Spatiotemporal Modeling of Longitudinal MRI with Causal Graph Reasoning for Breast Cancer pCR Prediction

Haiwei Lin[1], Bin Huang[1], Meng Wang[2], Ya Ren[2], Zhou Liu[2(✉)], and Bingsheng Huang[1(✉)]

[1] Medical AI Lab, School of Biomedical Engineering, Medical School, Shenzhen University, Shenzhen, China
huangbs@gmail.com

[2] Department of Radiology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, 113 Baohe Avenue, 518116, Shenzhen, China
zhou_liu8891@yeah.net

Abstract. Early and accurate prediction of pathological complete response (pCR) to neoadjuvant chemotherapy (NAC) is critical for personalized breast cancer treatment. However, modeling complex longitudinal dynamics remains challenging in small-sample settings. Although biomedical foundation models offer strong representations, most 2D architectures ignore 3D spatial continuity and multi-scale tumor morphology, and existing approaches mainly rely on statistical correlations without considering underlying causal mechanisms. To address these issues, we propose a novel framework that integrates spatiotemporal modeling leveraging a biomedical foundation model with causal graph reasoning for pCR prediction. Our method introduces a Spatiotemporal-aware Attention (STA) mechanism, consisting of Multi-scale Adaptive Spatial-aware Attention (MASA) to capture 3D volumetric tumor morphology and Temporal-aware Attention (TA) to model treatment-induced changes across longitudinal MRI scans. Furthermore, we incorporate Causal Graph Reasoning to uncover latent causal pathways among longitudinal imaging dynamics, clinical biomarkers, and pCR, enabling unbiased estimation of treatment effects. Experiments on a multicenter cohort of 438 cases show that our method outperforms existing approaches. The causal model further surpasses its non-causal counterpart and identifies consistent causal effects of key biomarkers (ER, PR, HER-2) across centers. These results suggest that our method improves both predictive accuracy and interpretability.

Keywords: Longitudinal MRI · Spatiotemporal Attention · Causal Graph Reasoning · Foundation Model · Breast Cancer

1 Introduction

Breast cancer is the most common malignancy among women [20]. Neoadjuvant chemotherapy (NAC) improves surgical outcomes, and pathological complete response (pCR) is a surrogate marker of favorable survival [9,22,4]. However, NAC response varies markedly, with only 19–30% achieving pCR [10]. As surgical pathology remains the gold standard, reliable preoperative prediction is lacking [23]. Therefore, early and non-invasive prediction of treatment response is critically needed.

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is widely used to assess NAC response in breast cancer. Longitudinal imaging characterizes lesion evolution across treatment, providing guidance for optimizing NAC and improving pCR [12]. However, the heterogeneity of longitudinal data challenges machine learning models [1]. Many studies rely on feature concatenation or difference-based fusion without explicitly modeling temporal dependencies [13,15]. Recurrent neural networks (RNNs) improve multi-time-point modeling but suffer from high complexity and limited robustness [5,17]. Transformers further enhance temporal modeling via attention mechanisms [26,21], yet these approaches still depend on conventional deep learning frameworks and large annotated datasets, limiting generalization in small-sample clinical settings.

Biomedical foundation models demonstrate promise in extracting generalizable representations via self-supervised learning, particularly for cancer imaging biomarker discovery [25,16]. However, many foundation models utilize 2D architectures for 3D medical images, risking loss of inter-slice spatial continuity [8]. Furthermore, substantial variability in breast cancer lesion size and irregular morphological evolution during treatment complicate robust representation learning [4].

Traditional machine learning often overlooks underlying causal mechanisms, potentially leading to misleading clinical conclusions [24,18]. By contrast, causal reasoning identifies graph structures from observational data to explicitly characterize logical relationships among variables. Through counterfactual reasoning and intervention analysis, it mitigates confounding bias and enables unbiased treatment effect estimation. This provides robust inductive constraints for data-scarce medical imaging and facilitates a deeper understanding of disease progression, thereby optimizing personalized treatment strategies [2,7].

To address the above limitations, this study makes the following contributions: 1) We propose a spatiotemporal-aware attention mechanism integrating a biomedical foundation model with longitudinal MRI to capture multi-scale 3D spatial features and dynamic temporal changes. 2) We develop a tumor response prediction framework based on causal graph reasoning, revealing latent causal pathways among imaging dynamics, clinical biomarkers, and outcome, thereby improving both predictive accuracy and interpretability.

