Henry Ford Health
Henry Ford Health Scholarly Commons

Radiation Oncology Articles | Radiation Oncology

9-12-2022

OpenKBP-Opt: an international and reproducible evaluation of 76 knowledge-based planning pipelines

Aaron Babier

Rafid Mahmood

Binghao Zhang

Victor G. L Alves

Ana Maria Barragán-Montero

See next page for additional authors

Follow this and additional works at: https://scholarlycommons.henryford.com/radiationoncology_articles

Recommended Citation

Babier A, Mahmood R, Zhang B, Alves VGL, Barragán-Montero AM, Beaudry J, Cardenas CE, Chang Y, Chen Z, Chun J, Diaz K, David Eraso H, Faustmann E, Gaj S, Gay S, Gronberg M, Guo B, He J, Heilemann G, Hira S, Huang Y, Ji F, Jiang D, Carlo Jimenez Giraldo J, Lee H, Lian J, Liu S, Liu KC, Marrugo J, Miki K, Nakamura K, Netherton T, Nguyen D, Nourzadeh H, Osman AFI, Peng Z, Darío Quinto Muñoz J, Ramsl C, Joo Rhee D, David Rodriguez J, Shan H, Siebers JV, Soomro MH, Sun K, Usuga Hoyos A, Valderrama C, Verbeek R, Wang E, Willems S, Wu Q, Xu X, Yang S, Yuan L, Zhu S, Zimmermann L, Moore KL, Purdie TG, McNiven AL, and Chan TCY. OpenKBP-Opt: an international and reproducible evaluation of 76 knowledge-based planning pipelines. Phys Med Biol 2022; 67(18).

This Article is brought to you for free and open access by the Radiation Oncology at Henry Ford Health Scholarly Commons. It has been accepted for inclusion in Radiation Oncology Articles by an authorized administrator of Henry Ford Health Scholarly Commons.

Authors

Aaron Babier, Rafid Mahmood, Binghao Zhang, Victor G. L Alves, Ana Maria Barragán-Montero, Joel Beaudry, Carlos E. Cardenas, Yankui Chang, Zijie Chen, Jaehee Chun, Kelly Diaz, Harold David Eraso, Erik Faustmann, Sibaji Gaj, Skylar Gay, Mary Gronberg, Bingqi Guo, Junjun He, Gerd Heilemann, Sanchit Hira, Yuliang Huang, Fuxin Ji, Dashan Jiang, Jean Carlo Jimenez Giraldo, Hoyeon Lee, Jun Lian, Shuolin Liu, Keng-Chi Liu, José Marrugo, Kentaro Miki, Kunio Nakamura, Tucker Netherton, Dan Nguyen, Hamidreza Nourzadeh, Alexander F. I Osman, Zhao Peng, José Darío Quinto Muñoz, Christian Ramsl, Dong Joo Rhee, Juan David Rodriguez, Hongming Shan, Jeffrey V. Siebers, Mumtaz H. Soomro, Kay Sun, Andrés Usuga Hoyos, Carlos Valderrama, Rob Verbeek, Enpei Wang, Siri Willems, Qi Wu, Xuanang Xu, Sen Yang, Lulin Yuan, Simeng Zhu, Lukas Zimmermann, Kevin L. Moore, Thomas G. Purdie, Andrea L. McNiven, and Timothy C. Y Chan

This article is available at Henry Ford Health Scholarly Commons: https://scholarlycommons.henryford.com/radiationoncology_articles/382

Physics in Medicine & Biology

IPEM
Institute of Physics and Engineering in Medicine

PAPER • OPEN ACCESS

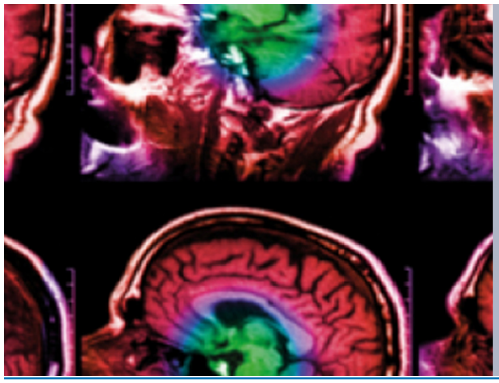
OpenKBP-Opt: an international and reproducible evaluation of 76 knowledge-based planning pipelines

To cite this article: Aaron Babier *et al* 2022 *Phys. Med. Biol.* **67** 185012

View the article online for updates and enhancements.

You may also like

- Automatic IMRT planning via static field fluence prediction (AIP-SFFP): a deep learning algorithm for real-time prostate treatment planning
Xinyi Li, Jiahan Zhang, Yang Sheng et al.
- Use of knowledge based DVH predictions to enhance automated re-planning strategies in head and neck adaptive radiotherapy
Elisabetta Cagni, Andrea Botti, Agnese Chendi et al.
- Predicting personalised optimal arc parameter using knowledge-based planning model for inoperable locally advanced lung cancer patients to reduce organ at risk doses
Nilesh S Tambe, Isabel M Pires, Craig Moore et al.



IPEM | IOP
Series in Physics and Engineering in Medicine and Biology

Your publishing choice in medical physics, biomedical engineering and related subjects.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 165.225.62.22 on 13/10/2022 at 20:04

IOP Publishing Phys. Med. Biol. 67 (2022) 185012 https://doi.org/10.1088/1361-6560/ac8044

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

CrossMark

OPEN ACCESS

RECEIVED
26 March 2022

REVISED
21 June 2022

ACCEPTED FOR PUBLICATION
11 July 2022

PUBLISHED
12 September 2022

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

OpenKBP-Opt: an international and reproducible evaluation of 76 knowledge-based planning pipelines

Aaron Babier[1,2,*], Rafid Mahmood[1], Binghao Zhang[1], Victor G L Alves[3], Ana Maria Barragán-Montero[4], Joel Beaudry[5], Carlos E Cardenas[6], Yankui Chang[7], Zijie Chen[8], Jaehee Chun[9], Kelly Diaz[10], Harold David Eraso[10], Erik Faustmann[11], Sibaji Gaj[12], Skylar Gay[13], Mary Gronberg[13], Bingqi Guo[14], Junjun He[15], Gerd Heilemann[16], Sanchit Hira[17], Yuliang Huang[18], Fuxin Ji[19], Dashan Jiang[19], Jean Carlo Jimenez Giraldo[10], Hoyeon Lee[20], Jun Lian[21], Shuolin Liu[19], Keng-Chi Liu[22], José Marrugo[10], Kentaro Miki[23], Kunio Nakamura[12], Tucker Netherton[13], Dan Nguyen[24], Hamidreza Nourzadeh[25], Alexander F I Osman[26], Zhao Peng[7], José Darío Quinto Muñoz[10], Christian Ramsl[11], Dong Joo Rhee[13], Juan David Rodriguez[10], Hongming Shan[27], Jeffrey V Siebers[3], Mumtaz H Soomro[3], Kay Sun[28], Andrés Usuga Hoyos[10], Carlos Valderrama[10], Rob Verbeek[29], Enpei Wang[8], Siri Willems[30], Qi Wu[19], Xuanang Xu[31], Sen Yang[32], Lulin Yuan[33], Simeng Zhu[34], Lukas Zimmermann[35,36], Kevin L Moore[37], Thomas G Purdie[38,39,40,41], Andrea L McNiven[38,39] and Timothy C Y Chan[1,2,40]

