Article Type: **Original Article**

Title: Clinical Validation of RlapsRisk BC in an international multi-cohorts setting

Authors: V. Gaury[1*], Z. Vaquette[1], V. Aubert[1], C. Barcenas[2], L.J. Jones[3], D. Jacobs[1], J. Guillon[1], A. Filiot[1], B. van der Vegt[4], E. Hocquet[1], C. Maisin[5], L.K. Spary[6], M. Bateson[1], A.-L. Martin[7], D. Drubay[8], S. Everhard[7], L. Guillou[1], M. Sefta[1], I. Garberis[9], F. André[9,10], A. Vincent Salomon[11,12], S. Krishnamurthy[13], M. Lacroix-Triki[14]

Affiliations:

1. Owkin, Paris, France.
2. Department of Breast Medical Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA.
3. Barts Cancer Institute, Queen Mary University of London.
4. Department of Pathology and Medical Biology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
5. Cypath & Cypath-RB, Villeurbanne, France.
6. Institute of Cancer & Genetics and Wales Cancer Bank, School of Medicine, Cardiff University, Velindre Cancer Centre, Whitchurch, Cardiff, CF14 2TL, UK.
7. Unicancer, Paris, France.
8. Gustave Roussy, Office of Biostatistics and Epidemiology, Université Paris-Saclay, Villejuif, France Inserm, Université Paris-Saclay, CESP U1018, Oncostat, labeled Ligue Contre le Cancer, Villejuif, France.
9. INSERM U981, Gustave Roussy, Paris-Saclay University, Villejuif, France.
10. Department of Cancer Medicine, Gustave Roussy, Paris-Saclay University, Villejuif, France.
11. Diagnostic and Theranostic Medicine Division, Institut Curie, PSL University, Paris, France
12. INSERM U934, CNRS UMR 3215, Paris, France
13. Department of Anatomical Pathology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
14. Department of Pathology, Gustave Roussy, Paris-Saclay University, Villejuif, France.

***Corresponding author:** Valentin Gaury**,** Owkin France, 14/16 Bd Poissonnière, 75009 Paris, valentin.gaury@owkin.com

Context

Key Objective

This study aimed to assess whether RlapsRisk BC, a digital pathology-based AI model, can provide clinically meaningful prognostic stratification in ER-positive, HER2-negative early breast cancer. The AI model combines standard histology exhibited in whole slide images of hematoxylin and eosin stained tissue sections and clinical data. The prognostic performance of RlapsRisk BC was evaluated across multiple independent patient cohorts.

Knowledge Generated

The study demonstrates that RlapsRisk BC offers independent prognostic value beyond established clinical variables and genomic assays. It consistently stratifies patients into high- and low-risk groups for distant recurrence, with reproducible performance across diverse and independent cohorts, supporting its potential integration into routine clinical decision-making.

Relevance

RlapsRisk BC may serve as a scalable alternative or adjunct to molecular assays, supporting more personalized and accessible treatment decisions in breast cancer, particularly in settings where genomic testing is unavailable, limited, or yields intermediate-risk results.

ABSTRACT

Purpose:
The purpose of this study was to evaluate the prognostic performance of RlapsRisk BC, a multimodal deep learning–based digital pathology aid developed to predict distant recurrence-free interval (dRFI) in patients with early-stage, ER-positive, HER2-negative, breast cancer. RlapsRisk BC relies on routine H&E-stained whole-slide images (WSI) and standard clinicopathologic features.

Methods:
RlapsRisk BC, developed and internally validated using seven retrospective cohorts totaling 6,039 patients, was assessed for its ability to stratify patients into high- and low-risk groups for distant recurrence-free interval (dRFI). A blinded, retrospective validation was performed across three independent international cohorts (UK, France, and USA), including 593 patients with non-metastatic, ER-positive, HER2-negative, breast cancer who received adjuvant endocrine therapy alone.

Results:
Across all validation cohorts, RlapsRisk BC consistently demonstrated significant prognostic performance for distant recurrence-free interval (dRFI). The model stratified patients into clearly distinct low- and high-risk groups, with hazard ratios (HRs) ranging from 3.93 to 9.05. At 5 years, the probability of distant recurrence in low-risk patient groups ranged from 0.86% to 4.7%, compared to 6.39% to 34.8% in high-risk patient groups. This separation remained robust across key clinical subgroups, including histological grade 2 tumors, pre- and postmenopausal patients, and both node-negative and node-positive disease. RlapsRisk BC was also an independent prognostic factor in each cohort and improved prognostic performance when combined with standard clinical variables (age, tumor size, nodal status). Across the three validation cohorts, the model increased the c-index value by 0, 0.19, and 0.20, respectively, compared to clinical variables alone, with the improvement being statistically significant in two out of three cohorts.

Compared to established genomic assays, RlapsRisk BC demonstrated complementary—and in some cases superior—performance, particularly in identifying patients at low risk of distant recurrence. At equivalent specificity levels, RlapsRisk BC achieved higher sensitivity: 0.85 vs. 0.33 compared to Oncotype DX, and 0.74 vs. 0.49 compared to EndoPredict, highlighting its enhanced ability to correctly identify both low- and high-risk patients.

Conclusion:
RlapsRisk BC demonstrates robust, independent prognostic value and may offer a scalable, accessible complement to genomic assays. Further studies are warranted to confirm clinical utility and support its integration into treatment decision-making.

Introduction

Estrogen receptor–positive (ER+), human epidermal growth factor receptor 2–negative (HER2−) breast cancer accounts for approximately 70% of early-stage breast cancer (eBC) cases and encompasses a heterogeneous group of tumors with varying risks of recurrence. While most patients benefit from endocrine therapy and have favorable short-term outcomes, particularly those with luminal A–like tumors, a substantial subset remains at risk for late distant relapse, often occurring more than a decade after initial treatment [1]. Accurately identifying patients at high or low risk of recurrence is critical to optimizing adjuvant treatment, avoiding unnecessary chemotherapy, and considering extended or intensified endocrine strategies when appropriate [2].

Risk stratification is currently guided by clinicopathologic features, integrated into tools such as the Nottingham Prognostic Index and PredictBreast [3], and increasingly supported by molecular assays like Oncotype DX, MammaPrint, EndoPredict, and Prosigna. While these multigene assays improve prognostic precision, they are associated with practical limitations, including cost, accessibility, and turnaround time, that may restrict their widespread use. Consequently, there is a need for scalable alternatives that maintain or exceed the performance of existing methods.

In a previous proof-of-concept study, we introduced RlapsRisk BC, a multimodal deep learning–based prognostic tool that integrates features extracted from hematoxylin and eosin (H&E)–stained whole-slide images (WSIs) with routine clinical data to estimate the risk of distant recurrence in ER+/HER2− eBC [4]. Initial findings demonstrated that image-based models can capture global biologically meaningful prognostic information ascertained from the entire tissue section and offer a viable path toward accessible, inherent pathology-based biomarkers.

In the present study, we expand upon that work by conducting a large-scale, blinded, retrospective validation of RlapsRisk BC in three independent, international cohorts comprising patients treated with endocrine therapy alone, a clinically important subgroup for considering the potential addition of chemotherapy. In addition to assessing the model's standalone prognostic performance, we directly compare RlapsRisk BC with some of the commercially available genomic assays, to evaluate whether this AI-based approach can match or complement the current molecular gold standards in identifying patients who may safely avoid chemotherapy.

Methods

Study Design & Setting

This study reports the blind external evaluation of RlapsRisk BC, a multimodal deep learning–based prognostic model developed to estimate the risk of distant recurrence in ER+/HER2− early breast cancer. The model was trained to directly predict dRFI using retrospective multicenter data combining digitized hematoxylin and eosin (H&E) stained slides with routinely collected clinicopathologic variables (age, tumor size, and number of invaded lymph nodes).

We performed a blinded evaluation of RlapsRisk BC across three independent retrospective cohorts: BARTS (UK), MDA (USA), and EP-GR (France). These cohorts were not involved in the model training and were selected to ensure geographic, institutional, and population diversity.

The analysis focused on patients who received endocrine therapy alone for five years following surgery. This targeted subgroup allows for a clearer interpretation of the model's prognostic performance, without confounding effects from adjuvant chemotherapy or other targeted therapies.

The primary endpoint was distant recurrence-free interval (dRFI) at five years. Secondary analysis included stratification by clinical subgroups and comparison with existing clinical and genomic prognostic tools, where available.

Endpoint Definition

dRFI is defined as the time from initial diagnosis or definitive surgery (whichever occurs later) to the first occurrence of distant metastatic recurrence, with patients censored at the time of death from other causes or last follow-up if no distant event has occurred. This endpoint specifically captures the development of metastatic disease and excludes locoregional recurrence or contralateral breast cancer, thereby providing a focused measure of distant disease control.

AI Model Development

A first version of the RlapsRisk BC was designed to predict 5-year distant relapse directly from hematoxylin, eosin, and saffron (HES) stained WSIs, without requiring pathologist annotations [4]. In the present study, a new architecture and training procedures are used, where optimized losses were introduced and a more performant and robust feature extractor implemented. The neural network produces a digital pathology score from WSIs, which is then combined with three clinicopathological variables – age, number of invaded lymph nodes, and tumor size – using a Weibull survival model to generate RlapsRisk BC score, a prognostic indicator of dRFI.

