Improving Neoadjuvant Therapy Response Prediction by Integrating Longitudinal Mammogram Generation with Cross-Modal Radiological Reports: A Vision-Language Alignment-guided Model

Yuan Gao[1,2], Hong-Yu Zhou[3], Xin Wang[1,2], Tianyu Zhang[2,4], Luyi Han[2,4], Chunyao Lu[2,4], Xinglong Liang[2,4], Jonas Teuwen[4,5] Regina Beets-Tan[1,2], Tao Tan[6,1*], and Ritse Mann[2,4]

[1] GROW School for Oncology and Development Biology, Maastricht University, 6200 MD, Maastricht, The Netherlands
[2] Department of Radiology, Netherlands Cancer Institute (NKI), 1066 CX, Amsterdam, The Netherlands
[3] Department of Biomedical Informatics, Harvard Medical School, MA 02115, Boston, USA
[4] Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, 6525 GA, Nijmegen, The Netherlands
[5] Department of Radiation Oncology, Netherlands Cancer Institute (NKI), 1066 CX, Amsterdam, The Netherlands
[6] Faculty of Applied Sciences, Macao Polytechnic University, 999078, Macao, China
Corresponding author: taotanjs@gmail.com

Abstract. Longitudinal imaging examinations are vital for predicting pathological complete response (pCR) to neoadjuvant therapy (NAT) by assessing changes in tumor size and density. However, quite-often the imaging modalities at different time points during NAT may differ from patients, hindering comprehensive treatment response estimation when utilizing multi-modal information. This may result in underestimation or overestimation of disease status. Also, existing longitudinal image generation models mainly rely on raw-pixel inputs while less exploring in the integration with practical longitudinal radiology reports, which can convey valuable temporal content on disease remission or progression. Further, extracting textual-aligned dynamic information from longitudinal images poses a challenge. To address these issues, we propose a longitudinal image-report alignment-guided model for longitudinal mammogram generation using cross-modality radiology reports. We utilize generated mammograms to compensate for absent mammograms in our pCR prediction pipeline. Our experimental result achieves comparable performance to the theoretical upper bound, therefore providing a potential 3-month window for therapeutic replacement. The code will be accessible to the public.

Keywords: pCR prediction · Longitudinal mammogram generation · Multi-modal data · Radiology report

1 Introduction

Neoadjuvant therapy (NAT) has become the common care for breast cancer patients with the goal of reducing the volume of the tumor and clinical stage, so that patients have more opportunities for breast preservation and less extensive surgery [23]. An early prediction of the pathological complete response (pCR) to NAT may facilitate the tailoring therapy for breast cancer patients, thus leading to an increased likelihood of achieving pCR [24, 22, 25]. However, the standard method of evaluating the tumor response to NAT heavily relies on the post-operative specimens collected at the end of treatment (as shown in Fig. 1), which leaves little room to adjust the NAT plan. Against this background, imaging-based technologies (e.g., digital mammography, magnetic resonance imaging (MRI)) have become a promising direction for estimating the patients' response to NAT [8, 14]. These are non-invasive and therefore flexible enough for dynamic monitoring. Nonetheless, assessment of treatment response remains difficult, as the different imaging modalities may result in underestimation or overestimation of disease status [13].

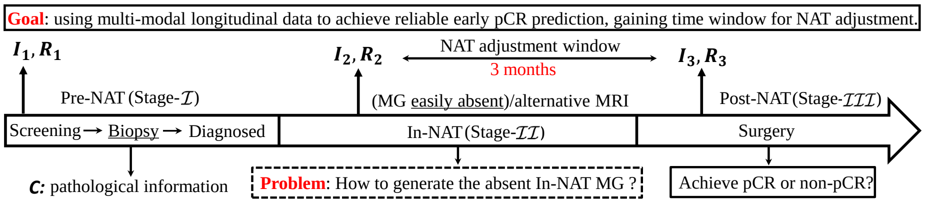
Goal: using multi-modal longitudinal data to achieve reliable early pCR prediction, gaining time window for NAT adjustment.
I_1, R_1
I_2, R_2
NAT adjustment window
3 months
I_3, R_3
Pre-NAT (Stage-$\mathcal{I}$)
(MG easily absent)/alternative MRI
Post-NAT (Stage-$\mathcal{III}$)
Screening → Biopsy → Diagnosed
In-NAT (Stage-$\mathcal{II}$)
Surgery
C: pathological information
Problem: How to generate the absent In-NAT MG ?
Achieve pCR or non-pCR?

