Deep learning dose prediction for IMRT of esophageal cancer: The effect of data quality and quantity on model performance

Ana M. Barragán-Montero^{a,*}, Melissa Thomas^{b,c}, Gilles Defraene^b, Steven Michiels^a, Karin Haustermans^{b,c}, John A. Lee^a, Edmond Sterpin^{a,b}

^a UCLouvain – Institut de Recherche Expérimentale et Clinique – Molecular Imaging Radiotherapy and Oncology (MIRO), Brussels, Belgium

^b KU Leuven – Department of Oncology – Laboratory of Experimental Radiotherapy, Leuven, Belgium

^c University Hospitals Leuven, Department of Radiation Oncology, 3000 Leuven, Belgium

ARTICLE INFO

Keywords:

Radiotherapy
Deep learning
IMRT
Automatic planning
Esophageal cancer

ABSTRACT

Purpose: To investigate the effect of data quality and quantity on the performance of deep learning (DL) models, for dose prediction of intensity-modulated radiotherapy (IMRT) of esophageal cancer.

Material and methods: Two databases were used: a variable database (VarDB) with 56 clinical cases extracted retrospectively, including user-dependent variability in delineation and planning, different machines and beam configurations; and a homogenized database (HomDB), created to reduce this variability by re-contouring and re-planning all patients with a fixed class-solution protocol.

Experiment 1 analysed the user-dependent variability, using 26 patients planned with the same machine and beam setup (E26-VarDB versus E26-HomDB). Experiment 2 increased the training set by groups of 10 patients (E16, E26, E36, E46, and E56) for both databases.

Model evaluation metrics were the mean absolute error (MAE) for selected dose-volume metrics and the global MAE for all body voxels.

Results: For Experiment 1, E26-HomDB reduced the MAE for the considered dose-volume metrics compared to E26-VarDB (e.g. reduction of 0.2 Gy for D95-PTV, 1.2 Gy for Dmean-heart or 3.3% for V5-lungs). For Experiment 2, increasing the database size slightly improved performance for HomDB models (e.g. decrease in global MAE of 0.13 Gy for E56-HomDB versus E26-HomDB), but increased the error for the VarDB models (e.g. increase in global MAE of 0.20 Gy for E56-VarDB versus E26-VarDB).

Conclusion: A small database may suffice to obtain good DL prediction performance, provided that homogenous training data is used. Data variability reduces the performance of DL models, which is further pronounced when increasing the training set.

1. Introduction

Despite the advances in radiotherapy, treatment planning is still one of the bottlenecks in clinical practice. In order to achieve the optimal radiation dose distribution for a certain patient, multiple interactions between the medical physicist or dosimetrist and the physician are needed, accompanied by an iterative manual tuning of the dose-volume objectives. This results in several hours of trial-and-error manual work. The variability in clinical expertise, together with the time limit imposed by tight clinical schedules, can sometimes lead to suboptimal radiotherapy plans that can comprise the treatment outcome [1,2]. Automating and standardizing the full treatment planning process is,

therefore, the key to improve clinical practice and ensure the best possible treatment for each patient.

Several strategies have been proposed to automate this manual process, the most popular ones being multi-criteria optimization (MCO) [3], wish-list iterative planning [4], and knowledge-based planning (KBP) [5]. The latter aims at extracting information and/or learning from previous clinical plans, so that the knowledge contained in clinical databases is used to generate the current treatment plan. The first KBP studies [5,6] focused on predicting dose-volume histogram (DVH) metrics for the new patient, by extrapolating information extracted from similar patients contained in a database of prior clinical plans [7–12]. The similarity between the new patient and those in the database was

* Corresponding author.

E-mail address: ana.barragan@uclouvain.be (A.M. Barragán-Montero).

<https://doi.org/10.1016/j.ejmp.2021.02.026>

Received 6 December 2020; Received in revised form 15 February 2021; Accepted 24 February 2021

Available online 10 March 2021

1120-1797/© 2021 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. All rights reserved.

typically measured using pre-selected (i.e. hand-crafted) features, such as the distance from organs to target. The predicted DVH metrics were later used to guide the optimization process in the treatment planning system (TPS). More recently, the rise of deep learning (DL) methods [13,14] has led to an alternative approach where the database of prior plans is used instead to train a DL model that directly predicts a three-dimensional dose distribution for the new patient [15–22]. DL is a sub-family of machine learning (ML), and its success lies in the fact that, in contrast to classical ML algorithms, they are able to learn without the need of hand-crafted features. This allows us to extract the most relevant global and local features from the patient's anatomy (input images) to later map each voxel to the optimal dose value. This voxel-wise dose map is then used to guide the optimization process in the TPS and generate the final treatment plan, which is known as the dose mimicking process [18,23,24]. Alternatively, DVH metrics can be computed from the predicted dose distribution and used as objectives for an inverse optimization problem to generate the treatment plan [25–27].

Since the first publications in 2017 [20,24], the research on dose prediction with deep learning has become a very popular topic. Numerous studies have investigated its application to different tumor locations and treatment modalities [17–22,24–26,28–35]. The main DL architectures investigated in these studies have been variants of U-Net (a type of Convolutional Neural Network with a downsampling and an upsampling part [36]) and Generative Adversarial Networks (GANs) [37]. Additionally, different ways to improve the performance of the models (e.g. dense connections [19], dilated convolutions [38], dvh-loss functions [39] or extra input channels [17,38]) have been explored. However, little attention has been paid to the influence of the training database on the model's performance. Provided that DL is a fully data-driven technology, the quality of the database used to train the model can significantly influence the results. In computer science, the term Garbage In Garbage Out (GIGO) describes the fact that a model trained with a low quality database will produce faulty outputs. A high quality database should then contain a large number of properly labelled samples that are representative of the problem to solve. For dose prediction problems, this implies a sufficiently large database to cover the patient population distribution (e.g. anatomical variability, tumor size, ...), where each treatment plan ensures the optimal radiation dose distribution for each patient according to specific clinical goals.

For computer vision tasks, the existing databases (e.g. ImageNet) contain thousands to millions of natural images (e.g. animals, objects, ...), and it has been shown that the model's performance increases logarithmically based on volume of training data size [40]. However, in the medical field, reaching a database of 50 patients with well-curated images and annotations is already a challenge for most institutions [41]. Given the ability of current state-of-the-art DL architectures used for dose prediction to learn from small databases [36], the unwritten rule is that a minimum of 50 to 100 patients is enough to build a good model. Hence, published studies so far trained models with a retrospective clinical database, containing a number of patients in the range of 80 to 200 [15–22]. However, it is unclear how the clinical variability in a retrospective database (e.g. different planners, treatment machines, physicians, ...) and the size of the database influence the results.

