Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Decoding Tumor Phenotypes: A Radiologist-Inspired Deep Learning Framework for Breast Cancer Recurrence Prediction

**Tao Tan**

taotanjs@gmail.com

The Netherlands Cancer Institute https://orcid.org/0000-0001-5403-0887

**Chunyao Lu**

The Netherlands Cancer Institute

**Tianyu Zhang**

Radboud University Medical Centre https://orcid.org/0009-0007-4724-719X

**Xinglong Liang**

The Netherlands Cancer Institute

**Antonio Portaluri**

Radboud University Medical Centre

**Luyi Han**

The Netherlands Cancer Institute https://orcid.org/0000-0003-4046-2763

**Yaqian Chen**

Duke University

**Nika Rasoolzadeh**

The Netherlands Cancer Institute

**Ruixiang Qi**

Affiliated Hangzhou First People's Hospital

**Yuan Gao**

Netherlands Cancer Institute https://orcid.org/0000-0001-6326-129X

**Xin Wang**

The Netherlands Cancer Institute https://orcid.org/0000-0001-9619-7503

**Yaofei Duan**

Macau Polytechnic University https://orcid.org/0009-0007-2562-5794

**Zahra Aghdam**

The Netherlands Cancer Institute

**Muzhen He**

Fujian Provincial Hospital

**Jonas Teuwen**

The Netherlands Cancer Institute https://orcid.org/0000-0002-1825-1428

**Maciej Mazurowski**

Duke University

**Ritse Mann**

Radboud UMC https://orcid.org/0000-0001-8111-1930

**Article**

**Keywords:**

**Posted Date:** April 15th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9185479/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

**Additional Declarations:** There is **NO** Competing Interest.

# Decoding Tumor Phenotypes: A Radiologist-Inspired Deep Learning Framework for Breast Cancer Recurrence Prediction

Chunyao Lu[1,2#], Tianyu Zhang[1,2#], Xinglong Liang[1,2#], Antonio Portaluri[1,2], Luyi Han[1,2], Yaqian Chen[3], Nika Rasoolzadeh[1,2], Ruixiang Qi[1,2,4], Yuan Gao[1,5], Xin Wang[1,5], Yaofei Duan[1,2,6], Zahra Aghdam[1,2], Muzhen He[1,7], Jonas Teuwen[2,8], Maciej A. Mazurowski[3,9,10,11], Tao Tan[6*], Ritse Mann[1,2]

[1]Department of Radiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands

[2]Department of Medical Imaging, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands

[3]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, United States

[4]Department of Ultrasound, Affiliated Hangzhou First People's Hospital, Westlake University, 310006, Hangzhou, China

[5]GROW School for Oncology and Development Biology, Maastricht University, P. O. Box 616, 6200 MD, Maastricht, The Netherlands

[6]Faculty of Applied Sciences, Macao Polytechnic University, 999078, Macao, China

[7]Department of Radiology, Shengli Clinical Medical College of Fujian Medical University, Fujian Provincial Hospital, Fuzhou University Affiliated Provincial Hospital, Fuzhou, Fujian, 350001, China

[8]Department of Radiation Oncology, The Netherlands Cancer Institute, Plesmanlaan 121, Amsterdam, 1066 CX, The Netherlands

[9]Department of Radiology, Duke University, Durham, NC 27710, United States

[10]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, United States

[11]Department of Computer Science, Duke University, Durham, NC 27708, United States

[#] Chunyao Lu, Tianyu Zhang, and Xinglong Liang are co-first authors.

* Correspondence: Tao Tan (taotanjs@gmail.com)

## Abstract

Accurate prognostication in breast cancer remains constrained by the limited resolution of conventional clinicopathological markers, often leading to overtreatment of indolent disease or missed opportunities for intensified therapy in aggressive cases. Here we present a radiologist-inspired multimodal deep learning framework that integrates whole-volume dynamic contrast-enhanced MRI (DCE-MRI), radiology reports and clinical variables to predict recurrence-free survival. The model is trained through cross-modal semantic alignment, in which report-derived representations guide the learning of prognostic imaging features. In a large development cohort from the Netherlands Cancer Institute (NKI, n = 3,266), the framework achieved strong discrimination (C-index 0.727) and identified high-risk patients with a hazard ratio of 7.32 ($P < 0.0001$). Robustness was further assessed in two independent public external cohorts (Duke and I-SPY 1; n = 1,054) using text-free inference, where the model maintained consistent performance (C-index 0.702 and 0.697, respectively). Prognostic accuracy remained stable across both short- and long-term horizons, with high discriminative power from 2 to 12 years after treatment (AUC 0.77–0.83). This temporal stability enabled reliable identification of ultra-low-risk subgroups (5-year Recurrence-free survival > 95%) as well as early detection of aggressive phenotypes prone to early recurrence. These findings support semantically guided multimodal learning as an interpretable and operationally robust imaging biomarker for personalized breast cancer management.

# Introduction

Breast cancer is the most diagnosed cancer and remains a leading cause of cancer-related mortality among women worldwide[1]. While advances in early detection and treatments have significantly improved survival, disease recurrence, particularly distant metastasis, remains the leading cause of death[2,3]. Breast cancer management is increasingly shifting from population-based treatment strategies toward precision oncology. Central to this transition is accurate risk stratification, which enables optimization of the therapeutic ratio[4]. Reliable prognostic assessment allows de-escalation of toxic chemotherapy in patients with indolent disease, thereby reducing avoidable morbidity and healthcare costs[5]. Conversely, it enables identification of patients with aggressive, therapy-resistant tumors who may benefit from intensified treatment strategies or enrolment in clinical trials[6,7].

At present, prognostication relies largely on clinicopathological factors (e.g., TNM staging, receptor status), which may not fully reflect the biological complexity and heterogeneity of breast cancer[8]. While genomic assays (e.g., Oncotype DX, MammaPrint) have improved the risk stratification for specific luminal subtypes, the clinical use remains limited due to several practical considerations[9-11]. First, cost-related burden and logistical requirements can restrict access in resource-constrained healthcare settings, potentially widening global health disparities[12,13]. Second, reliable genomic biomarkers remain scarce for aggressive subtypes such as triple-negative breast cancer (TNBC), where existing panels are often less informative[14].

In parallel, computational pathology has emerged as a promising approach for prognostic modelling, leveraging deep learning on digitized whole-slide images (WSIs) to capture morphological features beyond conventional histological grading[15-21]. Nevertheless, both genomic assays and WSI-based models remain susceptible to sampling bias[22,23]. Biopsy samples provide only a localized and static view of a spatially heterogeneous tumor ecosystem and may fail to capture distinct intratumoral niches and peritumoral microenvironments associated with drug resistance[24,25]. Consequently, there remains a clear need for non-invasive prognostic tools that capture the “whole-tumor” phenotype to enable personalized management without adding economic or procedural burden.

Radiological imaging, particularly Dynamic Contrast-Enhanced MRI (DCE-MRI), offers a unique opportunity to address these challenges. Unlike biopsy-based approaches, DCE-MRI provides a macroscopic and functional assessment of tumor biology, capturing perfusion dynamics and kinetic heterogeneity across the entire lesion[26]. In addition, radiology reports represent a rich repository of expert cognition, encapsulating semantic reasoning that pixel data alone cannot convey. Despite these advantages, a “cognitive gap” persists in current AI-based approaches. Most multimodal methods combine imaging and text using late-fusion strategies and require both modalities during inference, which limits robustness in real-world settings where radiology reports may be inconsistent or unavailable[27,28]. Moreover, few models explicitly align visual feature learning with the semantic reasoning used by radiologists, which restricts interpretability and clinical trust[29].

In this study, we propose a biomimetic multimodal deep learning framework designed to bridge this gap between imaging perception and expert semantic reasoning for recurrence prediction in breast cancer. The framework leverages radiology-report language during training to supervise whole-volume DCE-MRI representation learning via cross-modal knowledge distillation, thereby internalizing clinically meaningful semantics into the visual encoder. Moreover, this design enables report-free deployment when radiology reports are variable or unavailable, while maintaining individualized risk prediction from imaging and clinical inputs alone. We developed the model using a large institutional cohort from the Netherlands Cancer Institute (NKI, n = 3,266) and evaluated its robustness in two independent external datasets (Duke, n = 892 and I-SPY 1, n = 162). Despite the absence of radiology reports during external inference, the model preserved strong predictive performance. The framework reliably identifies ultra-low-risk patients who may be candidates for treatment de-escalation, while identifying patients with aggressive disease who may benefit from more intensive therapies. By integrating imaging perception with expert-derived semantics, this approach provides an interpretable, scalable and cost-neutral digital biomarker to support personalized breast cancer management across diverse clinical settings (Fig. 1).

