Predicting up to 10 year breast cancer risk using longitudinal mammographic screening history

Authors

Xin Wang,[1,2,3] Tao Tan,[1,4*] Yuan Gao,[1,2,3] Ruisheng Su,[5] Tianyu Zhang,[1,2,3] Luyi Han,[1,3] Jonas Teuwen,[6] Anna D'Angelo,[1,7] Caroline A. Drukker,[8] Marjanka K. Schmidt,[9,10] Regina Beets-Tan,[1,2] Nico Karssemeijer,[11] Ritse Mann[1,3]

Affiliations

[1] Department of Radiology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.

[2] GROW School, Maastricht University, Maastricht, 6202 AZ, The Netherlands.

[3] Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Nijmegen, 6525 GA, The Netherlands.

[4] Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, 999078, China.

[5] Erasmus Medical Center, Erasmus University, Rotterdam, 3015 GD, The Netherlands.

[6] Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.

[7] Dipartimento di Diagnostica per Immagini, Radioterapia Oncologica ed Ematologia, Fondazione Policlinico Universitario A. Gemelli, IRCCS, 00168 Rome, Italy.

[8] Department of Surgical Oncology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.

[9] Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.

[10] Department of Epidemiology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.

[11] Department of Medical Imaging, Radboud University Medical Center, Nijmegen, 6500 HB, The Netherlands

*Corresponding author

Abstract

Risk assessment of breast cancer (BC) seeks to enhance individualized screening and prevention strategies. BC risk informs healthy individuals of the short- and long-term likelihood of cancer development, also enabling detection of existing BC. Recent mammographic-based deep learning (DL) risk models outperform traditional risk factor-based models and achieve state-of-the-art (SOTA) at short-term risk prediction, but mainly use single-time exams, which seem to rely more on detecting existing lesions. We present a novel temporospatial and explainable deep learning risk model, the Multi-Time

Point Breast Cancer Risk Model (MTP-BCR), which learns from longitudinal mammography data to identify subtle changes in breast tissue that may signal future malignancy. Utilizing a large in-house dataset of 171,168 screening mammograms from 42,792 consecutive exams involving 9,133 women, our model demonstrates a significant improvement in long-term (10-year) risk prediction with an area under the receiver operating characteristics (AUC) of 0.80, outperforming the traditional BCSC 10-year risk model and other SOTA methods at 5-year AUC in various screening cohorts. Furthermore, MTP-BCR provides unilateral breast-level predictions, achieving AUCs up to 0.81 and 0.77 for 5-year risk and 10-year risk assessments, respectively. The heatmaps derived from our model may help clinicians better understand the progression from normal tissue to cancerous growth, enhancing interpretability in breast cancer risk assessment.

Teaser

MTP-BCR model uses multi-time points mammograms and rich risk factors to predict 10-year breast cancer risk more accurately.

MAIN TEXT

Introduction

Breast cancer (BC) is one of the most common cancers in the world and is the cause of a large fraction of cancer-related mortality among women *(1, 2)*. Studies have shown that age-based population-level BC screening programs, which aim to detect breast tumors at an early stage *(3)*, reduces breast cancer specific mortality *(4–7)*. However, the broad adoption of mammographic screening results also in high cost of imaging, false-positives and over-diagnoses, which explains the strong controversy of screening *(8, 9)*. Therefore, "personalized" BC screening regimens are advocated, based on the individual women's future risk of BC, which follows from demographic and genetic information, exposure to endogenous and exogenous risk factors, and also medical imaging *(10–12)*. Current BC risk assessment models are designed to be sensitive to the high-risk population who could benefit from more aggressive screening and prevention. At the same time these models could advocate less frequent screening for the low-risk population to reduce the harm and cost of screening, however this is less common.

Based upon the timespan for breast cancer prediction, risk models can be divided into short- and long-term risk models. Short-term risk models can be used to guide physicians in selecting and adding supplemental screening modalities for women at the time of screening. Long-term risk prediction helps determine risk-based screening regimens and eligibility for preventive treatment *(13)*. Many of the traditional risk models, such as Tyrer-Cuzick *(11)*, CANRISK *(14)*, National Cancer Institute Breast Cancer Risk Assessment Tool (BCRAT) *(12)*, and Breast Cancer Surveillance Consortium (BCSC) *(15)* investigate primarily long-term risk estimates. The performances of these risk models remain modest in clinical practice and are not very sensitive to short/middle-term cancer risk variation due to the shortage of individual-specific risk adaptation, for example through the incorporation of detailed imaging findings beyond breast density only. With the recent boost in deep learning (DL) methods, some studies that combined large screening mammography datasets with detailed risk factors have shown considerable promise to help balance the harm-to-benefit ratios of BC screening *(7, 16–22)* and were

even validated in clinical settings *(23)*. For example, a recent study *(19)* developed a risk model, MIRAI, that achieved state-of-the-art (SOTA) performance in five-year BC risk prediction and outperformed the clinically adopted traditional models *(19, 23–25)*.

However, most recent methods drive DL models to learn the risk output directly from an image or exam as single input without any historical reference *(7, 16, 19)*. It's like judging the motion trajectory of an object in a still frame of a video. We hypothesize that accurate estimation of the breast tissue changes may make the task of predicting long-term BC risk easier. In clinical practice, radiologists routinely compare mammography exams to identify developing abnormalities. Therefore, beyond learning risk features (e.g. breast density) from single-time point imaging, multi-time point learning may also be helpful in discovering the underlying dynamics of the risk pattern for BC development *(20, 26)*. Moreover, due to the lack of a long-term longitudinal screening mammogram dataset, the potential of image-based DL methods for longer-term (e.g., 10 years) BC risk prediction has been less explored. Only one recent research investigated the long-term performance of an image-based short-term risk model *(27)*.

Despite the development of promising risk models in BC screening programs, the interpretability of medical AI models is still difficult, whereas understanding the predicted outputs is essential for clinical acceptance. How to endow an existing risk model with explainability of the underlying reasoning remains the boundary to explore. Apart from being similar to what an actual radiologist does when searching for the sign of BC risk *(28, 29)*, the AI models must reasonably show radiologists more details during inference *(30)* for clinical acceptance. However, most recent studies only aim to predict the patient-level risk and do not produce a location-specific risk. Improvement of these specific predictions and visualizations could not only improve the interpretability of the model and make it easier for physicians to understand the model's decision-making but also inform doctors where to focus and then guide them when deciding on the most suitable targeted examinations and prevention strategies. An ideal risk model should therefore not only stratify high-risk groups but also focus the doctors attention to changing areas in the breast earlier.

We propose the Multi-Time-Points Breast Cancer Risk model (MTP-BCR), an end-to-end model that estimates the long-term future BC risk based on changes in breast tissue. Our contributions are as follows. First, our model leverages historical and current exams from a large in-house clinical mammogram dataset and obtains remarkable performance compared to other SOTA methods on patient-level BC risk prediction. Second, we explore and show that our image-based DL risk model outperforms clinical traditional BCSC risk models for long-term 10-year risk prediction. Third, we explore the unilateral breast level BC risk prediction and achieve similar performance to our risk models at the patient level. Fourth, we highlight suspicious areas in a longitudinal test dataset using the model’s heatmaps, which may illustrate the attention consistency of our model and improve its interpretability. Fifth, to demonstrate the robustness of our method in clinical settings, we perform a systematic subgroup analysis. The results imply that our model may improve upon traditional and other image-based DL risk models.

Results

Overview of algorithm

For investigating our hypothesis that the breast tissue changes can help in learning the tumor (including both invasive and ductal carcinoma in situ) development pattern better,

multiple time points of examinations from the longitudinal screening mammograms are required. Like radiologists, who typically identify developing abnormalities by looking at changes in longitudinal exams, we propose a novel end-to-end multi-time point network, MTP-BCR, shown in Fig. 1A, leveraging longitudinal mammograms and medical records to capture the features related to increased BC risk. We aim to predict the risks on a patient-based level as well as for a single breast. Briefly, the proposed risk model first utilizes the multi-level (breast and patient level risk) and multi-task learning to extract static risk features from a single time point exam and prior medical records. Then the features of five historic exams, obtained before the current exam are selected (to mimic the practical use scenario) for learning the dynamic risk features using a multi-time point fusion model. This is combined with the risk factors of patients for predicting the future risk. It means that our end-to-end model uses current and historic screening mammography exams and existing medical records and then predicts the future 10-year BC risk. Architectural details, contents of the medical records, and risk factors are presented in the method part.

Risk calculation can be treated as a multi-class classification problem *(16, 19)*, which is common in breast imaging, such as the classification of breast density *(31)*, the Breast Imaging Reporting and Data System (BI-RADS *(32)*) score *(33)*, the type of malignancy *(3)*, and the BC molecular subtype *(34)*. As shown in Fig. 1B, the risk of a patient getting BC within 5 or 10 years from the available data can be calculated as the cumulative sum of the probability from the first year up to the fifth or tenth year. Importantly, the prediction results of our model can guarantee that the risk is monotonically increasing and self-consistent. This avoids the situation, that can occur with separately trained models, where long-term risk may be lower than short-term risk. Moreover, this formulation also learns the inherent relationship between risks at different time points. In this study, the model is trained to predict the risk of BC at each of the 15 years and is validated by predicting BC within ten years. Therefore, our model can be easily extended for a longer than 15-year risk prediction when collecting enough longer-term follow-up data.

Screening cohorts for risk modeling & compared candidates

Fig. 1 shows the flowchart of the Screening Cohort selection in the inhouse dataset. The dataset contains 42,792 exams of 9,133 patients, split into 32,049 exams / 6,858 patients, 4,432 exams / 919 patients, and 6,311 exams / 1,356 for the train set, validation set, and test set, respectively. The proposed model aims to handle multiple tasks, including cancer detection and future risk prediction. This would facilitate implementation in an actual clinical BC screening program, where not only focusing on the stratification of the high-risk population, short-term risk and determining whether women should be recalled is also essential. In fact, detection of existing malignancies, can be considered as an extremely short-term BC risk *(35)*. Moreover, some of the model characteristics for the tasks of cancer detection and future risk prediction can be complementary *(19)*.

To evaluate the model's capability of longer term BC risk prediction, we use two standard screening test sets following the protocols of *(25)* and *(19)*. The first one is named as the biopsy-negative screening group (5,937 exams / 1,236 patients), which includes mammography exams with BI-RADS 1 and 2 scores, or other BI-RADS scores but with benign biopsy results within 90 days from the screening date. The second test set is called the normal BI-RADS screening group (5,139 exams / 1,157 patients), which only consists of the exams scored as BI-RADS 0, 1 and 2. It aims to explore how the model performs on high-risk population stratification when radiologists deem the exams not suspicious.

The inhouse mammogram dataset is coupled to patient derived classical risk factors that can be used in the clinical Breast Cancer Surveillance Consortium version 2 (BCSC, https://tools.bcsc-scc.org/BC5yearRisk/). The distribution of clinical risk factors is shown in Supplemental Table S1. The BCSC model can estimate five-year, and ten-year BC risk based on risk factors but requires excluding patients following exclusion criteria (previous diagnosis of BC, younger than age 35 or older than age 74, or missing density estimates). Although studies have shown that image-based DL risk models outperform traditional risk models in 5-year risk assessments *(19, 24, 25)*, the potential advantages of the former still need to be explored in longer-term 10-year risk assessments. Thus, we compare our model with not only the traditional BCSC 1-year and 5-year risk models but also with the BCSC predicted 10-year risk. To demonstrate the added value of the inclusion of patient based risk factors , we also build a similar multi-time point model without risk factors as MTP-BCR for comparison.

To demonstrate the risk prediction capacity, we compare it to the SOTA MIRAI (Massachusetts Institute of Technology, Boston, Massachusetts) model *(19)*. MIRAI is a mammogram-based risk model that can predict 5-year risk at multiple time points and outperforms traditional models. Moreover, this model includes a pretrained risk factor predictor that allows missing risk factors. In this study, we also explore MIRAI's ability of the longer-term 10-year risk estimate. We obtain the pretrained MIRAI model from their public GitHub (https://github.com/yala/Mirai). For a competitive comparison, we finetune the MIRAI model on the inhouse training dataset to alleviate the impact of domain shift. We conduct hyperparameter search to finetune MIRAI and select the model with the best concordance index (C-index) on the validate set. Similar to the research *(19)*, to investigate the performance of our method on BC detection, we also compare our model with the retrospective radiologist BI-RADS scores and the Globally Aware Multiple Instance (GMIC, New York University, New York) model *(29)*. The GMIC is another recent SOTA DL model which focuses on detecting BC within three months, and some researches also show its potential for BC risk prediction *(19,25)*. The pretrained GMIC model is obtained from the public GitHub repository (https://www.github.com/nyukat/gmic). For fairness, we re-implement the prepossessing from the raw DICOM format mammograms through their preset prepossessing pipeline and collect the ensembled predictions from the five pretrained models.

Note that, for full leverage of the mammogram examinations, we include all exams with at least one-year screening follow-up. To fairly compare five-year risk prediction with other SOTA methods and prove the contribution of our algorithm design of longitudinal input and multi-task learning, we also re-form the inhouse five-year risk dataset by excluding the examinations without at least five-year screening follow-up. Then we train our model from scratch on five-year risk prediction using the inhouse five-year risk dataset and we reperform all experiments (shown in the Supplemental Section).

Risk prediction on full inhouse test dataset

All concordance index (C-index) and Area Under the Receiver Operating Characteristics (AUC) results on the inhouse test dataset (6,311 examinations / 869 positives within 5 years / 1,132 positives within 10 years) are summarized in Table 2.

