Deep learning on histopathological images to predict breast cancer recurrence risk and chemotherapy benefit: a multicentre, model development and validation study

Gil Shamai, Shachar Cohen, Yoav Binenbaum, Edmond Sabo, Alexandra Cretu, Chen Mayer, Iris Barshack, Tal Goldman, Gil Bar-Sela, António Polónia, Dezheng Huo, Alexander T Pearson, Frederick M Howard, Joseph A Sparano, Ron Kimmel, Dvir Aran**

Summary

Background Genomic assays such as Oncotype DX have transformed adjuvant treatment selection for hormone receptor-positive, HER2-negative, early breast cancer but remain inaccessible to many patients because of high cost and logistical barriers. We aimed to develop and validate an artificial intelligence (AI) model that estimates Oncotype DX 21-gene recurrence scores directly from routine histopathology slides and clinicopathological variables.

Methods In this multicentre, model development and validation study, a multimodal deep-learning model was trained on digital whole-slide images and clinical features using a foundation model pre-trained on 171 189 histopathology slides for predicting Oncotype DX recurrence score. We included slides from patients with hormone receptor-positive, HER2-negative, invasive breast cancers and without scanning artifacts and with at least 100 tissue tiles (1·6 mm²). The model was fine-tuned and validated on the TAILORx randomised trial (8284 patients after quality control). Prognostic and predictive performance was assessed in the TAILORx-test set and externally validated in six independent cohorts (Carmel, Haemek, and Sheba medical centres [Israel], the University of Chicago Medical Center [USA], the Australian Breast Cancer Tissue Bank [Australia], and the Cancer Genome Atlas Breast Invasive Carcinoma project [USA]).

Findings In the TAILORx-test set (n=2407), the AI model classified 1097 (45·6%) patients as low risk, 1021 (42·4%) as intermediate risk, and 289 (12·0%) as high risk. For identifying high genomic-risk disease (recurrence score ≥26), the area under the curve (AUC) was 0·898 (95% CI 0·879–0·913). AI-based risk stratification was prognostic for recurrence-free interval (hazard ratio 2·61 [95% CI 1·68–4·04]), distant recurrence-free interval (2·88 [1·73–4·79]), and disease-free survival (1·32 [0·92–1·89]). Chemotherapy benefit was evident in premenopausal patients classified by AI as being at high risk (0·63 [0·46–0·86]) but absent in postmenopausal patients classified by AI as being at low risk (0·94 [0·78–1·12]). 151 (31·3%) clinically high-risk postmenopausal women (by MINDACT criteria) were reclassified as low AI risk with no chemotherapy benefit. Analysis on external cohorts (5497 patients) showed that the model is transferable to new data with high generalisability (recurrence score ≥26 AUC ranging from 0·858 to 0·903).

Interpretation These findings show that AI applied to routine histopathology can serve as a practical and scalable tool for guiding chemotherapy decisions in hormone receptor-positive, HER2-negative, early breast cancer. This approach has the potential to reduce unnecessary chemotherapy and broaden access to precision oncology, particularly in resource-limited settings where genomic testing remains unavailable or unaffordable.

Funding Israel Innovation Authority (Kamin), Zimin Institute for Artificial Intelligence Solutions in Healthcare, Israel Precision Medicine Partnership program, and Israel Cancer Research Fund.

Introduction

Breast cancer is the most frequently diagnosed cancer worldwide.[1] Among patients with hormone receptor-positive, HER2-negative, early-stage breast cancer, which accounts for approximately 70% of cases, a crucial challenge is identifying those who benefit from adjuvant chemotherapy.[2] Treatment decisions were traditionally based on clinicopathological features such as tumour size, receptor expression, grade, and nodal status, but these criteria alone lack sufficient accuracy for optimal stratification.[3] Consequently, some patients might have received chemotherapy without benefit, whereas others who could have benefited were missed.

Multigene expression assays have revolutionised treatment decision making in hormone receptor-positive, HER2-negative, early breast cancer.[4–7] Among these assays, the Oncotype DX (ODX) 21-gene recurrence score assay is the only test recommended by the US National Comprehensive Cancer Network guidelines as predictive of chemotherapy benefit.[8,9] Early validation studies (eg, NSABP-B-20 and SWOG-8814)[6,10] showed that a high ODX recurrence score identifies patients who benefit from chemotherapy, and subsequent reanalyses defined an ODX recurrence score of at least 26 as high risk.[11,12] The prospective TAILORx trial[13] confirmed that patients with an intermediate recurrence score (11–25) derive no

Lancet Oncol 2026

Published Online March 11, 2026 https://doi.org/10.1016/S1470-2045(25)00727-2

*Joint last authors

Taub Faculty of Computer Science (G Shamai PhD, S Cohen BSc, Prof R Kimmel PhD, Prof D Aran PhD), **Faculty of Electrical and Computer Engineering** (Prof R Kimmel), **and Faculty of Biology** (Prof D Aran), **Technion-Israel Institute of Technology, Haifa, Israel; Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Harvard Medical School, Boston, MA, USA** (Y Binenbaum MD PhD); **Wyss Institute for Biologically Inspired Engineering at Harvard University, Boston, MA, USA** (Y Binenbaum); **Department of Pathology, Carmel Medical Center, Haifa, Israel** (E Sabo MD, A Cretu MD); **Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel** (E Sabo); **Institute of Pathology, Sheba Tel Hashomer Medical Center, Ramat-Gan, Israel** (C Mayer MD, Prof I Barshack MD); **Department of Pathology, Emek Medical Center, Afula, Israel** (T Goldman MD); **Department of Oncology, Emek Medical Center, Afula, Israel** (Prof G Bar-Sela MD); **Technion Integrated Cancer Center, Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel** (Prof G Bar-Sela); **Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal** (A Polónia MD PhD); **School of Medicine and Biomedical Sciences, Fernando Pessoa University, Porto, Portugal** (A Polónia); **Department of Public Health Sciences** (Prof D Huo MD PhD), **Department of Medicine** (A T Pearson MD PhD,

F M Howard MD), **University of Chicago, Chicago, IL, USA; Chan Zuckerberg Biohub Chicago, Chicago, IL, USA** (A T Pearson); **Icahn School of Medicine at Mount Sinai, Tisch Cancer Institute, New York, NY, USA** (Prof J A Sparano MD)

Correspondence to: Dr Gil Shamai, Taub Faculty of Computer Science, Technion-Israel Institute of Technology, Haifa, 3200003, Israel
sgils@technion.ac.il

Research in context

Evidence before this study

We searched Google Scholar for studies published between Jan 1, 2016, and March 1, 2025, using the terms ("Hematoxylin" OR "Haematoxylin" OR "Eosin") AND "Breast cancer" AND "Deep learning" AND ("Oncotype" OR "Oncotype DX"), without language restrictions. The search identified 416 papers; 11 peer-reviewed studies specifically addressed the prediction of the Oncotype DX recurrence score from haematoxylin and eosin (H&E)-stained breast cancer slides. Review of reference lists found no additional relevant reports. These studies showed increasing interest in using machine-learning models to predict genomic recurrence scores, but all studies were based on retrospective observational data, where treatment assignment confounds prognostic and predictive evaluation. None were validated in randomised clinical trials, preventing assessment of chemotherapy benefit, the main clinical purpose of the Oncotype DX assay. These studies also lacked large-scale external validation across diverse populations and laboratory settings. A separate body of work has aimed to infer patient outcomes directly from H&E images, including regulatory-approved commercial tools (PreciseDX, Stratipath, Ataraxis, and RlapsRisk). Although these methods can outperform existing genomic assays in some datasets, their clinical impact remains uncertain because they were validated on observational data, where nearly all high-risk patients receive chemotherapy, precluding evaluation of treatment benefit. By contrast, predicting the Oncotype DX score itself allows AI models to inherit the assay's established clinical validation from randomised trials such as TAILORx. Across published studies, the risk of bias was high to moderate, driven by retrospective design, small sample size, and absence of external patient-level validation.

Added value of this study

To our knowledge, this study presents the first AI model based on H&E pathology images and clinicopathological variables to be assessed retrospectively using data from a randomised clinical trial. Leveraging this trial enabled assessment of the model's ability to predict chemotherapy benefit derived from randomised treatment assignment, rather than only correlate with recurrence scores or outcomes. The integration of a large-scale foundation model with a simple calibration method that does not rely on local genomic data shows that such an AI model can generalise across diverse populations and laboratory conditions. Extensive external validation across multiple independent cohorts, comprising one of the largest evaluations of an AI pathology model to date, provides first-of-its-kind evidence that combines retrospective assessment using randomised trial data with extensive real-world validation of recurrence score estimation and prognostication.

Implications of all the available evidence

This study presents a digital pathology AI model that advances towards real-world deployment to guide chemotherapy decisions in hormone receptor-positive, HER2-negative early breast cancer. The model's potential impact is particularly significant in low-income and middle-income countries, where genomic testing reaches fewer than 5% of patients. These findings show the potential for democratising precision oncology through widely accessible histopathological analysis, extending personalised treatment decisions to resource-limited settings globally. Future prospective and randomised validations across node-positive patients and populations from low-income and middle-income countries will further establish clinical utility and support global implementation.

benefit from chemotherapy, and RxPONDER extended use of recurrence scores to node-positive disease, showing differential benefit by menopausal status.[9,14–16] Despite these advances, global adoption of genomic testing remains limited.[17–20] Barriers include high cost (approximately US$3500 per test), extended turnaround times, and logistical challenges such as sample shipping and reimbursement procedures, despite potential downstream savings from more tailored therapy.[21]

Haematoxylin and eosin (H&E) staining is universally available and routinely used to assess tumour morphology. With the growth of digital pathology, including in resource-limited settings,[22] artificial intelligence (AI) has shown promise in analysing H&E whole-slide images to diagnose tumorus,[23] predict receptor status,[24–29] identify mutations,[30–32] and estimate outcomes.[33–36] Several studies reported that a high ODX recurrence score can be inferred from H&E images and clinicopathologic variables.[37–43] However, none were validated with randomised clinical trial (RCT) data, a crucial step to overcome treatment-related confounding and to show prediction of chemotherapy benefit, the central purpose of the ODX recurrence score. Moreover, these studies were limited by cohort size, diversity, and independent datasets, leaving generalisability uncertain, with most studies involving only a few hundred patients with up to one external cohort and one larger study trained on several thousand patients with two 500-patient external cohorts.[39] Hence, these previous studies represent important conceptual advances but stop short of producing models ready for clinical deployment and supporting treatment decisions in practice.