[1] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada
[2] Vector Institute, Toronto, ON, Canada
[3] Department of Radiation Oncology, University of Virginia Health System, Charlottesville, VA, United States of America
[4] Department of Molecular Imaging Radiation Oncology, UCLouvain, Louvain-la-Neuve, Belgium
[5] Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, United States of America
[6] Department of Radiation Oncology, The University of Alabama at Birmingham, Birmingham, AL, United States of America
[7] Department of Engineering and Applied Physics, University of Science and Technology of China, Hefei, People's Republic of China
[8] Shenying Medical Technology Co., Ltd., Shenzhen, Guangdong, People's Republic of China
[9] Department of Radiation Oncology, Yonsei University College of Medicine, Seoul, Republic of Korea
[10] Department of Physics, National University of Colombia, Medellín, Colombia
[11] Atominstitut, Vienna University of Technology, Vienna, Austria
[12] Department of Biomedical Engineering, Cleveland Clinic, Cleveland, OH, United States of America
[13] Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States of America
[14] Department of Radiation Oncology, Cleveland Clinic, Cleveland, OH, United States of America
[15] Department of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, People's Republic of China
[16] Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria
[17] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, United States of America
[18] Department of Radiation Oncology, Peking University Cancer Hospital and Institute, Beijing, People's Republic of China
[19] Department of Electrical Engineering and Automation, Anhui University, Hefei, People's Republic of China
[20] Department of Radiation Oncology, Massachusetts General Hospital, Boston, MA, United States of America
[21] Department of Radiation Oncology, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States of America
[22] Department of Medical Imaging, Taiwan AI Labs, Taipei, Taiwan
[23] Department Of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan
[24] Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, The University of Texas Southwestern Medical Center, Dallas, TX, United States of America
[25] Department of Radiation Oncology, Thomas Jefferson University, Philadelphia, PA, United States of America
[26] Department of Medical Physics, Al-Neelain University, Khartoum, Sudan
[27] Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai, People's Republic of China
[28] Studio Vodels, Atlanta, GA, United States of America
[29] Department Computer Science, Aalto University, Espoo, Finland
[30] Department of Electrical Engineering, KULeuven, Leuven, Belgium
[31] Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY, United States of America
[32] Tencent AI Lab, Shenzhen, Guangdong, People's Republic of China
[33] Department of Radiation Oncology, Virginia Commonwealth University Medical Center, Richmond, VA, United States of America
[34] Department of Radiation Oncology, Henry Ford Health System, Detroit, MI, United States of America
[35] Faculty of Health, University of Applied Sciences Wiener Neustadt, Wiener Neustadt, Austria
[36] Competence Center for Preclinical Imaging and Biomedical Engineering, University of Applied Sciences Wiener Neustadt, Wiener Neustadt, Austria
[37] Department of Radiation Oncology, University of California, San Diego, La Jolla, CA, United States of America
[38] Radiation Medicine Program, UHN Princess Margaret Cancer Centre, Toronto, ON, Canada
[39] Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada
[40] Techna Institute for the Advancement of Technology for Health, Toronto, ON, Canada
[41] Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
* Author to whom any correspondence should be addressed.

E-mail: ababier@mie.utoronto.ca

© 2022 The Author(s). Published on behalf of Institute of Physics and Engineering in Medicine by IOP Publishing Ltd

Keywords: knowledge-based planning, radiotherapy, optimization, inverse problem, inverse optimization, automated planning, open data

Abstract

Objective. To establish an open framework for developing plan optimization models for knowledge-based planning (KBP). *Approach.* Our framework includes radiotherapy treatment data (i.e. reference plans) for 100 patients with head-and-neck cancer who were treated with intensity-modulated radiotherapy. That data also includes high-quality dose predictions from 19 KBP models that were developed by different research groups using out-of-sample data during the OpenKBP Grand Challenge. The dose predictions were input to four fluence-based dose mimicking models to form 76 unique KBP pipelines that generated 7600 plans (76 pipelines $\times$ 100 patients). The predictions and KBP-generated plans were compared to the reference plans via: the dose score, which is the average mean absolute voxel-by-voxel difference in dose; the deviation in dose-volume histogram (DVH) points; and the frequency of clinical planning criteria satisfaction. We also performed a theoretical investigation to justify our dose mimicking models. *Main results.* The range in rank order correlation of the dose score between predictions and their KBP pipelines was 0.50–0.62, which indicates that the quality of the predictions was generally positively correlated with the quality of the plans. Additionally, compared to the input predictions, the KBP-generated plans performed significantly better ($P < 0.05$; one-sided Wilcoxon test) on 18 of 23 DVH points. Similarly, each optimization model generated plans that satisfied a higher percentage of criteria than the reference plans, which satisfied 3.5% more criteria than the set of all dose predictions. Lastly, our theoretical investigation demonstrated that the dose mimicking models generated plans that are also optimal for an inverse planning model. *Significance.* This was the largest international effort to date for evaluating the combination of KBP prediction and optimization models. We found that the best performing models significantly outperformed the reference dose and dose predictions. In the interest of reproducibility, our data and code is freely available.

1. Introduction

Automated radiotherapy planning is transforming clinical practice and personalized cancer treatment (Moore 2019). The most common type of automated planning is knowledge-based planning (KBP), which leverages knowledge derived from historical clinical treatment plans to generate new treatment plans without human intervention (Cornell *et al* 2020, Kaderka *et al* 2021, McIntosh *et al* 2021). Most common KBP methods are formulated as a two-stage pipeline (see figure 1) that first predicts the dose that should be delivered to a patient (Kearney *et al* 2018, Nguyen *et al* 2019) and then converts that prediction into a treatment plan via optimization (Babier *et al* 2021a, Eriksson and Zhang 2022). Both stages of this pipeline, which are active areas of research, can significantly affect the quality of generated treatment plans (Babier *et al* 2020). 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.

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 (Hussein *et al* 2018, Ge and Wu 2019, Wang *et al* 2020, Momin *et al* 2021). 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 (Babier *et al* 2021b). The OpenKBP dataset, which includes data for 340 patients with head-and-neck cancer who were treated with 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 (Craft *et al* 2014, Breedveld and Heijmen 2017). 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 small datasets (123 patients total) that span multiple sites (prostate, liver, and head-and-neck) and multiple modalities (CyberKnife, volumetric modulated arc therapy, proton therapy, and 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 (Boutilier *et al* 2016).

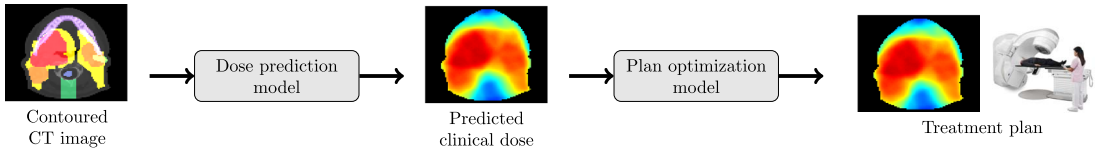


Figure 1. An overview of a complete knowledge-based planning pipeline.