Our method's performance: The performances of the MTP-BCR model with risk factors and without risk factors at patient level risk prediction (as shown in Table 2, and ROCs are shown in Fig. S3) obtained 10-year C-indices of 0.82 (95% CI, 0.81-0.84) and 0.77 (95%

CI, 0.75-0.78), with AUCs of 0.91 (95% CI, 0.89-0.92) and 0.87 (95% CI, 0.85-0.89) at easiest 1-year risk prediction and AUCs at the most difficult 10-year risk prediction of 0.80 (95% CI, 0.78-0.82) and 0.77 (95% CI, 0.75-0.79), respectively. The AUCs results of 1- to 10-year risk prediction show that the performances of the MTP-BCR with risk factors are significantly higher than those of MTP-BCR without risk factors (All *P* values < 0.05). We also evaluate the unilateral breast level cancer prediction of MTP-BCR models both with risk factors and without risk factors, as shown in Fig. S3. Similar performances of patient-level risk prediction are obtained, with the 10-year C-indices of 0.81 (95% CI, 0.79-0.82) and 0.76 (95% CI, 0.75-0.78) for our methods with and without risk factors, and 5-year C-indices of 0.82 (95% CI, 0.81-0.84) and 0.78 (95% CI, 0.76-0.80). The AUCs of unilateral breast-level cancer prediction without risk factors ranged from 0.76 to 0.87. When incorporating patient derived risk factors, the AUCs ranged from 0.77 to 0.89. Therefore, our MTP-BCR risk model can also accurately predict the risk of development of BC in a unilateral breast, and using risk factor information can further improve the performance of 1- to 10-year risk.

Moreover, we perform two ablation studies to choose the best design of the MTP-BCR model (Table S4 and S5). The first one is to investigate whether our multi-task and multi-level, and multi-time point learning strategies can improve the ability to extract risk-related features. The C-indices and AUCs show that the model with multi-task, multi-level, and multi-time-point learning is better than others alone. Using risk factors can further improve the performance of the risk models. Besides, the model achieves the best performance when using five time-point historic mammogram references.

Comparing with other methods (except BCSC): The 1-year AUCs of Radiologists BI-RADS assessments and the BC detection method GMIC are 0.83 (95% CI, 0.81-0.85) and 0.74 (95% CI, 0.72-0.77) respectively, which are significantly lower than both MTP-BCR risk models ($P < 0.001$). Therefore our MTP-BCR risk model outperforms radiologists and the SOTA BC detection model for BC detection and extremely short-term risk prediction, even when only using mammograms. Comparing to the SOTA DL-based MIRAI model, the 5-year C-indices of the MTP-BCR models (with/without risk factor) are 0.82 (95% CI, 0.81-0.84) and 0.77 (95% CI, 0.75-0.78) versus 0.73 (95% CI, 0.72-0.75). The AUC results show that both the MTP-BCR models significantly outperform the MIRAI model at all time points from 1-year to 5-year risk (All *P* values < 0.01). Also for 10-year risk prediction, the MTP-BCR methods still have a significant advantage (All *P* values < 0.05).

Comparing with BCSC risk model: Note that our study also include patients who are not eligible for risk calculation by the BCSC model as they are either out of the required age range of 35-74 or had a prior BC history. For this comparison, we conduct the experiments excluding the women that did not have scores from the BCSC model (Table S1). The AUC curves of all methods are shown in Fig. 3A. The AUC results show that both the MTP-BCR models significantly outperform the 1-year, 5-year, and 10-year BCSC risk models (All *P* values < 0.001). The latter obtains AUCs of 0.62 (95% CI, 0.58-0.64), 0.65 (95% CI, 0.62-0.68), and 0.71 (95% CI, 0.68-0.74), respectively. The 5-year and 10-year C-indices of BCSC models are 0.63 (95% CI, 0.61-0.65) and 0.64 (95% CI, 0.61-0.66), compared to C-indices of 0.91 (95% CI, 0.90-0.92) and 0.90 (95% CI, 0.88-0.91) by our MTP-BCR risk model with risk factors. Moreover, we also compare our risk models with the other methods in this specific population. The AUC results show that both the MTP-

BCR models significantly outperform all other models for 1-year to 10-year risk prediction (All P values < 0.001).

Performing experiments of 5-year risk prediction: Similar results are obtained for reperformed experiments on the formed 5-year risk prediction dataset (Table S6 and S7). Both the MTP-BCR models with and without risk factors are significantly better than other methods. Moreover, we also perform two same ablation studies on the inhouse 5-year risk dataset (Table S12 and S13). Similar results to those of 10-year risk ablation experiments also demonstrate the stability of our learning strategy.

Risk prediction in a healthy screening population

In biopsy negative screening population: we evaluate all methods using the inhouse negative biopsy screening test set including 5,937 examinations / 495 positives within 5 years / 758 positives within 10 years, as shown in Table 3. In this test set, results show that our MTP-BCR with risk factors holds superiority in C-index and AUC metrics. Extremely short-term (1 year) risk prediction can be equivalent to interval cancer detection. In this task, the AUC of our model without risk factors is 0.70 (95% CI, 0.65-0.74). The AUC increases to 0.77 (95% CI, 0.73-0.81) when using risk factors, significantly higher than that from BI-RADS scores with 0.61 (95% CI, 0.57-0.65), the BC detection method GMIC with 0.59 (95% CI, 0.54-0.65), and also the finetuned MIRAI with 0.65 (95% CI, 0.60-0.70). In Fig. 3B and Table S2, the clinical BCSC 1-year risk model obtains an AUC of 0.70 (95% CI, 0.63-0.76), which is significantly lower than the AUCs of our MTP-BCR models (with risk factor: 0.87 / without risk factor: 0.82). For the long-term 5-year risk prediction, our model reaches a C-index of 0.65 (95% CI, 0.63-0.68) without risk factors and a C-index of 0.74 (95% CI, 0.72-0.76) with risk factors, versus 0.64 (95% CI, 0.62-0.66) for the finetuned MIRAI model (Table 3). When comparing with the clinical BCSC 5-year risk model (as shown in Table S1), our MTP-BCR obtains a C-index of 0.79 (95% CI, 0.76-0.82) versus 0.69 (95% CI, 0.65-0.72). For the longer-term 10-year risk estimate, our only image based MTP-BCR model has a similar performance to 10-year risk BCSC model according by the 10-year C-index. But the results indicate that the MTP-BCR with risk factors model still significantly outperforms all other models by AUCs at each time point (All P values < 0.05).

In normal BI-RADS screening population: In the normal BI-RADS screening test set with 5,139 examinations / 404 positives within 5 years / 612 positives within 10 years, we investigate the potential of risk models when radiologists cannot find any suspicious findings on the mammograms. Thus, the C-indices and AUCs of the BI-RADS are close to 0.5, as shown in Table 3 and Table S2. Our MTP-BCR with risk factors is still significantly better than all other methods among the extremely short-term with a 1-year AUC of 0.74 (95% CI, 0.69-0.79), long-term with a 5-year AUC of 0.72 (95% CI, 0.70-0.75), and longer-term with 10-year AUC of 0.73 (95% CI, 0.71-0.76) risk predictions. Especially in the BCSC model target population (aged 35-74, without prior BC history), the 1-year AUC of our MTP-BCR risk-based model reaches 0.85 (95% CI, 0.78-0.91) while the radiologists and the cancer detection-based models fail to outperform random guessing.

Consistent with the findings on the full inhouse test set, AUC and C-index metrics show that our MTP-BCR model performs similarly for unilateral BC risk prediction and patient-level cancer risk prediction on both screening sets. We note that using the full training dataset to finetune MIRAI lead to poor performance of MIRAI on these two screening sets

(shown in Table S8 and S9, Replicate 5-Year BC Risk Prediction). Thus, we clean the training and validation sets using the same settings as for the two screening sets, and then re-finetune the MIRAI model and test it on the corresponding test set. Despite that, we find that the finetuned MIRAI model on the dataset with missing screening follow-up seems to be difficult. Specifically, the performances of the finetuned MIRAI from 3- to 10-year BC risk prediction become worse (shown in Fig. 3 B and C). Thus, for a meaningful comparison, it is necessary to re-implement these comparison experiments on the two recollected five-year screening datasets which cleaned exams with missing follow-up labels. As shown in Table S8 and S9, the AUCs and C-indices of MIRAI, after finetuning, reaches its own optimal, and outperforms the 5-year BCSC model. The finetuned MIRAI only achieves similar performance to our image-only MTP-BCR model from 3- to 5-year risk at both screening test sets (when only for women who can be scored by the BCSC model) (Table S9). Yet our method with risk factors still surpasses the MIRAI.

The ability of short and long future BC risk assessment

While the above results demonstrate the advantages of our methods in risk prediction, the ability to predict real long-term future BC risk after eliminating the biases of cancer detection and short-term to 5-year risk prediction from current mammograms requires further exploration. Thus, we compare the models' performance in 5-year and 10-year risk prediction in the different subgroups of the full inhouse test sets by excluding exams from women diagnosed with cancer within less than 1, 3, and 5 years. These results, as shown in Table 4, demonstrate that our methods could not only detect BC and improve the performance of 5-year risk prediction compared with other SOTA methods but also learn the features related to the real longer-term (10 years) risk. The results of replicate 5-year BC risk prediction are shown in Table S10.

Clinical sub-group analysis

To distinguish how our MTP-BCR model performs in different populations and to determine the potential population that can benefit most from it, we evaluate our risk model in different clinical subgroup, based on age, breast density, personal history, and future cancer sub-types (Fig. 4). We find that the MTP-BCR model performed similarly across different density groups and independent from future cancer sub-types. The C-indices for the MTP-BCR model for women aged <40, 40-60, and 60-80 are 0.87 (95% CI, 0.84-0.90), 0.81 (95% CI, 0.79-0.83) and 0.82 (95% CI, 0.79-0.84), respectively, which implies that our risk model performs better in younger cohorts. We also note that the risk model performs best in the female population without any prior personal BC history with C-index of 0.86 (95% CI, 0.84-0.87), compared to 0.76 (95% CI, 0.74-0.78), for women with a personal history of breast cancer. We also compare the C-index of 10-year risk prediction of different methods in different sub-groups, which are available in Table S3. Our results are further supported by the consistent performance from additional subgroup analyses in recollected 5-year risk dataset (Tables S11, supplemental). Due to a lack of race labels, the group analysis for different race groups is not performed.

Consistency of model attention in longitudinal images

To investigate how risk-related areas evolve on multiple-time mammograms that our MTP-BCR model focused on, we utilize the gradient-weighted class activation maps (Grad-CAM) *(36)*. Fig. 5 shows a visualization example of a BC patient. The heatmaps highlight potentially related regions where our proposed MTP-BCR model identifies predictive imaging features for BC risk. While this visualization is a preliminary process,

results show that the high-risk regions from multiple time point examinations that our model focuses on are relatively consistent. Moreover, the heatmaps show that our risk model could accurately figure out high-risk areas of short-term BC at both the craniocaudal (CC) and the mediolateral oblique (MLO) views of mammograms. The similar result is observed from the heatmaps of the retrained model on the inhouse five-year risk dataset (Fig. S5, supplemental).

Discussion

In this study, we develop a multi-time-point examination-based risk model, MTP-BCR, to assess 10-year breast cancer risk on patient and unilateral breast level using longitudinal screening mammograms and medical records. From extremely short-term risk (1-year, BC detection) to long-term (10-year) risk prediction, MTP-BCR outperforms radiologists' BI-RADS assessment, a SOTA BC detection method (GMIC), an extensively clinically validated single time point-based 5-year risk DL approach (MIRAI), and a traditional clinical risk model (BCSC). Apart from the patient level risk, our method could also estimate the risk base on unilateral breast level with comparable ability. Experiments on different screening subcohorts suggest that the longitudinal assessment of MTP-BCR is able to accurately identify longer-term, future risk-related features on mammograms, which is further supported by consistent heatmaps of multi-time point mammograms. Final sub-group analysis indicates that the proposed method performs consistently across subgroups of different breast densities and for future types of BC.

The motivation to develop BC risk models is for guiding personalized screening or triggering prevention regimens. The idea is to determine the screening frequency and the appropriate screening modality based on the individual risk of women, potentially also to recommend preventive therapy for women at a high-risk of developing breast cancer *(13).* Based on of risk factors such as age, genetic determinants, family history, previous benign biopsies, and recently considered image-based breast density, traditional risk models are used to globally assess 5-year, 10-year, or lifetime risk for large groups of women *(37).* However, our results show that density alone is not sufficient to represent all of the risk-related information within the mammograms. Moreover, ignoring the short-term risk of BC limits the value of these models in early BC detection. Recent DL-based risk models may fully utilize screening images but mainly target short-term risk prediction or interval cancer detection while ignoring long-term risk, which limits the ability to offer personalized screening recommendations and preventive interventions. In contrast our MTP-BCR risk model combines the advantages of both short- and long-term risk prediction strategies.

For short-term risk estimation, our risk model outperforms radiologists' BI-RADS and other recent DL methods. The results on the full test set demonstrate that our MTP-BCR risk model is more competitive than other risk models in indicating current or future BC risk in a realistic complex clinical screening setting independent of radiological interpretation. Other methods, especially traditional risk tools, can only work with modest performance to estimate women's future risk on the premise that the physician confirm that there is no existing cancer. Furthermore, in the normal BI-RADS screening group, our risk model has the highest AUC of 0.74 for the 1-year risk prediction, which can be regarded as an aid for radiologists to improve interval cancer detection in the whole screening population, including primary and recurring cancers. When detecting the primary interval cancers of the population that the BCSC model targets, our risk model reaches an AUC of 0.87. Aiming to facilitate radiologists' understanding of the model

decision-making, our risk model could also estimate the unilateral-breast level risk with similar performance to the patient-level risk prediction. Thus, our retrospective analysis and visualizations indicate that the MTP-BCR risk model could improve early detection and reduce interval cancer by guiding radiologists to identify high-risk regions on images.

