D GAN (e.g., Figure 3.4) with smoother dose gradients across the longitudinal axis. Lastly, we note that 2D-dose′ outperformed 3D-dose′ on total clinical criteria but the 3D-dose′ plans achieved the same criteria as the clinical plans more often than the 2D-dose′ plans. We conjecture that some criteria may need to fail in order to pass the criteria achieved by the clinical plans. The 3D-dose′ plans make the appropriate trade-offs; they replicate the clinical plan trade-offs more often than the 2D-dose′ plans which results in a small reduction on overall clinical criteria satisfaction.

Multiplicatively scaling the predictions before optimization: Multiplicative scaling greatly enhanced the final KBP plan quality. Scaled KBP plans satisfied the same criteria as clinical plans 66% more often than unscaled plans; scaled plans also satisfied 11% more criteria than the unscaled plans overall. The idea of scaling predictions prior to optimization is novel in the KBP literature and is a general tool that can be applied to other KBP methods. There are three points worth highlighting. First, scaling is done automatically, just like how our deep learning approach automates high-dimensional feature selection. Thus, our KBP pipeline remains automated. Second, we believe that scaling works because it corrects small inaccuracies that may arise when learning the absolute dose delivered. That is, the GAN seems to be more effective at learning how dose varies among different tissues rather than learning the exact dose that should be delivered to a tissue (otherwise, scaling would not make a difference). Third, the scaled plans achieve better target performance at the expense of OARs, and we suspect that target performance was easier to achieve by delivering more dose to certain OARs (e.g., larynx, mandible) as compared to their unscaled counterparts.

Finally, a minor enhancement was the representation of dose to each voxel as a scalar, instead of as an RGB triple. We hypothesized that predicting dose encoded in single value (i.e., grayscale) would be easier than predicting dose encoded as a 3-channel RGB value and our experiments confirmed this result. In particular, 2D-dose′ plans satisfied the same criteria as the clinical plans 74% more often across all ROIs as compared to 2D-RGB′ plans. The improvement due to this modification is likely because it is much easier to predict a single value rather than a triplet. Although predicting dose rather than an image is intuitive from a medical physics perspective, it reflects the fact that computer vision techniques, while useful in non-imaging applications, need to be appropriately modified.

In our experiments, we used clinical criteria as the primary performance measure to evaluate the plan quality of several deep learning architectures. Since it is generally impossible to develop plans that simultaneously achieve all clinical criteria—in our dataset of 217 clinically delivered plans, only 68.4% of criteria were achieved—our primary goal was to achieve as many criteria as possible. We were also interested in generating plans that met the same criteria that the original clinical plans achieved; presumably, these represent the criteria that clinicians originally believed to be the most important. We believe that an automated planning method that produces dose distributions that satisfy the same criteria as treatment plans that have already been delivered is more likely to be clinically implemented. Our best plans across all clinical criteria metrics were 3D-dose′ plans, which were significantly better than the plans generated from prediction methods in the literature (i.e., 2D-RGB[71] and DoseNet[61]).

Surprisingly, we observe that good KBP predictions with low error (e.g., DoseNet) do not necessarily lead to the treatment plans with the best performance on clinical criteria. In particular, five of the seven alternative prediction methods had significantly lower error than the 3D-dose′ predictions, yet they also produced plans with significantly worse clinical criteria satisfaction (i.e., 2D-RGB, DoseNet, 2D-dose, 3D-dose, and 2D-RGB′). We conjecture that this phenomenon results from predictions with errors that negatively impact the optimization stage but are masked by standard summary statistics. For example, some prediction models may make several unrealistic tradeoffs (e.g., sparing an OAR inside the target while also achieving perfect target coverage) that the optimization method cannot easily correct. This may be less of a problem for predictions from the 3D-dose′ model because it was trained with a discriminator, which encourages that model to make predictions that resemble a full 3D clinical dose distribution (e.g., realistic tradeoffs, no unrealistic artifacts) at the expense of slightly less accurate DVH metrics. In contrast, DoseNet is trained only with a mean-squared-error loss function, which encourages the model to predict a good average dose distribution rather than a plausible one because there is no discriminator to detect unrealistic artifacts. This result also suggests that prediction DVH error does not necessarily correlate well with clinical measures of KBP plan quality, such as clinical criteria satisfaction. Furthermore, it highlights the importance of reporting fluence map-optimized performance metrics.

There are several reasons why we believe GANs are a good choice for KBP. First, they have a history of performing well in applications that involve medical images; specifically in the detection of brain lesions[3] and image augmentation for liver lesion classification.[45] Second, we found that all of our GAN models performed well inside a KBP pipeline without significant parameter tuning and architecture modification, both of which are essential and potentially time consuming steps in conventional GAN implementations. We attribute this success to the application; the prediction of dose distributions is akin to the prediction of relatively smooth and uniform images with the same orientation. Third, in the KBP pipeline, the GAN produces images that are used as input to an optimization model in order to obtain treatment plans via a traditional inverse planning procedure. Thus, the GAN learns a simpler style mapping as compared to conventional applications, and the optimization phase acts as a safety net to correct potential errors.

Finally, it is interesting to note that the method to train a GAN conceptually mimics the iterative process between a treatment planner and an oncologist. The generator behaves as a treatment planner by proposing dose distributions while the discriminator behaves as an oncologist by determining whether the proposed dose distribution is suitable.

While other pipelines that predict dose directly have used voxel-based dose mimicking to construct the final plans,[76] we chose to do inverse planning using DVH-based objectives following prediction because it is in line with current clinical practice and is the most common approach used in the academic KBP literature.[7;44;109] We also emphasize that our method does not use hand-tailored feature engineering (e.g., features derived from overlap-volume histograms). Thus, as compared to existing KBP methods, we expect our pipeline to be easier to implement in practice and can result in more predictable results if custom treatment plans are desired. For example, institutions with specific clinical guidelines can train a GAN on images they deem indicative of an ideal treatment plan. In addition, in the future it may be possible that several medical centers combine data to form a large training set, which should further improve performance.

A limitation of our approach is that the prediction and optimization steps are separate stages in the overall pipeline. In theory, an integrated model that does both prediction and treatment plan optimization simultaneously should produce even better results. A second limitation is that our approach requires a clean, well-structured and high-quality dataset, where all images need to have a consistent size (in terms of number of pixels), coloring convention, and orientation. Finally, as with any neural network-based approach, GAN predictions suffer from a lack of interpretability. It is not straightforward to understand why the GAN makes certain predictions, effectively rendering it a black box. Consequently, a treatment planner may have more difficulty using this approach to understand when and why prediction errors occur.

3.5 Conclusion

We developed the first knowledge-based automated planning framework using a 3D generative adversarial network for prediction. Our results based on 217 oropharyngeal cancer treatment plans demonstrated superior performance in satisfying clinical criteria and generated more realistic predictions compared to the previous state-of-the-art. Our two primary contributions to the KBP framework are the generation of full 3D dose predictions using a generative adversarial network and the scaling of dose predictions pre-fluence map optimization. This allowed us to design high-quality fluence-based treatment plans without manual intervention.

Chapter 4

OpenKBP: The open dose prediction challenge

In the previous chapters, we developed new dose prediction models and evaluated their performance on a large private dataset. However, establishing high-quality benchmarks in those chapters was challenging because we needed to recreate complex models from the extent literature, which were also originally developed and evaluated on private datasets. In this chapter, we develop an open framework for future knowledge-based planning (KBP) research that fosters a collaborative environment where benchmarking is straightforward. Our framework involves a publicly available dataset and standardized evaluation metrics. To promote the adoption of our framework, we organize the first international competition for dose prediction where models are developed and compared using our standardized dataset and evaluation metrics.

4.1 Introduction

The increasing demand for radiation therapy to treat cancer has led to a growing focus on improving patient flow in clinics.[6] Knowledge-based planning methods promise to reduce treatment lead time by automatically generating patient-specific treatment plans, thereby streamlining the treatment planning process.[94] KBP methods are generally formulated as two-stage pipelines (see Figure 4.1). In most cases, the first stage is a machine learning (ML) method that predicts the dose distribution that should be delivered to a patient based on contoured CT images, and the second stage is an optimization model that generates a treatment plan based on the predicted dose distribution.[11;76]

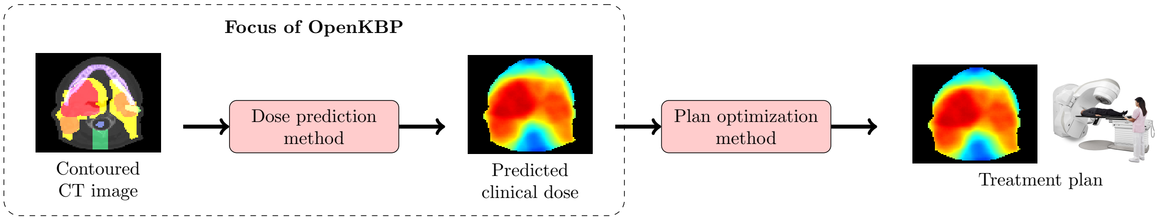


Figure 4.1: Overview of a complete knowledge-based planning pipeline.

Research into dose prediction has experienced major growth in the past decade,[46] in part due to the growing sophistication of machine learning and optimization methods in conjunction with advances in computational technology. There are two main branches of dose prediction methods: (1) those that predict summary statistics (e.g., dose-volume features)[5;7;113;116] and (2) those that predict entire 3D dose distributions.[10;61;71;74;83;95] Both branches of dose prediction methods use a wide range of methodologies, e.g., linear regression,[7] principal component analysis,[113;116] random forest,[74] neural networks.[10;61;71;83;95] All of this KBP research is performed in close collaboration with radiation therapy clinics using private clinical datasets that are generated via local planning protocols.[46]

Development of KBP models is further challenged by the lack of large open datasets and standardized evaluation metrics. Existing open radiation therapy datasets cater to optimization[25;36] or classification problems (e.g., segmentation, prognosis).[33] Researchers that develop dose prediction models must rely on their own private clinical datasets and different evaluation metrics, which makes it difficult to objectively and rigorously compare the quality of different prediction approaches at a meaningful scale.[46] As a result, researchers must attempt to reproduce published dose prediction models to benchmark their new models via a common dataset and set of evaluation metrics.[11;75] In contrast, open datasets and standardized metrics are staples of thriving artificial intelligence-driven fields, as evidenced by the uptake of the CIFAR[63] and ImageNet[91] datasets in the computer vision community over the past decade.

We launched the Open Knowledge-Based Planning (OpenKBP) Grand Challenge to advance knowledge-based planning by 1) providing a platform to enable fair and consistent comparisons of dose prediction methods and 2) developing the first open dataset for KBP. Participants of the Challenge used the dataset to train, test, and compare their prediction methods, using a set of standardized evaluation metrics. The data and accompanying code-base is freely available at `https://github.com/ababier/open-kbp` for KBP researchers to use going forward.

4.2 Methods and Materials

We first describe our process for building and validating the dataset for the Challenge. We then describe how the Challenge was organized and delivered. Finally, we provide an analysis of the Challenge results. This study was approved by the Research Ethics Board at the University of Toronto.

4.2.1 Data Processing

Figure 4.2 depicts our data processing approach at a high level, which consisted of four steps: (i) data acquisition, (ii) data cleaning, (iii) plan augmentation, and (iv) data partitioning.

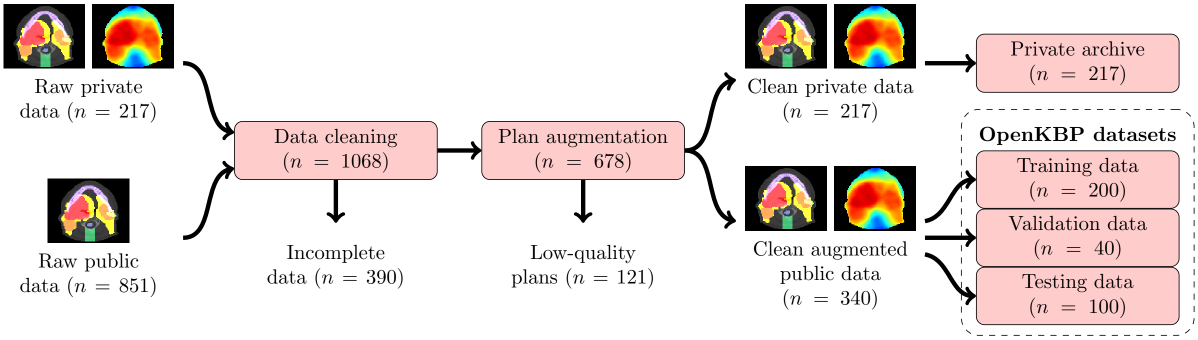

Raw private data (n = 217)
Raw public data (n = 851)
Data cleaning (n = 1068)
Incomplete data (n = 390)
Plan augmentation (n = 678)
Low-quality plans (n = 121)
Clean private data (n = 217)
Clean augmented public data (n = 340)
Private archive (n = 217)
OpenKBP datasets
Training data (n = 200)
Validation data (n = 40)
Testing data (n = 100)

Figure 4.2: Overview of the data processing pipeline. n represents the number of patients at each stage of the pipeline.

4.2.1.1 Acquiring the raw data

We obtained the Digital Imaging and Communications in Medicine (DICOM) files of 217 patients, which we call the raw private data (denoted by $\mathcal{P}^{\text{raw}}$), who were treated for oropharyngeal cancer at Princess Margaret Cancer Center. Each file included a treatment plan that was delivered from nine approximately equispaced coplanar fields with 6 MV, step-and-shoot, intensity-modulated radiation therapy (IMRT). Each patient was prescribed 70 Gy in 35 fractions, with 70 Gy to the high-dose planning target volume (PTV70), 63 Gy to the mid-dose planning target volume (PTV63), and 56 Gy to the low-dose planning target volume (PTV56); a PTV63 was only contoured in 130 of the patients. All plans included CT images, contours for regions-of-interest (ROIs), and the dose distributions based on a consistent set of planning protocols.

We also retrieved clinical DICOM data for 851 patients, which we call the raw public data (denoted by $\mathcal{O}^{\text{raw}}$), from four public data sources[22;50;102;119] hosted on The Cancer Imaging Archive (TCIA).[33] The data was originally sourced from twelve different institutions between 1999 and 2014. Each file contained CT images and contours for the regions of interest (ROIs). This collection of files contained several inconsistencies because the data originated from different institutions. For example, different institutions may have employed different dose levels, fractionation schemes, ROI naming conventions, languages (English versus French nomenclature), PTV margins (isotropic versus anisotropic margins from clinical target volume (CTV)), and treatment modalities (3D conformal radiation therapy (3DCRT) versus IMRT).

4.2.1.2 Data cleaning

In order to standardize and improve the homogeneity of the datasets, we employed a sequence of data cleaning procedures. First, we relabeled all of ROIs according to a consistent nomenclature. For each patient $p \in \mathcal{P}^{\text{raw}} \cup \mathcal{O}^{\text{raw}}$, we included organ-at-risk (OAR) contours for the brainstem, spinal cord, right parotid, left parotid, larynx, esophagus, and mandible; let $\mathcal{I}_p$ denote this set of OARs for a patient p. All other OAR contours were deleted. Also, an OAR was omitted from $\mathcal{I}_p$ if it was not contoured in the clinical plan (e.g., a patient whose left parotid was not contoured would not have it in the set $\mathcal{I}_p$). To construct the set of targets $\mathcal{T}_p$, we identified the low-, mid-, and high-dose targets based on their relative dose levels[73] and relabeled them as PTV56, PTV63, and PTV70, respectively. Any region with overlapping PTVs was relabeled as a single PTV with a dose-level equal to that of the highest dose-level of those overlapping PTVs.

Next, we modified target contours in the raw public dataset ($\mathcal{O}^{\text{raw}}$) to match the protocols from the private dataset ($\mathcal{P}^{\text{raw}}$). These modifications helped to fix some of the inconsistencies in contouring (e.g., no PTV margins, anisotropic PTV margins) that were present in the raw public dataset. Every PTV was expanded to include the voxels within 5 mm of its respective CTV; the PTV was left unchanged in cases where there was no CTV contour associated with the PTV. Every PTV was also clipped to be no closer than 5 mm from the surface of the patient.

We generated dose influence matrices for each patient in the public dataset $\mathcal{O}^{\text{raw}}$ based on 6 MV step-and-shoot IMRT with nine equispaced coplanar fields at $0°, 40°, \ldots, 320°$. Those fields were divided into a set of beamlets $\mathcal{B}$ that were each 5 mm × 5 mm. Every patient was also divided into a set of voxels $\mathcal{V}^p$ that were downsampled to fill axial slices of dimension 128×128. The relationship between the intensity w_b of beamlet b

and dose d_v deposited to voxel v was calculated in MATLAB using the IMRPT library in A Computational Environment for Radiotherapy Research,[38] which we used to form the elements $D_{v,b}$ of each patient's dose influence matrix. The dose to a voxel v was calculated as:

$$d_v = \sum_{b \in \mathcal{B}} D_{v,b} w_b, \ \forall v \in \mathcal{V}^p, \ \forall p \in \mathcal{P}.$$

To prepare the patient data for deep learning models, we framed each patient in a 128 × 128 × 128 voxel tensor in two steps. First, we calculated the weighted average position of each patient p, using $\sum\limits_{b \in \mathcal{B}} D_{v,b}$ as the weight for each voxel $v \in \mathcal{V}^p$. Second, we applied a bounding box centered on that weighted average position with dimensions of 128 × 128 × 128 voxels. We added zero-padding where necessary to ensure consistent tensor volumes. Over the course of the data cleaning phase, 390 patients were removed from the public dataset ($\mathcal{O}^{\text{raw}}$) for a variety of reasons including missing target contours and issues generating a valid dose influence matrix. No patients were removed from our private dataset. At the end of the data cleaning step, we had clean private $\mathcal{P}^{\text{clean}}$ and public $\mathcal{O}^{\text{clean}}$ datasets consisting of 217 and 461 patients, respectively.

4.2.1.3 Plan augmentation

Next, we generated synthetic plans for each patient in the clean public dataset and only retained the associated dose distribution. These synthetic plans were generated using a variation of a published automated KBP pipeline,[10] which was trained using the cleaned clinical plans from our private dataset $\mathcal{P}^{\text{clean}}$. Figure 4.3 illustrates the plan augmentation process.

The dose prediction model in the KBP pipeline was a conditional generative adversarial network (GAN)[57] with the same architecture as the 3D-GAN in Babier *et al.* 2020.[10] It uses two neural networks: (1) a generator that predicts the dose distribution

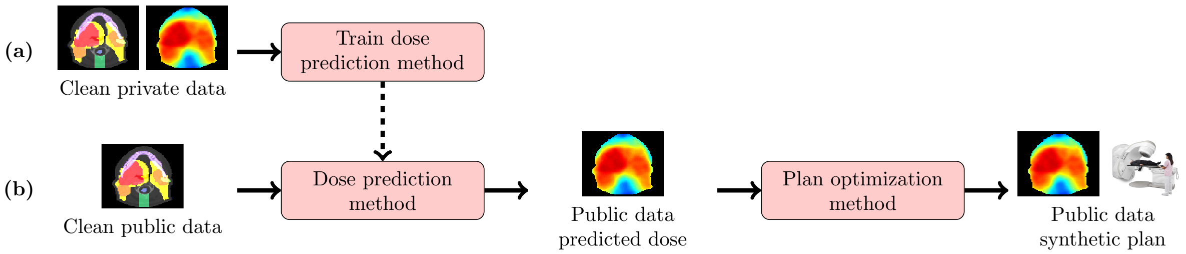


Figure 4.3: Overview of the plan augmentation process, which is a two phased approach: (a) the clean private clinical dataset is used to train the dose prediction method, and then (b) the trained method is used in a complete KBP pipeline that intakes the cleaned public data to generate synthetic plans.

for a contoured CT image; and (2) a discriminator that predicts whether the input is a predicted or clinical dose distribution. We trained the generator to minimize the mean absolute difference between the predicted and clinical dose distributions, which we regularized with the discriminator to encourage the generator to make predictions resembling clinical dose distributions. Between each batch update of the generator, we also trained the discriminator to minimize a binary-cross-entropy loss function. This GAN model was trained for 200 epochs using the clean private dataset of 217 treatment plans, and it was implemented in Tensorflow 1.12.3 on a Nvidia 1080Ti graphic processing unit (GPU) with 12 GB of video memory.

As part of the plan optimization, we added seven optimization structures to each patient in the public dataset to encourage high-quality synthetic plans. All of these optimization structures are based on structures that were used to optimize the plans in our private clinical dataset. These structures were not included in the final Challenge datasets. The first optimization structure was called limPostNeck, which is used to limit dose to the posterior neck. The limPostNeck includes all of the non-target voxels between the posterior aspect of a 3 mm expansion of the spinal cord and the patient posterior; there were 12 cases where no spinal cord was contoured where we extended the brainstem inferiorly to approximate the spinal cord to make the limPostNeck. All spinal cord and target voxels were removed from the limPostNeck. The other six optimization structures were PTV rings, which we added to encourage high dose gradients around the PTVs. We used 2 mm and 6 mm rings that include voxels within 0 mm to 2 mm and 2 mm to 6 mm of the PTV, respectively. Any overlap between rings was eliminated by removing voxels in those overlapping regions from the ring of PTV with the lower dose-level. All target voxels were also removed from the rings.

The plan optimization method was a two-stage approach to inverse planning.[8] In the first stage, we estimate the objective function weights for a conventional inverse planning model that makes a predicted dose distribution optimal. In the second stage, we use the estimated weights and solve the inverse planning model to generate a synthetic treatment plan. The objective of the planning model was to minimize the sum of 114 functions: seven per OAR, three per target, and seven per optimization structure. The functions for each OAR evaluated the mean dose; maximum dose; and average dose above 0.25, 0.50, 0.75, 0.90, and 0.975 of the maximum predicted dose to that OAR. The functions for each target evaluated the mean dose, maximum dose, average dose below the target dose level, and average dose 5% above the target dose level (e.g., average dose above 73.5 Gy in the PTV70). The functions for each optimization structure were the same as the OAR functions. To ensure that all plans had a similar degree of fluence complexity, all synthetic plans were constrained to a sum-of-positive-gradients (SPG) value of 65.[35] Both optimization models were solved in Python 3.6 using Gurobi 9.0.1 (Gurobi Optimization, TX, US) to generate a dose distribution $\hat{\mathbf{d}}_p$ for each patient in the clean public dataset.

We used Algorithm 1 to correct or remove low-quality plans that were generated by our plan augmentation process (i.e., the process in Figure 4.3). The algorithm curated a set of patients $\mathcal{O}^{\text{aug}}$ with high-quality dose distributions $\hat{\mathbf{s}}_p$, which were based on the dose distributions $\hat{\mathbf{d}}_p$ for the plans of patients in the clean public dataset $\mathcal{O}^{\text{clean}}$. The algorithm retained any patients that had plans with a high-dose target that received a higher mean dose or 1st percentile dose (D_{99}^{r}) than the mid-dose or low-dose targets (line 4). The entire dose was then multiplicatively scaled so that maximum dose to the high-dose target $\mathrm{D}^{\mathrm{PTV70}}_{\mathrm{max}}(\hat{\mathbf{s}}_p)$ was no lower than the lowest maximum dose ($\underline{\mathrm{D}}^{\mathrm{PTV70}}_{\mathrm{max}}$) observed in the plans of the patients from our clean private dataset $\mathcal{P}^{\mathrm{clean}}$ (line 5). For each instance where we scaled dose by a constant factor, we also scaled the dose by a random factor ε that was sampled from a uniform distribution between 1.00 and 1.05 (i.e., $\varepsilon \sim \mathcal{U}(1.00, 1.05)$) for two reason: (1) we did not want participants to learn a strict cutoff and (2) strict cutoffs are unrealistic. Next, we reduced the dose $\hat{\mathbf{s}}_p$ so that, for each ROI $r \in \mathcal{I}_p \cup \mathcal{T}_p$, it delivered a maximum dose $\mathrm{D}^r_{\mathrm{max}}$, mean dose $\mathrm{D}^r_{\mathrm{mean}}$, and dose to 99% of voxels D^r_{99} that was lower than the respective upper bound observed in the plans from our private clinical dataset (line 8). We denote the highest value observed in the clinical plans with a bar (e.g., $\overline{\mathrm{D}}^{\mathrm{r}}_{\mathrm{c}}$ for a criteria c and ROI r). Lastly, a patient p was added to $\mathcal{O}^{\mathrm{aug}}$ if that patient's respective dose $\hat{\mathbf{s}}_p$ had a maximum dose to the high-dose target that was between the lower and upper bounds that we observed in our private set of clinical plans (line 10). The final size of $\mathcal{O}^{\mathrm{aug}}$ was 340.

