Received: 10 May 2024 | Revised: 11 January 2025 | Accepted: 14 January 2025

DOI: 10.1002/mp.17655

RESEARCH ARTICLE

MEDICAL PHYSICS

# TransAnaNet: Transformer-based anatomy change prediction network for head and neck cancer radiotherapy

Meixu Chen[1] | Kai Wang[1,2] | Michael Dohopolski[1] | Howard Morgan[1,3] | David Sher[1] | Jing Wang[1]

[1] Medical Artificial Intelligence and Automation (MAIA) Lab, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, Texas, USA

[2] Department of Radiation Oncology, University of Maryland Medical Center, Baltimore, Maryland, USA

[3] Department of Radiation Oncology, Central Arkansas Radiation Therapy Institute, Little Rock, Arkansas, USA

**Correspondence**
Jing Wang, Medical Artificial Intelligence and Automation (MAIA) Lab, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX, 75235, USA.
Email: Jing.Wang@UTSouthwestern.edu

## Abstract

**Background:** Adaptive radiotherapy (ART) can compensate for the dosimetric impact of anatomic change during radiotherapy of head–neck cancer (HNC) patients. However, implementing ART universally poses challenges in clinical workflow and resource allocation, given the variability in patient response and the constraints of available resources. Therefore, the prediction of anatomical change during radiotherapy for HNC patients is of importance to optimize patient clinical benefit and treatment resources. Current studies focus on developing binary ART eligibility classification models to identify patients who would experience significant anatomical change, but these models lack the ability to present the complex patterns and variations in anatomical changes over time. Vision Transformers (ViTs) represent a recent advancement in neural network architectures, utilizing self-attention mechanisms to process image data. Unlike traditional Convolutional Neural Networks (CNNs), ViTs can capture global contextual information more effectively, making them well-suited for image analysis and image generation tasks that involve complex patterns and structures, such as predicting anatomical changes in medical imaging.

**Purpose:** The purpose of this study is to assess the feasibility of using a ViT-based neural network to predict radiotherapy-induced anatomic change of HNC patients.

**Methods:** We retrospectively included 121 HNC patients treated with definitive chemoradiotherapy (CRT) or radiation alone. We collected the planning computed tomography image (pCT), planned dose, cone beam computed tomography images (CBCTs) acquired at the initial treatment (CBCT01) and Fraction 21 (CBCT21), and primary tumor volume (GTVp) and involved nodal volume (GTVn) delineated on both pCT and CBCTs of each patient for model construction and evaluation. A UNet-style Swin-Transformer-based ViT network was designed to learn the spatial correspondence and contextual information from embedded image patches of CT, dose, CBCT01, GTVp, and GTVn. The deformation vector field between CBCT01 and CBCT21 was estimated by the model as the prediction of anatomic change, and deformed CBCT01 was used as the prediction of CBCT21. We also generated binary masks of GTVp, GTVn, and patient body for volumetric change evaluation. We used data from 101 patients for training and validation, and the remaining 20 patients for testing. Image and volumetric similarity metrics including mean square error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM),

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2025 The Author(s). *Medical Physics* published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine.

Dice coefficient, and average surface distance were used to measure the similarity between the target image and predicted CBCT. Anatomy change prediction performance of the proposed model was compared to a CNN-based prediction model and a traditional ViT-based prediction model.

**Results:** The predicted image from the proposed method yielded the best similarity to the real image (CBCT21) over pCT, CBCT01, and predicted CBCTs from other comparison models. The average MSE, PSNR, and SSIM between the normalized predicted CBCT and CBCT21 are 0.009, 20.266, and 0.933, while the average Dice coefficient between body mask, GTVp mask, and GTVn mask is 0.972, 0.792, and 0.821, respectively.

**Conclusions:** The proposed method showed promising performance for predicting radiotherapy-induced anatomic change, which has the potential to assist in the decision-making of HNC ART.

KEYWORDS
adaptive radiation therapy, anatomy change prediction, head and neck cancers, transformer network

## 1 | INTRODUCTION

Head and neck cancer (HNC) is one of the most common types of cancer in the United States, with an estimated 66 920 new cases anticipated in 2023.[1,2] In current clinical practice, the predominant approach to managing locally advanced HNC involves the application of radiotherapy (RT), with intensity-modulated radiotherapy (IMRT) serving as the standard treatment technique because of its highly conformal dose distribution and steep dose gradient. Nevertheless, a subset of HNC patients experience substantial anatomical and geometric variations in both target volumes and organs at risk (OARs) during IMRT. These deviations may lead to inadvertent overdosing of normal tissue and/or underdosing of target structures, thereby compromising the therapeutic benefits of this highly conformal approach. Previous studies have suggested that the affected HNC patient population can range from 21 to 66%.[3–6] While image-guided radiotherapy (IGRT) is routinely employed in most treatment centers to correct day-to-day positional deviations between planning computed tomography (CT) and daily images, it remains inadequate for addressing internal anatomical changes of target or normal tissues.

Adaptive radiotherapy (ART) is desired for mitigating unfavorable dosimetry outcomes for HNC patients experiencing anatomical changes during IMRT. The dosimetric benefits and favorable clinical outcomes stemming from ART have been extensively reported.[3,5,7,8] However, the extensive adoption of ART for all patients is challenging. As a resource-intensive technique, it requires time-consuming and labor-intensive procedures including repeated imaging, contouring, replanning, and dosimetric analysis.[3,6,9,10] Therefore, accurately predicting the anatomy change during HNC RT is crucial for identifying individuals who would benefit from adaptive replanning or receiving online ART treatment.

Numerous investigations have been conducted to explore various clinical and imaging factors that predict the necessity for replanning in ART. These studies have aimed to establish criteria for identifying anatomical changes that could result in suboptimal dosimetric outcomes in HNC RT. Statistical analyses have pinpointed several potential indicators, including tumor location, patient age, body mass index (BMI), intended dosage to the parotid glands, initial volume of the parotid glands, initial tumor and involved nodal volumes, and the extent of their overlap with the planning target volume (PTV).[3,4,7,11,12] More recent advances in radiomics and machine learning have led to the development of imaging-based multivariable binary prediction models specifically for selecting HNC ART patients.[13–17] These models, trained using pretreatment CT or MRI radiomics data, with or without additional clinical factors, have demonstrated superior binary prediction accuracy compared to models based solely on clinical factors. However, this binary prediction procedure simplifies the decision-making into a yes/no outcome, failing to capture the complex nature of anatomical and tumor changes. It does not provide insights into the primary causes of dose variation nor guidance for clinicians during the replanning process.

Different from binary prediction, image prediction of anatomy change could offer a more substantial and perceptible set of prognostic data for physicians in the decision-making of ART. Deep-learning-based image prediction and generation have recently emerged as a focal area of interest, spurred by the remarkable advancements in Convolution Neural Network (CNN)-based U-NET, Generative Adversarial Networks (GANs), and Vision Transformer (ViT)-based networks. Among these models, ViT-based models often outperform others due to their ability to capture global dependencies, robustness in training, scalability with large datasets, and flexibility for fine-tuning and constructing hybrid models.[18–21] These technological evolutions have culminated in the development of a spectrum of research models and even commercial products dedicated to tasks such as text-image translation, image inpainting, motion prediction, and next video frame prediction.[22–26] In the field of oncology, inspired by the achievement in general image prediction tasks, several studies have demonstrated promising image-based tumor anatomy change prediction results on different cancer sites, including brain, lung, and pancreas, for either tumor growth prediction or tumor treatment response prediction.[27–30] Their prediction results presenting in the form of 3D images show enhanced visual representation and can be used for image segmentation, tumor aggressiveness quantification, treatment response prediction, or assisting in determining the necessity for RT replanning.

In this work, inspired by the promising performance of the pretreatment image-based ART eligibility prediction models,[13,16] and the recent advance in ViT-based image generation and prediction studies, we propose to construct an image-to-image model for HNC patient anatomy change prediction. As patient initial anatomy, volume of treatment target, and treatment dose distribution are all clinically available and important factors contributed to anatomy change during RT,[3,4,17,31] we hypothesize that a ViT-based deep image prediction model (TransAnaNet) constructed by using multifaceted analyses of planning CT, daily Cone Beam CT (CBCT), and dosimetric data of HNC patients is indicative of patient anatomy change during RT, which can assist in the decision making of HNC ART. Previous studies indicated that the optimal timing for replanning is around the third and the fourth week of treatment on the offline adaptive setting.[32–34] Based on these findings, we selected the patient anatomy change at the end of the fourth week (Fraction 21st) as the prediction target of this study. We leveraged a retrospective database to investigate the accuracy and characteristics of the anatomy change prediction model. To the best of our knowledge, this is the first deep-learning model for predicting HNC patient anatomy change during RT.