A large dataset of histopathological whole-slide images (WSIs) paired with clinical and survival data from ER+/HER2− eBC patients diagnosed between 2000 and 2019 was collected, curated and leveraged to develop this new version. A total of 6,039 patients were included for model development (training and internal evaluation) (**Figure 1**). The goal was to train a multimodal deep learning model to predict dRFI directly from WSIs and standard prognostic clinical variables. Model training was performed using data from the GrandTMA [5] (n=1,881), CANTO [6] (n=1,704), CYPATH (n=199), and CARDIFF (n=876) cohorts. Cross-validation was used on these datasets for internal evaluation and model's parameters selection, 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. For inference on the external cohorts, predictions were generated by ensembling all 75 trained models through output averaging.

Internal performance assessment was conducted using three retrospective cohorts: TCGA-BRCA (n = 764), GRONINGEN (n = 258), and CURIE (n = 357). Prognostic performance was evaluated through stratified bootstrap evaluation (N=5,000). Full details on cohort characteristics, model architecture, and training procedures are provided in the Supplementary Materials.

Study population

All patients underwent surgical resection as initial treatment. Eligible patients had confirmed ER+/HER2- eBC. Follow-up data and at least one digitized H&E–stained WSI of the primary tumor were required for inclusion; when available, the diagnostic slide was used. Clinical variables necessary for risk computation -age, tumor size, and number of invaded lymph nodes- had to be available. A detailed inclusion and exclusion flow diagram for the evaluation cohorts is provided in **Supplementary Figure 1**.

Clinical performance validation was conducted on three independent external cohorts: MDA (n=154), BARTS (n=401), and EP-GR (n=381). Model performance was assessed on each individual cohort and all cohorts aggregated referred to as POOLED results (n=927).

Additionally, molecular signature test results were available in three cohorts: EndoPredict (EPclin) in EP-GR, Oncotype DX (RS) in MDA as well as in OD-GR, another independent cohort (n=161) – notably, the later cohort is the only one that does not include follow-up data.

The evaluation mainly focuses on patients that received endocrine therapy alone. Full details on cohort characteristics are provided in the Supplementary Materials.

Statistical Analyses

The primary endpoint was dRFI, defined per STEEP criteria [7] as the time from breast cancer diagnosis to the first distant recurrence or death from breast cancer. Outcomes such as local–regional recurrence, ipsilateral or contralateral ductal carcinoma in situ (DCIS), invasive contralateral breast cancer, and second non-breast primaries were not counted as events.

To assess the prognostic performance of the model, the inverse probability of censoring weighted (IPCW) concordance-index [17] (hereafter referred as C-index) from scikit-survival to address right-censoring and the AUC were used, computed with the scikit-survival package (0.19.0). P-values for comparing AUCs were computed using the method proposed by Blanche et al. [8], implemented as the compare function in the timeROC R package. To stratify patients into high- and low-risk groups, we used the threshold corresponding to a 5% probability of relapse computed from the training cohort as a cutoff and applied the same threshold across all validation cohorts. This 5% dRFI rate threshold corresponds to the 5-year interpolation of an exponential model from the 10-year dRFI of 10%, which is the most common output of the molecular signatures currently used in clinical practice [9]. Kaplan–Meier analysis and the log-rank test were used to compare survival distributions between risk groups. Multivariable analyses were performed using the Cox proportional hazards model to evaluate whether the model provided independent prognostic value beyond the clinical variables. We also used classification metrics including sensitivity, specificity, positive and negative predictive value (PPV and NPV) for comparison with clinical scores, implemented using the R package timeROC. These metrics were computed assuming a positive event corresponds to a distant recurrence occurring within 5 years, as defined by the dRFI endpoint. The Wald chi-square test was used to assess differences across variable groups, and hazard ratios (HRs) were reported to quantify the impact of each variable on survival outcomes.

Other prognostic methods used for comparison

The clinical and molecular signature tests included in our analysis were implemented and thresholded according to established methodologies from the literature.

The Clinical Treatment Score post-5 years (CTS5) was calculated using a Python implementation of the published formula, with patients classified as high risk if their CTS5 score was ≥3.86, following Dowsett et al. (J Clin Oncol, 2018) [10].

For the PREDICT Breast UK model (version 3), we used the official R implementation available at https://github.com/WintonCentre/predictv30r, adapted to our computational environment.

The Oncotype DX test results were obtained from certified clinical laboratories; patients were considered high risk based on standard thresholds: a recurrence score (RS) ≥25 for postmenopausal patients and ≥16 for premenopausal patients, as defined in the TAILORx trial [11]. For the analysis of the restratification of intermediary risk of RS, we also applied the intermediate recurrence score (RS) category of 11–25 for Oncotype DX, as defined in the RxPONDER trial [12].

For EndoPredictCLIN (EPclin), we used the validated cutoff of 3.86 to define the high risk subgroup [13] .

IRB approval

In each country, the ethical approvals or waivers were obtained from the relevant institutional review boards or ethics committees for each participating study. These retrospective studies were conducted across sites with varying regulatory and ethical requirements. Informed

consent was obtained from patients where required by local regulations; in some studies, consent was waived due to the retrospective nature of the data.

Results

Internal Evaluation

In the internal validation, RlapsRisk BC demonstrated consistent prognostic performance across the three internal cohorts. The concordance indices were 0.66 for TCGA-BRCA, 0.70 for CURIE, and 0.72 for GRONINGEN, indicating good discriminative ability. Stratification into high- and low-risk groups yielded hazard ratios of 1.82 (95% CI, 1.19–2.80) in TCGA-BRCA, 2.71 (95% CI, 1.56–4.72) in CURIE, and 1.98 (95% CI, 1.04–3.76) in GRONINGEN, supporting the reproducibility of RlapsRisk BC's risk predictions across independent internal datasets after models' parameters selection.

Cohort characteristics

To evaluate the prognostic performance of the RlapsRisk BC model, we performed an external validation using three independent retrospective cohorts with available outcome data, including a total of 927 patients with ER+/HER2– eBC: MDA (n = 154), BARTS (n = 401), and EP-GR (n = 381). Among them, 593 patients received adjuvant endocrine therapy alone. A fourth cohort, OGDR (n = 267), for which no follow-up data were available, was included separately only for further concordance assessment between RlapsRisk BC and Oncotype DX. The combined dataset of the three cohorts with survival data is referred to as the POOLED cohort. Molecular signature test results were available for EndoPredict in the EP-GR cohort, and for Oncotype DX in both the MDA and OD-GR cohorts. Summary statistics for full cohorts, including patients who received chemotherapy, are provided in Supplementary Material (**Table S1-S2**). The cohorts' characteristics presented in this section focus exclusively on patients treated with five years of adjuvant endocrine therapy alone; characteristics of chemotherapy-treated patients are reported in the Supplementary Materials.

In the BARTS cohort (n = 401), 262 patients (65.3%) received adjuvant endocrine therapy without chemotherapy. The mean age in this subgroup was 60.4 years (SD 11.5), and 74.0% of patients were postmenopausal. The mean tumor size was 22.5 mm (SD 13.2). 79.8% of cases had no lymph node involvement. Histological grade was predominantly low to intermediate: 19.5% of tumors were grade 1 and 65.6% were grade 2. During a median follow-up of 90.2 months (SD 34.0), 21 dRFI events occurred in this subgroup (recurrence rate: 8%).

The MDA cohort included 154 node-negative patients, all treated with adjuvant endocrine therapy alone. It was specifically enriched with patients who experienced distant metastasis within 5 years, each matched to a patient who remained metastasis-free during the same period. The mean age was 53.7 years (SD 10.1), and 55.8% of patients were premenopausal. The average tumor size was 19.6 mm (SD 12.3). Histological grading

showed 64.3% grade 1 and 32.5% grade 2 tumors. Over a median follow-up of 74.7 months (SD 52.0), 42 dRFI events were recorded (27.3%).

In the EP-GR cohort (n = 381), 180 patients (47.2%) were treated with endocrine therapy alone. Their mean age was 59.1 years (SD 10.5), and 75% were postmenopausal. The average tumor size was 18.7 mm (SD 7.5). 78.9% of patients were node-negative. Histological grades were distributed as follows: 6.7% grade 1, 80% grade 2, and 13.3% grade 3. The mean follow-up duration was 59.4 months (SD 17.6), and 5 dRFI events were observed (recurrence rate: 2.8%).

The OD-GR cohort (n = 267), for which follow-up data were not available, was used exclusively for a concordance analysis with Oncotype DX. Among these patients, 161 (60.2%) received endocrine therapy alone. Their mean age was 58.8 years (SD 7.8), and 81.4% were postmenopausal. The mean tumor size in this subgroup was 21.6 mm (SD 10.5). Node-negative disease was present in 36.6% of patients. Tumor grade was predominantly intermediate (77.6% grade 2, 13.0% grade 3).

All results presented in the main text refer to patients who received five years of adjuvant endocrine therapy alone and had sufficient follow-up, except for the final paragraph of the Results section, which reports findings in patients who also received chemotherapy. We focus on patients treated with endocrine therapy alone to ensure a clinically meaningful evaluation of the prognostic score. This approach enables a clear assessment of the score's ability to identify truly low-risk individuals who may safely forgo chemotherapy. For completeness, we also report results from the combined external validation cohorts,referred to as the POOLED dataset.