The present study aims to switch the attention from model to data, by analysing the effect of data quantity and quality in the performance of DL models for dose prediction, in the specific framework of neoadjuvant radiotherapy treatment for esophageal cancer. This experimental study compares the performance of a DL model for dose prediction trained with a retrospective clinical database, including typical variability (such as different physician contourers, physicist planners, treatment machines and planning class-solutions), with a homogenized database, specifically created to reduce all possible variabilities. Additionally, the effect of the size of the database was analysed.

2. Materials and methods

2.1. Patient database

Two different databases were used in this study, referred to as the “variable database” (VarDB) and “homogenized database” (HomDB).

The variable database was extracted retrospectively from our patient directory, and contained 56 patients with esophageal cancer treated with intensity modulated radiotherapy (IMRT) at our institution, from 2016 to 2020. Since the treatment protocol has evolved over time, this database (Fig. 1, left) contained three different treatment machines (Clinac 2100C/D, TrueBeam, and Halcyon, Varian Medical Systems, Palo Alto CA) and different beam configurations (from 5 to 9 coplanar beams) and beam energies (6 or 10 MV). In addition, it involved different physicians and medical physicists for contouring and planning, respectively. Dose calculation and optimization was done using the AAA algorithm (version 10.0.28 or 15.6.03), from the treatment planning system (Eclipse, Varian Medical Systems, Palo Alto CA). The version of the treatment planning system evolved during the time frame of the study from Eclipse 13, Eclipse 15.1 and Eclipse 15.6.

The homogenized database (Fig. 1, right) was intentionally built to reduce the variabilities in the variable database, by re-contouring and re-planning the same 56 patients from the variable database, carefully following organ at risk delineation guidelines [42–44] and a fixed planning protocol. The 56 patients were uniformly planned using a seven beam IMRT class-solution on Halcyon (beam configuration at 0°, 30°, 60°, 155°, 220°, 300°, 330° angles) and an updated list of dose constraints [45]. Optimisation was performed by the same observer (medical physicist with 10 years of experience in inverse planning optimization with Eclipse), starting from a fixed set of objectives and weights and clinical judgement by one radiation oncologist, for all plans. Higher priorities were set for the dose constraints in the lungs, since recent studies [46,47] have found more evidence for the correlation between lung dose volume parameters and pulmonary toxicity and survival [48–50]. Hence, all treatment plans in this database were guided towards the lowest possible lung V5, V10, V20, V40 and Dmean. The dose calculation algorithm was AAA version 15.6.03, from the treatment planning system (Eclipse, Varian Medical Systems, Palo Alto CA). The version of Eclipse was the same (15.6) for all HomDB plans.

Detailed information about the contour delineation and the replanning objectives to generate the HomDB can be found in the [Supplementary material](#). For the VarDB plans, the constraints, objectives and priorities evolved over time, but the exact information could not be retrieved due to the retrospective nature of the database.

For both the variable and homogenized databases, the prescribed total radiation dose to the planning target volume (PTV) was 45.0 Gy in fractions of 1.8 Gy. The dose map grid size was the same for both databases and equal to 2.5 mm × 2.5 mm (in plane) × 3mm (slice thickness).

The use of patient data for the study was approved by the Institutional Ethical Review Board of the University Hospitals Leuven (S59667).

3. Model architecture

The DL model architecture was inspired from the popular U-Net [36], a type of fully convolutional network widely used for medical image segmentation and other medical applications [19,51–53]. U-Net type of architectures have gained interest for the prediction of dose distributions in radiotherapy treatments, due to its ability to include both local and global features from the input images (i.e. the patient's anatomy) to generate a pixel-wise or voxel-wise prediction, with its two-dimensional (2D) or 3D variant, respectively [17,20,38]. In our case, we used a variant of the 3D U-Net that has been first developed by Nguyen et al. [19], and included dense connections [54] between the convolutional layers within the same resolution or “hierarchy” in the U-Net. This

Fig. 1. Illustration of the variability, with respect to the number of machines and beams, for the variable database (VarDB) versus the homogenized database (HomDB). Each marker represents a patient in the database, amounting a total of 56 patients in both databases. Note that the figure does not visually show the user-dependent variability (i.e. different physicians for contouring and different planners), which was additionally present in the VarDB, and reduced in the HomDB, since all operations were performed by the same observers.

Hierarchically Densely Connected U-Net or HD U-Net has demonstrated a more efficient feature propagation with respect to the standard U-Net [19]. The HD U-Net has been previously used for dose prediction in head and neck [19] and lung cancer patients [17]. Mean Squared Error (MSE) between the predicted dose and the ground truth dose was used as loss function. The detailed architecture has been described elsewhere [17,19] and it is presented in Fig. 2. The implementation was done in

Python, using dedicated DL libraries (i.e. Tensorflow and Keras).

For our particular case, all dose distributions and contour masks were resampled to a voxel resolution of $3 \times 3 \times 3 \text{ mm}^3$, and the model was trained using image patches of $128 \times 128 \times 96$ voxels. Eleven input channels were used, which comprised 3D binary masks for the available region of interests (ROIs): (1) one channel containing a mask with the prescription dose ($D_{pre} = 45.0 \text{ Gy}$) for voxels inside and 0 for voxels

Fig. 2. Architecture of the model used in this study. Black numbers on the left side of the model represent the volume shape and resolution at a specific hierarchy. Red numbers represent the number of feature maps at a particular layer, starting with the 11 input channels and 16 feature maps. Orange features represent the newly calculated features and trainable parameters to learn, while blue features are copied or max pooled features that do not need trainable parameters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Legend and Operations

outside the PTV; and (2) ten channels containing a mask with 1 for voxels inside and 0 for voxels outside relevant organs at risk for esophageal cancer (i.e. body, heart, lungs, liver, left and right kidney, spinal canal, spinal canal + 3 mm, fundus and left ventricle). The number of starting feature maps was 16.

4. Experiments

To analyse the effect of the quality and size of the database in the model's performance, two different experiments (Experiment, E) were conducted. Fig. 3 illustrates all considered models.

The first experiment (Experiment 1) aimed at investigating the effect of user-dependent variability (i.e. planning and delineation variability), for a given treatment machine and beam configuration. Hence, for this experiment, we used all patients planned with 7 beams for Halcyon (Fig. 1, blue circles), which amounted to a total of 26 patients. Two models were trained and tested with the patients coming from the variable (E26-VarDB) and the homogenized database (E26-HomDB), respectively. The training was performed using 20 patients and following a 5-fold cross-validation procedure, while the testing was done on the remaining 6 patients, using all 5 cross-validation models independently. The goal of this first experiment was to investigate the influence of the user-dependent variability (i.e. planning and delineation) on the model's performance, without being biased by other factors associated with the plan setup, such as the beam configuration and the choice of treatment machine.

The second experiment (Experiment 2) analysed the influence of the database size on the model's performance, for both the variable and the homogenized databases. Starting from the subset of 26 patients used for Experiment 1, the size of the training database was increased and decreased with groups of 10 patients. When decreasing the database, only one extra model was created, using 16 patients (E16-VarDB vs E16-HomDB). When increasing the database, three extra models were created until covering the full dataset for both the variable and homogenized databases, which were trained and tested with 36 (E36-VarDB and E36-HomDB), 46 (E46-VarDB and E46-HomDB), and 56 patients (E56-VarDB and E56-HomDB), respectively.