## 2 | METHODS

### 2.1 | Dataset and preprocessing

#### 2.1.1 | Patient and data

This retrospective study was reviewed and approved through the institutional review board (IRB# 082013-008) of the University of Texas Southwestern Medical Center. Patients treated between April 2014 and October 2019 at the University of Texas Medical Center were included if they were diagnosed with locally advanced HNC (including oropharynx, supraglottic, glottic, or hypopharynx) and completed a full course of conventionally fractionated definitive RT with daily or weekly CBCT imaging. Patients with all primary structures and nodal structures receiving 70-Gy dose were selected for evaluation. Those who had a prior history of RT to the head and neck, received prior induction chemotherapy, had distant metastases (DM), or had the presence of a separate active malignancy, were excluded.

We collected four sets of image data for each of the patients for analysis: the baseline CT simulation scan (CT), the 3D dose map from the treatment planning system, CBCT prior to the initial fraction (CBCT01), and CBCT prior to Fraction 21 (CBCT21). At baseline simulation, all patients were simulated on a Philips 16-slice Brilliance large-bore CT simulator with iodinated IV contrast. CBCT imaging was performed either daily or weekly on a Varian TrueBeam™ machine (Varian Medical Systems, Palo Alto, CA). See the supplementary file for more details about imaging parameters (Table S1).

#### 2.1.2 | Data preprocessing

All initial segmentations, including patient body contours, on CT were delineated by a board-certified radiation oncologist specializing in HNC RT. Nodal contours were combined into a single structure to simplify the analysis. For CBCT images, all segmentations were deformed from CT with rigid and/or deformable image registration in Velocity (Varian Medical Systems, Palo Alto, CA). All generated contours were manually edited by the physician to verify the inclusion of all affected mucosa if applicable and to exclude any incident bone, air, cartilage, or adipose tissue that may be overlapping the GTV boundaries. This process was repeated for both CBCT01 and CBCT21. Then, all the image data including the dose map and segmented RT structure masks were rigid registered to CBCT21, the voxel sizes were resampled to 2 × 2 × 4 mm$^3$, and the images were center cropped to 128 × 128 × 32 matrices for the model training process. CT and CBCT images were clipped to [−1000, 1000] HU and min–max normalized to [−1, 1], dose maps were *z*-score normalized, segmentation masks were binarized to 0 and 1.

### 2.2 | Transformer-based anatomy prediction network

#### 2.2.1 | Workflow

Recently, ViT architectures have been applied to medical image registration tasks. Compared with convolutional neural networks (CNNs), ViT models could predict more

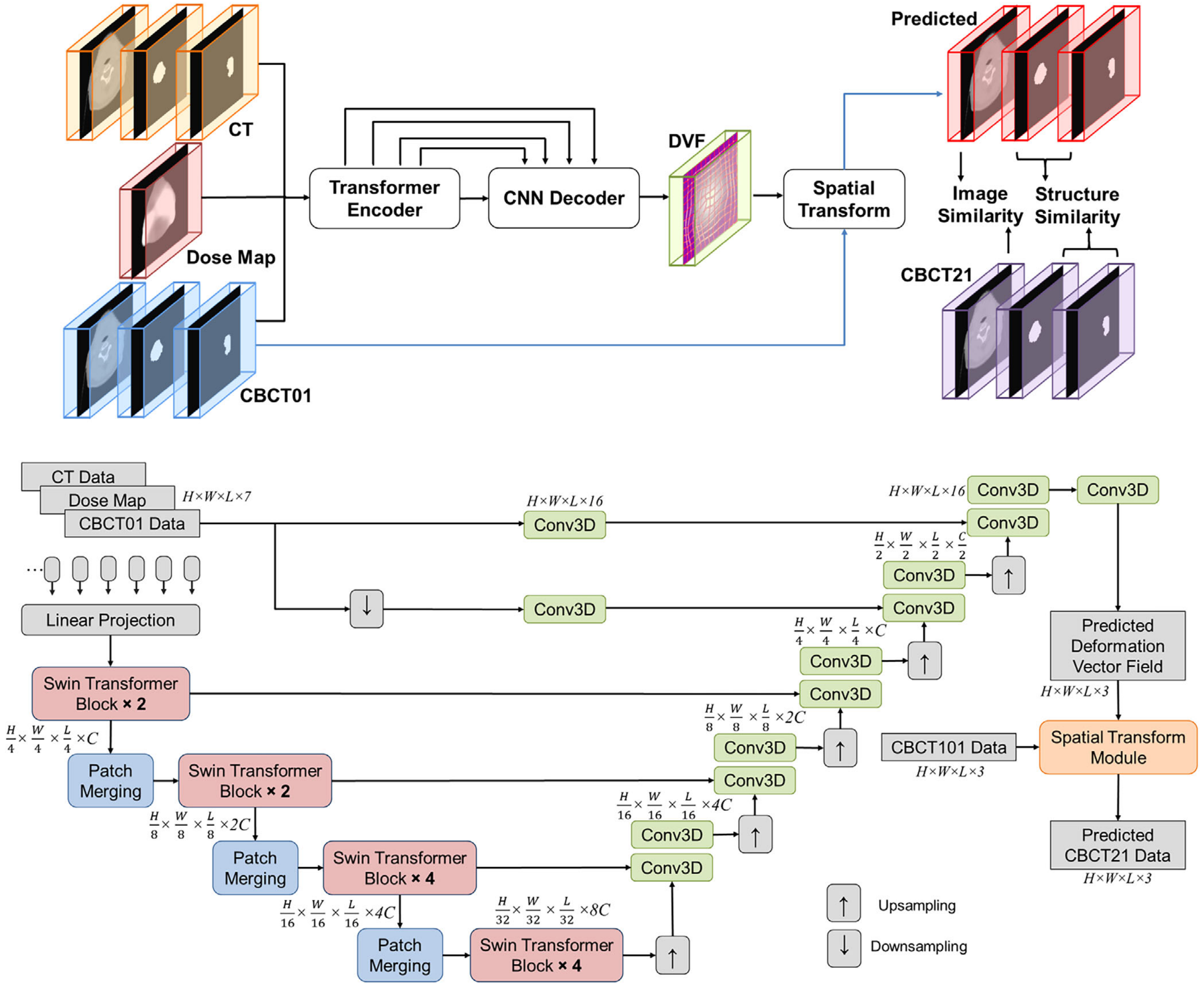
Predicted
CT
DVF
Transformer Encoder
CNN Decoder
Spatial Transform
Image Similarity
Structure Similarity
Dose Map
CBCT21
CBCT01
CT Data
Dose Map
CBCT01 Data
$H \times W \times L \times 7$
$H \times W \times L \times 16$
Conv3D
Linear Projection
Swin Transformer Block × 2
$\frac{H}{4} \times \frac{W}{4} \times \frac{L}{4} \times C$
Patch Merging
$\frac{H}{8} \times \frac{W}{8} \times \frac{L}{8} \times 2C$
Swin Transformer Block × 4
$\frac{H}{16} \times \frac{W}{16} \times \frac{L}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times \frac{L}{32} \times 8C$
$\frac{H}{2} \times \frac{W}{2} \times \frac{L}{2} \times \frac{C}{2}$
Predicted Deformation Vector Field
$H \times W \times L \times 3$
CBCT101 Data
Spatial Transform Module
Predicted CBCT21 Data
Upsampling
Downsampling

**FIGURE 1** The overall framework (a) and detailed architecture (b) of the proposed transformer-based head and neck cancer patient anatomy change prediction (TransAnaNet) model. The hybrid Transformer-ConvNet network takes seven inputs: planning CT image, GTVp mask on planning CT, GTVn mask on planning CT, planned dose map, initial fraction CBCT (CBCT01), and GTVp and GTVn mask on it. The network predicts a nonlinear warping deformation vector field (DVF), which is then applied to the CBCT01 image through a spatial transformation function to generate the predicted patient anatomy (Fraction 21). For training data, the image similarity and structure similarity between the predicted and real CBCT21 are used as part of the loss functions to update the model, they are also used to evaluate the performance of the constructed model.

precise deformation vector fields (DVFs) between the moving and fixed images due to the capability of capturing long-range spatial information.[35–37] Inspired by the success of ViT-based method for medical image registration, we expanded this idea from spatial relationship prediction between two known images, for example, image registration, to spatial–temporal relationship prediction between known images and future images, for example, anatomy change prediction.

