Chapter 4

Automatic IMPT planning: from a predicted dose to a robust treatment plan

The content presented in this chapter comes from a manuscript submitted to Medial Physics unther the title: "" and which is at the moment under revision. An alternative to online dose restoration is a fully automated planning performed without any previous information about the initial treatment plan. The feasibility of fully automated robust IMPT plans combining post-processed deep learning dose prediction and dose mimicking optimization was investigated. A special focus is put on the dose mimicking optimization step, being the key to ensure the robustness and accuracy of the automatically generated plans. This is especially challenging in head and neck cancer due to the presence of many radio-sensitive organs at risk close to the target. From the comparison of four different variation from the presented workflow, information could be extracted to derive the appropriate combination of post-processing and dose mimicking settings providing the most accurate and robust IMPT plan.

4.1 Introduction

Radiation therapy (RT) treatment planning in clinics is currently a time-consuming process involving tedious manual operations. Automated treatment planning workflows are being widely implemented in clinical settings to reduce inter-planner variability, speed up the optimization process, and improve plan quality [217].

Various approaches have recently been developed that use prior knowledge (e.g., knowledge from libraries of previously treated patients) to automate treatment planning (knowledge-based planning, or KBP)[218]. In one of its forms, KBP can be described as a two-stage pipeline. In the first stage, a trained machine learning (ML) model uses binary volumes of the regions of interest (ROIs) to predict the dose that should be delivered to each voxel of the patient. In the second stage, the dose prediction from the first stage is used in an optimization process to generate a deliverable treatment plan. After the user has chosen a beam configuration, the optimization process begins. The optimization step is crucial because it determines the machine parameters required to deliver the desired dose distribution. In the case of intensity-modulated proton therapy (IMPT) delivered with pencil beam scanning (PBS), the optimization will compute the number of energy layers needed to cover the tumor in-depth, the spot positions for each layer, and the spot weights for each beam.

Dose mimicking (DM) is one of the optimization processes that can follow the dose prediction step in the KBP pipeline [148, 216]. The optimization used for DM is the same as in regular treatment planning, with algorithms like gradient descent for optimization and analytical or Monte Carlo methods for dose calculation. Babier et al. [219] compared different combinations of two prediction models (a generative adversarial network and a random forest) and two optimization models (inverse planning and dose mimicking) in head and neck IMRT. According to the findings of this research, the performance of an automated KBP pipeline is determined by how well the prediction and optimization models work together.

Significant research has been conducted to advance KBP, with a primary emphasis on improving ML models like ResNet or UNet [220, 188, 221]. Most of the research has focused on photon radiation therapy techniques such as intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT), with only a few studies addressing proton therapy in the pelvic [222, 223] and abdominal [215] regions.

Optimization models, on the other hand, have received far less attention in the literature than dose prediction models. They are, however, particularly important

in proton therapy to ensure the robustness of the generated treatment plans. Proton dose distributions are more sensitive to treatment uncertainties than photon doses [11, 12]. This means that the robustness of proton dose distributions is not solely dependent on the dose shape, rendering the margin strategy suboptimal for IMPT. Robust optimization strategies have been proposed as a more appropriate strategy for the generation of high-quality proton plans [40, 35, 36, 224, 225].

Automated planning workflows are not always flawless. An ML model may predict suboptimal dose distributions, such as an overdose of an organs at risk (OAR), insufficient target coverage of the clinical target volume (CTV), or an unrealistic trade-off between OAR sparing and CTV coverage. These effects could be compensated for by post-processing the predicted dose and/or tweaking the setup of the dose mimicking problem to improve the deliverability of the final dose. The commercial treatment planning system (TPS) RayStation 11B includes an ML module for automatic IMPT treatment planning including both post-processing and dose mimicking. Before the plan optimization, an in-built post-processing step with pre-configured clinic-specific settings for each protocol runs in their KBP pipeline. The post-processing is applied to the predicted dose aiming to ensure enough target coverage and improve OAR sparing. Such post-processing, together with the configuration of the dose-mimicking parameters, are crucial to the KBP pipeline. Therefore, the impact of the post-processing and mimicking parameters on the quality of the automatically generated IMPT plans in the KBP pipeline must be investigated. This is especially important in head and neck regions, where the presence of multiple nearby OAR that are preferably spared as much as possible makes the implementation of these automated planning strategies even more complex and challenging.

