



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# Ordinal Learning: Longitudinal Attention Alignment Model for Predicting Time to Future Breast Cancer Events from Mammograms

Xin Wang<sup>1,2</sup>, Tao Tan<sup>3,1\*</sup>, Yuan Gao<sup>1,2</sup>, Eric Marcus<sup>4</sup>, Luyi Han<sup>1,5</sup>, Antonio Portaluri<sup>1,6</sup>, Tianyu Zhang<sup>1,2,5</sup>, Chunyao Lu<sup>1,5</sup>, Xinglong Liang<sup>1,5</sup>, Regina Beets-Tan<sup>1,2</sup>, and Jonas Teuwen<sup>4</sup> Ritse Mann<sup>1,5</sup>

<sup>1</sup> Department of Radiology, Netherlands Cancer Institute (NKI),  
1066 CX, Amsterdam, The Netherlands

<sup>2</sup> GROW School for Oncology and Development Biology, Maastricht University,  
6200 MD, Maastricht, The Netherlands

<sup>3</sup> Faculty of Applied Sciences, Macao Polytechnic University, 999078, Macao, China

<sup>4</sup> AI for Oncology, Netherlands Cancer Institute (NKI),  
1066 CX, Amsterdam, The Netherlands

<sup>5</sup> Department of Radiology and Nuclear Medicine, Radboud University  
Medical Centre, 6525 GA, Nijmegen, The Netherlands

<sup>6</sup> Department of Biomedical Sciences and Morphologic and Functional Imaging,  
AOU G. Martino, University of Messina, 98100, Messina, Italy

\* Corresponding author: [taotanjs@gmail.com](mailto:taotanjs@gmail.com)

**Abstract.** Precision breast cancer (BC) risk assessment is crucial for developing individualized screening and prevention. Despite the promising potential of recent mammogram (MG) based deep learning models in predicting BC risk, they mostly overlook the “time-to-future-event” ordering among patients and exhibit limited explorations into how they track history changes in breast tissue, thereby limiting their clinical application. In this work, we propose a novel method, named OA-BreaCR, to precisely model the ordinal relationship of the time to and between BC events while incorporating longitudinal breast tissue changes in a more explainable manner. We validate our method on public EMBED and inhouse datasets, comparing with existing BC risk prediction and time prediction methods. Our ordinal learning method OA-BreaCR outperforms existing methods in both BC risk and time-to-future-event prediction tasks. Additionally, ordinal heatmap visualizations show the model’s attention over time. Our findings underscore the importance of interpretable and precise risk assessment for enhancing BC screening and prevention efforts. The code will be accessible to the public.

**Keywords:** Breast cancer · Risk prediction · Longitudinal mammogram

## 1 Introduction

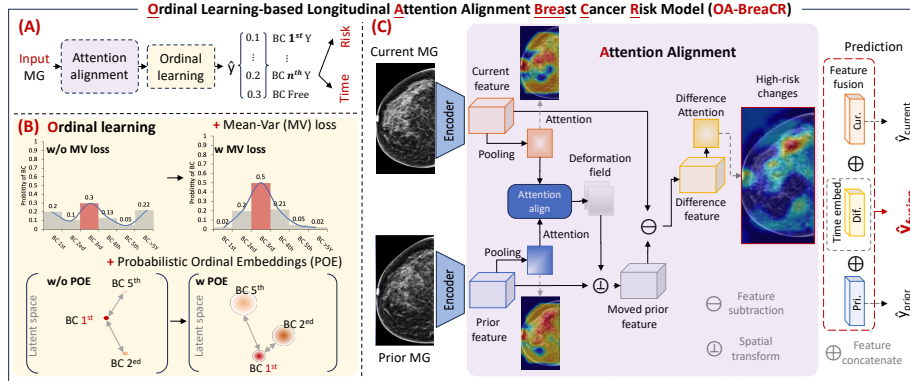
Breast cancer (BC) screening, aiming to detect tumors at an early stage, can efficiently reduce the mortality of BC in women [17]. However, the one-size-fits-all

BC screening policy generates millions of negative mammograms (MG) annually worldwide and requires substantial resources [21]. To optimize BC screening, risk prediction models have emerged as promising tools for stratifying the screening population by guiding high-risk patients toward supplemental imaging or more frequent screenings while recommending longer intervals for low-risk populations to mitigate overdiagnosis [17].

Recently, research has explored artificial intelligence (AI) based risk models to leverage risk-related predictive patterns from MG [17], which achieved superior BC risk stratification compared to only risk factors based traditional risk methods [4, 5, 20]. These methods focus primarily on identifying high-risk patients, where they deal with BC risk prediction as binary classification [2, 12, 22] or as multi-binary classification tasks [10, 18, 21]. Nevertheless, for practical medical applications, precise time-to-future BC prediction is more helpful for physicians in deciding prevention strategies or the timing of subsequent screenings. The existing models, however, cannot accurately predict when breast cancer occurs due to their lack of focus on the time-to-event estimation [19]. The time-to-event estimation task could guide the model to learn the natural ordering of the time to BC among the patients. Therefore, ideally, integrating both risk prediction and time estimation tasks enables not only the precision of time predictions but also augments the model’s capability to identify features indicative of BC development more effectively.

