Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review

M. Boulanger a, Jean-Claude Nunes a,*, H. Chourak a,c, A. Largent b, S. Tahri a, O. Acosta a, R. De Crevoisier a, C. Lafond a, A. Barateau a

a Univ. Rennes 1, CLCC Eugène Marquis, INSERM, LTSI - UMR 1099, F-35000 Rennes, France
b Developing Brain Institute, Department of Diagnostic Imaging and Radiology, Children’s National Hospital, Washington, DC, USA
c CSIRO Australian e-Health Research Centre, Herston, Queensland, Australia

ARTICLE INFO

Keywords:
Deep learning
MRI
Synthetic-CT
Radiation therapy
Dose calculation

ABSTRACT

Purpose: In radiotherapy, MRI is used for target volume and organs-at-risk delineation for its superior soft-tissue contrast as compared to CT imaging. However, MRI does not provide the electron density of tissue necessary for dose calculation. Several methods of synthetic-CT (sCT) generation from MRI data have been developed for radiotherapy dose calculation. This work reviewed deep learning (DL) sCT generation methods and their associated image and dose evaluation, in the context of MRI-based dose calculation.

Methods: We searched the PubMed and ScienceDirect electronic databases from January 2010 to March 2021. For each paper, several items were screened and compiled in figures and tables.

Results: This review included 57 studies. The DL methods were either generator-only based (45% of the reviewed studies), or generative adversarial network (GAN) architecture and its variants (55% of the reviewed studies). The brain and pelvis were the most commonly investigated anatomical localizations (39% and 28% of the reviewed studies, respectively), and more rarely, the head-and-neck (H&N) (15%), abdomen (10%), liver (5%) or breast (3%). All the studies performed an image evaluation of sCTs with a diversity of metrics, with only 36 studies performing dosimetric evaluations of sCT.

Conclusions: The median mean absolute errors were around 76 HU for the brain and H&N sCTs and 40 HU for the pelvis sCTs. For the brain, the mean dose difference between the sCT and the reference CT was <2%. For the H&N and pelvis, the mean dose difference was below 1% in most of the studies. Recent GAN architectures have advantages compared to generator-only, but no superiority was found in term of image or dose sCT uncertainties. Key challenges of DL-based sCT generation methods from MRI in radiotherapy is the management of movement for abdominal and thoracic localizations, the standardization of sCT evaluation, and the investigation of multicenter impacts.

Introduction

In radiation therapy, computed tomography (CT) is the standard imaging modality for treatment planning. Magnetic resonance imaging (MRI) is a complementary modality to CT providing better soft-tissue contrast without irradiation. MRI improves the delineation accuracy of the target volume and/or organs at risk (OARs) in the brain, head-and-neck (H&N), and lung or prostate radiotherapy [1–3]. However, MRI does not provide information on the electron density of the tissue, requires for accurate dose calculation. Most of the literature has proposed the generation of synthetic-CT (sCT) images for MRI-based dose planning. sCT (or pseudo-CT) is a synthetic image in Hounsfield Units (HU) generated from MRI data.

The methods for generating sCTs can be divided into three categories: bulk density, atlas-based and machine learning (ML) methods (including classical ML methods and deep learning methods [DLMs]). The bulk density methods consist of segmenting MRI images into several classes (usually air, soft-tissue, and bone). Each of these delineated volumes is assigned a homogeneous electron density, and the dose can then be calculated. This method has several drawbacks: it is tedious, time-consuming, operator-dependent, and does not consider tissue heterogeneity [4–8]. The atlas-based methods involve complex, non-
rigid registrations of one or several co-registered MRI-CT atlases with a target MRI. This registration step is followed by a fusion step to generate the sCT. The drawbacks of this method are the lack of robustness in the case of large anatomical variations and the need for computationally intensive pairwise registrations \[4,5,9,10\]. Among the classical ML methods, the patch-based methods (such as [4]) can be decomposed into four steps. The first step is interpatient rigid or affine registration with MR images. These methods involve inter-patient registration, feature extraction, and patch partitioning during the training step. The training patches closest to the patches of the target MRI are then selected for aggregation to generate the sCT [4]. The main drawbacks of this method are the imprecise interpatient registration and calculation time.

DLMs are models comprising multiple processing layers that learn multiscale representations of data through multiple levels of abstraction \[11\]. These methods have recently been introduced in radiotherapy for applications, including image segmentation, image processing and reconstruction, image registration, treatment planning, and radiomics \[12–19\]. DLMs have been proposed for sCT generation from MRI. They were trained to model the relationships between HU CT values and MRI intensities. Once the optimal DL parameters are estimated, the model can be applied to a test MRI to generate its corresponding sCT. DLMs have the advantage of being fast for sCT generation, and some do not require deformable inter-patient registration (only intra-patient registration) such as in \[20\].

Two reviews, both published in 2018, have already summarized sCT generation methods from MRI \[21,22\], they focused only on the bulk density, atlas-based, and voxel methods and did not include recent DLMs. Other studies have listed sCT generation methods from MRI in the context of MR-only radiotherapy \[22–25\]. More recently, Wang et al. \[26\] proposed a review on medical imaging synthesis using DL and Spadea and Maspero et al. \[27\] a review on sCT generation with DLM. These methods have recently been introduced in radiotherapy for applications or learning data. Several reviews have detailed the DL network architectures have been developed, depending on specific applications or learning data. Many DL network architectures containing multiple processing components with adjustable parameters to learn a representation of data. However, these architectures may be roughly divided into two classes: generator-only and generative adversarial network (GAN) and its variants (such conditional-GAN, Least square GAN and cycle-GAN). Fig. 2 shows the hierarchy of the DL architectures.

A. Common deep learning networks for sCT generation from MRI

Deep learning, as a mainstream of ML method, uses trainable computational models containing multiple processing components with adjustable parameters to learn a representation of data. Many DL network architectures have been developed, depending on specific applications or learning data. Several reviews have detailed the DL network architectures for radiotherapy or medical imaging \[12,26,27,34–37\]. The DL architecture for sCT generation from MRI can be roughly divided into two classes: generator-only and generative adversarial network (GAN) and its variants (such conditional-GAN, Least square GAN and cycle-GAN). Fig. 2 shows the hierarchy of the DL architectures.

