nature communications

Article

https://doi.org/10.1038/s41467-025-60824-z

Deep learning assessment of metastatic relapse risk from digitized breast cancer histological slides

Received: 10 October 2024

Accepted: 4 June 2025

Published online: 01 July 2025

Check for updates

I. Garberis[1,9] ✉, V. Gaury[2,9], C. Saillard[2], D. Drubay[3,4], K. Elgui[2], B. Schmauch[2], A. Jaeger[2], L. Herpin[2], J. Linhart[2], M. Sapateiro[5], F. Bernigole[5], A. Kamoun[2], A. Filiot[2], O. Tchita[2], R. Dubois[2], M. Auffret[2], L. Guillou[2], I. Bousaid[6], M. Azoulay[6], J. Lemonnier[7], M. Sefta[2], S. Everhard[7], A. Sarrazin[2], J-F Reboud[2], F. Brulport[2], J. Dachary[2], B. Pistilli[8], S. Delaloge[8], P. Courtiol[2], F. André[1,8], V. Aubert[2] & M. Lacroix-Triki[5]

Accurate risk stratification is critical for guiding treatment decisions in early breast cancer. We present an artificial intelligence (AI)-based tool that analyzes digitized tumor slides to predict 5-year metastasis-free survival (MFS) in patients with estrogen receptor-positive, HER2-negative (ER + /HER2 −) early breast cancer (EBC). Our deep learning model, RlapsRisk BC, independently predicts MFS and provides significant prognostic value beyond traditional clinico-pathological variables (C-index 0.81 vs 0.76, $p < 0.05$). Applying a 5% MFS event probability threshold stratifies patients into low- and high-risk groups. After dichotomization, combining RlapsRisk BC with clinico-pathological factors increases cumulative sensitivity (0.69 vs 0.63) and dynamic specificity (0.80 vs 0.76) compared to clinical factors alone. Expert analysis of high-impact regions identified by the model highlights well-established morphological features, supporting its interpretability and biological relevance.

Despite significant progress in classification and treatment over the past two decades, breast cancer (BC) remains the leading cause of cancer death for women worldwide[1]. Proposing an optimal therapeutic strategy to each patient requires systematic and accurate characterization of each disease. Specifically, estrogen receptor-positive (ER+), HER2-negative (HER2−) invasive BC, which accounts for approximately 70% of all invasive BC, is associated with a wide spectrum of outcomes and treatment requirements. For many of these women, a key question remains whether adjuvant chemotherapy with the burden of acute side effects and the potential long-term persistent quality of life (QoL) deterioration[2] can be safely avoided. Furthermore, women with a predicted high risk of metastatic relapse despite current standard treatment could be offered more intensive or extended adjuvant strategies, including the addition of a CDK4/6 inhibitor[3,4].

Prognosis definition has been traditionally based on clinical and histopathological factors, such as the patient's age and the histological classification and grade. Biomarker assessment, mainly by immunohistochemistry (ER, progesterone receptor (PR), HER2, and the proliferation marker KI67), was added to this estimation and later refined with the inclusion of molecular signatures. Results from the TAILORx trial showed that Oncotype DX®, a 21-gene test that predicts 10-year metastasis-free survival (MFS), could help avoid unnecessary

[1]INSERM U981, Gustave Roussy, Paris-Saclay University, Villejuif, France. [2]Owkin, Paris, France. [3]Gustave Roussy, Office of Biostatistics and Epidemiology, Université Paris-Saclay, Villejuif, France. [4]Inserm, Université Paris-Saclay, CESP U1018, Oncostat, labeled Ligue Contre le Cancer, Villejuif, France. [5]Department of Pathology, Gustave Roussy, Paris-Saclay University, Villejuif, France. [6]Department of Digital Transformation and Information Systems (DTNSI), Gustave Roussy, Villejuif, France. [7]Unicancer R&D, Unicancer, Paris, France. [8]Department of Cancer Medicine, Gustave Roussy, Paris-Saclay University, Villejuif, France. [9]These authors contributed equally: I. Garberis, V. Gaury. ✉e-mail: ingrid-judith.GARBERIS@gustaveroussy.fr

chemotherapy in up to 85% of women with early-stage ER+/HER2− node-negative BC, without affecting outcomes[5–7]. Currently, several gene expression signatures assessed on the primary tumor material are endorsed by international guidelines to support clinicians in refining the prognosis of patients with early BC (EBC) and taking adjuvant treatment decisions[8]. Besides this molecular characterization, prognostic tools using classical factors and embedded into publicly available websites may be used as an aid in clinical decision making. For instance, Predict Breast Cancer, a widely used online prognostication software, uses known prognostic factors such as tumor size, KI67 index, tumor grade, and lymph node status to predict overall survival at 5 and 10 years[9,10]. However, sensitive markers such as Ki67 may be subject to reproducibility and expertise biases[11–15].

Disease prognosis has advanced with the integration of sophisticated computational methods. Artificial intelligence (AI), particularly machine learning (ML), is increasingly used to tackle biological and clinical challenges. Recent studies demonstrate the potential of deep learning (DL) models applied to histopathological whole-slide images (WSI) to predict patient outcomes and identify prognostic features, particularly in BC[16–21].

In this study, we aimed to investigate whether AI applied to tumor WSI could: (i) provide additional prognostic information beyond clinico-pathological prognostic criteria, and (ii) identify patients who have a substantial risk of metastatic relapse despite receiving standard treatments. The ultimate goal was to develop an AI-based digital pathology tool to allow assessment of the risk of metastatic relapse.

Results

The primary objective of this study was to evaluate the additional 5-year MFS prognostic value of RlapsRisk BC score, an AI-derived prognostic score based exclusively on WSI, relative to that of the current clinico-pathological criteria, in patients with ER+/HER2− EBC. The secondary objective consisted of comparing the capacity of a model combining standard clinico-pathological criteria and RlapsRisk BC to dichotomize patients between high risk and low risk of developing 5-year MFS events to that of a model based on clinico-pathological factors only. This comparison was assessed on the entire population of the validation cohort and in different subgroups of clinical interest (histological grade 2, intermediate clinical risk of relapse as defined in the French guidelines displayed in Supplementary Table 4, pre- versus post-menopausal status, with or without lymph node invasion, treated with or without adjuvant chemotherapy).

Model development and datasets

To build our model, we used the GrandTMA cohort as a discovery dataset. This cohort has been collected retrospectively from patients diagnosed in the "One-stop breast clinic" program and treated at Gustave Roussy Cancer Center (Villejuif, France) between October 10, 2005 and February 7, 2013[22]. The cohort comprised 1802 patients diagnosed with early invasive BC (1429 ER+/HER2−, 110 ER+/HER2+, 70 ER−/HER2+, and 193 ER−/HER2− tumors) who underwent surgery as first treatment and had at least one available tumor slide from the surgery specimen. For the ER+/HER2− BC patients, we had for each case one hematoxylin, eosin, and saffron (HES) slide from primary diagnosis and one additional hematoxylin and eosin (H&E) slide.

For the purpose of an external validation of our model, we used a dataset from the French multicenter observational prospective CANTO cohort (NCT01993498)[23]. Out of 14,000 patients accrued in CANTO so far, 1703 ER+/HER2− EBC patients had a minimum follow-up of 5 years and were eligible for the present study. Tumor slides and full clinico-pathological features were exploitable from 1229 patients: 676 patients with HES-stained slides included at Gustave Roussy (hereafter CANTO HES), and 553 patients with H&E-stained slides coming from other UNICANCER centers (hereafter CANTO H&E). After applying the calibration protocol (see Supplementary Methods and Materials), CANTO H&E and CANTO HES were merged into one larger cohort, hereafter the CANTO cohort ($N = 1229$). There was no overlap of patients between the GrandTMA and the CANTO cohorts. See Supplementary Methods for additional information on the two cohorts.

Our approach consisted of developing an algorithm that learned from the WSIs to predict the 5-year MFS without any local annotation provided by pathologists. To build this model, we used a method composed of four steps: (i) tissue tiling, (ii) feature extraction, (iii) creation of a risk score, (iv) binary classification (Fig. 1). All these steps, including model architectures and training methods, are detailed in the Supplementary Methods. The RlapsRisk BC risk score was optimized using the discovery cohort GrandTMA (HES WSIs) with a stratified threefold cross-validation approach, repeated five times. Due to the limited number of events, stratification was performed based on event occurrence to ensure a minimum number of events in each fold.