Fig. 1. Clinical background. The general neoadjuvant therapy (NAT) pipeline, typically commences with the confirmed breast cancer diagnosis through biopsy (stage-$\mathcal{I}$), followed by the initiation of NAT (stage-$\mathcal{II}$), and the surgical intervention (stage-$\mathcal{III}$). Throughout this process, I_1, I_2, and I_3 stand for the breast imaging examinations in each stage to monitor the patient's response to NAT; R_1, R_2, and R_3 stand for paired radiology reports.

Recent studies have demonstrated the efficacy of deep learning in predicting pCR to NAT with breast imaging. Noteworthy, most of these prediction models only leveraged the single time-point or longitudinal breast MRI scans across NAT [15, 20], ignoring the incorporation of mammograms. As a complement to MRI, mammograms can provide reliable distinct pCR indicators regarding changes in tumor size, density, and calcifications [18, 4]. However, in-NAT mammograms are often unavailable in clinical practice (cf. stage-$\mathcal{II}$ in Fig. 1) to avoid radiation exposure [23]. It would benefit the precise prediction of pCR through multi-modal data integration, and minimize the radiation exposure to patients if we could generate longitudinal (i.e., in-NAT) mammograms with generative models. Existing longitudinal medical image generation methods [11, 10, 27, 19] mainly relied on raw-pixel inputs, less exploration in the knowledge provided by longitudinal radiology reports. In contrast, recent efforts [16, 9, 29] on medical visual representation learning have shown that radiology reports contain rich semantic information which may improve image generation tasks. However, deploying the longitudinal semantic representations from radiology reports to the longitudinal image generation framework remains a less explored topic. In this work, we propose a novel multi-modal pCR prediction framework that integrates longitudinal mammogram generation with cross-modal radiological reports. Leveraging the advantage of the conditional diffusion generative model over ordinary generative adversarial network (GAN) framework [3], our model explores the integration of MRI reports and pre-NAT mammograms to generate in-NAT mammograms for the patients who missed true in-NAT mammograms from our clinical practice. Meanwhile, it is constrained by the alignment to longitudinal reports in the latent space to enable reasonable image generation. Moreover, according to the response monitoring needs in various NAT scenarios, we establish experiments in each NAT stage, demonstrating the effectiveness of our generated mammograms for early accurate pCR prediction. The experiment at the time point of stage-$\mathcal{III}$ is considered the theoretical upper bound. Our primary contributions are as follows: (1) We investigate a novel approach for early NAT response prediction of BC patients by integrating the synthesized absent in-NAT mammograms, facilitating a comprehensive multi-modal analysis. (2) In the generation process, we introduce longitudinal image-report alignment (LIMRA) methodology to guide reasonable longitudinal image generation. (3) By employing the generated in-NAT mammogram within the pCR prediction pipeline, the model achieves comparable performance to the upper bound, thus potentially providing a 3-month therapeutic replacement window.

2 Method

The overall pipeline comprises two main parts. The first part is longitudinal mammogram generation, depicted in Fig. 2. We propose a report-guided denoising diffusion probabilistic model (RG-DDPM) that learns a conditional score function of the semantics between paired longitudinal reports; additionally, we introduce an U-shape image-report alignment model to promote the generation task. In the second part, we train the pCR prediction model based on multi-modal inputs from stage-($\mathcal{I}$, $\mathcal{II}$, $\mathcal{III}$), for quantifying the effectiveness of pCR prediction enhanced by our generated stage-$\mathcal{II}$ (i.e. in-NAT) mammograms.