2 Method

2.1 Overview

As illustrated in Fig. 1, we proposed a framework that integrates spatiotemporal modeling leveraging a biomedical foundation model with causal graph reasoning to predict breast cancer treatment response. The framework consisted of two key components: Spatiotemporal-aware Attention and Causal Graph Reasoning.

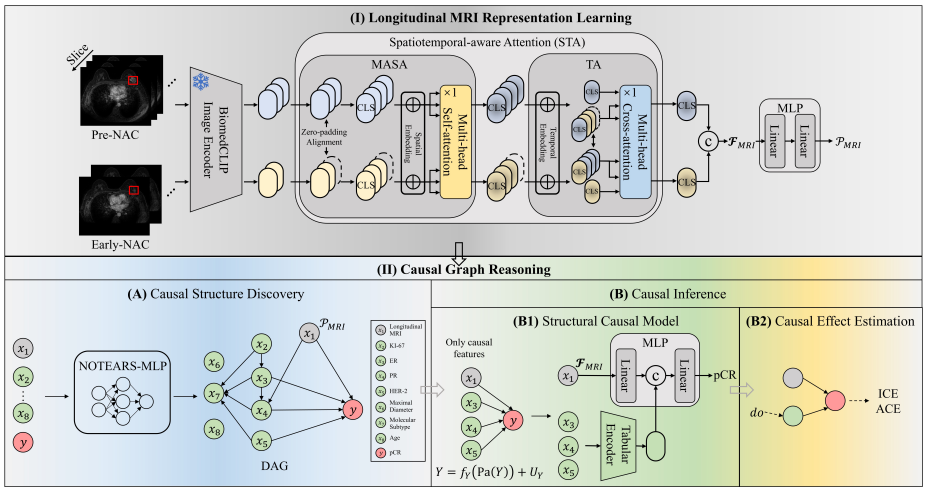

Fig. 1. Overview of the proposed method. MASA: Multi-scale Adaptive Spatial-aware Attention; TA: Temporal-aware Attention; CLS: classification token; $\mathcal{F}$: features; $\mathcal{P}$: probability of pCR; Pa: parent nodes; ICE: individual causal effect; ACE: average causal effect.

2.2 Spatiotemporal-aware Attention (STA)

Multi-scale Adaptive Spatial-aware Attention (MASA) For a given patient at time point t, the tumor volume is represented as a set of axial slices $\mathcal{I}^t = \{I_1^t, I_2^t, \ldots, I_N^t\}$, where $I_i^t \in \mathbb{R}^{H \times W}$ and N denotes the number of tumor-containing slices. Each slice is independently encoded by a shared-parameter BiomedCLIP [25] image encoder, producing a sequence of slice-level feature embeddings $\mathcal{F}^t = \{f_1^t, f_2^t, \ldots, f_N^t\}$ with $f_i^t \in \mathbb{R}^C$.

Due to substantial inter-patient and inter-timepoint variations in tumor size, the number of slices N varies across samples. To enable consistent multi-scale modeling of volumetric features, the feature sequence is aligned to a fixed length $N_{\max}$ via zero padding:

$$\mathcal{F}^t \leftarrow \mathrm{Pad}(\mathcal{F}^t) = \mathrm{Concat}\left(\mathcal{F}^t, \mathbf{0}_{(N_{\max}-N)\times C}\right). \tag{1}$$

To aggregate global spatial information from the 3D volume, a learnable classification token CLS $\in \mathbb{R}^{1\times C}$ is prepended to the padded sequence. In addition, a learnable spatial positional embedding SE $\in \mathbb{R}^{(N_{\max}+1)\times C}$ is incorporated to explicitly encode the relative spatial positions of slices:

$$\mathcal{F}^t \leftarrow \mathcal{F}^t + \mathrm{SE}. \tag{2}$$

The resulting sequence is then processed through multi-head self-attention (MHSA) to capture global inter-slice dependencies. Following a pre-normalization strategy, the spatial aggregation is performed as:

$$\mathcal{F}^t \leftarrow \mathcal{F}^t + \mathrm{MHSA}\big(\mathrm{LN}(\mathcal{F}^t)\big). \tag{3}$$

A feed-forward network (FFN) is subsequently applied in a residual manner, i.e., $\mathcal{F}^t \leftarrow \mathcal{F}^t + \mathrm{FFN}(\mathrm{LN}(\mathcal{F}^t))$, further enhancing nonlinear feature representations. Finally, the CLS token aggregates global contextual information from all slices, yielding a compact representation that jointly captures overall tumor morphology and fine-grained spatial details.