The current state-of-the-art in our clinic is to treat the esophageal cancer patients using 7 beams with Halcyon. Hence, for all experiments, the test is always performed in the same set of 6 patients (planned with 7 beams for Halcyon), since these are the plans we want to be able to predict for future patients. The training was done following a k-fold cross-validation process, which consists in splitting the training data into k subsets, to later run k models where one subset is used for validation (a different one each time) and the remaining k-1 subsets are used for training. A 5-fold cross-validation was used in our case, where the five models obtained were later used to predict the dose on the 6 patients from the test set. For each fold, the network was trained during 250 epochs and the weights corresponding to the epoch with the lowest validation loss were selected as the final model. All these operations were performed on an NVIDIA GeForce RTX 2080 Graphical Processing Unit (GPU) with 11 GB dedicated RAM. The training time for each model was depending on the number of patients used, taking up to 1.6, 2.2, 3.4, 4.7, and 6 h for E16, E26, E36, E46, and E56, respectively, for each fold in the cross-validation. The inference time was the same for all models and equal to an average value of 24 s.

The accuracy of the models was evaluated by analysing the difference between relevant dose-volume histogram (DVH) metrics for the predicted and the ground truth doses, using either box plots or the mean absolute error (MAE) for each of the considered metrics. In order to have a single value that allows us to rate each model with respect to the others, an average composite error was computed by averaging the

absolute errors from a set of DVH metrics including: PTV [D95, D5], lungs [Dmean, V5, V10, V20, V30, V40], heart [Dmean, V40], liver [Dmean, V20, V30], kidneys [Dmean], spinal canal + 3 [D2], left ventricle [Dmean], and fundus [Dmean]. The absolute error for the dose metrics was expressed as a percentage of the prescription dose, while the error for the volume metrics was expressed as volume percentage. In addition, the MAE between the predicted dose and the ground truth for all voxels in the body contour was computed.

5. Results

5.1. Experiment 1

Fig. 4 shows the results for Experiment 1 (model E26-VarDB versus E26-HomDB) on the test set, comparing the error for relevant DVH metrics for the model trained on the variable and the homogenized database, respectively. Although the median value is similar for both models in most of the organs, the box plots show a reduction of the error for the HomDB database, since they are narrower than those for E26-VarDB. This is confirmed by looking at the MAE of the DVH metrics (Table 1). On the one hand, the error on the PTV dose coverage (D95) for the test set decreased from $2.03 \pm 1.37\%$ of the prescribed dose to $1.59 \pm 1.41\%$, for the VarDB and HomDB models, respectively. On the other hand, the error for the considered organs at risk also improved for most DVH metrics. For instance, the error for the heart Dmean on the test set was reduced by 2.7% of the prescribed dose (1.2 Gy) for the HomDB model, in comparison with the VarDB model. The improvement on the predicted Dmean in the test set for other organs, such as the lungs or liver, was smaller (around 0.2% of the dose prescription). Table 1 also summarizes the results from the cross-validation, which shows an even larger difference between the VarDB and HomDB models for some specific cases. In particular, the HomDB model outperformed the VarDB model for the lung V5, which improved by 3.3% (absolute volume).

The composite error (Table 2) also reflects an overall improvement of the predicted DVH metrics for the HomDB model versus the VarDB model. For cross-validation, the error decreased around 1.3%, while for the test the reduction was slightly smaller (0.6%). Although the HomDB model clearly outperforms the VarDB model for predicting the DVH metrics, the MAE of all voxels within the body contour (Table 3) only showed a negligible improvement (less than 0.05 Gy).

Fig. 5 shows an example for the DVH (for the PTV and relevant organs) of a selected patient from the test set, for both the VarDB (top) and HomDB models (bottom). For this patient, the highest difference can be observed in the DVH for the heart, where the HomDB model clearly reduced the prediction error.

5.2. Experiment 2

For the models trained with the homogenized database, the MAE for all voxels within the body contour (Table 3) shows a trend to improve during cross-validation when increasing the size of the training database with respect to Experiment 1 (E36, E46, and E56), i.e. a decrease of 0.13 Gy for the E56-HomDB model with respect to the E26-HomDB. For the test set, the improvement is slightly smaller, with a reduction of 0.08 Gy and 0.05 Gy for the E46-HomDB and E56-HomDB models, respectively. In contrast, for the models trained with the variable database, the results show an inverse trend, with an increase in the MAE when sequentially adding more patients to the training set used in Experiment 1. For instance, the MAE for the E56-VarDB model increased by 0.37 Gy with respect to the E26-VarDB for the cross-validation set, and 0.20 Gy for the test set. Decreasing the size of the database with respect to Experiment 1 (E16-HomDB and E16-VarDB) resulted in worse results for the MAE for

Fig. 3. Summary of all considered models, trained with different subsets of patients from the variable (VarDB) and homogenized (HomDB) databases. Each marker represents a patient, the color and the shape defines the beam setup and the treatment machine, respectively. Experiment 1 analysed the user-dependent variability, for a given treatment machine and beam setup. Hence, it involved two models: E26-VarDB versus E26-HomDB. Experiment 2 analysed the size of the database, and involved four extra models for each database: E16-VarDB versus E16-HomDB, E36-VarDB versus E36-HomDB, E46-VarDB versus E46-HomDB, and E56-VarDB versus E56-HomDB.

Fig. 4. Comparison between model E26-VarDB and E26-HomDB (Experiment 1) for the evaluation on the test set (including all five cross-validation models). Box plots for the difference between relevant DVH metrics for the real and predicted doses on the PTV volume and the organs included in the model's input channels (i.e., lungs, heart, spinal canal + 3, liver, right and left kidney, left ventricle and fundus). VarDB refers to the variable database while HomDB refers to the homogenized database.

both cross-validation and testing (Table 3).

The MAE for HomDB models was always better than that of VarDB models, except for E16 testing results.

The error for relevant DVH metrics on the test set for the PTV and a few selected organs is shown with box plots in Fig. 6. For the homogenized database (Fig. 6, right), a slight improvement can be observed for the lungs and heart Dmean when increasing the size of the training set with respect to Experiment 1. However, for the rest of the plotted DVH metrics (PTV D95, Dmean liver), there is no clear trend. For the variable database (Fig. 6, left), increasing the size of the training set did not result in an improvement of the results for none of the plotted metrics. For the lungs and liver Dmean, the prediction error gradually increased from the E26-VarDB to the E56-VarDB model.

Fig. 6 also shows that decreasing the size of the database resulted in wider box plots for the OARs in both the variable (E16-VarDB) and the homogenized database (E16-HomDB). In contrast, the PTV D95 presented narrower boxplots when using this smaller patient group, especially for the homogenized database.