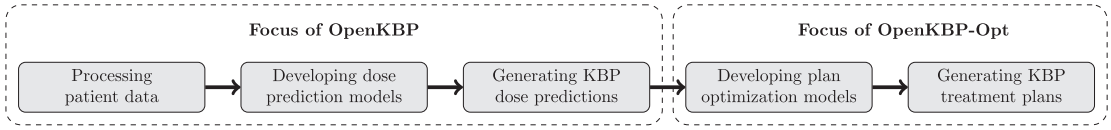


Figure 2. An overview of our methods. A full description of each component is provided in its corresponding subsection.

There are several types of KBP optimization models that translate dose predictions into treatment plans. One major type of KBP optimization model is dose mimicking, which generally generates a plan that is similar to an input prediction based on linear (Kierkels *et al* 2019) or quadratic (McIntosh *et al* 2017) differences. Another type of KBP optimization model is inverse planning weight estimation, which optimizes patient-specific parameters that make an input dose prediction optimal in a conventional planning model (Chan *et al* 2014). However, both types of models can also use information beyond a single dose prediction. For example, dose mimicking models can incorporate parameters that reflect the uncertainties in a predicted dose distribution (Zhang *et al* 2021). Similarly, inverse planning weight estimation models can incorporate an ensemble of dose predictions to leverage the combined wisdom of multiple predictions (Babier *et al* 2021a). Note that these optimziation models make dose predictions an intermediate step in a KBP pipeline.

Most KBP pipelines are developed as *fully-automated* pipelines that can replace human treatment planners in the planning process (McIntosh *et al* 2017, Fan *et al* 2019, Bai *et al* 2020, Wortel *et al* 2021). 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 (Cornell *et al* 2020, McIntosh *et al* 2021). 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 both data and human expertise.. In the KBP literature, however, there are relatively few papers that describe tools that humans can intuitively interact with in semi-automated KBP pipeline (Babier *et al* 2018, Bohara *et al* 2020, Kaderka *et al* 2021, Zhang *et al* 2022).

In this paper, we extend the results from the OpenKBP Grand Challenge with an international validation of 76 KBP pipelines. We made this extension, which we call OpenKBP-Opt, open to provide a benchmark for future 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 a conventional inverse planning process with good patient-specific parameters (i.e. objective weights). This relationship provides a mechanism for transforming some existing KBP optimization models, which are 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.

2. Materials and methods

Figure 2 separates our methods into five components. The first three components (processing patient data, developing dose prediction models, and generating KBP dose predictions) are based on the results from the OpenKBP Grand Challenge. The final two components (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.

2.1. Processing patient 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 the Computational Environment for Radiotherapy Research (Deasy *et al* 2003) using MATLAB, and it is given by $d_v = \sum_{b \in \mathcal{B}} D_{v,b} w_b$.

2.2. Developing dose prediction models

All dose prediction models used in this paper were developed in the OpenKBP Grand Challenge (Babier *et al* 2021b). During the challenge, teams developed dose prediction models using identical training and validation datasets with access only to ground truth data (i.e. reference dose) for the training set. Every dose prediction model used a neural network architecture that was based on either a U-Net (Ronneberger *et al* 2015), V-Net (Milletari *et al* 2016), or Pix2Pix (Isola *et al* 2017) architecture. Many of the best performing models also used other generalizable techniques like ensembles (Nguyen *et al* 2021), one-cycle learning (Zimmermann *et al* 2021), radiotherapy-specific loss functions (Gronberg *et al* 2021), and deep supervision (Liu *et al* 2021).

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. their KBP dose predictions). 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 D_{mean} and $D_{0.1\text{cc}}$, which was the mean dose delivered to the OAR and the maximum dose delivered to 0.1 cc of the OAR, respectively. The target DVH points were the D_1, D_{95}, and 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.

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 dose predictions from the participating teams for each patient in the test set to create a dataset with 2100 dose predictions (21 different predictions for each of the 100 patients). We observed that two models had 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 ($n = 1900$ dose predictions).

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 are described in sections 2.4.1 and 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 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.

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 $\mathcal{R}_p$ contained OARs, targets, and optimization structures. The OARs were the brainstem, spinal cord, right parotid, left parotid, larynx, esophagus, and mandible. Each target t was a planning target volume (PTV) with a dose level θ_t, and those targets were the PTV56, PTV63, and PTV70. The optimization structures 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 (Babier *et al* 2021b). Every ROI $r \in \mathcal{R}_p$ was also divided into a set of voxels $\mathcal{V}_r$.

2.4.2. Objective functions

Our models used the objective functions in table 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 dose objective function quantified the average dose and maximum dose delivered to an

Table 1. The formulations for our objective functions.

	Objective function
Average dose	$\underset{v\in\mathcal{V}_r}{\text{mean}}(d_v)$
Maximum dose	$\underset{v\in\mathcal{V}_r}{\max}(d_v)$
Average dose over threshold	$\underset{v\in\mathcal{V}_r}{\text{mean}}(d_v - f)^+$
Average dose under threshold	$\underset{v\in\mathcal{V}_r}{\text{mean}}(f - d_v)^+$

Table 2. The cost functions for each dose mimicking model that minimize mean absolute (MeanAbs), max absolute (MaxAbs), mean relative (MeanRel), and max relative (MaxRel) differences between all pairs of the optimized and predicted objective values $(g_m(\mathbf{w}), \hat{g}_m)$.

	Dose mimicking model cost function
MeanAbs	$\underset{m\in\mathcal{M}_p}{\text{mean}}(g_m(\mathbf{w}) - \hat{g}_m)^+ + \epsilon \underset{m\in\mathcal{M}_p}{\text{mean}}(g_m(\mathbf{w}) - \hat{g}_m)^-$
MaxAbs	$\underset{m\in\mathcal{M}_p}{\max}(g_m(\mathbf{w}) - \hat{g}_m)$
MeanRel	$\underset{m\in\mathcal{M}_p}{\text{mean}}\left(\frac{g_m(\mathbf{w}) - \hat{g}_m}{\hat{g}_m}\right)^+ + \epsilon \underset{m\in\mathcal{M}_p}{\text{mean}}\left(\frac{g_m(\mathbf{w}) - \hat{g}_m}{\hat{g}_m}\right)^-$
MaxRel	$\underset{m\in\mathcal{M}_p}{\max}\left(\frac{g_m(\mathbf{w}) - \hat{g}_m}{\hat{g}_m}\right)$

ROI r, respectively. The average dose over and under threshold objective functions quantified the average dose delivered to an ROI r that was over and under a dose threshold f, respectively. Our average dose over and under threshold objective functions are similar to *tail mean dose* (Romeijn *et al* 2006) and *conditional value-at-risk* (Rockafellar and Uryasev 2000), which are both defined on the percentiles of a distribution.

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 average dose over thresholds of 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, the average dose under a threshold f equal to the dose level of the target (i.e. $f = \theta_t$), and the average dose over a threshold f equal to five percent more than 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 the models associated with those patients had fewer than 107 objective functions.

