The supplementary file of “Decoding Tumor Phenotype: A Radiologist-Inspired Deep Learning System for Breast Cancer Recurrence Prediction.”

Distribution of clinicopathological characteristics across cohorts

Supplementary Fig. S1 presents the distribution of clinicopathological characteristics across the multi-center cohorts used in this study, including the NKI development cohort, the Duke external validation cohort, and the I-SPY1 external validation cohort. Stacked bar charts illustrate the proportions of recurrence and recurrence-free cases across categorical clinical variables, while box plots summarize the distributions of continuous variables such as patient age and time-to-event. Recurrence cases are indicated in red and recurrence-free cases in blue.

Statistical comparisons between recurrence and recurrence-free groups were performed using the Chi-square test for categorical variables and the Mann–Whitney U test for continuous variables. Several established clinical risk factors, including tumor stage (T stage) and nodal status (N stage), showed statistically significant associations with recurrence in univariable analyses in certain cohorts. However, a notable proportion of patients within clinically high-risk categories (e.g., higher T stage or node-positive disease) remained recurrence-free, suggesting that conventional clinicopathological variables alone provide limited discrimination for recurrence risk at the individual patient level.

In addition, the distributions of both recurrence prevalence and clinical characteristics varied across the three cohorts. For example, the I-SPY1 cohort exhibited a relatively higher proportion of recurrence events compared with the other cohorts. These differences highlight the presence of inter-cohort heterogeneity and potential distributional shifts between institutions, underscoring the importance of evaluating model robustness and generalizability across independent datasets.

