nature communications

Article

https://doi.org/10.1038/s41467-026-73088-y

Multi-modal AI for comprehensive breast cancer prognostication

Received: 5 May 2025

Accepted: 29 April 2026

Published online: 20 May 2026

Check for updates

Jan Witowski[1], Ken G. Zeng[1], Joseph Cappadona[1], Jailan Elayoubi[2], Khalil Choucair[2], Elena Diana Chiru[3], Nancy Chan[4], Young-Joon Kang[5], Frederick Howard[6], Irina Ostrovnaya[7], Carlos Fernandez-Granda[8,9], Freya Schnabel[4], Zoe Steinsnyder[1], Ugur Ozerdem[4], Kangning Liu[8], Waleed Abdulsattar[2], Yu Zong[2], Lina Daoud[2], Rafic Beydoun[2], Anas M. Saad[2], Nitya Thakore[4], Mohammad Sadic[4], Frank Yeung[4], Elisa Liu[4], Theodore Hill[4], Benjamin Swett[4], Danielle Rigau[4], Andrew J. Clayburn[4], Valerie Speirs[10], Marcus Vetter[3], Lina Sojak[3], Simone Muenst[11], Daniel Baumhoer[11], Jia-Wern Pan[12], Haslina Makmur[12], Soo-Hwang Teo[12], Linda M. Pak[4], Victor Angel[13], Dovile Zilenaite-Petrulaitiene[14], Arvydas Laurinavicius[14], Natalie Klar[4], Brian D. Piening[15], Carlo Bifulco[15], Sun-Young Jun[5], Jae Pak Yi[5], Su Hyun Lim[5], Adam Brufsky[16], Francisco J. Esteva[17], Lajos Pusztai[18], Yann LeCun[8,9,19] & Krzysztof J. Geras[1,4,8] ✉

Treatment selection in breast cancer is guided by risk assessment using molecular subtypes and clinicopathological characteristics. However, current approaches lack the precision required for optimal clinical decision-making. To address this, we use data from 8161 patients to develop and evaluate an AI test integrating digital pathology with clinical data. The AI test provides a robust method for predicting disease-free interval (C-index: 0.71 [0.68-0.75], HR: 3.63 [3.02-4.37, $p < 0.001$]). In a direct comparison, the AI test displays numerically higher discrimination (C-index: 0.67 [0.61–0.74]) than the standard-of-care 21-gene assay (C-index: 0.61 [0.49–0.73]). Across molecular subtypes, the AI test demonstrates robust prognostic performance, including in triple negative breast cancer (C-index: 0.71 [0.62-0.81], HR: 3.81 [2.35-6.17, p=0.02]), where no guideline-recommended assays currently exist. These findings highlight the potential of AI-based pathology tests as a promising tool for improved risk stratification across all major subtypes, with implications for clinical decision-making.

Over the past few decades, breast cancer treatment has evolved significantly with the introduction of various chemo, endocrine, and targeted therapies. Treatment selection is primarily informed by clinical guidelines that rely on clinicopathological factors such as tumor size, nodal status, and estrogen and HER2 receptor status. While this approach optimizes outcomes at the population level to a certain degree, it overlooks critical, actionable information for individual patients, exposing a significant gap in personalized care.

In an attempt to address this limitation, genomic assays were developed in the early 2000s, offering a more refined approach to patient risk stratification. Tools such as Oncotype DX[1], Mammaprint[2], and Prosigna[3] assess the risk of distant recurrence based on gene expression data. However, these assays were developed primarily for

A full list of affiliations appears at the end of the paper. ✉e-mail: krzysztof.geras@nyulangone.org

hormone receptor-positive (HR+) breast cancer patients, mainly to de-escalate adjuvant chemotherapy. In addition to being limited to certain cancer subtypes, their accuracy is modest, often on par with simple models based on clinical characteristics[4,5]. These tests also require the processing of physical tissue specimens, which creates additional work for pathology departments and depletes tissue that could be used in the future for advanced molecular profiling.

Complementing genomic assays, pathologists often evaluate the histopathological characteristics of a tumor to estimate the cancer aggressiveness. These features include histologic grade, tumor-infiltrating lymphocytes level, and Ki-67 expression. These primarily morphological features have been shown to add independent prognostic information to genomic scores, and recent prognostic models integrate both[6,7]. However, these features are limited in their scope and prognostic ability, which drives the need for stronger and more informative prognostic features. In recent years, advances in AI, particularly in self-supervised learning, have enabled the development of more effective methods for learning meaningful features from imaging data[8–13]. These features are not based on pre-specified characteristics, and extracting them does not require manual annotation by pathologists. Instead, AI models autonomously determine the most salient features by learning from millions of images. Early evidence shows that these AI-enabled features extracted from digital pathology images can be used to predict key patient characteristics, genomic signature scores, and long-term outcomes[14–20]. However, many of these approaches do not incorporate clinical data or undergo validation in non-HR+ subtypes.

In response to the gaps in today's treatment selection tools, we developed and evaluated an AI test as a prognostic marker for the risk of cancer recurrence in invasive breast cancer patients. It is built upon the latest advances in self-supervised learning and digital pathology. Specifically, our test extracts strongly generalizable morphological features from standard H&E-stained slides of core needle biopsies or surgical specimens utilizing Kestrel[21], a state-of-the-art pan-cancer AI foundation model. Kestrel is a variant of a vision transformer[22] trained with the self-supervised learning method DINOv2[13] using 400 million digital pathology image patches. The image patches used to train Kestrel were sourced from a large, heterogeneous database of pan-cancer whole slide images, including breast cancer images. Kestrel's hyperparameters were tuned using a methodology that evaluates model performance on a suite of digital pathology benchmarks[21]. This approach iteratively scales model and dataset size, in a chain of hyperparameter searches, yielding a model that is specially tuned for downstream tasks in digital pathology.

In this study, we present the application of a generalist AI foundation model for outcome prediction in breast cancer. Our multi-modal AI test integrates pathology-based features, extracted by Kestrel, with routinely collected clinical variables, namely cancer T and N stage, patient age, ER, PR, and HER2 status, and presence of ductal or lobular histology (Fig. 1). The test produces a continuous score between 0 and 1 that quantifies the risk of cancer recurrence. The test was developed and evaluated using data from a total of 8161 patients across 15 cohorts originating from seven countries. Of these, 3502 patients from five cohorts were used exclusively for external evaluation, while the remaining patients were used for model training. To our knowledge, this is the first large-scale real-world evaluation of an AI-based prognostic model for breast cancer conducted across subtypes. Detailed patient characteristics are described in Table 1. We evaluate our test's prognostic performance across a range of independent test cohorts and clinical subgroups that capture the clinical and demographic variability observed in real-world patient populations. We also compare our test's accuracy to a standard-of-care genomic assay, showing how the introduced multi-modal AI test enhances breast cancer prognostication. Our findings demonstrate that the AI test is able to discriminate recurrence risk across all major breast cancer subtypes, including TNBC, where no guideline-approved prognostic assays currently exist. Our AI test also demonstrates numerically higher performance than a standard-of-care genomic assay in both overall and intermediate-risk patients.

Results

Breast cancer cohorts

To execute this project, we assembled a diverse collection of both public datasets (TCGA-BRCA[23,24], METABRIC[25], BASIS [26], Providence[27]) and private datasets from several sources (NYU Langone Health, The Catholic University of Korea, Gundersen Health, National Center of Pathology in Lithuania, Breast Cancer Now Biobank, Cancer Research Malaysia, Omica.bio Cancer Atlas, Wales Cancer Biobank, Karmanos Cancer Institute, University Hospital Basel, and UChicago Medicine). We excluded patients without documented follow-up or available pathology slides, as well as those with stage IV disease or a history of breast cancer. Altogether, we collected 5162 slides from 4659 breast cancer patients across 10 datasets for training, and 3632 slides from 3502 breast cancer patients from 5 datasets for evaluation, making this one of the largest data collection efforts for predicting breast cancer recurrence. All patients in the data are women. Two of the evaluation datasets, Providence (n = 1733) and TCGA-BRCA (n = 911), represent a broad spectrum of invasive breast cancer patients. The remaining three test cohorts − Karmanos (n = 168), Basel (n = 269), and UChicago (n = 421) − included only HR+ HER2− patients who were previously tested with Oncotype DX. See Table 1 for more detailed statistics of the training and evaluation cohorts.

Study design and results

Unless explicitly stated otherwise, the presented results are from the multi-modal AI test. We first report its prognostic capability, i.e., whether our model accurately predicts the risk of cancer recurrence and distinguishes between high- and low-risk patients. Then, we compare its accuracy against a standard-of-care genomic risk signature in various molecular, histological, and clinical subgroups. Finally, we discuss potential clinical use cases. When reporting performance across multiple datasets, we present pooled results using a random effects model (Section SI.3). The additional results for specific datasets are in Figs. SI.3 and SI.4.

To evaluate how well the test orders patients according to their risk, we report the concordance index (C-index). Additionally, we estimate the hazard ratio (HR) for the AI test in univariate and multivariate Cox proportional hazards models. While the AI test takes values between 0 and 1, the HR presented is for every 0.2 unit increase. Unless explicitly stated otherwise, the HR is univariate and is computed using the continuous AI test risk score, rather than its dichotomized version. By using the continuous score, we highlight how incremental shifts in the score translate to gradual changes in risk, rather than the change in risk that occurs at one specific cutoff. Furthermore, as Oncotype DX's risk categories and corresponding cutoffs were determined using an entirely separate patient population, it is more appropriate to compare results without a specific cutoff. For additional analysis, we dichotomize the score from the AI test. We defined high- and low-risk groups using the 80th percentile cutoff in the HR+ HER2− subpopulation within our largest internal training dataset (NYU). We refer to Methods (Section 4) and Sections SI.1 and SI.2 for a detailed explanation of the performance metrics used in this paper.

The primary endpoint used for training and validation is the disease-free interval (DFI). When appropriate, we also evaluate our test on several secondary endpoints: distant recurrence-free interval (DRFI), late recurrence (LR), recurrence-free survival (RFS), and overall survival (OS). Endpoint definitions can be found in Section SI.4 and Table SI.1.