Advances in deep learning have dramatically improved the analysis of histological images through large-scale foundation models, which are particularly strong in generalisability to new data.[44–46] These models incorporate

self-supervised learning techniques and transformer architectures. Transformers,[47] originally developed for natural language processing, excel in capturing complex spatial patterns and long-range dependencies in histological slides, outperforming traditional convolutional neural networks. Self-supervised learning methods allow models to learn salient image features without relying on manual annotations, greatly expanding the diversity and volume of available training data.[48–50]

In this Article, we present a novel, multimodal, deep-learning model to guide chemotherapy decisions by inferring the ODX recurrence score from H&E images and clinicopathological features, leveraging data from the TAILORx trial. The intended use of this model is as a decision-support tool within routine digital pathology, estimating recurrence risk and chemotherapy benefit directly from H&E slides, to complement or substitute genomic testing when needed. We aimed to develop and validate this model, hypothesising that the model would match the prognostic and predictive value of genomic testing across diverse cohorts.

Methods

Study design and participants

In this retrospective, multicentre, model development and validation study, we investigated an AI model for predicting ODX recurrence score and chemotherapy benefit. We analysed a comprehensive dataset of H&E slides across seven independent cohorts: Carmel,[29] Haemek,[29] and Sheba medical centres (Israel), the University of Chicago Medical Center (UCMC; USA),[37] the Australian Breast Cancer Tissue Bank (ABCTB; Australia),[51] The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA; USA),[52] and the TAILORx RCT (multiple sites worldwide).[13] These cohorts span diverse populations, multiple geographical locations, diagnostic years, scanners, and staining protocols, representing real-world clinical settings. TAILORx served for model training and testing. External validation was done on the remaining cohorts, and two calibration sets (Carmel-calibration and Haemek-calibration; appendix p 12) supported model calibration. For TCGA-BRCA, we estimated the ODX recurrence score from mRNA expression using published formulae.[53] Inclusion criteria were hormone receptor-positive, HER2-negative, invasive breast cancers. Slides with scanning artifacts or fewer than 100 tissue tiles (1·6 mm²) were excluded (<0·1%; appendix p 20). Additional details on slide acquisition, scanning protocols, and cohort-specific inclusion criteria are provided in the appendix (p 3).

All data were collected under ethical committee approvals and data use agreements in compliance with the Declaration of Helsinki and respective Institutional Review Boards. ABCTB access followed its ethical and scientific approval policy. Data use agreements were signed with Eastern Cooperative Oncology Group and the American College of Radiology Imaging Network Cancer Research Group (ECOG-ACRIN) for TAILORx slides and with the National Clinical Trials Network (NCTN) and the National Cancer Institute (NCI) Community Oncology Research Program (NCORP) for clinical data. All datasets were fully deidentified.

Model development, training, and inference

TAILORx was randomly split into training (70%) and test sets (30%), with five-fold cross-validation applied to the training data. All splits were at the patient level to prevent data leakage. Predictions from the five-fold cross-validation were averaged to form an ensemble model applied to the TAILORx-test set and external cohorts. Further methodological details on tissue segmentation; tiling; feature extraction with the GigaPath foundation model, which was pretrained on 171 189 histopathology images; and the downstream multiple-instance learning (MIL) architecture are described in the appendix (pp 3–4; figure 1A). Clinicopathological variables were incorporated into the patient-level prediction, creating a multimodal AI model. Throughout this Article, TAILORx-test refers to the TAILORx-test set, and TAILORx-CV refers to the average performance of the five cross-validation AI models on their corresponding validation folds in the TAILORx training set.

For benchmarking, we did a comprehensive literature review of published models for inferring the ODX recurrence score from H&E images (appendix p 11). Most previous studies used a fully supervised pipeline consisting of a convolutional neural network feature extraction followed by an attention-based MIL framework trained directly on the ODX recurrence score labels. We implemented a supervised MIL model[29] representing this methodological class and trained it with TAILORx data. Furthermore, we benchmarked against the Deep Learning for Recurrence Score (DLRS) method,[37] the only publicly available recurrence score-prediction model.

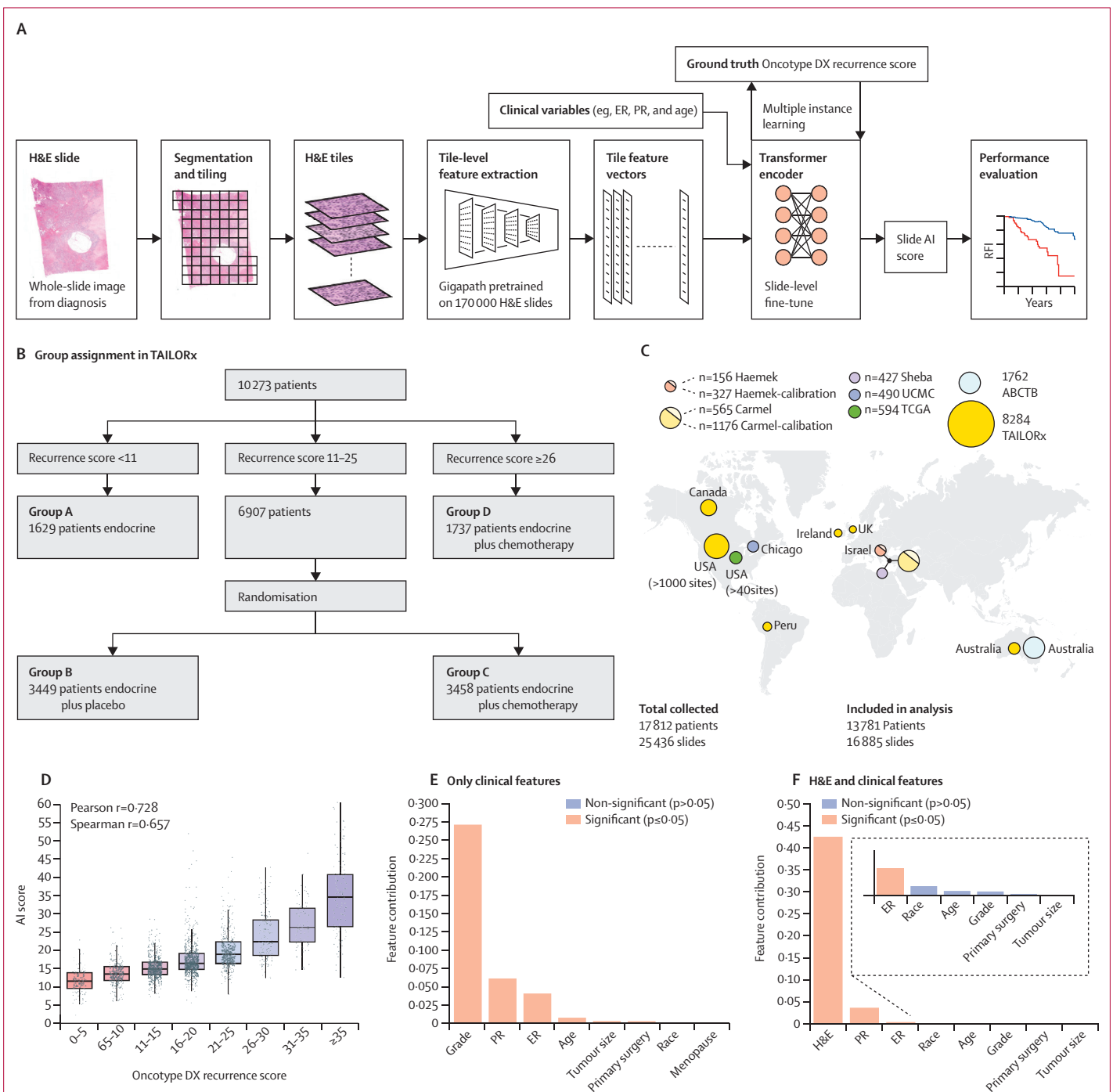
Feature importance was assessed by iteratively adding features to the model on the basis of their contribution, measured as the difference in r^2 performance on the test set when incorporating or not the feature during training.

Low and high AI thresholds were chosen from the standard TAILORx subgroup boundaries (ODX recurrence scores 11, 16, 21, 26) to maximise the proportion of classifications while minimising false-negative classifications relative to a recurrence score of 26 or greater and false-positive classifications relative to a recurrence score of 16 or greater.

We next aimed to show that chemotherapy improves outcomes for premenopausal patients identified by AI as being at high risk (AI ≥26), and that chemotherapy is not beneficial for postmenopausal patients identified by AI as being at low risk (AI <16). In TAILORx, randomisation occurred only for patients with an ODX recurrence score of 11–25 (figure 1B), posing a methodological limitation for chemotherapy benefit analysis for the AI-based

For more on **ABCTB** see https://nsw.biobanking.org/biobanks/view/7

stratification. To address this limitation, we quantified the proportion of discordant cases between AI and ODX-recurrence score and assessed whether these small discordant subsets could plausibly change the ODX-recurrence score-based chemotherapy-benefit estimates: we first re-evaluated the chemotherapy benefit in premenopausal and postmenopausal patients in TAILORx using arms B and C for disease-free survival, distant recurrence-free interval, and recurrence-free interval. Then, for premenopausal patients, we simulated the addition of the 1·4% of patients with a recurrence score below 16 by randomly sampling from the group with recurrence score below 16 (appendix pp 4–6). For postmenopausal patients, we simulated the addition of



(Figure 1 continues on next page)

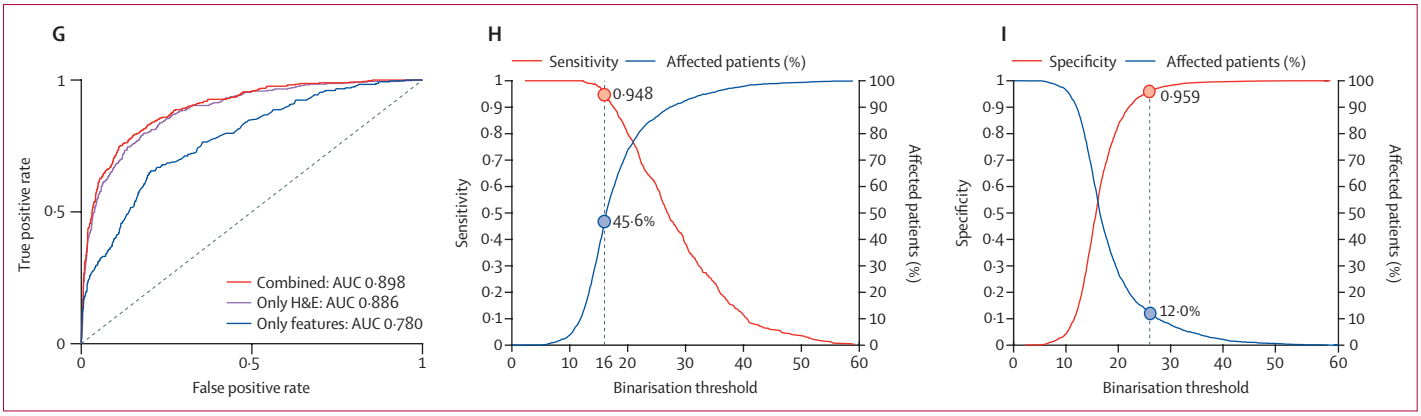


Figure 1: **Study cohorts, deep learning system architecture, and performance evaluation on TAILORx-Test**