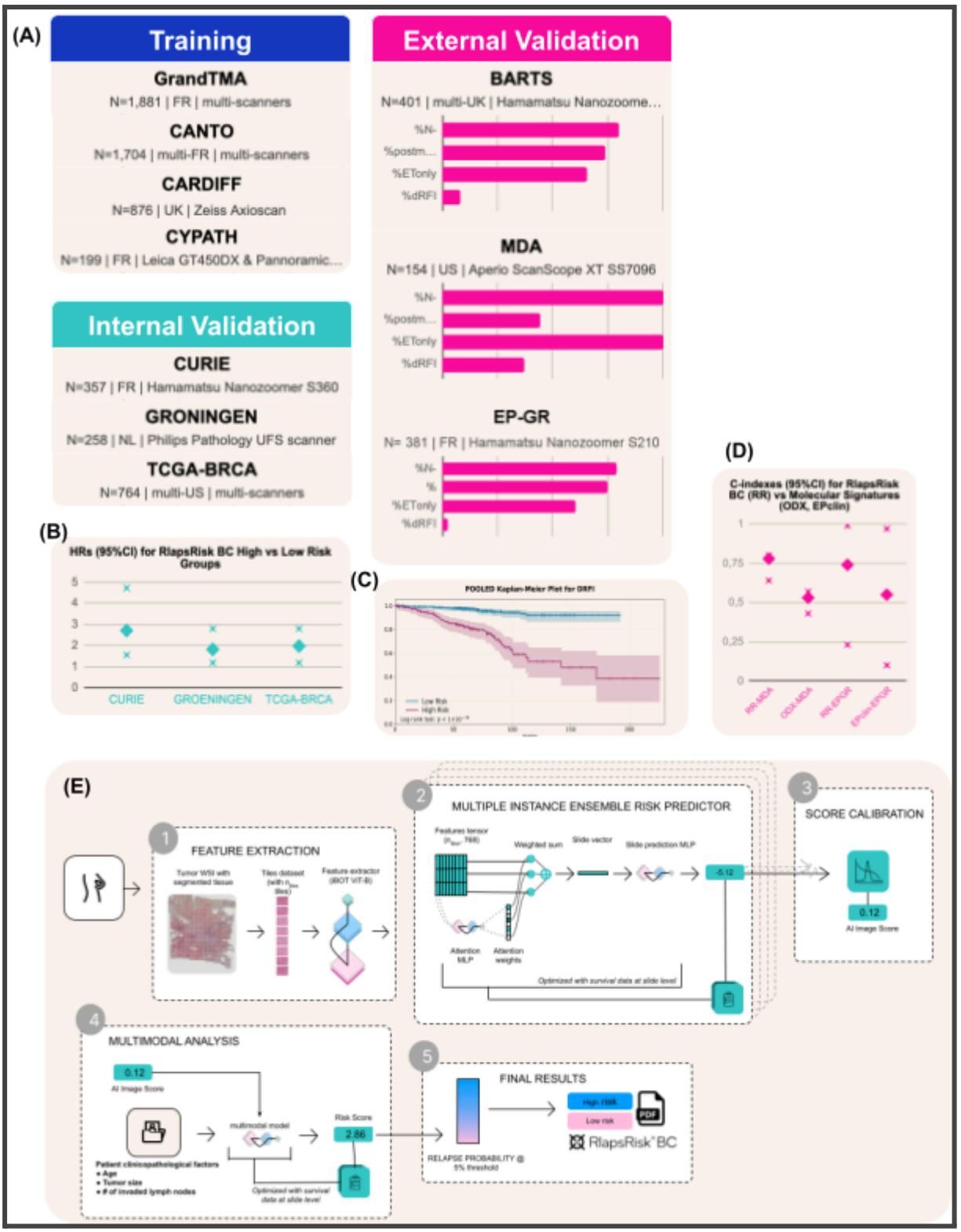

Figure 1: Schematic overview of the RlapsRisk BC pipeline cohorts, and results. A large dataset of histopathological whole-slide images (WSIs) paired with clinical and survival data from ER+/HER2− early breast cancer (eBC) patients diagnosed between 2000 and 2019 was collected and curated. A total of 6,039 patients were used for model development (training and internal validation), and 936 patients for external validation. **(A)** Cohorts overview: training (blue), internal validation (teal), and external validation (pink), with sample sizes, countries, scanner models, and key clinical characteristics (percentages of:

node-negative patients, post-menopausal patients, patients treated with endocrine therapy only, dRFI events). **(B)** Hazard ratios (HRs) comparing RlapsRisk BC high- vs. low-risk groups in internal validation sets. **(C)** Kaplan–Meier survival curve showing RlapsRisk BC stratification for dRFI on pooled external validation cohorts for patients treated with endocrine therapy only (N=591). **(D)** C-indices of RlapsRisk BC versus existing molecular scores for patients treated with endocrine therapy only on MDA (Oncotype Dx, N=154) and EP-GR (EPclin, N=180). **(E)** RlapsRisk BC's pipeline scheme: WSI-based feature extraction, multiple instance learning risk prediction, score calibration, integration with clinical variables, and final risk stratification.

Prognosis performance

Across the three independent external validation cohorts, RlapsRisk BC demonstrated a significant stratification of dRFI between low- and high-risk subgroups (P < .0005; **Figure 2**). The model achieved a concordance index (c-index) of 0.70 (95% CI, 0.56–0.78) with a hazard ratio (HR) of 3.93 (95% CI, 1.65–9.33) in the BARTS cohort; a c-index of 0.74 (95% CI, 0.62–0.81) and HR of 7.53 (95% CI, 3.34–16.9) in the MDA cohort; and a c-index of 0.79 (95% CI, 0.56–0.94) and HR of 9.05 (95% CI, 0.99–82.4) in the EP-GR cohort. In the pooled analysis, performance reached a c-index of 0.78 (95% CI, 0.69–0.81) and HR of 7.45 (95% CI, 4.30–12.85).

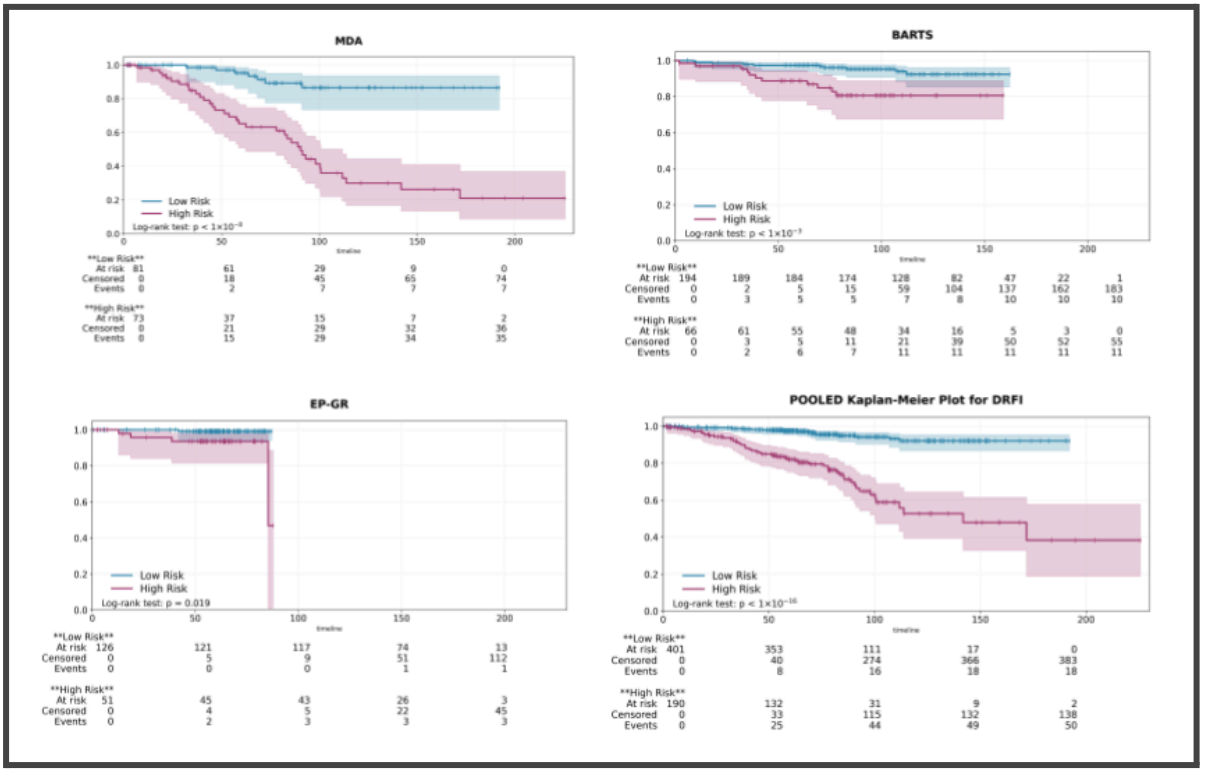

Figure 2: Kaplan-Meier curves for the dRFI endpoint in HR+/HER2- eBC treated with endocrine therapy only, for the high and low-risk subgroups provided by RlapsRisk BC. The upper left panel corresponds to the MDA cohort, the upper right panel corresponds to the BARTS cohort, the lower left panel corresponds to the EP-GR cohort, and the lower right panel corresponds to the POOLED cohort.

