Predicting breast cancer with AI for individual risk-adjusted MRI screening and early detection

Lukas Hirsch[1], Yu Huang[1], Hernan A. Makse[1], Danny F. Martinez[2], Mary Hughes[2], Sarah Eskreis-Winkler[2], Katja Pinker[2], Elizabeth Morris[2], Lucas C. Parra[1]*, Elizabeth J. Sutton[2,3]
1 City College of New York,
2 Memorial Sloan Kettering Cancer Center
3 University of California, Davis
* Corresponding author

Abstract

Objectives: Women with an increased life-time risk of breast cancer undergo supplemental annual screening MRI. We propose to predict the risk of developing breast cancer within one year based on the current MRI, with the objective of reducing screening burden and facilitating early detection.

Materials and Methods: An AI algorithm was developed on 53,858 breasts from 12,694 patients who underwent screening or diagnostic MRI and accrued over 12 years, with 2,331 confirmed cancers. A first U-Net was trained to segment lesions and identify regions of concern. A second convolutional network was trained to detect malignant cancer using features extracted by the U-Net. This network was then fine-tuned to estimate the risk of developing cancer within a year in cases that radiologists considered normal or likely benign. Risk predictions from this AI were evaluated with a retrospective analysis of 9,183 breasts from a high-risk screening cohort, which were not used for training. Statistical analysis focused on the tradeoff between number of omitted exams versus negative predictive value, and number of potential early detections versus positive predictive value.

Results: The AI algorithm identified regions of concern that coincided with future tumors in 52% of screen-detected cancers (60/115, CI: 42.7-61.6%). Upon directed review, a radiologist found that 71.3% of cancers (82/115, CI: 62.1-79.4%) had a visible correlate on the MRI prior to diagnosis, 65% of these correlates were identified by the AI model (53/82, CI: 53.3-74.9%). Reevaluating these regions in 10% of all cases with higher AI-predicted risk could have resulted in up to 33% early detections by a radiologist (56/167, CI: 26.4-41.2%). Additionally, screening burden could have been reduced in 16% of lower-risk cases (1,496/9,350, CI: 15.3-16.8%) by recommending a later follow-up without compromising current interval cancer rate.

Conclusions: With increasing datasets and improving image quality we expect this new AI-aided, adaptive screening to meaningfully reduce screening burden and improve early detection.

Key points

Retrospective analysis found that the machine identified higher-risk cases with regions of concern that coincided with future tumors in 52% of screen-detected cancers (60/115, CI: 42.7-61.6%). When re-evaluating 10% of the machine-predicted higher-risk cases, 33% of developing cancers can be detected one year early (56/167, CI: 26.4-41.2%). Following AI recommendations, 16% of patients can extend the interval of their next screening exam (1,496/9,350, CI: 15.3-16.8%).

Summary Statement

Predicting the short-term risk of developing breast cancer from MRI using AI has the potential to meaningfully improve early detection, while simultaneously reducing screening burden.

Introduction

A frequent question among women who undergo breast cancer screening is when they should return for their next examination? In the case of breast cancer supplemental screening with magnetic resonance imaging (MRI), this question is particularly pressing. In the U.S., more than 500,000 women undergo yearly supplemental screening breast MRI,[1] often beginning at 25-30 years of age, with some undergoing up to 40-50 MRIs in their lifetime. Women are enrolled in such supplemental screening breast MRI exams because they are at increased risk of breast cancer.[2] Yet, only a minority of these women will actually develop breast cancer in their lifetime. Indeed, the number of supplemental breast MRI screening is likely to increase given the recent recommendation to enroll women with extremely dense breasts.[3] based on the result of the DENSE[4,5] and ECOG-ACRIN[6] clinical trials. However, the incidence of breast cancer in this population is lower than in the high-risk population and thus many of these screening exams will remain negative.

There is an urgent unmet need for a new precision prediction model in women at increased risk of breast cancer. While risk can be stratified by considering individual genetic factors,[7] family history[8] or imaging information,[9,10] and there are a number of established risk models in use, (e.g., the Tyrer-Cuzick model or the updated Gail model).[11] The current risk models are largely static and do not take into account the current risk that may be gleaned from screening exams. Although a few models do include mammographic breast density as predictor,[12–14] MRI information is currently not used. Earlier efforts with deep-learning suggest that MRI has value in predicting risk of developing cancer within 5-years.[15] Thus, we suggest a new framework to determine an individual's probability of developing breast cancer within a defined period of time based on the current MRI. The goal of this framework is both early detection and individualized screening intervals based on risk.

After a negative screening mammogram, supplemental screening MRI can detect an additional 15-18 cancers per 1,000 high-risk women.[16] Due to the higher sensitivity of MRI over mammography,[17,18] we see a potential for estimating individual short-term risk within the high-risk screening population based on the most recent MRI. For instance, retrospective

studies suggest that 34-47% of detected cancers were present already in prior MRI exams.[19–21] In addition to local signs of cancers, one may be able to identify global features of general risk. For instance, fibroglandular breast tissue,[22] which is linked to the risk factor of breast density, can be assessed in MRI.[23] Background parenchymal enhancement may be an additional predictors of breast cancer risk.[24–27] But these MRI features, including dynamic contrast enhancement have not been systematically leveraged to predict individual risk.

Our proposed framework is based on the hypothesis that the current MRI exams already contain information about the outcome of the next yearly screening and AI will be able to predict the occurrence of breast cancer in the near future by evaluating the current breast MRI exam. We intend that the machine should evaluate all exams deemed normal and probably benign by the radiologists within the high-risk screening population, to suggest a longer follow-up period for lower-risk cases, and to suggest immediate follow-up for higher-risk cases. The objectives are 1) reduce the screening burden by identifying individuals who can be safely screened at longer intervals and 2) allow early detection by detecting cancers that are already present in the prior MRI. This could lead to decreased health care costs, minimized exposure to intravenous gadolinium, decreased false-positive biopsies and improved quality of life by decreasing anxiety.[28]

Methods:

Patient Sample

The evaluation used retrospective data from 12,694 women who underwent breast MRI at a tertiary Cancer Center in the United States, either for diagnostic or screening purposes, and who have been followed for up to 13 years (Fig. S2). The use of these retrospective data was approved by the institutional review board and the need for informed consent was waived, and all procedures were HIPAA compliant. All data had been previously de-identified by removing all patient information and saving exams with anonymized identifiers. The data included a total of 69,149 breasts with MRIs taken between 2002 and 2014. Complete data was available for 53,858 breasts (see Figure S2 for a full data chart). This sample size was the maximum accessible for this study.

Out of the 12,694 patients, 337 had screen-detected cancer. Exams from the preceding year were available for 193 of these cancers. To be precise, patients do not return necessarily within one year for their next screening. Here, any interval up to 15 months was considered “1 year” follow-up (see Table S1 for a summary of these numbers) while later follow-ups were excluded. Screening exams used for risk predictions included only normal or probably benign cases (BI-RADS ≤ 3, excluding BI-RADS 0), and were labeled (future) “benign” if the next scheduled screening yielded a BI-RADS ≤ 3 or a negative biopsy, otherwise they were labeled “malignant”. By this definition, there were 9,183 benign and 167 malignant breasts from the screening population. Ground-truth was defined from the outcome of clinical diagnosis and/or pathology.

Proposed framework of risk-adjusted screening

We first introduce here the proposed framework of risk-adjusted screening based on the current MRI. This framework involves using an AI-pipeline to review exams after the

radiologist has determined that the exam is cancer-free -- defined here as exams with a Breast Imaging Reporting and Data System (BI-RADS) assessment ≤ 3. The machine may predict with confidence that, in a year's time, the health status will not have changed (case #3 in Fig. 1). In this lower-risk case, the patient may be recommended to come back at a time point greater than a year. Otherwise, the machine may determine that there is a finite risk for the next exam to present with a malignant lesion (case #2). In this medium-risk case, the patient would be recommended to return to the regular yearly screen. Finally, the machine may predict a higher probability of developing cancer. In this higher-risk case, the recommendation may be to take another look at the exam and for the patient to undergo immediate assessment (case #1). If this is the case, the machine should point to the region of concern in the breast MRI and re-evaluation may possibly allow tumors to be detected at the time of the current MRI rather than at the next yearly MRI. In terms of current clinical practice, case #2 changes nothing; case #3 reduces the burden of screening, and case #1 has the potential to detect breast cancers earlier.

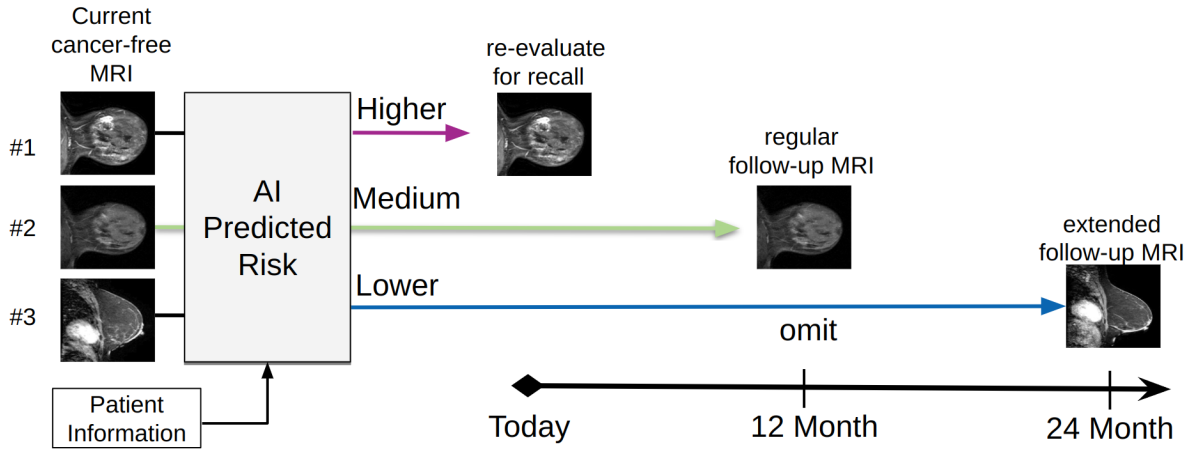

Figure 1: Framework of risk adjusted screening based on the current MRI exam. The risk of a future lesion is predicted based on the current MRI exam and patient information (age, family history, ethnicity, race). Example of three cases: The higher risk patient (#1) is referred to the radiologist for re-evaluation based on the identification of a suspicious lesion in the breast by the AI; medium risk patient (#2) is asked to return for regular follow-up MRI in a year; and the lower risk patient (#3) can skip one follow-up MRI and return for follow-up in two years.

In the proposed framework, the radiologists would evaluate the breast as usual, assigning a BI-RADS assessment to the breast. The machine would process only those breasts with BI-RADS 1 and 2 (Negative and Benign) and BI-RADS 3 (probably benign) assessments. Suspicious or highly suggestive of malignancy findings (BI-RADS 4 and 5, respectively), would proceed with biopsy recommendation without machine intervention. Higher risk prediction by the machine would prompt re-evaluation by a radiologist to decide whether to refer for additional imaging and/or biopsy. Lower risk prediction by the machine would also prompt a re-evaluation to decide on an extended screening interval. Given that we can only analyze retrospective data, we evaluated this proposed work-flow assuming that the radiologist accepts the machine's recommendation. While the machine does not take the radiologist assessment into account, the assessment does determine the overall clinical workflow.

The objective of reducing screening burden and the objective of early detection are not in opposition. We describe this in some detail in the Supplement (Fig. S1), but briefly, the two

criteria depend on two separate risk thresholds to determine lower and higher risk cases. As the lower-risk threshold is changed (square in Fig. S1A) the fraction of exams that could be omitted changes. This threshold also determines the precision of predicting health in a year's time changes, i.e. the negative predictive value (NPV, Eq. S1 in Supplement). The trade-off between NPV and reduced screening burden is captured by the resulting precision-effort curve (Fig. S1B). Similarly, as the higher-risk threshold changes (circle in Fig. S1A) the fraction of cases that need to be re-evaluated changes, along with the precision of predicting cancer in a year's time, i.e. the positive predictive value (PPV, Eq. S3). The trade-off between PPV and fraction that needs to be re-evaluated potential early detection is captured by the resulting precision-effort curve (Fig. S1C).