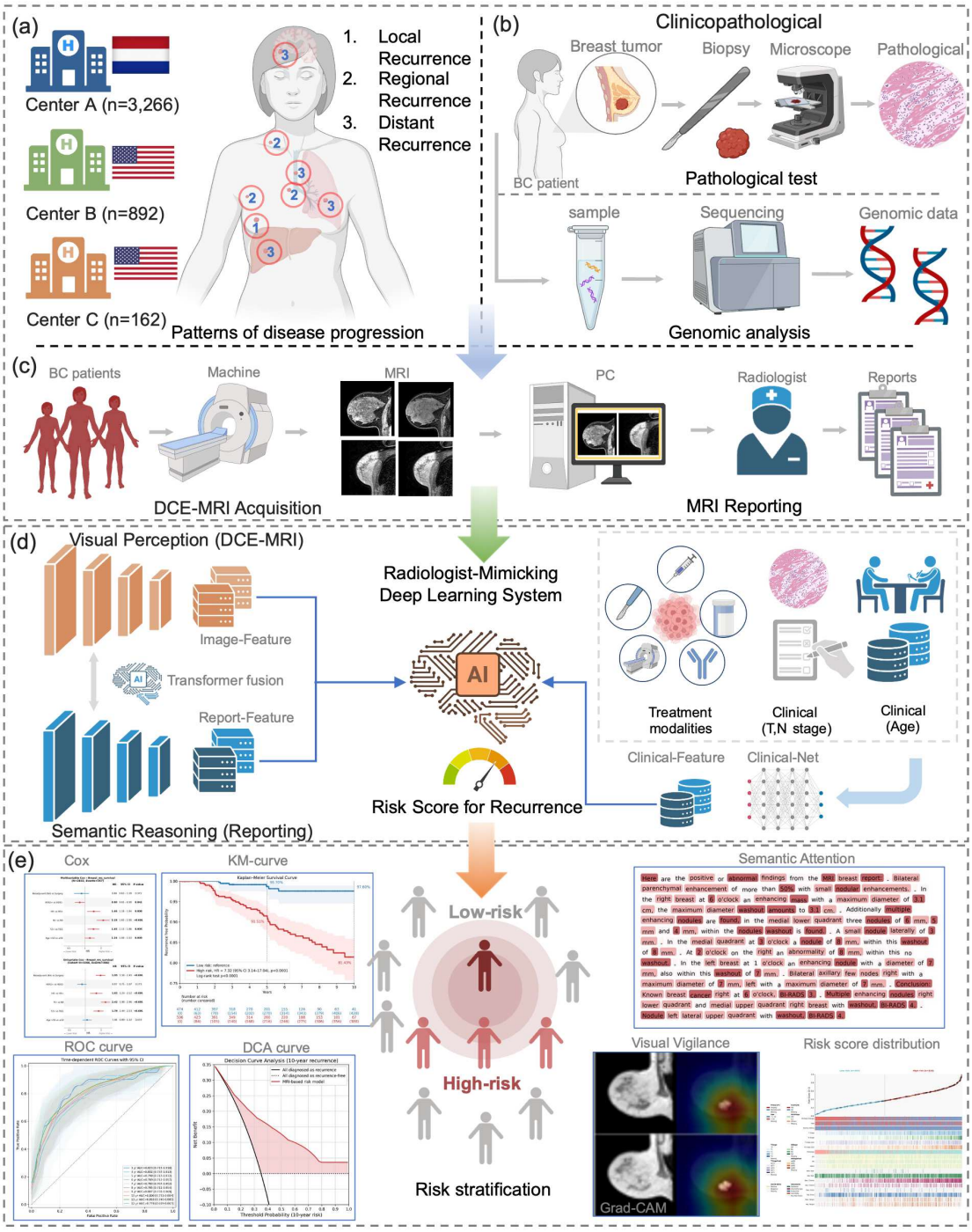



**Figure 1 | Study design and multimodal deep learning framework for breast cancer prognostication. a,** Overview of the multi-center study design, including one development cohort (Center A: NKI dataset) and two independent external validation cohorts (Center B: DUKE dataset and Center C: I-SPY1 dataset). The primary endpoint was recurrence-free survival, defined as the time from diagnosis to the first documented local, regional or distant recurrence. **b,** Standard clinical workflow for breast cancer assessment, including tissue biopsy, pathological evaluation and molecular profiling. These procedures provide the clinicopathological reference variables used in routine practice. **c,** Radiological workflow. Pre-treatment DCE-MRI was acquired as part of routine care, together with the

corresponding free-text radiology reports. **d,** Overview of the “Perception-to-Cognition” framework. The model integrates three complementary inputs: DCE-MRI-derived imaging features, report-derived semantic representations and structured clinical variables, including age, treatment and clinical T/N stage, to generate a continuous risk score. **e,** Model evaluation and interpretability. Patients were stratified into low-risk and high-risk groups, and prognostic performance was assessed using Kaplan–Meier analysis, time-dependent receiver operating characteristic curves and decision curve analysis. Model interpretability was examined using token-level attention maps for text and voxel-wise Grad-CAM heatmaps for MRI.

**Table 1 | Baseline demographic, clinical and pathological characteristics of the development and validation cohorts.** Data are presented as number of patients (n) with percentages (%) for categorical variables and mean ± standard deviation (s.d.) for continuous variables. The study includes the internal development cohort (Center A: NKI) and two independent external validation cohorts (Center B: Duke; Center C: I-SPY1). $P$-values indicate statistical differences across the three cohorts, calculated using Pearson's $\chi^2$ test for categorical variables and one-way analysis of variance (ANOVA) for age. Significant differences were observed across cohorts for most variables (all $P < 0.001$), except for T stage, adjuvant anti-HER2 therapy and neoadjuvant radiotherapy.

**Abbreviations**: NA, Neoadjuvant; ER, estrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; NKI, Netherlands Cancer Institute.

| Characteristics | Center A: NKI (n=3,266) | Center B: DUKE (n=892) | Center C: ISPY1 (n=162) | *P*-value |
|---|---|---|---|---|
| **Recurrence** | 386(11.82%) | 88(9.87%) | 45(27.78%) | <0.001 |
| **Age at Diagnosis** mean (SD) | 51.26±11.42 | 52.77±11.14 | 48.06±8.95 | <0.001 |
| ≤ 50 years | 1597(48.90%) | 375(42.04%) | 92(56.79%) | <0.001 |
| > 50 years | 1669(51.10%) | 517(57.96%) | 70(43.20%) | <0.001 |
| **Primary therapy** | | | | <0.001 |
| NA | 1768(54.13%) | 301(33.74%) | 162(100%) | <0.001 |
| Surgery | 1498(45.87%) | 591(66.26%) | 0 | <0.001 |
| **Adjuvant therapy** | 2793(85.52%) | 815(91.37%) | - | <0.001 |
| Adjuvant: Anti-HER2 | 521 | 137 | - | 0.964 |
| Adjuvant: Chemotherapy | 692 | 331 | - | <0.001 |
| Adjuvant: Endocrine | 1709 | 593 | - | <0.001 |
| Adjuvant: Radiation | 2402 | 602 | - | <0.001 |
| **Neoadjuvant therapy** | 1768(54.13%) | 301(33.74%) | 162(100%) | <0.001 |
| Neoadjuvant: Anti-HER2 | 448 | 82 | - | <0.001 |
| Neoadjuvant: Chemotherapy | 1502 | 283 | - | <0.001 |
| Neoadjuvant: Endocrine | 197 | 20 | - | <0.001 |
| Neoadjuvant: Radiation | 73 | 19 | - | 0.923 |
| **T Stage** | | | | 0.427 |
| T1 | 1244(38.09%) | 407(45.0%) | - | |
| T2 | 1272(38.95%) | 380(42.60%) | - | |
| T3 | 286(8.76%) | 85(9.50%) | - | |
| T4 | 45(1.38%) | 20(2.20%) | - | |
| Missing | 419(12.83%) | 6(0.67%) | 162(100%) | |
| **N stage** | | | | <0.001 |
| N0 | 2189(67.02%) | 523(58.63%) | - | |
| N+ | 1024(31.35%) | 352(39.46%) | - | |

| | | | | |
|---|---|---|---|---|
| Missing | 53(1.62%) | 17(2.13%) | - | |
| **Post T Stage** | | | | <0.001 |
| ypT1 | 1704(52.17%) | 401(45.0%) | - | |
| ypT2 | 534(16.35%) | 380(42.60%) | - | |
| ypT3 | 81(2.48%) | 85(9.53%) | - | |
| ypT4 | 5(0.15%) | 20(2.24%) | - | |
| Missing | 942(28.84%) | 6(0.67%) | 162(100%) | |
| **Post N stage** | | | | <0.001 |
| ypN0 | 1859(56.92%) | 523(58.63%) | - | |
| ypN+ | 915(28.02%) | 352(39.49%) | - | |
| Missing | 492(15.06%) | 17(1.91%) | - | |
| **ER** | | | | <0.001 |
| Positive | 2398(73.42%) | 668(74.89%) | 84(51.85%) | |
| Negative | 538(16.47%) | 224(25.11%) | 78(48.15%) | |
| Missing | 330(10.1%) | 0 | 0 | |
| **PR** | | | | <0.001 |
| Positive | 2117(64.82%) | 581(65.13%) | 67(41.35%) | |
| Negative | 829(25.08%) | 311(34.87%) | 95(58.64%) | |
| Missing | 330(10.1%) | 0 | 0 | |
| **HER2** | | | | <0.001 |
| Positive | 1358(41.58%) | 160(17.94%) | 52(31.10%) | |
| Negative | 1572(48.13%) | 732(82.06%) | 110(67.90%) | |
| Missing | 336(10.29%) | 0 | 0 | |
| **Histologic types** | | | | <0.001 |
| IDC | 2814(86.16%) | 892(100%) | 162(100%) | |
| ILC | 337(10.32%) | 0 | 0 | |
| Others | 115(3.53%) | 0 | 0 | |
| **Family history** | | | | <0.001 |
| Yes | 399(12.22%) | | | |
| No | 2217(67.88%) | 892(100%) | 162(100%) | |
| Missing | 650(19.9%) | | | |

# Results

## Patient cohorts and clinical heterogeneity across study centers

To evaluate the generalizability of the proposed framework, we assembled a multi-institutional cohort of 4,320 patients from three independent centers: the Netherlands Cancer Institute (NKI, n = 3,266), Duke University (Duke, n = 892) and the I-SPY1 clinical trial (I-SPY1, n = 162)[30-32]. Substantial differences in clinicopathological characteristics, treatment strategies and outcome distributions were observed across the three cohorts ($P < 0.001$ for most key variables; Table 1). This high heterogeneity creates a challenging and realistic setting to probe model robustness under real-world domain shifts.

Specifically, the NKI and DUKE cohorts served as representative samples of the general breast cancer population, capturing a broad spectrum of early-stage disease managed with standard-of-care protocols. In these cohorts, most tumors were early-stage (T1/T2: 77.0% in NKI, 87.6% in DUKE) and node-negative (67.0% in NKI, 58.6% in DUKE), reflecting a typical surveillance population. In contrast, the I-SPY1 cohort represents a clinically distinct population enriched for high-risk disease. All patients in this cohort received neoadjuvant chemotherapy (100% vs. 33.7% in DUKE, $P < 0.001$) as primary therapy and presented with more advanced disease at diagnosis. This difference was reflected in recurrence outcomes: recurrence occurred in 11.8% of patients in NKI (386/3,266) and 9.9% in Duke (88/892), compared with 27.8% in I-SPY1 (45/162). These differences create a challenging data heterogeneity gap, serving as an ideal platform to validate whether our AI model captures intrinsic biological phenotypes independent of population shifts. A comprehensive overview of the clinicopathological distributions and cross-cohort comparisons among the three datasets is provided in Supplementary Fig. S1.