https://doi.org/10.1038/s41467-026-73088-y

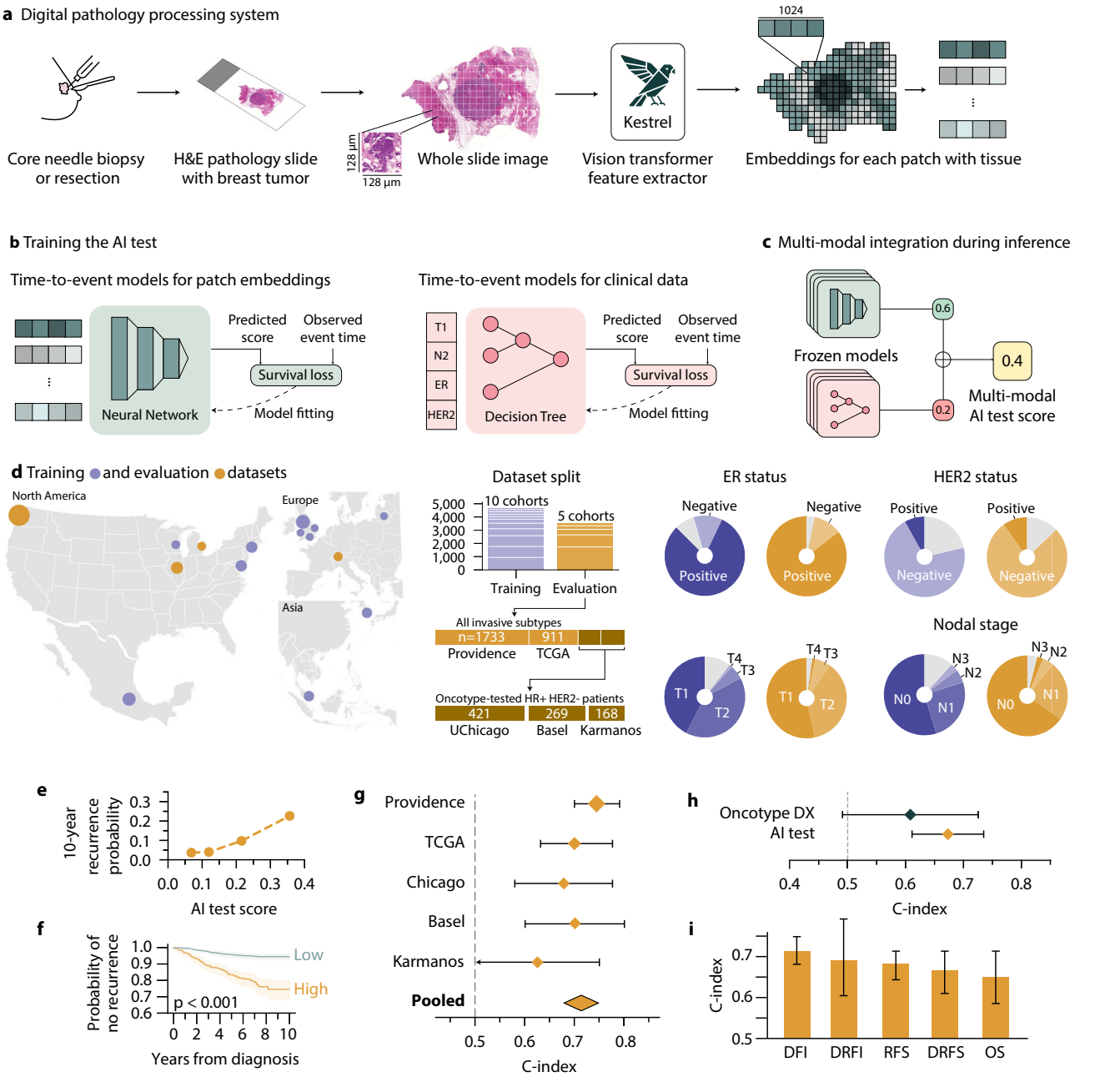

Fig. 1 | We present a multi-modal AI test for invasive breast cancer. a The key component of the test is a system which processes high-resolution digital images of breast cancer specimens. Features are extracted from the digitized slides using Kestrel, a foundation model trained using self-supervised learning on a pan-cancer dataset of 400 million pathology image patches. **b** Features extracted from digitized slides and clinical variables are used to train supervised time-to-event models predicting breast cancer recurrence or death. **c** The AI test produces a multi-modal risk score, integrating pathological and clinical risk scores. **d** We developed the AI test using 10 cohorts from six countries (highlighted in blue, $n = 4659$ patients), and evaluated it on five patient cohorts which were not used during training (highlighted in orange, $n = 3502$ patients). Evaluation sets consisted of two cohorts (Providence and TCGA) with all invasive breast cancer subtypes, and three cohorts (UChicago, Basel, and Karmanos) with only HR+ HER2− patients tested with Oncotype DX. **e** The 10-year recurrence probability increases monotonically as the AI test score increases ($n = 1792$ patients). The four dots represent the four quantiles of the AI test score. 10-year recurrence probabilities were estimated using a Kaplan-Meier estimator. **f** Patients with predicted high risk had significantly worse outcomes compared to patients with predicted low risk ($n = 1792$ patients, $p < 0.001$). The error bars represent the 95% CI. The p-values were computed from a univariate Cox proportional hazards model using a two-sided Wald test, with the AI test score dichotomized. **g** The AI test achieved strong prognostic results across all evaluation datasets ($n = 3502$ patients), as measured by the concordance-index (C-index). Estimates were pooled across datasets using a random effects model. Error bars represent the 95% CI. **h** In a direct comparison ($n = 858$ patients) to a standard-of-care genomic assay, Oncotype DX, the AI test was a better predictor of cancer recurrence, as measured by the C-index pooled across datasets using a random effects model. Error bars represent the 95% CI. **i** the AI test performs consistently across various endpoints (DFI disease-free interval, DRFI distant recurrence-free interval, RFS recurrence-free survival, DRFS distant recurrence-free survival, OS overall survival), as measured by the C-index ($n = 3502$ patients). Endpoint definitions are in Section SI.4. The error bars represent the 95% CI.

https://doi.org/10.1038/s41467-026-73088-y

Table 1 | Datasets used for training and evaluation

Category	Training cohorts (*n* = 4659)	Evaluation cohorts (*n* = 3502)
Age, median [IQR]	58 [49–68]	60 [51–68]
Race		
Asian	380 (8.16%)	157 (4.48%)
Black	71 (1.52%)	321 (9.17%)
White	923 (19.81%)	2461 (70.27%)
Unknown/other	3285 (70.49%)	563 (16.08%)
Follow-up time, median [IQR]	5.11 [3.62–8.31]	4.51 [2.42–7.15]
ER receptor status		
Positive	3748 (80.45%)	2992 (85.44%)
Negative	540 (11.59%)	399 (11.39%)
Unknown	371 (7.96%)	111 (3.17%)
PR receptor status		
Positive	2676 (57.44%)	2648 (75.61%)
Negative	1010 (21.68%)	728 (20.79%)
Unknown	973 (20.88%)	126 (3.60%)
HER2 receptor status		
Positive	377 (8.09%)	348 (9.94%)
Negative	3286 (70.53%)	2702 (77.16%)
Equivocal or unknown	996 (21.38%)	452 (12.91%)
T stage		
T1	1880 (40.35%)	1869 (53.37%)
T2	1780 (38.21%)	1287 (36.75%)
T3	252 (5.41%)	225 (6.42%)
T4	51 (1.09%)	40 (1.14%)
Unknown/TX	464 (9.96%)	81 (2.31%)
N stage		
N0	2550 (54.73%)	2284 (65.22%)
N1	1174 (25.20%)	839 (23.96%)
N2	251 (5.39%)	156 (4.45%)
N3	141 (3.03%)	86 (2.46%)
Unknown/NX	543 (11.65%)	137 (3.91%)
Ductal histology		
Yes	3567 (76.56%)	1392 (39.75%)
No	904 (19.40%)	377 (10.77%)
Unknown	188 (4.04%)	1733 (49.49%)
Lobular histology		
Yes	600 (12.88%)	359 (10.25%)
No	3871 (83.09%)	1410 (40.26%)
Unknown	188 (4.04%)	1733 (49.49%)
Grade		
1	523 (11.23%)	332 (9.48%)
2	1880 (40.35%)	1287 (36.75%)
3	1286 (27.60%)	722 (20.62%)
Unknown	970 (20.82%)	1161 (33.15%)
Any recurrence		
Yes	726 (15.58%)	279 (7.97%)
No	3933 (84.42%)	3223 (92.03%)
Distant recurrence		
Yes	457 (9.81%)	174 (4.97%)
No	4098 (87.96%)	3328 (95.03%)
Unknown	104 (2.23%)	0 (0.00%)
Death		
Yes	935 (20.07%)	303 (8.65%)
No	3710 (79.63%)	3199 (91.35%)
Unknown	14 (0.30%)	0 (0.00%)
Adjuvant chemo- and endocrine therapy		
Chemoendocrine therapy	1124 (24.13%)	768 (21.93%)
Chemotherapy only	564 (12.11%)	500 (14.28%)
Endocrine therapy only	1551 (33.29%)	1442 (41.18%)
Unknown	729 (15.65%)	368 (10.51%)

All values are *N* (%) unless indicated otherwise. Follow-up time statistics are computed taking into account all patients (censored and not censored). Clinical characteristics of the patients in the training and test cohorts are significantly different ($p < 0.001$). We used Chi-squared tests for all categorical variables and the Mann–Whitney U test for patient age. Additionally, the log-rank test[61] shows a significant difference in outcome distributions between the training and test cohorts ($p < 0.001$)

The AI test is prognostic of recurrence and survival

The AI test was found to be prognostic in all five patient cohorts used as external evaluation sets. For the primary endpoint, disease-free interval (DFI), the risk score generated by our model achieved a pooled C-index of 0.71 [0.68-0.75] and a pooled HR of 3.63 [3.02–4.37, $p < 0.001$].

Two cohorts — Providence and TCGA-BRCA — were composed of a broad population of invasive breast cancer patients. Our AI test achieved a C-index of 0.74 [0.70–0.79] and an HR of 4.02 [3.09-5.23, $p < 0.001$] in the Providence cohort, and a C-index of 0.70 [0.63–0.77] and an HR of 3.00 [2.10–4.28, $p < 0.001$] in the TCGA-BRCA cohort.

Three cohorts — Karmanos, Basel, and UChicago — included only HR+ HER2− patients who were previously tested with Oncotype DX. The AI test achieved a C-index of 0.62 [0.49–0.75] and a HR of 3.82 [1.33–10.98, $p = 0.013$] in the Karmanos cohort ($n = 168$), a C-index of 0.70 [0.60–0.80] and a HR of 3.98 [1.92–8.25, $p < 0.001$] in the Basel cohort ($n = 269$), and a C-index of 0.67 [0.58-0.77] and a HR of 3.26 [1.45-7.31, p = 0.004] in the UChicago cohort ($n = 421$).

To complement the C-index and HR analyses, we also examined the time-dependent AUC to evaluate how prognostic performance varies over time. The AI test demonstrated strong prognostic performance at 3, 5, and 7 years, with pooled time-dependent AUCs of 0.76 [0.70–0.82], 0.71 [0.66–0.76], and 0.73 [0.69–0.77], respectively.

The AI test was also prognostic for secondary endpoints (as defined in Section SI.4). For distant recurrence-free interval (DRFI), the test achieved a C-index of 0.70 [0.61–0.78] and an HR of 4.02 [2.64–6.13, $p = 0.001$]. For overall survival (OS), the test achieved a C-index of 0.65 [0.58–0.72] and an HR of 2.52 [1.88–3.38, $p = 0.001$]. For recurrence-free survival (RFS), the test achieved a C-index of 0.68 [0.65–0.71] and a HR of 2.65 [2.33–3.01, $p < 0.001$]. For distant recurrence-free survival (DRFS), the test achieved a C-index of 0.66 [0.61–0.71] and a HR of 2.64 [2.14–3.26, $p < 0.001$]. Forest plots for all endpoints and metrics are reported in Sections SI.6 and SI.8 and Figs. SI.1, SI.2 and SI.5.

AI test demonstrated higher accuracy than a genomic assay

We compared the prognostic performance and potential clinical utility of our model to Oncotype DX, a standard-of-care 21-gene assay developed to predict distant recurrence risk. Three external cohorts (Karmanos, Basel, UChicago) contain a total of 858 patients who were tested with Oncotype DX Recurrence Score assays performed as part of routine care.

In the Karmanos cohort, the AI test's C-index for DFI was 0.62 [0.49–0.76], and the HR was 3.82 [1.33–10.98, $p = 0.013$], compared with Oncotype DX's C-index of 0.54 [0.41–0.68] and HR of 1.36 [0.56–3.27, $p = 0.500$]. In the Basel cohort, the AI test's C-index was 0.70 [0.60–0.80], and the HR was 3.98 [1.92–8.25, $p < 0.001$],

compared with Oncotype DX's C-index of 0.55 [0.42–0.68] and HR of 1.76 [0.85–3.64, $p = 0.13$]. In the UChicago cohort, the AI test's C-index was 0.67 [0.58–0.77] and the HR was 3.26 [1.45–7.31, $p = 0.004$], compared with Oncotype DX's C-index of 0.71 [0.61–0.81] and HR of 2.78 [1.61–4.81, $p < 0.001$]. Together, the AI test achieved a pooled C-index of 0.67 [0.61–0.74] and HR of 3.67 [2.79–4.84, $p = 0.002$], compared to Oncotype DX's C-index of 0.61 [0.491–0.725] and HR of 2.09 [0.85–5.15, $p = 0.21$].

