Machine learning for automatic construction of pediatric abdominal phantoms for radiation dose reconstruction

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

DOI
10.1117/12.2548969

Publication date
2020

Document Version
Final published version

Published in
Medical Imaging 2020

Citation (APA)
Virgolin, M., Wang, Z., Alderliesten, T., & Bosman, P. A. N. (2020). Machine learning for automatic construction of pediatric abdominal phantoms for radiation dose reconstruction. In P.-H. Chen, & T. M. Deserno (Eds.), *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications* (Vol. 11318). Article 1131815 SPIE. https://doi.org/10.1117/12.2548969

Machine learning for automatic construction of pediatric abdominal phantoms for radiation dose reconstruction

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

Marco Virgolin, Ziyuan Wang, Tanja Alderliesten, Peter A. N. Bosman, "Machine learning for automatic construction of pediatric abdominal phantoms for radiation dose reconstruction," Proc. SPIE 11318, Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, 1131815 (2 March 2020); doi: 10.1117/12.2548969

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Machine Learning for Automatic Construction of Pediatric Abdominal Phantoms for Radiation Dose Reconstruction

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

[a]Life Sciences and Health Group, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, the Netherlands
[b]Department of Radiation Oncology, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands
[c]Software Technology Group, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, the Netherlands

ABSTRACT

The advent of Machine Learning (ML) is proving extremely beneficial in many healthcare applications. In pediatric oncology, retrospective studies that investigate the relationship between treatment and late adverse effects still rely on simple heuristics. To capture the effects of radiation treatment, treatment plans are typically simulated on virtual surrogates of patient anatomy called phantoms. Currently, phantoms are built to represent categories of patients based on reasonable yet simple criteria. This often results in phantoms that are too generic to accurately represent individual anatomies. We present a novel approach that combines imaging data and ML to build individualized phantoms automatically. We design a pipeline that, given features of patients treated in the pre-3D planning era when only 2D radiographs were available, as well as a database of 3D Computed Tomography (CT) imaging with organ segmentations, uses ML to predict how to assemble a patient-specific phantom. Using 60 abdominal CTs of pediatric patients between 2 to 6 years of age, we find that our approach delivers significantly more representative phantoms compared to using current phantom building criteria, in terms of shape and location of two considered organs (liver and spleen), and shape of the abdomen. Furthermore, as interpretability is often central to trust ML models in medical contexts, among other ML algorithms we consider the Gene-pool Optimal Mixing Evolutionary Algorithm for Genetic Programming (GP-GOMEA), that learns readable mathematical expression models. We find that the readability of its output does not compromise prediction performance as GP-GOMEA delivered the best performing models.

Keywords: machine learning, pediatric cancer, radiation treatment, dose reconstruction, phantom

1. INTRODUCTION

Virtual anthropomorphic phantoms are 3D representations of the human body that are used as surrogates for the anatomy of humans, to estimate the quantity and geometric distribution of radiation dose when having been exposed to radiation, e.g., in radiation treatment for cancer patients.[1,2]

Several sources of uncertainty exist in situations where phantoms are needed such as, e.g., retrospective radiation treatment dosimetry. Among these sources, the largest uncertainty is associated to the degree to which a phantom matches the anatomy of a patient.[3] Therefore, phantoms need to represent as closely as possible the anatomy they are used as surrogate for.

Current methods for phantom building have two major limitations. Firstly, building phantoms is a manual and time-consuming task. Therefore, a limited number of phantoms is typically produced, according to reasonable human-designed criteria based on population-based statistics, in order to describe categories of patients.[4–8] The second and perhaps more fundamental limitation is that it is unknown how to best define categorization

Further author information: (Send correspondence to M. Virgolin)
E-mail: marco.virgolin@cwi.nl, Telephone: +31 (0)6 33 85 22 33
T. Alderliesten is currently working at the Department of Radiation Oncology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, the Netherlands.

Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications,
edited by Po-Hao Chen, Thomas M. Deserno, Proc. of SPIE Vol. 11318, 1131815
© 2020 SPIE · CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2548969

criteria that describe resemblance of patient anatomy at a personalized level. So far, only simple criteria such as partitioning by percentiles of (combinations of) age, gender, height, and weight, have been explored.[6–9] Nevertheless, recent work has indicated that current phantom building methods result in limited anatomical resemblance,[2,7,10,11] ultimately leading to coarse dose estimations.[12]

We present a new take on phantom building, to overcome both aforementioned limitations. We propose a fully automatic phantom-construction pipeline that generates a pseudo-realistic phantom that is patient-specific. To overcome the need for laborious manual intervention, we propose to re-use 3D patient imaging (CT scans and organ segmentations) collected in a database, to assemble new anatomy combinations. To estimate how to best perform this assembling, i.e., to move beyond the use of too simplistic criteria, we rely on Machine Learning (ML): we train ML models to learn relationships between patient features and 3D metrics based on their internal anatomy.

We consider a relatively hard scenario where phantoms are needed and patient features are limited: dose reconstruction for patients treated in the pre-3D planning era, when radiation treatment plans were designed using 2D radiographs (historical patients). As no 3D imaging is available for historical patients to simulate the treatment and estimate the radiation dose distribution, phantoms are necessary to act as surrogate anatomies. We consider children between 2 to 6 years, and abdominal radiation treatment, because children are typically under-represented in existing phantom libraries,[6,7] and because abdominal radiation treatment is associated with high survival rates for several types of pediatric abdominal cancer (e.g., Wilms' tumor), but also with late adverse effects.[13,14]

2. MATERIALS AND METHODS

2.1 Problem decomposition and automatic pipeline

We decompose the problem of constructing a pediatric abdominal phantom into the following separate tasks: (1) prediction of a segmentation that is representative of the overall body (abdomen); (2) for each organ at risk (OAR) of toxic radiation exposure, (2.1) prediction of center of mass position according to Left-Right (LR), Anterior-Posterior (AP), and Superior-Inferior (SI) direction; (2.2) prediction of a representative segmentation.

We define an automatic pipeline as follows (see Figure 1). The features typically available of the patient treated in the pre-3D planning era (historical patient) are used as input for separate ML models. Each model is trained beforehand (see next section), to make a prediction that is associated with the aforementioned separate tasks: representative body segmentation ($\mathcal{M}_{\mathrm{S}}^{\mathrm{Body}}$), position of the center of mass of the OAR in three dimensions ($\mathcal{M}_{\mathrm{LR}}^{\mathrm{OAR}}$, $\mathcal{M}_{\mathrm{AP}}^{\mathrm{OAR}}$, $\mathcal{M}_{\mathrm{SI}}^{\mathrm{OAR}}$), and representative OAR segmentation ($\mathcal{M}_{\mathrm{S}}^{\mathrm{OAR}}$). In the following, we refer to these metrics regarding the (3D) phantom as "3D metrics".

The CT corresponding to the body predicted to be best, called "receiver", is retrieved, and each OAR is "resected" from it by setting its voxel values to Hounsfield units that represent generic abdominal soft tissue (78 as done by a state-of-the-art CT-based phantom construction method[7]). Subsequently, for each OAR, the segmentation predicted to be most similar is retrieved, and the respective OAR "transplanted" into the receiver CT from the corresponding "donor" CT, at the predicted center of mass position. The resulting CT is the phantom.

2.2 3D metrics and regression

ML models need to be trained before they can be utilized in the pipeline. To this end, we consider 60 recent pediatric cancer patients (age range 2-6 yrs) for which a CT acquired for radiation treatment planning purposes is available. The CTs were collected from the radiation oncology department of the Amsterdam UMC, location AMC, in Amsterdam, the Netherlands, and from the University Medical Center Utrecht/Princess Maxima Center for Pediatric Oncology in Utrecht, the Netherlands. Each CT included the abdomen in a common region of interest between the thoracic 10th vertebra and the sacral 1st vertebra. The median in-plane resolution is 1.0mm × 1.0mm, and the median axial thickness of the CTs is 2.5mm. For each CT, segmentations were prepared for the abdomen (in the common region of interest between vertebrae thoracic 10th and sacral 1st), and for the liver and the spleen (considered as OARs here). In part, the research software ADMIRE (v. 2.3.0, Elekta AB, Stockholm, Sweden) was used to provide initial segmentation estimates. Segmentations were manually checked