Furthermore, incorporating the historical images is vital for comprehensive learning of both the BC risk-related predictive patterns and time-to-cancer-related ordinal biomarkers. Several efforts have already demonstrated that multi-time-point (MTP) based methods [18, 22, 10, 2] can enhance the accuracy of BC risk assessment compared to single-time-point (STP) based approaches. Despite the advancement, most of them [10, 2] leverage implicit (unsupervised) multi-view attention mechanisms (e.g., transformer-based approaches) to infer information from previous scans indirectly with limited exploration of the explainable spatial-temporal analysis process. Furthermore, these methods may potentially introduce biases unrelated to the actual disease progress observed in multi-time-point images, such as variations in breast compression during imaging sessions. Hence, there still exists a significant need for the explicit monitoring and tracking of temporal variations related to BC risk, which increases not only the robustness of the risk prediction models but also their explainability.

In this paper, we address the limitations described above by introducing a novel ordinal-learning-based risk model that uses longitudinal attention alignment. Our contributions are summarized as follows: *(1)* We propose an ordinal learning framework for concurrently considering both time-to-event BC prediction and risk stratification tasks. *(2)* We incorporate attention alignment mechanisms to explicitly capture risk-related changes from multi-time point MG in an interpretable manner. *(3)* We validate the efficacy and robustness of our approach on two longitudinal MG datasets for future BC risk prediction. Using the ordinal learning and interpretable attention mechanisms, our model offers



**Fig. 1.** The schematic overview of the proposed OA-BreaCR. (A) Our method utilizes ordinal learning and longitudinal attention alignment for the risk and time-to-BC prediction tasks. (B) Ordinal learning for estimating the time to future BC events by leveraging Mean-variance (MV) Loss [15] and probabilistic ordinal embedding (POE) [11]. (C) The attention alignment model aims to learn temporal breast tissue changes from two-time point mammograms in an interpretable manner.

a promising avenue for improving the accuracy and interpretability of BC risk prediction—both vital for personalized screening.

## 2 Methodology

### 2.1 Definition of the risk and time-to-BC prediction tasks

**Risk prediction** This task aims to estimate the probability of future BC within a span of  $n$  years subsequent to a specified MG,  $\mathbf{x} \in \mathbb{R}^{H \times W}$ , with  $n = 5$  in this study. As depicted in Fig 1 A, our proposed OA-BCR model, is designed to predict an  $(n + 1)$ -dimensional vector  $\hat{\mathbf{y}}$  of mutually exclusive outcomes. The initial  $n$  outcomes indicate the probability of BC emergence in each subsequent year up to the  $n^{\text{th}}$  year, while the final outcome denotes the probability of remaining cancer-free over the  $n$ -year timeframe. The BC risk within  $m$  years ( $m \leq n$ ), denoted as  $\text{Risk}_m$ , is defined by  $\text{Risk}_m = \sum_{i=1}^m \hat{\mathbf{y}}_i$ , where  $\hat{\mathbf{y}}_c$  is the probability of BC occurrence in the  $i^{\text{th}}$  year. This approach builds upon methods found in prior research [18].

**Time-to-BC prediction** “If we identify the high-risk population, is it possible to estimate the onset time to future BC?” Inspired by classification-based regression for predicting age [15], we formulate time-to-event prediction as an ordinal regression problem. This is defined by,  $\hat{\mathbf{t}} = \sum_{i=1}^{n+1} i \hat{\mathbf{y}}_i$ , where  $\hat{\mathbf{t}}$  is estimated time to BC onset as determined by the risk model. To enhance the accuracy of the time prediction, we have integrated two regression-task-specific functions aimed at constraining the risk model to learn the ordinal dynamics inherent in the progression toward BC.

For each sample, the corresponding ground truth,  $\mathbf{y}$ , is an  $(n+1)$ -dimensional binary vector. The binary cross-entropy (BCE) loss for the  $i^{\text{th}}$  category of each sample is calculated as  $\ell^{\text{bce}}(i) = -\mathbf{y}(i) \log(\hat{\mathbf{y}}(i)) - (1 - \mathbf{y}(i)) \log(1 - \hat{\mathbf{y}}(i))$ . To accommodate samples with incomplete follow-up data, a masking function,  $\delta(i)$ , is employed to determine the inclusion of a sample at class  $i$  in the loss computation. It is defined as  $\delta(i) = 1$ , if the follow-up period  $\geq \min(i, n)$  or diagnosed within  $n$  years, and 0 otherwise. This method allows for the exclusion of periods with unknown outcomes, concentrating solely on accessible data. Consequently, the overall BCE loss,  $L_{\text{BCE}}$ , is computed as  $L_{\text{BCE}} = \sum_{i=1}^{n+1} \delta(i) \ell^{\text{bce}}(i)$ , ensuring the loss accurately reflects the available follow-up label for each sample.