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 $m \in \mathcal{M}_p$ index one of the 107 objective functions (as outlined in section 2.4.2), and let the elements in the vector $\mathbf{w}$ represent beamlet intensities w_b, $\forall\, b \in \mathcal{B}$. Let $g_m(\mathbf{w})$ 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 $g_m(\mathbf{w})$ were favored over higher values. Table 2 presents the cost functions of our dose mimicking models. Each model minimized either the mean or max difference between all corresponding pairs $(g_m(\mathbf{w}), \hat{g}_m)$ of the objective values, which were quantified via an absolute $(g_m(\mathbf{w}) - \hat{g}_m)$ or relative $((g_m(\mathbf{w}) - \hat{g}_m)/\hat{g}_m)$ 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 $g_m(\mathbf{w}) \leqslant \hat{g}_m$, $\forall\, m \in \mathcal{M}_p$).

The main constraint in all four models was a constraint to limit plan complexity. In particular, the sum-of-positive gradients (SPG) (Craft *et al* 2007) 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 (Babier *et al* 2021b). The remaining constraints were simply auxiliary constraints (including auxiliary variables) used to linearize both the objective and cost functions

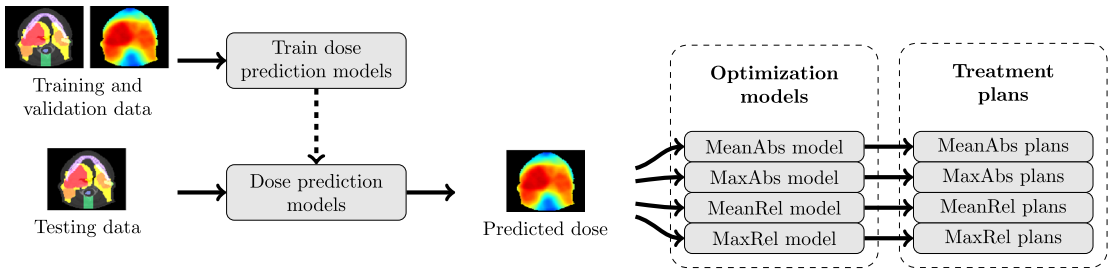


Figure 3. An overview of our process. First, dose prediction models were developed with training and validation data. Second, those models predicted dose for testing data that was used by the dose mimicking models to generate KBP plans.

(i.e. the formulations in table 1 and table 2). The optimization models were all formulated in Python 3.7 using OR-Tools 9.1 and solved using Gurobi 9.1 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.

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 3). We used these plans in our analysis to measure the quality of the respective KBP models. We refer to the plans generated by each dose mimicking model as MeanAbs, MaxAbs, MeanRel, and MaxRel plans.

Altogether, after completing the process in figure 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 distributions that were submitted by 19 teams in the final testing phase of the challenge. 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 multiple factors including noisy and undeliverable predictions. Differences between a KBP plan and its corresponding reference plan reflect different trade-offs in the cost function used to generate these plans.

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.

2.6.1. Dose score and error

We evaluated the KBP models using the dose score and dose error as defined in section 2.2. We calculated the Spearman rank order correlation of the dose score rank between the prediction models and corresponding KBP pipelines. The distribution of dose error was also visualized using a box plot. A one-sided Wilcoxon signed-rank test was used to evaluate whether the dose error of the optimization models was the same (null hypothesis) or lower (alternative hypothesis) than the dose prediction models. For all hypothesis tests in this paper, $P < 0.05$ was considered significant.

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 2.2 and visualized using boxplots. We used the one-sided Wilcoxon signed-rank test to evaluate 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 D_{mean}, $D_{0.1cc}$, and D_1; higher values were better for D_{95} and D_{99}.

2.6.3. Expected clinical 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 3). The target

Table 3. The clinical criteria that we used to evaluate dose distributions.

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

criteria were evaluated after overlap between targets, which was removed when processing patient data for the OpenKBP dataset, was reinstated. We tabulated the proportion of clinical 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.

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 (Bertsimas and Tsitsiklis 1997, Chapter 4). This analysis was based on previous literature that showed a connection between Benson's method (Benson 1978), which identifies efficient solutions to multi-objective optimization models, and estimating the weights for inverse planning (Chan *et al* 2014). We were motivated to conduct a similar analysis as in Chan *et al* (2014) because our dose mimicking models are similar to the formulations in Benson (1978). In particular, we linearized the dose mimicking models, took their duals, and related the dual variables to objective weights $\hat{\alpha}_m$ in a conventional multi-objective inverse planning problem depicted in model (1):

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \sum_{m \in \mathcal{M}_p} \hat{\alpha}_m g_m(\boldsymbol{w}), \\ \text{subject to} \quad & \text{SPG} \leqslant 65, \\ & \text{Auxiliary constraints to linearize functions in Table 1 and 2.} \end{aligned} \tag{1}$$

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. We also recorded the average, first quartile, and third quartile solve times.

3. Results

In this section, we summarize the performance of the 19 dose predictions models, four dose mimicking models, and 76 KBP pipelines. We also complete our theoretical analysis of dose mimicking models and summarize the metadata generated by our experiments.

3.1. Dose score and error

Table 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 was 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 dose prediction model that outperforms a competitor will not always generate better plans when it is used as input to a dose mimicking model. 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).

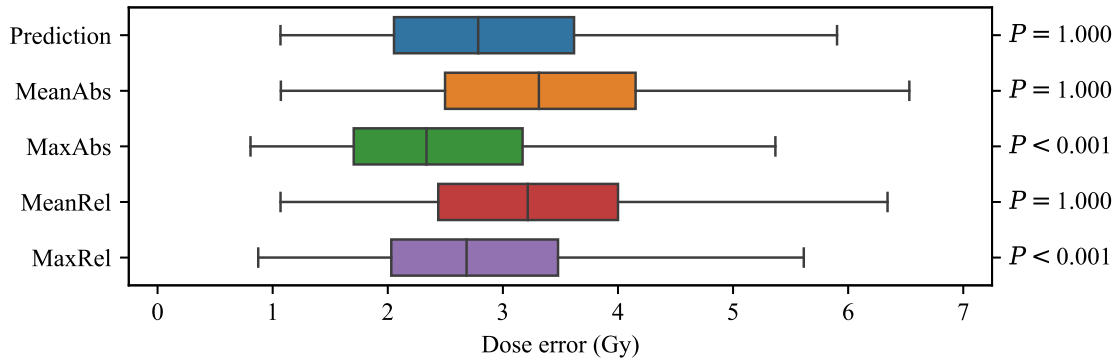


Figure 4. The distribution of dose error over all KBP-generated dose ($n = 1900$ points in each box). Boxes indicate median and interquartile range (IQR). Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier.

Table 4. Each dose mimicking model is compared to the predictions in terms of Spearman 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 4. Two of the four sets of KBP plans (those generated by MaxAbs and MaxRel) 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 ($P < 0.001$) than predictions. The MaxAbs model achieved the lowest median dose error (2.34 Gy).

3.2. DVH point differences

Figure 5 shows the DVH point differences between the reference dose and KBP-generated 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.