In this paper, we studied four complete KBP pipelines for IMPT plan generation in oropharyngeal cancer. The steps following dose prediction were the focus of this research. We investigated the effect of post-processing (PP) or no post-processing (NPP) on the final dose distribution when combined with two different dose mimicking algorithms: the one available in RayStation (RSM) and a simpler isodose-based mimicking (IBM). Overall, the four proposed KBP pipelines (PP-RSM, PP-IBM, NPP-RSM, and NPP-IBM) were built and benchmarked against the corresponding clinical plans using DVH-based metrics. The analysis of the robustness of the final treatment plans in the four settings is also given special attention.

4.2 Material and methods

4.2.1 Patient cohort and planning strategy

Our dataset included 60 patients with bilateral oropharyngeal cancer which were manually planned by the same dosimetrist using a treatment planning protocol to ensure consistency (Appendix **??**)(see supplementary material Table SM1). Local ethics committee approval was obtained for the use of all patient data. The prescribed dose (Dp) was 70 GyRBE to the high-dose CTV and 54.25 GyRBE to the low-dose CTV in 35 fractions with a constant relative biological effectiveness (RBE) factor of 1.1. The parotids, submandibular glands (SMGs), the pharyngeal constrictor muscles (PCMs), the oral cavity, the spinal cord and the brainstem were defined as the main OAR. Target and OAR delineations performed by experienced radiotherapy oncologists were available on all planning-CTs.

The same beam configuration was used for all patients: two anterior beams at (60º,10º), and (120º,10º) and two posterior beams at (240º, 350º) and (300º, 350º) for gantry and couch angles, respectively. For each patient, a robustly optimized treatment plan was created using Monte Carlo dose calculation and classical worst-case robust optimization (setup error (SE) = 4 mm, range error (RE) = 2.6%). Although 3% of range uncertainty is generally used in the clinic, when applying the 1.6% sigma to range uncertainty, as quantified by Paganetti [32], to Van Herk's margins formula at a 90% confidence level [30], the resulting range uncertainty can be estimated as 1.6%*1.64 = 2.6%.

SE scenarios included patient shifts along the three space axes x,y,z (±SE, 0, 0), (0, ±SE, 0), (0, 0, ±SE), and the nominal scenario (0,0,0). Range robustness was also included to account for uncertainties in the conversion of CT numbers to proton stopping power ratios. RE scenarios were modelled by scaling the mass density by ± 2.6%. The nominal scenario (i.e. 0%) was also included. The combination of seven SE scenarios and three RE scenarios resulted in a 21 error scenario optimization (7 SE scenarios x 3 RE scenarios). Robust optimization was followed by a robust evaluation using the same robustness parameters (SE = 4 mm, RE = 2.6%). The robust evaluation was performed by simulating 42 uncertainty scenarios (14 SE scenarios × 3 RE scenarios). According to Korevaar et al. [61], the 6 principal directions (±SE, 0, 0), (0, ±SE, 0), (0, ±SE, 0), (0, SE, 0) and the 8 diagonals from the center to the vertices (±SE/3, ±SE/3, ±SE/3) in combination with 0, ± RE% were the sampled error scenarios. Plans were manually adjusted until they met the list of clinical goals as closely as possible in both the nominal and worst-case scenarios. The resulting plan meeting the clinical goals included in the planning protocol was considered the ground truth of the study and will be referred to as the clinical plan from this point.

4.2.2 Dose prediction model

Spatial dose prediction is available in RayStation 11B using a UNet architecture that takes as input 3D binary matrices from the delineated volumes. The network is trained in a supervised manner to learn how to predict a dose value for each voxel. More information concerning the specifications of the prediction model can be found in Çiçek et al. [226]

In order to get the most out of the patient dataset containing 60 patients, 11 dose prediction models were trained using 47 patients for training, 5 patients for validation, and 5 patients for testing. Three patients (check set) were left out to determine whether the differences introduced by the different models were significant or negligible. The partition of the dataset is explained in Figure 4.1.