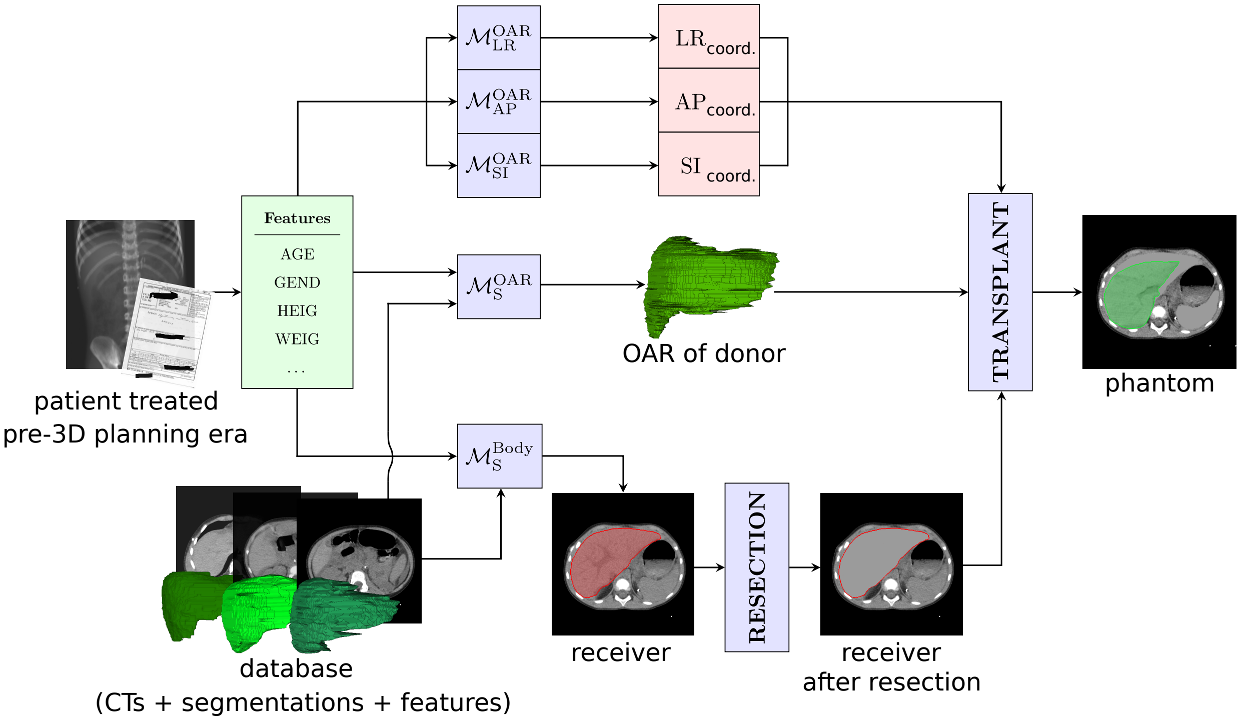
$\mathcal{M}^{\mathrm{OAR}}_{\mathrm{LR}}$
$\mathcal{M}^{\mathrm{OAR}}_{\mathrm{AP}}$
$\mathcal{M}^{\mathrm{OAR}}_{\mathrm{SI}}$
LR coord.
AP coord.
SI coord.
Features
AGE
GEND
HEIG
WEIG
...
patient treated
pre-3D planning era
$\mathcal{M}^{\mathrm{OAR}}_{\mathrm{S}}$
OAR of donor
TRANSPLANT
phantom
$\mathcal{M}^{\mathrm{Body}}_{\mathrm{S}}$
receiver
RESECTION
receiver
after resection
database
(CTs + segmentations + features)

Figure 1. Pipeline of automatic phantom construction. ML models $\mathcal{M}$ are used to predict the 3D metrics. Resection (in red) and transplant (in green) of the liver is shown.

and corrected by experienced radiation treatment technologists, using the software Velocity (v. 3.2.0, Varian Medical Systems, Inc).