3.3. Expected clinical criteria satisfaction

In table 5, we compare the percentage of criteria that were satisfied by the reference plans ($n = 100$), predictions ($n = 1900$), plans generated by each of the four dose mimicking models ($n = 4 \times 1900$), and 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 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 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 5. The right parotid, left parotid, and larynx were satisfied less than 40% of the time by the the two baselines. In contrast, each of our four dose mimicking 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.

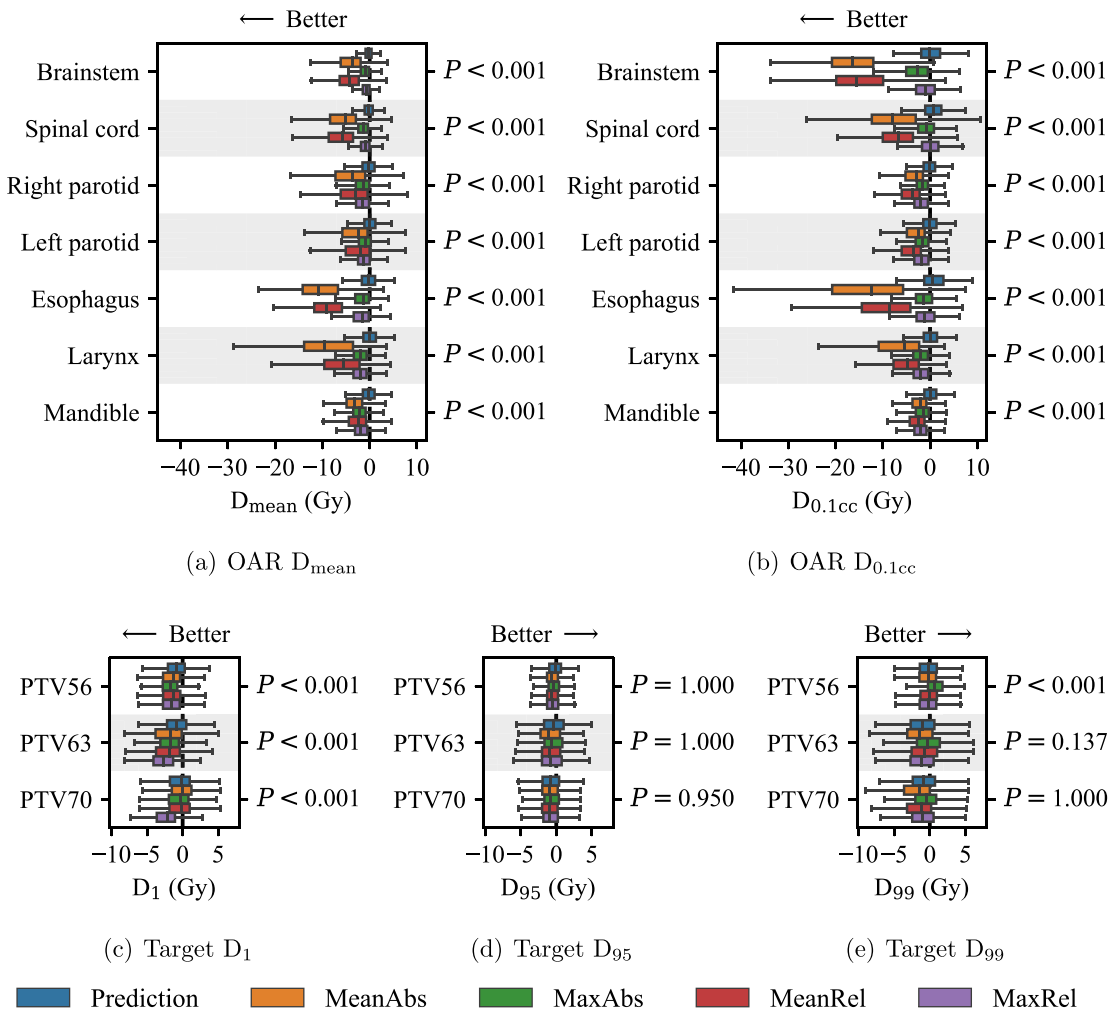


Figure 5. The distribution of DVH point differences between the reference dose and each set of KBP-generated dose. Negative differences indicate cases where the KBP-generated dose had a lower DVH point than the reference dose, and arrows indicate the direction where KBP-generated dose is considered better than reference dose for each DVH point. Boxes indicate median and IQR. Whiskers extend to the minimum of 1.5 times the IQR and the most extreme outlier.

Table 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 summarizes the clinical criteria that were satisfied by each of the 76 KBP pipelines that we evaluated. The spread in OAR criteria satisfaction across all 19 models (55.4%–82.1%) was lower than that of target criteria satisfaction (24.5%–89.7%), see figures 6(a) and (b), respectively. Overall, the MeanAbs model generated plans that satisfied more criteria than the other three dose mimicking models for 16 of the 19 dose prediction models (see figure 6(c)). Additionally, the pipelines that used better prediction models (i.e. lower dose score ranks) generally produced plans with higher criteria satisfaction. Interestingly, however, the best performing KBP pipeline (from the last column of table 5) used the dose prediction model that ranked 16th in terms of dose score. Note that the poor performing KBP pipelines used the 12th, 13th, 17th, 18th, and 19th ranked dose prediction models. Since the dose mimicking columns in table 5 included all KBP pipelines, these poor performing models contributed to low performance that was most pronounced 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.

3.4. Theoretical analysis of dose mimicking models

We use theoretical results from Chan *et al* (2014) to demonstrate the connection between our dose mimicking formulations and inverse planning. The inverse planning problem presented previously as model (1), is presented again in vector and matrix notation to follow Chan *et al* (2014). The objective functions are represented as the rows of the matrix $\mathbf{C}$ and the objective 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 $\mathbf{b}$. With this vector and matrix notation, we can write the inverse planning problem as model (2):

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \hat{\boldsymbol{\alpha}}'\mathbf{C}\mathbf{x}, \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geqslant \mathbf{0}. \end{aligned} \tag{2}$$

Table 6 presents the formulations of the four dose mimicking models and their respective dual models in vector and matrix notation. 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 by $\boldsymbol{\alpha}$ and $\mathbf{p}$. The vectors of all 0 and 1 are denoted by $\mathbf{0}$ and $\mathbf{e}$, respectively. The symbol $\odot$ denotes element-wise multiplication of two vectors and prime denotes the transpose operator.

Next, we complete our theoretical analysis. We first observe that the weight estimation technique developed in Chan *et al* (2014) is identical to our dual formulations (see table 6) except for the constraints related to the objective weights $\boldsymbol{\alpha}$, which prevent trivial solutions to the weight estimation technique. In the context of our models, proposition 5 from Chan *et al* (2014) establishes that an optimal decision vector $\mathbf{x}^*$ from each dose mimicking model is also optimal for the inverse planning model with objective weights equal to the optimal dual vector $\boldsymbol{\alpha}^*$, which is a byproduct of solving the corresponding dose mimicking model. This result means that the solution to each dose mimicking model is also optimal to an inverse planning model with a particular set of objective weights (i.e. $\mathbf{x}^*$ is an optimal solution for model (2) when $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^*$). Additionally, by complementary slackness, a plan generated by the MeanAbs or MeanRel model will achieve the same objective values (i.e. $\mathbf{C}\mathbf{x}^*$) as a plan that is optimal for its corresponding inverse planning model. These theoretical results were validated computationally but omitted for brevity.

