1	Predicting up to 10 year breast cancer risk using longitudinal
2	mammographic screening history
3	
4	
5	Authors \mathbf{X} in Wang 1/23 Tag Tag 1/4* Waan Cag 1/23 Division a Su 5 Tierry Zhang 1/23 Luci Hag 1/3
6 7 8	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}
9	Beets-Tail, Neo Karssenleijer, Ritse Main
10	
11	Affiliations
12 13	¹ Department of Radiology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.
14	² GROW School, Maastricht University, Maastricht, 6202 AZ, The Netherlands.
15 16	³ Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Nijmegen, 6525 GA, The Netherlands.
17 18	⁴ Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, 999078, China.
19	⁵ Erasmus Medical Center, Erasmus University, Rotterdam, 3015 GD, The Netherlands.
20 21	⁶ Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.
22 23	⁷ Dipartimento di Diagnostica per Immagini, Radioterapia Oncologica ed Ematologia, Fondazione Policlinico Universitario A. Gemelli, IRCCS, 00168 Rome, Italy.
24 25	⁸ Department of Surgical Oncology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.
26 27	⁹ Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.
28 29	¹⁰ Department of Epidemiology, The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands.
30 31	¹¹ Department of Medical Imaging, Radboud University Medical Center, Nijmegen, 6500 HB, The Netherlands
32	*Corresponding author
33	
34	
35	Abstract
36	Risk assessment of breast cancer (BC) seeks to enhance individualized screening and
37	prevention strategies. BC risk informs healthy individuals of the short- and long-term
38	likelihood of cancer development, also enabling detection of existing BC. Recent
59 40	based models and achieve state-of-the-art (SOTA) at short-term risk prediction, but
41	mainly use single-time exams, which seem to rely more on detecting existing lesions. We
42	present a novel temporospatial and explainable deep learning risk model, the Multi-Time

Point Breast Cancer Risk Model (MTP-BCR), which learns from longitudinal 43 44 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 45 42,792 consecutive exams involving 9,133 women, our model demonstrates a significant 46 improvement in long-term (10-year) risk prediction with an area under the receiver 47 operating characteristics (AUC) of 0.80, outperforming the traditional BCSC 10-year risk 48 model and other SOTA methods at 5-year AUC in various screening cohorts. 49 50 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 51 derived from our model may help clinicians better understand the progression from 52 normal tissue to cancerous growth, enhancing interpretability in breast cancer risk 53 assessment. 54

56 Teaser

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

MAIN TEXT

62 63

55

57

58 59 60

61

64 Introduction

Breast cancer (BC) is one of the most common cancers in the world and is the cause of a 65 large fraction of cancer-related mortality among women (1, 2). Studies have shown that 66 age-based population-level BC screening programs, which aim to detect breast tumors at 67 an early stage (3), reduces breast cancer specific mortality (4-7). However, the broad 68 adoption of mammographic screening results also in high cost of imaging, false-positives 69 and over-diagnoses, which explains the strong controversy of screening (8, 9). Therefore, 70 71 "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 72 endogenous and exogenous risk factors, and also medical imaging (10-12). Current BC 73 risk assessment models are designed to be sensitive to the high-risk population who could 74 benefit from more aggressive screening and prevention. At the same time these models 75 could advocate less frequent screening for the low-risk population to reduce the harm and 76 77 cost of screening, however this is less common.

Based upon the timespan for breast cancer prediction, risk models can be divided into 78 79 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 80 screening. Long-term risk prediction helps determine risk-based screening regimens and 81 eligibility for preventive treatment (13). Many of the traditional risk models, such as 82 Tyrer-Cuzick (11), CANRISK (14), National Cancer Institute Breast Cancer Risk 83 Assessment Tool (BCRAT) (12), and Breast Cancer Surveillance Consortium (BCSC) 84 (15) investigate primarily long-term risk estimates. The performances of these risk models 85 remain modest in clinical practice and are not very sensitive to short/middle-term cancer 86 risk variation due to the shortage of individual-specific risk adaptation, for example 87 through the incorporation of detailed imaging findings beyond breast density only. With 88 the recent boost in deep learning (DL) methods, some studies that combined large 89 screening mammography datasets with detailed risk factors have shown considerable 90 promise to help balance the harm-to-benefit ratios of BC screening (7, 16-22) and were 91

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 95 image or exam as single input without any historical reference (7, 16, 19). It's like judging 96 the motion trajectory of an object in a still frame of a video. We hypothesize that accurate 97 estimation of the breast tissue changes may make the task of predicting long-term BC risk 98 easier. In clinical practice, radiologists routinely compare mammography exams to 99 identify developing abnormalities. Therefore, beyond learning risk features (e.g. breast 100 density) from single-time point imaging, multi-time point learning may also be helpful in 101 discovering the underlying dynamics of the risk pattern for BC development (20, 26). 102 Moreover, due to the lack of a long-term longitudinal screening mammogram dataset, the 103 potential of image-based DL methods for longer-term (e.g., 10 years) BC risk prediction 104 has been less explored. Only one recent research investigated the long-term performance 105 of an image-based short-term risk model (27). 106

- Despite the development of promising risk models in BC screening programs, the 107 interpretability of medical AI models is still difficult, whereas understanding the predicted 108 outputs is essential for clinical acceptance. How to endow an existing risk model with 109 explainability of the underlying reasoning remains the boundary to explore. Apart from 110 being similar to what an actual radiologist does when searching for the sign of BC risk 111 (28, 29), the AI models must reasonably show radiologists more details during inference 112 (30) for clinical acceptance. However, most recent studies only aim to predict the patient-113 level risk and do not produce a location-specific risk. Improvement of these specific 114 predictions and visualizations could not only improve the interpretability of the model and 115 make it easier for physicians to understand the model's decision-making but also inform 116 doctors where to focus and then guide them when deciding on the most suitable targeted 117 examinations and prevention strategies. An ideal risk model should therefore not only 118 stratify high-risk groups but also focus the doctors attention to changing areas in the breast 119 earlier. 120
- We propose the Multi-Time-Points Breast Cancer Risk model (MTP-BCR), an end-to-end 121 model that estimates the long-term future BC risk based on changes in breast tissue. Our 122 contributions are as follows. First, our model leverages historical and current exams from 123 a large in-house clinical mammogram dataset and obtains remarkable performance 124 125 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 126 models for long-term 10-year risk prediction. Third, we explore the unilateral breast level 127 BC risk prediction and achieve similar performance to our risk models at the patient level. 128 Fourth, we highlight suspicious areas in a longitudinal test dataset using the model's 129 130 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 131 perform a systematic subgroup analysis. The results imply that our model may improve 132 upon traditional and other image-based DL risk models. 133

134 **Results**

135 **Overview of algorithm**

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

multiple time points of examinations from the longitudinal screening mammograms are 138 139 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, 140 MTP-BCR, shown in Fig. 1A, leveraging longitudinal mammograms and medical records 141 to capture the features related to increased BC risk. We aim to predict the risks on a 142 patient-based level as well as for a single breast. Briefly, the proposed risk model first 143 utilizes the multi-level (breast and patient level risk) and multi-task learning to extract 144 145 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 146 the practical use scenario) for learning the dynamic risk features using a multi-time point 147 fusion model. This is combined with the risk factors of patients for predicting the future 148 risk. It means that our end-to-end model uses current and historic screening 149 mammography exams and existing medical records and then predicts the future 10-year 150 BC risk. Architectural details, contents of the medical records, and risk factors are 151 presented in the method part. 152

Risk calculation can be treated as a multi-class classification problem (16, 19), which is 153 common in breast imaging, such as the classification of breast density (31), the Breast 154 Imaging Reporting and Data System (BI-RADS (32)) score (33), the type of malignancy 155 (3), and the BC molecular subtype (34). As shown in Fig. 1B, the risk of a patient getting 156 BC within 5 or 10 years from the available data can be calculated as the cumulative sum 157 of the probability from the first year up to the fifth or tenth year. Importantly, the 158 prediction results of our model can guarantee that the risk is monotonically increasing and 159 self-consistent. This avoids the situation, that can occur with separately trained models, 160 where long-term risk may be lower than short-term risk. Moreover, this formulation also 161 learns the inherent relationship between risks at different time points. In this study, the 162 model is trained to predict the risk of BC at each of the 15 years and is validated by 163 predicting BC within ten years. Therefore, our model can be easily extended for a longer 164 than 15-year risk prediction when collecting enough longer-term follow-up data. 165