(A) Pipeline of the deep learning system: H&E slides undergo automated tissue segmentation and tiling, followed by feature extraction using the GigaPath foundation model pre-trained on 171 189 H&E slides via self-supervision. Combined with clinicopathological variables, these features are then used for tuning a transformer-based multiple-instance learning model to generate slide-level scores. Clinical features include histological grade, hormone receptor status (progesterone receptor and oestrogen receptor), patient age, tumour size (dichotomised at 2 cm), surgery type (mastectomy or lumpectomy), self-reported race, and menopausal status (appendix pp 3–4). The final multimodal model included only ER status and PR status, as the contribution of the rest of the features was insignificant. (B) Design of the TAILORx trial, showing patient stratification by recurrence score and randomisation of patients at intermediate risk (recurrence score 11–25) to endocrine therapy alone or with chemotherapy. (C) Geographical distribution of study cohorts. (D) Distribution of the AI scores versus Oncotype DX recurrence score on the TAILORx held-out test set, showing high correlation (Pearson r=0·728 and Spearman r=0·657). (E, F) Quantitative assessment of feature contributions to recurrence score estimation using sequential forward selection, for the model using only clinical features (E) and the AI multimodal model incorporating both H&E images and clinical features (F). Bar heights represent additive r^2 values, with statistical significance indicated. (G) Receiver operating characteristic curves for identifying high genomic risk (recurrence score ≥26) for the three models: the AI model combining both H&E and clinical features (AUC 0·898), the image model using only H&E images (AUC 0·886), and the clinicopathological model using clinical features alone (AUC 0·780). (H) Sensitivity (red) and cumulative proportion of patients classified as low risk (blue) across binarisation thresholds for identifying high genomic risk (recurrence score ≥26). Percentage of affected patients refers to the cumulative proportion of patients below the given threshold—ie, those who would be triaged as low risk by the model. (I) Specificity (red) and cumulative proportion of patients classified as high risk (blue) across binarisation thresholds for identifying high genomic risk (recurrence score ≥26). Specificity was measured for the task of identifying a recurrence score of 26 or higher. ABCTB=Australian Breast Cancer Tissue Bank. AI=artificial intelligence. AUC=area under the curve. H&E=haematoxylin and eosin. UCMC=University of Chicago Medical Center. TCGA-BRCA=The Cancer Genome Atlas Breast Invasive Carcinoma.

the 1·8% of patients with a recurrence score of 26 or greater to arms B and C, while correcting for the potential benefit of chemotherapy.

To rigorously assess the model's generalisability and clinical readiness, we assessed six independent external cohorts (figure 1C; appendix pp 12–13). After applying hormone receptor-positive, HER2-negative, inclusion criteria and quality control, 7502 slides from 5497 patients were analysed (appendix p 20). To account for dataset-specific variations and facilitate robust deployment, we applied a one-parameter linear scaling calibration approach (appendix pp 6–7), which showed that clinical variables could accurately estimate the mean ODX recurrence score and the true calibration scale (appendix pp 18, 30). Empirical analysis across two independent calibration cohorts showed that using 100 patients sufficed for stable estimation of the scaling factor (appendix p 31).

To assess the prognostic generalisability of the multimodal AI model, we did survival analyses across three external cohorts with follow-up data: UCMC, ABCTB, and TCGA-BRCA. The UCMC cohort included both ODX recurrence scores and patient outcomes, enabling direct comparison between AI and the ODX recurrence score prognostication.

Finally, we assessed the potential clinical utility of the model by comparing AI-derived risk categories with clinical-risk groups defined by the MINDACT criteria (appendix p 4).[7] This analysis quantified the extent to which AI-based risk classification could modify treatment recommendations based on conventional clinico-pathologic assessment.

See **Online** for appendix

Outcomes

Time-to-event endpoints were defined as in the TAILORx study[13] and included distant recurrence-free interval, recurrence-free interval, disease-free survival (defined as freedom from invasive disease recurrence, second primary cancer, or death), overall survival, and breast cancer-specific survival.

Statistical analysis

Model performance was assessed using area under the receiver operating characteristic curve (AUC), Pearson and Spearman correlations, and area under the precision-recall curve for the positive and negative classes. Clinical utility was measured using sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), balanced accuracy, and F1 score. Details of the clinical risk stratification used for the clinical-utility analyses, based on the MINDACT criteria, are provided in the appendix (p 4). For the Kaplan–Meier curves, p values, hazard ratios (HRs), and 95% CIs were computed to quantify the effect size under the Cox proportional hazards assumption, with a significance level of 0·05. Bootstrapping with 1000 replicates estimated 95% CIs for the performance metrics and feature importance analysis, where no analytical variance estimates are available. The detailed methodology for the chemotherapy-benefit simulations based on TAILORx

randomisation is provided in the appendix (pp 4–6). The linear calibration procedure, including estimation of the cohort-specific scaling factor and its evaluation across calibration sample sizes, is described in the appendix (pp 6–7). Importantly, this calibration requires only an estimate of the mean ODX recurrence score for the target cohort. This mean recurrence score can be approximated from basic clinical features, allowing the calibration method to be applied even in regions where genomic testing is unavailable.

Statistical analysis was performed using MATLAB (version 2024b) and R (version 4.4.0). Preprocessing, training of the models, and inference were done using Python (version 3.9.18) and Pytorch library (version 2.0.0). Additional Python libraries used for database management, graphical plotting, scientific calculations, and other tasks were Numpy (version 1.26.4), Pandas (version 2.2.2), Scipy (version 1.13.1), and Openslide (version 1.3.1).

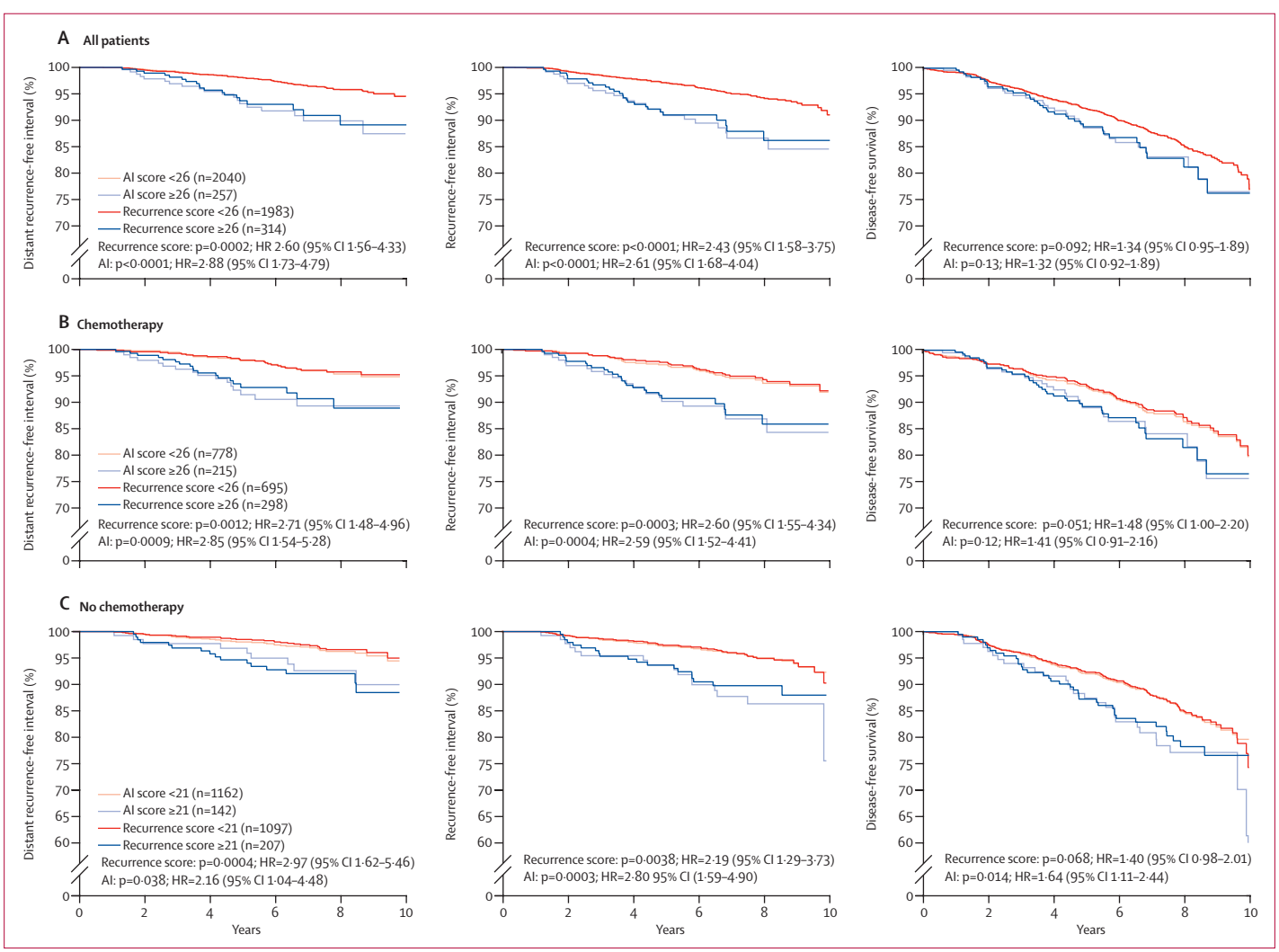
Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

From the TAILORx trial (n=10 273), we included 8284 patients (80·6%) after quality control, with 5877 in the training set and 2407 in the test set (figure 1B, C; appendix p 12). The median age was 56 years (IQR 49–63); 2590 (31%) were aged 50 years or younger, 2812 (34%) were premenopausal, median tumour size was 1·5 cm (1·2–2·1), 1449 (18%) had poor histological grade, 1398 (17%) had high ODX recurrence score (≥26), and 5531 (70%) had low clinical risk by the MINDACT criteria (appendix pp 4, 13).