Temporal-aware Attention (TA) To encode temporal order, a learnable time embedding $\mathrm{TE}_t \in \mathbb{R}^1$ is added to the feature sequence of each time point:

$$\mathcal{F}^t \leftarrow \mathcal{F}^t + \mathrm{TE}_t, \quad t = 0, 1. \tag{4}$$

For cross-time feature interaction, each time point's CLS token is concatenated with the slice-level features from the other time point to form a cross-time sequence:

$$\mathcal{F}^0 \leftarrow \mathrm{Concat}(\mathcal{F}^0_{\mathrm{CLS}}, \mathcal{F}^1_{\mathrm{slice}}), \quad \mathcal{F}^1 \leftarrow \mathrm{Concat}(\mathcal{F}^1_{\mathrm{CLS}}, \mathcal{F}^0_{\mathrm{slice}}). \tag{5}$$

Multi-head cross-attention (MHCA) is then applied to aggregate dynamic information. For each time point t, the CLS token $\mathcal{F}^t_{\mathrm{CLS}}$ serves as the query, while the corresponding cross-time sequence $\mathcal{F}^t$ serves as the key and value:

$$\mathcal{F}^t_{\mathrm{CLS}} \leftarrow \mathcal{F}^t_{\mathrm{CLS}} + \mathrm{MHCA}\big(\mathrm{LN}(\mathcal{F}^t_{\mathrm{CLS}}), \mathrm{LN}(\mathcal{F}^t)\big), \quad t = 0, 1. \tag{6}$$

The CLS tokens are then refined through an FFN with residual connection. The final temporal feature is obtained by concatenating the two CLS tokens, $\mathcal{F} = \mathrm{Concat}(\mathcal{F}^0_{\mathrm{CLS}}, \mathcal{F}^1_{\mathrm{CLS}}) \in \mathbb{R}^{2C}$, which is subsequently passed to a multi-layer perceptron (MLP) to predict the treatment outcome, $\hat{y} = \mathrm{MLP}(\mathcal{F})$.

2.3 Causal Graph Reasoning

Causal Structure Discovery To reduce the complexity of causal graph learning, we use the predicted probabilities from the longitudinal MRI model as image representations, combined with clinical data and labels, as input to NOTEARS [27] to learn a directed acyclic graph (DAG) over the variables. As an example, we

consider a linear structural equation model (SEM), where each variable X_j in the set $X = \{X_1, X_2, \ldots, X_d\}$ is expressed as a linear combination of the others, $X_j = \sum_{i \neq j} W_{ij} X_i + Z_j$, with $W \in \mathbb{R}^{d \times d}$ denoting the weighted adjacency matrix, W_{ij} representing the direct influence of X_i on X_j, and Z_j an independent noise term. The parameters are estimated via least squares with ℓ_1 regularization to promote sparsity, $F(W) = \frac{1}{2n}\|X - XW\|_F^2 + \lambda\|W\|_1$, and a smooth acyclicity constraint is imposed, $h(W) = \mathrm{tr}\big(\exp(W \circ W)\big) - d$, where $\circ$ denotes the Hadamard product. The causal structure learning problem is then formulated as a constrained optimization:

$$\min_{W \in \mathbb{R}^{d \times d}} F(W) \quad \text{subject to } h(W) = 0. \tag{7}$$

This nonconvex constrained optimization problem is solved using the augmented Lagrangian method:

$$\mathcal{L}^{\rho}(W, \alpha) = F(W) + \frac{\rho}{2} h(W)^2 + \alpha h(W), \tag{8}$$

where α is the Lagrange multiplier and ρ is the penalty coefficient. Optimization proceeds via alternating updates: 1) minimize $\mathcal{L}^{\rho}(W, \alpha)$ with respect to W; 2) update the multiplier $\alpha \leftarrow \alpha + \rho h(W)$; 3) gradually increase ρ until $h(W) < \varepsilon$. Each unconstrained subproblem is efficiently solved using a quasi-Newton method.

Finally, to enhance sparsity and remove spurious edges due to numerical error, any weight with $|W_{ij}| < \tau$ is set to zero, yielding the final DAG structure.