To assess prognostic accuracy at different time horizons, we computed pooled time-dependent AUCs, where the AI test achieved 0.71 [0.48–0.95], 0.66 [0.57–0.74], and 0.71 [0.59–0.83] at 3, 5, and 7 years, respectively. In comparison, Oncotype DX achieved 0.66 [0.35–0.98], 0.60 [0.22–0.99], and 0.65 [0.40–0.91].

To assess the clinical utility of the AI test and Oncotype DX, we performed a decision curve analysis adapted for survival data at 5 and 10 years (Section SI.11). While both Oncotype DX and the AI test exceed the baseline treat all and treat none strategies, the AI test consistently outperforms Oncotype DX and provides more clinical utility across threshold probabilities and both time points.

These comparisons are illustrated in Fig. 2a and further extended in Section SI.9 and Figs. SI.6 and SI.8.

Oncotype DX has been shown to predict the benefit of adjuvant chemotherapy in HR+ patients. Patients with high Oncotype DX scores had significant chemotherapy benefits, while patients with low Oncotype DX scores did not[28]. Two large randomized clinical trials, TAILORx[29] and RxPONDER[30], have shown that some patients with intermediate Oncotype DX scores might also safely avoid adjuvant chemotherapy. For patients tested with Oncotype DX, the AI model would re-classify 666 out of 858 (77.6%) patients into different risk categories (Fig. 2b). All intermediate Oncotype DX-risk patients (61.3% of all Oncotype DX patients) would be reclassified as either low- or high-risk by our AI test. 423 out of 526 (80.4%) intermediate Oncotype DX-risk patients would be classified as low-risk patients, and 103 of 526 (19.6%) as high-risk patients using the AI test. The outcomes for high/low risk patients are illustrated in Fig. 2c. Within this intermediate Oncotype DX-risk group, the continuous AI test score was significantly associated with recurrence (HR: 3.45 [1.85–6.42, $p < 0.001$]) in the Karmanos, Basel, and UChicago cohorts after adjusting for the dataset (see Table SI.2).

Finally, we demonstrate that the AI test provides prognostic value independent from Oncotype DX, as well as from other common clinical indicators. To do this, we incorporated the AI test score into a multivariate Cox proportional hazards model along with Oncotype DX score, Nottingham cancer grade, dataset, and race (see Section 4 for more details). After including all these factors, the AI test has an adjusted HR of 2.95 [1.82–4.79, $p < 0.001$]. In comparison, Oncotype DX has an adjusted HR of 1.43 [0.91–2.27, $p = 0.12$], which is not statistically significant. These results demonstrate that the AI test provides significant independent prognostic value beyond Oncotype DX, after adjustment for established clinical indicators. Race groups or grade were not significant. Key findings are presented in Fig. 2d. The full results are in Table SI.3 in Section SI.5. To further demonstrate that the AI test is not solely dependent on clinical covariates, we performed a multivariate analysis with the clinical and pathology scores within the AI test. As shown in Fig. 2e and Table SI.4, both the clinical and pathology scores are significantly associated with DFI, and the pathology score is significant after adjusting for the clinical score. Additional analyses demonstrate that the pathology score is significantly associated with DFI after further adjusting for Oncotype DX (see Table SI.5). In a separate analysis, we further verified that the pathology score is significant after adjusting for Oncotype DX and the clinical variables used by our model (see Table SI.6). Additionally, the multivariate Cox models with pathology score demonstrates a significantly better fit compared to models without the score ($p < 0.001$). These results show that the pathology score adds prognostic information that is independent of the information in clinical variables used in our AI test.

AI test performs well in clinically meaningful subgroups

Traditional genomic assays are typically developed for specific patient subgroups, such as Oncotype DX for HR+ HER2− patients, and cannot be used in patients with less common molecular, histological, or clinical characteristics. Given that the AI test was developed using data from non-metastatic breast cancer patients without additional exclusion criteria, we can evaluate its prognostic accuracy in various subgroups. This includes patients with varying age and menopausal status, nodal status, tumor size, estrogen and HER2 receptor status, race, and administered adjuvant therapy (Fig. 3). Furthermore, our analysis indicates that the AI test score is associated with known clinical factors that influence cancer prognosis, including higher scores in ER− and PR − patients, as well as patients with more advanced staging (both N and T) (Fig. 4).

Currently, there is no prognostic/predictive tool supported by NCCN guidelines for use in triple-negative and HER2+ breast cancer patients[31]. The AI test reliably predicted DFI in TNBC ($n = 230$, C-index: 0.71 [0.62–0.81], HR: 3.81 [2.35–6.17, $p = 0.02$]) and HER2+ patients ($n = 343$, C-index: 0.67 [0.55–0.80], HR: 2.22 [0.99–5.01, $p = 0.05$]) (Section SI.7, Figs. SI.3 and SI.4).

NCCN guidelines also do not currently recommend any tests for the selection of adjuvant systemic therapy in most premenopausal HR + patients (intermediate-risk node-negative and low- or intermediate-risk 1-3 LN+). The AI test reliably predicted DFI in HR+ premenopausal patients (if the menopausal status is unknown, we consider patients younger than 50 years as premenopausal), with a C-index of 0.66 [0.58–0.73] and an HR of 2.87 [1.81–4.55, $p = 0.003$] ($n = 677$).

Furthermore, genomic assays have been reported to perform poorly in Black patients[32,33]. As shown in Fig. 3, the AI test reliably predicted DFI in Black patients ($n = 299$), with a C-index of 0.72 [0.65–0.79] and an HR 4.24 [3.27–5.50, $p = 0.002$]. Among White patients ($n = 2461$), the AI test achieved a C-index of 0.73 [0.69–0.79] and an HR of 3.44 [2.18–5.42, $p = 0.003$].

Some prognostic genomic assays, including Oncotype DX, have also been shown to be predictive of adjuvant chemotherapy benefit[28,34] in HR+ patients. In this study, we evaluate only the prognostic ability of the AI test. As a feasibility analysis, we examine prognostic performance in several HR+ HER2− subgroups at distinct treatment decision points. Although capturing "residual" high-risk in these subgroups does not demonstrate predictive utility, which remains to be established, this analysis provides a foundation for future research.

Specifically, we found the AI test to be prognostic in HR+ HER2− patients who received either adjuvant endocrine therapy (ET) alone ($n = 1171$, C-index: 0.68 [0.61–0.76], HR: 5.01 [1.61–15.58, $p = 0.03$]) or adjuvant chemoendocrine therapy ($n = 469$, C-index: 0.70 [0.60–0.80], HR: 4.10 [2.09–8.02, $p = 0.007$]). Finally, the AI test was prognostic for late recurrence in HR+ HER2− patients ($n = 1158$, C-index 0.79 [0.71–0.86], HR 8.45 [3.11–22.93, $p < 0.001$]).

We completed a few exploratory analyses, in which we defined high- and low-risk groups using the 80th percentile cutoff from our largest training dataset (NYU). For analyses within HR+ HER2− and the triple-negative breast cancer (TNBC) subgroups, we applied the 80th percentile cutoff from the corresponding subpopulation. We selected this threshold as it yielded statistically significant hazard ratios in our training set. Additionally, this threshold balanced the distribution of high- and low-risk groups across all subgroups, consistently demonstrating worse outcomes for high-risk patients compared to low-risk patients. As demonstrated in Fig. 5, when we apply this cutoff to the Providence cohort, we see a significant difference in outcomes between the low- and high-risk groups. We selected the Providence cohort for this analysis as it contains the most patients and the largest

https://doi.org/10.1038/s41467-026-73088-y

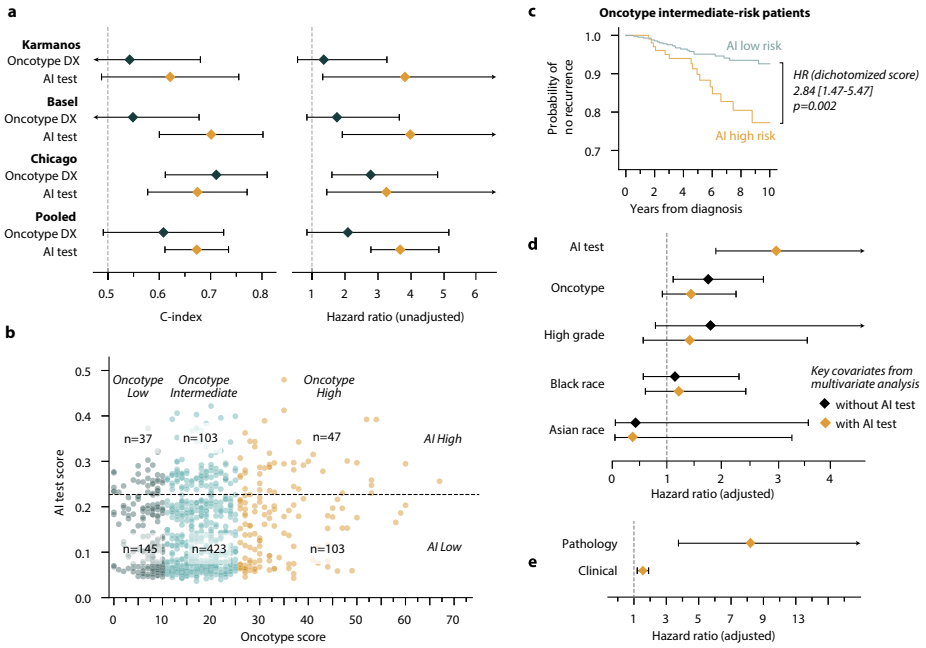

Fig. 2 | The AI test showed improved prognostic accuracy compared with a standard-of-care genomic assay for predicting cancer recurrence. Oncotype DX, a standard-of-care 21-gene assay, classifies patients into low-, intermediate-, and high-risk groups. The AI test demonstrated statistically significant discrimination between high- and low-risk patients without the need to introduce an intermediate-risk group, thereby enhancing clarity in decision-making. We analyze the differences in classification and patient outcomes between the two tests using data from the Karmanos, Basel, and UChicago cohorts. While the Oncotype DX score has also been shown to be associated with chemotherapy benefit in prospective studies, that evidence was based on a prognostic test, making similar predictions to the test presented in this paper. Although this study does not evaluate the ability of our test to predict treatment response directly, the strength of the prognostic signal suggests potential for informing treatment decisions. **a** Comparison of prognostic ability between Oncotype DX and the AI test in univariate models (n = 858 patients). We compare the hazard ratio for every 0.2 increase in our score, compared to a 20-point increase for Oncotype DX. Error bars represent 95% confidence intervals. **b** A scatter plot illustrating risk scores for Oncotype DX-tested patients. Each point represents one patient. The majority of patients with intermediate Oncotype DX scores were reclassified into the low-risk group by the AI test. **c** For intermediate-risk Oncotype DX patients (n = 526 patients), the AI test was able to accurately distinguish between low- and high-risk patients (HR = 2.84, 95% CI = 1.47–5.47, p = 0.002). Hazard ratios were estimated using a Cox proportional hazards model and p-values were computed using a two-sided Wald test. **d** Hazard ratios associated with common clinical variables in a multivariate Cox analysis, with and without including the AI test (n = 858 patients). The AI test is significantly associated with DFI after adjusting for Oncotype DX score, grade (based on the Nottingham grading system, categorized from 1 to 3), and race in a multivariate Cox regression model. Error bars represent 95% confidence intervals. **e** Comparing the AI test's modalities in a multivariate Cox model (n = 858 patients) shows that the pathology score was more informative than the clinical score within the Oncotype DX cohort (see Table SI.4). Error bars represent 95% confidence intervals.

number of observed events among the test cohorts. We acknowledge that the specific cutoff at the 80th percentile is preliminary and we anticipate to optimize it in future publications.