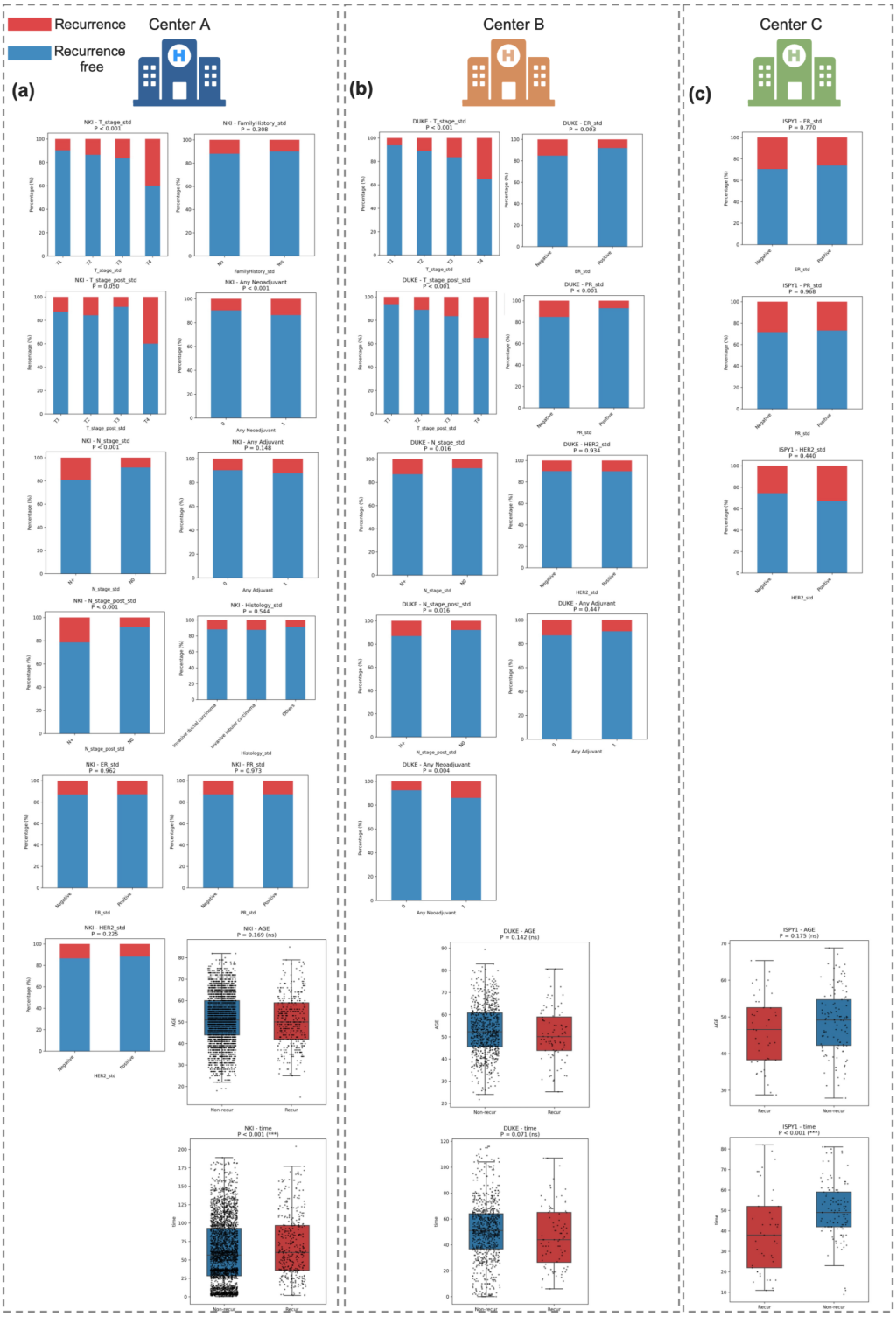

Fig. S1 | Clinicopathological characteristics and outcome distributions across multi-center cohorts. Stacked bar charts and box plots displaying the distribution of key clinical variables stratified by recurrence status (Red: Recurrence; Blue: Recurrence-free) for the NKI development cohort (a), Duke external validation cohort (b), and I-SPY1 external validation cohort (c). P-values were calculated using the Chi-square test for categorical

variables and the Mann-Whitney U test for continuous variables. While univariable analyses confirm that standard risk factors (e.g., T-stage, N-stage) are statistically associated with recurrence ($P < 0.05$), the substantial proportion of recurrence-free survivors within high-risk groups (e.g., T3/N+) illustrates the limited distinctiveness of clinical variables alone. Furthermore, the varying prevalence of recurrence across centers (e.g., higher event rate in I-SPY1) highlights the distributional shift challenges addressed by the AI model.

Subgroup survival analyses performed in the Duke external validation cohort

Supplementary Fig. S2 presents subgroup survival analyses performed in the Duke external validation cohort. Patients were stratified into model-predicted low-risk and high-risk groups based on the median risk score derived from the training cohort. Kaplan–Meier survival curves were then generated within clinically relevant subgroups, including age (<50 vs ≥50 years), molecular subtype (HR-/HER2-, TNBC, HER2+), treatment strategy (primary surgery vs neoadjuvant therapy), and pathological staging categories (T1 vs T2 stage and N0 vs N+ stage).

Across most clinical strata, the model-predicted risk groups remained significantly associated with recurrence-free survival, demonstrating consistent risk stratification performance in the independent Duke dataset. Notably, even within conventionally defined clinical categories, substantial heterogeneity in recurrence outcomes was observed between the predicted low- and high-risk groups. This finding suggests that the imaging-based model captures prognostic information beyond standard clinicopathological factors.