166 Screening cohorts for risk modeling & compared candidates

Fig. 1 shows the flowchart of the Screening Cohort selection in the inhouse dataset. The 167 dataset contains 42,792 exams of 9,133 patients, split into 32,049 exams / 6,858 patients, 168 4,432 exams / 919 patients, and 6,311 exams / 1,356 for the train set, validation set, and 169 test set, respectively. The proposed model aims to handle multiple tasks, including cancer 170 detection and future risk prediction. This would facilitate implementation in an actual 171 clinical BC screening program, where not only focusing on the stratification of the high-172 risk population, short-term risk and determining whether women should be recalled is also 173 essential. In fact, detection of existing malignancies, can be considered as an extremely 174 short-term BC risk (35). Moreover, some of the model characteristics for the tasks of 175 cancer detection and future risk prediction can be complementary (19). 176

177 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 178 biopsy-negative screening group (5,937 exams / 1,236 patients), which includes 179 mammography exams with BI-RADS 1 and 2 scores, or other BI-RADS scores but with 180 benign biopsy results within 90 days from the screening date. The second test set is called 181 the normal BI-RADS screening group (5,139 exams / 1,157 patients), which only consists 182 of the exams scored as BI-RADS 0, 1 and 2. It aims to explore how the model performs on 183 high-risk population stratification when radiologists deem the exams not suspicious. 184

The inhouse mammogram dataset is coupled to patient derived classical risk factors that 185 186 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 187 in Supplemental Table S1. The BCSC model can estimate five-year, and ten-year BC risk 188 based on risk factors but requires excluding patients following exclusion criteria (previous 189 diagnosis of BC, younger than age 35 or older than age 74, or missing density estimates). 190 Although studies have shown that image-based DL risk models outperform traditional risk 191 192 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 193 with not only the traditional BCSC 1-year and 5-year risk models but also with the BCSC 194 predicted 10-year risk. To demonstrate the added value of the inclusion of patient based 195 risk factors, we also build a similar multi-time point model without risk factors as MTP-196 BCR for comparison. 197

- To demonstrate the risk prediction capacity, we compare it to the SOTA MIRAI 198 (Massachusetts Institute of Technology, Boston, Massachusetts) model (19). MIRAI is a 199 mammogram-based risk model that can predict 5-year risk at multiple time points and 200 outperforms traditional models. Moreover, this model includes a pretrained risk factor 201 predictor that allows missing risk factors. In this study, we also explore MIRAI's ability of 202 the longer-term 10-year risk estimate. We obtain the pretrained MIRAI model from their 203 204 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. 205 We conduct hyperparameter search to finetune MIRAI and select the model with the best 206 concordance index (C-index) on the validate set. Similar to the research (19), to 207 investigate the performance of our method on BC detection, we also compare our model 208 with the retrospective radiologist BI-RADS scores and the Globally Aware Multiple 209 Instance (GMIC, New York University, New York) model (29). The GMIC is another 210 recent SOTA DL model which focuses on detecting BC within three months, and some 211 researches also show its potential for BC risk prediction (19,25). The pretrained GMIC 212 model is obtained from the public GitHub repository 213
- (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).

224 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 231 232 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 233 1- to 10-year risk prediction show that the performances of the MTP-BCR with risk 234 factors are significantly higher than those of MTP-BCR without risk factors (All P values 235 < 0.05). We also evaluate the unilateral breast level cancer prediction of MTP-BCR 236 models both with risk factors and without risk factors, as shown in Fig. S3. Similar 237 238 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 239 risk factors, and 5-year C-indices of 0.82 (95% CI, 0.81-0.84) and 0.78 (95% CI, 0.76-240 0.80). The AUCs of unilateral breast-level cancer prediction without risk factors ranged 241 from 0.76 to 0.87. When incorporating patient derived risk factors, the AUCs ranged from 242 0.77 to 0.89. Therefore, our MTP-BCR risk model can also accurately predict the risk of 243 development of BC in a unilateral breast, and using risk factor information can further 244 improve the performance of 1- to 10-year risk. 245

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 multilevel, and multi-time point learning strategies can improve the ability to extract riskrelated features. The C-indices and AUCs show that the model with multi-task, multilevel, 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-253 RADS assessments and the BC detection method GMIC are 0.83 (95% CI, 0.81-0.85) and 254 0.74 (95% CI, 0.72-0.77) respectively, which are significantly lower than both MTP-BCR 255 risk models (P < 0.001). Therefore our MTP-BCR risk model outperforms radiologists 256 and the SOTA BC detection model for BC detection and extremely short-term risk 257 prediction, even when only using mammograms. Comparing to the SOTA DL-based 258 MIRAI model, the 5-year C-indices of the MTP-BCR models (with/without risk factor) 259 are 0.82 (95% CI, 0.81-0.84) and 0.77 (95% CI, 0.75-0.78) versus 0.73 (95% CI, 0.72-260 0.75). The AUC results show that both the MTP-BCR models significantly outperform the 261 MIRAI model at all time points from 1-year to 5-year risk (All P values < 0.01). Also for 262 10-year risk prediction, the MTP-BCR methods still have a significant advantage (All P 263 values < 0.05). 264
- *Comparing with BCSC risk model:* Note that our study also include patients who are not 265 eligible for risk calculation by the BCSC model as they are either out of the required age 266 range of 35-74 or had a prior BC history. For this comparison, we conduct the experiments 267 excluding the women that did not have scores from the BCSC model (Table S1). The 268 AUC curves of all methods are shown in Fig. 3A. The AUC results show that both the 269 270 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), 0.65 271 (95% CI, 0.62-0.68), and 0.71 (95% CI, 0.68-0.74), respectively. The 5-year and 10-year 272 C-indices of BCSC models are 0.63 (95% CI, 0.61-0.65) and 0.64 (95% CI, 0.61-0.66), 273 compared to C-indices of 0.91 (95% CI, 0.90-0.92) and 0.90 (95% CI, 0.88-0.91) by our 274 MTP-BCR risk model with risk factors. Moreover, we also compare our risk models with 275 the other methods in this specific population. The AUC results show that both the MTP-276

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 5year risk dataset (Table S12 and S13). Similar results to those of 10-year risk ablation
experiments also demonstrate the stability of our learning strategy.

285 **Risk prediction in a healthy screening population**

- In biopsy negative screening population: we evaluate all methods using the inhouse 286 negative biopsy screening test set including 5,937 examinations / 495 positives within 5 287 years / 758 positives within 10 years, as shown in Table 3. In this test set, results show that 288 our MTP-BCR with risk factors holds superiority in C-index and AUC metrics. Extremely 289 short-term (1 year) risk prediction can be equivalent to interval cancer detection. In this 290 task, the AUC of our model without risk factors is 0.70 (95% CI, 0.65-0.74). The AUC 291 increases to 0.77 (95% CI, 0.73-0.81) when using risk factors, significantly higher than 292 that from BI-RADS scores with 0.61 (95% CI, 0.57-0.65), the BC detection method 293 GMIC with 0.59 (95% CI, 0.54-0.65), and also the finetuned MIRAI with 0.65 (95% CI, 294 295 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 296 models (with risk factor: 0.87 / without risk factor: 0.82). For the long-term 5-year risk 297 prediction. our model reaches a C-index of 0.65 (95% CI, 0.63-0.68) without risk factors 298 and a C-index of 0.74 (95% CI, 0.72-0.76) with risk factors, versus 0.64 (95% CI, 0.62-299 0.66) for the finetuned MIRAI model (Table 3). When comparing with the clinical BCSC 300 5-year risk model (as shown in Table S1), our MTP-BCR obtains a C-index of 0.79 (95%) 301 CI, 0.76-0.82) versus 0.69 (95% CI, 0.65-0.72). For the longer-term 10-year risk estimate, 302 our only image based MTP-BCR model has a similar performance to 10-year risk BCSC 303 model according by the 10-year C-index. But the results indicate that the MTP-BCR with 304 risk factors model still significantly outperforms all other models by AUCs at each time 305 point (All *P* values < 0.05). 306
- In normal BI-RADS screening population: In the normal BI-RADS screening test set 307 with 5,139 examinations /404 positives within 5 years /612 positives within 10 years, we 308 investigate the potential of risk models when radiologists cannot find any suspicious 309 310 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 311 significantly better than all other methods among the extremely short-term with a 1-year 312 AUC of 0.74 (95% CI, 0.69-0.79), long-term with a 5-year AUC of 0.72 (95% CI, 0.70-313 0.75), and longer-term with 10-year AUC of 0.73 (95% CI, 0.71-0.76) risk predictions. 314 Especially in the BCSC model target population (aged 35-74, without prior BC history), 315 316 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 317 guessing. 318
- 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 patientlevel 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 323 324 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 325 find that the finetuned MIRAI model on the dataset with missing screening follow-up 326 seems to be difficult. Specifically, the performances of the finetuned MIRAI from 3- to 327 10-year BC risk prediction become worse (shown in Fig. 3 B and C). Thus, for a 328 meaningful comparison, it is necessary to re-implement these comparison experiments on 329 330 the two recollected five-year screening datasets which cleaned exams with missing followup labels. As shown in Table S8 and S9, the AUCs and C-indices of MIRAI, after 331 finetuning, reaches its own optimal, and outperforms the 5-year BCSC model. The 332 finetuned MIRAI only achieves similar performance to our image-only MTP-BCR model 333 from 3- to 5-year risk at both screening test sets (when only for women who can be scored 334 by the BCSC model) (Table S9). Yet our method with risk factors still surpasses the 335 MIRAI. 336

