Machine Learning: Science and Technology

PAPER • OPEN ACCESS

Generalizability issues with deep learning models in medicine and their potential solutions: illustrated with cone-beam computed tomography (CBCT) to computed tomography (CT) image conversion

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
16 April 2020

REVISED
5 August 2020

ACCEPTED FOR PUBLICATION
24 August 2020

PUBLISHED
1 December 2020

PAPER

Generalizability issues with deep learning models in medicine and their potential solutions: illustrated with cone-beam computed tomography (CBCT) to computed tomography (CT) image conversion

Xiao Liang, Dan Nguyen and Steve B Jiang

Medical Artificial Intelligence and Automation Laboratory and Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, United States of America

E-mail: Steve.Jiang@UTSouthwestern.edu

Keywords: deep learning, medicine, generalizability, transfer learning

Supplementary material for this article is available online

Abstract

Generalizability is a concern when applying a deep learning (DL) model trained on one dataset to other datasets. It is challenging to demonstrate a DL model's generalizability efficiently and sufficiently before implementing the model in clinical practice. Training a universal model that works anywhere, anytime, for anybody is unrealistic. In this work, we demonstrate the generalizability problem, then explore potential solutions based on transfer learning by using the cone-beam computed tomography (CBCT) to computed tomography (CT) image conversion task as the testbed. Previous works only studied on one or two anatomical sites and used images from the same vendor's scanners. Here, we investigated how a model trained for one machine and one anatomical site works on other machines and other anatomical sites. We trained a model on CBCT images acquired from one vendor's scanners for head and neck cancer patients and applied it to images from another vendor's scanners and for prostate, pancreatic, and cervical cancer patients. We found that generalizability could be a significant problem for this particular application when applying a trained DL model to datasets from another vendor's scanners. We then explored three practical solutions based on transfer learning to solve this generalization problem: the target model, which is trained on a target dataset from scratch; the combined model, which is trained on both source and target datasets from scratch; and the adapted model, which fine-tunes the trained source model to a target dataset. We found that when there are sufficient data in the target dataset, all three models can achieve good performance. When the target dataset is limited, the adapted model works the best, which indicates that using the fine-tuning strategy to adapt the trained model to an unseen target dataset is a viable and easy way to implement DL models in the clinic.

1. Introduction

Deep learning (DL) has been increasingly applied in medicine because it can improve the accuracy of diagnosis, prognosis, and treatment decision making by retrieving hidden information from big clinical data, improve efficiency by automating or augmenting clinical procedures, and transfer expertise to less experienced clinicians by learning from experienced clinicians. However, the generalizability of any given DL model must be demonstrated before that model can be implemented in clinical practice (Rajpurkar *et al* 2020). Many researchers train and test their DL models with their own data, but this gives no indication about the model's generalizability to the datasets from other institutions that may have different distributions in the latent space. A better practice is to test the model with an external dataset, as some journals have recently started requiring for published DL research (David *et al* 2020). Although one external dataset is better than none, it still cannot sufficiently demonstrate the model's generalizability. Take, for example, a

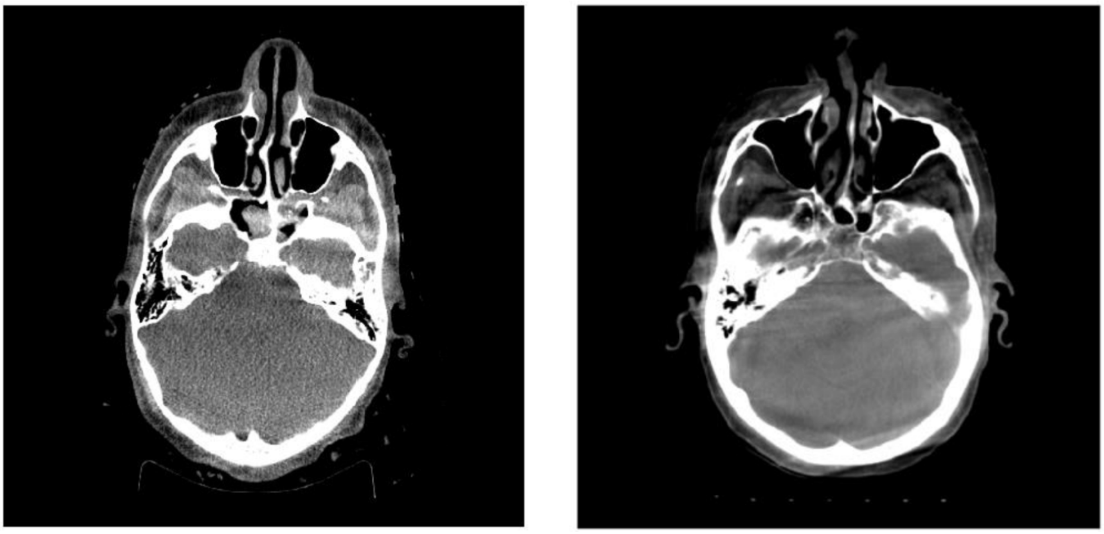


Figure 1. An illustration of different data distributions of CBCT datasets acquired with different scanning protocols from the same vendor's scanner. The CBCT images were acquired from a Varian OBI scanner with total exposure of 150 mAs (left) and 750 mAs (right). The left image is much noisier than the right one. The display window is from −200 HU to 400 HU.

recent study on DL-based breast cancer screening (McKinney *et al* 2020). The investigators trained the model on two datasets from the UK and applied it to a single dataset from the US, then claimed that the DL system can generalize from the UK to the US. However, the US dataset came from only one institution, and 99% of the data were acquired from the same vendor's scanners. The test result for one vendor's scanners and one institution cannot represent all the clinical settings and data distributions for different institutions, vendors' scanners, scanning protocols, and so on. To address the problem of model generalizability, many researchers try to collect as much and as diverse patient data as possible to train a DL model that works in any clinical scenario, anytime, anywhere, for anybody. This ambitious goal seems unrealistic, as it is very challenging, if not impossible, to collect patient data from enough medical institutions to represent all clinical scenarios.

In this paper, we first illustrate the problem of generalizing a DL model, then we explore a few practical solutions based on transfer learning. We also investigate how the dataset scale influences the performance of the proposed solutions. This study uses image conversion from cone-beam computed tomography (CBCT) to computed tomography (CT) as the testbed. In cancer radiation therapy, CBCT images acquired during the treatment course are commonly used for positioning patients, monitoring anatomical changes, segmenting organs, and calculating radiation doses. However, CBCT's image quality is far inferior to CT's, mainly because of scatter and other artifacts. To improve CBCT's image quality, we previously proposed a CycleGAN model to convert CBCT to synthetic CT (sCT) images for patients with head and neck (H&N) cancer (Liang *et al* 2019). Subsequently, Harms *et al* and Kida *et al* also obtained similar results when using CycleGAN for patients with brain, pelvis, and prostate cancers (Harms *et al* 2019, Kida *et al* 2019). These studies all focus on only one or two anatomical sites, and the CBCT images used for training and testing came from the same vendor's scanners. However, CBCT images might be scanned with different protocols or different machines, thus resulting in variant data distributions of image datasets and showing different displays in the same Hounsfield unit (HU) window, as illustrated in figures 1 and 2. Thus, it is unknown whether a DL model trained with one dataset coming from one protocol and machine will work on another dataset from different protocols or machines. It is also unclear whether a DL model trained for one disease site will work on another. In this work, we conduct an extensive set of experiments using our CycleGAN model to perform the CBCT-to-CT conversion task with seven different datasets from different anatomical sites and vendors' scanners. We first demonstrate the problem of model generalizability, taking our own model as exemplar, then we explore different methods to solve this problem.