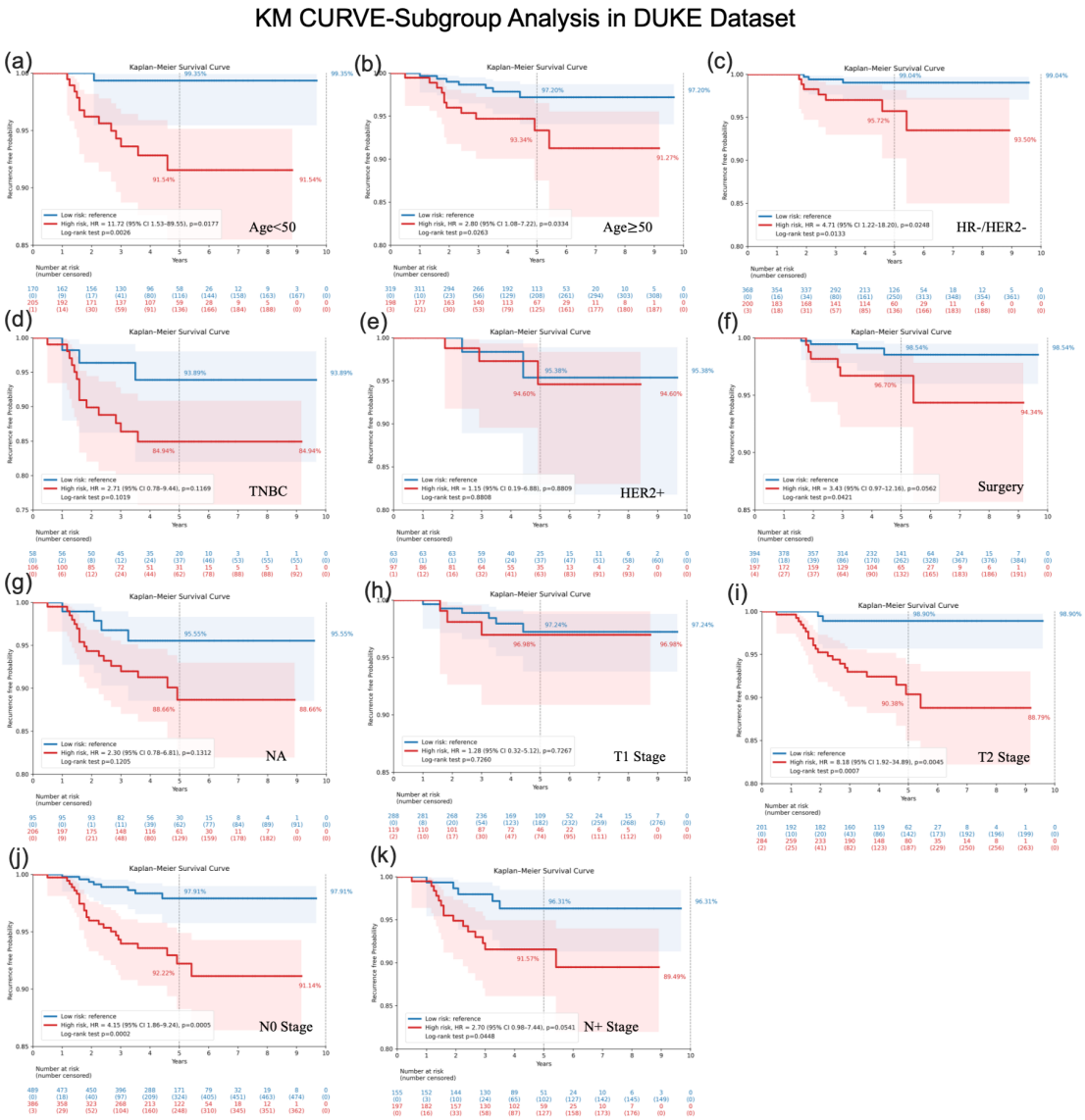

Fig. S2 | Stratified Kaplan-Meier analysis demonstrating model robustness across clinically relevant subgroups in the NKI cohort. Recurrence-free survival curves stratified by the AI Risk Score (Red: High-risk; Blue: Low-risk) within distinct patient subgroups. **a-j,** Subgroup analyses performed on 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 and T2, T3, T4. Nodal Status: N0 and N+.

Visual-cognitive alignment and interpretability of the multimodal framework from the external validation cohorts

Supplementary Fig. S3 shows representative visualization examples from the external validation cohorts (Duke and I-SPY1). For each case, the original DCE-MRI image, lesion segmentation, and corresponding model attention map are displayed. Low-risk and high-risk examples are presented separately.

In both external datasets, the model attention maps predominantly localize to tumor regions defined by the lesion masks. The spatial attention patterns observed in these cohorts are consistent with those observed in the development dataset, suggesting that the model focuses on similar lesion-related imaging features across different institutions.