Attention-map visualisations confirmed that the model focused on invasive tumour regions and captured both tumour-intrinsic and microenvironmental features



(Figure 2 continues on next page)

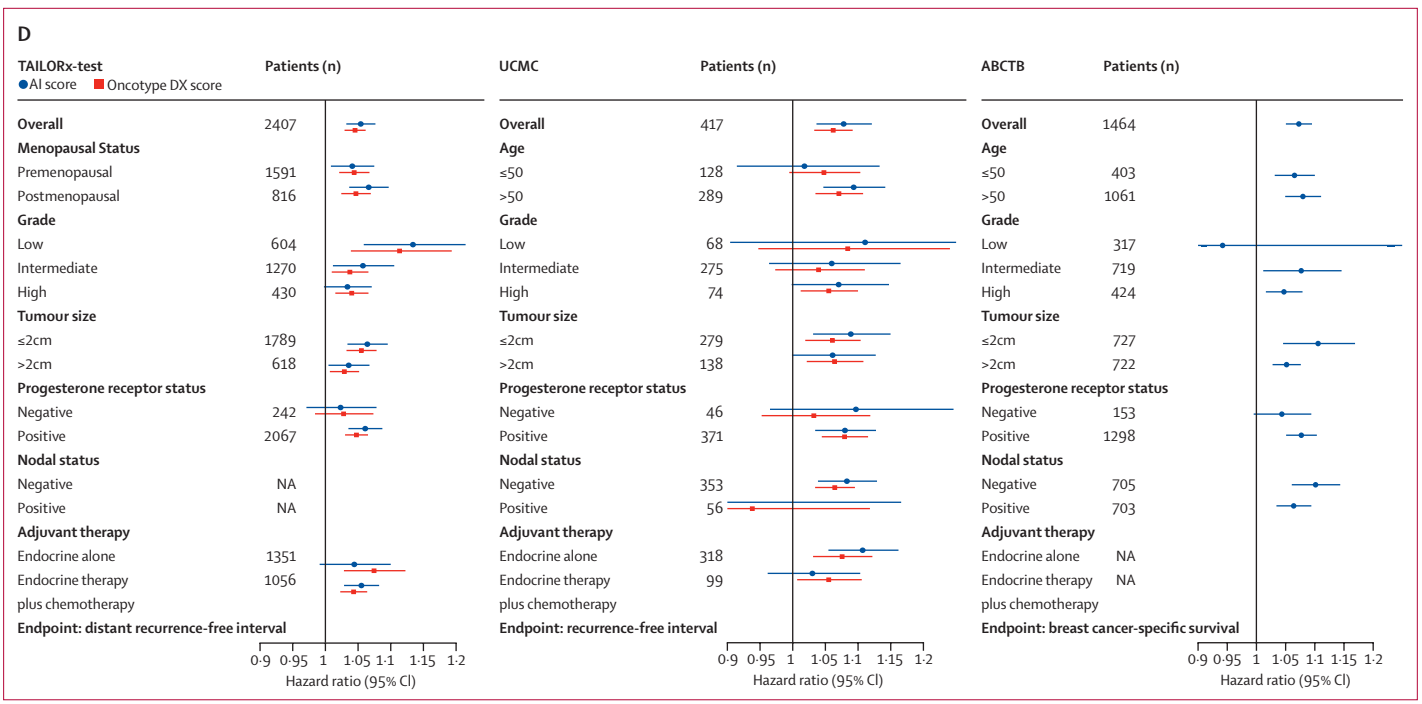


Figure 2: **Survival analysis comparing recurrence score and AI for prognostication in TAILORx-Test**
(A) Kaplan–Meier plots comparing patient stratification using recurrence score versus AI scores across TAILORx-Test patients, using distant recurrence-free, recurrence-free interval, and disease-free survival probabilities, and a recurrence score stratification threshold of 26. (B) Analysis of the same endpoints in patients who received chemotherapy. (C) Corresponding analysis for patients who did not receive chemotherapy, for a lower classification threshold of 21. For each subgroup analysis, patient numbers, p values for the stratification, and HRs with 95% CIs are shown. (D) Forest plots comparing HRs with 95% CIs for the AI scores and recurrence score across multiple patient subgroups in three cohorts with long-term outcome data. The number of patients in each subgroup is indicated. Note that TAILORx included only node-negative patients and that ABCTB did not have recurrence score data, showing only AI prognostic performance. Wald tests show that HRs derived from AI and recurrence score were not statistically different across all subgroups (p=0·11–0·97). ABCTB=Australian Breast Cancer Tissue Bank. HR=hazard ratio. UCMC=University of Chicago Medical Center.

associated with recurrence risk (appendix pp 8–10). These analyses suggest that the predictions were based on biologically meaningful morphology rather than spurious image artifacts.

On the TAILORx-test set, the AI model predictions correlated strongly with ODX recurrence score (r=0·728; p<0·0001; figure 1D). Feature-importance analyses showed that, when only clinical variables were used, grade contributed most, followed by progesterone receptor, oestrogen receptor, age, and tumour size. When H&E images were added, image-derived features became the dominant predictors, followed by progesterone receptor and oestrogen receptor, whereas the grade and other clinical variables added no additional information, showing that the image features captured most of the information associated with histological grade (figure 1E, F).

When evaluated for the clinically relevant task of identifying high genomic-risk disease (recurrence score ≥26; 383 [15·9%] patients), the AI model achieved an AUC of 0·898 (95% CI 0·879–0·913; figure 1G; appendix p 14) in the TAILORx-test cohort, outperforming both the clinicopathological-only and image-only models (p<0·0001). Complementary performance measures, including area under the precision-recall curve, balanced accuracy, F1 score, and Spearman correlation, supported these findings (appendix pp 14, 21). A baseline fully supervised model trained with TAILORx data achieved AUCs of 0·851 (95% CI 0·829–0·872) on the TAILORx-test set, whereas our foundation-model-based approach obtained higher discrimination (appendix p 15). Bench marking against DLRS achieved an AUC of 0·817 (95% CI 0·792–0·840) on the TAILORx-test set (appendix p 15).

To make the model clinically applicable, thresholds of 16 and 26 were selected to maximise NPV, PPV, and affected patients based on TAILORx-CV (appendix p 22). Using these cutoffs in TAILORx-test, the model classified 1097 (45·6%) patients as AI low-risk, 1021 (42·4%) as intermediate, and 289 (12·0%) as AI high-risk (figures 1H, I; appendix p 16). For the low-risk threshold, sensitivity was 0·948 and NPV was 0·982 against a recurrence score of 26 or greater, indicating that only 20 (1·8%) patients classified as AI low-risk had high genomic risk. For the high-risk threshold, specificity was 0·959 and PPV was 0·716 against a recurrence score of 26 or greater. Overall, for the 1386 (57·6%) patients classified as either AI high or low risk, the model accurately identified genomic high-risk or low-risk disease. Test performance was consistent with

cross-validation (appendix pp 14, 16, 23), indicating no overfitting.

Concordance between AI-based risk categories and the ODX recurrence score groups was high overall, with most discordance occurring near the classification boundaries (appendix p 24). Review of discordant cases by an experienced pathologist revealed that many of them had pronounced intratumoral heterogeneity. In such cases, different slides from the same tumour produced markedly different AI scores, reflecting this hetrogeneity. Consistent with this finding, in discordant cases, the reported ODX recurrence score often matched the AI score of one of the individual slides. Systematic evaluation revealed substantially higher within-patient variance in AI predictions in outliers than in non-outliers (p=0·0055; appendix p 25).

We assessed the prognostic performance of the AI model using long-term outcomes from the TAILORx-test set (figure 2). When stratified by a threshold of 26, survival curves based on AI scores closely mirrored those derived from the ODX recurrence score. This concordance was consistent across multiple endpoints, including distant recurrence-free interval (HR 2·88 [95% CI 1·73–4·79]), recurrence-free interval (2·61 [1·68–4·04]), and disease-free survival (1·32 [0·92–1·89]; figure 2A). Similar patterns were observed when including cross-validation patients and when assessing overall and breast-cancer-specific survival (appendix p 26). Wald tests confirmed no statistical difference between AI-derived and recurrence score-derived HRs across endpoints (appendix p 17).

The AI model captured prognostic information complementary to the ODX recurrence score: among patients with a recurrence score below 26, those reclassified as being at high risk by AI had significantly lower RFI and DRFI, whereas among patients with a recurrence score of 26 or greater, those reclassified by AI as being at low risk showed numerically higher survival outcomes, although this difference did not reach statistical significance (appendix p 27). Within each chemotherapy treatment subgroup (received chemotherapy or not), stratification by AI and ODX recurrence score maintained strong concordance across all endpoints, with similar HRs (figure 2B, C; appendix p 26). For the no-chemotherapy analysis, we used an AI threshold of 21 instead of 26, as all patients with a recurrence score of 26 or higher in TAILORx were assigned to chemotherapy. Other thresholds produced similar results (appendix p 28).

Across clinical subgroups, including menopausal status, tumour grade, tumour size, PR status, and treatment, the AI model's prognostic performance was similar to that of ODX recurrence score when measured by distant recurrence-free interval (figure 2D).

In the TAILORx-test group, AI high-risk patients are composed of three groups (figure 3A): a recurrence score of 26 or higher (207 [71·6%]), where NSABP-B-20 post-analysis established substantial chemotherapy benefit,[11,12] recurrence score 16–25 (78 [27·0%]), where chemotherapy benefit was shown in node-negative premenopausal women,[13] and a small discordant subset of patients with a recurrence score below 16 (4 [1·4%]). Patients deemed by AI to be at low risk were composed of two groups: those with a recurrence score below 26 (1077 [98·2%]), where previous analyses showed no chemotherapy benefit for postmenopausal patients,[13,15] and a small discordant subset of patients with a recurrence score of 26 or greater (20 [1·8%]).

