Surrogate-free machine learning-based organ dose reconstruction for pediatric abdominal radiotherapy

M Virgolin[1‡], Z Wang[2‡], B V Balgobind[2], I W E M van Dijk[2], J Wiersma[2], P S Kroon[3], G O Janssens[3,4], M van Herk[5], D C Hodgson[6], L Zadravec Zaletel[7], C R N Rasch[8], A Bel[2], P A N Bosman[1,9], T Alderliesten[2,8]

[1] Life Sciences and Health Group, Centrum Wiskunde & Informatica, the Netherlands

[2] Department of Radiation Oncology, Amsterdam UMC, University of Amsterdam, the Netherlands

[3] Department of Radiotherapy, University Medical Center Utrecht, the Netherlands

[4] Princess Màxima Center for Pediatric Oncology, the Netherlands

[5] Manchester Cancer Research Centre, Division of Cancer Sciences, University of Manchester, United Kingdom

[6] Department of Radiation Oncology, Princess Margaret Cancer Centre, Canada

[7] Department of Radiation Oncology, Institute of Oncology Ljubljana, Slovenia

[8] Department of Radiation Oncology, Leiden University Medical Center, the Netherlands

[9] Department of Software Technology, Algorithmics Group, Delft University of Technology, the Netherlands

E-mail: `marco.virgolin@cwi.nl, z.wang@amsterdamumc.nl`

[‡] Shared first author, the two authors contributed equally to this work

Abstract. To study radiotherapy-related adverse effects, detailed dose information (3D distribution) is needed for accurate dose-effect modeling. For childhood cancer survivors who underwent radiotherapy in the pre-CT era, only 2D radiographs were acquired, thus 3D dose distributions must be reconstructed from limited information. State-of-the-art methods achieve this by using 3D surrogate anatomies. These can however lack personalization and lead to coarse reconstructions. We present and validate a surrogate-free dose reconstruction method based on Machine Learning (ML). Abdominal planning CTs ($n = 142$) of recently-treated childhood cancer patients were gathered, their organs at risk were segmented, and 300 artificial Wilms' tumor plans were sampled automatically. Each artificial plan was automatically emulated on the 142 CTs, resulting in 42,600 3D dose distributions from which dose-volume metrics were derived. Anatomical features were extracted from digitally reconstructed radiographs simulated from the CTs to resemble historical radiographs. Further, patient and radiotherapy plan features typically available from historical treatment records were collected. An evolutionary ML algorithm was then used to link features to dose-volume metrics. Besides 5-fold cross validation, a further evaluation was done on an independent dataset of five CTs each associated with two clinical plans. Cross-validation resulted in mean absolute errors $\leq$0.6 Gy for organs completely inside or outside the field. For organs positioned at the edge of the field, mean absolute errors $\leq$1.7 Gy for $\mathrm{D_{mean}}$, $\leq$2.9 Gy for $\mathrm{D_{2cc}}$, and $\leq$13% for $\mathrm{V_{5Gy}}$ and $\mathrm{V_{10Gy}}$, were obtained, without systematic bias. Similar results were found for the independent dataset. To conclude, we proposed a novel organ dose reconstruction method that uses ML models to predict dose-volume metric values given patient and plan features. Our approach is not only accurate, but also efficient, as the setup of a surrogate is no longer needed.

Keywords: dose reconstruction, radiotherapy dosimetry, machine learning, plan emulation, childhood cancer, late adverse effects

Submitted to: *Phys. Med. Biol.*

1. Introduction

Patients undergoing radiotherapy (RT) are prone to develop radiation-related Adverse Effects (AEs) (Birgisson et al 2005, van Dijk et al 2010, Cheung et al 2017). To improve the design of future multi-modality treatments, clinicians are interested in better understanding the relationship between radiation dose and onset of AEs. Modern research efforts in this direction delve into dosimetric details, employing dose distribution metrics to a specific organ (or sub-volume) as explanatory variables. Such rich information is obtained by simulating the RT plan on 3D imaging of the patient (i.e., CT scans) with organ segmentations in a Treatment Planning System (TPS) (Donovan et al 2007, Feng et al 2007, Bölling et al 2011).

Unfortunately, when so-called *late* AEs (onset can be decades after RT) need to be studied, it is not always possible to straightforwardly obtain detailed information on dose distributions (Birgisson et al 2005). For patients who underwent RT before the use of planning CTs became commonplace (in the following, *historical patients*), 2D radiographs were used for treatment planning (e.g., this was the case until the 1990s in the Netherlands (van Dijk et al 2010)), meaning no 3D anatomical imaging is available. Consequently, no simulations can be performed in a TPS to estimate 3D dose distributions for these patients (Stovall et al 2006, Verellen et al 2008, Ng et al 2012). The information available for historical patients normally consists of what was reported in treatment records, e.g., features of the patient such as age and gender, and features of the plan such as prescribed dose, geometry of the plan, and the use of blocks. Additionally, 2D radiographs can be available, from which information can be gathered on the internal anatomy (mainly bony anatomy, as internal organs are normally not clearly distinguishable), and on the plan configuration with respect to the patient's anatomy (Leisenring et al 2009, van Dijk et al 2010).

To improve the understanding of late AEs, recent research is striving to develop increasingly accurate *dose reconstruction* methods, i.e., methods to estimate the 3D dose distribution received by historical patients (Stovall et al 2006, Ng et al 2012, Xu 2014, Lee et al 2015). State-of-the-art approaches employ *phantoms*, i.e., 3D surrogates of the human anatomy upon which the RT plan can be simulated, to compute the dose distribution. Phantoms exist in different forms: physical or virtual, made by simple geometrical shapes or by adopting and morphing actual CT scans and organ segmentations (Stovall et al 2006, Xu 2014, Lee et al 2015). Generally, phantoms are built to represent *average* anatomies, for categories of patients (e.g., for a certain age range), and are collected into so-called phantom libraries (Cassola et al 2011, Segars et al 2013, Geyer et al 2014). Whenever dose reconstruction for a historical patient is needed, the phantom that represents the category that the patient belongs to is retrieved from the library and used as surrogate for simulation of the RT plan.

As the largest source of error related to phantom-based dose reconstruction comes from the mismatch between the anatomy of the phantom and the true anatomy of the patient (Bezin et al 2017), it is important to define the best way to match phantoms to patients. This issue is still under research, and different approaches employ different heuristic matching criteria that are normally hand-crafted and based upon statistics and guidelines drawn from large population studies (e.g., ICRP89, NANTHES) (Valentin 2002, Cassola et al 2011, Segars et al 2013, Geyer et al 2014). However, the use of heuristic matching criteria has been hypothesized to be too simplistic to capture the high variability of internal human anatomy (de la Grandmaison et al 2001, Geyer et al 2014, Xu 2014, Virgolin et al 2018*b*, Wang et al 2019). For example, a popular phantom-based dose reconstruction approach uses solely age and gender for surrogate matching (Howell et al 2019). Our group's recent work focusing on Wilms' tumor (the most common type of kidney cancer for childhood cancer patients) irradiation for pediatric patients showed that utilizing surrogate CTs using age- and gender-based matching can lead to poor dose reconstruction quality in individual cases (Wang et al 2018).

To improve the resemblance of a surrogate phantom, there have been efforts to replace the normally hand-crafted heuristic matching criteria with data-driven decisions. For example, statistical models inferred from CTs and 3D organ segmentations of adult patients have been used to drive a deformable image registration procedure that adapted 3D organ segmentations to the 2D anatomy of a specific patient, given features of the latter as measurable from 2D radiographs (Ng et al 2012, Mishra et al 2013). Using a state-of-the-art Machine Learning (ML) algorithm, it has been shown that features typically available for historical patients treated for Wilms' tumor can be linked to different 3D anatomy similarity metrics based on organ segmentations and CTs (Virgolin et al 2018*b*). Our group recently proposed an automatic pipeline that uses ML to steer the assembling of a new original anatomy based on 3D CTs and organ segmentations of multiple patients using the features of a historical patient (Virgolin et al 2019, Virgolin et al 2020*b*). However, it is important to realize that maximizing some form of overall anatomical resemblance is difficult. Moreover, from the standpoint of optimizing dose reconstruction accuracy, it can be considered sub-optimal for RT dosimetry purposes. This is because in RT dosimetry, what part of anatomy is most meaningful largely depends on the particular RT plan (Wang et al 2019).

To the best of our knowledge, although *both* patient anatomy and plan geometry play a key role in determining dose-volume metrics for Organs At Risk (OARs), existing dose reconstruction approaches focused solely on patient anatomy information, to obtain a representative surrogate. Plan information is used only later, to calculate the dose on the surrogate. The purpose of this article is to develop and validate an ML approach to predict dose-volume metrics for OARs based on patient anatomy and plan geometry information. Specifically, we propose to use ML to directly learn what dose-volume metrics for an OAR are likely given information on the patient and on the plan, without the need to select or craft any surrogate anatomy. We argue that this is a sensible choice because ML can directly be trained upon what ultimately matters, i.e., dose reconstruction accuracy. In this article we present our ML-based organ dose reconstruction approach and its validation, that was performed on a relatively large dataset of artificial plans, as well as on a smaller dataset of clinical plans.

2. Materials & Methods

We considered pediatric flank RT, and in particular RT for Wilms' tumor, as an application for our dose reconstruction method, in continuity with our previous work. The choice to focus on pediatrics is because children are the most prone to develop late AEs (Cheung et al 2017), and are typically underrepresented in existing phantom libraries (Xu 2014). Moreover, more than 85% of pediatric patients survives Wilms' tumor five years or longer, but considerable chances of the onset of late AEs remain (van Dijk et al 2010).

2.1. Patient data

To be able to create a ground-truth to learn dose-volume metrics from, CT scans were needed. Hence, a total of 142 pediatric planning CTs were collected by involving the following institutes (number of CTs in brackets): Amsterdam University Medical Centers / Emma Children's Hospital ($n = 38$), University Medical Center Utrecht / Princess Máxima Center for Pediatric Oncology ($n = 42$), The Christie NHS Foundation Trust ($n = 33$), Princess Margaret Cancer Centre ($n = 18$), and Institute of Oncology Ljubljana ($n = 11$). Five further CTs were collected from the Amsterdam University Medical Centers and kept aside to be used exclusively for an additional validation step (Sec. 2.5).

The inclusion criteria were: patient age at scan acquisition between 1 to 8 years; the CT field of view including a common abdominal region from the tenth thoracic (T10) vertebral body to the first sacral (S1) vertebral body; presence of five lumbar vertebrae (rare cases of patients with six exist); patient scanned in supine position; quality of CT sufficient to perform organ segmentations. The patients underwent RT between 2002 and 2018, mostly but not exclusively for abdominal cancers. The median CT slice resolution was 0.94×0.94 mm, the median slice thickness was 3 mm.

As we focused on Wilms' tumor treatment, four OARs were considered: the liver, the spleen, the contralateral kidney (left or right, depending on the side of the tumor), and the spinal cord (between T10 and S1). To provide accurate and consistent OAR segmentations, we carefully prepared the OAR segmentations in all CTs ($n = 142 + 5$): for 60 CTs, pre-existing clinical segmentations of OARs were manually improved and approved (I.W.E.M. van Dijk, checked by B.V. Balgobind only for difficult cases). To aid the manual segmentation of the OARs for the remaining CTs (n=87), the software ADMIRE (research version 2.3.0) from Elekta (Elekta AB, Stockholm, Sweden) was used to generate multi-atlas based automatic segmentations using the previous 60 CTs as atlas. These segmentations were further manually checked slice-by-slice and possibly adapted (Z. Wang, I.W.E.M. van Dijk), and finally checked and approved (I.W.E.M. van Dijk, checked by B.V. Balgobind only for difficult cases). Some patients did not have both kidneys intact, due to nephrectomy prior to RT. The number of CTs that had only a complete right kidney, only a complete left kidney, and two complete kidneys were 36, 40, and 71, respectively.