At five years, the predicted low-risk subgroups exhibited low probabilities of dRFI events, always below a value of 5%: 2.61% (95% CI, 1.0%–6.0%) for the BARTS cohort, 4.7% (95% CI, 1.5%–14.0%) for the MDA cohort, 0.86% (95% CI, 0.12%–5.8%) for the EP-GR cohort, and 2.4% (95% CI, 1.26%–4.58%) in the pooled analysis. In contrast, the high-risk subgroups showed substantially higher 5-year dRFI event probabilities: 11.2% (95% CI, 5.5%–22.1%) for BARTS, 34.8% (95% CI, 23.0%–49.0%) for MDA, 6.39% (95% CI, 0.3%–18.5%) for EP-GR, and 17.9% (95% CI, 12.7%–24.7%) in the POOLED cohort.

When restricting the analysis to multiple clinically relevant subgroups, the difference in terms of dRFI events between the high- and low-risk groups remained statistically significant ($P \leq 0.005$, **Supplementary Figure 2**), notably for histological grade 2 patients, pre- and post-menopausal status, as well as N0/N+ (dRFI event rates respectively for BARTS, MDA, EP-GR, POOLED in % for Grade 2: 7.6%, 29.2%, 2.8%; pre-menopausal: 13.2%, 21.3%, 2.3%, 13.2%; post-menopausal: 7.3%, 31.2%, 3.0%, 11.4%; N0: 6.3%, 27.3%, 2.9%, 11.8%; N+: 11.4%, NA, 100%, 18.8%). The model demonstrated consistent prognostic value across these clinically relevant subgroups, including those with inherently greater risk estimation challenges. In the POOLED cohort, HRs across subgroups (**Figure 3**, left panel) ranged from 1.13 (95% CI, 0.17 to 2.42) to higher values, with an average HR of 2.00 (95% CI, 1.84 to 2.15). The concordance index (c-index) across subgroups ranged from 0.63 (95% CI, 0.46 to 0.81) to 0.93 (95% CI, 0.86 to 1.00), with an average c-index of 0.79 (95% CI, 0.77 to 0.81), supporting the model's robust discriminatory performance across diverse clinical subgroups (**Figure 3**, right panel).

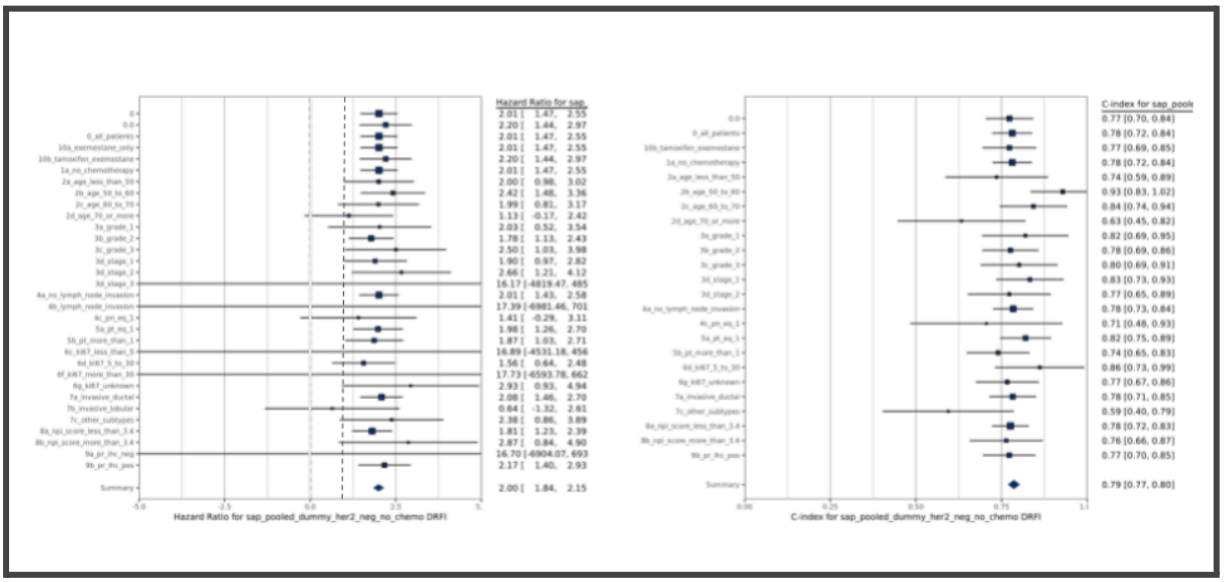

Figure 3: Prognostic performance of RlapsRisk BC across clinical subgroups in the pooled external validation cohort.
This figure presents two panels evaluating the prognostic value of the RlapsRisk BC score in predefined subgroups of patients with hormone receptor–positive, HER2-negative ($HR^+/HER2^-$) breast cancer.
The left panel shows a forest plot of the hazard ratio (HR) for distant recurrence-free interval (dRFI) comparing high- versus low-risk groups, also with 95% CIs, derived from univariable Cox proportional hazards models. The right panel shows a forest plot of RlapsRisk BC c-index when predicting distant recurrence-free interval (dRFI). Squares represent point

estimates, with their size proportional to the sample size of each subgroup and horizontal segments indicate 95% confidence intervals.

Comparison with molecular and clinical prognostic scores

Clinical factors

We found that the AI-derived score generated from Whole Slide Images (WSIs) provides independent prognostic information beyond established clinical variables, including histological grade, patient age, lymph node involvement, tumor size, and Ki-67 expression – although Ki-67 data were not available in the BARTS cohorts, missing values were replaced with the median values computed from the other cohorts ($P \leq 0.005$; **Supplementary Figure 3**). In assessing model performance, we observed that integrating the AI-based pathology score with key clinical variables (tumor size, number of invaded lymph node, age) significantly enhanced the prediction of 5-year distant recurrence-free interval (dRFI) in the MDA cohort, yielding a c-index of 0.74 with RlapsRisk BC (95% CI: 0.63–0.76), compared to 0.55 (95% CI: 0.48–0.66) for the clinical factors alone ($P = 0.013$; **Supplementary Table 3**). In the BARTS cohort, the predictive performance of RlapsRisk BC and the clinical factors alone were comparable, with c-indices of 0.70 (95% CI: 0.58–0.80) and 0.70 (95% CI: 0.60–0.78), respectively. In the EP-GR cohort, RlapsRisk BC achieved a c-index of 0.79 (95% CI: 0.56–0.94), while the clinical score attained a c-index of 0.59 (95% CI: 0.40–0.81). When evaluated on the POOLED dataset, the RlapsRisk BC model demonstrated superior performance, with a c-index of 0.78 (95% CI: 0.69–0.82), compared to 0.54 (95% CI: 0.50–0.63) for the clinical factors alone.

Comparison with established prognostic clinical models revealed a notable difference in prognostic performance. On the POOLED cohort, the RlapsRisk BC model achieved a c-index of 0.78 (95% CI: 0.70–0.80), outperforming both the CTS5 score, which reached a c-index of 0.60 (95% CI: 0.51–0.63), and PredictBreast, which yielded a c-index of 0.59 (95% CI: 0.44–0.59). RlapsRisk BC also outperformed clinical factors alone across multiple clinically relevant subgroups, including tumor grade, menopausal status, and nodal involvement (see **Supplementary Table 3** for details).

In the POOLED cohort, RlapsRisk BC demonstrated improved classification performance compared to CTS5. Specifically, RlapsRisk BC achieved a positive predictive value (PPV) of 0.183 and a negative predictive value (NPV) of 0.976, whereas CTS5 yielded a PPV of 0.156 and an NPV of 0.936. In terms of sensitivity and specificity, RlapsRisk BC provided a more balanced performance, with both sensitivity and specificity at 0.74, in contrast to CTS5, which showed a markedly lower sensitivity of 0.17 and a higher specificity of 0.92.

Molecular Signatures

We next compared the Oncotype DX Recurrence Score (RS) and the RlapsRisk BC score within the MDA cohort, where RS data were available. The two scores showed no significant correlation ($r = 0.15$, $p = 0.072$), and concordance in risk classification was modest at 56%, with high-risk RS patients defined using the thresholds defined in the Methods section (**Supplementary Figure 4**, upper panel). Similarly, in the OD-GR cohort, the scores were not correlated ($r = 0.15$, $p = 0.086$), with a concordance of 73%, higher than from the MDA cohort (**Supplementary Figure 4**, lower panel).

We compared the prognostic performance of RlapsRisk BC with that of the Oncotype DX (RS) classification for predicting 5-year dRFI in the MDA cohort, using the thresholds defined in the Methods section. C-index for continuous scores was significantly superior for RlapsRisk BC (0.78, 95% CI: 0.64–0.80) compared to RS (0.53, 95% CI: 0.43–0.57). Kaplan–Meier analysis demonstrated that both scores stratified patients into distinct risk groups; however, RlapsRisk BC provided a more pronounced separation between high- and low-risk patients (HR = 7.54; 95% CI, 3.34–16.90) compared to RS (HR = 1.01; 95% CI, 0.97–1.05) (**Figure 4** upper panel). RlapsRisk BC achieved a sensitivity of 0.85 and a specificity of 0.63, while RS reached a sensitivity of 0.33 and a specificity of 0.60. Positive predictive value (PPV) was at 0.34 for RlapsRisk and 0.16 for RS, yet both achieved high negative predictive values (NPV) of 0.95 and 0.80, respectively (**Supplementary Table 4**).

To assess the potential of RlapsRisk BC to refine risk assessment in patients classified as intermediate risk by Oncotype DX, we conducted a focused analysis within this subgroup. RlapsRisk BC further stratified these patients into two groups with significantly different outcomes (**Figure 4** lower panel, HR=10.5 (3.1–34.2), p< 0.005).

