Machine learning for the prediction of pseudorealistic pediatric abdominal phantoms for radiation dose reconstruction

Virgolin, Marco; Wang, Ziyuan; Alderliesten, Tanja; Bosman, Peter A.N.

DOI
10.1117/1.JMI.7.4.046501

Publication date
2020

Document Version
Final published version

Published in
Journal of Medical Imaging

Machine learning for the prediction of pseudorealistic pediatric abdominal phantoms for radiation dose reconstruction

Marco Virgolin,[a,*] Ziyuan Wang,[b] Tanja Alderliesten,[b,c] and Peter A. N. Bosman[a,d]

[a]Centrum Wiskunde and Informatica, Life Sciences and Health Group, Amsterdam, The Netherlands

[b]Amsterdam UMC, University of Amsterdam, Department of Radiation Oncology, Amsterdam, The Netherlands

[c]Leiden University Medical Center, Department of Radiation Oncology, Leiden, The Netherlands

[d]Delft University of Technology, Algorithmics Group, Delft, The Netherlands

Abstract

Purpose: Current phantoms used for the dose reconstruction of long-term childhood cancer survivors lack individualization. We design a method to predict highly individualized abdominal three-dimensional (3-D) phantoms automatically.

Approach: We train machine learning (ML) models to map (2-D) patient features to 3-D organ-at-risk (OAR) metrics upon a database of 60 pediatric abdominal computed tomographies with liver and spleen segmentations. Next, we use the models in an automatic pipeline that outputs a personalized phantom given the patient's features, by assembling 3-D imaging from the database. A step to improve phantom realism (i.e., avoid OAR overlap) is included. We compare five ML algorithms, in terms of predicting OAR left–right (LR), anterior–posterior (AP), inferior–superior (IS) positions, and surface Dice–Sørensen coefficient (sDSC). Furthermore, two existing human-designed phantom construction criteria and two additional control methods are investigated for comparison.

Results: Different ML algorithms result in similar test mean absolute errors: ∼8 mm for liver LR, IS, and spleen AP, IS; ∼5 mm for liver AP and spleen LR; ∼80% for abdomen sDSC; and ∼60% to 65% for liver and spleen sDSC. One ML algorithm (GP-GOMEA) significantly performs the best for 6/9 metrics. The control methods and the human-designed criteria in particular perform generally worse, sometimes substantially (+5-mm error for spleen IS, −10% sDSC for liver). The automatic step to improve realism generally results in limited metric accuracy loss, but fails in one case (out of 60).

Conclusion: Our ML-based pipeline leads to phantoms that are significantly and substantially more individualized than currently used human-designed criteria.

 [DOI: 10.1117/1.JMI.7.4.046501]

Keywords: machine learning; pediatric cancer; radiation treatment; dose reconstruction; phantom.

Paper 19232RR received Sep. 7, 2019; accepted for publication Jul. 15, 2020; published online Jul. 30, 2020.

1 Introduction

Virtual anthropomorphic phantoms are three-dimensional (3-D) representations of the human body that are used as surrogates for the anatomy of humans, to estimate the quantity and

*Address all correspondence to Marco Virgolin, E-mail: marco.virgolin@cwi.nl

2329-4302/2020/$28.00 © 2020 SPIE

geometric distribution of radiation dose when having been exposed to radiation, e.g., in radiation treatment for cancer patients.[1,2]

We consider the scenario of radiation dose reconstruction for childhood cancer survivors who were treated in the pre-3-D planning era, i.e., before computed tomography (CT) scans became commonplace in radiation treatment practice.[3,4] Reconstructing the radiation dose that these patients were subject to is key to study the possible late adverse effects of radiation treatment and consequently improve future treatments.[5–7] Here phantoms are needed to act as a surrogate for the unknown 3-D anatomy of the patient.[8]

Nowadays, several types of (pediatric and adult) phantom exist. For example, research groups have been building phantoms as nonrealistic but easily reconfigurable cuboids,[8,9] or, conversely, as extremely realistic anatomies, by morphing patient CT scans and respective organ segmentations.[10–12] Efforts have also been done to build 4-D phantoms, i.e., phantoms that include realistic motion simulations, as well as phantoms that include a large number (a thousand) of organ segmentations.[13,14]

In the scenario we consider, phantom building techniques that assume 3-D patient imaging to be available cannot be used.[15] This leads to a key issue that existing approaches have in common: they build phantom libraries using large population statistics (e.g., ICRP '89[16]) and categorize them by simple hand-crafted criteria, typically based on age, gender, and/or height and weight percentiles.[9–12,14] When dose reconstruction for a childhood cancer survivor is needed, the criterion is applied to choose the phantom that corresponds to the patient's category. Several studies have, however, indicated that such simple criteria are incapable of capturing the high variance in internal anatomy.[2,3,10,17] Since the lack of resemblance between the patient's and the phantom's anatomy is a primary source of error,[18] this can ultimately lead to inaccurate dose reconstructions.[4] We remark that, for the purpose of dose reconstruction, anatomy resemblance (configuration and properties of internal organs) does not necessarily imply anatomy realism.[8]

To improve phantom individualization for patients with no 3-D imaging, adaptation techniques have been studied such as morphing organ shapes[19,20] and organ repositioning.[21] In this work, we attempt to improve upon the use of simple hand-crafted criteria, and upon adaptation techniques, by proposing an end-to-end approach based on nontrivial criteria (models) discovered by machine learning (ML).[22,23] In particular, we design an automatic pipeline that predicts a pseudorealistic, individualized phantom for a childhood cancer survivor, given features that are typically available from the pre-3-D planning era and a database of recent patient 3-D imaging. We remark that we primarily focus on individualization and only secondarily on anatomical realism, since the latter aspect is desirable but not a necessary condition to perform dose reconstruction.

We consider pediatric patients between 2 to 6 years and focus on dose reconstruction for abdominal radiation treatment, for the following reasons. First, children are typically under-represented in existing phantom libraries, i.e., phantoms are available for few broad categories.[9,10] Second, the inclusion of radiation treatment in cancer care has led to high survival rates for several types of pediatric abdominal cancer (e.g., Wilms' tumor, the most common type of kidney cancer), but it is known to cause late adverse effects.[6,24]

2 Materials and Methods

2.1 *Data*

We built a database using data of 60 pediatric cancer patients, in the age range of 2 to 6 years, roughly half males and half females (details in Table 1). The patients were treated after 2002 at the Radiation Oncology Department of the Amsterdam UMC, location AMC, in Amsterdam, or at the University Medical Center Utrecht/Princess Màxima Center for Pediatric Oncology in Utrecht, mostly but not exclusively for Wilms' tumor (almost 2 out of 3 patients). Nephrectomy was performed for 32 out of 60 patients (13 left and 19 right). For each patient, a CT scan is available (median axial thickness: 2.5 mm, median in-plane resolution: 1.0 mm $\times$ 1.0 mm) of the patient positioned on a flat couch in supine orientation. The CT scan fully includes the lower

Table 1 Features of our cohort, typically available for patients treated in the pre-3-D planning era.

Feature name	Abbreviation	Unit	Source	Min	Max	Mean	St. Dev.
Age	AGE	Years	Records	2.0	6.0	3.8	1.2
Abdominal diameter in AP at typical isocenter	ADAP	mm	Records	11.1	16.0	13.3	1.2
Abdominal diameter in LR at middle of L2	ADLR	mm	Radiograph	16.3	23.5	19.4	1.4
Distance from top of iliac crest to spinal cord along LR	ICSC	mm	Radiograph	4.3	6.8	5.5	0.6
Gender	GEND	–	Records	33 females, 27 males			
Heart size along LR	HESZ	mm	Radiograph	6.8	9.9	8.5	0.7
Height	HEIG	cm	Records	86.0	123.0	103.0	10.7
Left diaphragm length along LR	LDLR	mm	Radiograph	6.5	10.7	8.4	0.9
Right diaphragm length along LR	RDLR	mm	Radiograph	6.2	10.5	8.3	0.8
Right diaphragm top to T12 distance along IS	RDIS	mm	Radiograph	4.0	7.8	5.9	1.0
Spinal cord length along IS from T12 to L4	SPIS	mm	Radiograph	7.0	10.9	9.3	0.8
Weight	WEIG	kg	Records	10.0	28.0	16.4	3.7

Note: Gender is categorical, other features are numerical.

part of the thorax to the lower part of the abdomen, from the top of the thoracic 10th vertebra (T10) to the bottom of the sacral 1st vertebra (S1).

We simulated the scenario of dose reconstruction for patients treated in the pre-3-D planning era using the database of recent patient data, by considering only features that were typically recorded in the past. Note that this is a necessary condition to validate our approach and to train ML models (explained in Sec. 2.3). We base features availability choices considering the Emma Children's Hospital/Academic Medical Center (EKZ/AMC) childhood cancer survivor cohort (treated between 1966 and 1996).[6] For these patients, 2-D coronal radiographs are available. Consequently, we could consider measurements along left–right (LR) and inferior–superior (IS) directions based on the reliably visible bony anatomy (see Fig. 1). We simulated the availability of historical radiographs using digitally reconstructed radiographs (DRRs). We manually tuned DRR quality to resemble the quality of historical radiographs of the EKZ/AMC cohort: window level of 1441 hounsfield unit (HU), window width of 3200 HU, no bone enhancement. Figures 1(a) and 1(b) show an example of a historical radiograph and a DRR.

Table 1 lists the features considered in this work. Features involving measurements from DRRs were collected after manual placement of landmarks (using 3D Slicer software[25]) exemplified in Fig. 1(c). The abdominal diameter in anterior–posterior (AP) direction was historically measured using rulers and calipers, at the center of the radiation treatment field (corresponding with the isocenter for the EKZ/AMC cohort). For our cohort, we measured the abdominal diameter along AP from the CTs using a typical isocenter position for abdominal flank irradiation, as described in our previous work.[3] Figure 2 shows the Pearson correlation coefficients between the considered features. Most features are moderately correlated and few are strongly correlated, e.g., height with age and weight. The distance along IS between the top of the right diaphragm and T12 (RDIS) has the lowest correlation with any other feature. This is likely due to RDIS being heavily influenced by the breathing state of the patient,[26,27] which is itself independent from the other features.