For 5-year risk prediction, our methods surpass the SOTA MIRAI risk method across all screening cohorts or subcohorts. Although the MIRAI model does not need to have the risk factors as inputs and performs similarly to our image-only MTP-BCR model at 5-year risk ($P > 0.05$) on the biopsy negative screening set and normal BI-RADS set, we note that the MIRAI model involves a pretrained risk factors predictor enabling it to benefit from the missing risk factors that reaches a similar performance of 5-year risk prediction as MIRAI with risk factors ($P = 0.27$) *(19)*. While our MTP-BCR model with risk factors is significantly better than all others (All P values < 0.05). Besides, the MIRAI is already trained on their large private MGH dataset. The training set includes 210,819 exams is 6.5 times larger than our Inhouse training set. Then we still fully finetuned the MIRAI with its weight in our In-house training set without freezing the encoder. On the other hand, our MTP-BCR model is only trained on the In-house dataset directly. As for long-term risk, the MTP-BCR risk model is more accurate than the BCSC 10-year risk model, which suggests our risk model has the potential for better decisions regarding a risk-adapted screening regimen and preventive therapy. At breast level, risk could be the foundation of more refined screening and prevention strategies. It should be underlined that in this study we also involve patients with prior BC history, which are not included in most of the risk models. Our risk model is, in fact, also designed to leverage the history information of prior-tumor and therapy, which the recurrence risk models use *(38)*, for recurrence cancer risk prediction.

The promising performance of MTP-BCR can be attributed to its capacity to capture unique BC risk-related characteristics. The multi-task learning strategy helps the model to fully extract risk-related features from images and also improves the generalization of the DL model *(39)*. The multi-level learning strategy enables our model to learn the relationship between local (unilateral breast) level risk and global (patient) level risk while keeping the local information as much as possible when combining the local features for the summary of the global features. Thus, the MTP-BCR risk model can consistently focus on similar regions on longitudinal mammograms without registration. For accurate longer-term BC risk prediction, we need not only the static risk-related features from the single time point exam but also the dynamic features from the multi-time point exams to indicate the development risk of BC. Recently, a study *(20)* also explored the potential of using longitudinal mammogram examinations to improve short-term risk prediction. But we note that it has been restricted to a small image-only dataset and a setting of a fixed number of input time points, which hinders its clinical application. On the contrary, our MTP-BCR model is built based on an extensive clinical screening mammogram dataset, enabling efficient use of risk factors, and has the flexibility to input 0-5 historical reference exams.

This research has limitations. Although we leverage the prior tumor information and therapy records (as explained in Section Methods) to improve our risk model's performance on BC recurrence risk, big gaps exist between the subgroups with/without prior BC. More efforts are needed to improve the performance of recurrence risk prediction. Further validation of our model is required before it can be broadly implemented in clinical practice. For instance, more detailed demographics (e.g., race) are

required to prove its generalizability. We note that the breast cancer incidence in our screening set is higher than that of the standard screening dataset. Therefore, in the future, we will also explore the clinical potential of our MTP-BCR model based on external standard screening mammography datasets from multiple hospitals. To validate the detection of extremely early signs of BC, a pixel-level annotated dataset is necessary. Moreover, a reader study for incorporation of the risk model in the radiologist workflow may also be a future direction for further demonstrating the benefits of risk models for personalization BC screening policy.

In conclusion, we propose a novel DL model using longitudinal mammogram examinations and history obtained from medical records that outperforms the SOTA MIRAI and traditional BCSC risk model by a large margin. The improvement is consistent across screening and future risk subgroups. These results support the hypothesis that longitudinal mammography contains informative spatiotemporal indicators of future breast risk that cannot be captured by the single-time point DL models. Multi-time point models based on longitudinal analysis strategies have the potential to replace single-time point based risk prediction models. Apart from increasing the accuracy for BC risk prediction, we also improve the interpretability of our risk model, which could potentially accelerate the translation of personalized AI-based risk stratification into routine BC screening policies.

Materials and Methods

Data collection

Our retrospective study was approved by the Institutional Review Board (IRB) of Netherlands Cancer Institute (NKI) with protocol numbers: IRBd21-060. A flowchart illustrating the construction of this large study dataset is shown in Fig. 2. We collect 37,517 patients recorded in our hospital between January 1, 2004, and December 31, 2020. Then we collect the longitudinal digital screening mammograms and exclude patients without at least one year of screening follow-up, in line with the research *(19).* Details about the distributions of the dataset are available in Fig. S1. Although part of the patients did not have 10-year screening follow-up, we also leverage their known outcomes and images to supervise the model. Therefore, we keep 9,133 patients consisting of 2,562 BC patients who were biopsy-proven within 10 years and 6,571 at intermediate risk who had at least 10 years of screening follow-up and did not receive a cancer diagnosis. All patients are randomly divided into training, validation, and test sets with a ratio of 7.5:1:1.5. The training, validation, and test sets include 6,858, 919, and 1,356 patients with 32,049, 4,432, and 6,311 examinations, respectively.

BC-relevant risk factors are already showing an essential role in both traditional *(12, 15)* and image-based DL methods *(16, 19)*. In our study, we collect the risk factors through electronic medical records from clinical radiology, tumor, therapy, and pathology reports from our hospital. The distribution of clinical risk factors in the inhouse dataset is shown in Table 1. Specifically, we obtain age, race, BI-RADS and breast density (ACR) grade, family history, genetic determinants, previous BC history, previous ovarian cancer history, self-reported menopausal status, and age of menarche. The BI-RADS and breast density (ACR) grade are estimated by radiologists during clinical interpretation. BI-RADS grades include additional imaging required (BI-RADS 0), normal (BI-RADS 1), benign (BI-RADS 2), probably benign (BI-RADS 3), suspicious for malignancy (BI-RADS 4), highly suggestive of malignancy (BI-RADS 5), and known biopsy-proven malignancy (BI-RADS

6). ACR class include mostly composed of fatty tissue (ACR 1), scattered fibroglandular tissue (ACR 2), heterogeneously dense (ACR 3), and extremely dense (ACR 4). Images with missing densities are interpolated by nearest neighbour interpolation with reference to the density estimates of screening images from adjacent years of the patient. Because the weights of patients are missing, we did not calculate the body mass index (BMI). In our study, we also include patients with prior BC. Thus, following the research of recurrence risk prediction *(38)*, we also leverage the information of prior tumor, which include pathologic tumor (pT)-stage, pathologic node (pN)-stage, hormone receptor status (estrogen receptor (ER)- and progesterone receptor (PR)-status), anti-hormonal therapy, human epidermal growth factor receptor 2 (HER2-status), type of surgery, adjuvant chemotherapy, adjuvant radiation therapy, antibody therapy and Pathologic Complete Response (pCR).

Problem formulation

For the risk prediction, we first divide the relevant time span into one-year time-slots and treat each slot as an independent class. To evaluate the overall risk of the first j years, the probabilities of each year are summed together from the first year up to the j_{th} year. The formulas are defined as follows Eq. 1:

$$Risk_j = \sum_{i=1}^{j} y_i = \sum_{i=1}^{j} Softmax\big(F(x)\big)_i \quad (1)$$

where, y_i means the predicted probability of an exam getting BC diagnosis at i_{th} year, which is calculated by inputting a sequence of a mammograms and corrected risk factors m into the model F and using the $Softmax$ function for the probability generalization. For example, as shown in Fig. 1B, to predict the $Risk$ of a patient getting BC within $j = 2$ years from the checked images, it can be calculated as the sum of the probability of the first year and second year.

Architectural details

As shown in Fig. 1, the MTP-BCR model consists of the network weights shared **Image Encoder ($\varphi_{encoder}$)** connected by **Side-Specific (Unilateral-based) Module ($\varphi_{unilateral}$)**, **Exam-based Module (φ_{exam})**, and finally, **Multi-Time-Point fusion Module (φ_{fusion})** combines with the inputted risk factors of patients. Moreover, to improve risk modeling performance and generalization, we also introduce **multi-task learning**, which could benefit from learning the domain-specific features from multiple BC risk-related tasks (detailed below).

Image Encoder: We employ ImageNet pre-trained ResNet-18, excluding the last full connection (FC), as the encoder ($\varphi_{encoder}$) to extract breast tissue features. The weights-shared encoders correspond to each image from a sequence of mammography exams. Each exam includes four images including craniocaudal (CC) view ($v = cc$) and mediolateral oblique (MLO) view ($v = mlo$) from the left($l = left$) and right($l = right$) side breast. Thus, shown in the Eq. 2, each input mammogram, $x_{v,l}^t$, from the 6 time point (t) exams is represented as the high-dimensional locally feature vector $\theta_{v,l}^t$ with the size of 512×1 by the encoder separately.

$$\theta_{v,l}^t = \varphi_{encoder}\big(x_{v,l}^t\big), v \in \{cc, mlo\}, l \in \{right, left\}, t \in \{0,1,2,3,4,5\} \quad (2)$$

Side-Specific Prediction Module: Ipsilateral CC and MLO views are different projection views of the same breast. Practically, they are combined to express the three-dimensional structure of the breast, which radiologists use to detect abnormalities. Moreover, most tumors only appear in one of the breasts *(40)*. In order to train the model to learn the three-dimensional structure of the breast fully and correspond to the previous left and right breast-specific tumor information, we concatenate ($\bigoplus$) the feature vectors of the ipsilateral view, combined with the side-specifically prior tumor information ($tumor_l^t$) and then place a Multi-layer Perceptron (MLP) for side-based multi-task learning. The MLP layer includes two FC layers with an input size of 1152 for the first FC layer and 512 output units each. A dropout layer with a rate of 0.5 between the two FC layers. Therefore, show in Eq. 3 features of the ipsilateral CC and MLO views and outputs a vector (ε_l^t) with the size of 512×1 representing the unilateral-based breast features.

$$\varepsilon_l^t = \varphi_{unilateral}\left(tumor_l^t \oplus \theta_{l,v=cc}^t \oplus \theta_{l,v=mlo}^t\right), l \in \{right, left\}, t \in \{0,1,2,3,4,5\} \quad (3)$$

Exam-Based Prediction Module: For a similar purpose, we also need to combine the information of bilateral breasts to predict the patient level risk. As in Eq. 4, we concatenate the feature vectors from the output bilateral breasts and feed them to another MLP layer with the same structure for the exam-based multi-task learning. Also, a size of 512×1 vector feature (δ^t), which combines the features from right and left breast, represents the global information of a four-view exam.

$$\delta^t = \varphi_{exam}\left(\varepsilon_{l=right}^t \oplus \varepsilon_{l=left}^t\right), t \in \{0,1,2,3,4,5\} \quad (4)$$

Multi-Time Point Fusion Model: For learning the risk development pattern from the longitudinal screening mammograms, five historic exams before the current exam are randomly selected as the reference for the comparison by the multi-time point fusion model. The current exam refers to the target exam for which we access future BC risk. A sequence of mammography exams serve as references along with the time intervals $(i_0, i_1, \cdots, i_5)$ to the current exam and are combined with the risk factors to predict the future likelihood of BC occurring after the current mammograms. Inspired by the research *(26)*, we leverage a sequence/time-aware transformer learning *(41)* to capture features about the temporal relations between multiple mammograms, which aims to disentangle the risk-relevant changing patterns from the normal breast tissue changing patterns. For embedding the spatiotemporal relationships of past-current exams into the continuous latent space, we employ Continuous Position Embedding (CPE) method *(26)*, which computes time continuous embedding e^t to condition the image features. Not that, to avoid ignoring local information of images during multi-time exam comparison, we combine both the local image features $\theta_{v,l}^t$, and global features δ^t. For patients without five history records for references, we select all history records and then mute the missing data by filling 0. At the same time, for the purpose of data augmentation, we randomly drop a subset of exams in the reference sequence to improve model robustness and avoid overfitting. The fusion model also includes the patient risk factors ($riskf$). Subsequently, a fused feature (τ, a vector size of 640×1) is obtained for the final multi-time fused-based multi-task learning, representing the patient's multi-time point screening information.

$$\tau = \varphi_{fusion}(\, riskf \oplus e^t \oplus \delta^t \oplus \theta_{l,v}^t) \quad (5)$$

Multi-task classifier for side-based, exam-based, and multi-time fused prediction:
During the multi-task learning, we constrain the feature extractor for BC risk-related prediction task learning, shown in Fig. 1C. For instance, the predictions of breast-based BC risk, history, tumor location, and tumor sub-type are included for the side-based multi-task classifier. For the exam-based multi-task classifier, we replace the breast-based BC risk prediction and history prediction with exam-based prediction. We also add the prediction of age, BI-RADS, race, density, and manufacture. And for the final multi-time fused classifier, we mainly focus on unilateral specific BC risk. We calculate the Binary Cross Entropy (BCE) for risk prediction and Cross Entropy (CE) Loss for other predictions. For all three classifiers, we mainly focus on the task of risk prediction thus risk-specific tasks have 5 times higher weight than other tasks during the training. For total loss computing, Eq. 6, we also allocate weight w_{fusion} = 1 to the final multi-time point fused classifier, 5 times higher than the other two classifiers ($w_{side} = 0.2,\ w_{exam} = 0.2$). We choose the weights of loss after the hyperparameter search.

$$L_{total} = L_{side} \times w_{side} + L_{exam} \times w_{exam} + L_{fusion} \times w_{fusion} \quad (6)$$

Implementation details
We use ResNet-18 initializing with ImageNet pre-trained weights as the backbone of all our methods. All methods are implemented in PyTorch (version 1.12.1) with the same training strategies. We use the Adam *(42)* optimizer and a rate of 0.5 for dropout *(43)* after every fully connected layer. We train models for 20 epochs with a batch size of 8 and an initial learning rate of 10^{-4}. The learning rate is decayed by a factor of 10 every five epochs. The best models of each method are chosen with the best AUC performance index on the validation set. The experiments are performed on a Quadro A6000 GPU (48GB). The source code is available at https://github.com/Netherlands-Cancer-Institute/MTP-BCR. Mammograms with standard DICOM format are pre-processed before being fed into the model. First, we convert the images into 16-bit PNG format and segment the whole breast region to exclude the background. Then, to unify the size of all images, we zero-pad and resize images to 512 by 1024 pixels while retaining the relative scale and aspect ratio. Finally, the image is normalized using the min-max method. We also employ standard data augmentation techniques (i.e., random flip, brightness, and contrast) during training for model robustness and overfitting prevention.