Discussion

In this study, we present a prognostic multi-modal AI test for invasive breast cancer. The overall performance of our test is robust across breast cancer patients. We also compared the AI test to Oncotype DX in three cohorts where the Oncotype DX score is available. The test achieved numerically higher performance metrics than Oncotype DX in two cohorts, Karmanos and Basel, and performed as well as Oncotype DX in one additional cohort, UChicago. Although the confidence intervals for the C-index estimates overlap, this should not be interpreted as evidence of equivalence. The C-index is known to have low statistical power, and prior methodological work has shown that the C-index often understates incremental improvements that are clinically significant[35]. We also acknowledge that not all evaluation datasets contained Oncotype DX scores. This is potentially causing a bias in this comparison.

The presented test is one of the few prognostic tools that can predict long-term outcomes of triple-negative and HER2+ breast cancer patients. This capability, combined with strong performance regardless of nodal status and patient age, supports the viability of the AI test as a single tool to inform treatment decisions in all breast cancer patients.

The proposed AI test is designed to predict cancer recurrence risk, similar to genomic assays such as Oncotype DX and MammaPrint. However, just like these assays, it is not explicitly trained to model treatment effects. We hypothesize that when validated using randomized prospective data, our prognostic test will show predictive capabilities in the same manner that the genomic assays do. That is, high-risk patients will benefit from more aggressive treatment options (such as adjuvant chemotherapy), while low-risk patients will have no benefit. While this study uses observational data and did not evaluate the test's predictive capabilities (i.e., its ability to determine whether a patient is likely to benefit from a specific treatment), we present several potential clinical implications of using such a test in a clinical setting.

The AI test has the potential to refine clinical decision-making in several key areas. In HR+ patients who are candidates for adjuvant chemotherapy, the AI test demonstrated numerically greater accuracy in predicting recurrence than Oncotype DX. This can potentially improve the selection of patients who may benefit from chemotherapy in addition to endocrine therapy. In a multivariate analysis, the AI test was shown to add independent prognostic information with respect to Oncotype DX and cancer grade, suggesting it captures information that is strongly associated with outcomes and independent of these factors. Although these findings are based solely on prognostic associations and do not yet confirm predictive value for treatment benefit, strong risk stratification alone might provide an additional data point

https://doi.org/10.1038/s41467-026-73088-y

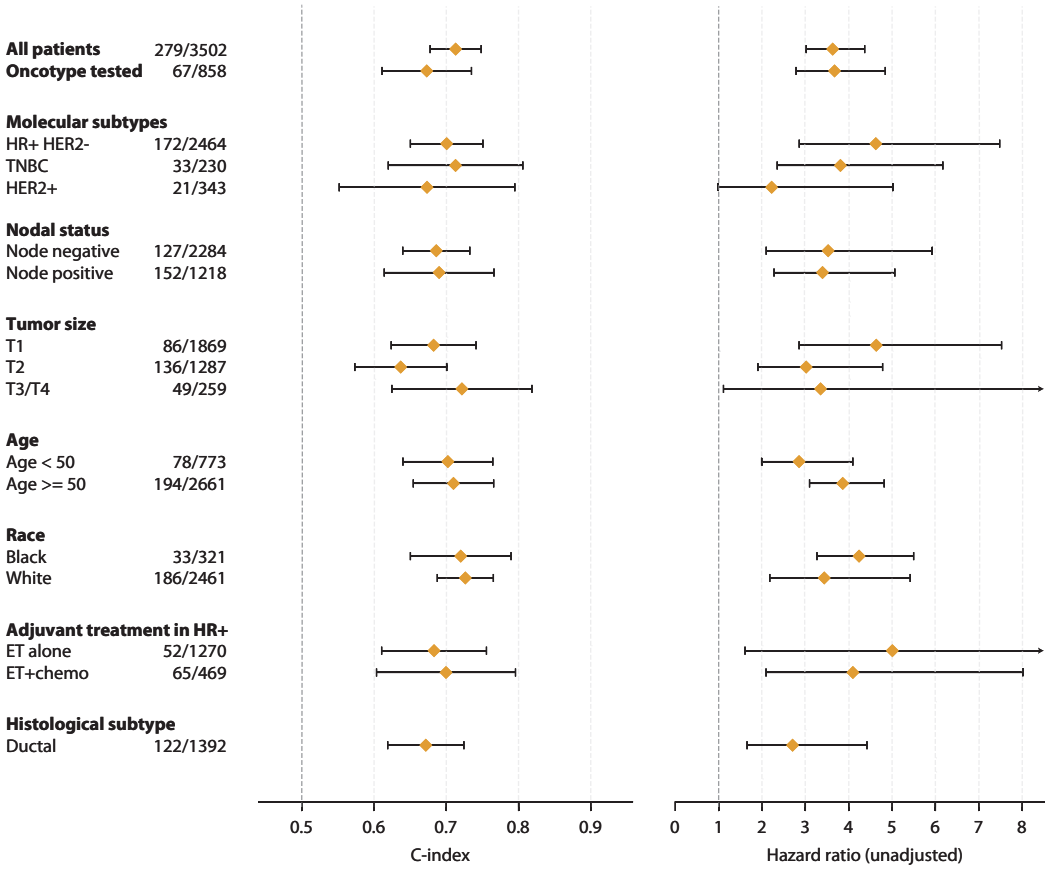

Fig. 3 | The AI test performs well in all major clinically relevant groups. Forest plots display our AI test's performance across clinical, molecular, and demographic groups (n = 3502 patients). Next to each group name, we report the number of disease-free interval events and the total number of patients. Subgroup performances were pooled across all evaluation cohorts using a random effects model. For subgroup performance based on histological subtype, we excluded patients from the Providence cohort because histological subtype data were unavailable. Error bars represent 95% confidence intervals.

in therapy escalation or de-escalation decisions. Importantly, the AI test was prognostic in premenopausal HR+ patients, expanding patient groups who could be evaluated for recurrence risk. We further showed that the AI test is prognostic in HR+ patients who received adjuvant chemoendocrine therapy, and identifies patients with high "residual" recurrence risk. Identifying patients who are still at high risk of recurrence despite chemoendocrine therapy treatment is important in evaluating potential alternative treatment options, such as adjuvant CDK4/6 inhibitors. Finally, the AI test can identify patients with high late recurrence risk, which might be informative when identifying patients who should be considered for extended endocrine therapy.

Besides HR+ patients, the test was prognostic in triple-negative and HER2+ patients, for which there are no NCCN guideline-supported tests for treatment selection. Current standard of care treatment in triple-negative breast cancer involves intense immunochemotherapy (KEYNOTE-522 regimen[36,37]). There are multiple ongoing clinical trials working on treatment de-escalation through avoiding adjuvant pembrolizumab in patients with pathological complete response (OptimICE-pCR[38,39]) or using shorter, anthracycline-free neoadjuvant chemotherapy regimen in patients with immune-enriched TNBC (NeoTRACT)[40].

Our test stands apart from existing genomic tests as it leverages pathology data and employs self-supervised learning. The previous generation of feature extractors for digital pathology relied on pre-specified "pathomic" features[19]. The self-supervised learning paradigm does not require any pre-specification of features; rather, it enables the model to learn the most salient features. Spratt et al. successfully applied this technique to prostate cancer, with their model validated through several randomized clinical trials[41,42]. Similarly, Garberis et al. developed an AI test for breast cancer using contrastive learning, yielding strong results[17], though their study was limited to a single ER+ HER2− cohort. Volinsky-Fremond et al. extended the use of self-supervised deep learning to endometrial tumor slides, outperforming conventional combined pathological and molecular analyses and demonstrating predictive utility for chemotherapy benefit in the PORTEC-3 trial[43]. Our test is developed using Kestrel, an AI foundation model for digital pathology trained using self-supervised learning to extract morphological features from digitized pathology slides. As Kestrel was developed using a pan-cancer dataset, it showed promising results for a wide range of tasks across different clinical indications[21].

Integration of digital pathology-based tests into clinical workflow offers several advantages over genomic and other molecular prognostic and predictive tests. First, as there is no need for complex laboratory equipment and workflows, including sequencing, the turnaround time can be significantly shorter, potentially allowing for more timely treatment decisions. The estimated time required for case accessioning, slide scanning, and inference is less than an hour. Furthermore, both slide scanning and inference can be done for hundreds of cases at a time, allowing for high-volume processing, and the digitization of slides is becoming a routine pathology workflow. This is compared to a turnaround time between 10 and 30 days for breast cancer genomic assays[44,45]. While not all pathology laboratories have digital capabilities yet, the slides can be shipped to centralized laboratories for scanning, which is still faster than molecular processing of the tissue in the lab, which itself takes over a week per case[45]. Second, by utilizing standard H&E-stained slides and not requiring

https://doi.org/10.1038/s41467-026-73088-y

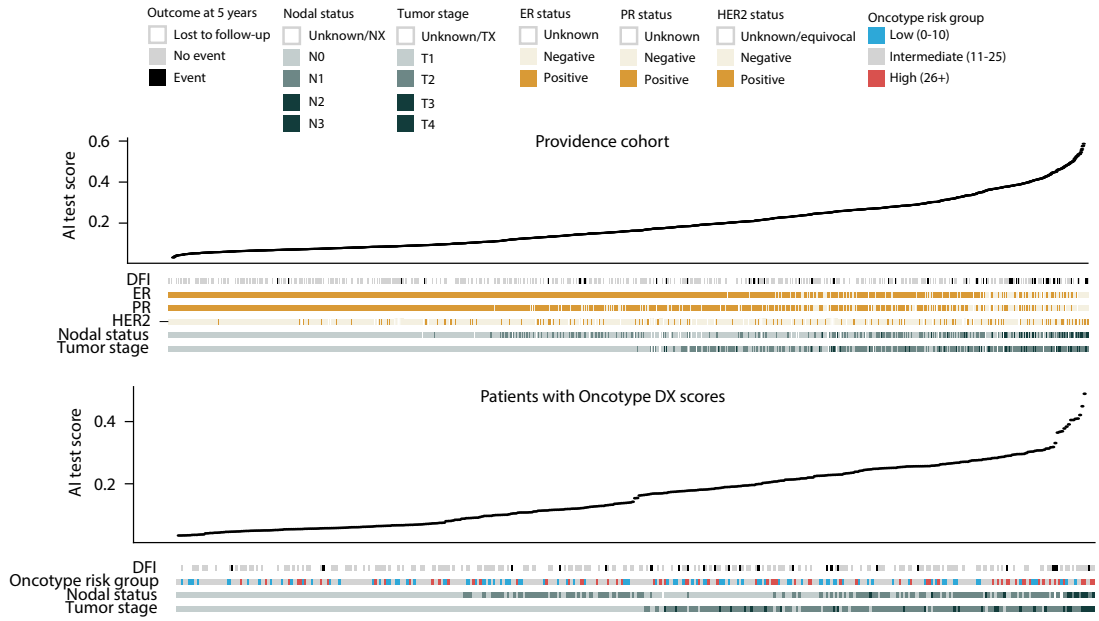

Fig. 4 | Relationship between predicted risk of recurrence and established prognostic factors. The top plot illustrates the relationship between the AI test score and established factors, such as ER/HER2 status and staging in the Providence cohort. The bottom plot combines the Basel, Karmanos, and UChicago cohorts and illustrates the Oncotype DX score in relation to the AI test score. Higher scores are observed in patients with HR− and HER2+ status and in patients with more advanced T or N staging. Although AI test scores are significantly associated with Oncotype DX score (two-sided Pearson correlation: $R^2 = 0.02$, $p < 0.001$), the strength of the association is weak, corresponding to a small Pearson correlation ($r \approx 0.16$). The similarity between the two panels shows that the associations between the AI score and clinical characteristics are consistent across both general invasive breast cancer patients and HR+HER2− patients tested with Oncotype DX.

specialized wet lab tissue processing, the test is more cost-effective to run than genomic assays, which will likely translate to significant savings for patients and for the health system. For example, Medicare has priced the first laboratory procedure codes for digital pathology-based tests at $706[46], compared to approximately $3800 for genomic assays[47]. Third, our test can be performed on core biopsy slides, which allows for earlier risk assessment and might inform neoadjuvant therapy decisions. Finally, this digital approach preserves valuable tumor samples, which might be needed for deep sequencing in cases of disease progression.