Table 1 presents a more complete overview of the MAE for the selected DVH metrics, for both the models trained with the variable and homogenized databases, which can be summarized in the average composite error presented in Table 2. Please note that the comparison of the cross-validation results should be done carefully, since different validation sets were used in E16 (2 patients), E26 (4 patients), and E36, E36, E56 (6 patients) models (Fig. 3). For the homogenized database, the composite error is slightly better for the patients with bigger training set

Table 1

Mean absolute error (MAE) and its standard deviation (mean $\pm$ SD) for relevant DVH metrics on the PTV and on several organs for cross-validation (average prediction on the validation set for all five folds), and testing (average prediction on the test set for all five folds), for all models from Experiment 1 and 2. The values are expressed as percentage of the prescription dose ($D_{pre} = 45.0$ Gy) for the metrics reporting the dose received by x% of volume (D_x), and as absolute difference for the metrics reporting the volume (in %) receiving a dose of y Gy (V_y).

Mean absolute error for DVH metrics						
		Cross-validation (mean $\pm$ SD)			Testing (mean $\pm$ SD)	
		VarDB	HomDB		VarDB	HomDB
PTV	D_{95} (% D_{pre})	E16	2.17 $\pm$ 0.99	1.15 $\pm$ 0.86	3.12 $\pm$ 1.80	1.39 $\pm$ 1.05
		E26	1.87 $\pm$ 0.85	1.45 $\pm$ 1.28	2.03 $\pm$ 1.37	1.59 $\pm$ 1.41
		E36	1.68 $\pm$ 1.40	0.95 $\pm$ 0.77	1.68 $\pm$ 2.35	0.95 $\pm$ 0.99
		E46	1.63 $\pm$ 1.19	1.11 $\pm$ 0.81	2.81 $\pm$ 2.30	1.13 $\pm$ 0.76
		E56	1.55 $\pm$ 1.30	1.98 $\pm$ 1.03	2.77 $\pm$ 2.21	1.93 $\pm$ 1.15
	D_5 (% D_{pre})	E16	1.37 $\pm$ 0.81	1.57 $\pm$ 0.95	1.36 $\pm$ 0.67	1.13 $\pm$ 0.84
		E26	2.87 $\pm$ 1.90	2.11 $\pm$ 2.05	2.94 $\pm$ 2.01	1.96 $\pm$ 1.73
		E36	1.77 $\pm$ 1.25	1.14 $\pm$ 0.77	1.77 $\pm$ 0.91	1.14 $\pm$ 0.63
		E46	2.78 $\pm$ 2.16	1.63 $\pm$ 1.19	2.48 $\pm$ 2.05	1.54 $\pm$ 1.19
		E56	2.24 $\pm$ 1.24	1.43 $\pm$ 1.03	1.95 $\pm$ 1.10	1.38 $\pm$ 0.87
Liver	D_{mean} (% D_{pre})	E16	2.71 $\pm$ 3.08	1.58 $\pm$ 1.10	2.38 $\pm$ 1.97	2.19 $\pm$ 1.58
		E26	2.26 $\pm$ 1.66	1.59 $\pm$ 1.19	1.86 $\pm$ 1.34	1.79 $\pm$ 1.05
		E36	2.96 $\pm$ 2.17	1.36 $\pm$ 0.94	2.96 $\pm$ 1.30	1.36 $\pm$ 0.92
		E46	3.75 $\pm$ 2.35	1.69 $\pm$ 0.96	3.99 $\pm$ 1.81	2.00 $\pm$ 0.96
		E56	3.22 $\pm$ 2.37	1.64 $\pm$ 1.03	4.02 $\pm$ 2.04	2.10 $\pm$ 1.32
	D_{mean} (% D_{pre})	E16	6.76 $\pm$ 4.97	3.15 $\pm$ 2.53	5.30 $\pm$ 3.24	3.59 $\pm$ 1.99
		E26	6.49 $\pm$ 4.04	3.50 $\pm$ 2.09	5.39 $\pm$ 3.75	2.69 $\pm$ 1.82
		E36	5.17 $\pm$ 4.05	3.75 $\pm$ 2.72	5.17 $\pm$ 3.57	3.75 $\pm$ 1.90
		E46	6.66 $\pm$ 3.98	3.31 $\pm$ 2.87	5.07 $\pm$ 3.52	3.02 $\pm$ 1.56
		E56	5.96 $\pm$ 4.23	2.82 $\pm$ 1.75	3.96 $\pm$ 3.02	2.35 $\pm$ 1.59
Heart	V_{40} (%)	E16	2.78 $\pm$ 2.52	1.55 $\pm$ 0.91	1.70 $\pm$ 0.90	1.67 $\pm$ 0.89
		E26	1.75 $\pm$ 1.12	1.52 $\pm$ 1.19	1.13 $\pm$ 0.56	1.64 $\pm$ 1.03
		E36	2.01 $\pm$ 2.48	0.92 $\pm$ 0.67	2.01 $\pm$ 1.16	0.92 $\pm$ 1.02
		E46	2.48 $\pm$ 1.99	0.88 $\pm$ 0.80	1.75 $\pm$ 0.87	1.24 $\pm$ 1.12
		E56	2.23 $\pm$ 1.50	0.86 $\pm$ 0.74	1.39 $\pm$ 0.64	1.39 $\pm$ 0.92
	D_2 (% D_{pre})	E16	5.07 $\pm$ 3.51	4.13 $\pm$ 3.83	6.76 $\pm$ 3.30	5.03 $\pm$ 3.25
		E26	4.70 $\pm$ 3.71	4.63 $\pm$ 2.88	5.26 $\pm$ 3.41	4.08 $\pm$ 3.33
		E36	5.32 $\pm$ 4.70	4.68 $\pm$ 3.35	5.32 $\pm$ 3.65	4.68 $\pm$ 3.32
		E46	8.66 $\pm$ 5.85	3.54 $\pm$ 2.01	9.75 $\pm$ 3.45	5.91 $\pm$ 4.21
		E56	7.27 $\pm$ 4.37	4.48 $\pm$ 3.07	7.62 $\pm$ 4.03	5.84 $\pm$ 4.24
SC + 3 mm	D_{mean} (% D_{pre})	E16	2.55 $\pm$ 1.71	1.42 $\pm$ 1.25	1.29 $\pm$ 0.85	1.18 $\pm$ 0.86
		E26	2.81 $\pm$ 2.20	1.53 $\pm$ 0.92	1.33 $\pm$ 1.07	1.07 $\pm$ 0.72
		E36	2.31 $\pm$ 1.66	1.34 $\pm$ 1.21	2.31 $\pm$ 1.17	1.34 $\pm$ 0.57
		E46	2.83 $\pm$ 1.93	1.35 $\pm$ 1.26	1.86 $\pm$ 1.15	0.86 $\pm$ 0.65
		E56	2.54 $\pm$ 1.87	1.35 $\pm$ 1.26	2.08 $\pm$ 1.35	0.75 $\pm$ 0.63
	V_5 (%)	E16	5.62 $\pm$ 3.39	3.10 $\pm$ 3.18	3.11 $\pm$ 2.67	4.53 $\pm$ 5.02
		E26	6.25 $\pm$ 4.72	2.98 $\pm$ 2.41	3.71 $\pm$ 2.61	3.37 $\pm$ 3.03
		E36	6.44 $\pm$ 5.68	2.99 $\pm$ 2.66	6.44 $\pm$ 2.43	2.99 $\pm$ 3.14
		E46	6.20 $\pm$ 5.50	2.96 $\pm$ 2.03	2.71 $\pm$ 2.32	3.4 $\pm$ 3.14
		E56	6.63 $\pm$ 6.65	2.90 $\pm$ 2.53	2.88 $\pm$ 2.28	3.27 $\pm$ 3.07
Lungs	V_{20} (%)	E16	4.78 $\pm$ 4.75	4.15 $\pm$ 2.73	2.71 $\pm$ 2.50	2.19 $\pm$ 1.79
		E26	5.57 $\pm$ 3.25	3.33 $\pm$ 2.17	2.68 $\pm$ 2.13	3.09 $\pm$ 2.07
		E36	4.65 $\pm$ 3.76	3.18 $\pm$ 2.64	4.65 $\pm$ 2.13	3.18 $\pm$ 1.37
		E46	6.50 $\pm$ 4.36	2.88 $\pm$ 3.18	5.21 $\pm$ 3.03	1.75 $\pm$ 1.31
		E56	5.45 $\pm$ 3.84	3.06 $\pm$ 2.50	5.23 $\pm$ 3.60	2.27 $\pm$ 1.33