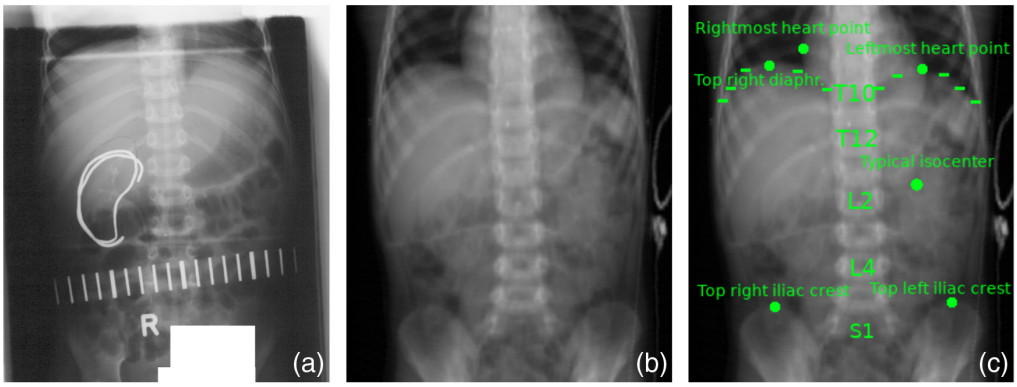

Fig. 1 (a) An example of a 2-D coronal radiograph, taken by a radiation treatment simulator used in the pre-3-D planning era, including annotations by medical personnel (sensitive information censored). (b) A DRR built from a CT. Images are acquired in AP setting. Liver and spleen are not clearly visible. (c) Example of manually placed landmarks used to measure features from radiographs. The length of the left and the right diaphragm along the LR direction is derived by fitting a cubic spline to the respective dashes.

Fig. 2 Pearson correlation coefficients (number and color-coded) between the considered features. See Table 1 for the meaning of abbreviations.

We consider two organs at risk (OARs), i.e., organs for which exposure to radiation is known to lead to adverse effects: the liver and the spleen. The shape and position of these OARs is known to vary substantially per individual, and they are considered hard to predict.[3] Moreover, the liver and the spleen are not (clearly) visible in historical radiographs (see Fig. 1). For each patient in the cohort, 3-D segmentations of the OARs and the external body (delimited along IS between T10 and S1) were first automatically generated (with ADMIRE research software, 2.3.0, Elekta AB, Stockholm, Sweden), then manually checked and corrected by a doctoral student (3 years of experience) and an experienced radiation treatment technologist (more than 25 years of experience), using Velocity software (version 3.2.0, Varian Medical Systems, Inc. Palo Alto, CA, US), and finally approved by a pediatric radiation oncologist (5 years experience).

2.2 *Pipeline for Automatic Phantom Construction*

Our pipeline takes as input features of a pre-3-D planning era pediatric patient and returns as output a personalized abdominal CT-based phantom. The phantom is built by assembling recent patient imaging data using ML predictions. The pipeline is summarized in Fig. 3. Given the patient features as input, the following phases follow.

Phase 1: Input features dispatching. The features of the patient are dispatched to ML models. These models were trained beforehand (as explained in Sec. 2.3). The models that predict similar OAR segmentations to retrieve ($\mathcal{M}_{\mathrm{S}}^{\mathrm{body}}$ and $\mathcal{M}_{\mathrm{S}}^{\mathrm{OAR}}$) also use features of patients in the database as input.

Phase 2: Receiver preparation. An ML model $\mathcal{M}_{\mathrm{S}}^{\mathrm{body}}$ is used to predict which CT is most resembling in terms of overall body shape (for the abdominal region). We call this CT the "receiver." Then the OARs of the receiver are "resected": the voxels of the OARs are set to HU values that represent generic soft abdominal tissues (we used 78 as done in the phantoms of the University of Florida/National Cancer Institute[10]). This way, no cavities are present that could harm dose calculations.

Phase 3: Prediction of similar OAR segmentation and its position. For each OAR, its center of mass position is predicted using independent models, i.e., one for the LR ($\mathcal{M}_{\mathrm{LR}}^{\mathrm{OAR}}$), one for the AP ($\mathcal{M}_{\mathrm{AP}}^{\mathrm{OAR}}$), and one for the IS position ($\mathcal{M}_{\mathrm{IS}}^{\mathrm{OAR}}$). A fifth model ($\mathcal{M}_{\mathrm{S}}^{\mathrm{OAR}}$) is used to predict which OAR segmentation to retrieve among the ones available based on a shape-focused metric (described in Sec. 2.3.1). We refer to the CT that is chosen by $\mathcal{M}_{\mathrm{S}}^{\mathrm{OAR}}$ to provide the OAR segmentation as a "donor."

Phase 4: OAR transplant. Each OAR segmentation is "transplanted" into the receiver, using its predicted position. Transplantation is achieved similarly to resection: we add the OAR segmentation to the set of segmentations of the receiver, placed in the predicted position, and we set the HU values of the voxels in the receiver that belong to the OAR segmentation to the respective HU values from the donor scan.

Phase 5: Anatomical inconsistency correction. Because ML predictions can in principle place OARs in nonrealistic positions, an anatomical inconsistencies correction (AIC) procedure is used to enhance phantom realism. AIC attempts to be minimal to ensure fidelity with respect to the individualized ML predictions (see Sec. 2.4).

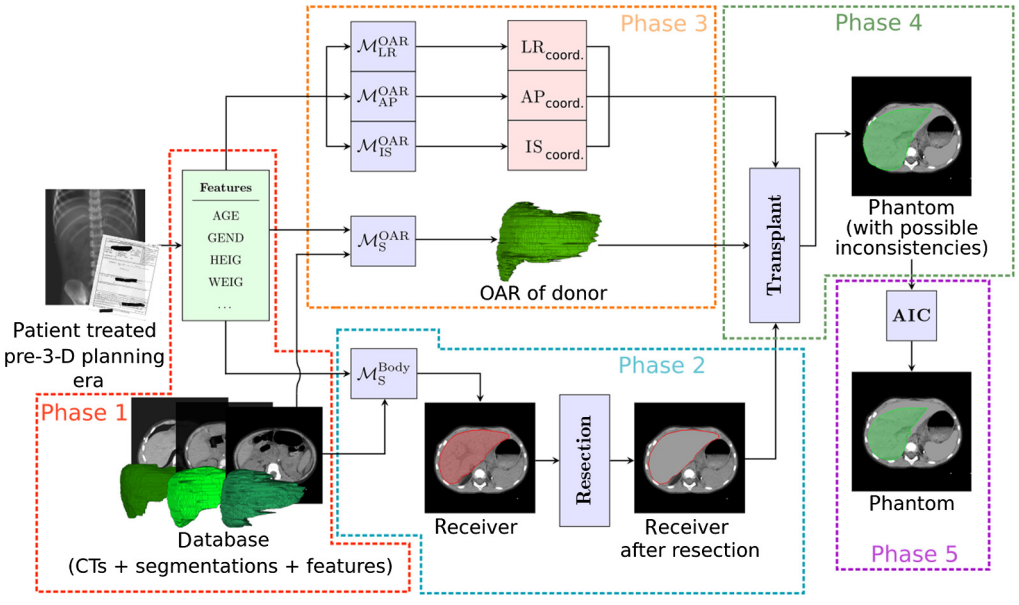

Fig. 3 Pipeline for automatic phantom building. Pretrained ML models are used to predict 3-D OAR positions as well as what OAR donor segmentation and what receiver CT to retrieve from the database to construct a patient-specific phantom CT (only the liver is considered as OAR in this example). The resected and transplanted OAR is highlighted in red and green, respectively.

To store CTs and segmentations sets, we relied on the Digital Imaging and Communications in Medicine (DICOM) standard. Segmentation sets were stored in DICOM RT-Structure sets. To handle this data format, we used the pydicom package for Python (https://pydicom.github.io/). For further details, we refer to our source code: http://github.com/marcovirgolin/APhA.

2.3 *Machine Learning*

2.3.1 *Datasets for training*

To train the ML models we prepare $1 + 4 \times n_{\text{OARs}}$ datasets: one to learn how to pick the receiver CT, and for each OAR three to predict the spatial position, and one to predict the segmentation to retrieve.

Features are normalized by z scoring[28] (subtraction of the mean and division by the standard deviation). Patient gender is set to a binary value (0 female and 1 male). As no correlation coefficient above 0.9 was found between the features (Fig. 2), we do not exclude any feature. z scoring is also applied to the target variable Y.

We construct two types of datasets. One type is used to learn OAR position (in LR, AP, and IS). The other type is used to learn a notion of shape similarity between segmentations. The preparation of the datasets for learning OAR position is straightforward. We define the OAR position relative to a common landmark: the center of the L2 vertebra. In other words, the OAR position in LR, AP, or IS direction is a signed, one-dimensional distance from the center of the L2 vertebra to the center of mass of the OAR's segmentation, measured in mm. We do not account for possible (but typically limited) patient tilting.

For the datasets concerning OAR segmentation similarity, a measure of similarity needs to be chosen to act as target variable. Note that a measure of similarity is defined between pairs of segmentations, hence each example needs to be defined over a pair of patients, leading to a total of $n_{\text{patients}}(n_{\text{patients}} - 1)/2$ training examples. A pairwise version of the features is considered here, defined as the absolute difference of the features of patients p and q: $x_{p,q} = |x_p - x_q|$, with x_p being the feature x of patient p. For the target variable, several possibilities to address segmentation similarity have been proposed in the literature. For example, a possibility is to consider similarity in OAR weight.[12] We do not rely on OAR weight nor volume because they do not account for similarity of shape. Another commonly adopted option to assess OAR similarity is the use of the volumetric Dice–Sørensen coefficient (DSC).[3,29,30] However, the DSC still has limitations, because it is segmentation-volume dependent. Similarly, we do not consider the Hausdorff distance because it is determined by a single point.[31] We choose to rely on the recently introduced surface Dice–Sørensen coefficient (sDSC),[32] which considers only significant millimetric deviations between the surfaces of the segmentations to evaluate similarity. The sDSC uses a threshold parameter τ that expresses what deviations are acceptable (e.g., as part of inter-observer variability). We set $\tau = 5.0$ mm, i.e., double the median CT slice thickness in our database. A valid alternative to the sDSC is the average symmetric surface distance.[33–35] We nevertheless preferred the use of the sDSC because the average symmetric surface distance considers all deviations as significant and is not a percentage. Because we predict OAR positions separately for each OAR, the sDSC is computed after aligning the segmentations on their center of mass, i.e., to maximally focus on shape similarity.

2.3.2 *Loss function for learning*

As loss function to train and validate the ML algorithms, we consider the mean absolute error (MAE), i.e.,

$$\text{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|,$$

where y and $\hat{y}$ are n-dimensional vectors of ground-truth values and ML predictions, respectively. We choose to use the MAE over, e.g., the (rooted) mean squared error, because it weighs

all errors equally, meaning it is relatively robust to outliers (important considering the limited number of patients).

Measuring the MAE in OAR positions committed by our pipeline is trivial. It corresponds to the MAE of the ML predictions. This is not the case for segmentation retrieval ($\mathcal{M}_S^{OAR}$ and $\mathcal{M}_S^{body}$ in Fig. 3). Given the predictions of sDSCs between the OAR segmentation of the test patient (O^t) and the OAR segmentations of the database, we retrieve from the database the segmentation with the largest sDSC prediction ($O^\star$). Now, the actual sDSC between O^t and $O^\star$ may be different from the predicted one, hence we calculate it anew: this is the sDSC we ultimately report. Note also that the maximum score of OAR shape similarity achievable is not necessarily 100%, rather, it is the maximum sDSC that can be obtained with the OAR segmentations available in the database.