2.2. Automatic generation of artificial Wilms' tumor plans

A method to automatically generate historical-like abdominal flank irradiation plans (i.e., artificial plans) for Wilms' tumor treatment based on information visible on 2D radiographs was created, in order to obtain large plan variations.

Figures 1(a) and 1(c) illustrate examples of actual historical plans on respective historical radiographs. As can be observed from the examples, a typical historical flank irradiation field is a rectangular area, with possible shielding blocks, that is located on the right or on the left flank. Flank irradiation is done by beams from anterior-posterior (AP) and posterior-anterior (PA) direction. Along right-left (RL), one field border is located at the edge of the patient's body contour, while the other is located as to include the vertebral column (van den Heuvel-Eibrink et al 2017). In some cases, blocks were placed to protect OARs from irradiation (Fig. 1(c)). In historical plans the isocenter was positioned in the center of the treatment field that is projected on the coronal plane (Fig. 1) and at the middle of the patient's AP abdominal diameter.

To generate artificial plans, two reference digitally reconstructed radiographs (DRRs) were considered, randomly selected from the data. One DRR was derived from a CT of a 5-year old female patient without nephrectomy (ref 1 in Fig. 2), and the other was derived from a CT of a 4-year old female patient with nephrectomy of the left kidney (ref 2 in Fig. 2). Upon these two DRRs, boundaries defining plan variability were identified by an experienced pediatric radiation oncologist (B. V. Balgobind), to ensure that generated plans would appear to be reasonable according to historical clinical guidelines. Note that historical clinical guidelines are slightly different from current ones (e.g., currently the iliac crests should be safeguarded, unlike in Fig. 1(c)). Figure 2 shows two examples of landmark locations identifying possible plan variations, on the two reference DRRs. Specifically, given the boundaries of possible isocenter positions and field borders, plans with a rectangular field were generated by sampling *uniformly* within those boundaries. For each plan generated, an additional version of that plan including one block was generated as well. A block was simulated as the area in the upper lateral corner enclosed by the border of the rectangular field and a line crossing two randomly sampled endpoints. The endpoints were sampled from two regions roughly covering the start and end points of rib 9 and rib 12 on the DRRs (regions indicated by the green boxes in Fig. 2). This way, a sampled block covered part of the liver (in right-sided plans) or part of the spleen (in left-sided plans). All plans consist of two opposing and symmetrical beams in AP-PA directions irradiating one side of the abdominal flank. Figures 1(b) and 1(d) illustrate two examples of sampled artificial plans (without or with a block) on respective DRRs.

A total of 300 artificial plans were generated automatically, of which 150 without a

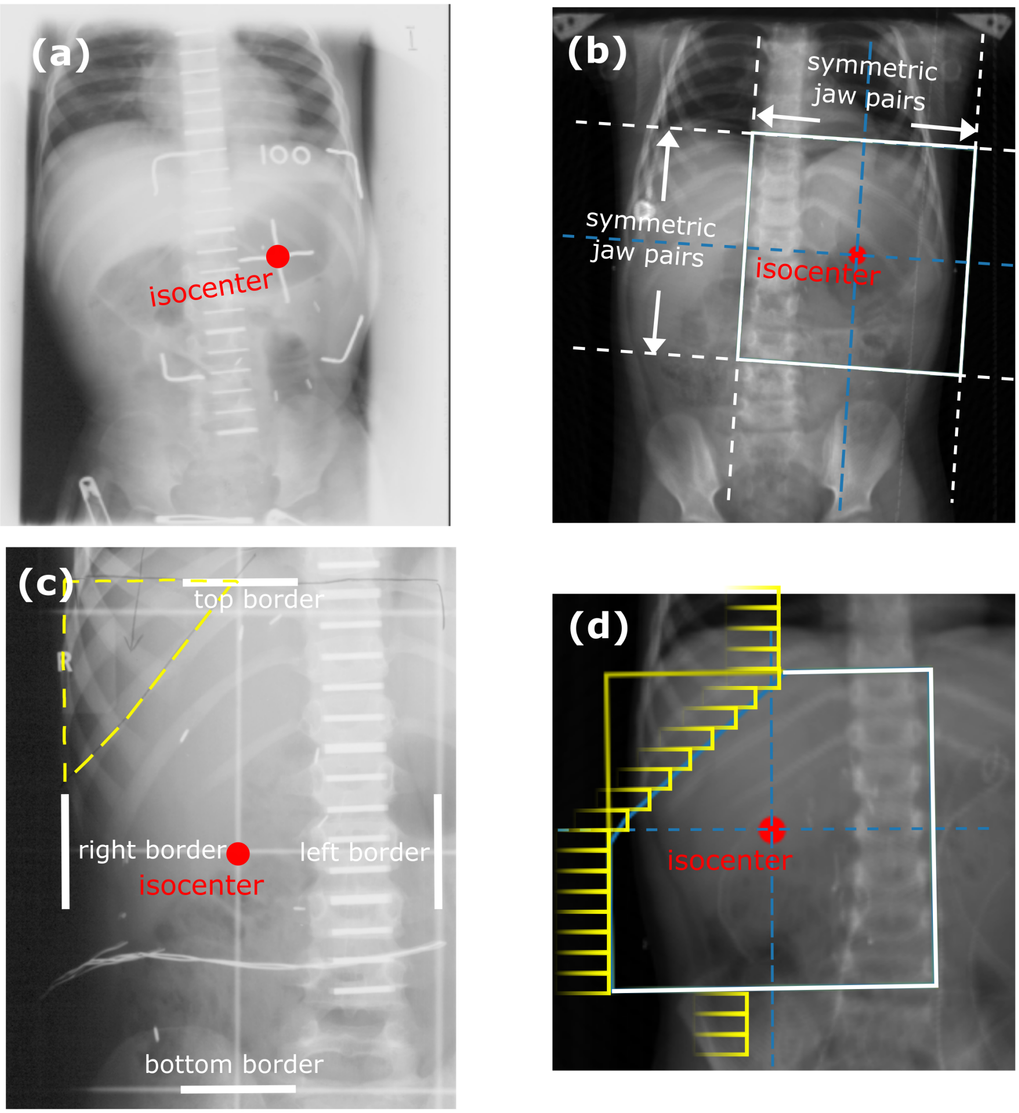

Figure 1. (a) An actual hand-drawn plan on a historical radiograph with a rectangular field (indicated by white corners). (b) An artificial plan with a rectangular field (in white lines) plotted on the DRR of a recent patient. (c) An actual hand-drawn plan on a historical radiograph with a rectangular field (in white bars) and an additional block (outlined by dashed yellow lines) to spare part of the liver. (d) An artificial plan plotted on the DRR of a recent patient with a rectangular field (in white bars) and an additional block (obtained by multi-leaf collimators, outlined by yellow lines) to spare part of the liver. For each plot, the isocenter is indicated by a red dot in the middle of the field.

block, and 150 with a block. The random sampling of the plan side led to 142 left-sided plans and 158 right-sided plans (roughly half-half). The same set of plan features used in our previous work was considered to generate plans in DICOM RTPLAN format (e.g., gantry and collimator angles, isocenter location, field sizes) (Wang et al 2020).

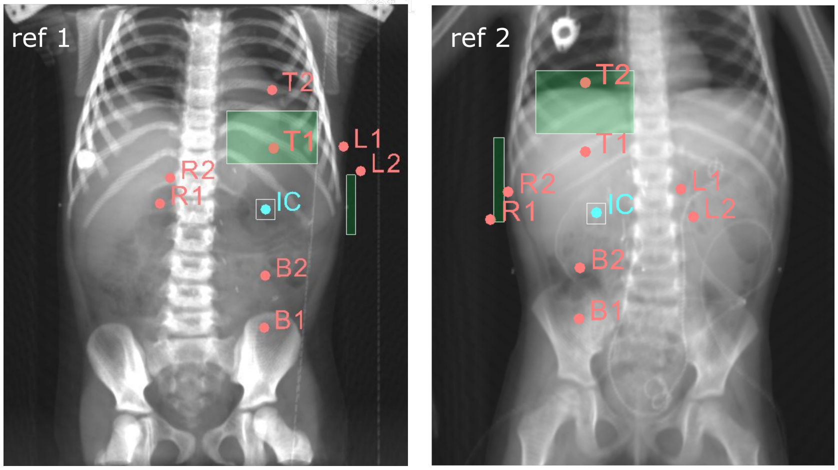

Figure 2. Examples of landmark locations to specify geometry variability of two types of artificial plans (left-sided plans in the left figure and right-sided plans in the right figure). Ref 1 is the DRR derived from the reference CT of a 5-year-old female patient and ref 2 is the DRR derived from the reference CT of a 4-year-old female patient. The box around the isocenter (IC) specifies the range of possible isocenter positions. The vertical position of T1/T2 and of B1/B2 specify the lowest/highest position of the upper and lower border of the field, respectively. The horizontal positions of R1/R2 and L1/L2 specify the rightmost/leftmost position of the right and left border of the field, respectively. The isocenter and artificial field border positions were sampled uniformly at random within the specified ranges. The green boxes indicate the regions where two endpoints of a line representing a block border can be sampled. This line, together with the upper and left/right field borders, encloses the block.

2.3. Generation of the dataset for ML

Figure 3 summarizes the pipeline used to generate the dataset for ML. Firstly, we emulated each of the 300 artificial plans on each of the 142 CT scans by the automatic plan emulation method proposed in our previous work (Wang et al 2020), leading to a total of 42,600 emulations. The method automatically transfers a plan prepared on one CT to another CT (with quality comparable to human experts), using landmark detection upon the respective DRRs. Secondly, for each of the 42,600 plan emulations, dose-volume metrics of interest (see Sec. 2.3.1) were collected for the different OARs by use of our automatic dose computation pipeline (Wang et al 2020). The pipeline used the collapsed cone dose calculation algorithm of Oncentra TPS (version 4.3, Elekta AB, Stockhom, Sweden). Thirdly, features that are plausible to be available for typical historical cases were collected from the anatomy of the included CTs as visible in the respective DRRs, from the artificial plans, and from the relationship between anatomy and plan geometry.

2.3.1. Response variables: dose-volume metrics To select dose-volume metrics to use as response variables for ML, we considered metrics typically used to validate state-of-the-art dose reconstruction approaches (Ng et al 2012, Lee et al 2015), and typically found

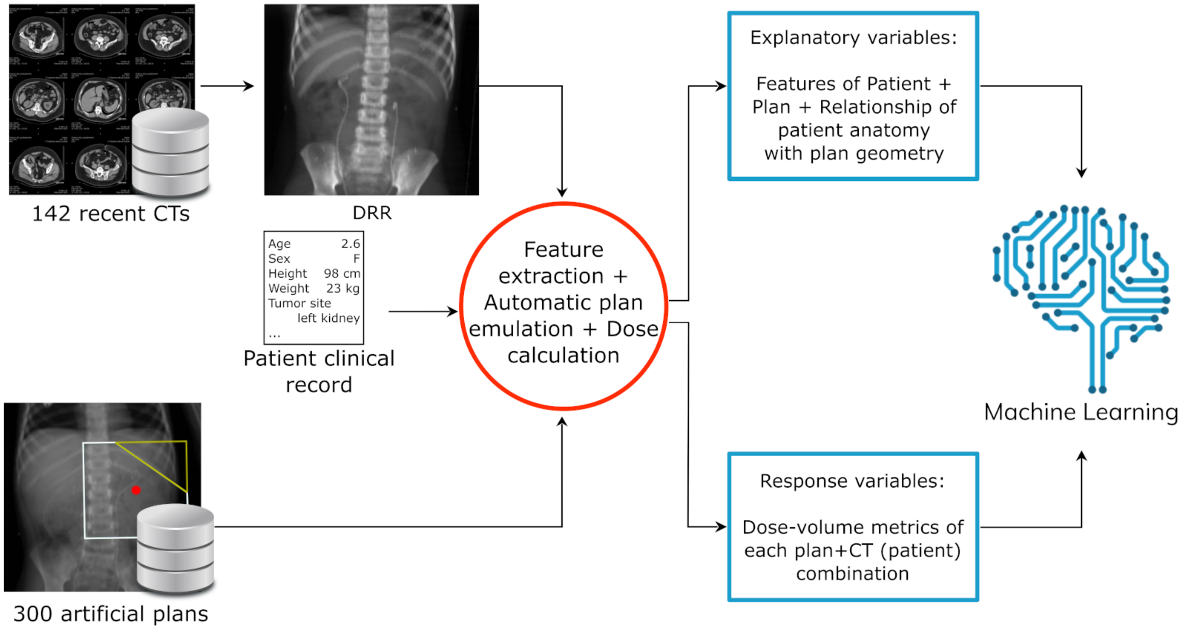
142 recent CTs
DRR
Age 2.6
Sex F
Height 98 cm
Weight 23 kg
Tumor site left kidney
...
Patient clinical record
300 artificial plans
Feature extraction + Automatic plan emulation + Dose calculation
Explanatory variables:
Features of Patient + Plan + Relationship of patient anatomy with plan geometry
Response variables:
Dose-volume metrics of each plan+CT (patient) combination
Machine Learning

Figure 3. Pipeline for data generation. Artificial plans are sampled automatically. The explanatory and response variables are used as input to train the ML model. The explanatory variables include features of the plan (e.g., isocenter location, field size), patient features (e.g., age, nephrectomy), and features on the relationship between the anatomy of the patient and the geometry of the plan (e.g., signed distance between the 2$^{\mathrm{nd}}$ lumbar vertebra and the plan isocenter). The response variables are dose-volume metrics for each OAR.

to be of clinical relevance in studies of AEs in adults (e.g., QUANTEC) (Leisenring et al 2009, Bölling et al 2011, Emami et al 1991). Studies on dose-volume response relationships for pediatric patients (so-called PENTEC studies (Constine et al 2019)) are currently limited.

