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.

Interpretable clinicogenomic model for risk-adapted therapeutic intervention in metastatic breast cancer

Pedram Razavi

razavip@mskcc.org

Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0003-4236-0576

Luca Boscolo-Bielo

Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0003-1026-3923

Francisco Sanchez-Vega

Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-2990-4893

Subhiksha Nandakumar

Memorial Sloan Kettering Cancer Center

Antonio Marra

European Institute of Oncology IRCCS https://orcid.org/0000-0002-7310-7824

Enrico Moiso

Memorial Sloan Kettering Cancer Center

Emanuela Ferraro

European Institute of Oncology IRCCS

Ino de Bruijn

Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-5427-4750

Karl Pichotta

Memorial Sloan Kettering Cancer Center

Bianca Malagutti

European Institute of Oncology IRCCS

Anton Safonov

Memorial Sloan Kettering Cancer Center

Michele Waters

Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-5416-5172

Julia Ah-Reum An

Memorial Sloan Kettering Cancer Center

Maureen Carey

Tempus AI Inc.

Sameet Sreenivasan

Clinical Data Sciences, Oncology R&D, AstraZeneca Pharmaceuticals LP

Gráinne Long
Clinical Data Sciences, Oncology R&D, AstraZeneca UK Ltd

Sajan Khosla
AstraZeneca

Ingrid Mayer
Vanderbilt University

Maurizio Scaltriti
Astra-Zeneca https://orcid.org/0000-0002-5522-1447

Nicola Fusco
Department of Oncology and Hemato-Oncology, University of Milan https://orcid.org/0000-0002-9101-9131

Elisabetta Munzone
European Institute of Oncology https://orcid.org/0000-0003-3371-3878

Fresia Pareja
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0003-3748-8049

Dara Ross
Memorial Sloan Kettering Cancer Center

Walid Chatila
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-5474-218X

Michael Berger
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0003-3882-5000

David Solit
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-6614-802X

Shanu Modi
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-6427-7373

Giuseppe Curigliano
European Institute of Oncology, IRCCS. Department of Oncology and Hemato-Oncology, University of Milano https://orcid.org/0000-0003-1781-2518

Mark Robson
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-3109-1692

Sohrab Shah
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-6402-523X

Sarat Chandarlapaty
Memorial Sloan Kettering Cancer Center, New York, NY, USA. https://orcid.org/0000-0003-4532-8053

Luc Cabel
Institut Curie https://orcid.org/0000-0001-5515-9180

Mark Donoghue
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0003-4538-6069

Yixiao Gong
Memorial Sloan Kettering Cancer Center, New York, NY, USA.
Matthew Zatzman
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-5235-4116
Nikolaus Schultz
Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-0131-4904

Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10176871/v2

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

Additional Declarations: The authors declare potential competing interests as follows: E.F. has received personal fee as speaker from AstraZeneca and Lilly. A.M. has received institutional grants or funding from Gilead Sciences; reports personal fee as speaker, consultant, or advisor from Menarini/Stemline, Roche, Lilly, AstraZeneca, Daiichi-Sankyo, Pfizer, StratifAI, outside the submitted work; reports travel support from Menarini/Stemline, AstraZeneca, Lilly, Pfizer, Daiichi-Sankyo. S.S., G.L., S.K., I.M., and M.S. are currently employed by AstraZeneca. M.C. is currently employed by Tempus. N.F. reports consulting or advisory roles with Abbvie, Alira Health, AstraZeneca, Biocartis, Epredia, Gilead, Menarini Group, Merck, MSD, Novartis, Owkin, Pfizer, Roche, Sakura, Signatur Bio, Sysmex, and Veracyte; reports personal fees as a speaker from AstraZeneca, Daiichi Sankyo, Epredia, Exact Sciences, Gilead, GSK, Johnson&Johnson, Leica Biosystems, Lilly, Menarini Group, MSD, Novartis, Roche, Sysmex, ThermoFisher, and Veracyte; reports research support from AstraZeneca, Gilead, GSK, Menarini, Novartis, Owkin, Pfizer, Pillar Biosciences, and Roche; and reports travel support from AstraZeneca, Novartis, and Roche. D.S.R. reports personal fees from AstraZeneca. M.F.B. received personal fees from AstraZeneca and possesses intellectual property rights with SOPHiA Genetics. D.B.S. has consulted/received honoraria from Pfizer, BridgeBio, Scorpion Therapeutics, FORE Therapeutics, Corramedical, Inc., and Meliora Therapeutics, Inc. G.C. has received institutional grants and/or funding from Astellas, AstraZeneca, Blueprint Medicine, BMS, Daiichi Sankyo, Kymab, Merck, Novartis, Philogen, Relay Therapeutics, Roche, and Sanofi; and consultation, ad board, or honoraria as a speaker or writer from AstraZeneca, BMS, Daiichi Sankyo, Exact Sciences, Gilead, Lilly, Menarini, Merck, Novartis, Pfizer, Roche, and Veracyte; reports non-financial leadership or advisory roles with ESMO, Europa Donna, EUSOMA, and Fondazione Beretta. M.E.R. reports honoraria from Research to Practice, Intellisphere, Physicians' Education Resource, MyMedEd, MJH Holdings; reports uncompensated consulting or advisory work with Artios Pharma, AstraZeneca, Exact

Sciences, Foundation Medicine, Merck; reports institutional research funding from Artios Pharma, AstraZeneca, Merck. S.C. has received institutional grant/funding from Pathos.ai and AstraZeneca, consultation/Ad board/Honoraria from Daiichi-Sankyo, AstraZeneca, Novartis, Casdin Capital, Nuvalent, Merck, Bayer, HighTop, PathosAI, and Revelio. P.R. has received institutional grants and/or funding from Grail, Novartis, AstraZeneca, Invitae, Biotheranostics, Tempus, Neogenomics, Guardant Health, Personalis, Myriad Genetics, Foresight Diagnostics, Natera, Biodesix, SAGA Diagnostics, SOPHiA Genetics, Haystack and Roche; and consultation, ad board or honoraria from Novartis, AstraZeneca, Lilly/Loxo, Stemline Therapeutics, Prelude Therapeutics, Neogenomics, Regor Pharmaceuticals, SAGA Diagnostics, SOPHiA Genetics, Tempus, Puma, Myriad Gen

Interpretable clinicogenomic model for risk-adapted therapeutic intervention in metastatic breast cancer

Luca Boscolo-Bielo[1,2,3], Francisco Sanchez-Vega[4]*, Subhiksha Nandakumar[4]*, Antonio Marra[3,5]*, Karl Pichotta[4], Enrico Moiso[2], Emanuela Ferraro[5], Ino de Bruijn[4], Luc Cabel[2], Bianca Malagutti[3,5], Anton Safonov[1,6], Michele Waters[4], Julia Ah-Reum An[2], Mark Donoghue[4], Yixiao Gong[4], Matthew Zatzman[4], Maureen Carey[7], Sameet Sreenivasan[8], Gráinne Long[9], Sajan Khosla[9], Ingrid Mayer[8], Maurizio Scaltriti[8], Nicola Fusco[3,5], Elisabetta Munzone[5], Fresia Pareja[10], Dara S. Ross[10], Walid K. Chatila[4], Michael F. Berger[10], David B. Solit[1,6,11], Shanu Modi[1], Nikolaus Schultz[4], Mark E. Robson[1,6], Giuseppe Curigliano[3,5], Sohrab Shah[4], Sarat Chandarlapaty[1,6,11#], Pedram Razavi[1,2,6#]

Affiliations

[1] Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, United States
[2] Translational Oncology Partnership Program, Memorial Sloan Kettering Cancer Center, New York, United States
[3] Department of Oncology and Hematology-Oncology, University of Milan, Milan, Italy
[4] Department of Epidemiology and Biostatistics, Computational Oncology Service, Memorial Sloan Kettering Cancer Center, New York, United States
[5] European Institute of Oncology IRCCS, Milan, Italy
[6] Weill Cornell Medical College, New York, United States
[7] Tempus AI Inc., Chicago, IL, United States
[8] AstraZeneca, Gaithersburg, MD, United States
[9] AstraZeneca, Cambridge, United Kingdom
[10] Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, United States
[11] Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, United States

* Equal secondary contributions

[#]Co-corresponding author: Sarat Chandarlapaty; chandars@mskcc.org
[#]Corresponding author: Pedram Razavi; razavip@mskcc.org

SUMMARY

Criteria guiding treatment intensification in oncology largely predate broad molecular profiling. This is exemplified in hormone receptor-positive metastatic breast cancer, where outcomes on first-line CDK4/6 inhibitors plus endocrine therapy (CDK4/6i+ET) vary widely, yet treatment escalation remains guided by crude clinical criteria. Drug resistance is difficult to forecast, as it arises through diverse tumor-intrinsic and host mechanisms, many occurring in only a minority of tumors. To better anticipate outcomes, we developed a machine-learning (ML) framework (ReCAST) integrating clinicopathologic and genomic features while retaining the contribution of rare yet mechanistically important genomic events. ReCAST benchmarked favorably against state-of-the-art ML architectures and foundation models designed for tabular data while retaining clinical interpretability. We applied ReCAST to develop CDKPredict, a model predicting outcomes on first-line CDK4/6i+ET trained on 1,264 densely annotated patients and validated in an independent multi-institutional cohort of 564. CDKPredict stratified patients into three risk groups spanning early treatment failure to durable disease control, outperformed current criteria for therapeutic escalation and held across clinically relevant and biomarker-selected populations. Stratification showed strong clinical and biological plausibility, with high-risk tumors enriched for actionable biomarkers linked to CDK4/6i+ET resistance and featuring reduced expression of estrogen-regulated transcriptional programs. Finally, a scalable version requiring a restricted set of features for resource-variable settings was validated across two external cohorts spanning 1,019 patients and made publicly available. Together, these findings provide a prognostic backbone for the prospective evaluation of risk-adapted, biomarker-informed therapy in breast cancer, and a generalizable approach for outcome prediction where clinicogenomic data can be sparse and clinical interpretability represents a priority.

INTRODUCTION

Risk-adapted therapeutic interventions have long been pursued in early-stage, hormone receptor-positive/HER2-negative (HR+/HER2–) breast cancer. A cornerstone of this paradigm was the implementation of multigene assays, which were initially developed as robust prognostic tools to estimate the risk of distant recurrence and were much later incorporated into prospective trials, thereby gaining predictive utility in guiding the use of adjuvant chemotherapy[1-5]. These approaches were subsequently refined to individualize risk estimates through integration with patient-specific clinicopathologic features[6-8].

The implementation of similar risk-adapted strategies in metastatic breast cancer (MBC) has been challenging. Despite significant improvements in progression-free survival (PFS) and overall survival with first-line CDK4/6 inhibitors plus endocrine therapy (CDK4/6i+ET), clinical outcomes remain highly heterogeneous, with approximately 20% of patients experiencing rapid progression, and a similar proportion achieving durable disease control[9-12]. We and others have characterized a wide array of resistance mechanisms underlying this variability. Within the cell-cycle machinery, these include *RB1* loss[13], overexpression of *CCNE1*[14,15], aberrant Hippo signaling (*FAT1*/YAP-TAZ) resulting in *CDK6* overexpression[16,17], *TP53*/*MDM2*-induced and DREAM/CDK2-mediated cell-cycle re-entry[18], *AURKA* overexpression[19], and *MYC* amplification driving resistance through KLHL42-mediated pRB1 degradation[20]. Other drivers of resistance to CDK4/6i and/or ET include alterations in *PTEN*, *FGFR1*, *ESR1*, *ERBB2*, *NF1*/MAPK, among other regulators of endocrine and growth-factor signaling[21-30]. Beyond individual biomarkers, several mechanisms of resistance have been identified, including germline *BRCA2*-mediated homologous recombination deficiency (HRD)[13,31], APOBEC3 mutagenesis[32], elevated tumor mutational burden (TMB)[32,33], as well as certain copy-number signatures resulting in *RB1* hemizygosity as determinants of poor outcomes[13,15,34]. Collectively, these overlapping and often competing mechanisms are individually infrequent yet common in aggregate, each carrying a large effect in the patients who harbor it, underscoring the limited power of any standalone biomarker to forecast treatment outcomes and the need for a multimodal approach for accurate baseline risk stratification.

This clinical need is further underscored by the expanding array of intensified therapeutic strategies being evaluated in the first-line setting for HR+/HER2– MBC to tackle resistant tumors. Notably, the addition of targeted therapies, such as PI3K-alpha-selective inhibitors, to standard CDK4/6i+ET has proven to be effective[35], with several ongoing phase III trials investigating such therapeutic approaches (NCT06380751, NCT06790693, NCT06065748, NCT04862663,

NCT04711252, NCT04546009). However, in the absence of robust and reliable risk stratification tools, high-risk patients are currently identified using historic and crude prognostic criteria, namely treatment-free interval (TFI) ≤12 months, defined as metastatic recurrence during or within 12 months from the end of adjuvant ET, that fail to recapitulate the full array of factors promoting endocrine and/or CDK4/6i resistance. Such criteria can lead to a significant misclassification problem. Tumors fulfilling historic high-risk criteria may retain endocrine sensitivity, resulting in overtreatment and avoidable health-related and financial toxicities. Conversely, biologically aggressive tumors lacking overt high-risk clinical features might be exposed to standard treatments, resulting in short-term clinical benefits, treatment attrition, and poorer cancer-related outcomes.

To address this, we developed ReCAST (Regularized Clinicogenomic Adaptive Survival Tool), a machine-learning framework designed to integrate clinicopathologic and genomic features for outcome prediction while preserving signal from rare but clinically relevant alterations. Applying this framework, we developed CDKPredict, an integrated clinicogenomic prognostication model trained on a large cohort of patients who received routine care at Memorial Sloan Kettering Cancer Center (MSK) and rigorously validated across multiple independent institutions and clinical settings. Our aim was to establish a pragmatic and interpretable baseline risk stratification framework to inform first-line risk-adapted therapeutic interventions in HR+/HER2– MBC.

RESULTS

Baseline clinical and genomic features and their prognostic associations

To develop a risk stratification framework, we leveraged the MSK Breast Translational Dataset, a prospectively collected and comprehensively annotated cohort of patients with breast cancer who received routine care at MSK. The final MSK-cohort comprised 1,264 patients with HR+/HER2– MBC who received first-line CDK4/6i+ET, with available baseline tumor-normal sequencing (on samples collected before and up to 28 days from treatment start) using an FDA-authorized clinical sequencing assay interrogating up to 505 cancer-associated genes[36] (**Figure 1A**, **Figure S1**).

The cohort was broadly representative of a real-world HR+/HER2– MBC population seen in a tertiary cancer center[37], with a median age at metastatic diagnosis of 57 years (interquartile range [IQR]: 49–65), 29.1% of patients (n=368) diagnosed with de novo MBC (**Figure 1B**), and visceral disease at presentation in 44.5% of cases (n=563). At a median follow-up of 50.4 months (IQR:

25.8–79.2), median progression-free survival (mPFS) was 18.6 months (95% confidence interval [CI] 17.5–20.5), with a 12-month PFS rate of 63.3% (95%CI: 60.6–66.1; **Figure 1C**), lower than that reported in pivotal first-line trials and consistent with the higher-risk profile of patients treated at tertiary cancer centers.

The vast majority of sequenced samples were obtained from the first metastatic biopsy (70.5%, n=891), and the genomic landscape aligned with that previously described for HR+/HER2– MBC candidates for front-line therapy[22], with *PIK3CA*, *CCND1*, and *TP53* representing the most common alterations (**Figure 1D**). We confirmed expected co-segregations between genomic and clinical features in our large cohort, including enrichment of *TP53* loss-of-function (LoF) among high-grade tumors (odds ratio [OR]=2.85, q=1.8x10^{-13}) and *ESR1* mutations in metastatic tumors with a TFI ≤12 months (OR=3.64, q=3.7x10^{-06}; **Figure 1E, Table S1**).

Our cohort further recapitulated established prognostic biomarkers reported in prospective randomized clinical trials[11,38,39], including shorter PFS in patients with liver metastasis (univariable hazard ratio [HR]=1.95, q=1.3x10^{-16}; **Figure 1F**), and in those with a TFI ≤12 months (univariable HR=1.69, q=5.2x10^{-14}; **Figure 1G**). Moreover, previously reported genomic alterations promoting CDK4/6i+ET resistance were significantly associated with poor outcomes (**Figure 1H**), including *TP53*[18] (HR=1.94, q=6.5x10^{-17}) and *RB1* LoF alterations[13,15,31,34] (HR=2.46, q=7.7x10^{-05}; **Figure 1I, Table S2**).

Development of a clinicogenomic model for risk stratification

We next sought to develop a risk stratification model leveraging baseline clinical and genomic data for individuals receiving first-line CDK4/6i+ET. Accordingly, we split the MSK-cohort into a training set (n=884, 70%) and a hold-out validation set (n=380, 30%) using a stratified sampling approach designed to balance the rate of PFS events, presence of liver metastasis, and TFI ≤12 months, resulting in a comparable representation of clinical and genomic features between the two sets (**Figure S2, Table S3**).

We first evaluated several interpretable machine learning architectures spanning linear models and ensemble trees[40,41] (**Figure 2A**). A critical consideration for CDK4/6i+ET prognostication was the risk of overlooking clinically and biologically relevant but low-prevalence alterations promoting CDK4/6i+ET resistance (e.g., *RB1*, *BRCA1/2,* and *PALB2*)[13,15,16,18,20,22,25,26], as standard penalized models compress their parameter contribution, potentially obscuring clinically meaningful

signals[42]. To address this, we developed ReCAST, a framework encompassing n=100 L1-penalized Cox models (base learners), each trained on a bootstrapped subset of the training cohort to promote feature diversity, reduce model variance and prevent overfitting[43] (**Methods**). Each base learner was designed to apply an adaptive penalty inversely proportional to feature importance as derived from an a priori L2-penalized Cox model fitted on the base-learner specific cohort, aimed to preserve the contribution of rare yet informative predictors. The regularization terms were further optimized using a 3-fold cross-validation strategy to maximize the discriminative accuracy of the model for the AUC[t] over a 3–36-month timeframe, reflecting a clinically meaningful and actionable timeframe for first-line therapeutic stratification in HR+/HER2– MBC. Ultimately, ReCAST results in an aggregated risk-score ranging from 0 to 100 derived from the mean risk estimation across base learners.