After the addition of patients with a recurrence score less than 16 to arms B and C, such that their proportion matched that of the small discordant group (1·4%), chemotherapy benefit remained significant for disease-free survival for premenopausal patients (HR 0·63 [95% CI 0·46–0·86]; chemotherapy *vs* no chemotherapy; figure 3B); and for distant recurrence-free survival and recurrence-free survival (appendix p 29), indicating

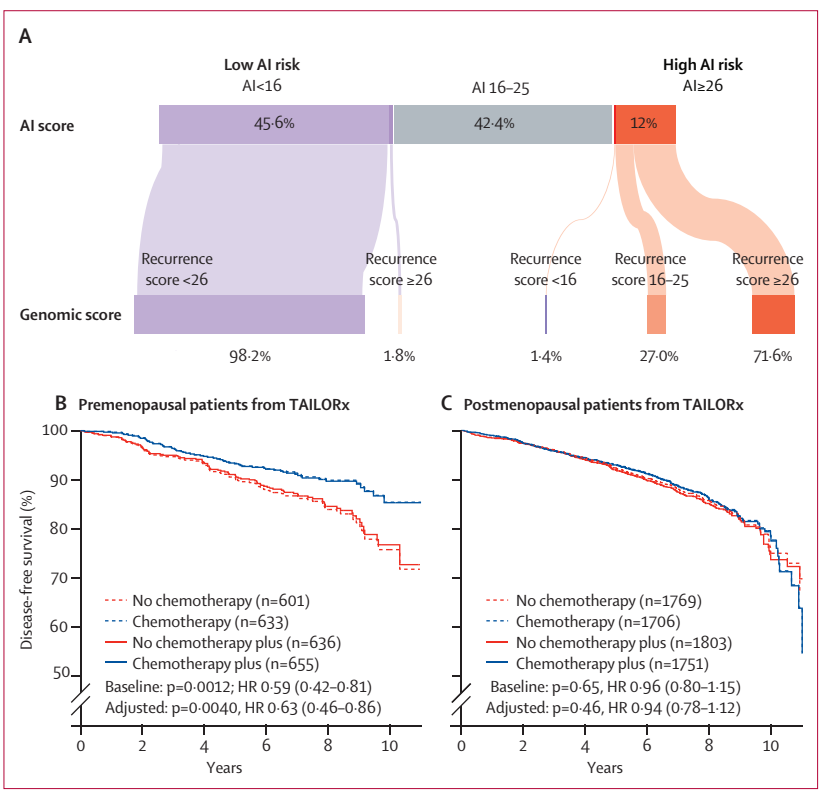
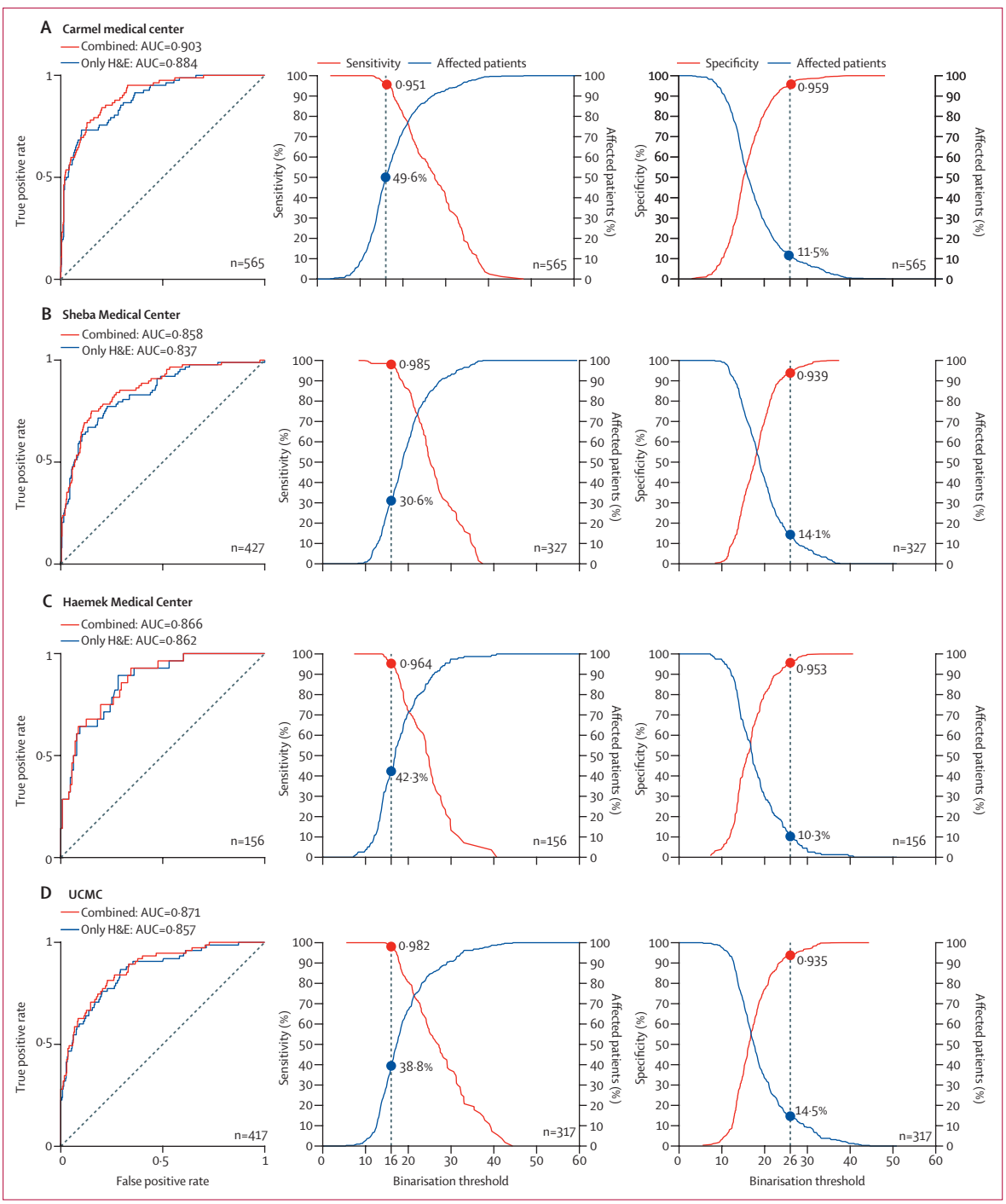


Figure 3: **AI prediction of chemotherapy benefit by menopausal status**
(A) Sankey diagram showing the relationship between AI score categories and genomic recurrence score categories on the TAILORx-Test set (n=2407). The width of the flows represents the proportion of patients in each category. (B) Kaplan–Meier curves of disease-free survival for premenopausal patients in the entire TAILORx cohort. Patients with a recurrence score of 16–25 were randomly assigned to endocrine therapy alone (dashed red) versus endocrine therapy plus chemotherapy (dashed blue). To assess the effect of discordant cases (AI ≥26 but recurrence score <16), we added a randomly sampled subset of patients with recurrence score below 16 to each group (solid lines). (C) Kaplan–Meier curves of disease-free survival for postmenopausal patients in the entire TAILORx cohort with recurrence score 11–25, randomly assigned to endocrine therapy alone (dashed red) versus endocrine therapy plus chemotherapy (dashed blue). To assess the effect of discordant cases (AI <16 but recurrence score ≥26), we added a randomly sampled subset of patients with a recurrence score of 26 or higher patients to each group while correcting for the expected chemotherapy benefit (solid lines). HR=hazard ratio.

chemotherapy benefit for premenopausal patients deemed by AI to be at high risk. After the addition of patients with a recurrence score of 26 or greater to arms B and C, such that their proportion matched that of the small discordant group (1·8%), while correcting for the expected chemotherapy benefit, no chemotherapy benefit was observed for disease-free survival for postmenopausal patients (HR 0·94 [0·78–1·12]; chemotherapy *vs* no chemotherapy) and for distant recurrence-free survival and recurrence-free

survival (appendix p 29). Importantly, the one-sided 10% type-I error corresponded to a 20% increase in disease-free survival risk without chemotherapy, well below the 32·2% non-inferiority margin established in



(Figure 4 continues on next page)

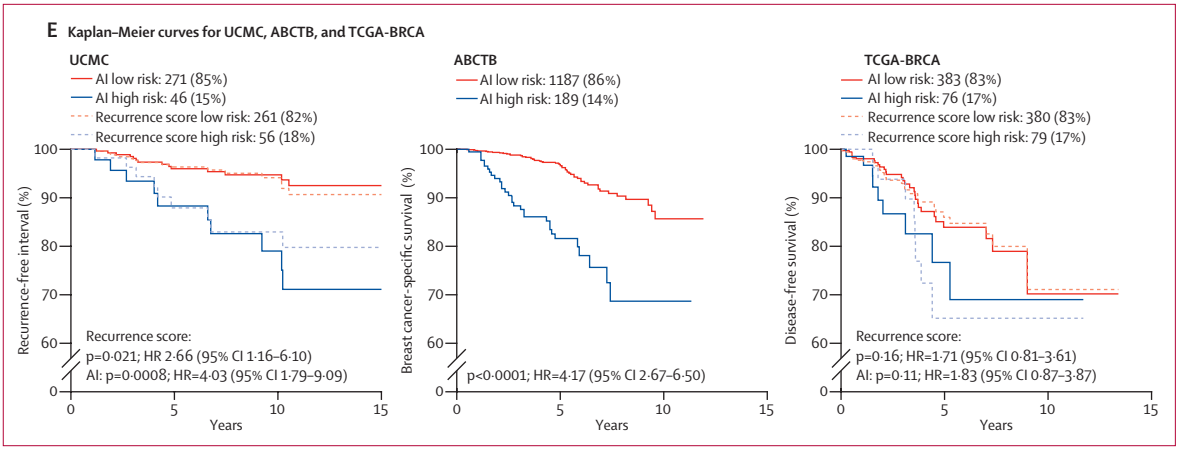


Figure 4: **External validation of the deep learning AI model across independent cohorts**
(A–D) AI model performance at Carmel Medical Center, Sheba Medical Center, Haemek Medical Center, and the UCMC, each with matched recurrence scores. For each institution the left panel shows ROC curves comparing the AI multimodal model using both H&E images and clinicopathological variables (combined) with the image model (only H&E) for identifying high genomic-risk disease (recurrence score ≥26); the middle panel shows sensitivity analysis showing the proportion of patients classified by AI as having low risk who could potentially avoid genomic testing (blue line), and the sensitivity for identifying a recurrence score of 26 or higher (red line) at different classification thresholds (values corresponding to the low classification threshold of 16 are indicated); and the right panel shows specificity analysis showing the proportion of patients classified by AI as having high risk who could potentially proceed directly to chemotherapy (blue line), and the specificity for identifying a recurrence score of 26 or higher (red line) at different classification thresholds. Values corresponding to the high classification threshold 26 are indicated. For sensitivity and specificity analyses at UCMC and Sheba, patients in the calibration set (n=100) were excluded. These patients were included in ROC analyses as calibration does not affect ranking-based metrics. (E) Prognostic performance of recurrence score and AI scores across external cohorts with follow-up data. Kaplan–Meier curves show patient stratification using the high-risk threshold (AI ≥26) for UCMC (endpoint: recurrence-free interval), ABCTB (endpoint: breast cancer-specific survival), and TCGA (endpoint: disease-free interval). For UCMC and TCGA, recurrence score-based stratification (recurrence score ≥26) is shown for direct comparison. For TCGA, recurrence score values were estimated from the gene expression data. HRs with 95% CIs and p values show the consistent prognostic value of AI-derived risk stratification across diverse patient populations. ABCTB=Australian Breast Cancer Tissue Bank. AUC=area under the curve. H&E=haematoxylin and eosin. ROC=receiver operating characteristic. TCGA-BRCA=The Cancer Genome Atlas Breast Invasive Carcinoma. UCMC=University of Chicago Medical Center.