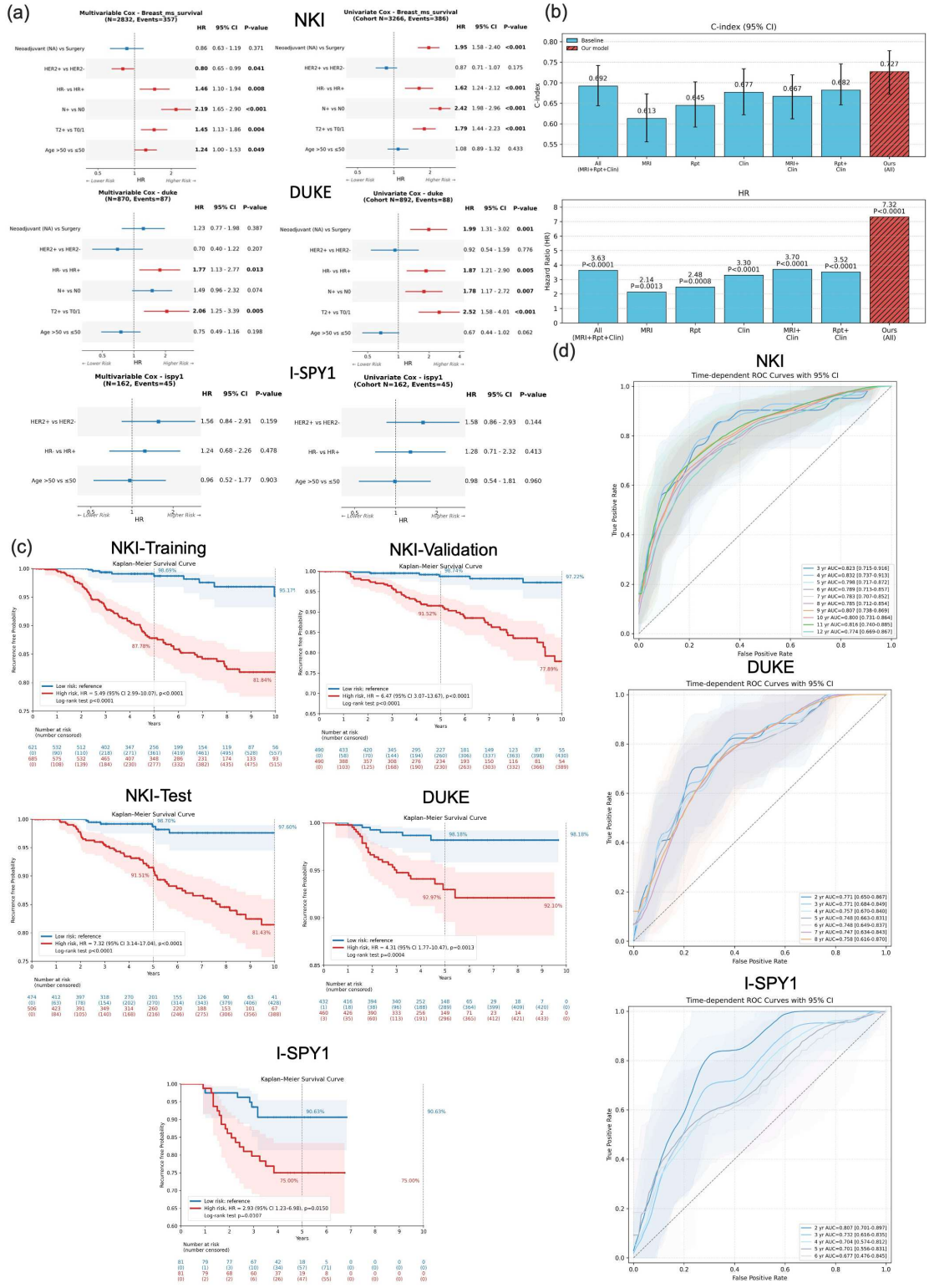



**Fig. 2 | Prognostic performance of the multimodal AI framework compared with standard clinical variables. a,** Prognostic effects of standard clinicopathological variables. Forest plots show univariable (right) and multivariable (left) Cox proportional hazards analyses for the NKI, DUKE and I-SPY1 cohorts. Hazard ratios (HRs) and 95% confidence intervals are displayed for each variable. These panels establish the clinical baseline,

showing that traditional risk factors (e.g., higher T/N stage) are generally associated with increased recurrence risk (HR > 1), though with varying significance levels across cohorts (e.g., Age is not significant in all settings). **b,** Model comparison and ablation analysis in the NKI internal test cohort. Prognostic performance of the proposed multimodal AI model ("Ours") compared with baseline models, including single-modality models (MRI-only, report-only and clinical-only) and simple multimodal combinations. Performance is reported using the concordance index (C-index; top) and hazard ratio between predicted high- and low-risk groups (bottom). **c,** Risk stratification using the AI-derived risk score. Kaplan–Meier curves of recurrence-free survival for the training, validation and internal test cohorts (NKI), as well as external validation cohorts (Duke and I-SPY1). Patients were stratified into high-risk and low-risk groups according to the predicted risk score. Statistical significance was assessed using the log-rank test. **d,** Time-dependent predictive performance. Time-dependent receiver operating characteristic (ROC) curves at 2-, 3-, 5- and 10-year prediction horizons for the internal (NKI) and external (Duke and I-SPY1) cohorts. Shaded regions indicate 95% confidence intervals.

## Multimodal learning improves prognostic stratification beyond standard clinical variables

We first evaluated the prognostic value of conventional clinicopathological variables, including age, TN stage and molecular subtype. Univariable and multivariable Cox regression analyses showed that these factors provided limited and inconsistent prognostic stratification across cohorts (Fig. 2a). Hazard ratios ranged from 0.67 to 2.52, and several variables did not reach statistical significance. Consistent with these findings, models based solely on clinical variables showed limited discriminative performance (Fig. 2b).

We next assessed whether multimodal integration could improve prognostic prediction. A standard multimodal fusion model that concatenated learned representations from MRI, radiology reports, and clinical variables achieved higher C-indices than unimodal models but provided only modest improvement (Fig. 2b). Our Biomimetic framework, however, significantly improved upon this all-modality benchmark. By leveraging semantic supervision to distill expert knowledge, our model outperformed other methods, achieving a peak C-index of 0.727 and a Hazard Ratio of 7.32 (Log-rank test, $P$<0.0001) in the internal development cohort (NKI, Red bars).

This quantitative superiority translated into robust clinical stratification. In the NKI cohort, Kaplan-Meier analysis (Fig. 2c) demonstrated a distinct and statistically significant separation between high-risk (red) and low-risk (blue) patient groups (Log-rank test, $P$<0.0001). Similar patterns were observed across training, validation and internal test sets, indicating stable performance within the development dataset.

## External validation demonstrates robust generalization across heterogeneous cohorts

Given the model's superiority in the internal cohort, we then assessed its robustness against real-world data heterogeneity. We hypothesized that the semantic supervision during training effectively performed "cross-modal knowledge distillation," forcing the visual encoder to "internalize" expert-level diagnostic features. To validate this, we evaluated the model on external cohorts (DUKE and I-SPY 1) in a strictly "text-free setting. Specifically, during inference, no radiology reports or text-derived features were provided; the NLP branch used only during training for representation supervision was omitted, and recurrence risk was predicted solely from the MRI and clinical branches.

Despite the absence of textual input and the intrinsic technical variability in scanner manufacturers (The acquisition parameters of the DCE-MRI data across datasets are summarized in Supplementary Table S1.), the model maintained consistent predictive performance (Fig. 2c). In the DUKE cohort, it yielded a C-index of 0.702 and a significant HR of 4.31 (95% CI:1.77-10.47, Log-rank test, $P$=0.0004). In the I-SPY1 cohort, which consists of patients receiving neoadjuvant chemotherapy, the model achieved a C-index of 0.697 and a HR of 2.93 (95% CI: 1.23-6.98, Log-rank test, $P$=0.0107).

This suggests that through multimodal training on the NKI dataset, the visual encoder effectively learned to identify prognostic semantic features directly from pixel data, retaining high discriminative power without explicit textual guidance. This generalization capability translated into distinct risk stratification across institutions. As shown in the Fig. 2(c), the model successfully separated high-risk (red) from low-risk (blue) patients in both external settings.

## Stable prognostic performance across multiple follow-up horizons

To distinguish whether the model is capturing stable prognostic signals, we performed time-dependent receiver operating characteristic (ROC) analysis (Fig. 2d). Predictive performance remained consistent across multiple follow-up intervals, indicating that the model retains prognostic value across both early and late recurrence periods.

In the NKI cohort, the model did not merely predict long-term outcomes but demonstrated sustained high-performance across the full surveillance period. It achieved a 3-year AUC of 0.823 (95% CI: 0.715–0.916) for early recurrence prediction and maintained comparable discrimination over long-term follow-up, with an 11-year AUC of 0.816 (95% CI: 0.740–0.885).

This temporal robustness also generalized well to heterogeneous external cohorts, while remaining informative across patient populations with distinct baseline risk profiles. In the DUKE cohort, In the Duke cohort, which reflects a typical post-treatment surveillance setting, the model maintained robust discrimination during the critical intermediate window, yielding a 3-year AUC of 0.771 (95% CI: 0.684–0.849). High-Risk Stratification (I-SPY1): Crucially, in the I-SPY 1 cohort—characterized by locally advanced disease where the "risk window" is aggressively compressed—the model exhibited precise early-detection capability. It achieved a peak 2-year AUC of 0.807 (95% CI: 0.701–0.897), validating its

utility in identifying patients with rapid progression who require immediate therapeutic intensification. Together, these results indicate that the model provides reliable risk prediction across both early and long-term follow-up intervals and maintains predictive performance across heterogeneous clinical cohorts.

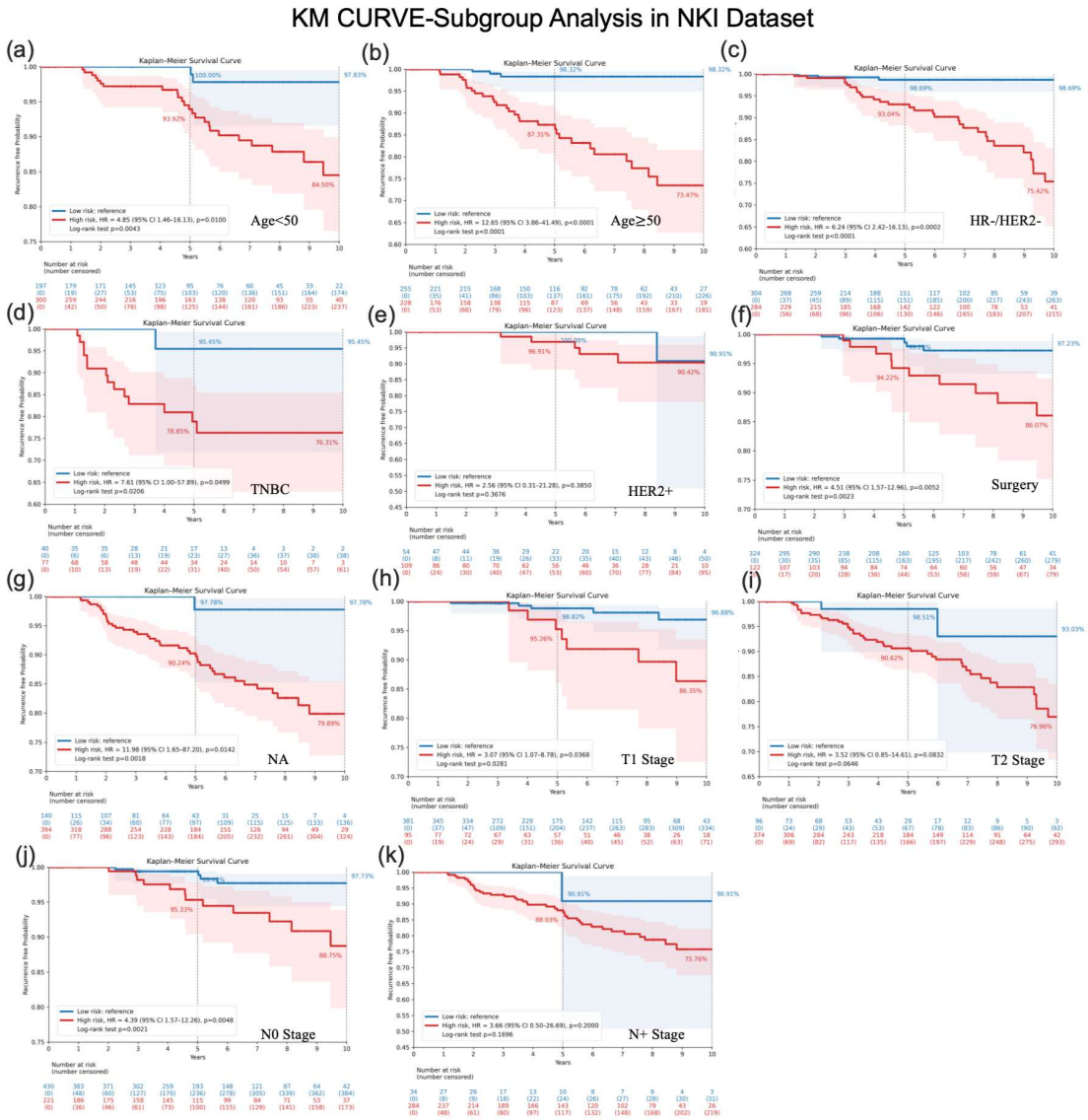



**Fig. 3 | Stratified Kaplan-Meier analysis demonstrating model robustness across clinically relevant subgroups in the NKI cohort.** Kaplan–Meier recurrence-free survival curves stratified by the AI-derived risk score (high risk, red; low risk, blue) within distinct patient subgroups. **a-k,** Subgroup analyses performed on the NKI internal test set, categorized by: Age: <50 years and ≥50 years. Molecular Subtype: HR+/HER2-, TNBC, and HER2+.Treatment: Surgery and Neoadjuvant Therapy (NA). Tumor Stage: T1-T2, T3, T4. Nodal Status: N0 and N+.