For benchmarking, we employed a 5-fold cross validation (CV) limited to the training set to avoid data leakage while estimating out-of-fold risk-predictions to mitigate biased results. Moreover, standard linear models and ensemble trees were optimized using a 3-fold CV using the 4-fold training partitions for hyperparameter tuning (nested CV). While all models performed comparably, ReCAST framework and random forest yielded the best absolute numerical and relative performances (**Table S4**). Next, we measured the sensitivity of predictors to rare but clinically validated genomic drivers of CDK4/6i+ET resistance using a cross-fitted feature-deletion analysis (**Methods**). ReCAST retained sensitivity to rare yet key genomic drivers of resistance associated with reduced PFS on CDK4/6i+ET (**Figure S3A**), while conventional penalized Cox models, and to a lesser extent ensemble trees, frequently attenuated or eliminated their contribution (**Figure 2B**).

We then expanded our benchmarking to include state-of-the-art transformers specifically designed for tabular data (**Methods**). These comprised transformers featuring a survival head for risk estimation fine-tuned directly in the MSK cohort (direct training) or involving a first step of self-supervised training followed by reinitialized training (gradual learning)[44], and foundation models including millions of parameters pretrained on more than 100 million synthetic datasets (TabPFN[45]) or hundreds of millions (TabFM[46]), performing an in-context learning[47] in the MSK cohort followed by a zero-shot prediction[46] (**Figure 2C**). Remarkably, although these architectures are recognized for their ability to capture non-linear effects and rare informative predictors through attention, ReCAST matched their sensitivity for rare yet mechanistically validated drivers of CDK4/6i+ET resistance (**Figure S3B**), with comparable weights supporting the plausibility of the

risk-score assignment by ReCAST; moreover, ReCAST achieved comparable overall discrimination (**Figure 2D**, **Figure S3C**) while retaining clinical and biological interpretability. This supports the adoption of ReCAST where interpretability and ease of deployment are prerequisites for wide clinical use, allowing implementation across resource-variable settings and locked, auditable use in prospective trials.

Building on these results, we employed ReCAST to design a clinical-only model incorporating baseline clinicopathological variables established to be associated with outcomes on front-line CDK4/6i+ET (**Figure 2E, Table S5**). In parallel, we developed a genomic-only model including oncogenic and likely oncogenic somatic and germline alterations according to OncoKB, an FDA-recognized precision oncology knowledge base[48,49], filtered for alterations present in at least 1.5% of the overall MSK-cohort (**Table S5**). This parallel development strategy was deliberately designed to determine whether clinical and genomic features capture distinct, non-redundant prognostic signals. To guard against the risk of overconfident estimates, risk-scores for the training set were derived exclusively from the out-of-bag risk estimation by base learners as inherently implemented in ReCAST (**Methods**).

In the training set, both the clinical-only (AUC[t]=0.69) and the genomic-only model (AUC[t]=0.70) demonstrated superior prognostication compared with standalone biomarkers (most discriminative clinical and genomic biomarkers: $AUC[t]_{TFI \leq 12\ months}$=0.62 and $AUC[t]_{TP53}$=0.61; **Figures 2F-G, Figure S4A-B**). To define clinically meaningful risk groups, we next fitted a two-component Gaussian-mixture model (GMM) to risk-scores (**Figure S4C, Methods**), identifying patients at high-risk of early progression who demonstrated poorer outcomes than those predicted to be low-risk cases in both frameworks ($HR_{Clinical}$=2.11, 95%CI: 1.79–2.47, P=1.4x10^{-19}; $HR_{Genomic}$=2.20, 95%CI: 1.87–2.59, P=9.8x10^{-22}; **Figure 2H**).

Evaluating whether the two models provided distinct prognostic information, we observed only fair agreement between the clinical and genomic risk classifications (Cohen's kappa=0.21; **Figure 2I, Figure S4D)**. Consistently, in a multivariable Cox regression including clinical and genomic predictions, each model retained independent prognostication (**Figure 2J**), emphasizing the non-redundant contribution of clinical and genomic features to predict CDK4/6i+ET outcomes. We next stratified the analyses combining clinical and genomic risks. Patients classified as clinical-low/genomic-low (i.e., predicted to be low-risk by both models) demonstrated prolonged disease control (n=356, mPFS=34.1 months, 95%CI: 29.0–44.6) compared with those classified as

clinical-high/genomic-high (n=192, mPFS=7.9 months, 95%CI: 6.4–9.5). By comparison, patients with discordant risk classification demonstrated intermediate and comparable outcomes (clinical-high/genomic-low: n=150, mPFS=16.7 months, 95%CI: 13.7–21.0; clinical-low/genomic-high: n=186, mPFS=15.5 months, 95%CI: 12.6–19.4; **Figure 2K, Figure S4E**). This emphasizes the importance of integrating genomic information along with clinical features, resulting in substantial reclassification that bears directly on treatment decisions, a property that global ranking metrics are not designed to capture and may therefore understate.

Based on these results, we next embarked on the development of CDKPredict, a unified prediction model building on the ReCAST framework integrating clinical and genomic variables (**Figure 2L**). As anticipated, the unified model demonstrated superior discrimination compared with each standalone model (AUC[t]=0.74 vs. 0.69 vs. 0.70; and C-index=0.68 vs. 0.64 vs. 0.64 for the clinicogenomic, clinical, and genomic model, respectively; **Figures 2M and 2N**), with cumulative risk-scores showing a strong association with shorter time to progression ($HR_{per\ 10\text{-}unit\ increase}$=1.31, 95%CI: 1.26–1.36, P=6.9x10^{-48}; **Figures S4F-G**). Finally, leveraging our observation that the clinical and genomic models frequently captured discordant risk profiles with similar PFS, we designed the clinicogenomic model to classify patients into low-, intermediate-, and high-risk groups using a three-component GMM (**Figure 2O, Methods**). Across strata, high-risk patients had significantly shorter PFS on first-line CDK4/6i+ET than those predicted at intermediate- (6.2 vs. 15.9 months, HR=2.32, 95%CI: 1.88–2.86, P=6.7x10^{-15}) and low-risk (33.6 months, HR=4.58, 95%CI: 3.66–5.74, P=5.1x10^{-40}; **Figure 2P**).

CDKPredict performance was robustly preserved in the holdout MSK-test set (n=380; **Figure 2Q, Figure S4H**), with high-risk patients demonstrating significantly shorter PFS on CDK4/6i+ET (n=67, mPFS=8.3 months) compared to intermediate- (n=154, mPFS=16.9 months, HR=1.68, 95%CI: 1.27–2.22, P=0.00026) and low-risk patients (n=159, mPFS=30.1 months, HR=3.91, 95%CI: 2.79–5.48, P=2.0x10^{-15}, **Figure 2R**), supporting the generalizability of the model for independent risk-prediction.

CDKPredict demonstrates preliminary specificity for endocrine-based regimens (ET +/- CDK4/6i)

A prognostic classifier that merely captured generic poor prognosis would be expected to perform similarly irrespective of the treatment received. We therefore evaluated whether CDKPredict discrimination differed according to the type of regimen, hypothesizing it would better forecast

outcomes in patients receiving endocrine-based regimens given the several model features linked to CDK4/6i and/or ET resistance. To this end, we leveraged two independent cohorts of patients with HR+/HER2– MBC who received routine care at MSK and were treated with first-line ET alone (n=402) or first-line chemotherapy (n=265), and compared CDKPredict performance with that in patients receiving CDK4/6i+ET in the internal MSK validation cohort (**Figure 3A, Figure S5**). As expected, patients treated with first-line chemotherapy carried poor prognostic clinicopathological variables, including a higher rate of liver metastasis (OR=3.15, q=5.7x10^{-10}), higher metastatic burden (≥4 metastatic organs, OR=2.8, q=7.6x10^{-05}), and ER≤10% (OR=4.82, q=4.0x10^{-06}) among others (**Figure 3B, Table S6**), consistent with the endorsed use of chemotherapy in patients with visceral crisis or carrying clinicopathological surrogates of endocrine resistance.

Across cohorts, CDKPredict showed superior discrimination in patients treated with CDK4/6i+ET (AUC[t]=0.74, 95%CI: 0.68–0.80) compared to chemotherapy (AUC[t]=0.65, 95%CI: 0.57–0.71, AUC[t]$_{1\text{-}12month}$ difference P=0.033, **Methods,** $P_{interaction}$=0.028), while showing intermediate performance in those receiving ET alone (AUC[t]=0.69, 95%CI: 0.64–0.75, AUC[t]$_{1\text{-}12month}$ difference P=0.25, $P_{interaction}$=0.069, **Figures 3C-D**). Cognizant of the adverse clinicopathological features of the chemotherapy cohort, potentially resulting in reduced case-mix heterogeneity that could depress model discrimination irrespective of treatment specificity[50], we repeated the analysis after propensity score matching. The prognostic effect for CDK4/6i+ET remained stronger compared to chemotherapy after balancing the representation of baseline clinical variables (n=195 pairs, AUC[t]$_{CDK4/6i+ET}$=0.74, 95%CI: 0.66–0.81 vs. AUC[t]$_{chemotherapy}$=0.62, 95%CI: 0.53–0.71, $P_{interaction}$=0.048, **Figure S6A, Table S6**).

To further ascertain the specificity of CDKPredict towards HR+/HER2– MBC and endocrine-based regimens, we analyzed two independent MSK cohorts of 1) patients with triple-negative MBC (TNBC) treated with first-line chemotherapy (n=547), and 2) patients with HER2+ MBC who received first-line THP (i.e., taxane-based regimen, trastuzumab and pertuzumab, n=315, **Figure S6B, Table S6**). CDKPredict demonstrated poor performance across these two cohorts (AUC[t]$_{TNBC}$=0.60, 95%CI: 0.55–0.65; P-interaction=0.01; and AUC[t]$_{HER2+}$=0.52, 95%CI: 0.44–0.59; P-interaction=4.1x10^{-05}, **Figures S6C-S6D**). Together, these results argue against CDKPredict as a generic marker of poor clinical outcomes and point instead in favor of preliminary specificity for endocrine-based therapy.

Independent external clinical validation of CDKPredict

To obtain the independent clinical validation of CDKPredict, we next leveraged a nationwide, external, multi-institutional cohort totaling 564 patients with HR+/HER2– MBC, who received first-line CDK4/6i+ET in community or academic settings and underwent baseline tumor genomic sequencing on samples collected before and up to 28 days from first-line CDK4/6i+ET initiation (**Figure 4A, Figure S7**). Genomic profiling was conducted using clinically validated platforms, including a tissue-based assay interrogating up to 648 cancer-related genes[51,52] (n=417), and a 105-gene circulating tumor DNA (ctDNA) panel[53] (n=147). Compared with the MSK-cohort, patients in the external cohort exhibited shorter PFS on first-line CDK4/6i+ET (HR 1.32 (95%CI 1.15–1.51), P=4.7x10^{-05}; **Figure S8A**), consistent with a higher representation of poor prognostic features, including a higher prevalence of liver (OR=2.06, q=4.7x10^{-10}) and central nervous system (CNS) metastasis (OR=4.29, q=7.5x10^{-11}), and a lower frequency of bone-only disease (OR=0.48, q=1.7x10^{-9}; **Figure S8B, Table S7**). Nonetheless, the genomic landscape of the external cohort was broadly comparable to that of the MSK-cohort (**Table S7**).

Before deployment, the clinicogenomic model was retrained on the overall MSK-cohort, where model parameters and risk-group cutoffs were fine-tuned and locked for all subsequent evaluations (**Figure 4A**). In the external cohort, CDKPredict risk-scores retained robust single-patient discrimination (AUC[t]=0.74, 95%CI: 0.68–0.77), correlating with shorter PFS on CDK4/6i+ET in a stepwise manner ($HR_{per\ 10\text{-}unit\ increase}$=1.28, 95%CI: 1.22–1.35, P=2.3x10^{-22}; **Figure 4B**), and demonstrated strong agreement between predicted probabilities and observed progression at 12 and 24 months, with no evidence of overfitting and risk underestimation in low-risk patients (calibration curve above the diagonal line) or risk overestimation in high-risk patients (calibration curve below the diagonal, **Figure S8C**).

High-risk cases demonstrated poorer outcomes on CDK4/6i+ET (mPFS=8.0 months, 95%CI: 6.4–10.0) compared with cases estimated at intermediate (mPFS=14.7 months, 95%CI: 12.3–19.5; HR=1.93, 95%CI: 1.48–2.52, P=1.2x10^{-06}) and low risk (mPFS=29.8 months, 95%CI: 21.8–42.1; HR=3.85, 95%CI: 2.83–5.26, P=1.5x10^{-17}; **Figure 4C**). The risk stratification further resulted in pronounced differences in the rate of durable disease control at 6 months (high-risk=61.0%, 95%CI: 52.9%–70.1%; intermediate-risk=74.4%, 95%CI: 68.6%–80.8%; low-risk=84.8%, 95%CI: 80.0%–90.2%), as well as likelihood of achieving a radiological tumor response across risk strata (P=0.0011 by Pearson's chi-square test; **Figure 4D**).

To ascertain the generalizability of CDKPredict to the general population of patients considered for first-line CDK4/6i+ET, we next evaluated its performance across clinically relevant subgroups. Robust model predictions were consistent across CDK4/6i compounds ($P_{Interaction}$=0.93) and ET backbone (i.e., fulvestrant vs. aromatase inhibitors, $P_{Interaction}$=0.34; **Figure 4E, Table S8**). Moreover, while the model was not trained on patients with baseline ctDNA profiling, it retained strong risk stratification on the subset of test cohort with baseline ctDNA sequencing ($HR_{High\ vs.\ Low}$=4.96, 95%CI: 2.59–9.51, P=1.4x10^{-06}; $HR_{Intermediate\ vs.\ Low}$=2.05, 95%CI: 1.11–3.81, P=0.023; **Figure 4F**).

We next focused on patients with tumors harboring *PIK3CA* mutations (n=226), considering the recent approval of inavolisib in combination with CDK4/6i+ET in the first-line setting[35], with additional ongoing trials investigating PI3K/AKT inhibitors in this therapeutic space (NCT06790693, NCT04862663, NCT07174336). Consistent with prior reports[16,33], *PIK3CA* mutation status was not associated with PFS (HR=1.02, 95%CI: 0.81–1.28, P=0.88). Conversely, CDKPredict yielded robust risk stratification across *PIK3CA* wild-type and mutant tumors (**Figure 4G**), identifying a subset of *PIK3CA*-mutant cases achieving durable disease control on CDK4/6i+ET alone (n=85, 37.6%, mPFS=32.9 months, 95%CI: 22.1–not reached), emphasizing the rationale for pursuing risk-adapted therapeutic strategies in this biomarker-selected population. Finally, CDKPredict retained robust risk stratification across clinically relevant subgroups (TFI ≤12 and >12 months, recurrent vs. *de novo* metastatic disease, bone-only disease, presence of liver metastasis, and age) both in the external and MSK-cohort (**Figure S8D**).

Lastly, we benchmarked CDKPredict against historic risk criteria currently employed for first-line therapeutic escalation in HR+/HER2– MBC, namely TFI ≤12 months. While TFI ≤12 months was associated with inferior outcomes in the external cohort (HR=1.81, 95%CI: 1.41–2.32, P=2.8x10^{-06}), and TFI high-risk patients (n=149) were largely predicted at high- or intermediate-risk by CDKPredict (**Figure 4H**), a critical limitation emerged among patients classified as low-risk (TFI >12 months, n=415). In this population, CDKPredict reclassified a substantial proportion of TFI low-risk patients as high- (n=73, 17.6%) or intermediate-risk (n=151, 36.4%) of early progression, who ultimately achieved short- to mid-term disease control (CDKPredict-high: mPFS=10.1 months, 95%CI: 8.5–13.9; CDKPredict-intermediate: mPFS=14.7 months, 95%CI: 12.4–19.5; **Figure 4H**). In an exploratory analysis, we assessed this question in the larger MSK-cohort using out-of-bag risk estimates from the locked ensemble model (**Methods**). In this analysis, a

substantial proportion of TFI low-risk patients were classified as intermediate or high risk by CDKPredict (n=330, 43.9%). Conversely, 16.6% (n=85) of TFI high-risk patients were reclassified as CDKPredict low risk and achieved durable disease control on CDK4/6i+ET (mPFS=31.0 months, 95%CI: 19.3–73.7; **Figure S8E**). Finally, CDKPredict demonstrated more accurate discrimination at both the 12- and 24-month landmark PFS analyses in the external validation cohort (**Figure 4I**).

Beyond discrimination, we evaluated the utility of CDKPredict using decision-curve analysis, where we estimated the net benefit by weighing patients who were hypothetically escalated correctly (progression within 12 months) against those escalated unnecessarily across threshold probabilities, i.e., the risk of progression above which a clinician would escalate. In the external cohort, selecting patients by CDKPredict high-risk status yielded a higher net benefit than the TFI criterion across all thresholds, with the highest net benefit across all strategies between 35% and 72%, and a similar advantage was observed in the MSK cohort at thresholds above 30% (**Figure S8F, Table S9**), reinforcing the clinical advantage of CDKPredict over currently adopted criteria for front-line risk stratification.

CDKPredict risk classifications reflect clinical and biological plausibility

To assess the clinical and biological plausibility of CDKPredict, we took two complementary approaches. First, we leveraged the natural interpretability of its linear ensemble framework to examine features driving risk predictions. Second, we tested whether risk classifications recapitulated established genomic and transcriptomic programs of CDK4/6i+ET resistance not prespecified a priori to the model.

For each variable, we quantified the selection frequency across the 100 base learners together with the mean contribution to patient risk-scores (**Figures 5A-B, Figure S9A**). Among clinical features, liver metastasis, low estrogen receptor expression (ER≤10%), lack of progesterone receptor expression (PR-negativity), higher number of metastatic organs (≥4), and TFI ≤12 months demonstrated the highest linear increment in CDKPredict risk-scores (+21.1, +18.2, +15.6, +12.3, and +11.8, respectively). Among genomic features, *TP53* LoF (+19.6) and *MDM2* amplification (+15.7) were among the highest-weighted contributors to risk predictions, consistent with prior work establishing their role as major drivers of CDK4/6i+ET resistance through abrogation of DREAM-mediated cell-cycle control[18]. Beyond these, several genomic features prioritized by CDKPredict recapitulated established mechanisms of resistance to CDK4/6i, ET, or

both, including *ESR1* mutations[23,54] (risk-score: +9.9) and *MYC* amplification[22,55] (+4.4). High TMB (≥10 mutations per megabase) also strongly correlated with risk (risk-score: +10.6; **Figure 5C**), consistent with its previously reported association with reduced benefit from CDK4/6i+ET in phase III randomized trials[33]. Conversely, predictions of durable disease control were supported by distinct genomic mechanisms, such as *GATA3* (risk-score: -9.6), which is known to promote luminal differentiation and endocrine dependency in breast cancer[56,57].