2.1 Longitudinal Mammogram Generation

RG-DDPM. Our report-guided diffusion-based model (RG−DDPM) follows the formulation of DDPMs given in [7, 21]. The DDPM is a generative model that aims to predict the added noise at each step of the diffusion process in order to generate images. Specifically, given an input image x_0, the model adds small amounts of noise to it in the forward diffusion process to generate a series of noisy images from $x_0, \ldots, x_t, \ldots, x_T$, where $0 < t \leq T$. The amount of noise added at each step t is determined by the parameter β_t, where $0 < \beta_t < 1$. The image at

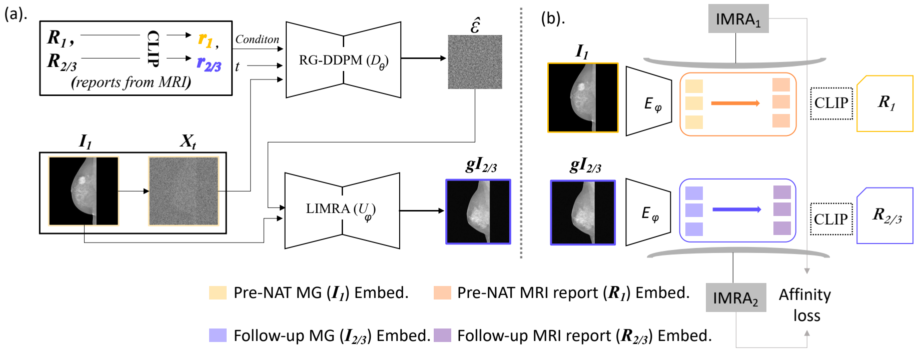

Fig. 2. Workflow of Longitudinal MG Generation Network: (a) Given longitudinal MRI reports ($R_1, R_{2/3}$, where '/' indicates either one), time step t, and the noisy pre-NAT mammogram X_t, the report-guided diffusion model (RG-DDPM) D_θ estimates the conditional score function (ϵ). The pre-NAT mammogram I_1 and ϵ are then provided to the U-shape generation model U_φ for synthesizing the follow-up mammogram ($gI_{2/3}$). A longitudinal image-report alignment (LIMRA) scheme is introduced during the generation process to enable reasonable image generation. (b) Affinity loss is employed between patient-wise image-report alignment matrices (IMRA). The image and report features are embedded from the encoder of U_φ and CLIP [17], respectively. The details of the $I_{1/2/3}$ and $R_{1/2/3}$ are provided Fig. 1.

step t is generated as follows: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\alpha_t}\epsilon$, where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$, $\epsilon \sim \mathcal{N}(0, I)$. In reverse diffusion, the model is trained to predict $x_{(t-1)}$ from x_t. With this iterative denoising process, it can generate a fake image x_0. Inspired by this, we employ the strategy of conditional diffusion models [3] to generate images with desired semantics. It compares the conditions of longitudinal reports input and delivers semantic information through conditional score function $\hat{\epsilon}$ to the next model. The mean squared error (MSE) loss used for training is $\mathcal{L}_d = \|\epsilon - D_\theta(x_t, t; r_1, r_{2/3})\|_2^2$, where D_θ stands for RG-DDPM model, and θ represents its weight. Textural feature maps (r_1 denotes embeddings from R_1, and $r_{2/3}$ representing embeddings from either R_2 or R_3) acquired through CLIP [17].

Longitudinal image-report alignment (LIMRA). Research has demonstrated that leveraging the representation alignment in the latent space between text and images improves the scalability of visual representations [30, 6, 5]. Inspired by this and recognizing the value of longitudinal reports in disease progress tracking, we introduce the longitudinal image-report alignment (LIMRA) model. This approach facilitates mammogram generation through two levels. At the feature level, it encompasses the alignment of latent space representations of the image and report at each time point. At the image level, it enables the generation of reasonably deformable mammograms.