2. Data

As an IRB-approved retrospective clinical study, we randomly selected seven datasets from the datasets collected as part of the routine clinical procedures for previously treated patients at our institution: two for patients with head and neck cancer (H&N1 and H&N2), three for patients with prostate cancer (Prostate1, Prostate2, and Prostate3), one for patients with cervical cancer (Cervix), and one for patients with pancreatic cancer (Pancreas), as shown in table 1. Each dataset includes both CT and CBCT scans. CT scans in all seven

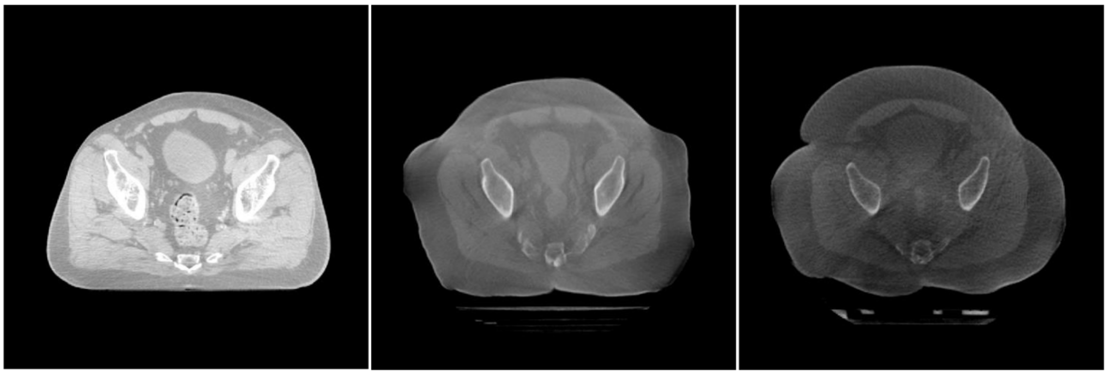


Figure 2. An illustration of different data distributions of CBCT datasets acquired from different vendors' scanners. The CBCT images were acquired with a Varian OBI scanner (left), Elekta XVI (Versa) scanner (middle), and Elekta XVI (Agility) scanner (right). In our institution, the XVI scanners on Versa and Agility machines have different hardware and software configurations, so we treat them as different scanners in this work. The display window is from −600 HU to 200 HU.

Table 1. CBCT image datasets used for experiments.

	Vender	Scanner	Scanning protocol (kVp/mAs^{-1})	No. of patients for training/validation/testing	No. of images for training/validation/testing
H&N1	Varian	OBI	100/150	83/9/23	6640/720/1840
H&N2	Varian	OBI	125/750	11/1/10	880/80/800
Prostate1	Varian	OBI	125/1070	15/3/11	1200/240/880
Prostate2	Elekta	XVI (Versa)	120/1600	39/4/11	2730/280/770
Prostate3	Elekta	XVI (Agility)	120/1600	15/2/10	1035/138/690
Cervix	Elekta	XVI (Agility)	120/1600	15/3/10	1035/207/690
Pancreas	Elekta	XVI (Versa)	120/1600	15/3/10	1050/210/700

datasets were acquired via Philips CT scanner with the same kVp. Information about the CBCT scanners, vendors, and scanning protocols is given in table 1. In summary, these seven CBCT datasets were collected on three scanners of two vendors, using four scanning protocols, and for four anatomical sites.

The H&N1, H&N2, Prostate1, Prostate2, Prostate3, Cervix, and Pancreas datasets contain 115, 22, 29, 54, 27, 28, and 28 patients, respectively. Each dataset is divided into training, testing, and validation sets. For all the following experiments, the validation dataset was used for hyperparameter tuning, and the test dataset was untouched during training and used to assess model performance finally. Their gender and age distribution is shown in figure 3. Each patient has a CT volume and a CBCT volume. For training and validation, each CT volume was resampled to its corresponding CBCT's voxel spacing and then cropped to the CBCT's dimension and number of slices. Each volume in the H&N1, H&N2, Prostate1, Prostate2, Prostate3, Cervix, and Pancreas datasets has 80, 80, 80, 70, 69, 69, and 70 image slices, respectively, with unified dimensions of 512 × 512. We count each 2D image slice as a single sample for the model training. Thus, there are totally 12 890 images used for training in this experiment. Because our proposed CycleGAN did not require paired CT and CBCT images for training (Liang *et al* 2019), there was no need for image registration between CT and CBCT images. However, we performed rigid image registration and deformable image registration (DIR) between CT and CBCT images in the testing dataset through commercial software (Velocity Oncology Imaging Informatics System, Varian Medical Systems, Inc.). An intensity-based DIR algorithm is used in Varian Velocity software package. A study quantifying the accuracy of CBCT-to-CT DIR registration algorithm from Varian Velocity with a physical phantom shows the Dice similarity coefficient were >0.8 and the mean distance to agreement <2 mm for head and pelvis (Wu *et al* 2019). Therefore, deformed CT (dCT) images are used as the ground truth in this study and serve as the basis for evaluating the quality of the sCT images generated from the CBCT images via the CycleGAN model. The numbers of patients and 2D CBCT/CT images in each dataset for training, validation, and testing are shown in table 1.

3. Methods

3.1. DL model and generalizability experiments

This study used the CycleGAN architecture we developed previously (figure 4). It consists of two discriminators (DiscriminatorA and DiscriminatorB) and two generators (GeneratorA and GeneratorB). GeneratorA is to generate sCT from CBCT and GeneratorB is to generate synthesized CBCT (sCBCT) from

CT, while DiscriminatorA is to distinguish between sCT and CT and DiscriminatorB is to distinguish between sCBCT and CBCT. Generators and discriminators are trained together and work against each other to minimize objectives. Besides typical adversarial loss in every GAN architecture, CycleGAN introduces the concept of cycle consistency loss to avoid paired dataset for training. A CBCT image going through GeneratorA and then GeneratorB is supposed to generate an image equal to the CBCT image ($\text{GeneratorB}\left[\text{GeneratorA}\left[\text{CBCT}\right]\right] = \text{CBCT}$), and vice versa ($\text{GeneratorA}\left[\text{GeneratorB}\left[\text{CT}\right]\right] = \text{CT}$). This cycle consistency helps to constrain the mapping from one domain to another. In this experiment, we use the CycleGAN architecture with U-Net for generators and pacthGAN for discriminators to convert CBCT images to sCT images with less noise and other artifacts. For more details, the reader is kindly referred to (Liang *et al* 2019).