In addition, we quantified feature importance using Shapley additive explanations (SHAP), which decompose each individual prediction into additive feature-level contributions, enabling assessment of population-level importance from the mean SHAP contribution across patients (**Figure 5B**). A set of highly prevalent clinical and genomic features consistently carried the highest population-level impact (top four by SHAP: *TP53* LoF: 0.23, PR-negativity: 0.22, liver metastasis: 0.21, and TFI ≤12 months: 0.17). In contrast, low-prevalence alterations with well-established mechanistic roles in CDK4/6i+ET resistance such as *RB1*[13,15,16,31,34], *BRCA2*[13], and *BRCA1*[13,58,59] demonstrated a modest population-level importance (SHAP: 0.02, 0.04, 0.01, respectively), yet conferred large per-patient risk-scores (*RB1*: +15.7, *BRCA2*: +14.6, *BRCA1*: +8.3), highlighting CDKPredict's ability to preserve the influence of rare but mechanistically validated drivers of resistance.

Beyond features directly implemented in CDKPredict, risk-scores demonstrated a significant correlation with broader markers of genomic instability among tumors carrying a sufficient number of copy-number alterations (n=1,085, 85.8%; **Methods, Figure 5D**), including fraction of genome altered (FGA; Spearman's rho=0.27, P=6.3x10^{-20}; **Figure 5E, Figure S9B**) and whole genome doubling (WGD) events (P=7.7x10^{-5} by Pearson's chi-square test; $OR_{High\ vs.\ Low}$=2.28, 95%CI: 1.53–3.40, P=2.7x10^{-05} by Fisher's exact test; **Figure 5F**). Evaluating tumor mutational signatures (**Figure 5G**), high-risk tumors featured a higher frequency of HRD ($OR_{High\ vs.\ Low}$=2.24, 95%CI: 1.43–3.53, P=0.0002; **Figure 5H**), consistent with its role in promoting CDK4/6i resistance through acquisition of *RB1* loss and other established resistance mechanisms[13]. High-risk tumors further presented an enrichment of dominant APOBEC mutagenesis ($OR_{High\ vs.\ Low}$=1.59, 95%CI: 1.10–2.30, P=0.0115), emphasizing its key role in CDK4/6i+ET resistance[32]. The distinct representation of genomic signatures across risk-strata reflected their co-segregation with CDKPredict biomarkers (**Figures S9C–S9D**), suggesting that the model inherently captures higher-order genomic alterations through its constituent features. Indeed, augmenting the model with dominant tumor mutational signatures (HRD and APOBEC3), genome-wide copy-number

features (WGD and FGA), and hemizygosity in tumor suppressor genes associated with CDK4/6i resistance (i.e., *RB1*, *BRCA1*, *BRCA2*, *TP53*, *LATS2*, *NF1*, *PTEN*)[13] yielded no discriminative improvements (augmented model AUC[t]=0.74, 95%CI: 0.699–0.811, vs. CDKPredict AUC[t]=0.74, 95%CI: 0.699–0.808; **Figure S10, Methods**).

Finally, we extended this validation to baseline transcriptome, an orthogonal data modality, leveraging the subset of patients from the external validation cohort with available baseline RNA sequencing[60] (n=450, 79.8%; **Figure 5I**). CDKPredict low-risk tumors were significantly enriched for PAM50[3] luminal subtypes ($OR_{Low\ vs.\ High}$=3.58, 95%CI: 2.01–6.50, P=3.6x10^{-06}; **Figure 5J**), whereas high-risk cases were enriched for basal-like and HER2-enriched tumors ($OR_{High\ vs.\ Low}$=5.89, 95%CI: 3.08–11.72, P=2.6x10^{-09}), consistent with established associations between genomic features included in CDKPredict and PAM50 transcriptomic profiles[56] (**Figure S9E, Table S10**). At the gene level, CDKPredict risk-scores showed the strongest correlation with progesterone receptor RNA expression (Spearman's rho=-0.44, q=9.0x10^{-19}; **Figure 5K, Table S11**), recapitulating its established role as a marker of endocrine sensitivity[61], and secondarily with *AURKA* among variables included in CDKPredict (Spearman's rho=0.31, q=1.1x10^{-08}; **Figure S9F, Table S11**), corroborating the role of AURKA overexpression in promoting CDK4/6i+ET resistance, as previously shown[19].

At the pathway level, gene set variation analysis (GSVA) using the Hallmark collection[62] revealed an overactivation of estrogen signaling pathways among low-risk tumors (Estrogen response early, ER_{early} q=9.5x10^{-07}; ER_{late} q=0.0137 by Mann-Whitney U) and a downregulation of *KRAS* signaling (q=0.0078), among others (**Figure S9G**). Remarkably, ER_{early} signaling showed the most significant correlation with CDKPredict risk-scores (Spearman's rho=-0.30, q=2.7x10^{-09}**; Figure 5L, Table S12**), resulting in a negative gene-set enrichment among high-risk patients (q=0.0010; **Figure 5M**). Notably, a small subset of high-risk tumors (12.5%, n=15) exhibited overactive estrogen signaling by GMM classification (**Figure 5N, Methods**), with a prominent overrepresentation of luminal PAM50 subtypes among these cases (13 of 15; OR=5.41, 95%CI: 1.14–51.73, P=0.023). Strikingly, these high-risk/estrogen response-high tumors featured a marked enrichment of *ESR1* mutations compared to high-risk/non-estrogen-high cases (53.3% vs. 7.0%, q=0.0013 by Fisher's exact test; **Figure 5O**), an association not observed among CDKPredict low- and intermediate-risk tumors (**Figure S9H**). Although based on a small subset and therefore hypothesis-generating, these findings define a specific subset of high-risk cases in which aberrant estrogen signaling, driven by baseline *ESR1* mutations, might represent the

primary mechanism of early progression, providing a biological rationale for prioritizing oral selective estrogen receptor degraders (SERDs) as an ET backbone in this population, an approach currently investigated in *ESR1*-selected and unselected first-line clinical trials (NCT04964934, NCT04711252, NCT04546009).

Collectively, these findings demonstrate that CDKPredict converges on the established biology of endocrine and CDK4/6i resistance. None of the weights were prespecified in advance, yet the model independently prioritized mechanistically validated drivers of resistance. In the external validation cohort, risk strata defined exclusively from clinical and DNA-level features also mapped onto transcriptional programs never provided to the model. This correspondence supports the biological plausibility of the model and is an assessment that our interpretable architecture readily permits.

Recurrent patterns of actionable alterations to tackle resistance

Building on the genomic variables for risk predictions, we next evaluated the representation of actionable alterations across risk strata (**Figure 6A**). Leveraging OncoKB[49] and the MSK-cohort, the frequency of established actionable alterations (OncoKB levels 1–2) increased gradually across risk classes (low-risk: 55.1%; intermediate-risk: 60.2%; high-risk: 67.0%; **Figure 6B**). This occurred despite a progressive decline in the frequency of *PIK3CA* mutations among higher risk-strata, reflecting instead an increase in non-*PIK3CA* level 1–2 alterations (**Figure 6C**), with 48.0% of high-risk tumors harboring level 1–2 alterations beyond *PIK3CA* compared with 17.0% of low-risk tumors (**Figure 6D**). Among these, besides *PTEN*, with proven clinical utility of AKT inhibitors in endocrine-resistant tumors[63], *ESR1* mutations represent a predictive biomarker for oral SERDs[64,65], and *ERBB2*-activating mutations have been shown to promote endocrine resistance[28], potentially reversible by HER2 tyrosine kinase inhibitors[66,67]. Moreover, higher-risk strata featured a higher frequency of high TMB (16.1% in high-risk vs. 1.8% in low-risk; $OR_{High\ vs.\ Low}$=10.58, 95%CI: 4.99–25.09, P=1.3x10^{-13}), a tumor-agnostic biomarker of immunotherapy efficacy[68-70].

Beyond level 1–2 alterations, high-risk tumors showed a higher representation of alterations with supportive clinical or preclinical evidence of actionability (OncoKB level 3–4, **Figure 6B**), with only 15.8% of high-risk tumors lacking any actionable alterations (vs. 35.4% in low-risk; **Table S13**). While the frequency of germline or somatic *BRCA2* alterations was significantly higher in intermediate- and high-risk tumors compared to low-risk cases (**Table S14**), *BRCA1* ($OR_{High\ vs.}$

$_{Low}$=11.59, *q*=0.00052) and *PALB2* ($OR_{High\ vs.\ Low}$=7.61, *q*=0.0125) were found almost exclusively among high-risk tumors. Given the established role of PARP inhibitors for somatic or germline *BRCA1/2* and germline *PALB2* alterations[71-73], and prior work supporting the clinical utility of prioritizing PARP inhibition in this context[13], these data further support the design of EvoPARBreast-01, a phase III randomized trial comparing front-line PARP inhibition plus ET with standard CDK4/6i+ET in tumors harboring germline or somatic *BRCA1/2* or *PALB2* alterations (NCT06380751).

Finally, *TP53* LoF demonstrated the starkest increase across risk-strata (61.1% in high-risk vs. 4.3% in low-risk; $OR_{High\ vs.\ Low}$=34.66, 95%CI: 20.82–59.29, *q*=6.8x10^{-69}; **Figure 6A**). *MDM2* amplification followed a similar pattern (8.4% in high-risk vs. 1.8% in low-risk; $OR_{High\ vs.\ Low}$=5.08, 95%CI: 2.22–12.64, *q*=0.0001). Notably, the addition of CDK2 inhibitors to CDK4/6i+ET has been previously shown to overcome CDK4/6i resistance in *TP53*/*MDM2*-altered breast cancer[18], providing a direct mechanistic rationale for evaluating the use of CDK2-targeted strategies in this high-risk subgroup.

Collectively, these findings highlight the potential value of CDKPredict as a backbone for risk-adapted, biomarker-directed therapeutic strategies, linking each risk stratum to a distinct landscape of actionable alterations with supporting clinical utility to tackle endocrine resistance.

Development of a scalable model for broad clinical deployment

We finally sought to develop a scalable and pragmatic model, termed CDKPredict-S, that could be utilized globally, both in routine clinical practice and as an enrichment tool for clinical trials (**Figure 7A**). Two key principles guided this effort: first, limiting the required input to clinical variables routinely and uniformly collected in clinical practice worldwide, and genomic variables to a minimal, essential gene panel accessible across diverse sequencing platforms; second, ensuring robustness against missing data to maintain model calibration at both the clinical and genomic levels. We therefore developed two versions of CDKPredict-S: a clinical-only model relying exclusively on limited clinicopathologic variables, and a clinicogenomic model that also included a restricted set of genomic biomarkers with established clinical relevance in HR+/HER2– MBC and broad availability across different sequencing panels (*PIK3CA*, *PTEN*, *AKT1*, *BRCA1*, *BRCA2*, *ESR1*, *TP53*, *RB1*). To further maximize scalability, both versions were designed to tolerate missing variables by using the average risk contribution as measured in the development set to mitigate systematic distortions of risk-scores (**Methods**).

Following evaluation in the internal hold-out set (**Figure 7B, Table S15**) and fine-tuning using the full MSK-cohort (**Figures S11A-S11B**), CDKPredict-S was evaluated across two independent test cohorts (**Figure 7C**). Besides the external nationwide multi-institutional cohort utilized for the clinical validation of CDKPredict, we leveraged 455 patients with HR+/HER2– MBC who received first-line CDK4/6i+ET at a single European institution (IEO-cohort), for which baseline genomic sequencing was available in 21.1% of patients (96 of 455) using a tissue-based panel interrogating 523 cancer-related genes[74] (12.7%, n=58) or a cell-free DNA (cfDNA)-based 50-gene assay[75] (8.4%, n=38; **Table S16**).

In the pooled external cohort, clinical (AUC[t]=0.68) and clinicogenomic (AUC[t]=0.71) risk-scores of CDKPredict-S demonstrated robust risk-discrimination (**Figures 7D-E, Figures S11C-S11D, Table S15)**. Across risk-strata, both the clinical model evaluated in 1,019 patients ($HR_{High\ vs.\ Intermediate}$=1.99, 95%CI: 1.65–2.41, P=1.0x10^{-12}; $HR_{High\ vs.\ Low}$=3.09, 95%CI: 2.50–3.82, P=2.4x10^{-25}) and the clinicogenomic model evaluated in 660 patients ($HR_{High\ vs.\ Intermediate}$=2.12, 95%CI: 1.68–2.68, P=3.0x10^{-10}; $HR_{High\ vs.\ Low}$=3.60, 95%CI: 2.75–4.72, P=1.1x10^{-20}) demonstrated robust baseline risk stratification (**Figures 7F-G**), confirming the clinical validation of the scalable model. Although its discrimination was modestly lower than that of the full CDKPredict model (AUC[t]=0.74), this reflected a deliberate trade-off, restricting input to a minimal set of widely available features in exchange for broad clinical implementation. Following validation, CDKPredict-S was deployed for academic and non-commercial use at https://cdkpredict.com/.

DISCUSSION

In the present work, leveraging a large and densely annotated clinicogenomic cohort together with validation in independent multi-institutional datasets, we developed ReCAST, a machine-learning framework for outcome prediction designed for sparse clinicogenomic data, and used it to build and clinically validate CDKPredict, a model designed to forecast treatment outcomes on front-line CDK4/6i+ET to inform risk-adapted therapeutic interventions in clinical practice.

Over the past decade, several individual genomic biomarkers linked to CDK4/6i+ET outcomes have been discovered; however, a dedicated effort to combine them with clinical biomarkers for robust baseline prognostication had not been undertaken. We now demonstrate that integrating clinical and genomic variables within a single prediction framework outperforms standalone biomarkers. In addition, we measured the magnitude of benefit of adding genomic biomarkers for

prognostication, which beyond incremental gains in global discrimination, reclassified the risk of a substantial proportion of patients compared to clinical features alone. Importantly, because treatment escalation or de-escalation hinges on the category a patient is assigned to, reclassification translates directly into clinical decision-making, a property that ranking metrics (e.g., AUC[t]) are not designed to capture and may substantially understate. Moreover, we developed an interpretable model matching the performance of state-of-the-art machine-learning architectures, while preserving sensitivity to rare but mechanistically validated drivers of resistance (e.g., *RB1*, *BRCA1/2*, *PALB2*, *PTEN*) that conventional penalized models tend to discard. This represents a consequential property given that several of these drivers are directly targetable. This performance was preserved across cohorts profiled with diverse tissue- and plasma-based sequencing panels and spanning academic and community practice settings, supporting robustness to real-world platforms and practice heterogeneity where most prediction models tend to fail[76].

In oncology practice, a growing range of therapeutic options are becoming available to providers and patients, often with limited guidance on the best available therapy for a given case. Target-specific predictive biomarkers enable matching molecular features to mechanism of therapy, but they may i) fail to identify patients at greatest risk of poor outcomes on standard of care; ii) lack the clinical validity (i.e., a biomarker defining separated clinical outcomes) on which to randomize large populations of patients; and iii) lack specificity for a given therapy. Moreover, certain escalation strategies, such as induction with chemotherapies, do not lend themselves to genomic selection. Thus, prior to the deployment of such predictive biomarkers, a clear strategy for prognostication is necessary. Such was the case for the development of prognostic multi-gene assays to identify early-stage breast cancer at high risk of relapse[1], which preceded their implementation as integral biomarkers in prospective trials to establish their predictive power[4,5]. This motivated us to develop a prognostic tool to establish outcomes in HR+/HER2– MBC where disparate outcomes on endocrine therapy are routinely observed and a variety of mechanism-based treatments are available beyond standard of care. Therefore, we clinically validated a prognostication tool addressing the acute need for reliable prognostication in HR+/HER2– MBC to guide risk-adapted prospective trials. To our knowledge, CDKPredict represents the first clinically validated AI-based biomarker for CDK4/6i+ET prognostication (EBAI class C1)[77], and in a head-to-head comparison, it outperformed the clinical criteria currently adopted for risk stratification in pivotal phase III clinical trials (i.e., TFI≤12 months, as in INAVO120[35]). Notably, about half (50.7%) of the patients classified as high-risk by the clinicogenomic model would have

been considered low-risk by this criterion and would therefore not have been eligible for escalation strategies intended to improve their outcomes. The attenuated performance of CDKPredict in patients treated with chemotherapy and in other breast cancer subtypes further indicates that it captures endocrine-resistance biology rather than generic poor prognosis, and establishing its predictive validity (EBAI class C2) will require randomized trials in which treatment is allocated by risk group.

Moving forward, CDKPredict demonstrated consistent capability to detect tumors that were less likely to derive benefit from standard therapy, thereby identifying a population for which intensified front-line treatments could potentially prevent rapid disease progression (**Figure S12**). In parallel, higher-risk strata were enriched for clinically actionable alterations, coupling risk stratification to biomarker-matched escalation, such as PARP inhibition for *BRCA1/2* or *PALB2* alterations, PI3K, AKT, and mTOR inhibitors for PI3K/AKT pathway alterations, CDK2 inhibitors for *TP53* LoF or *MDM2* amplification, and oral SERDs for *ESR1*-driven estrogen signaling. Moreover, CDKPredict readily identified a subgroup of patients exhibiting long-term benefits from standard therapy alone, for whom treatment escalation is unlikely to provide additional benefit while adding financial and health-related toxicities. In the same context, CDKPredict could potentially address the clinical task of identifying patients with tumors exhibiting pronounced endocrine sensitivity, who could represent an appropriate risk-defined population for de-escalated treatment approaches. Indeed, although comparable benefits from early versus deferred use of CDK4/6i have been observed[78], the lack of reliable biomarkers represents a key limitation to the confident adoption of such de-escalation strategies in clinical practice.

The framework underlying CDKPredict is not specific to this clinical setting. Sparsity is a frequent feature of tabular clinicogenomic data; in many tumor types, rare alterations with a clear mechanistic link to treatment resistance are observed in only a small fraction of patients. Transformer architectures are increasingly regarded as state of the art for prediction from tabular data, yet ReCAST matched their sensitivity to such uncommon events while remaining interpretable. Indeed, the lack of intrinsic interpretability of such models represents a constraint on clinical implementation. ReCAST addresses this directly by assigning feature-specific adaptive penalties within a bootstrapped ensemble, retaining rare variables that carry large per-patient effects while preserving the interpretability of a linear Cox framework. Here, this property allowed the model to leverage key drivers of CDK4/6i+ET resistance, such as *RB1*, *ESR1*, *PTEN*, *BRCA1/2*, and *PALB2*, among others, at magnitudes commensurate with their biological

importance despite low population-level frequencies. We anticipate the same approach may be applicable wherever outcome prediction must integrate common clinical variables with rare, mechanistically consequential molecular events, and we have made the framework openly available to enable such applications.