3.5. Optimization metadata

In table 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. 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 was also a wide range in average solve time between the models (222–393 s). On average, the MaxAbs model was the fastest.

4. Discussion

KBP 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

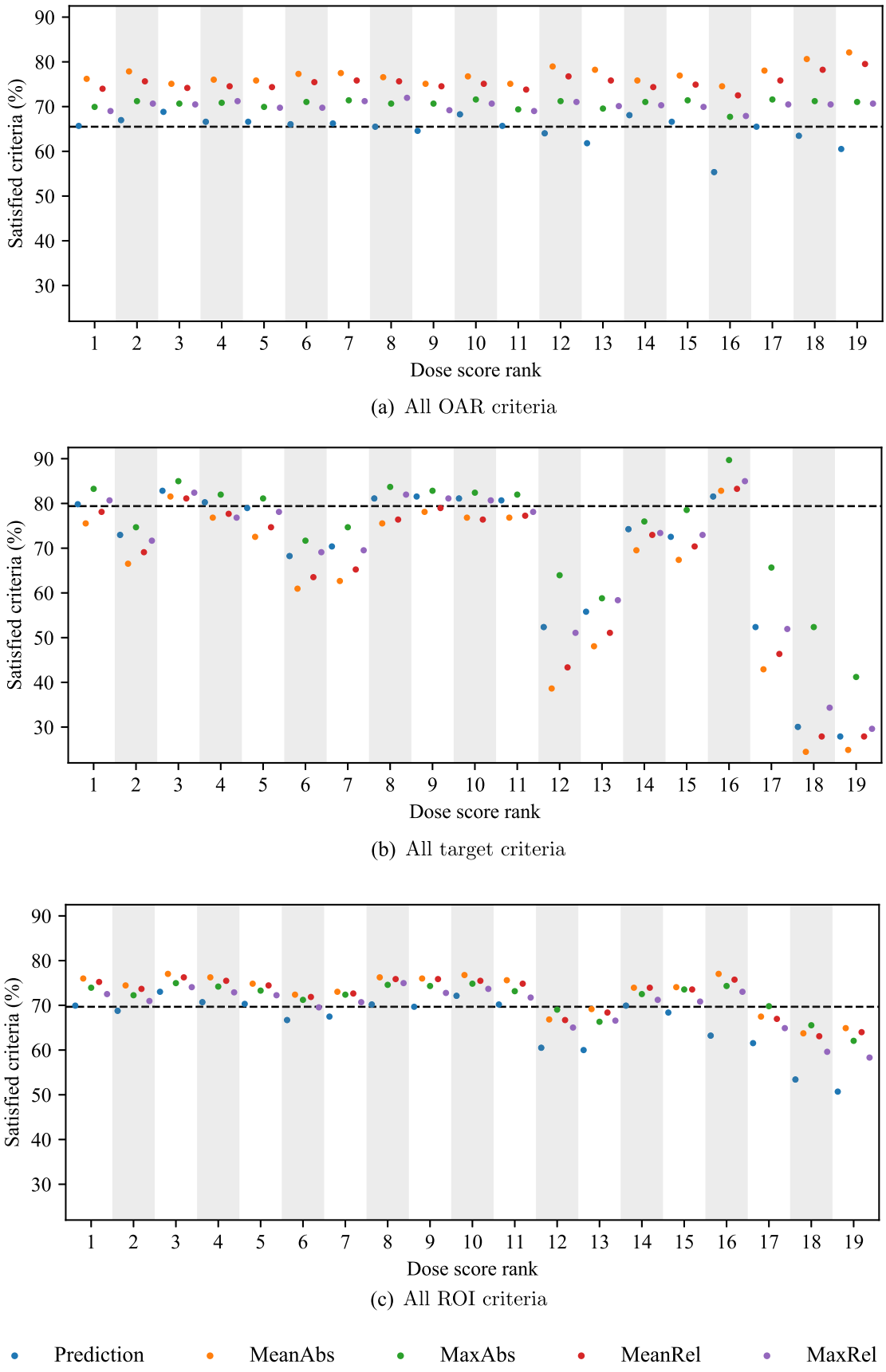


Figure 6. The percentage of all (a) OAR, (b) target, and (c) ROI clinical criteria that were satisfied by each KBP pipeline, which are labeled by their prediction dose score rank. The points indicate the percentage of satisfied criteria for $n = 100$ patients. A dashed line indicates the percentage of criteria satisfied by reference plans.

Table 6. The dose mimicking models presented in vector and matrix notation with their dual models. Terms that follow colons indicate the dual variables for that constraint.

	Dose mimicking model	Dual model
MeanAbs	$\underset{\mathbf{x},\boldsymbol{\sigma},\boldsymbol{\delta}}{\text{minimize}}\ \mathbf{e}'\boldsymbol{\sigma} + \epsilon\mathbf{e}'\boldsymbol{\delta}$ subject to $\mathbf{C}\mathbf{x} = \mathbf{C}\hat{\mathbf{x}} + \boldsymbol{\sigma} + \boldsymbol{\delta}$: $\boldsymbol{\alpha}$ $\mathbf{A}\mathbf{x} = \mathbf{b}$: $\mathbf{p}$ $\mathbf{x} \geqslant \mathbf{0}$ $\boldsymbol{\sigma} \geqslant \mathbf{0}$ $\boldsymbol{\delta} \leqslant \mathbf{0}$	$\underset{\boldsymbol{\alpha},\mathbf{p}}{\text{minimize}}\ \boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} - \mathbf{b}'\mathbf{p}$ subject to $\mathbf{C}'\boldsymbol{\alpha} \geqslant \mathbf{A}'\mathbf{p}$: $\mathbf{x}$ $\boldsymbol{\alpha} \leqslant \mathbf{e}$: $\boldsymbol{\sigma}$ $\boldsymbol{\alpha} \geqslant \epsilon\mathbf{e}$: $\boldsymbol{\delta}$
MaxAbs	$\underset{\mathbf{x},\zeta}{\text{minimize}}\ \zeta$ subject to $\mathbf{C}\mathbf{x} \leqslant \mathbf{C}\hat{\mathbf{x}} + \zeta\mathbf{e}$: $\boldsymbol{\alpha}$ $\mathbf{A}\mathbf{x} = \mathbf{b}$: $\mathbf{p}$ $\mathbf{x} \geqslant \mathbf{0}$	$\underset{\boldsymbol{\alpha},\mathbf{p}}{\text{minimize}}\ \boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} - \mathbf{b}'\mathbf{p}$ subject to $\mathbf{C}'\boldsymbol{\alpha} \geqslant \mathbf{A}'\mathbf{p}$: $\mathbf{x}$ $\boldsymbol{\alpha}'\mathbf{e} = 1$: ζ $\boldsymbol{\alpha} \geqslant \mathbf{0}$
MeanRel	$\underset{\mathbf{x},\boldsymbol{\sigma},\boldsymbol{\delta}}{\text{minimize}}\ \mathbf{e}'\boldsymbol{\sigma} + \epsilon\mathbf{e}'\boldsymbol{\delta}$ subject to $\mathbf{C}\mathbf{x} = \mathbf{C}\hat{\mathbf{x}} \odot (\mathbf{e} + \boldsymbol{\sigma} + \boldsymbol{\delta})$: $\boldsymbol{\alpha}$ $\mathbf{A}\mathbf{x} = \mathbf{b}$: $\mathbf{p}$ $\mathbf{x} \geqslant \mathbf{0}$ $\boldsymbol{\sigma} \geqslant \mathbf{0}$ $\boldsymbol{\delta} \leqslant \mathbf{0}$	$\underset{\boldsymbol{\alpha},\mathbf{p}}{\text{minimize}}\ \boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} - \mathbf{b}'\mathbf{p}$ subject to $\mathbf{C}'\boldsymbol{\alpha} \geqslant \mathbf{A}'\mathbf{p}$: $\mathbf{x}$ $\boldsymbol{\alpha} \odot \mathbf{C}\hat{\mathbf{x}} \leqslant \mathbf{e}$: $\boldsymbol{\sigma}$ $\boldsymbol{\alpha} \odot \mathbf{C}\hat{\mathbf{x}} \geqslant \epsilon\mathbf{e}$: $\boldsymbol{\delta}$
MaxRel	$\underset{\mathbf{x},\zeta}{\text{minimize}}\ \zeta$ subject to $\mathbf{C}\mathbf{x} \leqslant \mathbf{C}\hat{\mathbf{x}} \odot (\mathbf{e} + \zeta\mathbf{e})$: $\boldsymbol{\alpha}$ $\mathbf{A}\mathbf{x} = \mathbf{b}$: $\mathbf{p}$ $\mathbf{x} \geqslant \mathbf{0}$	$\underset{\boldsymbol{\alpha},\mathbf{p}}{\text{minimize}}\ \boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} - \mathbf{b}'\mathbf{p}$ subject to $\mathbf{C}'\boldsymbol{\alpha} \geqslant \mathbf{A}'\mathbf{p}$: $\mathbf{x}$ $\boldsymbol{\alpha}'\mathbf{C}\hat{\mathbf{x}} = 1$: ζ $\boldsymbol{\alpha} \geqslant \mathbf{0}$