1. Generator-only models

i. Basic concepts of convolutional neural networks (CNN)
For image applications, a convolutional neural network (CNN, or ConvNet) is a popular class of deep neural networks using a set of convolution kernels/filters for detecting image features. A CNN consists of an input layer, multiple hidden layers and an output layer. The hidden layers include layers that perform convolutions with trainable kernels. Nonlinear activation functions (Rectified Linear Units (ReLU) [38], Leaky-RELU [39], Parametric-ReLU (PReLU) or exponential linear unit (ELU) [40]) play a crucial role in discriminative capabilities of the deep neural networks. The ReLU layer preserves the input otherwise is the most commonly used activation layer due to its computational simplicity, representational sparsity, and linearity. It is commonly to periodically insert a pooling layer between successive convolutional layers in a CNN architecture. Pooling layers allow to reduce the dimension (subsampling) of the feature maps. These maps are generated by following the convolutional operations. The pooling methods performs down-sampling by dividing the input into rectangular pooling regions and computing the average, the maximum, or the minimum of each region represented by the filter (mean pooling, max-pooling, min-pooling). Batch normalization [41] layers are inserted after a convolutional or fully connected layer to improve the convergence of the loss function during gradient descent (optimizer). It prevents the problem of vanishing gradient from arising and significantly reduces the time required for network convergence. After several convolution and pooling layers, the CNN generally ends with several fully connected layers. Dropout is one of the most promising techniques for regularization of CNN. Softmax layer is typically the final output layer in a neural network that performs multi-class classification (for example: object recognition).

ii. Generator-only models

The generator model can be considered as representing a complex end-to-end mapping function that transforms an input MR image to its corresponding CT image. During the training phase, the generator tries to minimize an objective function called a loss function (voxel-wise loss function $L_o$), which is an intensity-based similarity measurement between the generated image (sCT) and the corresponding ground truth image (real CT). Fig. 3 presents the global architecture of generator-only model.

In sCT generation from MRI, the generator architectures are generally based on convolution encoder-decoder networks (CED). In the literature, the variants of generator model include deep CED network [42], deep embedding CNN (DECNN) or Embedded Net [30], fully convolutional network (FCN) [28], U-Net [20,42–57], efficient CNN (eCNN) model [60], ResNet [61], SE-ResNet [61,62], and DenseNet [63]. Fig. 4 presents some architectures of CED-based generators (Fig. 4).

The CED network consists of a paired encoder and decoder networks. CED have been extensively used in DL literature thanks its excellent performance. In the encoding part, low-level feature maps are downsampled to high-level feature maps. In the decoding part, the high-level feature maps are upsampled to low-level feature maps using the transposed convolutional layer to construct the prediction image (sCT).

The encoder network uses a set of combined 2D convolution filtering (no dilated convolutions) for detecting image features, followed by normalization (instance [66] or batch normalization [41]), a nonlinear activation function (ReLU [38], LeakyRELU [39], or PReLU), and max-pooling.

The decoder path combines the feature and spatial information through a sequence of symmetrical transpose convolutional layers (up-convolutions), up-sampling operators, concatenate layer (concatenations with high-resolution features), and convolutional layers with a ReLU activation function.

The most well-known and popular CED variants for biomedical...
image applications is the U-shaped CNN (U-Net) architecture proposed by Ronneberger et al. [67]. The U-Net [67] has a CED structure with direct skip connections between the encoder and decoder. Han et al. were the first to publish a sCT study with a U-Net architecture [44] that is similar to Ronneberger’s model. This 2D U-net model directly learns a mapping function to convert a 2D MR grayscale image to its corresponding 2D sCT image. Han et al. study [44] differs from the original U-net since the three fully connected layers were removed. Thus, the number of parameters is reduced by 90%, and the final model is easier to train. In Wang et al. [46], the U-net model used batch normalization [41] and leaky ReLU, which was different from the classical U-net [67].

The efficient CNN model proposed by Xiang et al. [30] is derived by inserting multiple embedding blocks into the U-net architecture. This embedding strategy helps to backpropagate the gradients in the CNN and also provides easier and more effective training of the end-to-end mapping from MR to CT with faster convergence.

The deep CED network [42] consists of a combined encoder network...
(the popular Visual Geometry Group [VGG] 16-layer net model) and a decoder network (reversed VGG16) with multiple symmetrical shortcut connections between layers.

Twenty-nine state-of-the-art sCT image generation methods have adopted a generator-only network [20,28,30,42–57,70–79]. The loss functions $L_G$ evaluating sCT and real CTs used in these generative models are:

- the mean square error (MSE), the L2-norm, or the Euclidean norm: only for sCT [20,42,46,47,55,57,78], for sCT and embedding blocks [30],
- the mean absolute error (MAE), mean absolute deviation (MAD), or L1-norm [43–45,49,52,53,70,71],
- a combined MAE and MSE loss [48],
- perceptual loss [20] based on VGG (the output of the 7th VGG16 convolutional layer).

The use of L2 distance as a loss function tends to produce blurry results. Perceptual loss is used to capture the discrepancy between the high frequency components within an image.

One limitation of generative models based on CNN is that they may lead to blurry results due to generally misalignment between MR and CT [80].

2. Generative adversarial network (GAN)

The following section summarizes GAN-based architectures to generate sCT from MRI. We introduce the GAN architecture and three most popular GAN-based extensions: least squares-GAN, conditional-GAN, and cycle-GAN.

i) GAN

The adversarial learning strategy was proposed by Goodfellow et al. [81] to generate better sCT images than previous generator-only models. The original way is to simultaneously train two separate neural networks (Fig. 5), the generator $G$ (one of the generator-only models described in i) and Fig. 4) and the discriminator $D$. These two neural networks form a two-player min–max game where $G$ tries to produce realistic images to fool $D$ while $D$ tries to distinguish between real and synthetic data [81,82]. Compared to generator-only models, GAN

---

**Fig. 4.** Representation of generator architecture for U-Net [64] and adapted implementations of DenseNet [63], SE-ResNet [65], and Embedded Net [30]. The size of the boxes indicates the relative resolutions of the feature maps. The green boxes represent convolutional layers and orange boxes represent transposed convolutional layers. Yellow, red and blue boxes represent the SE-ResNet, DenseNet, and Embedded blocks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
introduces a data-driven regularizer, the adversarial loss, to ensure that the learned distribution approaches the ground truth.