Algorithm 1: Improve low-quality plans where possible and construct the set of public patients with high-quality synthetic plan dose distributions $\hat{\mathbf{s}}_p$.

```
1  $\mathcal{O}^{\mathrm{aug}} \leftarrow \{\}$
2  for $p \in \mathcal{O}^{clean}$ do
3      $\hat{\mathbf{s}}_p \leftarrow \hat{\mathbf{d}}_p$
4      if $\mathrm{D}^{PTV70}_{\mathrm{mean}}(\hat{\mathbf{s}}_p) \geq \mathrm{D}^{t}_{\mathrm{mean}}(\hat{\mathbf{s}}_p)$ or $\mathrm{D}^{PTV70}_{99}(\hat{\mathbf{s}}_p) \geq \mathrm{D}^{t}_{99}(\hat{\mathbf{s}}_p),\ \forall t \in \mathcal{T}_p$ then
5          $\hat{\mathbf{s}}_p \leftarrow \hat{\mathbf{s}}_p \times \max(1,\ \underline{\mathrm{D}}^{\mathrm{PTV70}}_{\mathrm{max}} / \mathrm{D}^{PTV70}_{\mathrm{max}}(\hat{\mathbf{s}}_p) \times \varepsilon \sim \mathcal{U}(1.00, 1.05))$
6          for $r \in \mathcal{I}_p \cup \mathcal{T}_p$ do
7              for $c \in \{\mathrm{max},\ \mathrm{mean},\ 99\}$ do
8                  $\hat{\mathbf{s}}_p \leftarrow \hat{\mathbf{s}}_p \times \min(1,\ \mathrm{D}^{r}_{\mathrm{c}}(\hat{\mathbf{s}}_p) / \overline{\mathrm{D}}^{\mathrm{r}}_{\mathrm{c}} \times \varepsilon \sim \mathcal{U}(0.97, 1.00))$
9          if $\underline{\mathrm{D}}^{\mathrm{PTV70}}_{\mathrm{max}} > \mathrm{D}^{PTV70}_{\mathrm{max}}(\hat{\mathbf{s}}_p) > \overline{\mathrm{D}}^{\mathrm{PTV70}}_{\mathrm{max}}$ then
10             $\mathcal{O}^{\mathrm{aug}} \leftarrow \mathcal{O}^{\mathrm{aug}} \cup \{p\}$
```

4.2.1.4 Validation of final competition datasets

We evaluated the distribution of synthetic dose $\hat{\mathbf{s}}_p$ quality over every patient $p \in \mathcal{O}^{\mathrm{aug}}$ by comparing it to the distribution of the clinical dose quality over every patient $p \in$

$\mathcal{P}^{\text{clean}}$. We measured quality using the set of DVH criteria used in the Challenge. The distribution of DVH criteria over the population of synthetic doses and clinical doses was visualized with a box plot for each set of criteria. For each of the DVH criteria, we used a one-sided Mann-Whitney U test to determine whether the synthetic doses were inferior (null hypothesis) or non-inferior (alternative hypothesis) to the clinical doses, based on an equivalence interval of 2.1 Gy (i.e., 3% of the high-dose level).[4] Lower values were better for D_{mean}, $\text{D}_{0.1\text{cc}}$, and D_1; higher values were better for D_{95} and D_{99}. For these and all subsequent hypothesis tests, $P < 0.05$ was considered significant.

The final public dataset $\mathcal{O}^{\text{aug}}$ was randomly split into training $\mathcal{O}^{\text{train}}$, validation $\mathcal{O}^{\text{val}}$, and testing $\mathcal{O}^{\text{test}}$ datasets with 200, 40, and 100 patients, respectively. Every patient in these datasets had a synthetic dose distribution ($\hat{\mathbf{s}}_p$), CT images, structure masks, feasible dose mask (i.e., voxels $v \in \mathcal{V}^p$ such that $\sum_{b \in \mathcal{B}} D_{v,b} > 0$), and voxel dimensions. This data was released to the participants in phases as described in the next section. A detailed description of the data format and files is given in Appendix C.1.

4.2.2 Challenge Description

OpenKBP was hosted as an online competition using CodaLab (Microsoft Research, Redmond, WA). Participants could compete in the Challenge as a member of a team or as individuals (i.e., a team of one). The Challenge proceeded in two phases. In the first (validation) phase, teams developed their models and compared their performance in real time to other teams via a public leaderboard. In the second (testing) phase, teams submitted their dose predictions for a new unseen dataset, and we compared their performance to other teams via a hidden leaderboard to determine the final rankings for the Challenge.

4.2.2.1 Challenge timeline

The Challenge took place over four months in 2020. Individuals could register to participate in the Challenge anytime after it started on February 21, 2020, which is also when the training and validation data was released to start the first (validation) phase of the Challenge. Three months later, on May 22, 2020, the testing data was released to start the second (testing) phase of the Challenge, which ended ten days later on June 1, 2020. The final competition rankings (based on testing phase performance) were released four days later on June 5, 2020. The Challenge also coincided with the beginning of the COVID-19 pandemic.[41] As a result, we extended the validation phase to accommodate for the challenges posed by the pandemic. The result was about a one-month delay compared to the originally planned timeline.

4.2.2.2 Participants

OpenKBP was designed with a view towards having a low barrier to entry. Registration was free and open to anyone. We also offered comprehensive instructions to set up free, high-quality compute resources via Google Colab (Google Research, US), for those teams who did not have access to sufficient computational resources otherwise.[28]

To understand the make-up of the OpenKBP community, we collected demographic information from every participant via a two-part registration survey (see Appendix C.2). The first part of the survey, which was mandatory, collected professional information including their past KBP research experience, primary research area, and academic/industry affiliations. The second part of the survey, which was optional, collected equity, diversity, and inclusion (EDI) data including how participants self-identify in terms of gender, race, and disability status, using terminology from the United States Census Bureau.

4.2.2.3 Evaluation metrics

Teams predicted dose distributions or dose-volume histograms for a set of patients and submitted those predictions to our competition on CodaLab. For each patient p, we compared the submitted prediction $\mathbf{s}_p$ to the corresponding synthetic plan dose distribution $\hat{\mathbf{s}}_p$ via two error measures (1) *dose error*, α_p, which measures the mean absolute difference between a submission and its corresponding synthetic plan (i.e., mean absolute voxel-by-voxel difference in dose), and (2) *DVH error*, $\beta^r_{p,c}$, which measures the absolute difference in DVH criteria between a submission and its corresponding synthetic plan. The dose error α_p was chosen as a general measure of prediction quality that is not radiation therapy specific. It was only used to evaluate dose distributions (i.e., not DVH submissions), and it is calculated as

$$\alpha_p = \frac{||\mathbf{s}_p - \hat{\mathbf{s}}_p||_1}{|\mathcal{V}^p|},\ \forall\ p \in \mathcal{O}^{\text{val}} \cup \mathcal{O}^{\text{test}}. \tag{4.1}$$

The DVH error $\beta^r_{p,c}$ was chosen as a clinical measure of prediction quality that is radiation therapy specific. It involves a set of DVH criteria $\mathcal{C}_i$ and $\mathcal{C}_t$ for each OAR $i \in \mathcal{I}_p$ and target $t \in \mathcal{T}_p$, respectively. There were two OAR DVH criteria: D^i_{mean}, which is the mean dose received by OAR i; and $\text{D}^i_{0.1\text{cc}}$, which is the maximum dose received by 0.1cc of OAR i. There were also three target DVH criteria: D^t_1, D^t_{95}, and D^t_{99}, which are the doses received by 1% (99th percentile), 95% (5th percentile), and 99% (1st percentile) of voxels in target t, respectively. The DVH error was used to evaluate both dose distribution and DVH submissions, and it is calculated as

$$\beta^r_{p,c} = \left|\text{D}^r_{\text{c}}(\mathbf{s}_p) - \text{D}^r_{\text{c}}(\hat{\mathbf{s}}_p)\right|,\ \forall\ c \in \mathcal{C}_r,\ \forall\ r \in \mathcal{I}_p \cup \mathcal{T}_p,\ \forall\ p \in \mathcal{O}^{\text{val}} \cup \mathcal{O}^{\text{test}}. \tag{4.2}$$

We chose to make both error metrics absolute differences to reward models that learn to make realistic dosimetric trade-offs, as opposed to signed differences that may reward unrealistic dosimetric trade-offs (e.g., predict low dose to an OAR that is unachievable). This is critical because prediction models that make unrealistic trade-offs generally perform worse than models that make realistic trade-offs in full KBP pipelines.[7]

Building on these error metrics, we scored submissions using *dose score* A_h and *DVH score* B_h. Both scores are a variation of mean absolute error (MAE). The dose score is the mean dose error over all patients in a hold-out set (i.e., $\mathcal{O}^{\text{val}}$ or $\mathcal{O}^{\text{test}}$):

$$A_h = \frac{1}{|\mathcal{O}^h|} \sum_{p \in \mathcal{O}^h} \alpha_p, \ \forall \ h \in \{\text{val, test}\}. \tag{4.3}$$

The DVH score is the mean DVH error over all criteria from the patients in a hold-out set:

$$B_h = \frac{1}{\sum\limits_{p \in \mathcal{O}^h} \sum\limits_{r \in \mathcal{I}_p \cup \mathcal{T}_p} |\mathcal{C}_r|} \sum_{p \in \mathcal{O}^h} \sum_{r \in \mathcal{I}_p \cup \mathcal{T}_p} \sum_{c \in \mathcal{C}_r} \beta^r_{p,c}, \ \forall \ h \in \{\text{val, test}\}. \tag{4.4}$$

Using those scores, we ranked all of the submissions to the Challenge in two streams: (1) the dose stream where the team with the lowest (i.e., best) dose score won, and (2) the DVH stream where the team with the lowest (i.e., best) DVH score won.

4.2.2.4 Validation phase

At the start of the validation phase, the full training dataset $\mathcal{O}^{\text{train}}$ of 200 patients was released, and the teams used that data to train their models. An out-of-sample validation dataset $\mathcal{O}^{\text{val}}$, which included data for 40 patients without synthetic plan dose, was also released for teams to validate the out-of-sample performance of their models. Predictions made on the validation dataset were submitted directly to our competition on CodaLab where they were scored in the cloud using the held-back synthetic plan dose. The resulting scores populated a public leaderboard, but they were not used to determine the winners of the Challenge.

4.2.2.5 Testing phase

The testing dataset $\mathcal{O}^{\text{test}}$, which included data for 100 patients without synthetic plan dose, was released at the start of the testing phase. Teams used the models they developed during the validation phase and made predictions on this new unseen testing dataset. Similar to the validation phase, all predictions were submitted to our competition on CodaLab where they were scored in the cloud using the held-back synthetic plan dose. However, the resulting dose and DVH scores populated the testing leaderboard that we kept hidden until the competition finished. The team that performed best on the testing leaderboard with respect to the dose and DVH score was the winner of the dose and DVH stream, respectively. Teams that submitted to the testing leaderboard also responded to a model survey (see Appendix C.2) to summarize their models.

4.2.3 Analysis of Challenge Outcomes

We conducted four analyses. First, we summarized the demographics of the participants. Second, we evaluated the aggregate improvements made by the teams over the course of the validation phase. Third, we compiled and analyzed the final results from the testing phase. Fourth, we summarized common modeling techniques that were employed by the participants.

4.2.3.1 Participant information

We examined the registration information of all participants and calculated summary statistics for primary research area, past KBP research experience, country of work/study, and EDI data. We compared our aggregated EDI data to comparable data for the population of people who are employed in science and engineering (S&E) in the United States (US)[81] and the general US population.[99]

4.2.3.2 Performance over validation phase

As a retrospective analysis, we evaluated the aggregate improvement of all teams over the validation phase to measure their progress throughout the Challenge. We plotted the dose and DVH score against a relative measure of progress towards their final model, which we call the *normalized submission count* (NSC). The NSC is equal to the cumulative number of submissions a team made up to a certain point in time divided by the total number of submissions made in the validation phase. For example, if a team made 100 total submissions, the score at NSC = 0.5 represents that team's best recorded performance after their 50th submission. For each team, we recorded their best cumulative dose and DVH score in increments of 0.05 NSC. At each increment we plotted the average and the 95% confidence interval of those scores over all teams that made more than 20 total submissions.

4.2.3.3 Final results in testing phase

We used a one-sided Wilcoxon signed-rank test to determine whether the set of predictions of the best team in each stream had the same (null hypothesis) or lower (alternative hypothesis) error (i.e., α_p and $\beta^r_{p,c}$) than each set of predictions submitted by the other teams. To visualize the range of expected error differences, we plotted the difference in dose error over all patients ($n = 100$) and the difference in DVH error over all DVH criteria ($n = 1783$) between the winning submission and the runner-up submissions, for the dose stream and DVH stream, respectively.

As a retrospective sensitivity analysis, we evaluated the submissions according to an *alternative* scoring function with squared error terms (i.e., α_p^2 and ${\beta^r_{p,c}}^2$) instead of absolute error terms (i.e., α_p and $\beta^r_{p,c}$) to determine if the final competition standings would have changed. We refer to the competition and alternative scores as MAE-based and mean squared error (MSE)-based, respectively. As a quantitative measure of the alignment between the two ranking methods, we evaluated the rank-order correlation between the rankings for the MAE-based and MSE-based scores via Spearman's rank test.

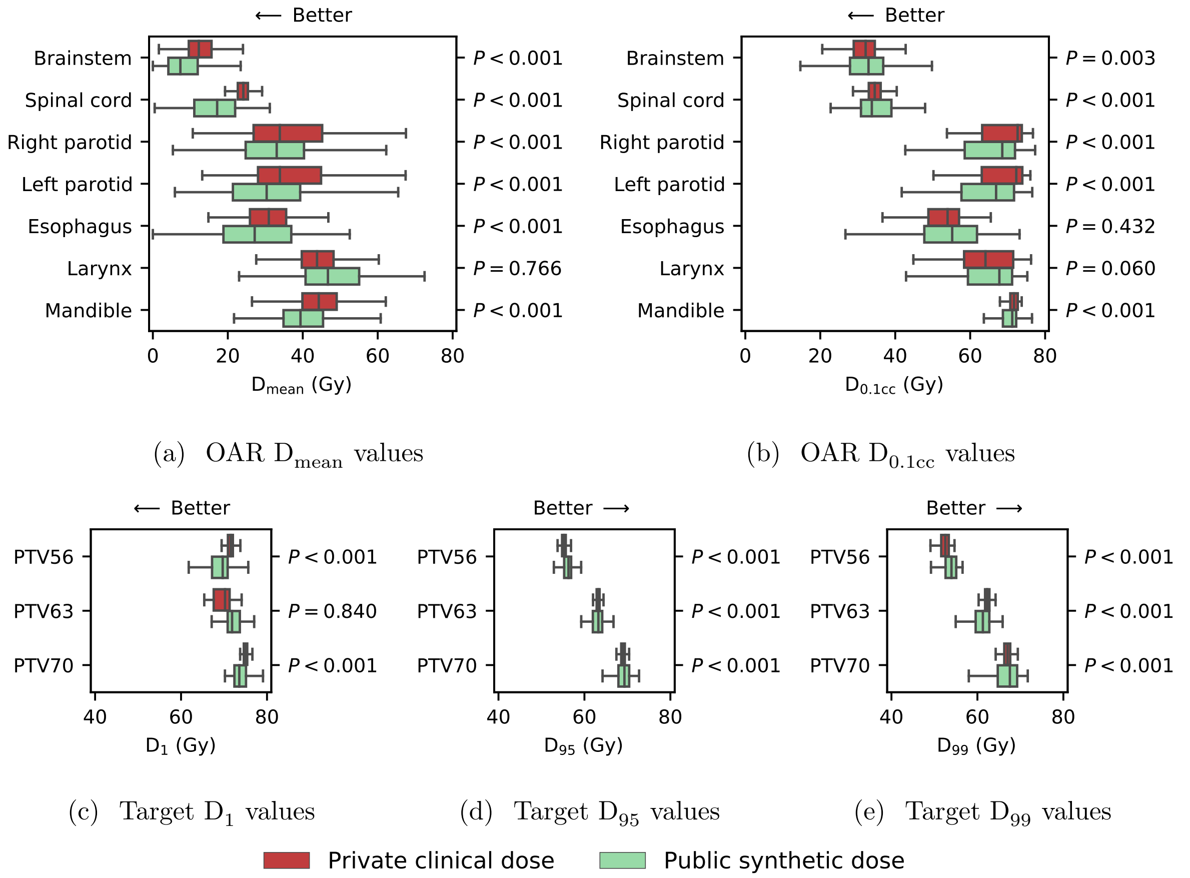
4.2.3.4 Common modeling decisions

Finally, we present a summary of the model survey information that teams submitted during the testing phase. We present common modeling choices (e.g., model architectures), hardware, and software that teams used. Lastly, we present a set of techniques that we believe are generalizable to most dose prediction frameworks, based on what teams commonly employed.

4.3 Results

4.3.0.1 Validation of final competition datasets

Figure 4.4 compares the quality of the public synthetic dose distributions to the private clinical dose distributions. The box plots in the top and bottom row summarize the performance across OAR and target DVH criteria, respectively. The public synthetic doses were non-inferior ($P < 0.05$) to the clinical doses on 19 of the 23 criteria. For the remaining four criteria, the synthetic dose was 2.1 Gy worse on average than the clinical dose (3.7% average relative difference). While the synthetic doses were not a perfect replication of the clinical doses, they are sufficiently close to representing clinical dose distributions for the purpose of this Challenge and future research that leverages this dataset.



(a) OAR $\mathrm{D_{mean}}$ values

(b) OAR $\mathrm{D_{0.1cc}}$ values

(c) Target $\mathrm{D_1}$ values

(d) Target $\mathrm{D_{95}}$ values

(e) Target $\mathrm{D_{99}}$ values

Figure 4.4: The distribution of DVH criteria from the private clinical dose and the public synthetic dose is plotted, and the corresponding P-values for each criterion are on the right axes. The boxes indicate median and interquartile range (IQR). Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier.

4.3.0.2 Participant information

Table 4.1 summarizes the participation in each phase of the Challenge. Overall, 195 people registered to participate, and 73 participants were active during the validation phase. A total of 1750 submissions were made to the validation phase, which is an average of 40 submission per team. There were 28 unique models submitted in the testing phase.

Table 4.1: Participation throughout each phase of the Challenge.

	Registration	Validation	Testing
Total participants	195	73	54
Total teams	129	44	28
Number of submissions	—	1750	28

Table 4.2 summarizes the participants' past KBP experience and primary area of research. Interestingly, 61.5% of participants had no prior KBP experience and less than half (42.6%) identified medical physics as their primary area of research. Machine learning researchers constituted the majority (50.3%) of the participants, and only about one third (33.3%) of those researchers had prior KBP experience.

Table 4.2: Distribution of participants by primary research area (rows) and whether they have past KBP research experience (columns).

	KBP Experience		
Primary Research Area	Yes	No	Total
Machine Learning	16.9%	33.3%	50.3%
Medical Physics	17.9%	24.6%	42.6%
Optimization	2.1%	2.6%	4.6%
Other	1.5%	1.0%	2.6%
Total	38.5%	61.5%	100.0%

Table 4.3 presents the proportion of participants by country of work or study. In total, 28 different countries were represented in the Challenge. The three countries with the most participants were the United States (32.8%), China (17.4%), and India (11.3%). Each of the other 25 counties that were represented had less than 5.1% of the participants.

Table 4.3: The proportion of participants based on country of work or study.

Australia	2.1%	Colombia	4.1%	Malaysia	0.5%	Sudan	0.5%
Austria	2.1%	Croatia	0.5%	Netherlands	1.0%	Sweden	1.5%
Bangladesh	0.5%	Finland	2.1%	Pakistan	1.0%	Taiwan	2.1%
Belgium	1.5%	France	3.6%	Poland	0.5%	Turkey	1.0%
Brazil	0.5%	Germany	1.0%	South Africa	0.5%	United Kingdom	1.0%
Canada	5.1%	India	11.3%	South Korea	3.1%	United States	32.8%
China	17.4%	Japan	1.0%	Spain	1.0%	Vietnam	0.5%

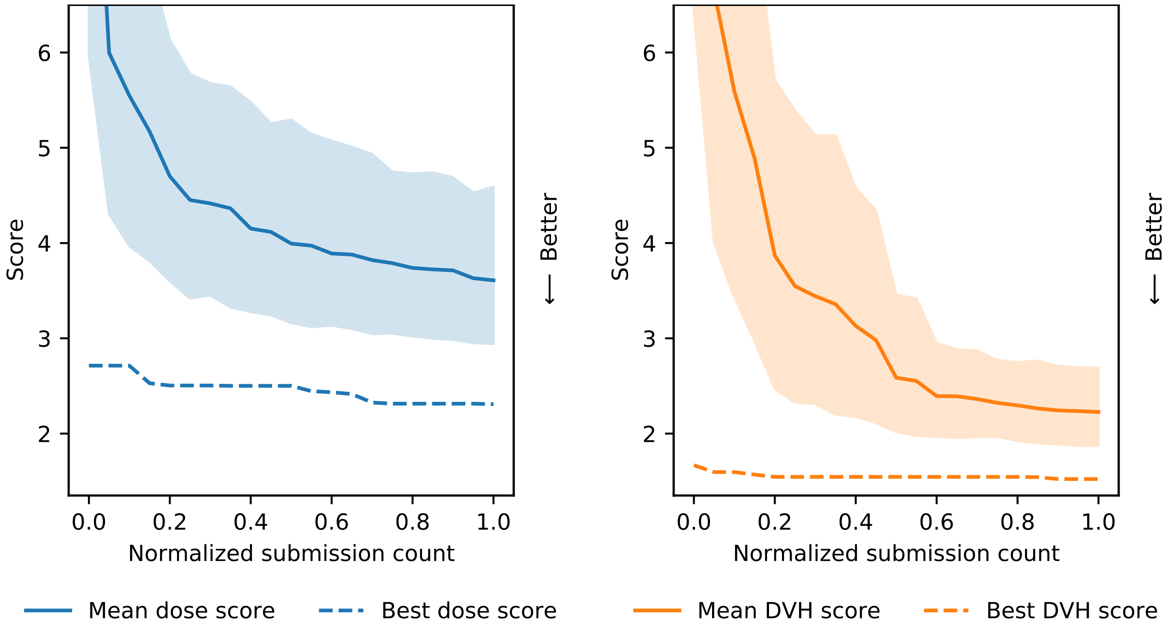
In Table 4.4, we present the aggregate data from our EDI survey. Men were overrepresented in OpenKBP (76.9%) compared to the science and engineering population (52.3%) and the general US population (49.2%). "Asian American/Asian" was the most common racial or ethnic identity (48.7%) in OpenKBP, which is much greater than the science and engineering population (13.0%) and the general US population (5.6%). On the other hand, individuals who identified as "White" were underrepresented in OpenKBP (21.5%) compared to the science and engineering (68.7%) and US (60.0%) populations. Individuals who identified as "African American/Black" (1.0%) and "Hispanic/Latinx" (4.1%) were also underrepresented relative to both baseline populations. A relatively large proportion (18.9%) of respondents chose not identify their racial or ethnic identity. Lastly, the proportion of OpenKBP participants who identified as having no disability (87.2%) was comparable to both the science and engineering (89.7%) and general US (87.3%) population. Fewer respondents identified with having a disability (2.1%) compared to both baselines (10.3% and 12.7%), and the remaining proportion of OpenKBP participants chose not to identify their disability status.