## 2.2 Ordinal learning for estimating the time to future BC events

**Mean-Variance (MV) Loss** Motivated by the existing ambiguity of the tumor stage when patients are diagnosed, the probability of getting a BC should increase as the diagnostic date nears. Therefore, the predictive model should learn the distribution of cancer occurrence times, aiming for the estimated probability distribution to be close to a normal distribution. As shown in Fig 1 B, the maximally probable time should approximate the actual time ( $\mathbf{t}$ ) of diagnosis; the predicted probabilities should diminish as it moves away from  $\mathbf{t}$ . Thus, we employ the MV loss function [15] to constrain the mean and variance of the predicted probability distribution. The MV loss is presented as  $\mathcal{L}_{\text{MV}} = \|\hat{\mathbf{t}} - \mathbf{t}\|_2^2 + \ell^{\text{var}}(\hat{\mathbf{y}})$ , where  $\ell^{\text{var}}$  is variance loss introduced by [15]. It aims to ensure that the estimated time aligns closely with the actual time of diagnosis constrains the probability distribution as concentrated as possible.

**Probabilistic ordinal embedding (POE)** In addition to constraining the predictive distributions, we also integrate the POE technique [11] to reinforce the ordinal relationships learning within the latent space. As shown in the bottom of Fig. 1 B, this method represents each sample as a multivariate Gaussian distribution rather than a deterministic point in the latent space. Furthermore, it takes into account the ordinal relationships among MGs of women diagnosed with BC at varying years post-scan (e.g., first, second, and fifth year), learning a distribution following the ordinal constraint (e.g.,  $1 < 2 < 5$ ). The POE loss can be defined as  $\mathcal{L}_{\text{POE}} = \ell^{\text{ord}} + \ell^{\text{KL}}$ . The triplet loss ( $\ell^{\text{ord}}$ ) [11] enforces the ordinal relationships within the embedding space, ensuring that the learned representations maintain the correct sequence order. The KL divergence term ( $\ell^{\text{KL}}$ ) is used to constrain probabilistic ordinal embeddings for minimizing the divergence between the predicted and standard normal distributions.

## 2.3 Risk-related attention alignment of prior-current mammograms

To learn both the BC risk-related predictive [10, 18, 20] and time-to-cancer-related ordinal patterns, the model must capture features related to breast tissue change from multi-time-point images. However, the inherent two-dimensional projection principle of mammography introduces challenges in spatial-temporal analysis due to significant deformations between two time points arising from

variations in breast compression during imaging sessions. Image-level registration methods may struggle to handle such large deformations accurately, and incorrect registration may mislead the prediction model [6, 3]. To address this challenge, we propose the attention alignment module based on tissue features.

**Attention alignment** Unlike transformer-based methods which implicitly fuse the prior and current features [10, 18], our method aims to explicitly compare the two images in an explainable way, as shown in Fig 1 C. Utilizing an attention-guided pooling method [14], we obtain attention maps  $\{a^{\text{cur}}, a^{\text{pri}}\}$  corresponding to the  $\{f^{\text{cur}}, f^{\text{pri}}\}$  for which both  $f \in \mathbb{F}^{h \times w \times c}$  derived from the encoder. The attention maps indicate the high-risk areas of the corresponding images. Subsequently, these paired prior-current attention maps are fused within an attention alignment block, which is a convolutional block (two convolutional layers with batch normalization and ReLU activation functions), designed to estimate the deformation field ( $\phi$ ) of them. Inspired by Voxelmorph [1], the alignment process involves registering the prior (moving) feature  $f^{\text{pri}}$  to the current (fixed) feature  $f^{\text{cur}}$  via a spatial transformation function [1], resulting in the moved feature  $\hat{f}^{\text{pri}}$ . Here, we leverage the  $L^2$  loss to constrain the registration processing, denoted as  $\mathcal{L}_{\text{reg}} = \|\hat{f}^{\text{pri}}, f^{\text{cur}}\|_2^2$ . Inspired by [2, 13], our method calculates the differential feature via subtraction of the current feature  $f^{\text{cur}}$  and deformed prior feature  $\hat{f}^{\text{pri}}$ , which already aligned to  $f^{\text{cur}}$ . The difference attention map  $a^{\text{dif}}$  can straightforwardly derived through the attention-guided pooling.

