On the feasibility of automatically selecting similar patients in highly individualized radiotherapy dose reconstruction for historic data of pediatric cancer survivors

Marco Virgolin[a)]
Centrum Wiskunde & Informatica, Amsterdam 1098XG, the Netherlands

Irma W. E. M. van Dijk and Jan Wiersma
Department of Radiation Oncology, Academic Medical Center, Amsterdam 1100DD, the Netherlands

Cécile M. Ronckers
Emma Children's Hospital/Academic Medical Center, Amsterdam 1100DD, the Netherlands

Cees Witteveen
Department of Software Technology, Technical University of Delft, Delft 2628CD, the Netherlands

Arjan Bel and Tanja Alderliesten
Department of Radiation Oncology, Academic Medical Center, Amsterdam 1100DD, the Netherlands

Peter A. N. Bosman
Centrum Wiskunde & Informatica, Amsterdam 1098XG, the Netherlands

(Received 4 July 2017; revised 5 January 2018; accepted for publication 22 January 2018; published xx xxxx xxxx)

Purpose: The aim of this study is to establish the first step toward a novel and highly individualized three-dimensional (3D) dose distribution reconstruction method, based on CT scans and organ delineations of recently treated patients. Specifically, the feasibility of automatically selecting the CT scan of a recently treated childhood cancer patient who is similar to a given historically treated child who suffered from Wilms' tumor is assessed.

Methods: A cohort of 37 recently treated children between 2- and 6-yr old are considered. Five potential notions of ground-truth similarity are proposed, each focusing on different anatomical aspects. These notions are automatically computed from CT scans of the abdomen and 3D organ delineations (liver, spleen, spinal cord, external body contour). The first is based on deformable image registration, the second on the Dice similarity coefficient, the third on the Hausdorff distance, the fourth on pairwise organ distances, and the last is computed by means of the overlap volume histogram. The relationship between typically available features of historically treated patients and the proposed ground-truth notions of similarity is studied by adopting state-of-the-art machine learning techniques, including random forest. Also, the feasibility of automatically selecting the most similar patient is assessed by comparing ground-truth rankings of similarity with predicted rankings.

Results: Similarities (mainly) based on the external abdomen shape and on the pairwise organ distances are highly correlated (Pearson $r_p \geq 0.70$) and are successfully modeled with random forests based on historically recorded features (pseudo-$R^2 \geq 0.69$). In contrast, similarities based on the shape of internal organs cannot be modeled. For the similarities that random forest can reliably model, an estimation of feature relevance indicates that abdominal diameters and weight are the most important. Experiments on automatically selecting similar patients lead to coarse, yet quite robust results: the most similar patient is retrieved only 22% of the times, however, the error in worst-case scenarios is limited, with the fourth most similar patient being retrieved.

Conclusions: Results demonstrate that automatically selecting similar patients is feasible when focusing on the shape of the external abdomen and on the position of internal organs. Moreover, whereas the common practice in phantom-based dose reconstruction is to select a representative phantom using age, height, and weight as discriminant factors for any treatment scenario, our analysis on abdominal tumor treatment for children shows that the most relevant features are weight and the anterior–posterior and left–right abdominal diameters. [https://doi.org/10.1002/mp.12802]

Key words: deformable image registration, dose reconstruction, late adverse effects, machine learning, pediatric cancer

1. INTRODUCTION

Every day, radiation oncologists working on the treatment of childhood cancer patients are faced with the challenge of designing individualized treatment plans which ensure that a sufficiently high dose is delivered to the tumor while the surrounding healthy organs are spared. An excessive exposure of sensitive tissues to radiation may compromise crucial physiological functions and lead to severe health complications. Although the absolute number of young patients undergoing radiation treatment is moderate,[1] the presence of malignancy in their developing bodies is likely to impact their entire life both physically and psychologically.[2–7] Moreover, children arguably are the most susceptible to adverse effects of radiation treatment and thus stand to benefit most from improvements in planning under the desired hypothesis of long-term survival, which is currently achieved for the treatment of Wilms' tumor, or nephroblastoma, the most common childhood abdominal malignancy.[8,9]

For adult patients, many follow-up studies exist where the relationship between radiation treatment with specific dose (fractions) and onset of adverse effects is analyzed (see, e.g., the work of QUANTEC[10]). Furthermore, detailed work has been done to understand which specific organ subvolumes are most sensitive to ionizing radiation, by observing fine-grained 3D dose distributions.[11–13] For children, however, the evidence collected so far is limited.[14] Incorporating detailed knowledge on the relationship between 3D dose distributions and late adverse effects may greatly improve the design of treatment plans, ultimately reducing posttreatment complications and improving pediatric cancer survivors' quality of life.

Currently, researchers willing to study possible relationships between detailed 3D dose distributions and the onset of *late* adverse effects in *long-term* pediatric cancer survivors face a major obstacle: the lack of 3D treatment data. In fact, 3D information about the anatomy of historically treated patients could not be acquired before the advent of computed tomography (CT) and 3D treatment planning. The available information consists of patient characteristics recorded in historical patient records, notes on the treatment, and, in some cases, two-dimensional (2D) simulator films used for planning at the time. The available data on late adverse effects collected from long-term follow-ups cannot be exploited to its full potential, as it cannot be related to fine-grained 3D dose information. Therefore, to enable accurate dose-risk modeling in retrospective studies, a method to accurately reconstruct 3D dose distributions is needed.

The current state-of-the-art method to bridge this gap is the so-called *phantom-based dose reconstruction.*[15,16] Phantoms are 3D representations of human bodies, constructed according to reference guidelines (e.g., ICRP 89[17]), stored in libraries in a gender-, age-, height-, and weight-dependent fashion. The doses delivered to the organs of a historically treated patient are estimated by simulating the original treatment on a phantom. The dose reconstruction procedure can be summarized in four fundamental steps: (a) *Selection* — a phantom from the library is chosen, which most closely resembles a patient's available features (this is typically done using gender, age, height, and weight); (b) *Adaptation* — the phantom is adapted (i.e., shrunken, stretched) according to other specific features, such as measurements from a 2D simulator film; (c) *Treatment simulation* — the original treatment is simulated, using the phantom's virtual anatomy as a surrogate for the original body; and (d) *Measurement* — parameters about the dose are measured. Clearly, the accuracy of the estimated dose relies on the completeness of the historical patient record considered, on the representativeness of the phantom library, and on the quality of each one of the four aforementioned steps. A poor selection based on irrelevant features, as well as an ineffective adaptation, may compromise the accuracy of dose estimation. For children, growth and development do not follow a standard age-related pattern; thus, selecting a representative phantom is especially difficult. Although rich libraries exist with reference phantoms for many height–weight combinations,[16] and more and more possibilities to adapt mesh-based models to improve patient individualization are under investigation,[18] an inherent limitation of phantom-based dose reconstruction is that it relies on *average* organ shapes and dimensions. Studies have however shown that there can be a great variability in internal organ shape among individuals with a similar body mass index and that it is practically impossible to establish reference organ anatomy.[16,19]

Recently, an alternative to phantom-based dose reconstruction has been proposed, based on the reconstruction of 3D organs for historically treated patients, using navigator channels and finite element modeling deformable image registration.[20–22] The feasibility of the method has been tested on 3D dose reconstruction for lungs, heart, and breasts of adult Hodgkin lymphoma patients. Relying on CT scans of recently treated patients, a deformation model was built which uses information from 2D simulator films (or digitally reconstructed radiographs) to synthesize the organs of a historically treated patient by deforming the organs of a recently treated patient. A primitive *selection* step is performed to match the historically treated patient to a recent representative patient. This selection is based on 2D thorax measurements and gender, but it is advised that taking a smarter approach, possibly relying on more or different features, may improve the overall outcome. Still, the resulting dose reconstruction method was found to be clinically acceptable (median mean dose difference $\leq$1 Gy and median V_5 and V_{20} differences $\leq$2%)[22] and has recently been adopted in practice.[23]

In this study, we present a novel approach to select a recently treated patient for whom a CT scan is available to match a historically treated patient, and apply it to the scenario in which the 3D dose distribution of historically treated children with Wilms' tumor needs to be reconstructed. In order to find the features that are important to select a recently treated patient that resembles the 3D anatomy of a historically treated patient, a ground-truth notion of *similarity* among patients is needed. To this end, we propose and analyze five different notions of similarity; each focused on

specific anatomical aspects. We choose to focus on anatomical similarity, rather than directly on similarity in 3D dose distributions, because the latter needs a specific treatment to be defined beforehand, whereas the former enables the reconstruction of different treatments on a region of interest. The notions here proposed can be computed in an automatic and reproducible way, starting from CT scans and 3D organ delineations. We compute the five similarities on a cohort of pediatric patients and study correlations among them. We then assess which of the features that are typically available from historical patients' records are the most relevant to explain the similarities. Consequently, a state-of-the-art machine learning model, *random forest*, is trained using the most relevant features and its performance is measured in terms of correctly predicting rankings of similar patients, which ultimately is the goal. Finally, we provide a case of comparison between dose reconstruction based on the most similar patient according to one of the proposed similarities, and based on the most similar patient according to age, height, and weight.