In the original version [81], the discriminator and generator are implemented as multilayer perceptrons (MLPs) and more recently implemented as CNNs. The architecture of the generator is often the conventional U-Net. Another proposed generator architecture in a GAN is ResNet [62] which is easy to optimize and can gain accuracy from considerably increased depth. The discriminator of the GAN [81] consists of six convolutional layers with different filter sizes but the same kernel sizes and strides, followed by five fully connected layers. ReLU was used as the activation function and a batch normalization layer for the convolutional layers. The dropout layer was added to the fully connected layers, and a sigmoid activation function was used in the last fully connected layer.

The discriminator used in [64] a convolutional “PatchGAN” classifier (markovian discriminator) models high frequency image structure in local patches and only penalizes structure at the scale of image patches.

Using adversarial loss $L_D$, the classical GAN model can generate high-quality sCT images with less blurry results [29,80] than generator-only models. The discriminator tries to maximize it while the generator tries to minimize it.

In this review, six studies used classical GAN-based architectures to generate sCT from MRI [20,29,62,77,83,84]. The overall loss functions intergrating the adversarial loss function $L_D$ and evaluating sCT and the original CTs used in these GANs are:

- L2-norm alone [20,84],
- perceptual loss [20,83] and the multiscale perceptual loss [20].

The adversarial loss function $L_D$ of the discriminator used in these GANs was generally the binary cross-entropy [29].

Perceptual regularization, used by Largent et al. [20], helps to prevent images over-smoothing and loss of structure details. The perceptual loss functions are based on high-level features extracted from pre-trained VGG network (7th VGG16 in [20]).

As shown by several studies [29,62,85], (1) the adversarial network prevents the generated images from blurring and better preserve details, especially edge features; (2) the accuracy of sCT within the bone region is increased; and (3) the discriminator detects patch features in both real and fake images, mitigating misregistration problem caused by an imperfect alignment between multi-parametric MRI and CT. General convergence in GANs is heavily dependent on hyperparameter tuning to avoid vanishing [86] or exploding gradients, and they are prone to mode collapse. To tackle the training instability of GANs, a plethora of
ii) Least Squares-GAN (LS-GAN)

Most GANs use the binary cross-entropy as the discriminator loss function. However, this cross-entropy loss function leads to the saturation problem in GANs learning (the well-known problem of vanishing gradients [86]). Least square loss function strongly penalized the fake samples away from decision boundary and improve the stability of learning process. Mao et al. [87] adopted the least-squares loss function for the discriminator and showed that minimizing the objective function of LS-GAN minimizes the Pearson \(\chi^2\) divergence [88]. Emami et al. [62] replaced the negative log-likelihood objective with a least square loss function (L2 loss), which was more stable during training and generated better CT quality.

iii) Conditional-GAN (cGAN)

Since the original GAN allows no explicit control on the actual data generation, Goodfellow et al. [81] proposed the conditional GAN (cGAN) to incorporate additional information such as class labels in the synthesis process. cGAN is an extension of the GAN model in which both the generator and the discriminator are conditioned on some additional information. The sCT image output is conditioned on the MR image input.

Different generator architectures in a cGAN have been proposed, including SE-ResNet [61,62], DenseNet [63], U-Net [56,58,59], Embedded Net [30], and the atrous spatial pyramid pooling (ASPP) method [56]. Fettty et al. [89] evaluated four different generator architectures: SE-ResNet, DenseNet, U-Net, and Embedded Net in a cGAN to generate sCT from T2 MRI. Olberg et al. [56] explored two generators: the conventional U-Net architecture implemented in the Pix2Pix framework [64] and the ASPP method [90,91]. The discriminator of the GAN framework was similar in both implementations.

Twelve studies used a cGAN architecture to generate sCT from MRI [31,33,50,56–59,88,89,92–101]. The overall loss functions \(L_D\) integrating the adversarial loss function \(L_D\) and evaluating sCT and real CTs used in these cGANs were as follows:

- adversarial loss function (binary cross entropy) [59,101],
- L1-norm (MAE) [92],
- least squares loss function (L2 loss) [88,101],
- mutual information (MI) [58,59],
- focal regression loss [102] used in [99],
- the combination of adversarial (binary cross-entropy) and L2-norm [56],
- the combination of L1-norm and PatchGAN loss (as proposed by Isola et al. [64]) used in [50,89,93],
- the combination of adversarial (binary cross-entropy) and term derived from the log-likelihood of the Laplace distribution [95],
- the combination of \(L_p\)-norm, adversarial and gradient [33],
- the combination of multiscale L1-norm, L1 norm and PatchGAN loss [64] used in [88].

The loss functions \(L_D\) of the discriminator evaluating sCT and real CTs used in these cGANs areas follows:

- the mostly used adversarial loss (binary cross entropy) [56,58,59,88,93],
- least squares loss function (L2 loss) [88,94,101],
- L1-norm [101].

The L2-based loss function of the generator can cause image blurring. To alleviate blurriness and improve the prediction accuracy, the L1 norm [46] makes the learning more robust to outliers in the training data, such as noise or other artifacts in the images or due to imperfect matching between MR and CT images. The Markovian Discriminator loss or Patch-GAN loss [64], which can be understood as a form of texture/style loss, effectively models the image as a Markov random field, assuming independence between pixels separated by more than a patch diameter.

Pix2Pix proposed by Isola et al. [64] is a successful cGAN variant for high-resolution image-to-image translation. Pix2Pix model generally uses Unet generator and PatchGAN discriminator. As investigated by Isola et al. [64], the use of a loss function based on L1 alone leads to reasonable but blurred results; while cGAN alone leads to sharp results but introduces image artifacts. The authors showed that training in an adversarial setting together with an L1 norm generated sharp images with few artefacts (tissue-classification errors, especially for bone and air differentiation).

In Hemsley et al. [95], the L1 term in cGAN loss function [64] is replaced by a term derived from the log-likelihood of the Laplace distribution to capture data dependent uncertainty.