Since the data distribution could be different for different anatomical sites, scanning protocols, and vendors' scanners, it is unknown whether a model trained on one dataset works for another dataset. To investigate the generalizability of the CycleGAN model trained on different CBCT datasets, we split the seven datasets into source dataset (H&N1 and H&N2) and target dataset (Prostate1, Prostate2, Prostate3, Cervix and Pancreas) to mimic a situation where CBCT scans come from different clinical environments. A flowchart of the generalizability experiments is shown in figure 5(a). The source dataset consists of the H&N1 and H&N2 datasets, which represented data collected from one circumstance. The model trained on the source dataset is called source model. We then directly applied the source model to five target datasets without re-training and compared the results with that from the source model applied in source datasets to illustrate the generalizability problem that one might encounter. The number of patients used for training the source model and the number of testing patients for each dataset can be referred to table 1.

3.2. Potential solutions to the generalizability problem

To solve the generalizability problem mentioned above, we investigated three potential solutions: target model, combined model, and adapted model, shown in figures 5(b)–(d) respectively. The target model is trained on a target dataset starting from scratch, so each target dataset has its own unique target model. The combined model is trained on the combined source and target datasets starting from scratch. The source model trained on the H&N1 and H&N2 datasets was fine-tuned on the Prostate1, Prostate2, Prostate3, Cervix, and Pancreas datasets separately to get an adapted model for each target dataset. No layers in the architecture were frozen, and all layers were updated in the fine-tuning process. This strategy is commonly used in transfer learning for adapting an old model to a new domain when the training data from the new domain is limited.

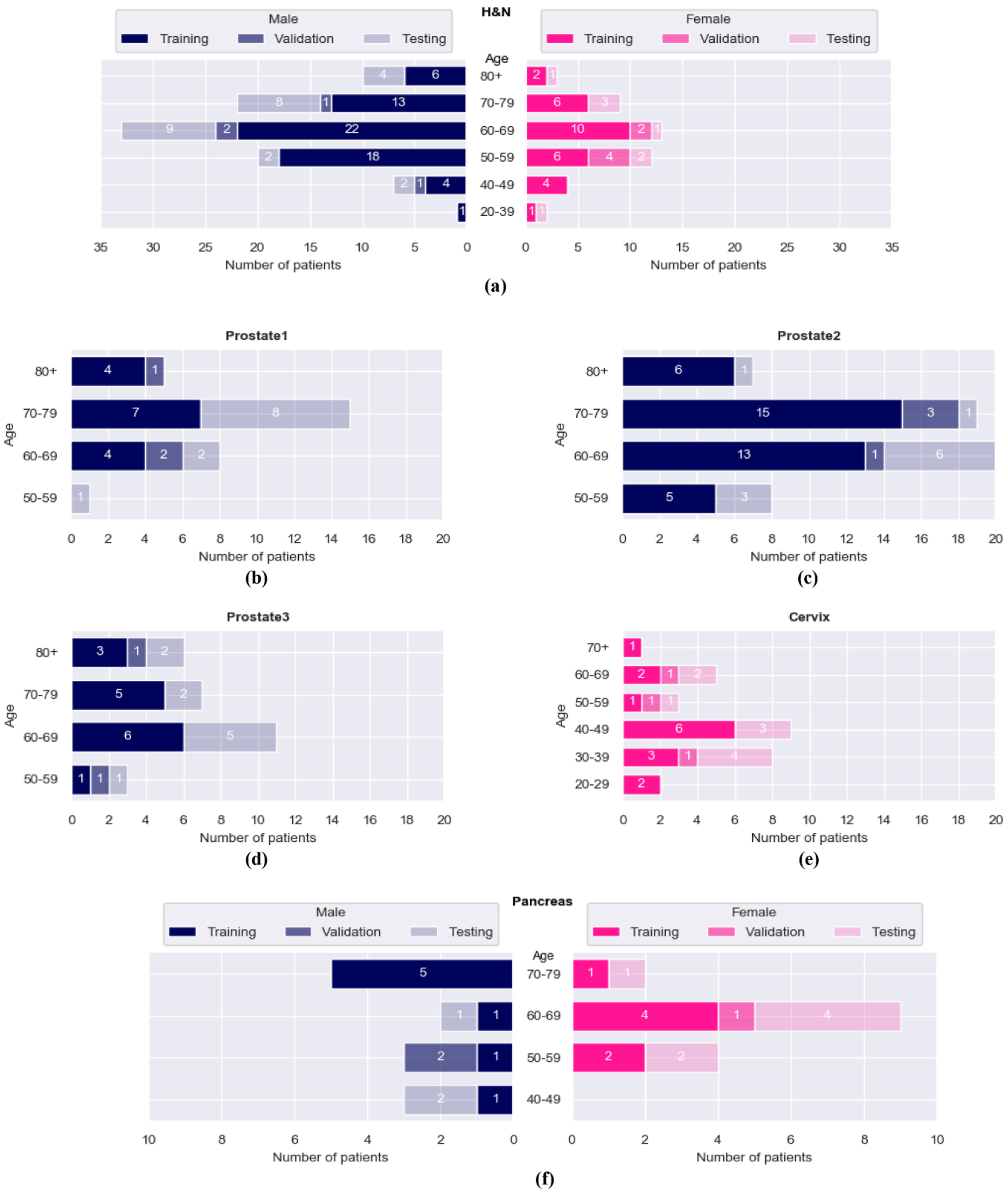
The number of training, validation, and testing patients for each dataset used in these experiments are shown in table 1 except Prostate2 training dataset. For fair comparison, we randomly selected 15 out of 39 patients from Prostate2 dataset for training in experiments 2–4. For all three solutions, all hyper-parameters—including layer architecture, batch size and number of training iterations—were kept the same, except the learning rate. We used the grid search method to find the optimal learning rate for each target dataset (Reed and Marksii 1999, Bengio 2012, Goodfellow *et al* 2016). The search range of the learning rate was from 9×10^{-3} to 2×10^{-6}.

3.3. Dependence of model performance on target dataset size

The performance of the target, combined, and adapted models might depend on the size of the training dataset in the target domain. We conducted an experiment using Prostate2 dataset to investigate this effect, illustrated in figure 5(e). We trained the target, combined, and adapted models by randomly picking 5, 10, 15, 27 and 39 patients from the Prostate2 training dataset. In the target and adapted models, the total number of patients for training was 5, 10, 15, 27, and 39 separately. In the combined model, because the target dataset is combined with the source dataset, the total number of patients for training was 99, 104, 109, 121, and 133 respectively. Eleven patients were used to test all the models, the same as in sections 3.1 and 3.2. We repeated the whole process including training and testing 10 times for statistical analysis. We then analyzed the model performance in correlation with training data size for each model.

3.4. Evaluation methods

In addition to visually evaluating the image quality, we also evaluated the model performance by using mean absolute error (MAE), root mean square error (RMSE), structural similarity index (SSIM), and signal-to-noise ratio (SNR), the four most widely used metrics (Wang *et al* 2004, Al-Obaidi 2015) for measuring the similarity between generated sCT images and the dCT images. Since lack of real ground truth, dCT images serve as the gold standard in this experiment to evaluate image quality by the four similarity measure metrics in the following sections. Assume $sCT(x,y)$ to be the HU value of the sCT images generated



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3. The flowcharts of all the experiments: (a) experiment 1: the generalizability problem of the DL model; (b) experiment 2: train CycleGAN on target datasets starting from scratch to get target models; (c) experiment 3: train CycleGAN on combined source and target datasets starting from scratch to get combined model; (d) experiment 4: fine tune the source model on target datasets to get adapted models; (e) experiment 5: dependence of model performance on target dataset size.

from the CycleGAN models and $dCT(x,y)$ to be the HU value of the deformably registered reference CT images. All the images have the same size [512 512].