In the EP-GR cohort, we compared RlapsRisk BC to EPclin scores in 180 patients treated with endocrine therapy only where both scores were available. The two scores were significantly correlated with a correlation coefficient r of 0.34 (p<0.05). The proportion of patients classified as high-risk was similar between models: RlapsRisk BC identified 28% of patients as high-risk, compared to 31% for EPclin (**Supplementary Figure 5**). Yet RlapsRisk BC demonstrated superior discriminative performance with a c-index of 0.74 (95% CI, 0.23–0.99) versus 0.55 (95% CI, 0.10–0.97) for EPclin.

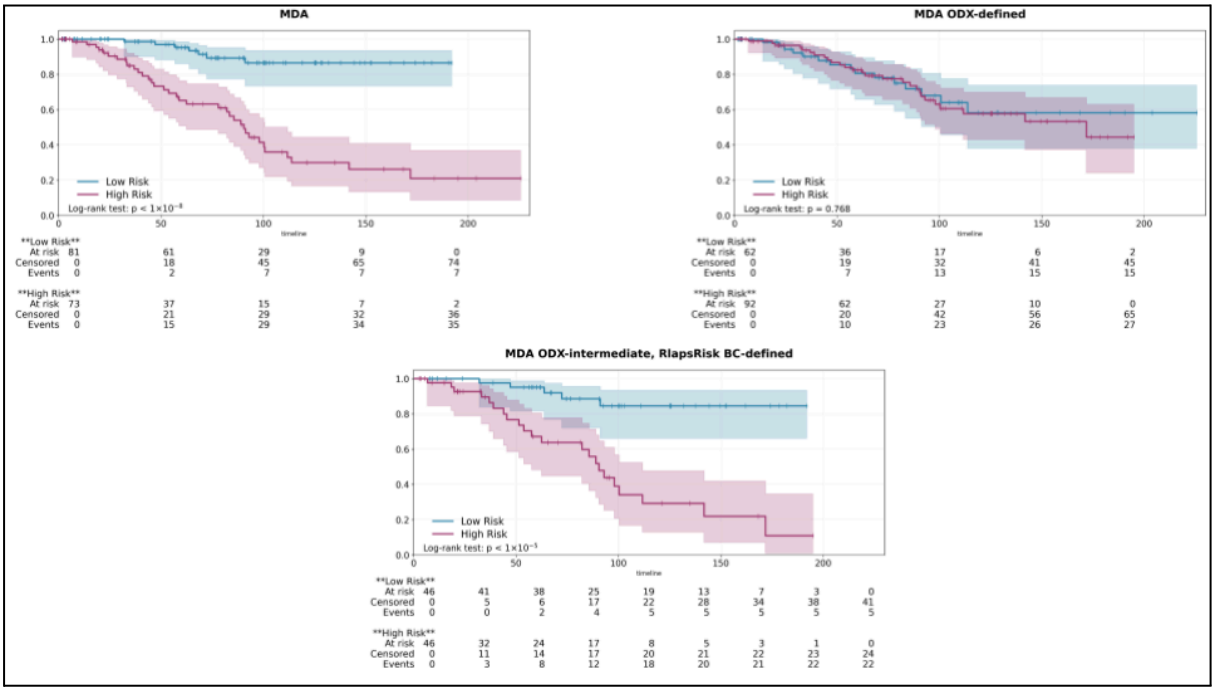

Figure 4: Kaplan-Meier curves for the dRFI endpoint in HR+/HER2- eBC treated with endocrine therapy only in the MDA cohort.
The upper left panel displays risk group stratification using RlapsRisk BC, while the upper right panel shows stratification based on the Oncotype DX Recurrence Score (RS).

The lower panel illustrates re-stratification of patients classified as intermediate-risk by RS using the RlapsRisk BC score.

Additionally, the two models differed in their classification performance.. RlapsRisk BC achieved a sensitivity of 0.74 and a specificity of 0.74, while EPclin reached a sensitivity of 0.49 and a specificity of 0.80. Positive predictive value (PPV) was modest for both scores, at 0.066 for RlapsRisk and 0.057 for EPclin, yet both achieved high negative predictive values (NPV) of 0.9915 and 0.9844, respectively (**Supplementary Table 4**).

Performance in Chemotherapy-Treated Patients

To further evaluate the robustness of RlapsRisk BC in a higher-risk clinical setting, we assessed its prognostic and classification performance in patients who received both endocrine therapy and chemotherapy within the POOLED cohort. In this population, RlapsRisk BC successfully stratified patients into two distinct risk groups with a statistically significant difference in distant recurrence risk (HR = 2.10; p = 0.037; **Supplementary Figure 6**). At 5 years, the incidence of distant recurrence was 2.7% (95% CI, 1.14%–6.3%) in the low-risk group compared to 7.0% (95% CI, 3.76%–13.25%) in the high-risk group. The overall prognostic performance of RlapsRisk BC was comparable to that of established clinical scores, with a C-index of 0.711 (95% CI, 0.60–0.84), compared to 0.72 (95% CI, 0.76–0.92) for CTS5 and 0.68 (95% CI, 0.63–0.85) for the pathology-based (PB) score.

In terms of classification performance, RlapsRisk BC achieved a sensitivity of 0.64 (± 0.12) and a specificity of 0.59 (± 0.03), while CTS5 showed a sensitivity of 0.56 (± 0.13) and a specificity of 0.69 (± 0.03). Positive predictive values were similarly low for both scores, with 0.071 (± 0.02) for RlapsRisk BC and 0.081 (± 0.02) for CTS5. Negative predictive values were high and nearly identical, at 0.972 (± 0.01) for RlapsRisk BC and 0.971 (± 0.01) for CTS5.

Discussion

In this study, we report the blind validation of RlapsRisk BC, an AI-based prognostic model that combines histopathological features exhibited in whole slide images with routinely available clinicopathologic variables (age, tumor size, and number of positive lymph nodes). The validation was performed across three independent retrospective cohorts of patients with ER-positive, HER2-negative early breast cancer, as well as a pooled analysis. The results reported in the main corpus focus specifically on patients treated with 5 years of adjuvant endocrine therapy alone, the standard baseline treatment in this population. This design provides a consistent and clinically relevant framework for interpreting the model's prognostic value and aligns with the real-world question of whether endocrine therapy alone is sufficient for a given patient.

Across all cohorts, RlapsRisk BC consistently demonstrated significant prognostic performance for dRFI. The model stratified patients into clearly distinct low- and high-risk groups, with hazard ratios (HRs) ranging from 3.93 to 9.05 across validation cohorts. At 5 years, the probability of distant recurrence in low-risk patients ranged from 0.86% to 4.7%, compared to 6.39% to 34.8% in high-risk patients. The separation of risk groups was consistent across key clinical subgroups, including histological grade 2 and pre-menopausal patients.

RlapsRisk BC integrates standard clinical variables that are robust and routinely collected, ensuring consistent applicability across healthcare settings. By combining these with digitized routinely available H&E slides, the model provides prognostic information that is complementary to traditional clinico-pathologic features. In two out of three cohorts, RlapsRisk BC achieved a superior c-index compared to the score based on clinical variables alone.

In comparative analyses with genomic assays, RlapsRisk BC further demonstrated clinical utility. It showed no correlation with the Oncotype DX recurrence score, indicating that it captures distinct prognostic signals. RlapsRisk BC provided additional granularity by reclassifying patients with intermediate Oncotype DX recurrence scores into groups with distinct outcomes. Compared to EndoPredict, it identified fewer patients as high-risk while maintaining a higher negative predictive value, suggesting improved ability to reassure patients with truly low risk of distant recurrence.

While these initial comparisons are promising, they should be interpreted with caution due to confounding bias in patients population, especially with regard to treatment decisions that might have been recommended based on the genomic assay results. Additional studies comparing RlapsRisk BC to molecular signatures will be important to precisely define its positioning within clinical decision workflows and to determine how best to use it as a surrogate or in conjunction to the currently used molecular assays. Several recent studies have further underscored the potential of histology-based AI models in risk stratification for ER+/HER2− early breast cancer. Lee et al. demonstrated that epithelial and stromal features extracted from H&E-stained slides could accurately predict high genomic risk, with strong validation on TCGA data [14]. The multimodal Ataraxis AI test, trained on over 8,000 patients from seven countries, reported a ~50% improvement in performance compared to standard

genomic assays in independent HR+ cohorts, showing that a transformer-based architecture can outperform Oncotype DX (C-index 0.67 vs. 0.61) [15], while the Orpheus model achieved an AUC of 0.89 in identifying patients with recurrence score (RS) >25, outperforming conventional clinicopathologic models [16]. These findings collectively highlight the growing maturity of AI-driven pathology models and support the feasibility of using AI based models using digitized histology as a surrogate for molecular assays.

These findings support the use of RlapsRisk BC as a decision-support tool to refine treatment pathways. By identifying patients at low risk of recurrence, the model may help avoid unnecessary chemotherapy or overtreatment. Conversely, it can help flag patients at higher risk who may benefit from therapeutic escalation. This risk-based personalization of treatment aligns with evolving clinical practice trends, which emphasize risk stratification to inform treatment pathways.

It is also worth noting that multigene signatures such as Oncotype DX or EndoPredict are currently recommended primarily for patients with ambiguous clinicopathologic profiles—their widespread adoption is often limited by cost, and implementation complexity. RlapsRisk BC, in contrast, leverages digitized H&E-stained pathology slides and clinical data, offering a scalable and accessible alternative that could broaden risk stratification capacity across healthcare systems. Its added value over clinical variables, combined with lower barriers to access, supports broad applicability across the ER+/HER2– population.