TAILORx for determining non-inferiority.[13] This finding shows that postmenopausal patients classified as low risk by AI did not benefit from chemotherapy. This group consisted of patients with a recurrence score below 26 and a small discordant subset with a recurrence score of 26 or greater. Inclusion of this small subset did not materially alter the treatment-effect estimates.

One-parameter linear scaling calibration reduced the root mean squared error between AI-predicted and measured recurrence scores across all datasets, confirming improved alignment beyond threshold-level stratification (appendix p 32). Across the external cohorts, the calibrated models showed a strong correlation between AI and ODX recurrence score (appendix p 33). The AUC for identifying a recurrence score of 26 or greater ranged from 0·858 to 0·903 (figure 4A–D; appendix p 14). Benchmarking showed that our model outperformed a fully supervised approach (average AUC 0·728), as well as the DLRS method (average AUC 0·812; appendix p 15).

Applying the same clinical thresholds (AI score <16 and ≥26) maintained consistent classification performance across cohorts (figure 4A–D; appendix pp 16, 34). 30–50% of patients were classified as being at low risk and 10–15% at high risk, supporting potential immediate chemotherapy decisions for these subsets. Comparing the AI model's performance before and after calibration showed improvement across cohorts, highlighting the importance of our calibration approach (appendix pp 35–36).

Notably, the performances and proportions of patients deemed to be at low risk and high risk by AI remained consistent across the cohorts, despite the dataset variations, and despite having node-positive patients, which were not included in the TAILORx data. In the UCMC, Sheba, and Haemek cohorts, we identified 983 patients (203 node-positive and 780 node-negative), and model discrimination for predicting high recurrence score was similar in node-positive (AUC 0·874 [95% CI 0·795–0·930]) and node-negative patients (AUC 0·861 [0·821–0·889]), indicating that nodal status was not a confounder for recurrence score prediction.

Additional validation in the TCGA cohort, using estimated recurrence score from RNA sequencing data, showed good discrimination (AUC 0·832 [95% CI 0·779–0·872]) despite estimating recurrence score and working with highly heterogeneous specimens, with 220 (44·7%) classified by AI as being at low risk and 82 (16·7%) as high risk (appendix pp 33–34, 37).

When stratified by the ODX recurrence score of 26 or higher, follow-up data from UCMC showed significant risk separation (HR 2·7 [95% CI 1·2–6·1]; figure 4E). Stratification using AI scores yielded similar HRs and identified a similar proportion of patients at high risk (HR 4·0 [1·8–9·1]). In the ABCTB cohort, the AI model significantly stratified breast cancer-specific

survival (HR 4·2 [2·7–6·5]). For UCMC, the AI model's prognostic HRs closely matched those across clinical subgroups (figure 2D). For ABCTB, the AI HRs matched those of TAILORx and UCMC. Adjusting for treatment did not change these results (appendix p 19). In TCGA, the AI model had numerically higher HRs than the ODX recurrence score, although the difference was not statistically significant. Wald tests confirmed no significant difference between AI-derived and recurrence score-derived HRs in UCMC (p=0·34) or TCGA (p=0·88).

Among node-negative patients in both TAILORx and external cohorts, 4·5–5·2% of premenopausal patients who were clinically at low risk were upgraded by AI to high risk (figure 5). More substantially, 151 (31·3%) of postmenopausal patients who were clinically at high risk were downgraded by AI to low risk. Chemotherapy benefit analysis within each MINDACT risk category confirmed significant chemotherapy benefit in patients deemed by AI to be at high risk despite being classified as being at low risk by conventional clinicopathological criteria, and confirmed no chemotherapy benefit in patients deemed by AI to be at low risk despite being clinically at high risk (appendix p 38). In node-positive postmenopausal patients, where there is no chemotherapy benefit for recurrence score below 26,[15] the model reclassified 350 (40·7%) of patients who were clinically at high risk as being at low risk. We observed that, of patients classified as being at low risk by our model, only 17 (3·4%) had a recurrence score of 26 or higher. Furthermore, the vast majority of patients had a recurrence score below 31 (NPV 0·994 for recurrence score <31, meaning that only three [0·6%] patients classified by AI as low risk had recurrence score ≥31), a range in which chemotherapy benefit remains uncertain.[54]

Discussion

In this study, we developed and validated a multimodal deep-learning model that estimates ODX recurrence score directly from H&E-stained slides and clinicopathological variables, leveraging the TAILORx RCT. To our knowledge, this is the first digital pathology

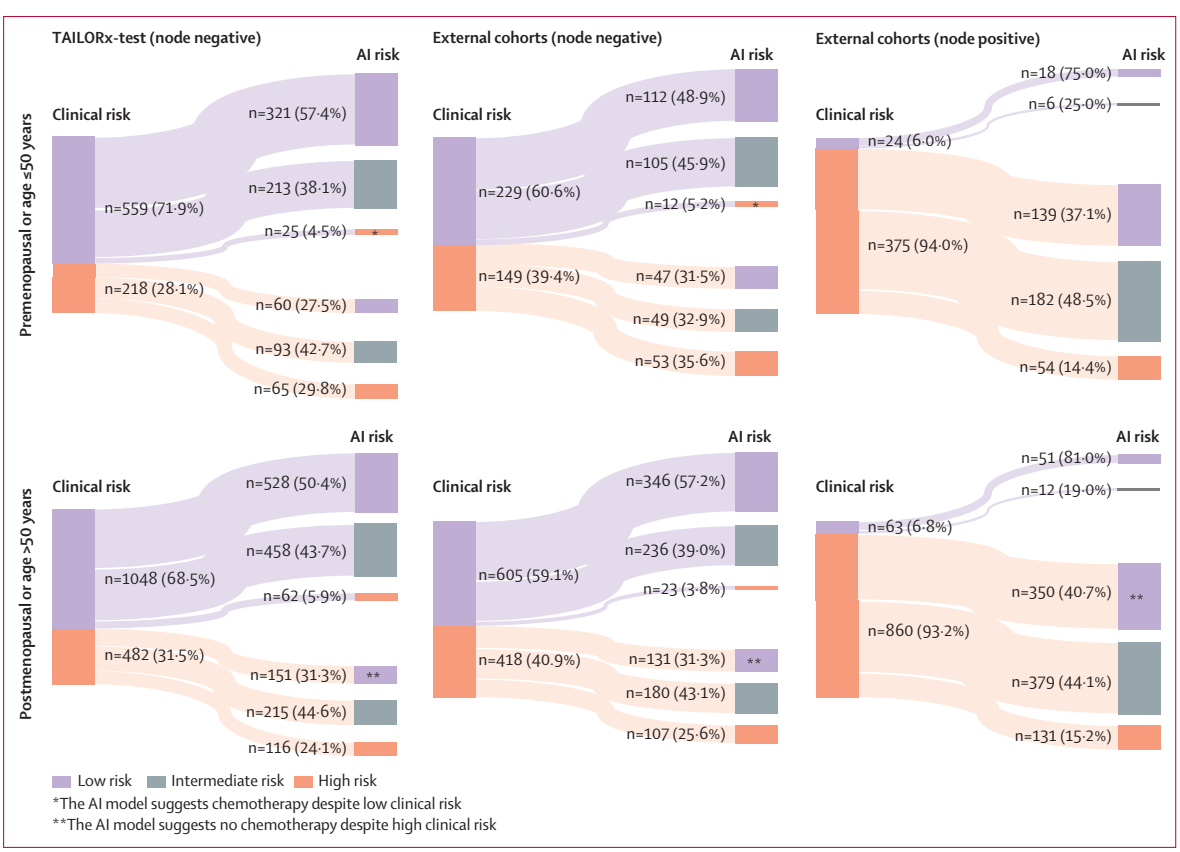


Figure 5: **Clinical risk reclassification**

Sankey diagrams showing how patients classified as high or low risk by traditional clinical risk assessment (MINDACT definition; appendix p 4) would be reclassified using the AI model across three patient groups: TAILORx-test set (left), node-negative patients in the external cohorts (middle), and node-positive patients in the external cohorts (right). Diagrams are stratified by menopausal status for TAILORx and by age for the external cohorts (premenopausal or age ≤50 years, top, and postmenopausal or age >50 years, bottom). The width of connecting flows indicates the proportion of patients, with exact numbers and percentages shown. Blue represents low-risk classification (low clinical risk or AI <16), red represents high-risk classification (high clinical risk or AI ≥26), and grey represents intermediate AI risk classification (AI 16–25). Asterisks indicate a potential impact on chemotherapy decisions that are based on clinical risk and would follow the AI risk. AI=artificial intelligence.

AI model for recurrence-score estimation to be assessed retrospectively using data from an RCT, leveraging randomised treatment assignment to assess chemotherapy benefit in predefined subgroups. Using a large-scale, pre-trained, foundation model and a distribution-matching calibration strategy that does not require genomic labels, the model achieved robust generalisation across six independent cohorts (>5000 patients) despite differences in populations, laboratory protocols, and scanners. Among node-negative postmenopausal patients with an AI threshold below 16, 1077 (98·2%) had a recurrence score below 26, consistent with TAILORx findings that such patients derive no chemotherapy benefit, and among node-negative premenopausal patients with an AI score of at least 26, 285 (98·6%) had a recurrence score of 16 or higher, consistent with observed benefit. Our model refines risk classification within standard clinicopathological groups, identifying around 5% of clinically low-risk premenopausal patients who might benefit from chemotherapy, while downgrading 30–40% of clinically high-risk postmenopausal patients who are likely to be overtreated. Together, these results show that our approach is accurate, generalisable, deployable, and clinically meaningful, complementing genomic testing and providing an affordable alternative where genomic testing is unavailable.

Based on our analyses, node-negative postmenopausal patients with low AI-derived scores (<16) should generally avoid chemotherapy, whereas node-negative premenopausal patients with high scores derived by AI (≥26) should receive chemotherapy, and might be candidates for intensified systemic strategies, including consideration of CDK4/6 inhibitors such as ribociclib. For all other node-negative patients, analysis of TAILORx data could not provide evidence for chemotherapy benefit; such patients should follow standard-of-care treatment (genomic testing when available and clinicopathologic criteria otherwise). This framework offers a pragmatic pathway for integrating AI-based risk assessment into standard decision making.

Our model also generalised to node-positive patients despite being trained exclusively on node-negative cases, indicating it captures biological signals reflective of genomic risk regardless of nodal status. RxPONDER[15] showed that all premenopausal node-positive patients benefit from chemotherapy, whereas postmenopausal node-positive patients with a recurrence score below 26 do not, and observational studies reported uncertain benefit for a recurrence score of 26–30.[54] While chemotherapy benefit could not be directly assessed in node-positive patients using TAILORx, only 17 (3·4%) patients classified as low risk by AI had a recurrence score of 26 or higher, and only three (0·6%) had a recurrence score of 31 or higher, indicating that node-positive postmenopausal patients classified as low risk by AI by our model would probably not benefit from chemotherapy.