Causal Inference Based on the learned DAG, we further perform causal inference. A structural causal model (SCM) is constructed to functionalize the DAG and enable explicit intervention-based effect estimation.

Formally, an SCM is defined as $\mathcal{M} := \langle X, U, F \rangle$, where X denotes the set of endogenous variables corresponding to all nodes in the DAG, U represents mutually independent exogenous noise variables, and F is the set of structural functions governing the causal mechanisms among variables. For each endogenous variable $X_j \in X$, the structural equation is given by $X_j = f_j(\mathrm{Pa}(X_j)) + U_j$, where $\mathrm{Pa}(X_j)$ denotes the set of parent nodes of X_j in the DAG.

Since this study focuses on treatment outcome prediction and its causal interpretation, we restrict the SCM construction to causal edges pointing to the outcome variable Y. The corresponding structural equation is expressed as:

$$Y = f_Y(\mathrm{Pa}(Y)) + U_Y. \tag{9}$$

After constructing the SCM, causal effects are estimated via explicit intervention operations. For a parent variable $X_k \in \mathrm{Pa}(Y)$ that is binary, the individual causal effect (ICE) for subject i is defined as:

$$\mathrm{ICE}_{i, X_k \to Y} = Y_i(\mathrm{do}(X_k = 1)) - Y_i(\mathrm{do}(X_k = 0)). \tag{10}$$

The average causal effect (ACE) is obtained by aggregating individual effects across the population:

$$\mathrm{ACE}_{X_k \to Y} = \mathrm{E}_i[Y_i(\mathrm{do}(X_k = 1))] - \mathrm{E}_i[Y_i(\mathrm{do}(X_k = 0))]\,. \tag{11}$$

3 Experiments and Results

3.1 Dataset and Evaluation

This retrospective multicenter study enrolled breast cancer patients receiving NAC across three centers. For each patient, pre-NAC and early-NAC DCE-MRI scans were acquired, along with clinical data including age, ER, PR, HER-2, Ki-67, molecular subtype, and maximum tumor diameter. Tumor regions of interest (ROIs) were manually segmented on DCE-MRI by experienced radiologists. pCR was determined via postoperative histopathology. A total of 438 cases were included (Center 1: $n = 237$, Center 2: $n = 119$, Center 3: $n = 82$), with pCR/non-pCR distributions of 81/156, 25/94, and 30/52, respectively. Center 1 was used for model development with five-fold cross-validation, while Centers 2 and 3 were used as independent external validation cohorts.

The model performance was evaluated using area under the receiver operating characteristic curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPE).

3.2 Implementation Details

Bilateral lesions were processed separately, and multiple unilateral lesions were merged. DCE-MRI at 2 minutes post-contrast injection was selected. Image preprocessing included: 1) resampling to $1 \times 1 \times 1$ mm^3; 2) intensity clipping to the 0.1–99.9th percentile; 3) min-max normalization to [0,1]; 4) cropping based on the ROI with a 10-pixel margin; 5) center cropping or zero-padding to $N \times 224 \times 224$; 6) z-score normalization. Missing values in clinical variables were imputed using Multiple Imputation by Chained Equations.

Gradient accumulation was applied with a step size of 16. The BiomedCLIP image encoder was frozen during training. Multi-scale feature sequences were aligned to a fixed length of $N_{\max} = 200$ via zero padding. Binary cross-entropy (BCE) loss was used. Our method was implemented in PyTorch, using an NVIDIA RTX A6000 GPU. The model was trained for 15 epochs using the Adam optimizer with cosine annealing to reduce the learning rate of newly added modules from 5×10^{-5} to 1×10^{-7}.

For causal structure discovery, we employed a non-linear NOTEARS-MLP model with default settings to learn the DAG.

3.3 Comparison with Existing Methods

We evaluated the proposed MASA and TA against existing spatial and temporal modeling methods. For spatial modeling comparison, MASA was compared with average pooling and multiple instance learning (MIL) methods, including ABMIL [14] and TransMIL [19], with all methods using simple concatenation for temporal modeling. For temporal modeling comparison, TA was compared with LSTM [11], BiLSTM, GRU [3], BiGRU, and iMRrhpc [6] (simple concatenation), while all methods used MASA for spatial modeling. As shown in Tables 1

Table 1. Comparison of MASA with existing methods for spatial modeling.