We model all 3D metric prediction tasks as supervised regression problems, and prepare a dataset for each. Features to be used in the datasets are measured for these patients based on the availability of data for patients treated historically. These are: age, gender, height, weight, and eight additional measurements that can be taken from historical 2D radiographs (here these are simulated from 3D CT). For each task, we use the 3D imaging (CT scans and segmentations) to define the target variable. For the estimation of OAR position, we measure the distance along each direction between the center of mass of the OAR segmentation and the 2nd lumbar vertebra (in mm), and craft data examples by linking such distances with the respective patient features. To model OAR and body shape similarity, we use the recently introduced surface Dice-Sørensen coefficient on OAR segmentations,[15] with a threshold of 5mm (twice the size of the median CT slice thickness), between all pairs of patients. The features for this type of dataset are defined as absolute difference values of features for pairs of patients. ML models are finally trained and validated on each dataset, using leave-one-out cross-validation.

2.3 Machine learning algorithms and human-designed phantom building criteria

We consider five ML algorithms: Least Angle RegreSsion (LARS),[16] Least Absolute Shrinkage and Selection Operator (LASSO),[17] Random Forest (RF),[18] and Genetic Programming in two versions: its traditional design (GP-Trad)[19,20] and a recent variant of the Gene-pool Optimal Mixing Evolutionary Algorithm (GP-GOMEA).[21,22]

LARS and LASSO can learn interpretable models as linear feature combinations. RF learns non-linear models, but they are not interpretable. GP-Trad and GP-GOMEA can learn non-linear but potentially interpretable models.[23] The first three algorithms are trained with traditional hyper-parameter tuning, the last two algorithms use a strict complexity limitation to avoid overfitting and enable interpretability. Potential interprability of the models obtained by GP-Trad and GP-GOMEA is because these models are computation graphs which can be represented as human-readable mathematical formulas. For these models to be interprable, they need to be sufficiently small in the number of operations they consider (the same holds for the models found by LARS and LASSO), and the compositions of operations need not to be excessively complex.[22,24]

Table 1. Mean training and test MAEs for the ML algorithms on the different OAR-specific regression tasks. Standard deviation is reported in subscript. Results in bold are best in that no other method delivers significantly better ones. The letter "S" stands for segmentation retrieval.

		Body	Liver				Spleen				#Best
		S	LR	AP	SI	S	LR	AP	SI	S	
Training	LARS	$16.63_{0.08}$	$7.89_{0.24}$	$4.27_{0.14}$	$7.10_{0.33}$	$17.02_{0.07}$	$\mathbf{3.99}_{0.21}$	$7.21_{0.13}$	$7.71_{0.19}$	$15.39_{0.10}$	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	LARS	$23.70_{15.31}$	$8.54_{6.83}$	$\mathbf{4.82}_{4.57}$	$9.10_{5.33}$	$\mathbf{38.90}_{17.84}$	$5.18_{3.33}$	$\mathbf{7.42}_{8.06}$	$8.52_{7.48}$	$35.12_{14.51}$	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
	HC1	$37.54_{10.82}$	$8.88_{6.90}$	$\mathbf{4.83}_{4.70}$	$9.10_{6.13}$	$43.19_{8.35}$	$5.65_{3.68}$	$7.96_{7.75}$	$9.78_{8.20}$	$37.58_{8.53}$	1
	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
	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
	sCT	$37.13_{22.00}$	$15.29_{10.02}$	$7.44_{5.91}$	$11.45_{9.2}$	$42.72_{18.30}$	$4.85_{3.43}$	$\mathbf{7.40}_{8.32}$	$\mathbf{7.84}_{9.37}$	$\mathbf{32.08}_{12.28}$	3

We compare the accuracy of predictions of the ML models with the application of state of the art human-designed phantom building criteria. We simulate the construction of a phantom with those criteria using our database. The first human-designed criterion (HC1), selects the CTs that share the same (category bin of) age and gender with the historical patient, and take, for each subtask, the mean of the 3D metrics measured on the selected CTs (i.e., as if a phantom were to be made that is an average anatomy of the selected CTs).[6,25] The second criterion (HC2) works similarly to HC1, but partitions the CTs into bins of gender, height, and weight.[7]

As control method, we also consider picking a CT completely at random (RAND). Lastly, to assess whether there is merit in assembling a new anatomy compared to selecting a single CT, we also consider a single CT selection strategy (sCT) that uses the predictions of ML models to retrieve the CT predicted to have the most accurate center of mass of both liver and spleen positions, in Euclidean norm.