2. MATERIALS AND METHODS

2.A. Patient data

The records and CT data of 37 children were included (17 males, 20 females) in the age range of 2–6 yr. Most of the patients suffered from Wilms' tumor (22); many underwent (partial) nephrectomy (21). All patients received chemotherapy prior to radiation treatment. The patients have been treated at the Academic Medical Center/Emma Childrens Hospital in Amsterdam (34) or at the University Medical Center Utrecht/Princess Máxima Center for Pediatric Oncology in Utrecht (3), all after January 2000. A CT scan in supine position of the abdomen, from the top of the 10th thoracic vertebra (T10) to the bottom of first sacral vertebra (S1), is available for each patient. The median voxel size is 0.977 mm along left–right (LR) and anterior–posterior (AP) directions and 2.5 mm along the superior–inferior (SI) direction (slice thickness). Also, delineations of the liver and the spleen and of the spinal cord and the outer body contour in the common region of interest from T10 to S1 are included. Delineations of kidneys are not considered due to (partial) nephrectomy (see Section 2.B.2), with exception for the right kidney of three patients, used for an example of dose reconstruction (see Section 2.E).

Features that are typically reported in historical treatment records have been gathered for our cohort, together with measurements from digitally reconstructed radiographs generated from the CT scans, under the reasonable assumption that the old 2D simulator films have been preserved. These features are reported in Tables I and II. Details are as follows. Age was recorded at CT acquisition; height, weight, and body mass index were recorded at intake in the radiotherapy department, which happened up to 3 months before CT acquisition. The distance between iliac crest and spinal cord is defined as the distance between the top point of the left iliac crest and the center of the spinal cord, along the line passing through both iliac crest top points. The LR diameter has been measured at the center of second lumbar. Historically, the AP diameter was measured at the isocenter. Because of the high conformity of the treatment for renal fossa irradiation, such isocenter would typically be located in an SI section within the top of first lumbar and the bottom of second lumbar. After inspection of historical treatments, an *average isocenter* has been set at the intersection of an SI line crossing the renal fossa with an LR line crossing the intervertebral disc between first and second lumbar. For our (recently treated) patients, we measured the AP diameter at isocenter on their CT scans. Assuming symmetry of the abdomen, the average isocenter is set either in the left or right renal fossa, according to ease in carrying out the measurement for each patient. We observed on a random sample of ten patients that the difference between the AP diameter measured on the average isocenter in the left renal fossa and the one in the right renal fossa is below 1 cm. For one patient, the height was missing from clinical records, so an age- and gender-matched estimate from the Dutch children growth chart of 2010 has been used.

2.B. Similarity notions

In the following, we present five different notions of anatomical similarity and how they can be computed from CT scans and organ delineations. We further describe how correlation among similarities is measured.

TABLE I. Numerical patient features.

Numerical feature (abbreviation)	Unit of measure	Min.	Max.	Median	Mean	St. Dev.
Age	yr	2.21	5.84	3.87	3.94	1.08
Height	cm	89.00	123.00	103.50	104.28	9.58
Weight	kg	10.00	28.00	16.55	16.88	3.75
Body mass index (BMI)	kg/m^2	10.90	18.50	15.36	15.40	1.81
Length spinal cord (T12-L4)	cm	7.00	10.90	9.30	9.33	0.89
Distance iliac crest–spinal cord (IC-SC)	cm	4.30	6.80	5.70	5.58	0.54
LR diameter at L2 (diam. LR)	cm	16.30	23.50	19.30	19.48	1.28
AP diameter at isocenter (diam. AP)	cm	11.30	16.00	13.20	13.37	1.53

TABLE II. Categorical patient features.

Categorical feature (abbreviation)	Categories (# patients)
Gender	Female (20), Male (17)
Diagnosis	Ependymoma (1), Medulloblastoma (2), Neuroblastoma (9), Rhabdomyosarcoma (3), Wilms' tumor (22)
Tumor Site	Ductus choledochus (1), Fourth ventricle (3), Left kidney (12), Left suprarenal gland (6), Pelvic region (1), Retroperitoneum (1), Right kidney (10), Right lower abdomen (1), Right suprarenal gland (2)
Partial Nephrectomy (part. Nephr)	Left (2), Right (1), None (34)
Radical Nephrectomy	Left (11), Right (10), None (16)

2.B.1. Deformation-based similarity

This first notion of similarity is based on the amount of deformation that is needed to register one CT to another, via intensity-based deformable image registration. This metric is inspired by previous applications.[24]

To compute this similarity, first scans are manually aligned on bony anatomy, using S1, the 5th lumbar vertebra, and the iliac crests as reference. Second, for each possible pair of children, deformable image registration is performed to deform the first patient's CT to match the CT of the second patient and vice versa. This two-way registration is performed because of the asymmetry of most practical deformable image registration software. The software *elastix*[25,26] has been adopted, with mostly standard parameters settings (adaptive stochastic gradient descent optimization, Mattes' mutual information metric, multiresolution B-spline transformation). For the finest resolution step, a coarse grid size of 28 mm has been chosen following the guidelines for the deformation of large structures as found in the manual of elastix, combined with visual inspection of registration outcomes for several grid sizes. This choice limits the amount of unrealistic deformation on internal anatomy when registering the whole abdominal area (from T10 to S1) at once.

After computing the two registrations, a measure of deformation magnitude can be computed based on the deformations. A deformation is described using a meshed cube C, where each cell is the 3D offset to apply to a specific B-spline control point in order to register the first image to the second. The deformation magnitude we compute from C considers only "stretching" and "shrinking" effects, disregarding translations. Specifically, for each cell $c \in C$, the differences of the offset of c and the ones of each adjacent cell are summed, and the result is normalized by the number of adjacent cells. The values obtained this way are lastly summed together. Formally, the deformation magnitude is thus:

$$D = \sum_{c \in C} \frac{1}{|A_c|} \sum_{a \in A_c} \|\vec{o}_c - \vec{o}_a\|,$$

where A_c is the subcube of cells centered at cell c and $\vec{o}_c$ is the 3D vector of offsets stored in cell c.

The two deformation magnitudes D_1 and D_2 computed from the two registrations are then averaged. Finally, to obtain a measure of similarity, the average magnitude needs to be inverted. We denote this similarity with S_{deform}:

$$S_{\text{deform}} = \left(\frac{D_1 + D_2}{2}\right)^{-1}$$

2.B.2. Organ overlap-based similarity

This similarity notion focuses on internal organ overlap and is based on the well-known Dice similarity coefficient[27] (DSC) that indicates the overlap of two volumes V_1 and V_2.

We denote this measure for a specific organ delineation o by S^o_{DSC}, which is computed after aligning the images on their centers of mass. Thus, $S^o_{\text{DSC}} = 100\frac{2|V^o_1 \cap V^o_2|}{|V^o_1|+|V^o_2|}$. A measure of similarity S_{DSC} is then computed by combining S^o_{DSC} for the various organs of interest. This is done by taking the Euclidean norm of the vector of all S^o_{DSC} components:

$$S_{\text{DSC}} = \sqrt{\sum_{o \in O} \left(S^o_{\text{DSC}}\right)^2}.$$

For the set of organ delineations O, the liver, the spleen, the part of the spinal cord from T10 to S1, and the section of the external body contour within the field T10-S1 (arms excluded) are considered. Kidneys are not taken into consideration because 21 of 37 patients of our cohort have been subject to (partial and/or radical) nephrectomy. If a similar patient needs to be found for a treatment that includes a kidney either as target (ipsilateral) or as organ at risk (contralateral), then patients who (partially) miss this kidney should be considered completely dissimilar. If the kidney is not interesting for the reconstruction, a patient without (part of) the kidney may still be a good candidate. Therefore, for practical use, the outcome of a matching based on this similarity should be twofold: a similar patient who necessarily shares the kidney configuration with the historical one, and a similar patient who does not. To be able to use the whole cohort in the analysis, we consider the scenario where kidneys are not relevant for the reconstruction.

2.B.3. Organ shape-based similarity

Different from DSC, the Hausdorff distance is another recognized metric used to compare organ shapes which is focused on outlier points.[28,29] Given two meshed surfaces A and B, the *directed* Hausdorff distance from A to B is defined as the maximum of the minimal Euclidean distances from A to B's vertices, that is, $h(\text{A},\text{B}) = \max_{a \in \text{A}} \min_{b \in \text{B}} \|a - b\|$. The Hausdorff distance between A and B is H (A, B) = max{h (A, B), h(B, A)}.