(e.g. 2.28 $\pm$ 2.00% for E56-HomDB versus 2.33 $\pm$ 2.18% for E26-HomDB for the test set), while the models trained with the variable database present a decreased accuracy when adding more training patients (e.g. 3.39 $\pm$ 2.66% for E56-VarDB on the test set versus 2.95 $\pm$ 2.86% for E26-VarDB).

6. Discussion

The present study aimed at analysing the influence of the size and quality of the training database on the DL model's performance, in the framework of dose prediction for esophageal cancer patients. For this purpose, two different experiments were designed, which compared models that were trained using a variable and a homogenized database independently (see Section 2). It is important to note that the data

quality for DL models refers to how well your training set represents your evaluation set (i.e. validation/test set). Thus, for our particular problem, a higher-quality database (i.e. the homogenized database, HomDB) is defined as the one having a higher consistency between the training and test set, with respect to both user-dependent variability (delineation and planning) and planning setup (class-solution protocol, treatment machine, beam configuration).

In the first experiment, which aimed at analysing the influence of the user-dependent variability (i.e. planning and delineation variability), the overall performance of the two considered models (E26-VarDB vs E26-HomDB) was very good, achieving a MAE below 1.3 Gy for all voxels within the body contour in the test patients (Table 3). However, the model trained with the homogenized database (E26-HomDB) showed a decrease of the prediction error with respect to the model

Table 2

Average composite error. It was computed by averaging the absolute errors from a set of DVH metrics including: PTV [D95, D5], lungs [Dmean, V5, V10, V20, V30, V40], heart [Dmean, V40], liver [Dmean, V20, V30], kidneys [Dmean], spinal canal + 3 [D2], left ventricle [Dmean], and fundus [Dmean]. The absolute error for the dose metrics was expressed as a percentage of the prescription dose, while the error for the volume metrics was expressed as volume percentage.

		Cross-validation (mean $\pm$ SD)[SPS]code="TW" instruction="Fix table(column) widths"[/SPS]		Testing (mean $\pm$ SD)	
		VarDB	HomDB	VarDB	HomDB
Composite error (%)	E16	3.84 $\pm$ 3.47	2.15 $\pm$ 2.02	2.99 $\pm$ 2.58	2.69 $\pm$ 2.80
	E26	3.69 $\pm$ 3.09	2.40 $\pm$ 2.20	2.95 $\pm$ 2.86	2.33 $\pm$ 2.18
	E36	3.56 $\pm$ 3.28	2.19 $\pm$ 2.14	3.56 $\pm$ 2.87	2.19 $\pm$ 2.07
	E46	4.42 $\pm$ 3.82	2.19 $\pm$ 2.09	3.82 $\pm$ 2.87	2.31 $\pm$ 2.33
	E56	3.93 $\pm$ 3.74	2.11 $\pm$ 1.94	3.39 $\pm$ 2.66	2.28 $\pm$ 2.00

trained with the variable database (E26-VarDB), indicating a more robust model. Specifically for the lungs, where the dose objectives (V5, V10, V20, V40 and Dmean) for the homogenized database were systematically pushed down towards the lowest possible value, the over-performance of the HomDB models is clear (Table 1). These results demonstrate that a higher consistency with respect to the user-dependent factors (delineation and planning) in the training database has a beneficial effect on the performance of the model. Note that this experiment analysed the user-dependent variability as a whole, but could not distinguish between the individual influence of the variability in delineation and planning, respectively, since this information was impossible to retrieve from the clinical data. Further investigation is required to analyse the individual influence of these two factors.

The improved performance of the HomDB model compared to the VarDB one brings up the popular debate about “data mining or data farming” for deep learning applications in the radiation oncology field [55], and reinforces the position for a data farming approach. Data mining refers to the approach in which the data is just extracted from retrospective databases and directly analysed, assuming the data is well-curated and ready to be used. In contrast, data farming aims to shape and build the databases according to the application under development. In this sense, our results highlight the importance of training DL models with well-curated data, since any variability, errors or bias in the training database will be reflected in the results. Hence, we recommend to carefully analyse the clinical requirements for the DL model, and build or “farm” the database accordingly. Moreover, the building of the database may allow to include newly gained clinical insights in the prediction model. For instance, in our particular application for dose prediction in esophageal cancer patients, special attention was put into correctly predicting the dose to the lungs, since recent studies have found a correlation between lung dose metrics, pulmonary toxicity and survival [46,47]. Thus, one needs to have this in mind when building the database and carefully check the optimality of the most relevant dose constraints.

The second experiment investigated the influence of the size of the training dataset on the models performance, by decreasing and increasing the database with respect to Experiment 1. This resulted in a total of five models, including the one generated in Experiment 1: E16, E26, E36, E46, and E56. For the homogenized database models, an improvement trend is observed when increasing the size of the training set from E26 to E56 models for both the global MAE (Table 3) and the DVH metrics (Tables 1 and 2). But surprisingly, despite having increased the training set by a factor 2, the improvement was small (e.g. improvement in MAE smaller than 0.15 Gy). In contrast, for the variable database models, the trend was inverted, and the prediction error increased when adding more patients to the training set (Tables 1–3). This can be explained by the fact that the increase in the number of

training patients for the clinical models goes with an increase in the database heterogeneity, including different treatment machines and beam configurations (Fig. 3). The larger the heterogeneity in the database, the harder it is to accurately predict the dose distributions, which again supports the need of using a consistent and high-quality database to get the most of your DL model. The results from Experiment 2 also showed that reducing the database too much (i.e. models E16-HomDB and E16-VarDB) can also lead to a decreased performance of the DL model.