A second barrier to deployment is data availability. Comprehensive genomic profiling remains inaccessible in many practice settings worldwide, and even routinely collected clinical variables are frequently incomplete or inconsistently recorded outside of academic centers. To address this, we additionally developed CDKPredict-S, a scalable version requiring routinely collected clinical variables and a minimal gene panel, which retained robust risk stratification across two large independent cohorts, positioning it as a practical enrichment tool for clinical trials and a decision-support aid in routine practice worldwide. Together, these models are well-suited for implementation in prospective studies investigating biologically informed, precision-based, and risk-adapted treatment interventions.

We acknowledge the limitations of our study. First, although we demonstrated an attenuated association of CDKPredict with outcomes in non-endocrine treatment arms, predictive validation warrants randomized prospective biomarker-stratified studies, as noted above. Second, while our clinicogenomic framework provides baseline risk stratification for CDK4/6i+ET, its value in informing alternative therapeutic strategies to improve clinical outcomes has not been addressed in the present study. Third, our metastatic cohorts did not include patients who received adjuvant CDK4/6i; although resistance biology is expected to be shared, performance after adjuvant CDK4/6i will require dedicated evaluation, which the framework supports through retraining. Finally, while ReCAST is designed as a general framework for outcome prediction from sparse clinicogenomic data, its performance has been established here in a single disease and treatment context; application to other tumor types and therapeutic settings will require dedicated training and independent validation.

In conclusion, we presented a validated translational framework for risk stratification from sparse clinicogenomic data and its application to first-line CDK4/6i+ET in HR+/HER2– MBC. In an era of increasingly complex and biologically driven therapeutic strategies, these findings provide a foundation for the prospective evaluation of risk-adapted, precision-based escalation and de-escalation therapeutic interventions for front-line therapy in MBC, and a generalizable approach for outcome prediction in oncology.

ACKNOWLEDGEMENTS

This work was supported by the Fondazione Gianni Bonadonna International Post-Doctoral Research Fellowship Program and ASCO Conquer Cancer Foundation Young Investigator Award (to L.B.B.); the American Association for Cancer Research and BCRF (AACR-BCRF, to F.S.V.); the Komen Career Catalyst Research Award (to P.R.); the MBCure Foundation (to P.R.); the Porat family (to P.R. and S.M.); the Burton family (to P.R.); the Breast Cancer Research Foundation (BCRF; to S.C.); the National Cancer Institute grants P50 CA247749 (to S.C. and P.R.) and R35 CA305347 (to S.C.); the Cancer Couch Fund (to S.C.); the Foss Family Foundation (to S.C. and P.R.); the Associazione Italiana Ricerca sul Cancro under Next Gen Clinician Scientist - ID. 31734 project (to A.M.); the NIH/NCI Cancer Center Support Grant P30 CQA008748 and the Breast Cancer Research Foundation (to M.E.R.). This work was also partially funded by AstraZeneca and SOPHiA GENETICS. The authors acknowledge Juan Arango Ossa for the critical input in the development of the cdkpredict.com website.

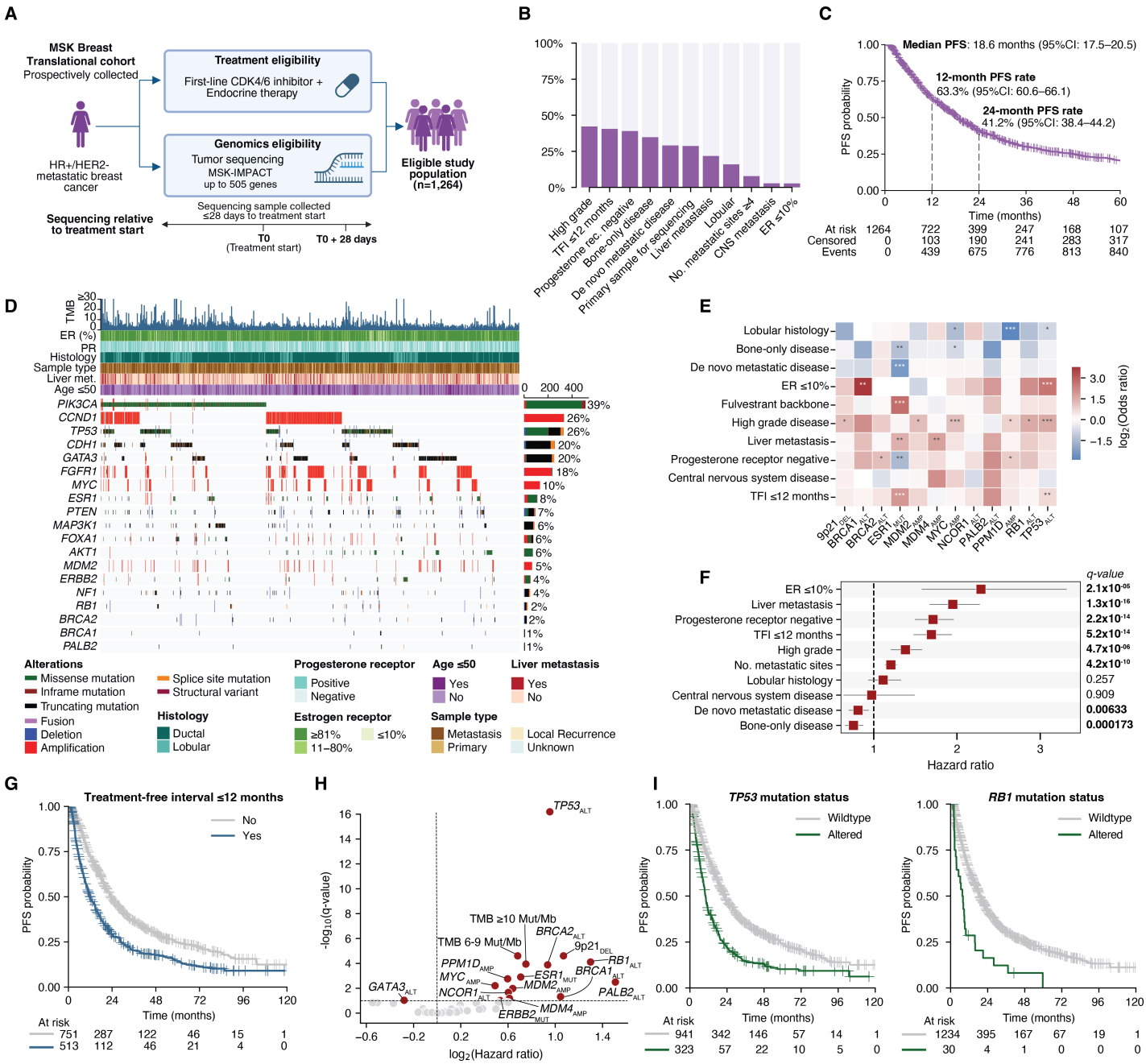
CONTRIBUTIONS

P.R., S.C. and L.B.B. conceived the study; L.B.B., P.R. and S.C. led the development of ReCAST and CDKPredict with secondary contributions from K.P.; L.B.B. and P.R. conceived the development and validation framework of CDKPredict; L.B.B., P.R. and K.P. designed and trained the benchmarking models; L.B.B. and P.R. led the biostatistical analysis and validation of CDKPredict; L.B.B., F.S.V., S.N., A.M., E.M., E.F., L.C., A.S., B.M. and M.C. contributed to data collection; S.N., M.D., and Y.G. performed quality control of the genomic sequencing data from the external multi-institutional cohort; M.Z. conducted quality control of the transcriptomic data; L.B.B., S.C., and P.R. led the clinicogenomic analysis with secondary contributions from S.N.; L.B.B. and P.R. conducted the transcriptomic analysis; M.F.B., D.B.S. and N.S. developed the MSK-IMPACT germline and somatic analytical platforms and the germline variant pathogenicity annotation pipeline; A.M. and N.F. contributed to generating genomic sequencing data and quality control from the IEO cohort; I.D.B. developed cdkpredict.com with secondary contributions from L.B.B. and P.R.; L.B.B., P.R. and S.C. wrote the original version of the manuscript; All authors provided critical revisions to the manuscript and accepted the final version; P.R. and S.C. supervised the overall study.

COMPETING INTERESTS

E.F. has received personal fee as speaker from AstraZeneca and Lilly. A.M. has received institutional grants or funding from Gilead Sciences; reports personal fee as speaker, consultant,

or advisor from Menarini/Stemline, Roche, Lilly, AstraZeneca, Daiichi-Sankyo, Pfizer, StratifAI, outside the submitted work; reports travel support from Menarini/Stemline, AstraZeneca, Lilly, Pfizer, Daiichi-Sankyo. S.S., G.L., S.K., I.M., and M.S. are currently employed by AstraZeneca. M.C. is currently employed by Tempus. N.F. reports consulting or advisory roles with Abbvie, Alira Health, AstraZeneca, Biocartis, Epredia, Gilead, Menarini Group, Merck, MSD, Novartis, Owkin, Pfizer, Roche, Sakura, Signatur Bio, Sysmex, and Veracyte; reports personal fees as a speaker from AstraZeneca, Daiichi Sankyo, Epredia, Exact Sciences, Gilead, GSK, Johnson&Johnson, Leica Biosystems, Lilly, Menarini Group, MSD, Novartis, Roche, Sysmex, ThermoFisher, and Veracyte; reports research support from AstraZeneca, Gilead, GSK, Menarini, Novartis, Owkin, Pfizer, Pillar Biosciences, and Roche; and reports travel support from AstraZeneca, Novartis, and Roche. D.S.R. reports personal fees from AstraZeneca. M.F.B. received personal fees from AstraZeneca and possesses intellectual property rights with SOPHiA Genetics. D.B.S. has consulted/received honoraria from Pfizer, BridgeBio, Scorpion Therapeutics, FORE Therapeutics, Corramedical, Inc., and Meliora Therapeutics, Inc. G.C. has received institutional grants and/or funding from Astellas, AstraZeneca, Blueprint Medicine, BMS, Daiichi Sankyo, Kymab, Merck, Novartis, Philogen, Relay Therapeutics, Roche, and Sanofi; and consultation, ad board, or honoraria as a speaker or writer from AstraZeneca, BMS, Daiichi Sankyo, Exact Sciences, Gilead, Lilly, Menarini, Merck, Novartis, Pfizer, Roche, and Veracyte; reports non-financial leadership or advisory roles with ESMO, Europa Donna, EUSOMA, and Fondazione Beretta. M.E.R. reports honoraria from Research to Practice, Intellisphere, Physicians' Education Resource, MyMedEd, MJH Holdings; reports uncompensated consulting or advisory work with Artios Pharma, AstraZeneca, Exact Sciences, Foundation Medicine, Merck; reports institutional research funding from Artios Pharma, AstraZeneca, Merck. S.C. has received institutional grant/funding from Pathos.ai and AstraZeneca, consultation/Ad board/Honoraria from Daiichi-Sankyo, AstraZeneca, Novartis, Casdin Capital, Nuvalent, Merck, Bayer, HighTop, PathosAI, and Revelio. P.R. has received institutional grants and/or funding from Grail, Novartis, AstraZeneca, Invitae, Biotheranostics, Tempus, Neogenomics, Guardant Health, Personalis, Myriad Genetics, Foresight Diagnostics, Natera, Biodesix, SAGA Diagnostics, SOPHiA Genetics, Haystack and Roche; and consultation, ad board or honoraria from Novartis, AstraZeneca, Lilly/Loxo, Stemline Therapeutics, Prelude Therapeutics, Neogenomics, Regor Pharmaceuticals, SAGA Diagnostics, SOPHiA Genetics, Tempus, Puma, Myriad Genetics, Foresight Diagnostics, Natera, Pathos AI and BioNTech. The other authors declare no competing interests.



676 **Figure 1. Clinical and genomic prognostic biomarkers for first-line CDK4/6i+ET. (A)**
677 Representation of the MSK-cohort, encompassing 1,264 individuals with breast cancer who
678 received first-line CDK4/6i+ET, together with tumor genomic sequencing at baseline. **(B)**
679 Distribution of clinical variables in the MSK-cohort. **(C)** Median PFS on first-line CDK4/6i+ET in
680 the MSK-cohort, together with the 12-month and 24-month PFS rate. **(D)** Genomic landscape of
681 the MSK-cohort, with the top annotation showing the co-segregation with relevant clinical
682 features. **(E)** Association between clinical and genomic variables in the MSK-cohort,

683 demonstrating expected associations measured using the Fisher's exact tests and log_2(odds-
684 ratios), with significance defined using the Benjamini-Hochberg procedure (* denotes q-value
685 <0.05, ** q-value <0.01, *** q-value <0.001). **(F)** Univariable Cox-regression between clinical
686 variables and PFS on first-line CDK4/6i+ET, showing the effect size and statistical significance
687 (adjusted for multiple testing using the Benjamini-Hochberg procedure). **(G)** Clinical outcomes
688 according to currently adopted criteria of high-risk in clinical trials investigating escalated first-line
689 therapeutic options in HR+/HER2– MBC, namely TFI ≤12 months. **(H)** Prognostic genomic
690 alterations for first-line CDK4/6i+ET, demonstrated with a univariable Cox regression with
691 significance determined using the Benjamini-Hochberg procedure. **(I)** Clinical outcomes
692 according to *TP53* (left) and *RB1* alteration status (right).

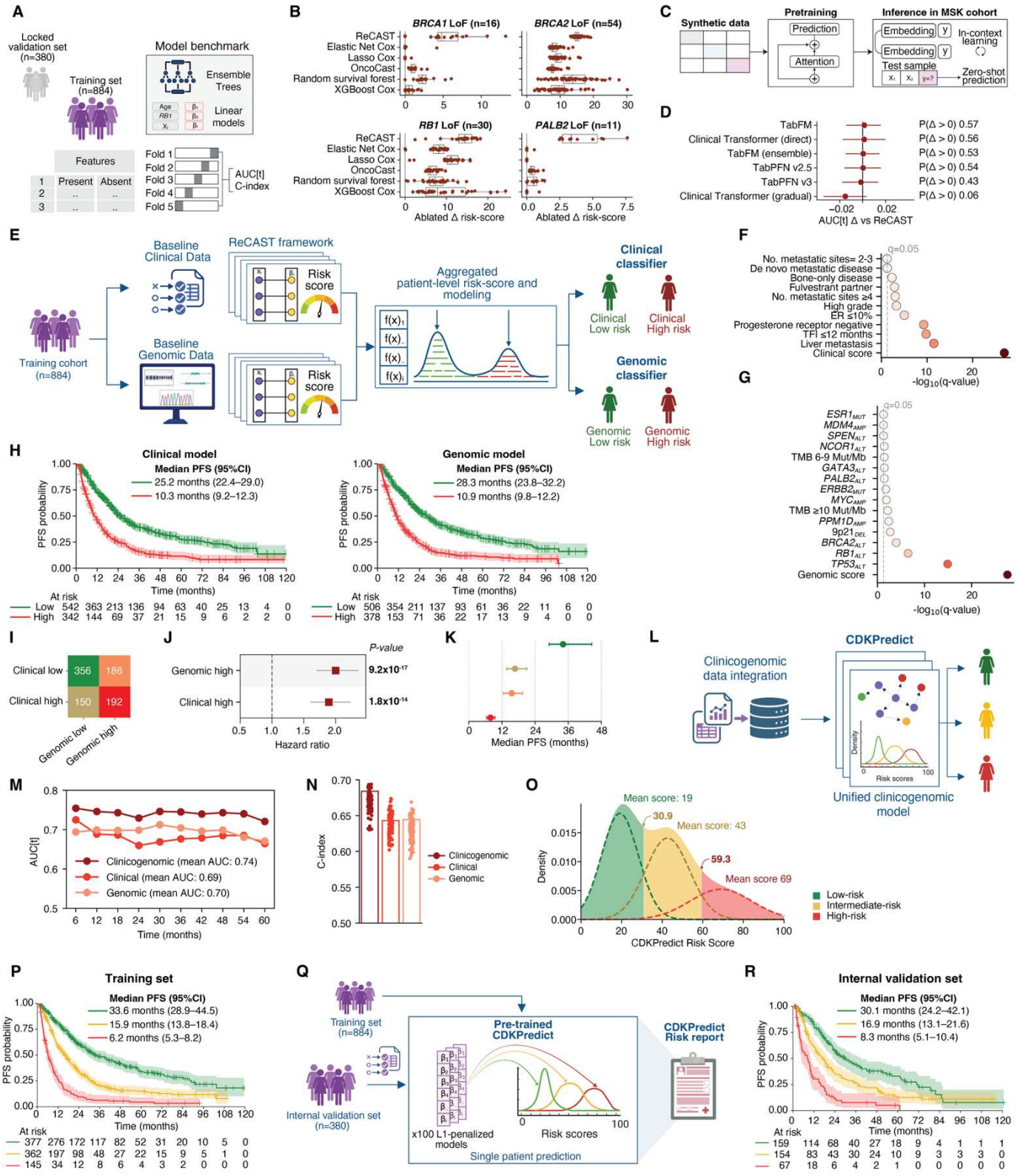

A
Locked validation set (n=380)
Training set (n=884)
Model benchmark
Ensemble Trees
Linear models
Features
Present
Absent
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
AUC[t] C-index
B
BRCA1 LoF (n=16)
BRCA2 LoF (n=54)
RB1 LoF (n=30)
PALB2 LoF (n=11)
ReCAST
Elastic Net Cox
Lasso Cox
OncoCast
Random survival forest
XGBoost Cox
Ablated Δ risk-score
C
Synthetic data
Pretraining
Prediction
Attention
Inference in MSK cohort
Embedding
In-context learning
Test sample
Zero-shot prediction
D
TabFM
Clinical Transformer (direct)
TabFM (ensemble)
TabPFN v2.5
TabPFN v3
Clinical Transformer (gradual)
P(Δ > 0) 0.57
P(Δ > 0) 0.56
P(Δ > 0) 0.53
P(Δ > 0) 0.54
P(Δ > 0) 0.43
P(Δ > 0) 0.06
AUC[t] Δ vs ReCAST
E
Baseline Clinical Data
ReCAST framework
Risk score
Aggregated patient-level risk-score and modeling
Clinical classifier
Clinical Low risk
Clinical High risk
Training cohort (n=884)
Baseline Genomic Data
Genomic classifier
Genomic Low risk
Genomic High risk
F
q=0.05
No. metastatic sites= 2-3
De novo metastatic disease
Bone-only disease
Fulvestrant partner
No. metastatic sites ≥4
High grade
ER ≤10%
Progesterone receptor negative
TFI ≤12 months
Liver metastasis
Clinical score
-log10(q-value)
G
ESR1MUT
MDM4AMP
SPENALT
NCOR1ALT
TMB 6-9 Mut/Mb
GATA3ALT
PALB2ALT
ERBB2MUT
MYCAMP
TMB ≥10 Mut/Mb
PPM1DAMP
9p21DEL
BRCA2ALT
RB1ALT
TP53ALT
Genomic score
H
Clinical model
Median PFS (95%CI)
25.2 months (22.4–29.0)
10.3 months (9.2–12.3)
Genomic model
28.3 months (23.8–32.2)
10.9 months (9.8–12.2)
PFS probability
Time (months)
At risk
Low 542 363 213 136 94 63 40 25 13 4 0
High 342 144 69 37 21 15 9 6 2 2 0
Low 506 354 211 137 93 61 36 22 11 6 0
High 378 153 71 36 22 17 13 9 4 0 0
I
Clinical low
Clinical high
356
186
150
192
Genomic low
Genomic high
J
P-value
Genomic high
Clinical high
9.2x10-17
1.8x10-14
Hazard ratio
K
Median PFS (months)
L
Clinicogenomic data integration
CDKPredict
Density
Risk scores
Unified clinicogenomic model
M
AUC[t]
Clinicogenomic (mean AUC: 0.74)
Clinical (mean AUC: 0.69)
Genomic (mean AUC: 0.70)
Time (months)
N
C-index
Clinicogenomic
Clinical
Genomic
O
Mean score: 19
30.9
Mean score: 43
59.3
Mean score 69
Density
CDKPredict Risk Score
Low-risk
Intermediate-risk
High-risk
P
Training set
Median PFS (95%CI)
33.6 months (28.9–44.5)
15.9 months (13.8–18.4)
6.2 months (5.3–8.2)
PFS probability
Time (months)
At risk
377 276 172 117 82 52 31 20 10 5 0
362 197 98 48 27 22 15 9 5 1 0
145 34 12 8 6 4 3 2 0 0 0
Q
Training set (n=884)
Internal validation set (n=380)
Pre-trained CDKPredict
x100 L1-penalized models
Risk scores
Single patient prediction
CDKPredict Risk report
R
Internal validation set
Median PFS (95%CI)
30.1 months (24.2–42.1)
16.9 months (13.1–21.6)
8.3 months (5.1–10.4)
PFS probability
Time (months)
At risk
159 114 68 40 27 18 9 4 1 1 1
154 83 43 30 24 10 9 3 3 3 0
67 18 6 4 2 1 0 0 0 0 0