337 The ability of short and long future BC risk assessment

While the above results demonstrate the advantages of our methods in risk prediction, the 338 ability to predict real long-term future BC risk after eliminating the biases of cancer 339 detection and short-term to 5-year risk prediction from current mammograms requires 340 further exploration. Thus, we compare the models' performance in 5-year and 10-year risk 341 prediction in the different subgroups of the full inhouse test sets by excluding exams from 342 women diagnosed with cancer within less than 1, 3, and 5 years. These results, as shown 343 in Table 4, demonstrate that our methods could not only detect BC and improve the 344 performance of 5-year risk prediction compared with other SOTA methods but also learn 345 the features related to the real longer-term (10 years) risk. The results of replicate 5-year 346 BC risk prediction are shown in Table S10. 347

348 Clinical sub-group analysis

To distinguish how our MTP-BCR model performs in different populations and to 349 determine the potential population that can benefit most from it, we evaluate our risk 350 model in different clinical subgroup, based on age, breast density, personal history, and 351 future cancer sub-types (Fig. 4). We find that the MTP-BCR model performed similarly 352 across different density groups and independent from future cancer sub-types. The C-353 indices for the MTP-BCR model for women aged <40, 40-60, and 60-80 are 0.87 (95%) 354 CI, 0.84-0.90), 0.81 (95% CI, 0.79-0.83) and 0.82 (95% CI, 0.79-0.84), respectively, 355 which implies that our risk model performs better in younger cohorts. We also note that 356 the risk model performs best in the female population without any prior personal BC 357 history with C-index of 0.86 (95% CI, 0.84-0.87), compared to 0.76 (95% CI, 0.74-0.78), 358 for women with a personal history of breast cancer. We also compare the C-index of 10-359 year risk prediction of different methods in different sub-groups, which are available in 360 Table S3. Our results are further supported by the consistent performance from additional 361 subgroup analyses in recollected 5-year risk dataset (Tables S11, supplemental). Due to a 362 lack of race labels, the group analysis for different race groups is not performed. 363

364 **Consistency of model attention in longitudinal images**

365To investigate how risk-related areas evolve on multiple-time mammograms that our366MTP-BCR model focused on, we utilize the gradient-weighted class activation maps367(Grad-CAM) (36). Fig. 5 shows a visualization example of a BC patient. The heatmaps368highlight potentially related regions where our proposed MTP-BCR model identifies369predictive 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 fiveyear risk dataset (Fig. S5, supplemental).

376 Discussion

In this study, we develop a multi-time-point examination-based risk model, MTP-BCR, to 377 assess 10-year breast cancer risk on patient and unilateral breast level using longitudinal 378 screening mammograms and medical records. From extremely short-term risk (1-year, BC 379 detection) to long-term (10-year) risk prediction, MTP-BCR outperforms radiologists' BI-380 RADS assessment, a SOTA BC detection method (GMIC), an extensively clinically 381 validated single time point-based 5-year risk DL approach (MIRAI), and a traditional 382 clinical risk model (BCSC). Apart from the patient level risk, our method could also 383 estimate the risk base on unilateral breast level with comparable ability. Experiments on 384 different screening subcohorts suggest that the longitudinal assessment of MTP-BCR is 385 able to accurately identify longer-term, future risk-related features on mammograms, 386 which is further supported by consistent heatmaps of multi-time point mammograms. 387 Final sub-group analysis indicates that the proposed method performs consistently across 388 subgroups of different breast densities and for future types of BC. 389

The motivation to develop BC risk models is for guiding personalized screening or 390 triggering prevention regimens. The idea is to determine the screening frequency and the 391 appropriate screening modality based on the individual risk of women, potentially also to 392 recommend preventive therapy for women at a high-risk of developing breast cancer (13). 393 Based on of risk factors such as age, genetic determinants, family history, previous benign 394 biopsies, and recently considered image-based breast density, traditional risk models are 395 used to globally assess 5-year, 10-year, or lifetime risk for large groups of women (37). 396 However, our results show that density alone is not sufficient to represent all of the risk-397 related information within the mammograms. Moreover, ignoring the short-term risk of 398 BC limits the value of these models in early BC detection. Recent DL-based risk models 399 may fully utilize screening images but mainly target short-term risk prediction or interval 400 cancer detection while ignoring long-term risk, which limits the ability to offer 401 personalized screening recommendations and preventive interventions. In contrast our 402 MTP-BCR risk model combines the advantages of both short- and long-term risk 403 404 prediction strategies.

For short-term risk estimation, our risk model outperforms radiologists' BI-RADS and 405 other recent DL methods. The results on the full test set demonstrate that our MTP-BCR 406 risk model is more competitive than other risk models in indicating current or future BC 407 risk in a realistic complex clinical screening setting independent of radiological 408 409 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 410 that there is no existing cancer. Furthermore, in the normal BI-RADS screening group, our 411 risk model has the highest AUC of 0.74 for the 1-year risk prediction, which can be 412 regarded as an aid for radiologists to improve interval cancer detection in the whole 413 screening population, including primary and recurring cancers. When detecting the 414 primary interval cancers of the population that the BCSC model targets, our risk model 415 reaches an AUC of 0.87. Aiming to facilitate radiologists' understanding of the model 416

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 421 screening cohorts or subcohorts. Although the MIRAI model does not need to have the 422 risk factors as inputs and performs similarly to our image-only MTP-BCR model at 5-year 423 risk (P > 0.05) on the biopsy negative screening set and normal BI-RADS set, we note that 424 the MIRAI model involves a pretrained risk factors predictor enabling it to benefit from 425 the missing risk factors that reaches a similar performance of 5-year risk prediction as 426 MIRAI with risk factors (P = 0.27) (19). While our MTP-BCR model with risk factors is 427 significantly better than all others (All P values < 0.05). Besides, the MIRAI is already 428 trained on their large private MGH dataset. The training set includes 210,819 exams is 6.5 429 times larger than our Inhouse training set. Then we still fully finetuned the MIRAI with its 430 weight in our In-house training set without freezing the encoder. On the other hand, our 431 MTP-BCR model is only trained on the In-house dataset directly. As for long-term risk, 432 the MTP-BCR risk model is more accurate than the BCSC 10-year risk model, which 433 suggests our risk model has the potential for better decisions regarding a risk-adapted 434 screening regimen and preventive therapy. At breast level, risk could be the foundation of 435 more refined screening and prevention strategies. It should be underlined that in this study 436 we also involve patients with prior BC history, which are not included in most of the risk 437 models. Our risk model is, in fact, also designed to leverage the history information of 438 prior-tumor and therapy, which the recurrence risk models use (38), for recurrence cancer 439 risk prediction. 440

The promising performance of MTP-BCR can be attributed to its capacity to capture 441 unique BC risk-related characteristics. The multi-task learning strategy helps the model to 442 443 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 444 relationship between local (unilateral breast) level risk and global (patient) level risk while 445 keeping the local information as much as possible when combining the local features for 446 the summary of the global features. Thus, the MTP-BCR risk model can consistently focus 447 on similar regions on longitudinal mammograms without registration. For accurate longer-448 term BC risk prediction, we need not only the static risk-related features from the single 449 time point exam but also the dynamic features from the multi-time point exams to indicate 450 451 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 452 note that it has been restricted to a small image-only dataset and a setting of a fixed 453 number of input time points, which hinders its clinical application. On the contrary, our 454 MTP-BCR model is built based on an extensive clinical screening mammogram dataset, 455 enabling efficient use of risk factors, and has the flexibility to input 0-5 historical 456 reference exams. 457

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 464 465 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 466 standard screening mammography datasets from multiple hospitals. To validate the 467 detection of extremely early signs of BC, a pixel-level annotated dataset is necessary. 468 Moreover, a reader study for incorporation of the risk model in the radiologist workflow 469 may also be a future direction for further demonstrating the benefits of risk models for 470 471 personalization BC screening policy.