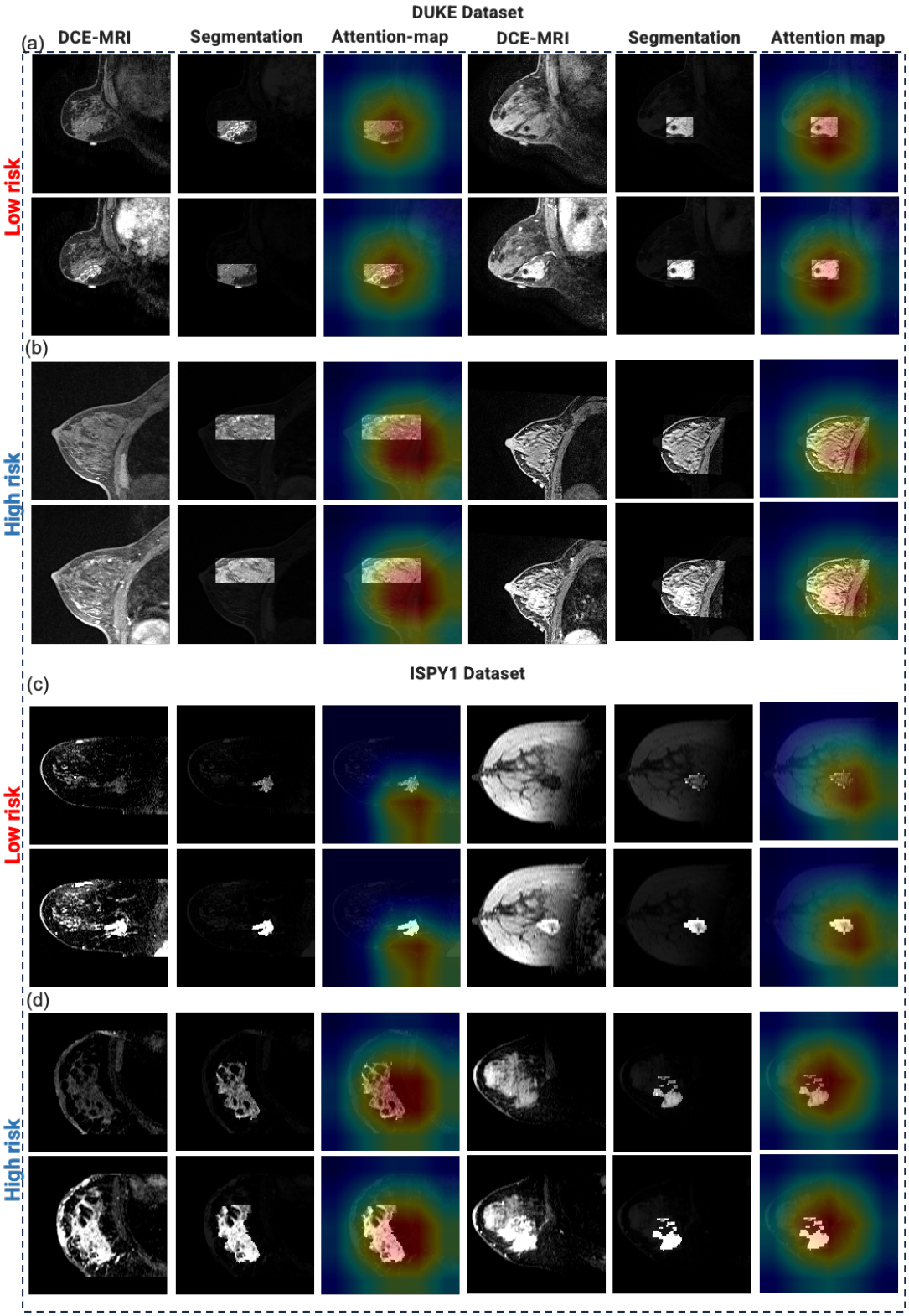

Fig. S3 | Cross-center visual explainability and phenotypic characterization in external validation cohorts. Representative examples of model interpretability applied to the Duke (Top panel) and I-SPY1 (Bottom panel) external validation datasets.

Distribution of predicted risk scores and associated clinicopathological characteristics within I-SPY1 cohort

Supplementary Fig. S4 illustrates the distribution of AI-derived risk scores in the independent I-SPY1 cohort, which served as an external test dataset. Patients are ranked according to the predicted risk score generated by the model. Recurrence events are indicated by black circles. The vertical dashed line represents the median risk score used to divide the cohort into low-risk and high-risk groups.

Despite the relatively small cohort size and the higher overall recurrence prevalence characteristic of the I-SPY1 neoadjuvant trial population, recurrence events are predominantly concentrated toward the higher end of the predicted risk spectrum. The lower panel displays the distribution of available clinicopathological variables across the ranked patients, including treatment strategy, age group, histological subtype, hormone receptor status, HER2 status, and systemic therapies. Many of these variables are either relatively homogeneous or partially missing within this cohort.