Naturally, our work is not without limitations. For example, the histopathology slides used as input might not fully capture the heterogeneity of breast cancer. Other sources of data, such as radiological imaging and gene expression profiles, could provide complementary information. Additionally, we made several methodological design choices, such as the use of multiple-source cross-validation, attention-based MIL, and model ensembling. While the efficacy of these choices is supported in the literature[48–50], comprehensive ablations would be beneficial to help isolate the effects of each on our final model performance. We leave this to future work.

In summary, this study underscores the potential of a digital pathology-based AI test in breast cancer prognosis. By delivering accurate, accessible, and cost-effective risk stratification in diverse patient populations, the test bridges significant gaps in current clinical workflows. With more rigorous validation through clinical trials, it could become an essential tool in personalized cancer care.

Methods

Data

The breakdown of patient demographic, molecular, and clinical characteristics of the training and test datasets is described in Table 1. In total, 3502 patients across five evaluation datasets were included. All patients were female, reflecting patient eligibility for this test. Race was self-reported, with the majority of patients identifying as White (70.27%). TCGA, METABRIC, and BASIS are publicly available datasets. The remaining datasets are private, and ethical approvals were provided as required by contributing institutions. This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and applicable regulatory requirements, including ICH-GCP, relevant US/Canadian human research protection regulations, and TCPS2, as applicable. Informed consent was waived due to minimal risk to subjects, retrospective data collection, and use of only de-identified data.

Task and evaluation

Our primary task was to accurately predict the risk of cancer-related events. To accomplish this, we trained time-to-event models using clinical variables and digital pathology slides. For both modalities, we trained several models and ensembled their predictions, as detailed in Section 3.3.1. We then created a multi-modal model by averaging the predictions of the two ensembles. We trained and evaluated the models using the endpoint definitions in Table SI.1.

Hyperparameter selection. In all instances, we selected the hyperparameters of the models through random search[51], using the training data, the evaluation cohorts were not used in this process. Unless otherwise explicitly stated we maximized the C-index.

In the hyperparameter search, the performance of pathology models was estimated using modified multiple-source cross-validation[48]. That is, we had two collections of datasets $\mathcal{D}^T = \{D_1, D_2, \ldots, D_J\}$ and $\mathcal{D}^V = \{D_{J+1}, D_{J+2}, \ldots, D_{J+K}\}$, where each D in $\mathcal{D}^T$ and in $\mathcal{D}^V$ denotes a separate cohort. Datasets in $\mathcal{D}^T$ were reserved only for training, while $\mathcal{D}^V$ could be used for training or validation. For a given hyperparameter setting, we trained K models, m_1, m_2, ..., m_K. The training set for the k^{th} model was $\mathcal{D}_k^T = \mathcal{D}^T \cup \mathcal{D}^V \backslash D_{J+k}$, and the validation set was $D_{J+k} \in \mathcal{D}^V$. The performances on the validation sets were averaged, weighted by the number of comparisons c_k used in the calculation of the validation C-index using the dataset D_{J+k}. This

https://doi.org/10.1038/s41467-026-73088-y

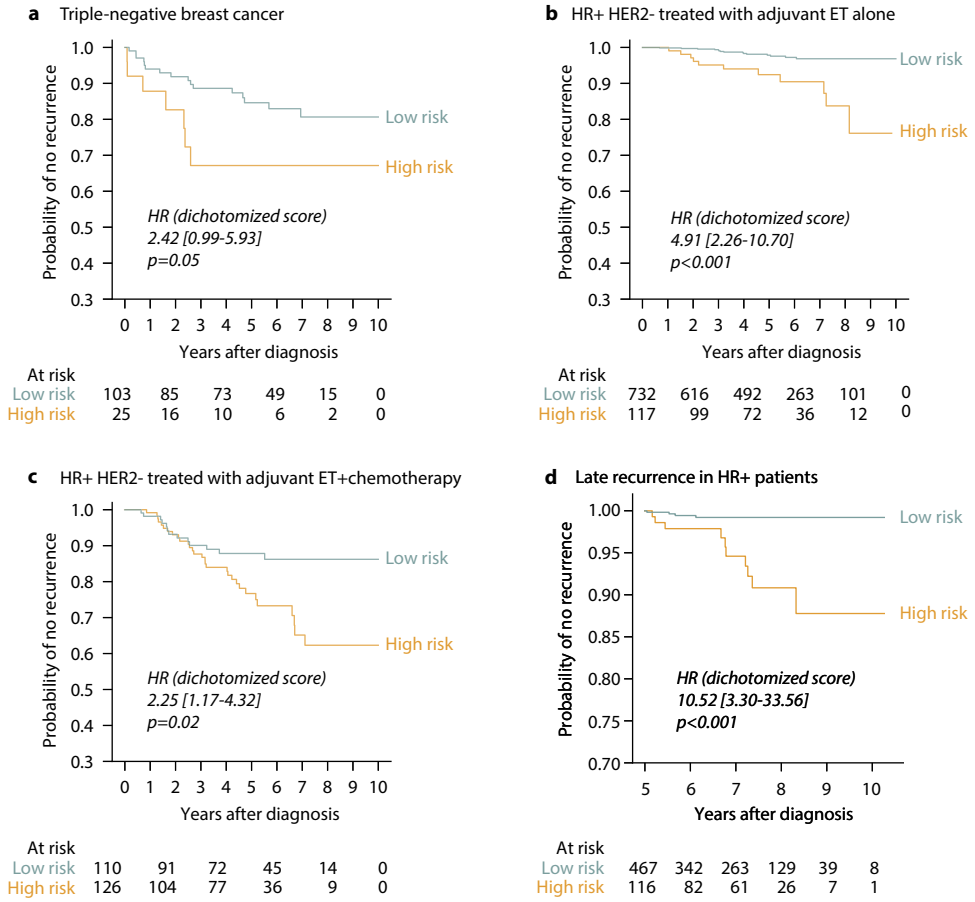

Fig. 5 | The AI test is prognostic in clinically meaningful subgroups. All plots in this figure are for the Providence cohort (the largest evaluation cohort, $n = 1792$ patients). Kaplan-Meier analyses reveal that the AI model effectively stratifies patients by recurrence risk within four distinct clinical subgroups. **a** Risk stratification in the triple-negative breast cancer (TNBC) subgroup ($n = 128$ patients) approaches statistical significance ($p = 0.05$). **b** The AI test stratifies HR+ HER2− patients treated with adjuvant endocrine therapy alone ($n = 849$ patients, $p < 0.001$). **c** Recurrence risk of HR+ HER2− patients treated with with adjuvant ET + chemotherapy ($n = 236$ patients) is stratified using the AI test ($p = 0.02$). **d** Risk stratification of late recurrence in HR+ patients ($n = 583$ patients) with the AI test is significant ($p < 0.001$). Hazard ratios were estimated from Cox proportional hazards models, and p-values were computed using a two-sided Wald test.

procedure was used for performance estimation and hyperparameter selection as formalized in Algorithm 1. To limit the variance in hyperparameter selection, for each set of hyperparameters, we trained and evaluated several models with different random seeds and averaged the validation performances.

Algorithm 1. Multiple-source cross-validation for pathology models

for n = 1 to N **do**
  Sample hyperparameters θ_n
  for k = 1 to K **do**
    Train a model $m_k^{\theta_n}$ using data $\mathcal{D}^T \cup \mathcal{D}^V \backslash D_{J+k}$.
    Evaluate the model $m_k^{\theta_n}$ using D_{J+k} to obtain the validation performance $v_k^{\theta_n}$.
  end for
    The overall validation score of hyperparameters θ_n is $v^{\theta_n} = \frac{1}{K}\sum_{k=1}^{K} c_k v_k^{\theta_n}$.
  The model obtained for θ_n is $m^{\theta_n} = \frac{1}{K}\sum_{k=1}^{K} m_k^{\theta_n}$.
end for
Select the top $L = 10$ hyperparameters, $\theta_1^*, \ldots, \theta_L^*$, yielding the best overall validation scores.
Ensemble models trained with the top $L = 10$ hyperparameters: $m = \frac{1}{L}\sum_{l=1}^{L} m^{\theta_l^*}$.

In our implementation, when training the pathology models, we set $\mathcal{D}^T = \{\text{Korea, Gundersen, Lithuania, BCN, Malaysia, METABRIC, BASIS}\}$ and $\mathcal{D}^V = \{\text{NYU, Omica, Wales}\}$.

As clinical models use much lower dimensional input than pathology models and their performance exhibits less variance, we adopted a simpler modified multi-source cross-validation procedure. We divided the data into K splits, that is, $\mathcal{D}^C = \{D_1^C, D_2^C, \ldots, D_K^C\}$. These splits are designed such that each fold contains data from distinct sources, preserving the multi-source structure. For each iteration, we trained the models using $\mathcal{D}^C \backslash D_k^C$ and selected the hyperparameters that perform the best for this split based on validation performance on D_k^C. This procedure was used for performance evaluation and hyperparameter selection as formalized in Algorithm 2.

Algorithm 2. Multiple-source cross-validation for clinical models

for k = 1 to K **do**
  for n = 1 to N **do**
    Sample hyperparameters θ_n and train a model $m_k^{\theta_n}$ using data $\mathcal{D}^C \backslash D_k^C$.
  end for
  Select the top $L = 10$ hyperparameters, $\theta_1^*, \ldots, \theta_L^*$, yielding the best overall validation scores $v^{\theta_1^*}, \ldots, v^{\theta_L^*}$ for split D_k^C.
  The model obtained for split k is $m^k = \sum_{l=1}^{L} m^{\theta_l^*}$.
end for
Obtain final model $m = \frac{1}{K}\sum_{k=1}^{K} m^k$.

In our implementation, when training the clinical models, we trained using three splits $\mathcal{D}^C = \{D_1^C, D_2^C, D_3^C\}$. We set D_1^C = NYU, D_2^C = METABRIC, and $D_3^C = \bigcup\{$Korea, Gundersen, Lithuania, BCN, Malaysia, BASIS, Wales$\}$. The Omica dataset was excluded from the development of the clinical model due to incomplete clinical data.

Models used to build the AI test

Our AI test makes predictions using both clinical variables and digital pathology images. Internally, it is composed of many models whose predictions are ensembled in two rounds. First, the models are ensembled within their data modality. Then, the ensemble of clinical models is combined with the ensemble of digital pathology models. The intuition behind ensembling lies in the idea that different models make errors that are not perfectly correlated. Ensembling several models cancels out some of these errors, adding accuracy and robustness to predictions. An additional advantage of ensembling at the level of modalities is that for every test example, it is easy to interpret the contribution of the clinical data and the digital pathology data to the final prediction. The process of generating predictions from the AI test is illustrated in Fig. 6.

Creating model ensembles. Ensembling models has become commonplace in machine learning. By averaging predictions, ensembling enables lower variance, more stable estimates, and improved robustness to label noise[50].