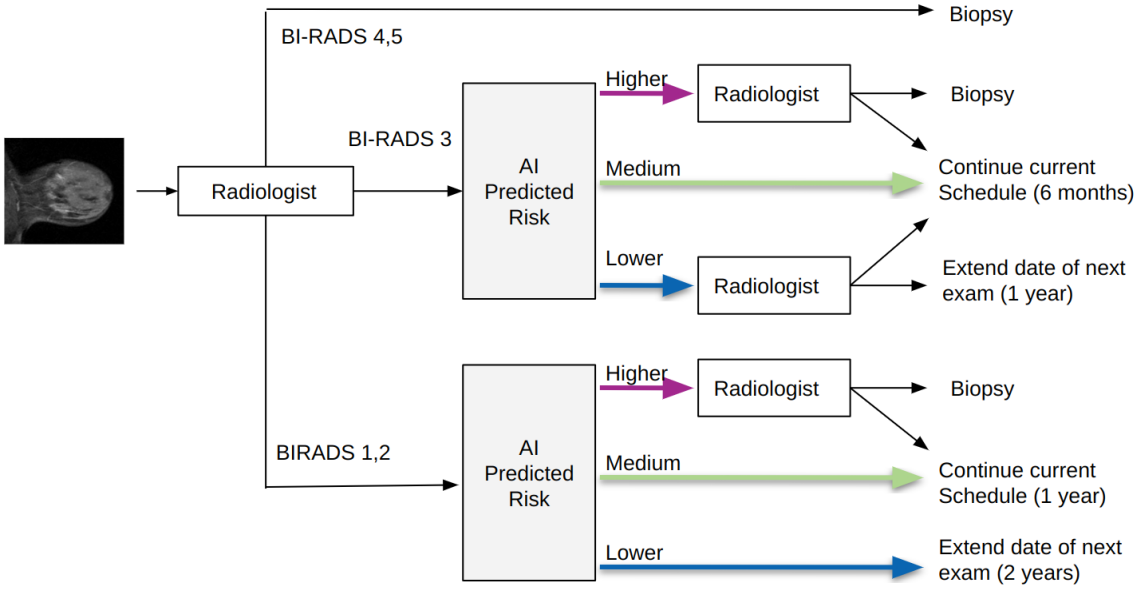

Figure 2: The proposed workflow integrates traditional radiologist's BI-RADS assessments and the AI predicted risk. Higher-risk patients are referred to the radiologist to evaluate suspicious lesions identified by the AI to consider possible biopsy. Medium-risk patients are asked to return for regular follow-up MRI (6 months for BI-RADS 3 and 12 months for BI-RADS 1 and 2). Finally, lower risk patients can extend the interval of the next scheduled MRI, after confirmation with the radiologist (12 months for BI-RADS 3 and 24 months for BI-RADS 1 and 2).

Development of a Risk Prediction Network

For a detailed description of the development of a risk prediction network, please refer to the Supplement. Note that the small number of screen-detected cancers in the screening population presents a clinical challenge.[29,30] It is also a challenge for machine learning which benefits from large datasets. To overcome this limitation the development of the risk prediction network was divided into three steps, with transfer learning used in each step. The first step involved using a *segmentation* network (developed previously[31]) to identify several regions of concern in each scan, i.e., regions with a high probability of belonging to a malignant lesion (see examples in Fig. 3A). Second, a *diagnostic network* was trained on a large dataset of breast MRIs including diagnostic and screening exams. Image features generated by the segmentation network from several regions of concern (n = 5) serve as input to this network to determine if the current exam has a biopsy-confirmed malignant lesion anywhere in the breast (Fig. 3B). Model development and selection was done using this diagnostic task. Third, a *risk prediction network* was obtained by fine-tuning the *diagnostic network*, to predict if a breast in the screening cohort that is currently cancer-free will or will not develop cancer within a year. Therefore, the diagnostic network was trained to

predict the immediate outcome of the current exam, while the risk prediction network, with identical architecture and input, was fine-tuned to predict outcomes of the next scheduled screening. This network outputs a prediction for the overall breast, based on the 5 most relevant regions as identified by the segmenter. Risk prediction is a much harder task than diagnosis as it operates on exams that the radiologist has deemed cancer-free, and the tumors become apparent only in a year's time in only 2% of cases.[32]

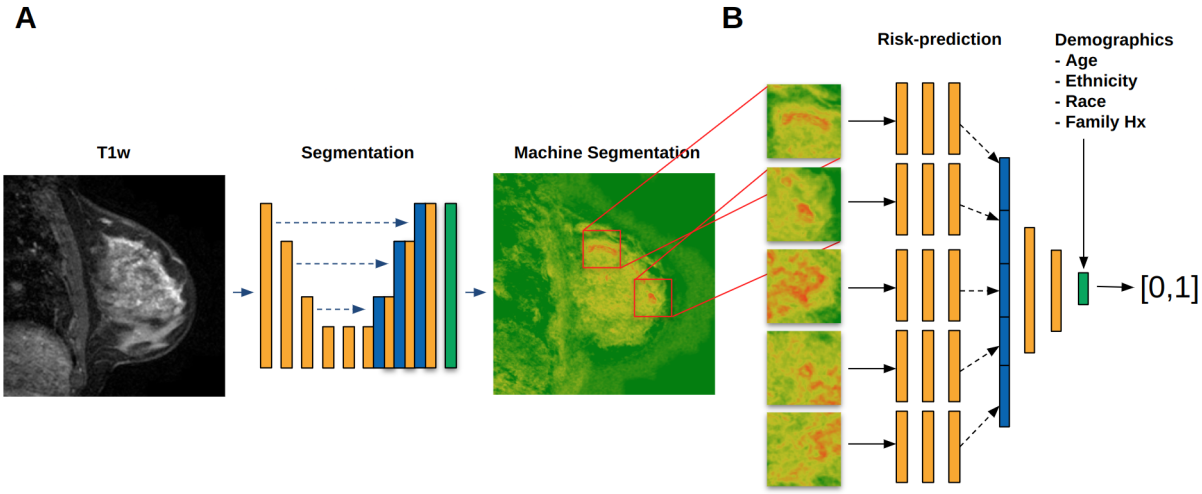

Figure 3: AI deep-network pipeline. A) A segmentation network (U-Net) extracts regions of concern from a breast MRI, which have a high likelihood of being part of a cancerous lesion. Here shown for an exam that develops cancer in a year . Two regions from a single slice are shown, but in general the regions of concern may come from different slices in the breast volume. Input to the risk-prediction network are features extracted by the segmentation network from aT1-weighted image (T1w) as well as dynamic contrast enhancement images. B) Architecture of the risk-prediction network (CNN) to predict exam outcomes. Convolutional layers (orange) include a 2D convolution, ReLU, batch-normalization and max-pooling. The feature maps from the 5 regions are concatenated (blue) and classified together with demographic information in a final dense layer with a sigmoid (green).

The segmentation network, previously developed,[31] was retrained on 30,253 breasts (including 1,220 segmented cancers). The diagnostic network used 9,006 breasts (including 2,331 cancers) for training and model selection. The risk prediction network was fine-tuned and evaluated with 5-fold cross-validation on a retrospective set of 9,350 MRI-screened breasts with 167 cancers (See Fig S3 and Table S1 for a summary and data partitions). The data from the training the segmentation and diagnostic networks came from 7,732 unique patients, while the data for fine-tuning the risk-prediction network came from a set of 4,962 patients. There is no overlap in exams between the two data sets. For acquisition, preprocessing and harmonization of the data as well as demographics see the Supplement.

Radiologist review of screen-detected tumors

A breast radiologist reviewed all MRIs for the 167 cancers in the risk-prediction test set. BI-RADS features could not be evaluated in 33 cases because the MRI at time of diagnosis was not available and 19 cases couldn't be evaluated due to post-lumpectomy change (n=5), axillary recurrence (n=3), post-treatment imaging (n=2), biopsy change obscured visualization (n=6) of there was no measurable disease (n=3). This left 115 cases for analysis of tumor location and BI-RADS features. To obtain an unbiased estimate of lesion size at both time points, we used automatic segmentation.[31] We selected a connected component at the location of the index lesion, and measured its length in the principal axis in

2D. We confirmed this metric on 4 cases measured by the radiologist using conventional clinical approach (see Fig. S10).

Results

Risk-adjusted screening

First we evaluated the feasibility of identifying individuals who can be safely screened at longer intervals on this retrospective cohort. We trained a risk prediction network to predict from the current cancer-free breast the outcome of the next scheduled screening (Fig. 3), i.e. predict 167 future screen-detections among 9,350 breasts. On this test set, the network achieves an area under the receiver operating characteristic curve (ROC-AUC) of 0.67 (CI: 0.63-0.70) (Fig. 4A). The next-year exams that could have been omitted are those below a low-risk threshold (Fig. 4B, square). As this threshold is increased, the number of omitted exams increases, while the negative predictive value (NPV) decreases (Fig. 4C). At an NPV of 100% (i.e., without missing a single future cancer), the network identified 3.44% (316/9,183) of breasts for which exams could have been omitted, corresponding to 316 next-year exams from 276 unique breasts. At an NPV of 99.5% (1,496/1,503, Fig. 4C, square), 16% (1,496/9,350, CI: 15.3-16.8%) of next-year exams could have been omitted (Fig. 4B, vertical arrow), with cancer in 7 of 9,350 breasts being missed. This corresponds to an interval cancer rate of 0.07%. In other words, 16% of next-year exams could be skipped while remaining below the interval cancer rate of 0.1% associated with the fixed, yearly screening schedule currently in use in the United States. Incidentally, had we omitted exams at random we would have missed 27 tumors (16% of 167 tumors), which is significantly more than with the AI recomendation ($Z=3.4$, $p=0.0007$).

Early detection

Next, we addressed the question of early detection. Cases that the network places above the highest-risk threshold could be referred for immediate follow-up (Fig. 4B, circle). As this threshold is reduced, the number of cancers potentially detected early increases, while the positive predictive value (PPV) drops (Fig. 4D). At a PPV of 25%, which is the current PPV of radiologists at our clinical site, the network recommends taking another look at 16 highest-risk cases. From these cases, 4 had a malignant exam in the following year. Compared to the total of 167 screen-detected cancers in the patient sample, this suggests that 2.4% (4/167, CI: 0.7-6.0%) of cancers might be detected one year earlier, at no additional cost compared to current clinical practice. If radiologists were to re-evaluate a larger fraction of high-risk cases for decision referral, the AI would have shown an enriched set with 1 tumor in every 20 cases (PPV=5%, Fig. 4D, circle) instead of 1 in 50 cases in the entire dataset. This would require reevaluation of 10% of all cases but it would flag 56 breasts that developed cancer within one year (Fig. 4D, vertical arrow), which is 33.5% (56/167, CI: 26.4-41.2%) of the total number of screen-detected cancers.