This reasoning led us to consider mean organ dose (D_{mean}), two levels of percentage of OAR volume receiving at least X Gy (V_{XGy}), and the minimum dose received by the maximally exposed 2 cubic centimeters of an OAR (D_{2cc}), the latter being similar but more robust than the maximum dose to a single point. Typically, V_{5Gy} and V_{20Gy} are considered. However, instead of V_{20Gy}, we decided to use V_{10Gy} in this work since our plans have a prescribed dose of 14.4 Gy (thus V_{20Gy} was always 0 for the OARs). Regarding D_{2cc}, we decided to include this metric because peak dose values to a small OAR portion may be relevant to explain late AEs related to OARs that work in a serial fashion (e.g., the spinal cord).

2.3.2. Explanatory variables: features of patients and plans To assess what information can be available for historical patients, we considered the Dutch records of the Emma Children's Hospital/Academic Medical Center childhood cancer survivor cohort, who underwent RT between 1966 and 1996 (van Dijk et al 2010). For this cohort, along with historical patient records and treatment plan details, 2D coronal radiographs were consistently taken, hence providing partial information on the anatomy.

The complete set of features considered in this work is reported in Table 1. Note the absence of height and weight (which are used by some phantom-based methods (Geyer et al 2014)). For 12% of the patients, height and/or weight data were missing and preliminary experiments using automatic imputation methods showed no benefit in including them.

For the features related to anatomical geometry, anatomical landmarks from DRRs were detected automatically using the landmark detection method in our previous work (Wang et al 2020). Note that the landmarks concerned only bony anatomy because other internal anatomy tissues are not reliably visible in historical radiographs. Importantly, we normalized features related to measurements of anatomy and anatomy-plan geometry configuration (e.g., rib-cage width, field sizes, distances between landmarks and the isocenter) by the width and height of the respective DRR they were measured from (after the DRRs were cropped to a same region of interest between T10 and S1). This was done because when plans are emulated, they are scaled based on proportions derived from the landmarks (Wang et al 2018, Wang et al 2020). Since differences in anatomy solely due to overall anatomy scaling do not result in different dose-volume metric values, these differences should not be accounted for by the explanatory variables (confirmed in preliminary experiments).

The abdominal diameter in AP ($Diam_{AP}^{IC}$) is the only anatomical feature not measurable from DRRs generated along AP/PA direction. In historical RT, it was measured using a ruler to determine the isocenter position along the AP axis, and was subsequently reported in the records. For our cohort, $Diam_{AP}^{IC}$ was not reported in the records because a CT scan was used for RT planning. We therefore measured $Diam_{AP}^{IC}$ automatically on the CT scans, by using a pre-determined isocenter position of typical abdominal flank irradiation plans. In particular, the intervertebral disk between the 1st and the 2nd lumbar vertebra (L1 and L2) was used to determine the isocenter position along the inferior-superior (IS) axis, and the center of mass of the kidney was used to determine the isocenter position along the RL axis (as the aim of Wilms' tumor plans is to irradiate the renal fossa). For CTs including both kidneys, two $Diam_{AP}^{IC}$ were measured, and the average was taken. Conversely, only one $Diam_{AP}^{IC}$ was measured for CTs including a single kidney.

In our simulations we used for all artificial plans the same fractionation scheme (8 × 1.8 Gy), beam energy (6 MV), and prescribed dose (14.4 Gy at isocenter). These settings are the most common in historical records, and are still valid in the current Wilms' tumor RT protocol (van den Heuvel-Eibrink et al 2017). Moreover, choosing a specific prescribed dose (e.g., 14.4 Gy) does not limit generalizability, since the dose distribution over the entire anatomy depends linearly on the prescribed dose. Thus, if dose reconstruction for a historical case using a different prescription is needed, the dose-volume metrics predicted by the models trained on plans with a 14.4 Gy prescribed dose can be re-scaled.

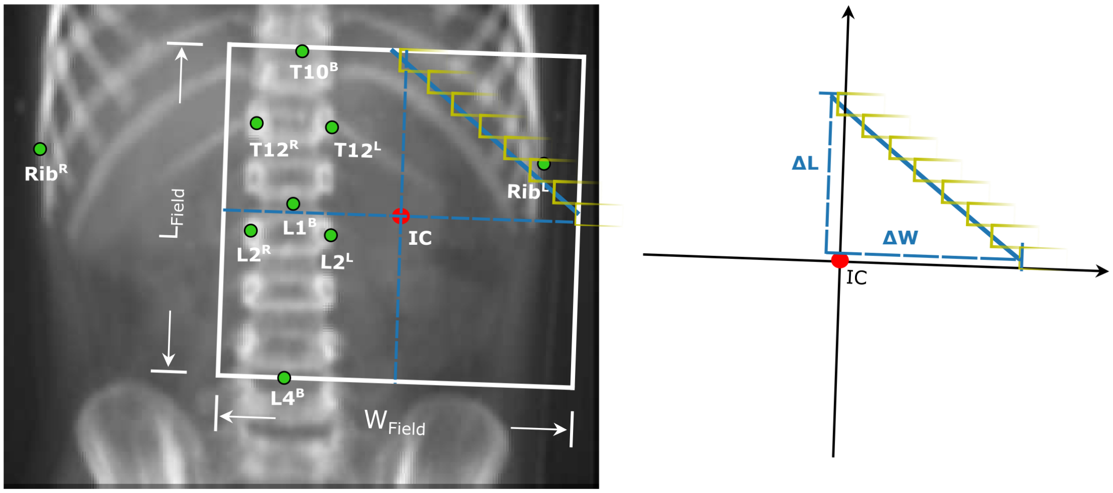

Figure 4. An example of the beam's eye view of a plan plotted on a DRR with the landmark locations used to compute the features concerning plan field configuration on top of the patient's anatomy. The plot next to the DRR illustrates how the block is simulated by aligning the center of the leaves with the boundary of the block and how the slope of an MLC-simulated block is calculated.

For both fields with and without a block, the features representing field sizes in RL and IS directions (W_{Field} and L_{Field}) were set by simply considering the full rectangular area (i.e., irrespective of blocking). For fields with a block, the slope of the block (note that the block is formed by Multi-Leaf Collimators (MLCs)) and the ratio between the blocked region and non-blocked region of the field ($Ratio_{Block}$) was computed. In addition, we considered features that relate to how the plan was configured with respect to the patient's anatomy, based on the position of the isocenter and of the bony landmarks. For instance, $\Delta^{IC}_{RL}(T10^B)$ links the bottom of the T10 vertebra to the position of the isocenter in RL direction. Figure 4 shows an example of the anatomical landmarks and plan geometrical borders used to calculate features describing plan configuration with respect to the patient's anatomy.

2.3.3. Dataset for supervised learning Features and dose-volume metrics were finally collected in a dataset. The dataset corresponded to a 2D matrix, where the rows represented patient-plan combinations, i.e., examples ($n = 42,600$), and the columns represented features (33) and response variables (4 for each OAR).

2.4. Machine learning

In the following sections we describe how ML was performed in terms of training and validation on the artificial plans. We further introduce the ML algorithm adopted, and describe an independent validation on clinical cases.

Table 1. Description of the 33 features considered as explanatory variables for ML.

Feature name	Origin	Unit	Description
Age	Records	years	patient age at CT scanning
$ArmsUp$	DRR	yes/no	whether the patient had arms in a raised position during scanning
$Diam^{IC}_{AP}$	Records	cm	patient AP diameter measured at isocenter
$Nephrectomy$	Records	yes/no	whether the patient underwent nephrectomy
W_{Rib^R}	DRR	cm	width (in RL) of right-part of the rib cage (from vertebral column to location of right-most rib)
W_{Rib^L}	DRR	cm	width (in RL) of left-part of the rib cage (from vertebral column to location of left-most rib)
W_{VC}	DRR	cm	average vertebral column width
L_{VC}	DRR	cm	length (in IS) of the vertebral column from T11 to L4
W_{Field}	Plan	cm	field width (in RL)
L_{Field}	Plan	cm	field length (in IS)
$FieldSide$	Plan	right/left	whether the plan concerns left-sided or right-sided flank irradiation
$Intercept_{Block}$	Plan	cm	distance (in RL) between isocenter and block endpoint of the top field border
$Ratio_{Block}$	Plan	%	$Area(Block)/Area(Rectangular\ field)$, 0 for block-free plans
$Slope_{Block}$	Plan	-	$\Delta L/\Delta W$ of the block (see Fig. 4); 0 for block-free plans
θ_C	Plan	°	angle of collimator system with respect to gantry system
$\Delta^{IC}_{RL}(T10^B)$	Plan + DRR	cm	RL distance between bottom of T10 and isocenter
$\Delta^{IC}_{IS}(T10^B)$	Plan + DRR	cm	IS distance between bottom of T10 and isocenter
$\Delta^{IC}_{RL}(T12^R)$	plan + DRR	cm	RL distance between right border of T12 and isocenter
$\Delta^{IC}_{IS}(T12^R)$	Plan + DRR	cm	IS distance between right border of T12 and isocenter
$\Delta^{IC}_{RL}(T12^L)$	Plan + DRR	cm	RL distance between left border of T12 and isocenter
$\Delta^{IC}_{IS}(T12^L)$	Plan + DRR	cm	IS distance between left border of T12 and isocenter
$\Delta^{IC}_{RL}(L1^B)$	Plan + DRR	cm	RL distance between bottom of L1 and isocenter
$\Delta^{IC}_{IS}(L1^B)$	Plan + DRR	cm	IS distance between bottom of L1 and isocenter
$\Delta^{IC}_{RL}(L2^R)$	Plan + DRR	cm	RL distance between right border of L2 and isocenter
$\Delta^{IC}_{IS}(L2^R)$	Plan + DRR	cm	IS distance between right border of L2 and isocenter
$\Delta^{IC}_{RL}(L2^L)$	Plan + DRR	cm	RL distance between left border of L2 and isocenter
$\Delta^{IC}_{IS}(L2^L)$	Plan + DRR	cm	IS distance between left border of L2 and isocenter
$\Delta^{IC}_{RL}(L4^B)$	Plan + DRR	cm	RL distance between bottom of L4 and isocenter
$\Delta^{IC}_{IS}(L4^B)$	Plan + DRR	cm	IS distance between bottom of L4 and isocenter
$\Delta^{IC}_{RL}(Rib^R)$	Plan + DRR	cm	RL distance between location of right-most rib and isocenter
$\Delta^{IC}_{IS}(Rib^R)$	Plan + DRR	cm	IS distance between location of right-most rib and isocenter
$\Delta^{IC}_{RL}(Rib^L)$	Plan + DRR	cm	RL distance between location of left-most rib and isocenter
$\Delta^{IC}_{IS}(Rib^L)$	Plan + DRR	cm	IS distance between location of left-most rib and isocenter

Abbreviations: R (in superscript): right, L (in superscript): left, RL: right-left, AP: anterior-posterior, IS: inferior-superior, IC: isocenter, VC: vertebral column, W: width, L in L_{VC} and L_{Field} : length.