We ensembled the clinical models in two steps. First, the output risk scores from all clinical models were scaled approximately to the interval [0, 1]. That is, for each model m and its outputs $\hat{y}$, we found the maximum over the entire validation set, $y_{\max} = \max_i\{\hat{y}_i\}$, and the minimum over the entire validation set, $y_{\min} = \min_i\{\hat{y}_i\}$. For the test examples, the normalized raw prediction y^* from any model was computed as $y = \frac{y^* - y_{\min}}{y_{\max} - y_{\min}}$. Using multiple-source cross validation for clinical models (see Algorithm 2), for each of the K partitions in cross-validation we select the top L = 10 hyperparameters, therefore selecting total of $K \cdot L$ models. For any test example, the aggregated clinical prediction, y^c, was computed as the average of the scaled predictions from all $K \cdot L$ models.

For pathology models, as the predictions were already in the interval [0, 1], we did not apply any normalization. Using multiple-source cross validation for pathology models (see Algorithm 1), we selected the top L = 10 sets of hyperparameters that resulted in the best performance across all K cross validation splits, providing us with $L \cdot K$ models. We generated the final pathology model by averaging these $L \cdot K$ models across different combinations of pooling method (mean, max, or multiple instance learning), training loss function (Cox or discrete-time), and feature extractor to form the final ensemble.

Finally, the clinical score and the pathology score were averaged using equal weights, $y = \frac{1}{2}(y^c + y^p)$, reflecting the absence of a strong a priori belief that either score would be more informative than the other.

Digital pathology models

Unlike many prior approaches [19,52,53], we did not use any hand-crafted pathomic features to build the pathology models. Instead, we obtained learned morphological features from Kestrel, our purpose-built foundation model trained on a pan-cancer dataset. These features were used to train time-to-event models. Below we present our model in greater detail.

Kestrel's architecture. Kestrel is a ViT-L which contains 303 million parameters. An overview of the vision transformer architecture is provided in Fig. 7.

Preprocessing histopathology slides. Digital pathology slides are extremely high resolution, often exceeding 1 billion pixels per image.

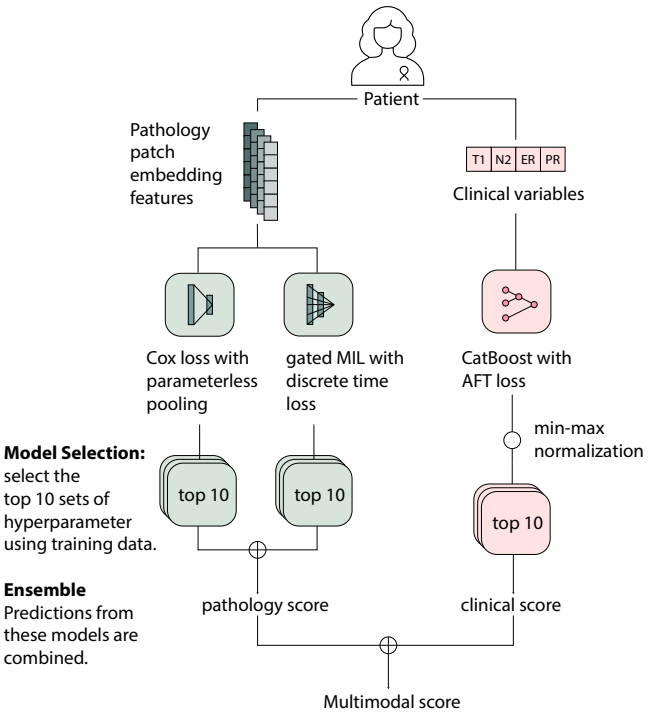

Fig. 6 | The process of generating predictions in the AI test. Two types of models were trained with pathology data: Cox proportional hazards models with parameterless pooling and discrete time models with gated attention MIL. Clinical variables were used in CatBoost models with an AFT loss. The models trained with the top 10 hyperparameter sets from each pipeline were ensembled. Then, the two pathology ensembles were ensembled, and finally the pathology ensemble was ensembled with the clinical ensemble to produce the final multimodal AI test score.

Additionally, large regions of these images are empty. Therefore, to extract only meaningful tissue, we used Otsu's method [54] to distinguish background from foreground. Subsequently, we patchify the foreground into non-overlapping 256 × 256 patches at 20 × magnification (0.5 microns per pixel). Thus, each slide is reduced to a set of patches. There are typically between 8000 to 16,000 patches per slide.

Training Kestrel. Kestrel[21] was trained using a self-supervised learning method DINOv2[13] on 400 million patches extracted from 45,000 histopathology slides. An overview of DINOv2 is provided in Fig. 8. We tuned the hyperparameters for training Kestrel by evaluating trained models on a suite of 16 benchmarks. These benchmarks come mostly from public challenges and are designed to test a model's ability to learn to recognize different characteristics of the tumor microenvironment, for example, tumor cellularity and histology. This has enabled us to train models that excel at generating meaningful features from tissue patches. An analysis of the representations produced by Kestrel can be found in Fig. SI.7B.

Extracting digital pathology features. Slides were preprocessed to extract only foreground patches. These foreground patches were passed through our foundation model to obtain patch-level embeddings. These patch-level embeddings are used as inputs to time-to-event models. During the training of time-to-event models, the gradient was not backpropagated through the foundation model.

Our model's predictions were generated using H&E-stained slides from the same tissue block that was previously used for Oncotype DX testing. For an overwhelming majority of cases (approximately 93% of the patients across all evaluation cohorts), we used only a single slide.

https://doi.org/10.1038/s41467-026-73088-y

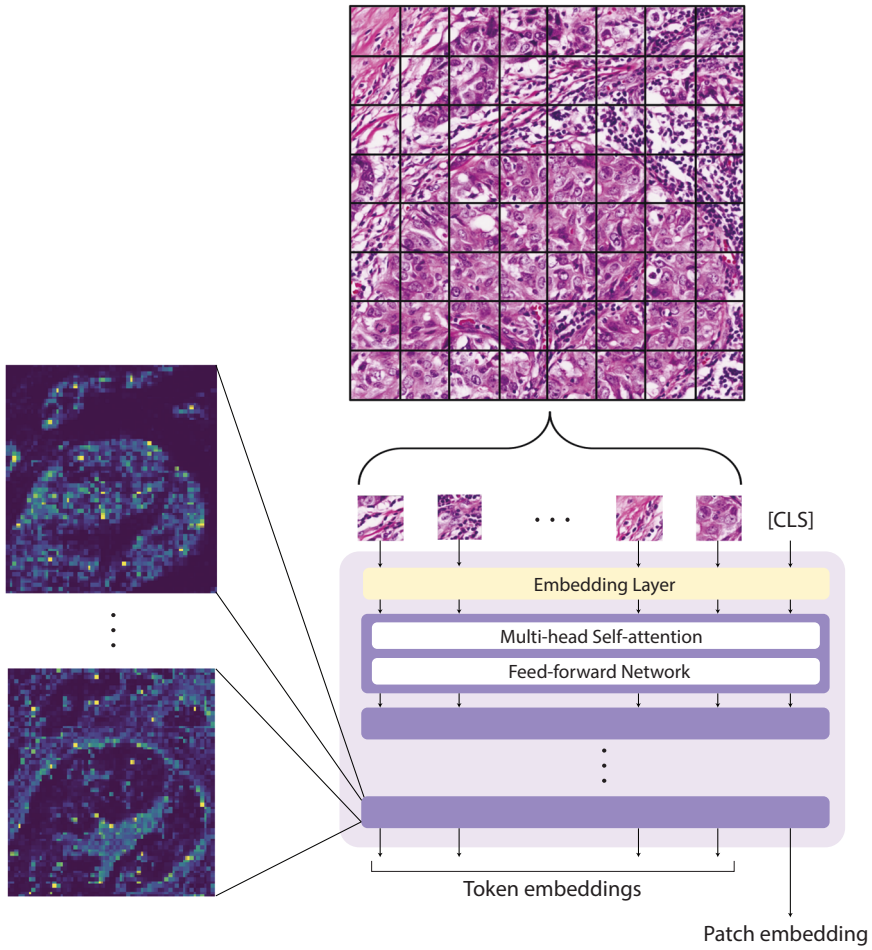

Fig. 7 | Overview of the vision transformer (ViT) architecture applied to digital pathology. Images are first broken into a grid of sub-patches, or "tokens". These tokens, along with a CLS token which serves to aggregate features across sub-patches, are vectorized via an embedding layer. These vectors then progress through a sequence of blocks which consist of multi-headed self-attention followed by a feed-forward network. The ViT output consists of token embeddings and an aggregate patch embedding. The patch embedding is the primary output used downstream, and both the patch and token embeddings are used for self-supervised learning. On the left, we display attention masks between the CLS embedding and token embeddings extracted from the final block. These masks provide insight into the types of features the vision transformer is using to create the patch embeddings.

For the remaining minority of cases for which we had more than one slide, we averaged the predictions across slides.

Training the digital pathology models. We trained two types of models using features extracted by Kestrel as input: Cox proportional hazard models and discrete-time models.

Cox proportional hazard model. Consider a model, m, parameterized by β, with the training dataset D_T and the validation dataset D_V. The training loss for this model is computed as

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{Cox}} + \alpha\mathcal{R}(\beta), \tag{1}$$

where $\mathcal{L}_{\text{Cox}}$ is the negative log of the partial likelihood computed for the training data, $\mathcal{R}$ is a regularizer, and α is a hyperparameter controlling the relative weight of $\mathcal{L}_{\text{Cox}}$ and $\mathcal{R}$.

Specifically, the training loss $\mathcal{L}_{\text{Cox}}$ is computed as

$$\mathcal{L}_{\text{Cox}} = \frac{1}{|D_T|}\sum_{i\in D_T}\delta_i \log\left(\sum_{j\in\mathcal{S}_i}\exp[g(x_j)-g(x_i)]\right), \tag{2}$$

where x_i represents image features for a whole slide image, g is a neural network transforming x_i, δ_i indicates whether the patient experienced an event, and $\mathcal{S}_i$ represents the risk set (the individuals that have not yet experienced an event at time t_i).

The regularizer $\mathcal{R}$ is computed as

$$\mathcal{R}(\beta) = \frac{(1-\gamma)}{2}||\beta||_2^2 + \gamma||\beta||_1, \tag{3}$$

where γ is a hyperparameter controlling the relative weight of the L1 and L2 regularization terms. The entire training loss is optimized with Adam[55], a variant of stochastic gradient descent.

Discrete-time model. We divide times from t_0 to t_J into J contiguous time intervals $(t_0, t_1], (t_1, t_2], \ldots, (t_{J-1}, t_J]$, where $t_0 = 0$ and $t_J = \infty$.