LIMRA and longitudinal affinity learning. In Fig. 2.(b), the visual feature map (i_1, $gi_{2/3}$) and textual feature map (r_1, $r_{2/3}$) are embedded from the encoder of U_φ and CLIP [17], respectively. Both have shapes $(b, c, w \times h)$ and (b, c, r_{num}), where b is the batch size, c is the channel, and w, h represent the width and height of image features. r_{num} corresponds to the number of words in the report. The detailed generation process of obtaining $gI_{2/3}$ is elaborated in the subsequent paragraph, focusing on image-level learning. Here, for latent space alignment at each time point t, we introduce the image-aligned report feature map $\mathcal{F}_t$. It is defined as $\mathcal{F}_t = i_t \odot r_t$, where $\odot$ represents element-wise multiplication. Consequently, we measure the image-report alignment matrix ($IMRA$) between image-weighted features (i.e., $\mathcal{F}_t \odot i_t$) and respective report features using cosine similarity, defined as: $IMRA_t = cos(\mathcal{F}_t \odot i_t, r_t)$. In this way, $IMRA$ can capture the relational associations across all regions of the image features $(w \times h)$ to each word (r_{num}) feature of the report, allowing for more effective image-report representation learning. Let $IMRA_1$ and $IMRA_2$ represent the pre-NAT and follow-up image-report alignment matrices, respectively. Ideally, for the same patient within a batch, their $IMRA$ matrices between two time points (i.e., $IMRA_1$ and $IMRA_2$) should be in affinity to each other for intra-patient agreement compared to different patients. Within a batch of N cases, we denote the affinity loss as $\mathcal{L}_a = \frac{1}{N}\sum_{n=1}^{N} \|(IMRA_1^n - IMRA_2^n)\|_2^2$.

Furthermore, regarding image-level learning, to ensure that the generated mammogram has similar tissue structure to the pre-NAT mammogram, meanwhile enabling deformation for progressive changes, we employ the decoder (D_φ) of U_φ to generate the deformed image according to the semantic information provided by $\hat{\epsilon}$ and pre-NAT mammogram. Precisely, the longitudinal mammogram is generated by warping the input I_1 with a predicated deformation field from D_φ for the alignment purpose. We use MSE loss $(\mathcal{L}_r)$ to encourage a reasonably deformable generation of the target mammogram. The ground truth of the target mammogram includes real-world images I_2 and I_3 in the generation task, with details provided in supplementary S.Tab.1. By minimizing the combined above three loss functions $(\mathcal{L}_{sum} = \mathcal{L}_d + \mathcal{L}_a + \mathcal{L}_r)$ during the generation training process, the model can effectively learn to generate high-quality longitudinal mammogram images that capture the multi-level tissue progress over time.

2.2 Longitudinal pCR prediction.

In this phase, our goal is to evaluate the performance and time-effectiveness of pCR prediction by using our generated in-NAT mammogram of stage-$\mathcal{II}$. Meanwhile, we leverage a number of pCR prediction experiments by using multi-modal information obtained from different therapy stages as comparing baselines. We employ the same settings in all experiments, with a binary classifier and late fusion strategy that combines various features, including image features from longitudinal mammograms, textual features output from CLIP by paired MRI reports, and clinical information features (more details are provided in Sec. 3.1). The clinical features were encoded from a linear layer followed by the Exponential Linear Unit (ELU) activation function. The binary cross-entropy between the target attributes and the predicted attribute probability vector is used as the classification loss.