Figure 2. Development of a clinicogenomic model designed for baseline risk stratification on first-line CDK4/6i+ET. (A) Comparison of interpretable models for CDK4/6i+ET prognostication, conducted in the training set using a 5-fold cross validation. **(B)** Model-reliance analysis investigating the sensitivity of predictors to key genomic drivers of CDK4/6i+ET resistance, showing the difference in risk-score (from 0 to 100 for all models) using a feature deletion strategy (**Methods**). **(C)** State of the art foundation model for tabular data, which underwent in-context learning in the MSK training cohort before zero-shot prediction. **(D)** Contrast score analysis using ReCAST as reference, with P(Δ>0) denoting the proportion of times the transformer comparator models outperformed the reference across n=1,000 bootstrap iterations of the training set. **(E)** Development strategy of a clinical and genomic model using the ReCAST framework. **(F)** Univariable Cox-regression using clinical variables included in the clinical model and the clinical risk-scores, with significance defined using the Benjamini-Hochberg procedure. **(G)** Same as above, but showing significant genomic variables (q<0.05) and the genomic risk-scores. **(H)** Clinical validity of the clinical (left) and genomic (right) classifiers in the MSK-training set. **(I)** Concordance between the clinical and genomic classifiers shown with a confusion matrix, demonstrating only fair agreement between the two models (Cohen's kappa=0.21). **(J)** Multivariable Cox-regression including genomic and clinical risk-groups, demonstrating the non-overlapping prognostication between the clinical and genomic risk estimation. **(K)** Forest plot showing the median PFS and 95% confidence intervals of risk-strata according to the level of agreement between the two classifiers. **(L)** Development of a clinicogenomic model integrating both clinical and genomic variables to result in a 3-risk strata classification. **(M)** Time-dependent AUC comparing the performance of the three models, integrating or not clinicogenomic variables. **(N)** C-index between the 3 independent classifiers. The bar value demonstrates the final C-index of the ensemble model, and the dots demonstrate the C-index of the n=100 base learners encompassed within each ensemble architecture. **(O)** Definition of 3-risk groups by the clinicogenomic model using a 3-component Gaussian mixture model, displaying the probability density function of the three components and cutoffs, the latter defined using Brent's method. **(P)** Clinical validity of the clinicogenomic model in the training set. **(Q)** Following development, the model was independently tested on n=380 (30%) patients who received routine care at MSK (holdout internal validation set). **(R)** Assessment of the clinicogenomic model in the holdout internal validation set.

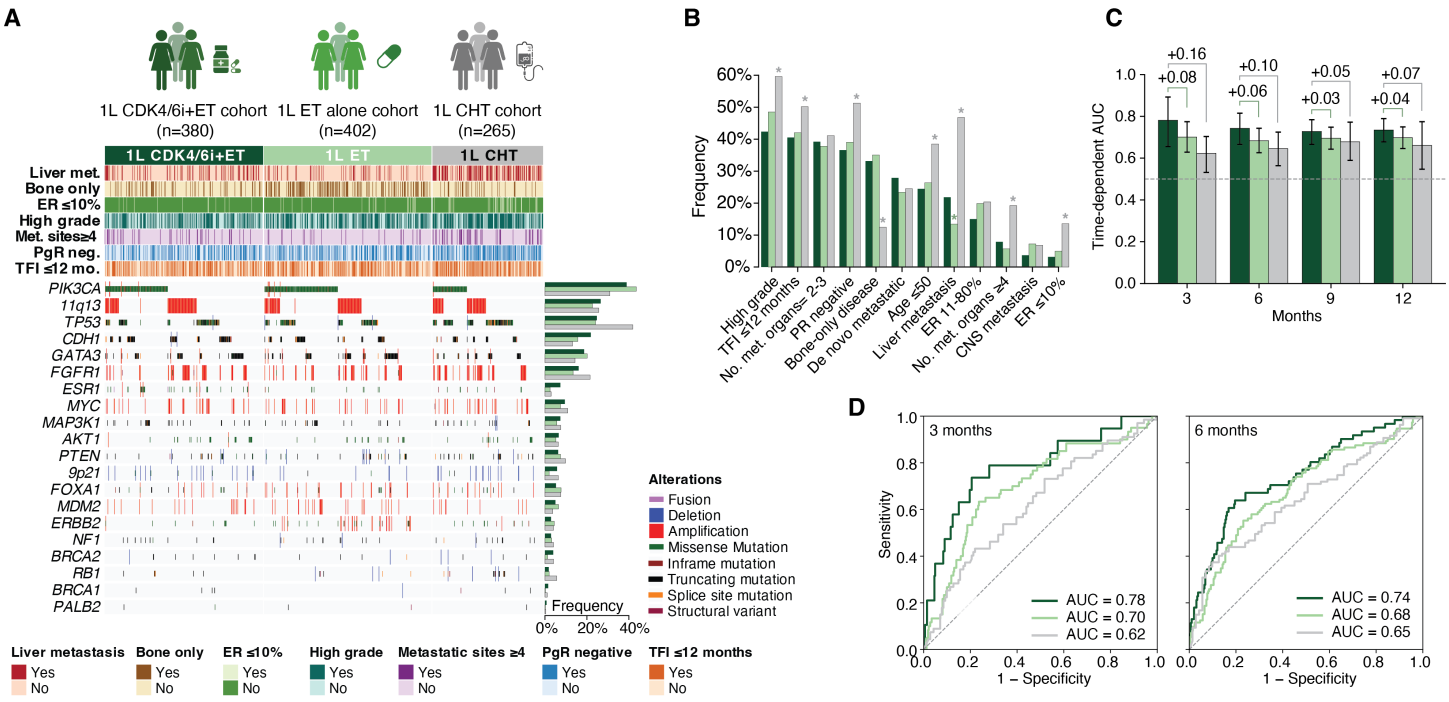
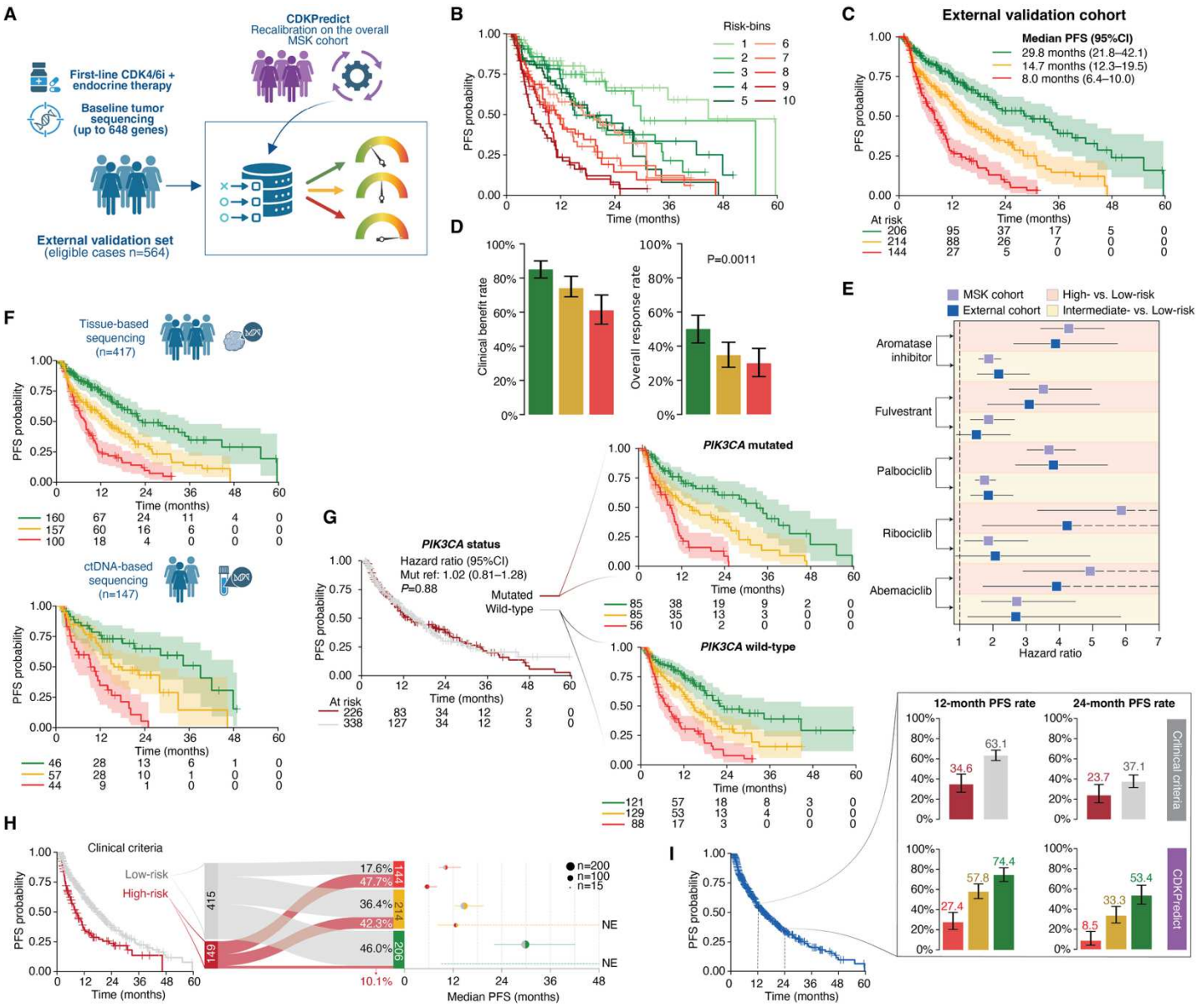
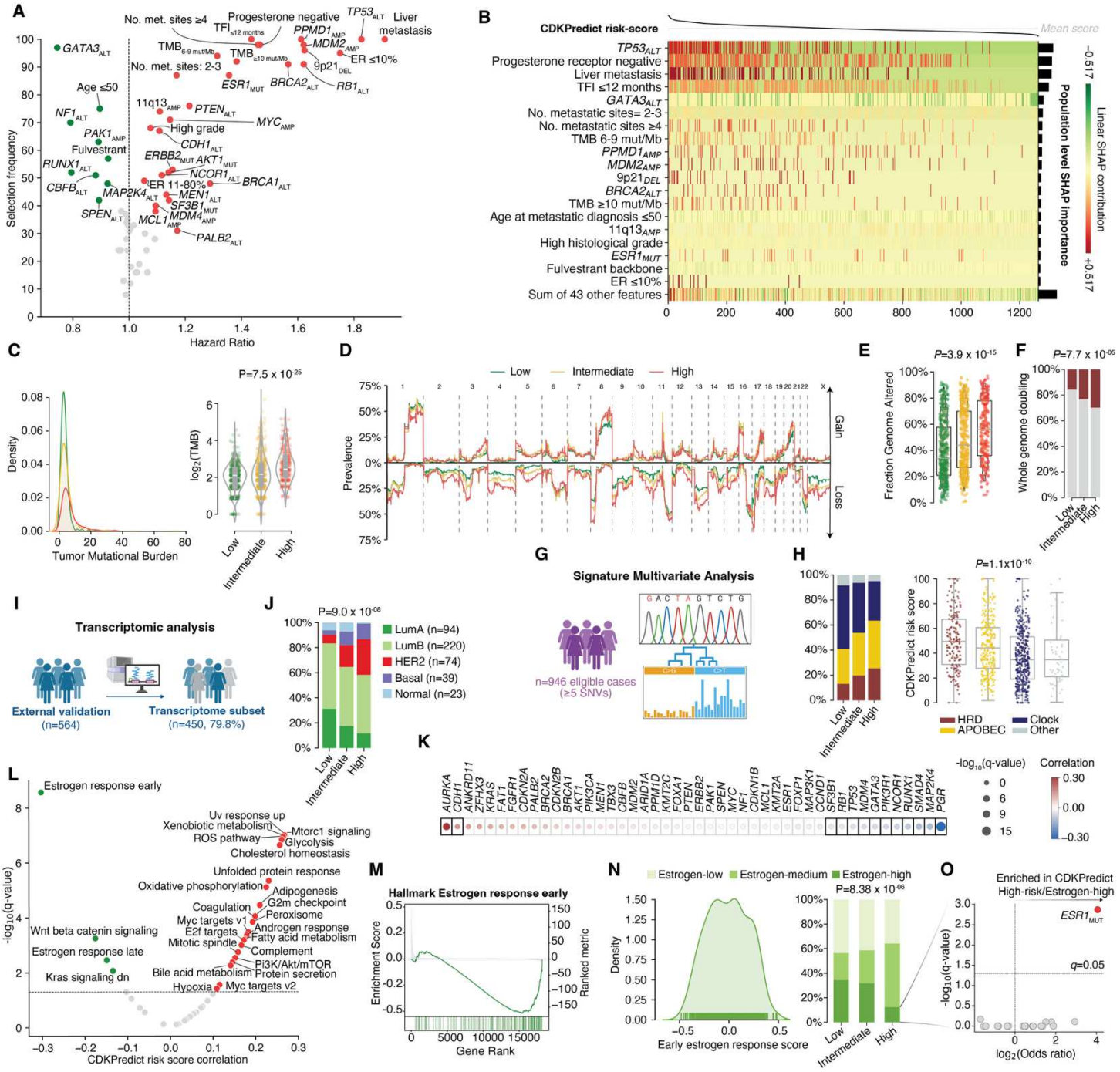


Figure 3. Preliminary demonstration of CDKPredict specificity for endocrine-based regimens (ET +/- CDK4/6i). **(A)** MSK cohorts of patients treated with first-line CDK4/6i+ET included in the internal validation cohort (n=380), patients treated with first-line ET alone (n=402), and patients with HR+/HER2– MBC who received first-line chemotherapy (n=265). **(B)** Comparison of clinicopathological variables across treatment cohorts, with the chemotherapy group showing poor prognostic variables; stars denote a *q*-value <0.01 using a Fisher's exact test conducted between each treatment arm using the CDK4/6i+ET group as comparator. **(C)** Time-dependent AUC comparison across treatment arms; bars represent the mean AUC[t] at the 12-month landmark analysis, error bars the 95%CI measured across 1,000 bootstrap iterations, and the top annotation showing the mean-delta difference across bootstrap iterations between the $AUC[t]_{CDK4/6i+ET}$ vs. $AUC[t]_{ET/chemotherapy}$ at the specific timepoint. **(D)** Time dependent receiver operating characteristic (ROC) curves to detect early progressors at 3 and 6 months across treatment arms, showing higher discrimination for CDK4/6i+ET and ET alone compared to chemotherapy.



740 **Figure 4. Independent clinical validation of CDKPredict in an external multi-institutional**
741 **cohort. (A)** Workflow of the clinical validation of CDKPredict, which, after fine-tuning on the overall
742 MSK-cohort, was tested using an external multi-institutional cohort of patients who received first-
743 line CDK4/6i+ET and tumor sequencing at baseline (n=564 cases). **(B)** Association between
744 continuous clinicogenomic risk-scores and median PFS, showing reduced time to progression
745 with increasing risk-scores (divided in 10-quantile bins). **(C)** Clinical validity of CDKPredict in the
746 external cohort, showing robust and clinically meaningful baseline risk stratification. **(D)** Disease
747 control at 6 months (*left*) and overall response rate by radiological evaluation (*right*), with
748 significance tested using the Pearson's chi-square test. **(E)** Clinical validity of CDKPredict across

FDA-authorized compounds, with each bin (red and yellow) demonstrating the univariable Cox-hazard ratio of the high (red) and medium (yellow) risk groups compared with the low-risk category, utilized as reference. **(F)** Performance of CDKPredict across distinct sequencing platforms, with model classifications retaining robust clinical validity for tumors sequenced using solid- and liquid-based panels. **(G)** Clinical validity of CDKPredict according to *PIK3CA* mutation status. **(H)** Sankey plot showing the level of agreement between historic clinical criteria of high-risk (TFI ≤12 months) and CDKPredict classifications, with the resulting progression-free survival according to the level of agreement demonstrated in the corresponding forest plot (*right*). **(I)** 12-month and 24-month progression-free survival according to historic clinical criteria of risk and CDKPredict classifications, demonstrating the superior clinical validity of the clinicogenomic model.