## Consistent risk stratification across clinical subgroups

To verify that our multimodal approach captures intrinsic tumor biology rather than acting as a simple proxy for demographic or anatomical features, we performed a comprehensive

stratified analysis within NKI and Duke cohort (Fig. 3 and Supplementary Fig. S2). The model maintained robust prognostic independence across all standard clinical variables. In terms of demographics, it yielded distinct risk separation in both younger (<50 years, HR=4.85, Log-rank test, $P$=0.0043) and older patients (≥50 years, HR=12.65, Log-rank test, $P < 0.0001$), indicating stable performance independent of age-related baseline risk. Notably, the model also identified risk heterogeneity within clinically defined subgroups. Among node-negative (N0) patients, who are typically considered lower risk based on anatomical staging, the model stratified patients into distinct risk groups (HR=4.39, Log-rank test, $P$=0.0021). This suggests that imaging-derived features may capture additional prognostic information beyond nodal status. Conversely, even within the biologically aggressive TNBC phenotype, the model successfully identified a distinct “ultra-low-risk” subset (HR = 7.61, log-rank test, P = 0.0206), with a 5-year recurrence-free survival rate exceeding 95%.Together, these results indicate that the multimodal model provides prognostic stratification across diverse clinical subgroups and may help identify risk heterogeneity that is not fully captured by standard clinicopathological factors.

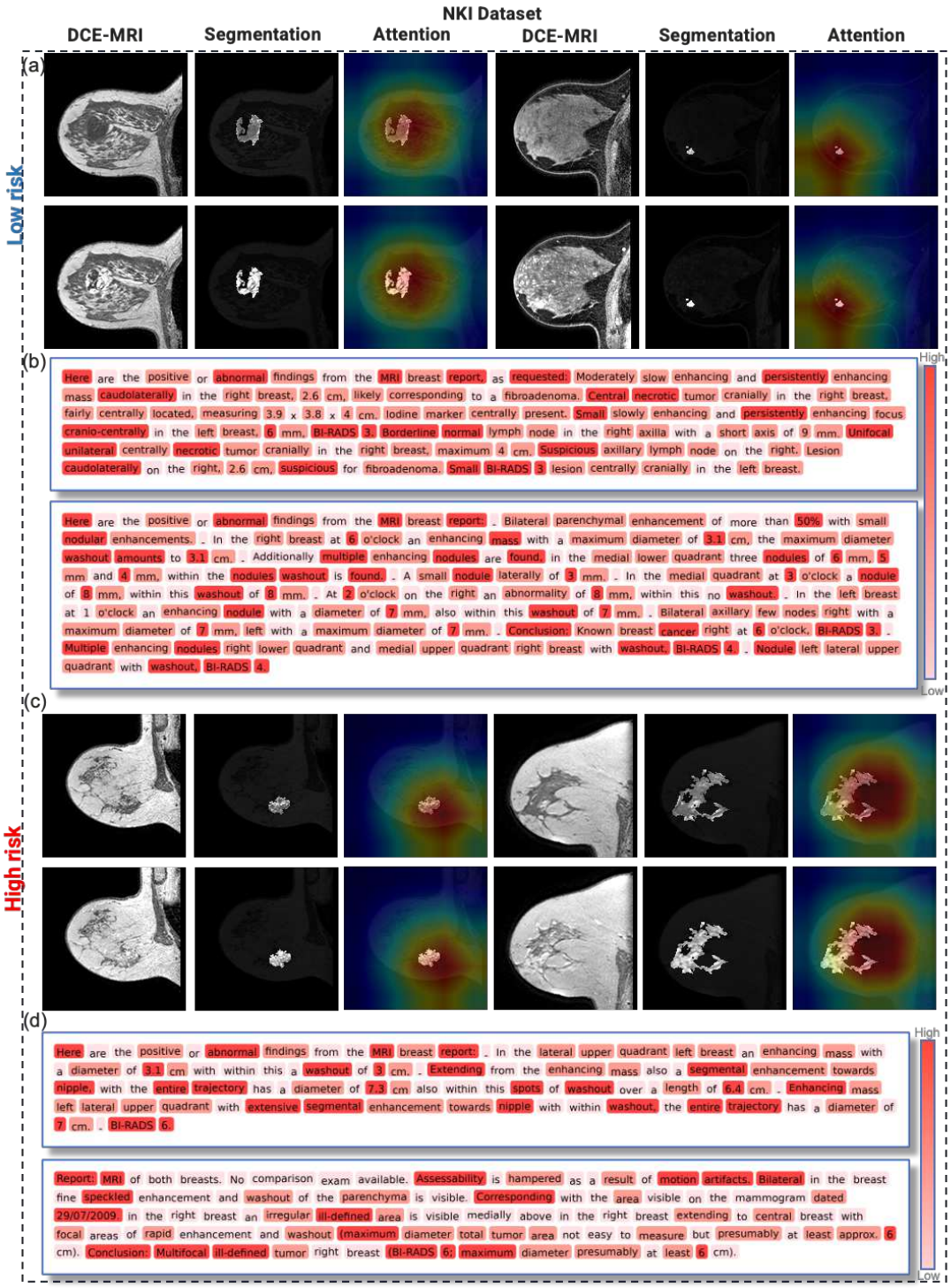



**Fig. 4 | Visual-cognitive alignment and interpretability of the multimodal framework.** **a, c,** Visual interpretability (Perception). Representative DCE-MRI slices, ground-truth tumor segmentations, and corresponding Gradient-weighted Class Activation Mapping (Grad-CAM) heatmaps for patients stratified as Low-risk (a) and High-risk (c). The attention heatmaps (Right column) demonstrate high spatial fidelity, accurately localizing the intratumoral regions driving the risk prediction without manual guidance. Notably, the model focuses on distinct morphological phenotypes: circumscribed, homogenous patterns in low-risk cases versus diffuse, heterogeneous patterns in high-risk cases. **b, d,**

Semantic interpretability (Cognition). Visualization of token-level self-attention weights from the radiology report encoder for the corresponding patients. Words receiving higher attention weights are highlighted, indicating textual features that contributed most strongly to the model prediction.

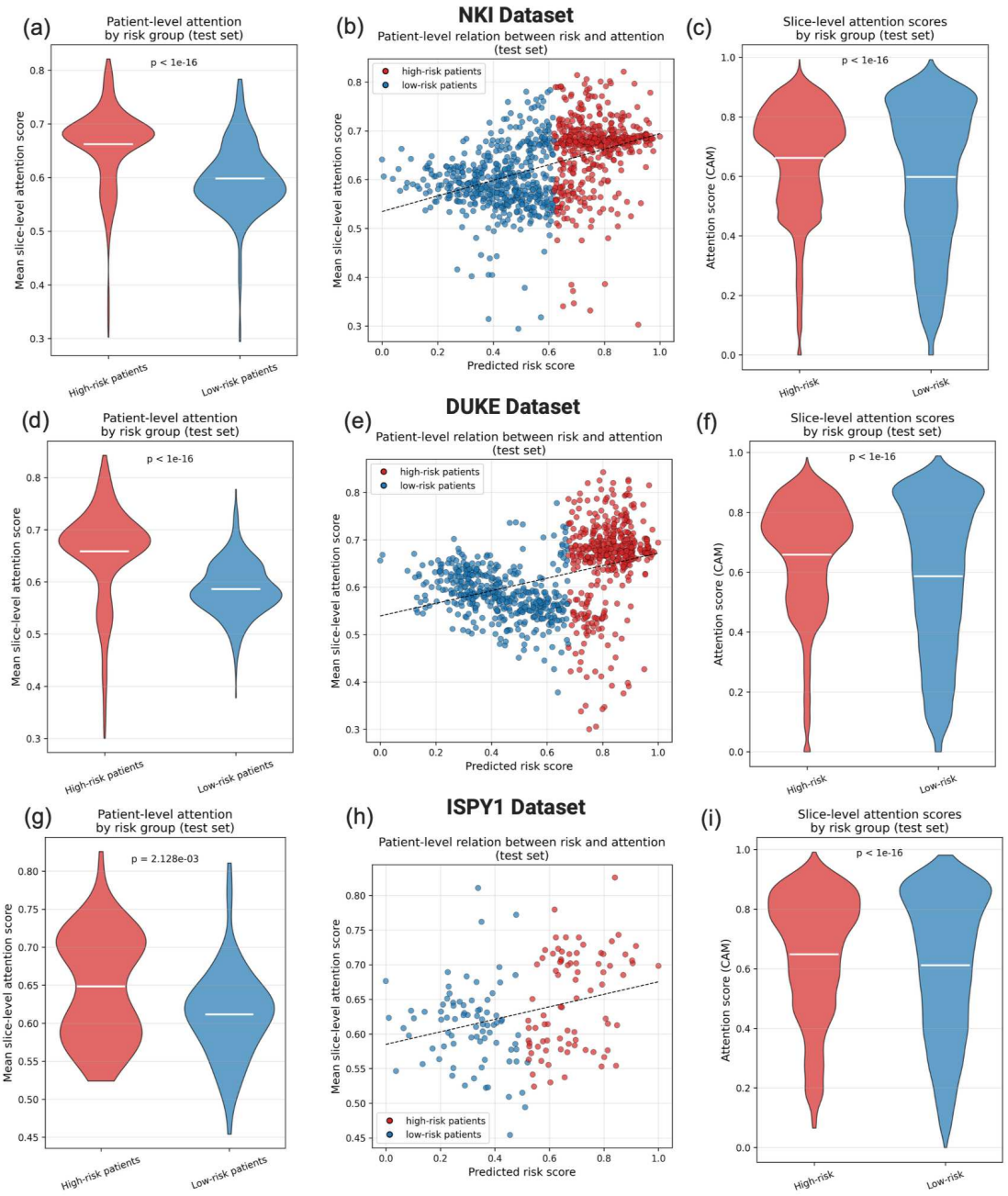



**Fig. 5 | Association between model attention scores and predicted recurrence risk across cohorts. a–c,** Analyses in the NKI cohort. **d–f,** Analyses in the Duke cohort. **g–i,** Analyses in the I-SPY1 cohort. **Left column:** Violin plots showing the distribution of patient-level mean attention scores derived from the visual encoder. Patients predicted as high risk (red) exhibit higher mean attention scores than those predicted as low risk (blue) across cohorts ( $P < 0.001$ ). **Middle column:** Scatter plots illustrating the positive correlation between the model's predicted risk score (x-axis) and the mean attention score

(y-axis). Dashed lines indicate linear regression trends. **Right column:** Violin plots of slice-level attention scores derived from Grad-CAM, showing the distribution of activation magnitudes across individual MRI slices for high- and low-risk groups.