The proposed framework of using transformer for patient anatomy change prediction is shown in Figure 1a. A ViT-encoder and a convolution decoder formed the hybrid Transformer-ConvVNet (TransAnaNet) for anatomy change prediction. The hybrid network takes seven inputs: planning CT image, GTVp mask on planning CT, GTVn mask on planning CT, planned dose map, initial fraction CBCT (CBCT01), and GTVp and GTVn mask on it. The network predicts a nonlinear warping DVF as the intermediate output, which is then applied to the CBCT01 image through a spatial transformation function to generate the predicted patient anatomy (CBCT 21). The GTVp and GTVn masks on CBCT21 are also predicted by deforming the corresponding masks on CBCT01. During the model training stage, the image similarity and structure similarity between the predicted and real CBCT21 are used as part of the loss functions to optimize the model, while in the model validation stage, they are also used to evaluate the performance of the constructed model. Of note, here we set CBCT01 as the baseline image for deformation since it shares the same imaging protocol as the prediction target CBCT21. It can be replaced with planning CT or any other early fraction CBCTs for anatomy change prediction as long as the dataset is consistent. On the other hand, the target image can be replaced by other late fraction CBCTs according to the needed prediction time point.

### 2.2.2 | Network structure

Figure 1b shows the network architecture of the proposed TransAnaNet. The model structure was modified based on a deep medical image registration model named TransMorph, which has state-of-the-art performance on image registration tasks.[37] We used the Swin Transformer as ViT encoder to capture the spatial correlation between different regions in the input dose map, planning CT, CBCT01 images, and their GTV masks. Compared to the original version of ViT module proposed by Dosovitskiy et al.,[38] Swin Transformer module can build hierarchical feature maps by merging image patches in deeper layers and has linear computation complexity to input image size, which is of high efficiency for model training.[38,39] Then a typical convolutional decoder is used to process the information provided by the transformer encoder into a dense displacement hidden field. Cross-layer connections were used to maintain the localization information between the encoder and decoder.

The encoder of the network first splits the input image, dose, and mask volumes into nonoverlapping 3D patches, each of size $7 \times P \times P \times P$, where 7 in the number of input channels corresponding to different modalities and GTVp and GTVn masks, $P$ is the side length of image patch in each channel. Then each multichannel patch is flattened and linear projected to a feature representation of dimension $K \times C$ via the linear projector, where $K = \frac{H}{P} \times \frac{W}{P} \times \frac{L}{P}$ is the total number of patches ($H \times W \times L \times 7 \rightarrow \frac{H}{4} \times \frac{W}{4} \times \frac{L}{4} \times C$, we set $P$ to 4 following the typical setting in Swin ViT), and $C$ is the dimension of the projected vector of each patch. Following the linear projection, several consecutive modules of patch merging and Swin Transformer blocks are adopted. The Swin Transformer blocks output the same number of features as the input, while the patch merging layers merge the features of each group of $2 \times 2 \times 2$ neighbors, thus they reduce the number of tokens by a factor of $2 \times 2 \times 2 = 8$. Then a linear layer is applied to produce features each of $2C$-dimension ($\frac{H}{4} \times \frac{W}{4} \times \frac{L}{4} \times C \rightarrow \frac{H}{8} \times \frac{W}{8} \times \frac{L}{8} \times 2C$). After four layers of Swin Transformer blocks and three patch merging modules in between, the features are encoded to the dimension of $\frac{H}{32} \times \frac{W}{32} \times \frac{L}{32} \times 8C$. The decoder consists of multiple upsampling and convolutional layers with the kernel size of $3 \times 3$. Each of the upsampled feature maps in the decoding stage was concatenated with the corresponding feature map from the encoding path via skip connections, then two convolutional layers were applied. We also employed two convolutional layers using the original and downsampled image pair as inputs to capture local information and generate high-resolution feature maps. The outputs of these layers were concatenated with the feature maps in the decoder to produce a DVF. Leaky rectified linear unit is adopted following each convolution layer except for the last DVF generation convolutional layer. The last DVF generation layer is a linear convolution layer, which can produce a wider range of values without constraints imposed by activation functions. Finally, a spatial transform module takes the predicted DVF and CBCT01 data as input and generates the predicted CBCT21 Data, which comprises the predicted CBCT21 image, the predicted GTVp mask, and the GTVn mask. During model training, the spatial transformation module applies a differentiable nonlinear warp using the predicted DVF to the deform input images (CBCT01 data), while it uses trilinear interpolation to deform images (CBCT21) and nearest-neighbor interpolation to deform mask images (GTVp and GTVn) during inference.[40]

### 2.2.3 | Loss functions

The loss function we used for model training is combined with similarity loss $L_{similarity}$ and diffusion loss $L_{diffusion}$:

$$Loss = L_{similarity}([\mathcal{D}(I_{01}, \phi), I_{21}], [\mathcal{D}(P_{01}, \phi), P_{21}],$$
$$\times [\mathcal{D}(N_{01}, \phi), N_{21}]) + \lambda L_{diffusion}(\phi)$$

where $I_n$ is the CBCT image collected at fraction #$n$, $P_n$ and $N_n$ are the corresponding GTVp volume and GTVn volume, $\phi$ is the estimated DVF, $\mathcal{D}(x, y)$ denotes the deformed moving image $x$ with DVF $y$. The similarity loss consists of overall image similarity loss between predicted image and target image, and structural similarity loss between the predicted GTV and GTV of target image. We used SSIM loss as the similarity loss for CBCT images, and Dice coefficient loss as the structural similarity loss of GTVp and GTVn. We assigned a weighting factor of value 1.0 to each of the similarity loss empirically. DVF gradient was used as the diffusion loss in our study to quantize the smoothness of the estimated DVF. The regularization parameter $\lambda$ was chosen as 0.01 to balance the data fidelity of the image deformed by the DVF estimator and the smoothness of the estimated DVF, which helps achieve high-quality, physically plausible, and robust DVF estimation.

### 2.2.4 | Model performance evaluation

We used a testing dataset for model performance evaluation. To demonstrate the effectiveness of the proposed model for anatomy change prediction, we compared the similarity between the target image (CBCT21) and the predicted image to the similarity between CBCT21 and other images, including planning CT and CBCT01. Two other deep prediction models using different encoders were also trained and validated for comparison. One of them replaced our Swin Transformer encoder with CNN to mimic the structure of VoxelMorph, which is a widely used CNN-based image registration network.[41] The other model replaced the Swin Transformer block with a basic ViT block.[38] The comparison was conducted quantitatively and qualitatively.

For quantitative evaluation, consistent with the design of our loss function, we evaluated the accuracy of anatomy prediction in two aspects, overall image similarity and structural similarity to CBCT21. Mean square error (MSE), peak signal-to-noise ratio (PSNR), and structure similarity index[42] (SSIM) were applied to the whole 3D image to quantify the similarity of the overall image. Dice coefficient (Dice) and average symmetric surface distance[43] (ASD) were applied to the patient body mask, GTVp, and GTVn masks to quantify the structural similarity. We averaged the scores of all testing patients for comparison. The mean and standard deviation of each metrics were compared across CT, CBCT01, and predicted images.

For qualitative evaluation, we visually checked the similarity and image/structure deformation between CT, CBCT01, and predicted CBCT21 to real CBCT21. An accurate anatomy change prediction is expected to generate an image, which looks more similar to CBCT21 anatomically but not CT or CBCT01, and it could reflect where significant anatomy change occurs. We showed the CT, CBCT01, CBCT21, and predicted CBCT21 to demonstrate that. The different images between the body masks of different images to that of CBCT21 were also listed for inspection.

A set of ablation studies was done to evaluate the contribution of each input image modality. We removed CT, dose, and GTV masks from the input data in succession and retained our model accordingly. The mean and standard deviation of each metric for them were reported and compared. In addition, we changed the baseline image from CBCT01 to CT to investigate the impact of the choice of baseline image on the prediction results.

### 2.2.5 | Implementation details

Training and validation code of the proposed method were implemented using PyTorch on a workstation with two NVIDIA RTX3090 GPUs. We trained all the models for 100 epochs using the Adam optimization algorithm with the ReduceLROnPlateau strategy, the initial learning rate is set as 0.001 and the training batch size is 4. The input images were augmented with flipping, shifting, rotating, and added Gaussian noise. Swin Transformer patch size was set to [2, 4, 4], window size was set as [5, 5, 5], block numbers were set as [2, 2, 4, 2], head number as [4, 4, 8, 8], and embedding dimension as 96. For the other two comparison deep learning models, we followed the default settings of model hyperparameters in their published implementations.[38,41]

# 3 | RESULTS

## 3.1 | Patient data

One-hundred twenty-one eligible patients were included in the current study, patients were randomly stratified into training ($n$ = 101) and testing ($n$ = 20) sets for model training and evaluation. Twenty patients were randomly selected from the training cohort for validation. The demographic, disease characteristics, treatment protocol, and tumor volume-related information for both training and testing sets are summarized in Table 1, and there is no significant difference between the two sets in terms of the listed characteristics. Examples of their planning CT, CBCT01, and CBCT21 are provided in Figure S1. Distribution of patient BMI and change of BMI are also presented in the supplementary (Figure S2). To demonstrate the range and distribution of tumor volumes identified on different images, we showed the scatter plots of tumor volume and tumor volume change between pretreatment and Fraction #1 to Fraction #21 for each individual patient (Figures S3 and S4). Of note, during the review of imaging, the image quality of the GTVp and/or GTVn on some CBCTs was found to be degraded by strong artifacts. Therefore, patients with CBCT01 or CBCT21 scans affected were replaced with a separate repeat scan done prior to the same fraction, preferably, or a CBCT +/−1 fraction if a repeat scan was not available.