**Multi-level (ML) joint learning** To ensure the consistency of attention maps across images taken at different time points, we employ a multi-level (ML) joint learning strategy. Specifically, current and prior features ( $f^{\text{cur}}, f^{\text{pri}}$ ) are also used to predict the risks associated with each individual time point, followed by the computation of loss based on the labels corresponding to each image. By doing so, the attention mechanism could learn the progress knowledge regarding breast cancer (BC) development. Subsequently, the final predictive modeling leverages the processed features derived from pooled vectors of  $f^{\text{cur}}, \hat{f}^{\text{pri}}$  and  $f^{\text{dif}}$ . Note that, before feature fusion, time gap information between images is integrated into the differential vectors using position embedding techniques [18], enhancing the model’s ability to account for temporal dynamics in risk assessment. This approach facilitates the comparison analysis of two-time point MG, thereby enabling an interpretable examination of temporal breast tissue changes.

### 3 Experimental

#### 3.1 Experimental settings

**Datasets** In this study, two longitudinal datasets were collected. Firstly, the EMBED dataset [9] is a large public screening dataset collected from two cohorts over an 8-year period. We included all images with enough five-year follow-up results according to the released data, including 53,324 images from 6,571 patients. We partitioned the dataset by the patient to create training, validation, and test sets with a ratio of 5:2:3. The test set contains 15,720 images, including

574 cancer diagnosed within 5 years. Secondly, the Inhouse dataset comprises 171,168 images from 9,133 women between 2004-2020, collected from a hospital with IRB approvals. The dataset is randomly split into training, validation, and testing sets at the patient level in an 8:1:1 ratio. The test dataset includes 25,244 images consisting of 1,706 images diagnosed with cancer in five years. Details of time-to-cancer label distribution and dataset split are provided in the supplementary.

**Evaluation** Following previous studies [10, 21], we utilized C-index [16] and the Area Under the Curve (AUC) to evaluate the performances of the methods in the 1-5 year risk prediction task. To evaluate the performance of time-to-event prediction, we used the mean absolute error (MAE) and cumulative score (CS) [11, 15]. The MAE quantifies the average deviation between predicted and actual event times. The CS evaluates the precision within a specified tolerance (e.g., 1 year) of absolute error. Due to the imbalance of the label distribution, we calculated both class-weighted MAE and CS, referred to as WMAE and WCS. Statistical significance is assessed through bootstrapping, with 1000 iterations, to establish confidence intervals as mean  $\pm$  1.96 standard deviations [10].

**Implementation details** All mammograms are preprocessed following the pipeline proposed by [12, 18], including segmenting the breast area, and resizing to 1024 $\times$ 512 pixels of the image size while maintaining aspect ratio. To avoid overfitting, we leveraged augmentations including random flipping, rotation, and affine transformations (e.g., translations, and scalings) for training. We adopted ResNet-18 [7], initialized with ImageNet pre-trained weights, as the backbone of our model. During the training, the Adam optimizer is used, with a learning rate of  $1 \times 10^{-4}$  and a batch size of 96. The learning rate is decayed with a rate of 0.5 and the patience is 5 continuous epochs that the metrics do not improve on the validation set. A total of 200 epochs for training and early stopping based on the validation c-index is utilized with patience of 15 continuous epochs. All deep

**Table 1.** Comparisons with other methods in time-to-event estimation and risk prediction tasks. The methods with **orange color** represent the single time point models, while **blue color** refers to the multi-time point models. The  $\pm$  refers to the 95% CI.