3 Experiments

3.1 Dataset and Metric

We deploy our model on the in-house dataset comprising 4,456 longitudinal mammogram exams (including bilateral multi-view mammograms) from 434 patients, along with paired radiology reports from MRI and clinical information, including pre-NAT T, N, M stage and molecular subtype. The dataset details are provided in S.Tab.1. Data were split 75%/10%/15% at the patient level for training, validation and test, respectively. Note that the reports analyzed in our study were derived from paired MRI, considering that in-NAT mammogram and paired report are usually unavailable for inference.

In our longitudinal image generation experiment, we utilize the learned perceptual image patch similarity (LPIPS) metric [28] to measure the quality of the generated images. This metric is specifically suitable for our study as it focuses on assessing perceptual similarity while reflecting semantic differences between the generated mammograms and real images across multiple latent spaces. Moreover, it avoids the potential impact of non-mammogram related changes (e.g., variations by patient positioning), which makes peak signal noise rate (PSNR) and structural similarity index measure (SSIM) inappropriate. For pCR prediction, we report the sensitivity, specificity, positive and negative predictive values (PPV, NPV), and area under the receiver operating curve (AUC), accompanied 95% confidence intervals, using the bootstrapping method [2]($N = 1000$ replicates).

3.2 Implementations

During the training of image generation, we adopt the architecture and hyperparameters of the diffusion model from the DDPM [7]. For our LIMRA model, we employ a U-shape network with the backbone of VoxelMorph [1]. The generation model was trained for 80,000 iterations with a batch size of 2 on Nvidia A6000 48GB GPU, which took approximately 180 hours. For the pCR prediction classifier, we choose the ResNet-50 architecture as the backbone. The Adam optimizer is employed with an initial learning rate of 10^{-4} and a weight decay of 10^{-6}. The maximum number of epochs and batch size were set to 200 and 8.

3.3 Comparisons and statistical analysis

Results on In-NAT mammogram synthesis. In Fig.3 (left), we present qualitative comparisons of the generated mammograms. Our results indicate that our model can effectively capture the distinct breast morphological characteristics changes between non-pCR and pCR cases. Specifically, for pCR case 1

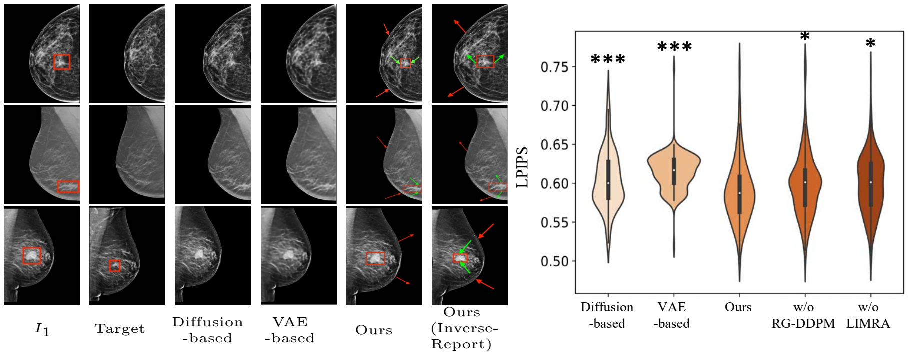

Fig. 3. Visualization and quantification of longitudinal mammogram generation. Left: mammogram examples of pCR (case 1, 2: top two rows) and non-pCR patients (case 3: the third row), with a red box indicating the tumor location. **Right:** the violin plots of learned perceptual image patch similarity (LPIPS) distribution. For detailed statistical improvements (*,**,***) are provided in Tab.2.