We then developed a clinical score to predict the 5-year MFS from a multivariable Cox model trained using the discovery dataset (hereafter the Clinical Score), based on the following clinical variables: age, tumor size, histological grade, lymph node invasion, and Ki67 expression. We used this score as the baseline reference and compared

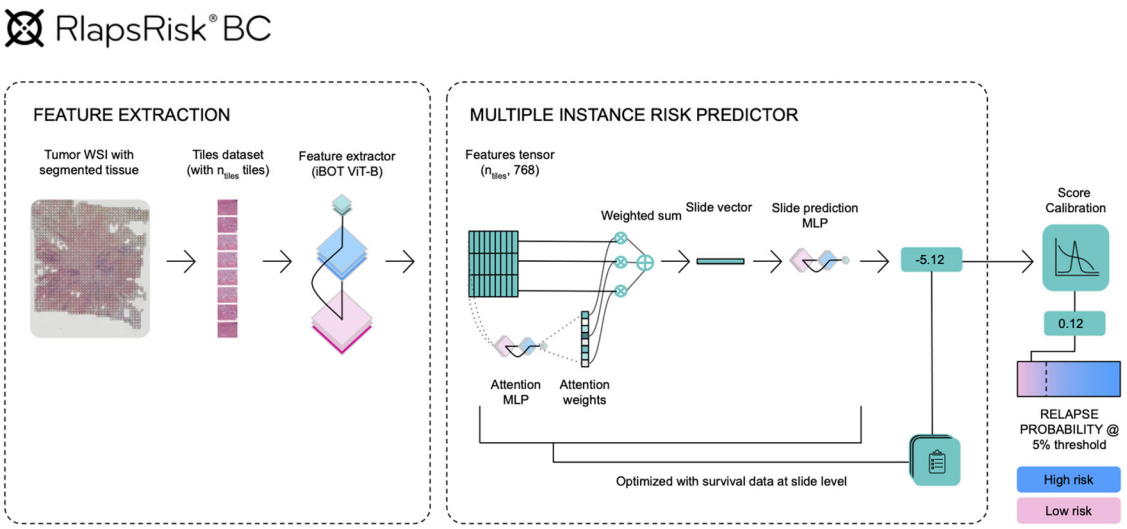

Fig. 1 | RlapsRisk BC algorithm overview. The left panel illustrates the algorithm's processing steps, while the right panel details the training procedure and the overall model architecture designed to predict the risk group.

Table 1 | Multivariable Cox proportional hazard models estimating the contribution of several prognostic variables on MFS on CANTO

CANTO *N* = 1169	Variable	Cox model without RlapsRisk BC		Cox model with RlapsRisk BC	
		Unit HR (95% CI)	*p*	Unit HR (95% CI)	*p*
	Histological grade 1–2	1 (ref)	N.A.	1 (ref)	N.A.
Multivariable Cox model	Histological grade 3	1.79 (0.94–3.41)	0.08	1.32 (0.69–2.54)	0.41
	Age	1.18 (0.90–1.55)	0.24	1.10 (0.84–1.45)	0.49
	Lymph node invasion	1.08 (1.03–1.14)	<0.005	1.08 (1.02–1.13)	<0.005
	Tumor size	1.02 (1.01–1.04)	<0.005	1.02 (1.00–1.03)	0.03
	Ki67 expression	1.28 (1.05–1.55)	0.01	1.23 (1.01–1.50)	0.04
	RlapsRisk BC score	N.A.	N.A.	1.67 (1.31–2.13)	<0.005

p Values correspond to two-sided Wald tests evaluating the null hypothesis that the hazard ratio (HR) for each variable is equal to 1. RlapsRisk BC scores are derived from inferences made on the CANTO external validation WSIs, while the model has been trained exclusively on the discovery cohort, remaining blind to the external validation data.

Table 2 | Performances in the C-index of the Clinical and RR Combined models

Cohort	Subgroup	*N*	C-index		Increment test
			Clinical model	RR Combined	*p* value
External validation dataset CANTO[a]	All population	1169	0.76 (0.69–0.81)	0.81 (0.76–0.86)	<0.05
External validation dataset CANTO[a]	Intermediate clinical risk	706	0.78 (0.70–0.84)	0.86 (0.81–0.91)	<0.05

The Clinical model contains the following variables: clinical variables: age, tumor size, histological grade, lymph node invasion, and Ki67 expression. The RR Combined model adds the RlapsRisk BC score to the clinical variables.

[a]Results were obtained from the C-index computed with the bootstrap resampling external validation procedure. The *p* value corresponds to the comparison of the C-index of the Clinical and RR Combined models using a permutation test.

it to a score derived from a Cox model adjusted for clinical factors and the RlapsRisk BC score (referred to as the RR Combined). The coefficients of this Cox model, which provides the RR Combined score, were also trained on the discovery dataset, combining the Clinical and RlapsRisk BC scores. This comparison allowed us to evaluate the added prognostic value of the RlapsRisk BC score beyond standard clinical factors. We excluded chemotherapy treatment from the clinical score because it was prescribed to higher-risk patients based on clinical best practices, introducing prescription bias. Since this risk was already captured through clinical variables, including chemotherapy would have led to redundancy and multicollinearity in the prognostic models.

RlapsRisk BC score and prognosis

On both the discovery cohort and the external validation cohort CANTO (HES and H&E WSIs), RlapsRisk BC score was an independent prognostic factor for MFS (Table 1 and Supplementary Table 5) (hazard ratio (HR) = 1.67 [1.31–2.13], p value < 0.005) in a multivariable Cox model that was fitted directly on the cohorts of interest, including histological grade, age, lymph node invasion, tumor size, and Ki67 expression. The results per staining for the CANTO cohort are also available in Supplementary Tables 7 and 8, showing notably equivalent performances on H&E and HES. Similarly, RlapsRisk showed strong performances in the intermediate clinical risk group (Supplementary Table 6) (HR = 1.81 [1.29–2.53], p value < 0.005). All variables, except histological grade, were considered continuous. Only patients with ER +/HER2− disease were included.

The discrimination power of the scores was then compared using the Harrell's C-index[24] (Table 2) on the external validation dataset. Data from the CANTO cohort were held out from the training dataset and were used only for external validation and the assessment of the discrimination of each model. The RR Combined model, obtained by integrating RlapsRisk BC with the Clinical Score, significantly outperformed the Clinical Score alone in the external validation dataset, demonstrating that RlapsRisk BC provides additional prognostic information beyond clinical factors. Importantly, RR Combined was entirely fitted on the training cohort before validation, without any recalibration on the external validation dataset. It achieved a Harrell's C-index of 0.81 (confidence interval (CI) = [0.76–0.86]), compared to 0.76 (CI = [0.69–0.81]) for the Clinical Score alone, reflecting an improvement of +0.05 (permutation test p value < 0.05). This added prognostic value of RlapsRisk BC was particularly notable in the intermediate clinical risk group, where RR Combined achieved an improvement of +0.08 (permutation test p value < 0.05).

We further compared the RR Combined model with the PREDICT Breast (PB) model, a widely used prognostic tool that estimates survival outcomes based on clinical and treatment-related factors[25]. Consistent with previous observations, RR Combined achieved superior prognostic performance, yielding the highest C-index of 0.81 (CI = [0.75–0.86]) (Supplementary Table 9). In contrast, PB models showed lower discrimination, with C-index values ranging from 0.70 to 0.75 across their different predicted outcomes.

We then assessed the prognostic performance of RlapsRisk BC score and the clinical factors in patients subgroups defined by standard clinico-pathological factors, by the clinical risk groups (Supplementary Table 4 for descriptions) and by adjuvant treatment regimen (Fig. 2). The assessment of potential heterogeneities in these subgroups was conducted by Cox regression analyses. When a factor was used to build a subgroup, it was removed from the associated Cox multivariable model. A higher RlapsRisk BC score was associated with an increased risk of distant recurrence (HR > 1, corrected p value < 0.05) in the majority of subgroups depicted in Fig. 2. However, for premenopausal patients, RlapsRisk was not associated with an increased risk of relapse. In patients with pN ≥ 2, those with pT ≥ 3, and histological grade 3, the sample sizes were insufficient to obtain significant results.

Clinical validity of RlapsRisk BC

To compare the capacity of the models to dichotomize patients between high risk and low risk of developing 5-year MFS events, thresholds corresponding to a probability of MFS event of 5% at 5 years