To overcome MR/CT registration issues, Kazemifar et al. [58,59] used a generator loss function based on the MI in cGAN. The MI loss allows the cGAN to use unregistered data to generate sCT and seems to accurately distinguish between air and bone regions.

Instead of the usual cross-entropy \(L_D\) loss in cGAN, Mao et al. [87] recommend the quadratic version of the least square GAN. Olberg et al. [56] evaluated a Pix2Pix framework with two different generators: the conventional U-Net and a proposed generator composed of stacked encoders and decoders separated by dilated convolutions applied to increase rates in parallel to encode large-scale features. The overall loss function was composed of adversarial (sigmoid cross-entropy) and MAE losses.

One of the main advantages of cGANs is that the networks learn reasonable image-to-image translations even if the training dataset size is small. However, cGANs require co-registered MR-CT image pairs for training except with MI as loss function [58,59].

iv) Cycle-GAN

For image-to-image translations between two modalities, the principles of the cycle-GAN are to extract characteristic features of both modalities and discover the underlying relationship between them [104]. The cycle-GAN involved two GANs: one to generate sCT from MRI and a second to generate synthetic-MRI (sMRI) from sCT (the output of the first GAN). These dual GANs learn simultaneously and a cyclic loss function minimizes the discrepancy between the original CT and the sCT obtained from the chained generators.

Cycle GAN-based framework does not require paired MRI/CT images [80,105]. Wolterink et al. [80] found that training using unpaired images could, in some cases, outperform a GAN-model on paired images.

Eleven studies used a cycle-GAN architecture to generate sCT from MRI [32,33,57,77,80,100,101,105–108]. The overall loss functions \(L_D\) integrating the adversarial loss function \(L_D\) and comparing the generated sCT and real CTs used in these cycle-GANs were:

- the combination of adversarial loss (cross-entropy) and L1-norm [33,101],
- the combination of the adversarial loss based on cross-entropy, the cycle consistency loss based on L1-norm, and the structural consistency loss based on L1-MIND [105] (the modality-independent neighborhood descriptor, MIND, introduced in [109]).
the combination of L2-norm, adversarial loss (binary cross-entropy), the gradient difference loss and cycle consistency loss (based on L1 norm) [80],
• the combination of Lp-norm (mean P distance, MPD), adversarial loss and gradient loss [32,33,77,107].

Loss functions Lp of the discriminator used in cycle-GAN are:
• L2-norm (least squares loss) [62] as proposed in [87,110],
• MAD (L1-norm) [32,77,101,107],
• Lp-norm (MPD) [108].

Since L2-based loss functions tend to generate blurry images and L1-based loss functions may introduce tissue-classification errors, some authors [32,33,77,107,108] used an Lp-norm (p = 1.5) distance, the MPD (Mean P distance). Using the MPD-based loss term, the authors also integrated an image gradient difference (GD) loss term (proposed in [29]) into the loss function [32,33,77,107,108], to retain sharpness in synthetic images, which maintain zones with strong gradients, such as edges. Cycle-GAN-based methods use MSE loss as distance loss function, which often leads to blurring and over-smoothing.

B. Data for sCT generation from MRI

1. MRI/CT image preprocessing and post-processing

In eighteen studies an MRI bias correction [20,30,32,33,43,44,47,49,78,83,89,92,94,101,105–108] was reported. In [30,32,44,47], intensity inhomogeneity (or non-uniformity) correction was performed in all MR images using the N3 bias field correction algorithm [111,112] to correct the bias field before training or synthesis. In [33,43,78,83,89,92,94,105–108], the authors reported that the intensity inhomogeneity of the MRI was corrected using the N4 bias field correction algorithm.

A 2D or 3D MRI geometry correction provided by the vendor was sometimes reported [49,57,70,105]. We can think that most of MR images had a geometry correction, but that it was not mentioned.

In [30,33,78,83,94], all MR images were normalized using a histogram-based intensity normalization [113] to minimize the inter-patient MR intensity variation. Intensity normalization was also used in [30,32]. In [44], all MR images were then histogram-matched to a randomly chosen template to help standardize image intensities across different patients using the method described by Cox et al. [114]. All MR volumes were normalized by aligning the white matter peak identified by fuzzy C-means in [105]. In [49,101], histogram standardizations performed using vendor-provided software (CLEAR) were applied as provided by the vendor.

In the study by Maspero et al. [93], the voxel intensity of CT was clipped within the interval HU to avoid an excessively large discretization step and the MR images were normalized to their 95% intensity interval over the whole patient. All the images were converted to 8-bits to conform to the Pix2Pix implementation [64]. Before training, the air cavities were filled in CT images and bulk-assigned (–1000 HU) as located in MR images using an automated method.

2. Training data characteristics

Compared to 2D CNN, 3D CNN can better model 3D spatial information (neighborhood information) owing to the use of 3D convolution operations [28] solving the discontinuity problem across slices, which are suffered by 2D CNN. However, the input type to DL models is mainly in 2D because fully 3D networks are much more difficult to train due to a large numbers of trainable parameters and requires exponentially more (GPU) memory and more data [28,44]. With the 2.5 D approach, Dinkla et al. [70] added 3D contextual information while maintaining a manageable number of trainable parameters. Furthermore, discontinuities across slices present in 2D methods, were decreased. Besides, the 2.5D approaches [45,70,71] include average axial, sagittal, and coronal images as input to train the CNN. In 3D (patch-based) CNN [28,32], an input MR image is first partitioned into overlapping patches. For each patch, the CNN is used to predict the corresponding CT patch and all predicted CT patches are merged into a single CT image by averaging the intensities of overlapping CT regions.

Most of the reviewed studies used one MRI sequence as input and generated one sCT as output; an architecture generally called single-input single-output (SISO). Four studies used several MRI sequences as input to generate one sCT in output [50,72,92,103], these architectures referred to as multi-input single-output (MISO) [50,72,92,103] or multi-input multiple-output (MIMO) [103]. Moreover, most studies used training and evaluation data from one MRI device while eight studies used multi-device MRI. One study reported use of MRI data from different centers [96] and two studies [88,89] used data from the Gold Atlas Data set [115]. Five studies used low MR field (0.35 T) as input images [31,33,56,73,89].