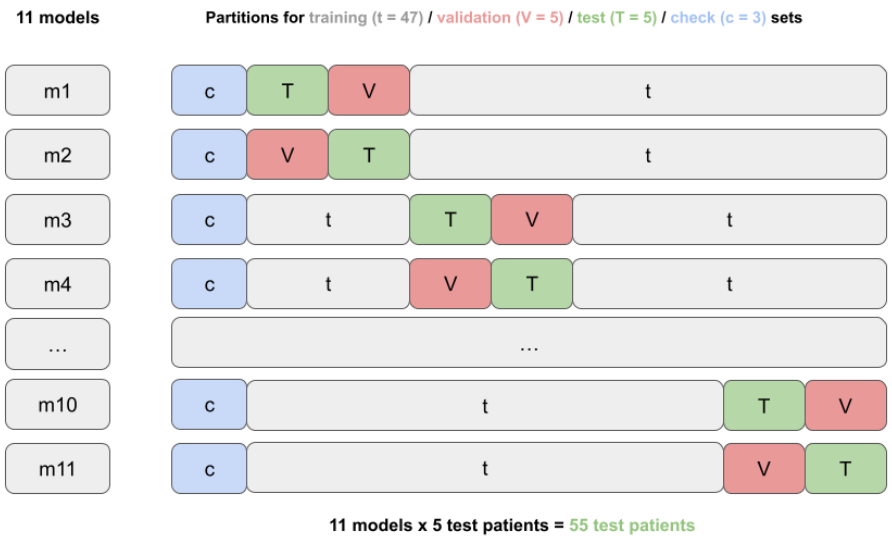

Figure 4.1: Database partitioned into training, validation, test, and check sets. The check set consists of 3 left-out patients to assess the difference between the 11 generated models. With this dataset partition a total of 55 test patients were obtained for this study. Abbreviations: m (model), c (check set), T (test set), V (validation set), t (training set).

4.2.3 Post-processing of the prediction

Before setting up the dose mimicking problem, the predicted spatial dose distribution of the UNet model goes through a post-processing (PP) phase. This phase is designed to ensure that the deliverable (mimicked) dose adheres to a set of pre-defined clinical goals. The post-processing is implemented in RayStation as a set of ROI-specific alterations of the DVHs of the predicted dose distribution, after which the spatial dose is resampled to fit the new DVHs. This step serves to drive the dose mimicking optimization towards a deliverable dose which is more clinically preferable than the raw UNet prediction.

The parameters of the post-processing step were chosen to be the same as the ones from the ML model 'RSL-IMPT-Oropharynx-7000-SIB (3.0)', which was trained and configured for a clinical protocol similar to the one of this study [227]. The parameters are presented in supplementary material Table SM2 and their detailed specifications can be found in the guide "RayStation 11B – Scripting Environments for Machine Learning". Once the post-processing settings are chosen for a specific clinic and protocol, the ML model will then be available for use in all subsequent patients without the need for manual tuning.

4.2.4 Dose mimicking strategies

The two mimicking methods described below used the RayStation worst-case optimizer with robustness parameters of 2.6% for the range error and 4 mm for the setup error as described in section 2.1. Both methods were fully integrated into the commercial TPS RayStation 11B.

Isodose based dose mimicking

Isodose Based Mimicking (IBM) used isodose volumes generated from the predicted dose every 2 Gy within 95%–107% of the prescription dose and 5 Gy outside as suggested by Bernatowicz et al. [16] for isodose volume dose restoration. Optimization was performed using minimum and maximum dose objectives on the CTVs and on each isodose volume.

Maximum and minimum dose objectives were set for isodoses above 70% of the prescribed dose, while isodoses below 70% of the prescribed dose only had a maximum objective. This allows placing the spots in a restricted area around the CTV to prevent the optimizer from setting spots in the whole patient volume. A total of 11 dose objectives were set as robust in order to ensure target coverage and avoid hot spots caused by treatment uncertainties. Contrary to other studies using IBM where the optimization objectives are patient-specific [16, 228], a

single optimization template was generated. The weights of the objectives were empirically chosen after manual tuning in the subset of the three left-out patients taking as a starting point the experience described in [228]. The final parameters were saved as the template that was used later to run the optimization (i.e. dose mimicking) for all patients. The IBM optimization objectives are detailed in the supplementary material Table SM3. IBM took around 10 minutes to complete the optimization step for the 21 robust scenarios

RayStation dose mimicking

RayStation Mimicking (RSM) is the dose mimicking method integrated in the ML model 'RSL-IMPT-Oropharynx-7000-SIB (3.0)'. It is formulated as a voxel-wise least squared error (LSE) minimization problem. The voxel weights are determined based on the geometric position of the voxel in relation to the ROIs, as well as by the assigned dose value of the voxel in the post-processed dose distribution. The LSE objective function is divided into two one-sided terms, enabling different weights depending on whether the optimized dose is greater or less than the post-processed reference dose – like the max and min dose strategy in IBM. For OAR, voxels with dose values lower than the reference dose are not penalized. In addition to the voxel-wise LSE objective, an additional term is added to the optimization problem. This term consists of a weighted sum of conventional radiotherapy objective functions, as implemented in RayStation 11B.