https://doi.org/10.1038/s41467-025-60824-z

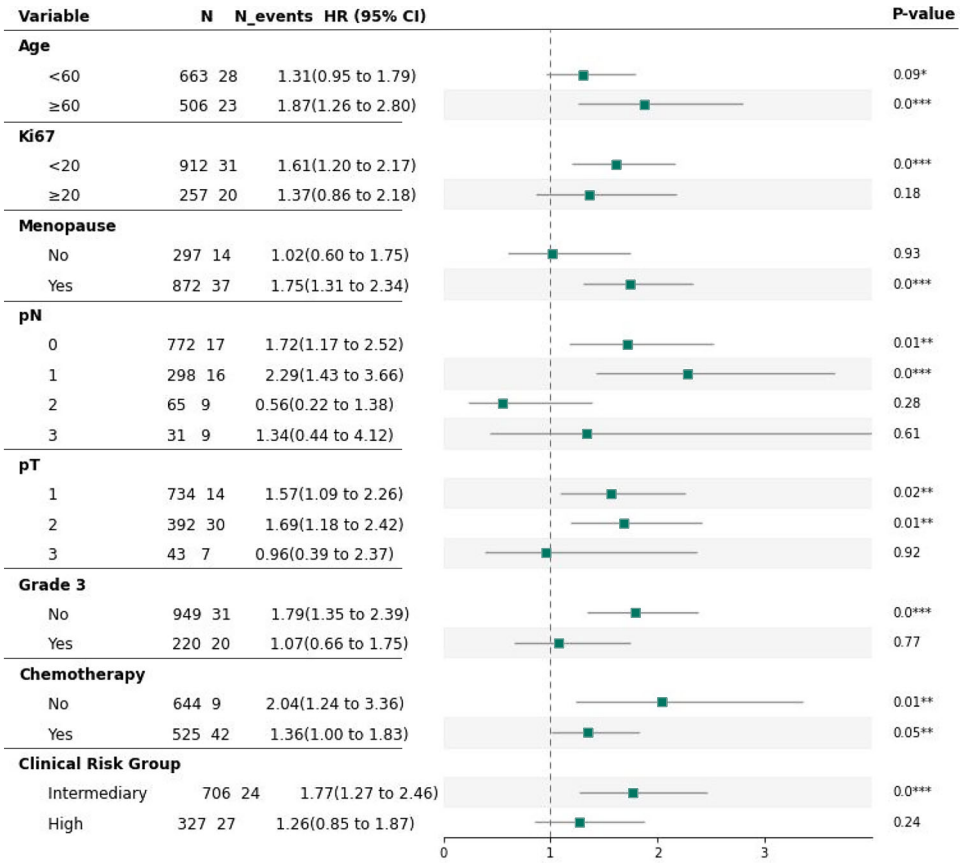

Fig. 2 | Forest plot of the adjusted RlapsRisk BC score HRs on the prediction of 5-year metastasis-free survival on the CANTO cohort. Each square of the forest plot represents the HR of the RlapsRisk BC score (a continuous variable) adjusted for the prognostic clinico-pathological factors in the subgroup of patients defined by the variable category in the first column of the table. The 95% HR confidence intervals, computed using the Wald method, are represented by the horizontal lines. Histological tumor grade 1 and low clinical risk group were removed as no MFS events were recorded in these subgroups. Hazard ratios (HRs) and 95% confidence intervals were estimated using a Cox proportional hazards model. *p* Values correspond to two-sided Wald tests evaluating the null hypothesis HR = 1 and were adjusted for multiple comparisons using the Benjamini–Hochberg procedure. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3 | Classification of patients according to each classifier (RlapsRisk BC, Clinical score, and RR Combined)

	CANTO validation cohort (*N* = 1169)		
	RlapsRisk BC classifier	**Clinical score classifier**	**RR Combined classifier**
Number at low risk[a]	893	879	910
Number at high risk[b]	276	290	259
% of patients with MFS events in the low-risk group	1.56%	1.59%	1.32%
% of patients with MFS events in the high-risk group	8.69%	8.28%	10.04%
Kaplan–Meier's hazard ratio	5.83	5.51	8.01
95% CI	(3.02–11.28)	(2.85–10.66)	(4.04–15.88)
Log-rank *p* value	<0.005	<0.005	<0.005

[a]Predicted probability of 5-year MFS event ≤5%.
[b]Predicted probability of 5-year MFS event >5%.

were set for each risk score in the training set and prespecified accordingly for validation (see "Methods" section for details). In the next section, when the scores are dichotomized, they are referred to as "classifiers" (e.g., RlapsRisk BC, Clinical Score, and RR Combined classifiers). To assess the prognostic power of the RlapsRisk BC classifier, we replicated selected multivariable analyses from the "RlapsRisk BC score and Prognosis" section, using the high/low classification as the input variable instead of the continuous score (Supplementary Table 10).

After applying the MFS risk stratification according to the previously defined classifiers, Kaplan–Meier analyses showed significant differences in distant recurrence events between low- and high-risk patients in the external validation dataset (Table 3 and Fig. 3). Kaplan–Meier analyses were also performed on the intermediate clinical risk subgroup and comparisons between low- and high-risk groups were also significant (Supplementary Table 11 and Fig. 4).

We also assessed the performance of RlapsRisk across subgroups of patients at different risk of recurrence, as defined according to known prognostic factors, such as menopausal status, presence of lymph node invasion vs. not, treatment by chemo-endocrine therapy vs. endocrine therapy alone. On the CANTO Cohort, the RR Combined

https://doi.org/10.1038/s41467-025-60824-z

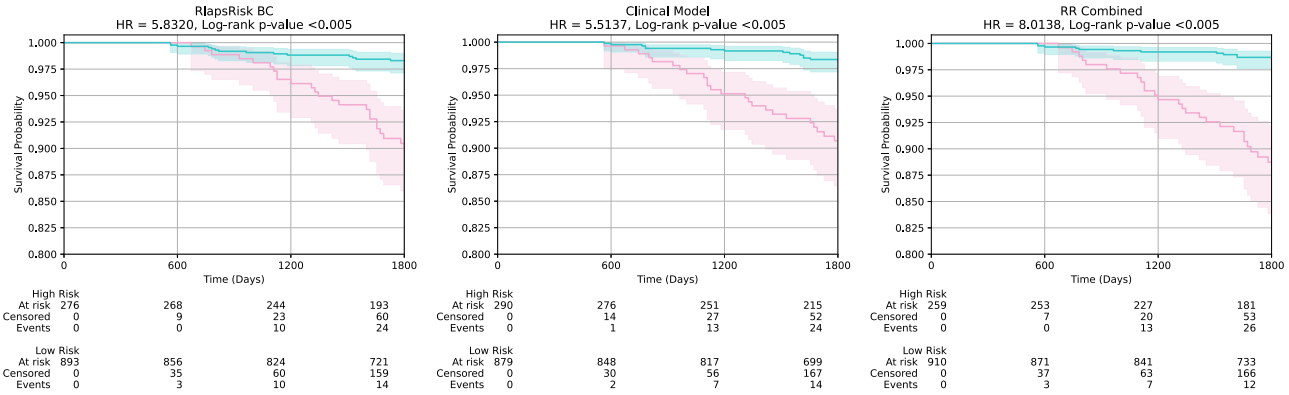

Fig. 3 | Metastases-free survival of patients stratified according to RlapsRisk BC classifier (left), Clinical score classifier (middle), and RR Combined classifier (right) among patients from the CANTO validation cohort. *p* Values were computed using the two-sided log-rank test to compare survival distributions between groups. Shaded areas represent 95% confidence intervals estimated using the Greenwood formula.

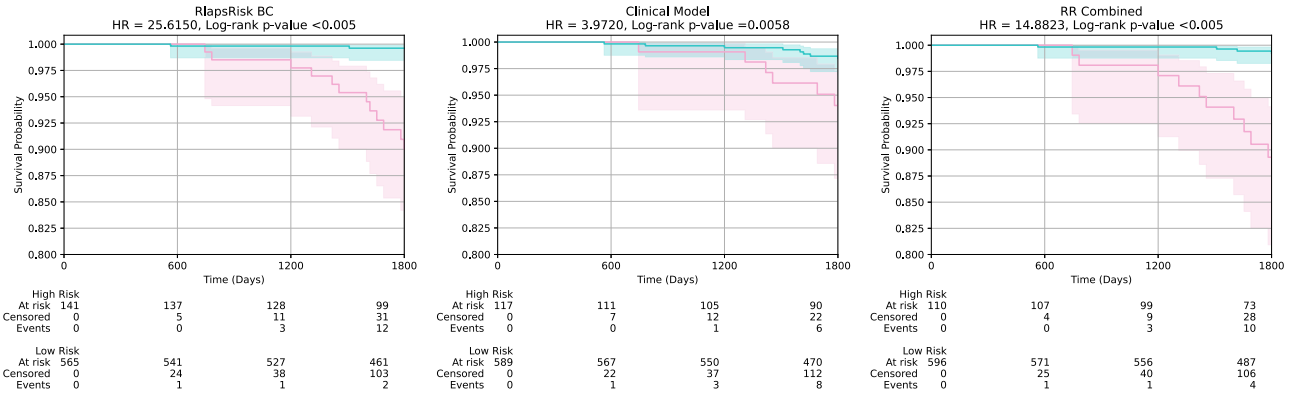

Fig. 4 | Metastases-free survival of patients stratified according to RlapsRisk BC classifier (left), Clinical score classifier (center), and RR Combined classifier (right) among patients from the CANTO intermediate clinical risk validation cohort. *p* Values were computed using the two-sided log-rank test to compare survival distributions between groups. Shaded areas represent 95% confidence intervals estimated using the Greenwood formula.

classifier increased the cumulative sensitivity by six points and the dynamic specificity by four points in comparison to the Clinical score classifier alone (Table 4). When looking at the histological grade 2 subgroup and the intermediate clinical risk subgroup (defined in the Supplementary Table 4), the RR Combined classifier outperformed the clinical score classifier by 21, resp. 28 points, in sensitivity and 2, resp. 3 points, in specificity. These figures show that adding RlapsRisk BC score to the current clinical factors improves prognosis within subgroups with a difficult prognosis estimation, except for premenopausal patients. Additional Kaplan–Meier analyses highlighted significant differences between high- and low-risk in all these subgroups (see Supplementary Information Figs. 3–6).

Model interpretability: histological feature assessment

To make sure our model leveraged reliable histological features, even though it was not based on handcrafted features, we computed the marginal contribution of clusters of similar tiles to the overall risk score to assess their positive or negative contribution to the final risk score predicted by the multilayer perceptron (MLP), on WSI from the CANTO cohort. The most impacting clusters were further reviewed and annotated by an expert pathologist blinded to the predicted outcome (see Fig. 5 for methodology and Supplementary Methods for more details).

Some histological patterns, as annotated by a pathologist, were associated with an increase of the predicted risk of relapse like the presence of high tumor cell content ($p < 0.05$), high degree of nuclear pleomorphism ($p < 0.05$), massive architecture ($p < 0.05$), low tubule formation ($p < 0.05$) and trabecular structures ($p < 0.05$) as well as mitotic activity ($p < 0.05$) (Fig. 6). Some features were also associated with a lower predicted risk of relapse such as fibrosis ($p < 0.05$), the presence of vascular structures ($p < 0.05$) or regions in the adjacent normal tissue ($p < 0.05$). The complete analysis is available in Supplementary Information Table 13. These results showed that our model was relying on well-established histological factors to predict patient outcomes. The majority of histological features significantly associated with an increase or decrease of the risk were statistically significant both on H&E and HES stainings individually (Supplementary Material Tables 14 and 15), supporting the robustness of our model across different staining platforms.