3. Training and evaluation of data size

The studies included in this review used several training strategies including k-fold cross-validation, single-fold validation, or leave-one-out. In k-fold cross-validation, the dataset is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k−1 subsets are combined to form a training set. The average error across all k trials is then computed. In single-fold validation, the dataset is separated into two sets, the training and testing sets. The leave-one out strategy consists on k-fold cross-validation taken to its logical extreme, with k equal to N, the number of data patients in the set.

Data size is a fundamental challenge for DL approaches. There is no reported minimal or optimal data size for DL training. In the head area, four studies assessed sCT image quality as a function of the number of available images for training, from 15 to 242 patients for Alvesares Andres et al. [43], from 5 to 47 patients for Gupta [48], from 34 to 135 patients for Peng et al. [100], and from 1 to 40 patients for Maspero et al. [96]. Better image results were found for higher numbers of available images. A minimum of 10 patients seems to be needed since it has shown similar performance than a training of 20, 30 or 40 patients. One effective way to improve model robustness is to enhance the diversity of the training dataset. Data augmentation is essential to teach the network the desired invariance and robustness properties when only a few training samples are available. One common data augmentation technique [32,44,92] is to apply random translations, rotations, zooms, and elastic deformations and adding low-level random noise to training images.

C. Evaluation metrics

sCT evaluation can be performed in terms of intensity, geometric fidelity, or dose metrics. A sCT evaluation was performed using intensity-based metrics for all reviewed studies and through dose criteria in 63% of the reviewed studies. The metrics used in the reviewed studies are listed in Table 1.
Advances in synthetic-CT generation from MRI have been achieved with deep learning methods. Results are divided into three categories: studies including less than 18 patients, studies including 19 to 40 patients and studies including more than 40 patients. Red dotted lines represent the median values. The median values are: 74.2 HU for the brain, 77.9 HU for H\&N, and 42.4 HU for the pelvis. The selected values are listed in the Additional tables 1 to 4 (Appendix A). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### Table 1

<table>
<thead>
<tr>
<th>Type of metrics</th>
<th>Metric</th>
<th>Definition</th>
<th>Ideal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image evaluation</td>
<td>Intensity-based metrics</td>
<td>ME: mean error</td>
<td>$\mu_x$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAE: mean absolute error</td>
<td>$\mu_y$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSNR: peak signal to noise ratio</td>
<td>$\sigma_x$, $\sigma_y$</td>
</tr>
<tr>
<td></td>
<td>Structural similarity metrics</td>
<td>SSIM: structural similarity metric</td>
<td>$K_1$, $K_2$, $\sigma_{pCT}$, $\sigma_{sCT}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE: mean square error</td>
<td>$\mu_{pCT}$, $\mu_{sCT}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE: root mean square error</td>
<td>$\sigma_{pCT}$, $\sigma_{sCT}$</td>
</tr>
<tr>
<td></td>
<td>Geometric fidelity metrics</td>
<td>NCC: normalized cross-correlation</td>
<td>$\mu_{pCT}$, $\mu_{sCT}$</td>
</tr>
<tr>
<td>Dose evaluation</td>
<td>Dose difference metrics</td>
<td>Voxel-to-voxel dose differences</td>
<td>$\delta_{pCT}$, $\delta_{sCT}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DVH difference</td>
<td>$\delta_{DVH}$</td>
</tr>
<tr>
<td>Gamma analysis metrics</td>
<td>Mean gamma</td>
<td>Percentage of pixels/voxels with a gamma value lower than 1</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: N: number of voxels; ME: Mean square error; Q: range of voxel value of sCT and reference CT; x: reference CT; y: sCT; $\mu_x$: mean value of x; $\mu_y$: mean value of y; $\delta^2_x$: variance of x; $\delta^2_y$: variance of y; $C_1$ and $C_2$ are expressed as $(k_1 Q)^2$ and $(k_2 Q)$; $\sigma_{pCT}$, $\sigma_{sCT}$: HU value of the reference CT, $sCT$: HU value of the sCT, $\mu_{pCT}$: mean intensity value of the reference CT, $\mu_{sCT}$: mean intensity value of the sCT, $\sigma_{pCT}$: standard deviation of the reference CT and sCT, $V$: volume on CT and sCT, $d_{ave}$: absolute Euclidean distance; $S_{ave}$: surface of the automated segmentation volume; $S_C$: surface of the reference organ delineation.

---

**Fig. 6.** Mean absolute error (MAE) results for body structure between reference CT and sCT generated with a deep learning method for studies including the brain, H\&N, liver, abdomen, and pelvis. Each marker represents a study result. Full markers represent generator-only models and empty markers generative models with adversarial. Results are divided into three categories: studies including less than 18 patients, studies including 19 to 40 patients and studies including more than 40 patients. Red dotted lines represent the median values. The median values are: 74.2 HU for the brain, 77.9 HU for H\&N, and 42.4 HU for the pelvis. The selected values are listed in the Additional tables 1 to 4 (Appendix A). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
1. **Intensity-based evaluation**

Only three studies of sCT generation from MRI did not report MAE values [56,73,98]. Some articles reported MAE in bone or soft tissue while others reported MAE in anatomical structures such as the kidneys, bladder, or rectum [20,45,47,76,88]. Fig. 6 summarizes the MAEs of the studies on brain, H&N, liver, abdomen, and pelvis sCT generation from MRI.

Eighteen studies reported mean error (ME) results. Fig. 7 details the MEs of the studies on brain, H&N, abdomen, and pelvis sCT generation from MRI. For the pelvis, three studies provided ME values for the bladder, rectum and soft tissue [20,47,88]. Some studies have illustrated MAE or ME for one or several slices. Such difference maps allow for qualitative comparisons and spatial analyses.

The peak signal to noise ratio (PSNR) is the simplest and most widely used fidelity measure (full-reference quality metric), which is related to the distortion metric, the MSE. Twenty-two studies on sCT generation from MRI reported PSNR results. Fig. 8 details PSNR results for the brain, H&N, and pelvis sCT generation from MRI studies.

Four studies reported MSE values in the brain and pelvis. Only three studies reported the root MSE (RMSE) value in the brain [96], the breast [56], and the abdomen [50]. Although the MSE is an attractive measure due to its simplicity of calculation, MSE/PSNR can be a poor predictor of visual fidelity in images [116].