and case 2, our model generates shrink mammograms, while for non-pCR case 3, the model generates a micro-altered mammogram, which is consistent with the clinical reality. Furthermore, we compare "Ours(Inverse-Report)" with the initial output, where the input involves replacing the original report with its opposite meaning. For instance, for case 2, when replacing the original pCR-related keyword "remission" with "residual", the output of Ours(Inverse-Report) shows more density in the tumor region, while the original output shows a shrinking tumor. This demonstrates our model can understand different semantic inputs from reports and generate semantically compliant mammograms. In contrast, the two baselines of Diffusion-based [10] and VAE-based [19] models, cannot generate different semantic mammogram outputs, and the input and output mammograms showed no obvious visual difference. This may be because their models are designed for generating specific patterns, such as next-frame reconstruction [10] and brain chronological aging [19], but the change of longitudinal mammogram during therapy is uncertain, making it less applicable. Then, we perform quantitative comparisons by calculating LPIPS between the target and the synthesized mammogram. The distribution plot in Fig.3 demonstrates that our approach yields more similar semantics between real and generated mammograms, with a larger number of points clustering towards the lower end. Statistical analysis shows the proposed method achieves significant improvements (Tab.2) on LPIPS over diffusion-based (p <0.05) and VAE-based (p <0.01) methods. Moreover, t-SNE [12] plots (S.Fig.1) show that our generated mammogram representations more closely match real mammogram representations in latent space.

Evaluation of the time-effectiveness to pCR prediction performance enhanced by our generated stage-$\mathcal{II}$ mammogram. We conducted exper-

Table 1. Results for pCR prediction, utilizing different therapy stage information. Each P-value is calculated on AUC by comparing it with the upper bound (the last column). Note that the experiment comparing the generated in-NAT MG (gI_2) with real-world MG is shown in S.Tab.2.

Scenarios	Pre-NAT evaluation		In-NAT evaluation		Post-NAT evaluation (before surgery)
Modalities	I_1+C	I_1+R_1+C	I_1+R_1+R_2+C	I_1+R_1+R_2+C+gI_2	I_1+R_1+I_3+R_3+C
Sensitivity	0.546 [0.469,0.627]	0.587 [0.505,0.663]	0.673 [0.605,0.752]	0.674 [0.607,0.752]	0.795 [0.691,0.848]
Specificity	0.600 [0.507,0.656]	0.686 [0.584,0.739]	0.672 [0.576,0.725]	0.767 [0.654,0.802]	0.719 [0.615,0.769]
PPV	0.474 [0.420,0.562]	0.569 [0.515,0.653]	0.571 [0.527,0.663]	0.658 [0.605,0.741]	0.676 [0.628,0.763]
NPV	0.671 [0.591,0.746]	0.675 [0.597,0.742]	0.736 [0.658,0.807]	0.756 [0.671,0.829]	0.827 [0.747,0.898]
AUC	0.567 [0.501,0.649]	0.609 [0.532,0.683]	0.669 [0.601,0.741]	0.710 [0.649,0.781]	0.765 [0.704,0.834]
P-value	1.406e-4	6.551e-4	3.204e-2	0.185	–

The details of the $I_{1/2/3}$, $R_{1/2/3}$ and C are provided in Fig. 1.

Table 2. For pCR prediction performance of comparing methods by integrating their generated in-NAT mammogram with the information in the fifth column of Tab. 1. Each P-value is calculated on LPIPS between ours and compared method.

Methods	Sensitivity	Specificity	PPV	NPV	AUC	LPIPS↓	P-value
Diffusion-based model [10]	0.621 [0.552,0.702]	0.756 [0.649,0.794]	0.646 [0.583,0.736]	0.726 [0.644,0.793]	0.672 [0.594,0.744]	0.604 ±0.069	1.378e-4
VAE-based model [19]	0.563 [0.498,0.643]	0.748 [0.638,0.785]	0.610 [0.564,0.715]	0.690 [0.617,0.760]	0.671 [0.593,0.749]	0.615 ±0.041	3.150e-14
w/o RG-DDPM	0.632 [0.563,0.715]	0.681 [0.598,0.742]	0.587 [0.533,0.686]	0.711 [0.632,0.788]	0.692 [0.619,0.761]	0.598 ±0.040	4.125e-2
w/o LIMRA	0.598 [0.539,0.689]	0.714 [0.613,0.761]	0.595 [0.548,0.696]	0.697 [0.616,0.761]	0.681 [0.606,0.747]	0.599 ±0.037	1.753e-2
Ours	**0.674** [0.607,0.752]	**0.767** [0.654,0.802]	**0.658** [0.605,0.741]	**0.756** [0.671,0.829]	**0.710** [0.649,0.781]	**0.589** ±0.039	–