Table 7. A summary of the average proportion of objective weight that was assigned to each group of ROI objectives and the solve time statistics of each dose mimicking model ($n = 1900$ plans in each column).

	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 structures	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

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 analysis to show that our KBP optimization 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 (Heil *et al* 2021). Our data includes the first open dataset with reference plans and predictions. 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 dose predictions. 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 (the 20th and 21st ranked models) were removed from our analysis because their dose scores

were poor, which we elaborated on in section 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 weights. This allows us to interpret previous weight estimation techniques (Chan *et al* 2014) 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 reflect the priorities of the input dose prediction, and those objective weights can be used in an inverse planning model (i.e. model (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 can 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 (Ahuja 2019).

Evaluating the performance of optimization models using many different dose predictions helps to identify interaction effects between these two stages of a KBP pipeline (Babier *et al* 2020). For example, the 16th ranked dose prediction 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. Note that the 16th ranked prediction model achieved the fewest OAR criteria (55.4%) and the third highest target criteria (81.5%), which suggests that the MeanAbs model was adept at correcting prediction errors related to under and over predicting OAR and target criteria satisfaction, respectively. 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 Grand 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 (Babier *et al* 2021b). 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.

5. Conclusion

In this paper, we combined the dose predictions contributed by a large international team with several KBP optimization models, resulting in 76 KBP pipelines. This was the largest international effort to date on KBP pipeline evaluation. We found that the best performing pipeline significantly outperformed the baselines. In the interest of reproducibility, our data and code is freely available.

Acknowledgments

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

ORCID iDs

Aaron Babier https://orcid.org/0000-0002-5949-2500
Rafid Mahmood https://orcid.org/0000-0003-1933-066X
Binghao Zhang https://orcid.org/0000-0001-5790-7875
Victor G L Alves https://orcid.org/0000-0001-7982-2039
Ana Maria Barragán-Montero https://orcid.org/0000-0002-9485-3076
Joel Beaudry https://orcid.org/0000-0002-8973-5139
Carlos E Cardenas https://orcid.org/0000-0003-1414-3849
Jaehee Chun https://orcid.org/0000-0002-9695-6079

Sibaji Gaj https://orcid.org/0000-0002-6997-5717
Skylar Gay https://orcid.org/0000-0003-4659-0766
Mary Gronberg https://orcid.org/0000-0002-0121-9579
Bingqi Guo https://orcid.org/0000-0003-1017-5237
Junjun He https://orcid.org/0000-0002-1813-1784
Gerd Heilemann https://orcid.org/0000-0002-7461-3956
Hoyeon Lee https://orcid.org/0000-0002-1165-1509
Jun Lian https://orcid.org/0000-0002-2041-9074
Keng-Chi Liu https://orcid.org/0000-0003-0415-6398
José Marrugo https://orcid.org/0000-0002-9226-6505
Kunio Nakamura https://orcid.org/0000-0002-7833-8138
Tucker Netherton https://orcid.org/0000-0003-1583-7121
Dan Nguyen https://orcid.org/0000-0002-9590-0655
Hamidreza Nourzadeh https://orcid.org/0000-0002-1686-4466
Alexander F I Osman https://orcid.org/0000-0002-1286-475X
Dong Joo Rhee https://orcid.org/0000-0003-0486-1556
Hongming Shan https://orcid.org/0000-0002-0604-3197
Jeffrey V Siebers https://orcid.org/0000-0002-9949-2863
Mumtaz H Soomro https://orcid.org/0000-0001-5354-7518
Kay Sun https://orcid.org/0000-0001-5136-5913
Siri Willems https://orcid.org/0000-0002-4269-1976
Xuanang Xu https://orcid.org/0000-0002-6045-8457
Sen Yang https://orcid.org/0000-0002-0639-4122
Lulin Yuan https://orcid.org/0000-0002-5085-419X
Simeng Zhu https://orcid.org/0000-0003-0850-6690
Lukas Zimmermann https://orcid.org/0000-0003-0099-2276
Thomas G Purdie https://orcid.org/0000-0003-4176-8457
Andrea L McNiven https://orcid.org/0000-0002-8775-9575
Timothy C Y Chan https://orcid.org/0000-0002-4128-1692