2.3.3 *Algorithms*

We compare a total of five regression ML algorithms: ridge regression (RidReg),[36,37] the least absolute shrinkage and selection operator (LASSO),[38] random forest (RF),[39] traditional genetic programming (GP-Trad),[40,41] and the genetic programming instance of the Gene-pool Optimal Mixing Evolutionary Algorithm (GP-GOMEA).[42,43]

RidReg and LASSO combine features linearly, include penalization metrics to avoid overfitting and build models that can be read as often interpretable mathematical expressions.[44] By "interpretable," we specifically refer to the concept of "global interpretability," meaning that the entirety of the logic of a model can be understood, and any outcome can be foreseen.[45] We use the implementation of RidReg and LASSO from the package glmnet in R.[46,47]

RF allows nonlinear modeling, by means of an ensemble of decision trees. Because the model is an ensemble, it is considered not interpretable. [44] We use the R implementation known as ranger.[48]

The GP algorithms are interesting because, like RidReg and LASSO, they deliver models in the form of mathematical expressions that can, however, include nonlinear feature combinations. GP-Trad is typically employed as a baseline algorithm, whereas GP-GOMEA was recently found to work particularly well when small, yet accurate expressions are needed.[43] We use the C++ implementation of GP-Trad and GP-GOMEA (https://github.com/marcovirgolin/GP-GOMEA).

Hyperparameter tuning is performed with fivefold grid-search cross validation to promote minimal MAE. More details are given in Appendix A.

2.3.4 *Validation strategy*

Because the cohort is relatively small (60 patients), we use leave-one-out cross validation, performed "patient-wise," i.e., all the examples relative to a particular patient p are removed from the training set and solely used for testing. For the datasets on OAR segmentation retrieval, where each row represents a pair of patients (see Sec. 2.3.1), all $n_{\text{patients}} - 1$ (59) rows, where p appears are removed from the training set and put in the test set. Note that this is necessary to avoid a positive bias in the test results.

Since RF, GP-Trad, and GP-GOMEA are stochastic algorithms, we repeat their execution 10 times and report the mean result.

2.4 *Anatomical Inconsistency Correction*

The last phase of our pipeline repairs possible anatomical inconsistencies present in the assembled phantoms. We automatically detect and correct for possible overlaps between the transplanted OARs (liver and spleen), and between the OARs and the spinal cord segmentation of the receiver, which is available for these patients. To have an additional margin, we enlarge the spinal cord segmentation by 10% uniformly in all dimensions. Furthermore, we assess if the transplanted OARs stick out of the segmentation of the body of the receiver. If a larger segmentation of the body exists than the common region of interest between T10 and S1 (used to assess

body shape similarity to train ML models), that segmentation is used here. Again, for robustness, the body segmentation is shrunk uniformly in all dimensions, by 2.5%.

We use a general purpose, derivative-free real-valued optimization algorithm to modify the transplanted OAR segmentations to eliminate the anatomical inconsistencies, with default parameter settings.[49,50] The algorithm modifies the OAR segmentations by expansion/shrinking (up to 1.25 and 0.25 of the original volume, respectively), and repositions their center of mass (up to 10 mm for each direction). We set anatomical inconsistencies as hard constraints to satisfy.

Modifying the OARs to satisfy the constraints deviates from the predictions of the ML models and can therefore result in making the phantom less individualized. Therefore, the objective function is set to minimize the effect of OAR modifications: for each OAR, we use the sDSC (again with a threshold of 5 mm) between its current modified state and its version premodifications. The final objective is given by summing the sDSC of the OARs to be maximized.

2.5 *Comparing to Phantom Selection Approaches*

As we mentioned in the introduction, current phantom-based dose reconstruction methods build phantom libraries and select a representative phantom according to a hand-crafted criterion. As CTs of actual patients can act as phantoms, we consider several approaches to compare with our pipeline: two methods that simulate state-of-the-art human-designed criteria used to build and select from phantom libraries, random selection, and one method to select a single CT from our database based on ML predictions.

2.5.1 *Human-designed criteria for phantom selection*

The first methods we consider are the criterion used by the University of Texas MD Anderson Cancer Center,[8,9] which we refer to as human criterion 1 (HC1), and the criterion used by the University of Florida/National Cancer Institute,[10] which we refer to as human criterion 2 (HC2). We further consider random selection to see if the other methods are better than random.

The phantoms on which HC1 has previously been used are virtual cuboid shapes with OARs represented as point clouds.[8,9] Only the age of the patient is considered as a feature to manipulate the phantom's representativeness by scaling the cuboids according to guidelines on population data. Furthermore, gender is used to exclude/include gender-specific OARs. To simulate HC1, i.e., age binning, we cluster our database into age bins, by rounding the age to years. This results in 5 bins, with the following distribution: 10 patients of age 2, 18 of age 3, 11 of age 4, 15 of age 5, and 6 of age 6.

For a test patient p, we consider the other patients from the database that share the same age bin with p. Then for each metric of interest, we report the average error given by comparing the metric value for p with the one for each other patient that has the same age as p. For example, for the assessment of liver segmentation similarity of a patient p that is 3 years old, we compute the sDSC between p and each other patient that is 3 years old and return the mean. By doing this, we simulate the fact that an average anatomy has been built using the anatomical information of all patients (different from p) that have the same age. We do this because phantoms are built to represent average anatomies.

The phantom library where HC2 is adopted comprises phantoms made by scaling segmentations acquired from actual patient CT scans, thus they are quite realistic.[10] For these phantoms, the features considered to build the library were gender, height, and weight (age was not used). HC2 uses these features to select a representative phantom. We simulate HC2 by clustering our database by gender, and by height and weight, using 5 bins for each of the latter two. We use the bin method of the R package binr (http://jabiru.github.io/binr) for this purpose, with default settings, which results in 5 bins of 12 patients each. Like for HC1, the error for a metric is given by comparing the metric value for the test patient p with the metric values of each patient that shares the same bin of p and taking the average.

We further consider a control method where the OAR positions and segmentations are retrieved uniformly at random from the set of patients, excluding the test patient p. Because of the stochastic nature of this approach, each iteration is repeated 10 times, and mean results are reported. In the following, we refer to this method as RAND.

As we propose a pipeline to build a phantom, it is interesting to assess how it fares against a simpler approach, e.g., to use ML predictions to select a single, overall most representative CT scan. This approach can be related to literature, where new ways to identify which phantom to pick are studied.[3,51,52]

Although our phantom construction pipeline can predict the different 3-D metrics independently (position for each direction, sDSC for each OAR segmentation), for a single CT approach, an overall score needs to be defined that expresses how representative a CT scan is. The design of such a score is not trivial.[3] For example, a choice needs to be made on whether 5 mm along the LR direction is more important or less important than an sDSC loss of 5%. For the sake of simplicity, we propose a score measure that considers only OAR positions, with equal importance. We choose to focus on OAR position because we recently found it to be the metric that is most correlated with dose accuracy for pediatric abdominal radiation treatment.[53] In detail, we take as the best CT the one that is closest to the predictions of the ML models in terms of squared position differences, i.e.,

$$\text{best CT} = \text{argmin}_{\text{CT}} \sum_{\mathcal{O} \in \text{OARs}} \sum_{\mathcal{D} \in \{\text{LR,AP,IS}\}} (\mathcal{D}^{\mathcal{O}}_{\text{ML}} - \mathcal{D}^{\mathcal{O}}_{\text{CT}})^2, \tag{1}$$

where $\mathcal{D}^{\mathcal{O}}_{\text{ML}}$ and $\mathcal{D}^{\mathcal{O}}_{\text{CT}}$ are, respectively, the ML-predicted position and the actually available position in the CT, of the OAR $\mathcal{O}$, along direction $\mathcal{D}$. We denote this approach with single computed tomography (sCT).

Note that sCT is essentially a hybrid between a human-designed criterion, represented by Eq. (1), and the use of ML, of which the predictions are used in this equation. Finally, since we found GP-GOMEA to be the overall best performing ML algorithm (shown later in Sec. 3.1), we used the models found by GP-GOMEA to provide the predictions for sCT.

2.6 *Experimental Setup*

We divide the experiments into two parts. In the first part, we compare the predictions of the ML algorithms in terms of MAE. In the second part, we compare the prediction of the overall best performing ML algorithm with the phantom selection approaches in a similar way. In this second comparison, we include the effect of anatomical inconsistency correction to assess how much it compromises the accuracy of the predictions of the ML models.

We run the ML algorithms on a machine with two Intel® Xeon® CPU E5-2699 v4 at 2.20 GHz. We assess statistical significance of the results as follows. Given the two 60-D vectors resulting from the leave-one-out cross validation performed using two different ML algorithms (or human-criteria), we run one statistical significance test to assess what algorithm performed best. Note that the patient tested in one round of leave-one-out is independent from the other patients tested in the other rounds. We use the Wilcoxon signed-rank test, paired by leave-one-out folds.[54] This means that we account for the effect of the same groups of patients being used in the training phase or in the test phase when comparing two methods. We also use the Bonferroni correction method to prevent type I errors.[55] In particular, since we perform pairwise tests between the algorithms for each metric, we assess whether the test p-value is below a confidence level of 0.05, further reduced by Bonferroni correction to contrast false positive outcomes due to chance.

3 Results

3.1 *Comparison of the Machine Learning Algorithms*

The mean (and standard deviation) training and test MAEs obtained from the (average across 10 repetitions) leave-one-out cross validation of the ML algorithms are reported in Table 2. If multiple results are not found to be statistically significantly worse than the best result, they are considered to be equally good. Note that for consistent use of error minimization, the task of segmentation retrieval is reported in terms of $\varepsilon_{\text{sDSC}} = 100 - \text{sDSC}$.

Table 2 Mean training and test MAEs for the ML algorithms on the different OAR-specific regression tasks. Standard deviation is reported in subscript. The MAE for OAR segmentation retrieval is a percentage, the MAE for OAR position estimation is in mm. Results in bold are best in that no other method delivers significantly better ones. The letter "S" stands for segmentation retrieval. Bonferroni correction coefficient = 180.