Method	Center 1 (Internal)				Center 2 (External)				Center 3 (External)				Mean AUC
	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	
Mean	0.742	0.717	0.642	0.756	0.646	0.588	0.800	0.532	0.772	0.695	0.800	0.635	0.720
ABMIL [14]	0.736	0.696	0.765	0.660	0.612	0.630	0.600	0.638	0.708	0.683	0.767	0.635	0.685
TransMIL [19]	0.743	0.713	0.605	0.769	0.671	0.656	0.720	0.638	0.815	0.793	0.767	0.808	0.743
MASA (Ours)	0.756	0.743	0.568	0.833	0.674	0.723	0.560	0.766	0.805	0.793	0.833	0.769	**0.745**

Table 2. Comparison of TA with existing methods for temporal modeling.

Method	Center 1 (Internal)				Center 2 (External)				Center 3 (External)				Mean AUC
	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	
LSTM [11]	0.722	0.726	0.593	0.795	0.675	0.613	0.720	0.585	0.810	0.793	0.900	0.731	0.736
BiLSTM	0.709	0.705	0.617	0.750	0.669	0.748	0.480	0.819	0.816	0.781	0.867	0.731	0.731
GRU [3]	0.726	0.726	0.605	0.789	0.687	0.765	0.480	0.840	0.817	0.793	0.900	0.731	0.743
BiGRU	0.732	0.722	0.593	0.789	0.686	0.647	0.720	0.628	0.815	0.781	0.900	0.712	0.744
iMRrhpc [6]	0.756	0.743	0.568	0.833	0.674	0.723	0.560	0.766	0.805	0.793	0.833	0.769	0.745
TA (Ours)	0.783	0.717	0.778	0.686	0.686	0.765	0.560	0.819	0.800	0.781	0.833	0.750	**0.756**

and 2, the proposed MASA and TA outperform existing methods in spatial and temporal modeling, respectively.

3.4 Ablation Study

To evaluate the effectiveness of STA and the contribution of each component, we conducted ablation studies under both unimodal and multimodal settings. In the unimodal setting, using only longitudinal imaging data, the baseline model aggregates spatial features via average pooling and fuses temporal features through simple concatenation, without attention mechanisms. We then progressively incorporated MASA and TA to analyze their respective contributions to spatial structure modeling and temporal dynamic modeling. In the multimodal setting, clinical tabular data were further integrated to assess the effectiveness of the proposed method in jointly modeling imaging and clinical information.

As shown in Table 3, in the unimodal setting, introducing MASA increased the mean AUC across the three centers from 0.720 to 0.745, and further adding TA raised it to 0.756, demonstrating the effectiveness of STA in modeling both spatial and temporal features. In the multimodal setting, incorporating clinical data improved the mean AUC from 0.792 to 0.812 with MASA and to 0.814 with TA. Although the additional gain from STA is slightly reduced when clinical information is included, it still provides consistent performance improvement.

3.5 Causal Graph Reasoning

The learned DAG in Fig. 1(II-A) shows that longitudinal imaging representations, along with ER, PR, and HER-2, directly influence pCR, suggesting po-

Table 3. Ablation study of STA.

MASA	TA	Clinical data	Center 1 (Internal)				Center 2 (External)				Center 3 (External)				Mean AUC
			AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	
×	×	×	0.742	0.717	0.642	0.756	0.646	0.588	0.800	0.532	0.772	0.695	0.800	0.635	0.720
✓	×	×	0.756	0.743	0.568	0.833	0.674	0.723	0.560	0.766	0.805	0.793	0.833	0.769	0.745
✓	✓	×	0.783	0.717	0.778	0.686	0.686	0.765	0.560	0.819	0.800	0.781	0.833	0.750	**0.756**
×	×	✓	0.811	0.734	0.827	0.686	0.794	0.740	0.800	0.723	0.771	0.744	0.667	0.789	0.792
✓	×	✓	0.814	0.722	0.877	0.641	0.800	0.731	0.840	0.702	0.821	0.781	0.833	0.750	0.812
✓	✓	✓	0.826	0.734	0.901	0.647	0.789	0.782	0.760	0.787	0.826	0.756	0.867	0.692	**0.814**

Table 4. Comparison of predictive performance between non-causal and causal methods.