The observed enrichment of recurrence events within the model-defined high-risk region suggests that the imaging-derived risk score retains prognostic relevance in this independent dataset, despite differences in cohort composition and clinical variable availability compared with the development cohort.

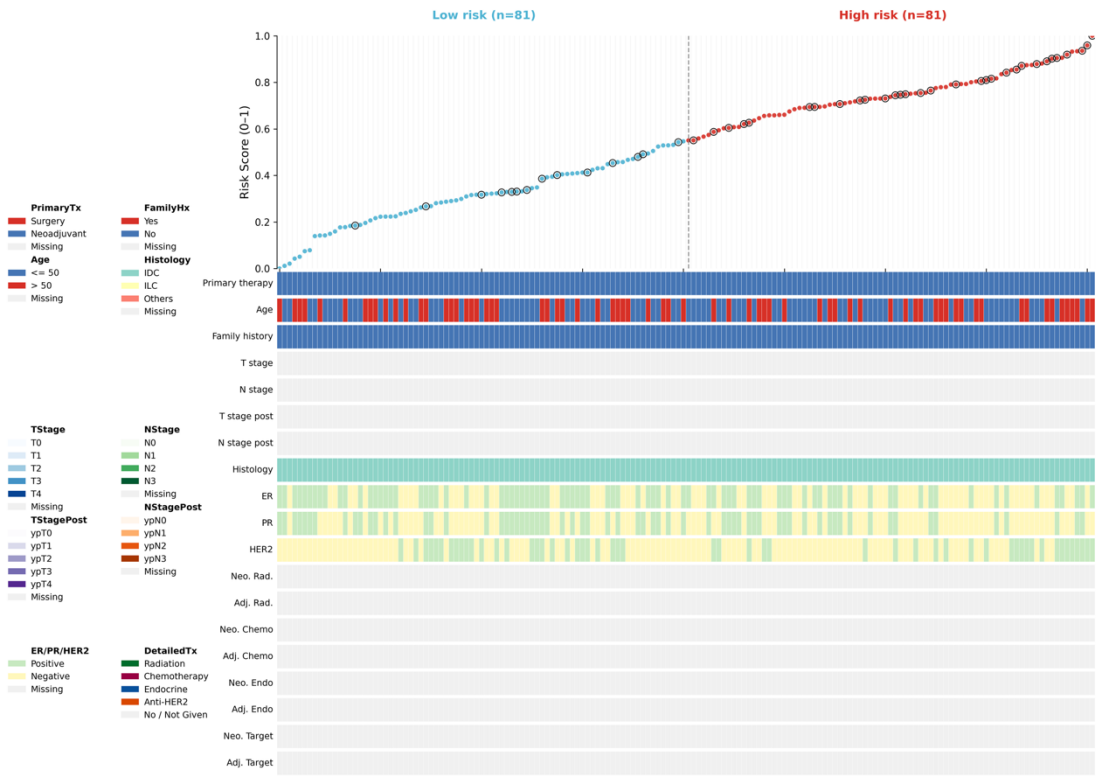

Fig. S4 | Risk landscape in the high-risk I-SPY1 neoadjuvant trial cohort. Patients are

ranked by AI Risk Score. Despite the cohort's high baseline risk (reflected in the higher overall density of recurrence events, black circles) and relatively small sample size, the model successfully enriches events towards the high-risk spectrum (right). The ability of the model to achieve risk stratification in this homogenous, data-scarce setting—where standard clinical discriminators are either uniform or unavailable—confirms that the prognostic signal is primarily driven by intrinsic imaging phenotypes captured by the Visual Encoder, validating the robustness of the "Zero-Shot Modality Dropout" protocol.

Detailed patient inclusion and exclusion criteria across cohorts

This study included one internal cohort for model development and validation, and two independent external cohorts for testing. The detailed patient inclusion and exclusion criteria for each cohort are presented in Supplementary Figure S5.

Development Cohort (Center A)

The primary cohort for model development, selection, and internal testing was retrospectively established from the Netherlands Cancer Institute (NKI) database. We initially identified 10,336 patients diagnosed with breast cancer between January 2007 and April 2024. Patients were excluded based on the following criteria: absence of pre-treatment Magnetic Resonance Imaging (MRI) (n=1,203), missing clinical information required for the model (n=1,801), or inadequate follow-up data for recurrence analysis (n=4,057). After applying these criteria, a final cohort of 3,266 patients remained for the analysis. This cohort was partitioned into a training set (n=1,306), a validation set used for hyperparameter tuning and model selection (n=980), and an internal test set for performance evaluation (n=980).