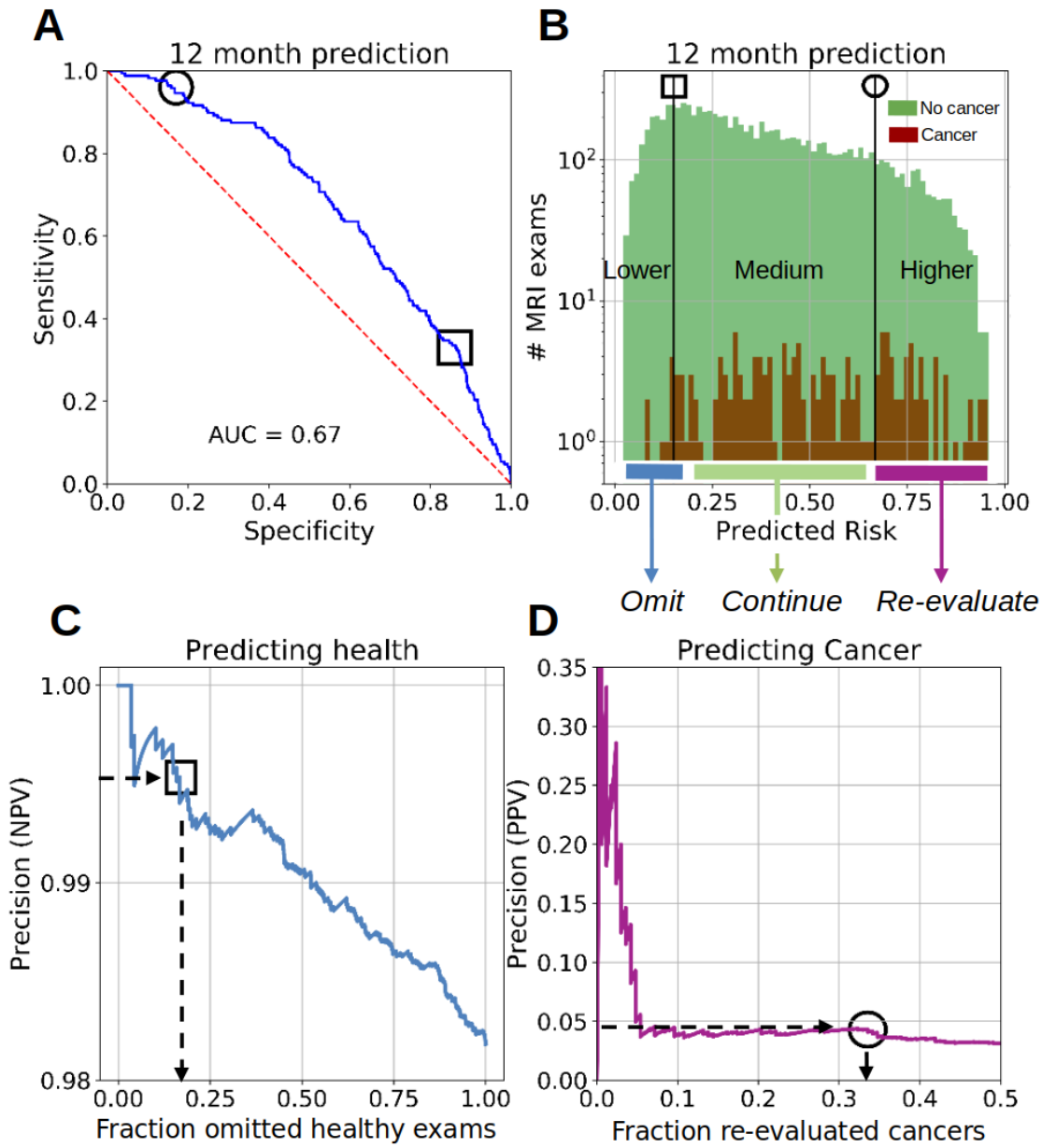

Figure 4: Prediction of developing breast cancer one year in advance from the current breast MRI in a clinical screening population. A) Distribution of future benign (green) and malignant (red) screening outcomes in terms of the AI algorithm-predicted risk based on the current cancer-free exam. Thresholds indicate suggested operating points for lower (square, risk=0.15) and higher risk (circle, risk=0.67). B) Cross-validation ROC curve for 12 month risk prediction (cross-validation performance). Operating points for higher and lower risk determination (sensitivity and specificity for circle: 96%, 17% and for square: 33%, 86%). C) Precision in predicting absence of cancer in a year's time versus fraction of exams that could have been omitted to reduce screening burden. D) Precision in predicting newly developed cancer in a year's time versus fraction of exams to be re-evaluated for early detection.

Tumor localization

The segmentation network highlights regions of concern in the breast (Fig. 3A), which could be used to direct the radiologist during re-evaluation. The risk-prediction network uses the top 5 regions detected by the segmentation network (Fig. 3B). We show examples of these

five suspicious regions for the 16 highest-risk cases (Fig. 5 for 4 future malignant cases, and Fig. S5 to S8 for 12 healthy cases). A trained breast radiologist reviewed these 16 highest-risk breasts and concurred that at least one of the regions selected by the network merits a biopsy. Of 167 screen-detected cancers, radiologist segmentations were available for 115 tumors of the next exam at time of diagnosis (For exclusion criteria, see Methods). We determined in this subset of cases whether one of the 5 regions of concern selected by the network overlapped with the radiologist segmentation. Examples of the regions for current breast and future malignancies are shown in Fig. 6. In 16% of cases (18/115, CI: 11.0-27.1%) the correct location is flagged and the breast receives a higher risk prediction (Fig. 6A). Remarkably, even when the predicted risk was lower, the location of the future cancer was correctly anticipated in about one third of all cancers (Fig. 6C). In total, the model correctly flagged the location of the future tumor in 52% (60/115, CI: 42.7-61.6%) of cases (Fig. 6A & C combined). There were a few cases where the regions of concern prompted a correct higher risk prediction (Fig. 6B) even though the future malignancy manifested elsewhere in the breast. Finally, there were a few cases with a lower risk estimate, and the region of concern did not match the location of the future tumor (Fig. 6D). These are genuinely hard cases with no obvious evidence of a future malignancy.

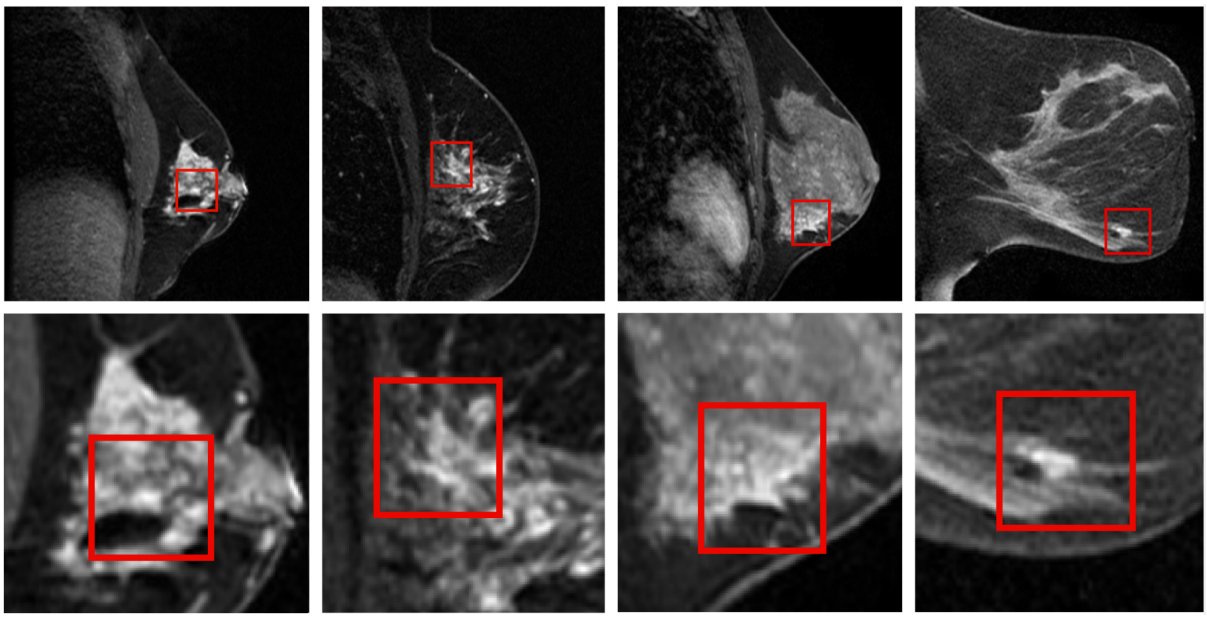

Figure 5: Four patients whose exams the network could have provided an early finding. Each column shows one of the top five regions of concern, proposed by the segmenter network (red square), which was confirmed by a radiologist to highlight the location of a future tumor.

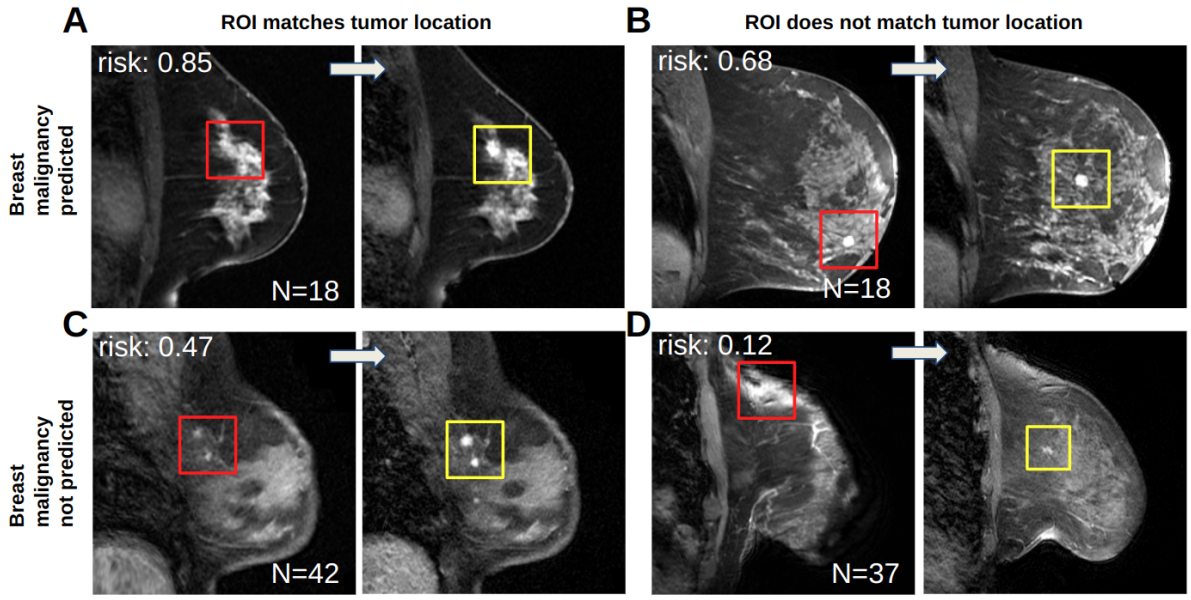

Figure 6: Localization of regions of concern and risk prediction in future tumors. Colored square indicates a region of concern flagged by the segmentation network (red) in the current MRI, and confirmed by a radiologist (yellow) in next year's MRI. All breasts here developed cancer within a year and had a radiologist segmentation (N=115), with "malignancy predicted" if the risk predicted by the AI for the whole breast was above the threshold that prompts re-evaluation (Fig. 4C, circle). N indicates the number of screen detected cancer in each category.

Characteristics of detected tumors

Next we asked how the regions highlighted by the network appeared to a human observer. To this end, a breast radiologist performed a retrospective directed review of all tumor cases (N=115; see exclusion criteria in Methods). The radiologist identified a visible correlate on the MRI prior to diagnosis in 71.3% of lesions (82/115, CI: 62.1-79.4%). 65% of these were also highlighted by the AI algorithm (53/82, CI: 53.3-74.9%). The radiologist confirmed no visible correlate in 28.7% of cases (33/115, CI: 20.6-37.9%) of cases, and the algorithm highlighted 21.2% of these cases (7/33, CI: 9.0-38.9%). Of the lesions identified on directed review 77.4% (89/115, CI: 68.7-84.7%) were less than 0.5 cm, and in average they were significantly smaller than at time of diagnosis (0.56±0.0.25 cm vs 1.02±0.56cm, t(156)=-5.5, p=1e-7). The radiologist also provided BI-RADS features for all visible lesions on MRI performed prior to and at diagnosis. Table S3 separates this based on the risk determined by the AI. The pathology findings separated by risk category are shown in Fig. S9.

Finally, we noted that age on its own was not predictive of cancer in this patient sample (age did not differ between cases that developed a tumor and those that remained healthy (rank-sum test, p = 0.09, W = 1.71), nor was family history of cancer (Chi-square test statistic = 0.43, p = 0.51, df = 1). Indeed, AUC-ROC is no different when demographic information is held constant (Fig. S9). This indicates that within our high-risk cohort, demographic information did not further stratify risk.

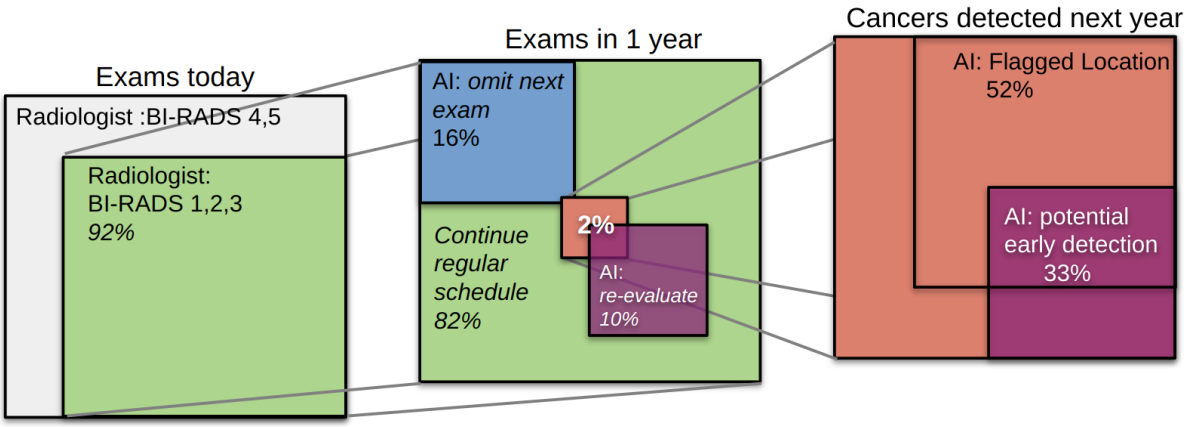

Figure 7: Summary of results on risk-adjusted screening, early detection and tumor localization. Areas are scaled in size to reflect fraction of cases - Left: After radiologist evaluation, BI-RADS 1-3 cases are read again by the AI algorithm (green). Center: The AI algorithm recommends omitting the next year's exam in lower-risk cases (blue), and re-evaluate higher risk cases (purple). Right: Fraction of future tumors potentially detected early (purple) assuming 10% of high-risk exams can be re-evaluated by a radiologist. AI flagged regions of concern that contained the location where the tumor was later detected (black rectangle).