Discussion

Digital pathology is increasingly being integrated into routine clinical practice, and workflows that include digitization of glass slides are no longer an exception in pathology laboratories. This ongoing transition is paving the way for the implementation of AI-based digital pathology medical devices in clinical practice.

In this study, we developed and validated a digital pathology score, which predicts distant recurrence at 5 years after surgery in adequately treated early-stage BC patients, bearing ER-positive and HER2-negative tumors. Our method uses just a standard-stained H&E or HES and a scanned tumor slide already available for diagnostic purposes in the pathology laboratory.

https://doi.org/10.1038/s41467-025-60824-z

Table 4 | Performance of the classifiers in different subgroups of the CANTO cohort (external validation)

Subgroup	N	RlapsRisk BC classifier		Clinical score classifier		RR Combined classifier	
		Cumulative sensitivity	Dynamic specificity	Cumulative sensitivity	Dynamic specificity	Cumulative sensitivity	Dynamic specificity
Full population	1169	0.64 (0.51–0.77)	0.79 (0.77–0.81)	0.63 (0.50–0.76)	0.76 (0.74–0.78)	0.69 (0.56–0.81)	0.80 (0.78–0.82)
Intermediate clinical risk group	706	0.86 (0.69–1.00)	0.82 (0.80–0.85)	0.43 (0.20–0.66)	0.84 (0.82–0.86)	0.71 (0.49–0.91)	0.87 (0.85–0.89)
Histological grade 2	720	0.81 (0.65–0.95)	0.79 (0.76–0.82)	0.54 (0.36–0.73)	0.81 (0.78–0.83)	0.75 (0.58–0.91)	0.83 (0.81–0.86)
Patients with node-positive disease	394	0.56 (0.40–0.72)	0.68 (0.64–0.73)	0.74 (0.60–0.88)	0.62 (0.58–0.67)	0.71 (0.56–0.85)	0.67 (0.63–0.72)
Patients with NO disease	772	0.82 (0.61–1.00)	0.84 (0.81–0.86)	0.35 (0.12–0.61)	0.83 (0.80–0.85)	0.63 (0.37–0.88)	0.86 (0.84–0.89)
Pre-menopausal patients	297	0.32 (0.00–0.59)	0.79 (0.74–0.83)	0.51 (0.23–0.77)	0.84 (0.79–0.87)	0.32 (0.07–0.59)	0.84 (0.80–0.87)
Post-menopausal patients	872	0.75 (0.61–0.88)	0.79 (0.76–0.82)	0.67 (0.52–0.82)	0.74 (0.71–0.77)	0.82 (0.69–0.93)	0.79 (0.76–0.82)
Patients who received adjuvant CT	525	0.56 (0.40–0.70)	0.68 (0.65–0.72)	0.65 (0.50–0.79)	0.61 (0.57–0.65)	0.65 (0.50–0.79)	0.68 (0.64–0.72)
Patients who did not receive adjuvant CT	644	1.00 (1.00–1.00)	0.87 (0.84–0.89)	0.56 (0.23–0.88)	0.88 (0.86–0.91)	0.85 (0.59–1.00)	0.90 (0.87–0.92)

RlapsRisk BC score has a strong prognostic value that is independent of established clinico-pathological factors, validated in an independent multicentric cohort. Although 55% of the patients of the validation cohort originate from the same hospital that provided the samples for the training of the algorithm, the performance obtained on the other 45% showed similar results. The RR Combined model, integrating clinico-pathological factors and RlapsRisk BC, effectively dichotomized patients into low- and high-risk groups with strong discriminative power across the entire population and most clinical subgroups, except for the premenopausal subgroup. The utilization of RlapsRisk BC demonstrated notable efficacy in enhancing risk assessment accuracy within the intermediate clinical risk group, a segment that poses considerable complexity for treatment decision-making.

Unlike earlier digital pathology approaches that primarily predict morphological features such as histological grade or Ki67 index[26–28] or leverage tumor microenvironment characteristics[20], we trained our DL model to directly predict 5-year MFS from WSI without requiring local annotations. Recent studies have demonstrated the feasibility of DL models for outcome prediction from histopathology images[29–31]. However, these models are built using predefined scores (Nottingham Grading System), predefined histological features, or rely on the integration of molecular recurrence scores as indirect proxies for prognosis[32–34]. In contrast, our model is initially trained to estimate MFS exclusively from WSI, without incorporating manual annotations or molecular data. While we later evaluate its integration with clinical prognostic factors to enhance prognostic performance, the WSI-based score is learned in an end-to-end manner, ensuring that prognostic patterns are directly extracted from histological slides under the supervision of a survival endpoint. This fully automated pipeline allows the model to independently extract relevant morphological patterns before combination with standard clinical risk factors.

Our approach, which directly predicts a survival endpoint from H&E slides, has already been successfully applied in prostate cancer. Notably, AI models using digital histopathology analysis have improved prognostic stratification and therapy personalization in prostate cancer, outperforming standard classification tools[35]. These findings support the relevance of applying a similar approach in BC, where integrating AI into digital pathology could further refine risk assessment.

A different approach presented by Wang et al.[30] attained a risk prediction from Nottingham histological grade through the re-stratification of the intermediate category (NHG2). NHG2, which encompassed the largest group of patients, was dichotomized into a low-grade subgroup and a high-grade subgroup, achieving a HR of 2.94 (95% CI 1.24–6.97, $p = 0.015$) for the stratification between the two groups according to recurrence-free survival[30]. Similarly, in our study, RlapsRisk BC classifier achieved on the validation cohort a high discriminative power on NHG2 patients with a HR of 5.67 (95% CI = [2.34–11.74], p value < 0.05) for MFS when stratifying patients into low- and high-risk groups (Supplementary Fig. 3).

The interpretability analysis of the pathological features associated with an increased or decreased risk of relapse, as predicted by RlapsRisk BC, supports the validity of our model. The model's reliance on well-established histological prognostic factors—such as nuclear pleomorphism, tumor architecture, and mitotic activity—demonstrates its alignment with known pathology-driven risk assessments. Our interpretability analysis confirmed that RlapsRisk BC relies on meaningful pathological features to predict an increased or decreased risk of relapse, supporting the robustness of our approach. However, a more comprehensive study would be required to establish the causal relationship between certain pathological features and the model's predictions. Aligned with other recent studies[36–39], our analysis highlights the potential role of tumor-adjacent microenvironment features, such as vascular structures and fibrosis, in prognosis assessment. Our analysis brought out into relief the microenvironment features that

https://doi.org/10.1038/s41467-025-60824-z

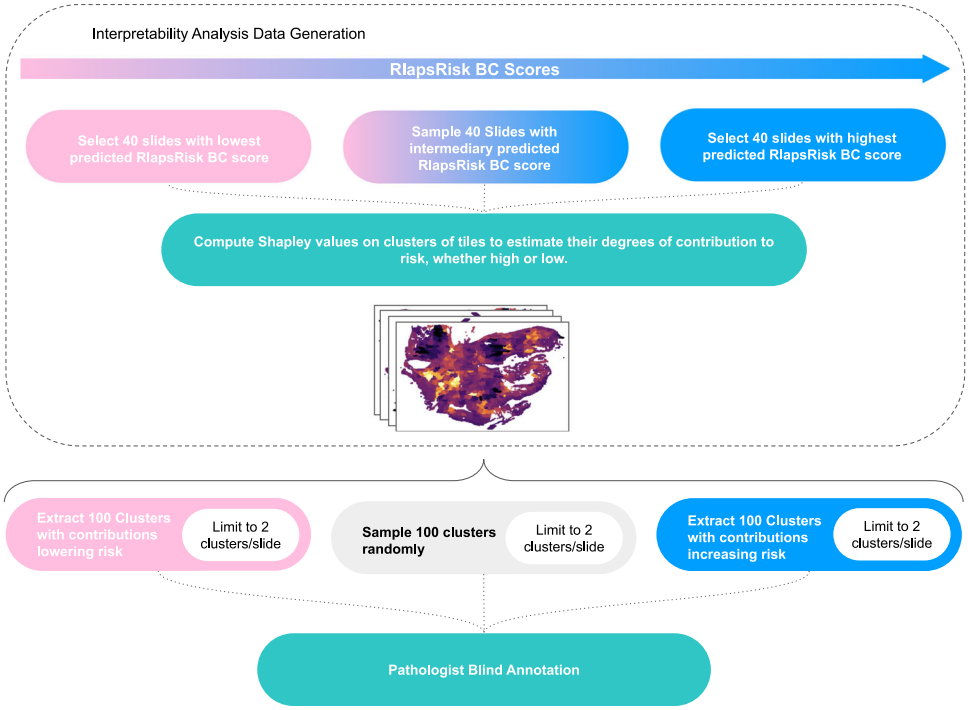

Fig. 5 | Interpretability of the RlapsRisk BC model. The methodology applied to generate the data used for this analysis is outlined. To streamline pathologist review, slides were sampled to express as much diversity as possible in the high, low, and intermediary predicted risks. Shapley values were computed to evaluate the positive or negative contributions of clusters of similar tiles. A selection of the most contributing clusters of tiles was made, and random clusters were also sampled from the distribution of Shapley values. An expert pathologist annotated the data, blinded to any information about contribution or risk.

could play a role in prognosis assessment, such as vascular structures or the presence of fibrosis. These findings support the approach of taking into consideration the tissue as a whole, rather than focusing on tumor cells only.