MAE measures the difference between two images, as follows:

$$MAE = \frac{1}{n_x n_y} \sum_{0}^{n_x-1} \sum_{0}^{n_y-1} \left| dCT(x,y) - sCT(x,y) \right| . \tag{1}$$

RMSE also measures the average errors between two images, but it gives relatively high weight to large errors, as follows:

$$RMSE = \sqrt{\frac{1}{n_x n_y} \sum_{0}^{n_x-1} \sum_{0}^{n_y-1} \left(dCT(x,y) - sCT(x,y) \right)^2 }. \tag{2}$$

Table 2. Evaluation of similarity measures of sCT (source model) against dCT on two source datasets: (1) H&N1 with 23 testing patients; (2) H&N2 with 10 testing patients. The mean and standard deviation (SD) of MAE, RMSE, SSIM, and SNR of sCT images were calculated using dCT images as reference.

Source dataset	Source model	MAE (HU)	RMSE (HU)	SSIM (%)	SNR (dB)
H&N1	sCT	36.81 ± 9.13	94.14 ± 14.57	86.11 ± 3.91	19.26 ± 1.67
H&N2	sCT	31.75 ± 7.27	83.42 ± 13.25	86.92 ± 3.90	20.50 ± 1.43

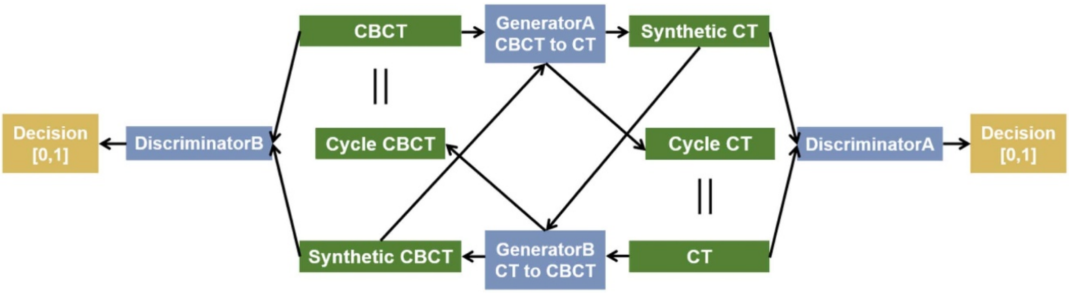


Figure 4. MAE, RMSE, SSIM, and SNR scores of sCT (source model) against dCT on two source datasets: (1) H&N1 with 23 testing patients; (2) H&N2 with 10 testing patients.

SSIM measures the similarity between two images based on the human visual system, (Wang *et al* 2003) that is

$$SSIM = \frac{[(2\mu_1\mu_2 + c_1)(2\sigma_{12} + c_2)]}{[(\mu_1^2 + \mu_2^2 + c_1)(\sigma_1^2 + \sigma_2^2 + c_2)]}, \tag{3}$$

where μ_1 is the average pixel value of the dCT image, μ_2 is the average pixel value of the sCT image, σ_1^2 is the pixel variance of the dCT, σ_2^2 is the pixel variance of the sCT, σ_{12} is the pixel covariance of the dCT and sCT, $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$, L is the dynamic range of the pixel values (L = 4095 in our case), $k_1 = 0.01$ and $k_2 = 0.03$. SSIM ranges from 0 to 1, and higher values indicate greater similarity between two images.

SNR compares the level of a desired signal to the level of background noise and is defined as

$$SNR = 10log_{10}\left[\frac{\sum_0^{n_x-1}\sum_0^{n_y-1}(dCT(x,y))^2}{\sum_0^{n_x-1}\sum_0^{n_y-1}(dCT(x,y) - sCT(x,y))^2}\right]. \tag{4}$$

MAE and RMSE are given in HU. SSIM is unitless. SNR is given in dB.

4. Results

4.1. Performance of the source, target, combined, and adapted models

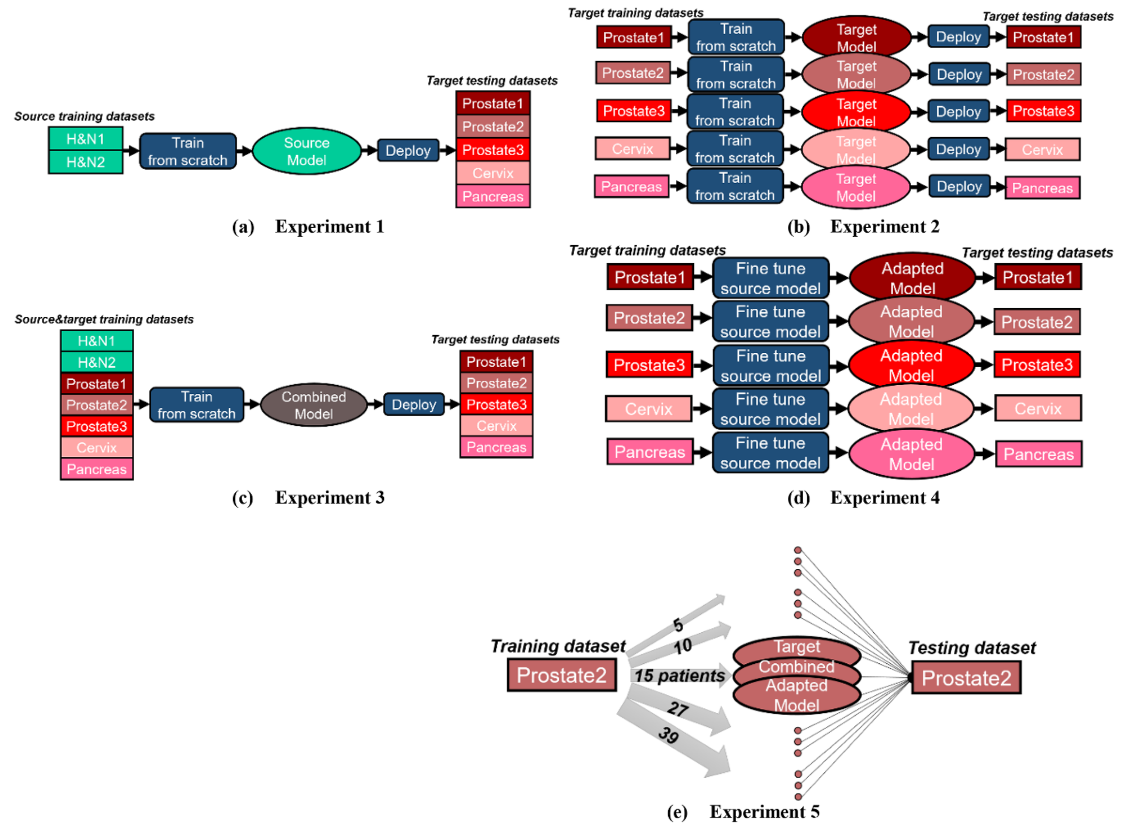
The quantitative evaluation results including MAE, RMSE, SSIM, and SNR of the sCT generated by the source model on source datasets are shown in table 2 and serve as a baseline of the performance of the source model on target datasets. The dispersion of evaluation results from the source model on source datasets are expressed in boxplots shown in figure 6. The generated sCT images and their corresponding CBCT and dCT images from the H&N1 and H&N2 testing datasets are shown in supplementary figures 1 and 2 (available online at stacks.iop.org/MLST/2/015007/mmedia) for visual evaluation. We can see that the trained CycleGAN model can generate sCT images that are very similar to the reference dCT images and much better than the CBCT images when the model is applied to the same dataset that it was trained on.