## Alignment between visual attention, radiology report semantics and AI risk prediction

To validate the fidelity of our "Perception-to-Cognition" architecture, we analyzed the spatial attention patterns of the visual encoder and the semantic attention patterns of the report encoder.

Visual vigilance and phenotypic complexity: Gradient-weighted class activation maps (Grad-CAM) demonstrated that the model's visual attention was spatially concentrated on biologically relevant intratumoral regions and peritumoral margins (Fig. 4a, c), while largely excluding background parenchymal signal[33,34]. Similar spatial attention patterns were observed across the development cohort and external validation cohorts (Supplementary Fig. S3), indicating consistent visual feature localization across institutions. To quantify these observations, we compared attention scores between predicted risk groups. As illustrated by the violin plots (Fig. 5a, c), patients classified as high risk exhibited significantly higher mean attention intensities than low-risk patients across the development and external validation cohorts ($P < 1\times10^{-16}$, Wilcoxon rank-sum test). In addition, predicted risk logits showed a positive correlation with the corresponding attention scores (Fig. 5b, e, h). These findings suggest that the model allocates greater visual attention to imaging patterns associated with more complex tumor phenotypes, such as heterogeneous enhancement or irregular tumor margins, while comparatively less attention is required for lesions with more circumscribed imaging characteristics.

Semantic Dichotomy: Consistent with expert radiological reporting patterns, the NLP attention mechanism revealed a clear semantic dichotomy (Fig. 4b, d). For high-risk predictions, the model assigned greater attention to descriptors commonly associated with aggressive tumor behavior (e.g., "ill-defined," "multifocal," and "washout"). In contrast, low-risk predictions were characterized by higher attention weights assigned to terms typically linked to less suspicious imaging appearances (e.g., "oval" and "circumscribed").

Importantly, this semantic attention pattern was consistently observed across the development cohort and the external validation cohorts (Duke and I-SPY1), suggesting that the textual reasoning component of the model captures clinically meaningful descriptors that align with radiologists' diagnostic language. Together, these findings provide an interpretable link between image-derived representations, radiological report semantics, and the resulting AI risk predictions.

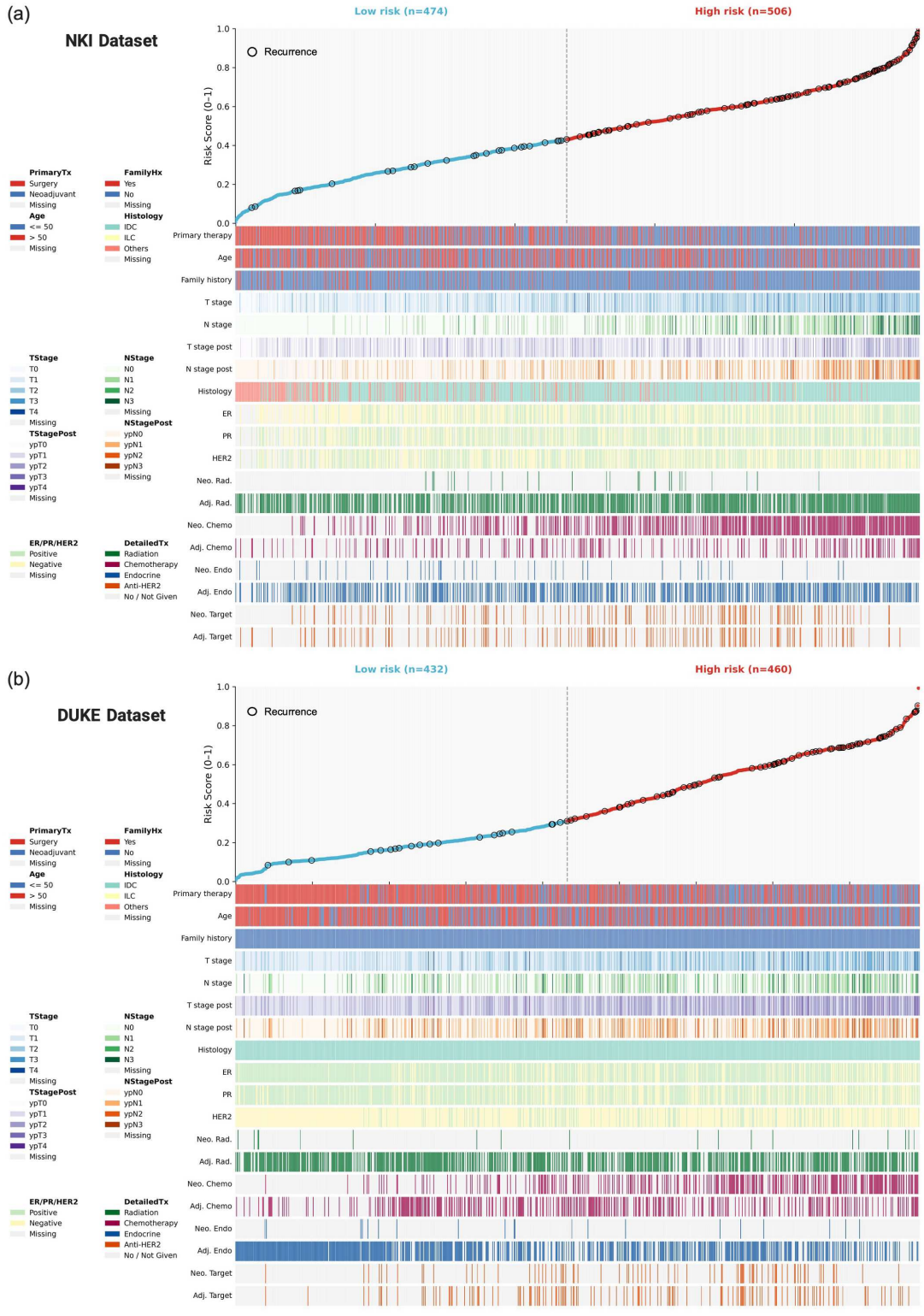



**Fig. 6 | Distribution of predicted risk scores and associated clinicopathological characteristics across cohorts. a, b**, Risk score distributions in the NKI development cohort (a) and the Duke external validation cohort (b). **Top panels** (risk curve): Patients are ordered according to their predicted AI risk score. The curve transitions from low risk (blue) to high-risk (red) based on the optimal stratification threshold. Black circles indicate patients with observed recurrence events. **Bottom panels** (clinical

heatmaps): Corresponding clinicopathological characteristics for each patient, aligned with the ranked risk scores.

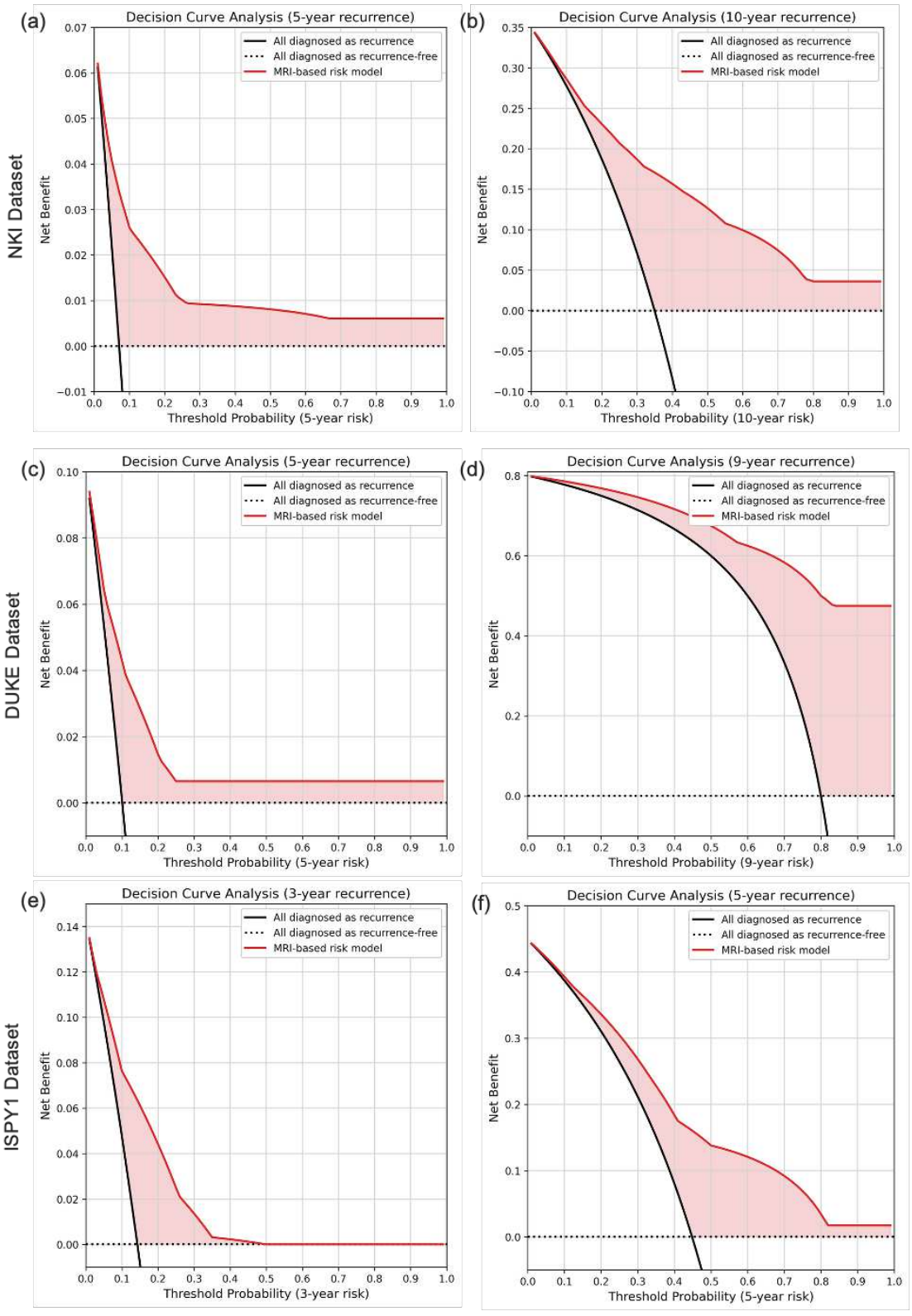
## Patient-level risk landscapes reveal associations with clinical characteristics

To visually interpret the determinants underlying the AI-derived risk predictions, we mapped the risk landscape of patients in the internal test cohort by ranking individuals according to their deep learning risk scores (Fig. 6a). Recurrence events (black circles) accumulated progressively along the increasing risk spectrum, whereas relatively few events were observed among patients with lower predicted risk, indicating effective stratification of recurrence risk.