More sophisticated measures have been developed to take advantage of the known characteristics of the human visual system. Wang et al. [117] proposed a structural similarity metric (SSIM) to capture the loss of image structure due to variations in lighting (contrast or brightness changes). The SSIM captures image distortion as a combination of three types of distortion: correlation, contrast, and luminance.

2. **Geometric fidelity evaluation**

Geometric fidelity is based on delineated structures. Nineteen articles reported dice score coefficients (DSCs) between sCT and reference CT for bone, air, or body structures. One study reported DSCs for the bladder and rectum [47]. DSCs were between 0.85 and 0.99 for body and were higher than 0.68 and up to 0.93 for bone structure.

Only two studies reported Hausdorff distance (HD) values for the H&N area [79] and the pelvis [77]. Only one study reported mean absolute surface distance (MASD) values for body, bone, bladder, and rectum volumes [47]. Five studies reported normalized cross-correlation
(NCC) values in the brain, liver, and pelvis [32,77,106–108].

The penultimate columns of Additional Tables 1, 2, 3, and 4 in Appendix A list the image results of sCT generated from MRI.

3. Dose evaluation

In MRI-only workflows for radiotherapy, a sCT is generated to perform dose calculation. In this context, studies have proposed dosimetric evaluation of dose calculation from sCT with DVH, voxel-to-voxel dose differences or gamma index analysis. Most studies evaluated dose calculation with photon particles, while nine studies investigates sCT dose uncertainties with protons [45,55,59,71,74,77,96,107,108].

v) Dose-volume histogram (DVH)

DVH is a widely used tool in routine clinical radiotherapy. All treatment planning systems allow for the analysis of dose distributions through DVHs. Twenty-two sCT studies reported dose differences at DVH specific points. Eighteen studies reported mean dose differences in selected volume (PTV, CTV, OAR).

vi) Voxel-to-voxel dose difference

The dose difference is defined as the difference between the dose distribution computed on the reference CT and the sCT. The dose difference can be expressed as absolute value (Gy) or relative to the reference dose (%).

Several studies reported mean absolute dose error to express dose uncertainties and mean dose error to express systematic dose uncertainties [47,49,50,70,77,96,101]. Some studies have provided dose differences using dose thresholds such as doses higher than 90% of the prescribed dose, while others have illustrated dose difference maps that allow qualitative and spatial analyses.

vii) Gamma index analysis

Gamma analyses allow spatial analysis (through gamma maps) of dose distributions calculated from sCT compared to those calculated from a reference CT [118]. Gamma analysis can be performed in two or three dimensions. This analysis combines dose and spatial criteria. Several parameters need to be set to perform a gamma analysis, including dose criteria, distance-to-agreement criteria, local or global analysis, and dose threshold. Interpretation and comparison between studies of gamma index results are challenging because they depend on
the chosen parameters, dose grid size, and voxel resolution [119]. The gamma results can be expressed as gamma pass-rate (percentage of pixels/voxels with a gamma value lower than 1) or mean gamma. Twenty-eight articles reported gamma pass-rate results. Only one study reported mean gamma values in the pelvis [20].

Fig. 9 summarizes the gamma pass-rate results between reference dose distribution and sCT dose distribution for several anatomical localizations. The mean gamma pass-rates were above 89% for all localizations and up to 100%, depending on gamma criteria.

viii) Specific metrics for proton dose calculation

Proton ranges along the beam paths were compared for dose distributions on the reference CT and sCT. In protontherapy, the range of the proton beam strongly depends on the stopping power ratio (SPR) of a given tissue relative to water, which can be determined using the electron density and effective atomic number through the Bethe-Bloch equation. The range is defined as at the 80% distal dose falloff along each beam direction. Several studies have reported the results of range shift or range difference (in mm) per beam [45,55,71,108].

The last columns of Additional Tables 1, 2, 3, and 4 in Appendix A summarize the dose results of the DL sCT generation studies for MRI dose calculation.

D. Image and dose results per anatomical localization

1. Brain

Twenty-four studies of the brain were performed between 2017 and 2021. Additional Table 1 (Appendix A) summarizes the DL networks and the image and dose metrics results of these studies on brain sCT generation from MRI in radiotherapy. T1-weighted (T1w) sequences were mostly used for generating sCTs 88% of the reviewed studies). For sCT evaluation, all brain studies reported MAEs, which varied from 44 to 129 HU for the whole brain (Fig. 6). For the brain, the MAEs for bone structure were above 174 HU and up to 399 HU in one study. Koike et al. [92] trained a cGAN network with only a T1 sequence or a combination of several sequences (T1w, T2w and FLAIR). The multi-sequence training showed a decrease in MAEs results for the body, soft tissue and bone (between −8 and −33 HU) [92]. Alvarez Andres et al. reported MAE values for CNN and U-Net networks, with higher values for the U-Net network than for the CNN network for the head (an increase of 9 HU) [43]. They also investigated the influence of several sequences (T1, T1-Gd, and T2 FLAIR images) as input in the CNN network. The MAEs were higher with a FLAIR sequence as input in the CNN network than those with T1 sequence (increase of 34 HU). The MAEs were also higher with contrast-enhanced T1-weighted MRI (T1-Gd) than those for T1w MRI (from +3 to +32 HU). Only four studies reported ME values [44,45,55,70], which ranged between −4 HU and
13 HU for the whole brain (Fig. 7). The PSNR values were above 24 dB for all brain studies [29,30,32,57,62,72,80,96,97,105] (Fig. 8). The DSCs were above 0.96 for the body, 0.69 for bone, and 0.70 for air structures [42,45,58,70,72]. For SSIM, values varied from 0.63 to 0.94 [57,62,72,96,105]. For NCC, two studies reported values of 0.96 [32,108]. Massa et al. [72] trained models on four different MR sequences: CUBE-FLAIR, T1, T1 post contrast and T2 fatsat. No sequence was statistically better on all the metrics (MAE, PSNR, SSIM, DSC).

