Supplementary Material for Longitudinal Mammogram Risk Prediction

Batuhan K. Karaman, MS[1,2], Katerina Dodelzon, MD[2], Gozde B. Akar, PhD[3], and Mert R. Sabuncu, PhD[1,2]

[1] Cornell University and Cornell Tech, New York, NY 10044, USA
[2] Weill Cornell Medicine, New York, NY 10021, USA
[3] Middle East Technical University, Ankara 06800, Turkey

S.1 Additional information about Karolinska dataset

Refer to Table ST-1.

Table ST-1. Distribution of follow-up times and times until cancer diagnosis for examinations in the Karolinska dataset.

Number of exams with minimum n years of screening followup					Number of exams followed by a cancer diagnosis within n years				
n=1	n=2	n=3	n=4	n=5	n=1	n=2	n=3	n=4	n=5
19328	16148	12873	9578	6530	517	681	1040	1181	1413

S.2 Image Processing

To ready images for the encoder, we resize them to 1664 by 2048 pixels and align them to the left for uniformity. We then normalize them with mean and standard deviation values defined for the image encoder, consistent across training, validation, and test sets. Lastly, we convert single-channel images to pseudo-RGB by replicating them across three channels.

S.3 Details of Image Encoder and Image Aggregator

The CNN and image aggregator employed in our model are identical to those described in Mirai, with the CNN being a ResNet-18 followed by a global max pooling layer to produce 1D image embeddings. At its input, the image aggregator enhances the embeddings by conditioning them on their specific views (CC/MLO) and laterality (left/right) using learned non-parameterized positional embeddings. An affine transformation is applied to each image embedding $\mathbf{x}$ as follows:

$$h = (W_{\text{scale}}\mathbf{e}) \odot \mathbf{x} + (W_{\text{shift}}\mathbf{e}), \tag{1}$$

where $\odot$ is the dot-product, and $\mathbf{e}$ represents the unique 1D positional embedding for each view and laterality, resulting in four distinct instances. The

matrices W_{scale} and W_{shift} are two-dimensional and fixed across all views and lateralities, ensuring uniform scaling and shifting of the embeddings. Following this conditioning, the image aggregator uses a self-attention block to process the embeddings and uses attention pooling implemented with a linear layer followed by softmax activation to produce a singular visit representation.

S.4 Additional Experimental Details

During training, we used Adam with a learning rate of $1e-3$ and implemented dropout with a probability of 0.25 at three points: before, within, and after the visit aggregator. The aggregator itself consists of a single self-attention block. For architectural fine-tuning and learning-related hyperparameter adjustments, we conducted a grid search over three key aspects: the embedding dimension and the number of attention heads within the aggregator, and the L2 regularization applied to the model's weights and biases. The values we explored were {128, 256, 512} for the embedding dimension, {1, 4, 8} for the number of heads, and {1e-4, 1e-5, 1e-6} for the L2 rate.

S.5 Additional Results

Refer to Table ST-2.

Table ST-2. C-index and ROCAUC scores for LoMaR and other existing models, excluding cancers confirmed within 6 months post-screening. Symbols * and † denote evaluations with annual and biennial longitudinal data frequencies, respectively.

Model	History duration	C-index	Follow-up year ROCAUC			
			2-year	3-year	4-year	5-year
Image-Only DL	0	0.67	0.66	0.68	0.66	0.64
Mirai	0	0.71	0.72	0.73	0.73	0.71
LoMaR (Ours)	0	0.70	0.73	0.73	0.73	0.72
LoMaR (Ours)	$1^\star$	0.71	0.73	0.74	0.73	0.72
LoMaR (Ours)	$2^\star$	0.71	0.73	0.73	0.73	0.75
LoMaR (Ours)	$3^\star$	0.72	0.73	0.73	0.75	0.79
LoMaR (Ours)	$4^\star$	0.74	0.74	0.75	0.79	0.81
LoMaR (Ours)	$2^\dagger$	0.71	0.73	0.73	0.73	0.75
LoMaR (Ours)	$4^\dagger$	0.73	0.73	0.75	0.76	0.78