		Body	Liver				Spleen				
		S	LR	AP	IS	S	LR	AP	IS	S	n_{best}
Training	RidReg	$16.63_{0.08}$	$7.89_{0.24}$	$4.27_{0.14}$	$7.10_{0.33}$	$17.02_{0.07}$	$\mathbf{3.99}_{0.21}$	$7.21_{0.13}$	$7.71_{0.19}$	$15.39_{0.10}$	1
	LASSO	$16.64_{0.08}$	$7.87_{0.36}$	$4.24_{0.10}$	$7.28_{0.19}$	$17.02_{0.07}$	$4.03_{0.10}$	$7.24_{0.13}$	$7.60_{0.13}$	$15.38_{0.09}$	0
	RF	$\mathbf{6.74}_{0.33}$	$8.36_{0.14}$	$4.87_{0.08}$	$8.74_{0.09}$	$\mathbf{12.37}_{1.08}$	$5.44_{0.07}$	$7.25_{0.14}$	$9.33_{0.14}$	$\mathbf{7.21}_{0.83}$	3
	GP-Trad	$19.55_{0.06}$	$\mathbf{6.97}_{0.11}$	$\mathbf{3.93}_{0.07}$	$6.42_{0.10}$	$15.97_{0.03}$	$4.01_{0.05}$	$6.56_{0.13}$	$7.06_{0.15}$	$14.08_{0.04}$	2
	GP-GOMEA	$19.55_{0.06}$	$\mathbf{6.97}_{0.11}$	$\mathbf{3.93}_{0.07}$	$\mathbf{6.38}_{0.09}$	$15.96_{0.03}$	$\mathbf{3.95}_{0.05}$	$\mathbf{6.47}_{0.12}$	$\mathbf{7.05}_{0.15}$	$14.07_{0.04}$	6
Test	RidReg	$23.70_{15.31}$	$8.54_{6.83}$	$\mathbf{4.82}_{4.57}$	$9.10_{5.33}$	$\mathbf{38.90}_{17.84}$	$5.18_{3.33}$	$\mathbf{7.42}_{8.06}$	$8.52_{7.48}$	$35.12_{14.51}$	3
	LASSO	$22.78_{15.34}$	$8.87_{6.91}$	$\mathbf{4.86}_{4.49}$	$8.98_{5.37}$	$\mathbf{38.98}_{19.11}$	$5.02_{3.21}$	$\mathbf{7.37}_{8.10}$	$8.68_{7.47}$	$36.82_{13.59}$	3
	RF	$\mathbf{21.38}_{14.59}$	$8.36_{6.55}$	$\mathbf{4.77}_{4.62}$	$7.94_{5.73}$	$41.12_{16.72}$	$4.98_{3.48}$	$7.83_{8.00}$	$8.80_{7.74}$	$\mathbf{33.98}_{15.47}$	3
	GP-Trad	$27.08_{16.67}$	$\mathbf{7.27}_{6.48}$	$\mathbf{4.68}_{4.30}$	$\mathbf{7.23}_{5.89}$	$\mathbf{40.61}_{14.25}$	$5.23_{3.40}$	$8.35_{8.22}$	$8.43_{9.73}$	$35.70_{10.57}$	4
	GP-GOMEA	$26.77_{16.75}$	$\mathbf{7.26}_{6.48}$	$\mathbf{4.78}_{4.36}$	$7.89_{6.03}$	$\mathbf{39.00}_{19.31}$	$\mathbf{4.56}_{3.49}$	$\mathbf{7.33}_{8.32}$	$\mathbf{8.21}_{9.78}$	$35.62_{11.43}$	6

Errors are typically of comparable magnitude. In terms of training performance, GP-GOMEA is overall the best algorithm, as it is not significantly worse than any other for six metrics, i.e., OAR position for all directions. RF follows with three top performances, in particular for the segmentation retrieval task (S) of all OARs, where it achieves markedly lower errors than all other algorithms. LASSO performs the worst, with at least one algorithm significantly outperforming it in each metric. Standard deviations are overall small in the training phase.

Regarding the test performance, GP-GOMEA significantly obtains again the best results for most metrics. GP-GOMEA also generalizes well on predicting which segmentation to select for the liver, but it is inferior to RF when it comes to selecting the segmentations for the body and the spleen.

Although RF performs better than any other ML algorithm for the body and spleen segmentations, the errors at test time are much larger than the ones found at training time. Note that this mismatch between training and test performance can be explained by possible overfitting, as well as by the fact that, at test time, error propagation can happen (as explained in Sec. 2.3.2). Further, standard deviations are much larger when testing than when training. In our case, this is because of having limited data available (see Appendix B). Note that the best performing algorithms according to statistical tests do not necessarily have the smallest deviations (see, e.g., spleen S).

3.2 *Machine Learning versus Phantom Selection Approaches*

Table 3 shows results comparing the ML algorithm found to perform best overall, GP-GOMEA, to the use of phantom selection approaches. The use of anatomical inconsistency correction upon GP-GOMEA's predictions is also included now (named $\mathrm{GP}-\mathrm{G}_{\mathrm{AIC}}$). We remark that GP-GOMEA and $\mathrm{GP}-\mathrm{G}_{\mathrm{AIC}}$ are the same in terms of error for body S because AIC does not introduce corrections to the body segmentation.

The correction-less predictions obtained with GP-GOMEA are generally best, with the only exceptions being the IS position and segmentation retrieval for the spleen (by relatively small errors on average). The use of anatomical inconsistency correction upon GP-GOMEA's predictions, $\mathrm{GP}-\mathrm{G}_{\mathrm{AIC}}$, can change shape and position of the OARs in both considerable (e.g., liver AP and liver S) and minor (e.g., liver LR and spleen IS) magnitude. Figure 4 shows box plots of the corrections on all phantoms. For our database, the AIC was triggered for 31/60 phantoms. The liver is subject to more corrections than the spleen: it is typically shrunk more than the spleen, and its position in AP is subject to large variations. This is not surprising, because the liver is a considerably larger organ than the spleen, and it is more likely to violate the anatomical consistency constraints we imposed.

Overall, correcting for inconsistencies typically comes at the cost of worsening the accuracy of the ML predictions. $\mathrm{GP}-\mathrm{G}_{\mathrm{AIC}}$ significantly has the best performance only on three metrics (body S, spleen AP, and spleen S). For spleen AP and spleen S, performing inconsistency correction leads to better test results compared to not applying corrections. However, the correction algorithm solely optimizes for resolving the inconsistencies (while attempting to retain maximum prediction fidelity). Moreover, because sCT also obtains good results on spleen AP and spleen S, it can be argued that these metrics are not modeled sufficiently by GP-GOMEA to begin with, and thus are more likely to be improved upon by chance.

Although the use of anatomical inconsistency correction after ML prediction typically leads to larger errors, it remains a valuable approach compared to the other methods. $\mathrm{GP}-\mathrm{G}_{\mathrm{AIC}}$ is the overall best when the correction-less GP-GOMEA predictions are excluded from the comparison. The human-designed criteria HC1 and HC2 perform overall worse than $\mathrm{GP}-\mathrm{G}_{\mathrm{AIC}}$. Despite its simplicity, HC1 performs well for AP and S of the liver. This may be because metrics related to the liver are harder to model by the ML algorithms and because the use of anatomical inconsistency correction compromises GP-GOMEA's prediction (particularly notable for liver AP). Despite the fact that HC2 is a somewhat more involved criterion compared to HC1 (HC2 considers gender, height, and weight, whereas HC1 considers only gender and age), it is never found to be competitive on any metric. This could mean that weight and height are not good features when accounting for similarity of internal anatomy in children. This result is in agreement with

Table 3 Mean test MAEs of the overall best performing ML algorithm GP-GOMEA, also including anatomical inconsistency correction (GP – G_{AIC}), and the phantom selection approaches. Standard deviation is reported in subscript. The MAE for OAR segmentation retrieval is a percentage, the MAE for OAR position estimation is in mm. Results in bold are best in that no other method delivers significantly better ones. Results italics and bold italics are best if GP-GOMEA is excluded from the comparison. The letter "S" stands for segmentation retrieval. Bonferroni correction coefficient = 135.

		Body	Liver				Spleen				n_{best}
		S	LR	AP	IS	S	LR	AP	IS	S	(Underlined)
Test	GP-GOMEA	$\mathbf{26.77}_{16.75}$	$\mathbf{7.26}_{6.48}$	$\mathbf{4.78}_{4.36}$	$\mathbf{7.89}_{6.03}$	$\mathbf{39.00}_{19.31}$	$\mathbf{4.56}_{3.49}$	$\mathbf{7.33}_{8.32}$	$8.21_{9.78}$	$35.62_{11.43}$	7 (–)
	GP – G_{AIC}	$\boldsymbol{26.77}_{16.75}$	$7.91_{6.27}$	$6.21_{5.42}$	$\boldsymbol{8.07}_{6.06}$	$42.20_{20.29}$	$5.18_{4.04}$	$\boldsymbol{7.35}_{8.30}$	$8.54_{9.67}$	$\boldsymbol{34.10}_{16.97}$	4 (7)
	HC1	$37.54_{10.82}$	$8.88_{6.90}$	$\boldsymbol{4.83}_{4.70}$	$9.10_{6.13}$	$\boldsymbol{43.19}_{8.35}$	$5.65_{3.68}$	$7.96_{7.75}$	$9.78_{8.20}$	$37.58_{8.53}$	2 (2)
	HC2	$36.83_{18.97}$	$9.84_{6.26}$	$5.18_{4.74}$	$9.20_{6.48}$	$49.67_{8.2}$	$5.54_{4.21}$	$8.02_{7.69}$	$13.91_{11.09}$	$34.76_{10.37}$	0 (0)
	RAND	$45.10_{13.14}$	$11.57_{6.19}$	$7.11_{4.01}$	$12.38_{4.48}$	$44.01_{8.96}$	$7.70_{3.29}$	$11.14_{7.22}$	$13.52_{7.05}$	$35.89_{8.11}$	0 (0)
	sCT	$37.13_{22.00}$	$15.29_{10.02}$	$7.44_{5.91}$	$11.45_{9.2}$	$42.72_{18.30}$	$4.85_{3.43}$	$\boldsymbol{7.40}_{8.32}$	$\boldsymbol{7.84}_{9.37}$	$\boldsymbol{32.08}_{12.28}$	3 (5)

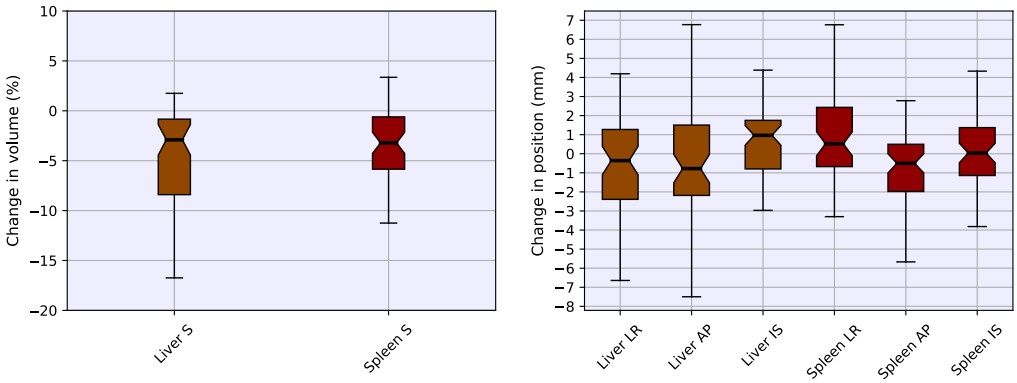

Fig. 4 Distribution of the effect of the anatomical inconsistency correction on all phantoms (29/60 are not corrected). OAR shape (S) is corrected by volume enlargement (if >0%) or shrinking (if <0%) uniformly along the three dimensions. Change of center of mass for AP, LR, and IS is in mm. Phantoms where correction is not needed, contribute to OAR shape modification with 0% (no enlargement nor shrinking), and to OAR position change with 0 mm (no repositioning). Boxes extend from the 25th to the 75th percentiles, inner bar is the median, and whiskers extend from the 10th to the 90th percentiles.

the previous work[53] where lack of correlations between dose reconstruction outcomes was found with respect to height and weight.