Similar to S^o_{DSC}, the notation $S^o_{\text{Hausdorff}}$ is used to indicate the organ-specific Hausdorff similarity:

$$S^o_{\text{Hausdorff}} = (H^o)^{-1} \times 10^p,$$

with p such that all $S^o_{\text{Hausdorff}} \geq 1$ (this ensures that $(S^o_{\text{Hausdorff}})^2 \geq S^o_{\text{Hausdorff}}$). The aggregated $S_{\text{Hausdorff}}$ is then computed as the Euclidean norm of the vector of all $S^o_{\text{Hausdorff}}$ components, where the delineations of liver, spleen, spinal cord (T10-S1), and external body (T10-S1) are considered:

$$S_{\text{Hausdorff}} = \sqrt{\sum_{o \in O} (S^o_{\text{Hausdorff}})^2}.$$

2.B.4. Organ constellation-based similarity

This similarity is proposed for the first time here. It is specifically aimed at capturing the variation in organ positions. Specifically, given a patient p, two organ delineations o_i,o_j, and the respective centers of mass $c^{o_i}(p)$, $c^{o_j}(p)$, let $d^{o_i,o_j}(p) = \|c^{o_i}(p) - c^{o_j}(p)\|$ be the center-of-mass distance. Again, $o_i \in \text{O} = \{$liver, spleen, spinal cord (T10-S1), external body (T10-S1)$\}$. Let E^{o}_{i} be the set of four points that are the projections of o_i's center of mass on the external body along anterior, posterior, left, and right directions. Then, the organ constellation-based similarity S_{const} for patients p_1 and p_2 is computed as follows. A first component $D^{\text{org}-\text{org}}_{\text{const}}(p_1, p_2)$ is calculated which represents the difference in pairwise organ distances, as:

$$D^{\text{org}-\text{org}}_{\text{const}}(p_1, p_2) = \sum_{o_i, o_j \in O, i \neq j} (d^{o_i,o_j}(p_1) - d^{o_i,o_j}(p_2))^2.$$

A second component represents the difference in distances between organs and the delineation of the external body:

$$D^{\text{org}-\text{ext}}_{\text{const}}(p_1, p_2) = \sum_{o_i \in O} \sum_{e \in E^{o_i}} (d^{o_i,e}(p_1) - d^{o_i,e}(p_2))^2.$$

Finally, $S_{\text{const}}(p_1, p_2)$ is:

$$S_{\text{const}}(p_1, p_2) = 1/\sqrt{D^{\text{org}-\text{org}}_{\text{const}}(p_1, p_2) + D^{\text{org}-\text{ext}}_{\text{const}}(p_1, p_2)}.$$

2.B.5. Overlap volume histogram-based similarity

The recently introduced overlap volume histogram (OVH)[30,31] was specifically designed to describe the position and shape of organs at risk near the tumor. In its original formulation, the OVH of an organ is computed by measuring the tumor organ overlap at each step of a discrete expansion (or shrinkage) of the 3D tumor delineation, centered at the tumor.

Here, the OVH is used to describe the shape and displacement of all organs at the same time. To this end, an artificial OVH is adopted, built using a sphere with a starting radius of 1 mm that expands from the center of mass of the body contour section within the T10-S1 region of interest. At each iteration, the sphere radius is expanded by 2.5 mm and the overlap with the organ delineation o of interest is computed. We denote with S^o_{OVH} the scaled inverse of the Manhattan distance of two patients' OVH for the organ delineation o. The Manhattan distance of two OVHs is the sum of absolute differences between each pair of histogram bins. A scaling is adopted similar to what is done for $S^o_{\text{Hausdorff}}$, to ensure that $(S^o_{\text{OVH}})^2 \geq S^o_{\text{OVH}}$. Furthermore, we denote with S_{OVH} the aggregated measure considering all organs at once, computed similar to how it was originally used in the work introducing the OVH[30]:

$$S_{\text{OVH}} = \sqrt{\sum_{o \in O} (S^o_{\text{OVH}})^2}.$$

2.B.6. Correlations of similarity notions

The correlation between the similarity measurements is assessed with Pearson r_p and Spearman r_s coefficients. The first assumes a linear relationship between the two variables and is sensitive to outliers, whereas the second focuses on monotonic relationships, by considering only ranks (i.e., from sorting) rather than actual data values.

2.C. Regression and feature relevance

The goal now is to reproduce the measurements of similarity among patients using a function of only the features described in Section 2.A. However, because a similarity is defined over pairs of patients, individual features cannot be used directly. Instead, pairwise versions of the features are considered. For a numerical feature, the absolute difference of the two individual feature values is taken. Pairwise versions of categorical features are Boolean values, indicating whether the two categories are the same (1) or different (0). In other words, for the ith feature f_i of patients 1 and 2, the corresponding pairwise feature is $g_i^{1,2} = |f_i^1 - f_i^2|$ if f_i is numeric (e.g., weight), and it is

$$g_i^{1,2} = \begin{cases} 1 & \text{if } f_i^1 = f_i^2 \\ 0 & \text{if } f_i^1 \neq f_i^2 \end{cases}$$

if f_i is categorical (e.g., gender).

A random forest algorithm is used to learn how features can explain similarities, that is, to learn a function representing similarity, given the features, and to compute feature relevance. Random forest is a widely adopted machine learning technique which is capable of performing nonlinear regression, and is robust in assessing feature relevance thanks to its instrinsic feature sampling.[32,33] In particular, the recent *cforest* implementation in R[34] is adopted in this work. Such technique uses conditional inference trees[35] to constitute the forest, providing an unbiased estimation of feature importance in scenarios where features have different scale of measurement or number of categories[36] (e.g., this study).

To understand which features are important for each similarity notion, a separate random forest is trained for each similarity. A split of data into separate training and testing sets is not necessary, since random forest inherently performs *bagging*, that is, each regression tree in the forest is trained on a

random subset of the data and tested on the remaining. Given the stochastic nature of the method, ten independent runs (i.e., training a random forest) are performed. The number of trees for the cforest is set to 100, and the number of random features to consider in the splits during tree construction is set to one-third of the total number of features (i.e., *mtry* = 1/3, guideline for regression). Feature relevance is investigated only if a forest out-of-bag pseudo-R^2 (i.e., the fit on the inherent test set, computed as 1 − mean squared error/variance of the ground-truth data) is high, since conclusions drawn from models with a low fit are generally false. Feature relevance is computed using the conditional variable importance method of cforest, which adjusts for correlations between predictor variables.[37] Successful forests are further refined by iteratively removing the least important feature, until a statistically significant increase in the out-of-bag mean-squared error of the regression is observed. Significance is assessed using the Mann–Whitney–Wilcoxon test with increasing Bonferroni correction at each iteration i, with p-value of $0.05 \times i$. At the end of the procedure, a trained model is obtained for each similarity which is explainable using a subset of salient features.

2.D. Prediction and automatic selection of similar patients

For each similarity notion, we investigate the capability of learned random forests to correctly predict rankings of patients, which is the ultimate goal.

For this purpose, recent patients are used instead of historically treated patients. This way, the prediction capability can be tested against the ground-truth similarities. This complete process is performed in a leave-one-out cross-validation fashion. Specifically, first, a test patient is removed from the cohort and a random forest is trained over the remaining patients using only the most relevant features. Second, the similarity between the test patient and each of the other patients is predicted by the random forest. This prediction is then sorted to obtain a ranking which is subsequently assigned a performance score. The overall prediction quality is the average of the scores obtained when repeating the steps above for each patient in the cohort. Moreover, runs for individual patients are repeated ten times to reduce stochastic noise in the forest training phases. The pseudocode of this procedure is illustrated in Fig. 1.

To score the performance in ranking prediction, the following four indicators are proposed: *head presence*: number of patients in the top k of the predicted ranking who are also in the top k of the ground-truth ranking, that is, the capability of correctly predicting the most similar patients; *tail presence*: analogous to head presence, but on the bottom k patients of the rankings, that is, the capability of correctly predicting the most dissimilar patients; *average displacement*: calculated for the patients who are wrongly predicted to be in the top k, the average displacement in positions between the kth position and the actual position in the ground-truth ranking; *worst displacement*: similar to average displacement, but calculated only for the worst case, that is, for the patient who is most dissimilar yet wrongly predicted to be in the top k. Note that, for $k = 1$, the average displacement is the same as the worst displacement. All the indicators are reported as percentages. A good prediction is one that reaches high head presence and tail presence and minimizes average displacement and worst displacement. In the experiments, the parameter k varies in $\{1, 3, 5\}$, where $k = 1$ means that the prediction is assessed only on the most similar patient (dissimilar for tail presence). This corresponds to evaluating an automatic selection which retrieves only one patient. By increasing k, it is possible to see if the prediction is generally good, noisy, or consistently poor. An example of the indicators is depicted in Fig. 2. With $k = 5$, three of five patients are correctly predicted as similar, that is, the head presence is 60%, while four are correctly predicted as very dissimilar (tail presence is 80%). Patients 5 and 9 are incorrectly predicted to be within the five most similar, whereas in the ground-truth ranking, they are respectively 3 and 5 positions away from the head, out of a total of 12 (= 17 − 5) dissimilar patients. Thus, the average displacement is 33% and the worst displacement is 42%.

2.E. Reconstruction case

Two illustrative dose reconstructions are performed for one patient p: one using a representative who is correctly predicted to be most similar by our model according to S_{deform}, and one using a representative who is the most similar

```
1  function scoreAutomaticSelection(cohort, similarityNotion, bestFeatures)
2      score ← 0
3      for i ∈ {1, ..., 10} do
4          for p ∈ cohort do
5              others ← cohort \ {p}
6              f ← trainForest(others, bestFeatures)
7              s ← f.predictSimilarity(p, others)
8              s* ← getSimilarity(p, similarityNotion)
9              r ← makeRanking(s)
10             r* ← makeRanking(s*)
11             score ← score + computeScore(r, r*)
12     score ← score/(10|cohort|)
13     return score
```

Fig. 1. Function assessing the quality of the automatic selection of similar patients.