References

Ahuja A S 2019 The impact of artificial intelligence in medicine on the future role of the physician *PeerJ* **7** e7702
Babier A, Boutilier J J, Sharpe M B, McNiven A L and Chan T C Y 2018 Inverse optimization of objective function weights for treatment planning using clinical dose-volume histograms *Phys. Med. Biol.* **63** 105004
Babier A, Chan T C Y, Lee T, Mahmood R and Terekhov D 2021a An ensemble learning framework for model fitting and evaluation in inverse linear optimization *INFORMS J. Optim.* **3** 119–38
Babier A, Mahmood R, McNiven A L, Diamant A and Chan T C Y 2020 The importance of evaluating the complete automated knowledge-based planning pipeline *Phys. Med.* **72** 73–9
Babier A, Zhang B, Mahmood R, Moore K L, Purdie T G, McNiven A L and Chan T C Y 2021b OpenKBP: the open-access knowledge-based planning grand challenge and dataset *Med. Phys.* **48** 5549–61
Bai P, Weng X, Quan K, Chen J, Dai Y, Xu Y, Lin F, Zhong J, Wu T and Chen C 2020 A knowledge-based intensity-modulated radiation therapy treatment planning technique for locally advanced nasopharyngeal carcinoma radiotherapy *Radiat. Oncol.* **15** 188
Benson H P 1978 Existence of efficient solutions for vector maximization problems *J. Optim. Theory Appl.* **26** 569–80
Bertsimas D and Tsitsiklis J N 1997 *Introduction to Linear Optimization* (Belmont, MA, USA: Athena Scientific)
Bohara G, Sadeghnejad Barkousaraie A, Jiang S and Nguyen D 2020 Using deep learning to predict beam-tunable pareto optimal dose distribution for intensity-modulated radiation therapy *Med. Phys.* **47** 3898–912
Boutilier J J, Craig T, Sharpe M B and Chan T C Y 2016 Sample size requirements for knowledge-based treatment planning *Med. Phys.* **43** 1212–21
Breedveld S and Heijmen B 2017 Data for TROTS—the radiotherapy optimisation test set *Data Brief* **12** 143–9
Chan T C Y, Craig T, Lee T and Sharpe M B 2014 Generalized inverse multiobjective optimization with application to cancer therapy *Oper. Res.* **62** 680–95
Cornell M, Kaderka R, Hild S J, Ray X J, Murphy J D, Atwood T F and Moore K L 2020 Noninferiority study of automated knowledge-based planning *versus* human-driven optimization across multiple disease sites *Int. J. Radiat. Oncol. Biol. Phys.* **106** 430–9
Craft D, Bangert M, Long T, Papp D and Unkelbach J 2014 Shared data for intensity modulated radiation therapy (IMRT) optimization research: the CORT dataset *Gigascience* **3** 37
Craft D, Suss P and Bortfeld T 2007 The tradeoff between treatment plan quality and required number of monitor units in intensity-modulated radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **67** 1596–605
Deasy J O, Blanco A I and Clark V H 2003 CERR: a computational environment for radiotherapy research *Med. Phys.* **30** 979–85
Eriksson O and Zhang T 2022 Robust automated radiation therapy treatment planning using scenario-specific dose prediction and robust dose mimicking *Med. Phys.* (https://doi.org/10.1002/mp.15622)
Fan J, Wang J, Chen Z, Hu C, Zhang Z and Hu W 2019 Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique *Med. Phys.* **46** 370–81
Ge Y and Wu Q J 2019 Knowledge-based planning for intensity-modulated radiation therapy: a review of data-driven approaches *Med. Phys.* **46** 2760–75

Gronberg M P, Gay S S, Netherton T J, Rhee D J, Court L E and Cardenas C E 2021 Technical note: dose prediction for head and neck radiotherapy using a three-dimensional dense dilated U-Net architecture *Med. Phys.* **48** 5567–73
Heil B J, Hoffman M M, Markowetz F, Lee S-I, Greene C S and Hicks S C 2021 Reproducibility standards for machine learning in the life sciences *Nat. Methods* **18** 1132–5
Hussein M, Heijmen B J M, Verellen D and Nisbet A 2018 Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations *Br. J. Radiol.* **91** 20180270
Isola P, Zhu J-Y, Zhou T and Efros A A 2017 Image-to-image translation with conditional adversarial networks *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Honolulu, HI, USA , July 21-26, 2017)* 5967–76
Kaderka R, Hild S J, Bry V N, Cornell M, Ray X J, Murphy J D, Atwood T F and Moore K L 2021 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** 705–15
Kearney V, Chan J W, Haaf S, Descovich M and Solberg T D 2018 Dosenet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks *Phys. Med. Biol.* **63** 235022
Kierkels R G J, Fredriksson A, Both S, Langendijk J A, Scandurra D and Korevaar E W 2019 Automated robust proton planning using dose-volume histogram-based mimicking of the photon reference dose and reducing organ at risk dose optimization *Int. J. Radiat. Oncol. Biol. Phys.* **103** 251–8
Liu S, Zhang J, Li T, Yan H and Liu J 2021 Technical note: a cascade 3D U-Net for dose prediction in radiotherapy *Med. Phys.* **48** 5574–82
McIntosh C, Welch M, McNiven A, Jaffray D A and Purdie T G 2017 Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method *Phys. Med. Biol.* **62** 5926–44
McIntosh C *et al* 2021 Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer *Nat. Med.* **27** 999–1005
Milletari F, Navab N and Ahmadi S-A 2016 V-Net: fully convolutional neural networks for volumetric medical image segmentation *2016 Fourth International Conference on 3D Vision (3DV) (Stanford, CA, USA, 25–28 October, 2016)* 565–71
Momin S, Fu Y, Lei Y, Roper J, Bradley J D, Curran W J, Liu T and Yang X 2021 Knowledge-based radiation treatment planning: a data-driven method survey *J. Appl. Clin. Med. Phys.* **22** 16–44
Moore K L 2019 Automated radiotherapy treatment planning *Semin. Radiat. Oncol.* **29** 209–18
Nguyen D, Jia X, Sher D, Lin M-H, Iqbal Z, Liu H and Jiang S 2019 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-Net deep learning architecture *Phys. Med. Biol.* **64** 065020
Nguyen D, Sadeghnejad Barkousaraie A, Bohara G, Balagopal A, McBeth R, Lin M-H and Jiang S 2021 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** 054002
Rockafellar R T and Uryasev S 2000 Optimization of conditional value-at-risk *J. Risk* **2** 21–42
Romeijn H E, Ahuja R K, Dempsey J F and Kumar A 2006 A new linear programming approach to radiation therapy treatment planning problems *Oper. Res.* **54** 201–16
Ronneberger O, Fischer P and Brox T 2015 U-Net: convolutional networks for biomedical image segmentation *International Conference on Medical image computing and computer-assisted intervention* 9351 *(Munich, Germany, October 5-9, 2015)* 234–41
Wang M, Zhang Q, Lam S, Cai J and Yang R 2020 A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning *Front. Oncol.* **10** 580919
Wortel G, Eekhout D, Lamers E, van der Bel R, Kiers K, Wiersma T, Janssen T and Damen E 2021 Characterization of automatic treatment planning approaches in radiotherapy *Phys. Imaging Radiat. Oncol.* **19** 60–5
Zhang T, Bokrantz R and Olsson J 2021 Probabilistic feature extraction, dose statistic prediction and dose mimicking for automated radiation therapy treatment planning *Med. Phys.* **48** 4730–42
Zhang T, Bokrantz R and Olsson J 2022 Probabilistic pareto plan generation for semiautomated multicriteria radiation therapy treatment planning *Phys. Med. Biol.* **67** 045001
Zimmermann Lukas, Faustmann Erik, Ramsl Christian, Georg Dietmar and Heilemann Gerd 2021 Technical note: dose prediction for radiation therapy using feature-based losses and one cycle learning *Med. Phys.* **48** 5562–6