Similar to the post-processing parameters, the settings of the dose mimicking optimization only need to be configured once per ML Model. For this study, these settings were based on the 'RSL-IMPT-Oropharynx-7000-SIB (3.0)' model, with a change in the relative importance of the brainstem. Details on the used parameters are found in supplementary material Table SM4. The optimization procedure consists of two separate runs, between which the clinical dose is calculated accurately. These two runs consist of 125 and 80 iterations respectively, and the full optimization procedure is completed in approximately 20 minutes for the 21 robust scenarios.

4.2.5 Automatic planning strategies

In collaboration with RaySearch, 11 prediction models were generated and imported into the ML module. The four KBP pipelines (PP-RSM, PP-IBM, NPP-RSM, and NPP-IBM) shown in Figure 4.2 were fully integrated into RayStation. They could be easily run by the user with a single mouse click. PP-RSM and NPP-RSM could also be launched directly from RayStation's GUI, while scripting was necessary for PP-IBM and NPP-IBM methods. The same robustness evaluation described in 4.2.1 for the clinical dose was applied to the dose distributions

automatically generated using the proposed KBP pipelines. All dose calculations were run in RayStation 11B, using the Monte Carlo algorithm on a 3.0 x 3.0 x 3.0 mm3 dose grid.

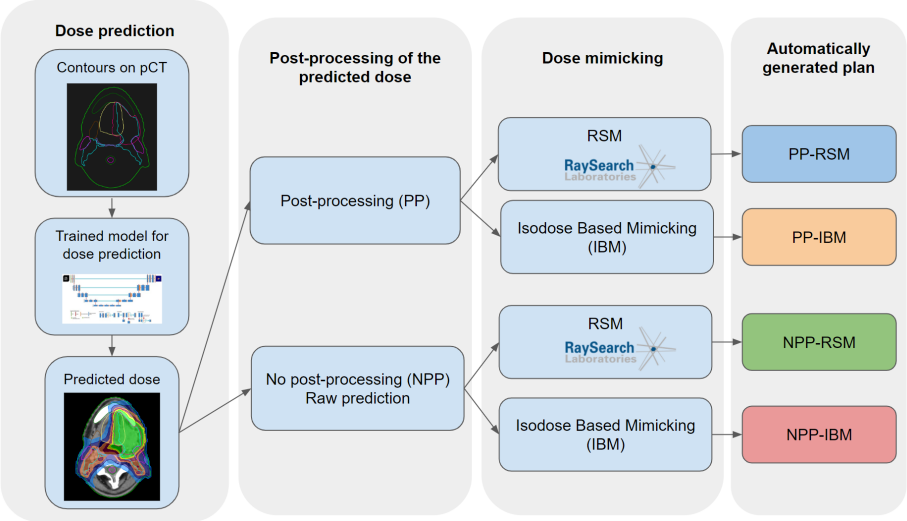

Figure 4.2: Overview of the automated knowledge-based planning pipelines evaluated in our study.

4.2.6 Evaluation metrics

Clinically relevant DVH-parameters of OAR and CTVs were compared between the four KBP pipelines and the clinical plan. Differences with respect to the clinical dose in D98 values (dose received by 98% of the volume) were calculated for low-dose and high-dose CTVs, while differences in Dmean (average dose) were reported for the OAR. These metrics, evaluated on the 55 patients, were displayed using boxplots. The median and interquartile range (IQR) values from the boxplots illustrate how much these metrics differ from the clinical dose for each KBP method.

For low-dose and high-dose CTV, nominal scenario (with no uncertainty) and worst-case scenario values were calculated from the 42 uncertainty explored scenarios (SE = 4mm, RE = 2.6%) during robustness evaluation as described in section 2.1. Only nominal values for the OAR average dose were reported.

Grouped data were tested for statistical significance using two-sided Wilcoxon-Mann-Whitney tests using SciPy version 1.5.2. The differences were considered statistically significant for p-values ¡ 0.05.

4.3 Results

4.3.1 Target coverage

For the four KBP pipelines, the clinical goals for target coverage (i.e., D98% 95%) in both high-dose and low-dose CTVs were met for all the patients in the nominal scenario. Figure 4.3 shows the difference in DVH-metrics between the clinical dose and the four investigated methods.