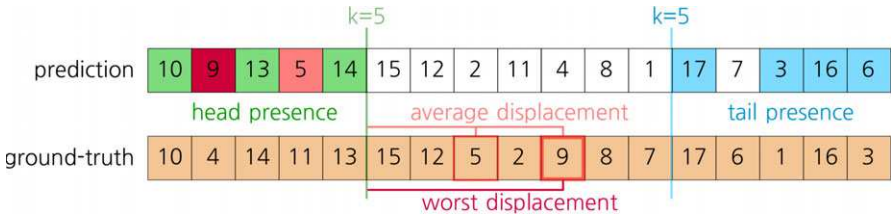

FIG. 2. Example to illustrate the computation of the four indicators of prediction performance. The prediction and ground-truth rankings contain the IDs of 17 patients instead of 37 for ease of representation. In both rankings, the leftmost IDs are the most similar patients, the rightmost the most dissimilar. Patients correctly predicted in the head (tail) are depicted in green (blue). Patients 5 and 9 are wrongly predicted to be in the head and determine the average displacement. Patient 9 is the worst to be predicted in the head, being the most dissimilar in the ground-truth ranking, and determines the worst displacement.

according to age, height, and weight. Specifically, the latter match is performed by taking the patient q with lowest rooted sum of squared age, height, and weight differences (after normalizing all differences to the interval [0, 1]):

$$\sqrt{(\text{age}^p-\text{age}^q)^2+(\text{height}^p-\text{height}^q)^2+(\text{weight}^p-\text{weight}^q)^2}.$$

The reconstruction is performed by applying the treatment plan of patient p to the other two patients, using the treatment planning system Oncentra (version 4.3, Elekta, Stockholm, Sweden). The treatment plan is a real clinical plan for left renal fossa irradiation (Wilms' tumor) using an Elekta Linac with a multileaf collimator beam limiting device, energy: 6 MV. Under the hypothesis that p is a historically treated patient, a digitally reconstructed radiograph of p is generated displaying the borders of the treatment field. Consequently, radiographs are also generated for the two matched patients and are used to adjust the field border of the plan to correct for evident discrepancies in the bony anatomy. The monitor units of the original plan are also scaled to keep the dose point in the middle of the field (isocenter) as close as possible to its value before adjustment. This work has been assessed by an experienced pediatric radiation oncologist. Finally, the treatment is simulated and the following metrics are recorded: mean dose D_{mean}, max. dose $D_{2\text{cc}}$, and the dose volume histograms (DVHs), for right kidney, liver, spleen, and spinal cord (T10-S1).

3. RESULTS

3.A. Correlations of similarity notions

Pairwise Pearson and Spearman correlation coefficients between the similarity measures S_{deform}, S_{DSC}, $S_{\text{Hausdorff}}$, S_{const}, and S_{OVH} are reported in Table III. The two coefficients r_p and r_s show good agreement in general. Although S_{deform} and S_{const} are moderately correlated (r_p of 0.64, r_s of 0.55), S_{DSC}, $S_{\text{Hausdorff}}$, and S_{OVH} are more independent. It is crucial to recall that these latter similarities are highly dependent on organ shape. The correlation coefficients of S_{deform} and S_{const} with S^o_{DSC}, $S^o_{\text{Hausdorff}}$, and S^o_{OVH}, that is, the latter similarities separately computed for each organ o, are represented in Fig. 3 (only Pearson correlation is reported since Spearman leads to very similar results). These results show that moderate to substantial correlations are present among S_{deform}, S_{const}, $S^{\text{body}}_{\text{DSC}}$, $S^{\text{body}}_{\text{Hausdorff}}$. Moreover, $S^{\text{liver}}_{\text{DSC}}$ is highly correlated with $S^{\text{liver}}_{\text{Hausdorff}}$, and $S^{\text{spleen}}_{\text{DSC}}$ is highly correlated with $S^{\text{spleen}}_{\text{Hausdorff}}$. However, these latter similarities are weakly correlated with the former ones, based on deformable image registration, disposition of the internal organs, and shape of the abdomen. The fact that $S^{\text{spinal cord}}_{\text{DSC}}$ and $S^{\text{spinal cord}}_{\text{Hausdorff}}$ are not clearly correlated is likely due to the elongated shape and different bending of this organ (see, e.g., Fig. 4).

3.B. Regression and feature relevance

The magnitude of Pearson correlation coefficients between all pairs of features is depicted in Fig. 5. The largest correlation coefficients were found for combinations of age and height, LR diameter and weight, and tumor site and radical nephrectomy.

Figure 6 shows the pseudo-R^2 of the random forest method (averaged over ten runs), for each similarity notion. Although with random forest it is possible to learn nonlinear interactions among features, only few similarities are modeled with a high pseudo-R^2: S_{deform}, S_{const}, $S^{\text{body}}_{\text{DSC}}$, and $S^{\text{body}}_{\text{OVH}}$ [see Figs. 6(a) and 6(b)]; with, respectively, pseudo-R^2 of 0.70, 0.69, 0.87, and 0.85. In particular, the variation in measurements for the notions aggregating (in Fig. 6(a), S_{DSC}, $S_{\text{Hausdorff}}$, S_{OVH}) or specifically focused on internal organ shapes [all similarities in Figs. 6(c)–6(e)] cannot be modeled well (i.e., low pseudo-R^2). This result clearly shows that the features at hand do not provide enough information to grasp the large variability in the internal anatomy of our young cohort. On the other hand, it is the similarities mainly or specifically focusing on the overall abdomen (S_{deform}, S_{const}, $S^{\text{body}}_{\text{DSC}}$, $S^{\text{body}}_{\text{OVH}}$) that are decently modeled. Not surprisingly, these similarities are found to be correlated among themselves (Section 3.A).

The feature relevance of the best modeled similarities S_{deform}, S_{const}, $S^{\text{body}}_{\text{DSC}}$, and $S^{\text{body}}_{\text{OVH}}$ is reported in Fig. 7. The abdominal AP and LR diameters clearly stand out as common relevant features for the four similarities. Although not salient in three out of four cases, weight is always among the predictors with a statistically significant relevance. It is also worth noticing how nephrectomy has a slight, yet relevant influence on the organs' constellation, which may be linked to a possible shift of organs after kidney resection.

TABLE III. Pearson r_p and Spearman r_s correlation coefficients for the five (aggregated) similarity notions.

	Pearson r_p					Spearman r_s				
	S_{deform}	S_{DSC}	$S_{\text{Hausdorff}}$	S_{const}	S_{OVH}	S_{deform}	S_{DSC}	$S_{\text{Hausdorff}}$	S_{const}	S_{OVH}
S_{deform}	1.00	0.24	0.32	0.64	0.22	1.00	0.18	0.35	0.55	0.16
S_{DSC}	0.24	1.00	0.39	0.49	0.34	0.18	1.00	0.46	0.50	0.31
$S_{\text{Hausdorff}}$	0.32	0.39	1.00	0.37	0.14	0.35	0.46	1.00	0.48	0.14
S_{const}	0.64	0.49	0.37	1.00	0.52	0.55	0.50	0.48	1.00	0.42
S_{OVH}	0.22	0.34	0.14	0.52	1.00	0.16	0.31	0.14	0.42	1.00

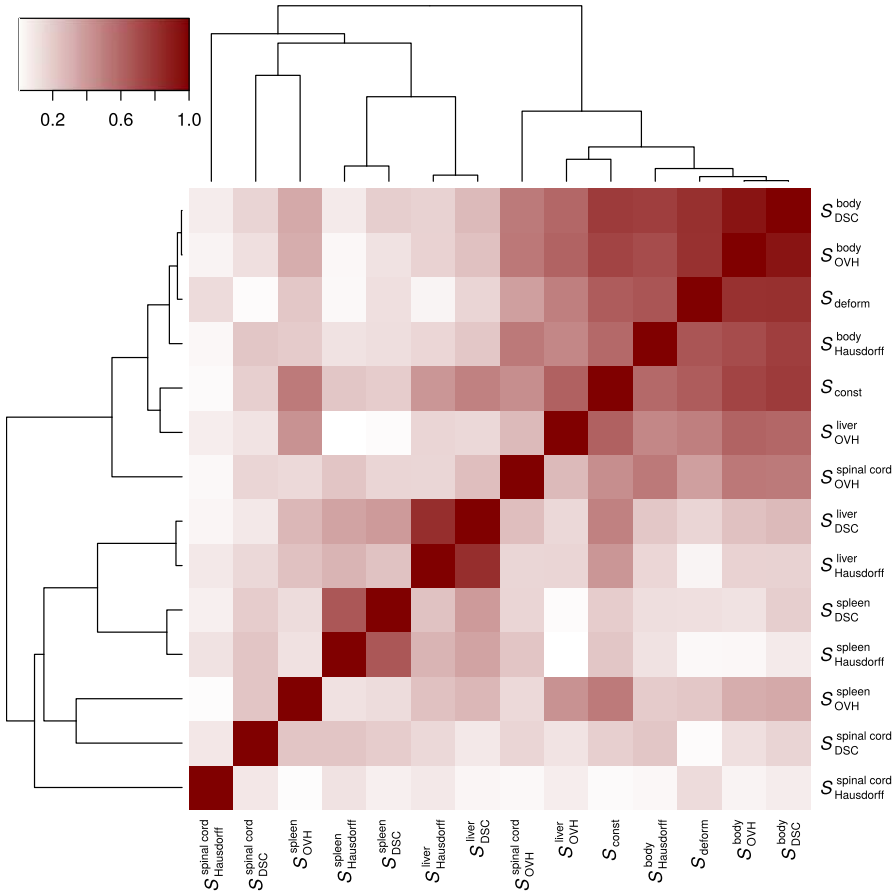

FIG. 3. Heatmap and hierarchical-clustering dendrogram based on absolute values of Pearson correlation coefficients between S_{deform}, S_{const}, and the organ-specific S^o_{DSC}, $S^o_{\text{Hausdorff}}$, and S^o_{OVH}, with $o \in \{\text{liver, spleen, spinal cord (T10-S1), body (T10-S1)}\}$.