## 3.2 | Validation of TransAnaNet model

Table 2 summarizes the image and structure similarity between CBCT21 to planning CT, CBCT01, and predicted CBCT21 from different anatomy change prediction models. The mean value and standard deviation of each evaluation metric on the testing cohort were

**TABLE 1** Patient characteristics, treatment protocol, and tumor volume information.

| | | Train and validation | Test | |
|---|---|---|---|---|
| **Characteristics** | | **Number/Median [IQR]** | **Number/Median [IQR]** | ***p*-value** |
| Number | | 101 | 20 | — |
| Gender | Male | 84 | 16 | 0.75[†] |
| | Female | 17 | 4 | |
| Ethnicity | White | 65 | 11 | 0.48[†] |
| | African American | 17 | 3 | |
| | Hispanic | 7 | 2 | |
| | Asian | 1 | 1 | |
| | Other and unknown | 11 | 3 | |
| Age | | 60 [54–67] | 60 [52, 67] | 0.43[§] |
| Pre-RT BMI | | 28.6 [25.6–32.2] | 28.3 [25.4, 31.1] | 0.61[§] |
| Disease site | Oropharynx | 54 | 13 | 0.81[†] |
| | Larynx | 32 | 5 | |
| | Oral cavity | 2 | 0 | |
| | Other | 13 | 2 | |
| Disease laterality | Left | 28 | 7 | 0.82[†] |
| | Right | 31 | 4 | |
| | Central | 18 | 4 | |
| | Bilateral | 24 | 5 | |
| T stage | 1 | 12 | 5 | 0.62[†] |
| | 2 | 24 | 4 | |
| | 3 | 38 | 8 | |
| | 4 | 17 | 3 | |
| N stage | 0 | 16 | 4 | 0.91[†] |
| | 1 | 25 | 5 | |
| | 2 | 56 | 10 | |
| | 3 | 4 | 1 | |
| HPV-P16 status | Positive | 49 | 11 | 0.80[†] |
| | Negative | 18 | 4 | |
| | Unknown | 34 | 5 | |
| Smoking status | Never smoker | 29 | 7 | 0.64[†] |
| | Former smoker | 45 | 9 | |
| | Current smoker | 18 | 4 | |
| | Unknown | 9 | 0 | |
| Treatment paradigm | Chemoradiotherapy | 96 | 17 | 0.12[†] |
| | Radiotherapy | 5 | 3 | |
| RT modality | IMRT | 100 | 21 | — |
| Dose and fraction | 7000 cGy in 35 Fx | 57 | 13 | 0.62[†] |
| | 6996 cGy in 33 Fx | 44 | 7 | |
| Days between CT-Sim to RT | | 12 [7, 18] | 12 [8, 13] | 0.12[§] |
| Days between Fx#1 and Fx#21 | | 28 [28, 29] | 28 [28, 28] | 0.28[§] |
| GTVp volume at Fx#1 (cc) | | 20.2 [13.6, 36.0] | 19.4 [13.0, 31.6] | 0.60[§] |
| GTVp volume at Fx#21 (cc) | | 18.0 [12.6, 29.6] | 18.2 [12.6, 28.4] | 0.37[§] |
| GTVp volume change between Fx#1 and Fx#21 (cc) | | 1.6 [0, 4.8] | 0.2 [–0.4, 3.4] | 0.76[§] |
| GTVn volume at Fx#1 (cc) | | 14.2 [2.2, 38.0] | 7.6 [1.4, 29.4] | 0.60[§] |

(Continues)

**TABLE 1** (Continued)

| Characteristics | Train and validation<br>Number/Median [IQR] | Test<br>Number/Median [IQR] | *p*-value |
|---|---|---|---|
| GTVn volume at Fx#21 (cc) | 12.0 [1.8, 31.8] | 7.2 [1.2, 21.8] | 0.37§ |
| GTVn volume change between Fx#1 and Fx#21 (cc) | 1.2 [0, 6.4] | 0.8 [0, 6.0] | 0.62§ |
| GTV volume at Fx#1 | 40.4 [24.8, 83.0] | 33.9 [21.3, 52.6] | 0.95§ |
| GTV volume at Fx#21 | 33.0 [21.2, 68.9] | 30.1 [20.2, 41.9] | 0.89§ |
| GTV volume change between Fx#1 and Fx#21 (cc) | 4.1 [1.0, 11.4] | 3.7 [0.4, 10.5] | 0.81§ |
| GTVp center of mass change from Fx#1 to Fx#21 (mm) | 3.5 [2.2, 4.7] | 3.5 [2.0, 4.9] | 0.67§ |
| GTVn center of mass change from Fx#1 to Fx#21 (mm) | 3.3 [2.4, 4.6] | 3.0 [1.6, 3.9] | 0.17§ |

Abbreviation: IQR, interquartile range.
All statistical analyses were performed in Matlab with significance defined as a *p*-value < 0.05 on a two-sided test, with either the Fisher's exact test† or an unpaired *t*-test§.

**TABLE 2** Image and structure similarity between CBCT21 to planning CT, CBCT01, and predicted CBCT21 from different anatomy change prediction models.

**(a)**

| Image/Model | Image similarity<br>MSE | PSNR | SSIM |
|---|---|---|---|
| Planning CT | 0.016 ± 0.007 | 18.223 ± 1.938 | 0.877 ± 0.048 |
| CBCT01 | 0.013 ± 0.004 | 19.021 ± 1.492 | 0.919 ± 0.028 |
| TransAnaNet | **0.009 ± 0.003** | **20.266 ± 1.410** | **0.933 ± 0.020** |
| CNN | **0.009 ± 0.003** | 20.164 ± 1.402 | 0.932 ± 0.021 |
| Original ViT[38] | 0.010 ± 0.003 | 20.126 ± 1.376 | 0.930 ± 0.020 |

**(b)**

| | Structure similarity | | | | | |
|---|---|---|---|---|---|---|
| | Body | | GTVp | | GTVn | |
| Image/Model | Dice | ASD | Dice | ASD | Dice | ASD |
| Planning CT | 0.948 ± 0.019 | 2.914 ± 1.062 | 0.731 ± 0.083 | 1.480 ± 0.390 | 0.760 ± 0.164 | 1.194 ± 0.880 |
| CBCT01 | 0.961 ± 0.014 | 2.210 ± 0.684 | 0.741 ± 0.099 | 1.462 ± 0.562 | 0.786 ± 0.159 | 1.084 ± 0.836 |
| TransAnaNet | **0.972 ± 0.008** | **1.910 ± 0.318** | **0.792 ± 0.090** | **1.338 ± 0.398** | **0.821 ± 0.130** | **0.922 ± 0.720** |
| CNN | 0.968 ± 0.009 | 1.946 ± 0.328 | 0.778 ± 0.088 | 1.402 ± 0.454 | 0.814 ± 0.134 | 0.984 ± 0.768 |
| Original ViT[38] | 0.969 ± 0.008 | 1.924 ± 0.294 | 0.781 ± 0.094 | 1.338 ± 0.474 | 0.815 ± 0.143 | 0.974 ± 0.782 |

(a) The image similarity is quantified with mean square error (MSE), peak signal-to-noise ratio (PSNR), and structure similarity index (SSIM). (b) The structure similarity is quantified with Dice coefficient (Dice) and average symmetric surface distance (ASD, unit: mm) for binary masks of patients' head and neck regions (body), primary tumors (GTVp), and involved lymph nodes (GTVn). Anatomy change prediction models built with Swin Transformer-based encoder (TransAnaNet), convolutional neural network-based encoder (CNN), and basic vision-transformer-based encoder (ViT) are listed for comparison. The best performance for each metric is bolded.

reported. Example images and GTV contours of three sample testing patients (patients with the smallest, median, and the largest volumetric change between planning CT and CBCT21) are shown in Figure 2. More of the predicted images for testing samples are provided in Figure S7. Difference images between body masks and GTV masks of each image to those of CBCT21 were also measured to show the anatomy change. Volume and volume prediction errors of GTVs for each testing sample are presented via scatter plots in Figure 3. Distribution of the evaluation metrics on testing cohort are summarized in Figure S6. The center of mass shift distribution between CBCT01 and predicted image to CBCT21 is summarized in Figure S5.