Our findings build on and extend previous studies that showed the feasibility of identifying high ODX recurrence score from histopathology images using various computational approaches.[37–43,55–59] Although these studies have laid important groundwork, our approach distinguishes itself in several key aspects. First, validation on TAILORx enabled evaluation on one of the largest breast cancer RCTs designed for ODX recurrence score. Crucially, TAILORx enabled addressing treatment confounders and assessing chemotherapy benefit in patient subgroups, suggesting applicability across diverse patient populations, including those with higher-grade, larger tumours, and later stage at presentation, as are more commonly encountered in resource-limited settings. Second, extensive external validation showed robust generalisation across clinical settings and addressed calibration issues. Although in TCGA stratification by recurrence score produced weaker separation, this finding probably reflects dataset heterogeneity and outdated treatment regimens relative to current practice. Third, our foundation model approach showed significantly superior performance compared with the state-of-the-art fully supervised learning methods across all validation cohorts, and particularly in external validation, where domain shifts typically challenge generalisation. Analysis of discordant cases revealed that intratumoral heterogeneity is a key contributor to differences between AI and genomic scores. In such cases, the reported ODX recurrence score often matched the AI prediction from one of the individual slides, with significantly higher within-patient variance among outliers (p=0·0055; appendix p 25). Whereas genomic testing assays a single tissue sample, AI predictions are aggregated across multiple slides, potentially capturing a more comprehensive assessment of tumour heterogeneity.

Several commercial AI pathology tools have sought to infer outcomes directly from H&E images.[33,35,60–65] Although such models might hold important advantages, potentially outperforming existing genomic tests and enabling broader access to molecular insights, their clinical impact remains uncertain because they have not been validated in RCTs. Without such validation, these models can show prognostic but not predictive value, meaning they cannot establish whether the inferred risk actually corresponds to chemotherapy benefit. Such randomised trials are rare, and observational datasets further limit evaluation, as few untreated high-risk patients exist, making direct estimation of treatment benefit infeasible. In our own ABCTB and UCMC analyses, emulating randomisation was impossible because nearly all patients deemed at high risk by AI received chemotherapy, eliminating overlap between treated and untreated groups. This finding underscores the challenge of evaluating treatment benefit in

observational cohorts. By contrast, training models to infer recurrence score leverages the established clinical utility of genomic assays, enabling indirect yet clinically meaningful estimation of benefit.

Several limitations warrant consideration. First, chemotherapy-benefit analyses were constrained by the TAILORx design, which randomly assigned only patients with a recurrence score of 11–25, limiting direct evaluation in patients with a recurrence score of 26 or higher. Future validation in trials such as NSABP-B-20, which included randomisation of patients at high risk, could strengthen predictive evidence and refine threshold selection. Second, although the model performed well in node-positive patients, validation on randomised data such as RxPONDER would confirm its applicability to this group. Third, our study was retrospective; prospective implementation studies are needed to confirm real-world effectiveness. Fourth, accurate progesterone receptor, oestrogen receptor, and HER2 testing are required to identify eligible hormone receptor-positive, HER2-negative patients, which might be inconsistent in low-income and middle-income countries (LMICs). Integration of AI-based molecular subtype classifiers might help address this gap. Finally, although our datasets span multiple regions, some populations, particularly from LMICs, remain under-represented. Differences in demographics, tumour characteristics, and slide preparation could bias the results and affect generalisability, although our subgroup analyses by age, grade, and tumour size suggest stable performance across relevant clinical contexts. Future validation in these settings will be essential.

Translation into practice will depend on technical performance, adoption, cost, and regulation. ODX use exceeds 70% in the USA and is nearly universal in Israel, but remains uneven elsewhere: broadly reimbursed in Western Europe yet limited in Eastern regions,[20] and minimal in low-resource settings (around 5% in India, <10% across most of Asia, negligible in Africa[19]). In the USA, although cost-effective versus empiric chemotherapy (US$10–15 000 per patient), ODX testing remains expensive (approximately $3500 per assay) and slow, with insurers spending $300–600 million annually.[66] By contrast, digital slide scanning costs less than $1 per slide, and an AI test could deliver results within days[67] at far lower cost.[68] Regulatory approval is feasible: the US Food and Drug Administration has already cleared predictive pathology AI tools,[34] and a Clinical Laboratory Improvement Amendments-based pathway similar to ODX is viable. In LMICs, where chemotherapy costs around US$1000 per patient[69] and 80–85% of patients with hormone receptor-positive, HER2-negative disease still receive the treatment,[70] our model could provide a cost-effective alternative to reduce overtreatment and toxicity.

In conclusion, deep learning applied to routine H&E slides enables rapid, cost-effective, risk stratification in hormone receptor-positive, HER2-negative early breast cancer with accuracy similar to genomic assays. Our model has been validated across diverse populations, requires no specialised genomic infrastructure, and could help in democratising precision oncology. Future work should focus on prospective validation to further establish this approach as an accessible, efficient, and globally scalable tool for breast cancer treatment decisions.

Contributors

GS conceptualised the study and designed the experiments with input from YB and RK. GS and DA performed the formal analyses and led the investigation. GS, SC, RK, and DA developed the methodology. GS and SC developed the software and created the models. JAS provided expert guidance on the interpretation of the TAILORx trial data and contributed to the validation and critical revision of the manuscript. Data curation, annotation, and pathology interpretation were contributed by GS, SC, ES, AC, TG, GBS, CM, IB, AP, FMH, ATP, and DH. Resources were provided by GS, ES, AC, TG, GBS, CM, IB, AP, FMH, ATP, DH, JAS, and RK. GS, RK, and DA performed the visualidation and validation. GS and RK contributed to project administration. GS and DA prepared the original draft; writing, review, and editing were done by all authors. GS and RK supervised the project and secured funding. GS, RK, and DA had full access to the raw data in the study and take responsibility for the integrity of the data and the accuracy of the analyses. GS and DA verified the data in the study. All authors had access to the data reported in the study and approved the decision to submit the manuscript.

Data sharing

The source code used in this study has been archived at https://zenodo.org/records/15422423, and can be accessed at the GitHub repository at https://github.com/shachar5020/TransformerWSI4OncoDXPrediction. The data were composed of seven independent cohorts. The TCGA-BRCA dataset is publicly available at https://portal.gdc.cancer.gov. The ABCTB dataset is accessible from the Australian Breast Cancer Tissue Bank, subject to ethical and scientific approvals as described in their access policy at https://nsw.biobanking.org/biobanks/view/7. The TAILORx dataset can be requested from the ECOG-ACRIN Cancer Research Group and the NCTN/NCORP Data Archive. The remaining data collected from medical centres are not available for public access due to privacy and ethical considerations, in alignment with the Helsinki agreements and institutional policies. Interested researchers may request access directly from the respective institutions. Sample sizes and descriptions for each cohort are provided throughout the manuscript. Data used in this study were collected in accordance with relevant regulations and ethics approvals from the respective institutions.

Declaration of interests

ATP reports personal fees from the Prelude Therapeutics Advisory Board, Elevar Advisory Board, AbbVie consulting, Ayala Advisory Board, ThermoFisher Advisory Board, Break Through Cancer Scientific Advisory Board, Daiichi-Sankyo Advisory Board, Merck research funds, Kura Oncology research funds, and EMD Serono research funds. FMH reports receiving personal fees from Novartis and Leica Biosystems outside the submitted work. AP reports a relationship with Indica Labs that includes consulting. JAS reports consulting, advisory, or other professional relationships with Roche/Genentech, Novartis, AstraZeneca, Pfizer, UpToDate, the American Association for Cancer Research, and Physician Education Resource. DA is a consultant to Link Cell Therapies. All other authors declare no competing interests.

Acknowledgments

This research was supported by the Israel Innovation Authority (Kamin; grant 69997), the Zimin Institute for Artificial Intelligence Solutions in Healthcare grant, the Israel Precision Medicine Partnership programme grant 3864/21, and the Israel Cancer Research Fund (grant 1281495). We would like to thank Karin Stoliar for helping with the data acquisition

and quality assurance, Hen Davidov for supporting the deep learning experiments, and Liat Dizengoff for managing the Helsinki approvals in Carmel Medical Center. YB, DH, JAS, and FMH receive support from the US National Institutes of Health. This manuscript was prepared using data from Datasets [PACCT-1] from the NCTN/NCORP Data Archive of the NCI. Data were originally collected from clinical trial NCT00310180, Program for the Assessment of Clinical Cancer Tests (PACCT-1): Trial Assigning Individualized Options for Treatment: The TAILOR Trial. All analyses and conclusions in this manuscript are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, the NCORP or the NCI. TAILORx was conducted by the ECOG-ACRIN Cancer Research Group (Peter J O'Dwyer and Mitchell D Schnall, Group Co-Chairs) and supported by the NCI of the NIH under award numbers U10CA180820 and U10CA180794. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Additional support was provided by the Breast Cancer Research Foundation under awards CONS-21-007, CONS-20-006, CONS-19-006, and CONS-18-007. For the ABCTB data, tissues and samples were received from the ABCTB, which is generously supported by the National Health and Medical Research Council of Australia, the Cancer Institute NSW, and the National Breast Cancer Foundation. The tissues and samples are made available to researchers on a non-exclusive basis.