We quantitatively evaluated the performance of the source, target, combined, and adapted models in terms of MAE, RMSE, SSIM and SNR methods for each of the following target datasets: Prostate1, Prostate2, Prostate3, Cervix, and Pancreas. The geometric mean and SD of the similarity scores for every case and its relative improvement from source model are shown in table 3. Two-tailed Wilcoxon signed-rank tests were used for pairwise patient samples comparison due to non-normally distributed data and small sample size. The calculated p-values are shown with boxplots in figures 7–8, and supplementary figures 8–10. The sCT images generated by the source, target, combined, and adapted models and their corresponding CBCT and dCT images for each target dataset are shown in supplementary figures 3–7 for visual evaluation.

In the Prostate1 dataset, which comes from Varian OBI scanners, like the source datasets (H&N1 and H&N2), the sCT images generated by the source model tend to have higher MAE geometric mean of 21.9

Table 3. Evaluation of similarity measures of sCT (source model), sCT (target model), sCT (combined model), and sCT (adapted model) against dCT on five target datasets: (1) Prostate1 with 11 testing patients; (2) Prostate2 with 11 testing patients; (3) Prostate3 with 10 testing patients; (4) Cervix with 10 testing patients; (5) Pancreas with 10 testing patients. The values in the table are mean ± SD (relative improvement). The mean and SD of MAE, RMSE, SSIM, and SNR of sCT images were calculated using dCT images as reference. Relative improvement is calculated using sCT (source model) as baseline.

Target datasets	Model	MAE (HU)	RMSE (HU)	SSIM (%)	SNR (dB)
Prostate1	sCT (source model)	21.9 ± 7.6	55.3 ± 10.9	88.8 ± 5.5	23.3 ± 2.0
	sCT (target model)	20.1 ± 5.7 (*8%*)	57.3 ± 10.8 (*−4%*)	91.2 ± 3.0 (*3%*)	23.0 ± 1.9 (*−1%*)
	CT (combined model)	19.3 ± 5.3 (*12%*)	56.5 ± 11.4 (*−2%*)	91.9 ± 2.6 (*4%*)	23.1 ± 2.0 (*−1%*)
	sCT (adapted model)	19.7 ± 5.8 (*10%*)	57.1 ± 10.9 (*3%*)	91.6 ± 2.9 (*3%*)	23.0 ± 2.0 (*−1%*)
Prostate2	sCT (source model)	46.8 ± 5.3	112.4 ± 8.2	75.1 ± 3.7	17.4 ± 1.0
	sCT (target model)	28.4 ± 4.4 (*39%*)	82.5 ± 6.8 (*27%*)	85.9 ± 4.1 (*14%*)	20.1 ± 0.9 (*15%*)
	sCT (combined model)	28.0 ± 4.5 (*40%*)	84.3 ± 5.8 (*25%*)	86.3 ± 3.9 (*15%*)	20.0 ± 0.9 (*14%*)
	sCT (adapted model)	22.4 ± 4.7 (*52%*)	62.9 ± 6.7 (*44%*)	88.4 ± 4.4 (*18%*)	21.5 ± 1.3 (*24%*)
Prostate3	sCT (source model)	60.5 ± 10.4	132.4 ± 12.9	74.5 ± 5.5	16.0 ± 1.4
	sCT (target model)	31.5 ± 8.1 (*47.9%*)	100.9 ± 16.5 (*24%*)	84.6 ± 4.1 (*14%*)	18.4 ± 1.9 (*15%*)
	sCT (combined model)	33.2 ± 9.2 (*45%*)	92.9 ± 12.8 (*30%*)	82.4 ± 5.9 (*11%*)	19.1 ± 1.6 (*20%*)
	sCT (adapted model)	27.0 ± 8.5 (*55%*)	76.9 ± 15.4 (*42%*)	85.3 ± 4.7 (*14%*)	20.8 ± 2.2 (*31%*)
Cervix	sCT (source model)	49.0 ± 9.4	115.1 ± 13.3	79.5 ± 4.1	17.4 ± 1.4
	sCT (target model)	27.7 ± 9.9 (*44%*)	94.1 ± 13.7 (*18%*)	86.0 ± 6.7 (*8%*)	19.3 ± 1.8 (*11%*)
	sCT (combined model)	26.5 ± 8.5 (*46%*)	83.9 ± 14.9 (*27%*)	87.1 ± 5.0 (*10%*)	20.2 ± 1.9 (*16%*)
	sCT (adapted model)	22.1 ± 7.4 (*55%*)	75.8 ± 13.8 (*34%*)	88.9 ± 4.1 (*12%*)	21.1 ± 2.1 (*22%*)
Pancreas	sCT (source model)	37.2 ± 6.4	94.2 ± 10.6	79.7 ± 4.9	19.4 ± 1.2
	sCT (target model)	25.2 ± 6.1 (*32%*)	77.3 ± 11.3 (*18%*)	86.8 ± 3.9 (*9%*)	20.2 ± 1.1 (*4%*)
	sCT (combined model)	26.3 ± 4.9 (*29%*)	85.6 ± 8.0 (*9%*)	86.7 ± 3.8 (*9%*)	20.2 ± 1.1 (*4%*)
	sCT (adapted model)	23.6 ± 5.0 (*36%*)	75.8 ± 9.8 (*20%*)	87.4 ± 3.8 (*10%*)	21.3 ± 1.4 (*10%*)



(a) **Experiment 1**

(b) **Experiment 2**

(c) **Experiment 3**

(d) **Experiment 4**

(e) **Experiment 5**

Figure 5. Evaluation of similarity measures of sCT (source model), sCT (target model), sCT (combined model), and sCT (adapted model) against dCT on a target testing dataset (Prostate1). (a) MAE, (b) RMSE, (c) SSIM, and (d) SNR of sCT images were calculated using dCT images as reference. P-values were calculated using two-tailed Wilcoxon signed rank tests.