761 **Figure 5. Clinical and biological plausibility of CDKPredict. (A)** Feature importance across
762 the CDKPredict framework, with the selection frequency by base learners displayed on the y-axis,
763 and the mean effect size on the x-axis. **(B)** Interpretable features contribution for CDKPredict risk-
764 score using Shapley additive explanations (SHAP); colors denote the linear contribution of
765 features in the risk-score estimation, with the overall population-level importance by SHAP values

shown by the histogram on the right; the top annotation demonstrates the overall CDKPredict risk-scores of predicted samples, sorted in descending order. **(C)** Probability density distribution (*left*) of TMB and log_2-scale distribution (*right*) across risk-groups in the MSK-cohort, with significance calculated using the Kruskal-Wallis test. **(D)** Genome-wide copy-number profile across risk-strata, showing an overall increasing burden of copy number alterations in higher risk-classes. **(E)** Boxenplot of fraction genome altered across risk groups, with significance calculated using the Kruskal-Wallis test. **(F)** Whole-genome doubling events across risk-strata, with significance calculated using the Pearson's chi-square test. **(G)** Assessment of tumor mutational signatures using SigMA for evaluable tumors carrying at least n=5 single-nucleotide variants (n=946, 74.8%). **(H)** Representation of dominant tumor mutational signatures across risk groups (*left*), and CDKPredict risk-scores across dominant tumor mutational signatures (*right*), with significance measured using the Kruskal-Wallis test. **(I)** Study schema for the transcriptomic analysis, including patients from the external cohort who received whole-exome transcriptomic sequencing at baseline (n=450 of 564, 79.8%). **(J)** PAM50 intrinsic subtypes across CDKPredict risk-strata, with significance determined using the Pearson's chi-square test. **(K)** Correlation between clinicogenomic risk-scores and RNA expression of variables included in CDKPredict, calculated using the Spearman's correlation, with significance measured with the Benjamini-Hochberg procedure, with black boxes denoting significant correlations (q<0.05). **(L)** Spearman's correlation between clinicogenomic risk-scores and pathway-level activity utilizing the Hallmark gene set collection. **(M)** Gene-set enrichment analysis of the Hallmark Estrogen Response Early gene set. Genes were ranked based on differential expression between low- and high-risk tumors as determined by DESeq2. The gene set was significantly negatively enriched in high-risk tumors (normalized enrichment score=-1.90, q=0.0010). **(N)** Probability density distribution of the Hallmark estrogen response early pathway in the study population, with estrogen-high tumors showing a relative enrichment among lower risk classes. **(O)** Gene enrichment analysis in CDKPredict high-risk/estrogen-high compared with CDKPredict high-risk/non-estrogen-high tumors, measured with a Fisher's exact test.

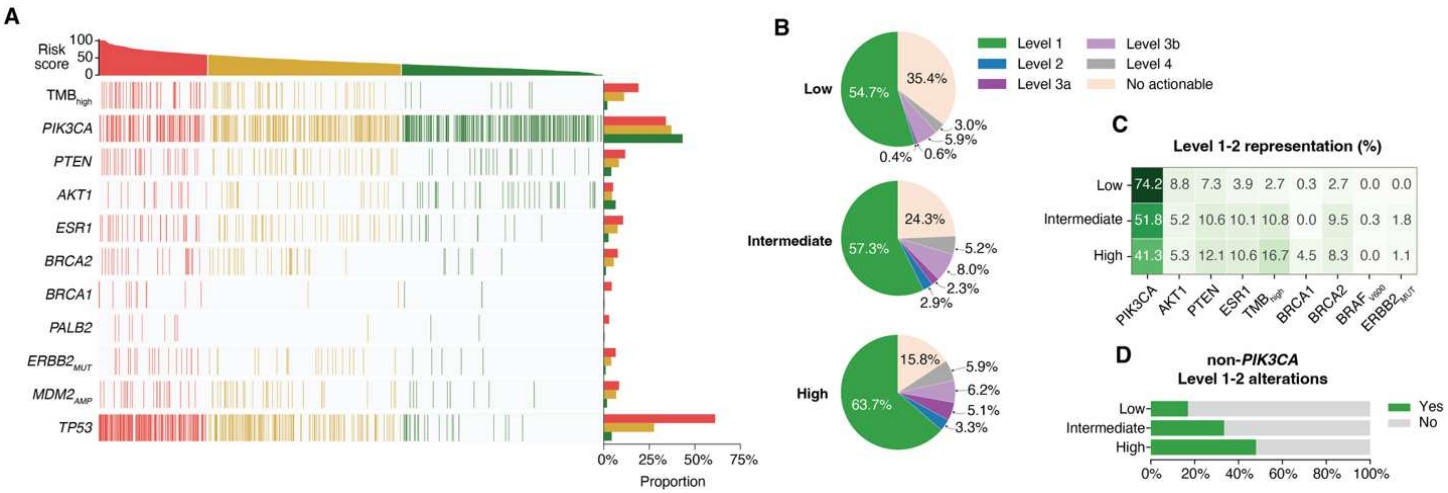


Figure 6. Actionable targets across CDKPredict risk-classes. (A) Oncoprint of variables included in CDKPredict with relevant actionability in breast cancer, with the top annotation demonstrating CDKPredict risk-scores in decreasing order, and the right barplot showing feature frequency across CDKPredict risk classes. **(B)** Overall representation of the highest-level of actionability per patient across risk-strata, as annotated in OncoKB, an FDA-Recognized Human Genetic Variant Database; overall, a higher prevalence of actionable alterations was observed across higher-risk groups. **(C)** Representation of level 1 and level 2 alterations across risk-strata, including as the denominator the frequency of level 1-2 alterations per risk-group, showing the higher representation of additional standard-of-care actionable targets among higher risk-groups. **(D)** Level 1 and level 2 alterations once excluding *PIK3CA* alterations across CDKPredict risk-groups.

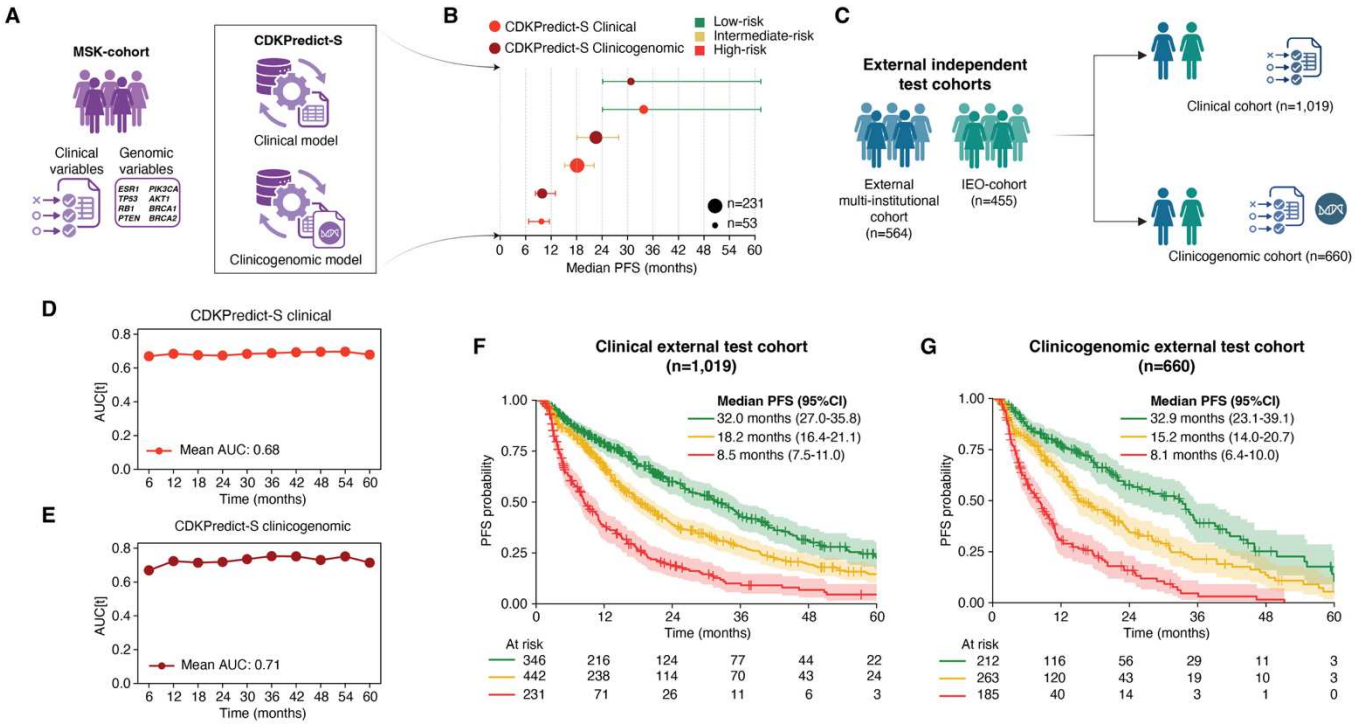


Figure 7. Development and clinical validation of CDKPredict-S. (A) Framework for developing CDKPredict-S, a scalable model intended for variable resource settings, using a restricted subset of variables to facilitate its broad clinical implementation. **(B)** Performance of CDKPredict-S in the hold-out MSK validation set (n=380), with the forest plot showing the median PFS and the confidence intervals according to CDKPredict-S clinical (*orange*) and CDKPredict-S clinicogenomic (*dark red*) risk-estimation. **(C)** External independent clinical validation of CDKPredict-S across two external cohorts of patients treated with first-line CDK4/6i+ET, with or without available genomic sequencing at baseline. Overall, 1,019 patients were available for the clinical validation of the clinical model, while 660 patients received baseline tumor-sequencing and were included for the clinical validation of the clinicogenomic model. **(D)** Time dependent AUC of the clinical model in the pooled external validation cohort (n=1,019). **(E)** Time dependent AUC of the clinicogenomic model in the pooled external validation cohort of patients with available baseline genomic sequencing (n=660). **(F)** Clinical validation of the clinical CDKPredict-S, demonstrating a statistically significant and clinically meaningful separation of outcomes according to the baseline risk stratification. **(G)** Same, showing the clinical validation of the clinicogenomic model in patients with available genomic sequencing at baseline.

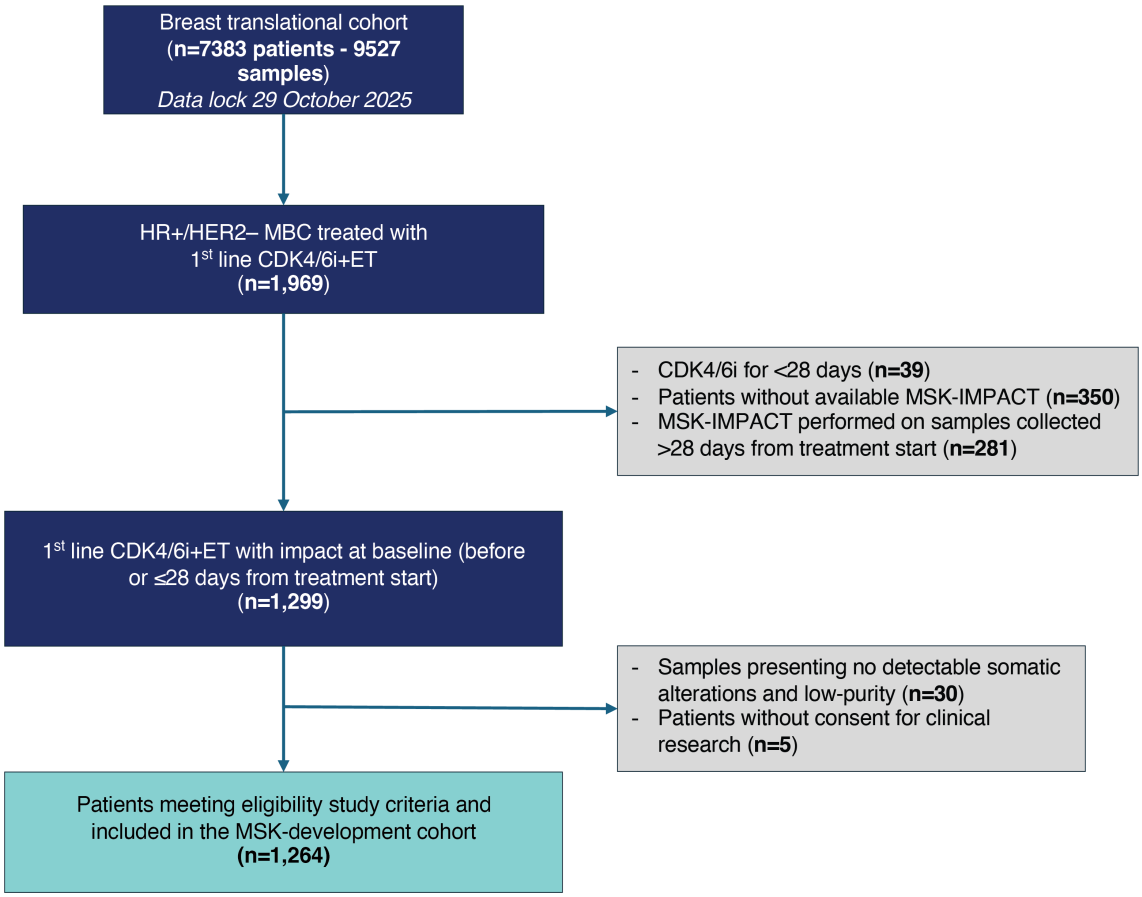


Figure S1. Consort diagram of the MSK-cohort. Consort diagram demonstrating inclusion and exclusion criteria of individuals with breast cancer included in the Memorial Sloan Kettering Breast Translational Dataset for the definition of the development cohort. Ultimately, 1,264 patients affected by HR+/HER2– breast cancer who received first-line CDK4/6i+ET together with baseline tumor-genomic sequencing were included in the study population.

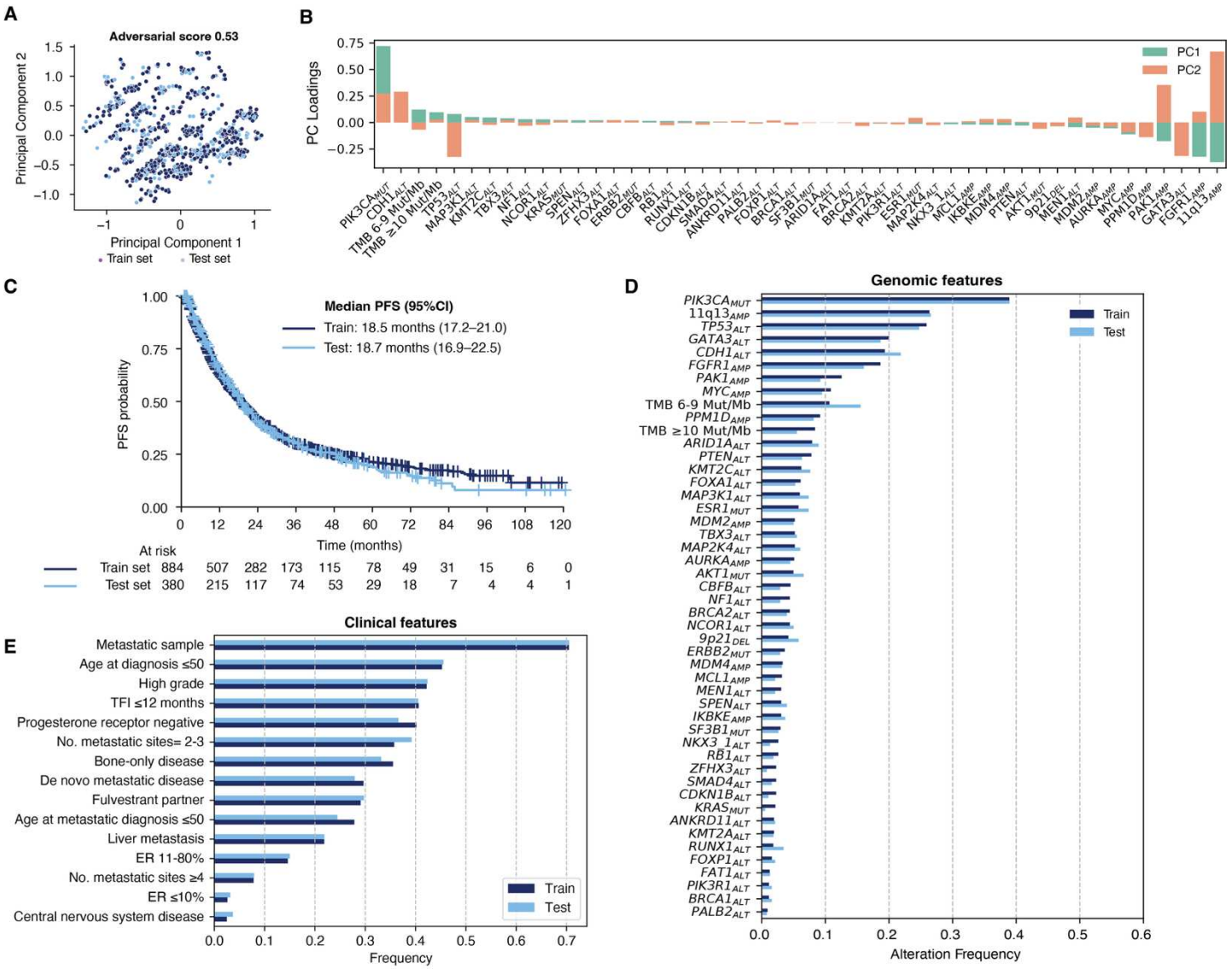


Figure S2. Comparison of genomic and clinical features between the training and the holdout MSK validation set. (A) Principal component analysis between the training and the internal holdout validation set in the MSK-cohort using genomic variables and the n=2 principal components, showing an overall overlapping high-dimensional embedding of genomic features (Adversarial score of a random forest classifier of 0.53, **Methods**). **(B)** Principal component loadings showing the contribution of the original genomic features in the definition of the n=2 principal components (**Methods**). **(C)** Clinical outcomes on first-line CDK4/6i+ET between the training and validation set. **(D)** Frequency of genomic alterations between the training and MSK-validation set. **(E)** Same, using clinical features.