iments to predict pCR at the patient level using inputs from different therapy stages, as defined in Fig. 1. Multi-view mammograms were utilized and combined to form a comprehensive imaging representation for each patient [26, 5]. Tab.1 demonstrates that incorporating reports enhances pCR prediction AUC by approximately 4% than single-modal input, while incorporating longitudinal imaging improves AUC by around 11%. Further, our generated in-NAT mammogram (gI2), besides eliminating the statistical difference between stage-$\mathcal{II}$ and the upper bound of stage-$\mathcal{III}$ (reducing the difference from 3.204e-2 to 0.185), achieves the best pCR prediction performance among the generation models (Tab.2). This implies that our model could potentially allow a 3-month window for therapeutic replacement for predicted non-pCR patients with a specificity of 76.8%. Moreover, S.Tab.2 provides a comparison of pCR prediction results in sub-populations using real-world in-NAT mammograms.

Ablations. In this section, we conduct an ablation study to analyze the impact of the RG-DDPM model and LIMRA model on our model's performance. Evidently, as depicted in Fig.3 (right) and Tab. 2, the w/o LIMRA model performs poorly compared to the w/o RG-DDPM model. This is not surprising since the diffusion-only (RG-DDPM) model is not specifically designed for longitudinal image generation and is unable to capture semantic changes over time. However, despite lacking the LIMRA model, it is still guided by reports as conditions during training, thereby enabling our RG-DDPM-only model to exhibit competitive performance to that of the diffusion-based model as well as VAE-based model.

4 Conclusion

In this paper, we propose the longitudinal image-report alignment (LIMRA) for guiding longitudinal mammogram generation from cross-modal radiological reports. Additionally, we establish a novel approach for early NAT response prediction of BC patients by incorporating generated absent mammograms within the framework of longitudinal multi-modal modeling. Importantly, utilizing our generated in-NAT mammogram achieves comparable performance to the upper bound, thus providing a potential 3-month window for therapeutic replacement. Future work could integrate longitudinal mammogram generation with other breast imaging modalities, such as MRI and ultrasound, to facilitate multi-modal learning in pCR prediction and other clinical downstream tasks.

Acknowledgments. The authors thank the support from China Scholarship Council.

Disclosure of Interests. The authors declare no competing interests.

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9252–9260 (2018)
2. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. Statistics in medicine **19**(9), 1141–1164 (2000)
3. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
4. Di Cosimo, S., Depretto, C., Miceli, R., Baili, P., Ljevar, S., Sant, M., Cappelletti, V., Folli, S., Gennaro, M., De Braud, F., et al.: Mammographic density to predict response to neoadjuvant systemic breast cancer therapy. Journal of Cancer Research and Clinical Oncology pp. 1–7 (2022)
5. Gao, Y., Zhou, H.Y., Wang, X., Zhang, T., Tan, R., Han, L., Estacio, L., D'Angelo, A., Teuwen, J., Mann, R., et al.: Visualize what you learn: a well-explainable joint-learning framework based on multi-view mammograms and associated reports (2023)
6. Heiliger, L., Sekuboyina, A., Menze, B., Egger, J., Kleesiek, J.: Beyond medical imaging-a review of multimodal deep learning in radiology (2022)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)