High-dose CTV target coverage decreased by less than 0.50 Gy in median for the four strategies. PP-IBM plans presented the largest decrease (i.e lowest D98 values) with a reduction in the median value of -0.42 (iqr = 0.27) Gy. NPP-RSM plans achieved the best coverage with an increase in the median value of 0.05 (iqr = 0.20) Gy with respect to the clinical dose.

For the low-dose CTV, D98 values were closer to no difference or even higher than clinical D98 values. All KBP methods had less than a 0.12 Gy reduction in the median target coverage. The difference in D98 values for the low-dose CTV between the PP-IBM and clinical plans was not statistically significant.

In contrast to the nominal scenario, in the worst-case scenario, not all the clinical goals were met for all patients. Especially in the high-dose CTV, plans where post-processing was applied, PP-RSM and PP-IBM, experienced a significant drop in target coverage with -1.75 (iqr = 1.24) Gy and -2.35 (iqr = 1.65) Gy, respectively. 65% (35/55) and 82% (45/55) of patients failed to achieve high-dose CTV coverage in the worst-case scenario for PP-RSM and PP-IBM, respectively. Without post-processing, only 5% (3/55) and 11% (6/55) of patients failed to meet the robustness criteria for NPP-RSM and NPP-IBM, respectively.

A higher robustness level was observed for the low-dose CTV. All patients achieved target coverage in the worst-case scenario for PP-RSM and NPP-RSM plans. For PP-RSM, there was no statistically significant difference in the worst-case scenario values with respect to clinical plans. However, PP-IBM and NPP-IBM did not meet adequate low-dose CTV coverage in the worst-case scenario in 5% (3/55) and 10% (6/55) of the patients.

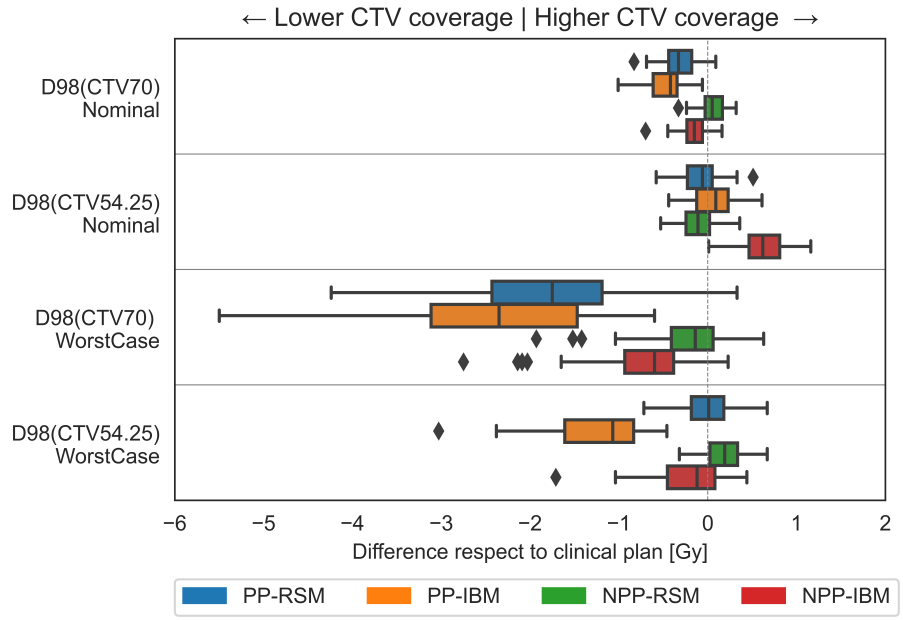

Figure 4.3: The Differences in terms of target coverage (D98 - dose received by 98% of the CTV) between ATP plans and the clinical plans in nominal and worst-case scenarios. Negative differences imply that the clinical plan was better than the ATP method for that metric. The boxes indicate the median and interquartile range (IQR). The whiskers extend to a minimum of 1.5 times the IQR and diamonds represent the outliers.

4.3.2 Organs at risk sparing

Figure 4.4 presents the difference in the average dose received by the OAR between the clinical dose and the four studied KBP methods. For all OAR metrics, the differences with respect to the clinical plan and between methods were statistically significant. In most of the cases, the plans generated with KBP pipelines achieved superior OAR sparing than the clinical plans. Median values of the average doses were lower than in the clinical plan for most OAR. The few exceptions were the esophagus in NPP-RSM plans and the SMGs in NPP-IBM plans. As expected, OAR sparing on the oral cavity, the parotids, and PCMs was especially aggressive for plans including post-processing according to the pre-defined settings. PP-RSM and PP-IBM plans saved with respect to the clinical plan median values of more than 6 Gy to the oral cavity, more than 5 Gy to the parotids, and between 4.5 and 5 Gy to the PCMs. An illustrative example is shown in Figure 4.5.