In conclusion, we propose a novel DL model using longitudinal mammogram 472 examinations and history obtained from medical records that outperforms the SOTA 473 MIRAI and traditional BCSC risk model by a large margin. The improvement is 474 consistent across screening and future risk subgroups. These results support the hypothesis 475 that longitudinal mammography contains informative spatiotemporal indicators of future 476 breast risk that cannot be captured by the single-time point DL models. Multi-time point 477 models based on longitudinal analysis strategies have the potential to replace single-time 478 point based risk prediction models. Apart from increasing the accuracy for BC risk 479 prediction, we also improve the interpretability of our risk model, which could potentially 480 accelerate the translation of personalized AI-based risk stratification into routine BC 481 screening policies. 482

483

484 Materials and Methods

485 **Data collection**

Our retrospective study was approved by the Institutional Review Board (IRB) of 486 Netherlands Cancer Institute (NKI) with protocol numbers: IRBd21-060. A flowchart 487 illustrating the construction of this large study dataset is shown in Fig. 2. We collect 488 37,517 patients recorded in our hospital between January 1, 2004, and December 31, 489 490 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). 491 Details about the distributions of the dataset are available in Fig. S1. Although part of the 492 patients did not have 10-year screening follow-up, we also leverage their known outcomes 493 and images to supervise the model. Therefore, we keep 9,133 patients consisting of 2,562 494 BC patients who were biopsy-proven within 10 years and 6,571 at intermediate risk who 495 had at least 10 years of screening follow-up and did not receive a cancer diagnosis. All 496 patients are randomly divided into training, validation, and test sets with a ratio of 497 498 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. 499

BC-relevant risk factors are already showing an essential role in both traditional (12, 15) 500 and image-based DL methods (16, 19). In our study, we collect the risk factors through 501 electronic medical records from clinical radiology, tumor, therapy, and pathology reports 502 from our hospital. The distribution of clinical risk factors in the inhouse dataset is shown 503 in Table 1. Specifically, we obtain age, race, BI-RADS and breast density (ACR) grade, 504 family history, genetic determinants, previous BC history, previous ovarian cancer history, 505 self-reported menopausal status, and age of menarche. The BI-RADS and breast density 506 (ACR) grade are estimated by radiologists during clinical interpretation. BI-RADS grades 507 include additional imaging required (BI-RADS 0), normal (BI-RADS 1), benign (BI-508 RADS 2), probably benign (BI-RADS 3), suspicious for malignancy (BI-RADS 4), highly 509 suggestive of malignancy (BI-RADS 5), and known biopsy-proven malignancy (BI-RADS 510

6). ACR class include mostly composed of fatty tissue (ACR 1), scattered fibroglandular 511 512 tissue (ACR 2), heterogeneously dense (ACR 3), and extremely dense (ACR 4). Images with missing densities are interpolated by nearest neighbour interpolation with reference 513 to the density estimates of screening images from adjacent years of the patient. Because 514 the weights of patients are missing, we did not calculate the body mass index (BMI). In 515 our study, we also include patients with prior BC. Thus, following the research of 516 recurrence risk prediction (38), we also leverage the information of prior tumor, which 517 include pathologic tumor (pT)-stage, pathologic node (pN)-stage, hormone receptor status 518 (estrogen receptor (ER)- and progesterone receptor (PR)-status), anti-hormonal therapy, 519 human epidermal growth factor receptor 2 (HER2-status), type of surgery, adjuvant 520 chemotherapy, adjuvant radiation therapy, antibody therapy and Pathologic Complete 521 Response (pCR). 522

523 **Problem formulation**

528

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:

$$Risk_{j} = \sum_{i=1}^{j} y_{i} = \sum_{i=1}^{j} Softmax (F(x))_{i}$$
(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 = 2years from the checked images, it can be calculated as the sum of the probability of the first year and second year.

535 Architectural details

536As shown in Fig. 1, the MTP-BCR model consists of the network weights shared Image537Encoder ($\varphi_{encoder}$) connected by Side-Specific (Unilateral-based) Module538($\varphi_{unilateral}$), Exam-based Module (φ_{exam}), and finally, Multi-Time-Point fusion539Module (φ_{fusion}) combines with the inputted risk factors of patients. Moreover, to540improve risk modeling performance and generalization, we also introduce multi-task541learning, which could benefit from learning the domain-specific features from multiple542BC risk-related tasks (detailed below).

Image Encoder: We employ ImageNet pre-trained ResNet-18, excluding the last full 543 connection (FC), as the encoder ($\varphi_{encoder}$) to extract breast tissue features. The weights-544 shared encoders correspond to each image from a sequence of mammography exams. 545 Each exam includes four images including craniocaudal (CC) view (v = cc) and 546 mediolateral oblique (MLO) view (v = mlo) from the left(l = left) and right(l = right) 547 side breast. Thus, shown in the Eq. 2, each input mammogram, $x_{v,l}^t$, from the 6 time point 548 (t) exams is represented as the high-dimensional locally feature vector $\theta_{\nu,l}^t$ with the size of 549 512×1 by the encoder separately. 550

551
$$\theta_{v,l}^{t} = \varphi_{encoder}(x_{v,l}^{t}), v \in \{cc, mlo\}, l \in \{right, left\}, t \in \{0, 1, 2, 3, 4, 5\}$$
(2)

12 / 28

Side-Specific Prediction Module: Ipsilateral CC and MLO views are different projection 552 553 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 554 tumors only appear in one of the breasts (40). In order to train the model to learn the three-555 dimensional structure of the breast fully and correspond to the previous left and right 556 breast-specific tumor information, we concatenate (\bigoplus) the feature vectors of the 557 ipsilateral view, combined with the side-specifically prior tumor information $(tumor_i^t)$ 558 and then place a Multi-layer Perceptron (MLP) for side-based multi-task learning. The 559 MLP layer includes two FC layers with an input size of 1152 for the first FC layer and 512 560 output units each. A dropout layer with a rate of 0.5 between the two FC layers. Therefore, 561 show in Eq. 3 features of the ipsilateral CC and MLO views and outputs a vector (ε_l^t) with 562 the size of 512×1 representing the unilateral-based breast features. 563

$$\varepsilon_{l}^{t} = \varphi_{unilateral}(tumor_{l}^{t} \oplus \theta_{l,\nu=cc}^{t} \oplus \theta_{l,\nu=mlo}^{t}), l \in \{right, left\}, t \in \{0, 1, 2, 3, 4, 5\}$$
(3)

565 **Exam-Based Prediction Module:** For a similar purpose, we also need to combine the 566 information of bilateral breasts to predict the patient level risk. As in Eq. 4, we 567 concatenate the feature vectors from the output bilateral breasts and feed them to another 568 MLP layer with the same structure for the exam-based multi-task learning. Also, a size of 569 512×1 vector feature (δ^t), which combines the features from right and left breast, 570 represents the global information of a four-view exam.