Evaluation metrics and statistical analysis
In this study, the tasks of prediction of 1- to 10-year risk are categorical classification tasks, in which positive samples are the patients diagnosed with BC within 1 to 10 years while negative samples are women who stayed healthy for at least 1 to 10 year-screening follow-ups. The performances of the different methods are evaluated by the area under the receiver operating characteristic curve (AUC, calculated by scikit-learn, version: 1.1.2, https://scikit-learn.org). To generally evaluate AUCs across all times (from 1- to 10-year risk), the Uno's C-index *(44)* is calculated using scikit-survival (version 0.18.0, https://scikit-survival.readthedocs.io/en/stable/). The 95% confidence intervals (CI) of AUC and C-index matrices are estimated by bootstrapping with 1,000 bootstraps for each measure. Statistical significance among different methods is assessed using DeLong's test *(45)*, with the significant level predefined as $P < 0.05$.

References

1. B. O. Anderson, A. M. Ilbawi, E. Fidarova, E. Weiderpass, L. Stevens, M. Abdel-Wahab, B. Mikkelsen, The global breast cancer initiative: a strategic collaboration to strengthen health care for non-communicable diseases. *Lancet Oncol.* **22**, 578–581 (2021).
2. L. E. Pace, N. L. Keating, A systematic assessment of benefits and risks to guide breast cancer screening decisions. *JAMA* **311**, 1327–1335 (2014).
3. S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S.Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J.Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. D. Fauw, S. Shetty, International evaluation of an ai system for breast cancer screening. *Nature* **577**, 89–94 (2020).
4. M. G. Marmot, D. Altman, D. Cameron, J. Dewar, S. Thompson, M. Wilcox, The benefits and harms of breast cancer screening: an independent review. *Br. J. Cancer* **108**, 2205–2240 (2013).
5. K. J. Geras, R. M. Mann, L. Moy, Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. *Radiology* **293**, 246 (2019).
6. B. Lauby-Secretan, C. Scoccianti, D. Loomis, L. Benbrahim-Tallaa, V. Bouvard, F. Bianchini, K. Straif, B, Breast-cancer screening—viewpoint of the iarc working group. *N. Engl. J. Med.* **372**, 2353–2358 (2015).
7. M. Eriksson, S. Destounis, K. Czene, A. Zeiberg, R. Day, E. F. Conant, K. Schilling, P. Hall, A risk model for digital breast tomosynthesis to predict breast cancer and guide clinical care. *Sci Transl Med* **14**, eabn3971 (2022).
8. A. Gastounioti, S. Desai, V. S. Ahluwalia, E. F. Conant, D. Kontos, Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. *Breast Cancer Res.* **24**, 1–12 (2022).
9. E. S. McDonald, A. S. Clark, J. Tchou, P. Zhang, G. M. Freedman, Clinical diagnosis and management of breast cancer. *J. Nucl. Med.* **57**, 9S–16S (2016).
10. N. Pashayan, A. C. Antoniou, U. Ivanus, L. J. Esserman, D. F. Easton, D. French, G. Sroczynski, P. Hall, J. Cuzick, D. G. Evans, J. Simard, M. Garcia-Closas, R. Schmutzler, O. Wegwarth, P. Pharoah, S. Moorthie, S. De Montgolfier, C. Baron, Z. Herceg, C. Turnbull, C. Balleyguier, P. G. Rossi, J. Wesseling, D.Ritchie, M. Tischkowitz, M. Broeders, D. Reisel, A. Metspalu, T. Callender, H. D. Koning, P. Devilee, S. Delaloge, M. K. Schmidt, M. Widschwendter, Personalized early detection and prevention of breast cancer: Envision consensus statement. *Nat. Rev. Clin. Oncol.* **17**, 687–705 (2020).
11. J. Tyrer, S. W. Duffy, J. Cuzick, A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med* 23, 1111–1130 (2004).
12. M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, J. J. Mulvihill, Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81**, 1879–1886 (1989).
13. A. R. Brentnall, J. Cuzick, D. S. Buist, E. J. A. Bowles, Long-term accuracy of breast cancer risk assessment combining classic risk factors and breast density. *JAMA Oncol.* **4**, e180174–e180174 (2018).
14. T. Carver, S. Hartley, A. Lee, A. P. Cunningham, S. Archer, C. B. D. Villiers, J. Roberts, R. Ruston, F. M. Walter, M. Tischkowitz, D. F. Easton, A. C. Antoniou, Canrisk tool—a web interface for the prediction of breast and ovarian cancer risk and the likelihood of carrying genetic pathogenic variants. *Cancer Epidemiol. Biomarkers Prev.* **30**, 469–473 (2021).

15. J. A. Tice, S. R. Cummings, R. Smith-Bindman, L. Ichikawa, W. E. Barlow, K. Kerlikowske, Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann. Intern. Med.* **148**, 337–347 (2008).
16. A. Yala, C. Lehman, T. Schuster, T. Portnoi, R. Barzilay, A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**, 60–66 (2019).
17. K. Dembrower, Y. Liu, H. Azizpour, M. Eklund, K. Smith, P. Lindholm, F. Strand, Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology* **294**, 265–272 (2020).
18. M. Eriksson, K. Czene, F. Strand, S. Zackrisson, P. Lindholm, K. Lång, D. Förnvik, H. Sartor, N. Mavaddat, D. Easton, P. Hall, Identification of women at high risk of breast cancer who need supplemental screening. *Radiology* **297**, 327–333 (2020).
19. A. Yala, P. G. Mikhael, F. Strand, G. Lin, K. Smith, Y.-L. Wan, L. Lamb, K. Hughes, C. Lehman, R. Barzilay, Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* **13**, eaba4373 (2021).
20. S. Dadsetan, D. Arefan, W. A. Berg, M. L. Zuley, J. H. Sumkin, S. Wu, Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. *Pattern Recognit.* p. 108919 (2022).
21. A. Yala, P. G. Mikhael, C. Lehman, G. Lin, F. Strand, Y.-L. Wan, K. Hughes, S. Satuluru, T. Kim, I. Banerjee, J. Gichoya, H. Trivedi, R. Barzilay, Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat. Med.* **28**, 136–143 (2022).
22. A. J. T. Wanders, W. Mees, P. A.M. Bun, N. Janssen, A. Rodríguez-Ruiz, M. U. Dalmış, N. Karssemeijer, C. H. van Gils, I. Sechopoulos, R. M. Mann, C. J. V. Rooden, Interval cancer detection using a neural network and breast density in women with negative screening mammograms. *Radiology* **303**, 269–275 (2022).
23. A. Yala, P. G. Mikhael, F. Strand, G. Lin, S. Satuluru, T. Kim, I. Banerjee, J. Gichoya, H. Trivedi, C. D. Lehman, K. Hughes, D. J. Sheedy, L. M. Matthis, B. Karunakaran, K. E. Hegarty, S. Sabino, T. B. Silva, M. C. Evangelista, R. F. Caron, B. Souza, E. C. Mauad, T. Patalon, S. Handelman-Gotlib, M. Guindy, R. Barzilay, Multi-institutional validation of a mammography-based breast cancer risk model. *J. Clin. Oncol.* pp. JCO–21 (2021).
24. C. D. Lehman, S. Mercaldo, L. R. Lamb, T. A. King, L. W. Ellisen, M. Specht, R. M. Tamimi, Deep learning vs traditional breast cancer risk models to support risk-based mammography screening. *J. Natl. Cancer Inst.* **114**, 1355–1363 (2022).
25. V. A. Arasu, L. A. Habel, N. S. Achacoso, D. S. Buist, J. B. Cord, L. J. Esserman, N. M. Hylton, M. M. Glymour, J. Kornak, L. H. Kushi, D. A. Lewis, V. X. Liu, D. L. Miglioretti, D. A. Navarro, W. Sieh, L. Shen, O. Sofrygin, H.-C. Yoon, C. Lee, Comparison of mammography artificial intelligence algorithms for 5-year breast cancer risk prediction. *medRxiv* pp. 2022–01 (2022).
26. A. Sriram, Muckley, K. Sinha, F. Shamout, J. Pineau, K. J. Geras, L. Azour, Y. Aphinyanaphongs, N. Yakubova, W. Moore, Covid-19 prognosis via self-supervised representation learning and multi-image prediction. *arXiv preprint arXiv:*2101.04909 (2021).
27. M. Eriksson, K. Czene, C. Vachon, E. F. Conant, P. Hall, Long-term performance of an image-based short-term risk model for breast cancer. *J. Clin. Oncol.* pp. JCO–22 (2023).
28. Liu, F. Zhang, C. Chen, S. Wang, Y. Wang, Y. Yu, Act like a radiologist: Towards reliable multi-view correspondence reasoning for mammogram mass detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 5947–5961 (2021).
29. Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho, K. J. Geras, An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med Image Anal.* **68**, 101908 (2021).

30. A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, C. Rudin, A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nat. Mach. Intell.* **3**, 1061–1070 (2021).
31. C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, R. Barzilay, Mammographic breast density assessment using deep learning: clinical implementation. *Radiology* **290**, 52–58 (2019).
32. D. A. Spak, J. Plaxco, L. Santiago, M. Dryden, B. Dogan, Bi-rads® fifth edition: A summary of changes. *Diagn. Interv. Imaging* **98**, 179–190 (2017).
33. S. M. Castro, E. Tseytlin, O. Medvedeva, K. Mitchell, S. Visweswaran, T. Bekhuis, R. S. Jacobson, Automated annotation and classification of bi-rads assessment from radiology reports. *J. Biomed. Inform.* **69**, 177–187 (2017).
34. W. Ma, Y. Zhao, Y. Ji, X. Guo, X. Jian, P. Liu, S. Wu, Breast cancer molecular subtype prediction by mammographic radiomic features, *Academic radiology* **26**, 196–201 (2019).
35. X. Qian, J. Pei, H. Zheng, X. Xie, L. Yan, H. Zhang, C. Han, X. Gao, H. Zhang, W. Zheng, Q. Sun, L. Lu, K. K. Shung, Prospective assessment of breast cancer risk from multimodal Multiview ultrasound images via clinically applicable deep learning. *Nat. Biomed. Eng.* **5**, 522–532 (2021).
36. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* (2017), pp. 618–626.
37. J. Louro, M. Posso, M. Hilton Boon, M. Rom´an, L. Domingo, X. Castells, M. Sala, A systematic review and quality assessment of individualised breast cancer risk prediction models. *Br. J. Cancer* **121**, 76–85 (2019).
38. V. V¨olkel, T. A. Hueting, T. Draeger, M. C. van Maaren, L. de Munck, L. J. Strobbe, G. S. Sonke, M. K. Schmidt, M. van Hezewijk, C. G. Groothuis-Oudshoorn, S. Siesling, Improved risk estimation of locoregional recurrence, secondary contralateral tumors and distant metastases in early breast cancer: the influence 2.0 model. *Breast Cancer Res. Treat.* **189**, 817–826 (2021).
39. Y. Zhang, Q. Yang, A survey on multi-task learning. *IEEE Trans Knowl Data Eng.* **34**, 5586–5609 (2021).
40. C. Wang, J. Li, F. Zhang, X. Sun, H. Dong, Y. Yu, Y. Wang, Bilateral asymmetry guided counterfactual generating network for mammogram classification. *IEEE Trans Image Process* **30**, 7980–7994 (2021)
41. N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer. *International conference on machine learning (PMLR, 2018)*, pp. 4055–4064.
42. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. *arXiv preprint arXiv:*1412.6980 (2014).
43. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:*1207.0580 (2012).
44. H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, L.-J. Wei, On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117 (2011).
45. X. Sun, W. Xu, Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* **21**, 1389–1393 (2014).

Acknowledgments

The authors thank to the support from the Chinese Scholarship Council scholarship (CSC) (X.W., Y.G., and L.H.: 202107720016, 202006930001, and 202006240065) and Guangzhou Elite Project (T.Z.: TZ–JY201948).