Table 4.4: The equity, diversity, and inclusion data (rows) of three populations of people (columns). In order, the columns correspond to the population people who participated in the Challenge (OpenKBP), are employed in the United States in science and engineering (S&E), and live in the United States (US). A dash (—) indicates that the data is unavailable.

	OpenKBP	S&E[81]	US[99]
Number of people (n)	195	27,274	328,239,523
Gender identity			
Man	76.9%	52.3%	49.2%
Woman	12.8%	47.7%	50.8%
Prefer not to say	6.7%	—	—
No answer	3.6%	—	—
Racial or ethnic identity			
African American/Black	1.0%	7.3%	12.4%
Asian American/Asian	48.7%	13.0%	5.6%
Hispanic/Latinx	4.1%	8.6%	18.4%
Middle Eastern/North African	2.1%	—	—
Native American/Indigenous	0.5%	0.3%	0.7%
Native Hawaiian/Other Pacific Islander	0.0%	0.3%	0.2%
White	21.5%	68.7%	60.0%
Other	3.1%	1.8%	2.8%
Prefer not to say	13.3%	—	—
No answer	5.6%	—	—
Identify as having a disability			
No	87.2%	89.7%	87.3%
Yes	2.1%	10.3%	12.7%
Prefer not to say	7.2%	—	—
No answer	3.6%	—	—

4.3.0.3 Performance over validation phase

In Figure 4.5, we plot the distribution of team scores against normalized submission count. The plots show that teams generally improved their model throughout the validation phase, however, most teams made the largest improvements early on. Overall, the average team improved their dose and DVH score by a factor of 2.7 and 5.7, respectively, over the course of the validation phase. Over all of the NSC bins, the best dose and DVH scores were achieved by two and four different teams, respectively. There were a total of seven lead changes throughout the validation phase.



(a) Dose score

(b) DVH score

Figure 4.5: The distribution of the dose and DVH scores across all teams. The solid lines indicate the mean score and the shaded regions indicate the 95% confidence interval. A dash lined indicates the best score.

4.3.0.4 Final results in testing phase

Figure 4.6 shows the distribution of error differences between the winning team (i.e., Team 1) and the top 23 runners-up (Teams 2-24). Compared to each of the other teams, Team 1 achieved significantly lower dose error over all 100 patients in the testing set ($P < 0.05$) and significantly lower DVH error over all 1783 DVH criteria ($P < 0.05$). Additionally, when compared to any other team, Team 1 achieved a lower dose and DVH error over at least 75% and 52% of patients and criteria, respectively.

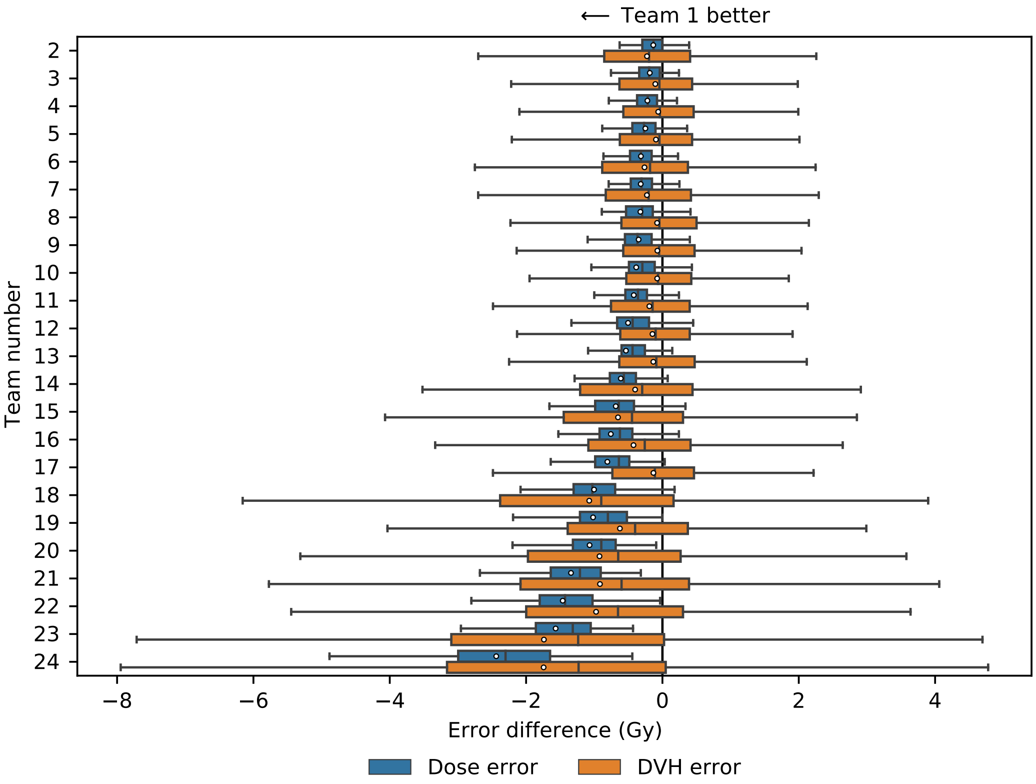


Figure 4.6: The distribution of dose and DVH error differences between the winning team (Team 1) and the top 23 runners-up, ranked by dose score. The boxes indicate median and IQR, and a circle indicates the mean. Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier.

Table 4.5 summarizes the relative performance of each team under the MAE-based and MSE-based scores. The winner and first runners-up according to MAE-based score would have finished in the same place under the MSE-based score. The average absolute rank difference between the two scoring approaches was one. The maximum rank difference was four and five for dose score and DVH score, respectively. The Spearman's rank-order correlation coefficient for the MAE-based and MSE-based score ranks was 0.983 ($P < 0.001$) and 0.981 ($P < 0.001$) for the dose and DVH score, respectively. Thus, the results of the competition would have been nearly identical had the MSE-based score been used.

Table 4.5: The score and rank that each team achieved in the testing phase according to an MAE-based (i.e., the score used in the Challenge) and MSE-based (i.e., an alternative score) score. A positive difference in rank implies that a team performed better on the MSE-based score than on the MAE-based score.

	Dose score					DVH score				
Team	MAE		MSE		Rank	MAE		MSE		Rank
	Score	Rank	Score	Rank	difference	Score	Rank	Score	Rank	difference
1	2.429	1	15.488	1	0	1.478	1	5.913	1	0
2	2.564	2	16.550	2	0	1.704	12	6.812	10	2
3	2.615	3	17.774	3	0	1.582	7	6.961	12	-5
4	2.650	4	18.091	5	-1	1.539	2	6.031	2	0
5	2.679	5	18.023	4	1	1.573	6	6.525	6	0
6	2.745	6	19.213	7	-1	1.741	14	7.088	14	0
7	2.748	7	18.916	6	1	1.706	13	7.083	13	0
8	2.753	8	19.356	8	0	1.556	5	6.372	3	2
9	2.778	9	19.469	9	0	1.551	3	6.634	7	-4
10	2.814	10	23.417	13	-3	1.555	4	6.433	4	0
11	2.850	11	20.432	10	1	1.669	11	6.904	11	0
12	2.934	12	25.266	14	-2	1.624	10	6.699	8	2
13	2.965	13	25.564	16	-3	1.611	9	6.745	9	0
14	3.040	14	23.192	12	2	1.878	15	7.648	15	0
15	3.114	15	22.454	11	4	2.130	18	9.445	17	1
16	3.186	16	25.427	15	1	1.902	16	8.622	16	0
17	3.237	17	33.642	20	-3	1.610	8	6.441	5	3
18	3.432	18	27.046	17	1	2.551	22	12.602	22	0
19	3.447	19	31.769	18	1	2.101	17	10.830	18	-1
20	3.498	20	33.459	19	1	2.401	20	11.324	19	1
21	3.771	21	37.980	22	-1	2.394	19	11.734	20	-1
22	3.892	22	37.439	21	1	2.451	21	12.436	21	0
23	3.996	23	43.654	23	0	3.216	23	19.856	23	0
24	4.867	24	55.435	24	0	3.220	24	20.178	24	0
25	5.446	25	70.694	25	0	3.712	25	22.335	25	0
26	8.165	26	130.212	26	0	10.362	28	143.358	27	1
27	11.818	27	266.484	27	0	7.713	26	100.080	26	0
28	14.660	28	341.024	28	0	8.379	27	155.465	28	-1

4.3.0.5 Common modelling decisions

According to the model survey, every team in the testing phase trained neural networks to predict dose distributions. The majority of those models had architectures based on U-Net,[90] V-Net,[78] and Pix2Pix[57] models. All models were built using either a TensorFlow

(Google AI, US) or PyTorch (Facebook AI Research, US) framework, but many teams also reported using higher-level libraries like fast.ai[55] to simplify model development. To train and develop the models quickly, teams generally used a GPU (e.g., NVIDIA 1080Ti, NVIDIA Titan V); seven teams also reported that they used Google Colab.

Many teams used generalizable techniques to get better model performance. For example, 22 of the 28 teams used some form of data augmentation in their training process, and 15 teams combined two or more augmentation methods. Common forms of data augmentation were rotations, flips, crops, and translations. Most teams also reported that they normalized dose and CT Hounsfield units. Additionally, most teams used a standard loss function, e.g., MAE, MSE, GAN loss. There were also some teams that developed radiation therapy specific loss functions (e.g., functions that prioritized regions-of-interest more than the unclassified tissue). Lastly, ensemble methods were used by several of the top teams. Those methods used multiple neural networks to predict candidate dose distributions that were combined by taking the average prediction. The exact techniques and methodology used by each of the top three teams are provided in: Liu et al. 2021,[66] Gronberg et al. 2021,[49] and Zimmermann et al. 2021.[118]

4.4 Discussion

There is widespread research interest in knowledge-based planning (KBP) dose prediction methods. However, the lack of standardized metrics and datasets make it difficult to measure progress in the field. In this paper, we present the first set of standardized metrics and the first open dataset for KBP research as part of the OpenKBP Grand Challenge, the first competition for KBP research. The Challenge democratizes KBP research by enabling researchers without access to clinical radiation therapy plans to develop state-of-the-art dose prediction methods. This spurred the development of 28 unique models and will serve as an important benchmark as the field of KBP continues to grow.

Our open dataset contains real patient images that were contoured by clinicians at twelve institutions with different planning protocols. There are two major differences in protocol that introduce some variance in how PTVs were drawn. First, the raw public clinical data included plans with multiple radiation therapy modalities. For example, some of the institutions delivered hybrid-IMRT/3DCRT plans, and those plans had no PTV margins on the lower neck target volumes. Second, some of the raw public clinical data is from multiple trials with unconventional contouring in the extent of the target volumes. For example, we observed some anisotropic PTV margins that were clipped to omit the OARs. These variations are non-existent in the raw private clinical dataset, which contains plans from a single institution where all planning and contouring was done according to a standard process, that was used to create the dose distributions for the competition. This variation may have been a factor in the public synthetic dose being non-inferior to the private clinical dose on 19 of 23 dose-volume criteria.

We proposed two new metrics that quantify the general performance (i.e., dose score) and the clinical performance (i.e., DVH score) of dose prediction methods. These two metrics may help measure progress in KBP research, and they will complement other metrics that are typically used in the literature to quantify strengths and weaknesses of a model. Other metrics are still important because our scoring metrics are unable to quantify every facet of radiation therapy dose quality. For example, the DVH criteria evaluated for the DVH score have varying degrees of clinical importance (e.g., $\mathrm{D}_{\mathrm{max}}^{\mathrm{mandible}}$ is much more important than $\mathrm{D}_{\mathrm{mean}}^{\mathrm{mandible}}$). We chose to weigh all errors equally because quantifying relative clinical importance is non-trivial and largely dependent on the institution. Additionally, since the scores are unweighted it is straightforward to use the scores for all other sites that have OARs and targets (e.g., prostate).

We aimed to make OpenKBP as accessible as possible in order to build a large and inclusive community, which was especially difficult because the Challenge started at the beginning of the COVID-19 pandemic when individuals around the world worked remotely. By building a large and inclusive community we ensure that underrepresented populations can contribute to KBP research, which should both accelerate innovation[53] and improve the quality of healthcare.[47] As part of this Challenge, we released all competition data in a non-proprietary format (comma-separated value) and a well-documented code repository that helped participants use the data easily and efficiently in Python without costly commercial software. The code repository also had instructions to give all participants access to high-quality computational resources at zero cost (i.e., Google Colab). In an effort to keep the data manageable for all participants, we also opted to use relatively large voxels (e.g., 3mm × 3mm × 2mm voxels) to ensure that the dose prediction problem was tractable for anyone using Google Colab. This manageable data size likely also helped the teams iterate and improve their models, which is reflected by the number of submissions made by teams in the validation phase (40 submissions on average). We conjecture that a successful model that was developed using the OpenKBP dataset should also succeed on other less accessible datasets (e.g., clinical datasets with smaller voxels and more ROIs).

A limitation of this work is that it uses synthetic dose distributions to augment the real clinical data. Those dose distributions were generated by a published KBP pipeline[7] and filtered via Algorithm 1, however, they underwent less scrutiny than clinical plans. Extensions of this work should ensure that the top performing models on this dataset also perform well with clinical dose distributions. A second limitation is that we can only report commonalities between the top models, which are correlated attributes rather than causal attributes. Future work should do ablation testing to isolate exactly what attributes contribute to a good dose prediction model. Lastly, all dose predictions were evaluated and ranked based on two scores. These scores do not capture all of the strengths and weaknesses of the models submitted to the Challenge.

4.5 Conclusion

OpenKBP democratizes knowledge-based planning research by making it accessible to everyone. It is also the first platform that researchers can use to compare their KBP dose prediction methods in a standardized way. The Challenge helps validate our platform and provides a much needed benchmark for the field. This new platform should help accelerate the progress in the field of KBP research, much like how ImageNet helped accelerate the progress in the field of computer vision.

4.6 Acknowledgments

OpenKBP was supported by the American Association of Physicists in Medicine and its Working Group on Grand Challenges. We also gratefully acknowledge everybody who participated in the Challenge including the following participants who competed in the final (testing) phase of the Challenge: Rob Verbeek (Aalto University); Alexander F.I. Osman (Al-Neelain University); Fuxin Ji, Dashan Jiang, Shuolin Liu, Qi Wu (Anhui University); Sibaji Gaj (Cleveland Clinic); Simeng Zhu (Henry Ford Health System); Kentaro Miki (Hiroshima University); Jingguang Liu (Hunan University); Sanchit Hira (Johns Hopkins University); Siri Willems (KU Leuven); Hoyeon Lee (Massachusetts General Hospital); Gerd Heilemann, Christian Ramsl, Lukas Zimmermann (Medical University of Vienna); Joel Beaudry (Memorial Sloan Kettering); Aditya Gautam, Anil Kumar, Supriya Mandal (Mirafra Technologies); Juan David, Harold David Eraso, Kelly Marixa, José Marrugo, José Darío Quinto Muñoz, Andrés Usuga Hoyos, Carlos Valderrama, Jean Carlo Jimenez Giraldo (National University of Colombia); Yuliang Huang (Peking University Cancer Hospital & Institute); Hongming Shan (Rensselaer Polytechnic Institute); Junjun He (Shanghai Jiao Tong University); Sen Yang (Sichuan University); Kay Sun (Studio Vodels); Keng-Chi Liu (Taiwan AI Labs); Jeevanshu Jain (Tata Memorial Centre); Carlos Cardenas, Skylar Gay, Mary Gronberg, Tucker Netherton, Dong Joo Rhee

(The University of Texas MD Anderson Cancer Center); Ana María Barragán Montero (UCLouvain); Chen Gefei (University of Macau); Jun Lian, Xuanang Xu (University of North Carolina at Chapel Hill); Yankui Chang, Mitty Meng, Zhao Peng (University of Science and Technology of China); Mumtaz Hussain Soomro (University of Virginia Health System); Dan Nguyen (UT Southwestern Medical Center); Erik Faustmann (Vienna University of Technology); Lulin Yuan (Virginia Commonwealth University Medical Center); Zijie Chen, Enpei Wang (WolHelp Technology (Shenzhen) Co Ltd); Nuo Tong (Xidian University); Jaehee Chun (Yonsei University).

Chapter 5

Evaluating complete automated planning pipelines

In the previous chapters we focused on improving one stage (i.e., the dose prediction model) of knowledge-based planning (KBP) while holding the other stage constant (i.e., the optimization model). In this chapter, we investigate the impact of changing both stages of KBP to explore the interaction effects between the two stages. We compare the performance of four KBP pipelines that are assembled from the four possible combinations of two high-quality dose prediction models and two high-quality plan optimization models. We evaluate the performance of each pipeline on our large private dataset of 217 oropharyngeal cancer treatment plans.

5.1 Introduction

Automated knowledge-based planning is a data-driven approach that uses previous radiation therapy treatments to generate high quality plans for patients diagnosed with cancer. KBP is typically conceptualized as a two-stage pipeline (see Figure 5.1). In the first stage, a machine learning (ML) model uses contoured CT images to predict the dose that should be delivered to a patient. In the second stage, an optimization model uses the dose prediction from the first stage to generate fluence maps or a set of beam apertures. In the past decade, there has been significant research in improving KBP, focusing either on advancing the machine learning or the optimization stage independently of each other.[5;44;61;83;95;96;107;111;113;116] In this work, we explore whether the interaction between the prediction and optimization model affects the overall quality of the final plans.

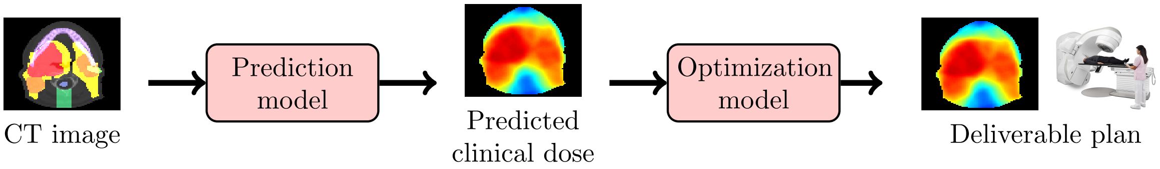


Figure 5.1: Overview of the automated knowledge-based planning pipeline.

Several prediction models have been developed that can accurately predict aspects of a clinical dose distribution from a patient's anatomy. Originally, these prediction models required the engineering of useful features from patient information and only predicted simple summaries (e.g., desirable dose-volume histograms) that could be used to design objectives for inverse planning software.[5;7;95;96;107;111;113;116] However, modern deep learning techniques can learn useful features in order to predict dose directly from CT images.[10;44;61;71;83] These high-dimensional predictions contain more information and permit better integration into more sophisticated automated KBP pipelines.[71]

The dominant optimization models for KBP are inverse planning (IP)[8] and dose mimicking (DM).[87] The choice of the parameters, objectives, and constraints in these models can lead to final treatment plans with characteristics that differ significantly from the initial predictions. For example, a prediction model may produce dose distributions that consistently predict excess dose to an OAR, but an optimization model with an objective to minimize the dose to that OAR may be able to correct for this bias. As a result, prediction models that produce dose distributions with good criteria satisfaction may not necessarily produce final plans with the same properties. Constructing effective automated KBP pipelines, therefore, requires careful selection of both the prediction and optimization model.

In this paper, we perform the first comparison between different combinations of prediction and optimization models in KBP; each model was previously validated in a KBP pipeline.[10;75] In total, we consider two dose prediction methods—a generative adversarial network[10] and a random forest[76]—and two optimization methods—inverse planning[8] and dose mimicking.[87] We then evaluate the four corresponding KBP pipelines (see Figure 5.2) using a large dataset of 217 patients with oropharyngeal cancer. We observe that the choice of both the prediction and optimization model can significantly affect the quality of the final plans generated by a KBP pipeline.

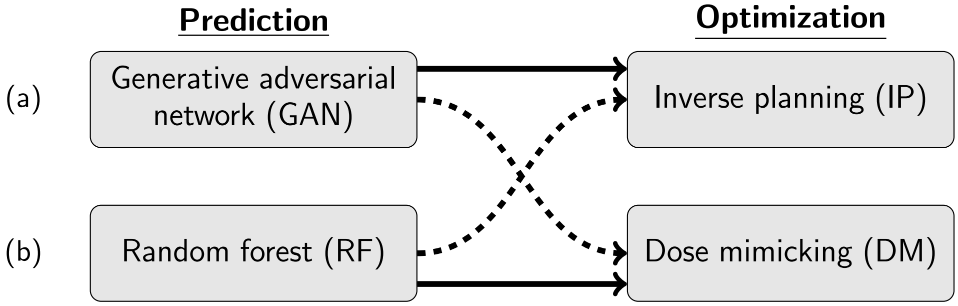


Figure 5.2: Overview of the automated knowledge-based planning pipelines evaluated in this paper. Solid lines connect the prediction and optimization methods that have been tested together in (a) Babier et al.[10] and (b) McIntosh and Purdie;[75] dashed lines connect the methods that have not been tested in the extant literature.

5.2 Methods and Material

We used CT images with contours, which highlight the regions-of-interest (ROIs), and dose distributions from clinically accepted treatment plans to train two dose prediction models that were then tested on out-of-sample clinical plans. The resulting predicted dose distributions were then passed through each optimization model to generate fluence-based treatment plans. Figure 5.2 gives an overview of the pipelines, which were then evaluated in terms of the quality of plans they produced.

5.2.1 Data

For this research ethics board approved study, we obtained plans for 217 oropharyngeal cancer treatments delivered at a single institution with 6 MV, step-and-shoot, intensity-modulated radiation therapy. All plans were prescribed 70 Gy and 56 Gy in 35 fractions to the gross disease (PTV70) and elective target volumes (PTV56), respectively; in 130 plans there was also a prescription of 63 Gy to the intermediate-risk target volume (PTV63). The organs-at-risk (OARs) were the brainstem, spinal cord, right parotid, left parotid, larynx, esophagus, mandible, and the limPostNeck, which is an artificial structure used to limit dose to the posterior neck.

5.2.2 Prediction models

We trained two state-of-the-art dose prediction models with the same 130 plans from our dataset and used the remaining 87 for out-of-sample testing.

5.2.2.1 Generative adversarial network

The conditional generative adversarial network (GAN) model[57] is based on Babier et al.[10] and uses two convolutional neural networks: (1) a generator that produces a dose distribution from a contoured CT image; and (2) a discriminator that tries to differentiate between the artificially generated dose and the actual clinical dose (see Figure 5.3). The generator is trained to minimize the mean absolute difference between the artificially generated image and the ground truth (i.e., clinical dose). The objective is regularized by the discriminator to make the output of the generator indistinguishable from a real clinical dose distribution. We then normalize the resulting dose generated by GAN so that it satisfies all target criteria.

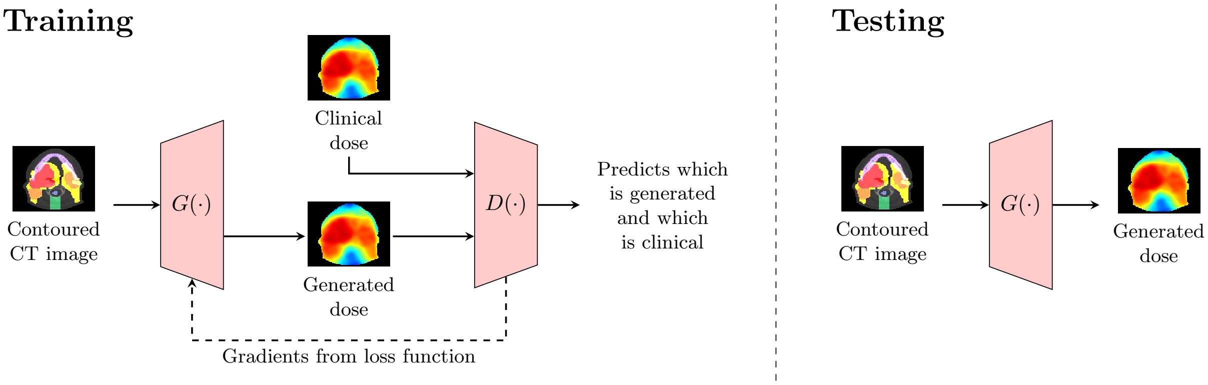


Figure 5.3: Overview of GAN training and testing phases.