2.4.1. Training and evaluation of ML models Since dose metrics are scalars, we treated the learning problem as a regression problem. We trained a separate ML model for each combination of dose-volume metric and OAR.

Preliminary analysis showed that right-sided plans and left-sided plans led to markedly different distributions of possible dose-volume metric values for all OARs except for the spinal cord. Thus, ML models were set to be composed of two sub-models, each to be trained independently on a particular sub-set of the data based on plan side (right or left).

The quality of the models was estimated with a 5-fold cross-validation. This means that a random partition of 1/5th of the total number of patients and plans was held out (test set), and training was performed on the remaining data. Then, the prediction error was measured on the test set. This process was repeated five times, each time considering a different data partition for the test set. No patient nor plan that was in the test set was included in the data at training time.

Each training step included hyper-parameter tuning by grid-search with internal 5-fold cross-validation (upon the training set), as well as feature selection (which resulted in eight features being systematically discarded, see the supplementary material A). For each dose-volume metric $k \in \{D_{\text{mean}}, D_{\text{2cc}}, V_{\text{5Gy}}, V_{\text{10Gy}}\}$, the Root Mean Square Error (RMSE) loss was used, i.e.,

$$RMSE(Y^k, \hat{Y}^k) = \sqrt{\frac{1}{\nu}\sum_{i=1}^{\nu}\left(Y_i^k - \hat{Y}_i^k\right)^2}, \tag{1}$$

where Y^k are the ground truth values and $\hat{Y}^k$ are the model predictions for the dose-volume metric k, and ν is the total number of rows in the training set. The RMSE was chosen to regularize ML, i.e., to penalize larger errors more (Bishop 2006).

To account for the stochastic nature of the ML algorithm employed (see Sec. 2.4.2) and for the random partitioning of the data, the 5-fold cross-validation was repeated ten times. The averages and standard deviations over the 5×10 validation results (five folds repeated ten times) were considered.

To put the results of ML into perspective, for each cross-validation, a baseline prediction was considered that simply used the average dose-volume metric observed in the training data. The Wilcoxon signed-rank test was used to assess whether the results obtained by ML and by this baseline are significantly different (p-value < 0.05).

2.4.2. Machine learning algorithm The recently introduced Genetic Programming version of the Gene-pool Optimal Mixing Evolutionary Algorithm (GP-GOMEA) was considered as ML algorithm, as it was found to achieve competitive performance on a variety of benchmark problems (Virgolin et al 2017, Virgolin et al 2020*a*), as well as in previous work concerning radiotherapy (Virgolin et al 2018*a*, Virgolin et al 2019).

In addition, GP-GOMEA can perform symbolic regression, i.e., it can generate a regression ML model in the form of a (symbolic) mathematical expression. GP-GOMEA incorporates Information Theory methods that enable it to synthesize fairly accurate expressions of particularly compact size, an aspect that makes these expressions lightweight and fast to execute, and that can aid human-interpretability. Details on the hyper-parameters (and their tuning) adopted in GP-GOMEA are described in the supplementary material B.

2.5. Independent evaluation on clinical plans

As aforementioned, the 300 plans used to cross-validate our approach were generated with an automatic sampling procedure (Sec. 2.2). To assess whether our results on artificial plans can be valid for clinically-used plans, we further evaluated our approach on an independent dataset for which clinical plans were crafted manually.

For this validation, we trained ML models (as a reminder, one model per OAR - dose-volume metric combination) on the dataset using the 142 CTs and the 300 artificial plans, and evaluated their prediction accuracy on a separate set of five CTs each associated with two clinical plans. We gathered five clinical plans (three right-sided, two left-sided) for these five CTs. Under the supervision of an experienced pediatric radiation oncologist (B.V. Balgobind), two adapted versions of each plan were manually created that both had the isocenter in the middle of the fields. In one plan no block was used and in the other plan a block was introduced to protect part of the liver or spleen, depending on the plan side. Training (including 5-fold cross-validation to determine the best hyper-parameter settings) and validation were repeated ten times to account for the stochastic nature of GP-GOMEA. Averages and standard deviations were computed over these ten repetitions.

3. Results

3.1. Dose-volume metric data distribution

Among the 300 artificial plans, plan side and OAR type was found to influence the distribution of a dose-volume metric considerably. To illustrate the effect of OAR type and plan side on the dose, Figure 5 shows the distributions found for D_{mean} and D_{2cc} for the liver and the spleen, in case of left- and right-sided plans. For D_{mean} for the liver, distributions approximately resembling the normal distribution were obtained (in case of right-sided plans with particular high variance and long left tail). The distribution in case of right-sided plans had a mean of 9.5 Gy (typically a major part of the liver was in-field), the distribution in case of left-sided plans had a mean of 3.4 Gy (typically a minor part of the liver was in-field). In terms of D_{2cc}, for the liver we observed values close to the prescribed dose (14.4 Gy) both in case of left- and right-sided plans. The distributions of D_{mean} and D_{2cc} for the spleen associated with the different plan sides had more marked differences than the ones for the liver. In case of right-sided plans, values close to 0 Gy were obtained for both metrics (typically the spleen was outside the field). For left-sided plans, large values of D_{mean} were found to be much more frequent than low values. The distribution of D_{2cc} exhibited a peak around the prescribed dose. For the contralateral kidney and for the spinal cord the distributions are similar for both plan sides, as the contralateral kidney should be outside the field and the spinal cord should be included within the field (according to protocol).

For all OARs, distributions obtained for V_{5Gy} and V_{10Gy} largely resembled the ones obtained for D_{mean}. In fact, Pearson correlation coefficients above 98% were found when comparing D_{mean} with V_{5Gy} and V_{10Gy} for almost all OARs. Smaller (yet still large) correlation coefficients were found between D_{mean} and V_{10Gy} for the left and right kidney, with values of 96% and 91% respectively.

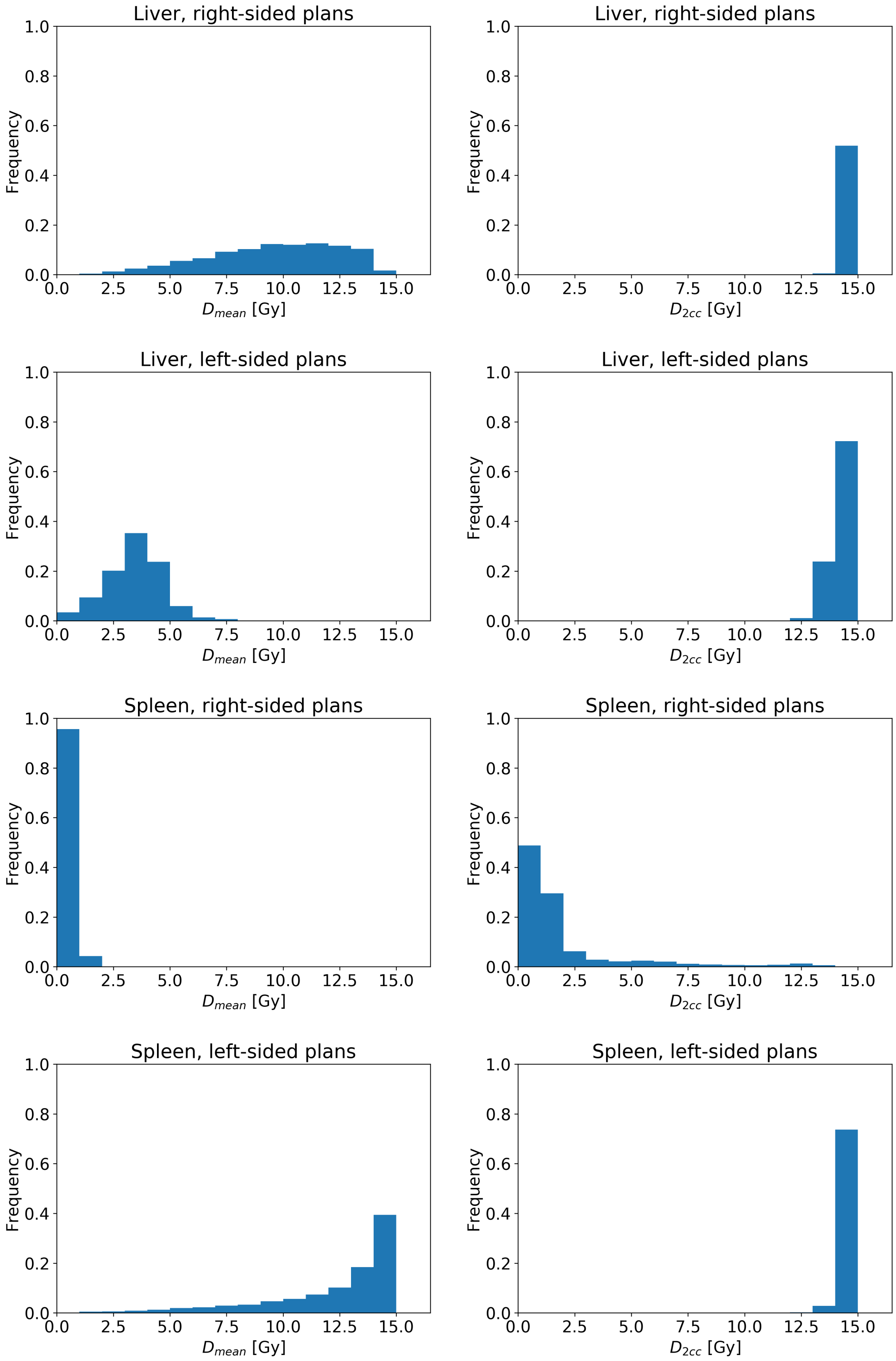

Figure 5. Distributions for the liver and the spleen of D_{mean} and D_{2cc} obtained by the automatic plan sampling procedure used to generate artificial plans and by applying the plans to the CT scans.

3.2. Validation on artificial plans

For each considered OAR, the Mean Absolute Errors (MAEs) (and standard deviation) at validation time for D_{mean}, D_{2cc}, V_{5Gy}, and V_{10Gy} from the ten repetitions of the 5-fold cross-validation procedure using the artificial plans are reported in Table 2. We further present the average decrease in MAE when using ML compared to when using the baseline (i.e., the effect size), as well as the outcome of statistical significance tests (Wilcoxon signed-rank) comparing ML to the baseline.

The errors for D_{mean} and D_{2cc} were generally below 2 Gy, which corresponds to approximately 14% of the prescribed dose of 14.4 Gy. For all OARs but for the spinal cord, the plan side had considerable impact on the magnitude of the errors. As the spinal cord in RL direction was in-field no matter the plan side, the MAEs of dose-volume metrics predictions were found to be small: < 1 Gy for D_{mean} and D_{2cc}, $< 4\%$ for V_{5Gy} and V_{10Gy}. For OARs that were almost out-of-field, e.g., the spleen in case of right-sided plans, small MAEs of D_{mean} were found (< 0.1 Gy), as very low values were obtained across all patient-plan combinations (see Fig. 5). Note that in this case (and also for the D_{2cc} for the liver in both left- and right-sided plans), ML performs significantly but not substantially better than the baseline.