3.C. Prediction and automatic selection of similar patients

The capability of the models trained using the most important features (obtained in Section 3.B) to perform automatic selection is now assessed. Table IV shows the quality of the prediction in terms of the four proposed indicators head presence, tail presence, average displacement, and worst displacement, for the similarity notions that could be reliably modeled: S_{deform}, S_{const}, $S^{\text{body}}_{\text{DSC}}$, and $S^{\text{body}}_{\text{OVH}}$.

The results are averages over ten repetitions. When considering only the most similar patient ($k = 1$), the best choice is predicted correctly 22.37% of the times (averaged over all similarity measures, where the worst is S_{deform} with only 16.22%, and the best is S_{const} with 30.00%). When such prediction is wrong, the patient misclassified as most similar is approximately the fifth most similar, that is, the patient is displaced within the top 11% of the ground-truth ranking (the worst is S_{const} with 13.73%, the best is $S^{\text{body}}_{\text{DSC}}$ with 7.71%). Now, by increasing k to 3 and 5, it can be seen that the head presence increases to roughly 50%, that is, half of the most similar patients become correctly predicted to be in the head of the ranking. A similar behavior can be observed on the tail presence indicator, that is, the accuracy in predicting the most dissimilar patients. While the head and the tail presence increases considerably with k, the average displacement

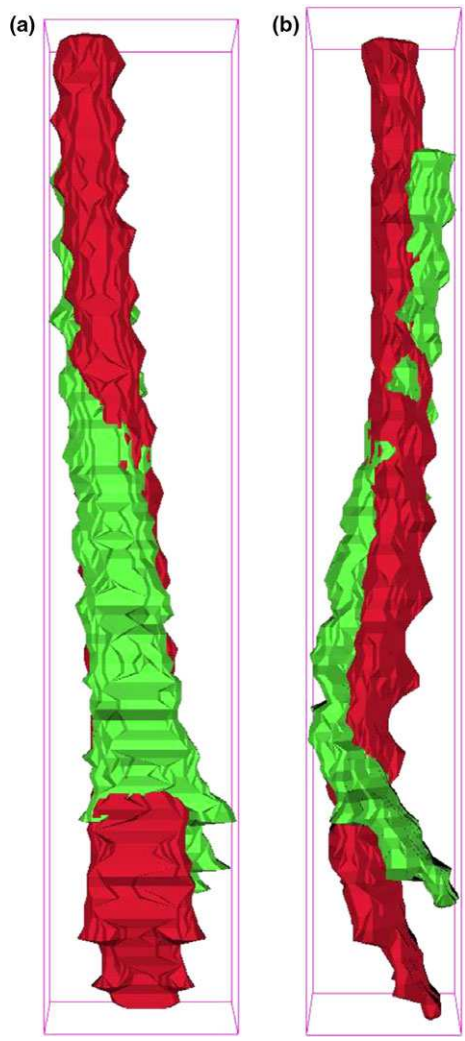

FIG. 4. Two spinal cords aligned on the center of mass for DSC and Hausdorff distance computation, from AP (a) and LR (b) perspective.

oscillates slightly. For the best-modeled $S_{\text{DSC}}^{\text{body}}$ (pseudo-R^2 of 0.87 with the random forest trained using only the most important features), the worst displacement increase is also particularly limited when increasing k. This trend shows that the models are reliably able to find a coarse notion of similarity, with a quite good capability of identifying which patients constitute a cluster of most similar, and which constitute a cluster of most dissimilar, but with limitations in terms of accurately ordering the most similar patients.

3.D. Reconstruction case

Patients with ID 6, 18, and 34 are the ones used to perform an illustrative reconstruction. The right kidney is intact in all three patients. Patient 6 (age: 2.51 yr, gender: female, height: 93.0 cm, weight: 14.0 kg) is hypothesized to be a historical patient for whom a dose reconstruction is needed. Random forest correctly predicts patient 34 (age: 2.21 yr, gender: male, height: 90.0 cm, weight: 15.0 kg) to be the closest match to 6, according to S_{deform}. Patient 18 (age: 2.58 yr, gender: female, height: 92.0 cm, weight: 13.0 kg) is the most similar to 6 according to age, height, and weight. However, patient 18 is ranked as 16th in terms of S_{deform} similarity to patient 6.

The outcome of the dose reconstruction is presented in terms of D_{mean} and $D_{2\text{cc}}$ in Table V and qualitatively in terms of DVHs in Fig. 8. Recurring to patient 34 as reference leads to markedly better dose reconstruction for right kidney and liver, while patient 18 is slightly preferable for spleen and spinal cord (with exception of $D_{2\text{cc}}$ for the latter). In particular, patient 34 excels against patient 18 when comparing D_{mean} of the right kidney, with a relative error of 12.78% for the former and 63.89% for the latter (51.11% difference). Instead, the case where 18 is mostly preferable is the relative error on D_{mean} of spinal cord, with 7.39% for patient 18 and 11.71% for patient 34 (only 4.32% difference). A qualitative inspection of the DVHs points at a similar conclusion; while patient 18 may seem to be preferable for the reconstruction of spleen and spinal cord [Figs. 8(c) and 8(d)], patient 34 is markedly better for right kidney and liver [Figs. 8(a) and 8(b)] reconstruction.

4. DISCUSSION

To the best of our knowledge, this study represents a first attempt to understand what are the key anatomical characteristics to represent similarity among childhood cancer patients and to assess the feasibility of performing an automatic selection of a representative patient for a highly individualized CT-based 3D dose reconstruction method.

To establish a ground-truth notion of similarity between patients, we have proposed and studied five possible measures of similarity. It has been found that the DSC and Hausdorff-based similarities of the same organ are highly correlated for liver and spleen, but not for the spinal cord (likely due to its elongated shape). Furthermore, the two new measures we proposed, S_{deform} and S_{const}, are correlated with the similarities that are based on established shape descriptors (DSC, Hausdorff, and OVH) when using the contour of the external abdomen as shape.

Random forest has been adopted to relate (pairwise) historically available features with the ground-truth notions of similarity. This allowed for the modeling of complex, nonlinear interactions and the assessment of feature relevance. We found that aggregated similarities focusing on general aspects of the whole abdomen (S_{deform}, S_{const}) as well as the ones focusing specifically on the shape of the external abdomen ($S_{\text{DSC}}^{\text{body}}$, $S_{\text{OVH}}^{\text{body}}$) were decently modeled. However, a relationship between the features at hand and internal organ-specific shape-based similarities could not be learned. Such result is not surprising, given previous literature studies on organ variability (e.g., correlating organ volumes with BMI in adults[19]). Internal organ variability is possibly even further increased by the disease these children suffer from, together with prior drug treatments (resulting in, e.g., possible hepatosplenomegaly). This means that more features are needed to predict the internal anatomy. To this end, we plan to harvest more information from images, that is, by means of