571

594

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

$$\tag{4}$$

Multi-Time Point Fusion Model: For learning the risk development pattern from the 572 longitudinal screening mammograms, five historic exams before the current exam are 573 randomly selected as the reference for the comparison by the multi-time point fusion 574 575 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 576 (i_0, i_1, \dots, i_5) to the current exam and are combined with the risk factors to predict the 577 future likelihood of BC occurring after the current mammograms. Inspired by the research 578 (26), we leverage a sequence/time-aware transformer learning (41) to capture features 579 about the temporal relations between multiple mammograms, which aims to disentangle 580 the risk-relevant changing patterns from the normal breast tissue changing patterns. For 581 embedding the spatiotemporal relationships of past-current exams into the continuous 582 latent space, we employ Continuous Position Embedding (CPE) method (26), which 583 computes time continuous embedding e^t to condition the image features. Not that, to 584 avoid ignoring local information of images during multi-time exam comparison, we 585 combine both the local image features $\theta_{v,l}^t$, and global features δ^t . For patients without 586 five history records for references, we select all history records and then mute the missing 587 data by filling 0. At the same time, for the purpose of data augmentation, we randomly 588 drop a subset of exams in the reference sequence to improve model robustness and avoid 589 overfitting. The fusion model also includes the patient risk factors (riskf). Subsequently, 590 a fused feature (τ , a vector size of 640 \times 1) is obtained for the final multi-time fused-591 based multi-task learning, representing the patient's multi-time point screening 592 information. 593

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

Multi-task classifier for side-based, exam-based, and multi-time fused prediction:

596 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 597 BC risk, history, tumor location, and tumor sub-type are included for the side-based multi-598 task classifier. For the exam-based multi-task classifier, we replace the breast-based BC 599 risk prediction and history prediction with exam-based prediction. We also add the 600 prediction of age, BI-RADS, race, density, and manufacture. And for the final multi-time 601 fused classifier, we mainly focus on unilateral specific BC risk. We calculate the Binary 602 Cross Entropy (BCE) for risk prediction and Cross Entropy (CE) Loss for other 603 predictions. For all three classifiers, we mainly focus on the task of risk prediction thus 604 risk-specific tasks have 5 times higher weight than other tasks during the training. For 605 total loss computing, Eq. 6, we also allocate weight $w_{fusion} = 1$ to the final multi-time 606 point fused classifier, 5 times higher than the other two classifiers ($w_{side} = 0.2$, $w_{exam} =$ 607 0.2). We choose the weights of loss after the hyperparameter search. 608

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

610 **Implementation details**

We use ResNet-18 initializing with ImageNet pre-trained weights as the backbone of all 611 our methods. All methods are implemented in PyTorch (version 1.12.1) with the same 612 training strategies. We use the Adam (42) optimizer and a rate of 0.5 for dropout (43) after 613 every fully connected layer. We train models for 20 epochs with a batch size of 8 and an 614 initial learning rate of 10^{-4} . The learning rate is decayed by a factor of 10 every five 615 epochs. The best models of each method are chosen with the best AUC performance index 616 on the validation set. The experiments are performed on a Ouadro A6000 GPU (48GB). 617 The source code is available at https://github.com/Netherlands-Cancer-Institute/MTP-618 BCR. Mammograms with standard DICOM format are pre-processed before being fed 619 into the model. First, we convert the images into 16-bit PNG format and segment the 620 whole breast region to exclude the background. Then, to unify the size of all images, we 621 zero-pad and resize images to 512 by 1024 pixels while retaining the relative scale and 622 aspect ratio. Finally, the image is normalized using the min-max method. We also employ 623 standard data augmentation techniques (i.e., random flip, brightness, and contrast) during 624 training for model robustness and overfitting prevention. 625

626 Evaluation metrics and statistical analysis

In this study, the tasks of prediction of 1- to 10-year risk are categorical classification 627 tasks, in which positive samples are the patients diagnosed with BC within 1 to 10 years 628 while negative samples are women who stayed healthy for at least 1 to 10 year-screening 629 follow-ups. The performances of the different methods are evaluated by the area under the 630 receiver operating characteristic curve (AUC, calculated by scikit-learn, version: 1.1.2, 631 https://scikit-learn.org). To generally evaluate AUCs across all times (from 1- to 10-year 632 risk), the Uno's C-index (44) is calculated using scikit-survival (version 0.18.0, 633 https://scikit-survival.readthedocs.io/en/stable/). The 95% confidence intervals (CI) of 634 AUC and C-index matrices are estimated by bootstrapping with 1,000 bootstraps for each 635 measure. Statistical significance among different methods is assessed using DeLong's test 636 (45), with the significant level predefined as P < 0.05. 637

638

595

- 639
- 640

641 **References**

- 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).
- L. E. Pace, N. L. Keating, A systematic assessment of benefits and risks to guide breast cancer screening decisions. *JAMA* 311, 1327–1335 (2014).
- 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).
- 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.*661 *Engl. J. Med.* 372, 2353–2358 (2015).
- 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).
- 668
 9. E. S. McDonald, A. S. Clark, J. Tchou, P. Zhang, G. M. Freedman, Clinical diagnosis and 669 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).
- I1. J. Tyrer, S. W. Duffy, J. Cuzick, A breast cancer prediction model incorporating familial
 and personal risk factors. *Stat Med* 23, 1111–1130 (2004).
- 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. 690 Kerlikowske, Using clinical factors and mammographic breast density to estimate breast 691 cancer risk: development and validation of a new predictive model. Ann. Intern. Med. 148, 692 337-347 (2008). 693
- 16. A. Yala, C. Lehman, T. Schuster, T. Portnoi, R. Barzilay, A deep learning mammography-694 based model for improved breast cancer risk prediction. Radiology 292, 60–66 (2019). 695
- 17. K. Dembrower, Y. Liu, H. Azizpour, M. Eklund, K. Smith, P. Lindholm, F. Strand, 696 697 Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. Radiology 294, 265–272 (2020). 698
- 18. M. Eriksson, K. Czene, F. Strand, S. Zackrisson, P. Lindholm, K. Lång, D. Förnvik, H. 699 700 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). 701
- 19. A. Yala, P. G. Mikhael, F. Strand, G. Lin, K. Smith, Y.-L. Wan, L. Lamb, K. Hughes, C. 702 703 Lehman, R. Barzilay, Toward robust mammography-based models for breast cancer risk. Sci. Transl. Med. 13. eaba4373 (2021).

704

705

706

707

708

709

710

715

716

717

718

719

720

721

- 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, 711 N. Karssemeijer, C. H. van Gils, I. Sechopoulos, R. M. Mann, C. J. V. Rooden, Interval 712 cancer detection using a neural network and breast density in women with negative 713 screening mammograms. Radiology 303, 269–275 (2022). 714
 - 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. 723 Hylton, M. M. Glymour, J. Kornak, L. H. Kushi, D. A. Lewis, V. X. Liu, D. L. 724 725 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 726 risk prediction. medRxiv pp. 2022–01 (2022). 727
- 26. A. Sriram, Muckley, K. Sinha, F. Shamout, J. Pineau, K. J. Geras, L. Azour, Y. 728 Aphinyanaphongs, N. Yakubova, W. Moore, Covid-19 prognosis via self-supervised 729 representation learning and multi-image prediction. arXiv preprint arXiv:2101.04909 730 (2021).731
- 27. M. Eriksson, K. Czene, C. Vachon, E. F. Conant, P. Hall, Long-term performance of an 732 image-based short-term risk model for breast cancer. J. Clin. Oncol. pp. JCO-22 (2023). 733
- 28. Liu, F. Zhang, C. Chen, S. Wang, Y. Wang, Y. Yu, Act like a radiologist: Towards 734 reliable multi-view correspondence reasoning for mammogram mass detection. IEEE 735 Trans. Pattern Anal. Mach. Intell. 44, 5947–5961 (2021). 736
- 29. Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. 737 738 Cho, K. J. Geras, An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med Image Anal. 68, 101908 (2021). 739

- 30. A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, C. Rudin, A case-based 740 interpretable deep learning model for classification of mass lesions in digital 741 mammography. Nat. Mach. Intell. 3, 1061–1070 (2021). 742
- 31. C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, R. Barzilay, 743 Mammographic breast density assessment using deep learning: clinical implementation. 744 Radiology 290, 52–58 (2019). 745
- 32. D. A. Spak, J. Plaxco, L. Santiago, M. Dryden, B. Dogan, Bi-rads® fifth edition: A 746 747 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. 748 Jacobson, Automated annotation and classification of bi-rads assessment from radiology 749 750 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). 752
- 753 35. X. Qian, J. Pei, H. Zheng, X. Xie, L. Yan, H. Zhang, C. Han, X. Gao, H. Zhang, W. Zheng, O. Sun, L. Lu, K. K. Shung, Prospective assessment of breast cancer risk from 754 multimodal Multiview ultrasound images via clinically applicable deep learning. Nat. 755 Biomed. Eng. 5, 522–532 (2021). 756
 - 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 ölkel, 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, O. Yang, A survey on multi-task learning. *IEEE Trans Knowl Data Eng.* 34, 768 769 5586-5609 (2021).
- 40. C. Wang, J. Li, F. Zhang, X. Sun, H. Dong, Y. Yu, Y. Wang, Bilateral asymmetry guided 770 counterfactual generating network for mammogram classification. IEEE Trans Image Process 30, 7980–7994 (2021) 772
- 41. N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image 773 transformer. International conference on machine learning (PMLR, 2018), pp. 4055– 774 775 4064.
- 42. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv preprint 776 arXiv:1412.6980 (2014).
- 43. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving 778 neural networks by preventing co-adaptation of feature detectors. arXiv preprint 779 arXiv:1207.0580 (2012). 780
- 44. H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, L.-J. Wei, On the c-statistics for 781 evaluating overall adequacy of risk prediction procedures with censored survival data. 782 Stat. Med. 30, 1105–1117 (2011). 783
- 45. X. Sun, W. Xu, Fast implementation of delong's algorithm for comparing the areas under 784 correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* 21, 1389– 785 1393 (2014). 786
- 787 788

751

757

758

759

760

761

762

763

764

765

766

767

771

777

790 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.
- 801 802 803

794

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



- 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.