5.2.2.2 Random forest

The random forest model is a slight variation of the RF from McIntosh and Purdie.[75] It uses the 148 features summarized in Table 5.1 to predict the dose delivered to each voxel independently. Of these features, 122 were generated by applying Gaussian filters (GFs) to the grayscale CT images. One of the GFs was isotropic ($\sigma = 10$) and a second was the Laplacian of the Gaussian ($\sigma = 10$). The remaining 120 filters were made from all combinations of the following four parameters: (a) first and second order GFs; (b) $\sigma = 4, 12, 24, 48$, and 64; (c) rotations of $0, 90, 180$, and 270 degrees; and (d) rotations in each of the three axes. RF was trained to minimize the mean squared difference between the prediction and the ground truth using the default settings of randomForestRegressor from `scikit-learn`.

Table 5.1: The features used in RF to predict the dose for each voxel.

Feature	Quantity	Description
Structure	11	Structure voxel is classified as (one-hot-encoded)
x-coordinate	1	Voxel's positions on the x-axis in a slice
y-coordinate	1	Voxel's positions on the y-axis in a slice
z-coordinate	1	Plane of voxel's slice
ROI distance	11	Voxel's distance from surface of each ROI
CT gray-scale	1	Voxel's gray-scale in the CT image
GF CT gray-scale	122	Voxel's gray-scale in CT image post GF

5.2.3 Optimization models

For all out-of-sample patients, the output from each prediction model was passed to both optimization models which were solved using `Gurobi 7.5`. The complexity of all generated treatment plans was constrained to a sum-of-positive-gradients (SPG) value of 55.[35] SPG was used since it is a convex surrogate for the physical deliverability of a plan and the parameter 55 was chosen as it is two standard deviations above the average clinical SPG.[7] Both optimization models used the same set of targets $\mathcal{T}$ and healthy structures $\mathcal{I}$. Each target $t \in \mathcal{T}$ was a planning target volume (PTV) with a prescribed dose θ^t. The healthy structures contained in $\mathcal{I}$ were the brainstem, spinal cord, right parotid, left parotid, larynx, esophagus, mandible, and limPostNeck. Each target structure $t \in \mathcal{T}$ and healthy structure $i \in \mathcal{I}$ was divided into a set of voxels $\mathcal{O}^t$ and $\mathcal{O}^i$, respectively.

The KBP-generated plans were delivered from nine equidistant coplanar beams at angles 0°, 40°, …, 320°. Those beams were divided into a set of beamlets $\mathcal{B}$, which make up one fluence map at each beam angle. The relationship between the intensity w_b of beamlet b and dose d_v deposited to voxel v was determined using the influence matrix $D_{v,b}$ generated by the IMRTP library from `A Computational Environment for Radiotherapy Research`,[38] and it is given by

$$d_v = \sum_{b \in \mathcal{B}} D_{v,b} w_b.$$

5.2.3.1 Inverse planning

We followed a previously developed two-stage approach to inverse planning.[8] In the first step, we estimate the objective weights for a conventional inverse planning model that makes a predicted dose distribution optimal. In the second step, the estimated weights are used to re-solve the conventional inverse planning optimization model and construct a treatment plan. The objective to be minimized was a sum of 65 functions: seven per OAR and three per target. The objectives for the OARs were the mean dose, maximum dose, and the average dose above 0.25, 0.50, 0.75, 0.90, and 0.975 of the maximum predicted dose to the OAR. The objectives for the target were the maximum dose, average dose below prescription, and average dose above prescription.

5.2.3.2 Dose mimicking

Our dose mimicking (DM) model minimized the sum of one-sided penalties to generate a plan that performs as close as possible to the predicted dose on several voxel- and structure-based objectives. Two types of OAR objectives were used. The first was a voxel-based objective that minimizes the dose d_v that exceeds the predicted dose $\hat{d}_v$ for each voxel v:

$$x_v = \max\left\{0,\ d_v - \hat{d}_v\right\}, \quad \forall v \in \mathcal{O}^i, \forall i \in \mathcal{I}. \tag{5.1}$$

The second was a structure-based objective that minimizes the maximum dose until it no longer exceeds the maximum predicted dose:

$$y^i = \max\left\{0,\ \max_{v\in\mathcal{O}^i}\{d_v\} - \max_{v\in\mathcal{O}^i}\{\hat{d}_v\}\right\}, \quad \forall i \in \mathcal{I}. \tag{5.2}$$

Three types of target objectives were also used. The first was a voxel-based objective to minimize the average deviation *below* the prescribed target dose θ^t, which is the average *underdose* to target t. Specifically, the objective function l_v penalizes dose until the plan underdose is no worse than what was predicted for each voxel v. The objective is formulated as:

$$l_v = \max\left\{0,\ \theta^t - d_v - \max\{0,\ \theta^t - d_v\}\right\}, \quad \forall v \in \mathcal{O}^t, \forall t \in \mathcal{T}. \tag{5.3}$$

Similarly, the second objective was also voxel-based, however, it minimizes the average deviation *above* the prescribed target dose θ^t, which is the average *overdose* to target t.

Specifically, the objective function u_v penalizes dose until the plan overdose is no worse than what was predicted for each voxel v. The objective is formulated as:

$$u_v = \max\left\{0,\ d_v - \theta^t - \max\{0,\ \hat{d}_v - \theta^t\}\right\}, \quad \forall v \in \mathcal{O}^t, \forall t \in \mathcal{T}. \tag{5.4}$$

The final target objective was structure-based. It maximizes the minimum dose to the target until it exceeds the minimum dose that was predicted for the target:

$$z^t = \min\left\{0,\ \min_{v \in \mathcal{O}^i}\{\hat{d}_v\} - \min_{v \in \mathcal{O}^i}\{d_v\}\right\}, \quad \forall t \in \mathcal{T}. \tag{5.5}$$

To formulate the dose mimicking optimization problem, we used an objective function that was the summation of (1)-(5) with the voxel-based objectives divided by the number of voxels in each respective structure. We then added the appropriate auxiliary variables and constraints.[18] That is, the conceptual DM model can be written as

$$\begin{aligned} \underset{x,y,l,u,z,w}{\text{minimize}} \quad & \sum_{i \in \mathcal{I}} \left(\frac{1}{|\mathcal{O}^i|} \sum_{v \in \mathcal{O}^i} x_v^2 + (y^i)^2 \right) + \sum_{t \in \mathcal{T}} \left(\frac{1}{|\mathcal{O}^t|} \sum_{v \in \mathcal{O}^t} (l_v^2 + u_v^2) + (z^t)^2 \right), \\ \text{subject to} \quad & \text{Equations (5.1)} - \text{(5.5)}, \\ & SPG \leq 55. \end{aligned}$$

5.2.4 Performance analysis

We evaluated four distinct KBP pipelines based on the plans they produced; predictions were also evaluated because they are an important intermediate step. We refer to the four sets of KBP plans as GAN-IP, RF-IP, GAN-DM, and RF-DM. We evaluated the predicted and plan dose distributions in terms clinical criteria, the difference in the performance of each optimization model when the same set of predictions is used as input, and the prediction error. Details of these performance metrics are presented below.

Criteria Satisfaction

We quantified the quality of KBP plans by how often they satisfied the same clinical criteria presented in Table 5.2. Specifically, we examined how often the plans satisfied the same criteria as the clinical plans in each class of criteria (i.e., OARs, targets, and all ROIs, which includes both OARs and targets). We also evaluated the quality of the prediction models to determine whether criteria satisfaction in the predicted dose distribution is an early indicator of final plan quality. Finally, we evaluated the difference in dose criterion between our KBP dose distributions and the reference clinical plans; the differences are visualized with box plots.

Table 5.2: The planning criteria used for evaluation: $\mathcal{D}_{99}$ is the minimum dose to 99% of the structure volume, $\mathcal{D}_{mean}$ is the mean dose to a structure, and $\mathcal{D}_{max}$ is the maximum dose to a structure.

Structure	Criteria
Brainstem	$\mathcal{D}_{max} \leq 54$ Gy
Spinal Cord	$\mathcal{D}_{max} \leq 48$ Gy
Right Parotid	$\mathcal{D}_{mean} \leq 26$ Gy
Left Parotid	$\mathcal{D}_{mean} \leq 26$ Gy
Larynx	$\mathcal{D}_{mean} \leq 45$ Gy
Esophagus	$\mathcal{D}_{mean} \leq 45$ Gy
Mandible	$\mathcal{D}_{max} \leq 73.5$ Gy
PTV56	$\mathcal{D}_{99} \geq 53.2$ Gy
PTV63	$\mathcal{D}_{99} \geq 59.9$ Gy
PTV70	$\mathcal{D}_{99} \geq 66.5$ Gy

Optimization performance differences

For each clinical planning criterion (Table 5.2), we evaluated the difference in dose between plans generated with an identical set of predictions but a different optimization model; the differences between the two optimization models (i.e., IP and DM) are visualized with a box plot. We then used a two-sided Mann-Whitney U test to determine if plans generated by IP were the same (null hypothesis) or different (alternative hypothesis) from those generated by DM for the population of plans generated from each set of predictions. For these and all subsequent hypothesis tests, $p < 0.01$ was considered significant.

Prediction performance differences

We evaluated the error of each prediction method by evaluating the median absolute difference between the predicted and clinical dose distributions across each ROI for every out-of-sample plan. The error is visualized with a box plot and we used a two-sided Mann-Whitney U test to determine if GAN had the same (null hypothesis) or a different (alternative hypothesis) prediction error than RF.

5.3 Results

Criteria Satisfaction

Table 5.3 summarizes the performance of the predicted and plan dose distributions. RF-DM plans achieved the same OAR criteria as the clinical plans most often (83.9%). However, GAN-IP plans satisfied the target criteria 28.8% more often than RF-DM plans, and achieved close to RF-DM performance on the OAR criteria (4.6 percentage points less). Across all ROIs, the proportion of GAN-IP plans satisfied the same criteria as the corresponding clinical plans 17% more often than its closest competitor (RF-IP). Additionally, while GAN-IP performed better than RF-IP, GAN-DM performed worse than RF-DM, which suggests that there is an interaction effect between the prediction and optimization model that must be accounted for.

In Table 5.3, we also compare the predictions to the clinical plans. We emphasize that unlike the generated plans, i.e., IP and DM plans, the predictions are only an intermediate step in the KBP pipeline. Here, we found that GAN predictions exhibited poor performance on all OAR criteria (24.1%) which we attribute to the poor performance on the mandible criteria (13.6%). The performance of RF and GAN predictions over all target criteria was similar. Overall, RF predicted that plans could satisfy the same criteria as the clinical plans in 78.2% of cases, which far exceeded GAN predictions (24.1%). Most importantly, however, these results do not carry through to the final plans. That is, only GAN-IP plans achieved the same proportion of All ROI criteria (78.2%) that was predicted by RF.

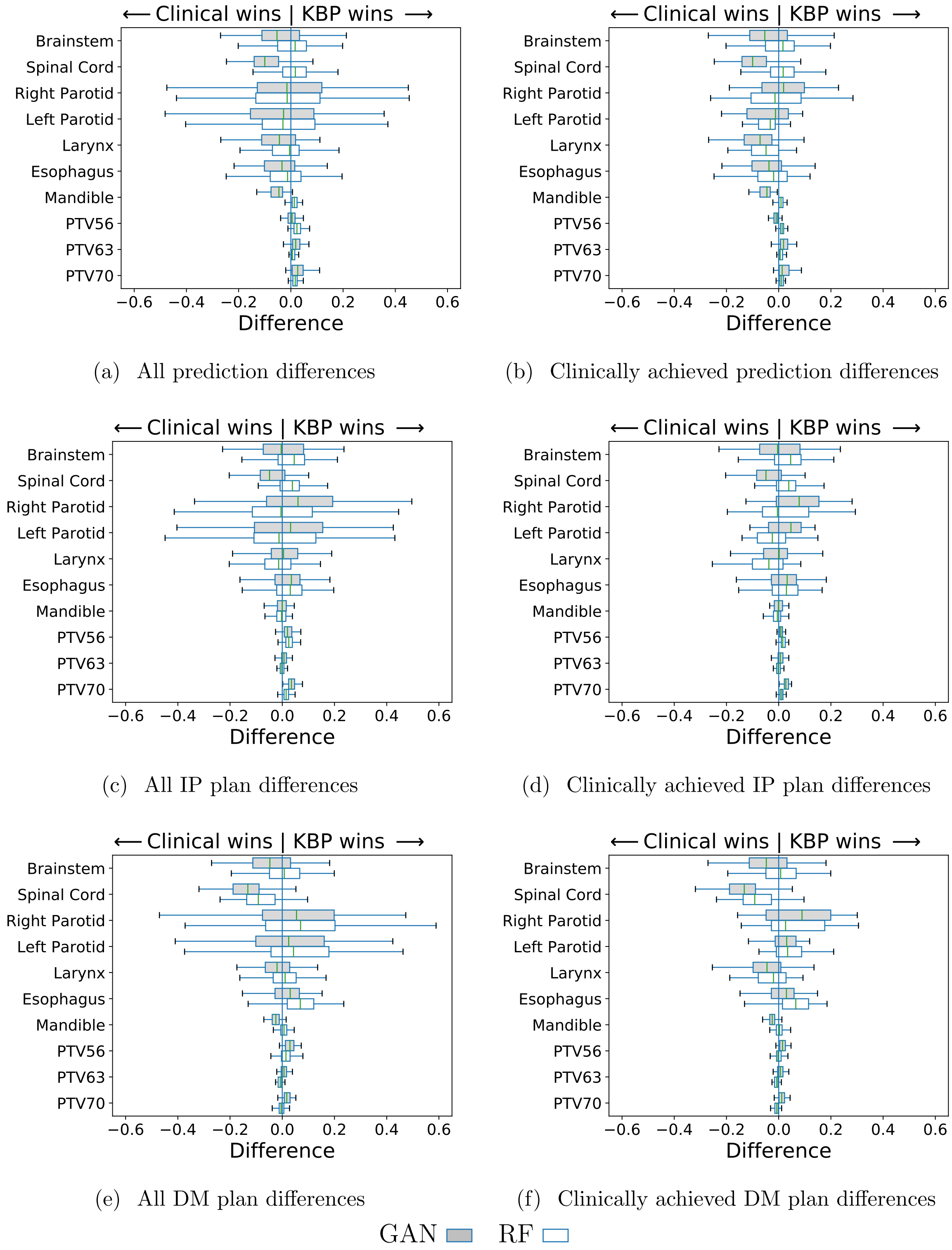
Table 5.3: The percentage of clinical plans that satisfied each criteria is summarized in the column "Clinical". The other columns summarize the percentage of KBP dose distributions that satisfied the same clinical criteria as the clinical plans. The rows under the "All" heading summarize the the percentage of KBP dose distributions that satisfied all clinical criteria in the corresponding group that were satisfied by the clinical plans. Only IP and DM plans use the full KBP pipeline.

		Predictions		IP plans		DM plans	
	Clinical	GAN	RF	GAN-IP	RF-IP	GAN-DM	RF-DM
OARs							
Brainstem	100.0	100.0	98.9	100.0	100.0	100.0	98.9
Spinal Cord	100.0	97.7	100.0	100.0	100.0	90.8	97.7
Right Parotid	20.5	64.7	64.7	94.1	76.5	76.5	82.4
Left Parotid	12.8	81.8	54.5	81.8	54.5	81.8	90.9
Larynx	61.3	71.4	89.8	91.8	87.8	81.6	93.9
Esophagus	94.0	100.0	100.0	100.0	100.0	100.0	100.0
Mandible	75.9	13.6	100.0	81.8	74.2	36.4	93.9
Targets							
PTV56	52.9	100.0	97.8	97.8	95.7	100.0	93.5
PTV63	100.0	100.0	98.0	100.0	98.0	100.0	100.0
PTV70	66.7	100.0	100.0	100.0	98.3	98.3	58.6
All							
OARs	–	24.1	80.5	79.3	69.0	41.4	83.9
Targets	–	100.0	97.7	98.9	95.4	98.9	70.1
ROIs	–	**24.1**	**78.2**	**78.2**	**66.7**	**41.4**	**56.3**

In Figure 5.4, we present six box plots to compare the difference between the KBP dose distributions and their respective clinical plans for each criterion; the reported differences are relative to the dose threshold of that criterion. We include plots for two categorizations: Figure 5.4 (a, c, e) the differences over all criteria and Figure 5.4 (b, d, f) the differences over only criteria that the clinical plans achieved. We found that IP is generally better than DM at transforming poor predictions into competitive plans. For example, the median GAN prediction for the mandible criterion was 0.045 (3.3 Gy) worse than the clinical plans. However, GAN-IP used those predictions to generate plans that were only 0.001 (0.1 Gy) worse than clinical plans; this was 0.024 (1.8 Gy) better than the GAN-DM plans. These plots also show the interquartile range (IQR) of differences in criteria satisfaction. They demonstrate that smaller IQR values occur over criteria that the clinical plans achieved, i.e., criteria that are high priority in practice. This suggests that the models will learn whether there is implicit consensus amongst oncologists as to the relative importance of criteria.

Optimization performance differences

In Figure 5.5, we present two box plots to compare the quality of plans from different optimization models when the same prediction model was used as input. The plots show how the plans generated by IP differ from those generated by DM in terms of the dose delivered to each clinical planning criterion relative to the dose threshold of that criterion. On average, IP was better than DM by 2% when GAN predictions were used as input. However, we found no difference between plans generated by IP and DM when RF predictions were used as input. We also found that IP performed better than DM in 69.5% and 50.8% of all evaluation criteria when the inputs were from GAN and RF, respectively. Statistically, when the GAN predictions were used as input, the plans generated by IP and DM performed differently on the clinical criteria ($p < 0.001$). However, we observed no difference ($p = 0.045$) when RF predictions were used as input to the optimization models. Overall, we observed that the performance of each optimization model was dependent on the prediction model that was used.



(a) All prediction differences

(b) Clinically achieved prediction differences

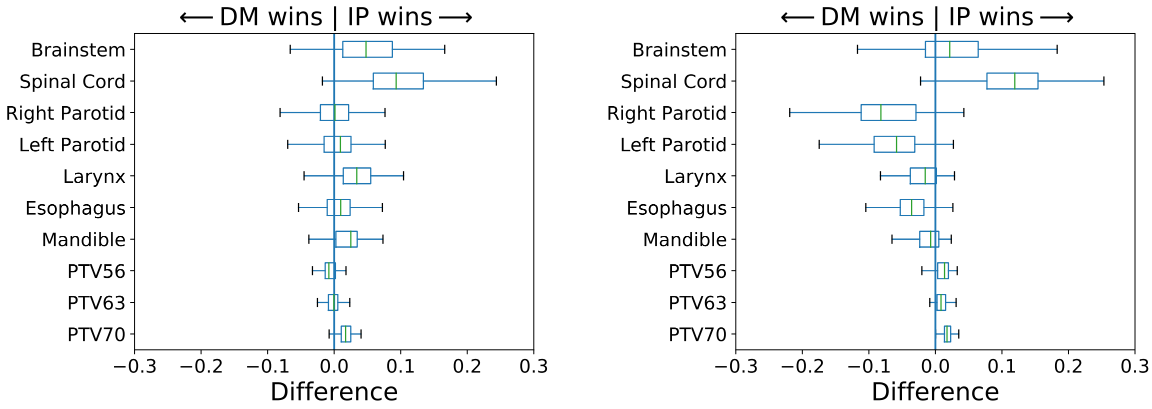
(c) All IP plan differences

(d) Clinically achieved IP plan differences

(e) All DM plan differences

(f) Clinically achieved DM plan differences

Figure 5.4: The distribution of clinical criteria differences between KBP dose and clinical dose over all ROIs for (a) predictions, (c) IP plans, and (e) DM plans; and only ROIs where clinical plan achieved the criteria for (b) predictions, (d) IP plans, and (f) DM plans. The boxes indicate median and IQR. Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier.



(a) GAN plan differences

(b) RF plan differences

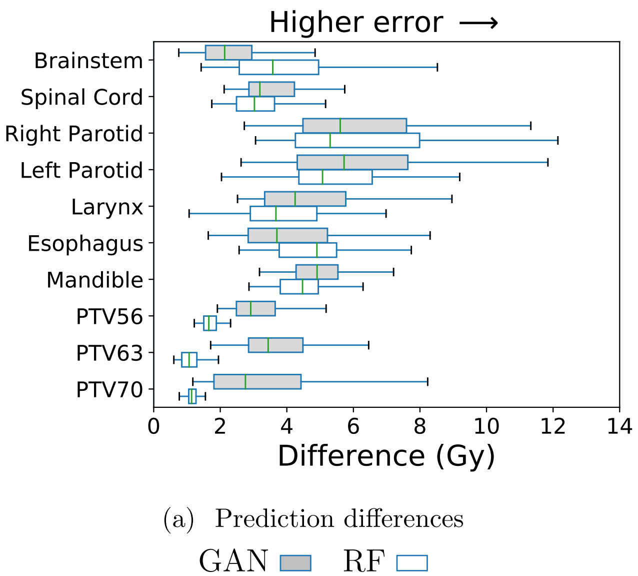
Figure 5.5: The difference in terms of clinical planning criteria between plans generated by IP and DM where the input to both models are (a) GAN predictions and (b) RF predictions. Positive differences imply that the IP plan was better than the DM plan in that criterion. The boxes indicate median and IQR. Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier.

Prediction performance differences

In Figure 5.6, we present the distribution of mean absolute differences between the predicted and clinical dose over the regions of interest (i.e., the mean absolute error between the predictions and clinical plans). Although both models had the same median prediction error across all OARs (4.3 Gy), RF error across targets (1.3 Gy) was much lower than GAN error (3.0 Gy). Overall, GAN predictions had higher median error across all ROIs (3.9 Gy) than RF predictions (3.6 Gy), and these predictions errors were significantly different ($p < 0.001$).

5.4 Discussion

Historically, each stage of KBP has been developed in isolation with a focus on improving the prediction stage. In this paper, we show that there are interaction effects between the prediction and optimization stages of KBP that significantly affect the quality of the generated plans. Our experimental setup consists of four KBP pipelines that were assem-



(a) Prediction differences

Figure 5.6: The distribution of average dose differences between KBP prediction and clinical dose over all ROIs. The boxes indicate median and IQR. Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier.

bled from two existing KBP methods, i.e., Babier et al.[10] and McIntosh and Purdie[75] (see Figure 5.2). Overall, the best performing combination of prediction and optimization methods was the GAN and IP. However, we also demonstrate that predictions that produce good plans with one optimization model (e.g., GAN-IP) do not always produce good plans with another optimization model (e.g., GAN-DM).

Although both RF and GAN predict 3D dose distributions, they differ in their approach. RF predicts the dose to each voxel independently of every other voxel. In contrast, GAN predicts the dose to all voxels simultaneously, thereby making predictions that are conditioned on the predictions of neighboring voxels. RF generally produces predictions that are more similar to clinical plans on summary statistics like mean absolute dose difference (Figure 5.6). This is likely because GAN optimizes a regularized loss function that encourages realistic looking images. This results in predictions that have worse performance on summary statistics as compared to a model like RF that minimizes the squared difference between predictions and the ground truth without regularization.