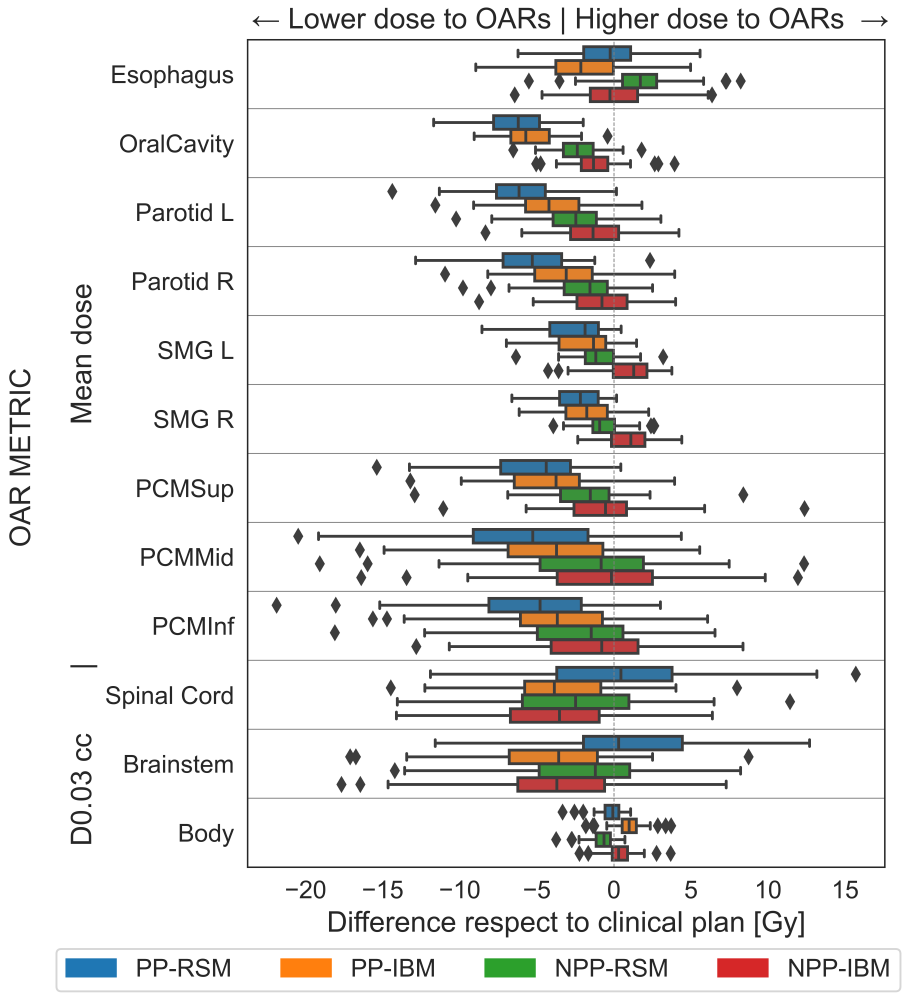

Figure 4.4: The differences in terms of clinical planning criteria for the OAR between ATP plans and the clinical plans Negative differences imply that the clinical plan was better than the ATP method for that metric. The boxes indicate the median and interquartile range (IQR). The whiskers extend to a minimum of 1.5 times the IQR and diamonds represent the outliers.