Author contributions: Conceptualization: X.W., T.T., and R.M. Data Collection: X.W., T.Z., L.H., and Y.G. Methodology: X.W., T.T., Y.G., .R.M, and R.S. Investigation: X.W., Y.G., and A.D.A. Visualization: X.W., Y.G., and R.S. Supervision: T.T., R.M., and R.B.T. Writing—original draft: X.W., Y.G., R.S., T.Z., and L.H. Writing—review & editing: R.M., N.K., X.W., C.A.D., M.K.S., and J.T.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All custom codes related to training and developing the deep learning models is available on https://github.com/Netherlands-Cancer-Institute/MTP-BCR. The trained MTP-BCR model is available after the publication of this paper. The inhouse datasets are used under license to the respective hospital system for the current study and are not publicly available. All data associated with this study are present in the paper or the Supplementary Materials.

Figures and Tables

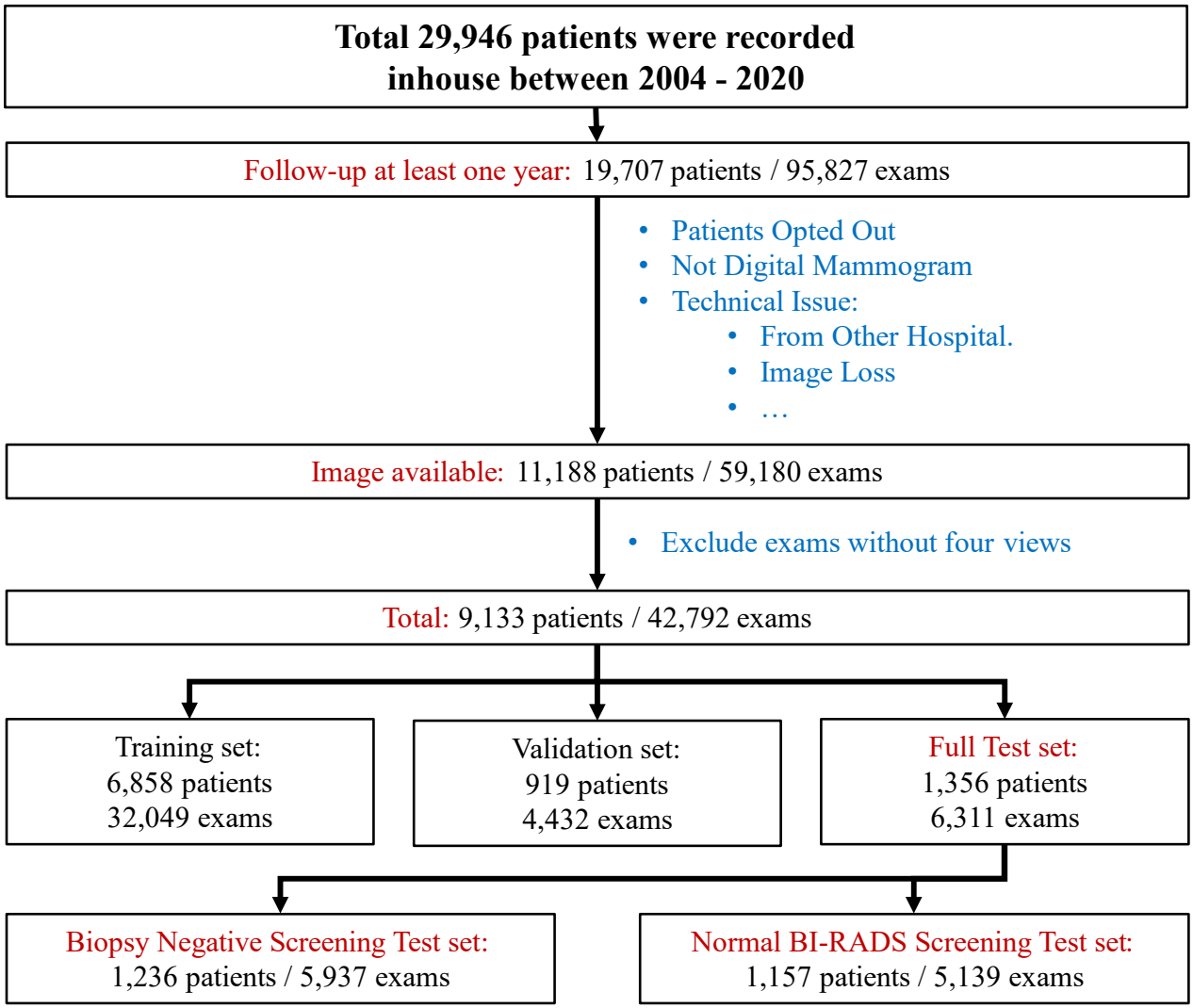

Fig. 1. The flowchart of Inhouse mammogram dataset collection for 10-year risk prediction.

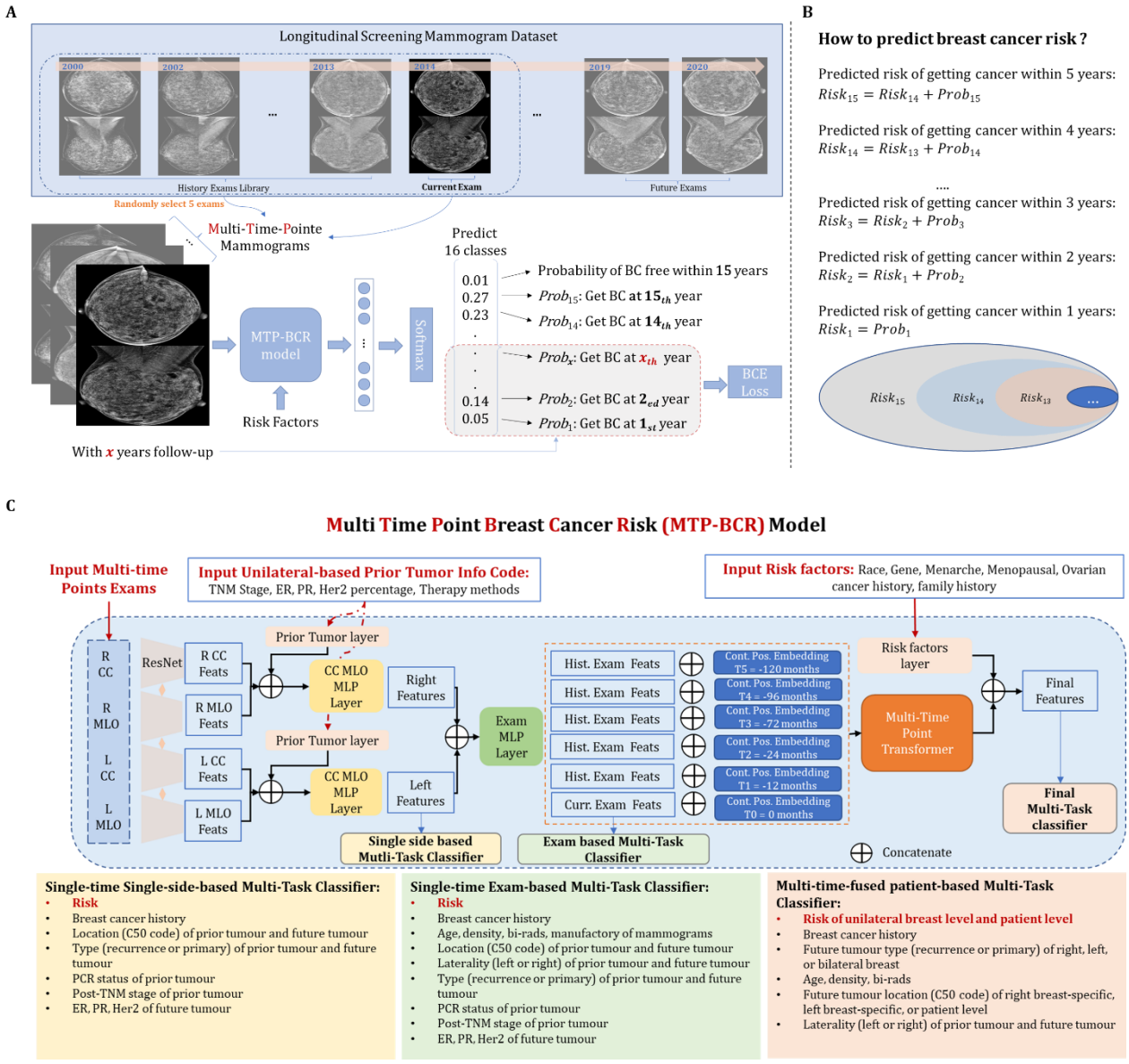

Fig. 2. Schematic description of Multi-Time-Point Breast Cancer Risk (MTP-BCR) model. A. Overview of selecting the multi-time points mammograms for training the MTP-BCR model. B. how to calculate the BC risk C. The details of the MTP-BCR model

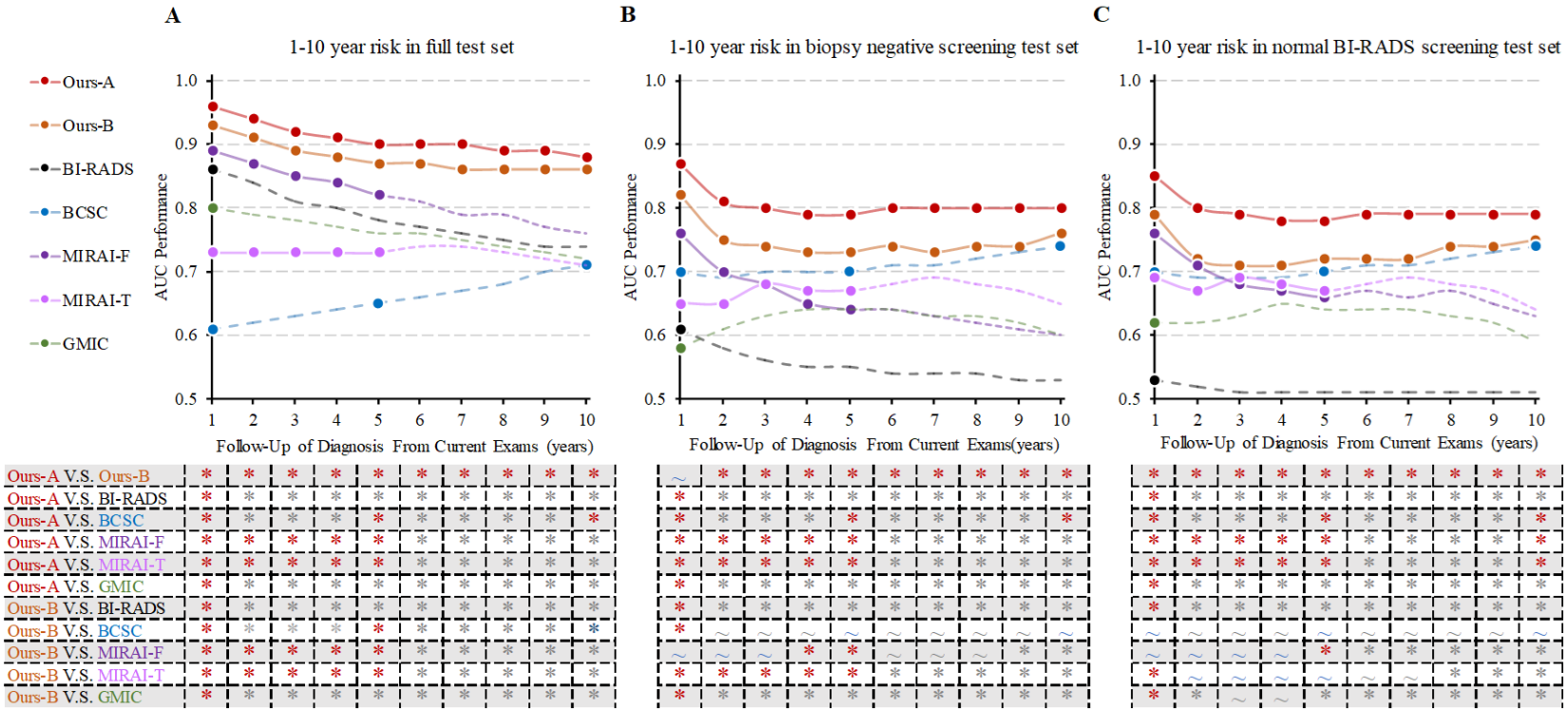

Fig. 3. Cumulative risk at multiple time points (only for women who are completely scored across the BCSC model). Results on the full test set (A), biopsy negative screening test set (B), and normal BI-RADS screening test set (C) are shown in the left, middle, and right, respectively. Ours A: MTP-BCR with risk factors; Ours B: MTP-BCR without risk factors. The color dots represent the performance of the target tasks for which different methods are originally designed to. For instance, the MIRAI risk model is designed to predict 5-year BC risk at multiple time points, GMIC and BI-RADS focus on detecting BC within three months, and the BCSC risk model is used to predict 1-, 5-, and 10-year BC risk. Our methods are designed to predict 10-year BC risk at multiple time points. The dashed curves with different colors represent the AUC metric of 1- to 10- years BC risk for these methods we explored. Asterisk (*) means that there is a significant difference (DeLong's test, $P < 0.05$) in two AUCs corresponding to the methods, whereas tilde (~) indicates that there is no significant difference (DeLong's test, $P > 0.05$) in AUCs of the two methods.