3. RESULTS AND DISCUSSION

Table 1 shows the results in terms of tasks as Mean Absolute Error (MAE) between predicted and true OAR position (for LR, AP, and SI, in mm), and the mean absolute surface Dice-Sørensen coefficient error (%). Between the ML algorithms, GP-GOMEA achieves overall best training and test performance according to pairwise Wilcoxon statistical tests with 95% confidence interval with Bonferroni correction. The human-designed criteria are inferior to the use of ML models in general, confirming that they are too simplistic to capture internal anatomy complexity. Nonetheless, they are typically better than RAND. The use of ML to predict a single CT works well for the spleen, but not for the liver, indicating that defining an overall similarity score remains a hard problem also when ML models are employed. These results markedly indicate the superiority of our approach.

The models learned by GP-GOMEA are interesting in that they can often be interpreted. For example, the model that predicts what spleen segmentation to retrieve is: $0.057 \times \exp(\mathit{age}) \times (\mathit{length\ of\ spinal\ cord})$ (the features and the output were standardized by z-scoring[26]). The exponentiation of age is reasonable because young children grow rapidly (our cohort is 2-6 yrs). The length of the spinal cord (measured in SI between the 12th thoracic vertebra and the 4th lumbar vertebra) is reasonably related as well, since the spleen is located close to the vertebral column.

Lastly, we consider the phantoms from a qualitative perspective. Examples of obtained phantoms are shown in Figure 2. We find that, in 2/3 of the cases, phantoms have limited or no anatomical inconsistencies. However, in the remaining cases, moderate or large inconsistencies are present. These are, e.g., OARs positioned in such a way that they overlap with neighboring OARs (slightly or abudantly), or exceed body boundaries. We believe that extending the database with more CTs will enlarge the anatomical variation upon which the ML models are trained, and consequently lead to more accurate predictions, with smaller chances of creating large anatomical inconsistencies. However, getting more data is not always easy. We will therefore also explore the use of optimization algorithms to correct the inconsistencies with minimal transformations.