Discussion

We proposed a new framework to determine an individual's probability of developing breast cancer within one year based on their current cancer-free MRI. The overall results are summarized in Fig. 7. In 1 of 2 screen-detected cancers the AI algorithm found a region of concern that matched the location of the tumor on the next scheduled MRI. Re-evaluation by a radiologist focusing on these high-risk areas may yield a meaningful number of early detections. Re-evaluating only 1 in 20 cases would place one third of future malignancies in front of the radiologists. These could be abbreviated readings as the AI algorithm already highlights the regions of concern within the breast. Re-evaluations might prompt, for instance, a shorter term follow-up, or a biopsy. On the other hand, the network identified lower-risk cases (16% of all cases) for which one could have recommended extending the screening interval without compromising the existing rate of interval cancers (0.1%). Overall, we suggest that instead of a fixed one-size-fits all schedule, the radiologist could adjust the time for the next follow-up, using a numerical risk score estimated by the AI algorithm based on the current cancer-free MRI exam.

Upon directed review, the radiologist identified a visible correlate on the MRI performed prior to diagnosis in 71.3% percent cases, with 77.4% of these measuring less than 0.5 cm. However, these should not be definitively considered "misses" or "false negatives" MRIs since a critical part of screening for early detection is diagnosing an interval change. Such interval change is considered suspicious and guides management for all of these cases.

The need for adaptive screening is likely to increase in the future given the recent recommendation to enroll women with extremely dense breasts into supplemental MRI screening.[3] In these women risk of developing cancer is elevated, while chances of detecting it with mammography are reduced. The DENSE and ECOG-ACRIN clinical trials

demonstrated the benefit of supplemental[4,5] and abbreviated MRI screening[6]. The size of this population with dencense breasts is substantial, making the need for adaptive screening even more urgent.

A common problem of artificial intelligence research is its focus on algorithm performance instead of clinically relevant outcomes.[33] For instance, most studies on cancer diagnosis only report the ROC-AUC. More recent studies on risk prediction report the concordance index (c-index).[34] These metrics quantify the performance of the AI algorithm, but do not directly quantify screening burden or address early detection. Studies focusing on risk prediction on MRI also often report AUC-ROC[15] and have not considered the effect on the clinical workflow. We suggest that the objectives of reducing screening burden and increasing early detection are best evaluated in terms of the desired NPV and PPV, which results in precision-effort curves for predicting health and disease respectively. Similar to precision-recall curves, precision-effort curves are preferable over receiver operating characteristic curves in the context of rare events.[35] Here, they allow an obvious choice for the operating point in terms of NPV and PPV. At those points, our risk prediction network identified a fraction of cases that remained healthy for at least one year, without compromising detection of future malignancies. Perhaps more importantly, the network found a handful of cases that were higher-risk and could have been referred to an immediate biopsy, without incurring a burden beyond current clinical practice. While the numbers of cases are small, they are important to the women involved.

Recent studies on predicting outcomes 1-5 years in advance using mammography[34,36,37] report ROC-AUC values slightly higher than the 0.67 we report here with MRI. We believe that this is due to the different populations involved. For reference, an influential study[38] reported an AUC of 0.68 for predicting malignancy at any time within 5 years when using mammography alone. In that study, the Tyrer-Cuzick model, which relies mostly on demographic data, had an AUC of 0.67. This is well in line with risk models based on demographic and genetic information which have AUCs in the range of 0.6-0.71.[39] Similar results of AUC in the range of 0.68-0.73 have been obtained in a multi-institutional validation study with the same publicly available network.[40] In contrast, in our high-risk patient sample, family history and age had no additional predictive value within the high-risk sample. Thus, stratifying risk within the high-risk population may be more challenging because this population is already being screened with MRI, which is highly sensitive,[41] and tumors have already been removed earlier from this population as compared to the broader mammographic screening population. Indeed, a previous study focusing on risk prediction using MRI in a high-risk population[15] reports an AUC of 0.49 for the Tyrer-Cuzick model and 0.63 for the MRI-based predictor, reinforcing the notion that this is a difficult population compared to the broader mammography screening population.

Limitations of this study were a relatively small number of screen-detected cancers, which included only sagittal scans from a single clinical site. AI performance and robustness is likely to improve in the future with higher resolution axial exams that are now routine in clinical practice; with the use of the prior year's exam to determine changes in the appearance of individual lesions; and with increasing datasets from multiple sites, which are required for robust deep-learning. Nevertheless, the present work provided proof-of-principle and baseline performance on early detection and adaptive screening.

The precision-effort analysis we have presented here also applies to the concept of triage. In that case, the classification network is trained to predict the outcome of the breast in the *current* exam. Those breasts with low probability of a malignant finding (at 100% NPV) could be omitted from the radiologist's workload (in the present dataset, that would be 4.6% of all breasts). This type of triage has been suggested for mammography[42,43] as well as breast MRI.[44,45] However, triage is controversial, because it is hard to justify the risk of not reading an exam when the patient has already been burdened, and the cost of scanning has already been incurred. Ultimately, it is important to realize that all current screening policies have implicitly selected a balance in terms of cost vs benefit. Dedicating screening to those that need it most while sparing those that need it less can only be an improvement over the current clinical practice of a fixed schedule.

Acknowledgement

This work was supported by a grant from NIH with grant number R01CA247910 and R01EB028157. We want to thank Joanne Chin for extensive and thorough proofreading of this manuscript.

Author contribution

LH designed the computational methods, analyzed the data, programmed the network, generated figures and wrote the manuscript. YH performed all the image preprocessing. HM edited the paper. MH segmented images. DM provided imaging and clinical data. LP designed the overall approach and analysis methods and wrote the manuscript. ES conceived the approach of image-based risk-adjusted screening, provided imaging and clinical data, evaluated the predictions of the network and edited the manuscript. SEW and KP provided extensive input to the manuscript. EM contributions include formulating overall study design, data anonymization and curation, result interpretation and manuscript review

Competing Interests

All authors declare no financial or non-financial competing interests.

Data availability

The datasets analyzed during the current study are not publicly available due to patient confidentiality. However, risk prediction and outcome information for statistical evaluation of the results are available together with the code.

Code availability

The underlying code for the network, trained parameters, and code and data needed for statistical analysis of results are available on GitHub and can be accessed via this link [insert persistent URL to code].

References

1. Wernli, K. J. *et al.* Patterns of Breast Magnetic Resonance Imaging Use in Community Practice. *JAMA Intern. Med.* **174**, 125–132 (2014).
2. Bevers, T. B. *et al.* Breast Cancer Screening and Diagnosis, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw. JNCCN* **16**, 1362–1389 (2018).
3. Mann, R. M. *et al.* Breast cancer screening in women with extremely dense breasts recommendations of the European Society of Breast Imaging (EUSOBI). *Eur. Radiol.* **32**, 4036–4045 (2022).
4. Bakker, M. F. *et al.* Supplemental MRI Screening for Women with Extremely Dense Breast Tissue. *N. Engl. J. Med.* **381**, 2091–2102 (2019).
5. Veenhuizen, S. G. A. *et al.* Supplemental Breast MRI for Women with Extremely Dense Breasts: Results of the Second Screening Round of the DENSE Trial. *Radiology* **299**, 278–286 (2021).
6. Comstock, C. E. *et al.* Comparison of Abbreviated Breast MRI vs Digital Breast Tomosynthesis for Breast Cancer Detection Among Women With Dense Breasts Undergoing Screening. *JAMA* **323**, 746–756 (2020).
7. Schmutzler, R. K. *et al.* Risk-Adjusted Cancer Screening and Prevention (RiskAP): Complementing Screening for Early Disease Detection by a Learning Screening Based on Risk Factors. *Breast Care* **17**, 208–223 (2022).
8. Pharoah, P. D. P., Day, N. E., Duffy, S., Easton, D. F. & Ponder, B. A. J. Family history and the risk of breast cancer: A systematic review and meta-analysis. *Int. J. Cancer* **71**, 800–809 (1997).
9. Kuhl, C. K. & Baltzer, P. You Get What You Pay For: Breast MRI Screening of Women With Dense Breasts Is Cost-effective. *JNCI J. Natl. Cancer Inst.* **113**, 1439–1441 (2021).
10. Geuzinge, H. A. *et al.* Cost-Effectiveness of Magnetic Resonance Imaging Screening for Women With Extremely Dense Breast Tissue. *JNCI J. Natl. Cancer Inst.* **113**, 1476–

1483 (2021).

11. Gail, M. H. *et al.* Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81**, 1879–1886 (1989).
12. Boyd, N. F., Martin, L. J., Yaffe, M. J. & Minkin, S. Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Res.* **13**, 223 (2011).
13. Brentnall, A. R. & Cuzick, J. Risk Models for Breast Cancer and Their Validation. *Stat. Sci.* **35**, 14–30 (2020).
14. Brentnall, A. R. *et al.* A Case-Control Study to Add Volumetric or Clinical Mammographic Density into the Tyrer-Cuzick Breast Cancer Risk Model. *J. Breast Imaging* **1**, 99–106 (2019).
15. Portnoi, T. *et al.* Deep Learning Model to Assess Cancer Risk on the Basis of a Breast MR Image Alone. *Am. J. Roentgenol.* **213**, 227–233 (2019).
16. Chiarelli, A. M. *et al.* Effectiveness of screening with annual magnetic resonance imaging and mammography: results of the initial screen from the ontario high risk breast screening program. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **32**, 2224–2230 (2014).
17. Roganovic, D., Djilas, D., Vujnovic, S., Pavic, D. & Stojanov, D. Breast MRI, digital mammography and breast tomosynthesis: Comparison of three methods for early detection of breast cancer. *Bosn. J. Basic Med. Sci.* **15**, 64–68 (2015).
18. Zhang, Y. & Ren, H. Meta-analysis of diagnostic accuracy of magnetic resonance imaging and mammography for breast cancer. *J. Cancer Res. Ther.* **13**, 862–868 (2017).
19. Vreemann, S. *et al.* The frequency of missed breast cancers in women participating in a high-risk MRI screening program. *Breast Cancer Res. Treat.* **169**, 323–331 (2018).
20. Yamaguchi, K. *et al.* Breast Cancer Detected on an Incident (Second or Subsequent) Round of Screening MRI: MRI Features of False-Negative Cases. *AJR Am. J. Roentgenol.* **201**, 1155–63 (2013).
21. Pages, E. B., Millet, I., Hoa, D., Doyon, F. C. & Taourel, P. Undiagnosed Breast Cancer

at MR Imaging: Analysis of Causes. *Radiology* **264**, 40–50 (2012).

22. Mammary Gland Mass and Breast Cancer Risk on JSTOR. https://www.jstor.org/stable/3702346#metadata_info_tab_contents.
23. Wengert, G. J. *et al.* Inter- and intra-observer agreement of BI-RADS-based subjective visual estimation of amount of fibroglandular breast tissue with magnetic resonance imaging: comparison to automated quantitative assessment. *Eur. Radiol.* **26**, 3917–3922 (2016).
24. King, V. *et al.* Background parenchymal enhancement at breast MR imaging and breast cancer risk. *Radiology* **260**, 50–60 (2011).
25. Pike, M. C. & Pearce, C. L. Mammographic density, MRI background parenchymal enhancement and breast cancer risk. *Ann. Oncol.* **24**, viii37–viii41 (2013).
26. Dontchos, B. N. *et al.* Are Qualitative Assessments of Background Parenchymal Enhancement, Amount of Fibroglandular Tissue on MR Images, and Mammographic Density Associated with Breast Cancer Risk? *Radiology* **276**, 371–380 (2015).
27. Hu, X., Jiang, L., You, C. & Gu, Y. Fibroglandular Tissue and Background Parenchymal Enhancement on Breast MR Imaging Correlates With Breast Cancer. *Front. Oncol.* **11**, (2021).
28. Fazeli, S. *et al.* Patient-Reported Testing Burden of Breast Magnetic Resonance Imaging Among Women With Ductal Carcinoma In Situ: An Ancillary Study of the ECOG-ACRIN Cancer Research Group (E4112). *JAMA Netw. Open* **4**, e2129697 (2021).
29. Laws, A. *et al.* Baseline Screening MRI Uptake and Findings in Women with ≥ 20% Lifetime Risk of Breast Cancer. *Ann. Surg. Oncol.* **27**, 3595–3602 (2020).
30. Ghoncheh, M., Pournamdar, Z. & Salehiniya, H. Incidence and Mortality and Epidemiology of Breast Cancer in the World. *Asian Pac. J. Cancer Prev. APJCP* **17**, 43–46 (2016).
31. Hirsch, L. *et al.* Deep learning achieves radiologist-level performance of tumor segmentation in breast MRI. *ArXiv200909827 Phys. Stat* (2020).