Currently, one of the challenges in ER+/HER2− EBC management is the adaptation of the treatment according to the risk of the patient. Reducing the number of unnecessary adjuvant chemotherapies or the length of endocrine therapies to improve QoL while maintaining an equivalent survival rate of the patients is the key challenge. This group is currently the target population for molecular signatures, where different genomic scores aim to guide the decision to avoid chemotherapy in certain patients[5,31]. However, variations in test results, restrictive indications, and limited reimbursement in many healthcare systems present challenges to their widespread use[40,41].

Multimodal histopathologic models have been investigated as a means to stratify hormone receptor-positive EBC and could serve as pre-screening tools to identify patients who may or may not require more expensive genomic testing[42]. While our study does not provide a direct comparison with genomic assays such as Oncotype DX, our results suggest that a WSI-based approach could serve as a complementary tool in clinical decision-making. Head-to-head comparison of RlapsRisk BC and molecular signatures in a cohort similar to TRANSATAC would provide further insights into its potential clinical utility, including defining optimal thresholds for genomic test endorsement[43]. In addition, assessment of the impact of a decision-making strategy based on RlapsRisk BC would require a randomized clinical trial to eliminate potential biases that may arise from observational studies.

While our study introduces an innovative and valuable tool for patient stratification, it also has certain limitations.

While the training and validation cohorts include a limited number of premenopausal patients, RlapsRisk BC appears to exhibit suboptimal performance within this subgroup. Addressing this issue requires dedicated studies with larger sample sizes to thoroughly investigate its clinical implications. Of note, current prognostic tools to stratify risk such as gene expression signatures have also shown limited performance in this premenopausal setting. Future studies incorporating comparisons to molecular test scores would be beneficial to strengthen the evidence supporting our tool's utility and to offer complementary insights on its prognostic accuracy relative to established standards.

To use RlapsRisk BC in clinical practice, the calibration protocol requires 30 BC slides, which may pose challenges, particularly in smaller pathology laboratories. While calibration is a standard procedure for many medical devices, it could impair the clinical deployment of such AI tools. Furthermore, variations in staining protocols, scanner settings, and other technical factors between healthcare centers could introduce domain shifts, potentially affecting the model's reliability. Here, prognosis results were robust with respect to the staining protocols (H&E and HES) since concordant prognostic results were found in two centers using two different staining protocols. However, the diversity of worldwide staining protocols is larger than the two protocols tested here, stressing the importance of analytical validations.

In conclusion, RlapsRisk BC resulted to be an independent prognostic factor of MFS and added significant prognostic information to clinico-pathological variables. After patients' dichotomization into low- and high-risk groups, RR Combined with classical clinico-pathological risk factors showed higher discrimination power compared to clinico-pathological risk factors alone. A prospective observational study comparing RlapsRisk BC to molecular prognostic signatures is currently ongoing and will determine the impact of the implementation of this AI-based tool into the practice workflow. Future extensions of our research include the development of novel algorithms, notably adapted to biopsy specimens. To deepen interpretability issues, an exhaustive analysis of tiles is contemplated with a focus on the spatiality notion that could provide novel insights into tumor biology.

https://doi.org/10.1038/s41467-025-60824-z

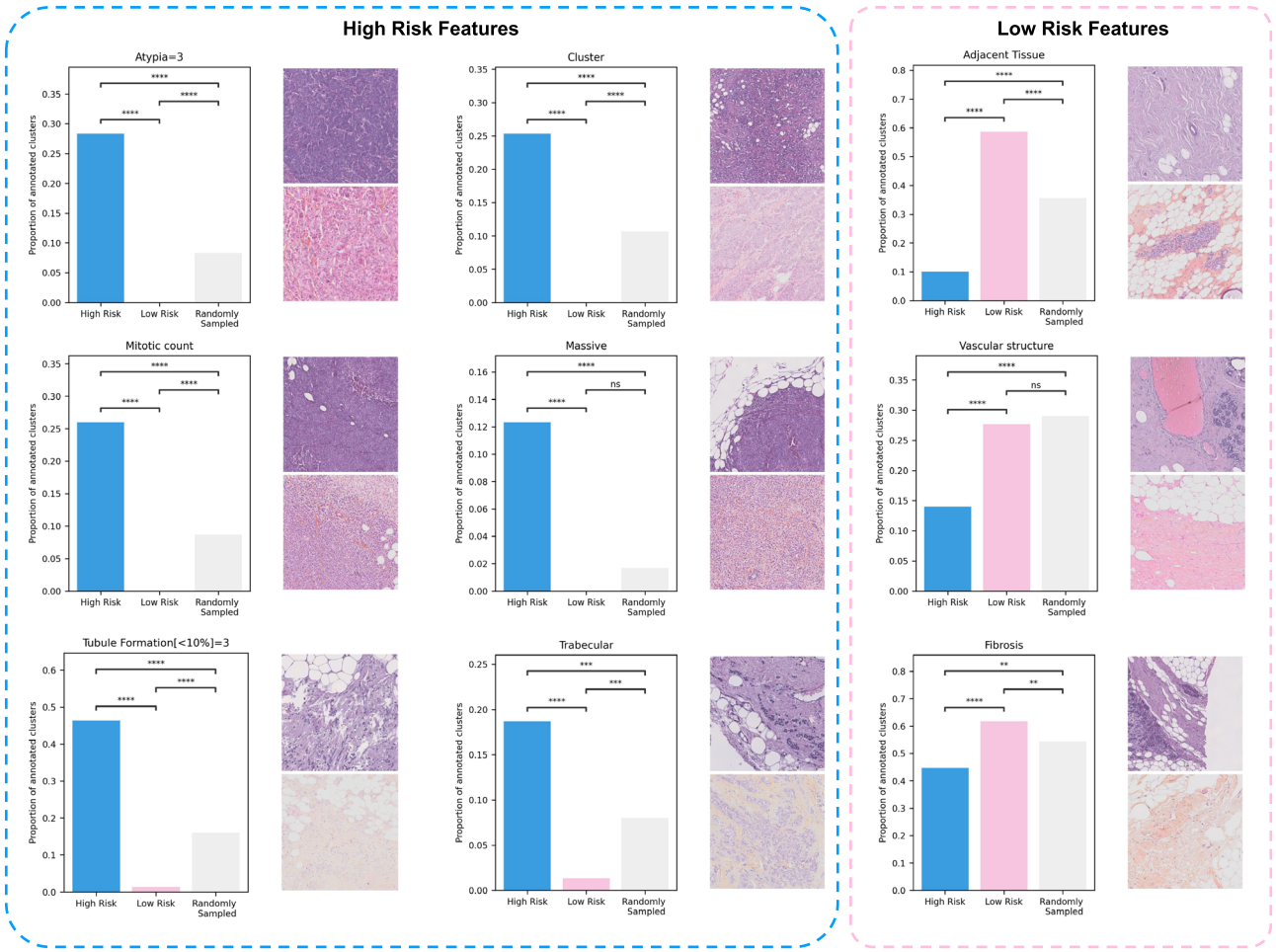

Fig. 6 | A selection of the histological features annotated by an expert pathologist. The proportion of tiles annotated as containing a particular histological feature within clusters of tiles linked to high and low risk, as well as randomly selected ones, is presented alongside the significance of p values from the Chi-square test of independence computed between each group, adjusted for multiple testing using the Benjamini–Hochberg procedure. For each histological feature, the region of a WSI containing the annotated cluster of tiles is displayed, extracted both from the CANTO H&E dataset (upper histological image) and from the CANTO HES dataset (lower histological image). ns $p > 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$.

Methods

This study complies with all relevant ethical regulations. The need for informed consent was waived due to the retrospective nature of the study and anonymization of all data.

Dataset description

GrandTMA. To build our models, we used a discovery dataset collected retrospectively from patients treated at Gustave Roussy in France and included in the "GrandTMA" cohort. This cohort comprises all patients newly diagnosed with a breast carcinoma as part of the "One-stop breast clinic" program at Gustave Roussy between October 10, 2005 and February 7, 2013[22]. The inclusion criteria for the present study were (i) diagnosis of invasive breast carcinoma, with or without associated in situ carcinoma, (ii) any type of treatment but neoadjuvant chemotherapy, (iii) availability of a surgical specimen with a formalin-fixed, paraffin-embedded (FFPE) tumor sample available, (iv) complete clinical and therapeutic data, (v) follow-up over at least 4 years and updated annually. The exclusion criteria were (i) exclusive non-invasive tumors, (ii) cytology-only available cases, (iii) absence of follow-up, (iv) other non-adenocarcinomatous lesions of the breast. This led to the inclusion of 1802 patients diagnosed with early invasive BC (1429 ER+/HER2−, 110 ER+/HER2+, 70 ER−/HER2+, 193 ER−/HER2−), with at least 1 available HES-stained tumor slide from the surgical specimen at the Pathology Department (details in Supplementary Fig. 1). For the model training, we used slides from the primary diagnosis whenever possible. For each of the 1429 ER+/HER2− patients, we also recut the FFPE block to obtain a new H&E slide that was digitized for the purpose of the study. Biomarker status (ER, PR, and Ki67 immunohistochemistry (IHC) expression, and HER2 protein expression/gene amplification) was defined and determined locally according to the current recommendations of the College of American Pathologists and the American Society of Clinical Oncology[44,45], and the French recommendations of the Study Group on Immunohistochemical Prognostic Factors in Breast Cancer[46]. ER and PR expression positivity was defined as an IHC staining of at least 10% of tumor cells, as is standard in several European countries. Lesions were considered positive for HER2 (score 3+) if the number of tumor cells with a complete and intense membrane IHC staining exceeded 10% of the whole invasive tumor cells population; equivocal (score 2+) if the number of tumor cells showing a complete and moderately intense membrane IHC staining, or an incomplete basolateral membrane IHC staining of moderate to severe intensity exceeded 10% of the total infiltrating tumor population; and negative in the remaining cases (scores 0 and 1+). When dichotomized, Ki67 cut-off was defined according to the current recommendations (i.e., cut-off set at 20%)[45,47].