Tasks	Methods	Time to future BC event		Risk prediction						
		WMAE $\downarrow$	WCS $\uparrow$	C-index $\uparrow$	1-Y AUC $\uparrow$	2-Y AUC $\uparrow$	3-Y AUC $\uparrow$	4-Y AUC $\uparrow$	5-Y AUC $\uparrow$	
EMBED Dataset	Risk Prediction (R)	Binary Classification	2.339 $\pm$ 0.010	34.86% $\pm$ 0.42%	0.677 $\pm$ 0.008	0.739 $\pm$ 0.011	0.708 $\pm$ 0.010	0.692 $\pm$ 0.009	0.680 $\pm$ 0.009	0.680 $\pm$ 0.008
		Cumulative hazard	2.405 $\pm$ 0.005	33.59% $\pm$ 0.09%	0.678 $\pm$ 0.008	0.731 $\pm$ 0.011	0.702 $\pm$ 0.010	0.684 $\pm$ 0.009	0.678 $\pm$ 0.009	0.680 $\pm$ 0.008
	Time to event (T)	PRIME+	2.380 $\pm$ 0.009	34.57% $\pm$ 0.41%	0.693 $\pm$ 0.008	0.734 $\pm$ 0.011	0.711 $\pm$ 0.010	0.699 $\pm$ 0.009	0.696 $\pm$ 0.008	0.696 $\pm$ 0.008
		LRP-NE	2.351 $\pm$ 0.008	34.08% $\pm$ 0.35%	0.681 $\pm$ 0.008	0.718 $\pm$ 0.011	0.690 $\pm$ 0.010	0.683 $\pm$ 0.009	0.683 $\pm$ 0.009	0.684 $\pm$ 0.008
		CASurv	1.784 $\pm$ 0.011	48.93% $\pm$ 1.18%	0.656 $\pm$ 0.009	0.683 $\pm$ 0.012	0.669 $\pm$ 0.010	0.662 $\pm$ 0.009	0.657 $\pm$ 0.009	0.659 $\pm$ 0.009
		POE	2.161 $\pm$ 0.016	38.06% $\pm$ 0.74%	0.695 $\pm$ 0.008	0.746 $\pm$ 0.011	0.717 $\pm$ 0.010	0.705 $\pm$ 0.009	0.699 $\pm$ 0.009	0.697 $\pm$ 0.008
		MV loss	1.886 $\pm$ 0.031	43.82% $\pm$ 1.46%	0.696 $\pm$ 0.008	0.738 $\pm$ 0.011	0.722 $\pm$ 0.010	0.709 $\pm$ 0.009	0.699 $\pm$ 0.009	0.699 $\pm$ 0.008
		R & T	Ours (STP O-BCR)	1.865 $\pm$ 0.035	45.30% $\pm$ 1.44%	0.706 $\pm$ 0.008	0.759 $\pm$ 0.011	<b>0.739<math>\pm</math>0.009</b>	<b>0.722<math>\pm</math>0.009</b>	0.711 $\pm$ 0.009
		<b>Ours (OA-BCR)</b>	<b>1.740<math>\pm</math>0.033</b>	<b>51.55%<math>\pm</math>1.52%</b>	<b>0.713<math>\pm</math>0.008</b>	<b>0.768<math>\pm</math>0.011</b>	0.735 $\pm$ 0.010	0.721 $\pm$ 0.009	<b>0.717<math>\pm</math>0.009</b>	<b>0.716<math>\pm</math>0.008</b>
	Inhouse Dataset	Risk Prediction (R)	Binary Classification	1.922 $\pm$ 0.010	42.50% $\pm$ 0.53%	0.748 $\pm$ 0.005	0.851 $\pm$ 0.005	0.813 $\pm$ 0.005	0.777 $\pm$ 0.006	0.748 $\pm$ 0.005
Cumulative hazard			2.244 $\pm$ 0.008	36.15% $\pm$ 0.26%	0.759 $\pm$ 0.005	0.845 $\pm$ 0.006	0.810 $\pm$ 0.006	0.779 $\pm$ 0.006	0.757 $\pm$ 0.005	0.748 $\pm$ 0.005
Time to event (T)		PRIME+	2.177 $\pm$ 0.008	38.05% $\pm$ 0.32%	0.760 $\pm$ 0.005	0.858 $\pm$ 0.005	0.818 $\pm$ 0.005	0.786 $\pm$ 0.005	0.763 $\pm$ 0.005	0.745 $\pm$ 0.005
		LRP-NE	2.007 $\pm$ 0.009	40.42% $\pm$ 0.48%	0.747 $\pm$ 0.005	0.834 $\pm$ 0.006	0.799 $\pm$ 0.005	0.770 $\pm$ 0.005	0.750 $\pm$ 0.005	0.733 $\pm$ 0.005
		CASurv	1.386 $\pm$ 0.007	55.40% $\pm$ 0.47%	0.668 $\pm$ 0.006	0.679 $\pm$ 0.009	0.653 $\pm$ 0.008	0.660 $\pm$ 0.007	0.668 $\pm$ 0.006	0.678 $\pm$ 0.006
		POE	1.847 $\pm$ 0.012	44.96% $\pm$ 0.58%	0.765 $\pm$ 0.005	0.853 $\pm$ 0.006	0.816 $\pm$ 0.006	0.788 $\pm$ 0.006	0.768 $\pm$ 0.005	0.751 $\pm$ 0.005
		MV loss	1.894 $\pm$ 0.014	44.32% $\pm$ 0.58%	0.746 $\pm$ 0.005	0.829 $\pm$ 0.006	0.796 $\pm$ 0.006	0.770 $\pm$ 0.006	0.756 $\pm$ 0.005	0.748 $\pm$ 0.005
		R & T	Ours (STP O-BCR)	1.564 $\pm$ 0.021	56.06% $\pm$ 0.90%	0.760 $\pm$ 0.005	0.842 $\pm$ 0.005	0.814 $\pm$ 0.006	0.793 $\pm$ 0.005	0.777 $\pm$ 0.005
		<b>Ours (OA-BCR)</b>	<b>1.346<math>\pm</math>0.016</b>	<b>60.92%<math>\pm</math>0.80%</b>	<b>0.784<math>\pm</math>0.005</b>	<b>0.879<math>\pm</math>0.005</b>	<b>0.840<math>\pm</math>0.005</b>	<b>0.812<math>\pm</math>0.005</b>	<b>0.795<math>\pm</math>0.005</b>	<b>0.779<math>\pm</math>0.005</b>