Method	Center 1 (Internal)				Center 2 (External)				Center 3 (External)				Mean AUC
	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	
Non-causal	0.826	0.734	0.901	0.647	0.789	0.782	0.760	0.787	0.826	0.756	0.867	0.692	0.814
Causal	0.835	0.738	0.889	0.660	0.810	0.765	0.840	0.745	0.824	0.756	0.833	0.712	**0.823**

tential causal relationships. As shown in Table 4, the causal model built from these causal features outperforms the non-causal model that uses all features, by effectively removing redundant information. Fig. 2 presents the distribution of ICE, revealing heterogeneity at the patient level. ACE estimates are consistent across centers: ER (−0.134, −0.110, −0.130) and PR (−0.101, −0.080, −0.097) negatively impact pCR, whereas HER-2 (0.082, 0.063, 0.086) exerts a positive effect. For example, in Center 1, ER positivity reduces the probability of pCR by 0.134 on average compared with ER negativity.

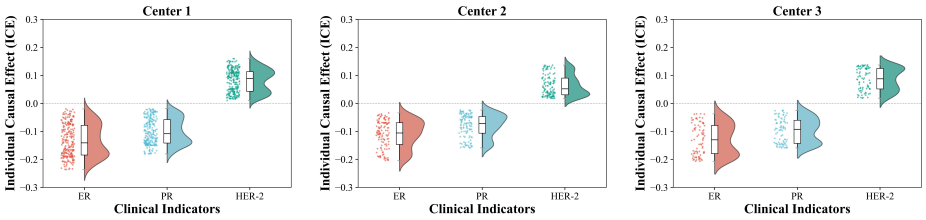

Fig. 2. Distribution plots of individual causal effect (ICE) for clinical indicators causally associated with pCR across three centers.

4 Discussion and Conclusions

The proposed STA effectively captures 3D spatial information and temporal dynamics, while causal graph reasoning highlights longitudinal imaging dynamics

and key biomarkers that have direct causal effects on pCR, providing interpretable insights into treatment response. The inferred causal graph aligns with established clinical knowledge of ER, PR, and HER-2 effects, supporting the plausibility of the causal relationships.

This study presented a framework that integrates spatiotemporal modeling leveraging a biomedical foundation model with causal graph reasoning to predict breast cancer treatment response. We proposed a Spatiotemporal-aware Attention (STA) mechanism, which utilized MASA for multi-scale 3D spatial modeling and TA to capture treatment dynamics across time points. Furthermore, we incorporated Causal Graph Reasoning to uncover latent causal pathways among longitudinal imaging dynamics, clinical biomarkers, and outcome. Multi-center validation demonstrates that our approach outperforms existing methods and provides interpretable causal insights, offering significant potential for personalized clinical decision support.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Nos. 82572308, 62371303, 82302153), the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515220117), the Shenzhen Science and Technology Program (No. JCYJ20250604181405007), the National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College (Nos. E010224007, E010324007), the Cooperation Fund of CHCAMS and SZCH (No. CFA202202022), the Beijing Medical Award Foundation (Nos. YXJL-2024-0350-0267, YXJL-2024-0350-0268), the Shenzhen High-level Hospital Construction Fund, and the Shenzhen Clinical Research Center for Cancer (No. 2021-287).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cascarano, A., Mur-Petit, J., Hernandez-Gonzalez, J., Camacho, M., de Toro Eadie, N., Gkontra, P., Chadeau-Hyam, M., Vitria, J., Lekadir, K.: Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artificial Intelligence Review **56**(Suppl 2), 1711–1771 (2023)
2. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 3673 (2020)
3. Cho, K., Van Merriënboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1724–1734 (2014)
4. Cortazar, P., Zhang, L., Untch, M., Mehta, K., Costantino, J.P., Wolmark, N., Bonnefoi, H., Cameron, D., Gianni, L., Valagussa, P., et al.: Pathological complete response and long-term clinical benefit in breast cancer: the ctneobc pooled analysis. The Lancet **384**(9938), 164–172 (2014)
5. Dadsetan, S., Arefan, D., Berg, W.A., Zuley, M.L., Sumkin, J.H., Wu, S.: Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. Pattern recognition **132**, 108919 (2022)