8. Jin, C., Yu, H., Ke, J., Ding, P., Yi, Y., Jiang, X., Duan, X., Tang, J., Chang, D.T., Wu, X., et al.: Predicting treatment response from longitudinal images using multi-task deep learning. Nature communications **12**(1), 1–11 (2021)
9. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022)
10. Kim, B., Han, I., Ye, J.C.: Diffusemorph: Unsupervised deformable image registration using diffusion model. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI. pp. 347–364. Springer (2022)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
13. Portnow, L.H., Kochkodan-Self, J.M., Maduram, A., Barrios, M., Onken, A.M., Hong, X., Mittendorf, E.A., Giess, C.S., Chikarmane, S.A.: Multimodality imaging review of her2-positive breast cancer and response to neoadjuvant chemotherapy. RadioGraphics **43**(2), e220103 (2023)
14. Qu, Y.H., Zhu, H.T., Cao, K., Li, X.T., Ye, M., Sun, Y.S.: Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using a deep learning (dl) method. Thoracic Cancer **11**(3), 651–658 (2020)
15. Rabinovici-Cohen, S., Tlusty, T., Abutbul, A., Antila, K., Fernandez, X., Rejo, B.G., Hexter, E., Cubelos, O.H., Khateeb, A., Pajula, J., et al.: Radiomics for predicting response to neoadjuvant chemotherapy treatment in breast cancer. In: Medical imaging 2020: imaging informatics for healthcare, research, and applications. vol. 11318, pp. 333–341. SPIE (2020)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
18. Romeo, V., Accardo, G., Perillo, T., Basso, L., Garbino, N., Nicolai, E., Maurea, S., Salvatore, M.: Assessment and prediction of response to neoadjuvant chemotherapy in breast cancer: A comparison of imaging modalities and future perspectives. Cancers **13**(14), 3521 (2021)
19. Sauty, B., Durrleman, S.: Progression models for imaging data with longitudinal variational auto encoders. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I. pp. 3–13. Springer (2022)
20. Skarping, I., Larsson, M., Förnvik, D.: Analysis of mammograms using artificial intelligence to predict response to neoadjuvant chemotherapy in breast cancer patients: proof of concept. European Radiology pp. 1–11 (2022)
21. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
22. Spring, L.M., Fell, G., Arfe, A., Sharma, C., Greenup, R., Reynolds, K.L., Smith, B.L., Alexander, B., Moy, B., Isakoff, S.J., et al.: Pathologic complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: A comprehensive meta-analysispcr and association with clinical outcomes in breast cancer. Clinical Cancer Research **26**(12), 2838–2848 (2020)

23. Thompson, A., Moulder-Thompson, S.: Neoadjuvant treatment of breast cancer. Annals of oncology **23**, x231–x236 (2012)
24. Von Minckwitz, G., Untch, M., Blohmer, J.U., Costa, S.D., Eidtmann, H., Fasching, P.A., Gerber, B., Eiermann, W., Hilfrich, J., Huober, J., et al.: Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. J Clin oncol **30**(15), 1796–1804 (2012)
25. Wang, X., Moriakov, N., Gao, Y., Zhang, T., Han, L., Mann, R.M.: Artificial intelligence in breast imaging. Breast Imaging: Diagnosis and Intervention pp. 435–453 (2022)
26. Wang, X., Tan, T., Gao, Y., Su, R., Zhang, T., Han, L., Teuwen, J., D'Angelo, A., Drukker, C.A., Schmidt, M.K., et al.: Predicting up to 10 year breast cancer risk using longitudinal mammographic screening history. medRxiv pp. 2023–06 (2023)
27. Yoon, J.S., Zhang, C., Suk, H.I., Guo, J., Li, X.: Sadm: Sequence-aware diffusion model for longitudinal medical image generation. arXiv preprint arXiv:2212.08228 (2022)
28. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
29. Zhou, H.Y., Chen, X., Zhang, Y., Luo, R., Wang, L., Yu, Y.: Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. Nature Machine Intelligence **4**(1), 32–40 (2022)
30. Zhou, H.Y., Lian, C., Wang, L., Yu, Y.: Advancing radiograph representation learning with masked record modeling. arXiv preprint arXiv:2301.13155 (2023)