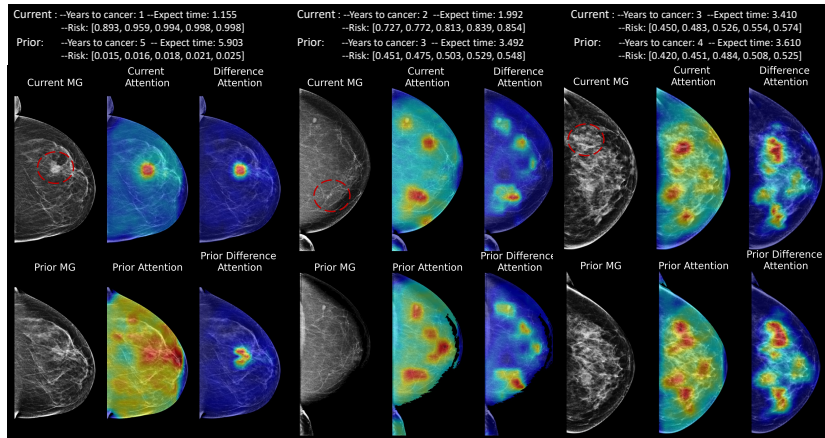
**Table 2.** Ablation results in time-to-event estimation and risk prediction tasks. The methods with **orange color** represent the STP models, while **blue color** refers to the MTP-based methods. The  $\pm$  refers to the 95% CI. MV, Mean-variance loss; POE, Probabilistic ordinal embedding; A2, Attention alignment; and ML, Multi-level.

		Time to future BC event				Risk prediction				
		WMAE↓	WCS↑	C-index↑	1-Y AUC↑	2-Y AUC↑	3-Y AUC↑	4-Y AUC↑	5-Y AUC↑	
EMBED Dataset	Single time baseline	2.265±0.015	36.28%±0.54%	0.666±0.009	0.741±0.011	0.705±0.010	0.681±0.010	0.669±0.009	0.668±0.009	
	√ √ × ×	2.161±0.016	38.06%±0.74%	0.695±0.008	0.746±0.011	0.717±0.010	0.705±0.009	0.699±0.009	0.697±0.008	
	√ √ × ×	1.886±0.031	43.82%±1.46%	0.696±0.008	0.738±0.011	0.722±0.010	0.709±0.009	0.699±0.009	0.699±0.008	
	√ √ × ×	<b>1.865±0.035</b>	<b>45.30%±1.44%</b>	<b>0.706±0.008</b>	<b>0.759±0.011</b>	<b>0.739±0.009</b>	<b>0.722±0.009</b>	<b>0.711±0.009</b>	<b>0.708±0.008</b>	
	Multi-time baseline	2.223±0.017	37.05%±0.96%	0.678±0.009	0.708±0.011	0.679±0.011	0.677±0.009	0.678±0.009	0.681±0.009	
	√ √ × ×	2.358±0.010	34.99%±0.34%	0.679±0.009	0.753±0.012	0.704±0.011	0.687±0.010	0.684±0.009	0.682±0.009	
	√ √ × ×	1.875±0.032	45.76%±1.41%	0.697±0.008	0.737±0.012	0.713±0.010	0.701±0.009	0.699±0.009	0.699±0.008	
	√ √ × ×	1.815±0.032	49.20%±1.46%	0.707±0.008	0.758±0.011	0.727±0.010	0.714±0.009	0.711±0.009	0.709±0.008	
	√ √ × ×	<b>1.740±0.033</b>	<b>51.55%±1.52%</b>	<b>0.713±0.008</b>	<b>0.768±0.011</b>	<b>0.735±0.010</b>	<b>0.721±0.009</b>	<b>0.717±0.009</b>	<b>0.716±0.008</b>	
	Inhouse Dataset	Single time baseline	2.195±0.007	36.34%±0.31%	0.751±0.005	0.820±0.006	0.785±0.006	0.760±0.006	0.746±0.005	0.746±0.005
√ √ × ×		1.847±0.012	44.96%±0.58%	0.765±0.005	0.853±0.006	0.816±0.006	0.788±0.006	0.768±0.005	0.751±0.005	
√ √ × ×		1.894±0.014	44.32%±0.58%	0.746±0.005	0.829±0.006	0.796±0.006	0.770±0.006	0.756±0.005	0.748±0.005	
√ √ × ×		<b>1.564±0.021</b>	<b>56.06%±0.90%</b>	<b>0.760±0.005</b>	<b>0.842±0.005</b>	<b>0.814±0.006</b>	<b>0.793±0.005</b>	<b>0.777±0.005</b>	<b>0.762±0.005</b>	
Multi-time baseline		2.147±0.008	37.28%±0.32%	0.744±0.005	0.839±0.006	0.797±0.006	0.765±0.006	0.749±0.006	0.735±0.005	
√ √ × ×		2.096±0.010	39.35%±0.41%	0.751±0.005	0.864±0.005	0.817±0.005	0.788±0.005	0.772±0.005	0.758±0.005	
√ √ × ×		1.720±0.016	50.28%±0.69%	0.755±0.005	0.847±0.005	0.809±0.005	0.787±0.005	0.777±0.005	0.769±0.005	
√ √ × ×		1.461±0.017	57.51%±0.80%	0.772±0.005	0.872±0.005	0.833±0.005	0.803±0.005	0.782±0.005	0.763±0.005	
√ √ × ×		<b>1.346±0.016</b>	<b>60.92%±0.80%</b>	<b>0.784±0.005</b>	<b>0.879±0.005</b>	<b>0.840±0.005</b>	<b>0.812±0.005</b>	<b>0.795±0.005</b>	<b>0.779±0.005</b>	