References

1 Siegel RL, Kratzer TB, Giaquinto AN, Sung H, Jemal A. Cancer statistics, 2025. *CA Cancer J Clin* 2025; **75:** 10–45.
2 Huppert LA, Gumusay O, Idossa D, Rugo HS. Systemic therapy for hormone receptor-positive/human epidermal growth factor receptor 2-negative early stage and metastatic breast cancer. *CA Cancer J Clin* 2023; **73:** 480–515.
3 Andre F, Ismaila N, Allison KH, et al. Biomarkers for adjuvant endocrine and chemotherapy in early-stage breast cancer: ASCO guideline update. *J Clin Oncol* 2022; **40:** 1816–37.
4 Wallden B, Storhoff J, Nielsen T, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med Genomics* 2015; **8:** 54.
5 Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004; **351:** 2817–26.
6 Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* 2023; **41:** 3565–75.
7 Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N Engl J Med* 2016; **375:** 717–29.
8 Henry NL, Somerfield MR, Abramson VG, et al. Role of patient and disease factors in adjuvant systemic therapy decision making for early-stage, operable breast cancer: update of the ASCO endorsement of the cancer care Ontario guideline. *J Clin Oncol* 2019; **37:** 1965–77.
9 National Comprehensive Cancer Network. NCCN clinical practice guidelines in oncology: breast cancer. Version 4.2025. https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1419 (accessed May 7, 2025).
10 Albain KS, Barlow WE, Shak S, et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *Lancet Oncol* 2010; **11:** 55–65.
11 Geyer CE Jr, Tang G, Mamounas EP, et al. 21-Gene assay as predictor of chemotherapy benefit in HER2-negative breast cancer. *NPJ Breast Cancer* 2018; **4:** 37.
12 Sparano JA, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. *J Clin Oncol* 2008; **26:** 721–28.
13 Sparano JA, Gray RJ, Makower DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N Engl J Med* 2018; **379:** 111–21.
14 Raab R, Ismaila N, Andre F, Stearns V, Kalinsky K. Biomarkers for adjuvant endocrine and chemotherapy in early-stage breast cancer: ASCO guideline update Q and A. *JCO Oncol Pract* 2022; **18:** 646–48.
15 Kalinsky K, Barlow WE, Gralow JR, et al. 21-gene assay to inform chemotherapy benefit in node-positive breast cancer. *N Engl J Med* 2021; **385:** 2336–47.
16 Sparano JA, Gray RJ, Ravdin PM, et al. Clinical and genomic risk to guide the use of adjuvant therapy for breast cancer. *N Engl J Med* 2019; **380:** 2395–405.
17 Horgan D, Hofman P, Giacomini P, et al. Challenges and barriers for the adoption of personalized medicine in Europe: the case of Oncotype DX Breast Recurrence Score test. *Diagnosis* 2024; **12:** 175–81.
18 O'Neill SC, Isaacs C, Chao C, et al. Adoption of gene expression profiling for breast cancer in us oncology practice for women younger than 65 years. *J Natl Compr Canc Netw* 2015; **13:** 1216–24.
19 Batra A, Patel A, Gupta VG, et al. Oncotype DX: Where does it stand in India? *J Glob Oncol* 2019; **5:** 1–2.
20 Jarząb M, Litwiniuk M, Innis P, et al. The utility of the 21-gene Oncotype DX Breast Recurrence Score® assay in node-negative breast cancer patients - the final analysis of the Polish real-life survey PONDx. *Contemp Oncol* 2024; **28:** 245–52.
21 de Jongh FE, Efe R, Herrmann KH, Spoorendonk JA. Cost and clinical benefits associated with Oncotype DX test in patients with early-stage HR+/HER2- node-negative breast cancer in the Netherlands. *Int J Breast Cancer* 2022; **2022:** 5909724.
22 Gupta R, Nandgaonkar S, Kurian N, et al. EGFR mutation prediction of lung biopsy images using deep learning. In: Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies SCITEPRESS, 2023: 102–09.
23 Ström P, Kartasalo K, Olsson H, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol* 2020; **21:** 222–32.
24 Shamai G, Binenbaum Y, Slossberg R, et al. Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer. *JAMA Netw Open* 2019; **2:** e197700.
25 Naik N, Madani A, Esteva A, et al. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. *Nat Commun* 2020; **11:** 5727.
26 Gamble P, Jaroensri R, Wang H, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Commun Med* 2021; **1:** 14.
27 Bychkov D, Linder N, Tiulpin A, et al. Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. *Sci Rep* 2021; **11:** 4037.
28 Shamai G, Livne A, Polónia A, et al. Deep learning-based image analysis predicts PD-L1 status from H&E-stained histopathology images in breast cancer. *Nat Commun* 2022; **13:** 6753.
29 Shamai G, Schley R, Cretu A, et al. Clinical utility of receptor status prediction in breast cancer and misdiagnosis identification using deep learning on hematoxylin and eosin-stained slides. *Commun Med* 2024; **4:** 276.
30 Qu H, Zhou M, Yan Z, et al. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. *NPJ Precis Oncol* 2021; **5:** 87.
31 Chen M, Zhang B, Topatana W, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis Oncol* 2020; **4:** 14.
32 Hoang D-T, Dinstag G, Shulman ED, et al. A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. *Nat Cancer* 2024; **5:** 1305–17.
33 Amgad M, Hodge JM, Elsebaie MAT, et al. A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer. *Nat Med* 2024; **30:** 85–97.
34 Spratt, DE, Tang S, Sun Y, et al. Artificial intelligence predictive model for hormone therapy use in prostate cancer. *NEJM Evid* 2023; **2:** 8.
35 Chen Y, Li H, Janowczyk A, et al. Computational pathology improves risk stratification of a multi-gene assay for early stage ER+ breast cancer. *NPJ Breast Cancer* 2023; **9:** 40.
36 Boehm KM, Aherne EA, Ellenson L, et al, and the MSK MIND Consortium. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer* 2022; **3:** 723–33.

37 Howard FM, Dolezal J, Kochanny S, et al. Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *NPJ Breast Cancer* 2023; **9:** 25.

38 Goyal M, Marotti JD, Workman AA, et al. A multi-model approach integrating whole-slide imaging and clinicopathologic features to predict breast cancer recurrence risk. *NPJ Breast Cancer* 2024; **10:** 93.

39 Boehm KM, El Nahhas OSM, Marra A, et al. Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. *Nat Commun* 2025; **16:** 2106.

40 Cho SY, Lee JH, Ryu JM, et al. Deep learning from HE slides predicts the clinical benefit from adjuvant chemotherapy in hormone receptor-positive breast cancer patients. *Sci Rep* 2021; **11:** 17363.

41 Li H, Wang J, Li Z, et al. Deep learning-based pathology image analysis enhances Magee feature correlation with oncotype DX breast Recurrence Score. *Front Med* 2022; **9:** 886763.

42 Whitney J, Corredor G, Janowczyk A, et al. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. *BMC Cancer* 2018; **18:** 610.

43 Romo-Bucheli D, Janowczyk A, Gilmore H, Romero E, Madabhushi A. Automated tubule nuclei quantification and correlation with oncotype DX risk categories in ER+ breast cancer whole slide images. *Sci Rep* 2016; **6:** 32706.

44 Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. *Nat Med* 2024; **30:** 850–62.

45 Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* 2024; **630:** 181–88.

46 Ding T, Wagner SJ, Song AH, et al. A multimodal whole-slide foundation model for pathology. *Nat Med* 2025; **31:** 3749–61.

47 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017: 5998–6008.

48 Oquab M, Darcet T, Moutakanni T, et al. DINOv2: learning robust visual features without supervision. *arXiv* 2023; published online April 14. https://doi.org/10.48550/arXiv.2304.07193 (preprint).

49 Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021.

50 Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations (SimCLR). *arXiv* 2020; published online Feb 13. https://doi.org/10.48550/arXiv.2002.05709 (preprint).

51 Carpenter JE, Marsh D, Mariasegaram M, Clarke CL. The Australian Breast Cancer Tissue Bank (ABCTB). *Open J Bioresources* 2014; **1:** e1.

52 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490:** 61–70.

53 Habel LA, Shak S, Jacobs MK, et al. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. *Breast Cancer Res* 2006; **8:** R25.

54 Rotem O, Peretz I, Leviov M, et al. Clinical outcomes in estrogen receptor-positive early-stage breast cancer patients with Recurrence Score 26-30: observational real-world cohort study. *NPJ Breast Cancer* 2023; **9:** 49.

55 Li H, Whitney J, Bera K, et al. Quantitative nuclear histomorphometric features are predictive of Oncotype DX risk categories in ductal carcinoma in situ: preliminary findings. *Breast Cancer Res* 2019; **21:** 114.

56 Su Z, Niazi MKK, Tavolara TE, et al. BCR-Net: a deep learning framework to predict breast cancer recurrence from histopathology images. *PLoS One* 2023; **18:** e0283562.

57 Romo-Bucheli D, Janowczyk A, Gilmore H, Romero E, Madabhushi A. A deep learning based strategy for identifying and associating mitotic activity with gene expression derived risk categories in estrogen receptor positive breast cancers. *Cytometry A* 2017; **91:** 566–73.

58 Su Z, Rosen A, Wesolowski R, et al. Deep-ODX: an efficient deep learning tool to risk stratify breast cancer patients from histopathology images. In: Tomaszewski JE, Ward AD, eds. Medical Imaging 2024: Digital and Computational Pathology. SPIE, 2024. 1293307

59 Shamai G, Schley R, Kimmel R, Balint-Lahat N, Barshack I, Mayer C. Abstract 5354: prediction of OncotypeDX high risk group for chemotherapy benefit in breast cancer by deep learning analysis of hematoxylin and eosin-stained whole slide images. *Cancer Res* 2023; **83:** 5354–5354.

60 Witowski J, Zeng K, Cappadona J, et al. Multi-modal AI for comprehensive breast cancer prognostication. *arXiv* 2024; published online Oct 28. https://doi.org/10.48550/arxiv.2410.21256 (preprint).

61 Wahab N, Toss M, Miligy IM, et al. AI-enabled routine H&E image based prognostic marker for early-stage luminal breast cancer. *NPJ Precis Oncol* 2023; **7:** 122.

62 Wang Y, Sun W, Karlsson E, et al. Clinical evaluation of deep learning-based risk profiling in breast cancer histopathology and comparison to an established multigene assay. *Breast Cancer Res Treat* 2024; **206:** 163–75.

63 Shi Y, Olsson LT, Hoadley KA, et al. Predicting early breast cancer recurrence from histopathological images in the Carolina Breast Cancer Study. *NPJ Breast Cancer* 2023; **9:** 92.

64 Garberis I, Gaury V, Saillard C, et al. Deep Learning allows assessment of risk of metastatic relapse from invasive breast cancer histological slides. *bioRxiv* 2022; published online Dec 5. https://doi.org/10.1101/2022.11.28.518158 (preprint).

65 Fernandez G, Zeineh J, Prastawa M, et al. Analytical validation of the PreciseDx digital prognostic breast cancer test in early-stage breast cancer. *Clin Breast Cancer* 2024; **24:** 93–102.e6.

66 Berdunov V, Cuyun Carter G, Laws E, et al. Cost-effectiveness analysis of the Oncotype DX Breast Recurrence Score test from a US societal perspective. *Clinicoecon Outcomes Res* 2024; **16:** 471–82.

67 Hung C, Sin HG, Wu J. Optimizing workflow for OncotypeDx result turnaround time from surgery to report at safety net hospital. *JCO Oncol Pract* 2024; **20:** 317–317.

68 Ardon O, Klein E, Manzo A, et al. Digital pathology operations at a tertiary cancer center: Infrastructure requirements and operational cost. *J Pathol Inform* 2023; **14:** 100318.

69 Wadasadawala T, Mohanty SK, Sen S, et al. Out-of-pocket payment and financial risk protection for breast cancer treatment: a prospective study from India. *Lancet Reg Health Southeast Asia* 2024; **24:** 100346.

70 Bajpai J, Ventrapati P, Joshi S, et al. Unique challenges and outcomes of young women with breast cancers from a tertiary care cancer centre in India. *Breast* 2021; **60:** 177–84.