32. Gao, Y. *et al.* Magnetic Resonance Imaging in Screening of Breast Cancer. *Radiol. Clin. North Am.* **59**, 85–98 (2021).
33. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *Npj Digit. Med.* **5**, 1–8 (2022).
34. Yala, A. *et al.* Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* **13**, (2021).
35. Ozenne, B., Subtil, F. & Maucort-Boulch, D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol.* **68**, 855–859 (2015).
36. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
37. Lotter, W. *et al.* Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat. Med.* **27**, 244–249 (2021).
38. Yala, A., Lehman, C., Schuster, T., Portnoi, T. & Barzilay, R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology* **292**, 60–66 (2019).
39. Louro, J. *et al.* A systematic review and quality assessment of individualised breast cancer risk prediction models. *Br. J. Cancer* **121**, 76–85 (2019).
40. Yala, A. *et al.* Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. *J. Clin. Oncol.* **40**, 1732–1740 (2022).
41. Kriege, M. *et al.* Efficacy of MRI and Mammography for Breast-Cancer Screening in Women with a Familial or Genetic Predisposition. *N. Engl. J. Med.* **351**, 427–437 (2004).
42. Rodriguez-Ruiz, A. *et al.* Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur. Radiol.* **29**, 4825–4832 (2019).
43. Dembrower, K. *et al.* Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective

simulation study. *Lancet Digit. Health* **2**, e468–e474 (2020).

44. Verburg, E. *et al.* Deep Learning for Automated Triaging of 4581 Breast MRI Examinations from the DENSE Trial. *Radiology* **302**, 29–36 (2022).
45. Bhowmik, A. *et al.* Automated Triage of Screening Breast MRI Examinations in High-Risk Women Using an Ensemble Deep Learning Model. *Invest. Radiol.* 10.1097/RLI.0000000000000976 doi:10.1097/RLI.0000000000000976.
46. Kim, G. R., Cho, N., Kim, S.-Y., Han, W. & Moon, W. K. Interval Cancers after Negative Supplemental Screening Breast MRI Results in Women with a Personal History of Breast Cancer. *Radiology* **300**, 314–323 (2021).
47. Modat, M. *et al.* Fast free-form deformation using graphics processing units. *Comput. Methods Programs Biomed.* **98**, 278–284 (2010).

Supplement

Metrics for evaluating reduction of screening burden and improved early detection

The objective of reducing screening burden and the objective of early detection of the proposed risk-adjusted screening are not in opposition. To see this, let's assume the AI-algorithm estimates the risk of developing cancer in the following year based on the current MRI exam as shown schematically in Fig. S1A. In this illustration we can clearly distinguish lower, medium and higher risk cases. What is important to note is that the two thresholds in this diagram can be selected independently from one another. In this view, the objectives of saving lives vs reducing screening burden are independent and not in competition.

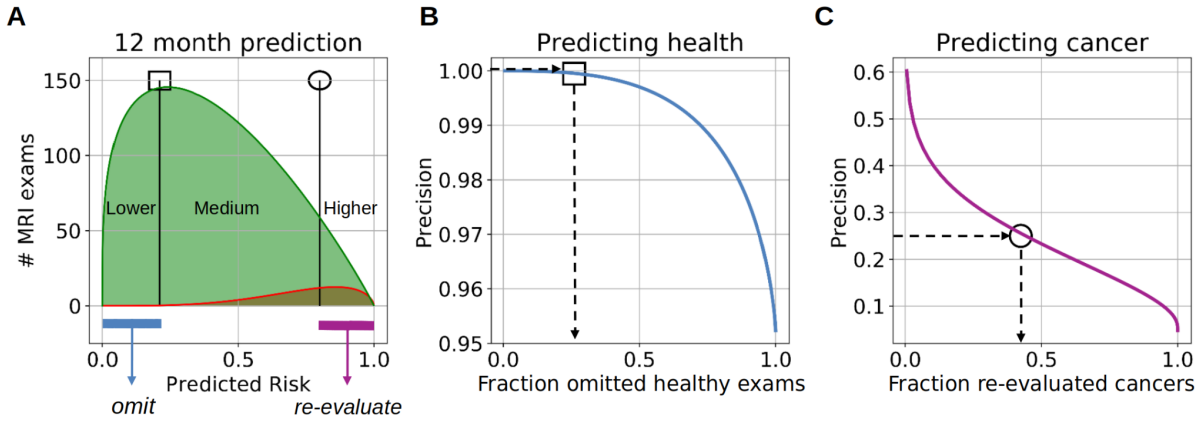

Figure S1: Schematic of possible 12-month outcomes as a function of machine-predicted risk. The AI-algorithm predicts the risk of developing cancer next year, based on the current exam. (A) All exams are currently cancer-free, and remain cancer-free (green) or will have a newly detected cancer (red) at the next yearly screening. Cases with a predicted risk below the lower threshold (square) are considered lower risk and can have an extended follow-up, while cases above the high threshold (circle) are higher risk and may be recalled early for immediate assessment. (B) Precision in finding healthy/lower-risk scans as a function of the fraction of exams that would be omitted due to extended follow-up. (C) Precision in finding scans that will develop cancer as a function of the fraction of exams that are re-evaluated, i.e., potential early detections relative to the total number of screen-detected cancers.

In the proposed risk-adjusted screening approach, medium-risk cases would continue to be screened according to current practice and having the higher-risk cases referred to the radiologist for immediate review should not be controversial. However, the lower-risk cases should be handled with care. To allow the recommendation that the next yearly screening interval can be extended and thus meet the objective of reducing screening burden, there must be absolute confidence that a low-risk prediction by the AI-algorithm would in fact result in a negative exam in 12 months' time. This confidence can be quantified as the "negative predictive value" (NPV, i.e. the number of true negatives over all negative predictions, see Eq. S1 in the Supplement). To avoid unnecessary risk, the low-risk threshold in Fig. S1A (square) can be set so that NPV=100%, i.e. there is 100% confidence that no cancer will occur within one year. As the low-risk threshold is moved rightward, the number of negative predictions that would omit the next yearly exam increases. Fig. S1B shows the tradeoff between confidence (NPV) and reduced screening burden (fraction of omitted exams relative to total number of exams, Eq. S2). The benefit of risk-adjusted screening can be read from this graph as the point of highest savings that still has 100% confidence (square). In this illustrative example, 25% of exams can potentially be omitted (vertical blue arrow) without

missing a single malignant exam. More realistically, in the current clinical practice, the number of tumors that are detected in the interval between two yearly screenings is exceedingly small. The number of such “interval cancers” is as low as 0.1-1.5% of the total number of yearly screening exams.[41,46] This corresponds to an NPV of 98.5-99.9%.

Meanwhile, the second objective of early detection can be evaluated in terms of the “positive predictive value” (PPV, Eq. S3), namely, the likelihood that a predicted cancer is in fact detected with an immediate assessment, e.g. with a biopsy. With PPV, we can be more permissive than with NPV, as it should not be controversial to send a few extra cases for immediate assessment, provided a few tumors may be detected earlier than under the current screening regime. At present, radiologists at our clinical tertiary cancer care center have a PPV of 25% (i.e., three in four biopsies are benign) when they recommend biopsy, and this is an acceptable “cost” of screening under current practice. Fig. S1C shows the tradeoff between the "cost" of extra biopsies (1-PPV) and the benefit of further increasing the sensitivity of a radiologist (tumors potentially detected one year earlier). A reasonable operating point therefore would be to set the high-risk threshold such that PPV=25% to match the performance of the radiologist under the current clinical practice. In the illustrative example of Fig. S1C the fraction of potential early detections (over total number of screen-detected cancer, Eq. S4) at this operating point is 45% (vertical blue arrow). In practice, radiologists would be asked to perform a supplementary reading of the higher-risk exams, which is a lower burden than performing a biopsy. Radiologists may be willing to perform a supplemental reading if they can catch, say, one extra tumor in 20 supplemental readings, i.e., PPV=5%, provided it is a small number of cases and the AI-algorithm can pinpoint where exactly to take a second look.

The following section explains that both the above mentioned trade-offs constitute precision-effort curves. In the case of Fig. S1B (Eq. S1 vs S2) the curve evaluates the “cost/benefit” of predicting health, and in the case of Fig. S1C (Eq. S3 vs S4), the curve evaluates the “cost/benefit” of predicting disease. We suggest that adaptive screening should be evaluated with these precision-effort curves, rather than ROC curves.

Positive Predictive Value, Negative Predictive Value, Sensitivity and Specificity
A variety of measures are used to quantify failure and success in binary classification problems. This variety can be confusing at times, and so we reproduce here the definitions relevant for the present work. All measures are defined based on the four possible outcomes of a binary classification, namely, the number of classifications that are true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In our work, “positive” means that cancer is present and “negative” means that cancer is not present. One important criterion is the number of true negatives, which also can be quantified relative to the total number of negative classifications, or relative to the total number of actual negative cases:

$$\text{Negative predictive value} = TN/(TN+FN) \quad (1)$$
$$\text{Specificity} = \text{True negative rate} = TN/(TN+FP) \quad (2)$$
$$\text{Fraction omitted} = (TN+FN)/(TN+FN+TP+FP) \quad (3)$$

The other important criterion is the number of true positives, which can be quantified relative to the total number of positive classifications, or relative to the total number of actual positive

cases. These two quantities go by different names:

$$\text{Positive predictive value} = \text{TP}/(\text{TP}+\text{FP}) \quad (4)$$
$$\text{Sensitivity} = \text{True positive rate} = \text{TP}/(\text{TP}+\text{FN}) \quad (5)$$
$$\text{Fraction re-evaluated} = (\text{TP}+\text{FP})/(\text{TN}+\text{FN}+\text{TP}+\text{FP}) \quad (6)$$

We use sensitivity and specificity when referring to the current diagnosis, with the trade-off between these two captured by the conventional ROC curve (Fig. S1). On the other hand, we use positive predictive value (PPV) vs fraction omitted from reading for anticipating cancer at one year, and negative predictive value (NPV) vs fraction re-evaluated for anticipating health at one year. Note that these pairs represent precision and effort; they capture the trade-off between precision of predicting health and the corresponding effort saved (Fig. S1B) or the precision in predicting cancer and the effort exerted (Fig. S1C). We refer to them as precision-effort curves and suggest that they constitute the preferred way of evaluating triage, risk-adjusted screening and early detection.

MRI Acquisition, Preprocessing, and Harmonization

The acquisition, preprocessing and harmonization of the MRIs used in this evaluation have been detailed in a previous publication.[31] Briefly, exams were acquired in the sagittal plane at varying in-plane resolutions, 2–4 mm slice thickness, and varying repetition times and echo times. The sequences used here included pre-contrast, fat-saturated T1-weighted images, and a variable number (n = 3–8) of post-contrast fat-saturated T1-weighted images to capture dynamic contrast enhancement. In-plane sagittal resolution was harmonized by upsampling low-resolution images. Image intensity from different scanners were harmonized by dividing with the 95th percentile of pre-contrast T1 intensity. To summarize the variable number of dynamic contrast-enhanced images, we measured the volume transfer constant for the initial uptake and subsequent washout, DCE-in and DCE-out, respectively (example of a T1-weighted image and DCE-in in Fig. 3A).

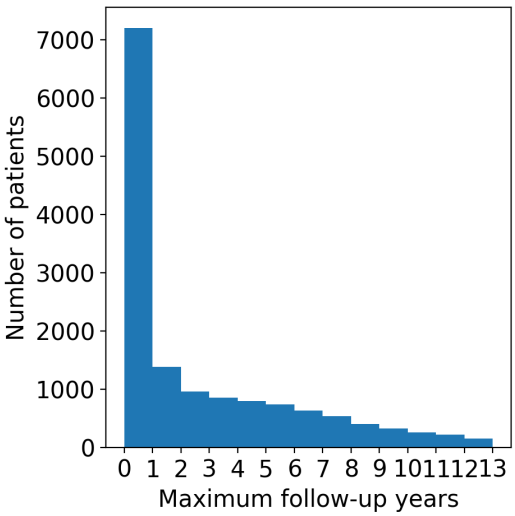

Figure S2: The patient sample included patients that have been followed in the screening program for a varying amount of time. Here is the time period of data available with sagittal MRI exams.