According to Table 2, the predicted CBCT21 images from the proposed TransAnaNet have the best overall image similarity to real CBCT21 scans in terms of MSE (0.009 ± 0.003), PSNR (20.266 ± 1.410), and SSIM (0.933 ± 0.020). For structure similarity, the results from the proposed TransAnaNet still outperform other scans or predicted images in body contour (Dice = 0.972 ± 0.008, ASD = 1.910 ± 0.318), GTVp contour (Dice = 0.792 ± 0.090, ASD = 1.338 ± 0.398), and GTVn contour (Dice = 0.821 ± 0.130,

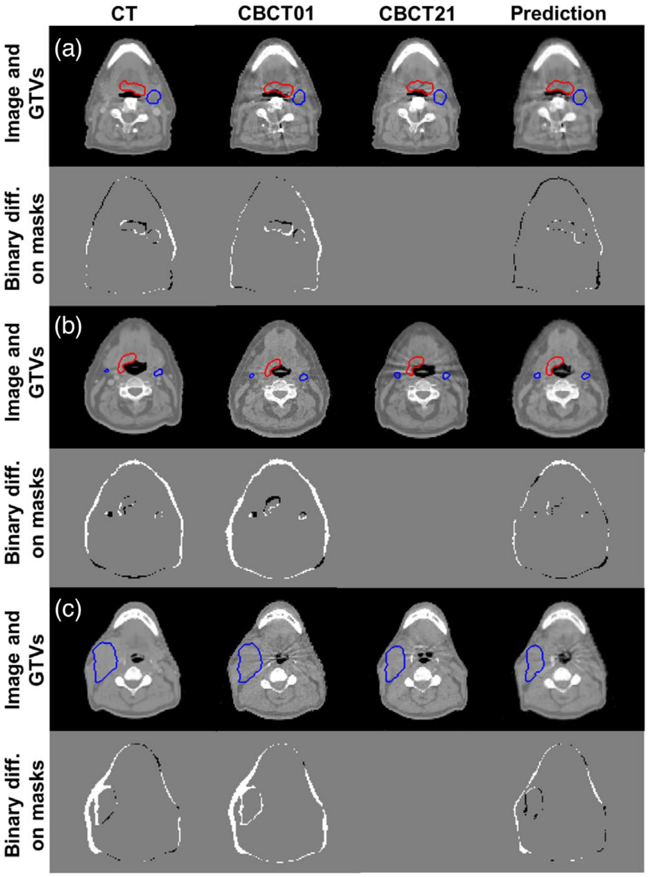

**FIGURE 2** Qualitative performance evaluation of the proposed method for anatomy change prediction (four example patients out of the 21 testing cohort patients). Testing patients with (a) the smallest, (b) median, and (c) the largest volumetric change between planning CT and CBCT21 are selected for comparison. CT, CBCT01, CBCT21, and predicted CBCT21 images are listed for comparison, original and predicted GTVp and GTVn are delineated in red and blue, respectively. Difference images between the body masks, GTVp and GTVn masks of planning CT, CBCT01, and predicted CBCT21 to those from the real CBCT21 are shown in the second row for each patient.

ASD = 0.922 ± 0.720). The predicted images from CNN have the same performance on MSE image similarity as TransAnaNet, but lower performance on other metrics.

Different degrees of anatomy change can be identified from Figure 2 when comparing CT/CBCT01 to CBCT21 or checking the difference of body mask images between CBCT21 and CT/CBCT01. Compared with CT, the predicted CBCT21 images from the proposed TansAnaNet have similar image visual quality to CBCT21 as they were deformed from CBCT01, which were collected with the same machine and protocol as CBCT21. Compared with both CT and CBCT01, the predicted CBCT21 images have less body volume prediction errors than CBCT21. A similar observation can be made from the GTV volume and volume prediction errors scatter plot (Figure 3), the predicted CBCT21 images via TransAnaNet have volumes similar to CBCT21 and demonstrate less volume prediction errors compared to CT, CBCT01, and predicted CBCT21 images via CNN and Original ViT. The narrower GTV center of mass shift distribution shown in Figure S5 also demonstrates the better geometrical similarity of the predicted image to real CBCT21.

## 3.3 | Ablation study

Table 3 compares the performance of anatomy change prediction using different baseline images and a combination of image modalities as input for the proposed TransAnaNet model. The result shows that all of the collected image modalities have a positive contribution to the anatomy change prediction using the proposed model. Missing either the CT image, GTV mask, planned dose, or CBCT01 as input data will lead to degraded performance. When adopting CBCT01 as the baseline image for deformation and using the combination of 3D dose map, planning CT, CBCT01, and GTV contours on CT and CBCT01 as model input, the proposed network has the best performance on overall image similarity (MSE: 0.009 ± 0.003, PSNR 20.266 ± 1.410, SSIM: 0.933 ± 0.020), and it has the superior head and neck body mask structure similarity to CBCT21.

The best GTV structure similarity performance was achieved when using the same input image to the model but adopting planning CT as the baseline image for deformation. The resulting GTVp similarity performance is 0.807 ± 0.084 and 1.296 ± 0.364 for Dice coefficient and ASD, respectively. The resulting GTVn similarity performance is 0.829 ± 0.128 and 0.904 ± 0.688 for Dice coefficient and ASD, respectively. Some testing samples generated following this input modality combination and baseline image are shown in Figure 4. The generated CBCT21 images have similar widths of body contours as CBCT21, and some of the volumetric changes of substructures can be clearly identified from the images (pointed out by red arrows in Figure 4).

# 4 | DISCUSSION

During HNC IMRT courses, patients may undergo substantial anatomical changes, resulting in a divergence of the actual administered dose from the initial treatment plan during the later stages of therapy, which might lead to unexpected side effects or loss of tumor control. In this exploratory study, we proposed the first deep-learning model prognosticate patient anatomy change in HNC RT, thereby facilitating the early detection of patients susceptible to significant anatomical changes. We demonstrated that a Swin-Transformer-based ViT model (TransAnaNet) with planned dose, planning CT, early fraction CBCT, and GTV masks as input possessed high predictive ability for patient anatomy change in