(SD 7.6) compared to the target model with 20.1 (5.7), the combined model with 19.3 (5.3), and the adapted model with 19.7 (5.8). In SSIM evaluation, they tend to have lower SSIM geometric mean of 88.8 (5.5) compared to the target model with 91.2 (53.0), the combined model with 91.9 (2.6), and the adapted model

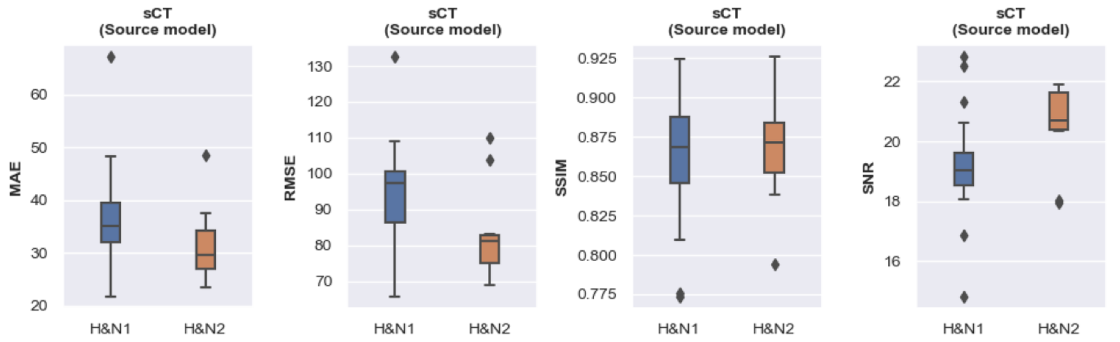


Figure 6. Evaluation of similarity measures of sCT (source model), sCT (target model), sCT (combined model), and sCT (adapted model) against dCT on a target testing dataset (Prostate2). (a) MAE, (b) RMSE, (c) SSIM, and (d) SNR of sCT images were calculated using dCT images as reference. P-values were calculated using two-tailed Wilcoxon signed rank tests.

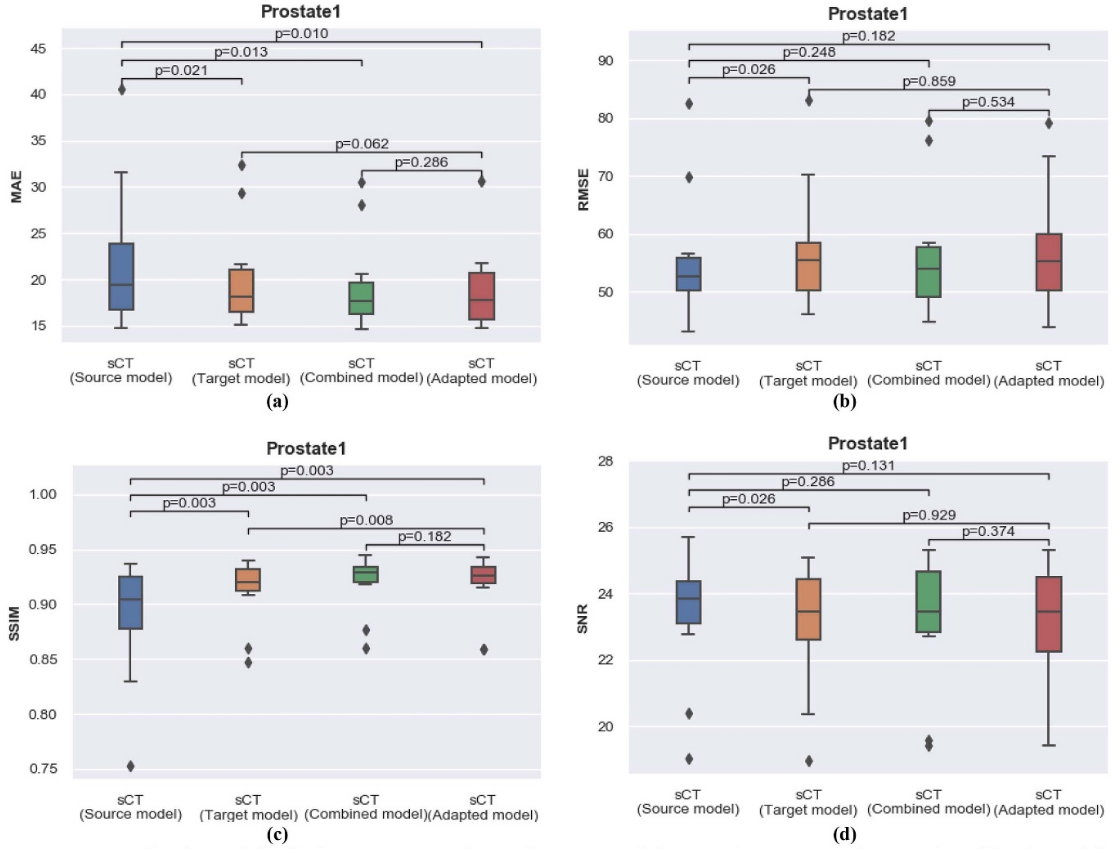


Figure 7. The influence of training data size on the performance of target, combined, and adapted models in Prostate2 dataset with similarity measures: (a) MAE, (b) RMSE, (c) SSIM, and (d) SNR.

with 91.6 (2.9). The *p*-values calculated between source vs target model, source vs combined model, and source vs adapted model in both MAE and SSIM evaluations are less than 0.05. However in RMSE and SNR evaluations, no statistically significant difference were found in these group comparisons. This indicates that the source model performed reasonably well on this target dataset, and the target, combined and adapted models can only improve a little. Among the three updated models, no statistically significant difference were found between adapted vs target model in RMSE and SNR evaluations, and between adapted vs combined model in all four similarity measures. These results suggest that the three updated models performed comparably. These results are compatible with supplementary figure 3 for visual evaluation. Thus, when applying the source model to a dataset which it has never seen before (different anatomical sites), but coming from the same vendor's scanners, the source model can generate good quality sCT images from CBCT, while the target model, combined model, and adapted model slightly improve upon this performance by 8.2%, 12.0%, and 10.3% respectively, if evaluated in MAE similarity measure.