For the liver in case of right-sided plans, and for the spleen in case of left-sided plans, larger MAEs were found (liver: 1.7 Gy for D_{mean}, 12.1% for V_{5Gy}, 12.6% for V_{10Gy}; spleen: 1.5 Gy for D_{mean}, 9.3% for V_{5Gy}, 10.7% for V_{10Gy}). These errors can be attributed to the particular configuration of the position of these OARs and the field of the plans.

Among the dose-volume metrics, D_{2cc} for the (partly) in-field OARs had low variability, with a D_{2cc} close to the prescribed dose (14.4 Gy). For example, small errors were obtained for the D_{2cc} for the liver (< 0.4 Gy), as we consistently obtained a large D_{2cc} value for both left- and right-sided plans (see Fig. 5). In contrast, D_{2cc} was harder to predict when the OAR was contralateral to the plan side. The MAEs obtained for D_{2cc} for the spleen in case of right-sided plans was 1.6 Gy. This was 2.9 Gy for the left kidney, and 1.4 Gy for the right kidney.

The largest average error was found for D_{2cc} for the left kidney, amounting to 20% of the prescribed dose. For all dose-volume metrics for the right kidney, and for D_{2cc} for the spleen, we found that ML predictions were slightly worse compared to using the baseline (note the negative effect sizes), but not significantly so. Lastly regarding the kidneys, although errors in D_{2cc} were relatively large, errors in V_{10Gy} were relatively small (compared with V_{10Gy} for the other OARs). In fact, only a small percentage of the contralateral kidney, from 0 to less than 3% typically received at least 10 Gy.

Although not reported in Table 2, we remark that the errors were found to be unbiased: no systematic over- nor under-estimations of dose-volume metrics were found on average, with the mean (non-absolute) error being close to zero for all metrics.

Table 2. Mean test MAE ± standard deviation and effect size (in small font) against the baseline (MAE of baseline - MAE of ML), of ten repetitions of 5-fold cross-validation for each OAR and dose-volume metric on the artificial plans. Bold results are significantly better than the baseline (the opposite is never found).

Side	OAR	$\boldsymbol{D}_{\mathbf{mean}}$ [Gy]	$\boldsymbol{D}_{\mathbf{2cc}}$ [Gy]	$\boldsymbol{V}_{\mathbf{5Gy}}$ [%]	$\boldsymbol{V}_{\mathbf{10Gy}}$ [%]
Right (22436 plans)	**Liver**	**1.7 ± 0.2** 0.7	**0.2 ± 0.0** 0.0	**12.1 ± 1.5** 5.1	**12.6 ± 1.8** 5.1
	Spleen	**0.1 ± 0.0** 0.0	1.6 ± 0.5 −0.1	0.9 ± 0.2 0.0	0.4 ± 0.1 0.0
	Left kidney	0.6 ± 0.5 0.1	**2.9 ± 0.2** 0.5	4.4 ± 0.5 0.1	2.5 ± 0.4 0.0
	Spinal cord	**0.4 ± 0.1** 0.7	0.2 ± 0.0 0.0	**3.2 ± 0.4** 5.6	**3.3 ± 0.4** 5.8
Left (20164 plans)	**Liver**	**0.8 ± 0.0** 0.2	**0.4 ± 0.1** 0.0	**5.8 ± 0.4** 1.1	**5.5 ± 0.3** 1.4
	Spleen	**1.5 ± 0.2** 0.7	0.3 ± 0.0 0.0	**9.3 ± 1.1** 5.2	**10.7 ± 1.4** 6.0
	Right kidney	0.6 ± 1.0 −0.3	1.4 ± 0.4 −0.1	2.3 ± 1.4 −0.4	0.8 ± 0.7 −0.1
	Spinal cord	**0.4 ± 0.0** 0.9	0.2 ± 0.0 0.0	**3.1 ± 0.3** 7.5	**2.9 ± 0.3** 7.4

3.3. Independent validation on clinical plans

Figure 6 and Figure 7 show, for each clinical case, the ground truth dose-volume metric values and the predictions obtained by the ML models (trained on the artificial plans). Results for D_{mean} and D_{2cc} are presented in Figure 6, and results for V_{5Gy} and V_{10Gy} are presented in Figure 7.

The errors in D_{mean} between predictions and ground truth values were generally low, totalling an average of 1.0 Gy (with a range of 0.0-4.9 Gy) across all OARs. Compared to the results obtained in the cross-validation using the artificial plans (Table 2), for the liver in case of right-sided plans, the average error on the clinical plans was found to be smaller (1.2 Gy vs. 1.7 Gy). Similar average errors in D_{mean} were found for the kidneys (0.8 Gy vs. 0.6 Gy for the left kidney and 0.5 Gy vs. 0.6 Gy for the right kidney), the liver in case of right-sided plans (1.0 Gy vs. 1.7 Gy), and the spinal cord (0.4 Gy vs. 0.4 Gy). Larger average errors in D_{mean} were found for the spleen (0.5 Gy vs. 0.1 Gy in case of the right-sided plans and 3.6 Gy vs. 1.5 Gy in case of left-sided plans). The largest error of 4.9 Gy was found for the spleen of case P_{L2B} (1.9 Gy error found for the spleen of case P_{L2}), indicating that the impact of the block on the plan was not well modeled. Furthermore, the error in spleen D_{mean} of P_{L1} and P_{L1B} was large (2.8 Gy and 4.7 Gy, respectively), indicating both field types (without and with a block, respectively) were not well modeled for this case (see the discussion in Sec. 4).

Regarding D_{2cc}, similar results to the ones obtained on artificial plans were found for the spinal cord, where an average error of 0.1 Gy was obtained in D_{2cc} (with a range of 0.0-0.4 Gy). For the spleen of cases P_{R1} and P_{R1B}, large errors in D_{2cc} were found (12.3 Gy on average), as ML predictions essentially wrongly represented the spleen to be out-of-field. Whereas this was the case for P_{R2} and P_{R2B}, and for P_{R3} and P_{R3B}, where small errors were obtained (1.1 Gy on average). Similarly, large errors in D_{2cc} were found in some cases for the contralateral kidneys (e.g., the left kidney: 5.1 Gy on average, with a range of 0.5-7.6 Gy).

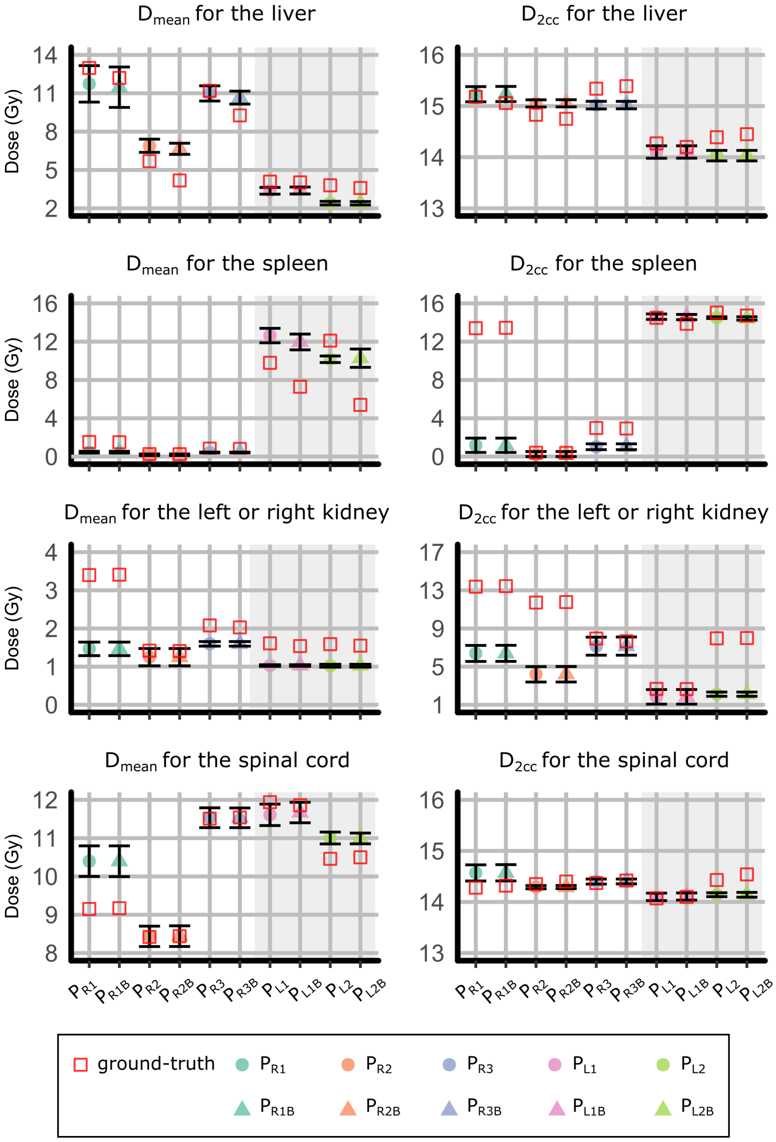

Figure 6. Mean and standard deviation of predictions across ten repetitions for the dose-volume metrics D_{mean} and D_{2cc} of the ten clinical plans. A plan-patient combination is encoded by color, and presence or absence of blocking is encoded by marker shape. Note that different plots have different scales of the vertical axis. For each case, the ground-truth dose-volume metric is indicated by a red square. In each plot, the first six cases (white background, plan subscripts starting with 'R') are right-sided plans, the last four cases (gray background, plan subscripts starting with 'L') are left-sided plans.

Results for V_{5Gy} (average error 8% with a range of 0-35%) and V_{10Gy} (average error 7% with a range of 0-49%) mostly followed the trend of the errors for D_{mean}, as these metrics were found to be correlated for most OARs.

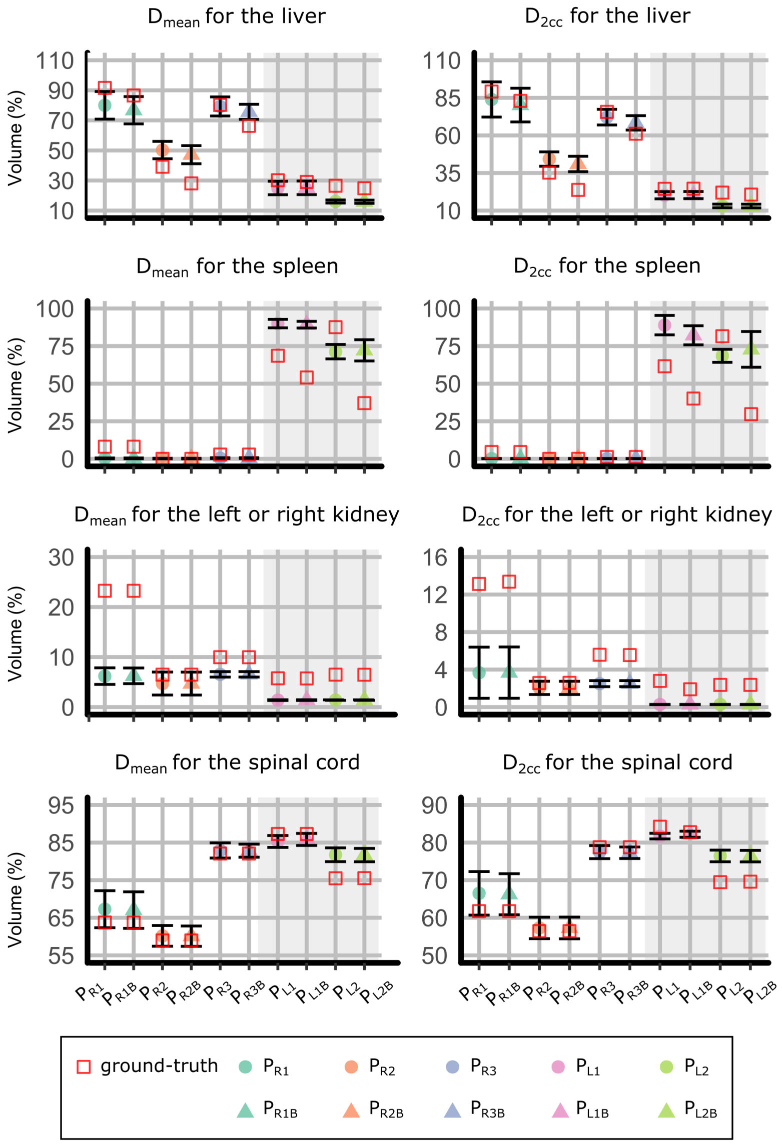

Figure 7. Mean and standard deviation of predictions across ten repetitions for the dose-volume metrics $V_{5\text{Gy}}$ and $V_{10\text{Gy}}$ of the ten clinical plans. A plan-patient combination is encoded by color, and presence or absence of blocking is encoded by marker shape. Note that different plots have different scales of the vertical axis. For each case, the ground-truth dose-volume metric is indicated by a red square. In each plot, the first six cases (white background, plan subscripts starting with 'R') are right-sided plans, the last four cases (gray background, plan subscripts starting with 'L') are left-sided plans.