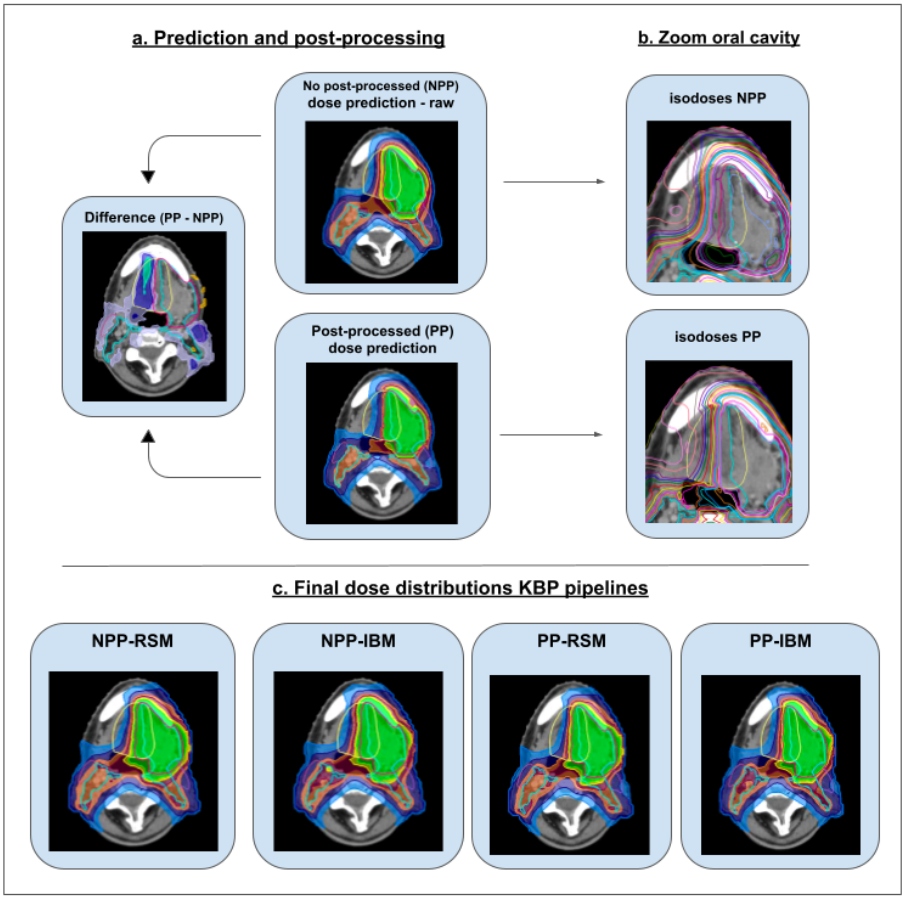

Figure 4.5: Different steps of the proposed KBP pipelines in a patient example. The image on the left (a) shows the differences between the raw predicted dose (NPP) and the post-processed predicted dose (PP): light blue (- 4 Gy), dark blue (-7 Gy), and turquoise blue (-15 Gy). In the middle (b) the isodose contours generated from the PP and NPP doses that are used during isodose based mimicking are presented. Specifically, a zoomed image of the oral cavity region is shown. On the left (c), the final result of the four proposed KBP workflows for this patient can be seen. The delineated contours are high-dose CTV (magenta), low-dose CTV (turquoise blue) and the oral cavity (yellow). The isodoses characterizing the dose distributions are represented with different colours as a percentage of the prescribed dose of 70 Gy: light blue (25%), dark blue (50%), orange (74%), dark red (81%), yellow (90%), green (95%) and red (105%). The green isodose corresponds to 95%(70 Gy) and the white isodose is 95%(54.25 Gy).

4.4 Discussion

Automated KBP pipelines for IMPT planning necessitate a comprehensive evaluation of all the steps included in the workflow for a safe clinical implementation

[219]. In this study, the clinical IMPT plans of 55 oropharyngeal patients were compared to automated IMPT plans generated using four different KBP pipelines (PP-RSM, PP-IBM, NPP-RSM, and NPP-IBM). The impact of a post-processing strategy combined with two different robust dose mimicking strategies was analyzed. These two steps of the KBP pipelines, coming after the dose prediction (post-processing and dose mimicking), have not been fully addressed in the literature, yet they are essential to judge the clinical acceptability of automatically generated IMPT plans.

The four proposed KBP pipelines, generated nominal dose distributions meeting all the clinical goals. In many cases, they even achieved superior quality in the OAR sparing compared to manually planned clinical doses, especially both KBP pipelines that use post-processing before dose mimicking: PP-RSM and PP-IBM. The potential of reduction of the dose to OAR in HNC patients is especially important in order to reduce radiation-induced side effects, in particular xerostomia and dysphagia, which have a major impact on quality of life [229, 230].