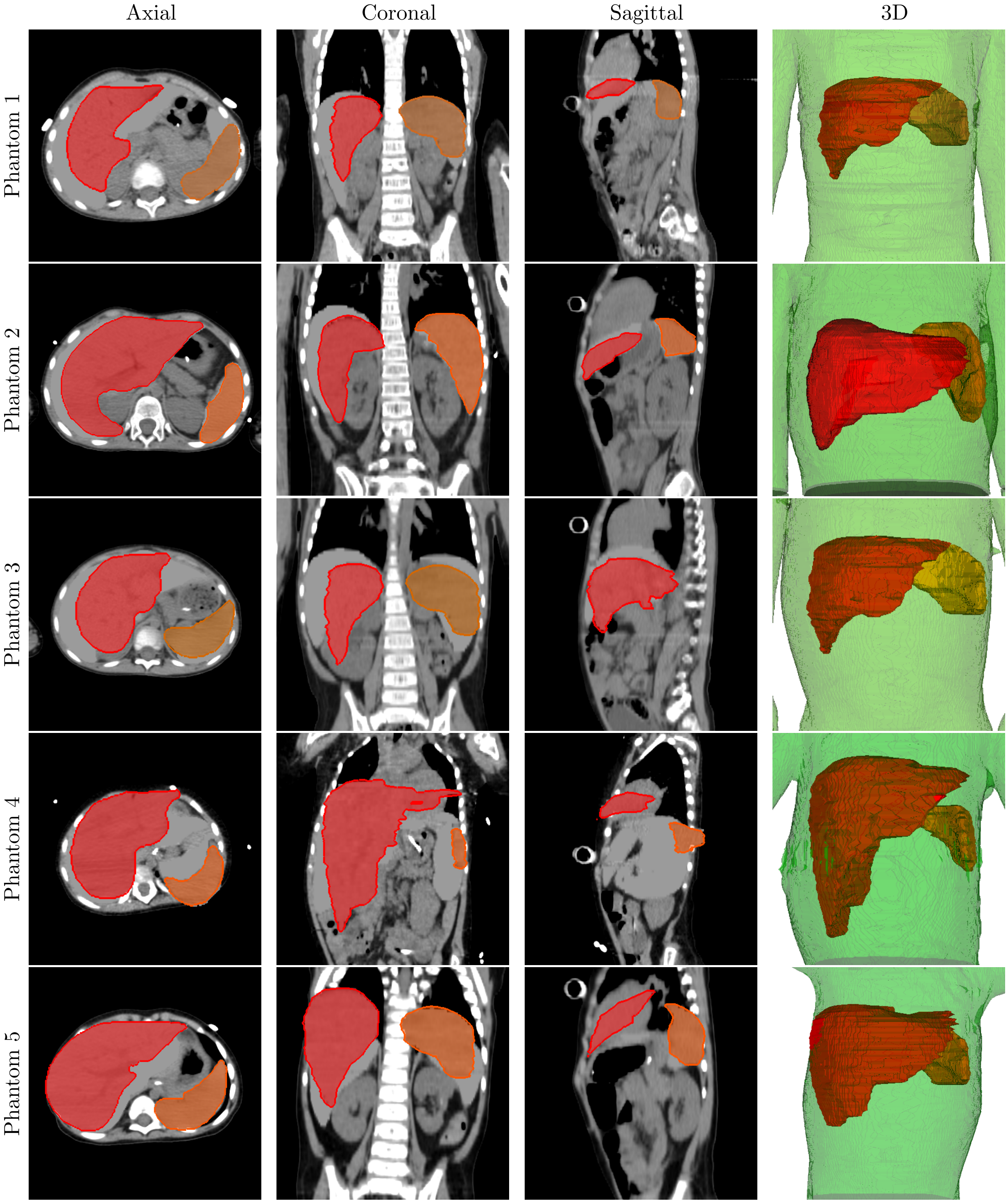

Figure 2. Examples of phantoms constructed with our pipeline where liver (in red) and spleen (in orange) are transplanted. Axial views are in SI, coronal views and 3D views are in AP, sagittal views are in LR.

4. CONCLUSION

We introduced a novel phantom construction method that builds phantoms automatically, based on imaging data and ML. Our method employs an automatic pipeline that, contrary to existing approaches, requires no time-consuming manual intervention, except for the initial effort of 3D imaging collection, segmentation, and the measurement of few features of the historical patient on the radiograph. The pipeline employs ML models that are each specialized to predict a well-defined 3D metric specific to an OAR (e.g., the shape or the position along a direction). This way it is no longer necessary to define a single metric that represents an overall notion of anatomical representativeness. The experimental results show that our method improves upon the lack of individualization that current phantom-building criteria suffer from: for the shape and position of two considered OARs (liver and spleen) and for the shape of the abdomen, our method builds pediatric abdominal phantoms that are significantly more representative than the ones built with current criteria. However, in some cases our phantoms can include anatomical inconsistencies. To this end, we plan to extend the database, and include automatic correction methods. In future work, more types of OARs will be included (e.g., the heart and the kidneys).

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 and the Princess Maxima Center of Pediatric Oncology, 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 & Genetics, Rockville, Maryland, U.S.A., for details on the phantom library of the University of Florida/National Cancer Institute.

REFERENCES

[1] Lee, C., Jung, J. W., Pelletier, C., Pyakuryal, A., Lamart, S., Kim, J. O., and Lee, C., "Reconstruction of organ dose for external radiotherapy patients in retrospective epidemiologic studies," *Physics in Medicine and Biology* **60**(6), 2309–2324 (2015).

[2] Xu, X. G., "An exponential growth of computational phantom research in radiation protection, imaging, and radiotherapy: a review of the fifty-year history," *Physics in Medicine and Biology* **59**(18), R233–R302 (2014).

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

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

[5] Kuczmarski, R. J., Flegal, K. M., Campbell, S. M., and Johnson, C. L., "Increasing prevalence of overweight among us adults: the national health and nutrition examination surveys, 1960 to 1991," *JAMA* **272**(3), 205–211 (1994).