As with any machine learning–based model, the potential for bias and overfitting must be carefully considered. In the case of RlapsRisk BC, these risks were mitigated through rigorous design choices, including strict separation of training and validation datasets and evaluation across diverse, multi-institutional cohorts representative of real-world clinical settings. The consistent prognostic performance observed across all external validation cohorts supports the model’s robustness and generalizability.

Beyond prognostic performance, an important dimension for clinical integration is the interpretability of the model's predictions. Preliminary analyses in prior work showed that the AI model relies on morphological patterns associated with recurrence risk, including nuclear atypia, mitotic activity, and stromal characteristics [4]. A dedicated pathology-focused study is underway to deepen model interpretability by identifying image regions most relevant to predictions, linking them to known histological features, and potentially uncovering novel tissue-based biomarkers. These efforts aim to enhance biological understanding and strengthen confidence in the use of AI models in pathology workflows.

In summary, this blind validation study provides robust evidence that RlapsRisk BC delivers clinically actionable, and reproducible risk information in early-stage ER+/HER2– breast cancer. Its ability to refine patient stratification, outperform clinical models, and complement or substitute genomic assays, positions it as a valuable tool for guiding adjuvant treatment decisions. Further studies will be necessary to confirm its impact and optimize its integration into routine care. These results highlight the potential of AI-based pathology to support more personalized, data-driven treatment pathways for early ER+/HER2- breast cancer patients.

Data Sharing Statement

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 training 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 submitted to https://redcap.link/DataRequestClinicalTrialsGustaveRoussy for GrandTMA or directed to s-everhard@unicancer.fr for CANTO cohorts.

The datasets used in this study, including GrandTMA, CANTO, CYPATH, CARDIFF, GRONINGEN, CURIE, BARTS, MDA, EP-GR, and OD-GR, are proprietary data of the respective medical centers and institutions. Access to these datasets, or subsets thereof, may be available upon request and subject to ethical approvals and data use agreements with the originating centers. Interested researchers should contact the individual institutions for specific data access procedures.

Author Contributions

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

Acknowledgments

The authors thank the patients who participated and who did not oppose this additional research.

RlapsRisk BC is being developed under the guidelines of PortrAIt, a french consortium financed by the government within the framework of France 2030 and by the European Union - Next Generation EU within the framework of the France Relance Plan.

The authors 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.

Anonymised data were provided by the Wales Cancer Bank (DOI: http://doi.org/10.5334/ojb.46), which is supported by Health and Care Research Wales. It is possible that other researchers have also received materials from the same donors.

The TGCA-BRCA dataset cited in this study was generated by the TCGA Research Network: https://www.cancer.gov/tcga.

The authors gratefully acknowledge the Barts Cancer Institute Breast Biobank for their role in the collection and provision of tissue samples and associated data. We are especially thankful to the patients who generously donated their tissue and shared their data, enabling the completion of this work.

The authors gratefully acknowledge MD Anderson Cancer Centre for their role in the collection and provision of tissue samples and associated data.

COI Disclosures

V.G., Z.V., V.A., J.G, A.F, M.B, E.H, D.J, L.G, M.S are employees of Owkin Inc.

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.

All other co-authors declare no competing interests (D.D., C.B., L.J., B.V.D.V, C.M, L.K.S, A.-L. M., S.E, I.G, A.V-S, S.K

References

1. Pan H, Gray R, Braybrooke J, Davies C, Taylor C, McGale P, Peto R, Pritchard KI, Bergh J, Dowsett M, Hayes DF; EBCTCG. 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. N Engl J Med. 2017 Nov 9;377(19):1836-1846. doi: 10.1056/NEJMoa1701830. PMID: 29117498; PMCID: PMC5734609.

2. 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)

3. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, Zackrisson S, Senkus E; ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. Ann Oncol. 2019 Aug 1;30(8):1194-1220. doi: 10.1093/annonc/mdz173. Erratum in: Ann Oncol. 2019 Oct 1;30(10):1674. doi: 10.1093/annonc/mdz189. Erratum in: Ann Oncol. 2021 Feb;32(2):284. doi: 10.1016/j.annonc.2020.08.2158. PMID: 31161190.

4. Garberis, I., Gaury, V., Saillard, C. et al. Deep learning assessment of metastatic relapse risk from digitized breast cancer histological slides. Nat Commun 16, 5876 (2025). https://doi.org/10.1038/s41467-025-60824-z

5. 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).

6. 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).

7. Tolaney SM, Garrett-Mayer E, White J, et al. Updated Standardized Definitions for Efficacy End Points (STEEP) in Adjuvant Breast Cancer Clinical Trials: STEEP Version 2.0. Journal of Clinical Oncology. 2021;39(24):2720-2731. doi:https://doi.org/10.1200/JCO.20.03613

8. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Statistics in Medicine. 2013;32(30):5381-5397. doi:https://doi.org/10.1002/sim.5958

9. Klein, E., Kiechle, M., Josipovic, A. et al. Long-term prospective outcome data using EndoPredict as risk stratification and chemotherapy decision biomarker in hormone receptor-positive, HER2-negative early breast cancer. Breast Cancer Res Treat 207, 119–127 (2024). https://doi.org/10.1007/s10549-024-07346-2

10. Mitch Dowsett et al. Integration of Clinical Variables for the Prediction of Late Distant Recurrence in Patients With Estrogen Receptor–Positive Breast Cancer Treated With 5 Years of Endocrine Therapy: CTS5. JCO 36, 1941-1948(2018). doi:10.1200/JCO.2017.76.4258

11. 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)

12. Kalinsky, K. et al. 21-Gene Assay to Inform Chemotherapy Benefit in Node-Positive Breast Cancer. N. Engl. J. Med. 385, 2336-2347 (2021)

13. Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, Dietze O, Greil R, Jelen A, Sevelda P, Freibauer C, Müller V, Jänicke F, Schmidt M, Kölbl H, Rody A, Kaufmann M, Schroth W, Brauch H, Schwab M, Fritz P, Weber KE, Feder IS, Hennig G, Kronenwett R, Gehrmann M, Gnant M; EP Investigators. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res. 2011 Sep 15;17(18):6012-20. doi: 10.1158/1078-0432.CCR-11-0926. Epub 2011 Aug 1. PMID: 21807638.

14. Lee DN, Li Y, Olsson LT, Hamilton AM, Calhoun BC, Hoadley KA, Marron JS, Troester MA. Image analysis-based identification of high risk ER-positive, HER2-negative breast cancers. Breast Cancer Res. 2024 Dec 4;26(1):177. doi: 10.1186/s13058-024-01915-5. PMID: 39633505; PMCID: PMC11616316.

15. Jan Witowski et al. Clinical validation of a multi-modal Ataraxis AI platform for recurrence prediction in early-stage breast cancer across multiple patient cohorts. JCO 43, 549-549(2025). DOI:10.1200/JCO.2025.43.16_suppl.549

16. Boehm, K.M., El Nahhas, O.S.M., Marra, A. et al. Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. Nat Commun 16, 2106 (2025). https://doi.org/10.1038/s41467-025-57283-x

17. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data". Statistics in Medicine, 30(10), 1105–1117.

Tables and Figures

Tables

Variable	Characteristics	BARTS		MDA		EP-GR		POOLED	
		260	100,00%	154	100,00%	180	295,08%	591	100,00%
Age (in years)	[18; 50[	38	14,62%	61	39,61%	45	73,77%	143	24,20%
	50-59	83	31,92%	50	32,47%	42	68,85%	175	29,61%
	60-69	85	32,69%	33	21,43%	68	111,48%	185	31,30%
	>70	54	20,77%	10	6,49%	25	40,98%	88	14,89%
	mean ± SD	60.39 ± 11.34		53.67 ± 10.06		59.13 ± 10.60		58.25 ± 11.10	
	[min ; max]	[31.00 ; 86.00]		[28.00 ; 82.00]		[27.44 ; 86.72]		[27.44 ; 86.72]	
	median (5th, 95th)	60.00 (41.95, 81.00)		52.00 (38.00, 71.00)		60.67 (42.83, 76.09)		58.00 (41.03, 77.09)	
Histological Grade	Grade 1	50	19,23%	35	22,73%	12	19,67%	97	16,41%
	Grade 2	171	65,77%	89	57,79%	144	236,07%	402	68,02%
	Grade 3	39	15,00%	30	19,48%	24	39,34%	92	15,57%
Nodal status	N-	208	80,00%	154	100,00%	142	232,79%	501	84,77%
	N+	52	20,00%	0	0,00%	38	62,30%	90	15,23%
	pN1	44	16,92%		0,00%	31	50,82%	75	12,69%
Tumor Size	pT1	176	67,69%	99	64,29%	125	204,92%	399	67,51%
	pT2/3/4	84	32,31%	55	35,71%	55	90,16%	192	32,49%
	mean ± SD	22.49 ± 13.19		19.59 ± 12.29		18.71 ± 7.47		20.61 ± 11.63	
	[min ; max]	[6.00 ; 81.00]		[6.00 ; 90.00]		[6.00 ; 50.00]		[6.00 ; 90.00]	
	median (5th, 95th)	18.00 (9.00, 50.00)		17.00 (8.65, 38.70)		18.00 (10.00, 32.15)		18.00 (9.00, 45.00)	
dRFI events	Number of dRFI events	21	8,08%		0,00%	5	8,20%		0,00%
	Number of dRFI events at 5 years	12	4,62%	22	14,29%	4	6,56%	38	6,43%
	Number of dRFI events at 10 years	21	8,08%	40	25,97%	5	8,20%	66	11,17%
dRFI duration	mean ± SD	90.48 ± 33.95		74.69 ± 52.00		59.57 ± 17.68		77.11 ± 38.38	
	[min ; max]	[2.00 ; 162.00]		[1.64 ; 225.59]		[0.69 ; 87.67]		[0.69 ; 225.59]	
	median (5th, 95th)	88.00 (32.95, 148.05)		67.81 (4.85, 175.61)		61.51 (16.53, 82.57)		72.00 (11.75, 148.50)	
IDFS events	Number of IFFS events	55	21,15%		0,00%	8	13,11%		0,00%
	Number of IDFS events at 5 years	31	11,92%	24	15,58%	6	9,84%	61	10,32%
	Number of IDFS events at 10 years	50	19,23%	42	27,27%	8	13,11%	100	16,92%
IDFS duration	mean ± SD	90.33 ± 34.23		73.46 ± 51.46		59.57 ± 17.68		76.72 ± 38.30	
	[min ; max]	[2.00 ; 162.00]		[1.64 ; 225.59]		[0.69 ; 87.67]		[0.69 ; 225.59]	
	median (5th, 95th)	88.00 (27.95, 148.05)		67.05 (4.85, 175.61)		61.51 (16.53, 82.57)		71.61 (12.98, 147.70)	