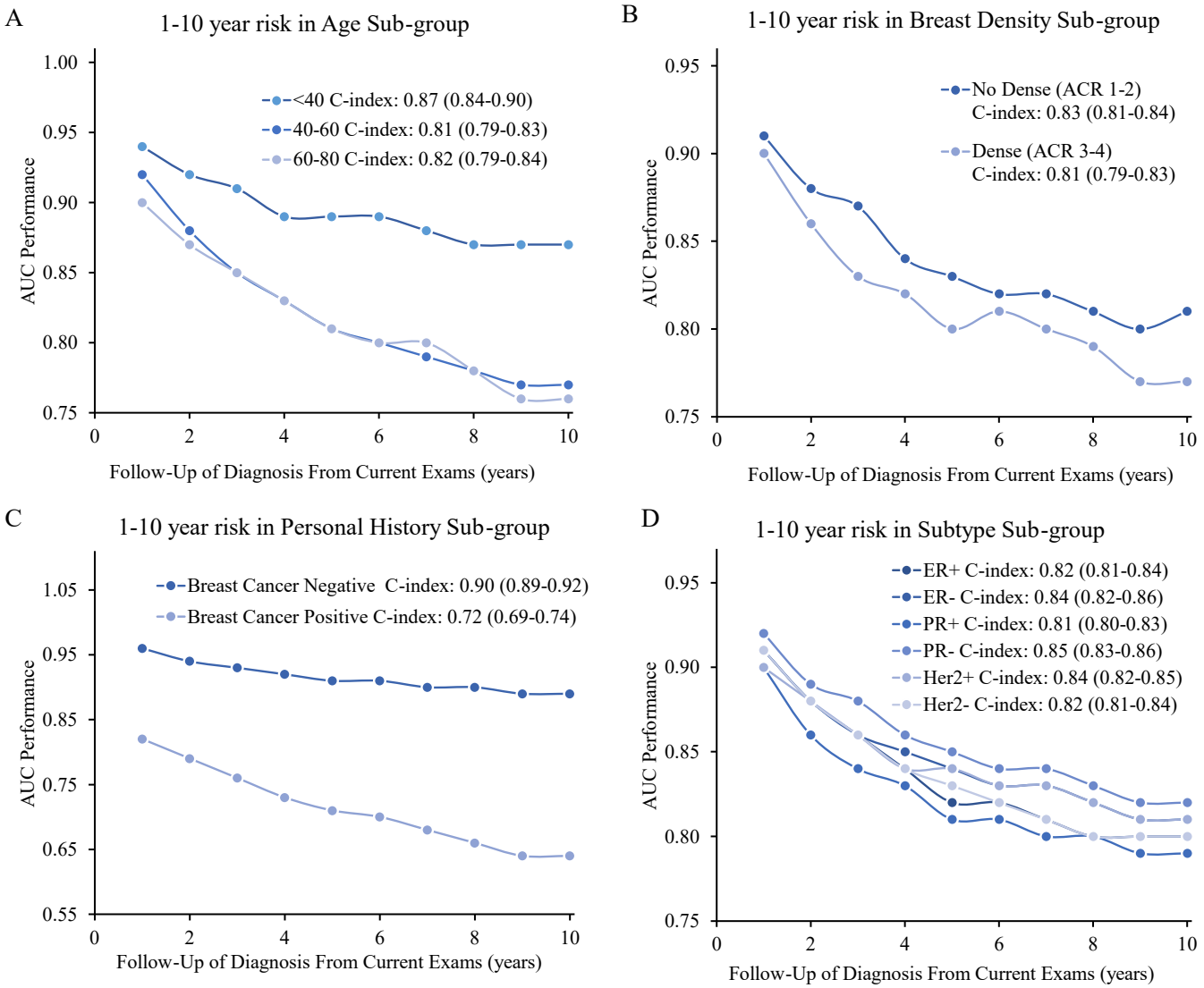

Fig. 4. Cumulative risk at multiple time points on different sub-groups.

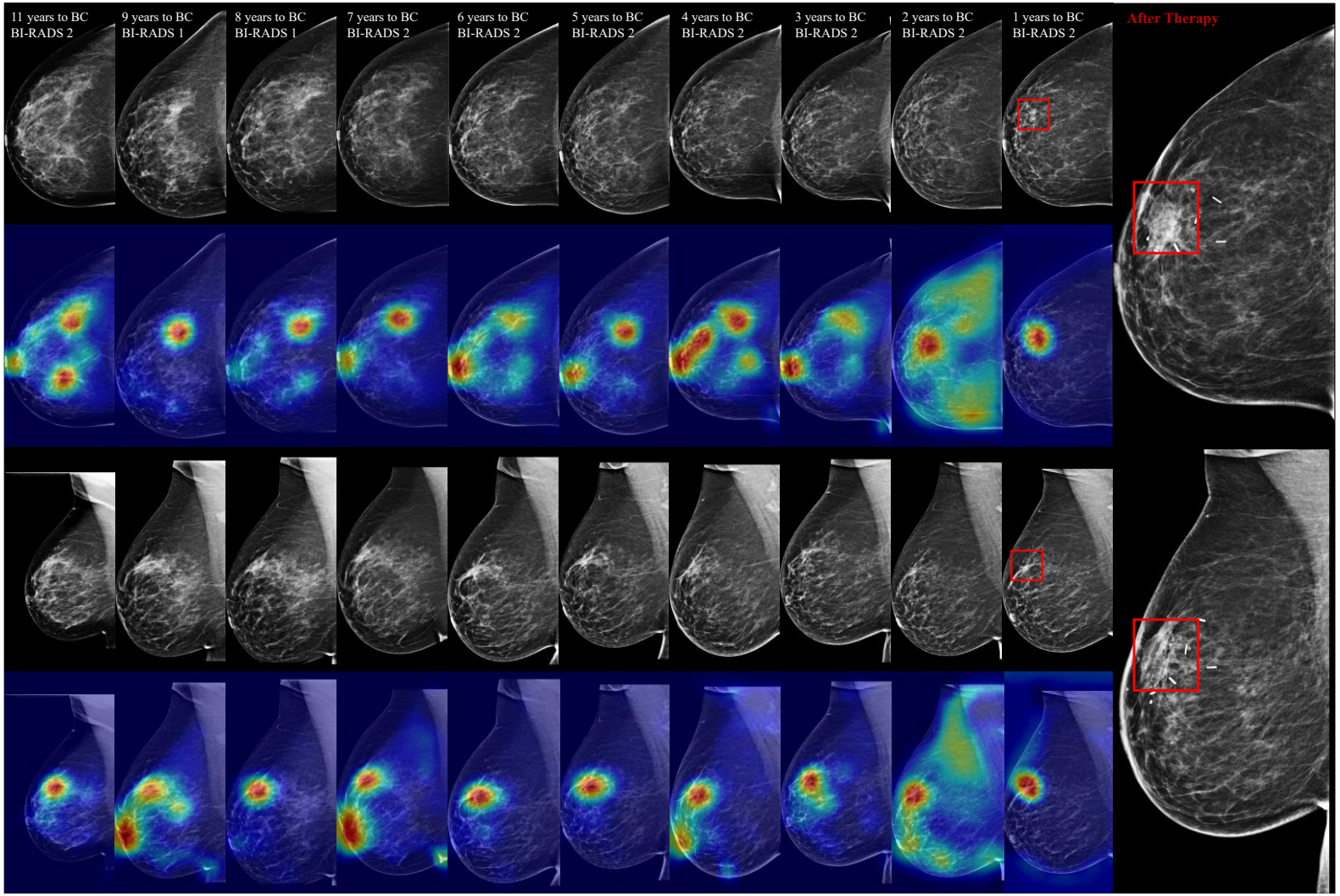

Fig. 5. An example of class activation map (CAM) visualization. The longitudinal craniocaudal (CC) and mediolateral oblique (MLO) mammograms were acquired from a patient who participated in ten consecutive breast cancer screening from 2005 to 2015, culminating in a breast cancer diagnosis during the last screening (invasive ductal and lobular carcinoma located at C50.4, exhibiting positive expression of estrogen receptor (ER+), progesterone receptor (PR+), and human epidermal growth factor receptor 2 (Her2Neu+)). The closer to red, the more relevant the pixel is to the risk prediction.

Table 1. Detailed demographics for the Inhouse dataset. We categorize the number of mammography examinations into different demographic subgroups (rows) and different health subgroups (columns). The reported percentages are the number of examinations as a proportion of the total number of exams in the corresponding health conditions (Num/ColSum) and as a ratio of all exams in the same demographic sub-group (Num/RowSum).

Sub-group	Healthy in 5Y (Num/ColSum) (Num/RowSum)	Healthy in 10Y (Num/ColSum) (Num/RowSum)	Get BC in 1Y (Num/ColSum) (Num/RowSum)	Get BC in 5Y (Num/CoSum) (Num/RowSum)	Get BC in 10Y (Num/ColSum) (Num/RowSum)	All (Num/ColTotal) (Num/ RowSum)
Total:	36,522 (100%)(85.3%)	34,769 (100%)(81.3%)	3,742 (100%)(8.7%)	6,270 (100%)(14.7%)	8,023 (100%)(18.7%)	42,792 (100%)(100%)
Age:						
<40	4,473 (12.2%)(90.1%)	4,290 (12.3%)(86.4%)	327 (8.7%)(6.6%)	492 (7.8%)(9.9%)	675 (8.4%)(13.6%)	4,965 (11.6%)(100%)
40-50	9,371 (25.7%)(87.0%)	8,983 (25.8%)(83.4%)	850 (22.7%)(7.9%)	1,405 (22.4%)(13.0%)	1,793 (22.3%)(16.6%)	10,776 (25.2%)(100%)
50-60	10,580 (29.0%)(85.6%)	10,017 (28.8%)(81.1%)	1,092 (29.2%)(8.8%)	1,776 (28.3%)(14.4%)	2,339 (29.2%)(18.9%)	12,356 (28.9%)(100%)
60-70	7,295 (20.0%)(81.6%)	6,828 (19.6%)(76.4%)	902 (24.1%)(10.1%)	1,647 (26.3%)(18.4%)	2,114 (26.3%)(23.6%)	8,942 (20.9%)(100%)
70-80	3,212 (8.8%)(83.6%)	3,070 (8.8%)(79.9%)	360 (9.6%)(9.4%)	630 (10.0%)(16.4%)	772 (9.6%)(20.1%)	3,842 (9.0%)(100%)
>80	279 (0.8%)(64.7%)	273 (0.8%)(63.3%)	102 (2.7%)(23.7%)	152 (2.4%)(35.3%)	158 (2.0%)(36.7%)	431 (1.0%)(100%)
Unknow	1,312 (3.6%)(88.6%)	1,308 (3.8%)(88.4%)	109 (2.9%)(7.4%)	168 (2.7%)(11.4%)	172 (2.1%)(11.6%)	1,480 (3.5%)(100%)
BI-RADS：						
BI-RADS 0	306 (0.8%)(83.8%)	296 (0.9%)(81.1%)	39 (1.0%)(10.7%)	59 (0.9%)(16.2%)	69 (0.9%)(18.9%)	365 (0.9%)(100%)
BI-RADS 1	6,126 (16.8%)(95.5%)	5,913 (17.0%)(92.2%)	66 (1.8%)(1.0%)	290 (4.6%)(4.5%)	503 (6.3%)(7.8%)	6,416 (15.0%)(100%)
BI-RADS 2	25,363 (69.4%)(91.1%)	24,174 (69.5%)(86.8%)	603 (16.1%)(2.2%)	2,492 (39.7%)(8.9%)	3,681 (45.9%)(13.2%)	27,855 (65.1%)(100%)
BI-RADS 3	1,052 (2.9%)(76.1%)	991 (2.9%)(71.7%)	245 (6.5%)(17.7%)	331 (5.3%)(23.9%)	392 (4.9%)(28.3%)	1,383 (3.2%)(100%)
BI-RADS 4	242 (0.7%)(28.2%)	230 (0.7%)(26.8%)	590 (15.8%)(68.8%)	616 (9.8%)(71.8%)	628 (7.8%)(73.2%)	858 (2.0%)(100%)
BI-RADS 5	21 (0.1%)(2.5%)	21 (0.1%)(2.5%)	826 (22.1%)(97.3%)	828 (13.2%)(97.5%)	828 (10.3%)(97.5%)	849 (2.0%)(100%)
BI-RADS 6	0 (0.0%)(0.0%)	0 (0.0%)(0.0%)	917 (24.5%)(100%)	917 (14.6%)(100%)	917 (11.4%)(100%)	917 (2.1%)(100%)
None	3,412 (9.3%)(82.2%)	3,144 (9.0%)(75.8%)	456 (12.2%)(11.0%)	737 (11.8%)(17.8%)	1,005 (12.5%)(24.2%)	4,149 (9.7%)(100%)
Density:						
ACR 1	2,340 (6.4%)(86.8%)	2,250 (6.5%)(83.5%)	208 (5.6%)(7.7%)	355 (5.7%)(13.2%)	445 (5.5%)(16.5%)	2,695 (6.3%)(100%)
ACR 2	21,163 (57.9%)(84.7%)	20,051 (57.7%)(80.2%)	2,308 (61.7%)(9.2%)	3,829 (61.1%)(15.3%)	4,941 (61.6%)(19.8%)	24,992 (58.4%)(100%)
ACR 3	9,782 (26.8%)(87.1%)	9,385 (27.0%)(83.6%)	812 (21.7%)(7.2%)	1,448 (23.1%)(12.9%)	1,845 (23.0%)(16.4%)	11,230 (26.2%)(100%)
ACR 4	3,137 (8.6%)(85.7%)	2,987 (8.6%)(81.6%)	310 (8.3%)(8.5%)	523 (8.3%)(14.3%)	673 (8.4%)(18.4%)	3,660 (8.6%)(100%)
unknow	100 (0.3%)(46.5%)	96 (0.3%)(44.7%)	104 (2.8%)(48.4%)	115 (1.8%)(53.5%)	119 (1.5%)(55.3%)	215 (0.5%)(100%)
Race:						
White	10,820 (29.6%)(88.2%)	10,133 (29.1%)(82.6%)	659 (17.6%)(5.4%)	1,446 (23.1%)(11.8%)	2,133 (26.6%)(17.4%)	12,266 (28.7%)(100%)
African	141 (0.4%)(88.1%)	135 (0.4%)(84.4%)	11 (0.3%)(6.9%)	19 (0.3%)(11.9%)	25 (0.3%)(15.6%)	160 (0.4%)(100%)
Asian	270 (0.7%)(91.2%)	265 (0.8%)(89.5%)	9 (0.2%)(3.0%)	26 (0.4%)(8.8%)	31 (0.4%)(10.5%)	296 (0.7%)(100%)
Other Race	94 (0.3%)(88.7%)	90 (0.3%)(84.9%)	7 (0.2%)(6.6%)	12 (0.2%)(11.3%)	16 (0.2%)(15.1%)	106 (0.2%)(100%)
Unknow	25,197 (69.0%)(84.1%)	24,146 (69.4%)(80.6%)	3,056 (81.7%)(10.2%)	4,767 (76.0%)(15.9%)	5,818 (72.5%)(19.4%)	29,964 (70.0%)(100%)