Image acquisition was performed using an Olympus VS120 slide scanner at ×20 magnification for the HES slides and Olympus VS200 at ×20 for the H&E slides. In order to avoid scanning issues that might

affect subsequent image analysis, all slides were checked by a pathologist after digitization to discard slides with insufficient quality and rescanned when necessary (blurred images, need for slides re-mounting when the coverslip was damaged).

This work was carried out in accordance with the provisions of the Public Health Code applicable to research not involving the human person (Public Health Code - Article R1121-1 amended by Decree no. 2017-884, May 9, 2017) and therefore it does not come under the jurisdiction of a Committee for the Protection of Persons. It obtained the favorable opinion of the expert Committee for Research, Studies and Evaluations in the field of breast pathology, as well as of the Ethics Committee (Data Protection Office, Gustave Roussy). It has been submitted to the National Commission for Computing and Liberties under reference No. F20220121170839 and has been declared in accordance with the reference methodology MR-004. The patients involved were informed of the research via an information letter distributed by postal mail with the possibility of opposing the study.

CANTO. For the purpose of an external validation of our model, we used a dataset from the French observational and prospective CANTO cohort (NCT01993498)[23], enrolling patients from 26 cancer centers. In this cohort, patients were included at diagnosis of their invasive BC, before any treatment, following the given criteria: (i) women only, (ii) aged over 45 years old, (iii) HER2− and ER+ (same definition as for the GrandTMA cohort), (iv) with a histologically invasive BC diagnosed, (v) with no clinical evidence of metastasis at the time of inclusion. Out of 12,000 patients accrued in CANTO so far, 1703 ER+/HER2− EBC patients had a minimum follow-up of 5 years and were eligible for the present study (707 patients from Gustave Roussy and 997 patients from other cancer centers of the UNICANCER group). None of these patients were also included in the GrandTMA cohort. Thirty-one patients from Gustave Roussy had incomplete data and were excluded from the study, resulting in 676 HES slides available from primary diagnosis that were digitized for the purpose of the study with an Olympus VS200 scanner at ×20 magnification. From the other centers of the CANTO cohort, we had access to 553 H&E slides from resection that were recut, restained, and digitized for the purpose of the study with a Hamamatsu Nanozoomer S60 at a ×20 magnification (details in Supplementary Fig. 2). In total, 1229 patients had exploitable WSI from HES or H&E slides together with full clinico-pathological features (described in Supplementary Table 1).

All data collections were performed in the framework of the CANTO clinical trial (NCT01993498), in compliance with all legal requirements. All patients included in the study were informed through the website https://mesdonnees.unicancer.fr/ on the reuse of their data for a separate objective with the possibility of opposing the study.

Endpoint. The chosen endpoint for survival data analysis was MFS at 5 years, defined as the time from initial surgery to the occurrence of a metastatic event or death before 5 years. Operable local relapse or axillary lymph node recurrence events were ignored. Patient's follow-up was censored at the time of contralateral BC, second non-breast primary cancer, or last available date of follow-up.

Model description

To develop our risk score from histology slides, we used a method composed of three steps: (i) tissue tiling, (ii) feature extraction, (iii) creation of a risk score. The transformation of the score into a probability of occurrence of a MFS event before 5 years and the selection of a threshold are two additional steps, detailed in "Statistical analysis."

Tissue segmentation and tiling. Each of the WSI was first divided into small squares, 76 × 76 micrometers in size (224 × 224 pixels) called "tiles." This tiling was performed by first segmenting the tissue, using a pre-trained U-Net neural network[48] that discarded the background and artifacts of scanning or preparation. This segmented tissue was then divided into N (ranging from 10,000 to 75,000) tiles.

Feature extraction. The N tiles were embedded into D-dimensional feature vectors using a pre-trained Vision Transformer (Fig. 1). We implemented iBOT ViT-B, a self-supervised learning transformer framework that improved performance for various prediction tasks in previous studies[49] trained on the Cancer Genome Atlas PanCancer40M dataset, which covers 13 anatomic sites and 16 cancer subtypes for 5558 patients, representing a total of 6093 slides. Multiple data augmentations (random cropping, random flips, color jitter, random grayscale, Gaussian blur) were applied while the model was optimized for 153,000 iterations (approximately 50 h) on 32 NVIDIA Tesla V100 graphics processing units (GPU). The following hyperparameters were used to train the model: teacher temperature was set to 0.04 with an initial value of 0.04 and 30 warm-up epochs. AdamW optimizer[50] was used and learning rate linearly ramped up during the first ten epochs to its base value scaled with the total batch size according to: $0.0005 \times \text{batch-size}/256$[51]. The final learning rate was set to 0.000002 through a cosine schedule. This frozen pre-trained algorithm was then used to extract features during training and inference.

Risk prediction using multiple instance learning (MIL). The N feature vectors were then aggregated using a MIL model trained to predict MFS at 5 years with a stratified threefold cross-validation approach, repeated five times. For each split, five models were also trained with random initialization of the weights. Given the limited number of events, stratification was performed based on event occurrence to ensure a minimum number of events in each fold (Fig. 1). For inference on the external cohorts, predictions were generated by ensembling all 75 trained models through output averaging.

We reimplemented the attention-based model called DeepMIL proposed by Ilse et al.[52]. DeepMIL is a deep MIL framework designed for weakly supervised classification. It aggregates instance-level (the WSI's tiles in our case) features using an attention-based pooling mechanism, which assigns learned importance weights to each instance within a bag. This allows the model to focus on the most relevant instances for prediction. The weighted instance representations are then combined into a single bag-level feature vector, which is processed by a fully connected layer to generate the final classification output. This approach enhances interpretability while effectively handling variability within each bag. A linear layer with L neurons ($L = 256$ here) was applied to the embedded features followed by a Gated Attention layer with L hidden neurons. A MLP with 128 neurons was then applied to the output. To speed-up training and fit all the data in memory, only a random subset of 8000 tiles per WSI was used, while all tiles of a slide are processed for inference. DeepMIL was trained utilizing a loss function calculated with a smoothed variant of the concordance index, as suggested by Mayr and Schmid in this study[53]. We used a smooth, differentiable loss function based on a sigmoid approximation of the standard C-index indicator function. This smoothing introduces a parameter σ that controls the transition sharpness, enabling gradient-based optimization. The loss function is integrated into a gradient boosting framework, iteratively minimizing the smoothed empirical risk to improve model discrimination. This approach ensures that the learned biomarker combinations are directly aligned with survival prediction performance. The model was trained with the following hyperparameters: batch size = 128, learning rate = 0.001, and MLP dropout = 0.4 (a comprehensive list of hyperparameters is provided in the Supplementary Methods). Given the 7.69% relapse rate in our training cohort, this resulted in an average of 9.84 events per minibatch, providing adequate supervision for model optimization. While minibatches without observed events were

possible, their occurrence was rare (~0.0036% probability). Although such batches do not directly contribute to differentiating survival outcomes, their impact was mitigated by the smoothed C-index loss, which allows gradient updates even when minibatches contain few or no events.

The models were trained for 15 epochs with a batch size of 128. Each epoch required approximately 26 s to complete, resulting in a total training time of 6 min and 30 s per fold. The entire training of the models with a five-times-repeated threefold cross-validation process with multiple random weight initializations took approximately 1 h 38 min. Training was conducted on a Tesla T4 Nvidia GPU using the PyTorch framework. The CPU memory usage for loading all features of the training data was 45GB, with an additional 13GB of GPU memory utilized during training.

Threshold determination. For the RlapsRisk BC score, as well as the clinical and RR Combined risk scores, we fitted a Weibull AFT (Accelerated Failure Time) model on the training dataset to transform risk scores into probabilities of occurrence of MFS event before 5 years. This step was used to identify the threshold of each risk score corresponding to a probability of 5-year MFS event of 5% defined by the Weibull Models. This 5% MFS rate threshold corresponds to the 5-year interpolation of an exponential model from the 10-year MFS of 10%, which is the most common output of the molecular signatures currently used in clinical practice[54]. The identified threshold was fixed before applying RlapsRisk on the validation dataset.

Method for interpretability feature assessment

Training an AI model on digital slides from diagnosis to predict metastatic relapse is an original approach compared to recent research works in Digital Pathology that generally predict well-known pathological features (e.g., histological grading or KI67 index). However, bypassing human knowledge in the training phase requires even more explanations on the functioning of the model. We detail herein our method to identify and characterize typical areas on a well-defined set of slides that had extreme risk scores. Interpretability relies on the possibility of accessing the relevant information for our model that is learned by the model itself.