4. Discussion

In this article we presented a new and different paradigm in organ dose reconstruction. By leveraging the modeling power of ML, we showed how patient and plan features can be used to predict organ dose-volume metrics directly, without the need of adopting a surrogate anatomy. Once the ML models are trained, they can readily be used to compute dose-volume metric predictions for a new historical patient and plan, by using their features as input.

Key to obtaining a decent amount of data to perform ML were the collaboration of five international institutes to gather pediatric patient CTs (147), the development of a new automatic sampling procedure yielding artificial Wilms' tumor RT plans, and the creation of an automatic dose reconstruction pipeline to calculate the dose for all patient and plan combinations. We validated our approach on 300 automatically generated artificial plans, and on ten manually created clinical plans, to assess whether the results of the validation on the artificial plans generalize well in practice. Our approach showed promising levels of accuracy in dose reconstruction in both settings.

Errors were found to be overall similar between the validation on the ten clinical plans and the validation on the artificial plans. However, for some metrics, errors were larger for the ten clinical plans. This may be due to chance, because ten is a small number to validate upon. Another possibility is that the artificial plan generation method needs to be improved. Artificial plans were generated by sampling geometry properties *uniformly* within predefined boundaries on two reference DRRs. Uniform sampling might not be representative of the distribution clinical plans have. Moreover, we consulted a single radiation oncologist to define clinically acceptable boundaries to use in the sampling of artificial plans. Consulting multiple experts and allowing for a larger variation might better help covering the extent of variation that is present in historical plans (Sec. 2.2). For example, the isocenter locations of artificial left-sided plans were never sampled below the 1st lumbar vertebra (see Fig. 2) and approximately half of the D_{mean} values for the spleen in case of the artificial left-sided plans were close to the prescribed dose (14.4 Gy, see Fig. 5), which means that the spleen was often almost completely in-field in our artificially generated set of left-sided plans. When a block was applied, only a small part of the spleen was spared. However, in clinical practice, isocenter locations can be lower, and a larger part of the spleen might actually be outside the field (see Fig. 8). This might explain the relatively large errors observed in Figure 6 for P_{L1B} and P_{L2B} where the isocenter location is lower than the sampled range. Ultimately, effort should be done to improve the sampling of artificial plans.

In the validation performed upon artificial plans as well as in the one performed upon clinical plans, a main result that emerges is that dose-volume metrics for an organ are hard to predict when, due to the field setup, it is unclear whether the OAR is (partially) included in the field or not. For example, consider the $D_{2\text{cc}}$ as opposed to the D_{mean} for the spleen for P_{R1} and P_{R1B} in Figure 6: a tiny part of the spleen being inside the field causes a large $D_{2\text{cc}}$ (wrongly predicted to be small), and small D_{mean}

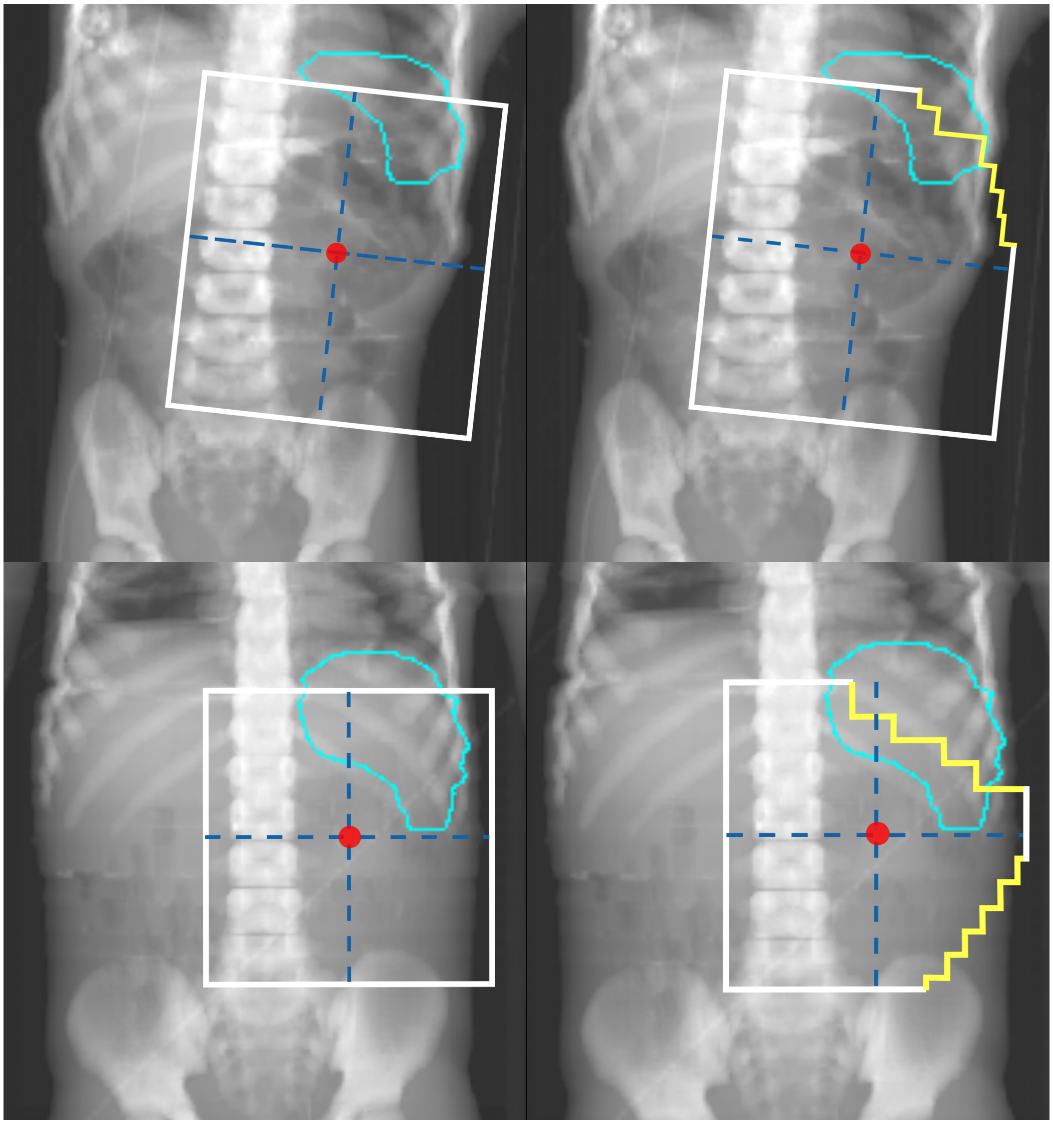

Figure 8. Effective field shape of plans P_{L1} (on the upper left), P_{L1B} (on the upper right), P_{L2} (on the lower left) and P_{L2B} (on the lower right) plotted on top of the associated DRR. The fields are placed lower than most of the sampled artificial plans (see Fig. 2) and consequently a (large) part of the spleen (indicated by the light blue contours) is outside the field (for P_{L1B} and P_{L2B} an even larger part of the spleen was blocked).

(correctly predicted to be small). As experimentally observed in prior work (Virgolin et al 2018*b*, Virgolin et al 2019), 2D bony anatomy provides only coarse information on OAR shape and position even for ML algorithms (e.g., an MAE of 6.4 mm for the prediction of the liver position along the IS axis was reported (Virgolin et al 2019)). Yet, because bony anatomy is the only structure that is reliably visible in historical radiographs, most of the anatomical features rely on it. Patients with similar anatomical features derived from bony anatomy may have different OAR shape and position, and thus different dose-volume metrics. Furthermore, impreciseness in feature values due to e.g., uncertainties in landmark detection and plan emulation, aggravate the situation.

Compared to conventional dose reconstruction methods (that use surrogate anatomies and heuristics to decide what surrogate to use), we considered a relatively large number of features: 33. Phantom-based methods consider, e.g., only age and gender (Stovall et al 2006, Howell et al 2019), or gender together with height and weight percentiles (Geyer et al 2014). However, if a 2D radiograph is available, the added value of this information should be exploited. In our work, the majority of the features we considered, i.e., 23 out of 33 (minus eight due to automatic feature selection, see supplementary A), regarded patient anatomy as visible on a 2D radiograph, which we simulated with DRRs. Our DRRs were generated in a conformal fashion, e.g., the abdomen was always fully included in RL direction. The automatic landmark detection that was used to generate features expects this conformity to achieve precise detection (Wang et al 2020). When dealing with actual historical radiographs, however, several challenges need to be taken into account. For example, our automatic landmark detection method requires further development to account for noise in the radiograph (e.g., the presence of hand-writing on the radiograph). Moreover, educated guesses of landmark locations may be needed in some cases, as some historical radiographs do not include the entire abdomen (see Fig. 1(c)). Nevertheless, as long as the features are somehow collected (e.g., manually), they can be used as input for the ML models to get respective dose-volume metric predictions.

There are disadvantages of our approach compared to conventional dose reconstruction methods that use surrogate anatomies beyond the need for patient radiographs, which are not always available in retrospective data. In particular, a key limitation is that ML models do not predict the entire 3D dose distribution an organ receives, but only the metrics they were trained for. Potentially useful information to link to AEs may be contained in 3D dose distributions. To predict 3D dose distributions, the ML models would need to be trained to predict a 3D output. Surrogate-based methods do allow to obtain the entire 3D dose distribution to an organ, since the distribution can be visualized on the organ of the surrogate anatomy, after plan simulation. However, considering the magnitude and variations of the errors of organ mean dose obtained by conventional approaches (Wang et al 2018), it is questionable whether the full 3D distribution will be sufficiently reliable. Our approach, as currently proposed, can straightforwardly be extended to predict any (scalar) dose-volume metric that is suspected to be useful to study AEs (it suffices to train ML on that metric).

Another limitation of our approach is that it does not take into consideration uncertainties related to OAR motion. For validation, we aimed at reconstructing the dose based on the particular snapshot of anatomy at the moment the CT (ground-truth) / respective DRR (to simulate historical radiographs) was taken. Yet, OAR motion plays a key role in the uncertainty of organ positioning at the edge of the field, which can lead to a discrepancy between the planned dose and the actual delivered dose. In RT practice, radiation delivery is performed over a number of days, with fractionation schemes. The OAR position can therefore vary (i.e., inter-fractional position variation). Intra-fractional organ motion due to, e.g., respiration variation, contributes to the difference between planned dose and delivered dose as well (Huijskens et al 2015).