Continued Table 1

Sub-group	Healthy in 5Y (Num/ColSum) (Num/RowSum)	Healthy in 10Y (Num/ColSum) (Num/RowSum)	Get BC in 1Y (Num/ColSum) (Num/RowSum)	Get BC in 5Y (Num/CoSum) (Num/RowSum)	Get BC in 10Y (Num/ColSum) (Num/RowSum)	All (Num/ColTotal) (Num/ RowSum)
Gene:						
BRCA 1 mutation	1,321 (3.6%)(83.8%)	1,229 (3.5%)(77.9%)	103 (2.8%)(6.5%)	256 (4.1%)(16.2%)	348 (4.3%)(22.1%)	1,577 (3.7%)(100%)

BRCA 2 mutation	1,183 (3.2%)(86.3%)	1,115 (3.2%)(81.3%)	75 (2.0%)(5.5%)	188 (3.0%)(13.7%)	256 (3.2%)(18.7%)	1,371 (3.2%)(100%)
Positive	2,722 (7.5%)(84.6%)	2,529 (7.3%)(78.6%)	197 (5.3%)(6.1%)	497 (7.9%)(15.4%)	690 (8.6%)(21.4%)	3,219 (7.5%)(100%)
Negative	3,120 (8.5%)(78.5%)	2,653 (7.6%)(66.8%)	367 (9.8%)(9.2%)	852 (13.6%)(21.5%)	1,319 (16.4%)(33.2%)	3,972 (9.3%)(100%)
Unknow	30,680 (84.0%)(86.2%)	29,587 (85.1%)(83.1%)	3,178 (84.9%)(8.9%)	4,921 (78.5%)(13.8%)	6,014 (75.0%)(16.9%)	35,601 (83.2%)(100%)
Menopausal Status：						
Pre-menopausal	5,570 (15.3%)(88.7%)	5,333 (15.3%)(85.0%)	431 (11.5%)(6.9%)	707 (11.3%)(11.3%)	944 (11.8%)(15.0%)	6277 (14.7%)(100%)
Peri-menopausal & Unknow	18,599 (50.9%)(86.9%)	17,825 (51.3%)(83.3%)	1,769 (47.3%)(8.3%)	2,792 (44.5%)(13.1%)	3,566 (44.4%)(16.7%)	21,391 (50.0%)(100%)
Post- menopausal	12,353 (33.8%)(81.7%)	11,611 (33.4%)(76.8%)	1,542 (41.2%)(10.2%)	2,771 (44.2%)(18.3%)	3,513 (43.8%)(23.2%)	15,124 (35.3%)(100%)
Personal History:						
Breast Cancer Positive	18,149 (49.7%)(87.2%)	17,058 (49.1%)(82.0%)	985 (26.3%)(4.7%)	2,656 (42.4%)(12.8%)	3,747 (46.7%)(18.0%)	20,805 (48.6%)(100%)
Breast Cancer Negative	18,373 (50.3%)(83.6%)	17,711 (50.9%)(80.6%)	2,757 (73.7%)(12.5%)	3,614 (57.6%)(16.4%)	4,276 (53.3%)(19.4%)	21,987 (51.4%)(100%)
Ovarian Cancer Positive	401 (1.1%)(79.2%)	368 (1.1%)(72.7%)	42 (1.1%)(8.3%)	105 (1.7%)(20.8%)	138 (1.7%)(27.3%)	506 (1.2%)(100%)
Ovarian Cancer Negative	36,121 (98.9%)(85.4%)	34,401 (98.9%)(81.4%)	3,700 (98.9%)(8.7%)	6,165 (98.3%)(14.6%)	7,885 (98.3%)(18.6%)	42,286 (98.8%)(100%)
Menarche Age：						
<12	854 (2.3%)(91.6%)	819 (2.4%)(87.9%)	30 (0.8%)(3.2%)	78 (1.2%)(8.4%)	113 (1.4%)(12.1%)	932 (2.2%)(100%)
12-15	4,981 (13.6%)(90.5%)	4,744 (13.6%)(86.2%)	202 (5.4%)(3.7%)	521 (8.3%)(9.5%)	758 (9.4%)(13.8%)	5,502 (12.9%)(100%)
>15	228 (0.6%)(90.8%)	223 (0.6%)(88.8%)	9 (0.2%)(3.6%)	23 (0.4%)(9.2%)	28 (0.3%)(11.2%)	251 (0.6%)(100%)
Unknow	30,459 (83.4%)(84.4%)	28,983 (83.4%)(80.3%)	3,501 (93.6%)(9.7%)	5,648 (90.1%)(15.6%)	7,124 (88.8%)(19.7%)	36,107 (84.4%)(100%)
Family History：						
Breast Cancer Positive	25,687 (70.3%)(83.2%)	24,026 (69.1%)(77.8%)	2,908 (77.7%)(9.4%)	5,191 (82.8%)(16.8%)	6,852 (85.4%)(22.2%)	30,878 (72.2%)(100%)
Breast Cancer Negative	10,835 (29.7%)(90.9%)	10,743 (30.9%)(90.2%)	834 (22.3%)(7.0%)	1,079 (17.2%)(9.1%)	1,171 (14.6%)(9.8%)	11,914 (27.8%)(100%)
Ovarian Cancer Positive	2,155 (5.9%)(85.5%)	2,009 (5.8%)(79.8%)	150 (4.0%)(6.0%)	364 (5.8%)(14.5%)	510 (6.4%)(20.2%)	2,519 (5.9%)(100%)
Ovarian Cancer Negative	34,367 (94.1%)(85.3%)	32,760 (94.2%)(81.3%)	3,592 (96.0%)(8.9%)	5,906 (94.2%)(14.7%)	7,513 (93.6%)(18.7%)	40,273 (94.1%)(100%)
Manufacturer：						
Hologic Selenia	5,648 (15.5%)(81.2%)	5,220 (15.0%)(75.1%)	880 (23.5%)(12.7%)	1,307 (20.8%)(18.8%)	1,735 (21.6%)(24.9%)	6,955 (16.3%)(100%)
Lorad Selenia	9,543 (26.1%)(82.2%)	8,870 (25.5%)(76.4%)	1,352 (36.1%)(11.6%)	2,072 (33.0%)(17.8%)	2,745 (34.2%)(23.6%)	11,615 (27.1%)(100%)
Selenia Dimensions	21184 (58.0%)(89.6%)	20,539 (59.1%)(86.8%)	1,100 (29.4%)(4.7%)	2,465 (39.3%)(10.4%)	3,110 (38.8%)(13.2%)	23,649 (55.3%)(100%)
Train Val Test：						
Train	27,294 (74.7%)(85.2%)	25,991 (74.8%)(81.1%)	2,853 (76.2%)(8.9%)	4,755 (75.8%)(14.8%)	6,058 (75.5%)(18.9%)	32,049 (74.9%)(100%)
Valid	3,786 (10.4%)(85.4%)	3,599 (10.4%)(81.2%)	378 (10.1%)(8.5%)	646 (10.3%)(14.6%)	833 (10.4%)(18.8%)	4,432 (10.4%)(100%)
Test	5,442 (14.9%)(86.2%)	5,179 (14.9%)(82.1%)	511 (13.7%)(8.1%)	869 (13.9%)(13.8%)	1,132 (14.1%)(17.9%)	6,311 (14.7%)(100%)

852

Table 2. Comparison of 10-year risk predictions on full test set. C-index and AUC results are presented with 95% Confidence Interval. Note that results of the BCSC model are based on the part of the full inhouse test set, as they are out of the age range of 35-74 or with prior BC history. For a fair comparison, we also implemented the comparison experiments excluding the women that did not have scores from the BCSC model, Shown in Fig. 3. The black fonts represent the performance of the target tasks for which different methods were originally designed to. The gray fonts represent the AUC metric of 1- to 10- years BC risk for these methods we explored. Bold: $P <$ 0.05, the AUCs of our methods are significantly higher than all other model for the same time horizon (except BCSC risk model).

	BI-RADS	*BCSC**	GMIC	MIRAI Test	MIRAI Finetune	MTP-BCR (Ours) Patient Level		MTP-BCR (Ours) Unilateral Breast Level	
Use Risk Factors	-	*Yes*	No	-	-	No	Yes	No	Yes
Full inhouse test set: 6,311 exams, 511 followed by cancer diagnosis within 1 years; 869 diagnosis within 5 years; 1,132 diagnosis within 10 years.									
5-Year C-Index	0.71 (0.70-0.73)	*0.63 (0.61-0.65)*	0.7 (0.68-0.72)	0.67 (0.65-0.69)	0.75 (0.73-0.77)	0.78 (0.76-0.80)	0.84 (0.82-0.85)	0.78 (0.76-0.80)	0.82 (0.81-0.84)
10-Year C-Index	0.69 (0.67-0.70)	*0.64 (0.61-0.66)*	0.69 (0.67-0.70)	0.67 (0.65-0.68)	0.73 (0.72-0.75)	0.77 (0.75-0.78)	0.82 (0.81-0.84)	0.76 (0.75-0.78)	0.81 (0.79-0.82)
1-Year AUC	0.83 (0.81-0.85)	*0.62 (0.58-0.64)*	0.74 (0.72-0.77)	0.70 (0.68-0.73)	0.84 (0.81-0.86)	0.87 (0.85-0.89)	0.91 (0.89-0.92)	0.87 (0.85-0.89)	0.89 (0.87-0.91)
2-Year AUC	0.78 (0.76-0.80)	*0.62 (0.59-0.65)*	0.72 (0.70-0.75)	0.69 (0.66-0.71)	0.80 (0.78-0.82)	0.83 (0.81-0.85)	0.88 (0.86-0.89)	0.83 (0.81-0.85)	0.87 (0.85-0.88)
3-Year AUC	0.74 (0.72-0.76)	*0.63 (0.61-0.66)*	0.72 (0.69-0.74)	0.69 (0.67-0.72)	0.78 (0.76-0.80)	0.80 (0.79-0.82)	0.86 (0.84-0.87)	0.80 (0.79-0.82)	0.84 (0.83-0.86)
4-Year AUC	0.72 (0.70-0.73)	*0.64 (0.61-0.67)*	0.71 (0.69-0.73)	0.69 (0.67-0.71)	0.76 (0.74-0.78)	0.79 (0.77-0.80)	0.84 (0.82-0.85)	0.79 (0.77-0.81)	0.82 (0.81-0.84)
5-Year AUC	0.70 (0.68-0.71)	*0.65 (0.62-0.68)*	0.70 (0.68-0.72)	0.70 (0.68-0.72)	0.74 (0.72-0.76)	0.77 (0.75-0.79)	0.82 (0.81-0.84)	0.77 (0.75-0.79)	0.81 (0.79-0.82)
6-Year AUC	0.68 (0.67-0.70)	*0.66 (0.63-0.69)*	0.70 (0.68-0.72)	0.70 (0.68-0.72)	0.73 (0.71-0.75)	0.77 (0.75-0.79)	0.82 (0.80-0.83)	0.76 (0.75-0.78)	0.80 (0.78-0.81)
7-Year AUC	0.67 (0.66-0.69)	*0.67 (0.65-0.70)*	0.70 (0.68-0.72)	0.70 (0.68-0.72)	0.73 (0.71-0.75)	0.76 (0.74-0.78)	0.81 (0.80-0.83)	0.76 (0.74-0.77)	0.79 (0.77-0.80)
8-Year AUC	0.66 (0.65-0.68)	*0.69 (0.66-0.71)*	0.69 (0.67-0.71)	0.69 (0.67-0.71)	0.72 (0.70-0.74)	0.76 (0.74-0.78)	0.80 (0.79-0.82)	0.75 (0.74-0.77)	0.78 (0.76-0.79)
9-Year AUC	0.65 (0.64-0.67)	*0.70 (0.67-0.73)*	0.69 (0.67-0.71)	0.68 (0.66-0.70)	0.71 (0.69-0.73)	0.76 (0.74-0.78)	0.80 (0.78-0.81)	0.75 (0.74-0.77)	0.77 (0.75-0.78)
10-Year AUC	0.65 (0.63-0.66)	*0.71 (0.68-0.74)*	0.69 (0.66-0.71)	0.67 (0.65-0.70)	0.71 (0.68-0.73)	0.77 (0.75-0.79)	0.80 (0.78-0.82)	0.76 (0.74-0.77)	0.77 (0.75-0.78)

Table 3. Comparison of 10-year risk predictions on Screening Test sets. C-index and AUC results are presented with 95% Confidence Interval. Note that results of the BCSC model are based on part of the test sets- as they are out of the age range of 35-74 or with prior BC history. For fair comparison we also impended the comparison experiments excluding the women that did not score across the BCSC model Show as Fig. 3. The black fonts represent the performance of the target tasks for which different methods were originally designed to. The gray fonts represent the AUC metric of 1- to 10- years BC risk for these methods we explored. Bold: $P < 0.05$, the AUCs of our methods are significantly higher than all other model for the same time horizon (except BCSC risk model).