Importantly, in some cases, HC1 and HC2 perform particularly poorly. HC1 is particularly inaccurate for spleen S compared to the other approaches, and HC2 leads to notably large errors for spleen IS. In the latter case, HC2 is not found to be better than RAND. Furthermore, both criteria perform poorly on body S (as well as sCT). Figure 5 shows the reliability function of the considered approaches for body S. This function shows the likelihood of committing errors of a certain magnitude (for segmentation retrieval, in percentage). A curve is better than the others if it is more on the left and if it decreases more rapidly since this means that the probability of any error magnitude is lower than the ones of the other methods. This figure illustrates that the model learned by GP-GOMEA achieves this behavior. For example, the probability of a GP-GOMEA's model to predict a body segmentation that has an $\varepsilon_{\text{sDSC}} > 20\%$ (sDSC < 80%) with the actual body of the patient is 60%. For HC2 and sCT, that magnitude of error happens more frequently, i.e., in almost 80% of the cases. HC1 performs worse since errors above 20% are almost certain. On the other hand, rarely, (probabilities around 5%), HC2 and sCT can retrieve body segmentations that have errors above 80%.

Regarding sCT, the results indicate that this approach is generally worse than $\text{GP} - \text{G}_{\text{AIC}}$, but better than the human-designed criteria to predict metrics related to the spleen. sCT performs

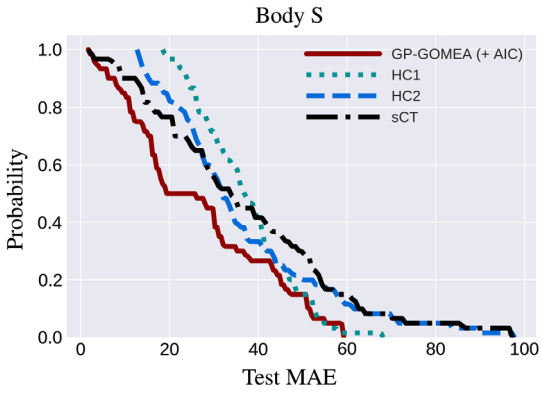

Fig. 5 Reliability functions for prediction of segmentation retrieval (S) of the external body segmentations. The *y* value is the probability of committing an error equal or greater than the *x* value. The test MAE is in $\varepsilon_{\text{sDSC}}$. Note that the anatomical inconsistency correction does not alter the body segmentation. GP-GOMEA leads to much better performance compared to the other approaches.

particularly poorly with respect to metrics regarding the liver (note in particular liver LR). This is an interesting result because the CT selected by sCT weighs OAR positions equally for liver and spleen [see Eq. (1)]. We remark that sCT uses the ML models (found by GP-GOMEA) to predict which single CT scan (and accompanying segmentations) to select by means of a hand-designed metric. Essentially, the results confirm our hypothesis that designing a good score to select one CT is not trivial and that there is added value in constructing a new anatomy by assembling different components into one. sCT may, however, still be preferable if anatomical consistency is desired, which is proven to be a challenging task to achieve when assembling a phantom based on multiple CTs (see Sec. 4).

3.3 *Model Interpretability*

As we mentioned before, the added value of utilizing the linear ML algorithms (RidReg and LASSO) and the GP algorithms (GP-Trad and GP-GOMEA) is that their models are mathematical expressions that, if simple enough, can be globally interpretable. [45] Interpretability can play a crucial role in determining whether ML can be applied in some clinical settings. For RF, the interpretation of the model is essentially impossible, as it returns an ensemble of (500) decision trees. Similarly, this would not be possible with other popular techniques like deep learning[56] and boosting algorithms.[57]

We report all the best-at-test-time models found by GP-GOMEA in Appendix C. Here we report two of those models that are remarkably simple. For the prediction of which body segmentation to retrieve, the model found most frequently in 10 repetitions is (assuming the target variable and the features are normalized):

$$0.420 \times (\mathrm{ADAP} + \mathrm{ADLR} + \mathrm{SCIS}).$$

Notably, this model is just a scaled sum of the abdominal diameters (ADAP and ADLR) and the spinal cord length (SCIS). Essentially, the model uses an equal contribution (features are normalized) of features capturing information of size relative to the 3-D LR, AP, and IS, to predict which body segmentation to retrieve. This seems a reliable, simple, and reasonable method to select a representative body segmentation.

A second model we showcase here is the one for the prediction of what spleen segmentation to retrieve:

$$2.718^{\mathrm{AGE}} \times 0.057 \times \mathrm{SCIS}.$$

It is interesting to see that spleen shape is found to be related to age by an exponentiation. This can be considered reasonable for our cohort, because the age of our patients is between 2 to 6 years, where anatomical development is rapid. The length of the spinal cord in IS further weighs the prediction. This feature seems also reasonable to consider, because it captures information relative to the size of the abdomen in IS, and because the spleen is located nearby the spinal cord. Albeit understandable, reasonable, and well-performing, it is arguably unlikely for humans to invent models like these.

4 Examples of Automatically Constructed Phantoms

Figure 6 shows qualitative results: five example phantoms generated with our pipeline (using GP-GOMEA, which was used to train all models). The images are created by using 3D Slicer[25] with module SlicerRT.[58] Note that the segmentations of the external body and the spinal cord shown in this figure can extend beyond the commonly available region of interest used for training (from T10 to S1). The receiver CT's original liver and spleen can also be identified (in part), as their not overridden voxels are uniformly set to a specific value (78 HUs as done for the phantoms of the University of Florida/National Cancer Institute[10]) in the resection step. Overlaps of the transplanted OAR can still happen with respect to other organs that are not considered in the training and correction process (e.g., spleen overlapping with the kidneys). Note also the presence of some further remaining limited anatomical inconsistencies, which are not detected by our algorithm (e.g., small overlaps with bony anatomy).

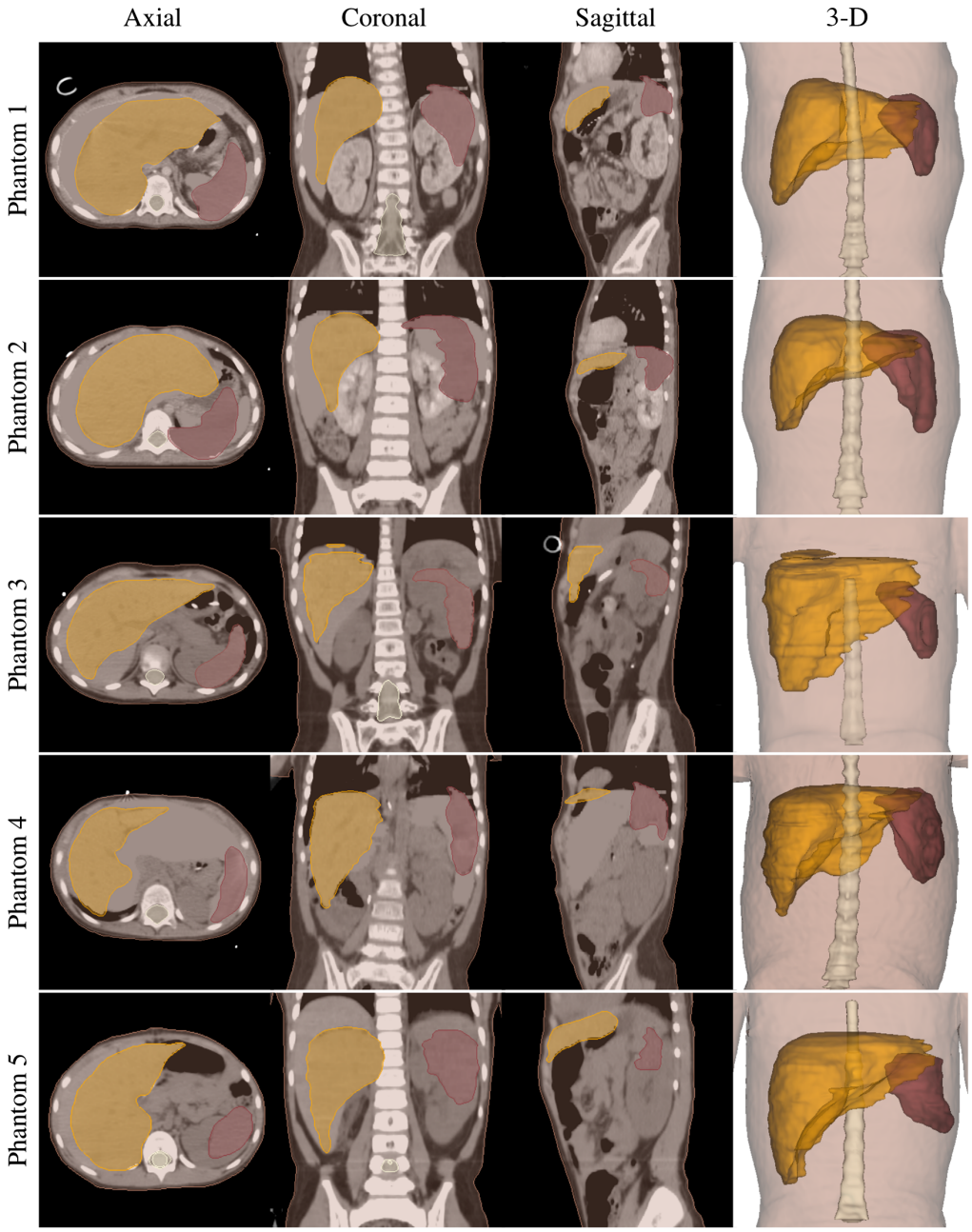

Fig. 6 Examples of phantoms constructed with our pipeline. CT snapshots are chosen to attempt to display both liver (in ocher) and spleen (in crimson red). Axial views are in IS, coronal views and 3-D views are in AP, and sagittal views are in LR.