Among the 24 brain studies, only 14 reported a dose evaluation [42,43,45,48,55,58,59,70,84,92–98,108]. Five studies reported results for proton dose planning [45,55,59,96,108]. All reported DVH mean dose differences were below 2% [42,45,48,59,70,84,92–98,108]. For most of the studies, gamma pass-rates were above 89% with the most restrictive criterion (1%/1 mm), and except for one study above 95% for other criteria (Fig. 9). One study [55] reported a mean gamma pass-rate of 89% with 2%/2 mm criteria. With the multi-sequence training, Koike et al. showed an increase in gamma pass-rates (between 0.1% and 1%), compared to single sequence training [92].

2. Head and Neck (H&N)

Nine DL sCT generation studies were performed in H&N radiotherapy. Additional Table 2 (Appendix A) summarizes the DL networks and the image and dose metrics results of these studies. The MRI sequences used in the H&N sCT studies were T1 and T2. Four studies used the Dixon reconstruction [49,50,79,101]. MAE and ME metrics have been widely reported in the literature. MAEs varied from 65 to 131 HU for the body or head structures (Fig. 6). For the H&N, the MAEs for bone were above 166 HU and up to 357 HU in one study [46]. Qi et al. [50] used multi-sequence input (T1w, T2w, contrast-enhanced T1, and contrast-enhanced T1 Dixon water) images to train a cGAN. The multi-sequence training showed a decrease in MAE for the body, soft tissue, and bone and an increase in PSNR, SSIM, and DSC. They also compared cGAN and U-Net networks. With cGAN and single-sequence input, the MAE, PSNR, and DSC were higher than those obtained using U-Net. For the body, the MEs were mostly around 0 HU, above –6 HU and up to 37 HU (Fig. 7). Five studies reported ME for air, bone, or soft tissue [9,46,49,79,100]. For bone structure, MEs were higher up to 247 HU. PSNR and SSIM were only reported in two studies [50,94]. The PSNR results were approximately 28 dB (Fig. 8). The SSIMs were between 0.78 and 0.92. For bone structure, DSC values were between 0.70 and 0.89 [50,70,71,79,94].

DVH dose difference was performed in nine studies, with a mean difference less than 1.6%. Klages et al. reported a mean dose (Dmean) to the parotid glands below 1% and the maximum dose (Dmax) to the spinal cord below 1.5% [101]. In the only one protontherapy study, the dose differences reached 8% for some OARs [71].

The gamma pass-rates were above 95% for the most restrictive criterion (2%/2mm) and above 98% for the other criterion (3%/3 mm). With the multi-sequence training, Qi et al. showed non-significant gamma pass-rate results [50]. In the same study, they also found higher gamma pass-rates for cGAN than for U-Net architectures.

3. Breast

Two DL sCT generation studies were carried out for breast radiotherapy. Additional Table 3 (Appendix A) summarizes the DL networks and the image and dose metrics results of the studies for breast sCT generation from MRI in radiotherapy. These two studies were based on MR images from a low field (0.35 T) MRI device. Jeon et al. [73] only reported DSC values for two patients. Olberg et al. reported a PSNR of 72 dB, an SSIM of 0.999, and an RMSE of 17 HU [56]. For dose results, they reported gamma pass-rate higher than 98% with 2%/2 mm criteria.

4. Liver

Three DL sCT generation studies were carried out for liver radiotherapy. Additional Table 3 (Appendix A) summarizes the DL networks and the image and dose metrics results of the studies for liver sCT generation from MRI in radiotherapy.

Two of the three liver studies used T1 sequence. The MAEs varied from 72 to 94 HU for body structure between studies. Fu et al. [33] compared cGAN and cycle-GAN DLMs for data from three patients. The MAEs were higher for cycle-GAN than for cGAN. The PNSR values were above 22 dB. The NCC values were 0.92 in two liver studies [106,107].

DVH dose difference were calculated in the three liver studies, with the mean differences below 1%. In one study, the dose difference in OARs was less than 0.6% [33]. The dose difference in PTV (Dmax) was less than 1.1% [106,107]. The gamma pass-rates were above 90% for the most restrictive criterion (1%/1mm) and above 95% for the other criteria (Fig. 9). In the study by Fu et al. [33], the gamma pass-rates were higher for a cGAN DLM than those for a cycle-GAN DLM.

5. Abdomen

Six DL sCT generation studies were carried out for abdomen radiotherapy. Additional Table 3 (Appendix A) summarizes the DL networks and the image and dose metrics results of the studies for abdomen sCT generation from MRI in radiotherapy. Acquisitions were performed in breath hold inspiration for two studies on 0.35 T MRI device [31,33].

The MAEs varied from 55 to 94 HU for body structure between abdomen studies. MAEs in lungs were 105 HU in two studies [74,76]. In Florkov et al. [74], PSNR value was 30 dB and DSC values were 0.76 for bone and 0.92 for lungs. Mean dose differences were lower than 1% and gamma pass-rate above 98% in 2%/2mm.

6. Pelvis

Eighteen DL sCT generation studies for the pelvis were performed between 2016 and 2020. Additional Table 4 (Appendix A) summarizes the DL networks and the image and dose metrics results of these studies. Most MRI sequences in these pelvis studies were T2 sequences. T1 sequences [30,53,78] or Dixon reconstruction [52,54,93] were also used to generate sCT from MRI in the pelvis.

The reported MAEs were 27–65 HU for body structure (Fig. 6) and around 120 HU and up to 250 HU for bone [20,31,47,51,53,78,99]. Fu et al. compared training in 2D and 3D in four patients [78], reporting higher MAEs for 2D than for 3D training (+2–5 HU). Largent et al. compared U-Net and GAN networks with different loss functions [20]. With the L2 loss function, U-Net showed lower MAEs than those for GAN. For all studies, the MEs were generally near to 0 HU for the whole body structure (Fig. 7). One multicenter study reported an ME of –18 HU [88]. For the pelvis, the MEs in the bone were up to 141 HU in one study [31]. The reported PSNRs were between 24 and 34 dB (Fig. 8). Only one study reported SSIM in the pelvic area [75]. Only two studies reported DSC for the body (0.85 and 0.99) [47,77]. The DSCs for bone ranged between 0.70 and 0.93 [47,52,53,75,78]. Only one study reported a DSC for the bladder and rectum of 0.9 [47]. Among the 18 pelvic studies, only nine reported dose evaluations. Liu et al. [107] performed proton dose planning. Most studies reported dose differences below 1.5% for target volumes and OARs. Arabi et al. reported maximum dose differences below 0.5% for the bladder and rectum between 1.1% and 2.9% for the CTV and PTV. Some studies reported a very low dose difference (less than 0.6%) for PTV, bladder,
rectum, and femoral heads [20,51]. However, Liu et al. reported dose differences up to 5% in the rectum and up to 11% in the bladder [77].