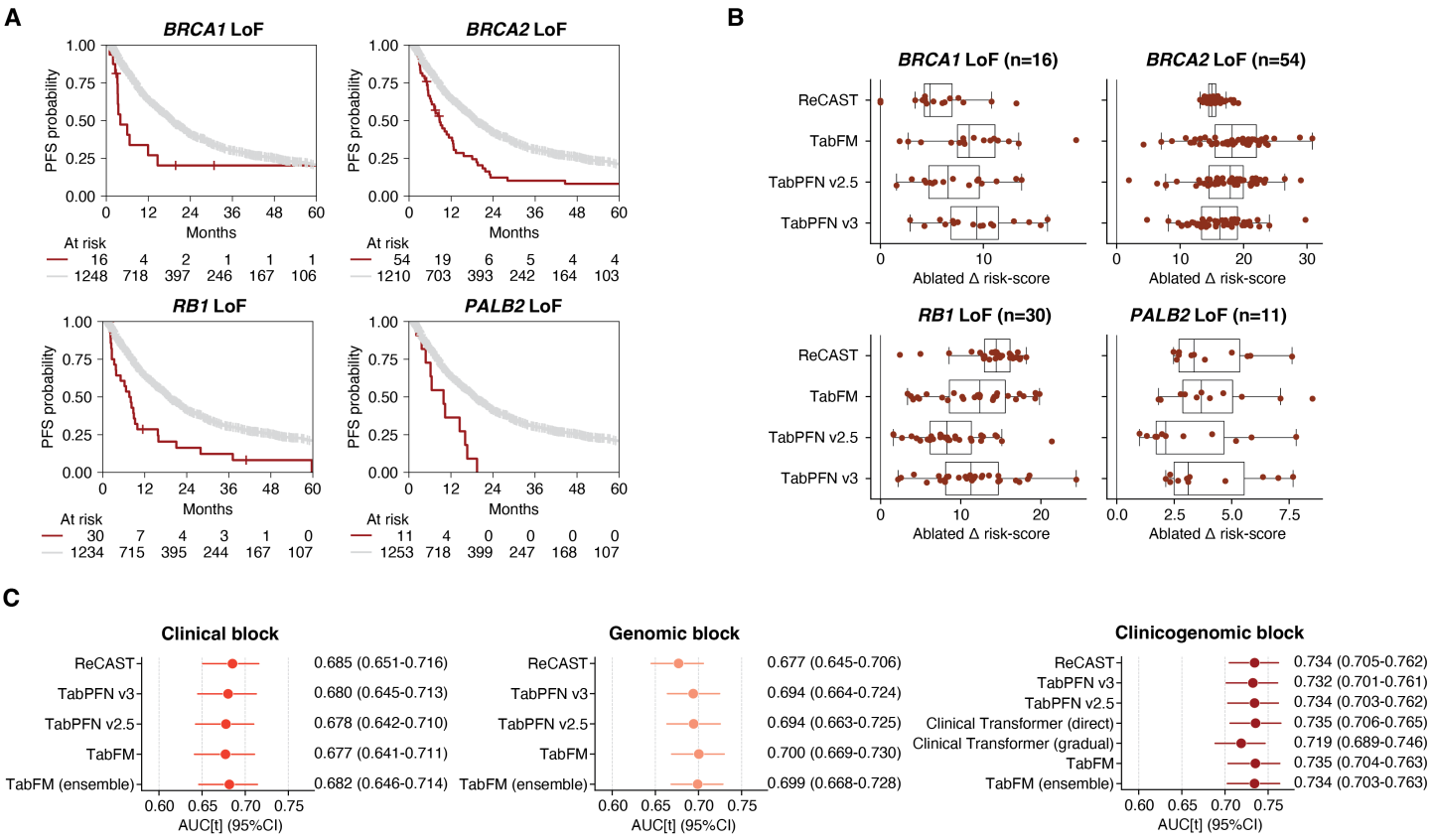
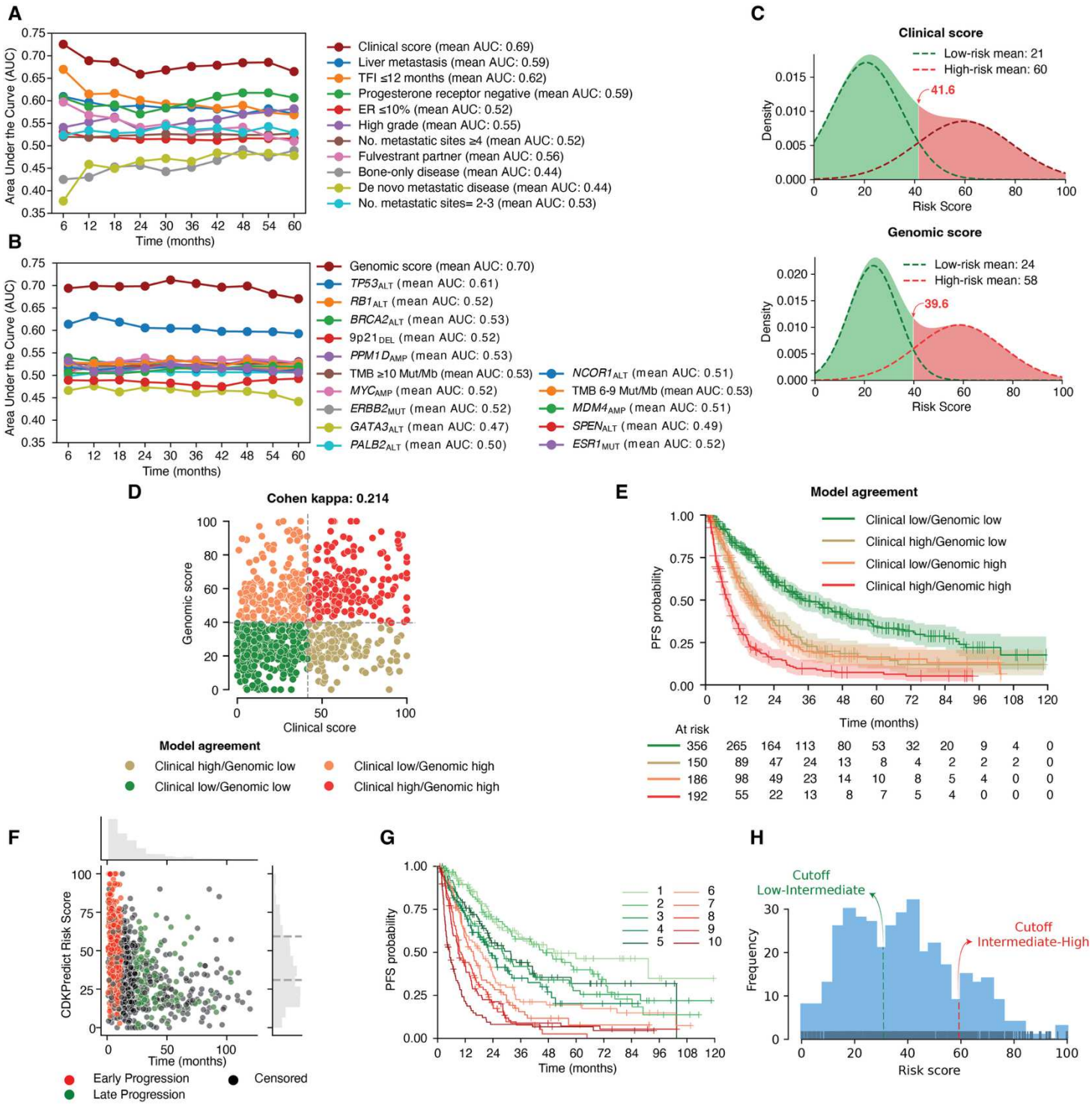
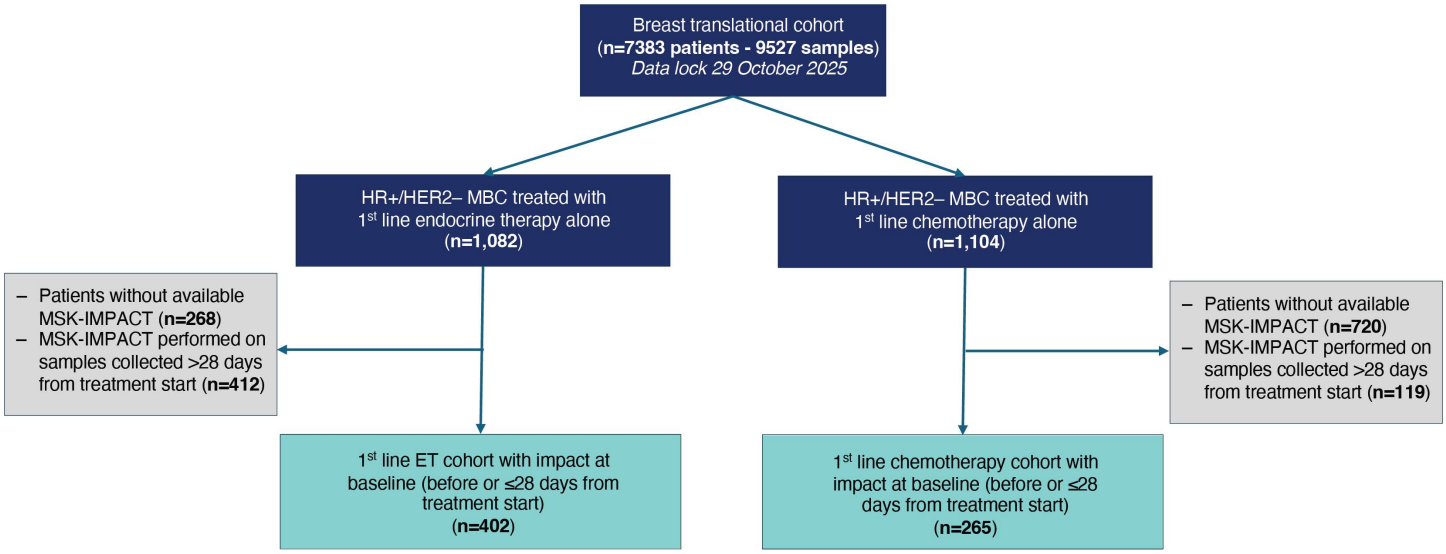


Figure S3. Model benchmark for CDK4/6i+ET prognostication. **(A)** Prognostication according to validated genomic drivers of CDK4/6i+ET resistance in the MSK-cohort (with patients carrying loss of function mutations [LoF] displayed in dark red). **(B)** Features-deletion analysis of predictors using rare genomic drivers of CDK4/6i+ET resistance, comparing ReCAST to tabular foundation models using a transformer architecture. **(C)** Comparison of ReCAST to state-of-the-art transformer architectures specifically designed for tabular data, separately displayed for predictions employing clinical-only, genomic-only, and combined clinicogenomic features; predictions were computed within a 5-fold cross validation framework, where fine-tuning was conducted in the 4 training partitions and risk estimation on the validation fold; 95%CI were measured across n=1,000 bootstrap resamples, with AUC[t] measured as the integrative mean from 6 to 60 months (using 10 time bins).



847 **Figure S4. Clinical validity of the prediction model. (A)** Comparison of the time-dependent
848 AUC at the 60-month landmark analysis including clinical risk-scores and standalone clinical
849 variables showing a significant association with PFS (not corrected for family wise error, P<0.05).
850 **(B)** same as above but using genomic-risk-scores and standalone genomic variables, showing

significant genomic variables as determined with a univariable Cox-regression corrected for multiple testing (q<0.05 by Benjamini-Hochberg procedure). **(C)** Representation of the probability density function using a 2-component Gaussian-mixture model for defining low- and high-risk tumors in the clinical (*above*) and genomic (*bottom*) model. **(D)** Correlation between genomic and clinical risk-scores, showing an overall low agreement between the clinical and genomic risk estimation. **(E)** Clinical outcomes according to the level of agreement between the two models, showing a markedly improved baseline risk stratification compared with standalone models. **(F)** Correlation between clinicogenomic risk-scores and PFS on CDK4/6i+ET in the training set, showing a progressively shorter time to progression with increasing risk-scores. Cases showing a PFS event before 12 months were colored in red, cases showing a PFS event ≥24 months in green, and censored patients in black. The top histogram denotes the time distribution, while the right histogram denotes the distribution of clinicogenomic risk-scores. **(G)** Association between continuous risk-scores of the clinicogenomic model and PFS on CDK4/6i+ET in the training set, categorized in n=10 quantile bins. **(H)** Risk-score distribution of patients included in the hold-out validation set related to the cutoffs between risk-strata established in the training cohort, showing a clear data-driven separation of the 3-risk groups to emphasize the plausibility of the model's classifications.



868 **Figure S5. Consort diagram of patients with HR+/HER2– MBC treated with first-line ET**
869 **alone or chemotherapy.** Additional independent MSK cohorts collected for evaluating the
870 predictive specificity of CDKPredict, including patients treated with first-line ET alone (n=402) and
871 patients who received first-line chemotherapy (n=265).

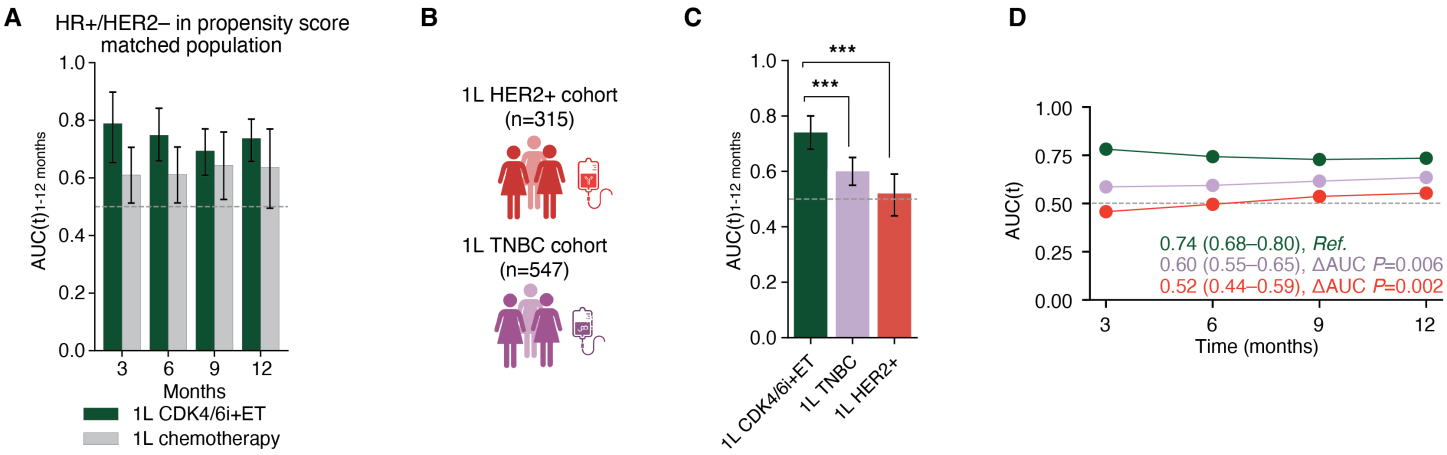


Figure S6. Evaluation of CDKPredict in propensity score matched population and different receptor subtypes. **(A)** Time dependent AUC (1-12 months) in the propensity score matched population of patients with HR+/HER2– MBC treated with first-line CDK4/6i+ET vs. chemotherapy to account for unbalanced clinical variables (n=195 pairs); once corrected for baseline clinical variables, CDKPredict continued to demonstrate significantly superior prognostication in those treated with CDK4/6i+ET compared to chemotherapy ($AUC[t]_{CDK4/6i+ET}$=0.74, 95%CI: 0.66–0.81 vs. $AUC[t]_{chemotherapy}$=0.62, 95%CI: 0.53–0.71, $P_{interaction}$=0.048). **(B)** Additional MSK-cohorts of patients with HER2+ breast cancer who received first-line THP regimens (n=315), and patients with TNBC treated with first-line chemotherapy (n=547). **(C)** Barplot comparing the 12-month time-window (3, 6, 9 and 12 months) AUC[t] across treatment groups, with *** denoting a *P*-value <0.01. **(D)** 12-month landmark analysis of AUC[t] differences between treatment arms, with CDKPredict showing poor predictive accuracy among TNBC and HER2+ breast cancers, arguing against its value as a generic marker of poor prognosis; *P*-value denotes mean AUC[t] differences (**Methods**).

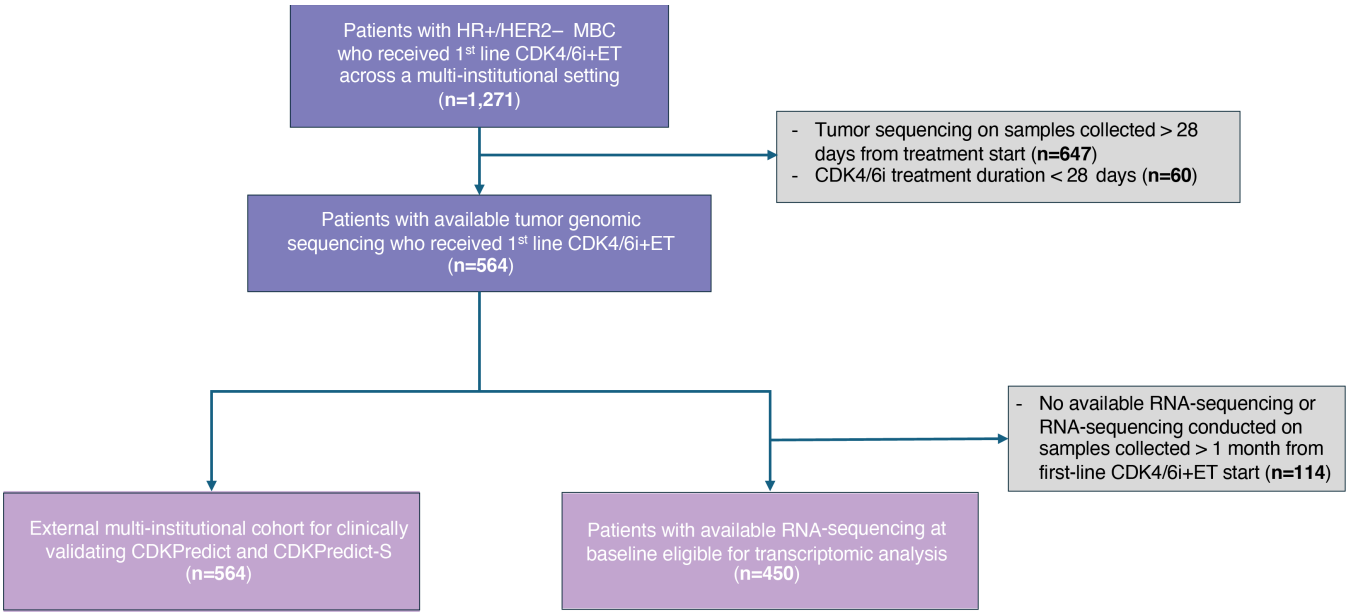


Figure S7. Consort diagram of the external multi-institutional cohort. Consort diagram representing inclusion and exclusion criteria for patients included in the external multi-institutional cohort employed for the independent clinical validation of CDKPredict. Ultimately, 564 individuals with breast cancer who received first-line CDK4/6i+ET and baseline tumor genomic sequencing were included for the clinical validation of the clinicogenomic model. Of them, 450 (79.8%) cases received baseline transcriptomic sequencing, and were considered for the transcriptomic analysis aimed at evaluating the biological plausibility of CDKPredict.