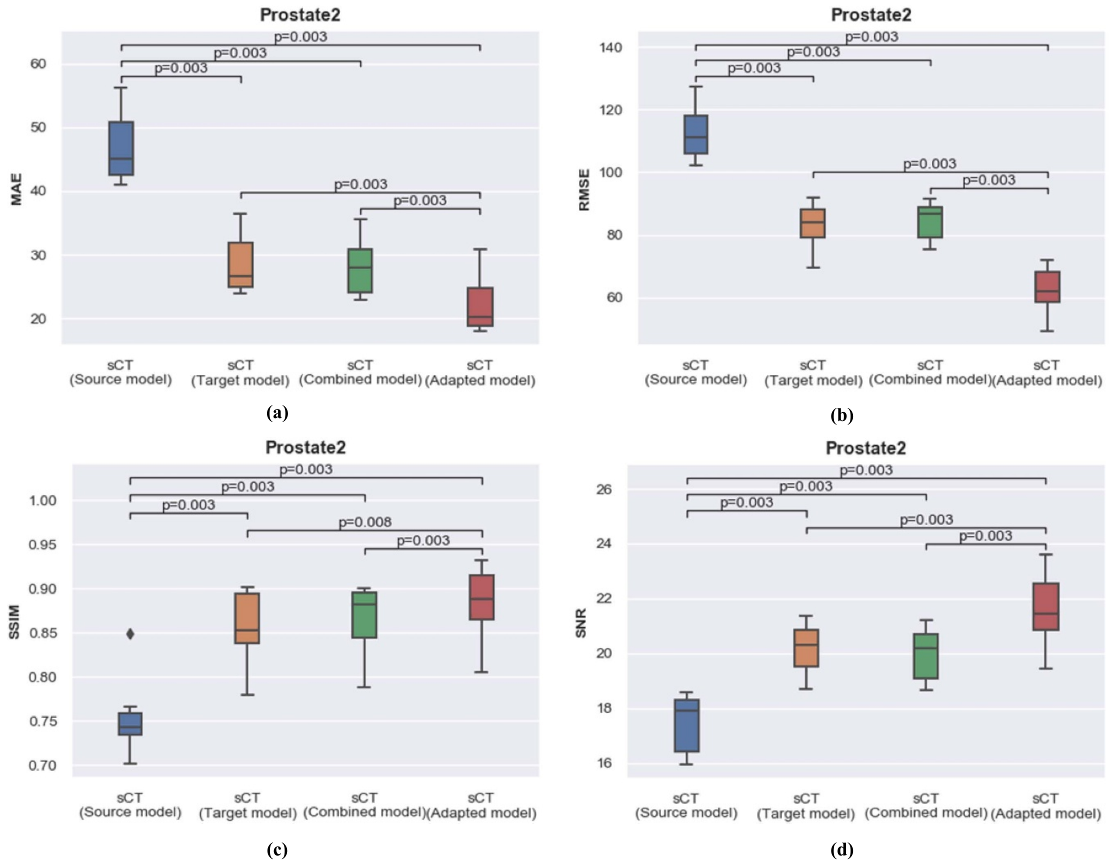


Figure 8. Evaluation of similarity measures of sCT (source model), sCT (target model), sCT (combined model), and sCT (adapted model) against dCT on a target testing dataset (Prostate2). (a) MAE, (b) RMSE, (c) SSIM, and (d) SNR of sCT images were calculated using dCT images as reference. P-values were calculated using two-tailed Wilcoxon signed rank tests.

In Prostate2 dataset, which come from Elekta XVI (Versa) scanners, the source model has the highest MAE and RMSE geometric means of 46.8 (SD 5.3) and 112.4 (8.2), and the lowest SSIM and SNR geometric means of 75.1 (3.7) and 17.4 (1.0) among all the models. Statistical tests between source vs target model, source vs combined model, and source vs adapted model in four similarity measures all show p-values less than 0.05. These results indicate that the source model performed much worse in this target dataset (Prostate2) than in the Prostate1 target dataset. In contrast, the adapted model has the lowest MAE and RMSE geometric means of 22.4 (4.7), and highest SSIM and SNR geometric means of 88.4 (4.4) and 21.5 (1.3) compared to the target model and combined model with MAE of 28.4 (4.4) and 28.0 (4.5), RMSE of 82.5 (6.8) and 84.3 (5.8), SSIM of 85.9 (4.1) and 86.3 (3.9), and SNR of 20.1 (0.9) and 20.0 (0.9) respectively. Statistical tests between adapted vs target model, and adapted vs combined models in four similarity measures all show p-values less than 0.05. All these results lead us to conclude that the adapted model has the best performance among the three updated models. Upon visual evaluation of supplementary figure 4, it also indicates that the source model performs poorly on the Prostate2 dataset, while the target, combined, and adapted models can improve its performance by 39.4%, 40.3%, and 52.2% respectively, if evaluated in MAE similarity measure. Among all the updated models, the adapted model performs the best. In the Prostate3, Cervix, and Pancreas datasets, which come from Elekta XVI (Versa) or XVI (Agility) scanners, similar patterns were also noted for the four models' performance in supplementary figures 8–10. For visual evaluation, one can refer to supplementary figures 5–7. Thus, when applying the source model to datasets which it has never seen before and been collected from different anatomical sites and different vendors' scanners, the source model fails to generate good quality sCT images, and its MAE, RMSE, SSIM, SNR scores are substantially worse than in the previous scenario (Prostate1). All three updated models greatly outperform the source model, and the adapted model always performs best.

4.2. Dependence of model performance on target dataset size

Figure 9 shows that the performance of the target, combined, and adapted models depends on the size of the training dataset in the target domain (Prostate2). The three models performed similarly with a large number (2730) of training images. As the number of training images decreased, the accuracy of the combined model

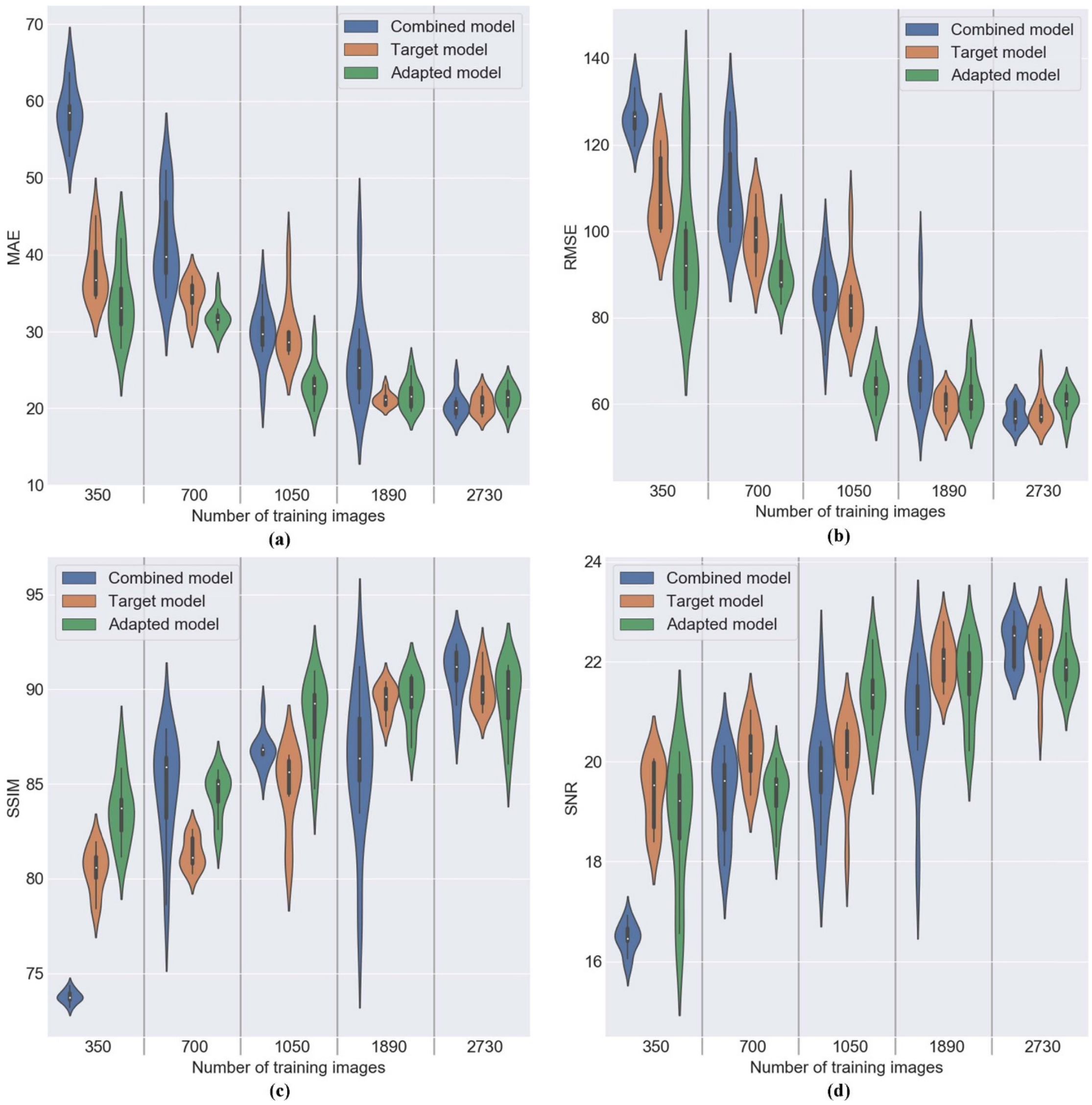


Figure 9. The influence of training data size on the performance of target, combined, and adapted models in Prostate2 dataset with similarity measures: (a) MAE, (b) RMSE, (c) SSIM, and (d) SNR.