Our current Python implementation takes about 5 min to generate a phantom if no anatomical inconsistencies are found, with resection and transplant taking most of the time (we expect that parallelizing voxels' HU value overwrites will markedly reduce running time). If anatomical inconsistency correction needs to be performed, the whole pipeline can take from a few minutes to a few hours, depending on how complex the correction is for the optimizer, and on the hardware (the optimizer is highly parallelizable). As mentioned in Sec. 2.3, for our database, the anatomical inconsistency correction triggered on half the phantoms. However, in our opinion, only roughly half of those cases (so 1/4 of the total number of phantoms) had really noticeable anatomical inconsistencies, e.g., large OAR overlaps or OARs exceeding the body boundaries, that required corrections of large magnitude. In the other cases, inconsistencies were more subtle (e.g., small OAR overlap) and caused corrections of small magnitude.

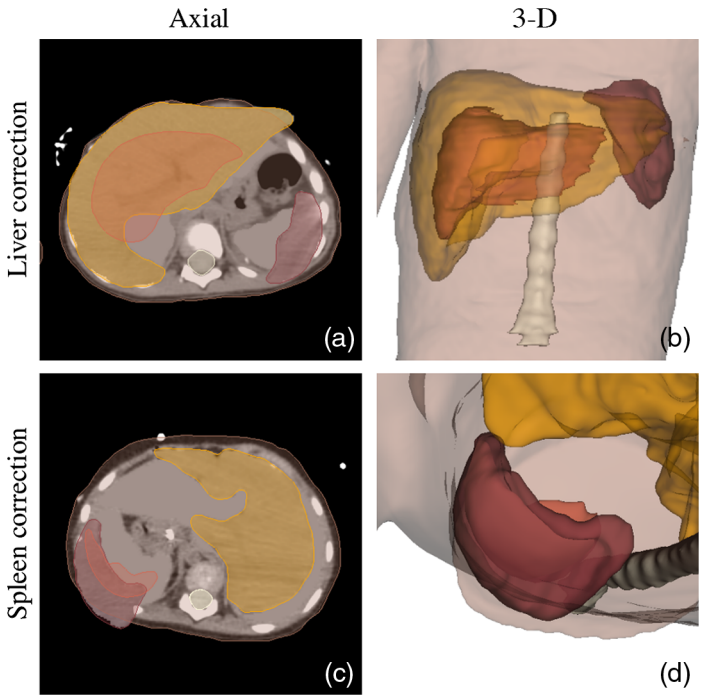

Fig. 7 Examples of anatomical inconsistency correction tackling OARs being partly positioned outside of the body segmentation. Precorrection liver is in ocher, postcorrection in orange; precorrection spleen is in crimson red, postcorrection is in salmon pink. (a) Axial and (b) 3-D view of large anatomical inconsistency involving the liver. (c) Axial and (d) 3-D view of large anatomical inconsistency involving the spleen. Note that this axial view is different from the ones in other figures as it is taken in SI direction (spleen displayed on the left), to be consistent with the 3-D visualization. The 3-D view shows the back of the patient from head to toes.

Figure 7 shows two examples of anatomical inconsistency correction for cases where inconsistency can be considered large, displaying phantoms pre- and postanatomical inconsistency correction, one to correct the liver, and one to correct the spleen.

5 Discussion

We have presented a new take on phantom construction: a fully automatic ML-based pipeline that predicts how to assemble a patient-specific phantom using a database of delineated 3-D CT scans, given the features of long-term childhood cancer survivors. We performed experiments upon data of 60 pediatric patients including imaging of the abdomen and focused on tailoring the position and shape of the liver and the spleen.

Our experimental results strongly suggest that our approach leads to phantoms that are quantitatively more representative of the unknown 3-D anatomy of the patient compared to the use of current human-designed criteria and to the use of an ML-based overall anatomical similarity metric to pick a single CT. The use of our method for dose reconstruction for childhood cancer survivors is, therefore, expected to provide high accuracy. A collaborative study where our method (extended to more OARs) is compared with the methods of the University of Texas/MD Anderson Cancer Center[8] and the University of Florida/National Cancer Institute[10] is currently in preparation.

One clear limitation of our work is that we could only employ a small database of 60 patients. It can be reasonably expected that, by increasing the database size, both the errors of the ML models and the onset of large anatomical inconsistencies will be reduced. Appendix B includes an experiment measuring ML performance for increasing database size. We are currently working on expanding the number of institutes contributing to the database. Furthermore, we are working on extending the number of OARs to build the phantom (e.g., heart and kidneys) and

on extending the region of interest (e.g., abdomen and thorax), for a cohort including also slightly older children (2 to 8 years). With more data available, we could consider including models for the prediction of more complex OAR transformations (e.g., resizing and rotations). In the future work, it is of interest to also compare methods across more OAR similarity metrics than the sDSC and center of mass position by considering the metrics proposed in segmentation challenges.[59]

To improve realism, we proposed a simple AIC method, which proved sufficient to resolve the constraint we imposed, often without excessively compromising the predictions of the ML models. An alternative is to impose restrictions to the ML predictions, e.g., by designing (OAR-specific) penalization metrics. However, designing competent penalization metrics is a nontrivial effort.

Even if AIC resolves all the automatically identified inconsistencies, other inconsistencies can still be present. This is because the constraints are not comprehensive enough. Figure 8 shows an example of phantom where none of our constraints is violated (liver and spleen do not overlap with each other nor with the spinal cord, and the organs do not exceed the body contour), yet the anatomy remains unrealistic because the organs are placed too high with respect to the receiver CT. In fact, this is because it is the prediction of the receiver CT that does not work well for this patient, as it retrieves a body which is quite shorter (14-cm smaller SCIS, sDSC of 60%) compared to the actual body of the patient. Fortunately, this is the only phantom out of 60 where such an evident inconsistency was found and likely the probability of this happening only reduces with a growing database size. Still, we intend to study how to further improve our correction method by attempting to craft more constraints and allowing for nonuniform OAR deformations.[15]

In our comparison with HC1 and HC2, we simulated the process of phantom selection in state-of-the-art phantom libraries using our own database of CT scans and organ segmentations. For HC2, we actually did further investigate the use of the library of the University of Florida/National Cancer Institute, which is of public access. We, however, found that selecting those phantoms leads to no better results than using our database (the body segmentation was not considered because a region of interest between T10 and S1 is not readily available for those phantoms). In particular, it was found to be significantly better for half of the 3-D metrics (LR and S for the liver, LR and IS for the spleen), but worse for the other half. We also found the use of the actual library to be always significantly worse than using GP-GOMEA, with exception for LR of the spleen, where it was equivalent. This may be because the statistics on which those phantoms are based, which come from the United States, do not accurately represent the patients of our cohort, who are Dutch.

In future work, we will include more OARs, as well as consider porting our approach to different types of regions of interest (e.g., head for brain tumors) and cohorts (e.g., older patients). Ultimately, we are interested in generating an entire body anatomy for any patient.

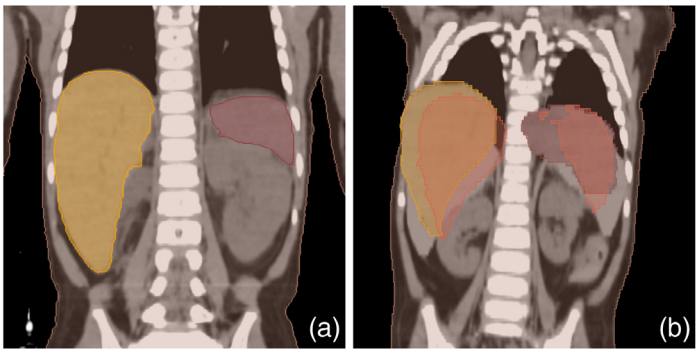

Fig. 8 Example of the limitations of anatomical inconsistency correction when applied to a coarse prediction of the body segmentation. Colors as in Fig. 7. (a) Actual anatomy of the patient and (b) proposed phantom, including corrections. The receiver CT is quite smaller than the actual CT in IS dimension. The anatomical inconsistency correction shrunk and relocated the liver that was exceeding the body boundaries and moved the spleen away from the spinal cord. However, liver and spleen are placed too high with respect to the underlying anatomy for it to be realistic.

We believe our approach is very promising because, once appropriate features are defined, no modifications in how to train ML algorithms, nor in how the pipeline works, are needed to obtain new phantoms. Moreover, since the availability of different OAR segmentations is currently limited to what is available in the database, and since delineating new OARs requires specialization, experience, and time, ways to use ML to deform an existing OAR template into a patient-specific segmentation could be worth investigating.[19] Furthermore, it will be interesting to investigate whether the single CT selection approach (sCT) can be improved, by understanding how to best combine more metrics than organ positions. With more data available, the use of deep convolutional neural networks could be explored[56] to replace the use of features extracted from the radiographs with the use of the radiographs themselves, potentially by proposing 2-D OAR segmentations that can be used to reconstruct 3-D versions.[60,61] Last but not least, in addition to organ dose reconstruction for historical radiation treatment, other uses of our approach could potentially be identified in the field of radiology: for example, phantoms could be built to estimate 3-D anatomical information of a patient from 2-D scout images, to minimize radiation exposure.[62–64]

6 Conclusion

We have presented a new take on phantom construction that leverages ML to assemble existing 3-D patient imaging into a new anatomy. Contrary to existing approaches, the pipeline we propose requires no manual intervention except for the initial effort of assembling a database of 3-D patient imaging (CTs, segmentations, and patient features), and the measurement of few features of the historical patients on their radiographs. With our approach, the problem of finding a globally good metric to represent anatomical categorization, typically faced by phantom libraries, is shifted to train ML models for parts of phantoms based on specific 3-D metrics. Our experimental results on a database of 60 pediatric cancer patients, focused on liver and spleen, showed that this approach can lead to significantly better anatomical resemblance compared to the use of phantom building criteria that are currently common practice. Positive results were still found after applying a simple but automatic anatomical inconsistency correction method to enhance realism. Regarding the ML algorithm used in the pipeline, we found that GP-GOMEA, a state-of-the-art genetic programming approach, can deliver models that are both accurate and readable. This aspect can be of added value as such models increase the chances of clinicians understanding them better and trusting their use.

7 Appendix A: Hyperparameters and Tuning

Table 4 shows the hyperparameters used by the ML algorithms. For RidReg and LASSO, we optimize λ to penalize complex models. Note that both algorithms perform squared-error minimization (along with penalization handling) via cyclical coordinate descent.[47] However, the hyperparameter tuning grid-search cross-validation procedure we use is aligned to MAE minimization, and this procedure is wrapped around the squared residual-based optimization.

For RF, we use the default, relatively large number of trees (given the datasets at hand) of 500 as advised by the literature.[65] We optimize how many features are randomly chosen when building the nodes that compose each decision tree ("mtry") and the minimum number of data samples a node should represent ("min. node size"), the same way we optimize λ for RidReg and LASSO.