For subject i with covariates x_i (here, features extracted from a whole slide image), the hazard in interval $A_j = (t_{j-1}, t_j]$ is computed as

$$\lambda_{ij}(x_i) = \Pr(T_i \in A_j | T_i > t_{j-1}, x_i), \tag{4}$$

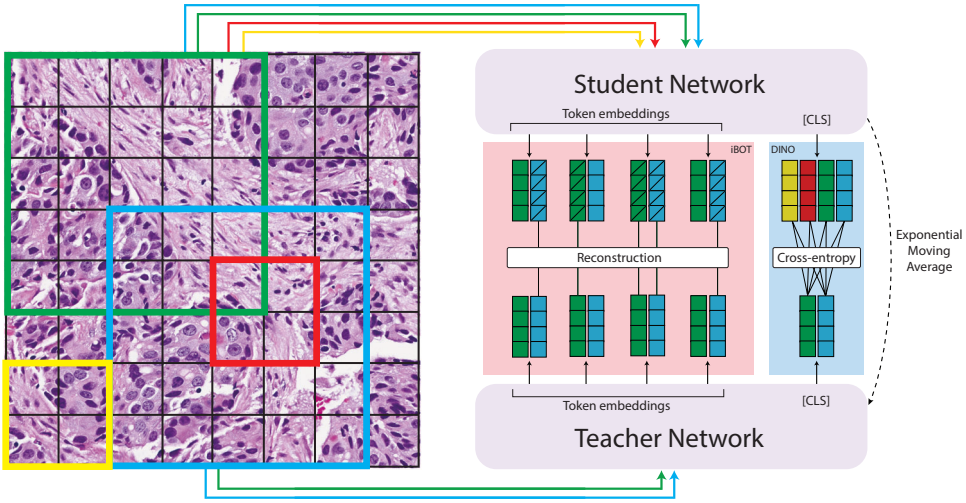

Fig. 8 | Overview of the self-supervised learning method DINOv2 applied to digital pathology. First, several larger (global) and smaller (local) crops are sampled from an image and transformed with image augmentations such as blurring, rotation, scaling, and stain color jitter [62]. Global and local crops are passed through the student network, while only global crops are passed through the teacher network. For the DINO loss component, cross-entropy is computed between the student's CLS token outputs and the teacher's CLS token outputs. For the iBOT loss component, input tokens are randomly masked before passing global crops through the student network, and a reconstruction loss is computed between the corresponding student network token embeddings and the corresponding teacher network token embeddings.

and survival probability is computed as

$$S_i(t|x_i) = \Pr(T_i > t|x_i) = \prod_{j:t_j \le t} (1 - \lambda_{ij}(x_i)). \tag{5}$$

We introduce an indicator $d_{ij} = \mathbb{1}[T_i \in A_{ij}] = \mathbb{1}[t_{j-1} < T_i \le t_j]$, which for censored subjects is given by $(d_{i1}, \ldots, d_{ij_i}) = (0, \ldots, 0)$ and for subjects that experience the event is given by $(d_{i1}, \ldots, d_{ij_i}) = (0, \ldots, 0, 1)$. The likelihood can then be defined as

$$\mathcal{L}_{\mathrm{DT}} = \prod_{i=1}^{n} \lambda_{ij_i}(x_i)^{d_{ij_i}} \prod_{j=1}^{j_i-1} (1 - \lambda_{ij}(x_i))^{1-d_{ij}}, \tag{6}$$

and the log-likelihood is

$$\log \mathcal{L}_{\mathrm{DT}} = \sum_{i=1}^{n} \left[d_{ij_i} \log \lambda_{ij_i}(x_i) + \sum_{j=1}^{j_i} (1 - d_{ij}) \log(1 - \lambda_{ij}(x_i)) \right]. \tag{7}$$

Then, we can model the score in the j^{th} time interval using a neural network as $\phi_j(x_i)$. To put the hazard in the [0, 1] interval, we set

$$\lambda_{ij}(x_i) = \frac{1}{1 + \exp[-\phi_j(x_i)]}. \tag{8}$$

We can estimate $\phi_j(x_i)$ by minimizing the negative log-likelihood. Regularization is added to $\mathcal{L}_{\mathrm{DT}}$ in the same way it is added to $\mathcal{L}_{\mathrm{Cox}}$ described in the previous section.

Aggregation of patch embeddings. In order to train time-to-event neural networks with Cox or discrete-time losses, it is necessary to aggregate patch embeddings into a single vector representation that can be passed into a standard feed-forward neural network. We used two strategies to do this: parameterless pooling and attention-weighted multiple instance learning (MIL) pooling.

Parameterless pooling. It has been shown that simple forms of pooling, such as mean- and max-pooling, perform well on digital pathology tasks where aggregation is needed. Thus, one set of the trained time-to-event models in our pathology ensemble consists of networks trained on mean- or max-pooled patch embeddings with a Cox loss.

Attention-weighted MIL. It has been shown that simple attention-based networks, such as gated attention[56], are also effective ways of aggregating patch embeddings in digital pathology[49]. Thus, we trained a complementary set of time-to-event models using a gated attention network with a discrete-time loss. In particular, for a single slide, the MIL network combines K patch embeddings $[\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{R}^{K \times D}$ into a single embedding $\mathbf{z} \in \mathbb{R}^D$ according to

$$\mathbf{z} = \sum_{k=1}^{K} a_k \mathbf{h}_k, \tag{9}$$

where

$$a_k = \frac{\exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_k^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_k^\top))\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_j^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_j^\top))\}}, \tag{10}$$

and $\mathbf{w} \in \mathbb{R}^{L \times 1}$, $\mathbf{V} \in \mathbb{R}^{L \times D}$, and $\mathbf{U} \in \mathbb{R}^{L \times D}$ are learnable parameters, $\odot$ denotes element-wise multiplication, and sigm($\cdot$) is the sigmoid function. A visualization of regions assigned close attention weight can be found in Section SI.10 in Fig. SI.7A.

Clinical models

For building a model for clinical variables, we utilized CatBoost[57], a gradient boosting decision tree ensemble. Its primary advantage over other possible models for this application, e.g., neural networks, is how it handles categorical variables, which are inevitable in clinical data. Here, CatBoost is specifically trained with the AFT (accelerated failure time) loss[58]. We utilized AFT models with three possible distributions (selected in hyperparameter search): normal, logistic, and extreme value.

We used eight routinely collected variables (see Table 2) that characterize breast cancer and are typically used to guide treatment decisions. For instance, ER and PR statuses help determine tumor hormone receptor sensitivity, which is critical for hormone-based therapies. HER2 status is important for targeted therapies such as trastuzumab, which is effective in HER2+ cancers.

Table 2 | Possible values for the clinical variables used as inputs for our clinical model

Clinical variable	Possible values
Patient age	[0, 100]
Estrogen receptor (ER) status	positive, negative, unknown
Progesterone receptor (PR) status	positive, negative, unknown
HER2 biomarker status	positive, negative, equivocal, unknown
T stage (tumor size/extent)	T1mi, T1a, T1b, T1c, T2, T3, T4, TX, unknown
N stage (nodal involvement)	N0, N1, N2, N3, NX, unknown
Has IDC component	yes, no, unknown
Has ILC component	yes, no, unknown

Statistics and reproducibility

Performance metrics. The two primary metrics we use are the concordance index (C-index) and hazard ratio (HR). Apart from computing these metrics on each individual dataset, we use a random effects model to obtain pooled metrics which summarize performance over multiple datasets, as explained in Section SI.3. In the following paragraphs, we justify the use of these metrics in our application.

C-index. The C-index measures how well the predicted risk ranking of patients aligns with the actual order in which they experienced events or are censored[59]. A C-index of 1.0 indicates that patients are perfectly ordered according to predicted risk, whereas a C-index of 0.5 indicates that the model's ordering is no better than random. The advantage of the C-index is that it is easy to interpret, as it is analogous to AUROC in how it is computed, while accommodating censoring. A detailed explanation of how the C-index is computed is in Section SI.1.

Hazard ratio of a dichotomized score. For a dichotomized score, patients are stratified into two categories: high-risk and low-risk. The hazard ratio is computed using a Cox proportional hazard model with the dichotomized score as the only covariate, with the low-risk group as the reference. The p-values are computed using the Wald test.

Hazard ratio of a continuous score. For a continuous score, the hazard ratio represents the relative increase or decrease in the hazard of the event associated with a one-unit increase in that score. In our analysis, we compare a 0.2 unit increase in our AI test (which ranges from 0 to 1) to a 20 unit increase in the Oncotype DX score (which ranges from 0 to 100). A detailed explanation of how the hazard ratio is computed is in Section SI.2. When available, it is often preferable to report hazard ratios for continuous scores, as dichotomization discards substantial information and decreases the power to detect relationships between the score and patient outcome[60].

Multivariate Cox proportional hazards model. To assess the simultaneous impact of multiple variables on the outcome, we fit a multivariate Cox proportional hazards model with all variables of interest. For categorical variables in this analysis, we assign a binary indicator to all categories except one, which serves as a reference. This analysis allows us to compute the hazard ratio associated with specific covariates, for example, risk scores, after adjusting for additional factors. For instance, multivariate Cox analysis allows us to estimate the hazard ratio after adjusting for the varying patient outcome distributions in different datasets. We have a separate indicator variable for each dataset other than Basel, which is used as the reference group. All multivariate Cox analyses performed in this paper adjusted for the differences between the datasets involved.

In order to evaluate the effect of adding new covariates to an existing Cox model, we use the likelihood ratio test (LRT). LRT is used to compare the goodness of fit between two Cox models, where the covariates of one model is a subset of the covariates of the other.

Recurrence rate as a function of the AI test score. To assess how the AI test score corresponds to patient recurrence, we divided patients based on their AI test scores into groups using quartiles. We then used the Kaplan-Meier estimator to compute 10-year recurrence rates (with confidence intervals) across these groups of patients. The aim was to compare the average score within each group to the estimated recurrence rate within that group. This approach helped us evaluate how well the AI test score corresponded to the observed risk of recurrence, providing a clear way to assess whether higher test scores consistently correspond to higher recurrence rates. The results are illustrated in Fig. 1e.

Study design and statistical analysis. This was an observational, non-randomized retrospective study. The investigators were not blinded during experiments or data analysis. Sample sizes were determined by cohort availability and predefined data splits, and no statistical method was used to predetermine sample size. Subgroup analyses were pre-specified based on clinical relevance. All statistical tests were two-sided.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

TCGA-BRCA data can be accessed from The Cancer Imaging Archive under doi:10.7937/K9/TCIA.2016.AB2NAZRP. The BASIS dataset can be requested from the Wellcome Sanger Institute and was accessed from the European Genome-Phenome Archive (EGA) under accession number EGAS00001001178. The METABRIC dataset can be requested from the METABRIC Committee and was accessed from EGA under accession number EGAD00010000270. The Providence dataset can be accessed from Nightingale via doi:/10.48815/N5159B. The remaining datasets were obtained from third-party collaborators and are subject to institutional, ethical, and data governance restrictions. These datasets are not publicly available, and the authors do not have the authority to grant access. Researchers seeking access to these data should contact the respective data-owning institutions. Access requests will be subject to the policies, approvals, and data use agreements of those institutions, including compliance with applicable privacy and ethical regulations. Researchers may contact the corresponding author for general guidance on the appropriate institutional points of contact. Source data are provided with this paper.

Code availability

The full codebase underlying the AI model is proprietary and cannot be publicly released. The model and selected components of the code may be made available to academic researchers for non-commercial research purposes, subject to institutional approval and execution of appropriate data use and/or material transfer agreements. Requests will be evaluated on a case-by-case basis to ensure compliance with data privacy, intellectual property, and commercial considerations. To facilitate reproducibility, we provide detailed methodological descriptions in the Methods section, including model architectures, training procedures, and hyperparameter selection strategies, and we reference all open-source libraries and frameworks used in the implementation. In addition, the code used to generate the main figures and aggregate results presented in this study can be made available upon reasonable request. DINOv2 model architecture for self-supervised training (https://github.com/facebookresearch/dinov2). PyTorch library for training and inference (https://github.com/pytorch/pytorch). For preprocessing whole slide images: OpenSlide (https://openslide.org), wsidicom (https://github.com/imi-bigpicture/

wsidicom). For training and models for survival analysis: CatBoost (https://github.com/catboost/catboost), PyCox (https://github.com/havakv/pycox), Scikit-learn (https://github.com/scikit-learn/scikit-learn), Torch Tuples (https://github.com/havakv/torchtuples). For result analysis: NumPy (https://github.com/numpy/numpy), Metagen (https://cran.r-project.org/web/packages/meta/meta.pdf), Lifelines (https://github.com/CamDavidsonPilon/lifelines), Concordance (https://cran.r-project.org/web/packages/survival/vignettes/concordance.pdf), timeROC (https://cran.r-project.org/web/packages/timeROC/index.html), dcurves (https://pypi.org/project/dcurves/). For visualizations: Seaborn (https://github.com/mwaskom/seaborn), Matplotlib (https://github.com/matplotlib/matplotlib).