6. Gao, Y., Ventura-Diaz, S., Wang, X., He, M., Xu, Z., Weir, A., Zhou, H.Y., Zhang, T., van Duijnhoven, F.H., Han, L., et al.: An explainable longitudinal multi-modal fusion model for predicting neoadjuvant therapy response in women with breast cancer. Nature communications **15**(1), 9613 (2024)
7. Glocker, B., Musolesi, M., Richens, J., Uhler, C.: Causality in digital medicine. Nature Communications **12**(1) (2021)
8. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. Medical Image Analysis **98**, 103324 (2024)
9. Gradishar, W.J., Anderson, B.O., Balassanian, R., Blair, S.L., Burstein, H.J., Cyr, A., Elias, A.D., Farrar, W.B., Forero, A., Giordano, S.H., et al.: Breast cancer, version 4.2017, nccn clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network **16**(3), 310–320 (2018)
10. Haque, W., Verma, V., Hatch, S., Suzanne Klimberg, V., Brian Butler, E., Teh, B.S.: Response rates and pathologic complete response by breast cancer molecular subtype following neoadjuvant chemotherapy. Breast cancer research and treatment **170**(3), 559–567 (2018)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
12. Huang, X., Xu, Z., Zhao, Y., Wang, Y., Liu, Y., Hu, W., Zhao, K., Yao, L., He, J., Yu, Y., et al.: Longitudinal mri-based deep learning model for predicting pathological complete response in breast cancer: a multicenter, retrospective cohort study. npj Precision Oncology (2026)
13. Huang, Y., Zhu, T., Zhang, X., Li, W., Zheng, X., Cheng, M., Ji, F., Zhang, L., Yang, C., Wu, Z., et al.: Longitudinal mri-based fusion novel model predicts pathological complete response in breast cancer treated with neoadjuvant chemotherapy: a multicenter, retrospective study. EClinicalMedicine **58** (2023)
14. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
15. Li, W., Huang, Y.H., Zhu, T., Zhang, Y.M., Zheng, X.X., Zhang, T.F., Lin, Y.Y., Wu, Z.Y., Liu, Z.Y., Lin, Y., et al.: Noninvasive artificial intelligence system for early predicting residual cancer burden during neoadjuvant chemotherapy in breast cancer. Annals of surgery **281**(4), 645–654 (2025)
16. Pai, S., Bontempi, D., Hadzic, I., Prudente, V., Sokač, M., Chaunzwa, T.L., Bernatz, S., Hosny, A., Mak, R.H., Birkbak, N.J., et al.: Foundation model for cancer imaging biomarkers. Nature machine intelligence **6**(3), 354–367 (2024)
17. Rahim, N., El-Sappagh, S., Ali, S., Muhammad, K., Del Ser, J., Abuhmed, T.: Prediction of alzheimer's progression based on multimodal deep-learning-based fusion and visual explainability of time-series data. Information Fusion **92**, 363–388 (2023)
18. Richens, J.G., Lee, C.M., Johri, S.: Improving the accuracy of medical diagnosis with causal machine learning. Nature communications **11**(1), 3923 (2020)
19. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
20. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer statistics, 2022. CA: a cancer journal for clinicians **72**(1), 7–33 (2022)
21. Sun, Y., Li, K., Chen, D., Hu, Y., Zhang, S.: Lomia-t: a transformer-based longitudinal medical image analysis framework for predicting treatment response of esophageal cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 426–436. Springer (2024)

22. 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. Journal of clinical oncology **30**(15), 1796–1804 (2012)
23. Wolf, D.M., Yau, C., Wulfkuhle, J., Brown-Swigart, L., Gallagher, R.I., Lee, P.R.E., Zhu, Z., Magbanua, M.J., Sayaman, R., O'Grady, N., et al.: Redefining breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies. Cancer cell **40**(6), 609–623 (2022)
24. Wu, X., Peng, S., Li, J., Zhang, J., Sun, Q., Li, W., Qian, Q., Liu, Y., Guo, Y.: Causal inference in the medical domain: a survey. Applied Intelligence **54**(6), 4911–4934 (2024)
25. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. Nejm Ai **2**(1), AIoa2400640 (2025)
26. Zhang, Y., Hu, C., Zhong, L., Song, Y., Sun, J., Li, M., Dai, L., Zhou, Y., Yang, W.: Spatiotemporal attention for early prediction of hepatocellular carcinoma based on longitudinal ultrasound images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 534–543. Springer (2022)
27. Zheng, X., Aragam, B., Ravikumar, P.K., Xing, E.P.: Dags with no tears: Continuous optimization for structure learning. Advances in neural information processing systems **31** (2018)