Table 1: Characteristics of the three cohorts used to evaluate RlapsRisk BC performances, for the population treated with adjuvant endocrine therapy only. Detailed description of the cohorts stratified by treatment regimen can be found in the supplementary material.

Figures

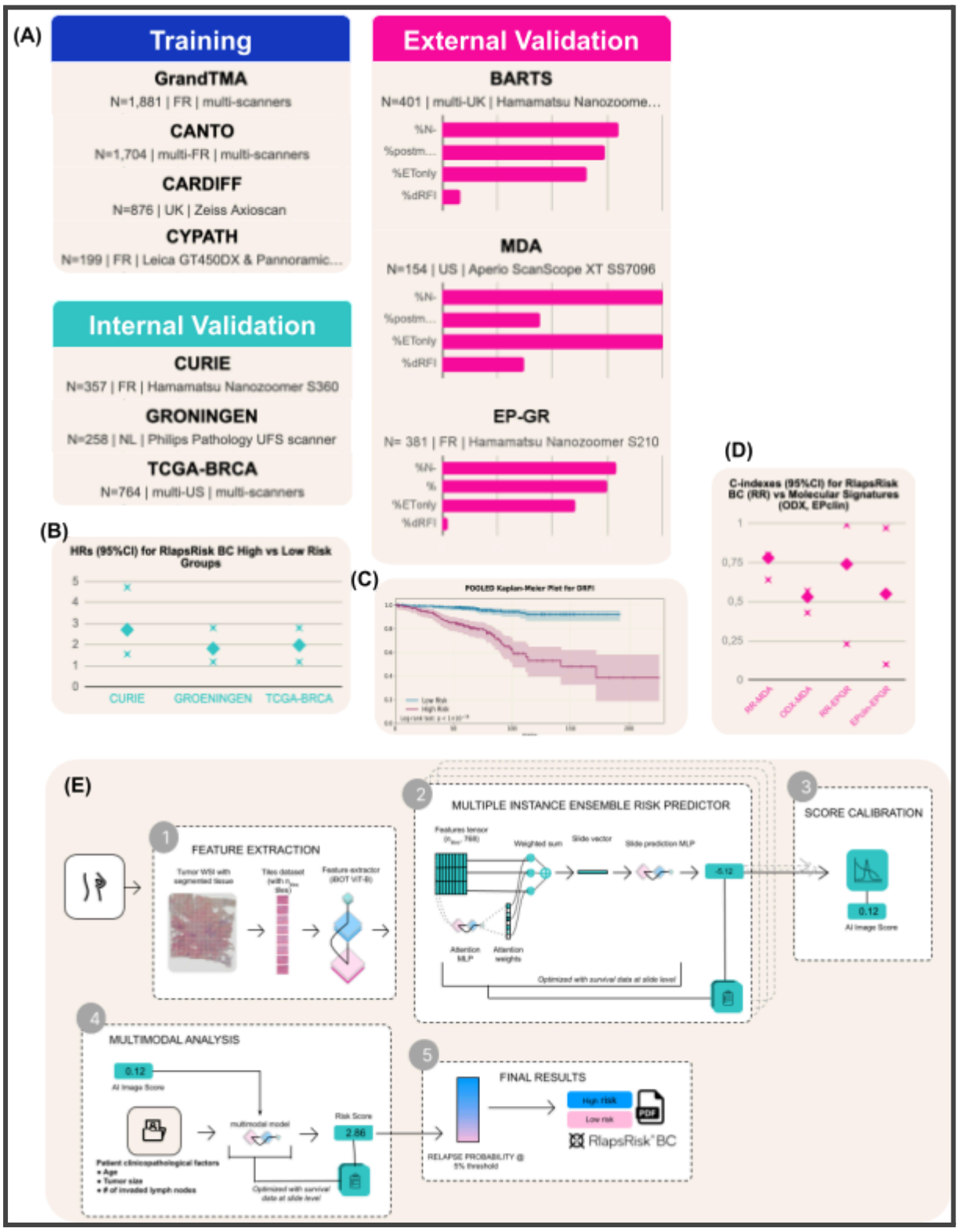
(A)
Training
GrandTMA
N=1,881 | FR | multi-scanners
CANTO
N=1,704 | multi-FR | multi-scanners
CARDIFF
N=876 | UK | Zeiss Axioscan
CYPATH
N=199 | FR | Leica GT450DX & Pannoramic...
Internal Validation
CURIE
N=357 | FR | Hamamatsu Nanozoomer S360
GRONINGEN
N=258 | NL | Philips Pathology UFS scanner
TCGA-BRCA
N=764 | multi-US | multi-scanners
External Validation
BARTS
N=401 | multi-UK | Hamamatsu Nanozoome...
%N-
%postm...
%ETonly
%dRFI
MDA
N=154 | US | Aperio ScanScope XT SS7096
%N-
%postm...
%ETonly
%dRFI
EP-GR
N= 381 | FR | Hamamatsu Nanozoomer S210
%N-
%
%ETonly
%dRFI
(B)
HRs (95%CI) for RlapsRisk BC High vs Low Risk Groups
5
4
3
2
1
0
CURIE
GROENINGEN
TCGA-BRCA
(C)
POOLED Kaplan-Meier Plot for DRFI
Low Risk
High Risk
(D)
C-indexes (95%CI) for RlapsRisk BC (RR) vs Molecular Signatures (ODX, EPclin)
1
0,75
0,5
0,25
0
RR-MDA
ODX-MDA
RR-EPGR
EPclin-EPGR
(E)
1
FEATURE EXTRACTION
Tumor WSI with segmented tissue
Tiles dataset
Feature-extractor (DINOT ViT-B)
2
MULTIPLE INSTANCE ENSEMBLE RISK PREDICTOR
Features tensor
Weighted sum
Slide vector
Slide prediction MLP
-5.12
Attention MLP
Attention weights
Optimized with survival data at slide level
3
SCORE CALIBRATION
0.12
AI Image Score
4
MULTIMODAL ANALYSIS
0.12
AI Image Score
multimodal model
Risk Score
2.86
Patient clinicopathological factors
• Age
• Tumor size
• # of invaded lymph nodes
Optimized with survival data at slide level
5
FINAL RESULTS
High risk
Low risk
PDF
RlapsRisk BC
RELAPSE PROBABILITY @ 5% threshold

Figure 1: Schematic overview of the RlapsRisk BC pipeline cohorts, and results. A large dataset of histopathological whole-slide images (WSIs) paired with clinical and survival data from ER+/HER2− early breast cancer (eBC) patients diagnosed between 2000 and 2019 was collected and curated. A total of 6,039 patients were used for model development (training and internal validation), and 936 patients for external validation. **(A)** Cohorts overview: training (blue), internal validation (teal), and external validation (pink), with sample sizes, countries, scanner models, and key clinical characteristics (percentages of: node-negative patients, post-menopausal patients, patients treated with endocrine therapy only, dRFI events). **(B)** Hazard ratios (HRs) comparing RlapsRisk BC high- vs. low-risk groups in internal validation sets. **(C)** Kaplan–Meier survival curve showing RlapsRisk BC stratification for dRFI on pooled external validation cohorts for patients treated with endocrine therapy only (N=591). **(D)** C-indices of RlapsRisk BC versus existing molecular scores for patients treated with endocrine therapy only on MDA (Oncotype Dx, N=154) and EP-GR (EPclin, N=180). **(E)** RlapsRisk BC's pipeline scheme: WSI-based feature extraction, multiple instance learning risk prediction, score calibration, integration with clinical variables, and final risk stratification.