External Test Cohort 1 (Center B)

The first external test cohort was derived from the publicly available DUKE dataset, comprising patients treated between 2000 and 2014. From an initial set of 922 patients, 30 were excluded due to the lack of follow-up data. The final cohort for external testing consisted of 892 patients.

External Test Cohort 2 (Center C)

The second external test cohort was sourced from the publicly available ISPY1 dataset, which also included patients from 2000 to 2014. The initial cohort consisted of 222 patients. We excluded 9 patients without follow-up data and 51 patients who lacked tumor annotations necessary for feature extraction. This resulted in a final cohort of 162 patients for the second external test.

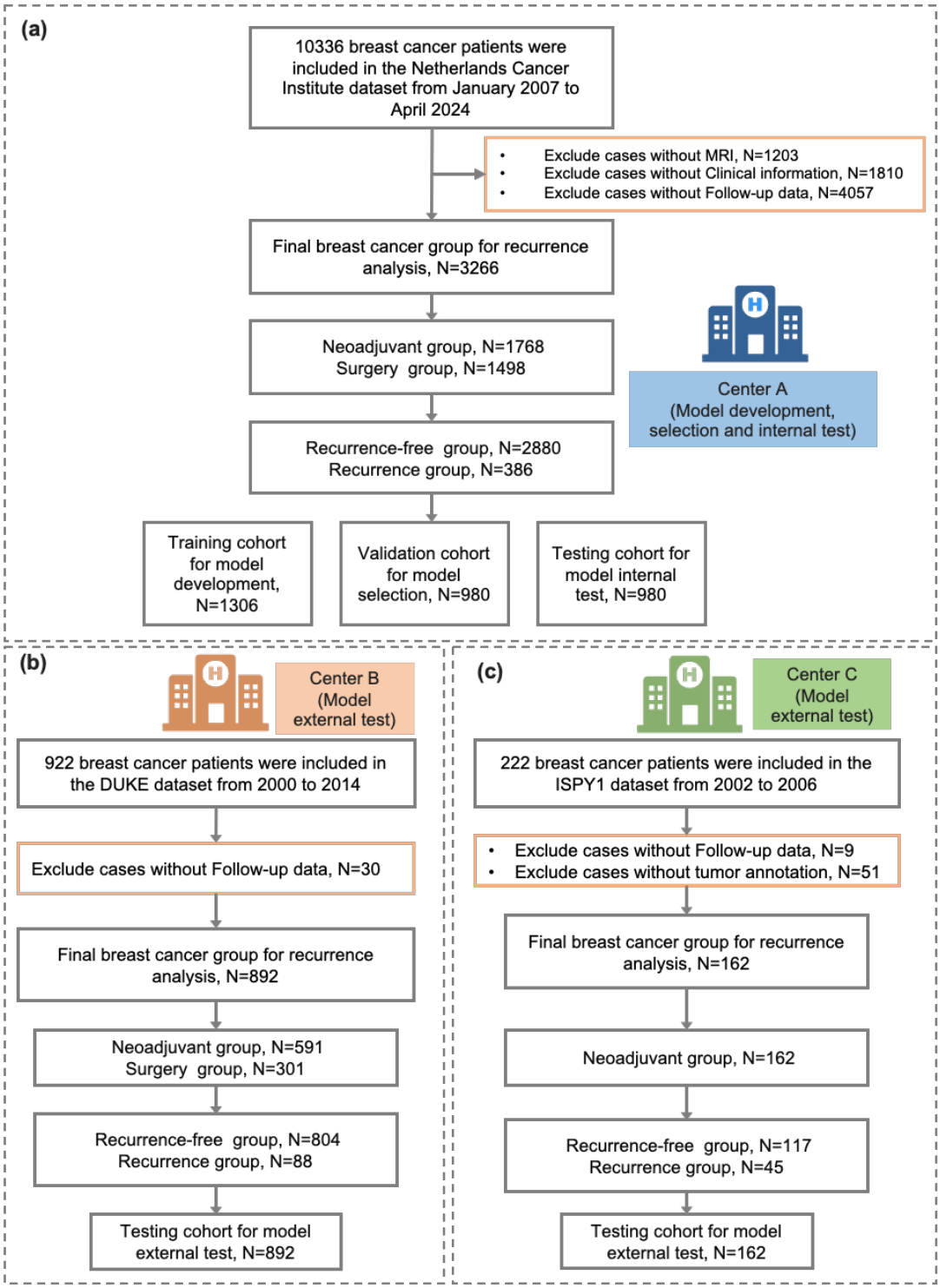

Fig. S5 | CONSORT flow diagrams illustrating patient inclusion, exclusion, and cohort stratification. a, Derivation of the primary development cohort (Center A: NKI). A total of 10,336 breast cancer patients diagnosed between January 2007 and April 2024 were initially screened. Following stringent quality control, 7,061 patients were excluded due to missing pre-treatment MRI (N=1,203), unavailable clinical information (N=1,801), or insufficient follow-up data (N=4,057). The final NKI cohort comprised 3,266 patients, including both neoadjuvant (N=1,768) and upfront surgery (N=1,498) groups. This dataset

was randomly split into independent training (N=1,306), validation (N=980), and internal testing (N=980) sets for model development and hyperparameter tuning. b, Selection of the Duke external validation cohort (Center B). Out of 922 patients screened from the Duke dataset (2000–2014), 30 were excluded for missing follow-up. The final cohort consisted of 892 patients (591 neoadjuvant, 301 surgery), serving as an independent test set to evaluate cross-center generalizability. c, Selection of the I-SPY1 external validation cohort (Center C). From the I-SPY1 clinical trial dataset (2002–2006), 60 patients were excluded due to missing follow-up or lack of tumor annotations. The final cohort comprised 162 high-risk patients, all of whom underwent neoadjuvant chemotherapy, providing a rigorous testbed for the model's performance in a specialized trial setting.