The corresponding clinical heatmaps provide additional context for the predicted risk gradient. As the risk score increases, a shift in clinicopathological characteristics is observed, with early-stage disease features (for example T1/N0) more common in the lower-risk range and more advanced tumor burden (for example T3/N+) appearing more frequently among higher-risk patients.

Treatment patterns also provide clinically relevant context. Patients located toward the higher end of the risk spectrum were more frequently treated with neoadjuvant systemic therapy and multimodal treatment strategies. Despite receiving these intensified treatment regimens, recurrence events remained concentrated within this group, suggesting that the model-derived risk score may capture additional prognostic information beyond conventional staging factors. Conversely, patients classified as low risk exhibited relatively favorable recurrence-free outcomes and were more commonly managed with less intensive treatment strategies.

Similar risk stratification patterns were observed in external validation cohorts. When patients in the Duke dataset were ranked according to predicted risk scores (Fig. 6b), recurrence events also accumulated toward the higher-risk range. Comparable trends were also observed in the I-SPY1 external validation cohort (Supplementary Fig. S4), despite its distinct clinical characteristics and higher baseline recurrence prevalence.




374

375 **Fig. 7 | Decision curve analysis of the clinical utility of the AI risk model. a, b,** Decision
376 curve analysis (DCA) for the NKI cohort at 5-year (a) and 10-year (b) recurrence horizons.
377 **c, d,** DCA for the Duke cohort at 5-year (c) and 9-year (d) recurrence horizons. **e, f,** DCA
378 for the I-SPY1 cohort at 3-year (e) and 5-year (f) recurrence horizons. In each panel, the

net benefit of the AI risk model is compared with the default strategies of treating all patients (solid black line) or treating no patients (horizontal dotted line) across a range of threshold probabilities.

## Decision curve analysis supports potential clinical utility across cohorts

To assess the potential clinical utility of the model, we performed decision curve analysis (DCA) across the development and validation cohorts (Fig. 7). Across all datasets, the multimodal AI model showed higher net benefit than the default strategies of treating all patients or treating no patients over a range of threshold probabilities.

In the NKI and Duke cohorts (Fig. 7a-d), the analysis revealed distinct utility patterns across temporal frames. For short-term prediction (5-year recurrence), the "treat-all" strategy showed a negative net benefit at thresholds exceeding 0.10, whereas the AI model maintained positive net benefit over the same range. This indicates that the model can effectively prevent overtreatment in early-stage disease where non-selective intervention offers no net benefit. For long-term prediction (9-to-10-year recurrence), the model still demonstrated sustained incremental benefit even at stringent decision thresholds ($P_t > 0.40$), supporting its value in guiding extended surveillance.

In the I-SPY1 cohort (Fig. 7e, f), which represents a higher-risk population with a shorter follow-up window (3–5 years), the model provided additional net benefit relative to default strategies, particularly within the intermediate threshold range where treatment decisions may be less certain. Together, these results support that the AI-derived risk score may offer clinically relevant decision support across heterogeneous risk subgroups, ranging from mitigating overtreatment in broader populations to refining stratification among aggressive, high-risk phenotypes.

## Discussion

The transition from population-based treatment strategies to precision oncology in breast cancer requires improved characterization of tumor heterogeneity beyond standard clinicopathological parameters. In this study, we developed and validated a biomimetic "Perception-to-Cognition" multimodal framework that reconstructs a holistic tumor phenotype from routine clinical data and predicts breast cancer recurrence. By integrating whole-volume DCE-MRI representations with clinical variables and semantically aligned visual embeddings learned under text supervision, the model achieved consistent performance across three independent cohorts with distinct clinical settings: a Dutch routine surveillance population (NKI), a U.S. real-world clinical cohort (Duke), and a high-risk neoadjuvant trial cohort (I-SPY1). Despite substantial differences in patient populations, treatment strategies and imaging protocols, the model maintained stable discriminative performance (C-index 0.727 in NKI, 0.702 in Duke, 0.697 in I-SPY 1) and consistent risk stratification in external validation (HR: 7.32 in NKI, 4.31 in Duke, 2.93 in I-SPY 1). These findings suggest that the framework captures transferable imaging phenotypes that remain informative across heterogeneous clinical environments.

From a clinical perspective, accurate risk stratification is essential for guiding both treatment intensification and de-escalation[35]. Among patients conventionally considered low risk, our model identified a subgroup with underestimated recurrence risk, suggesting a role in flagging individuals who may benefit from closer monitoring or more intensive systemic treatment. Conversely, among patients generally considered high risk, including those with triple-negative breast cancer, the model identified an ultra-low-risk subset with a 5-year recurrence-free survival exceeding 95%. If prospectively validated, this non-invasive imaging-based approach could help avoid overtreatment in selected patients and complement, or in some settings potentially reduce reliance on, genomic assays[36,37]. Because the method is based on routinely acquired imaging, it may also offer a more accessible and cost-effective alternative in healthcare settings where genomic testing is less available. Furthermore, the sustained improvement over clinical nomograms (ΔC-index ≈ 0.05) highlights the incremental prognostic value of whole-tumor imaging phenotypes beyond conventional clinicopathological variables.

Interpretability analyses further support the biological plausibility of the model predictions. Attention and attribution analyses showed that high-risk predictions were associated with imaging patterns commonly linked to aggressive tumor behavior, including heterogeneous enhancement, ill-defined margins and peritumoral abnormalities, whereas low-risk predictions were associated with more circumscribed and homogeneous imaging appearances. This concordance with established radiological semiotics supports that the model's "visual vigilance" is not simply exploiting site-specific noise but rather identifying phenotypic biomarkers. Importantly, radiology reports were not used as inputs during external validation; text served solely as a training-time supervisory signal to guide

representation learning. Therefore, the observed external generalization does not depend on the style or completeness of reports at the validation sites.

From a methodological perspective, achieving robust multi-center generalization in MRI-based models remains challenging because of variability in scanner manufacturers, field strength, temporal resolution and acquisition protocols[38]. The preservation of predictive accuracy across geographically and technically diverse cohorts without retraining or explicit domain adaptation suggests that semantically supervised multimodal alignment may help constrain visual feature learning toward biologically meaningful patterns rather than scanner-specific artefacts[39]. This observation provides a potential strategy for improving the robustness and deploy ability of medical imaging AI systems.

Despite these promising results, our study has some limitations. First, despite multi-center external validation, the retrospective design necessitates prospective evaluation to establish real-time utility and clinical net benefit, ideally within embedded workflows and with clinically relevant endpoints and decision-curve analyses. Second, local, regional, and distant recurrences were combined into a single composite recurrence endpoint. Although these failure patterns differ in biological basis, clinical consequences, and prognosis, this choice was motivated by the low proportion of recurrence events relative to the overall study population. Importantly, even after combining recurrence modes, the dataset remained highly imbalanced, reflecting the intrinsic distribution of the cohorts rather than an arbitrary analytical choice. Separate modeling of individual recurrence types would therefore have been limited by small event numbers and reduced statistical robustness. Accordingly, the model should be interpreted as predicting overall recurrence risk rather than site-specific patterns of failure. Future studies with larger, uniformly annotated datasets should evaluate these endpoints separately. Third, while we intentionally avoided using reports at inference to lower deployment barriers, the use of text supervision during training may still imprint cohort-specific semantic anchors; future work could incorporate large language models to normalize heterogeneous reporting into a unified semantic space or explore self-supervised/weakly supervised alternatives to further reduce reliance on a particular semantic schema. Fourth, the current model focuses on macroscopic tumor phenotypes captured by imaging. Integrating imaging-derived risk signatures with molecular or microenvironmental data, such as genomics, spatial transcriptomics or immune profiling, may help clarify the biological mechanisms underlying the observed risk patterns.

In conclusion, we demonstrate that a biomimetic, multimodal deep learning system can non-invasively decode the tumor phenotype to predict breast cancer recurrence with high accuracy and interpretability. By combining imaging perception with semantically guided feature learning, the proposed approach provides an interpretable and scalable method for risk stratification across heterogeneous clinical settings. Such models may help refine therapeutic decision-making by identifying patients who may benefit from treatment de-escalation as well as those requiring more aggressive management, thereby supporting more personalized breast cancer care.

# Methods

## Study design and cohorts

This retrospective multi-center study aimed to develop and validate a multimodal deep learning model for breast cancer prognostication using pre-treatment DCE breast MRI, unstructured radiology reports, and structured clinicopathological variables. The study was approved by the Institutional Review Board of the Netherlands Cancer Institute (NKI; IRBd21-058) with a waiver of informed consent due to the retrospective design. Corresponding approvals were obtained at the external validation institutions.

**Development Cohort: The NKI Dataset.** The development cohort was derived from the NKI. A total of 10,336 women diagnosed with breast cancer between January 2007, and April 2024 were screened. Patients were excluded if they (i) lacked pre-treatment DCE-MRI (n=1203) (ii) had no corresponding radiology report (n=1810) or (iii) had insufficient follow-up information (minimum follow-up <2 years for patients without recurrence, n=4057). After applying these criteria, 3,266 patients were included in the final cohort.

The NKI cohort reflects routine clinical practice at a tertiary cancer center and therefore comprises a relatively high-risk patient population. It includes patients treated with neoadjuvant systemic therapy as well as those undergoing primary surgery. For model development, the cohort was randomly divided at the patient level into a training set (n = 1,306), a validation set (n = 980) and an internal test set (n = 980) to avoid information leakage across datasets. The detailed patient selection process is shown in Supplementary Fig. S5(a).

**External Validation Cohorts: Duke and ISPY1 cohorts.** To assess generalizability, we performed external validation on two independent cohorts derived from large US-based datasets.

**Duke cohort.** The Duke Breast MRI cohort was retrospectively collected from the Duke Hospital, including 922 patients diagnosed with breast cancer between 1 January 2000 to 23 March, 2014[32]. Patient selection followed the same inclusion and exclusion criteria as the internal cohort. After excluding cases lacking follow-up data (n=30), 892 patients were included for external validation. Within this cohort, 591 patients received neoadjuvant therapy, while 301 patients underwent primary surgery. During follow-up, 88 patients experienced recurrence whereas 804 remained recurrence-free. As a large, real-world, single-institution academic cohort with heterogeneous imaging protocols and clinical pathways, this dataset provides a stringent external assessment of model robustness across institutional settings.

**ISPY1 cohort.** The second external validation cohort was derived from the I-SPY 1 Trial (Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis 1), a prospective, multi-center clinical trial enrolling 222 patients

from May 2002 to March 2006[30-31]. Patients without tumor annotations (n=51) or follow-up data (n=9) were excluded. Only patients with baseline breast MRI and complete clinical variables were included. A total of 162 patients met these criteria and were included for external testing. All patients in the I-SPY 1 cohort received neoadjuvant chemotherapy. During follow-up, 45 patients developed recurrence while 117 remained recurrence-free. Owing to its prospective design, standardized imaging acquisition, and rigorous data quality control, the I-SPY 1 cohort provides a complementary external validation setting that evaluates model performance under highly harmonized clinical trial conditions.