The quality of deliverable plans depends heavily on the combination of the prediction and optimization components used to construct the KBP pipeline. For example, combining GAN and IP results in plans that perform well on average in terms of satisfying clinical criteria. Interestingly, the KBP pipelines that perform the best contain stages that use same order loss and objective functions (i.e., linear-linear or quadratic-quadratic). Namely, GAN (trained with mean absolute loss) and IP (optimized with a linear objective function) produce the best IP plans. Similarly, RF (trained with mean squared loss) and DM (optimized with a quadratic objective function) produce the best DM plans.

When considering OAR and target criteria satisfaction as the two key components, we observe that there is no single two-stage KBP pipeline that dominates all others. While GAN-IP performs at least as well as RF-IP and GAN-DM on both metrics, RF-DM outperforms GAN-IP on OAR criteria satisfaction. We conjecture that because the PTV70 and the mandible are physically close together and their respective criterion compete with each other (i.e., the PTV70 criterion needs a high dose while the mandible criterion needs a low dose), the GAN struggles to learn an acceptable compromise. Variation in this compromise across the clinical plans may also add to the struggle. As a consequence, the model generally performs well on the PTV70 criterion but poorly on the mandible criterion.

Inverse planning includes a specific objective that minimizes the dose to the mandible, so even if the predictions (incorrectly) assume that mandible criterion satisfaction is unimportant, the mandible objective in IP improves mandible criterion performance. In contrast, dose mimicking attempts to construct dose distributions that are no worse than the predictions (in terms of Equations (5.1)–(5.5)), which generally leads to less improvement on the mandible criterion. Due to the biased nature of GAN predictions towards the mandible, IP can help to improve the single weak criterion with minimal expense to the other criteria. On the other hand, IP is unable to correct any significant under-performance of the RF predictions, which generally already perform well across most criteria (i.e., it achieves the same criteria as the clinical plans).

A limitation of our work is that, although we identified that the prediction and optimization stages of KBP affect the overall quality of the plans they generate, we were unable to isolate the root cause of those effects. Additionally, the data for this work is from a cohort of patients treated at a single institution. Due to variability in clinical practice between institutions, the performance of these models may vary for a different cohort of patients. A second limitation is that the computational resources required for this analysis scales exponentially with the number of prediction and optimization models considered. As a result, it is computationally intensive to determine what existing optimization model should be paired with a new approach to dose prediction (and vice versa).

5.5 Conclusion

This study demonstrates that the performance of an automated KBP pipeline is dependent on how well the prediction and optimization models perform together. As a result, we recommend that new prediction methods should be tested with multiple optimization models before they are considered to be state-of-the-art (and vice versa).

Chapter 6

OpenKBP: An open framework for plan optimization

In Chapter 5 we determined that the choice of both stages in knowledge-based planning (KBP) can contribute to considerable variation in quality. In this chapter, we explore these interaction effects further with the dose prediction models that were developed in the OpenKBP Grand Challenge, which we described in Chapter 4. Additionally, we develop four new plan optimization models and demonstrate a clear link between dose mimicking and inverse optimization. The data and code for this project is published to enable other researchers to test new KBP plan optimization models on a large set of dose prediction models.

6.1 Introduction

Automated radiotherapy planning is transforming clinical practice and personalized cancer treatment.[80] The most common type of automated planning is knowledge-based planning, which leverages knowledge derived from historical clinical treatment plans to generate new treatment plans without human intervention.[34;59;77] Most common KBP methods can be thought of as a two-stage pipeline that first predicts the dose that should be delivered to a patient, and then converts that prediction into a treatment plan via optimization (Figure 6.1). Both stages of this pipeline, which are active areas of research, can significantly affect the quality of generated treatment plans.[11] The contributions of this paper are twofold: 1) to provide data that supports KBP optimization research at scale and 2) to establish a connection between dose mimicking (a type of KBP optimization) and conventional planning methods. We expand on the impact of these contributions throughout this paper.

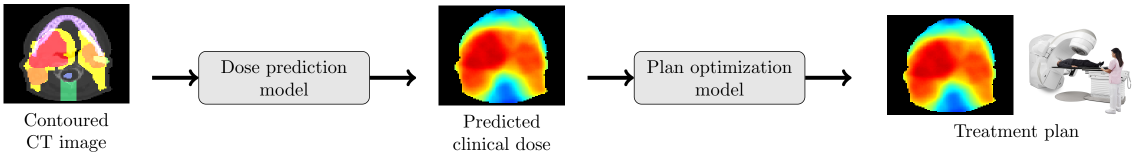


Figure 6.1: Overview of a complete knowledge-based planning pipeline.

Comparing the quality of competing KBP models from the research community is difficult because the vast majority of research is conducted with large private datasets, as noted in several reviews.[46;56;79;103] To help address this issue, the Open Knowledge-Based Planning (OpenKBP) Grand Challenge was organized to facilitate the largest international effort to date for developing and comparing dose prediction models on a single open dataset[13] The OpenKBP dataset, which includes data for 340 head-and-neck patients undergoing intensity modulated radiotherapy (IMRT), is limited to dose prediction research (i.e., it is incompatible with KBP optimization research). Although there are still no open datasets for KBP optimization research, there are two open datasets that support research in other areas of plan optimization.[25;36] However, it is challenging to use these datasets in KBP plan optimization research for two reasons. First, neither dataset includes dose predictions, which are the input to KBP plan optimization models. Second, they are smaller (123 patients across both datasets), span multiple sites (prostate, liver, head-and-neck), and multiple modalities (CyberKnife, volumetric modulated arc therapy, proton therapy, IMRT). While such a diversity in cases is important to demonstrate the robustness and generalizability of optimization algorithms across sites and modalities, this same diversity is a disadvantage when it comes to training dose prediction models, since there is insufficient data for any one site-modality pair.[23]

Most KBP pipelines are developed as fully-automated pipelines that can replace human treatment planners in the planning process.[15;44;76;106] These approaches have demonstrated promising results in prospective research studies where a sizeable portion of KBP-generated plans were considered inferior to human-generated plans, which suggests that there is an opportunity for improvement.[34;77] In those cases, making manual adjustments to the KBP-generated plan is non-trivial because they are generated by *fully-automated* pipelines that rely on the quality of the data. In contrast to fully automated pipelines, *semi-automated* pipelines rely on both the quality of data and human expertise, which puts less reliance on the data. For example, a semi-automated KBP pipeline could enable human planners to improve upon a KBP-generated plan via an intuitive process (e.g., inverse planning) and thereby provide a pipeline that leverages human expertise, models, and data. In the KBP literature, however, there are relatively few papers that describe tools that humans can intuitively interact with in semi-automated KBP pipeline.[8;20;64;114]

In this paper, we extend the results from the OpenKBP Grand Challenge, which we call OpenKBP, with an international validation of 76 KBP pipelines. We made this extension, which we call OpenKBP-Opt, open to provide a benchmark for KBP optimization research and to lower the barriers for contributing to this research area. We also demonstrate how KBP plan optimization models can be used to initialize the conventional planning process (i.e., inverse planning) with good patient-specific parameters (i.e., objective weights) and provide the means for a semi-automated KBP pipeline. Identifying this relationship provides a mechanism for transforming existing KBP optimization models, which are generally fully-automated pipelines that impede manual intervention, into semi-automated pipelines that promote human planners to improve upon a KBP-generated plan via inverse planning (i.e., a familiar and intuitive process). The data and code to reproduce this paper is publicly available at https://github.com/ababier/open-kbp-opt.

6.2 Methods and materials

Figure 6.2 summarizes the overall methodological approach into five components. The first three components (i.e., processing data, developing dose prediction models, and generating KBP dose predictions) are based on the results of the OpenKBP Grand Challenge. The final two components (i.e., developing plan optimization models and generating KBP treatment plans) are an extension of the OpenKBP Grand Challenge and the focus of this paper. Below, we describe all five components and our analysis.

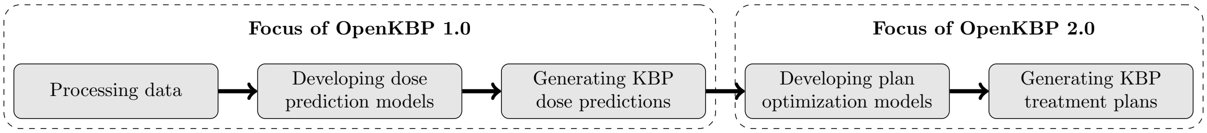

Focus of OpenKBP 1.0
Focus of OpenKBP 2.0
Processing data
Developing dose prediction models
Generating KBP dose predictions
Developing plan optimization models
Generating KBP treatment plans

Figure 6.2: An overview of our methods. A full description of each component is provided in the corresponding subsection.

6.2.1 Processing data

We obtained data for 340 patients ($n = 340$) with head-and-neck cancer from the OpenKBP Grand Challenge. The data consisted of a training set ($n = 200$), a validation set ($n = 40$), and a testing set ($n = 100$). The plans were delivered via 6 MV step-and-shoot IMRT from nine equidistant coplanar beams at angles $0°$, $40°$, $\ldots$, $320°$. Those beams were divided into a set of beamlets $\mathcal{B}$, which make up a fluence map. The relationship between the intensity w_b of beamlet b and dose d_v deposited to voxel v was determined using the influence matrix $D_{v,b}$ generated by the IMRTP library from A Computational Environment for Radiotherapy Research[38] using MATLAB, and it is given by

$$d_v = \sum_{b \in \mathcal{B}} D_{v,b} w_b. \tag{6.1}$$

6.2.2 Developing dose prediction models

All dose prediction models used in this paper were developed in the OpenKBP Grand Challenge.[13] During the challenge, teams developed dose prediction models using identical training and validation datasets with access only to ground truth data (i.e., dose) for the training set. Every dose prediction model used a neural network architecture that was based on either a U-Net,[90] V-Net,[78] or Pix2Pix[57] architecture. Many of the best performing models also used other generalizable techniques like ensembles,[85] one-cycle learning,[118] radiotherapy-specific loss functions,[49] and deep supervision.[66]

All teams competed to develop models that minimize one of two pre-defined error metrics that quantified the difference between the reference dose and a KBP-generated dose (i.e., KBP prediction or plan dose). The metrics were: 1) dose error, which was the mean absolute voxel-by-voxel difference between two dose distributions, and 2) dose-volume histogram (DVH) error, which was the absolute difference between a DVH point from two dose distributions. The DVH error was evaluated on two and three DVH points for each organ-at-risk (OAR) and target, respectively. The OAR DVH points were the $\mathrm{D_{mean}}$and $\mathrm{D_{0.1cc}}$, which was the mean dose delivered to the OAR and the maximum dose delivered to 0.1cc of the OAR, respectively. The target DVH points were the $\mathrm{D_{1}}$, $\mathrm{D_{95}}$, and $\mathrm{D_{99}}$, which was the dose delivered to 1% (99th percentile), 95% (5th percentile), and 99% (1st percentile) of voxels in the target, respectively. The models were ranked according to: 1) dose score, which was the average dose error of a model, and 2) DVH score, which was the average DVH error of a model.

6.2.3 Generating KBP dose predictions

In this paper, the OpenKBP organizers collaborated with teams that competed in the OpenKBP Grand Challenge. The 28 teams that completed the final phase of the OpenKBP Grand Challenge were invited to participate in the OpenKBP-Opt project, and 21 of those teams agreed to participate. We obtained the dose predictions from all teams for each patient in the test set to create a set of 2100 dose predictions (21 different predictions for each of the 100 patients). We observed that two models produced dose scores that were over two standard deviations (6.3 Gy) above the mean (4.0 Gy), whereas the rest were within half a standard deviation (1.6 Gy) of the mean. Thus, we omitted those two outlier models and proceeded with only 19 KBP models ($n = 1900$ predictions).

6.2.4 Developing plan optimization models

Next, we formulated four dose mimicking models, which are a type of KBP optimization model. Each model used the same set of structures and objective functions that we described in Section 6.2.4.1 and Section 6.2.4.2, respectively. However, they differ in how they mimic (i.e., penalize differences) a specific dose distribution. In particular, they each have a different cost function, outlined in Section 6.2.4.3. Note that in this paper the terms "objective function" and "cost function" refer to distinct concepts, and the cost functions in this paper are functions of objective functions.

6.2.4.1 Structures

All of our optimization models used the same set of regions-of-interest (ROIs) $\mathcal{R}_p$ for each patient $p \in \mathcal{P}$ in our test set. The set $\mathcal{R}_p$ contains OARs $\mathcal{I}_p$, targets $\mathcal{T}_p$, and optimization structures $\mathcal{O}_p$. The OARs contained in $\mathcal{I}_p$ were the brainstem, spinal cord, right parotid, left parotid, larynx, esophagus, and mandible. Each target $t \in \mathcal{T}_p$ was a planning target volume (PTV) with a dose level θ_t, and those targets were the PTV56, PTV63, and PTV70. The optimization structures contained in $\mathcal{O}_p$ were the limPostNeck, which was used to limit dose to the posterior neck, and six PTV ring structures (a 3 mm ring and a 6 mm ring for each target). These were the same structures used to generate the plans in the original OpenKBP dataset.[13] Every ROI $r \in \mathcal{R}_p$ was also divided into a set of voxels $\mathcal{V}_r$.

6.2.4.2 Objective functions

Our models used the objective functions in Table 6.1. Each objective function quantified a different measure of the dose delivered to a single ROI $r \in \mathcal{R}_p$ in a patient $p \in \mathcal{P}$, which we call an objective value. Specifically, the average and maximum objective values quantified the average dose and maximum dose delivered to an ROI r, respectively. The high and low conditional value at risk (CVaR) objective values quantified the average dose in ROI r that was higher and lower, respectively than the dose threshold f.

Table 6.1: The formulations for objective functions in our models.

Name	Objective function
Average dose	$\underset{v\in\mathcal{V}_r}{\text{mean}}\{d_v\}$
Maximum dose	$\max_{v\in\mathcal{V}_r}\{d_v\}$
High CVaR dose	$\underset{v\in\mathcal{V}_r}{\text{mean}}\{\max\{0,\ d_v - f\}\}$
Low CVaR dose	$\underset{v\in\mathcal{V}_r}{\text{mean}}\{\max\{0,\ f - d_v\}\}$

In total, we considered 107 objectives functions: seven per OAR, three per target, and seven per optimization structure. The objective functions for each OAR were the mean dose; maximum dose; and high CVaR dose with thresholds f equal to 0.25, 0.50, 0.75, 0.90, and 0.975 of the maximum predicted dose to that structure. The objective functions for each target were the maximum dose, low CVaR dose with a threshold equal to the dose level of the target (i.e., $f = \theta_t$), and a high CVaR dose with a threshold f equal to 1.05 of the dose level of the target (i.e., $f = 1.05\theta_t$). The objective functions for each optimization structure were the same as the OAR objective functions. Not all patients had all ROIs, so some models had fewer than 107 objective functions.

6.2.4.3 Model formulations

Our KBP optimization models performed dose mimicking to generate plans with optimized objective values that closely matched the input objective values from a dose prediction. To streamline our model formulation, let each $m \in \mathcal{M}_p$ denote one of the 107 objective functions (as outlined in Section 6.2.4.2). Let g_m and $\hat{g}_m$ be objective values of their corresponding objective functions evaluated over the optimized plan and predicted dose, respectively. In all models, the cost functions were formulated such that lower values of g_m were favored over higher values.

Table 6.2 presents the cost functions of our dose mimicking models. Each model minimized either the mean or max difference between all corresponding pairs $\{g_m, \hat{g}_m\}$ of the objective values, which were quantified via an absolute (e.g., $g_m - \hat{g}_m$) or relative (e.g., $(g_m - \hat{g}_m)/\hat{g}_m$) difference measure, resulting in four dose mimicking models. In the mean difference models, we chose to prioritize the positive differences (i.e., where the optimized plan objective value was higher than the predicted dose objective value) more than the negative differences, which we assigned a small positive weight ϵ ($\epsilon = 0.0001$ in our experiments). This was done to incentivize the model to do at least as well as the dose prediction before striving to outperform the dose prediction on other objective functions. In contrast, the max difference models used only a single term because the max difference naturally incentivizes the model to outperform the prediction only once the plan outperforms the prediction across all objective values (i.e., when $g_m \leq \hat{g}_m, \forall m \in \mathcal{M}_p$).

The main constraint in all four models was a constraint to limit plan complexity. In particular, the sum-of-positive gradients (SPG)[35] of all plans generated by the models was constrained to be less than or equal to 65, which was a constraint in the reference plans.[13] The remaining constraints were simply auxiliary constraints (including auxiliary

Table 6.2: The cost functions for each model that minimize mean absolute (MeanAbs), max absolute (MaxAbs), mean relative (MeanRel), and max relative (MaxRel) differences between the optimized and predicted objective values $\{g_m, \hat{g}_m\}$.

	Optimization model cost function
MeanAbs	$\underset{m\in\mathcal{M}_p}{\text{mean}}\,(g_m - \hat{g}_m)^{+} \;+\; \epsilon\, \underset{m\in\mathcal{M}_p}{\text{mean}}\,(g_m - \hat{g}_m)^{-}$
MaxAbs	$\max_{m\in\mathcal{M}_p}\,(g_m - \hat{g}_m)$
MeanRel	$\underset{m\in\mathcal{M}_p}{\text{mean}}\left(\frac{g_m-\hat{g}_m}{\hat{g}_m}\right)^{+} \;+\; \epsilon\, \underset{m\in\mathcal{M}_p}{\text{mean}}\left(\frac{g_m-\hat{g}_m}{\hat{g}_m}\right)^{-}$
MaxRel	$\max_{m\in\mathcal{M}_p}\left(\frac{g_m-\hat{g}_m}{\hat{g}_m}\right)$

variables) used to linearize both the objective and cost functions (i.e., the formulations in Table 6.1 and Table 6.2). The optimization models were all formulated in Python 3.7 using OR-Tools 8.2 and solved using Gurobi 9.1 (Gurobi Optimization, TX, US) on a single computer with an Intel i7-8700K (6-Core 3.7 GHz) CPU and 16 GB of random access memory. Default parameters were used with the Gurobi solver except for *Crossover* set to 0, *Method* set to 2, and *BarConvTol* set to 0.0001, which were selected based on past experience to improve solve time without compromising solution quality.

6.2.5 Generating KBP treatment plans

Next, we assembled 76 KBP pipelines by combining the 19 dose prediction models with each of the four dose mimicking models. Each pipeline was applied to the 100 patients in the testing set, resulting in 7600 KBP plans (see Figure 6.3). We used these plans in our analysis to measure the quality of the respective KBP models. We refer to the four plans generated from each prediction as MeanAbs, MaxAbs, MeanRel, and MaxRel plans.

Altogether, after completing the process in Figure 6.3, we had dose distributions for a set of reference plans ($n = 100$), predictions ($n = 1900$), and KBP plans generated by four dose mimicking models ($n = 4 \times 1900$). The reference plans are the plans that were released as part of the OpenKBP Grand Challenge, and the predictions are dose

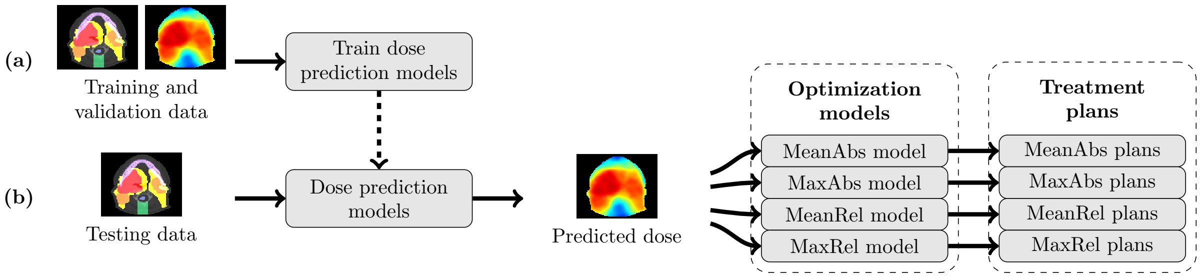


Figure 6.3: An overview of our process where (a) dose prediction models were trained on out-of-sample data and (b) those models were used to predict dose for input to dose mimicking optimization models to generate KBP plans.

distributions that were submitted by 19 teams in the final testing phase of OpenKBP. In general, there will be differences between the reference plan, prediction, and KBP plan dose distributions. Differences between a dose prediction and its corresponding KBP plan are due to factors including prediction noise and deliverability of the dose prediction. Differences between a KBP plan and its corresponding reference plan reflect different trade-offs in the cost function used to generate these plans.

6.2.6 Analysis

We conducted three analyses to measure model performance in terms of dose error, DVH point differences, and clinical criteria satisfaction. We also investigated the theoretical connection between our dose mimicking models and inverse planning. Finally, we summarized empirical optimization metadata.

6.2.6.1 Dose score and error

We evaluated the KBP models using the dose score and dose error as defined in Section 6.2.2. We calculated the Spearman rank order correlation of the dose score between the prediction models and corresponding KBP pipelines. The distribution of dose error was visualized using a box plot. A one-sided Wilcoxon signed-rank test was used to determine whether the dose error of the optimization models was the same (null hypothesis) or lower (alternative hypothesis) than the dose predictions models. For all hypothesis tests in this paper, $P < 0.05$ was considered significant.

6.2.6.2 DVH point differences

To measure the relative quality of dose distributions from a clinical perspective, we examined the distribution of DVH point differences between the reference and KBP-generated dose. The differences were evaluated over the DVH points listed in Section 6.2.2 and visualized using boxplots. We used the one-sided Wilcoxon signed-rank test to determine whether the dose generated by all optimization models performed the same (null hypothesis) or better (alternative hypothesis) than the dose predictions. This test was chosen to evaluate the aggregate performance of all optimization models relative to the predictions. Lower values were better for $\mathrm{D}_{\mathrm{mean}}$, $\mathrm{D}_{0.1\mathrm{cc}}$, and D_1; higher values were better for D_{95} and D_{99}.

6.2.6.3 Expected criteria satisfaction

As another measure of plan quality, we examined the proportion of clinical criteria that were satisfied by the reference plans and KBP-generated dose. One criterion was evaluated for each ROI (see Table 6.3). We tabulated the proportion of criteria that were satisfied by the reference plans, dose predictions, MeanAbs plans, MaxAbs plans, MeanRel plans, MaxRel plans, and the plans from the KBP pipeline that satisfied the most clinical criteria overall. We also plotted the proportion of OAR, target, and all ROI clinical criteria that each of the 76 KBP pipelines achieved.

6.2.6.4 Theoretical analysis of dose mimicking models

To justify our choice of dose mimicking models, we conducted a theoretical analysis into their structure using linear programming duality theory.[18] This analysis was based on previous literature that showed a connection between Benson's method,[17] which identifies

Table 6.3: The clinical criteria that we used to evaluate dose distributions. Before evaluating these criteria, we reinstated any overlap between targets that was removed.

Structures	Criteria
OARs	
Brainstem	$D_{0.1cc} \leq 50.0$ Gy
Spinal cord	$D_{0.1cc} \leq 45.0$ Gy
Right parotid	$D_{mean} \leq 26.0$ Gy
Left parotid	$D_{mean} \leq 26.0$ Gy
Esophagus	$D_{mean} \leq 45.0$ Gy
Larynx	$D_{mean} \leq 45.0$ Gy
Mandible	$D_{0.1cc} \leq 73.5$ Gy
Targets	
PTV56	$D_{99} \geq 53.2$ Gy
PTV63	$D_{99} \geq 59.9$ Gy
PTV70	$D_{99} \geq 66.5$ Gy

efficient solutions to multi-objective optimization models, and estimating the weights for inverse planning.[30] We were motivated to conduct a similar analysis as in Chan *et al.*[30] because our dose mimicking models are similar to the formulations in Benson 1978.[17] In particular, we linearized the dose mimicking models, took their duals, and related the dual variables to objective function weights in a conventional multi-objective planning problem depicted in model (6.2).

$$\begin{aligned} \underset{g}{\text{minimize}} \quad & \sum_{m \in \mathcal{M}_p} \hat{\alpha}_m g_m, \\ \text{subject to} \quad & \text{SPG} \leq 65, \\ & \text{Auxiliary constraints to linearize functions in Table 6.1 and 6.2.} \end{aligned} \tag{6.2}$$

6.2.6.5 Optimization metadata

Lastly, we summarized the metadata that each optimization model generated. In particular, we evaluated the average proportion of objective weight that each model assigned to OAR, target, and optimization structure objective functions. Additionally, we recorded the average, first quartile, and third quartile solve time.