847 848

851

849 850 mammography examinations into different demographic subgroups (rows) and different health subgroups (columns). The reported percentages are the number of examinations as a proportion of

Table 1. Detailed demographics for the Inhouse dataset. We categorize the number 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)
Tatal							
1 otal:		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%)
A							
nge.	<40	4,473	4,290	327	492	675	4,965
	40.50	(12.2%)(90.1%) 9,371	(12.3%)(86.4%) 8,983	(8.7%)(6.6%) 850	(7.8%)(9.9%) 1,405	(8.4%)(13.6%) 1,793	(11.6%)(100%) 10,776
	40-50	(25.7%)(87.0%)	(25.8%)(83.4%)	(22.7%)(7.9%)	(22.4%)(13.0%)	(22.3%)(16.6%)	(25.2%)(100%)
:	50-60	(29.0%)(85.6%)	(28.8%)(81.1%)	(29.2%)(8.8%)	(28.3%)(14.4%)	(29.2%)(18.9%)	(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	3,070	360	630	772	3,842
	<u>\</u> 80	(8.8%)(85.6%) 279	(8.8%)(79.9%) 273	(9.6%)(9.4%) 102	(10.0%)(10.4%) 152	(9.6%)(20.1%) 158	431
	200	(0.8%)(64.7%)	(0.8%)(63.3%) 1 308	(2.7%)(23.7%) 109	(2.4%)(35.3%)	(2.0%)(36.7%) 172	(1.0%)(100%) 1 480
Un	hknow	(3.6%)(88.6%)	(3.8%)(88.4%)	(2.9%)(7.4%)	(2.7%)(11.4%)	(2.1%)(11.6%)	(3.5%)(100%)
BI-RADS:							
BLBA	0 204	306	296	39	59	69	365
DIR		(0.8%)(83.8%) 6,126	(0.9%)(81.1%) 5,913	(1.0%)(10.7%) 66	(0.9%)(16.2%) 290	(0.9%)(18.9%) 503	(0.9%)(100%) 6,416
BI-RA	ADS I	(16.8%)(95.5%)	(17.0%)(92.2%)	(1.8%)(1.0%)	(4.6%)(4.5%)	(6.3%)(7.8%)	(15.0%)(100%)
BI-RA	ADS 2	25,363 (69.4%)(91.1%)	24,174 (69.5%)(86.8%)	(16.1%)(2.2%)	(39.7%)(8.9%)	(45.9%)(13.2%)	(65.1%)(100%)
BI-RA	ADS 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-RA	ADS 4	242 (0.7%)(28.2%)	230	590 (15.8%)(68.8%)	616 (9.8%)(71.8%)	628 (7.8%)(73.2%)	858 (2.0%)(100%)
BI-RA	ADS 5	21	21	826	828	(1.8%)(75.2%) 828	(2.0%)(100%) 849
		(0.1%)(2.5%) 0	(0.1%)(2.5%) 0	(22.1%)(97.3%) 917	(13.2%)(97.5%) 917	(10.3%)(97.5%) 917	(2.0%)(100%) 917
DI-KA	1050	(0.0%)(0.0%) 3.412	(0.0%)(0.0%) 3.144	(24.5%)(100%) 456	(14.6%)(100%) 737	(11.4%)(100%) 1.005	(2.1%)(100%) 4.149
	None	(9.3%)(82.2%)	(9.0%)(75.8%)	(12.2%)(11.0%)	(11.8%)(17.8%)	(12.5%)(24.2%)	(9.7%)(100%)
Density:							
А	ACR 1	2,340	2,250	208	355	445	2,695
		(6.4%)(86.8%) 21,163	(6.5%)(83.5%) 20,051	(5.6%)(7.7%) 2,308	(5.7%)(13.2%) 3,829	(5.5%)(16.5%) 4,941	(6.3%)(100%) 24,992
A	ACK 2	(57.9%)(84.7%) 9 782	(57.7%)(80.2%) 9 385	(61.7%)(9.2%) 812	(61.1%)(15.3%) 1 448	(61.6%)(19.8%) 1 845	(58.4%)(100%) 11.230
A	ACR 3	(26.8%)(87.1%)	(27.0%)(83.6%)	(21.7%)(7.2%)	(23.1%)(12.9%)	(23.0%)(16.4%)	(26.2%)(100%)
А	ACR 4	3,137 (8.6%)(85.7%)	2,987 (8.6%)(81.6%)	(8.3%)(8.5%)	523 (8.3%)(14.3%)	673 (8.4%)(18.4%)	3,660 (8.6%)(100%)
un	nknow	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%)
A	frican	141	135	11	19	25	160
	Acion	(0.4%)(88.1%) 270	(0.4%)(84.4%) 265	(U.3%)(6.9%) 9	(0.3%)(11.9%) 26	(0.3%)(15.6%) 31	(0.4%)(100%) 296
	Asidii	(0.7%)(91.2%) 94	(0.8%)(89.5%) 90	(0.2%)(3.0%) 7	(0.4%)(8.8%) 12	(0.4%)(10.5%) 16	(0.7%)(100%) 106
Other	r Race	(0.3%)(88.7%)	(0.3%)(84.9%)	(0.2%)(6.6%)	(0.2%)(11.3%)	(0.2%)(15.1%)	(0.2%)(100%)
Un	nknow	25,197 (69.0%)(84.1%)	24,146 (69.4%)(80.6%)	5,056 (81.7%)(10.2%)	4,767 (76.0%)(15.9%)	5,818 (72.5%)(19.4%)	29,964 (70.0%)(100%)
							Continued Table 1
Gene:							
BRCA 1 mu	itation	1,321	1,229	103	256	348	1,577

(3.6%)(83.8%)

(3.5%)(77.9%)

(2.8%)(6.5%)

(4.1%)(16.2%)

(3.7%)(100%)

(4.3%)(22.1%)