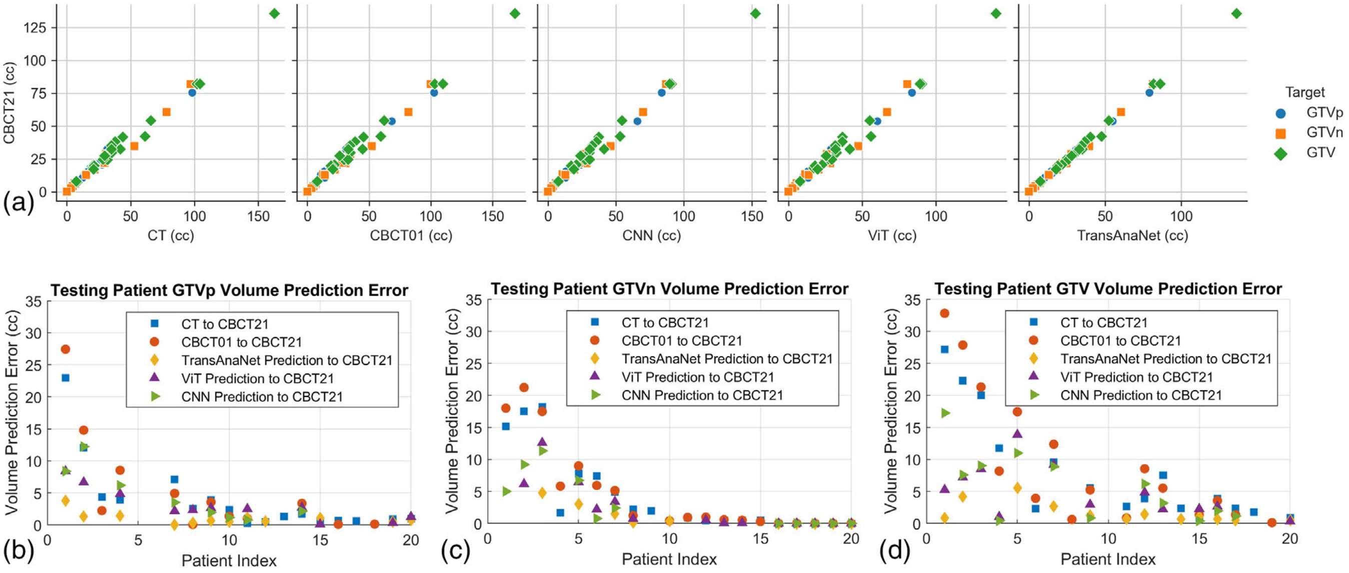

**FIGURE 3** Distribution of volumes and volume prediction errors of gross tumor volumes (GTVs) of testing data patients. Volumes are measured on CT, CBCT01, CBCT21, and predicted CBCT21 via TransAna, Original ViT, and CNN, respectively. Volume difference and prediction error between CT, CBCT01, and predicted CBCT21 to CBCT21 are measured. Positive value means volume decrease. Distribution of volumes of GTVs measured on CBCT21 versus CT, CBCT01, CNN predicted CBCT21, original ViT predicted CBCT21, and TransAnaNet predicted CBCT21 are shown in (a). Primary tumor volume (GTVp) prediction errors are shown in (b), patients are ranked according to GTVp volume measured on CBCT01. Volume prediction errors of involved lymph nodes (GTVn) are shown in (c), patients are ranked according to GTVn volume measured on CBCT01. Volume prediction errors of the total GTVs are shown in (d), patients are ranked according to GTV volume measured on CBCT01.

**TABLE 3** Performance comparison of using different baseline image and combination of image modalities as input for the proposed anatomy prediction model.

**(a)**

| | Image similarity | | |
|---|---|---|---|
| **Image modality** | **MSE** | **PSNR** | **SSIM** |
| CBCT01 + GTV + Dose (baseline: CBCT01) | 0.011 ± 0.004 | 19.576 ± 1.473 | 0.925 ± 0.027 |
| CT + CBCT01 + GTV (baseline: CBCT01) | 0.010 ± 0.003 | 19.829 ± 1.415 | 0.929 ± 0.022 |
| CT + CBCT01 + Dose (baseline: CBCT01) | 0.010 ± 0.003 | 20.001 ± 1.411 | 0.929 ± 0.022 |
| CT + CBCT01 + GTV + Dose (baseline: CT) | 0.014 ± 0.004 | 19.164 ± 1.961 | 0.902 ± 0.035 |
| CT + CBCT01 + GTV + Dose (baseline: CBCT01) | **0.009** ± **0.003** | **20.266** ± **1.410** | **0.933** ± **0.020** |

**(b)**

| | Structure similarity | | | | | |
|---|---|---|---|---|---|---|
| | **Body** | | **GTVp** | | **GTVn** | |
| **Image/Model** | **Dice** | **ASD** | **Dice** | **ASD** | **Dice** | **ASD** |
| CBCT01 + GTV + Dose (baseline: CBCT01) | 0.960 ± 0.017 | 2.187 ± 0.697 | 0.749 ± 0.091 | 1.436 ± 0.404 | 0.787 ± 0.160 | 1.050 ± 0.802 |
| CT + CBCT01 + GTV (baseline: CBCT01) | 0.971 ± 0.010 | 2.004 ± 0.357 | 0.775 ± 0.087 | 1.422 ± 0.430 | 0.805 ± 0.145 | 0.992 ± 0.779 |
| CT + CBCT01 + Dose (baseline: CBCT01) | 0.970 ± 0.010 | 2.024 ± 0.346 | 0.744 ± 0.091 | 1.486 ± 0.512 | 0.782 ± 0.155 | 1.092 ± 0.816 |
| CT + CBCT01 + GTV + Dose (baseline: CT) | 0.971 ± 0.009 | 1.954 ± 0.342 | **0.807** ± **0.084** | **1.296** ± **0.364** | **0.829** ± **0.128** | **0.904** ± **0.688** |
| CT + CBCT01 + GTV + Dose (baseline: CBCT01) | **0.972** ± **0.008** | **1.910** ± **0.318** | 0.792 ± 0.090 | 1.338 ± 0.398 | 0.821 ± 0.130 | 0.922 ± 0.720 |

(a) The image similarity is quantified with mean square error (MSE), peak signal-to-noise ratio (PSNR), and structure similarity index (SSIM). (b) The structure similarity is quantified with Dice coefficient (Dice) and average symmetric surface distance (ASD, unit: mm) for binary masks of patients' head and neck regions (body), primary tumors (GTVp), and involved lymph nodes (GTVn). The best performance for each metric is bolded.

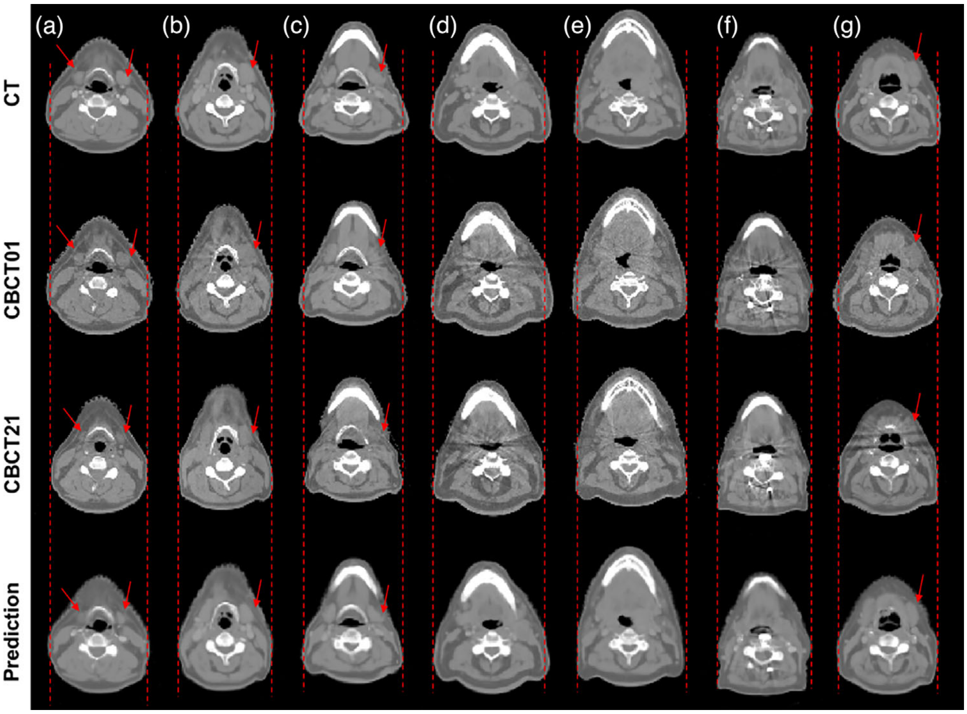

**FIGURE 4** Qualitative performance evaluation of the proposed method for anatomy change prediction with the planning CT image as the baseline image. Planning CT, CBCT01, CBCT21, and predicted CBCT21 from the proposed method for seven different testing patients (columns a–g) are listed as examples. The dash lines aside of each set of images show the width of the corresponding CBCT21 body contours, the red arrows point to the regions of obvious volumetric changes.

later treatment fraction. The proposed method shows great potential in assisting clinicians to identify HNC patients who are likely to derive greater benefit from ART. Dosimetric evaluations can be performed based on the predicted images and previous treatment plan parameters. If a significant decrease in target coverage or overdose of OARs is identified, ART can be recommended timely. The threshold for target coverage or OAR dose variation can be determined per institutional protocol. The code to implement our method is publicly available on GitHub (link will be released after acceptance).

We propose to use initial CBCT (CBCT01) as the baseline image and predict the DVF between initial CBCT and late fraction CBCT (CBCT21). Subsequently, CBCT01 is deformed according to the predicted DVF to generate the predicted CBCT21. Comparative analysis with the planning CT and CBCT01 indicates that the TransAnaNet predicted CBCT21 more closely approximates the actual CBCT21 in terms of overall image similarity and structural similarity (Table 2 and Figures 2 and 3). The MSE, PSNR, and SSIM between the predicted CBCT21 to real CBCT21 are 0.009, 20.266, and 0.933, respectively. The Dice coefficient and ASD (unit: mm) for the body mask, GTVp mask, and GTVn mask are 0.972/1.910, 0.792/1.338, and 0.821/0.922, respectively. We also compared the performance derived from other deep learning models, and the proposed Swin-Transformer-based model has the best performance for all metrics. In Figure 2, we used the predicted images, patient body contours, and GTVs' contours to assess the anatomy change prediction ability of our method qualitatively. We utilized the binarized difference maps between masks on CT, CBCT01, and predicted CBCT21 to actual CBCT21 to present the change or prediction errors of boundaries of the target or organ of interest. We did not use the pixel-to-pixel image difference in our study, as it is sensitive to imaging modality, image registration, and the noise and artifacts within each scan. As we deform CBCT01 to generate the predicted CBCT21, any imperfectly aligned artifacts will be amplified in the difference imaging. As such, comparing the difference map between predicted CBCT21 and real CBCT21 may not be a good way to demonstrate the effectiveness of the presented method. Additionally, the CT numbers in CBCT differ significantly from those in planning CT. Directly subtracting planning CT from CBCT makes it even more difficult to distinguish differences caused by anatomical changes.