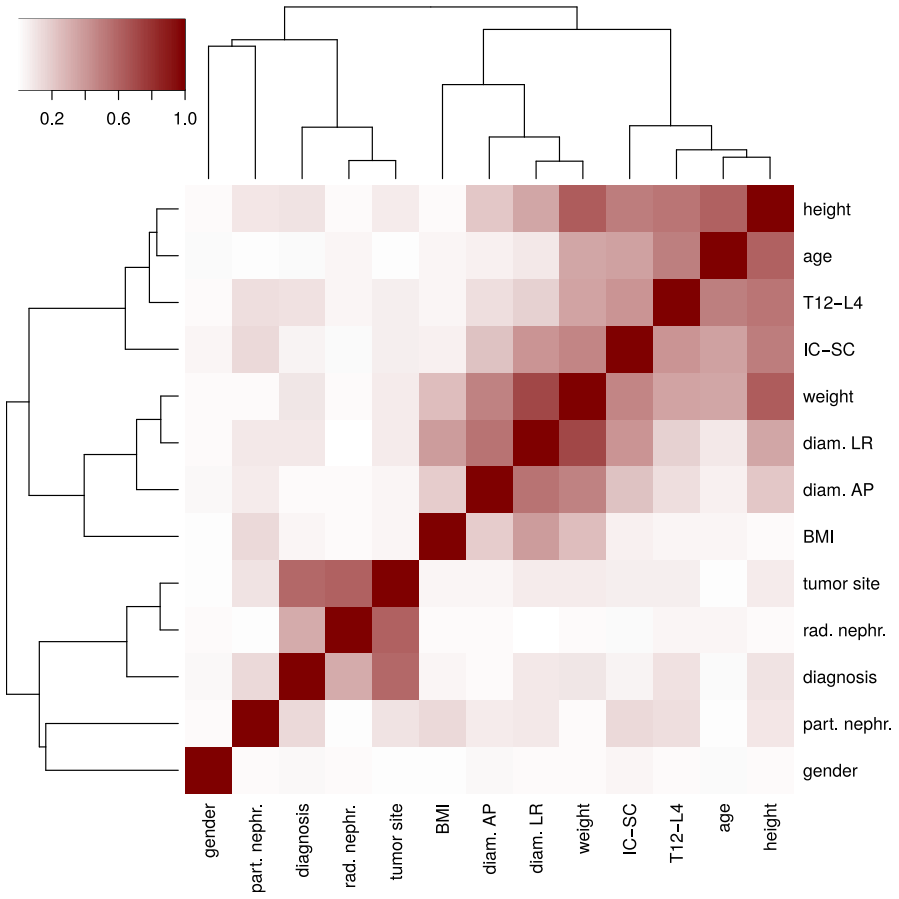

FIG. 5. Heatmap and hierarchical-clustering dendrogram representing (absolute) Pearson correlation among pairwise features (abbreviations as introduced in Table I and II).

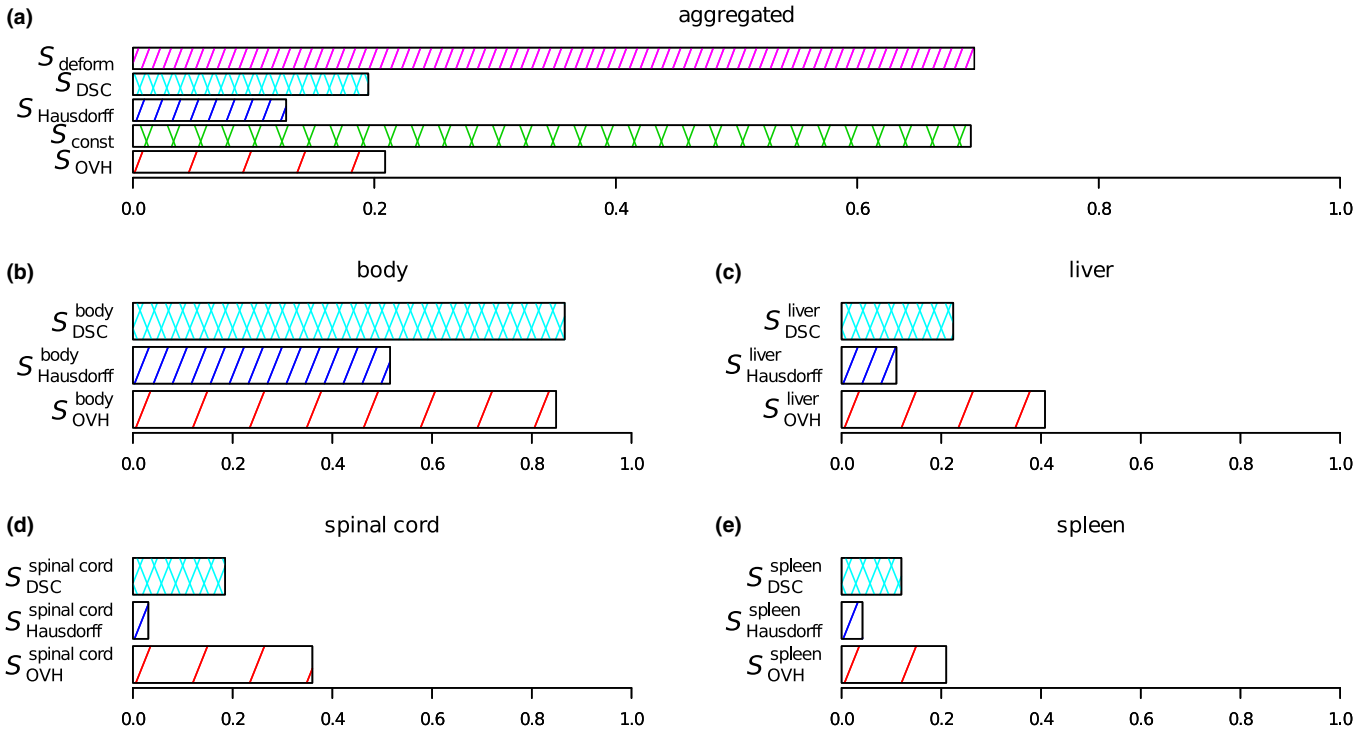

FIG. 6. Pseudo-R^2 of trained random forests for all the similarities, both aggregated (a), and for individual organs: body (b), liver (c), spinal cord (d), spleen (e).

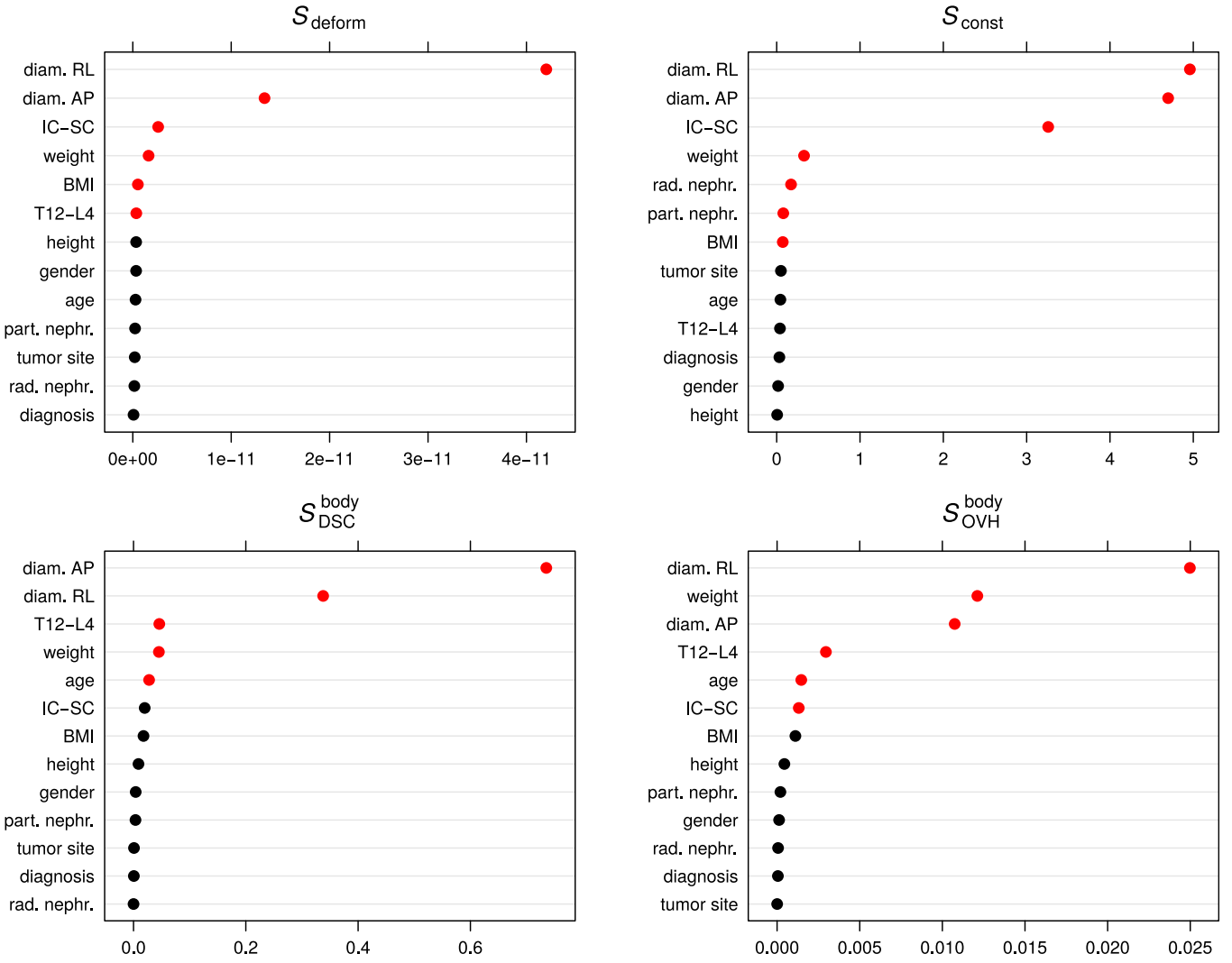

FIG. 7. Feature relevance from random forest for the similarities that could be modeled with high pseudo-R^2. Dots in red represent the minimal subset of features with which it is possible to obtain a random forest with no significant loss in mean-squared error. Note: different scales are used on the x-axes.

TABLE IV. Predicted ranking scores for the explainable similarities, for $k \in \{1, 3, 5\}$. Results in bold are the best scores among similarities for a fixed k.