Regarding the metrics used for evaluation, it is important to discuss the fact that although the MAE is used in most publications about dose prediction studies and can be a general indication of the performance of the model, the final evaluation of the model should be based on clinically relevant metrics for the treatments under study. Indeed, the MAE takes into account a large part of the dose distribution that is outside relevant organs. Thus, in this study, the main conclusions are drawn based on the DVH metrics for relevant organs (Table 2, Fig. 3) in the test set, which included exactly the same patients in all models.

We are aware of the limitations of this study, since an increase in the number of training patients in steps of 10 and a maximum of 56 might seem small, rather than using increments of 50 or 100 to reach a few hundreds patients. The latter would probably give us more solid statistics about the influence of the database size in the performance of the model. However, we believe that an increase of 10 to 20 patients in the training database is the order of magnitude that most clinics can achieve in the short term, especially for well-curated and consistent data as it is the case of the homogenized database. The same applies to the size of the testing set, since again, having a much larger set would help to gather more solid statistics about the accuracy of the models. However, the reality is that most clinics could only afford to use about 10–20% of the collected data for the testing phase (i.e. around 5 to 10 patients for a database of 50). It is important to note that the model trained with the subset of 26 patients performed already very well, in comparison to previous studies for dose prediction with deep learning. Recently, Zhang et al. [56] developed a dose prediction model using a database of 78 esophageal cancer patients, which achieved a MAE in the body of 3.4% of the prescription dose (i.e. 1.5 Gy for a prescription dose of 45 Gy). Our VarDB models achieved a MAE that was always smaller than 1.3 Gy. But regardless of external comparisons with other models, which might be biased by the difference in databases and DL architectures, the results from the second experiment suggest that blindly increasing the training database will not bring much improvement to the performance of the model. Instead, more sophisticated solutions, such as active learning, should be used to efficiently train DL models. Active learning [57] is a type of iterative supervised learning where the algorithm itself query the user to obtain new data points where they are most needed, in order to build up an optimally balanced training dataset.

Table 3

Mean absolute error (MAE) for all voxels within the body contour corresponding to all models under study (E16, E26, E36, E46, and E56), trained with the variable (VarDB) and homogenized (HomDB) database, averaged over the validation and test sets. The results are expressed as the mean value over all patients for the five cross-validation models, together with the standard deviation in Gy (mean $\pm$ SD).

Mean absolute error (MAE)				
	Cross-validation mean $\pm$ SD (Gy)		Testing mean $\pm$ SD (Gy)	
	VarDB	HomDB	VarDB	HomDB
E16	1.51 $\pm$ 0.30	1.44 $\pm$ 0.27	1.34 $\pm$ 0.42	1.48 $\pm$ 0.40
E26	1.39 $\pm$ 0.39	1.34 $\pm$ 0.33	1.28 $\pm$ 0.43	1.27 $\pm$ 0.32
E36	1.76 $\pm$ 0.54	1.24 $\pm$ 0.30	1.46 $\pm$ 0.37	1.23 $\pm$ 0.33
E46	1.77 $\pm$ 0.48	1.22 $\pm$ 0.31	1.50 $\pm$ 0.42	1.19 $\pm$ 0.35
E56	1.76 $\pm$ 0.50	1.21 $\pm$ 0.27	1.48 $\pm$ 0.44	1.24 $\pm$ 0.32

Another important point to be discussed is how the predicted dose distributions can be used to generate the final treatment plan in an end-to-end automatic planning workflow. As previously mentioned, the two main approaches used so far have been either dose mimicking strategies [18,23,24] or inverse optimization [25–27]. A recently published paper from Babier et al. [58] reported that the quality of the final treatment plan depends on how well the prediction and plan generation models (dose mimicking or inverse optimization) perform together. This demonstrates the importance of evaluating the full pipeline when aiming for clinical implementation of automatic planning workflows based on DL

dose prediction. However, they were unable to isolate the root cause of those effects. Thus, we believe that further research is needed in order to analyse the source of errors for the prediction and plan generation models, independently. In this context, our article brings some light into the possible ways to increase the accuracy of DL methods, by carefully designing homogeneous and high-quality databases. In addition, it is important to mention that very recently, there have been few publications that have investigated how to directly predict the fluence maps with ML/DL models [59–63], without the need of using dose mimicking or inverse optimization for the final plan generation. Future research should focus on further investigating and comparing this novel approach with the existing strategies for automatic planning, including those based on DL dose prediction.

7. Conclusion

To summarize, this study aims to raise awareness of the importance of the training database for deep learning applications, and in particular, for the dose prediction of radiation therapy treatments. The experiments conducted in this study quantified the effect of the data quality and quantity on the model's performance, in the specific framework of dose prediction for esophageal cancer patients treated with IMRT. On the one hand, our results demonstrate that a higher-quality database, with reduced user-dependent variability in delineation and planning, leads to an improvement in the performance of the DL model (E26-HomDB versus E26-VarDB). On the other hand, the results also suggest that,

Fig. 5. Example of the DVH for a selected patient from the test set for Experiment 1. Top figure shows the DVH corresponding to the model trained with the variable database (VarDB), for the predicted (dash line) and ground truth doses (solid line); while the bottom figure shows the DVH for the model trained with the homogenized database (HomDB).

Fig. 6. Comparison between all the models trained with an increasing database of 16, 26, 36, 46, and 56 patients (Experiment 2), for the variable (left) and the homogenized (right) databases. The box plots present the results for the evaluation on the test set (including all five cross-validation models). Each box plot shows the difference between relevant DVH metrics for the real and predicted doses on the PTV volume and selected organs (i.e., lungs, heart, spinal canal + 3, liver).

increasing the size of the database may either have a detrimental effect in the model's performance when the newly added data has a low-quality (E26-, E36-, E46-, and E56-VarDB), or either have a negligible positive effect when adding high-quality data (E26-, E36-, E46-, and E56-HomDB). The latter motivates further research to find strategies that allow us to build and expand medical databases in a more efficient way.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Ana Barragán is funded by the Walloon region in Belgium (PRO-THERWAL/CHARP, grant number 7289). Melissa Thomas was supported by Kom op tegen Kanker (Stand up to Cancer), the Flemish cancer society. Gilles Defraene is postdoctoral fellow of the Research Foundation Flanders (FWO, project 1292021 N). Karin Haustermans was a senior clinical investigator at the Research Foundation Flanders (FWO). John A. Lee is a Senior Research Associate with the F.R.S.-FNRS.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ejmp.2021.02.026>.