BRCA 2 mutation	BRCA 2 mutation 1,183 1,115 (3.2%)(86.3%) (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	2,653	367 (9.8%)(9.2%)	852	1,319	3,972
Unknow	30,680	29,587	3,178	4,921	6,014	35,601
UIKIOW	(84.0%)(86.2%)	(85.1%)(83.1%)	(84.9%)(8.9%)	(78.5%)(13.8%)	(75.0%)(16.9%)	(83.2%)(100%)
Menopausal Status:						
Pre-menopausal	5,570	5,333	431	707	944	6277
Peri-menopausal	(13.5%)(88.7%) 18,599	(13.3%)(83.0%) 17,825	(11.3%)(0.9%)	2,792	3,566	21,391
& Unknow	(50.9%)(86.9%)	(51.3%)(83.3%)	(47.3%)(8.3%)	(44.5%)(13.1%)	(44.4%)(16.7%)	(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	18 149	17 058	985	2,656	3 747	20.805
Positive	(49.7%)(87.2%)	(49.1%)(82.0%)	(26.3%)(4.7%)	(42.4%)(12.8%)	(46.7%)(18.0%)	(48.6%)(100%)
Breast Cancer	18,373	17,711	2,757	3,614	4,276	21,987
Negative Overien Concer	(50.3%)(83.6%)	(50.9%)(80.6%)	(73.7%)(12.5%)	(57.6%)(16.4%)	(53.3%)(19.4%)	(51.4%)(100%)
Positive	401 (1.1%)(79.2%)	308 (1.1%)(72.7%)	42 (1.1%)(8.3%)	(1.7%)(20.8%)	(1.7%)(27.3%)	(1.2%)(100%)
Ovarian Cancer	36,121	34,401	3,700	6,165	7,885	42,286
Negative	(98.9%)(85.4%)	(98.9%)(81.4%)	(98.9%)(8.7%)	(98.3%)(14.6%)	(98.3%)(18.6%)	(98.8%)(100%)
Menarche Age:						
<12	854	819	30	78	113	932
12	(2.3%)(91.6%) 4 981	(2.4%)(87.9%) 4 744	(0.8%)(3.2%)	(1.2%)(8.4%)	(1.4%)(12.1%) 758	(2.2%)(100%) 5 502
12-15	(13.6%)(90.5%)	(13.6%)(86.2%)	(5.4%)(3.7%)	(8.3%)(9.5%)	(9.4%)(13.8%)	(12.9%)(100%)
>15	(0.6%)(90.8%)	(0.6%)(88.8%)	(0.2%)(3.6%)	(0.4%)(9.2%)	(0.3%)(11.2%)	(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%)
		(0000000)(0000000)				
Family History:						
Breast Cancer	25,687	24,026	2,908	5,191	6,852	30,878
Positive Breast Cancer	(70.3%)(83.2%)	(69.1%)(77.8%)	(77.7%)(9.4%)	(82.8%)(16.8%)	(85.4%)(22.2%)	(72.2%)(100%)
Negative	(29.7%)(90.9%)	(30.9%)(90.2%)	(22.3%)(7.0%)	(17.2%)(9.1%)	(14.6%)(9.8%)	(27.8%)(100%)
Ovarian Cancer	2,155	2,009	150	364	510	2,519
Positive Ovarian Cancer	(5.9%)(85.5%)	(5.8%)(79.8%)	(4.0%)(6.0%)	(5.8%)(14.5%)	(6.4%)(20.2%)	(5.9%)(100%)
Negative	(94.1%)(85.3%)	(94.2%)(81.3%)	(96.0%)(8.9%)	(94.2%)(14.7%)	(93.6%)(18.7%)	(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	8,870	1,352	2,072	2,745	11,615
Salania Dimonsions	21184	20,539	(30.1%)(11.0%) 1,100	2,465	(34.2%)(25.6%) 3,110	23,649
	(58.0%)(89.6%)	(59.1%)(86.8%)	(29.4%)(4.7%)	(39.3%)(10.4%)	(38.8%)(13.2%)	(55.3%)(100%)
Train Val Test:						
Train	27,294	25,991	2,853	4,755	6,058	32,049
118111	(74.7%)(85.2%)	(74.8%)(81.1%)	(76.2%)(8.9%) 378	(75.8%)(14.8%)	(75.5%)(18.9%) 833	(74.9%)(100%) 4 432
Valid	(10.4%)(85.4%)	(10.4%)(81.2%)	(10.1%)(8.5%)	(10.3%)(14.6%)	(10.4%)(18.8%)	(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%)

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).

				MIRAI	MIRAI	MTP-BC	R (Ours)	MTP-BC	R (Ours)
	BI-RADS	BCSC*	GMIC	Test	Finetune	Patient	Level	Unilateral E	Breast Level
Use Risk									
Factors	-	Yes	No	-	-	No	Yes	No	Yes
Full inhouse	test set: 6,311 e	exams, 511 foll	owed by cance	r diagnosis wit	hin 1 years; 86	59 diagnosis wi	thin 5 years; 1,	132 diagnosis	within 10
years.									
5-Year	0.71	0.63	0.7	0.67	0.75	0.78	0.84	0.78	0.82
C-Index	(0.70 - 0.73)	(0.61-0.65)	(0.68 - 0.72)	(0.65 - 0.69)	(0.73-0.77)	(0.76 - 0.80)	(0.82 - 0.85)	(0.76 - 0.80)	(0.81 - 0.84)
10-Year	0.69	0.64	0.69	0.67	0.73	0.77	0.82	0.76	0.81
C-Index	(0.67-0.70)	(0.61-0.66)	(0.67-0.70)	(0.65-0.68)	(0.72-0.75)	(0.75-0.78)	(0.81-0.84)	(0.75-0.78)	(0.79-0.82)
1-Year	0.83	0.62	0.74	0.70	0.84	0.87	0.91	0.87	0.89
AUC	(0.81 - 0.85)	(0.58-0.64)	(0.72 - 0.77)	(0.68-0.73)	(0.81-0.86)	(0.85 - 0.89)	(0.89-0.92)	(0.85 - 0.89)	(0.87-0.91)
2-Year	0.78	0.62	0.72	0.69	0.80	0.83	0.88	0.83	0.87
AUC	(0.76 - 0.80)	(0.59-0.65)	(0.70 - 0.75)	(0.66 - 0.71)	(0.78 - 0.82)	(0.81 - 0.85)	(0.86 - 0.89)	(0.81 - 0.85)	(0.85 - 0.88)
3-Year	0.74	0.63	0.72	0.69	0.78	0.80	0.86	0.80	0.84
AUC	(0.72 - 0.76)	(0.61-0.66)	(0.69 - 0.74)	(0.67 - 0.72)	(0.76 - 0.80)	(0.79 - 0.82)	(0.84 - 0.87)	(0.79 - 0.82)	(0.83-0.86)
4-Year	0.72	0.64	0.71	0.69	0.76	0.79	0.84	0.79	0.82
AUC	(0.70 - 0.73)	(0.61-0.67)	(0.69 - 0.73)	(0.67 - 0.71)	(0.74 - 0.78)	(0.77 - 0.80)	(0.82 - 0.85)	(0.77 - 0.81)	(0.81 - 0.84)
5-Year	0.70	0.65	0.70	0.70	0.74	0.77	0.82	0.77	0.81
AUC	(0.68 - 0.71)	(0.62-0.68)	(0.68 - 0.72)	(0.68 - 0.72)	(0.72-0.76)	(0.75 - 0.79)	(0.81 - 0.84)	(0.75 - 0.79)	(0.79-0.82)
6-Year	0.68	0.66	0.70	0.70	0.73	0.77	0.82	0.76	0.80
AUC	(0.67 - 0.70)	(0.63-0.69)	(0.68 - 0.72)	(0.68 - 0.72)	(0.71 - 0.75)	(0.75 - 0.79)	(0.80 - 0.83)	(0.75 - 0.78)	(0.78-0.81)
7-Year	0.67	0.67	0.70	0.70	0.73	0.76	0.81	0.76	0.79
AUC	(0.66 - 0.69)	(0.65-0.70)	(0.68 - 0.72)	(0.68 - 0.72)	(0.71 - 0.75)	(0.74 - 0.78)	(0.80 - 0.83)	(0.74 - 0.77)	(0.77 - 0.80)
8-Year	0.66	0.69	0.69	0.69	0.72	0.76	0.80	0.75	0.78
AUC	(0.65 - 0.68)	(0.66-0.71)	(0.67 - 0.71)	(0.67 - 0.71)	(0.70 - 0.74)	(0.74 - 0.78)	(0.79-0.82)	(0.74 - 0.77)	(0.76 - 0.79)
9-Year	0.65	0.70	0.69	0.68	0.71	0.76	0.80	0.75	0.77
AUC	(0.64 - 0.67)	(0.67-0.73)	(0.67 - 0.71)	(0.66 - 0.70)	(0.69-0.73)	(0.74 - 0.78)	(0.78 - 0.81)	(0.74 - 0.77)	(0.75 - 0.78)
10-Year	0.65	0.71	0.69	0.67	0.71	0.77	0.80	0.76	0.77
AUC	(0.63-0.66)	(0.68-0.74)	(0.66-0.71)	(0.65-0.70)	(0.68-0.73)	(0.75-0.79)	(0.78-0.82)	(0.74-0.77)	(0.75-0.78)

869

- 870
- 871

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).