The gamma pass-rates were above 89% for the most restrictive criterion and generally above 95% for the other criteria (Fig. 9). In a study using prostate data in training to generate sCT of the rectum and cervix [93], the gamma pass-rates were around 91% for gamma criteria of 2%/2 mm.

Discussion

This article reviewed deep learning methods used to generate sCT from MRI in radiation therapy, and their associated image and dose uncertainties. Two types of DL architectures are widely used; generator-only, and GAN. The most recent DLMs were cGAN and cycle-GAN. A variety of metrics for image evaluation (image intensity and geometric fidelity) has been proposed. The median MAE results were 76 HU for head localization (brain and H&N) and liver, and 42 HU for the pelvic area. Dose evaluations consisted in DVH comparisons, voxel-to-voxel dose differences, or gamma index analyses. The mean dose differences were below 1% in the H&N, liver, breast, and pelvis sCT studies. In brain sCT studies, the mean dose difference was below 2%. For most of the studies, the gamma pass-rates were above 95% (with 2%/2 mm and 3%/3 mm criteria) (Fig. 9).

In radiotherapy, the first sCT generation methods from MRI were bulk density and atlas-based. Other ML methods (non-DLM) have also been investigated, including patch-based or random forest [10,120]. This review focused on DLMs which are the most recent methods with the first study in the pelvic area reported in 2016. Different neural network architectures have been used in the literature with multiple parameters to be set. Compared to other sCT generation methods, DLMs have fast computation times, and do not necessarily require deformable inter-patient registration. sCT generation DLMs have just been commercially available for a clinical use [79,121]. To our knowledge, no open source software is available for sCT generation from MRI with a DLM. Each research team has developed their own DLM with hyper-parameter tuning. This review was not able to identify the most “accurate” DL architecture. Although GAN DLMs are the most recent, for now they do not outperform generator-only DLMs (Figs. 6, 7, 8, and 9). Moreover, we acknowledge that studies are not directly comparable due to the great disparities in input data (imaging protocol, scanner parameters, etc.).

Training strategy depends on the number of available data. If you have few data (less than 19 patient data), a leave-one-out strategy is recommended. Stronger results are obtained if you have paired data (same data in open access from the Gold Atlas Project) and generally above 95% for the other criteria (Fig. 9). In a study using prostate data in training to generate sCT of the rectum and cervix [93], the gamma pass-rates were around 91% for gamma criteria of 2%/2 mm.

Recent studies reported global imaging evaluation metrics, without local or spatial analysis. Hemsley et al. [95] proposed a detailed sCT evaluation with uncertainty heatmaps. Models were proposed to spatially quantify intrinsic and parameter sCT uncertainties [122]. With this method, uncertainty maps are a second output of the DL network [123]. Moreover, an analysis based on image gradient could be performed.

Reviewed articles aimed to show accuracy of sCT generation from MRI compared to a reference CT. In the future, we can imagine an MRI-only workflow without any CT acquisition. In this case, image evaluation metrics without reference need to be developed. Before any use in clinical practice, commissioning and quality assurance process must be implemented [124]. Practical guidelines on the use of MRI for external radiotherapy treatment planning were recently proposed by a multidisciplinary working group of the Institute of Physics and Engineering in Medicine (IPEM) [125]. This document overviews all the aspects of MRI implementation for radiotherapy are described (MR safety, training and education, patient set-up, MRI sequence, MR quality assurance, etc.).

To date, few DL studies have been carried out on abdomen, liver, breast, or H&N radiotherapy. This limited number may be due to the small number of patients undergoing MRI for liver or abdomen radiotherapy compared to brain or prostate radiotherapy. Moreover, the standard acquisition of breast MRI is not in radiotherapy treatment position. Number of breast studies are increasing with the availability of MRI acquisitions from MRI-linac device. The lack of sCT generation from H&N and abdomen MRI may be due to the complexity of these anatomical localizations with the large part of heterogeneities. MRI in the treatment position can be challenging for H&N acquisitions because specific coils are used [83]. No study has yet investigated lung sCT generation from MRI with DLM. Movement is a huge challenge for MR imaging.

Several MRI sequences have been used to generate sCT from MRI in radiotherapy, with T2w sequences the most common. Some studies used specific reconstruction techniques such as mDixon or FLAIR. The FLAIR sequence is an inversion-recovery sequence. This sequence improves the detection of lesions of the cerebral parenchyma and enables visualization of edemas. It also facilitates the detection of white matter pathologies (softening, demyelination process), which appear as hypointensities.

Three studies investigated the impact of MISO, compared to a SISO [50,92,103]. MISO has the advantage of better tissue description [127,128]. Moreover, three studies evaluated the impact of generating a sCT from a device other than the one used during training [88,89,96]. Such “multidevice” or “multicenter” impact is a key challenge to a commercial development.

The emergence of linacs combining MRI in the treatment room (MRI-linacs) increase the willingness of MRI-only workflow radiotherapy [126]. Some reviewed studies already used DLM for sCT generation from this device [31,33,56,73]. In this context, dose planning need to consider the presence of magnetic field, with the electron return effect [127,128]. Moreover, on MRI-linac, an MR image is acquired for each
fraction. This image could be used to perform dose monitoring or replanning with the use of a DLM, in the context of MR-guided adaptive radiotherapy [129].

Conclusions

The emergence of DL allows the fast and accurate generation of sCT from MRI in radiotherapy. In the literature, a variety of DLMs have been applied, mainly for brain and pelvis cancer, and also for H&N and liver. Each DL study has showed particularities in terms of hyperparameters or loss functions. Different MRI sequences are used depending on the clinical setting. Among Patch-Based, Atlas-Based, and Bulk Density Methods. Int J Radiat Oncol Biol Phys 2019;103:479–86. 10.1016/j.ijrobp.2018.08.094.