For GP-Trad and GP-GOMEA, the function set $\mathcal{F}$ defines which functions to use as model components (tree nodes). The division operator $\div_A$ is the analytic quotient,[66] which not only guarantees that the divider can never be null, but also ensures smoothness (in contrast with the protected division operator,[41] which can harm generalization[66]). The logarithm operator is protected to avoid infeasible computations.[40] The ephemeral random constants are constants for which the value is set by uniform sampling from a defined interval.[41] Mathematical expressions are encoded as parse trees in GP-Trad and GP-GOMEA. We set a small tree height to keep the resulting mathematical expressions short and readable (we found that the larger three heights can result in hard to read expressions) and to prevent overfitting.

Table 4 Hyperparameters of the ML algorithms. The subscript "tune" means that the hyperparameter setting is subject to optimization with fivefold cross-validation grid-search among the listed values.

Algorithm	Hyperparameter	Settings
RidReg and LASSO	λ_{tune}	$10^{-10}, 10^{-9}, \ldots, 10^{10}$
RF	nr. trees	500
	min. node size$_{\text{tune}}$	$5, 10, \ldots, 20, 25$
	mtry$_{\text{tune}}$	$1, 2, \ldots, \frac{\#\text{features}}{2}$
GP-Trad and GP-GOMEA	$\mathcal{F}$	$\{+, -, \times, \div_{\text{A}}, \exp, \log_{\text{P}}\}$
	ERC	$\mathbb{U}[-10, 10]$
	Tree height	2
	g_{IMS}	4
	Time limit	60 s

The number of candidate expressions to evolve, i.e., the population size, is a sensitive parameter for GP algorithms. We run GP-Trad and GP-GOMEA using the interleaved multistart scheme (IMS), a method that interleaves multiple runs with increasing population size. We set the number of subiterations between runs g_{IMS} to 4 as has been reported to work well on benchmark problems.[42,43] Since the IMS can in principle run forever, we set a time limit of 60 s. We found this limit to be reasonable because the datasets are small and evaluations are fast and because the other ML algorithms take only a few seconds to execute. We also preliminarily observed that increasing the time limit (e.g., 5 or 10 min) does not alter the results in a significant way. For further details on GP-Trad, GP-GOMEA, the IMS, and other hyperparameters, the reader is referred to the seminal paper on GP-GOMEA for regression.[43]

For RidReg, LASSO, and RF, we perform grid-search hyperparameter tuning with fivefold cross validation upon the training data, to determine the best hyperparameter values. We use the R package caret for this purpose,[67] focused on minimizing the MAE. Once the best hyperparameter settings are found, we train the ML algorithm on the training set using those settings and test it on the test set. For GP-Trad and GP-GOMEA, we take the best expression found by the interleaved runs started by the IMS.

8 Appendix B: ML Performance for Increasing Database Size

Perhaps the biggest limitation of our study is the limited size of the database, consisting of only 60 patients. To put this number in perspective and understand whether increasing the database size will likely improve ML performance, we simulated the effect of having smaller (training) databases.

Figure 9 shows the effect of using GP-GOMEA (the overall best performing algorithm) when learning ML models of the liver position using different training set sizes in a traditional fivefold cross validation. The test set size consists of $\frac{1}{5}$th of the total number of patients ($60/5 = 12$ patients), whereas the maximum number of patients to be considered at training time is varied between $\frac{1}{5}$th (12 patients) and $\frac{4}{5}$th (48 patients). Patients are picked at random (taking care that patients in the test set never appear in the training set). This is repeated 10 times for each size to account for stochasticity.

Mean (of the 10 repetitions) test MAEs behave as can be expected: the larger the training set size is, the lower the test MAE is. The fact that means training MAEs increases with the number of patients to be considered is due to the fact that smaller sets are easier to (over)fit.

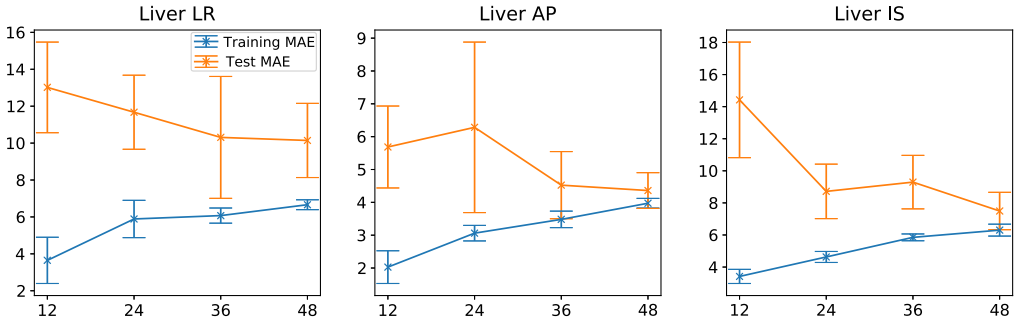

Fig. 9 Effect of increasing the database size on learning the liver position using GP-GOMEA. Center points represent the mean, and error bars represent the standard deviation, of the MAEs obtained from an average (across 10 repetitions) fivefold cross validation.

Note that the errors obtained in Table 2 are substantially different (smaller and with larger standard deviation) than the ones obtained here. This is because leave-one-out cross validation was applied in that case, which is precisely expected to be a better representative of the true generalization error but typically exhibits larger variations.[68]

Overall, these results allow us to conclude that the inclusion of more patients will most likely result in further improving the (test) performance of ML. Moreover, the more patients will be included, the more it will be possible to enlarge the set of features to consider: this will enable to better fit both the training set (and reduce the training MAE) and the test set (and further reduce the test MAE).

9 Appendix C: Examples of Models Found by GP-GOMEA

Examples of the mathematical expression models found by GP-GOMEA for the automatic construction of pediatric abdominal phantoms are reported below. Each model pertains to a particular metric and has been manually rearranged to aid readability. The features and the target metrics are normalized by z scoring. To apply the model in practice, each feature needs to be denormalized, i.e., scaled by the standard deviation and translated by the mean. Furthermore, the output of the model needs to be denormalized as well, i.e., scaled and translated by the standard deviation and the mean of the target variable, respectively (this information is available in Table 1). We do not include these denormalization coefficients in the models for the sake of readability. Also since normalized features are approximately in the same scale, the magnitude of the coefficients included in the model can be associated with feature relevance.

Body S:

$$0.420 \times (\mathrm{ADAP} + \mathrm{ADLR} + \mathrm{SCIS}).$$

Liver LR:

$$(\mathrm{GEND} + \mathrm{ADAP}) \frac{\sqrt{0.100 + \mathrm{RDLR}^2}}{8.264 + 0.100 \times \mathrm{RDLR}^2}.$$

Liver AP:

$$\frac{0.169 \times \mathrm{ADAP}}{\sqrt{0.100 + (\mathrm{AGE} - \mathrm{HEIG})^2}}$$

Liver IS:

$$0.179 \times \mathrm{ADLR} + \mathrm{GEND} \times \mathrm{RDIS}$$

Liver S:

$$\frac{\text{ADLR} + \text{SCIS}}{\sqrt{0.100 + (4.389 + \text{LDLR})^2}}.$$

Spleen LR:

$$\frac{-0.821 \times \text{LDLR}}{\sqrt{\text{LDLR}^2 + 0.100 \times \text{ICSC}^2 + 0.010}}.$$

Spleen AP:

$$\frac{-0.285 \times \text{ICSC}}{\sqrt{0.100 + (\text{AGE} + \text{ADAP})^2}}.$$

Spleen IS:

$$\frac{\text{AGE}}{\sqrt{(0.100 + \text{HEIG}^2) \times (9.673 - 6.188 \times \text{WEIG} + \text{WEIG}^2)}}.$$

Spleen S:

$$2.718^{\text{AGE}} \times 0.057 \times \text{SCIS}.$$

As mentioned in Sec. 3.3, the model for body S is linear. The other models found by GP-GOMEA are nonlinear. The recurrence of square roots and squaring terms is due to the use of the analytic quotient, which prevents dividing by 0. For the sake of intuition, $\frac{x}{\sqrt{0.1+y^2}}$ can be considered as approximately $\frac{x}{|y|}$ ($|\cdot|$ takes the absolute value).

Let us begin by considering the models that predict the LR position of the liver and the spleen. The model for liver LR will always return a positive number, whereas the one for the spleen will always return a negative number. This is reasonable because the center of mass of the liver is normally on the right of the reference point we used to determine the organ position, i.e., the second lumbar vertebra and the opposite holds for the spleen. Moreover, for the prediction of the liver position along LR, the model combines the AP direction, by means of the abdominal diameter in AP (ADAP), and the LR direction, by means of the right diaphragm length along LR (RDLR). Note that since GEND can either have value 0 (female) or 1 (male), the left multiplicand term is bigger for males than females. Thus for males, the model predicts bigger shifts in LR. Differently from the liver, the LR position of the spleen relies on the LR size of the left diaphragm rather than on the right diaphragm. This aspect reflects the fact that the liver and spleen are, respectively, placed below the right and the left diaphragm.

Disclosures

Tanja Alderliesten and Peter A. N. Bosman are involved in projects supported by Elekta (Elekta AB, Stockholm, Sweden). Elekta had no involvement in the study design, data collection, analysis and interpretation, and writing of this article.

Acknowledgments

The authors acknowledge the Kinderen Kankervrij Foundation for financial support (Project #187) and the Maurits and Anna de Kock Foundation for financing a high-performance computing system. We thank Dr. Brian V. Balgobind, Dr. Irma W. E. M. van Dijk, and Dr. Jan Wiersma from the Department of Radiation Oncology of Amsterdam UMC, location AMC, Amsterdam, the Netherlands, and Dr. Geert O. R. Janssens and Dr. Petra Kroon from the Department of Radiation Oncology of UMC Utrecht Cancer Center, Utrecht, the Netherlands, for providing help in the collection and/or in the assessment of the imaging data used in this work. The authors

are grateful to Elekta for providing ADMIRE research software for automatic organ segmentation. We further acknowledge Dr. Choonsik Lee from the National Cancer Institute, Division of Cancer Epidemiology and Genetics, Rockville, Maryland, USA, for details on the phantom library of the University of Florida/National Cancer Institute. This article is based upon and extends (by introducing the anatomical inconsistency correction method and related results), an SPIE proceedings paper of which an abstract has been recently submitted for consideration to the SPIE conference on Medical Imaging (2020), titled "*Machine Learning for Automatic Construction of Pediatric Abdominal Phantoms for Radiation Dose Reconstruction*," authored by the same authors of this article.