In the model, each tile was associated with an attention score that was used in the final weighted average to obtain the input vector for the risk predictor (see Fig. 4). However, this score did not provide information about the impact on prognosis of the tile. To overcome this limitation, we aggregated clusters of tiles based on features' similarity (using SLIC algorithm[55]) on all slides of the dataset and computed the Shapley value[56] associated with each cluster. We thus measured the marginal contribution of each cluster to the overall risk score to assess its positive or negative contribution to the final risk score predicted by the final MLP. We aggregated tile features to accelerate the computation of Shapley values, ensuring that our methodology maintained coherence between the contribution of clusters of tiles' features and the contribution of individual tiles alone on the RlapsRisk BC output.

For each staining of the CANTO dataset (HE and HES), we sampled 120 slides (40 classified with the highest RlapsRisk BC scores, 40 classified with the lowest RlapsRisk BC scores by our model, and 40 slides with intermediary scores predicted risk by RlapsRisk BC). We computed the Shapley values of each cluster and extracted those with the 2 highest computed contributions from the 50 slides with the highest predicted risk available (10 slides are taken from the intermediary scores here) and those with the 2 lowest computed contributions from the 50 slides with the lowest predicted risk available (10 slides are taken from the intermediary scores here). We also sampled 100 clusters randomly from the entire set of clusters generated from the 120 slides. For each dataset, we thus obtained a total of 300 clusters of tiles, 100 associated with high risk, 100 associated with low risk and 100 randomly selected. This sampling method allowed us to get a broad representation of which features were associated with a higher or lower predicted risk in slides with high, intermediate or low predicted risk.

Those clusters of tiles were further reviewed and annotated by an expert pathologist blinded to the predicted outcome.

Forty-two histological features were recorded, encompassing tumor architecture patterns, stroma features, presence of different cell types and tiles' localization. The proportions of appearance of each feature in the highest and lowest contribution groups were compared with the Chi-square test of independence applied to contingency tables derived from the annotated data. Statistical significance was calculated using the Benjamini–Hochberg procedure adjustment and was also assessed when comparing high and low contribution groups to the randomly sampled clusters group.

Calibration protocol

To guarantee the robustness of our stratification in high or low risk across diverse data acquisition protocols like in CANTO H&E or CANTO HES, we implemented Uniform Piecewise Approximation (UPA) as a calibration strategy[57]. Inspired by image processing's histogram matching, UPA aligns the unseen dataset's model prediction distribution with a predefined reference distribution.

The reference distribution was defined as the set of scores obtained from inference of RlapsRisk BC on the H&E samples of the training cohort, from which we derived the cumulative distribution.

For each new dataset or new laboratory setting, a set of 30 samples is selected, comprising 10, 15, and 5 samples, respectively, from histological grades 1, 2, and 3. This selection aims to mirror the histological grade distribution observed in the discovery dataset. Histological grade was considered as the main prognostic factor that is routinely captured by the pathologist's analysis of the tumor slide only. Thus, we ensure a good representativity of morphologies and an ease of implementation in clinical practice. These 30 samples are used to generate a new set of RlapsRisk BC predictions, from which we derived an estimated cumulative distribution of the unseen dataset (corresponding to the data acquisition protocol of the laboratory).

Next, a linear function is fitted to map the unseen set's cumulative distribution onto the reference distribution, essentially aligning the prediction "shapes," without impacting the relative order of patients' risk. This mapping function was then applied to calibrate any new prediction from the unseen set, ensuring consistency with the desired reference distribution and enhancing generalizability across acquisition protocols. By leveraging UPA, we could achieve consistent model predictions even when applied to data acquired differently from the reference set.

One-shot validation and bootstrapping

For the purposes of validation of the UPA calibration, we performed the calibration from only one subsample (randomly selecting the 10 grade 1 patients, 5 grade 2 patients and 5 grade 3 patients) to calibrate the score for each subcohort (CANTO H&E and CANTO HES) corresponding to two different data acquisition protocols. Once calibrated, the scores were merged as only one larger set forming the CANTO cohort (including both WSI from H&E and HES).

To assess the impact of the sample selection for the calibration step, we performed a repeated sensitivity analysis, reported in Supplementary Table 12 and Supplementary Fig. 3. CIs are derived from the point estimates obtained from the repetitions of the resampling ($N = 1000$).

Statistical analysis

The survival analyses were performed using the Cox model considering the well-known BC risk factors as a reference score. We compared the performance for the prediction of the MFS of this "clinical score," RlapsRisk BC, and a model combining RlapsRisk BC and the clinical factors. We used Harrell's C-index to assess the discrimination performance of the different scores. The comparison of the C-indices of the different models was based on permutation tests using 10,000 permutations. We computed the cumulative sensitivity as well as the dynamic specificity[58] of each model for the 5-year MFS prediction. Those metrics are natural extensions of the so-called sensitivity/specificity to the particular setting of time-to-event outcomes that may be censored. CIs were computed using bootstrapping with nonparametric, unstratified resampling (10,000 samples). To illustrate the discrimination on the survival for the test dataset, we used the Kaplan–Meier estimate based on the binarized scores using the threshold defined previously and compared the risk groups using the log-rank test. All tests were two-tailed, and p values < 0.05 were considered statistically significant. The C-index, Kaplan–Meier, and multivariable Cox proportional hazards models were implemented in the lifelines (0.27.4) package of Python. Cumulative sensitivities and dynamic specificities were computed using the scikit-survival package (0.19.0). We followed the recent MI-Claim checklist to improve the reporting of our ML methods. The checklist is available in the Supplementary Information.

Data availability

The GrandTMA dataset that supports the findings of this study is available from Gustave Roussy but restrictions apply to the availability, which was used with permission for the current study, and so is not publicly available. The CANTO dataset used for external validation is available from UNICANCER but restrictions apply to the availability of data, which was used with permission for the current study, and so is not publicly available. The datasets, or a test subset, may be available from Gustave Roussy or UNICANCER, subject to ethical approvals. Requests for data access can be directed to https://redcap.link/DataRequestClinicalTrialsGustaveRoussy for GrandTMA or s-everhard@unicancer.fr for CANTO cohorts.

Code availability

The code used for training the models has a large number of dependencies on internal tooling and its release is therefore not feasible. However, all experiments and implementation details are described thoroughly in the "Methods" and Supplementary Information (Cross-validation experiments and model hyperparameters) so that they can be independently replicated with non-proprietary libraries.

References

1. Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Ferreira, A. R. et al. Differential impact of endocrine therapy (ET) and chemotherapy (CT) on quality of life (QoL) of 4,262 breast cancer (BC) survivors: a prospective patient-reported outcomes (PRO) analysis. *J. Clin. Oncol.* **37**, 512–512 (2019).
3. Mastro, L. D. et al. Extended therapy with letrozole as adjuvant treatment of postmenopausal patients with early-stage breast cancer: a multicentre, open-label, randomised, phase 3 trial. *Lancet Oncol.* **22**, 1458–1467 (2021).
4. Harbeck, N. et al. Adjuvant abemaciclib combined with endocrine therapy for high-risk early breast cancer: updated efficacy and Ki-67 analysis from the monarchE study. *Ann. Oncol.* **32**, 1571–1581 (2021).
5. Sparano, J. A. et al. Prospective validation of a 21-gene expression assay in breast cancer. *N. Engl. J. Med.* **373**, 2005–2014 (2015).
6. 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).
7. Sparano, J. A. et al. Clinical and genomic risk to guide the use of adjuvant therapy for breast cancer. *N. Engl. J. Med.* **380**, 2395–2405 (2019).
8. DePolo, J. Oncotype DX tests for early-stage breast cancer and DCIS. Breastcancer.org. https://www.breastcancer.org/screening-testing/oncotype-dx (2022).
9. Wishart, G. C. et al. PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2. *Br. J. Cancer* **107**, 800–807 (2012).
10. Wishart, G. C. et al. A population-based validation of the prognostic model PREDICT for early breast cancer. *Eur. J. Surg. Oncol.* **37**, 411–417 (2011).
11. Polley, M.-Y. C. et al. An international Ki67 reproducibility study. *J. Natl. Cancer Inst.* **105**, 1897–1906 (2013).
12. Casterá, C. & Bernet, L. HER2 immunohistochemistry inter-observer reproducibility in 205 cases of invasive breast carcinoma additionally tested by ISH. *Ann. Diagn. Pathol.* **45**, 151451 (2020).
13. Gown, A. M. Current issues in ER and HER2 testing by IHC in breast cancer. *Mod. Pathol.* **21** (Suppl. 2), S8–S15 (2008).
14. NHS.UK. Predict breast cancer. https://breast.predict.nhs.uk/.
15. Wishart, G. C. et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res.* **12**, R1 (2010).
16. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. USA.* **115**, E2970–E2979 (2018).
17. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
18. Wulczyn, E. et al. Interpretable survival prediction for colorectal cancer using deep learning. *npj Digit. Med.* **4**, 71 (2021).
19. Ibrahim, A. et al. Artificial intelligence in digital breast pathology: techniques and applications. *Breast* **49**, 267–273 (2020).
20. Amgad, M. et al. A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer. *Nat. Med.* **30**, 85–97 (2024).
21. Ogier du Terrail, J. et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat. Med.* **29**, 135–146 (2023).
22. Delaloge, S. et al. The challenge of rapid diagnosis in oncology: diagnostic accuracy and cost analysis of a large-scale one-stop breast clinic. *Eur. J. Cancer* **66**, 131–137 (2016).
23. Vaz-Luis, I. et al. UNICANCER: French prospective cohort study of treatment-related chronic toxicity in women with localised breast cancer (CANTO). *ESMO Open* **4**, e000562 (2019).
24. Harrell, F. E. Jr, Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).
25. Down, S. K., Lucas, O., Benson, J. R. & Wishart, G. C. Effect of PREDICT on chemotherapy/trastuzumab recommendations in HER2-positive patients with early-stage breast cancer. *Oncol. Lett.* **8**, 2757–2761 (2014).
26. Couture, H. D. et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *npj Breast Cancer* **4**, 30 (2018).
27. Koopman, T., Buikema, H. J., Hollema, H., de Bock, G. H. & van der Vegt, B. Digital image analysis of Ki67 proliferation index in breast cancer using virtual dual staining on whole tissue sections: clinical validation and inter-platform agreement. *Breast Cancer Res. Treat.* **169**, 33–42 (2018).
28. Stålhammar, G. et al. Digital image analysis of Ki67 in hot spots is superior to both manual Ki67 and mitotic counts in breast cancer. *Histopathology* **72**, 974–989 (2018).