6.3 Results

In this section, we summarize the performance of the 19 dose predictions models, four dose mimicking models, and 76 KBP pipelines.

6.3.1 Dose error and score

Table 6.4 summarizes the rank order correlation between the dose prediction models and their corresponding KBP pipelines. We found that the rank of a prediction model is positively correlated with its corresponding KBP pipeline rank. However, there was a wide range in correlation from 0.50 to 0.62. This demonstrates that high quality predictions are correlated with high quality plans, but this result also indicates that a prediction model that outperforms a competitor will not always generate better plans. Additionally, the KBP plans generated by an optimization model that evaluated relative differences (i.e., MeanRel and MaxRel) achieved higher rank order correlations than their counterparts that evaluated absolute differences (i.e., MeanAbs and MaxAbs).

Table 6.4: Each KBP optimization model is compared to the predictions in terms of median rank change and rank order correlation.

	MeanAbs	MaxAbs	MeanRel	MaxRel
Rank order correlation	0.53	0.50	0.62	0.59
Rank order P-value	0.019	0.030	0.005	0.008

The dose errors of predictions and KBP plans are shown in Figure 6.4. Two of the four sets of KBP plans had a median dose error that was lower than the median dose error of the predictions (2.79 Gy), implying that it is possible for optimization models to generate dose distributions that more closely resemble the reference plan dose, compared to dose predictions. These two models also achieved a significantly lower error ($P \leq 0.001$) than predictions. The MaxAbs model achieved the lowest median dose error (2.34 Gy).

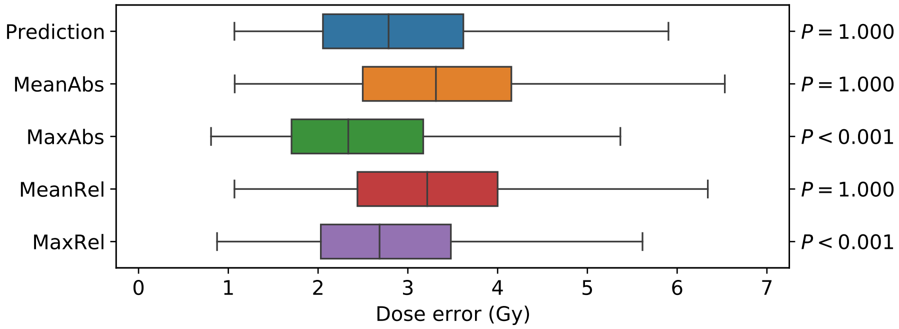


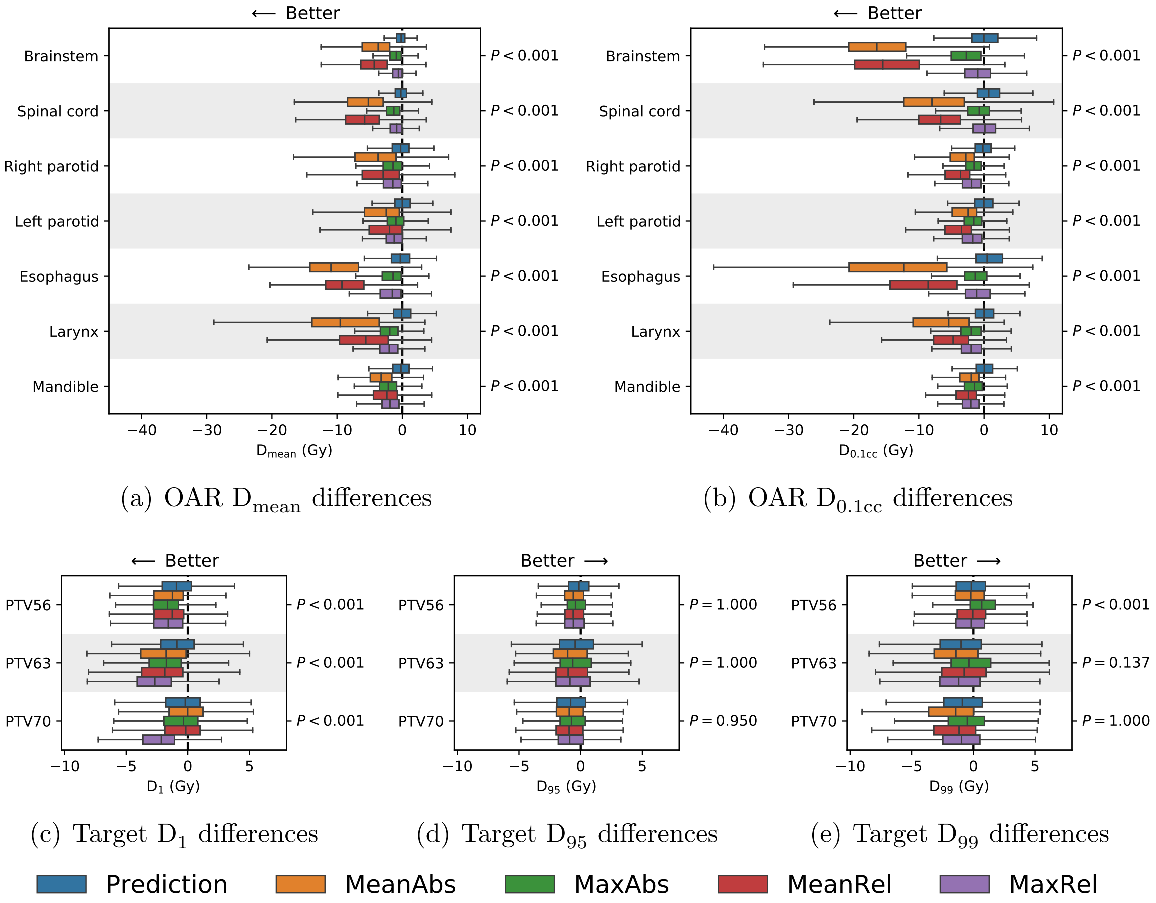
Figure 6.4: The distribution of dose error over all KBP-generated dose ($n = 1900$ points in each box). Boxes indicate median and interquartile range (IQR). Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier.

6.3.2 DVH point differences

Figure 6.5 shows the DVH point differences between the reference dose and either the predicted dose or KBP plan dose. In general, dose mimicking tends to produce a plan dose that is significantly better than the dose it received as input from a dose prediction model. In particular, the KBP plan dose is significantly better on 18 of the 23 DVH points than the predicted dose (all OAR points and four target points). The five DVH points where the plans were not significantly better are the three D_{95} points and two D_{99} points.

6.3.3 Expected criteria satisfaction

In Table 6.5, we compare the percentage of criteria that were satisfied by the reference plans ($n = 100$), the predictions ($n = 1900$), the plans generated by each of the four dose mimicking models ($n = 4 \times 1900$), and the plans generated by the top performing KBP pipeline ($n = 100$). We use the term baselines to refer to the reference dose and dose predictions collectively. The top performing KBP pipeline (denoted “Best” in Table 6.5) was defined as the single pipeline (i.e., the combination of one dose prediction model and one dose mimicking model) whose plans satisfied the most clinical criteria. Of all dose mimicking models, the MaxRel and MeanAbs models generated plans that satisfied



(a) OAR D_{mean} differences

(b) OAR $D_{0.1cc}$ differences

(c) Target D_1 differences

(d) Target D_{95} differences

(e) Target D_{99} differences

Figure 6.5: The distribution of DVH point differences between the reference dose and each set of KBP-generated dose. Negative differences indicate cases where the KBP-generated dose had a lower DVH points than the reference dose. The boxes indicate median and IQR. Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier.

the fewest (69.8%) and most (72.9%) ROI clinical criteria, respectively. For comparison, predictions only satisfied 66.2% of all clinical criteria, which was 3.5 percentage points lower than the reference plans (69.7%). The best KBP pipeline, which used the MeanAbs model and one of the 19 prediction models (discussed later), satisfied 77.0% of all ROI clinical criteria.

In general, clinical criteria satisfaction varied across each ROI criterion. The brainstem, spinal cord, esophagus, and mandible criteria were each satisfied more than 85% of the time across all the baselines and our dose mimicking models in Table 6.5. The right parotid, left parotid, and larynx were satisfied less than 40% of the time for the two baselines. In contrast, each of our four KBP models generated a higher average criteria satisfaction for these ROIs compared to the baselines. In fact, some were substantially higher. For example, the average criteria satisfaction of the MeanAbs model on the larynx was 71.5%, compared to an average of 36.2% for the baselines. In aggregate over all 19 prediction models, the performance of the four dose mimicking model was comparable or slightly worse than the reference dose in terms of criteria satisfaction in the targets. However, the best KBP pipeline outperformed the baselines on all criteria.

Table 6.5: The percentage of clinical criteria satisfied in each set of KBP-generated dose. Note that "Best" is defined as the top performing KBP pipeline that generated plans that satisfied the most ROI clinical criteria. The highest percentage of satisfied criteria is bolded in each row.

	Baselines		Dose mimicking models				
	Reference	Prediction	MeanAbs	MaxAbs	MeanRel	MaxRel	Best
OARs							
Brainstem	96.6	97.3	**100.0**	99.5	**100.0**	98.5	**100.0**
Spinal cord	95.5	92.7	99.7	97.3	**100.0**	95.6	**100.0**
Right parotid	32.3	32.7	**46.1**	38.9	45.0	38.0	41.4
Left parotid	30.6	30.1	**43.7**	35.0	41.9	35.0	40.8
Esophagus	93.0	92.7	**100.0**	95.2	**100.0**	97.3	**100.0**
Larynx	37.7	34.7	**71.5**	44.9	58.8	44.6	67.9
Mandible	87.5	89.4	**99.6**	98.7	99.2	99.0	93.1
Targets							
PTV56	91.2	85.8	83.3	91.8	84.1	84.6	**96.7**
PTV63	90.5	86.2	82.2	89.6	84.8	84.8	**92.9**
PTV70	64.0	45.7	37.2	51.6	40.1	47.7	**66.0**
All							
OARs	65.5	65.1	**77.1**	70.6	75.3	70.2	74.5
Targets	79.4	68.7	63.3	74.2	65.3	68.8	**82.8**
ROIs	69.7	66.2	72.9	71.7	72.3	69.8	**77.0**

Figure 6.6 summarizes the clinical criteria that were satisfied by each of the 76 KBP pipelines that we evaluated. The MeanAbs model generated plans that satisfied more criteria than the other three optimization models for 16 of the 19 dose prediction models

(see Figure 6.6(c)). Additionally, the pipelines that used better prediction models (i.e., dose score rank closer to 1) generally produced plans with higher criteria satisfaction. Interestingly, the best performing KBP pipeline (the last column of Table 6.5) used the dose prediction model that ranked 16th in terms of dose score. The spread in OAR criteria satisfaction across all 19 models (55.4% to 82.1%) was lower than that of target criteria satisfaction (24.5% to 89.7%), see Figure 6.6(a) and Figure 6.6(b), respectively. Note that the poor performing KBP pipelines used the 12th, 13th, 17th, 18th, and 19th ranked dose prediction models. Since the columns in Table 6.5 included all KBP pipelines, these poor performing models contributed to low performance on the target criteria. In contrast, many of the KBP pipelines that used the top ranked models prediction models clearly performed much better on target criteria.

6.3.4 Theoretical analysis of dose mimicking models

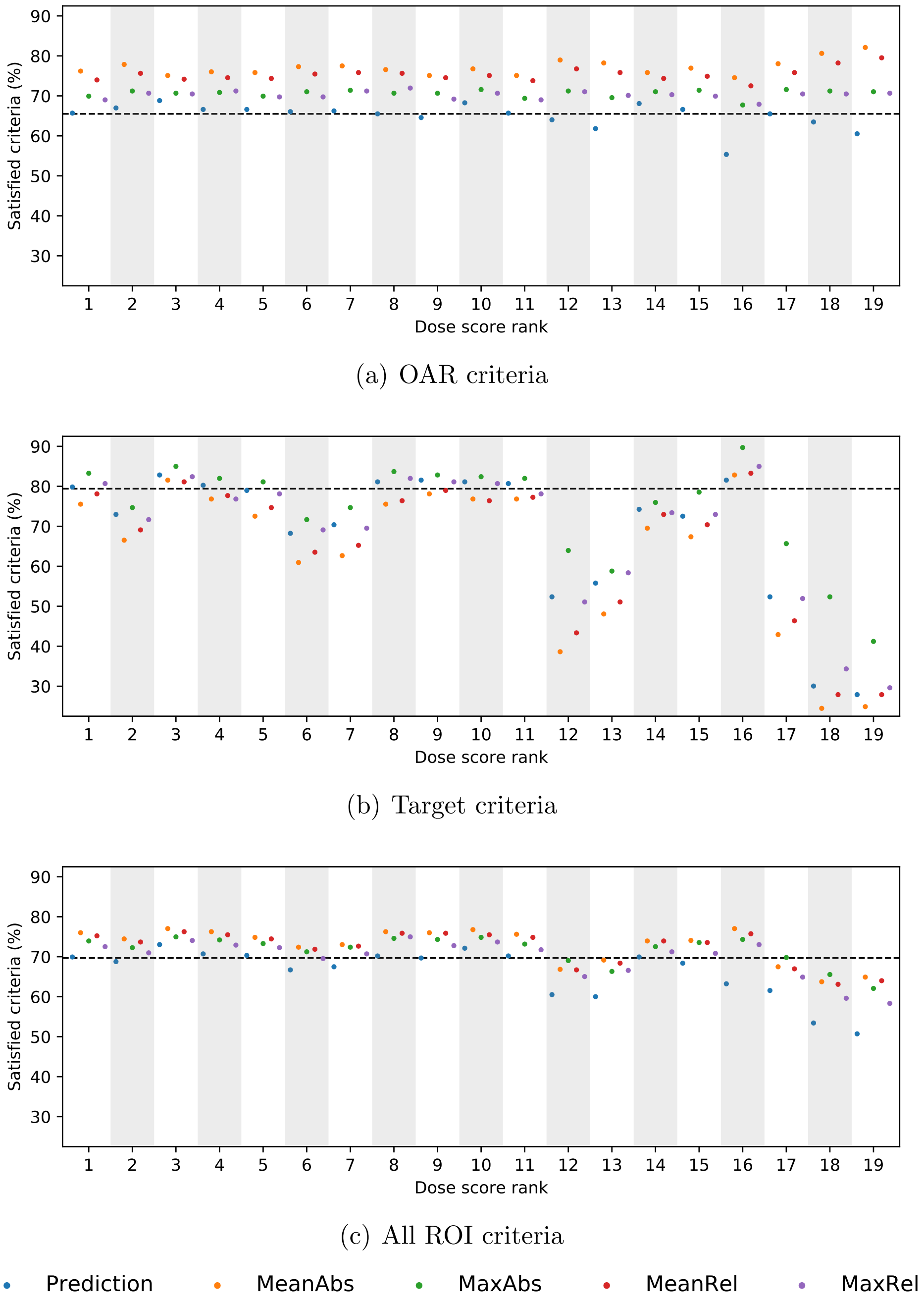
The inverse planning model (6.2) is shown in model (6.3) in vector and matrix notation following Chan *et al.*[30]

$$\begin{aligned}\underset{\mathbf{x}}{\text{minimize}} \quad & \hat{\boldsymbol{\alpha}}'\mathbf{C}\mathbf{x}\\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b},\\ & \mathbf{x} \geq \mathbf{0}.\end{aligned} \tag{6.3}$$

The objective functions are the rows of matrix $\mathbf{C}$ and the objective function weights are represented by the vector $\hat{\boldsymbol{\alpha}}$. The decision variables, which include the fluence variables ($w_b \ \forall b \in \mathcal{B}$) and auxiliary variables are represented by vector $\mathbf{x}$. The SPG and auxiliary constraints are encoded in the matrix $\mathbf{A}$ and vector $\mathbf{b}$.

Table 6.6 presents the formulations of the four dose mimicking models and their respective dual models. The positive and negative differences between optimized objective values $\mathbf{C}\mathbf{x}$ and predicted objective values $\mathbf{C}\hat{\mathbf{x}}$ are represented by vectors $\boldsymbol{\sigma}$ and $\boldsymbol{\delta}$, respectively. The max difference between the optimized and predicted objective values is expressed as a scalar ζ. The dual variables of the dose mimicking models are denoted $\boldsymbol{\alpha}$

and **p**. The vectors of all 0 and 1 are denoted by **0** and **e**, respectively. The symbol ⊙ denotes element-wise multiplication of vectors and prime denotes the transpose operator.



(a) OAR criteria

(b) Target criteria

(c) All ROI criteria

Figure 6.6: The percentage of (a) OAR, (b) Target, and (c) all ROI clinical criteria that were satisfied by each KBP pipeline. The points indicate the percentage of satisfied criteria. A dashed line indicates the percentage of criteria satisfied by reference plans.

Next, we complete our theoretical analysis. By Proposition 5 from Chan *et al.*,[30] it follows that an optimal decision vector $\mathbf{x}^*$ from each dose mimicking model is also optimal for the inverse planning model (6.3) with an optimal dual vector $\boldsymbol{\alpha}^*$ as objective weights (i.e., $\mathbf{x}^*$ is an optimal solution for model (6.3) when $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^*$). This result means that the solution to each dose mimicking model is also optimal to the inverse planning model with a particular set of objective function weights.

Table 6.6: The four KBP optimization models used in this paper in matrix notation with their corresponding dual models. Terms that follow colons indicate the dual variables for that constraint.

	KBP optimization model	Dual model
MeanAbs	$\min_{\mathbf{x},\boldsymbol{\sigma},\boldsymbol{\delta}} \quad \mathbf{e}'\boldsymbol{\sigma} + \epsilon\mathbf{e}'\boldsymbol{\delta}$ s.t. $\quad \mathbf{C}\mathbf{x} = \mathbf{C}\hat{\mathbf{x}} + \boldsymbol{\sigma} + \boldsymbol{\delta} \quad : \boldsymbol{\alpha}$ $\mathbf{A}\mathbf{x} = \mathbf{b} \quad : \mathbf{p}$ $\mathbf{x} \geq \mathbf{0}$ $\boldsymbol{\sigma} \geq \mathbf{0}$ $\boldsymbol{\delta} \leq \mathbf{0}$	$\min_{\boldsymbol{\alpha},\mathbf{p}} \quad \boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} - \mathbf{b}'\mathbf{p}$ s.t. $\quad \mathbf{C}'\boldsymbol{\alpha} \geq \mathbf{A}'\mathbf{p} \quad : \mathbf{x}$ $\boldsymbol{\alpha} \leq \mathbf{e} \quad : \boldsymbol{\sigma}$ $\boldsymbol{\alpha} \geq \epsilon\mathbf{e} \quad : \boldsymbol{\delta}$
MaxAbs	$\min_{\mathbf{x},\zeta} \quad \zeta$ s.t. $\quad \mathbf{C}\mathbf{x} \leq \mathbf{C}\hat{\mathbf{x}} + \zeta\mathbf{e} \quad : \boldsymbol{\alpha}$ $\mathbf{A}\mathbf{x} = \mathbf{b} \quad : \mathbf{p}$ $\mathbf{x} \geq \mathbf{0}$	$\min_{\boldsymbol{\alpha},\mathbf{p}} \quad \boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} - \mathbf{b}'\mathbf{p}$ s.t. $\quad \mathbf{C}'\boldsymbol{\alpha} \geq \mathbf{A}'\mathbf{p} \quad : \mathbf{x}$ $\boldsymbol{\alpha}'\mathbf{e} = 1 \quad : \sigma$ $\boldsymbol{\alpha} \geq \mathbf{0}$
MeanRel	$\min_{\mathbf{x},\boldsymbol{\sigma},\boldsymbol{\delta}} \quad \mathbf{e}'\boldsymbol{\sigma} + \epsilon\mathbf{e}'\boldsymbol{\delta}$ s.t. $\quad \mathbf{C}\mathbf{x} = \mathbf{C}\hat{\mathbf{x}} \odot (\mathbf{e} + \boldsymbol{\sigma} + \boldsymbol{\delta}) \quad : \boldsymbol{\alpha}$ $\mathbf{A}\mathbf{x} = \mathbf{b} \quad : \mathbf{p}$ $\mathbf{x} \geq \mathbf{0}$ $\boldsymbol{\sigma} \geq \mathbf{0}$ $\boldsymbol{\delta} \leq \mathbf{0}$	$\min_{\boldsymbol{\alpha},\mathbf{p}} \quad \boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} - \mathbf{b}'\mathbf{p}$ s.t. $\quad \mathbf{C}'\boldsymbol{\alpha} \geq \mathbf{A}'\mathbf{p} \quad : \mathbf{x}$ $\boldsymbol{\alpha} \odot \mathbf{C}\hat{\mathbf{x}} \leq \mathbf{e} \quad : \boldsymbol{\sigma}$ $\boldsymbol{\alpha} \odot \mathbf{C}\hat{\mathbf{x}} \geq \epsilon\mathbf{e} \quad : \boldsymbol{\delta}$
MaxRel	$\min_{\mathbf{x},\zeta} \quad \zeta$ s.t. $\quad \mathbf{C}\mathbf{x} \leq \mathbf{C}\hat{\mathbf{x}} \odot (\mathbf{e} + \zeta\mathbf{e}) \quad : \boldsymbol{\alpha}$ $\mathbf{A}\mathbf{x} = \mathbf{b} \quad : \mathbf{p}$ $\mathbf{x} \geq \mathbf{0}$	$\min_{\boldsymbol{\alpha},\mathbf{p}} \quad \boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} - \mathbf{b}'\mathbf{p}$ s.t. $\quad \mathbf{C}'\boldsymbol{\alpha} \geq \mathbf{A}'\mathbf{p} \quad : \mathbf{x}$ $\boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} = 1 \quad : \sigma$ $\boldsymbol{\alpha} \geq \mathbf{0}$

6.3.5 Optimization metadata

In Table 6.7, we present metadata that was generated by each optimization model, which assigned a different proportion of weight to the objectives for each group of ROIs (i.e., OARs, targets, optimization structures). The models that evaluate relative differences (i.e., MeanRel and MaxRel) spread the proportion of weight relatively evenly between the OAR and target objectives, but the other two models assigned the majority of the weight to target objectives with no more than 0.018 weight to OARs. Additionally, the optimization structures generally received the smallest proportion of weight with the exception of the MaxAbs model, which assigned more weight to optimization structure objectives (0.170) than OAR objectives (0.011). There is also a wide range in average solve time of the models (222 seconds to 393 seconds). On average, the MaxAbs model was the fastest.

Table 6.7: A summary of the metadata that each optimization model generated after optimizing 1900 plans.

	MeanAbs	MaxAbs	MeanRel	MaxRel
Objective weight				
OARs	0.018	0.011	0.554	0.417
Targets	0.976	0.819	0.418	0.569
Optimization	0.006	0.170	0.028	0.014
Solve time (s)				
Average	389	222	367	393
First quartile	192	107	183	188
Third quartile	502	261	481	507

6.4 Discussion

Knowledge-based planning research is flourishing. However, optimization models for KBP (e.g., dose mimicking) have received much less attention in the literature than dose prediction models. In this paper, we developed four dose mimicking models and evaluated their performance with 19 different dose prediction models, which were inputs to the optimization models. We showed that both the dose prediction model and optimization model contributed to considerable variation in the quality of plans generated by the corresponding KBP pipeline. Additionally, we conducted a theoretical investigation to show that our dose mimicking models generate plans that are optimal for a multi-objective inverse planning model with particular weights.

