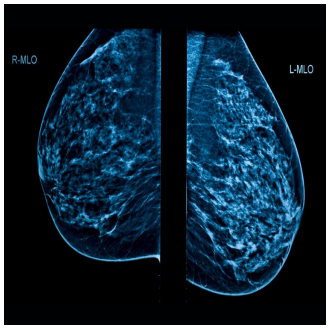
HealthManagement
Promoting Management and Leadership

DL Risk Scores Flag Interval Cancers in Breast Screening



Personalised breast cancer screening relies on accurate near-term risk assessment to balance early detection with the harms of overdiagnosis. Interval cancers (ICs) that emerge between routine screens are linked to delayed diagnosis, larger tumours and poorer prognoses compared with screen-detected cancers. A mammography-based deep learning model, Mirai, was evaluated for its ability to predict 1-, 2- and 3-year IC risk within the UK programme that invites women aged 50–70 years for triennial mammography. The evaluation analysed whether risk scores separated future ICs from noncancers and how performance varied by age and breast density. It also explored operating thresholds to inform potential recall strategies for supplemental imaging or shorter screening intervals.

Large Triennial Cohort and Evaluation Approach

Consecutive negative screening mammograms were retrospectively assembled from two UK screening sites using two primary mammography systems. Ethical approvals were in place and informed consent was waived. Examinations with implants, nonstandard views, less than 40 months of follow-up, screen-detected cancers at baseline and next-round cancers were excluded. The reference standard for normal mammograms was absence of a cancer diagnosis within 40 months, confirmed histopathologically. One baseline negative mammogram per woman was included.

Must Read: ***AI in Diagnostic Mammography Matches Radiologist Accuracy***

Mirai processed images locally using a validated containerised implementation, producing yearly risk scores interpretable as the likelihood of developing breast cancer within the subsequent 1, 2 or 3 years. Risk score distributions for women who later developed ICs were compared with those who did not at each time point. Receiver operating characteristic analyses generated AUCs with DeLong comparisons. A Cox model provided the Harrell concordance index (C index) as a time-to-event summary. Subgroup analyses assessed age quartiles and BI-RADS density categories derived from AI-estimated composite scores calibrated to match known population distributions.

The analysed cohort comprised 134 217 examinations from the same number of women (mean age 59.1 years ± 7.9), including 524 ICs. Most examinations at site 1 used Philips systems, while site 2 used GE HealthCare systems. Readers at participating screening units adhered to UK quality assurance processes, including double reading and annual performance evaluation.

Predictive Performance Across Time, Age and Density

Women who later developed an interval cancer tended to have higher Mirai scores at baseline than those who did not, and this pattern held at 1, 2 and 3 years. Overall discrimination was moderate and consistent across the three time windows, with summary measures around the high 0.6 to low 0.7 range. Looking at the two sites separately, performance was broadly similar, although one site showed somewhat stronger results for the longer time horizons.

The model's behaviour was steady across age groups. Younger and older women showed comparable separation between higher and lower risk, with concordance measures rising slightly from the early fifties into older age bands but without clear evidence of a marked difference. A similar picture emerged for breast density. Accuracy was higher for the least dense category and gradually lower for denser categories, yet when densities were grouped as lower versus higher, overall performance remained in a similar range. Taken together, these observations suggest that Mirai maintained a stable level of risk discrimination within the UK's triennial screening context, regardless of whether women were younger or older or had lower or higher breast density.

Thresholds, Systems and Practical Considerations

Operating thresholds simulated recall based on the highest proportions of 3-year scores within the screening round. Recalling women in the highest 1% of scores would have identified 3.6% of ICs. Expanding recall to the highest 5%, 10% and 20% of scores would have identified 14.5%, 26.1% and 42.4% of ICs, respectively. In this cohort, these thresholds corresponded to additional cancer detection rates of 0.1, 0.6, 1.0 and 1.7 per 1000. Using the Youden index to maximise sensitivity and specificity, recalling the highest 35.2% of scores would have captured 62.0% of ICs.

System-specific analyses suggested higher AUCs for examinations acquired on GE HealthCare systems than on Philips systems for years 2 and 3, while pooled performance remained comparable to prior evaluations conducted in different settings. As Mirai was originally trained on Hologic images, similarity between acquisition characteristics may partly explain system-level differences observed here.

Several limitations frame interpretation. The evaluation was retrospective, so whether ICs could have been detected at baseline with supplemental imaging is unknown. Next-round screen-detected cancers were excluded, which likely underestimates potential gains from additional imaging. Ethnicity data were sparsely available and most women were older than 50 years, limiting generalisability to younger populations or specific ethnic groups. Density estimates were AI-derived and calibrated to population distributions, which may not replicate all aspects of real-world BI-RADS assessment. Some subgroup comparisons were underpowered. Mirai processes only standard four-view mammograms, which complicates application in women requiring additional views or after mastectomy.

Within a large UK triennial screening cohort, a mammography-based deep learning model stratified near-term IC risk with consistent performance across time points, age quartiles and density categories. Threshold analyses showed that prioritising women with higher 3-year scores could recover a substantial share of ICs, supporting investigation of shorter screening intervals or supplemental imaging for those at highest risk. System-specific calibration, prospective evaluation, appropriate threshold selection and post-market surveillance will be important to define how model-driven risk complements or replaces density measures to improve population screening pathways.

Source: Radiology

Image Credit: iStock

Published on : Mon, 10 Nov 2025