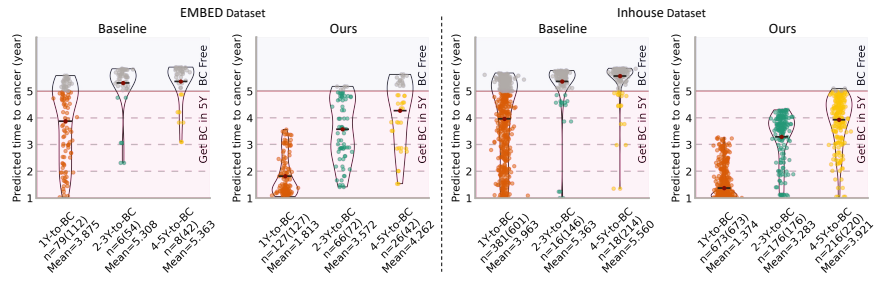
learning models are implemented in the Pytorch framework and an NVIDIA RTX A6000 GPU (48GB), which takes 16-32 hours for training for each model.

### 3.2 Experimental results

**Comparison of risk and time-to-cancer prediction.** The comparisons are shown in Table 1, including the comparison of both STP and MTP methods. Within the STP framework, we compare our method with existing BC risk prediction methods, including binary classification [12], and cumulative haz-



**Fig. 2.** Heatmaps of attention alignment on prior-current MG. The **red** circles indicate the the existing tumors or high-risk areas that developed the cancer within five years.



**Fig. 3.** Violin plots show the distributions of the expected time to cancer in stratified three high-risk populations. All dots indicate true high-risk patients whom the risk model successfully identified. Colored dots represent patients identified by the risk model with an estimated time to cancer within 5 years. While grey dots represent cases where the estimated time-to-cancer exceeds five years, indicating failures to accurately estimate time-to-cancer.  $n$  is the number of color dots, with the total count of color and grey dots provided in parentheses.

ard method proposed by [21]. The time prediction models are compared with CASurv [8], POE [11], and MV [15] methods. For the MTP-based method, we compare our approach with a transformer-based approach (PRIME+) [10] and a pre-registration based RNN based method (LRP-NE) [2]. The results show for the STP setup, our proposed ordinal learning yields an observable improvement on the time prediction with a stable advantage for the risk prediction. More importantly, our proposed method, OA-BreaCR, demonstrates a better performance across both the risk and time prediction tasks, as well as across both evaluated datasets, compared to all other considered methods.

**Visualization of attention alignment.** Fig. 2 displays multiple attention-aligned cases. The attention of the baseline method is globally on the dense breast tissue, consistent with findings of previous research [12]. However, after the attention alignment module, even without pre-registration, the heatmaps show that the attention of the model accurately focuses on the local changes in the breast tissue.

**Ablation study.** To better understand the effectiveness of the proposed module in this study, we also conduct an ablation study, as shown in Table 2. For the SPT-based method, we ablate utilized MV and POE of the proposed ordinal learning method. For the MTP-based method, we ablate both the ordinal learning part and the attention alignment module. The attention alignment module includes attention alignment (A2) and multi-level (ML) learning. The results show that our proposed ordinal learning model improves the precision of the time prediction. The advantage is further improved when cooperating with the prior image through our proposed attention alignment method. The conclusion is consistent on both the public and inhouse datasets.

**Comparison of time-to-future-BC estimation in the detected true high-risk groups.** To further investigate the abilities of the time to cancer



estimation, we plot the distribution of the expected time for the detected high-risk population. Fig 3 shows that our proposed ordinal learning not only could detect more high-risk patients but also could improve the model’s precision for time-to-cancer estimates on both two datasets.