[6] Stovall, M., Donaldson, S. S., Weathers, R. E., Robison, L. L., Mertens, A. C., Winther, J. F., Olsen, J. H., and Boice Jr, J. D., "Genetic effects of radiotherapy for childhood cancer: Gonadal dose reconstruction," *International Journal of Radiation Oncology · Biology · Physics* **60**(2), 542–552 (2004).

[7] Geyer, A. M., O'Reilly, S., Lee, C., Long, D. J., and Bolch, W. E., "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," *Physics in Medicine and Biology* **59**(18), 5225–5242 (2014).

[8] Alziar, I., Bonniaud, G., Couanet, D., Ruaud, J. B., Vicente, C., Giordana, G., Ben-Harrath, O., Diaz, J. C., Grandjean, P., Kafrouni, H., et al., "Individual radiation therapy patient whole-body phantoms for peripheral dose evaluations: method and specific software," *Physics in Medicine and Biology* **54**(17), N375–N383 (2009).

[9] Xie, T., Kuster, N., and Zaidi, H., "Computational hybrid anthropometric paediatric phantom library for internal radiation dosimetry," *Physics in Medicine and Biology* **62**(8), 3263–3283 (2017).

[10] de la Grandmaison, G. L., Clairand, I., and Durigon, M., "Organ weight in 684 adult autopsies: new tables for a caucasoid population," *Forensic Science International* **119**(2), 149–154 (2001).

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

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

[13] Breslow, N., Olshan, A., Beckwith, J. B., and Green, D. M., "Epidemiology of Wilms tumor," *Medical and Pediatric Oncology* **21**(3), 172–181 (1993).

[14] van Dijk, I. W. E. M., Oldenburger, F., Cardous-Ubbink, M. C., Geenen, M. M., Heinen, R. C., de Kraker, J., van Leeuwen, F. E., van der Pal, H. J. H., Caron, H. N., and Koning, C. C. E., "Evaluation of late adverse events in long-term Wilms' tumor survivors," *International Journal of Radiation Oncology · Biology · Physics* **78**(2), 370–378 (2010).

[15] Nikolov, S., Blackwell, S., Mendes, R., De Fauw, J., Meyer, C., Hughes, C., Askham, H., Romera-Paredes, B., Karthikesalingam, A., Chu, C., et al., "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," *arXiv preprint arXiv:1809.04430* (2018).

[16] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., "Least angle regression," *Annals of Statistics* **32**(2), 407–499 (2004).

[17] Tibshirani, R., "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society (Series B)* **58**, 267–288 (1996).

[18] Breiman, L., "Random forests," *Machine Learning* **45**(1), 5–32 (2001).

[19] Koza, J. R., "Genetic programming as a means for programming computers by natural selection," *Statistics and Computing* **4**(2), 87–112 (1994).

[20] Poli, R., Langdon, W. B., and McPhee, N. F., [*A Field Guide to Genetic Programming*], Lulu Enterprises, UK Ltd (2008).

[21] Virgolin, M., Alderliesten, T., Witteveen, C., and Bosman, P. A. N., "Scalable genetic programming by gene-pool optimal mixing and input-space entropy-based building-block learning," in [*Proceedings of the Genetic and Evolutionary Computation Conference*], 1041–1048, ACM (2017).

[22] Virgolin, M., Alderliesten, T., Witteveen, C., and Bosman, P. A. N., "Improving model-based genetic programming for symbolic regression of small expressions," *arXiv preprint arXiv:1904.02050* (2019).

[23] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D., "A survey of methods for explaining black box models," *ACM Computing Surveys (CSUR)* **51**, 93:1–93:42 (Aug. 2018).

[24] Virgolin, M., Alderliesten, T., Bel, A., Witteveen, C., and Bosman, P. A. N., "Symbolic regression and feature construction with GP-GOMEA applied to radiotherapy dose reconstruction of childhood cancer survivors," in [*Proceedings of the Genetic and Evolutionary Computation Conference*], 1395–1402, ACM (2018).

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

[26] Jain, A., Nandakumar, K., and Ross, A., "Score normalization in multimodal biometric systems," *Pattern Recognition* **38**(12), 2270–2285 (2005).