References

- [1] Moore KL, Schmidt R, Moiseenko V, Olsen LA, Tan J, Xiao Y, et al. Quantifying unnecessary normal tissue complication risks due to suboptimal planning: A Secondary Study of RTOG 0126. *Int J Radiat Oncol Biol Phys* 2015;92:228–35.
- [2] Marcello M, Ebert M, Haworth A, Steigler A, Kennedy A, Joseph D, et al. Association between treatment planning and delivery factors and disease progression in prostate cancer radiotherapy: Results from the TROG 03.04 RADAR trial. *Radiother Oncol* 2018;126:249–56.
- [3] Craft D, McQuaid D, Wala J, Chen W, Salari E, Bortfeld T. Multicriteria VMAT optimization. *Med Phys* 2012;39:686–96. <https://doi.org/10.1118/1.3675601>.
- [4] Breedveld S, Storch PRM, Voet PWJ, Heijmen BJM. iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans. *Med Phys* 2012;39:951–63.
- [5] Ge Y, Wu QJ. Knowledge-based planning for intensity-modulated radiation therapy: A review of data-driven approaches. *Med Phys* 2019;46:2760–75.
- [6] Hussein M, Heijmen BJM, Verellen D, Nisbet A. Automation in intensity modulated radiotherapy treatment planning – A review of recent innovations. *British J Radiol* 2018;91:20180270. <https://doi.org/10.1259/bjr.20180270>.
- [7] Zarepisheh M, Long T, Li N, Tian Z, Romeijn HE, Jia X, et al. A DVH-guided IMRT optimization algorithm for automatic treatment planning and adaptive radiotherapy replanning. *Med Phys* 2014;41:061711.
- [8] Fogliata A, Nicolini G, Bourcier C, Clivio A, De Rose F, Fenoglietto P, et al. Performance of a knowledge-based model for optimization of volumetric modulated Arc therapy plans for single and bilateral breast irradiation. *PLoS One* 2015;10:e0145137.
- [9] Fogliata A, Reggiori G, Stravato A, Lobefalo F, Franzese C, Franceschini D, et al. RapidPlan head and neck model: The objectives and possible clinical benefit. *Radiat Oncol* 2017;12:73.
- [10] Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WFAR. Evaluation of a knowledge-based planning solution for head and neck cancer. *Int J Radiat Oncol Biol Phys* 2015;91:612–20.
- [11] Wu B, Ricchetti F, Sanguineti G, Kazhdan M, Simari P, Jacques R, et al. Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys* 2011;79:1241–7.
- [12] Valdes G, Simone 2nd CB, Chen J, Lin A, Yom SS, Pattison AJ, et al. Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiother Oncol* 2017;125:392–7.
- [13] Boldrini L, Bibault J-E, Masciocchi C, Shen Y, Bittner M-I. Deep learning: A review for the radiation oncologist. *Front Oncol* 2019;9:977.
- [14] Shen C, Nguyen D, Zhou Z, Jiang SB, Dong B, Jia X. An introduction to deep learning in medical physics: Advantages, potential, and challenges. *Phys Med Biol* 2020;65:05TR01..
- [15] Shao Y, Zhang X, Wu G, Gu Q, Wang J, Ying Y, et al. Prediction of three-dimensional radiotherapy optimal dose distributions for lung cancer patients with asymmetric network. *IEEE J Biomed Health Inform* 2020. <https://doi.org/10.1109/JBHI.2020.3025712>.
- [16] Zhou J, Peng Z, Song Y, Chang Y, Pei X, Sheng L, et al. A method of using deep learning to predict three-dimensional dose distributions for intensity-modulated radiotherapy of rectal cancer. *J Appl Clin Med Phys* 2020;21:26–37.
- [17] Barragán-Montero AM, Nguyen D, Lu W, Lin M-H, Norouzi-Kandalan R, Geets X, et al. Three-dimensional dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations. *Med Phys* 2019;46:3679–91. <https://doi.org/10.1002/mp.13597>.
- [18] Fan J, Wang J, Chen Z, Hu C, Zhang Z, Hu W. Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Med Phys* 2019;46:370–81.
- [19] Nguyen D, Jia X, Sher D, Lin M-H, Iqbal Z, Liu H, et al. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. *Phys Med Biol* 2019;64:065020.
- [20] Nguyen D, Long T, Jia X, Lu W, Gu X, Iqbal Z, et al. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning 2017.
- [21] Chen X, Men K, Li Y, Yi J, Dai J. A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning. *Med Phys* 2019;46:56–64. <https://doi.org/10.1002/mp.13262>.
- [22] Kearney V, Chan JW, Haaf S, Descovich M, Solberg TD. DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Phys Med Biol* 2018;63:235022.
- [23] Petersson K, Nilsson P, Engström P, Knöös T, Ceberg C. Evaluation of dual-arc VMAT radiotherapy treatment plans automatically generated via dose mimicking. *Acta Oncol* 2016;55:523–5.
- [24] McIntosh C, Welch M, McNiven A, Jaffray DA, Purdie TG. Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. *Phys Med Biol* 2017;62:5926–44.
- [25] Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. Knowledge-based automated planning with three-dimensional generative adversarial networks. *Med Phys* 2020;47:297–306.
- [26] Automated Treatment Planning in Radiation Therapy using Generative Adversarial Networks 2018;vol. 85:484–99.
- [27] Babier A, Boutilier JJ, Sharpe MB, McNiven AL, Chan TCY. Inverse optimization of objective function weights for treatment planning using clinical dose-volume histograms. *Phys Med Biol* 2018;63:105004.
- [28] Liu Z, Fan J, Li M, Yan H, Hu Z, Huang P, et al. A deep learning method for prediction of three-dimensional dose distribution of helical tomotherapy. *Med Phys* 2019;46:1972–83.
- [29] Song Y, Hu J, Liu Y, Hu H, Huang Y, Bai S, et al. Dose prediction using a deep neural network for accelerated planning of rectal cancer radiotherapy. *Radiother Oncol* 2020;149:111–6.
- [30] Murakami Y, Magome T, Matsumoto K, Sato T, Yoshioka Y, Oguchi M. Fully automated dose prediction using generative adversarial networks in prostate cancer patients. *PLoS One* 2020;15:e0232697.
- [31] Kandalan RN, Nguyen D, Rezaeian NH, Barragán-Montero AM, Breedveld S, Namuduri K, et al. Dose prediction with deep learning for prostate cancer radiation therapy: Model adaptation to different treatment planning practices. *Radiother Oncol* 2020;153:228–35.
- [32] Kearney V, Chan JW, Wang T, Perry A, Descovich M, Morin O, et al. DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. *Sci Rep* 2020;10:11073.
- [33] Ma M, Kovalchuk N, Buyyounouski MK, Xing L, Yang Y. Incorporating dosimetric features into the prediction of 3D VMAT dose distributions using deep convolutional neural network. *Phys Med Biol* 2019;64:125017.
- [34] Kajikawa T, Kadoya N, Ito K, Takayama Y, Chiba T, Tomori S, et al. A convolutional neural network approach for IMRT dose distribution prediction in prostate cancer patients. *J Radiat Res* 2019;60:685–93.
- [35] Guerreiro F, Seravalli E, Janssens GO, Maduro JH, Knopf AC, Langendijk JA, et al. Deep learning prediction of proton and photon dose distributions for paediatric abdominal tumours. *Radiother Oncol* 2020;156:36–42.
- [36] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing; 2015, p. 234–41.
- [37] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. *arXiv [statML]* 2014.
- [38] Zhang J, Liu S, Li T, Mao R, Du C, Liu J. Voxel-level radiotherapy dose prediction using densely connected network with dilated convolutions. *Art Intell Radiat Therapy* 2019;70–7. https://doi.org/10.1007/978-3-030-32486-5_9.
- [39] Nguyen D, McBeth R, Sadeghnejad Barkousaraie A, Bohara G, Shen C, Jia X, et al. Incorporating human and learned domain knowledge into training deep neural networks: A differentiable dose-volume histogram and adversarial inspired framework for generating Pareto optimal dose distributions in radiation therapy. *Med Phys* 2020;47:837–49.
- [40] Sun C, Shrivastava A, Singh S, Gupta A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017. <https://doi.org/10.1109/iccv.2017.97>.
- [41] Willeminck MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing Medical Imaging Data for Machine Learning. *Radiology* 2020;295:4–15.
- [42] Jabbar SK, Hashem SA, Bosch W, Kim TK, Finkelstein SE, Anderson BM, et al. Upper abdominal normal organ contouring guidelines and atlas: a Radiation Therapy Oncology Group consensus. *Pract Radiat Oncol* 2014;4:82–9.
- [43] Feng M, Moran JM, Koelling T, Chughtai A, Chan JL, Freedman L, et al. Development and validation of a heart atlas to study cardiac exposure to radiation following treatment for breast cancer. *Int J Radiat Oncol Biol Phys* 2011;79:10–8.