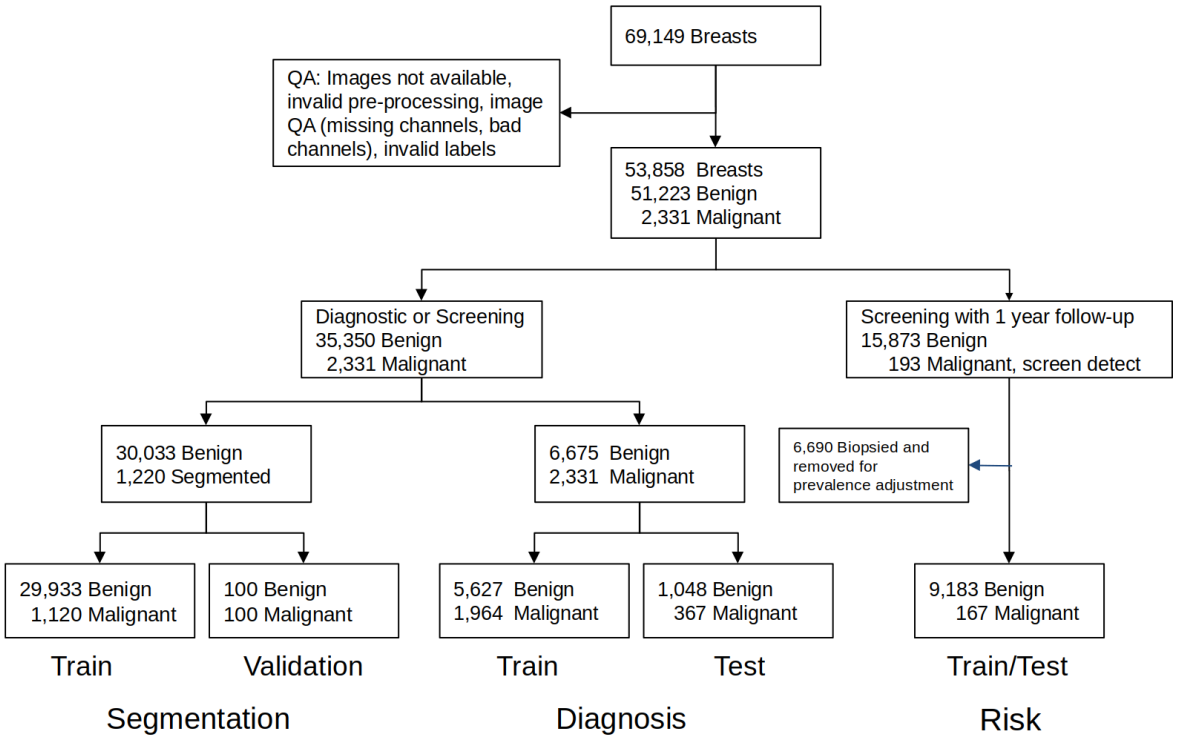

Figure S3: Chart of data used for training and testing. Numbers here refer to individual breasts, and we did not always have data for two breasts for each exam date. The segmentation and diagnostic network were trained with partially overlapping data. The risk-reduction network did not overlap with the other two, and only included screening exams.

Number of breasts (benign/malignant)	**Segmentation**	**Diagnosis**	**Risk**
Train	31053 (29,933/1120*)	6773 (4909/1864)	7,351** (7,206/145)**
Validation	200 (100/100)	818 (718/100)	****
Test	***	1415 (1048/367)	1986** (1977/22)**
Total	31,253 (30,033/1,220)	9,006 (6,675/2,331)	9,350 (9,183/167)

Table S1: Numbers of breasts used for training, validation and testing. *Each malignant case had a 2D segmentation for one of the slices in the MRI volume. **This was the size of data for each fold in 5-fold cross validation, for a total of 9,350 exams. ***We did not evaluate test set performance of segmentation in the present work. **** we did not perform model selection for the risk prediction network, so no validation set was needed

	Segmentation	Diagnosis	Risk
Family History	14%	10%	15%
Mean Age [min - max]	52 [17 - 93]	52 [18 - 93]	52 [18 - 88]
Ethnicity: Hispanic or Latino	5%	5%	4%
Ethnicity: Not Hispanic	71%	71%	71%
Ethnicity: Unknown	23%	23%	24%
Race: Asian East/Indian Subcontinental	3%	3%	3%
Race: Black or African American	5%	5%	5%
Race: White	24%	24%	24%
Race: Unknown / Other	67%	67%	68%

Table S2: Summary statistics of demographics per partition.

Extraction of candidate regions of concern using a 3D segmenter and training

The total dataset was partitioned for training and testing of the segmentation and diagnostic network on the one hand, and on the other hand, for training and testing of the risk prediction network (see Fig. S3). The latter only included screening exams. These partitions are disjointed at the image and exam level.

To identify regions of concern and to extract image features, we used a U-Net that had been previously developed for 3D segmentations.[31] This network computes for each pixel in the MRI volume the probability that the pixel is part of a malignant lesion (Fig. 4A). The network was trained on pixels drawn from tumor and non-tumor regions of 2D manual segmentations, but also on pixels drawn from anywhere on benign exams. As such, it was trained to distinguish pixels belonging to malignant lesions from that of normal breast tissue or benign lesions. Training and validation using the current data (Table S1) proceeded as in previous work.[31] Candidate regions of concern (49 x 49 pixels in size) were sampled on a fixed grid with half overlap from all slices in the 3D volume. The mean probability of malignancy was computed over all pixels in each region, and the top 5 regions were selected for further analysis in the diagnostic network.

Architecture and training of a diagnostic network

The diagnostic network is a convolutional neural network as shown in Fig. 3B. It estimates the probability that the current breast MRI contains a malignant lesion, based on 5 regions of concern. The inputs to the network for each region are 32 features from the last layer of the segmentation network (16 features computed separately for the ipsilateral and contralateral

breast). All five regions of concern are included as input to produce a single prediction of the outcome. The network contains 5 blocks of 2D convolutional layers with 3x3 kernels, followed by a ReLU non-linearity, batch-normalization layer and a max-pooling layer by a factor of 2. Prior to running through the segmentation network the contralateral breast was coregistered to the ipsilateral breast with NiftyReg,[47] so that the 5 regions can serve as a reference in the corresponding locations. Demographic information used as input by the network included ethnicity, race, and history of breast cancer in the family. This network was trained and tested on a combination of diagnostic and screening exams (Table S1). A subset of the training data (10%) was used for validation and model selection. The best validation-set performance was obtained for a model using the top 5 ROIs instead of top 1 (ROC-AUC 0.72 vs 0.76) and without extra penalty-weight on the low-prevalence class (AUC-ROC 0.74 vs 0.76). Subsequent fine-tuning for risk prediction used this final model.

Fine-tuning for risk prediction

We fine-tuned the diagnostic network on the prediction of future malignancy in current cancer-free exams. The resulting risk-prediction network has the same architecture as the diagnostic network, but its parameters are optimized to predict future tumors. Due to the small number of screen-detected cancer, we evaluated test-set performance using 5-fold cross validation. The test set was selected to have a natural prevalence of screen-detected cancers versus healthy breasts of 2%. All cross-validation folds were trained with the same number of epochs (n = 5), otherwise we found that risk distributions varied between folds, potentially requiring individual risk thresholds.

ROC performance of diagnostic and risk prediction network.

We tested the performance of the diagnostic network at detecting a malignant lesion in the current exam (Fig. S4A). The AUC-ROC on test data was 0.87 [CI: 0.86, 0.89] (n = 367 malignant, n = 1,048 benign). A much harder task was predicting future malignancy, on all currently cancer-free exams, as radiologists have already identified malignancies with high sensitivity. On this task, the risk prediction network reached a test-set performance of AUC-ROC of 0.67 [CI: 0.63, 0.70] (Fig. 4, 5-fold cross validation on n = 9,350).

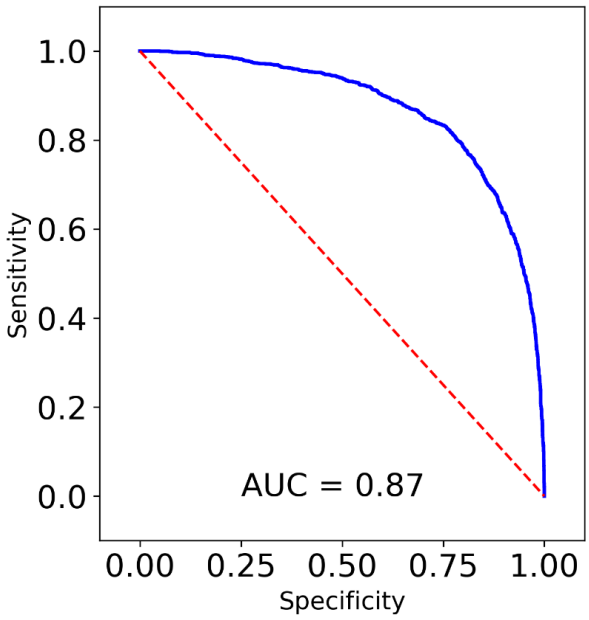

Figure S4: Performance in diagnosing current exam outcome. Test-set performance of the diagnostic network at classifying the current exams as “benign” or “malignant”.

To understand the increased difficulty between the conventional diagnostic task and the risk prediction task as we have formulated it here consider the following: Conventional diagnosis with AI uses all exams (only excluding BI-RADS 0 and 6) to predict outcome. Outcome is “benign” for all cases that remain cancer free for 2 years (including benign biopsy), and “malignant” for cases with an malignant biopsy. Importantly, difficult cases that are BI-RADS ≤ 3 but end up developing cancer in 1-2 years, are simply excluded from the analysis in most AI literature. The one-year risk prediction task here, instead only includes BI-RADS ≤ 3 cases. It is much harder to find tumors in that subset, as all the likely or certain malignant tumors have already been removed. Then, the network has to predict what will happen at the next screen, which might detect cancer with a mammogram, supplemental MRI, biopsy or any other approach. We are intentionally not excluding patients that present with cancer in clinical follow-up. In fact, it is those difficult cases we aim to anticipate.

Threshold values for lower and higher risk determination

Thresholds for the lower and higher risk category were set at 20th and 90th percentile of risk in the training data in each fold. Applying these thresholds to the left-out test data and aggregating across folds resulted in a NPV=99.5% and PPV=5% respectively, as reported in the main text. Selecting the threshold on the percentile of the total data is statistically more robust than selecting it directly on PPV given the small number of malignant cases. For the PPV=25% we selected the threshold directly on the test data. This value should be regarded with care as a large dataset would be required for a statistically robust choice of this higher threshold value.

Confidence intervals

All confidence intervals represent 95% confidence. For ratios they were computed using the Clopper-Pearson exact method. For AUC values they were computed using bootstrapping with resampling.

Software and libraries

Model design, training and evaluation was done using the Python deep-learning library Tensorflow version 1.14 using as backend Keras version 2.3.1.

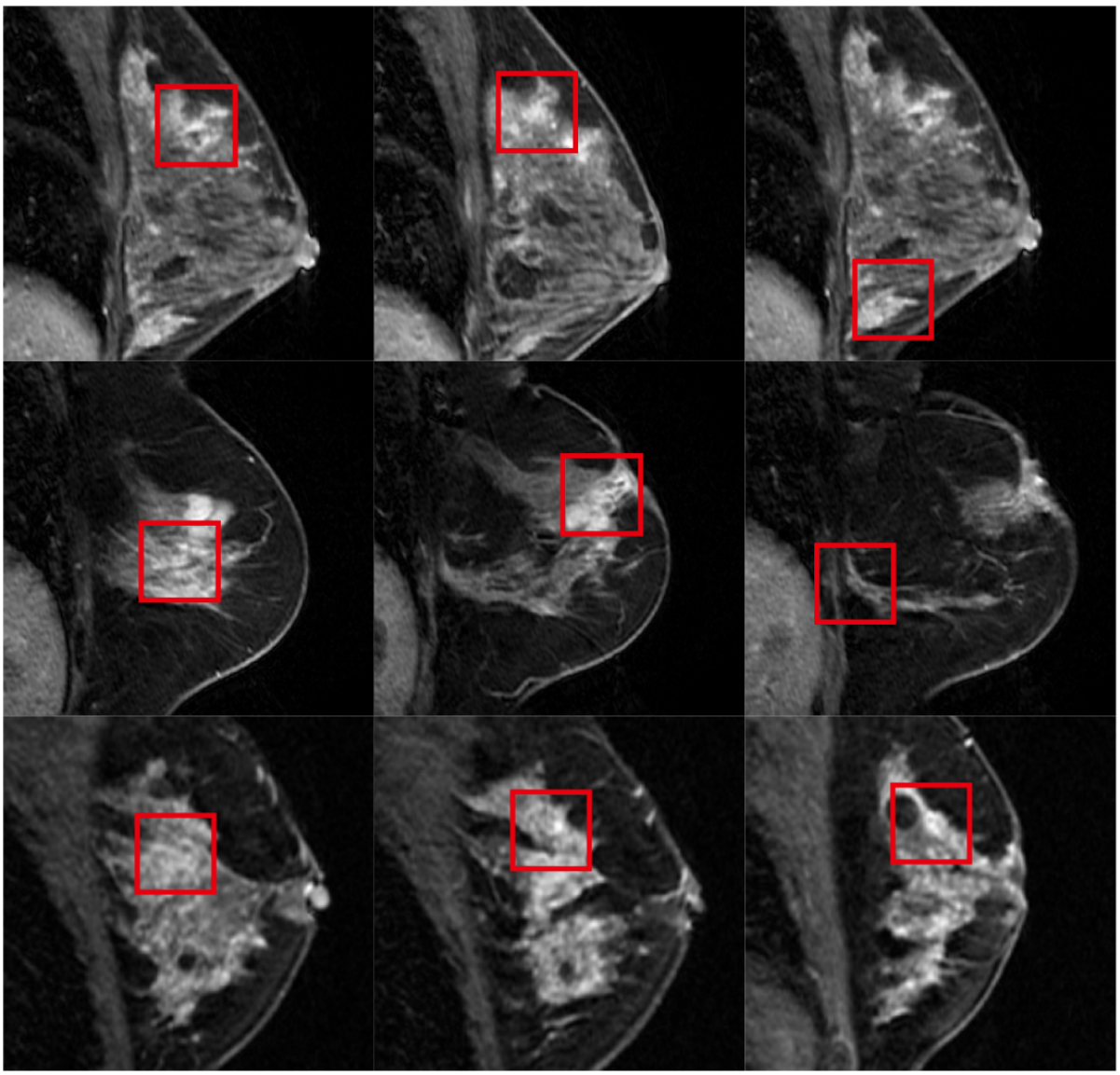

Figure S5: Six out of the twelve benign breasts for which the model assigned highest risk of cancer development within one year. Each row shows three (out of five) regions of concern per breast as identified by the network (blue box).