References

1. C. Lee et al., "Reconstruction of organ dose for external radiotherapy patients in retrospective epidemiologic studies," *Phys. Med. Biol.* **60**(6), 2309–2324 (2015).
2. X. G. Xu, "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 (2014).
3. M. Virgolin et al., "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 (2018).
4. Z. Wang et al., "Are age and gender suitable matching criteria in organ dose reconstruction using surrogate childhood cancer patients' CT scans?" *Med. Phys.* **45**(6), 2628–2638 (2018).
5. A. C. Paulino et al., "Late effects in children treated with radiation therapy for Wilms' tumor," *Int. J. Radiat. Oncol. Biol. Phys.* **46**(5), 1239–1246 (2000).
6. I. W. E. M. van Dijk et al., "Evaluation of late adverse events in long-term Wilms' tumor survivors," *Int. J. Radiat. Oncol. Biol. Phys.* **78**(2), 370–378 (2010).
7. Y. T. Cheung et al., "Chronic health conditions and neurocognitive function in aging survivors of childhood cancer: a report from the childhood cancer survivor study," *J. Natl. Cancer Inst.* **110**(4), 411–419 (2017).
8. R. M. Howell et al., "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 (2019).
9. M. Stovall et al., "Genetic effects of radiotherapy for childhood cancer: gonadal dose reconstruction," *Int. J. Radiat. Oncol. Biol. Phys.* **60**(2), 542–552 (2004).
10. A. M. Geyer et al., "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 (2014).
11. I. Alziar et al., "Individual radiation therapy patient whole-body phantoms for peripheral dose evaluations: method and specific software," *Phys. Med. Biol.* **54**(17), N375–N383 (2009).
12. T. Xie, N. Kuster, and H. Zaidi, "Computational hybrid anthropometric paediatric phantom library for internal radiation dosimetry," *Phys. Med. Biol.* **62**(8), 3263–3283 (2017).
13. W. P. Segars et al., "Development and application of the new dynamic NURBS-based Cardiac-Torso (NCAT) phantom," *J. Nucl. Med.* **42**(5), 23 (2001).
14. W. P. Segars et al., "The development of a population of 4D pediatric XCAT phantoms for imaging research and optimization," *Med. Phys.* **42**(8), 4719–4726 (2015).
15. D. J. Tward et al., "Generating patient-specific dosimetry phantoms with whole-body diffeomorphic image registration," in *IEEE 37th Annu. Northeast Bioeng. Conf. (NEBEC)*, IEEE, pp. 1–2 (2011).
16. J. Valentin, "Basic anatomical and physiological data for use in radiological protection: reference values: ICRP publication 89," *Ann. ICRP* **32**(3–4), 1–277 (2002).
17. G. L. de la Grandmaison, I. Clairand, and M. Durigon, "Organ weight in 684 adult autopsies: new tables for a caucasoid population," *Forensic Sci. Int.* **119**(2), 149–154 (2001).
18. J. V. Bezin et al., "A review of uncertainties in radiotherapy dose reconstruction and their impacts on dose–response relationships," *J. Radiol. Prot.* **37**(1), R1 (2017).

19. A. Ng et al., "Reconstruction of 3D lung models from 2D planning data sets for Hodgkin's lymphoma patients using combined deformable image registration and navigator channels," *Med. Phys.* **37**(3), 1017–1028 (2010).
20. A. Ng et al., "Navigator channel adaptation to reconstruct three dimensional heart volumes from two dimensional radiotherapy planning data," *BMC Med. Phys.* **12**(1) (2012).
21. S. Lamart et al., "Prediction of the location and size of the stomach using patient characteristics for retrospective radiation dose estimation following radiotherapy," *Phys. Med. Biol.* **58**(24), 8739–8753 (2013).
22. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, Berlin, Heidelberg (2006).
23. Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *New Engl. J. Med.* **375**(13), 1216–1219 (2016).
24. N. Breslow et al., "Epidemiology of Wilms tumor," *Med. Pediatric Oncol.* **21**(3), 172–181 (1993).
25. A. Fedorov et al., "3D Slicer as an image computing platform for the quantitative imaging network," *Magn. Reson. Imaging* **30**(9), 1323–1341 (2012).
26. S. C. Huijskens et al., "Magnitude and variability of respiratory-induced diaphragm motion in children during image-guided radiotherapy," *Radiother. Oncol.* **123**(2), 263–269 (2017).
27. M. Xi et al., "Analysis of abdominal organ motion using four-dimensional CT," *Chin. J. Cancer* **28**(9), 989–993 (2009).
28. A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.* **38**(12), 2270–2285 (2005).
29. L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology* **26**(3), 297–302 (1945).
30. K. H. Zou et al., "Statistical validation of image segmentation quality based on a spatial overlap index: scientific reports," *Acad. Radiol.* **11**(2), 178–189 (2004).
31. D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(9), 850–863 (1993).
32. S. Nikolov et al., "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," https://lmb.informatik.uni-freiburg.de/Publications/2018/Ron18a/ (2018).
33. T. Heimann et al., "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imaging* **28**(8), 1251–1265 (2009).
34. L. Ruskó, G. Bekes, and M. Fidrich, "Automatic segmentation of the liver from multi- and single-phase contrast-enhanced CT images," *Med. Image Anal.* **13**(6), 871–882 (2009).
35. X. Chen et al., "Medical image segmentation by combining graph cuts and oriented active appearance models," *IEEE Trans. Image Process.* **21**(4), 2035–2046 (2012).
36. A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, V. H. Winston & Sons, Washington (1977).
37. A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics* **12**(1), 55–67 (1970).
38. R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Stat. Soc. (Ser. B)* **58**, 267–288 (1996).
39. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
40. J. R. Koza, "Genetic programming as a means for programming computers by natural selection," *Stat. Comput.* **4**(2), 87–112 (1994).
41. R. Poli, W. B. Langdon, and N. F. McPhee, *A Field Guide to Genetic Programming*, Lulu Enterprises, UK Ltd. (2008).
42. M. Virgolin et al., "Scalable genetic programming by gene-pool optimal mixing and input-space entropy-based building-block learning," in *Proc. Genet. and Evol. Comput. Conf.*, ACM, pp. 1041–1048 (2017).
43. M. Virgolin et al., "Improving model-based genetic programming for symbolic regression of small expressions," *Evol. Comput.* 1–27 (2020).
44. R. Guidotti et al., "A survey of methods for explaining black box models," *ACM Comput. Surv.* **51**, 93:1–93:42 (2018).

45. A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE Access* **6**, 52138–52160 (2018).
46. R. C. Team, "R: a language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria (2013).
47. J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.* **33**(1), 1–22 (2010).
48. M. N. Wright and A. Ziegler, "Ranger: a fast implementation of random forests for high dimensional data in C++ and R," *J. Stat. Softw.* **77** (1), 1–17 (2017).
49. P. A. N. Bosman, J. Grahl, and D. Thierens, "Enhancing the performance of maximum-likelihood Gaussian EDAs using anticipated mean shift," in *Int. Conf. Parallel Prob. Solving Nat.*, Springer, pp. 133–143 (2008).
50. P. A. N. Bosman, J. Grahl, and D. Thierens, "Benchmarking parameter-free AMALGAM on functions with and without noise," *Evol. Comput.* **21**(3), 445–469 (2013).
51. A. N. I. Badouna et al., "Total heart volume as a function of clinical and anthropometric parameters in a population of external beam radiation therapy patients," *Phys. Med. Biol.* **57**(2), 473–484 (2012).
52. E. J. Stepusin et al., "Assessment of different patient-to-phantom matching criteria applied in Monte Carlo-based computed tomography dosimetry," *Med. Phys.* **44**(10), 5498–5508 (2017).
53. Z. Wang et al., "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 (2019).
54. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.* **7**, 1–30 (2006).
55. J. M. Bland and D. G. Altman, "Multiple significance tests: the Bonferroni method," *BMJ* **310**(6973), 170 (1995).
56. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**(7553), 436 (2015).
57. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.* **29**(5), 1189–1232 (2001).
58. C. Pinter et al., "SlicerRT: radiation therapy research toolkit for 3D Slicer," *Med. Phys.* **39**(10), 6332–6338 (2012).
59. G. Litjens et al., "Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge," *Med. Image Anal.* **18**(2), 359–373 (2014).
60. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
61. Y. Wang, Z. Zhong, and J. Hua, "DeepOrganNet: on-the-fly reconstruction and visualization of 3D/4D lung models from single-view projections by deep deformation network," *IEEE Trans. Vis. Comput. Graphics* **26**(1), 960–970 (2019).
62. O. Christianson et al., "Automated size-specific CT dose monitoring program: assessing variability in CT dose," *Med. Phys.* **39**(11), 7131–7139 (2012).
63. J. Hoye et al., "Organ doses from CT localizer radiographs: development, validation, and application of a Monte Carlo estimation technique," *Med. Phys.* **46**(11), 5262–5272 (2019).
64. S. L. Cohen, T. J. Ward, and M. D. Cham, "The relationship between CT scout landmarks and lung boundaries on chest CT: guidelines for minimizing excess z-axis scan length," *Eur. Radiol.* **30**(1), 581–587 (2020).
65. P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest," *J. Mach. Learn. Res.* **18**(1), 6673–6690 (2017).
66. J. Ni, R. H. Drieberg, and P. I. Rockett, "The use of an analytic quotient operator in genetic programming," *IEEE Trans. Evol. Comput.* **17**(1), 146–152 (2013).
67. M. Kuhn, "The caret package," *J. Stat. Software* **28**(5), 1–26 (2008).
68. J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, Vol. **1**, Springer, New York (2001).

Marco Virgolin is a postdoctoral fellow at Chalmers University of Technology, Sweden. He conducted this study while being appointed at Centrum Wiskunde and Informatica, Amsterdam, the Netherlands. He is mostly interested in evolutionary and explainable machine learning, with a special focus on genetic programming and symbolic regression.

Ziyuan Wang is working as a PhD student in the Radiation Oncology Department of Amsterdam UMC at the University of Amsterdam. Her research interests include medical imaging analysis and physics of radiation treatment. She has a background in applied physics and is currently working on the project "3D dose reconstruction for children with long-term follow-up: toward improved decision making in radiation treatment for children with cancer."

Tanja Alderliesten received her PhD in computer science in 2004 from Utrecht University, the Netherlands. Currently, she is an associate professor in the Department of Radiation Oncology, Leiden University Medical Center, the Netherlands. The focus of her research is translational in nature and primarily concerns the development of state-of-the-art methods and techniques from the fields of mathematics and computer science (including image processing, biomechanical modeling, and optimization) for radiation oncology.

Peter A. N. Bosman is a senior researcher in the Life Sciences and Health (LSH) Research Group at the Dutch National Research Institute for Mathematics and Computer Science (Centrum Wiskunde and Informatica) and a professor of evolutionary algorithms at Delft University of Technology. His research concerns the design of scalable model-based EAs and their application, primarily in the LSH domain. He has (co-)authored more than 100 refereed publications, out of which 4 received best paper awards.