MRI acquisition parameters of the in-house and external datasets

DCE-MRI data were acquired at three centers with center-specific imaging protocols (Table S1). The in-house NKI cohort was acquired on 1.5-T and 3-T Philips scanners, with a repetition time (TR) of 3.66–6.33 ms, an echo time (TE) of 1.72–2.30 ms, and a flip angle of 10–12°. The in-plane spatial resolution was ≤1.0 × 1.0 mm, with a slice thickness of 1.20–1.80 mm. The public Duke cohort was acquired on 1.5-T and 3-T GE and Siemens scanners, with TR ranging from 4.1 to 6.0 ms, TE from 1.2 to 2.5 ms, and flip angle from 7 to 12°. The in-plane resolution ranged from 0.6 × 0.6 mm to 1.0 × 1.0 mm, and slice thickness ranged from 1.1 to 2.5 mm. The public ISPY-1 cohort included scans acquired on 1.5-T GE, Siemens, and Philips systems, with TR ≤20 ms, TE of approximately 4.5 ms, and flip angle ≤45°. The in-plane resolution was ≤1.0 × 1.0 mm and the slice thickness was ≤2.5 mm.

Table S1. MRI acquisition parameters of the in-house and external datasets. **Note:** DCE-MRI = dynamic contrast-enhanced magnetic resonance imaging; TR = repetition time; TE = echo time. Values are presented as ranges, approximate values, or upper limits according to the acquisition protocols reported by each center/dataset. NKI represents the in-house dataset, whereas Duke and ISPY-1 represent publicly available external datasets.

Parameter	NKI (in-house)	Duke (public, external)	ISPY-1 (public, external)
Vendor	Philips	GE, Siemens	GE, Siemens,Philips
Magnetic Field Strength	1.5 T, 3 T	1.5 T, 3 T	1.5 T
Sequence	DCE-MRI	DCE-MRI	DCE-MRI
TR (ms)	3.66 - 6.33	4.1 - 6.0	≤ 20
TE (ms)	1.72 - 2.30	1.2 - 2.5	~4.5
Flip angle (°)	10 - 12	7 - 12	≤ 45
In-plane resolution (mm)	≤ 1.0 × 1.0	0.6 × 0.6 - 1.0 × 1.0	≤ 1.0 × 1.0
Slice thickness (mm)	1.20 - 1.80	1.1 - 2.5	≤ 2.5