	Head pres.			Tail pres.			Avg. disp.			Worst disp.		
k	1	3	5	1	3	5	1	3	5	1	3	5
S_{deform}	16.22	34.14	44.54	49.73	**82.52**	76.70	13.44	12.60	10.79	13.44	23.78	28.20
S^{body}_{DSC}	25.95	**52.07**	**65.35**	65.14	75.50	86.00	**7.71**	**5.61**	**5.07**	**7.71**	**12.94**	**18.08**
S_{const}	**30.00**	44.23	50.81	27.57	65.95	78.86	13.73	10.85	11.35	13.73	22.38	32.91
S^{body}_{OVH}	17.30	41.44	56.43	**83.24**	79.64	**86.92**	10.73	8.11	7.14	10.73	16.91	21.55
Mean	22.37	42.79	54.28	56.42	75.90	82.12	11.40	9.29	8.57	11.40	18.75	25.19

2D/3D registrations,[38] and the usage of navigator channels on 2D (digitally reconstructed) radiographs, which have been proven capable of enabling decent organ shape reconsruction[20,21] in the adaptation step. Furthermore, the models learned by random forest are complex to interpret. Adopting other machine learning methods may generate equally powerful models of easier interpretation and hopefully provide more insight on the problem (e.g., genetic programming[39]).

For the well-modeled similarities, results show that the most salient features are the abdominal diameters. This is in contrast with the common practice in phantom-based dose reconstruction of using age, height, and weight (gender is typically not considered for young children) to select a representative phantom for any treatment scenario. We hypothesize that it may be necessary to define a different set of relevant features depending on the specific treatment to

TABLE V. The upper part of the table shows D_{mean} and $D_{2\text{cc}}$ in cGy for right kidney, liver, spleen, and spinal cord (T10-S1) for patients 6, 18, and 34. The lower part reports the relative error of the reconstruction, with lowest errors in bold.

	Right kidney		Liver		Spleen		Spinal cord	
Patient ID	D_{mean}	$D_{2\text{cc}}$	D_{mean}	$D_{2\text{cc}}$	D_{mean}	$D_{2\text{cc}}$	D_{mean}	$D_{2\text{cc}}$
6	1.80	12.29	4.83	13.84	13.96	14.86	11.36	13.80
18	2.95	11.19	4.19	13.63	13.83	14.24	10.52	13.99
34	2.03	11.53	4.44	13.89	13.76	14.14	10.03	13.63
% relative error from patient 6								
18	63.89	8.95	13.25	1.52	**0.93**	**4.17**	**7.39**	1.38
34	**12.78**	**6.18**	**8.07**	**0.36**	1.43	4.84	11.71	1.23

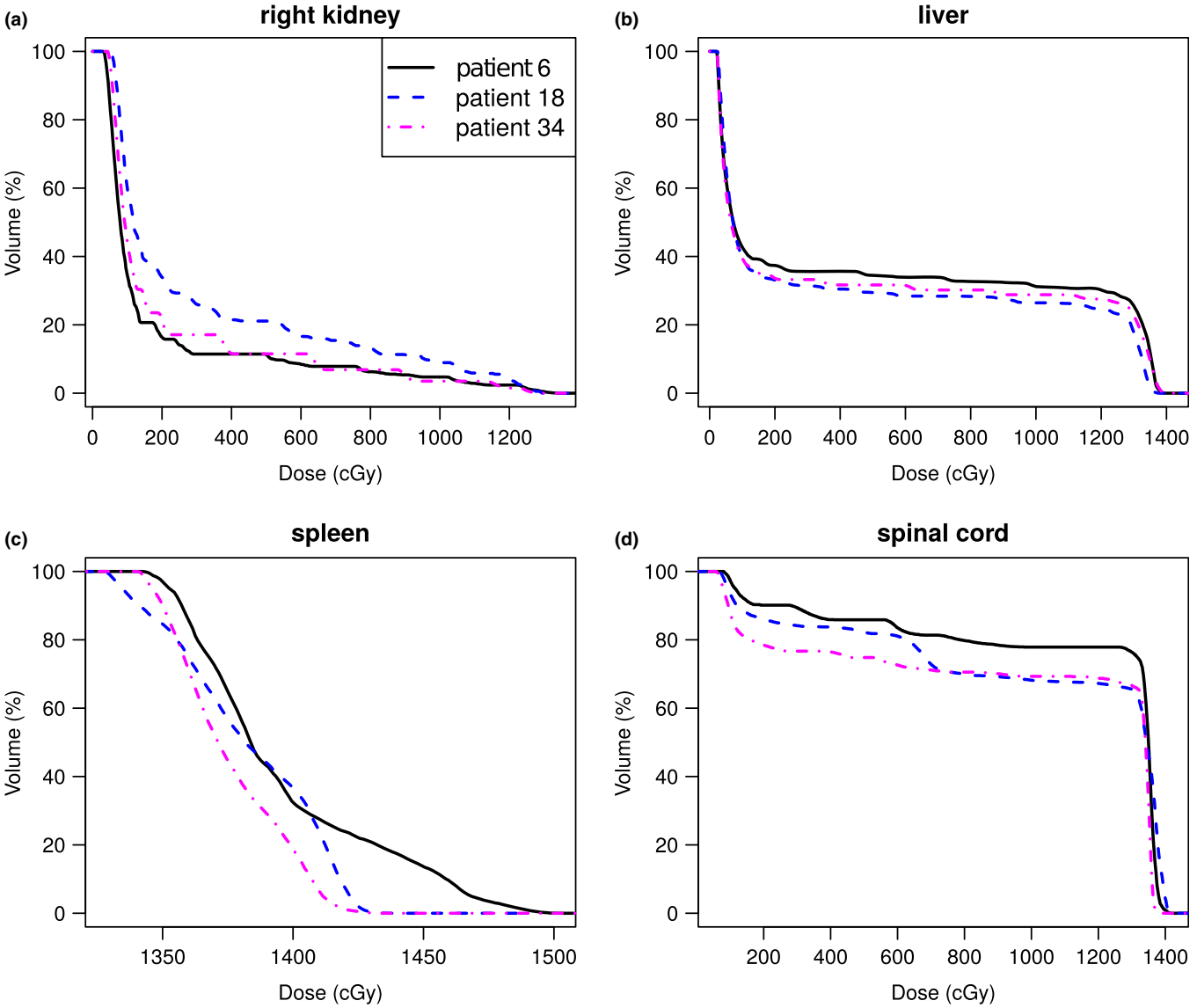

FIG. 8. DVHs of right kidney (a), liver (b), spleen (c), and spinal cord (T10-S1) (d) of patients 6, 18, and 34. Using patient 34 as reference for the dose reconstruction of patient 6 leads to highly similar DVHs for the right kidney (a) and the liver (b).

perform a very accurate selection. The model-predicted similarity rankings are noisy, but roughly coincide with the ground-truth ones. If the first retrieved patient would be taken as the singular best match ($k = 1$), the choice would often be wrong (e.g., three out of four times for $S_{\text{DSC}}^{\text{body}}$), yet the error would be limited (e.g., within the three most similar for

$S_{\mathrm{DSC}}^{\mathrm{body}}$). Thus, although suboptimal, the resulting learned models can be considered robust. When k is increased, the head presence increases much more than the average displacement. Therefore, it may be interesting to assess whether computing the 3D dose distributions for a small number of k patients, and then take the average as the final result, may lead to more accurate reconstructions.

It will be important to understand if and how these similarities can be combined into a single ground-truth similarity. Between highly correlated notions, it would be natural to look for an aggregated compromise that expresses all of them at once. Contrary, highly uncorrelated ones should probably be kept separated. This would mean that actually a multiobjective selection approach is sought after, that is able to retrieve a limited set of different patients who are similar to the historically treated one according to different notions of similarity. Eventually, a physician could decide which CT to use for reconstruction, or, as mentioned above, multiple reconstructions could be performed and the average dose distribution could be taken as final result.

Besides a new perspective on the selection of a reference patient, this work presents some limitations. A first limitation is that we chose to consider the scenario of Wilms' tumor treatment in children and focused on anatomical similarity of the abdomen. Thus, the results presented in this work are valid within the domain of the chosen region of interest (abdomen) and the characteristics of the cohort (Caucasian children between approximately 2–6 yr old). However, the approach presented here is general and can be applied to other regions of interests and cohorts. Furthermore, note that the choice of trying to machine learn similar anatomy rather than directly machine-learn similar dose distributions overcomes the limitations of the latter: a specific treatment does not need to be defined and (manually) simulated beforehand on each available CT scan. In fact, our results can be used for any abdominal treatment (e.g., neuroblastoma), within the cohort characteristics (Caucasian children between approximately 2–6 yr old). Nonetheless, it is well possible to define a similarity based on 3D dose distributions for a specific treatment and learn a model capable of retrieving similar patients in that sense. A second limitation is the lack of an analysis of the relationship between similarity notions and dose outcomes (e.g., D_{mean}, $D_{2\mathrm{cc}}$, and DVHs). This work showed one exemplary reconstruction, but a validation study involving a statistically relevant number of patients should be performed. Such work may be realized as shown in our illustrative reconstruction, using a number of recently treated patients instead of historically treated ones, as follows. (a) A treatment plan should be simulated on the patient and measurements from the 3D dose distribution should be recorded; (b) using (historically plausible) features of the patient, a representative patient from a cohort of candidates should be selected according to a similarity notion; (c) dose measurements should be taken on the representative patient and compared with the ones taken in the first step. The outcome of such a study may tell which similarity notion(s) is preferable, that is, leads to more accurate dose reconstruction. However, we remark that selecting a similar anatomy is but the first step of a dose reconstruction method. In order to comprehensively validate the contribution of this work in dose reconstruction, the reference anatomy retrieved by our method should undergo an adaptation step, to increase its resemblance with the historically treated patient, and the original treatment should be simulated as close as possible. Consequently, a fair comparison with state-of-the-art phantom-based dose reconstruction methods will be possible. This is however outside the scope of this article. A third limitation is the relatively small size of the cohort examined in this study, and, in general, the availability of data to specific institutes. As generally true with machine learning approaches, we expect that including more data will result in improved models and prediction capabilities. We plan to expand our cohort by including anonymized patient data provided by radiotherapy departments of other institutes.