	DIDADO	Daga		MIRAI	MIRAI	MTP-BC	R (Ours)	MTP-BCR (Ours)	
U. D. D.	BI-RADS	BCSC *	GMIC	Test	Finetune	Patient	t Level	Unilateral E	Breast Level
Factors	-	Yes	No	-	-	No	Yes	No	Yes
Inhouse Biop	sy Negative Sc	reening Test se	et: 5,937 exams	s, 137 followed	by cancer dia	gnosis within 1	years; 495 dia	gnosis within 5	5 years; 758
diagnosis with	hin 10 years.	0		,		6	J	0	yy
5-Year	0.54	0.69	0.61	0.63	0.64	0.65	0.74	0.65	0.73
C-index	(0.53-0.56)	(0.65-0.72)	(0.59 - 0.64)	(0.60-0.65)	(0.62 - 0.66)	(0.63 - 0.68)	(0.72 - 0.76)	(0.63 - 0.67)	(0.71 - 0.75)
10-Year	0.53	0.69	0.61	0.62	0.64	0.65	0.74	0.65	0.72
C-index	(0.52-0.55)	(0.66-0.72)	(0.59-0.63)	(0.60-0.64)	(0.62-0.66)	(0.63-0.67)	(0.72-0.76)	(0.63-0.67)	(0.71 - 0.74)
1-Year	0.61	0.70	0.59	0.63	0.65	0.70	0.77	0.68	0.76
AUC	(0.57-0.65)	(0.63-0.76)	(0.54 - 0.65)	(0.58-0.68)	(0.60-0.70)	(0.65 - 0.74)	(0.73-0.81)	(0.62-0.73)	(0.71-0.81)
2-Year	0.58	0.69	0.61	0.62	0.64	0.67	0.75	0.66	0.75
AUC	(0.55 - 0.60)	(0.64-0.75)	(0.57 - 0.65)	(0.58-0.66)	(0.60-0.67)	(0.63 - 0.70)	(0.71-0.78)	(0.62 - 0.70)	(0.72-0.79)
3-Year	0.55	0.70	0.62	0.65	0.65	0.66	0.75	0.66	0.74
AUC	(0.54 - 0.58)	(0.65-0.74)	(0.59 - 0.65)	(0.62 - 0.68)	(0.62 - 0.68)	(0.63 - 0.69)	(0.72 - 0.77)	(0.63-0.69)	(0.71-0.77)
4-Year	0.54	0.70	0.63	0.64	0.65	0.66	0.74	0.66	0.72
AUC	(0.53-0.56)	(0.65-0.74)	(0.60-0.65)	(0.62-0.67)	(0.62-0.68)	(0.63-0.68)	(0.71-0.76)	(0.63-0.68)	(0.70 - 0.75)
5-Year	0.54	0.70	0.62	0.65	0.66	0.66	0.73	0.66	0.72
AUC	(0.52-0.55)	(0.05-0.74)	(0.60-0.65)	(0.63-0.68)	(0.63-0.68)	(0.63-0.68)	(0./1-0./6)	(0.63-0.68)	(0.69-0.74)
6-Year	0.53	0./1	0.63	0.65	0.67	0.66	0.74	0.00	0.72
AUC 7 Veen	(0.52-0.55)	(0.07-0.74)	(0.00-0.05)	(0.03-0.08)	(0.65-0.70)	(0.64-0.69)	(0./1-0./0)	(0.64-0.68)	(0.69-0.74)
/-Year	(0.53)	0.71	(0.60.0.65)	(0.62.0.68)	(0.65, 0.70)	(0.64.0.60)	0.74	(0.64.0.68)	0.71
AUC 8 Voor	0.52	(0.08-0.75)	(0.00-0.03)	(0.03-0.08)	(0.03-0.70)	(0.04-0.09)	(0.72-0.70)	(0.04-0.08)	(0.09-0.73)
	(0.53)	(0.69-0.75)	(0.60-0.64)	(0.62 - 0.67)	(0.65 - 0.70)	(0.65-0.69)	(0.73 (0.71-0.75)	$(0.64_{-}0.69)$	(0.68 - 0.72)
9-Year	0.52	0.73	0.62	0.64	0.68	0.68	0.73	0.67	0.69
AUC	(0.51-0.53)	(0.69-0.76)	(0.60-0.65)	(0.61-0.66)	(0.65-0.70)	(0.65-0.70)	(0.70-0.75)	(0.65-0.69)	(0.67-0.71)
10-Year	0.52	0.74	0.63	0.63	0.68	0.70	0.73	0.68	0.69
AUC	(0.51-0.53)	(0.70-0.77)	(0.60-0.66)	(0.60-0.66)	(0.66-0.71)	(0.67-0.72)	(0.71-0.75)	(0.66-0.70)	(0.67-0.72)
Inhouse Norn	nal BI-RADS S	Screening Test	set: 5,139 exar	ns, 102 follow	ed by cancer di	agnosis within	1 years; 404 d	iagnosis within	5 years;
612 diagnosis	within 10 yea	rs.							
5-Year	0.51	0.68	0.61	0.62	0.62	0.64	0.73	0.65	0.73
C-index	(0.50-0.51)	(0.64-0.72)	(0.59-0.64)	(0.60-0.65)	(0.59-0.64)	(0.62 - 0.67)	(0.71-0.75)	(0.63-0.68)	(0.71-0.75)
10-Year	0.50	0.69	0.61	0.62	0.62	0.64	0.73	0.65	0.73
C-index	(0.50-0.51)	(0.66-0.72)	(0.59-0.64)	(0.60-0.64)	(0.60-0.65)	(0.62-0.67)	(0.71-0.75)	(0.63-0.67)	(0.71-0.75)
1-Year	0.53	0.70	0.61	0.62	0.63	0.66	0.74	0.68	0.77
AUC	(0.50-0.55)	(0.63-0.77)	(0.54-0.67)	(0.57-0.68)	(0.57-0.68)	(0.60-0.71)	(0.69-0.79)	(0.63-0.73)	(0.72-0.81)
2-Year	0.51	0.69	0.60	0.62	0.61	0.64	0.73	0.6/	0.75
AUC	(0.50-0.53)	(0.03-0.75)	(0.56-0.65)	(0.58-0.66)	(0.57-0.65)	(0.60-0.68)	(0.69-0.76)	(0.63-0.71)	(0.72 - 0.79)
3-Year	0.51	0.70	(0.52)	(0.65)	(0.63)	0.64	0.73	(0.60)	0.74
AUC 4 Voor	(0.30-0.32)	(0.04-0.74)	(0.38-0.00)	(0.01-0.08)	(0.39-0.00)	(0.01-0.08)	(0.70-0.70)	(0.02-0.09)	(0./1-0.//)
	(0.50, 0.51)	(0.65, 0.74)	(0.60.0.66)	(0.61.0.67)	(0.03)	(0.62, 0.67)	(0.72)	(0.63.0.68)	(0.73)
5-Year	0.51	0.70	0.62	0.65	0.64	0.65	0.72	0.66	0.72
AUC	(0.50-0.51)	(0.65 - 0.74)	(0.59 - 0.65)	(0.62 - 0.68)	(0.61 - 0.67)	(0.62 - 0.67)	(0.70 - 0.75)	(0.63-0.68)	(0.69-0.74)
6-Year	0.50	0.71	0.63	0.66	0.66	0.65	0.73	0.66	0.72
AUC	(0.50-0.51)	(0.66-0.75)	(0.60-0.65)	(0.63-0.68)	(0.63-0.68)	(0.63-0.68)	(0.71-0.76)	(0.63-0.68)	(0.70 - 0.74)
7-Year	0.50	0.71	0.63	0.66	0.66	0.66	0.73	0.66	0.71
AUC	(0.50-0.51)	(0.67-0.75)	(0.60-0.66)	(0.63-0.69)	(0.64-0.69)	(0.63-0.68)	(0.71-0.76)	(0.64-0.68)	(0.69-0.73)
8-Year	0.50	0.72	0.63	0.66	0.67	0.67	0.73	0.67	0.71
AUC	(0.50-0.51)	(0.68-0.76)	(0.60-0.65)	(0.63-0.68)	(0.65-0.70)	(0.64-0.69)	(0.71-0.75)	(0.64-0.69)	(0.68-0.73)
9-Year	0.50	0.73	0.63	0.65	0.67	0.67	0.73	0.67	0.70
AUC	(0.50-0.51)	(0.69-0.76)	(0.60-0.66)	(0.62-0.68)	(0.64-0.70)	(0.65-0.70)	(0.70-0.75)	(0.65-0.69)	(0.67-0.72)
10-Year	0.50	0.74	0.63	0.64	0.67	0.69	0.73	0.68	0.70
AUC	(0.50-0.51)	(0.70-0.77)	(0.60-0.66)	(0.61-0.67)	(0.65 - 0.70)	(0.67 - 0.72)	(0.71-0.76)	(0.66 - 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.

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