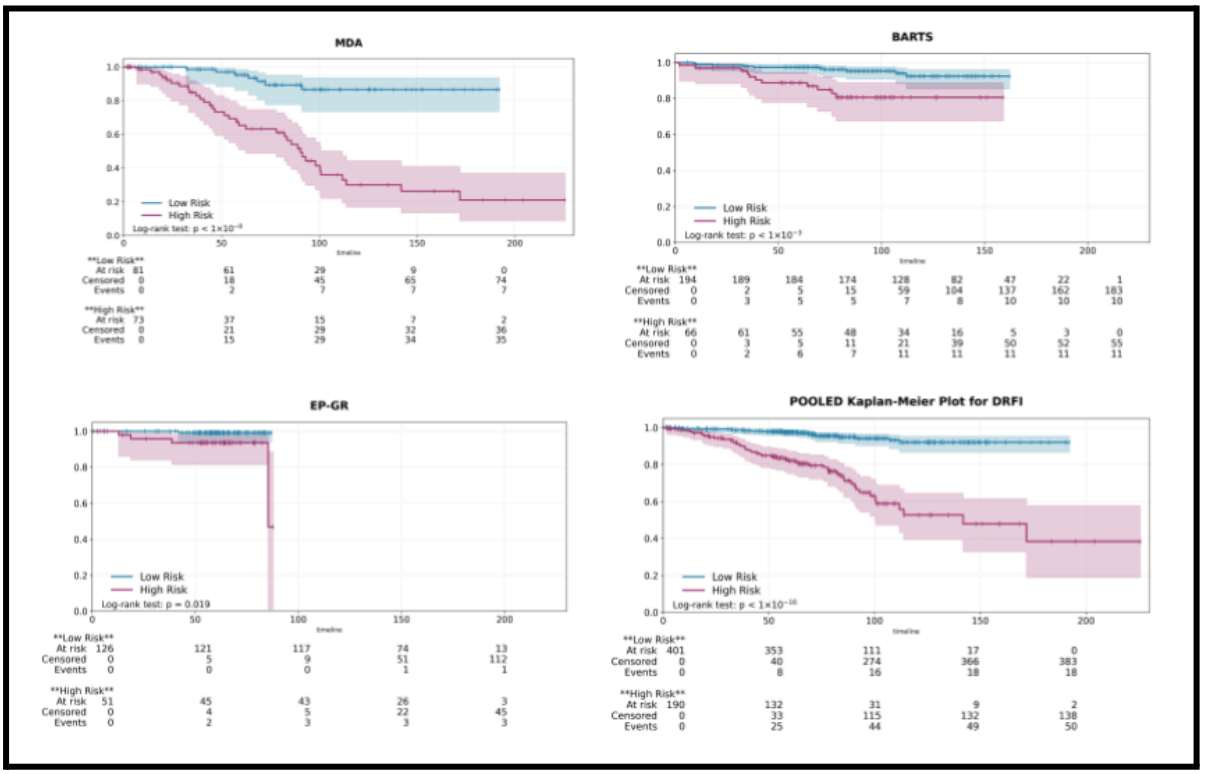

Figure 2: Kaplan-Meier curves for the dRFI endpoint in HR+/HER2- eBC treated with endocrine therapy only, for the high and low-risk subgroups provided by RlapsRisk BC. The upper left panel corresponds to the MDA cohort, the upper right panel corresponds to the BARTS cohort, the lower left panel corresponds to the EP-GR cohort, and the lower right panel corresponds to the POOLED cohort.

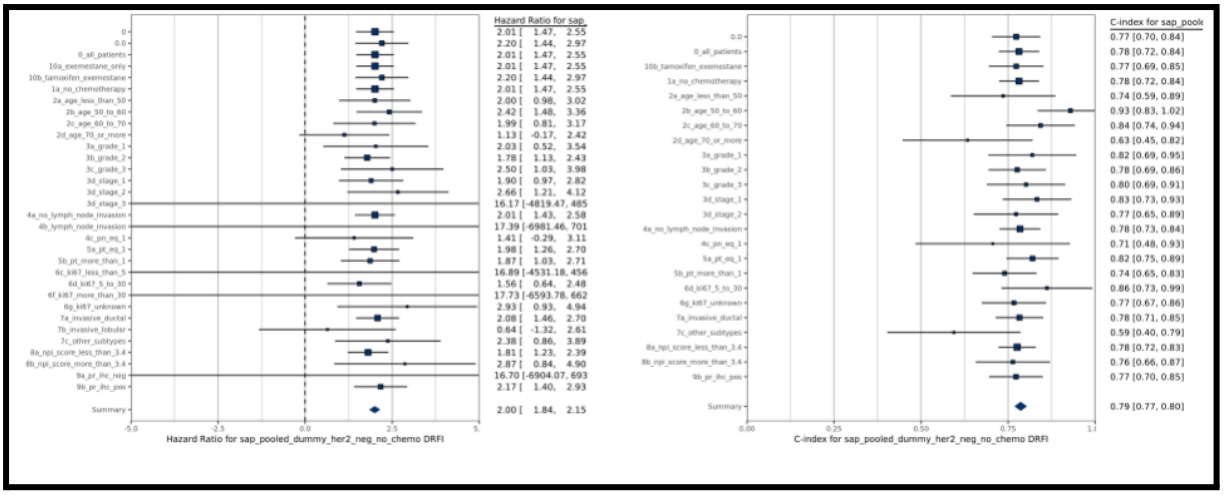

Figure 3: Prognostic performance of RlapsRisk BC across clinical subgroups in the pooled external validation cohort.

This figure presents two panels evaluating the prognostic value of the RlapsRisk BC score in predefined subgroups of patients with hormone receptor–positive, HER2-negative ($HR^+/HER2^-$) breast cancer.

The left panel shows a forest plot of the hazard ratio (HR) for distant recurrence-free interval (dRFI) comparing high- versus low-risk groups, also with 95% CIs, derived from univariate Cox proportional hazards models.

The right panel displays a forest plot of the concordance index (C-index) with corresponding 95% confidence intervals (CIs), indicating the discriminative accuracy of the score within each subgroup.

Subgroups analyzed include tumor size (pT1 vs. pT2+), nodal status (N0 vs. N+), histologic grade (I–II vs. III), age at diagnosis (<50 vs. ≥50 years), and receipt of adjuvant chemotherapy (yes vs. no). RlapsRisk BC demonstrated consistent prognostic value across these clinically relevant strata, with high C-index and significantly elevated HRs for high-risk patients across most subgroups.

Abbreviations: HR, hazard ratio; CI, confidence interval; C-index, concordance index; dRFI, distant recurrence-free interval; $HR^+/HER2^-$, hormone receptor–positive, HER2-negative.

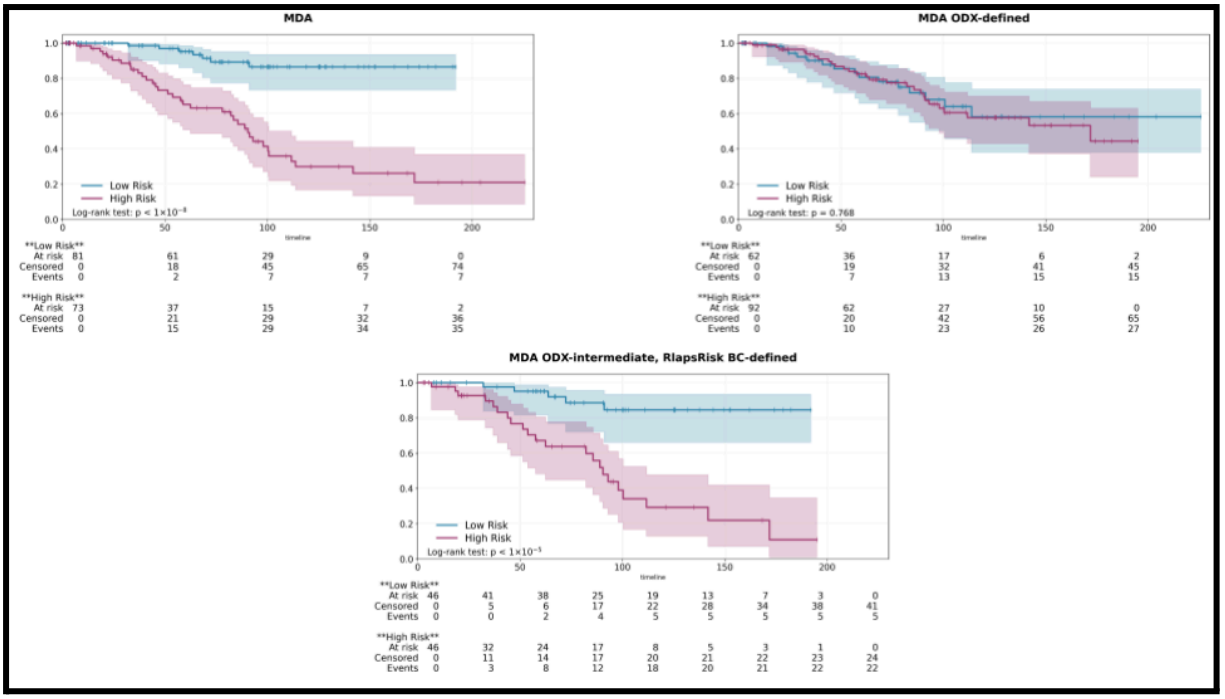

Figure 4: Kaplan-Meier curves for the dRFI endpoint in HR+/HER2- eBC treated with endocrine therapy only in the MDA cohort.
The upper left panel displays risk group stratification using RlapsRisk BC, while the upper right panel shows stratification based on the Oncotype DX Recurrence Score (RS).
The lower panel illustrates re-stratification of patients classified as intermediate-risk by RS using the RlapsRisk BC score.