5. CONCLUSION

This study presents a novel, machine learning-based approach to an important part in the process of 3D dose reconstruction for historically treated patients using recent real patient data rather than phantoms: selecting a good representative recently treated patient.

Similarity measures that consider the overall abdomen and the position of internal organs can be decently modeled (pseudo-$R^2 \geq 0.7$), and automatic selection based on such models reaches a coarse but robust performance. However, it was not possible to find a relationship between features available for historically treated patients and specific organ shapes. All in all, our novel approach shows potential in using CT scans of actual, recent patients directly to perform dose reconstruction, and a number of future research steps are possible to gain substantial improvements, for example, extending the data with more patients, exploiting available 2D image data such as simulator films to extend the feature set and exploring combinations of similarity notions.

ACKNOWLEDGMENTS

The authors would like to thank Brian V. Balgobind, Koen F. Crama, Rianne M.A.J. de Jong, Jorrit Visser, and Ziyuan Wang (Department of Radiation Oncology, Academic Medical Center, Amsterdam, the Netherlands) for their help with the retrieval and preprocessing of data. We further thank Raquel Dávila Fajardo (department of Radiation Oncology, UMC Utrecht Cancer Center, the Netherlands) for sharing the data of three patients treated at the UMC Utrecht for inclusion in this study. This work is part of the research project *3D dose reconstruction for children with long-term follow-up — Toward improved decision making in radiation treatment for children with cancer* with project number 187, which is financed by Stichting Kinderen Kankervrij (KiKa). Dr. C. M. Ronckers is supported by the Dutch Cancer Society (KWF), grant number UVA2012-5517.

CONFLICTS OF INTEREST

Dr. T. Alderliesten, Dr. A. Bel, and Dr. P.A.N. Bosman are involved in projects supported by Elekta.

[a)]Author to whom correspondence should be addressed. Electronic mail: marco.virgolin@cwi.nl

REFERENCES

1. Jairam V, Roberts KB, Yu JB. Historical trends in the use of radiation therapy for pediatric cancers: 1973–2008. *Int J Radiat Oncol Biol Phys*. 2013;85:151–155.
2. Sklar CA, Constine LS. Chronic neuroendocrinological sequelae of radiation therapy. *Int J Radiat Oncol Biol Phys*. 1995;31:1113–1121.
3. Mulhern RK, Merchant TE, Gajjar A, Reddick WE, Kun LE. Late neurocognitive sequelae in survivors of brain tumours in childhood. *Lancet Oncol*. 2004;5:399–408.
4. Geenen MM, Cardous-Ubbink MC, Kremer LC, et al. Medical assessment of adverse health outcomes in long-term survivors of childhood cancer. *J Am Med Assoc*. 2007;297:2705–2715.
5. Tsai YL, Tsai SC, Yen SH, et al. Efficacy of therapeutic play for pediatric brain tumor patients during external beam radiotherapy. *Child's Nerv Syst*. 2013;29:1123–1129.
6. van Dijk IW, Oldenburger F, Cardous-Ubbink MC, et al. Evaluation of late adverse events in long-term wilms' tumor survivors. *Int J Radiat Oncol Biol Phys*. 2010;78:370–378.
7. Thouvenin-Doulet S, Fayoux P, Broucqsault H, Bernier-Chastagner V. Neurosensory, aesthetic and dental late effects of childhood cancer therapy. *Bull Cancer*. 2015;102:642–647.
8. Breslow N, Olshan A, Beckwith JB, Green DM. Epidemiology of Wilms tumor. *Med Pediatr Oncol*. 1993;21:172–181.
9. Dome JS, Graf N, Geller JI, et al. Advances in Wilms tumor treatment and biology: progress through international collaboration. *J Clin Oncol*. 2015;33:2999–3007.
10. Bentzen SM, Constine LS, Deasy JO, et al. Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues. *Int J Radiat Oncol Biol Phys*. 2010;76(3 Suppl):3–9.
11. Boersma LJ, Damen EM, de Boer RW, et al. Dose-effect relations for local functional and structural changes of the lung after irradiation for malignant lymphoma. *Radiother Oncol*. 1994;32:201–209.
12. Cao Y, Pan C, Balter JM, et al. Liver function after irradiation based on computed tomographic portal vein perfusion imaging. *Int J Radiat Oncol Biol Phys*. 2008;70:154–160.
13. Buettner F, Gulliford SL, Webb S, et al. Assessing correlations between the spatial distribution of the dose to the rectal wall and late rectal toxicity after prostate radiotherapy: an analysis of data from the MRC RT01 trial (ISRCTN 47772397). *Phys Med Biol*. 2009;54:6535–6548.
14. Constine L, Hodgson D, Bentzen S. MO-D-BRF-01: pediatric treatment planning ii: the PENTEC report on normal tissue complications. *Med Phys*. 2014;41:419–419.
15. Stovall M, Weathers R, Kasper C, et al. Dose reconstruction for therapeutic and diagnostic radiation exposures: use in epidemiological studies. *Radiat Res*. 2006;166:141–157.
16. Geyer AM, O'Reilly S, Lee C, Long DJ, Bolch WE. The UF/NCI family of hybrid com putational phantoms representing the current US population of male and female children, adolescents, and adults–application to CT dosimetry. *Phys Med Biol*. 2014;59:5225–5242.
17. ICRP. Basic anatomical and physiological data for use in radiological protection: reference values. *ICRP Publication*. 2002;32:1–277.
18. Lamart S, Imran R, Simon SL, et al. Prediction of the location and size of the stomach using patient characteristics for retrospective radiation dose estimation following radiotherapy. *Phys Med Biol*. 2013;58:8739–8753.
19. de la Grandmaison GL, Clairand I, Durigon M. Organ weight in 684 adult autopsies: new tables for a Caucasoid population. *Forensic Sci Int*. 2001;119:149–154.
20. Ng A, Nguyen TN, Moseley JL, 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*. 2010;37:1017–1028.
21. Ng A, Nguyen TN, Moseley JL, et al. Navigator channel adaptation to reconstruct three dimensional heart volumes from two dimensional radiotherapy planning data. *BMC Med Phys*. 2012;12:1.
22. Ng A, Brock KK, Sharpe MB, et al. 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*. 2012;84:557–563.
23. Zhou R, Ng A, Constine LS, et al. A comparative evaluation of normal tissue doses for patients receiving radiation therapy for hodgkin lymphoma on the childhood cancer sur vivor study and recent children's oncology group trials. *Int J Radiat Oncol Biol Phys*. 2016;95:707–711.
24. Klein S, Loog M, van der Lijn F, et al. Early diagnosis of dementia based on intersubject whole-brain dissimilarities. In Proc, IEEE I S Biomed Imaging. ISBI'10. Piscataway, NJ, USA: IEEE Press. 2010;249–252.
25. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging*. 2010;29:196–205.
26. Shamonin DP, Bron EE, Lelieveldt BP, et al. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Front Neuroinform*. 2013;7:50.
27. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*. 2004;11:178–189.
28. Huttenlocher DP, Klanderman GA, Rucklidge WA. Comparing images using the Hausdorff distance. *IEEE Trans Pattern Anal Mach Intell*. 1993;15:850–863.
29. Xu Z, Panjwani SA, Lee CP, et al. Evaluation of body-wise and organ-wise registrations for abdominal organs. In Proc SPIE Int Soc Opt Eng. 2016; 97841O–97841O-6.
30. Kazhdan M, Simari P, McNutt T, et al. A shape relationship descriptor for radiation therapy planning. *Med Image Comput Comput Assist Interv*. 2009;12:100–108.
31. Wu B, Ricchetti F, Sanguineti G, et al. Patient geometry-driven information retrieval for IMRT treatment plan quality control. *Med Phys*. 2009;36:5497–5505.
32. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
33. Cutler A, Cutler DR, Stevens JR. Tree-based methods. In *High-Dimensional Data Analysis in Cancer Research*. New York: Springer. 2009:1–19.
34. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. 2016.
35. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat*. 2006;15:651–674.
36. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*. 2007;8:25.
37. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics*. 2008;9:307.
38. Markelj P, Toma˘zevi˘c D, Likar B, Pernu˘s F. A review of 3D/2D registration methods for image-guided interventions. *Med Image Anal*. 2012;16:642–661.
39. Koza JR. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press; 1992.