- [44] Kong FM, Quint L, Machtay M, Bradley J. Atlas for organs at risk (OARS) in thoracic radiation therapy 2013.
- [45] Ajani JA, D'Amico TA, Bentrem DJ, Chao J, Corvera C, Das P, et al. Esophageal and Esophagogastric Junction Cancers, Version 2.2019, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2019;17:855–83.
- [46] Beukema JC, Kawaguchi Y, Sijtsma NM, Zhai T-T, Langendijk JA, van Dijk LV, et al. Can we safely reduce the radiation dose to the heart while compromising the dose to the lungs in oesophageal cancer patients? *Radiother Oncol* 2020;149:222–7.
- [47] Thomas M, Defraene G, Lambrecht M, Deng W, Moons J, Naftaux P, et al. NTCP model for postoperative complications and one-year mortality after trimodality treatment in oesophageal cancer. *Radiother Oncol* 2019;141:33–40.
- [48] Lee HK, Vaporciyan AA, Cox JD, Tucker SL, Putnam JB, Ajani JA, et al. Postoperative pulmonary complications after preoperative chemoradiation for esophageal carcinoma: correlation with pulmonary dose–volume histogram parameters. *Int J Radiat Oncol Biol Phys* 2003;57:1317–22. [https://doi.org/10.1016/s0360-3016\(03\)01373-7](https://doi.org/10.1016/s0360-3016(03)01373-7).
- [49] Wang S-L, Liao Z, Vaporciyan AA, Tucker SL, Liu H, Wei X, et al. Investigation of clinical and dosimetric factors associated with postoperative pulmonary complications in esophageal cancer patients treated with concurrent chemoradiotherapy followed by surgery. *Int J Radiat Oncol Biol Phys* 2006;64:692–9.
- [50] Tucker SL, Liu HH, Wang S, Wei X, Liao Z, Komaki R, et al. Dose-volume modeling of the risk of postoperative pulmonary complications among esophageal cancer patients treated with concurrent chemoradiotherapy followed by surgery. *Int J Radiat Oncol Biol Phys* 2006;66:754–61.
- [51] Liu L, Cheng J, Quan Q, Wu F-X, Wang Y-P, Wang J. A survey on U-shaped networks in medical image segmentations. *Neurocomputing* 2020;409:244–58. <https://doi.org/10.1016/j.neucom.2020.05.070>.
- [52] Wu C, Nguyen D, Xing Y, Barragan A, Schuemann J, Shang H, et al. Improving proton dose calculation accuracy by using deep learning. *Mach Learn: Sci Technol* 2020. <https://doi.org/10.1088/2632-2153/abb6d5>.
- [53] Xing Y, Zhang Y, Nguyen D, Lin M-H, Lu W, Jiang S. Boosting radiotherapy dose calculation accuracy with deep learning. *J Appl Clin Med Phys* 2020. <https://doi.org/10.1002/acm2.12937>.
- [54] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017. <https://doi.org/10.1109/cvpr.2017.243>.
- [55] Mayo CS, Kessler ML, Eisbruch A, Weyburne G, Feng M, Hayman JA, et al. The big data effort in radiation oncology: Data mining or data farming? *Adv Radiat Oncol* 2016;1:260–71. <https://doi.org/10.1016/j.adro.2016.10.001>.
- [56] Zhang J, Liu S, Yan H, Li T, Mao R, Liu J. Predicting voxel-level dose distributions for esophageal radiotherapy using densely connected network with dilated convolutions. *Phys Med Biol* 2020;65:205013.
- [57] Blanch MG. Active Deep Learning for Medical Imaging Segmentation. 2017.
- [58] Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. The importance of evaluating the complete automated knowledge-based planning pipeline. *Phys Med* 2020;72:73–9.
- [59] Wang W, Sheng Y, Wang C, Zhang J, Li X, Palta M, et al. Fluence Map Prediction Using Deep Learning Models – Direct Plan Generation for Pancreas Stereotactic Body Radiation Therapy. *Front. Artif Intell* 2020;3. <https://doi.org/10.3389/frai.2020.00068>.
- [60] Sheng Y, Li T, Yoo S, Yin F-F, Blitzblau R, Horton JK, et al. Automatic planning of whole breast radiation therapy using machine learning models. *Front Oncol* 2019;9. <https://doi.org/10.3389/fonc.2019.00750>.
- [61] Ma L, Chen M, Gu X, Lu W. Deep learning-based inverse mapping for fluence map prediction. *Phys Med Biol* 2020. <https://doi.org/10.1088/1361-6560/abc12c>.
- [62] Li X, Zhang J, Sheng Y, Chang Y, Yin F-F, Ge Y, et al. Automatic IMRT planning via static field fluence prediction (AIP-SFFP): a deep learning algorithm for real-time prostate treatment planning. *Phys Med Biol* 2020;65:175014.
- [63] Lee H, Kim H, Kwak J, Kim YS, Lee SW, Cho S, et al. Fluence-map generation for prostate intensity-modulated radiotherapy planning using a deep-neural-network. *Sci Rep* 2019;9:15671.