A set of ablation studies was done to evaluate the contributory significance of each input image modality (Table 3). The result shows that every image modality employed enhanced the anatomy change prediction accuracy of the proposed model. Omitting any single modality results in diminished performance. While the overall image similarity and body mask similarity are not sensitive to input modalities, the GTVp and GTVn mask similarity are sensitive to input image data. Notably, the most significant decrease in performance is observed when the input lacks the preceding GTV mask. In this case, the predicted DVF only deforms the boundary of the patient body, without predicting the anatomy change of internal substructures. This outcome underscores the essentiality of incorporating substructure similarity within the loss function to refine model optimization.

Of note, due to the heavy workload of contouring OARs on CBCTs, in current study, we constructed our loss function and evaluated our model based on the head and neck 3D image and GTV masks. However, existing literature suggests that anatomical changes result in greater dose variation in OARs as opposed to target volumes.[4,6,12] The dose coverage of the GTV is typically more resilient to these changes due to the incorporation of the PTV concept. On the other hand, the planning volumes at risk (PRV) margins are only commonly employed for the spinal cord and brain stem, not for most of the other main OARs in HNC RT, such as the parotid glands, which are consistently reported to shrink and/or appear orientation shift during treatment.[3,4,44,45] Therefore, the incorporation of OAR contours and the refinement of our model to include these structures could enhance the efficacy of the proposed methodology, which is worthy of exploring in a future study.

In our experiment, we also validate the performance of the proposed model when planning CT is used as the baseline image (Table 3 and Figure 4), where the model predicts the DVF between planning CT and late fraction CBCT. It turns out that using planning CT as the baseline image achieved the best subprediction accuracy. The resulting Dice coefficient and ASD (unit: mm) are 0.807/1.296 and 0.829/0.904 for GTVp and GTVn, respectively. Concurrently, using CBCT01 as the baseline has better performance in terms of overall image and body mask similarity. The presence of artifacts in CBCT and the more pronounced soft tissue contrast in CT may account for the enhanced performance in substructure anatomical prediction when CT is utilized as the baseline image. Additionally, the consistency of imaging protocols could explain the heightened overall image similarity when CBCT01 was used as the baseline. Selection of baseline image merits additional investigation.

Previous studies have introduced various machine learning and radiomics techniques for predicting ART eligibility of HNC patients with the hypothesis that features from multiple modalities contain predictive biomarkers for tumor and OAR shrinkage following cancer treatment.[13–16] Brown et al. evaluated the predictive ability of patient demographics and tumor characteristics for predicting the need for HNC ART via logistic regression analysis, they identified that nodal disease stage, pretreatment node size, diagnosis, and initial weight are the significant factors for ART inclusion determination.[17] Without using patient electronic health record (EHR) data, Yu et al. proposed a pretreatment MR image feature-based HNC ART eligibility model, a T1–T2 joint radiomics feature set was selected via the least absolute shrinkage and selection operator (LASSO) logistic regression method, and the proposed model achieved an area under receiver operating characteristic curve (AUROC) of 0.85 for binary prediction.[14] Lam et al. proposed a pretreatment CT image feature-based model to identify nasopharyngeal carcinoma (NPC) patients who will experience ill-fitted thermoplastic mask (IfTM) event during RT, which may trigger ART to ensure treatment safety.[16] The model was evaluated via a multicenter setting, and the performance outperformed a clinical feature-based model and a clinic-radiomics fusion model in terms of AUROC. However, despite their high accuracy in binary prediction, these methods face challenges in quantifying the anatomical changes of tumors and/or OARs. The absence of a model prediction visual explanation might limit their reliability and reduce the confidence of their implementation in clinic.[46–48] In contrast, we aimed to build an anatomy change prediction model that does not directly predict the necessity for ART but provides the possible anatomy change of the treatment volume through image data, thereby assisting the decision-making for ART.

The proposed method facilitates the visualization of potential anatomical changes, which is a pivotal advantage over previous methods. Through our method, the predicted patient anatomy change can be visualized via the predicted DVFs, the predicted images, or the structure mask difference maps (Figures 2 and 3). Independent of whether CBCT01 or CT serves as the baseline image, the predicted image can distinctly delineate the location and extent of anatomy changes. This demonstrated the effectiveness of the proposed anatomy change prediction method qualitatively. However, our study did not provide the fundamental understanding of how anatomical changes can be predicted only from the initial CT/CBCT and the dose distribution. To the best of our knowledge, there is no existing research on this topic, and exploring the underlying mechanisms is beyond the scope of this study. Nevertheless, when combined with dose distribution, the changes between pCT and initial CBCT could inform the features in the latent space learned by ViT, potentially predicting future changes in patients.

The application of the proposed model extends beyond predicting later fraction CBCT images. Its primary function is to identify optimal candidates for ART by evaluating the magnitude of anatomical changes. In addition, with an accurate prediction of patient anatomy change, the current treatment plan can be dosimetrically evaluated on the predicted patient anatomy, thereby informing physicians’ decisions regarding the necessity for resimulation or replanning to align with clinical objectives. On the other hand, for online ART, the predicted image could potentially streamline the preparation of a new plan for the next treatment, which might reduce the duration of patient wait times on the treatment couch and enhance the efficiency of the clinical workflow. Besides, the predicted image can be used to evaluate the robustness of treatment plans, and their sensitivity to potential anatomic changes can be evaluated via dose calculation using the predicted images and previous treatment plans.

As a proof-of-concept study, our work has several limitations. First, the anatomy changes were predicted at a fixed fraction (Fraction 21), which might not be the optimal time point decision-making or replanning. Constructing a longitudinal prediction model using recurrent neural network to predict the anatomy change in time series will be one of the main directions of our future work. Additionally, the model's construction and validation were based on a limited patient cohort, and there is currently no independent data from an external institution or collected prospectively for evaluation. Conducting a prospective study for model evaluation is envisaged for our future work, and the code for implementation of the proposed method is publicly available for external use. Third, the target delineation quality, interobserver variability, and their potential influence were not evaluated. Moreover, this study focused on predicting anatomy changes without directly comparing our method to existing binary predictors for ART eligibility. As ART eligibility criteria could be physician and institution-dependent, establishing an optimal threshold for ART eligibility selection based on predicted anatomical changes and performing a comparison study to related work warrants further investigation. Finally, although anatomy change prediction offers physicians more detailed information to assist in the decision-making of replanning or ART than binary prediction methods, the prediction uncertainty of the proposed TransAnaNet has yet to be scrutinized. Our future research efforts may also involve exploring the model's interpretability and prediction uncertainty estimation to achieve more reliable predictions, which can aid in identifying potential biases and increase trust and understanding of the deep model.

The performance of the proposed method might be further improved in several ways. In the current study, we utilized a variety of image modalities as model input to predict patient anatomy change, regardless of the available patient demographic and disease characteristics data (Table 1). In addition, other information such as patient diet and chemotherapy can be potential factors affecting patients' anatomy change. Integrating both image data and patients' EHR data with a multimodal network might lead to a better performance. Recently, the advent of MR-guided RT in various institutions has significantly enhanced the visibility of soft tissue changes during treatment. The integration of MR data collected during RT could facilitate the development of a more precise and inclusive model for predicting anatomical changes, thereby aiding the decision-making process for ART.

## 5 | CONCLUSION

We constructed a transformer-based model (TransAnaNet) using a planning dose map, initial treatment radiological image, and targets' structures from planning CT and initial CBCT to predict the anatomy change of the treated region in later fractions. The proposed TransAnaNet has demonstrated promising capability in predicting the patient anatomy on later CBCT in the form of a 3D image, which could be valuable to assist in the optimization of the workflow and the use of clinic resources related to ART. Future work is needed to incorporate OARs that are susceptible to notable anatomy change and of high radiosensitivity into our model, construct longitudinal prediction model, and validate the proposed methods in a prospective manner.

## ACKNOWLEDGMENTS

We acknowledge the support from NIH (R01 CA251792).

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

Research data are not readily available because of institution regulations. Requests to access the datasets should be directed to the corresponding author.