	BI-RADS	*BCSC **	GMIC	MIRAI Test	MIRAI Finetune	MTP-BCR (Ours) Patient Level		MTP-BCR (Ours) Unilateral Breast Level	
Use Risk Factors	-	*Yes*	No	-	-	No	Yes	No	Yes
Inhouse Biopsy Negative Screening Test set: 5,937 exams, 137 followed by cancer diagnosis within 1 years; 495 diagnosis within 5 years; 758 diagnosis within 10 years.									
5-Year C-index	0.54 (0.53-0.56)	*0.69 (0.65-0.72)*	0.61 (0.59-0.64)	0.63 (0.60-0.65)	0.64 (0.62-0.66)	0.65 (0.63-0.68)	0.74 (0.72-0.76)	0.65 (0.63-0.67)	0.73 (0.71-0.75)
10-Year C-index	0.53 (0.52-0.55)	*0.69 (0.66-0.72)*	0.61 (0.59-0.63)	0.62 (0.60-0.64)	0.64 (0.62-0.66)	0.65 (0.63-0.67)	0.74 (0.72-0.76)	0.65 (0.63-0.67)	0.72 (0.71-0.74)
1-Year AUC	0.61 (0.57-0.65)	*0.70 (0.63-0.76)*	0.59 (0.54-0.65)	0.63 (0.58-0.68)	0.65 (0.60-0.70)	0.70 (0.65-0.74)	**0.77 (0.73-0.81)**	0.68 (0.62-0.73)	0.76 (0.71-0.81)
2-Year AUC	0.58 (0.55-0.60)	*0.69 (0.64-0.75)*	0.61 (0.57-0.65)	0.62 (0.58-0.66)	0.64 (0.60-0.67)	0.67 (0.63-0.70)	**0.75 (0.71-0.78)**	0.66 (0.62-0.70)	0.75 (0.72-0.79)
3-Year AUC	0.55 (0.54-0.58)	*0.70 (0.65-0.74)*	0.62 (0.59-0.65)	0.65 (0.62-0.68)	0.65 (0.62-0.68)	0.66 (0.63-0.69)	**0.75 (0.72-0.77)**	0.66 (0.63-0.69)	0.74 (0.71-0.77)
4-Year AUC	0.54 (0.53-0.56)	*0.70 (0.65-0.74)*	0.63 (0.60-0.65)	0.64 (0.62-0.67)	0.65 (0.62-0.68)	0.66 (0.63-0.68)	**0.74 (0.71-0.76)**	0.66 (0.63-0.68)	0.72 (0.70-0.75)
5-Year AUC	0.54 (0.52-0.55)	*0.70 (0.65-0.74)*	0.62 (0.60-0.65)	0.65 (0.63-0.68)	0.66 (0.63-0.68)	0.66 (0.63-0.68)	**0.73 (0.71-0.76)**	0.66 (0.63-0.68)	0.72 (0.69-0.74)
6-Year AUC	0.53 (0.52-0.55)	*0.71 (0.67-0.74)*	0.63 (0.60-0.65)	0.65 (0.63-0.68)	0.67 (0.65-0.70)	0.66 (0.64-0.69)	**0.74 (0.71-0.76)**	0.66 (0.64-0.68)	0.72 (0.69-0.74)
7-Year AUC	0.53 (0.52-0.54)	*0.71 (0.68-0.75)*	0.63 (0.60-0.65)	0.66 (0.63-0.68)	0.67 (0.65-0.70)	0.67 (0.64-0.69)	**0.74 (0.72-0.76)**	0.66 (0.64-0.68)	0.71 (0.69-0.73)
8-Year AUC	0.53 (0.51-0.54)	*0.72 (0.69-0.75)*	0.62 (0.60-0.64)	0.65 (0.62-0.67)	0.68 (0.65-0.70)	0.67 (0.65-0.69)	**0.73 (0.71-0.75)**	0.67 (0.64-0.69)	0.70 (0.68-0.72)
9-Year AUC	0.52 (0.51-0.53)	*0.73 (0.69-0.76)*	0.62 (0.60-0.65)	0.64 (0.61-0.66)	0.68 (0.65-0.70)	0.68 (0.65-0.70)	**0.73 (0.70-0.75)**	0.67 (0.65-0.69)	0.69 (0.67-0.71)
10-Year AUC	0.52 (0.51-0.53)	*0.74 (0.70-0.77)*	0.63 (0.60-0.66)	0.63 (0.60-0.66)	0.68 (0.66-0.71)	0.70 (0.67-0.72)	**0.73 (0.71-0.75)**	0.68 (0.66-0.70)	0.69 (0.67-0.72)
Inhouse Normal BI-RADS Screening Test set: 5,139 exams, 102 followed by cancer diagnosis within 1 years; 404 diagnosis within 5 years; 612 diagnosis within 10 years.									
5-Year C-index	0.51 (0.50-0.51)	*0.68 (0.64-0.72)*	0.61 (0.59-0.64)	0.62 (0.60-0.65)	0.62 (0.59-0.64)	0.64 (0.62-0.67)	**0.73 (0.71-0.75)**	0.65 (0.63-0.68)	0.73 (0.71-0.75)
10-Year C-index	0.50 (0.50-0.51)	*0.69 (0.66-0.72)*	0.61 (0.59-0.64)	0.62 (0.60-0.64)	0.62 (0.60-0.65)	0.64 (0.62-0.67)	**0.73 (0.71-0.75)**	0.65 (0.63-0.67)	0.73 (0.71-0.75)
1-Year AUC	0.53 (0.50-0.55)	*0.70 (0.63-0.77)*	0.61 (0.54-0.67)	0.62 (0.57-0.68)	0.63 (0.57-0.68)	0.66 (0.60-0.71)	**0.74 (0.69-0.79)**	0.68 (0.63-0.73)	0.77 (0.72-0.81)
2-Year AUC	0.51 (0.50-0.53)	*0.69 (0.63-0.75)*	0.60 (0.56-0.65)	0.62 (0.58-0.66)	0.61 (0.57-0.65)	0.64 (0.60-0.68)	**0.73 (0.69-0.76)**	0.67 (0.63-0.71)	0.75 (0.72-0.79)
3-Year AUC	0.51 (0.50-0.52)	*0.70 (0.64-0.74)*	0.62 (0.58-0.66)	0.65 (0.61-0.68)	0.63 (0.59-0.66)	0.64 (0.61-0.68)	**0.73 (0.70-0.76)**	0.66 (0.62-0.69)	0.74 (0.71-0.77)
4-Year AUC	0.51 (0.50-0.51)	*0.69 (0.65-0.74)*	0.63 (0.60-0.66)	0.64 (0.61-0.67)	0.63 (0.60-0.66)	0.64 (0.62-0.67)	**0.72 (0.70-0.75)**	0.66 (0.63-0.68)	0.73 (0.70-0.75)
5-Year AUC	0.51 (0.50-0.51)	*0.70 (0.65-0.74)*	0.62 (0.59-0.65)	0.65 (0.62-0.68)	0.64 (0.61-0.67)	0.65 (0.62-0.67)	**0.72 (0.70-0.75)**	0.66 (0.63-0.68)	0.72 (0.69-0.74)
6-Year AUC	0.50 (0.50-0.51)	*0.71 (0.66-0.75)*	0.63 (0.60-0.65)	0.66 (0.63-0.68)	0.66 (0.63-0.68)	0.65 (0.63-0.68)	**0.73 (0.71-0.76)**	0.66 (0.63-0.68)	0.72 (0.70-0.74)
7-Year AUC	0.50 (0.50-0.51)	*0.71 (0.67-0.75)*	0.63 (0.60-0.66)	0.66 (0.63-0.69)	0.66 (0.64-0.69)	0.66 (0.63-0.68)	**0.73 (0.71-0.76)**	0.66 (0.64-0.68)	0.71 (0.69-0.73)
8-Year AUC	0.50 (0.50-0.51)	*0.72 (0.68-0.76)*	0.63 (0.60-0.65)	0.66 (0.63-0.68)	0.67 (0.65-0.70)	0.67 (0.64-0.69)	**0.73 (0.71-0.75)**	0.67 (0.64-0.69)	0.71 (0.68-0.73)
9-Year AUC	0.50 (0.50-0.51)	*0.73 (0.69-0.76)*	0.63 (0.60-0.66)	0.65 (0.62-0.68)	0.67 (0.64-0.70)	0.67 (0.65-0.70)	**0.73 (0.70-0.75)**	0.67 (0.65-0.69)	0.70 (0.67-0.72)
10-Year AUC	0.50 (0.50-0.51)	*0.74 (0.70-0.77)*	0.63 (0.60-0.66)	0.64 (0.61-0.67)	0.67 (0.65-0.70)	0.69 (0.67-0.72)	**0.73 (0.71-0.76)**	0.68 (0.66-0.70)	0.70 (0.67-0.72)

Table 4. Comparison of future risk predictions. AUC results are presented with 95% Confidence Interval. BCSC: Breast Cancer Surveillance Consortium; GMIC: Globally-Aware Multiple Instance Classifier. Note that results of the BCSC model are based on part of the full inhouse test set, as they are out of the age range of 35-74 or with prior BC history. For fair comparison we also impended the comparison experiments excluding the women that did not score across the BCSC model. Bold: $P < 0.05$, the AUCs of our methods are significantly higher than all other model for the same time horizon.

	BI-RADS	BCSC	GMIC	MIRAI Test	MIRAI Finetune	MTP-BCR (Ours) Patient Level		MTP-BCR (Ours) Unilateral Breast Level	
Use Risk Factors	-	Yes	No	-	-	No	Yes	No	Yes
Full test set									
2-5 Year AUC	0.51 (0.50-0.52)	-	0.63 (0.60-0.66)	0.64 (0.61-0.67)	0.60 (0.57-0.63)	0.64 (0.61-0.66)	**0.72 (0.69-0.74)**	0.65 (0.62-0.67)	0.71 (0.68-0.73)
4-5 Year AUC	0.50 (0.49-0.52)	-	0.61 (0.57-0.65)	0.62 (0.58-0.66)	0.58 (0.53-0.63)	0.63 (0.59-0.66)	**0.72 (0.68-0.75)**	0.63 (0.59-0.67)	0.69 (0.66-0.73)
2-10 Year AUC	0.50 (0.49-0.51)	-	0.63 (0.60-0.65)	0.62 (0.59-0.65)	0.60 (0.57-0.62)	**0.68 (0.66-0.71)**	**0.73 (0.70-0.75)**	0.67 (0.65-0.69)	0.69 (0.66-0.71)
4-10 Year AUC	0.50 (0.49-0.51)	-	0.61 (0.58-0.64)	0.60 (0.57-0.63)	0.59 (0.56-0.62)	**0.67 (0.64-0.70)**	**0.73 (0.70-0.75)**	0.66 (0.64-0.68)	0.68 (0.66-0.71)
6-10 Year AUC	0.50 (0.49-0.52)	-	0.60 (0.57-0.64)	0.58 (0.55-0.62)	0.60 (0.56-0.63)	**0.67 (0.63-0.70)**	**0.73 (0.70-0.76)**	0.65 (0.62-0.68)	0.69 (0.66-0.72)
Part of Full test set: only women who completed scored across the BCSC model									
2-5 Year AUC	0.52 (0.49-0.55)	0.69 (0.64-0.74)	0.66 (0.61-0.71)	0.67 (0.62-0.71)	0.59 (0.54-0.65)	0.69 (0.64-0.74)	**0.76 (0.72-0.80)**	0.69 (0.65-0.74)	0.74 (0.70-0.78)
4-5 Year AUC	0.50 (0.47-0.53)	0.67 (0.60-0.74)	0.64 (0.58-0.71)	0.64 (0.57-0.70)	0.57 (0.49-0.64)	0.68 (0.61-0.74)	**0.76 (0.71-0.81)**	0.66 (0.60-0.73)	0.73 (0.68-0.78)
2-10 Year AUC	0.52 (0.50-0.54)	0.73 (0.69-0.77)	0.63 (0.59-0.67)	0.64 (0.60-0.69)	0.57 (0.53-0.62)	0.73 (0.69-0.77)	**0.79 (0.75-0.82)**	0.72 (0.68-0.75)	0.72 (0.68-0.75)
4-10 Year AUC	0.51 (0.49-0.53)	0.72 (0.67-0.76)	0.62 (0.57-0.67)	0.62 (0.57-0.67)	0.56 (0.51-0.61)	0.72 (0.68-0.76)	**0.79 (0.75-0.82)**	0.70 (0.66-0.74)	0.72 (0.68-0.75)
6-10 Year AUC	0.51 (0.49-0.54)	0.71 (0.66-0.76)	0.61 (0.55-0.67)	0.61 (0.55-0.66)	0.56 (0.50-0.63)	0.70 (0.65-0.75)	**0.79 (0.75-0.83)**	0.68 (0.64-0.73)	0.73 (0.68-0.77)