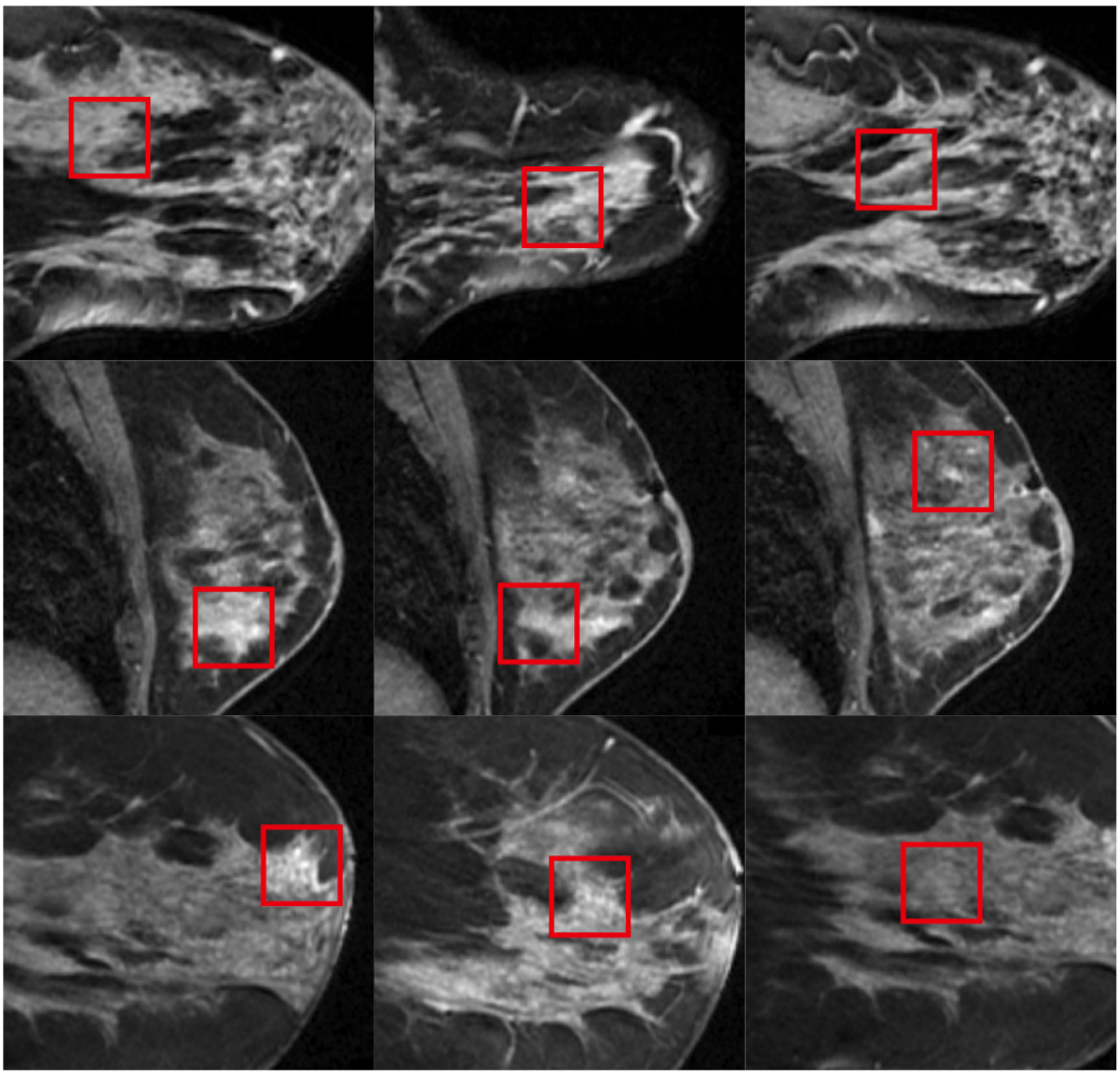

Figure S6: Six out of the twelve benign breasts for which the model assigned highest risk of cancer development within one year. Each row shows three (out of five) regions of concern per breast as identified by the network (blue box).

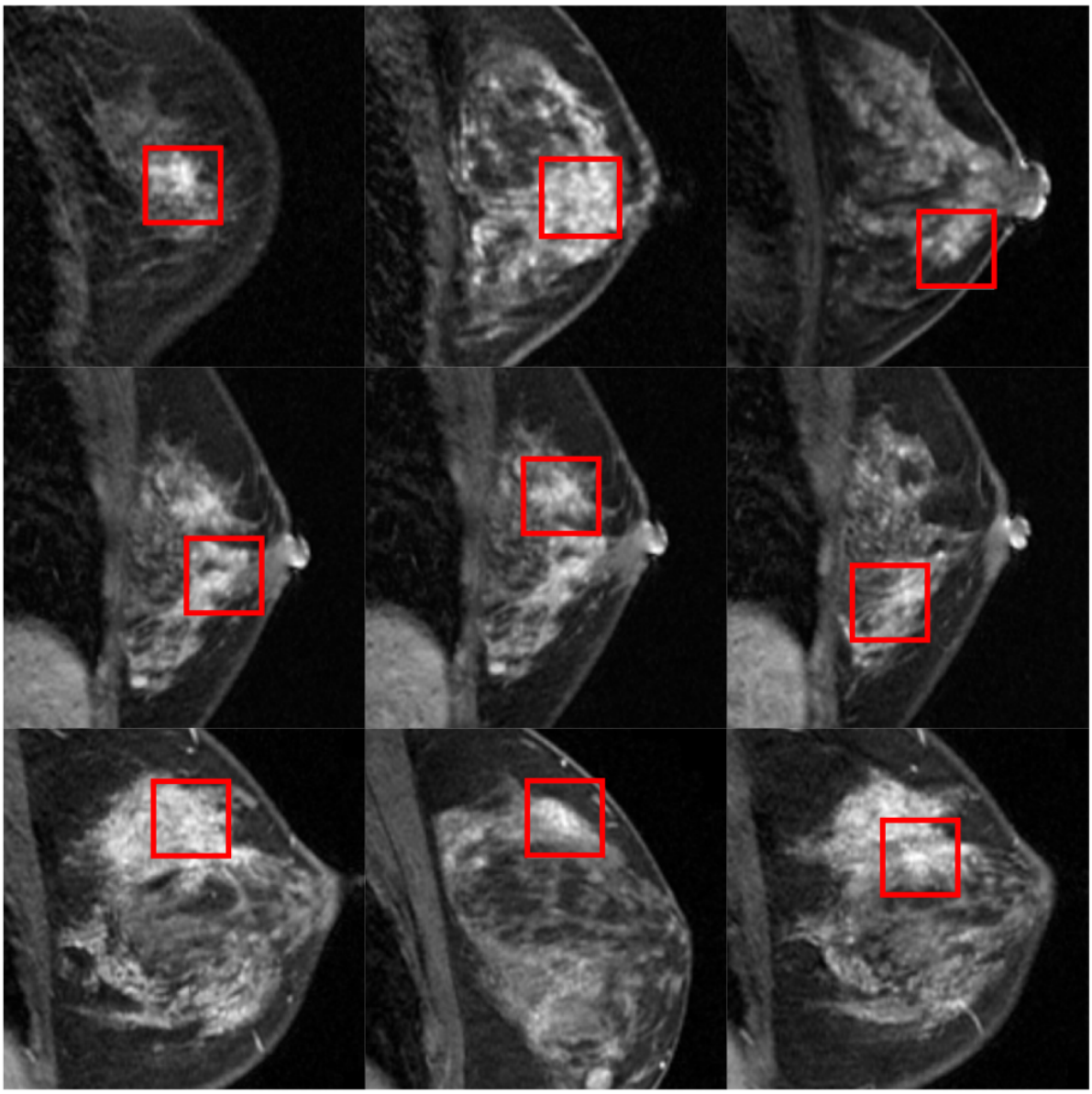

Figure S7: Further six out of the twelve examples of benign breasts for which the model assigned highest risk of cancer development within one year. Each row shows three (out of five)regions of concern per breast as identified by the network (blue box).

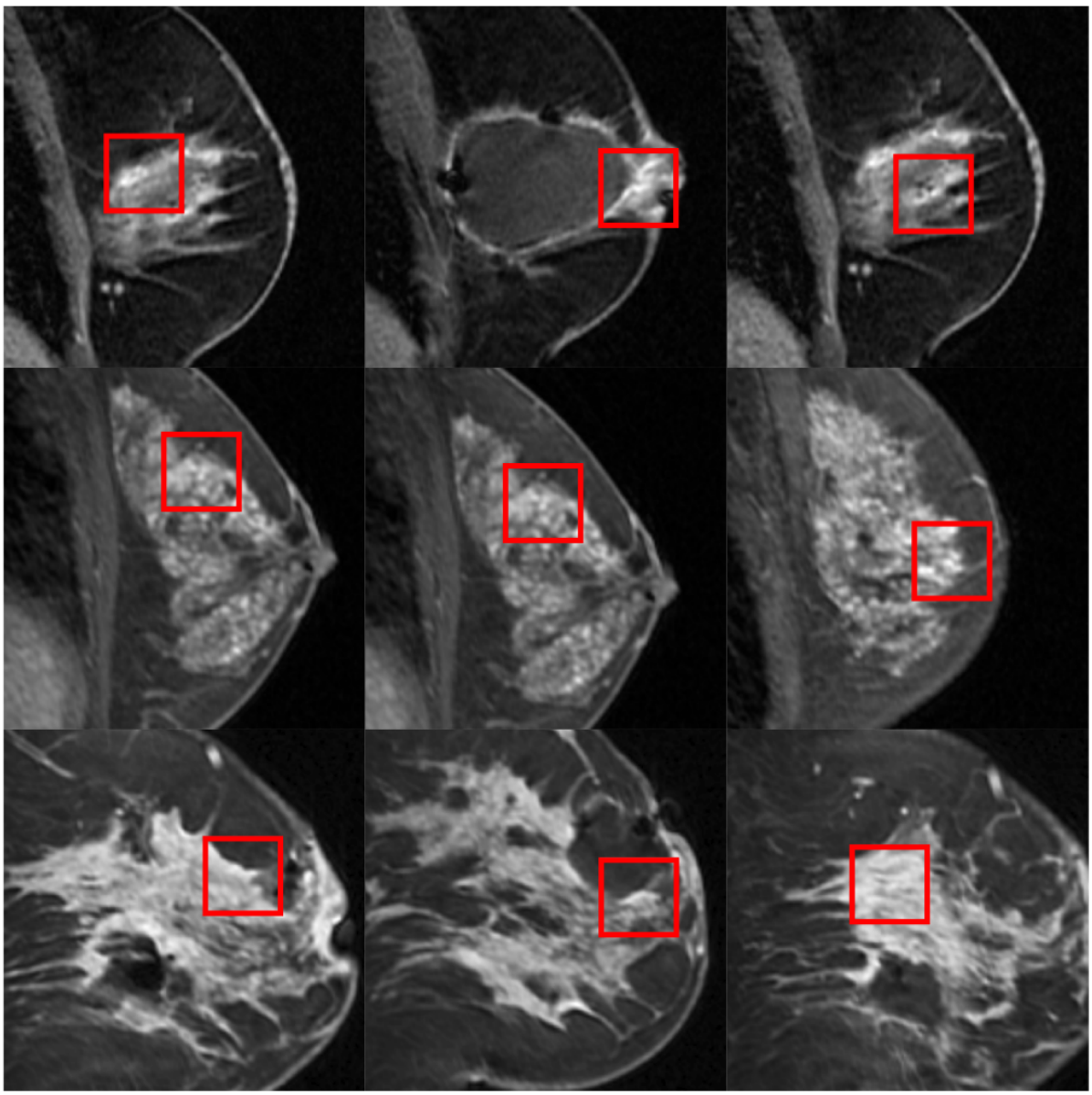

Figure S8: Six out of the twelve benign breasts for which the model assigned highest risk of cancer development within one year. Each row shows three (out of five) regions of concern per breast as identified by the network (blue box).