## 4 Conclusion

In this study, we introduce an ordinal learning for explainable multi-time-point BC risk prediction model, OA-BreaCR. Our method incorporates both prior mammograms and current mammograms with an attention alignment module to explicitly capture changes in breast tissue over time. More importantly, ordinal learning could further improve the precision of risk prediction, enhancing the ability to predict time-to-events. The experiments on two large datasets demonstrate the robustness of our method. Our findings highlight the importance of interpretable and precise risk assessment for enhancing BC screening and prevention efforts. The accuracy of the high-risk area detection can be further validated, which may improve the credibility of the risk model in the clinical practice of BC screening.

**Acknowledgments.** Xin Wang is funded by Chinese Scholarship Council scholarship (CSC) and this work was also funded by the Science and Technology Development Fund of Macau SAR (Grant number 0021/2022/AGJ). The authors would like to acknowledge the Research High Performance Computing (RHPC) facility of the Netherlands Cancer Institute (NKI).

**Disclosure of Interests.** The authors declare no competing interests.

## References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging* **38**(8), 1788–1800 (2019)
2. 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)
3. Dong, Q., Du, H., Song, Y., Xu, Y., Liao, J.: Preserving tumor volumes for unsupervised medical image registration. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 21208–21218 (2023)
4. Eriksson, M., Czene, K., Vachon, C., Conant, E.F., Hall, P.: Long-term performance of an image-based short-term risk model for breast cancer. *Journal of Clinical Oncology* **41**(14), 2536–2545 (2023)
5. Gastouniotti, A., Desai, S., Ahluwalia, V.S., Conant, E.F., Kontos, D.: Artificial intelligence in mammographic phenotyping of breast cancer risk: A narrative review. *Breast Cancer Research* **24**(1), 1–12 (2022)
6. Han, L., Tan, T., Zhang, T., Gao, Y., Wang, X., Longo, V., Ventura-Díaz, S., D’Angelo, A., Teuwen, J., Mann, R.: To deform or not: treatment-aware longitudinal registration for breast dce-mri during neoadjuvant chemotherapy via unsupervised keypoints detection. *arXiv preprint arXiv:2401.09336* (2024)

7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Hermoza, R., Maicas, G., Nascimento, J.C., Carneiro, G.: Censor-aware semi-supervised learning for survival time prediction from medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 213–222. Springer (2022)
9. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Illies, G., Smith, G., et al.: The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence* **5**(1), e220047 (2023)
10. Lee, H., Kim, J., Park, E., Kim, M., Kim, T., Kooi, T.: Enhancing breast cancer risk prediction by incorporating prior images. arXiv preprint arXiv:2303.15699 (2023)
11. Li, W., Huang, X., Lu, J., Feng, J., Zhou, J.: Learning probabilistic ordinal embeddings for uncertainty-aware regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13896–13905 (2021)
12. Liu, Y., Azizpour, H., Strand, F., Smith, K.: Decoupling inherent risk and early cancer signs in image-based breast cancer risk models. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. pp. 230–240. Springer (2020)
13. Loizidou, K., Skouroumouni, G., Pitris, C., Nikolaou, C.: Digital subtraction of temporally sequential mammograms for improved detection and classification of microcalcifications. *European radiology experimental* **5**(1), 1–12 (2021)
14. Mikhael, P.G., Wohlwend, J., Yala, A., Karstens, L., Xiang, J., Takigami, A.K., Bourgouin, P.P., Chan, P., Mrah, S., Amayri, W., et al.: Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *Journal of Clinical Oncology* **41**(12), 2191–2200 (2023)
15. Pan, H., Han, H., Shan, S., Chen, X.: Mean-variance loss for deep age estimation from a face. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5285–5294 (2018)
16. Uno, H., Cai, T., Pencina, M.J., D’Agostino, R.B., Wei, L.J.: On the  $c$ -statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine* **30**(10), 1105–1117 (2011)
17. 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)
18. 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)
19. Xiao, L., Yu, J.G., Liu, Z., Ou, J., Deng, S., Yang, Z., Li, Y.: Censoring-aware deep ordinal regression for survival prediction from pathological images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23. pp. 449–458. Springer (2020)
20. Yala, A., Hughes, K.S.: Rethinking risk modeling with machine learning. *Annals of Surgical Oncology* **30**(12), 6950–6952 (2023)
21. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Smith, K., Wan, Y.L., Lamb, L., Hughes, K., Lehman, C., Barzilay, R.: Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine* **13**(578), eaba4373 (2021)

22. Yeoh, H.H., Liew, A., Phan, R., Strand, F., Rahmat, K., Nguyen, T.L., Hopper, J.L., Tan, M.: Radifusion: A multi-radiomics deep learning based breast cancer risk prediction model using sequential mammographic images with image attention and bilateral asymmetry refinement. arXiv preprint arXiv:2304.00257 (2023)