Lastly, a main limitation of our approach is that the ML models we generated are specific to pediatric patients (1 to 8 years of age) and Wilms' tumor RT plans: they can only predict reliable dose-volume metrics of specific OARs they were explicitly trained for. The RT plans we have sampled were also restricted to a standard AP-PA setup without considering wedges, boost fields, or other radiation sources such as Cobalt-60. Moreover, the predictions of the ML models (as well as the validation performed in this study) are based on the dose calculation algorithm we adopted when preparing training data, which has inaccuracies. Specifically, we used a collapsed cone algorithm available in Oncentra TPS. Though good accuracy was reported in the in-field and near-field region ($<$ 5 cm from the field borders, achieves an error of 1-2% of the prescribed dose), in low dose regions (10-15 cm from the field border) an underestimation of 10% of the dose in the region was reported (Krieger and Sauer 2005). We remark that the OARs we considered in this study were mostly within 5 cm near the field border (except for the spleen in case of the right-sided plans). To make the method more general for OARs far from field borders, more advanced Monte Carlo dose calculation algorithms should be applied in future implementations. However, we believe that the core ideas of our work can be replicated for other cohorts and other types of plans. Essentially, as long as a sufficient number of anatomies and plans are collected or generated, and a large number of dose reconstructions are performed to be used as examples, new ML models can be trained to predict how the dose-volume metrics are linked to anatomy-plan configurations. As was the case in our study, the collection and preparation of sufficient data for ML is likely to be the largest required effort.

Our proposed approach presents several advantages compared to traditional dose reconstruction methods. First of all, we found our validation results to compare favorably with respect to our recent work considering dose reconstruction for a similar childhood cancer cohort (Wang et al 2018). The work considered 31 patients aged 2 to 6, 12 Wilms' tumor clinical plans, and a total of 50 dose reconstruction combinations, which were performed by matching a surrogate CT based on age and gender. The work reported an MAE for the D_{mean} for the liver of 1.6 Gy (average across both left- and right-sided plans), and an MAE for the D_{mean} for the spleen of 2.6 Gy. For the liver, we obtained an MAE of 1.3 Gy when validating on artificial plans, and of 1.1 Gy when validating on ten clinical plans. For the spleen, we obtained an MAE of 0.8 Gy on artificial plans, and of 1.7 Gy for the clinical plans. Furthermore, our ML-based predictions resulted in much smaller variations. The inter-quantile range (25th to 75th percentile) of the (non-absolute) prediction error of our previous work was 3.6 Gy for the liver, and 4.7 Gy for the spleen (Wang et al 2018). On the artificial plans, we obtained a range of 2.0 Gy for the liver, and of 1.2 Gy for the spleen. On the clinical plans, the range for the liver was 1.9 Gy, and the one for the spleen was 2.2 Gy. We remark that since the dose reconstruction accuracy is largely influenced by the particular plans considered, these values may not be a fair comparison. We are currently working on a multi-institute study to compare our approach with two state-of-the-art, phantom-based dose reconstruction approaches (Lee et al 2015, Howell et al 2019). In that study, a same set of patients and plans will be used for validation.

Finally, a benefit of having ML models is that, once features are collected, they can be used as inputs for the model to obtain the prediction of a dose-volume metric immediately. Running a model on a computer simply means to follow the steps encoded by the formula the model represents, which takes a few milliseconds. Conversely, in a surrogate-based approach, the features are used to craft or select a surrogate anatomy. Then, effort and time must be put to emulate the plan on the surrogate anatomy, calculate the dose, and obtain the dose-volume metrics (Lee et al 2015, Howell et al 2019, Wang et al 2020).

5. Conclusion

We presented the first surrogate-free organ dose reconstruction method based on ML. Our method was enabled by the collection of large amounts of patient and CT data, and the automatic generation of artificial plans and of dose distribution data. We assembled a dataset of dose-volume metrics corresponding to features of patient anatomy and plan geometry, and subsequently trained ML models to predict how features of patient anatomy and of treatment plans influence dose-volume metrics. The predictions were validated upon both artificial and clinical RT plans, and achieved good accuracy in both cases.

Acknowledgements

The authors acknowledge Stichting Kinderen Kankervrij (KiKa) for financial support (project #187), and the Maurits and Anna de Kock foundation for financing a high performance computing system. Elekta is acknowledged for providing the research software ADMIRE for automatic segmentation. The authors thank Abigail Bryce-Atkinson for her help in data preparation and feature extraction, and Dr. Cécile M. Ronckers for sharing her expertise.

References

Bezin J V, Allodji R S, Mège J P, Beldjoudi G, Saunier F, Chavaudra J, Deutsch E, de Vathaire F, Bernier V, Carrie C et al 2017 A review of uncertainties in radiotherapy dose reconstruction and their impacts on dose–response relationships *J. Radiol. Prot.* **37**(1), R1.

Birgisson H, Påhlman L, Gunnarsson U and Glimelius B 2005 Adverse effects of preoperative radiation therapy for rectal cancer: long-term follow-up of the Swedish Rectal Cancer Trial *J. Clin. Oncol.* **23**(34), 8697–8705.

Bishop C M 2006 *Pattern recognition and machine learning* Springer.

Bölling T, Ernst I, Pape H, Martini C, Rübe C, Timmermann B, Fischedick K, Kortmann R D and Willich N 2011 Dose–volume analysis of radiation nephropathy in children: Preliminary report of the risk consortium *Int. J. Radiat. Oncol. Biol. Phys.* **80**(3), 840–844.

Cassola V F, Milian F M, Kramer R, de Oliveira Lira C A B and Khoury H J 2011 Standing adult human phantoms based on 10th, 50th and 90th mass and height percentiles of male and female Caucasian populations *Phys. Med. Biol.* **56**(13), 3749–3772.

Cheung Y T, Brinkman T M, Li C, Mzayek Y, Srivastava D, Ness K K, Patel S K, Howell R M, Oeffinger K C, Robison L L et al 2017 Chronic health conditions and neurocognitive function in aging survivors of childhood cancer: A report from the childhood cancer survivor study *J. Nat. Cancer. Inst.* **110**(4), 411–419.

Constine L, Ronckers C M, Hua C H, Olch A, Kremer L C M, Jackson A and Bentzen S M 2019 Pediatric Normal Tissue Effects in the Clinic (PENTEC): An international collaboration to analyse normal tissue radiation dose–volume response relationships for paediatric cancer patients *Clin. Oncol.* **31**(3), 199–207.

de la Grandmaison G L, Clairand I and Durigon M 2001 Organ weight in 684 adult autopsies: new tables for a Caucasoid population *Forensic Sci. Int.* **119**(2), 149–154.

Donovan E, Bleakley N, Denholm E, Evans P, Gothard L, Hanson J, Peckitt C, Reise S, Ross G, Sharp G et al 2007 Randomised trial of standard 2D radiotherapy (RT) versus intensity modulated radiotherapy (IMRT) in patients prescribed breast radiotherapy *Radiother. Oncol.* **82**(3), 254–264.

Emami B, Lyman J, Brown A, Cola L, Goitein M, Munzenrider J E, Shank B, Solin L J and Wesson M 1991 Tolerance of normal tissue to therapeutic irradiation *Int. J. Radiat. Oncol. Biol. Phys.* **21**(1), 109–122.

Feng F Y, Kim H M, Lyden T H, Haxer M J, Feng M, Worden F P, Chepeha D B and Eisbruch A 2007 Intensity-modulated radiotherapy of head and neck cancer aiming to reduce dysphagia: early dose–effect relationships for the swallowing structures *Int. J. Radiat. Oncol. Biol. Phys.* **68**(5), 1289–1298.

Geyer A M, O'Reilly S, Lee C, Long D J and Bolch W E 2014 The UF/NCI family of hybrid computational phantoms representing the current US population of male and female children, adolescents, and adults—application to CT dosimetry *Phys. Med. Biol.* **59**(18), 5225–5242.

Howell R M, Smith S A, Weathers R E, Kry S F and Stovall M 2019 Adaptations to a generalized radiation dose reconstruction methodology for use in epidemiologic studies: An update from the MD Anderson late effect group *Radiat. Res.* **192**(2), 169–188.

Huijskens S C, van Dijk I W E M, de Jong R, Visser J, Fajardo R D, Ronckers C M, Janssens G O, Maduro J H, Rasch C R, Alderliesten T et al 2015 Quantification of renal and diaphragmatic interfractional motion in pediatric image-guided radiation therapy: a multicenter study *Radiother. Oncol.* **117**(3), 425–431.

Krieger T and Sauer O A 2005 Monte carlo-versus pencil-beam-/collapsed-cone-dose calculation in a heterogeneous multi-layer phantom *Phys. Med. Biol.* **50**(5), 859.

Lee C, Jung J W, Pelletier C, Pyakuryal A, Lamart S, Kim J O and Lee C 2015 Reconstruction of organ dose for external radiotherapy patients in retrospective epidemiologic studies *Phys. Med. Biol.* **60**(6), 2309–2324.

Leisenring W M, Mertens A C, Armstrong G T, Stovall M A, Neglia J P, Lanctot J Q, Boice Jr J D, Whitton J A and Yasui Y 2009 Pediatric cancer survivorship research: experience of the childhood cancer survivor study *J. Clin. Oncol.* **27**(14), 2319.

Mishra P, Li R, James S S, Mak R H, Williams C L, Yue Y, Berbeco R I and Lewis J H 2013 Evaluation of 3D fluoroscopic image generation from a single planar treatment image on patient data with a modified XCAT phantom *Phys. Med. Biol.* **58**(4), 841–858.

Ng A, Brock K K, Sharpe M B, Moseley J L, Craig T and Hodgson D C 2012 Individualized 3D reconstruction of normal tissue dose for patients with long-term follow-up: a step toward understanding dose risk for late toxicity *Int. J. Radiat. Oncol. Biol. Phys.* **84**(4), e557–e563.

Segars W P, Bond J, Frush J, Hon S, Eckersley C, Williams C H, Feng J, Tward D J, Ratnanather J T, Miller M I et al 2013 Population of anatomically variable 4D XCAT adult phantoms for imaging research and optimization *Med. Phys.* **40**(4), 043701.

Stovall M, Weathers R, Kasper C, Smith S A, Travis L, Ron E and Kleinerman R 2006 Dose reconstruction for therapeutic and diagnostic radiation exposures: use in epidemiological studies *Radiat. Res.* **166**(1), 141–157.

Valentin J 2002 Basic anatomical and physiological data for use in radiological protection: reference values: ICRP publication 89 *Annals of the ICRP* **32**(3-4), 1–277.

van den Heuvel-Eibrink M M, Hol J A, Pritchard-Jones K, van Tinteren H, Furtwängler R, Verschuur A C, Vujanic G M, Leuschner I, Brok J, Rübe C et al 2017 Position paper: rationale for the treatment of Wilms tumour in the UMBRELLA SIOP–RTSG 2016 protocol *Nat. Rev. Urol.* **14**(12), 743–752.

van Dijk I W E M, Oldenburger F, Cardous-Ubbink M C, Geenen M M, Heinen R C, de Kraker J, van Leeuwen F E, van der Pal H J, Caron H N, Koning C C et al 2010 Evaluation of late adverse events in long-term Wilms' tumor survivors *Int. J. Radiat. Oncol. Biol. Phys.* **78**(2), 370–378.

Verellen D, De Ridder M and Storme G 2008 A (short) history of image-guided radiotherapy *Radiother. Oncol.* **86**(1), 4–13.

Virgolin M, Alderliesten T, Bel A, Witteveen C and Bosman P A N 2018*a* Symbolic regression and feature construction with GP-GOMEA applied to radiotherapy dose reconstruction of childhood cancer survivors *in* 'Proc. Genetic and Evolutionary Computation Conference' GECCO '18 ACM New York, NY, USA pp. 1395–1402.

Virgolin M, Alderliesten T, Witteveen C and Bosman P A N 2017 Scalable genetic programming by gene-pool optimal mixing and input-space entropy-based building-block learning *in* 'Proc. Genetic and Evolutionary Computation Conference' GECCO '17 ACM New York, NY, USA pp. 1041–1048.

Virgolin M, Alderliesten T, Witteveen C and Bosman P A N 2020*a* Improving model-based genetic programming for symbolic regression of small expressions. Preprint arXiv:1904.02050. (Accepted for publication with minor revisions in *Evol. Comput.*).

Virgolin M, van Dijk I W E M, Wiersma J, Ronckers C M, Witteveen C, Bel A, Alderliesten T and Bosman P A N 2018*b* On the feasibility of automatically selecting similar patients in highly individualized radiotherapy dose reconstruction for historic data of pediatric cancer survivors *Med. Phys.* **45**(4), 1504–1517.