## REFERENCES

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin*. 2023;73(1):17-48.
2. Chow LQ. Head and neck cancer. *N Engl J Med*. 2020;382(1):60-72.
3. Morgan HE, Sher DJ. Adaptive radiotherapy for head and neck cancer. *Cancers Head Neck*. 2020;5(1):1-16.
4. Brouwer CL, Steenbakkers RJ, Langendijk JA, Sijtsema NM. Identifying patients who may benefit from adaptive radiotherapy: does the literature on anatomic and dosimetric changes in head and neck organs at risk during radiotherapy provide information to help? *Radiother Oncol*. 2015;115(3):285-294.
5. Alves N, Dias JM, Rocha H, et al. Assessing the need for adaptive radiotherapy in head and neck cancer patients using an automatic planning tool. *Rep Pract Oncol Radiother*. 2021;26(3):423-432.
6. Hansen EK, Bucci MK, Quivey JM, Weinberg V, Xia P. Repeat CT imaging and replanning during the course of IMRT for head-and-neck cancer. *Int J Radiat Oncol* Biol* Phys*. 2006;64(2):355-362.
7. Belshaw L, Agnew CE, Irvine DM, Rooney KP, McGarry CK. Adaptive radiotherapy for head and neck cancer reduces the requirement for rescans during treatment due to spinal cord dose. *Radiat Oncol*. 2019;14(1):1-7.
8. Brouwer CL, Steenbakkers RJ, van der Schaaf A, et al. Selection of head and neck cancer patients for adaptive radiotherapy to decrease xerostomia. *Radiother Oncol*. 2016;120(1):36-40.

9. Shen C, Chen L, Zhong X, et al. Clinical experience on patient-specific quality assurance for CBCT-based online adaptive treatment plan. *J Appl Clin Med Phys*. 2023;24(4):e13918.
10. Lim-Reinders S, Keller BM, Al-Ward S, Sahgal A, Kim A. Online adaptive radiation therapy. *Int J Radiat Oncol* Biol* Phys*. 2017;99(4):994-1003.
11. Noble DJ, Yeap P-L, Seah SY, et al. Anatomical change during radiotherapy for head and neck cancer, and its effect on delivered dose to the spinal cord. *Radiother Oncol*. 2019;130:32-38.
12. Wu Q, Chi Y, Chen PY, Krauss DJ, Yan D, Martinez A. Adaptive replanning strategies accounting for shrinkage in head and neck IMRT. *Int J Radiat Oncol* Biol* Phys*. 2009;75(3):924-932.
13. Alves N, Dias J, Ventura T, et al. Predicting the need for adaptive radiotherapy in head and neck patients from CT-based radiomics and pre-treatment data. In: *Computational Science and Its Applications – ICCSA 2021*. Springer; 2021:429-444.
14. T-T Yu, S-K Lam, L-H To, et al. Pretreatment prediction of adaptive radiation therapy eligibility using MRI-based radiomics for advanced nasopharyngeal carcinoma patients. *Front Oncol*. 2019;9:1050.
15. Lam S-K, Zhang Y, Zhang J, et al. Multi-organ omics-based prediction for adaptive radiation therapy eligibility in nasopharyngeal carcinoma patients undergoing concurrent chemoradiotherapy. *Front Oncol*. 2022;11:792024.
16. Lam S-K, Zhang J, Zhang Y-P, et al. A multi-center study of CT-based neck nodal radiomics for predicting an adaptive radiotherapy trigger of ill-fitted thermoplastic masks in patients with nasopharyngeal carcinoma. *Life*. 2022;12(2):241.
17. Brown E, Owen R, Harden F, et al. Predicting the need for adaptive radiotherapy in head and neck cancer. *Radiother Oncol*. 2015;116(1):57-63.
18. Chen J, Lu Y, Yu Q, et al. TransUNet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:210204306*. 2021.
19. Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021:6881-6890.
20. Chen J, Du Y, He Y, Segars WP, Li Y, Frey EC. TransMorph: transformer for unsupervised medical image registration. *Medical Image Analysis*. 2022;82:102615.
21. Jiang Y, Chang S, Wang Z. TransGAN: two transformers can make one strong GAN. *Advances in Neural Information Processing System*. 2021;34:14745-14758.
22. Choi H, Bajić IV. Affine transformation-based deep frame prediction. *IEEE Trans Image Process*. 2021;30:3321-3334.
23. Wang S, Saharia C, Montgomery C, et al. Imagen editor and editbench: advancing and evaluating text-guided image inpainting. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023:18359-18369.
24. Frolov S, Hinz T, Raue F, Hees J, Dengel A. Adversarial text-to-image synthesis: a review. *Neural Netw*. 2021;144:187-209.
25. Gui J, Sun Z, Wen Y, Tao D, Ye J. A review on generative adversarial networks: algorithms, theory, and applications. *IEEE Trans Knowl Data Eng*. 2021;35(4):3313-3332.
26. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: a survey. *ACM Comput Surv (CSUR)*. 2022;54(10s):1-41.
27. Zhang L, Lu L, Wang X, et al. Spatio-temporal convolutional LSTMs for tumor growth prediction by learning 4D longitudinal patient data. *IEEE Trans Med Imaging*. 2019;39(4):1114-1126.
28. Elazab A, Wang C, Gardezi SJS, et al. GP-GAN: brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR Images. *Neural Netw*. 2020;132:321-332.
29. Li R, Roy A, Bice N, Kirby N, Fakhreddine M, Papanikolaou N. Managing tumor changes during radiotherapy using a deep learning model. *Med Phys*. 2021;48(9):5152-5164.
30. Ebadi N, Li R, Das A, Roy A, Nikos P, Najafirad P. CBCT-guided adaptive radiotherapy using self-supervised sequential domain adaptation with uncertainty estimation. *Med Image Anal*. 2023;86:102800.
31. Gan Y, Langendijk JA, van der Schaaf A, et al. An efficient strategy to select head and neck cancer patients for adaptive radiotherapy. *Radiother Oncol*. 2023;186:109763.
32. Gan Y, Langendijk JA, Oldehinkel E, Lin Z, Both S, Brouwer CL. Optimal timing of re-planning for head and neck adaptive radiotherapy. *Radiother Oncol*. 2024;194:110145.
33. Yan D, Yan S, Wang Q, Liao X, Lu Z, Wang Y. Predictors for replanning in loco-regionally advanced nasopharyngeal carcinoma patients undergoing intensity-modulated radiation therapy: a prospective observational study. *BMC Cancer*. 2013;13:1-9.
34. Figen M, Çolpan Öksüz D, Duman E, et al. Radiotherapy for head and neck cancer: evaluation of triggered adaptive replanning in routine practice. *Front Oncol*. 2020;10:579917.
35. Jia X, Bartlett J, Zhang T, Lu W, Qiu Z, Duan J. U-net vs transformer: Is U-Net outdated in medical image registration? In: *Machine Learning in Medical Imaging*. Springer; 2022:151-160.
36. Shi J, He Y, Kong Y, et al. XMorpher: full transformer for deformable medical image registration via cross attention. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer; 2022:217-226.
37. Chen J, Frey EC, He Y, Segars WP, Li Y, Du Y. Transmorph: transformer for unsupervised medical image registration. *Med Image Anal*. 2022;82:102615.
38. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. *arXiv preprint arXiv:201011929*. 2020.
39. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*. 2021:10012-10022.
40. Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. In: Advances in Neural Information Processing Systems. Vol. 28. 2015:1-9.
41. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans Med Imaging*. 2019;38(8):1788-1800.
42. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600-612.
43. Heimann T, Van Ginneken B, Styner MA, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans Med Imaging*. 2009;28(8):1251-1265.
44. Castelli J, Simon A, Louvel G, et al. Impact of head and neck cancer adaptive radiotherapy to spare the parotid glands and decrease the risk of xerostomia. *Radiat Oncol*. 2015;10:1-10.
45. Bhide SA, Davies M, Burke K, et al. Weekly volume and dosimetric changes during chemoradiotherapy with intensity-modulated radiation therapy for head and neck cancer: a prospective observational study. *Int J Radiat Oncol* Biol* Phys*. 2010;76(5):1360-1368.
46. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762. doi:10.1038/nrclinonc.2017.141
47. Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *Eur J Nucl Med Mol Imaging*. 2019;46(13):2638-2655.

48. Yip SS, Aerts HJ. Applications and limitations of radiomics. *Phys Med Biol*. 2016;61(13):R150-R166.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Chen M, Wang K, Dohopolski M, Morgan H, Sher D, Wang J. TransAnaNet: Transformer-based anatomy change prediction network for head and neck cancer radiotherapy. *Med Phys*. 2025;52:3015–3029. https://doi.org/10.1002/mp.17655