Our data and code is published at https://github.com/ababier/open-kbp-opt. to enable others to reproduce our results, which meets the gold standard in reproducibility.[51] Our data includes the first open dataset of predictions and reference plans to accompany CT images. We hope that this effort produces a common resource and lowers the barriers for future KBP optimization research, given that researchers must currently acquire their own private datasets and develop in-house prediction models before they can start testing new KBP optimization models.

Our open dataset contains the data for 100 patients who were treated with IMRT and a sample of high quality dose predictions for those same patients. The dataset was curated for the purpose of developing new fluence-based KBP optimization models that use ROI masks, dose influence matrices, and a dose prediction. The dose predictions were generated by 21 dose prediction models that were developed by an international group of researchers, which provided a diverse sample of realistic inputs for a KBP optimization model. Two of those prediction models (20th and 21th ranked model) were removed from our analysis because their dose scores were low, which we elaborated on in Section 6.2.3. For completeness, however, those 200 predictions are also available as part of our dataset.

We also performed a theoretical analysis to justify our dose mimicking models. Our key theoretical finding was that dose mimicking and conventional inverse planning are equivalent under certain specifications of the objective function weights. This allows us to interpret previous weight estimation techniques[30] through the more intuitive lens of dose mimicking models. Finally, by connecting dose mimicking to inverse planning, there is the potential to convert fully-automated KBP pipelines into semi-automated pipelines. Specifically, we use dose mimicking to generate a high-quality plan with its corresponding objective weights, which can be used in an inverse planning model (i.e., model (6.3)). This is advantageous because it enables human planners to improve the quality of plans generated by KBP via a conventional inverse planning process. By enabling this intuitive human interaction, we create a semi-automated KBP pipeline that is aligned with a common belief that AI will augment, rather than replace, the duties of healthcare practitioners[1].

Evaluating the performance of optimization models using many different dose predictions helps to identify interaction effects between these two stages of a KBP pipeline.[11] For example, the 16th ranked model generated lower quality predictions (in terms of dose error) than most of its competitors. However, when used in a KBP pipeline with the right optimization model, in this case the MeanAbs model, it generated high quality plans that achieved more clinical criteria than any other KBP pipeline. In other words, the errors made by the 16th ranked model that contribute to its low prediction quality were corrected by the KBP optimization model. Since these interaction effects contribute to considerable variation in quality, it is important to evaluate KBP optimization models across a diverse set of dose prediction models. Additionally, if we can understand what types of prediction error are most highly correlated with KBP plan quality we could propose better evaluation metrics to drive KBP prediction research towards making predictions that consistently translate into higher quality plans.

As in the original OpenKBP challenge, a limitation of this work is that we use synthetic dose distributions (i.e., the reference dose) as a substitute for real clinical dose. Although these dose distributions were subject to less quality assurance than clinical plans, they were previously shown to be of similar quality.[13] A second limitation of this work is that the dose prediction models were developed with the goal of optimizing the dose and DVH scores. There may be other scoring metrics that are better suited for developing a dose prediction model that excels in a KBP pipeline. This is a possible direction for future research. Lastly, this work only covers a single site and treatment modality. There is no guarantee that KBP optimization models that are developed with this dataset can generalize to other sites or treatment modalities.

6.5 Conclusion

Our large international experiment demonstrates that optimization models can consistently improve upon the predictions in a KBP pipeline. We also demonstrate that dose mimicking models can be reformulated for inverse planning, which provides the means for practitioners to improve upon KBP-generated plans via a familiar and intuitive process. The code and data to reproduce these results was made available in an effort to encourage more collaborative research in this field, which still has many unanswered questions.

Chapter 7

Conclusion

Incorporating artificial intelligence (AI) tools into radiotherapy is an effective means for improving cancer care, however, development of these tools is impeded by a lack of standardized metrics and datasets. In this thesis, we address these problems and improve upon existing AI tools for knowledge-based planning (KBP). Our tools involve the first implementation of computer vision in a KBP pipeline and improvements to existing KBP optimization models. We also develop standardized metrics and datasets for KBP research that advance innovation and support future reproducible research. In this chapter, we present a summary of our contributions, future directions for KBP research, and final remarks on the impact of this work.

7.1 Contributions

Throughout this thesis we made made several contributions that were often improved upon in successive chapters. In this section, we highlight the five main contributions of this thesis and indicate the chapters where these contributions were made.

In Chapters 2 and 3, we developed the first dose prediction models that use generative adversarial networks. These models were among the first dose prediction models that eschewed the paradigms of site-specific feature engineering and predicting low-dimensional representations of dose distributions. However, like all dose prediction models our GAN models only provide an intermediate step (i.e., the dose prediction) in a full KBP pipeline.

In Chapter 2 and 3, we also implemented the first KBP pipelines that use computer vision and optimization. These were the first studies that compared the performance of computer vision models in a full KBP pipeline to several baselines from the extent literature. Our dose prediction models outperformed the baseline models on several clinical metrics. However, these models were all developed on a private dataset with institution-specific evaluation metrics, which prevents reproducible research.

In Chapter 4, we garnered widespread adoption of standardized metrics for KBP research. This was accomplished by organizing a large competition that tasked participants to develop the best dose prediction model. The models were evaluated using standardized metrics that can be used to compare models developed in the future to a large set of baselines models.

In Chapter 4 and 6, we published the first open datasets for KBP research. Following the OpenKBP Grand Challenge, we published the corresponding datasets that we used to enable anyone to contribute to KBP research. A dataset was also released for plan optimization research. This contribution gave the field its first public dataset to encourage reproducible research on the full KBP pipeline.

In Chapter 5 and 6, we identified interaction effects between the stages of KBP that affect performance. Although the optimization model was largely held constant throughout this thesis, we found that the choice of both the prediction and the optimization model can contribute to considerable variation in quality. After making this observation we proposed a new set of optimization models and demonstrated that the two dominant KBP optimization models (i.e., inverse optimization and dose mimicking) are actually equivalent under certain conditions.

7.2 Future research directions

There are still several underexplored areas in KBP research. In this section, we highlight five major gaps in the KBP literature and propose future research directions for the field.

7.2.1 KBP plan optimization models

We need to improve KBP plan optimization methods to get better utility out of existing dose prediction models. Most of the attention in this field has been directed at improving dose prediction models, and the improvements to those models are becoming increasingly incremental. Although conventional plan optimization methods have also been well researched,[26] it is non-trivial to adapt them for KBP plan optimization. For example, there are unpredictable interaction effects between dose prediction and KBP plan optimization models,[11] which suggests that KBP plan optimization models may need to be tailored to perform well with specific dose prediction methods. Other efforts could be directed at exploring how to leverage the combined power of multiple dose predictions in a KBP optimization model, which is an idea that has already been shown to outperform models that consider only a single dose prediction.[12]

7.2.2 Extensions to other sites and modalities

Another future direction is the translation of knowledge-based planning approaches to new types of cancer and modalities. About 70% of all KBP research is conducted for prostate, lung, or head-and-neck cancers; and the majority of those studies use intensity modulated radiotherapy or volumetric modulated arc therapy.[79] This is a problem because not all KBP models will perform consistently across all cancer types and modalities.[75] Some KBP models have also been developed for other modalities like Gamma Knife,[67] brachytherapy,[86] and proton therapy.[39] However, there have been limited efforts to tailor KBP methods to the unique attributes of those modalities. For example, proton therapy is generally planned via *robust optimization*[101] to account for uncertainties in the treatment (e.g., range uncertainty),[100] but approaches for incorporating the elements of robust optimization into knowledge-based planning for proton therapy have only recently been considered.[42]

7.2.3 Improving KBP performance on small datasets

Knowledge-based planning is also a field that will likely be limited by small datasets for the foreseeable future. Although more clinical data may become available through efforts like federated learning,[89] the plans that oncologists approve are based on ever changing clincian-specific protocols that are driven by new scientific evidence and institution-specific standards. These changing protocols will continue to limit the pool of acceptable training data, which restricts the power of machine learning models in this field. There are three approaches that could address this challenge: (i) models that can learn to adjust to the new planning protocols without new data,[9] (ii) use of active learning to identify small subsets of data that can be used to retrain existing dose prediction methods to better adhere to new protocols,[72] or (iii) adopt more tools that enable semi-automated planning that can leverage the expertise of trained clinicians to adjust KBP-generated plans.[114]

7.2.4 Predicting fluence maps

Another interesting development is the rise in methods that predict fluence-based plans directly. Those methods bypass the optimization stage in KBP pipelines by predicting the decision variables of the optimization model (i.e., they predict beamlet intensities within a fluence map).[69;105] To date, these techniques perform poorly on complex sites like head-and-neck, but there are several areas for improvement (e.g., model architecture).[65] Models that predict fluence-based plans are also promising because they circumvent the unpredictable interaction effects between the dose prediction and plan optimization stages from the more common predict-then-optimize KBP pipeline.

7.2.5 Collaborating with open data

Finally, we should promote more open and collaborative research to make efficient progress like other successful AI-driven fields. Currently, most methods are developed on private datasets with institution specific metrics that leads to two major problems. First, doing fair comparisons between competing models in the literature that are developed and tested on different datasets is virtually impossible, which makes it difficult to track the meaningful improvements to models in our field. Second, there are large barriers in many areas of this field where only researchers with access to large private datasets can contribute. These issues can be addressed by simply publishing more open datasets for KBP research. Fortunately, there are small steps being made to improve access to data. For example, the Task Group 263 report from American Association of Physicists in Medicine introduces the concept of standardized nomenclature,[73] which should enable big data initiatives like the pooling of clinical data from several institutions. Such initiatives will reduce the number of issues (e.g., inconsistent nomenclature) associated with the multi-institutional data sharing efforts that are critical for improving the rate of innovation in knowledge-based planning research.

7.3 Final remarks

Overall, this thesis improves our understanding of KBP techniques and promotes more collaboration within our research community. We hope that this work encourages the field to do more reproducible research that produces useful AI tools for the radiotherapy treatment planning process. Although we are optimistic about the clinical impact of KBP, we also hope more attention will be given to developing semi-automated tools that encourage clinicians to work with KBP rather than relying on it as an infallible black-box.

Bibliography

[1] A. S. Ahuja. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, 7:e7702, 2019.

[2] R. K. Ahuja and J. B. Orlin. Inverse optimization. *Oper Res*, 49(5):771–783, 2001.

[3] V. Alex, M. K. P. Safwan, S. S. Chennamsetty, and G. Krishnamurthi. Generative adversarial networks for brain lesion detection. *Med Imaging*, 10133:113–121, 2017.

[4] T. A. Althunian, A. de Boer, R. H. H. Groenwold, and O. H. Klungel. Defining the noninferiority margin and analysing noninferiority: An overview. *Br J Clin Pharmacol*, 83(8):1636–1642, 2017.

[5] L. M. Appenzoller, J. M. Michalski, W. L. Thorstad, S. Mutic, and K. L. Moore. Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Med Phys*, 39(12):7446–7461, 2012.

[6] R. Atun, D. A. Jaffray, M. B. Barton, F. Bray, M. Baumann, B. Vikram, T. P. Hanna, F. M. Knaul, Y. Lievens, T. Y. M. Lui, M. Milosevic, B. O'Sullivan, D. L. Rodin, E. Rosenblatt, J. Van Dyk, M. L. Yap, E. Zubizarreta, and M. Gospodarowicz. Expanding global access to radiotherapy. *Lancet Oncol*, 16(10):1153–86, 2015.

[7] A. Babier, J. J. Boutilier, A. L. McNiven, and T. C. Y. Chan. Knowledge-based automated planning for oropharyngeal cancer. *Med Phys*, 45(7):2875–2883, 2018.

[8] A. Babier, J. J. Boutilier, M. B. Sharpe, A. L. McNiven, and T. C. Y. Chan. Inverse optimization of objective function weights for treatment planning using clinical dose-volume histograms. *Phys Med Biol*, 63(10):105004, 2018.

[9] A. Babier, T. C. Y. Chan, A. Diamant, and R. Mahmood. Learning to optimize with hidden constraints. *arXiv:1805.09293*, 2020.

[10] A. Babier, R. Mahmood, A. L. McNiven, A. Diamant, and T. C. Y. Chan. Knowledge-based automated planning with three-dimensional generative adversarial networks. *Med Phys*, 47(2):297–306, 2020.

[11] A. Babier, R. Mahmood, A. L. McNiven, A. Diamant, and T. C. Y. Chan. The importance of evaluating the complete automated knowledge-based planning pipeline. *Phys Med*, 72:73–79, 2020.

[12] A. Babier, T. C. Y. Chan, T. Lee, R. Mahmood, and D. Terekhov. An ensemble learning framework for model fitting and evaluation in inverse linear optimization. *INFORMS Journal on Optimization*, 3(2):119–138, 2021.

[13] A. Babier, B. Zhang, R. Mahmood, K. L. Moore, T. G. Purdie, A. L. McNiven, and T. C. Y. Chan. OpenKBP: The open-access knowledge-based planning grand challenge and dataset. *Med Phys*, 48(9):5549–5561, 2021.

[14] A. Babier, R. Mahmood, B. Zhang, V. G. L. Alves, A. M. Barragán-Montero, J. Beaudry, C. E. Cardenas, Y. Chang, Z. Chen, J. Chun, K. Diaz, H. D. Eraso, E. Faustmann, S. Gaj, S. Gay, M. Gronberg, B. Guo, J. He, G. Heilemann, S. Hira, Y. Huang, F. Ji, D. Jiang, J. C. J. Giraldo, H. Lee, J. Lian, S. Liu, K.-C. Liu, J. Marrugo, K. Miki, K. Nakamura, T. Netherton, D. Nguyen, H. Nourzadeh, A. F. I. Osman, Z. Peng, J. D. Q. Muñoz, C. Ramsl, D. J. Rhee, J. D. Rodriguez, H. Shan, J. V. Siebers, M. H. Soomro, K. Sun, A. U. Hoyos, C. Valderrama, R. Verbeek, E. Wang, S. Willems, Q. Wu, X. Xu, S. Yang, L. Yuan, S. Zhu,

L. Zimmermann, K. L. Moore, T. G. Purdie, A. L. McNiven, and T. C. Y. Chan. OpenKBP-Opt: An international and reproducible evaluation of 76 knowledge-based planning pipelines. *arXiv:2202.08303*, 2022.

[15] P. Bai, X. Weng, K. Quan, J. Chen, Y. Dai, Y. Xu, F. Lin, J. Zhong, T. Wu, and C. Chen. A knowledge-based intensity-modulated radiation therapy treatment planning technique for locally advanced nasopharyngeal carcinoma radiotherapy. *Radiat Oncol*, 15(1):188, 2020.

[16] M. B. Barton, S. Jacob, J. Shafiq, K. Wong, S. R. Thompson, T. P. Hanna, and G. P. Delaney. Estimating the demand for radiotherapy from the evidence: a review of changes from 2003 to 2012. *Radiother Oncol*, 112(1):140–4, 2014.

[17] H. P. Benson. Existence of efficient solutions for vector maximization problems. *J Optim Theory Appl*, 26(4):569–580, 1978.

[18] D. Bertsimas and J. N. Tsitsiklis. Introduction to linear optimization. *Athena Scientific*, 1997.

[19] D. Bodensteiner. Raystation: External beam treatment planning system. *Med Dosim*, 43(2):168–176, 2018.

[20] G. Bohara, A. Sadeghnejad Barkousaraie, S. Jiang, and D. Nguyen. Using deep learning to predict beam-tunable pareto optimal dose distribution for intensity-modulated radiation therapy. *Med Phys*, 47(9):3898–3912, 2020.

[21] J. M. Borras, Y. Lievens, P. Dunscombe, M. Coffey, J. Malicki, J. Corral, C. Gasparotto, N. Defourny, M. Barton, R. Verhoeven, L. van Eycken, M. Primic-Zakelj, M. Trojanowski, P. Strojan, and C. Grau. The optimal utilization proportion of external beam radiotherapy in European countries: An ESTRO-HERO analysis. *Radiother Oncol*, 116(1):38–44, 2015.

[22] W. R. Bosch, W. L. Straube, J. W. Matthews, and J. A. Purdy. Data from head-neck cetuximab. *The Cancer Imaging Archive*, 2015.

[23] J. J. Boutilier, T. Craig, M. B. Sharpe, and T. C. Y. Chan. Sample size requirements for knowledge-based treatment planning. *Med Phys*, 43(3):1212–21, 2016.

[24] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, 127(16):3029–3030, 2021.

[25] S. Breedveld and B. Heijmen. Data for TROTS - the radiotherapy optimisation test set. *Data Brief*, 12:143–149, 2017.

[26] S. Breedveld, D. Craft, R. van Haveren, and B. Heijmen. Multi-criteria optimization and decision-making in radiotherapy. *Eur J Oper Res*, 277(1):1–19, 2019.

[27] C. E. Cardenas, J. Yang, B. M. Anderson, L. E. Court, and K. B. Brock. Advances in auto-segmentation. *Semin Radiat Oncol*, 29(3):185–197, 2019.

[28] T. Carneiro, R. V. Medeiros Da Nobrega, T. Nepomuceno, G.-B. Bian, V. H. C. De Albuquerque, and P. P. Reboucas Filho. Performance analysis of google colaboratory as a tool for accelerating deep learning applications. *IEEE Access*, 6: 61677–61685, 2018.

[29] T. C. Y. Chan and T. Lee. Trade-off preservation in inverse multi-objective convex optimization. *Eur J Oper Res*, 270(1):25–39, 2018.

[30] T. C. Y. Chan, T. Craig, T. Lee, and M. B. Sharpe. Generalized inverse multiobjective optimization with application to cancer therapy. *Oper Res*, 62(3):680–95, 2014.

[31] T. C. Y. Chan, R. Mahmood, and I. Y. Zhu. Inverse optimization: Theory and applications. *arXiv 2109.03920*, 2021.

[32] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun. Generating multi-label discrete electronic health records using generative adversarial networks. *arXiv:1703.06490*, 2017.

[33] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J Digit Imaging*, 26(6):1045–57, 2013.

[34] M. Cornell, R. Kaderka, S. J. Hild, X. J. Ray, J. D. Murphy, T. F. Atwood, and K. L. Moore. Noninferiority study of automated knowledge-based planning versus human-driven optimization across multiple disease sites. *Int J Radiat Oncol Biol Phys*, 106(2):430–439, 2020.

[35] D. Craft, P. Suss, and T. Bortfeld. The tradeoff between treatment plan quality and required number of monitor units in intensity-modulated radiotherapy. *Int J Radiat Oncol Biol Phys*, 67(5):1596–605, 2007.

[36] D. Craft, M. Bangert, T. Long, D. Papp, and J. Unkelbach. Shared data for intensity modulated radiation therapy (IMRT) optimization research: the CORT dataset. *Gigascience*, 3(1):37, 2014.

[37] I. J. Das, V. Moskvin, and P. A. Johnstone. Analysis of treatment planning time among systems and planners for intensity-modulated radiation therapy. *J Am Coll Radiol*, 6(7):514–7, 2009.

[38] J. O. Deasy, A. I. Blanco, and V. H. Clark. CERR: a computational environment for radiotherapy research. *Med Phys*, 30(5):979–85, 2003.

[39] A. R. Delaney, L. Dong, A. Mascia, W. Zou, Y. Zhang, L. Yin, S. Rosas, J. Hrbacek, A. J. Lomax, B. J. Slotman, M. Dahele, and W. F. A. R. Verbakel. Automated

knowledge-based intensity-modulated proton planning: An international multicenter benchmarking study. *Cancers*, 10(11):420, 2018.

[40] G. Delaney, S. Jacob, C. Featherstone, and M. Barton. The role of radiotherapy in cancer treatment. *Cancer*, 104(6):1129–1137, 2005.

[41] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*, 20(5):533–534, 2020.

[42] O. Eriksson and T. Zhang. Predicting scenario doses for robust automated radiation therapy treatment planning. *arXiv:2110.09984*, 2021.

[43] C. Esteban, S. L. Hyland, and G. Rätsch. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv:1706.02633*, 2017.

[44] J. Fan, J. Wang, Z. Chen, C. Hu, Z. Zhang, and W. Hu. Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Med Phys*, 46(1):370–381, 2019.

[45] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[46] Y. Ge and Q. J. Wu. Knowledge-based planning for intensity-modulated radiation therapy: A review of data-driven approaches. *Med Phys*, 46(6):2760–2775, 2019.

[47] L. E. Gomez and P. Bernet. Diversity improves performance and outcomes. *J Natl Med Assoc*, 111(4):383–392, 2019.

[48] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Proceedings of NIPS*, 2: 2672–2680, 2014.

[49] M. P. Gronberg, S. S. Gay, T. J. Netherton, D. J. Rhee, L. E. Court, and C. E. Cardenas. Technical note: Dose prediction for head and neck radiotherapy using a three-dimensional dense dilated U-Net architecture. *Med Phys*, 48(9):5567–5573, 2021.

[50] A. J. Grossberg, A. S. R. Mohamed, H. Elhalawani, W. C. Bennett, K. E. Smith, T. S. Nolan, S. Chamchod, M. E. Kantor, T. Browne, K. A. Hutcheson, G. B. Gunn, A. S. Garden, S. J. Frank, D. I. Rosenthal, J. B. Freymann, and C. D. Fuller. Data from head and neck cancer CT atlas. *The Cancer Imaging Archive*, 2017.

[51] B. J. Heil, M. M. Hoffman, F. Markowetz, S.-I. Lee, C. S. Greene, and S. C. Hicks. Reproducibility standards for machine learning in the life sciences. *Nat Methods*, 18(10):1132–1135, 2021.

[52] R. Hermoza and I. Sipiran. 3D reconstruction of incomplete archaeological objects using a generative adversary network. *arXiv:1711.06363*, 2017.

[53] B. Hofstra, V. V. Kulkarni, S. Munoz-Najar Galvez, B. He, D. Jurafsky, and D. A. McFarland. The diversity-innovation paradox in science. *Proc Natl Acad Sci U S A*, 117(17):9284–9291, 2020.

[54] T. S. Hong, W. A. Tomé, and P. M. Harari. Heterogeneity in head and neck IMRT target design and clinical practice. *Radiother Oncol*, 103(1):92–8, 2012.

[55] J. Howard and S. Gugger. Fastai: A layered API for deep learning. *Inf*, 11(2):108, 2020.

[56] M. Hussein, B. J. M. Heijmen, D. Verellen, and A. Nisbet. Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations. *Br J Radiol*, 91(1092):20180270, 2018.

[57] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv:1611.07004*, 2017.

[58] D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis. Head and neck squamous cell carcinoma. *Nat Rev Dis Primers*, 6 (1):92, 2020.

[59] R. Kaderka, S. J. Hild, V. N. Bry, M. Cornell, X. J. Ray, J. D. Murphy, T. F. Atwood, and K. L. Moore. Wide-scale clinical implementation of knowledge-based planning: An investigation of workforce efficiency, need for post-automation refinement, and data-driven model maintenance. *Int J Radiat Oncol Biol Phys*, 111(3): 705–715, 11 2021.

[60] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, and A. Zhavoronkov. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol Pharm*, 14(9): 3098–3104, 2017.

[61] V. Kearney, J. W. Chan, S. Haaf, M. Descovich, and T. D. Solberg. Dosenet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Phys Med Biol*, 63(23):235022, 2018.

[62] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[63] A. Krizhevsky. Learning multiple layers of features from tiny images. *Tech Report*, 2009.

[64] N. Li, R. Carmona, I. Sirak, L. Kasaova, D. Followill, J. Michalski, W. Bosch, W. Straube, L. K. Mell, and K. L. Moore. Highly efficient training, refinement, and validation of a knowledge-based planning quality-control system for radiation therapy clinical trials. *Int J Radiat Oncol Biol Phys*, 97(1):164–172, 2017.

[65] X. Li, Q. J. Wu, Q. Wu, C. Wang, Y. Sheng, W. Wang, H. Stephens, F.-F. Yin, and Y. Ge. Insights of an AI agent via analysis of prediction errors: a case study of fluence map prediction for radiation therapy planning. *Phys Med Biol*, 66(23), 2021.