Virgolin M, Wang Z, Alderliesten T and Bosman P A N 2019 Machine learning for automatic construction of pseudo-realistic pediatric abdominal phantoms. Preprint arXiv:1909.03723.

Virgolin M, Wang Z, Alderliesten T and Bosman P A N 2020*b* Machine learning for automatic construction of pediatric abdominal phantoms for radiation dose reconstruction. (To appear in *Proc. SPIE*).

Wang Z, Balgobind B, Virgolin M, van Dijk I W E M, Wiersma J, Ronckers C M, Bosman P A N, Bel A and Alderliesten T 2019 How do patient characteristics and anatomical features correlate to accuracy of organ dose reconstruction for Wilms' tumor radiation treatment plans when using a surrogate patient's CT scan? *J. Radiol. Prot.* **39**(2), 598–619.

Wang Z, van Dijk I W E M, Wiersma J, Ronckers C M, Oldenburger F, Balgobind B V, Bosman P A N, Bel A and Alderliesten T 2018 Are age and gender suitable matching criteria in organ dose reconstruction using surrogate childhood cancer patients' CT scans? *Med. Phys.* **45**(6), 2628–2638.

Wang Z, Virgolin M, Bosman P A N, Crama K, Balgobind B V, Bel A and Alderliesten T 2020 Automatic generation of three-dimensional dose reconstruction data for two-dimensional radiotherapy plans for historically treated patients *J. Med. Imaging* **7**(1), 015001.

Xu X G 2014 An exponential growth of computational phantom research in radiation protection, imaging, and radiotherapy: a review of the fifty-year history *Phys. Med. Biol.* **59**(18), R233–R302.

Supplementary Material: Surrogate-free machine learning-based organ dose reconstruction for pediatric abdominal radiotherapy

M Virgolin, Z Wang, B V Balgobind, I W E M van Dijk, J Wiersma, P S Kroon, G O Janssens, M van Herk, D C Hodgson, L Zadravec Zaletel, C R N Rasch, A Bel, P A N Bosman, T Alderliesten

(Shared first authorship between M Virgolin and Z Wang)

A. Feature selection

An automatic feature selection step was performed before training the ML models, as follows.

(i) Compute the absolute Pearson correlation coefficient $|\rho|$ between all pairs of features based on the values of the training set.

(ii) If $|\rho| > 0.95$ (highly positive or negative correlation) between two features, discard the second feature.

(iii) Repeat (ii) until no two features that have $|\rho| > 0.95$ remain.

Note that, at (ii), one could discard a feature at random. We systematically discarded the second feature to obtain a deterministic outcome.

As explained in Section 2.4.1 of the manuscript, a model for the prediction of the dose-volume metric of an organ was composed of two sub-models, trained independently based on the side of the plan. We found that partitioning the dataset by plan side does not influence feature selection significantly, i.e., the pairs of features that have $|\rho| > 0.95$ remain the same. Figure 1 shows the value $|\rho|$ between features across the entire dataset (averaged between left- and right-sided plans). The following eight features were systematically discarded in our experiments:

$\Delta^{IC}_{RL}(T12^R)$ – highly correlated with $\Delta^{IC}_{RL}(T10^B)$ (and others);
$\Delta^{IC}_{IS}(T12^R)$ – highly correlated with $\Delta^{IC}_{IS}(T10^B)$ (and others);
$\Delta^{IC}_{RL}(L2^R)$ – highly correlated with $\Delta^{IC}_{RL}(L1^B)$ (and others);
$\Delta^{IC}_{IS}(L2^R)$ – highly correlated with $\Delta^{IC}_{IS}(L1^B)$ (and others);
$\Delta^{IC}_{RL}(L2^L)$ – highly correlated with $\Delta^{IC}_{RL}(L1^B)$ (and others);
$\Delta^{IC}_{IS}(L2^L)$ – highly correlated with $\Delta^{IC}_{IS}(L1^B)$ (and others);
$\Delta^{IC}_{RL}(Rib^L)$ – highly correlated with $\Delta^{IC}_{RL}(L1^B)$ (and others);
$\Delta^{IC}_{IS}(Rib^L)$ – highly correlated with $\Delta^{IC}_{IS}(L1^B)$ (and others).

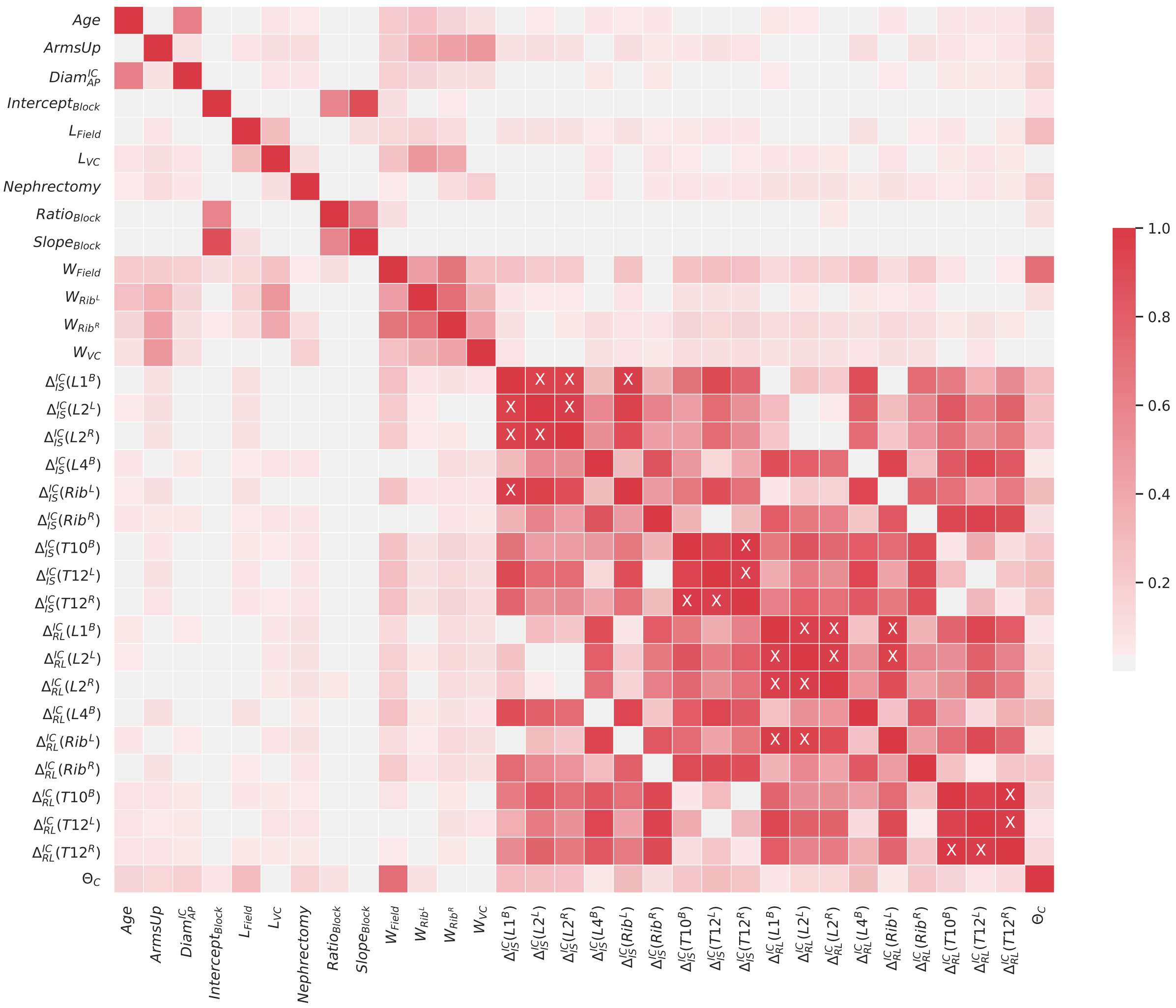

Figure 1. Absolute Pearson correlation coefficient ($|\rho|$) between features (average between left- and right-sided plans). Cells marked by a cross have $|\rho| > 0.95$ (excluding cells in the diagonal).

B. Hyper-parameters of GP-GOMEA and their tuning

The hyper-parameters of GP-GOMEA are reported in Table 1. A short description follows.

Tree height. GP-GOMEA encodes machine learning models with symbolic trees (Virgolin et al 2017). In our case, trees were binary. The maximal tree height of a tree limits how complex the encoded machine learning model can be. For instance, a tree with height 2 can contain up to 7 nodes, a tree with height 4 can contain up to 31 nodes. Larger trees can encode relatively complex models (but risk overfitting), smaller trees can encode relatively simple formulas (but risk underfitting).

Evaluations limit. The evaluations limit is used to terminate GP-GOMEA. An evaluation is the computation of the loss function of a model. Terminating early/late can be useful to prevent overfitting/underfitting.

Hyper-parameter	Setting(s)
Tree height*	$\{2,3,4\}$
Evaluations limit*	$\{10^4, 10^5, 10^6\}$
Function set*	$\{[+,-,\times,\div_p], [+,-,\times,\div_p,\cdot^2,\sqrt{\lvert\cdot\rvert},\sin,\cos,\exp]\}$
Interleaving generations g	6
Starting population size	500

Table 1. Hyper-parameters of GP-GOMEA and their settings. Starred hyper-parameters are subject to grid-search tuning, using the settings reported within curly braces.

Function set. The atomic components that can be instantiated as nodes can be features (of Table 1 of the manuscript), random constants, or functions from a pre-defined function set. Random constants are sampled uniformly at random between $\{-\rho, +\rho\}$, where $\rho := 5 \times \max(\mathbf{x})$, and $\mathbf{x}$ is the 2D matrix containing all numerical feature values for all patients and plans (in the training set). We considered a simpler function set of algebraic functions ($[+,-,\times,\div_p]$), and a more complex one including trigonometric and transcendental functions ($[+,-,\times,\div_p,\cdot^2,\sqrt{\lvert\cdot\rvert},\sin,\cos,\exp]\}$). More complex functions can help fit the data better, at the risk of overfitting.

Interleaving generations & starting population size. Whereas typical GP algorithms perform a single evolutionary run, GP-GOMEA uses a scheme of multiple runs of increasing evolutionary budget. Larger-budget runs are started later in the scheme, and executed in an interleaved fashion such that smaller-budget runs perform more iterations (called *generations*) than larger-budget runs. The first run r_1 uses the starting population size and executes g generations before the next run r_2 is started. The run r_2 has a population size that is double of that of r_1 (1000), and performs 1 generation every g generations of r_1. The first time r_2 has executed g generations, r_3 is started, with double the population of r_2 (2000). Subsequently, r_3 will execute 1 generation every time r_2 has executed g generations. If a run converges to all identical machine learning models, then it is terminated. A run is also terminated if a larger-budget run exists that has found a better machine learning model.

As shown in Table 1, we used a starting population size of 500 and a number of generations for interleaving $g = 6$. These choices are based on what is reported in our previous work (Virgolin et al 2020), except for the starting population size (set to 50 in the article), as the article shows that runs with population sizes below a thousand are typically terminated anyway within the first couple of minutes. Hence, computations performed by those runs are essentially wasted. We set $g = 6$ because it is the intermediate value (between 4 and 8) considered in our previous work (Virgolin et al 2020), and because GP-GOMEA is fairly robust to the choice of g, i.e., the results for $g = 4$ are similar to the ones of $g = 6$ and $g = 8$.

References

Virgolin M, Alderliesten T, Witteveen C and Bosman P A N 2017 Scalable genetic programming by gene-pool optimal mixing and input-space entropy-based building-block learning *in* 'Proc. Genetic and Evolutionary Computation Conference' GECCO '17 ACM New York, NY, USA pp. 1041–1048.

Virgolin M, Alderliesten T, Witteveen C and Bosman P A N 2020 Improving model-based genetic programming for symbolic regression of small expressions. Preprint arXiv:1904.02050. (Accepted for publication with minor revisions in *Evol. Comput.*).