			Current Exam (n=115)		Future Exam (n=115)
			High Risk (n=36)	Medium and Lower Risk (n=79)	
Lesion Type		Focus	8 (22%)	22 (28%)	8 (7%)
		Mass	6 (17%)	6 (8%)	57 (50%)
		Non Mass Enhancement	10 (28%)	30 (38%)	50 (42%)
		None	12 (33%)	21 (27%)	0
Size		Mean	0.62 cm	0.53 cm	1.02 cm
		STD	0.20 cm	0.27 cm	0.56 cm
Focus/ Mass	Shape	Round	9 (64%)	17 (61%)	39 (60%)
		Oval	1 (7%)	6 (21%)	4 (6%)
		Irregular	4 (29%)	5 (17%)	22 (34%)
	Margin	Circumscribed	5 (36%)	19 (67%)	16 (25%)
		Irregular	9 (64%)	9 (32%)	49 (75%)
		Spiculated	0	0	0
	Internal Enhancement	Homogeneous	12 (86%)	21 (75%)	50 (77%)
		Heterogeneous	2 (14%)	7 (25%)	13 (20%)
		Rim Enhancement	0	0	2 (3%)
		Dark internal Septations	0	0	0
	T2	Isointense	14 (100%)	26 (93%)	62 (95%)
		Hyperintense	0	0	1 (2%)
		Heterogeneous	0	0	1 (2%)
Non Mass Enhancement	Distribution	Focal	7 (70%)	30 (100%)	43 (86%)
		Linear	0	0	0
		Segmental	2 (20%)	0	4 (8%)
		Regional	1 (10%)	0	4 (8%)
	Internal Enhancement	Homogeneous	2 (20%)	19 (63%)	24 (48%)
		Heterogeneous	8 (80%)	11 (37%)	27 (54%)
		Rim Enhancement	0	0	0
	T2	Isointense	10 (100%)	29 (97%)	48 (96%)
		Hyperintense	0	1 (3%)	2 (4%)

Table S3: BI-RADS features for the 115 cancers before and after detection.

Significance of demographics

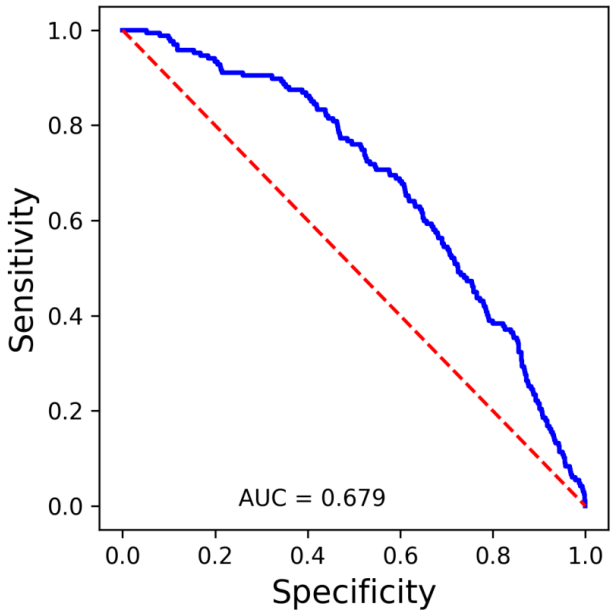

Figure S9: When replacing demographics and clinical information with the median value, we observe a numerically higher AUC-ROC although this is not significant (Delong test, z=1.42, p = 0.15). This suggests that demographic information has little predictive values in this high-risk population, and the network may have indeed overfitted on demographic data.

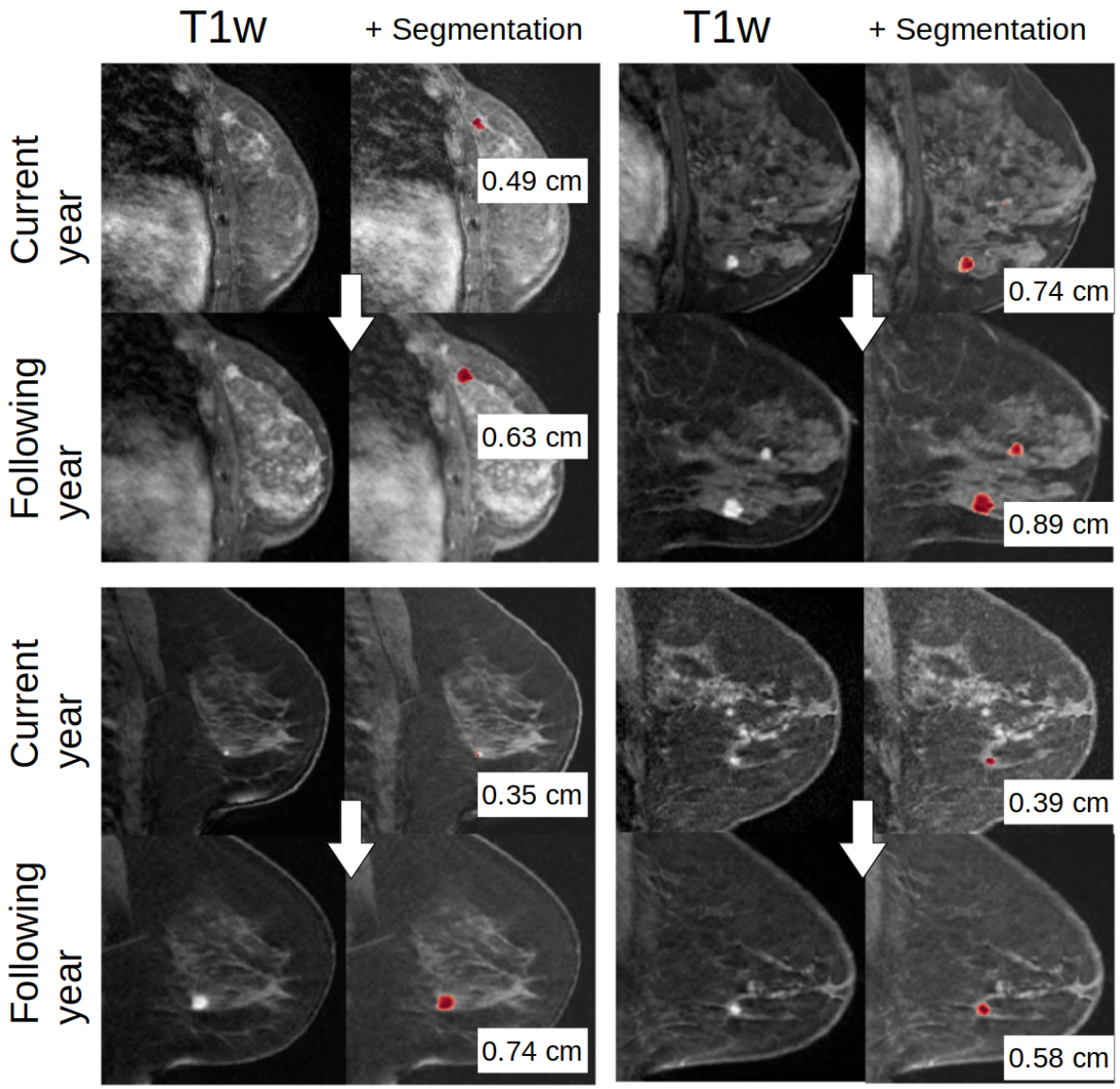

Figure S10: Automatic measuring of lesions in current and future breast. Four examples of lesion measuring in the current and following year examination where a cancer was detected. Automatic volumetric segmentation of both breasts was done with a pre-trained machine segmenter. A radiologist segmentation of the index lesion was used to mark the location of the developed cancer. A machine segmentation was used in the previous year exam, while keeping only the top 5 regions of interest. Spatial alignment of the subsequent exam was done to find the corresponding region in the breast, by evaluating if there is an overlap between the radiologist segmentation and one of the top 5 regions segmented by the machine. For measuring the size, the binarized segmented area was projected along its first component, and the range of the projection was measured and multiplied with the image resolution. In order to provide a robust estimate, the 95 percentile of the projection length was used. If there are multiple lesions present only the largest one is reported.

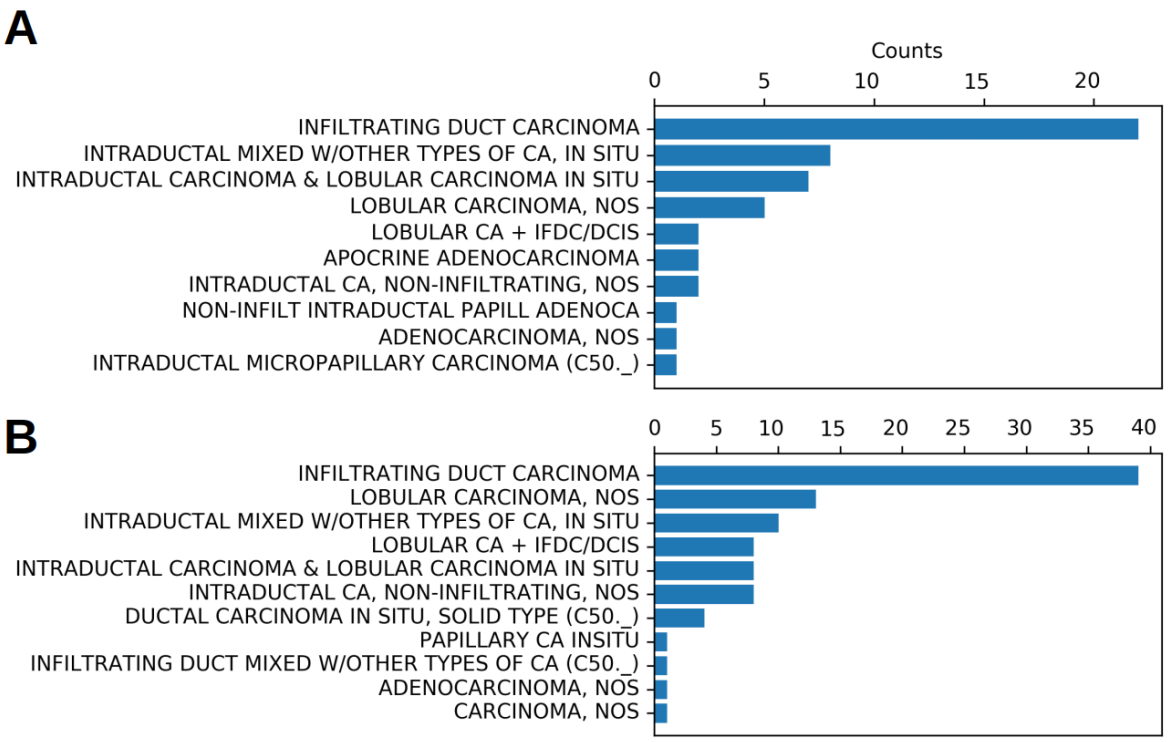

Figure S11: Cancer types in the higher (A) vs medium and lower risk groups (B) as stratified by the AI algorithm.

			Future Exam (n=5)
Lesion Type		Focus	0
		Mass	2 (40%)
		Non Mass Enhancement	3 (60%)
		None	0
Size		Mean	1.01 cm
		STD	0.82 cm
Focus/Mass	Shape	Round	1 (50%)
		Oval	
		Irregular	1 (50%)
	Margin	Circumscribed	1 (50%)
		Irregular	1 (50%)
		Spiculated	0
	Internal Enhancement	Homogeneous	1 (50%)

		Heterogeneous	1 (50%)
		Rim Enhancement	0
		Dark internal Septations	0
	T2	Isointense	1 (50%)
		Hyperintense	0
		Heterogeneous	1 (50%)
Non Mass Enhancement	Distribution	Focal	2 (66%)
		Linear	0
		Segmental	1 (33%)
		Regional	0
	Internal Enhancement	Homogeneous	0
		Heterogeneous	3 (100%)
		Rim Enhancement	0
	T2	Isointense	3 (100%)
		Hyperintense	0

Table S4: BI-RADS features for the 5 cancers that were assigned a lower-risk by the AI algorithm. A total of 7 cancers were assigned a lower risk assessment by the AI algorithm. Two of these were excluded from analysis for BI-RADS features due to axillary recurrence and negative breast, and due to post-biopsy changes.