[66] S. Liu, J. Zhang, T. Li, H. Yan, and J. Liu. Technical note: A cascade 3D U-Net for dose prediction in radiotherapy. *Med Phys*, 48(9):5574–5582, 2021.

[67] Y. Liu, C. Shen, T. Wang, J. Zhang, X. Yang, T. Liu, S. Kahn, H.-K. Shu, and Z. Tian. Automatic inverse treatment planning for gamma knife radiosurgery via deep reinforcement learning. *arXiv:2109.06813*, 2021.

[68] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy. A technique for the quantitative evaluation of dose distributions. *Med Phys*, 25(5):656–661, 1998.

[69] L. Ma, M. Chen, X. Gu, and W. Lu. Deep learning-based inverse mapping for fluence map prediction. *Phys Med Biol*, 2020.

[70] R. Mahmood. *Learning to Solve Optimization Problems with Hidden Components: Applications in Automated Treatment Planning*. PhD thesis, University of Toronto, 2020.

[71] R. Mahmood, A. Babier, A. Mcniven, A. Diamant, and T. Chan. Automated treatment planning in radiation therapy using generative adversarial networks. *Proceedings of Machine Learning Research*, 85:1–15, 2018.

[72] R. Mahmood, S. Fidler, and M. T. Law. Low budget active learning via wasserstein distance: An integer programming approach. *arXiv:2106.02968*, 2021.

[73] C. S. Mayo, J. M. Moran, W. Bosch, Y. Xiao, T. McNutt, R. Popple, J. Michalski, M. Feng, L. B. Marks, C. D. Fuller, E. Yorke, J. Palta, P. E. Gabriel, A. Molineu, M. M. Matuszak, E. Covington, K. Masi, S. L. Richardson, T. Ritter, T. Morgas,

S. Flampouri, L. Santanam, J. A. Moore, T. G. Purdie, R. C. Miller, C. Hurkmans, J. Adams, Q.-R. Jackie Wu, C. J. Fox, R. A. Siochi, N. L. Brown, W. Verbakel, Y. Archambault, S. J. Chmura, A. L. Dekker, D. G. Eagle, T. J. Fitzgerald, T. Hong, R. Kapoor, B. Lansing, S. Jolly, M. E. Napolitano, J. Percy, M. S. Rose, S. Siddiqui, C. Schadt, W. E. Simon, W. L. Straube, S. T. St James, K. Ulin, S. S. Yom, and T. I. Yock. American association of physicists in medicine task group 263: Standardizing nomenclatures in radiation oncology. *Int J Radiat Oncol Biol Phys*, 100(4):1057–1066, 2018.

[74] C. McIntosh and T. G. Purdie. Contextual atlas regression forests: Multiple-atlas-based automated dose prediction in radiation therapy. *IEEE Trans Med Imaging*, 35(4):1000–12, 2016.

[75] C. McIntosh and T. G. Purdie. Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Phys Med Biol*, 62 (2):415–431, 2017.

[76] C. McIntosh, M. Welch, A. McNiven, D. A. Jaffray, and T. G. Purdie. Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. *Phys Med Biol*, 62(15):5926–5944, 2017.

[77] C. McIntosh, L. Conroy, M. C. Tjong, T. Craig, A. Bayley, C. Catton, M. Gospodarowicz, J. Helou, N. Isfahanian, V. Kong, T. Lam, S. Raman, P. Warde, P. Chung, A. Berlin, and T. G. Purdie. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat Med*, 27(6):999–1005, 2021.

[78] F. Milletari, N. Navab, and S. Ahmadi. V-net: Fully convolutional neural networks

for volumetric medical image segmentation. *In 2016 fourth international conference on 3D vision (3DV)*, pages 565–571, 2016.

[79] S. Momin, Y. Fu, Y. Lei, J. Roper, J. D. Bradley, W. J. Curran, T. Liu, and X. Yang. Knowledge-based radiation treatment planning: A data-driven method survey. *J Appl Clin Med Phys*, 22(8):16–44, 2021. doi: 10.1002/acm2.13337.

[80] K. L. Moore. Automated radiotherapy treatment planning. *Semin Radiat Oncol*, 29(3):209–218, 2019.

[81] National Science Foundation, National Center for Science and Engineering Statistics. Women, minorities, and persons with disabilities in science and engineering: 2019. Technical report, National Science Foundation, 2019.

[82] D. Nguyen, T. Long, X. Jia, W. Lu, X. Gu, Z. Iqbal, and S. Jiang. Dose prediction with U-Net: A feasibility study for predicting dose distributions from contours using deep learning on prostate IMRT patients. *arXiv:1709.09233*, 2017.

[83] D. Nguyen, X. Jia, D. Sher, M.-H. Lin, Z. Iqbal, H. Liu, and S. Jiang. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-Net deep learning architecture. *Phys Med Biol*, 64(6):065020, 2019.

[84] D. Nguyen, T. Long, X. Jia, W. Lu, X. Gu, Z. Iqbal, and S. Jiang. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci Rep*, 9(1):1076, 2019.

[85] D. Nguyen, A. Sadeghnejad Barkousaraie, G. Bohara, A. Balagopal, R. McBeth, M.-H. Lin, and S. Jiang. A comparison of monte carlo dropout and bootstrap aggregation on the performance and uncertainty estimation in radiation therapy dose prediction with deep learning neural networks. *Phys Med Biol*, 66(5):054002, 2021.

[86] A. Nicolae, G. Morton, H. Chung, A. Loblaw, S. Jain, D. Mitchell, L. Lu, J. Helou, M. Al-Hanaqta, E. Heath, and A. Ravi. Evaluation of a machine-learning algorithm for treatment planning in prostate low-dose-rate brachytherapy. *Int J Radiat Oncol Biol Phys*, 97(4):822–829, 2017.

[87] K. Petersson, P. Nilsson, P. Engström, T. Knöös, and C. Ceberg. Evaluation of dual-arc VMAT radiotherapy treatment plans automatically generated via dose mimicking. *Acta Oncologica*, 55(4):523–525, 2016.

[88] T. G. Purdie, R. E. Dinniwell, A. Fyles, and M. B. Sharpe. Automation and intensity modulated radiation therapy for individualized high-quality tangent breast treatment plans. *Int J Radiat Oncol Biol Phys*, 90(3):688–95, 2014.

[89] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso. The future of digital health with federated learning. *NPJ Digit Med*, 3:119, 2020.

[90] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *Proceedings of MICCAI*, 9351:234–241, 2015.

[91] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *Int J Comput Vis*, 115:211–252, 2015.

[92] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays. Scribbler: Controlling deep image synthesis with sketch and color. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5400–5409, 2017.

[93] M. H. F. Savenije, M. Maspero, G. G. Sikkes, J. R. N. van der Voort van Zyp, A. N. T J Kotte, G. H. Bol, and C. A. T van den Berg. Clinical implementation of MRI-

based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. *Radiat Oncol*, 15(1):104, 2020.

[94] M. B. Sharpe, K. L. Moore, and C. G. Orton. Point/counterpoint: Within the next ten years treatment planning will become fully automated without the need for human intervention. *Med Phys*, 41(12):120601, 2014.

[95] S. Shiraishi and K. L. Moore. Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy. *Med Phys*, 43(1):378, 2016.

[96] S. Shiraishi, J. Tan, L. A. Olsen, and K. L. Moore. Knowledge-based prediction of plan quality metrics in intracranial stereotactic radiosurgery. *Med Phys*, 42(2):908, 2015.

[97] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 71(3): 209–249, 2021.

[98] Z. C. Taskin, J. C. Smith, H. E. Romeijn, and J. F. Dempsey. Optimal multileaf collimator leaf sequencing in IMRT treatment planning. *Oper Res*, 58(3):674–690, 2010.

[99] United States Census Bureau. 2019 american community survey 1-year estimates, 2019. URL https://data.census.gov/cedsci/table?tid=ACSST1Y2019.DP05.

[100] J. Unkelbach, T. C. Y. Chan, and T. Bortfeld. Accounting for range uncertainties in the optimization of intensity modulated proton therapy. *Phys Med Biol*, 52(10): 2755–73, 2007.

[101] J. Unkelbach, M. Alber, M. Bangert, R. Bokrantz, T. C. Y. Chan, J. O. Deasy, A. Fredriksson, B. L. Gorissen, M. van Herk, W. Liu, H. Mahmoudzadeh, O. No-

hadani, J. V. Siebers, M. Witte, and H. Xu. Robust radiotherapy planning. *Phys Med Biol*, 63(22):22TR02, 2018.

[102] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, N. Khaouam, P. F. Nguyen-Tan, C.-S. Wang, and K. Sultanem. Data from head-neck-PET-CT. *The Cancer Imaging Archive*, 2017.

[103] M. Wang, Q. Zhang, S. Lam, J. Cai, and R. Yang. A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning. *Front Oncol*, 10:580919, 2020.

[104] N. Wang, W. Zha, J. Li, and X. Gao. Back projection: an effective postprocessing method for GAN-based face sketch synthesis. *Pattern Recognition Letters*, 2017.

[105] W. Wang, Y. Sheng, C. Wang, J. Zhang, X. Li, M. Palta, B. Czito, C. G. Willett, Q. Wu, Y. Ge, F.-F. Yin, and Q. J. Wu. Fluence map prediction using deep learning models - direct plan generation for pancreas stereotactic body radiation therapy. *Front Artif Intell*, 3:68, 2020.

[106] G. Wortel, D. Eekhout, E. Lamers, R. van der Bel, K. Kiers, T. Wiersma, T. Janssen, and E. Damen. Characterization of automatic treatment planning approaches in radiotherapy. *Phys Imaging Radiat Oncol*, 19:60–65, 2021.

[107] B. Wu, F. Ricchetti, G. Sanguineti, M. Kazhdan, P. Simari, M. Chuang, R. Taylor, R. Jacques, and T. McNutt. Patient geometry-driven information retrieval for IMRT treatment plan quality control. *Med Phys*, 36(12):5497–505, 2009.

[108] B. Wu, F. Ricchetti, G. Sanguineti, M. Kazhdan, P. Simari, R. Jacques, R. Taylor, and T. McNutt. Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys*, 79(4):1241–7, 2011.

[109] B. Wu, M. Kusters, M. Kunze-Busch, T. Dijkema, T. McNutt, G. Sanguineti, K. Bzdusek, A. Dritschilo, and D. Pang. Cross-institutional knowledge-based planning (KBP) implementation and its performance comparison to auto-planning engine (APE). *Radiother Oncol*, 123(1):57–62, 2017.

[110] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. *Proceedings of NIPS*, pages 82–90, 2016.

[111] T. Yang, E. C. Ford, B. Wu, M. Pinkawa, B. van Triest, P. Campbell, D. Y. Song, and T. R. McNutt. An overlap-volume-histogram based method for rectal dose prediction and automated treatment planning in the external beam prostate radiotherapy following hydrogel injection. *Med Phys*, 40(1):011709, 2013.

[112] K. C. Younge, R. B. Marsh, D. Owen, H. Geng, Y. Xiao, D. E. Spratt, J. Foy, K. Suresh, Q. J. Wu, F. Yin, S. Ryu, and M. M. Matuszak. Improving quality and consistency in NRG oncology radiation therapy oncology group 0631 for spine radiosurgery via knowledge-based planning. *Int J Radiat Oncol Biol Phys*, 100(4): 1067–1074, 2018.

[113] L. Yuan, Y. Ge, W. R. Lee, F. F. Yin, J. P. Kirkpatrick, and Q. J. Wu. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. *Med Phys*, 39(11):6868–78, 2012.

[114] T. Zhang, R. Bokrantz, and J. Olsson. Probabilistic pareto plan generation for semi-automated multicriteria radiation therapy treatment planning. *arxiv:2110.05410*, 2021.

[115] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv:1703.10593*, 2017.

[116] X. Zhu, Y. Ge, T. Li, D. Thongphiew, F. Yin, and Q. J. Wu. A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Med Phys*, 38(2):719–26, 2011.

[117] B. P. Ziemer, S. Shiraishi, J. A. Hattangadi-Gluth, P. Sanghvi, and K. L. Moore. Fully automated, comprehensive knowledge-based planning for stereotactic radiosurgery: Preclinical validation through blinded physician review. *Pract Radiat Oncol*, 7(6):e569–e578, 2017.

[118] L. Zimmermann, E. Faustmann, C. Ramsl, D. Georg, and G. Heilemann. Technical note: Dose prediction for radiation therapy using feature-based losses and one cycle learning. *Med Phys*, 48(9):5562–5566, 2021.

[119] M. L. Zuley, R. Jarosz, S. Kirk, Y. Lee, R. Colen, K. Garcia, and N. D. Aredes. Radiology data from the cancer genome atlas head-neck squamous cell carcinoma [TCGA-HNSC] collection. *The Cancer Imaging Archive*, 2016.

Appendix A

Supplement to Chapter 2

A.1 Network architecture

The general network architecture was adapted from Isola et al.[57]. Contoured CT slices were used as input to the generator as 3-channel, 128 × 128 images. We used a U-net architecture, where the generator was comprised of an encoder and a decoder stage. We used 4 × 4 2D convolutions with stride 2 and padding 1. Each convolution layer was followed by a leaky ReLU and batch normalization. Deconvolution layers were followed by 50% dropout, ReLU, and batch normalization.

The encoder consisted of four downsampling layers. The first generated 64 channels, and each subsequent layer downsampled by a factor of 2. This was followed by 2 bottleneck layers, before the data was then passed through 4 upsampling layers. The output of each downsample layer was concatenated to the input of the corresponding upsample layer. The final output was a 3-channel, 128 × 128 slice.

The decoder consisted of five convolution layers, where the first four each downsample the output by 2. The fifth, and last layer, mapped to a scalar output. Once again, we applied batch normalization and leaky ReLU after the first four layers. The final layer was passed through sigmoid activation.

A.2 Random forest architecture

The random forest used ten custom features outlined in Table A.2 to predict the dose delivered to each voxel in the patient. The RF was trained with ten trees, and default settings with the `randomForestRegressor` from `scikit-learn`.

Table A.1: The ten features used in the RF to predict the dose for any voxel.

Feature	Description
Structure	Structure that the voxel is classified as
y-coordinate	Voxel's positions on the y-axis in a slice
z-coordinate	Plane of voxel's slice
Distance to larynx	Shortest path between voxel and the surface of the larynx
Distance to esophagus	Shortest path between voxel and the surface of the esophagus
Distance to limPostNeck	Shortest path between voxel the surface of the limPostNeck
Distance to PTV56	Shortest path between voxel and the surface of the PTV56
Distance to PTV63	Shortest path between voxel and the the surface of PTV63
Distance to PTV70	Shortest path between voxel and the the surface of PTV70
Influence	Sum of influence matrix elements for the voxel

Appendix B

Supplement to Chapter 3

The generator and discriminator architectures for the 2D-dose and 3D-dose models are summarized in Table B.1 and B.2, respectively.

Table B.1: Overview of the generator architecture. 'BN' refers to batch normalization; 'LR', 'R', and 'Tanh' refer to Leaky ReLU (0.2 slope), ReLU, and Tanh activations, respectively; and 'D' refers to dropout.

Layer	Concatenate input with	2D-dose		3D-dose		Processing
		Blocks	Input size	Blocks	Input size	
1	—	Conv2d	128x128x3	Conv3d	128x128x128x3	BN-LR
2	—	Conv2d	64x64x64	Conv3d	64x64x64x64	BN-LR
3	—	Conv2d	32x32x128	Conv3d	32x32x32x128	BN-LR
4	—	Conv2d	16x16x256	Conv3d	16x 16x16x256	BN-LR
5	—	Conv2d	8x8x512	Conv3d	8x8x8x512	BN-LR
6	—	Conv2d	4x4x512	Conv3d	4x4x4x512	BN-LR
7	—	Conv2d	2x2x512	Conv3d	2x2x2x512	LR
8	—	Deconv2d	4x4x512	Deconv3d	4x4x4x512	BN-R
9	layer 6 output	Deconv2d	8x8x1024	Deconv3d	8x8x8x1024	BN-D-R
10	layer 5 output	Deconv2d	16x16x1024	Deconv3d	16x16x16x1024	BN-D-R
11	layer 4 output	Deconv2d	32x32x1024	Deconv3d	32x32x32x1024	BN-R
12	layer 3 output	Deconv2d	64x64x512	Deconv3d	64x64x64x512	BN-R
13	layer 2 output	Deconv2d	128x128x256	Deconv3d	128x128x128x256	BN-R
14	layer 1 output	Deconv2d	128x128x128	Deconv3d	128x128x128x128	Tanh
Output	—	—	128x128x1	—	128x128x128x1	—

Table B.2: Overview of the discriminator architecture. 'BN' refers to batch normalization while 'LR' and 'Sig' refer to Leaky ReLU (0.2 slope) and sigmoid activations, respectively.

Layer	2D-dose		3D-dose		Processing
	Blocks	Input size	Blocks	Input size	
1	Conv2d	128x128x4	Conv3d	128x128x128x4	LR
2	Conv2d	64x64x64	Conv3d	64x64x64x64	BN-LR
3	Conv2d	32x32x128	Conv3d	32x32x32x128	BN-LR
4	Conv2d	16x16x256	Conv3d	16x16x16x256	Sig
Output	—	1	—	1	—

Appendix C

Supplement to Chapter 4

C.1 Data Format

The data for OpenKBP is structured to facilitate the development and validation of dose prediction models. In this section, we describe how the data is stored and formatted.

C.1.1 Summary of data

The data for each patient is provided as comma-separated values (CSV) files, which are separated into directories with the corresponding patient number. The files for each patient include:

dose.csv the full 3D dose distribution that was used to treat the patient (in units of Gy).

ct.csv grey-scale images of the patient prior to treatment (in Hounsfield units). There is a mix of 12-bit and 16-bit formats, and we recommend clipping the CT values to be between 0 and 4095 (inclusive) to convert them all to 12-bit number formats, which is the more common convention.

voxels.csv The dimensions of the patient voxels (in units of mm).

possible_dose_mask.csv a mask of voxels that can receive dose (i.e., the dose will always be zero where this mask is zero).

Structure masks a mask that labels any voxel that is contained in the respective structure. The tensor for each structure is stored as a CSV file under its respective structure name. Only structures that were contoured in the patient have CSV files.

Brainstem.csv mask of brainstem voxels.

SpinalCord.csv mask of spinal cord voxels.

RightParotid.csv mask of right parotid voxels.

LeftParotid.csv mask of left parotid voxels.

Esophagus.csv mask of esophagus voxels.

Larynx.csv mask of larynx voxels.

Mandible.csv mask of mandible voxels.

PTV56.csv mask of PTV56 voxels.

PTV63.csv mask of PTV63 voxels.

PTV70.csv mask of PTV70 voxels.

C.1.2 Data format

Other than the file voxels.csv, which contains a list of only three numbers, all of the CSV data in OpenKBP is saved as sparse tensors (i.e., only non-zero values are stored). The advantage of this sparse format, compared to dense tensors (i.e., all values are stored), is that the data size is smaller and thus loads into memory faster, which leads to faster model training. The disadvantage, is that working with sparse tensors is less intuitive than working with dense tensors. In general, we recommend converting the data into dense tensors once it is loaded.

All of the sparse tensors are stored in CSV files with two columns. The first column contains a list of indices. The second column contains either a list of values for the corresponding indices, or it contains no values if the tensor is a mask (i.e., where all corresponding values are 1). All indices are stored as single numbers that unravel into a 3D (i.e., x-y-z) coordinate system via C-contiguous ordering. We provide Python code in our repository to load the data as dense tensors.

C.2 Surveys

In this section, we present the two mandatory surveys that we released during the Challenge. In each survey, respondents answered questions either by writing free-text or by selecting option(s) from a list.

Mandatory questions are marked with an asterisk (*).

C.2.1 Registration

All participants completed the following two part survey to register for OpenKBP.

C.2.1.1 Part 1: Professional information

Please complete this form to be given access to the OpenKBP competition.

First Name*

Short-answer text

Last Name*

Short-answer text

E-mail (must be the same address used for your CodaLab account)*

Short-answer text

Institution name without acronyms (University, Hospital, Company, etc.)*

Short-answer text

Department (Computer Science, Medical Biophysics, Radiation Oncology, Machine Learning, Industrial Engineering, etc.)*

Short-answer text

Primary research area*

○ Medical Physics

○ Machine Learning

○ Optimization

○ Other...

Have you done research in knowledge-based planning in the past?*

○ Yes

○ No

Position*

○ Student

○ Professor

○ Post doctoral fellow

○ Medical physicist

○ Radiation oncologist

○ Industry research

○ Other...

C.2.1.2 Part 2: Equity, diversity, and inclusion

To help us learn how to support the diversification of researchers in the OpenKBP Grand Challenge, we ask that all applicants complete an equity survey at the time of their registration. Equity is one of our competition's goals. We seek to remove barriers to participation for all people including women, LGBTQ individuals, persons with disabilities, Indigenous People and racialized persons and persons of colour.

Your participation is voluntary and your responses are confidential. We hope you will choose to answer these questions to help us bring you an even better competition next time. The information we receive from your responses will be used to better understand who has access to the competition, to identify barriers that may exist and areas to develop and/or improve in our rules and procedure to achieve more diversity and equity in the application process. All responses will be kept strictly confidential and will be reported only in aggregate so that you cannot be personally identified by your characteristics. **Do you self-identify as (choose all that apply):**

□ Man

□ Women

□ Transgender

□ Prefer not to say

□ Other...

Please indicate the racial or ethnic groups with which you identify (check all that apply):

□ African American/Black

□ Asian American/Asian

□ Hispanic/Latinx

□ Middle Eastern/North African

□ Native American/Indigenous

□ Native Hawaiian/Other Pacific Islander

□ White

□ Prefer not to say

□ Other...

Do you identify as a person with a disability? This may mean that either you: (i) have a long-term or recurring condition or health problem which limits the kind or amount of work you can do in the workplace; OR (ii) feel that you may be perceived as limited in the kind or amount of work which you can do because of a physical, mental, sensory, psychiatric, or learning impairment.

○ Yes, I identify as a person with a disability

○ No, I do not identify as a person with a disability

○ Prefer not to say

C.2.2 Model summary

Every team that competed in the testing phase of the Challenge also completed the following one part survey to summarize their model.

C.2.2.1 Model survey

Please describe your final dose prediction model using this survey. We will consider your submission complete only if this survey is submitted. Any submission made on CodaLab that is not associated with a survey response will be considered void, and it will not be ranked in the final leaderboard. We may also reach out to you for more information.

This survey includes 5 long answer questions, and we expect the cumulative word count of your responses to be about 350 words. We provide an estimated word count for your response to each question. These estimates are only a guide and you may provide more detail where you see fit. Please reach out if you have any questions or need clarification.

For teams, only one team member should submit this form.

Username on CodaLab*

Short-answer text

Team Name on CodaLab (enter N/A if you have no team)*

Short-answer text

Broadly speaking, how would you describe your model?*

○ Linear regression

○ Random forest

○ Neural network

○ Gradient boosted trees

○ Support vector machines

○ Other...

Briefly describe your approach. (∼ 150 words)*

Long-answer text

What would you say is the biggest contributing factor(s) to your models efficacy? (∼ 100 words)*

Long-answer text

Did you use transfer learning?*

○ Yes

○ No

Please describe any data augmentation methods that you used (e.g., rotations, private clinical dataset)? ($\sim$ 50 words)*

Long-answer text

Briefly describe your loss function. Did you use radiation therapy specific metrics in your loss function (e.g., max dose to PTV)? ($\sim$ 50 words)*

Long-answer text

Briefly describe the hardware (e.g., GPU model, CPU model) or cloud resources (e.g., Google Colab) that you used. ($\sim$ 25 words)*

Long-answer text

Please leave any other comments about your process here.

Long-answer text

Provide a link to the code repository that will recreate your model. We will include links to all the provided repositories from our existing OpenKBP Github to enable new users to build on a library of existing models. You may also provide a repository link at a later date, but it is not required.

Short-answer text