The detailed cohort construction process for the external datasets is provided in Supplementary Fig. S5 (b)(c).

## Clinical endpoint and follow-up

The primary endpoint was recurrence-free survival (RFS), defined as the interval from initiation of primary treatment (neoadjuvant therapy or surgery) to the first documented recurrence or censoring at last follow-up. Recurrence events were categorized as local (ipsilateral breast or chest wall), regional (ipsilateral axillary, supraclavicular, infraclavicular, or internal mammary lymph nodes), or distant (metastatic disease involving non-regional lymph nodes and/or distant organs). Local and regional recurrences required histopathological confirmation. Distant recurrence was confirmed histopathologically when tissue sampling was available, or otherwise by unequivocal radiologic progression on serial imaging (e.g., contrast-enhanced CT, PET–CT, or MRI). Patients without documented recurrence were required to have a minimum follow-up of 2 years. Information on diagnosis and treatment was collected from the electronic health record, based on structured oncologic documentation, pathology results, and treatment records.

## Data harmonization and extraction

**DCE-MRI preprocessing.** Breast DCE-MRI volumes were loaded from Neuroimaging Informatics Technology Initiative (NifTI) files. For each patient, two pre-treatment phases—pre-contrast and early post-contrast—were used as paired inputs. The two volumes were concatenated along the channel dimension to form a two-channel 3D tensor for the visual encoder. The acquisition parameters of the DCE-MRI data from the in-house and external datasets are summarized in Supplementary Table S1.

**Intensity normalization.** To reduce inter-scan intensity variability, we applied percentile-based scaling jointly to the DCE pair. Specifically, the 99th percentile intensity of the early post-contrast volume, denoted as $p_{99}$, was used as a shared upper bound. Both volumes were clipped to $[0, p_{99}]$ and linearly rescaled to $[0, 1]$ using the same $p_{99}$.

**Lesion emphasis.** Lesion masks were available for all cohorts and were used to enhance lesion-related signal. The lesion masks were generated by a custom-designed segmentation model developed in-house, and all segmentation outputs were independently reviewed by two radiologists with an average of 7 years of experience to

ensure segmentation quality and clinical reliability. A voxel-wise weighting map $W$ was constructed by assigning a weight of 10 to voxels inside the lesion mask and 1 to voxels outside. This weighting map was applied multiplicatively to both the pre-contrast and early post-contrast volumes prior to channel concatenation.

**Radiology reports.** Unstructured radiology reports were retrieved for each patient (originally authored in Dutch) and standardized prior to model input. Newline characters were removed and boilerplate text was trimmed when present. In reports where a “Klinische gegevens” (“clinical information”) section preceded the narrative, only the text starting from the “Verslag” (“findings” / “report”) header was retained; the “Klinische gegevens” section typically contains the patient history and the indication for the examination. If no such section was present, the full report was used. The resulting text was then tokenized using a domain-specific RoBERTa tokenizer, with truncation and padding to a maximum sequence length of 512 tokens, yielding integer token IDs and an attention mask indicating non-padding tokens.

**Structured clinical variables.** Structured clinicopathological variables were extracted from the electronic health record and encoded as a fixed-length clinical feature vector. Variables included age (binned), family history, receptor status reflecting molecular phenotype (ER, PR, and HER2), histological subtype, and TNM staging information available at the pre-treatment and post-treatment time points. Categorical variables were encoded using one-hot representations, and continuous variables were discretized when appropriate (e.g., age bins). Treatment-related variables were represented as binary indicators capturing major therapy categories (neoadjuvant/adjuvant radiotherapy, chemotherapy, endocrine therapy, and anti-HER2 therapy), together with surgery-related flags. Missing values were handled conservatively using explicit “unknown” categories. For one-hot encoded fields, an additional “unknown” category was appended to each variable; when a field was unavailable, the corresponding “unknown” indicator was set to 1 (with all other categories set to 0). For therapy-related variables, we distinguished documented absence from missingness by encoding each therapy as a three-state representation (yes / no / unknown), where missing or undocumented treatments were assigned to the “unknown” state rather than being set to 0.

## Multimodal model architecture (“perception-to-cognition”)

We designed a biomimetic perception-to-cognition architecture to reflect the staged reasoning workflow in breast MRI interpretation. In contrast to conventional flat-fusion strategies that directly concatenate heterogeneous features, our framework adopts a cascaded, Transformer-based hierarchical fusion mechanism that performs (i) report-guided perceptual integration, aligning imaging evidence with radiologist-described semantics, followed by (ii) clinical-contextual cognition, modulating the integrated imaging–text representation according to the patient’s clinicopathological profile.

**Modality-Specific Feature Encoding.** Three specialized neural branches were deployed to project heterogeneous data streams into a unified high-dimensional latent space (d=256).

**Visual encoder.** Imaging features were extracted using a customized **3D ResNet-18** backbone to capture volumetric tumor heterogeneity[40]. The input layer was modified to accept dual-channel volumes (encoding distinct contrast phases) via a 7×7×7 convolutional kernel (stride 2). The network contains four stages of 3D residual blocks, followed by 3D adaptive average pooling and a linear projection to produce a 256-dimensional visual embedding.

**Semantic encoder.** Radiology reports were encoded using RadioLOGIC, a domain-specific RoBERTa model pre-trained on breast imaging reports[41]. To preserve knowledge learned during pre-training, the language model parameters were frozen during training. The contextual embedding corresponding to the [CLS] token was extracted and projected to the shared 256-dimensional latent space through a non-linear projection layer.

**Clinical encoder.** Structured clinicopathological variables were processed using a multilayer perceptron (Clinical Net) with Batch Normalization and Dropout (rate 0.2). This branch generated a compact 256-dimensional clinical embedding representing the patient's clinical profile.

**Hierarchical Transformer Fusion Strategy.** To integrate information from imaging, radiological reports, and clinical data, we designed a hierarchical fusion architecture based on Transformer encoders. The fusion process is structured in two sequential stages, progressively combining the different data modalities to generate a comprehensive patient-level representation for risk prediction.

A standardized Transformer-based fusion block was employed for both fusion stages. This block consists of a Transformer encoder with two stacked encoder layers. Each layer utilizes a multi-head self-attention mechanism with eight attention heads and a position-wise feed-forward network with a hidden dimensionality of 1024. A dropout rate of 0.4 was applied within each encoder layer for regularization. All feature embeddings and intermediate representations throughout the fusion process were maintained at a fixed dimensionality of 256.

**Stage 1: Report-Guided Visual-Semantic Alignment.** The initial stage aimed to create a unified representation of lesion characteristics by aligning visual features with their semantic descriptions from radiological reports. First, a 256-dimensional visual feature vector ($z_v$) was extracted from the imaging modality, and a corresponding 256-dimensional semantic vector ($z_t$) was derived from the text report.

These two embedding vectors were then structured as a sequence of two tokens. This sequence was processed by the first Transformer fusion block, where the self-attention mechanism enabled each representation to be updated based on the context provided by

the other. This process facilitates a guided refinement of the visual features under the supervision of the semantic context. The resulting two contextually aware output vectors were then aggregated into a single vector by computing their mean, yielding a unified visual-semantic representation ($z_{vt}$).

**Stage 2: Cognitive fusion (clinical-contextual integration).** The second stage integrated the patient's systemic clinical context with the fused visual-semantic representation derived from Stage 1. The 256-dimensional visual-semantic vector ($z_{vt}$) and a 256-dimensional clinical feature vector ($z_c$) were structured into a new two-token sequence. This sequence was subsequently processed by a second, architecturally identical Transformer fusion block. This step allows the model to modulate the prognostic relevance of the imaging and report findings based on the patient's broader clinical profile. Finally, the two output vectors from this block were aggregated via mean pooling to produce the final, comprehensive 256-dimensional patient-level multimodal representation.

**Prognostic Prediction and Optimization.** The final multimodal representation was aggregated via global average pooling and passed through a linear regressor to generate a continuous risk score. The entire pipeline was trained end-to-end by optimizing the negative log partial-likelihood (Cox) loss, ensuring that the learned features are directly discriminative for recurrence risk.

For RFS stratification analysis, patients were dichotomized into high- and low-risk groups using the median risk score derived from the internal validation set. This prespecified cutoff was applied to the internal test set and the external DUKE cohort. For the I-SPY1 cohort, application of the internal validation-set cutoff classified all patients as high risk; therefore, a cohort-specific median risk score was used for Kaplan–Meier stratification in this cohort.

## Text-free inference during external validation (zero-shot generalization)

To assess generalizability in settings where radiology reports are unavailable or linguistically incompatible with the training corpus (e.g., the Duke and I-SPY 1 cohorts), we evaluated a text-free inference protocol during external testing. Specifically, the semantic branch was disabled by replacing the text input with a masked/empty token (i.e., no report content was provided), while keeping the visual and clinical branches unchanged. Because the model was trained with cross-modal fusion, the visual representation learned to align with report-derived semantics during training, enabling the downstream fusion and risk head to remain functional when text is absent. Under this stress test, risk predictions were generated using visual and structured clinical features only, providing an estimate of model performance independent of explicit report guidance.

## Training and Statistical Analysis

All models were trained on an NVIDIA RTX A6000 GPU (48 GB VRAM) for up to 50 epochs with a batch size of 16. Optimization was performed using Adam with module-specific

learning rates (parameter groups) to stabilize multimodal training: the visual encoder was trained with $\mathrm{lr}_{\mathrm{image}} = 1 \times 10^{-3}$, while the remaining components (fusion/head, clinical branch, and report projection head) used $\mathrm{lr}_{\mathrm{fusion}} = \mathrm{lr}_{\mathrm{clinical}} = \mathrm{lr}_{\mathrm{report}} = 5 \times 10^{-4}$. Learning rates were decayed by a factor of 0.9 every 5 epochs. To mitigate overfitting, we applied early stopping based on the validation-set concordance index (C-index), and the checkpoint with the best validation C-index was selected for final evaluation.

Model performance was primarily assessed using the C-index, which measures the agreement between predicted risk scores and observed recurrence-free survival outcomes under right-censored data. For survival stratification, patients were dichotomized into high- and low-risk groups using the median risk score derived from the internal validation set. This prespecified cutoff was subsequently applied to the internal test set and the external Duke cohort. For the I-SPY1 cohort, application of the internal validation-set cutoff classified all patients as high risk, reflecting the substantially enriched risk profile of this cohort; therefore, a cohort-specific median risk score was used for Kaplan–Meier stratification in I-SPY1.

Kaplan–Meier survival curves were constructed for the resulting high- and low-risk groups, and differences between groups were evaluated using the log-rank test. Hazard ratios (HRs) and corresponding 95% confidence intervals (CIs) were estimated using Cox proportional hazards regression. To further assess prognostic performance over time, time-dependent receiver operating characteristic (ROC) analysis was performed, and the corresponding area under the curve (AUC) was calculated at multiple follow-up time points.