References

1. Paik, S. et al. A multigene assay to predict recurrence Of Tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
2. Mook, S., Van't Veer, L. J., Rutgers, E. J., Piccart-Gebhart, M. J. & Cardoso, F. Individualization of therapy using Mammaprint: from development to the MINDACT Trial. *Cancer Genomics Proteom.* **4**, 147–155 (2007).
3. Wallden, B. et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genom.* **8**, 1–14 (2015).
4. Sparano, J. A. et al. Clinical outcomes in early breast cancer with a high 21-gene recurrence score of 26 to 100 assigned to adjuvant chemotherapy plus endocrine therapy: a secondary analysis of the TAILORx randomized clinical trial. *JAMA Oncol.* **6**, 367–374 (2020).
5. Esserman, L. J. et al. Use of molecular tools to identify patients with indolent breast cancers with ultralow risk over 2 decades. *JAMA Oncol.* **3**, 1503–1510 (2017).
6. Pusztai, L. et al. Development and validation of RSClin N+ tool for hormone receptor-positive (HR+), HER2-negative (HER2-), node-positive breast cancer. *J. Clin. Oncol.* **42**, 508 (2024).
7. Miglietta, F. et al. 242MO Association of tumor-infiltrating lymphocytes (TILs) with recurrence score (RS) in patients with hormone receptor-positive (HR+)/HER2-negative (HER2-) early breast cancer (BC): A translational analysis of four prospective multicentric studies. *Ann. Oncol.* **34**, S280 (2023).
8. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 1597–1607 (2020).
9. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *CVPR* (2020).
10. Grill, J.B. et al. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, **33**, 21271–21284 (2020).
11. He, K. et al. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 16000–16009 (2022).
12. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proc.IEEE/CVF International Conference on computer vision*, 965–9660 (2021).
13. Oquab, M. et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024).
14. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924–2935 (2024).
15. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
16. Wang, X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 1–9 (2024).
17. Garberis, I. et al. Deep learning allows assessment of risk of metastatic relapse from invasive breast cancer histological slides. *Nat. Commun.* **16**, 5876 (2025).
18. Sharma, A. et al. Validation of an AI-based solution for breast cancer risk stratification using routine digital histopathology images. *Breast Cancer Res.* **26**, 123 (2024).
19. Fernandez, G. et al. Development and validation of an AI-enabled digital breast cancer assay to predict early-stage breast cancer recurrence within 6 years. *Breast Cancer Res.* **24**, 1–11 (2022).
20. Boehm, K. M. et al. Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. *Nat. Commun.* **16**, 2106 (2025).
21. Cappadona, J. et al. Squeezing performance from pathology foundation models with chained hyperparameter searches. In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice* (2024).
22. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations* https://openreview.net/forum?id=YicbFdNTTy (2021).
23. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
24. Ciriello, G. et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015).
25. Curtis, C. et al. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
26. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
27. Bifulco, C. et al. Identifying high-risk breast cancer using digital pathology images: a Nightingale Open Science dataset. Nightingale Open Science https://doi.org/10.48815/N5159B (2021).
28. Albain, K. 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.* **11**, 55–65 (2010).
29. Sparano, J. A. et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N. Engl. J. Med.* **379**, 111–121 (2018).
30. Kalinsky, K. et al. 21-gene assay to inform chemotherapy benefit in node-positive breast cancer. *N. Engl. J. Med.* **385**, 2336–2347 (2021).
31. Waks, A. et al. Assessment of the HER2DX assay in patients with ERBB2-positive breast cancer treated with neoadjuvant Paclitaxel, Trastuzumab, and Pertuzumab. *JAMA Oncol.* **9**, 835–840 (2023).
32. Flanagin, A., Frey, T. & Christiansen, S. L. Updated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA* **326**, 621–627 (2021).
33. Satpathy, Y. et al. Comparison of capture rates of the national cancer database across race and ethnicity. *JAMA Netw. Open* **6**, e2350237 (2023).
34. Paik, S. et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. *J. Clin. Oncol.* **24**, 3726–3734 (2006).
35. Hartman, N., Kim, S., He, K. & Kalbfleisch, J. D. Pitfalls of the concordance index for survival outcomes. *Stat. Med.* **42**, 2179–2190 (2023).
36. Takahashi, M. et al. Pembrolizumab Plus Chemotherapy Followed by Pembrolizumab in patients with early triple-negative breast cancer: a secondary analysis of a randomized clinical trial. *JAMA Netw. Open* **6**, e2342107 (2023).
37. Schmid, P. et al. Event-free survival with Pembrolizumab in early triple-negative breast cancer. *N. Engl. J. Med.* **386**, 556–567 (2022).
38. Santa-Maria, C. A. Optimizing and refining immunotherapy in breast cancer. *JCO Oncol. Pract.* **19**, 190–191 (2023).
39. Tolaney, S. et al. OptimICE-pCR: De-escalation of therapy in early-stage TNBC patients who achieve pCR after neoadjuvant chemotherapy with checkpoint inhibitor therapy (Alliance A012103)

[abstract]. In *Proceedings of the 2023 San Antonio Breast Cancer Symposium*, vol. 84 (AACR, 2024).
40. Stecklein, S. R. et al. NeoTRACT: Phase II trial of neoadjuvant tumor-infiltrating lymphocyte- and response-adapted chemoimmunotherapy for triple-negative breast cancer (TNBC). *J. Clin. Oncol.* **42**, http://www.clinicaltrials.gov/ct2/show/NCT05645380 (2024).
41. Spratt, D. E. et al. Artificial intelligence predictive model for hormone therapy use in prostate cancer. *NEJM Evid.* **2**, EVIDoa2300023 (2023).
42. Spratt, D. E. et al. An AI-derived digital pathology-based biomarker to predict the benefit of androgen deprivation therapy in localized prostate cancer with validation in NRG/RTOG 9408. *J. Clin. Oncol.* **40**, 223 (2022).
43. Volinsky-Fremond, S. et al. Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nat. Med.* **30**, 1962–1973 (2024).
44. Podany, E. L., Goldberg, P., Ma, C. X. & Davis, A. A. Improving the OncotypeDX ordering process in patients with ER+ HER2-early-stage breast cancer: A longitudinal QI project. *JCO Oncol Pract*. **20**, 292 (2024).
45. Hartzell, M. et al. Improving oncotype dx turnaround time for timeliness of adjuvant therapy: A single-system, multi-site quality improvement study in early-stage hr-positive breast cancer. *JCO Oncol Pract*. **21**, 464 (2025).
46. Freedman, B. Q. Cms posts final prices for new lab tests (November 2023). https://www.discoveriesinhealthpolicy.com/2023/11/cms-posts-final-prices-for-new-lab.html (2023).
47. Berdunov, V. et al. Cost-effectiveness analysis of the Oncotype Dx Breast Recurrence Score® test from a US societal perspective. *ClinicoEconomics Outcomes Res.* **16**, 471–482 (2024).
48. Geras, K. & Sutton, C. Multiple-source cross-validation. In *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 of *Proceedings of Machine Learning Research* (2013).
49. Shao, D. et al. Do multiple instance learning models transfer? In *the Forty-second International Conference on Machine Learning* https://openreview.net/forum?id=hfLqdquVt3 (2025).
50. Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, 1-15 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2000).
51. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
52. Amgad, M. et al. A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer. *Nat. Med.* **30**, 85–97 (2024).
53. Nimgaonkar, V. et al. Development of an artificial intelligence-derived histologic signature associated with adjuvant gemcitabine treatment outcomes in pancreatic cancer. *Cell Rep. Med.* **4**, 101013 (2023).
54. Otsu, N. A Threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybernet.* **9**, 62–66 (1979).
55. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. In *ICLR* (2015).
56. Ilse, M., Tomczak, J. & Welling, M. Attention-based Deep Multiple Instance Learning. In *International Conference on Machine Learning*, 2127–2136 (PMLR, 2018).
57. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: unbiased boosting with categorical features. In *Proc. Advances in Neural Information Processing Systems*. **31**, 6639–6649 (2018).
58. Wei, L.J. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat. Med.* **11**, 1871–1879 (1992).
59. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).
60. Altman, D. G. & Royston, P. The cost of dichotomising continuous variables. *Bmj* **332**, 1080 (2006).
61. Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50**, 163–170 (1966).
62. Shen, Y., Luo, Y., Shen, D. & Ke, J. RandStainNA: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In *MICCAI* (2022).

Acknowledgements

DZP acknowledges funding from the Research Council of Lithuania (LMTLT), agreement No. S-PD-22-86. For METABRIC cohort: This study makes use of data generated by the Molecular Taxonomy of Breast Cancer International Consortium. Funding for the project was provided by Cancer Research UK and the British Columbia Cancer Agency Branch. For TCGA-BRCA cohort: The results shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. For the Wales cohort: Biosamples were obtained from the Wales Cancer Bank (https://doi.org/10.5334/ojb.46), which is funded by Health and Care Research Wales. Other investigators may have received specimens from the same subjects. For Breast Cancer Now Biobank cohort: the authors wish to acknowledge the roles of the Barts Cancer Now Tissue Bank in collecting and making available the samples and/or data, and the patients who have generously donated their tissues and shared their data to be used in the generation of this publication.

Author contributions

J.W. and K.J.G. contributed to the study conception. C.F.G., L.P., Y.L.C., and K.J.G. contributed as research advisors. N.C., F.S., U.O., V.S., L.M.P., N.K., A.B., and F.J.E. provided clinical guidance. J.W., K.Z., J.C., and K.L. wrote code, developed infrastructure, and trained models throughout the study. J.W., K.Z., J.C., J.E., E.D.C., Y.J.K., F.H., Z.S., N.T., M.S., F.Y., E.L., T.H., B.S., D.R., A.J.C., V.S., M.V., L.S., S.M., D.B., K.C., Y.Z., L.D., A.M.S., W.A., R.B., J.W.P., H.M., S.H.T., V.A., D.Z.P, A.L., B.D.P., C.B., S.Y.J., J.P.Y., S.H.L., and K.J.G. worked on data preparation. J.W., K.Z., J.C., I.O., C.F.G, and K.J.G. performed evaluation and analysis. IO contributed as a statistical advisor. J.W., K.Z., J.C, I.O., C.F.G., and K.J.G. worked on drafting and revising the manuscript. All authors critically reviewed the paper and the results and approved the final version.

Competing interests

J.W., K.Z., J.C., C.F.G., F.S., Z.S., A.B., F.J.E., Y.L.C., and K.J.G. are equity holders of Ataraxis AI. IO served as a consultant for Ataraxis AI. New York University (NYU) maintains financial and intellectual property interests in Ataraxis AI that are pertinent to the research presented in this manuscript. J.W. and K.J.G. are inventors on a US patent application filed corresponding to some of the methodological aspects of this work. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-73088-y.

Correspondence and requests for materials should be addressed to Krzysztof J. Geras.

Peer review information *Nature Communications* thanks Stacey Winham and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

[1]Ataraxis AI, New York, NY, USA. [2]Karmanos Cancer Institute, Detroit, MI, USA. [3]Cancer Center Baselland, Basel, Liestal, Switzerland. [4]NYU Grossman School of Medicine, New York, NY, USA. [5]The Catholic University of Korea, Seoul, South Korea. [6]UChicago Medicine, Chicago, IL, USA. [7]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [8]NYU Center for Data Science, New York, NY, USA. [9]NYU Courant Institute of Mathematical Sciences, New York, NY, USA. [10]The University of Aberdeen, Aberdeen, Scotland. [11]University Hospital Basel, Basel, Liestal, Switzerland. [12]Cancer Research Malaysia, Subang Jaya, Malaysia. [13]Omica.bio, Mexico City, Mexico. [14]Vilnius University, Vilnius, Lithuania. [15]Providence Health, Renton, Washington, US. [16]University of Pittsburgh Medical Center, Pittsburgh, PA, USA. [17]Northwell Health, New York, NY, USA. [18]Yale School of Medicine, New Haven, CT, USA. [19]Meta AI, New York, NY, USA.
✉e-mail: krzysztof.geras@nyulangone.org