decreased the fastest, while the accuracy of the adapted model decreased the slowest. Because the combined model was trained on the Prostate2, H&N1 and H&N2 datasets, the training data for the combined model becomes unbalanced when there is much less Prostate2 data than H&N1 and H&N2 data. The knowledge previously learned from the H&N1 and H&N2 data gives the adapted model a good starting point to update features, so it requires less training data to achieve good performance. Therefore, with a small number of training images, the adapted model is clearly superior to the target and combined models. The adapted model still achieved good performance when trained on only 1050 images from 15 patients. These results suggest that model fine-tuning is the best option to guarantee good performance and robust training regardless of the amount of data available.

5. Discussion and conclusions

In this paper, we have illustrated the problem of generalizability for a DL model in a CBCT-to-CT image conversion application, and we showed that the model only trained on the source datasets (source model) does perform differently in different target datasets. In Prostate1 target dataset coming from the same vendor's scanner with source datasets but different disease sites, updating models can only improve upon the source model performance by up to 3.3% with RMSE evaluation. However, in Prostate2, Prostate3, Cervix, and Pancreas target datasets coming from different vendor's scanners and different disease sites, updating models can improve upon the source model performance largely by up to 44.0%, 41.9%, 34.2%, and 19.6% in each target dataset with RMSE evaluation. Thus, in our application, disease site was a minor influence on the source model's performance, but vendor's scanner was a major influence that could dramatically decrease the accuracy of the source model.

To solve this generalizability problem, we compared three solutions—target, combined, and adapted models—in each target dataset. The source model is only trained on source datasets starting from scratch. The target model is only trained on a target dataset starting from scratch. The combined model is trained on a combined source and target dataset starting from scratch. The adapted model fine-tunes the trained source model on a target dataset. We found that all three updated models modestly outperform the source model when the source model already performs well on a target dataset from the same vendor's scanners, but they significantly outperform the source model when the source model performs poorly on a target dataset from a different vendor's scanners. Among the three updated models, the adapted model works the best with lowest MAE means of 22.4, 27.0, 22.1, and 23.6 HU in Prostate2, Prostate3, Cervix, and Pancreas compared to the other models.

By analyzing the change in the models' performance with different numbers of training images, we found that the target, combined, and adapted models perform comparably with a large number of training images, but the adapted model significantly outperforms the other two with smaller numbers of training images. Therefore, we suggest using the fine-tuning strategy to solve the generalization problem when deploying a DL model in clinical settings.

The generalizability of DL models is a challenging issue in medicine and is important for clinical implementation, where accuracy and precision are required. Even though this study only uses data from one institution, it has already shown a significant problem within one institution, so more significant problems should be expected for datasets from different institutions. Future studies should investigate this.

The generalizability of DL models is also task-specific. In this paper, our task is to decrease the noise and other artifacts from CBCT images and convert CBCT's HU values to CT's HU values. Our assumption is that gender, age, and other background illnesses have ignorable effects on noise distribution and CBCT-to-CT HU relationship. Thus, we have taken patients randomly without taking into account gender, age, and other medical conditions. However, if the task is detection and classification of diseases, factors like gender, age, and other background illnesses are parameters needed to be considered in the generalizability of DL models.

Another limitation of this work is that we only study one clinical task, CBCT-to-CT image conversion, to illustrate the problem and test three solutions. The number of training images needed to train or fine-tune a model is task-specific. However, we have shown the problem and solutions clearly, and we can still draw general conclusions from this one task.

We promote using the fine-tuning strategy for commissioning a model before implementing it in clinical practice, as it is very difficult to include enough data types from enough institutions to train a universal model. Future studies can focus on implementing the fine-tuning strategy through an automatic workflow that can be used easily in clinical environments by clinicians.

Acknowledgments

We thank Dr Jonathan Feinberg for editing the manuscript and Varian Medical Systems, Inc., for funding part of the project.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Code availability

The CycleGAN model algorithm is free to download for non-commercial research purposes on GitHub (https://github.com/lxaibl/CycleGAN-CBCT-to-CT).

ORCID iDs

Xiao Liang https://orcid.org/0000-0002-2472-2396
Dan Nguyen https://orcid.org/0000-0002-9590-0655

References

Al-Obaidi F E 2015 Image quality assessment for defocused blur images *Am. J. Signal Process.* **5** 51–55
Bengio Y 2012 *Practical Recommendations for Gradient-based Training of Deep Architectures* (Berlin: Springer)

David A B, Linda M, Miriam A B, Birgit B E-W, Kathryn J F, Vicky J G, Elkan F H, Christopher P H, Mark L S and Clifford R W 2020 Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the Radiology Editorial Board *Radiology* **294** 487–9

Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT Press)

Harms J, Lei Y, Wang T, Zhang R, Zhou J, Tang X, Curran W J, Liu T and Yang X 2019 Paired cycle-GAN based image correction for quantitative cone-beam CT *Med. Phys.* **46** 3998–4009

Kida S, Kaji S, Nawa K, Imae T, Nakamoto T, Ozaki S, Ohta T, Nozawa Y and Nakagawa K 2019. Cone-beam CT to Planning CT synthesis using generative adversarial networks *preprint arXiv:1901.05773*

Liang X, Chen L, Nguyen D, Zhou Z, Gu X, Yang M, Wang J and Jiang S 2019 Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy *Phys. Med. Biol.* **64** 125002

McKinney S M, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado G C and Darzi A 2020 International evaluation of an AI system for breast cancer screening *Nature* **577** 89–94

Rajpurkar P, Joshi A, Pareek A, Chen P, Kiani A, Irvin J, Ng A Y and Lungren M P 2020 CheXpedition: investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting preprint (arXiv:2002.11379)

Reed R and Marksii R J 1999 *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks* (Cambridge, MA: MIT Press)

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P 2004 Image quality assessment: from error visibility to structural similarity *IEEE Trans. Image Process.* **13** 600–12

Wang Z, Simoncelli E P and Bovik A C 2003 Multiscale structural similarity for image quality assessment *The Thrity-Seventh Asilomar Conf. on Signals, Systems & Computers, 2003* vol **2** pp 1398–402

Wu R Y, Liu A Y, Williamson T D, Yang J, Wisdom P G, Zhu X R, Frank S J, Fuller C D, Gunn G B and Gao S 2019 Quantifying the accuracy of deformable image registration for cone-beam computed tomography with a physical phantom *J. Appl. Clin. Med. Phys.* **20** 1526–9914