Univariable and multivariable Cox regression analyses were additionally performed to evaluate the prognostic value of conventional clinicopathological variables, including age, TN stage, and molecular subtype. For multimodal model comparisons, the same evaluation framework was applied across all candidate models and cohorts. Unless otherwise specified, all statistical tests were two-sided, with $P < 0.05$ considered significant. Analyses were performed using Python (v3.8, lifelines and scikit-survival libraries).

## Interpretability and Visual-Cognitive Alignment.

To examine how the model generates predictions, we applied interpretability analyses for both imaging and textual components.

**Visual attention analysis (Grad-CAM).** Grad-CAM was used to visualize the image regions within the 3D MRI volumes that contributed most strongly to the predicted risk score. Let $A \in \mathbb{R}^{H \times W \times D}$ denote the Grad-CAM activation map generated from the final convolutional layer of the visual encoder, where $H$, $W$, and $D$ are the spatial dimensions. The activation map was first upsampled to the original image resolution and normalized to the range $[0, 1]$. To quantify attention intensity, all voxel values in $A$ were sorted in descending order, and the attention score $S$ was defined as the mean activation value of the top 10% highest-activation voxels:

$$S = \frac{1}{K}\sum_{i=1}^{K} A_{(i)}\,, K = \lceil 0.1N \rceil,$$

where $N = H \times W \times D$ is the total number of voxels in the Grad-CAM map, $A_{(i)}$ denotes the $i$-th largest voxel activation value, and $\lceil \cdot \rceil$ denotes the ceiling operation. This score was used for statistical comparisons between risk groups and for correlation analyses with predicted risk scores.

**Semantic attention analysis.** To identify report tokens most strongly associated with model predictions, we extracted attention weights from the first-stage Transformer fusion module. Let $Q$, $K$, and $V$ denote the query, key, and value matrices derived from the fused visual–text token sequence. The attention weight matrix was computed as:

$$W = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right),$$

where $d_k$ is the key dimension. Token-level attention scores for report words were then obtained by aggregating the corresponding attention weights across attention heads and relevant token interactions. These scores were used to characterize visual–textual interactions and to assess whether clinically relevant radiological descriptors received greater attention in patients with higher predicted risk scores.

## Data availability

The external validation datasets analysed during the current study are publicly available in The Cancer Imaging Archive (TCIA) repository. The Duke Breast MRI dataset is accessible at *https://doi.org/10.7937/TCIA.e3sv-re93*, and the I-SPY1 dataset is accessible at *https://doi.orsg/10.7937/K9/TCIA.2016.HdHpgJLK*.

The internal NKI dataset is not publicly available due to European General Data Protection Regulation (GDPR) and patient privacy restrictions. However, de-identified data supporting the findings of this study may be made available from the corresponding author upon reasonable request and subject to the approval of the Institutional Review Board of the Netherlands Cancer Institute and the execution of a Data Use Agreement.

Source data underlying the main figures and tables are provided with this paper.

## Code availability

The codes of this paper are available at *https://github.com/Robin54223/MRI_RFS*. In addition to the complete modeling pipeline, the code for the segmentation model used to generate the masks will also be made publicly available.

**Reference:**


1. Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., & Jemal, A. (2025). Cancer statistics, 2025. Ca, 75(1), 10.
2. Duffy, S. W., Tabár, L., Yen, A. M. F., Dean, P. B., Smith, R. A., Jonsson, H., ... & Chen, T. H. H. (2021). Beneficial effect of consecutive screening mammography examinations on mortality from breast cancer: a prospective study. Radiology, 299(3), 541-547.
3. Gennari, A., André, F., Barrios, C. H., Cortes, J., de Azambuja, E., DeMichele, A., ... & Harbeck, N. (2021). ESMO Clinical Practice Guideline for the diagnosis, staging and treatment of patients with metastatic breast cancer☆. *Annals of oncology*, *32*(12), 1475-1495.
4. Loibl, S., André, F., Bachelot, T., Barrios, C. H., Bergh, J., Burstein, H. J., ... & Harbeck, N. (2024). Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up☆. *Annals of Oncology*, *35*(2), 159-182.
5. Piccart, M., van't Veer, L. J., Poncet, C., Cardozo, J. M. L., Delaloge, S., Pierga, J. Y., ... & Rutgers, E. J. (2021). 70-gene signature as an aid for treatment decisions in early breast cancer: updated results of the phase 3 randomised MINDACT trial with an exploratory analysis by age. The Lancet Oncology, 22(4), 476-488.
6. Schmid, P., Cortes, J., Pusztai, L., McArthur, H., Kümmel, S., Bergh, J., ... & O'Shaughnessy, J. (2020). Pembrolizumab for early triple-negative breast cancer. *New England Journal of Medicine*, *382*(9), 810-821.
7. Tutt, A. N., Garber, J. E., Kaufman, B., Viale, G., Fumagalli, D., Rastogi, P., ... & Geyer Jr, C. E. (2021). Adjuvant olaparib for patients with BRCA1-or BRCA2-mutated breast cancer. *New England Journal of Medicine*, *384*(25), 2394-2405.
8. Pan, H., Gray, R., Braybrooke, J., Davies, C., Taylor, C., McGale, P., ... & Hayes, D. F. (2017). 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *New England Journal of Medicine*, *377*(19), 1836-1846.
9. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., ... & Wolmark, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, *351*(27), 2817-2826.
10. Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., ... & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, *415*(6871), 530-536.
11. Giachetti, P. P. M. B., Schianca, A. C., Trapani, D., Marra, A., Toss, A., Marchiò, C., ... & Curigliano, G. (2025). Current controversies in the use of Oncotype DX in early breast cancer. *Cancer Treatment Reviews*, *135*, 102887.
12. Coles, C. E., Earl, H., Anderson, B. O., Barrios, C. H., Bienz, M., Bliss, J. M., ... & Zikmund-Fisher, B. (2024). The lancet breast cancer commission. *The Lancet*, *403*(10439), 1895-1950.
13. Dutta, R., Vallurupalli, M., McVeigh, Q., Huang, F. W., & Rebbeck, T. R. (2023). Understanding inequities in precision oncology diagnostics. *Nature Cancer*, *4*(6), 787-794.
14. Bianchini, G., De Angelis, C., Licata, L., & Gianni, L. (2022). Treatment landscape of

triple-negative breast cancer—expanded options, evolving needs. Nature reviews Clinical oncology, 19(2), 91-113.

**15.** Volinsky-Fremond, S., Horeweg, N., Andani, S., Barkey Wolf, J., Lafarge, M. W., de Kroon, C. D., ... & Bosse, T. (2024). Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nature medicine*, *30*(7), 1962-1973.

**16.** Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering, 5(6), 555-570.

**17.** Chen, R. J., Lu, M. Y., Williamson, D. F., Chen, T. Y., Lipkova, J., Noor, Z., ... & Mahmood, F. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer cell, 40(8), 865-878.

**18.** Hoang, D. T., Dinstag, G., Shulman, E. D., Hermida, L. C., Ben-Zvi, D. S., Elis, E., ... & Ruppin, E. (2024). A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. *Nature cancer*, *5*(9), 1305-1317.

**19.** Yan, C., Li, L., Qian, X., Ou, Y., Huang, Z., Ruan, Z., ... & Liu, J. (2025). Predicting MammaPrint Recurrence Risk from Breast Cancer Pathological Images Using a Weakly Supervised Transformer. *Advanced Science*, e10307.

**20.** Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V., & Madabhushi, A. (2019). Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, *16*(11), 703-715.

**21.** Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., ... & Clozel, T. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, *25*(10), 1519-1525.

**22.** Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., ... & Campbell, P. J. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature medicine*, *21*(7), 751-759.

**23.** Glaser, A. K., Reder, N. P., Chen, Y., McCarty, E. F., Yin, C., Wei, L., ... & Liu, J. T. (2017). Light-sheet microscopy for slide-free non-destructive pathology of large clinical specimens. *Nature biomedical engineering*, *1*(7), 0084.

**24.** Zhang, A., Miao, K., Sun, H., & Deng, C. X. (2022). Tumor heterogeneity reshapes the tumor microenvironment to influence drug resistance. *International journal of biological sciences*, *18*(7), 3019.

**25.** Ramón y Cajal, S., Sesé, M., Capdevila, C., Aasen, T., De Mattos-Arruda, L., Diaz-Cano, S. J., ... & Castellví, J. (2020). Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of Molecular Medicine*, *98*(2), 161-177.

**26.** Mann, R. M., Kuhl, C. K., & Moy, L. (2019). Contrast-enhanced MRI for breast cancer screening. *Journal of Magnetic Resonance Imaging*, *50*(2), 377-390.

**27.** Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature medicine*, *28*(9), 1773-1784.

**28.** Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, *616*(7956), 259-265.

**29.** Rudin, C. (2019). Stop explaining black box machine learning models for high stakes

decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.

**30.** Chitalia, R., Pati, S., Bhalerao, M., Thakur, S. P., Jahani, N., Belenky, V., ... & Bakas, S. (2022). Expert tumor annotations and radiomics for locally advanced breast cancer in DCE-MRI for ACRIN 6657/I-SPY1. *Scientific data*, *9*(1), 440.

**31.** Hylton, N. M., Gatsonis, C. A., Rosen, M. A., Lehman, C. D., Newitt, D. C., Partridge, S. C., ... & ACRIN 6657 Trial Team and I-SPY 1 TRIAL Investigators. (2016). Neoadjuvant chemotherapy for breast cancer: functional tumor volume by MR imaging predicts recurrence-free survival—results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL. *Radiology*, *279*(1), 44-55.

**32.** Saha, A., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., & Mazurowski, M. A. (2018). A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *British journal of cancer*, *119*(4), 508-516.

**33.** Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).

**34.** Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv. *arXiv preprint arXiv:1711.05225*, *10*.

**35.** Meattini, I., Tecic Vuger, A., Becherini, C., Epstein, E. J., Garcia, J. P., Berger, E. R., & Harbeck, N. (2025). Multimodal de-escalation strategies in early breast cancer. *American Society of Clinical Oncology Educational Book*, *45*(3), e473462.

**36.** Sparano, J. A., Gray, R. J., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F., ... & Sledge Jr, G. W. (2018). Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, *379*(2), 111-121.

**37.** Cardoso, F., van't Veer, L. J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., ... & Piccart, M. (2016). 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *New England Journal of Medicine*, *375*(8), 717-729.

**38.** Kushol, R., Parnianpour, P., Wilman, A. H., Kalra, S., & Yang, Y. H. (2023). Effects of MRI scanner manufacturers in classification tasks with deep learning models. *Scientific Reports*, *13*(1), 16791.

**39.** Schouten, D., Nicoletti, G., Dille, B., Chia, C., Vendittelli, P., Schuurmans, M., ... & Khalili, N. (2025). Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications. *Medical image analysis*, *105*, 103621.

**40.** Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 6546-6555).

**41.** Zhang, T., Tan, T., Wang, X., Gao, Y., Han, L., Balkenende, L., ... & Mann, R. M. (2023). RadioLOGIC, a healthcare model for processing electronic health records and decision-making in breast disease. *Cell Reports Medicine*, *4*(8).

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Thesupplementaryfile.pdf
- rawdata.xlsx