29. Jaroensri, R. et al. Deep learning models for histologic grading of breast cancer and association with disease prognosis. *npj Breast Cancer* **8**, 113 (2022).
30. Wang, Y. et al. Improved breast cancer histological grading using deep learning. *Ann. Oncol.* **33**, 89–98 (2022).
31. Cardoso, F. et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.* **375**, 717–729 (2016).
32. Howard, F. M. 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* **9**, 25 (2023).
33. Li, H. et al. Deep learning-based pathology image analysis enhances Magee feature correlation with oncotype DX breast recurrence score. *Front. Med.* **9**, 886763 (2022).
34. Su, Z. et al. BCR-Net: A deep learning framework to predict breast cancer recurrence from histopathology images. *PLoS ONE* **18**, e0283562 (2023).
35. Esteva, A. et al. Prostate cancer therapy personalization via multimodal deep learning on randomized phase III clinical trials. *npj Digit. Med.* **5**, 71 (2022).
36. Xu, Q. et al. Prognostic significance of the tumor-stromal ratio in invasive breast cancer and a proposal of a new Ts-TNM staging system. *J. Oncol.* **2020**, 9050631 (2020).
37. Zhao, H. et al. Identifying tumour microenvironment-related signature that correlates with prognosis and immunotherapy response in breast cancer. *Sci. Data* **10**, 119 (2023).
38. Li, H. et al. Collagen fiber orientation disorder from H&E images is prognostic for early stage breast cancer: clinical trial validation. *npj Breast Cancer* **7**, 104 (2021).
39. Nederlof, I. et al. Spatial interplay of lymphocytes and fibroblasts in estrogen receptor-positive HER2-negative breast cancer. *npj Breast Cancer* **8**, 56 (2022).
40. Marrone, M., Stewart, A. & Dotson, W. D. Clinical utility of gene-expression profiling in women with early breast cancer: an overview of systematic reviews. *Genet. Med.* **17**, 519–532 (2015).
41. Whitney, J. et al. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. *BMC Cancer* **18**, 610 (2018).
42. Boehm, K. M. et al. Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. *Nat. Commun.* **16**, 2106 (2025).
43. Sestak, I. et al. Comparison of the performance of 6 prognostic signatures for estrogen receptor-positive breast cancer: a secondary analysis of a randomized clinical trial. *JAMA Oncol.* **4**, 545–553 (2018).
44. Wolff, A. C. et al. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline focused update. *J. Clin. Oncol.* **36**, 2105–2122 (2018).
45. Dowsett, M. et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J. Natl. Cancer Inst.* **103**, 1656–1664 (2011).
46. Franchet, C. et al. Mise à jour 2021 des recommandations du GEF-PICS pour l'évaluation du statut HER2 dans les cancers infiltrants du sein en France. *Ann. Pathol.* **41**, 507–520 (2021).
47. Tashima, R. et al. Evaluation of an optimal Cut-Off point for the Ki-67 index as a prognostic factor in primary breast cancer: a retrospective study. *PLoS ONE* **10**, e0119565 (2015).
48. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds Navab, N., Hornegger, J., Wells, W. & Frangi, A.) 234–241 (Springer, 2015).
49. Filiot, A. et al. Scaling self-supervised learning for histopathology with masked image modeling. Preprint at *medRxiv* https://doi.org/10.1101/2023.07.21.23292757 (2023).
50. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://arxiv.org/abs/1711.05101 (2017).
51. Goyal, P. et al. Accurate, large minibatch SGD: training ImageNet in 1 hour. Preprint at https://arxiv.org/abs/1706.02677 (2017).
52. Ilse, M., Tomczak, J. & Welling, M. Attention-based Deep Multiple Instance Learning. PMLR (2018).
53. Mayr, A. & Schmid, M. Boosting the concordance index for survival data - a unified framework to derive and evaluate biomarker combinations. *PLoS ONE* https://doi.org/10.1371/journal.pone.0084483 (2014).
54. Kwa, M., Makris, A. & Esteva, F. J. Clinical utility of gene-expression signatures in early stage breast cancer. *Nat. Rev. Clin. Oncol.* **14**, 595–610 (2017).
55. Achanta, R. et al. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 2274–2282 (2012).
56. Shapley, L. S. 17. in *Contributions to the Theory of Games (AM-28)*, Vol. II (eds Kuhn, H. W. & Tucker, A. W.) 307–318 (Princeton University Press, 1953).
57. Roschewitz, M. et al. Automatic correction of performance drift under acquisition shift in medical image classification. *Nat. Commun.* **14**, 6608 (2023).
58. Uno, H., Cai, T., Tian, L. & Wei, L. J. Evaluating prediction rules for t-year survivors with censored regression models. *J. Am. Stat. Assoc.* **102**, 527–537 (2007).

Acknowledgements

We thank the patients who participated in both studies and who did not oppose this additional research. We also thank all the participating centers in the CANTO trial within the UNICANCER R&D frame: Gustave Roussy, Institut Jean Godinot, Institut de Cancérologie de Montpellier, Centre Oscar Lambret, Centre GF Leclerc, Institut de Cancérologie de Lorraine, Centre Léon Bérard, Centre François Baclesse, Centre Jean Perrin, Centre Antoine Lacassagne, Centre Paul Strauss, Institut Ste Catherine. The CANTO research project was supported by the French Government under the "Investment for the Future" program managed by the National Research Agency (ANR), grant no. ANR-10-COHO-0004. We would also like to acknowledge the support provided by Région Île de France and the impulsion given to this study by organizing the AI for Health Data Challenge in 2019.

Author contributions

M.L.-T., F.A., I.G., V.G., V.A., and C.S. conceived the project. C.S., V.G., K.E., B.S., and P.C. designed the methodology. Data collection was performed by I.G., A.J., D.D., and M.L.-T. Formal analysis was conducted by V.G. D.D. contributed with statistical analysis support. Medical expertise was provided by M.L.-T., F.A., S.D., B.P., and I.G. The original draft was written by I.G. and V.G., and all authors reviewed and edited the manuscript. L.H., J.L., M.S., F.B., A.K., R.D., M.A., L.G., I.B., M.A., J.L., M.S., S.E., A.S., J.-F.R., F.B., J.D., and P.C. contributed to manuscript revisions and approved the final version. All authors read and approved the final manuscript.

Competing interests

V.G., C.S., K.E., B.S., A.J., L.H., R.D., M.A., L.G., M.S., A.S., J.R., F.B., J.D., and V.A. are employees of Owkin Inc. S.D. reports grants and non-financial support from Pfizer, grants from Novartis, grants and non-financial support from AstraZeneca, grants from Roche Genentech, grants from Lilly, grants from Orion, grants from Amgen, grants from Sanofi, grants from Exact Sciences, grants from Servier, grants from MSD, grants from BMS, grants from Pierre Fabre, grants from Exact Sciences, grants from Besins, grants from European Commission grants, grants from French government grants, grants from Fondation ARC grants, grants from Taiho, grants from Elsan, outside the submitted work. F.A. declares institutional financial interests, research grants with

Novartis, Pfizer, AstraZeneca, Eli Lilly, Daiichi, Roche, and Sanofi. B.P. reports Consulting fees from Astra Zeneca (institutional), Seagen (institutional), Gilead (institutional), Novartis (institutional), Lilly (institutional), MSD (institutional), Pierre Fabre (personal), Daiichi-Sankyo (institutional/personal); research funding (institutional) from Astra Zeneca, Daiichi-Sankyo, Gilead, Seagen, MSD, and Fondation ARC. Travel support: Astra Zeneca; Pierre Fabre; MSD; Daiichi-Sankyo. M.L.-T. reports consulting fees from Astra Zeneca (institutional/personal), Seagen (personal), Lilly (personal), MSD (institutional/personal), Pierre Fabre (personal), Daiichi-Sankyo (institutional/personal), Myriad Genetics (personal), Exact Sciences (personal), Roche Diagnostics ((institutional/personal); research funding (institutional) from Roche Diagnostics, Daiichi-Sankyo, and Pierre Fabre. Travel support: AstraZeneca, Seagen, and Daiichi-Sankyo. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-60824-z.

Correspondence and requests for materials should be addressed to I. Garberis.

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