On the other hand, a dosimetric gain in the OAR can come at the cost of losing robustness. The results show that in some cases, our post-processing parameters compromise the robustness of the plans, especially for the high-dose CTV. This is probably due to the reduction of the dose to the OAR during the post-processing step. In particular, the dose reduction of 20% to the oral cavity from the predicted value, combined with the push for target homogeneity in the high-dose CTV, generates unrealistic dose gradients where both regions of interest meet (Figure 5a. and 5c.). These unrealistic dose gradients are then propagated to the isodose volumes, which end up being very close to each other (Figure 5b.). In such cases, the optimization problem becomes extremely challenging for PP-IBM plans. IBM could not compensate for the loss in robustness in the high-dose CTV but was able to maintain correct target coverage for the low-dose CTV, where the dose gradients were not so steep. A more sophisticated mimicking process, such as RSM, seemed to partially recover the robustness. However, PP-RSM still failed to satisfy high-dose CTV robustness in 65% of the cases. The lack of robustness in the target coverage could be translated into cold spots (i.e underdosage) during the treatment delivery, which could lead to a drastic reduction of the tumor control probability [231].

It was shown that the construction of a KBP pipeline, including post-processing and dose mimicking, is an iterative process that requires adjustment of the settings provided by the commercial solution. With the current parameters of our study, NPP-RSM plans seem to achieve the best trade-off between robustness and OAR sparing. They present comparable target coverage as in the reference clinical plan and higher OAR sparing, except for the esophagus. In order to improve the

generated plans using NPP-RSM, specific objectives aiming at reducing the dose to the esophagus could be introduced in the mimicking step. Another interesting option would be to adjust the post-processing parameters for PP-RSM plans to maximize the potential of the OAR sparing.

It should be pointed out that the post-processing parameters were not configured specifically for this study, but were taken from another (similar) model which might have been less strict in the desired robustness. It is likely that one would find a good set of post-processing parameters for this study as well, if iterated a few times. Therefore, it would be worth exploring if less aggressive post-processing can maybe still improve OAR sparing while not compromising so much on the robustness of the plans. In any case, these results highlight the difficulty of generating universal KBP strategies (combining dose prediction, post-processing and dose mimicking) that are interchangeable between clinical protocols, even if they are similar. Although the same prediction model could be used, the post-processing and dose mimicking parameters would highly rely on specific clinical preferences.

The only input to the ML dose prediction used in our proposed KBP pipelines was the contours. Recent studies about dose prediction for radiotherapy have explored the use of auxiliary information (e.g., non-modulated beam doses) to improve the robustness of the ML model [189, 232]. Assuming that the robustness of a dose distribution relies on the IMPT plan, another option would be to directly predict those plan parameters (e.g. energy layers, spot positions and spot weights). Approaches suggesting the latter have been described in the literature using fluence map prediction to generate photon plans [233, 234, 235] but still need to be investigated for protons. However, even if ML models are excellent methods for fast, automatic, and free-of-any human intervention tasks, they are also well known for their low interpretability [150]. For the moment, there is no evidence suggesting that ML methods could also outperform the already available, fast, automatic and well-studied analytical computing models for specific tasks like dose calculation or treatment plan optimization. Therefore, with the current technology, ML models might not be suitable to replace every step in the treatment planning workflow, but they are excellent tools to ease and automate most manual steps.

Eleven dose prediction models were used in our study to increase the number of testing patients instead of a single model. Therefore, a bias towards each model for each subgroup of five testing patients could be derivated from this procedure. Nevertheless, the difference between models in the three left-out patients seem to be insignificant in the target volumes and most of the OAR and in any case not higher than 7 Gy for any structure (supplementary material Figure SM1). In some cases, it might be that for some OAR (e.g the PCMs) the differences shown with

respect to the ground truth were also influenced by the prediction model.

It is also important to mention the wide range of applications for automatic RT planning. First, RT plans could be inferred in a few minutes for different treatment modalities (e.g., IMPT versus conventional RT), in order to refer the patient to the most optimal treatment [222, 215, 236]. Second, automatically generated plans could be also used to reduce inter-planner and inter-center variability. Moreover, this approach could be used to inform the planner about where there is room for improvement [237]. Third, the planning workflow which can take several hours or even days, could maybe take an automatically generated plan as a starting point and apply fine-tuning if necessary. Finally, automatic planning could allow for wide implementation of online adaptive IMPT on high-quality daily images (e.g repeated CT images or corrected CBCT images [100, 238, 239]). In all previously mentioned scenarios, automatic planning allows for huge time savings, efficient resource usage, or improved plan quality.

In conclusion, this study highlights the crucial necessity of finding the appropriate combination of post-processing settings and dose mimicking strategy to achieve robust deliverable IMPT doses when using automatic planning in oropharyngeal cancer. We stress the importance of a comprehensive robust evaluation of the final dose calculations in the feedback loop to configure a KBP pipeline with the desired clinical specifications.