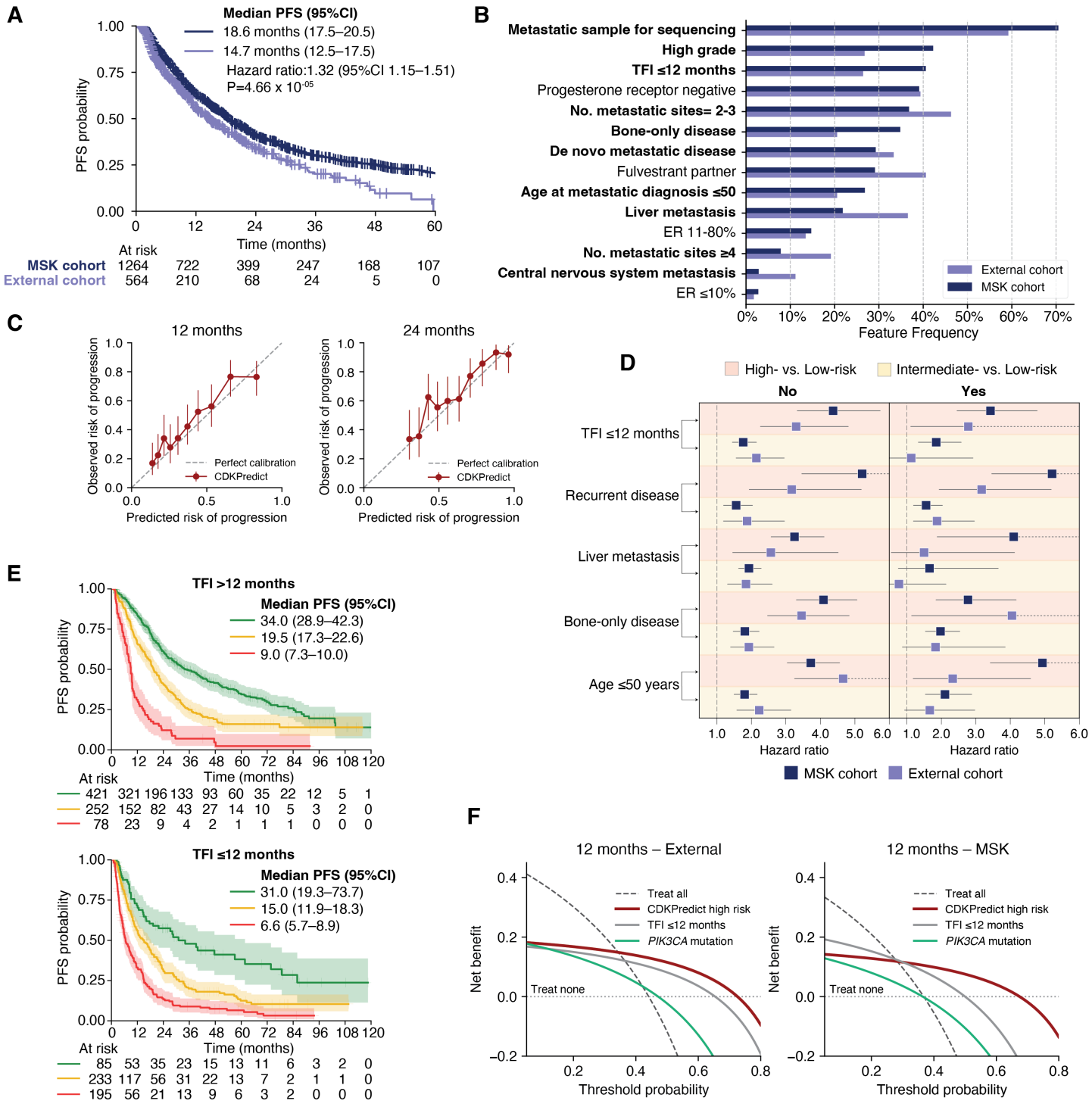


Figure S8. External multi-institutional cohort and clinical validation of CDKPredict. (A) Comparison of PFS on first-line CDK4/6i+ET between the MSK-cohort and external validation cohort. **(B)** Distribution of clinical variables between the overall MSK-cohort and the external validation set, with labels in bold representing clinical variables showing a significantly different

frequency between the two datasets (q<0.05 by Benjamini-Hochberg procedure). **(C)** Calibration plots comparing predicted and observed PFS at the 12-month and 24-month landmark analysis; the dashed diagonal line represents perfect calibration (i.e., predicted probabilities exactly match observed 12-month and 24-month PFS rates across risk-bins). **(D)** Subgroup analysis demonstrating the validity of CDKPredict across clinically relevant subgroups separately in the external and MSK-cohort, with the red and yellow rows demonstrating the hazard ratio of high-risk and intermediate-risk cases, respectively, relative to the low-risk group, utilized as reference, emphasizing the generalizability of the model across the overall population with breast cancer considered for first-line CDK4/6i+ET. **(E)** Clinical outcomes according to CDKPredict risk stratification relative to a treatment-free interval above (*top*) and equal or below 12 months (*bottom*), recapitulating the clinical observations from the external test cohort in the larger MSK-population. **(F)** Decision curve analysis showing greater net benefit of CDKPredict compared to TFI≤12 months across threshold probabilities (i.e., the lowest progression risk at which a clinician would escalate treatment; higher thresholds mean fewer unnecessary escalations are accepted for each correct one, e.g., three accepted unnecessary escalations for each one correct at 25%; one accepted unnecessary escalation for each one correct at 50%); a curve above treat none (net benefit >0) means correct escalations outweigh unnecessary ones at that threshold.

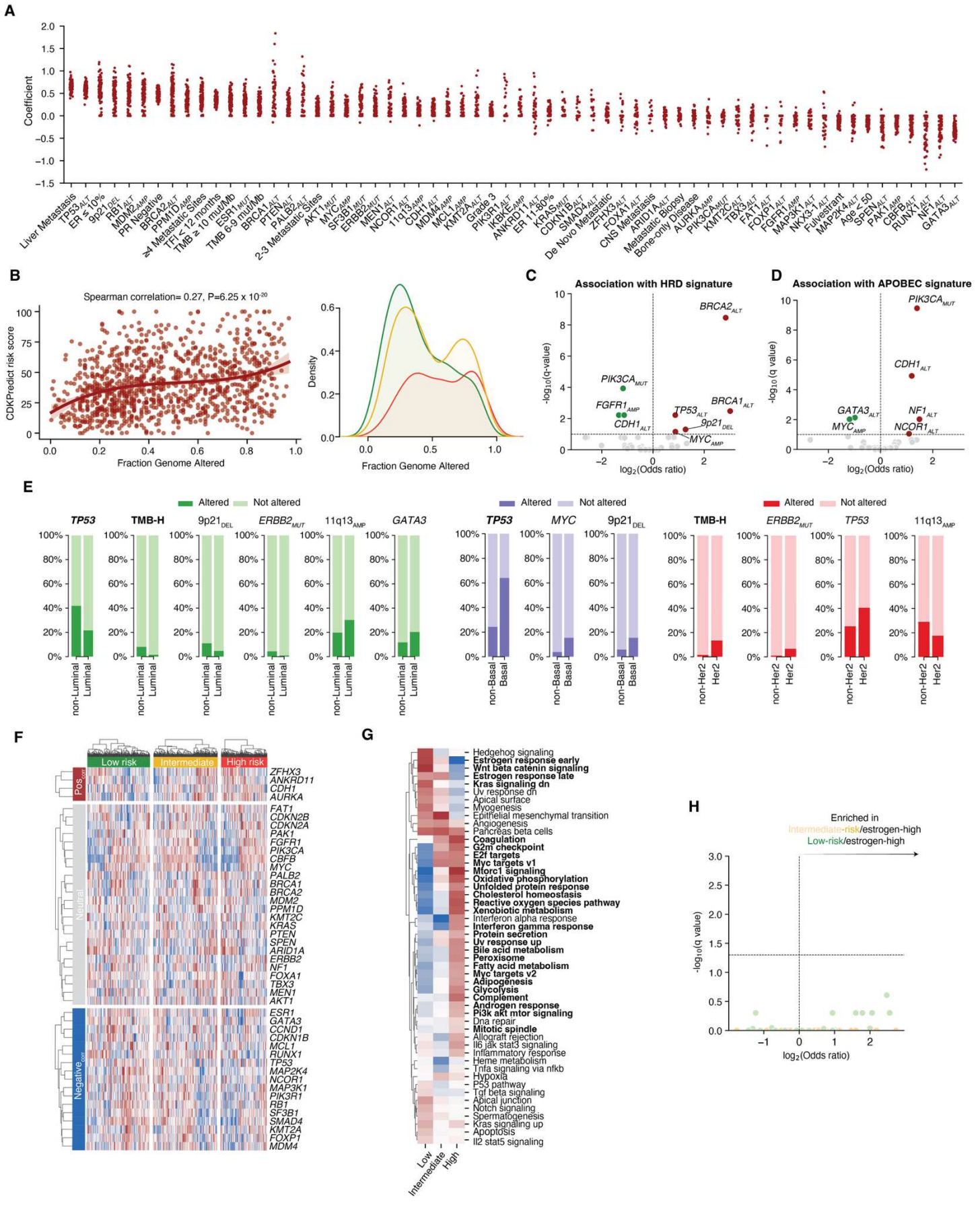
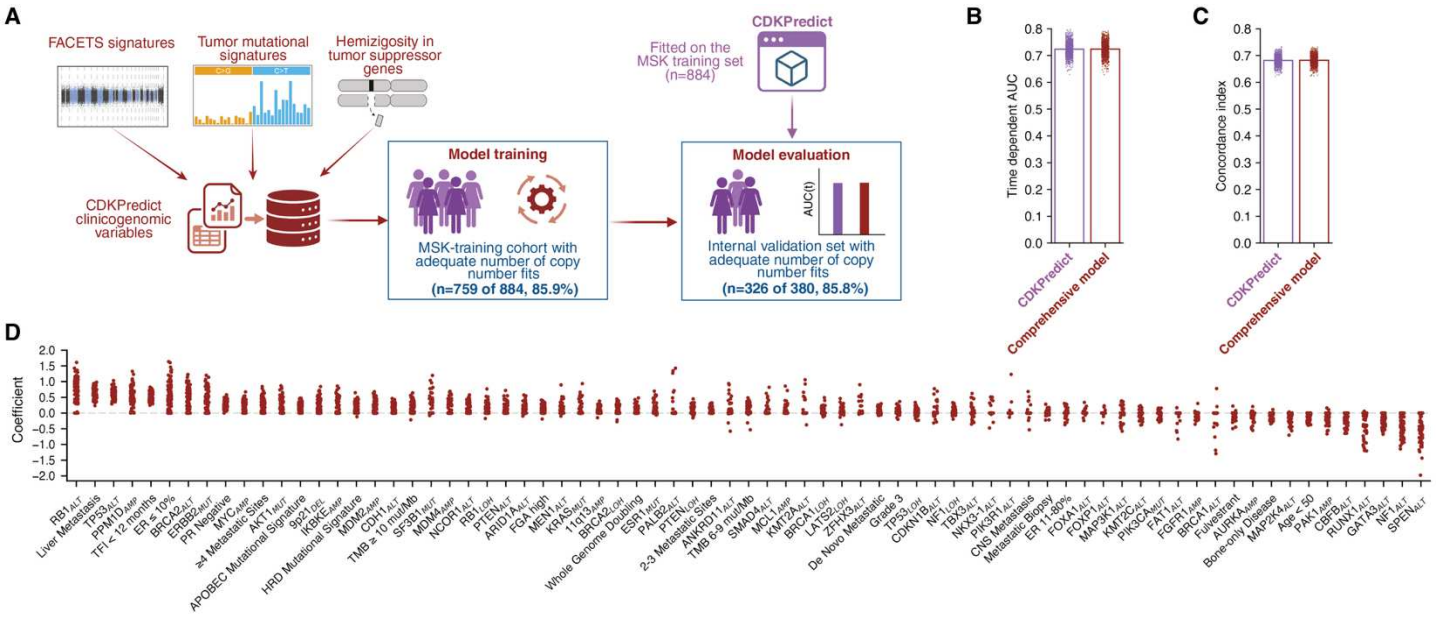
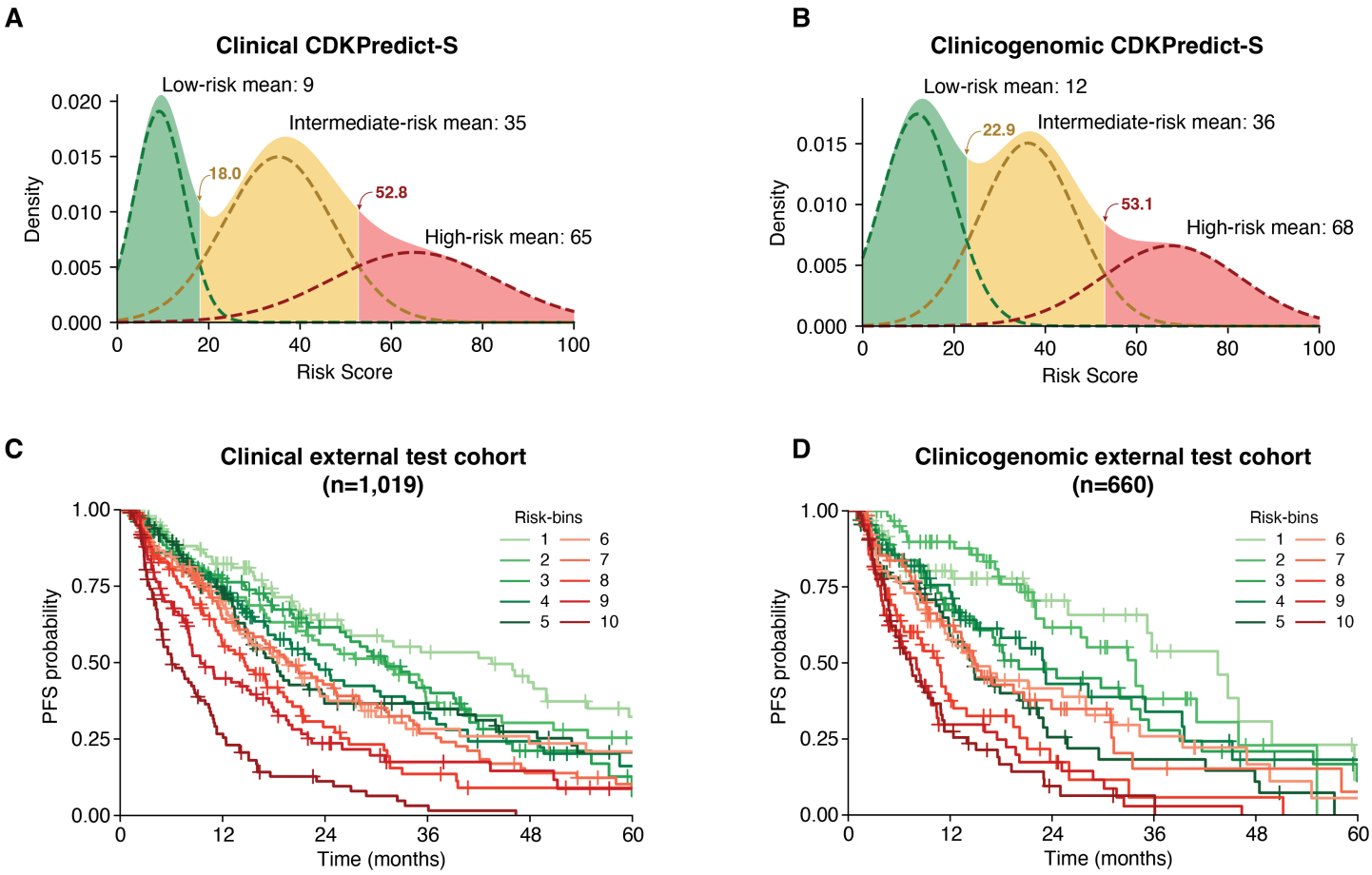

A
Coefficient
Liver Metastasis
TP53ALT
ER ≤ 10%
9p21DEL
RB1ALT
MDM2AMP
PR Negative
BRCA2ALT
PPM1DAMP
≥4 Metastatic Sites
TFI < 12 months
TMB ≥ 10 mut/Mb
ESR1MUT
TMB 6-9 mut/Mb
BRCA1ALT
PTENALT
PALB2ALT
2-3 Metastatic Sites
AKT1MUT
MYCAMP
SF3B1MUT
ERBB2MUT
MEN1ALT
NCOR1ALT
11q13AMP
CDH1ALT
MDM4AMP
MCL1AMP
KMT2AALT
Grade 3
PIK3R1ALT
IKBKEAMP
ANKRD11ALT
ER 11-80%
KRASMUT
CDKN1BALT
SMAD4ALT
De Novo Metastatic
ZFHX3ALT
FOXA1ALT
CNS Metastasis
ARID1AALT
Metastatic Biopsy
Bone-only Disease
AURKAAMP
PIK3CAMUT
KMT2CALT
TBX3ALT
FAT1ALT
FOXP1ALT
FGFR1AMP
MAP3K1ALT
NKX3-1ALT
Fulvestrant
MAP2K4ALT
Age ≤ 50
SPENALT
PAK1AMP
CBFBALT
RUNX1ALT
NF1ALT
GATA3ALT
B
Spearman correlation= 0.27, P=6.25 x 10-20
CDKPredict risk score
Fraction Genome Altered
Density
Fraction Genome Altered
C
Association with HRD signature
-log10(q-value)
log2(Odds ratio)
BRCA2ALT
PIK3CAMUT
FGFR1AMP
CDH1ALT
TP53ALT
BRCA1ALT
9p21DEL
MYCAMP
D
Association with APOBEC signature
-log10(q value)
log2(Odds ratio)
PIK3CAMUT
CDH1ALT
GATA3ALT
MYCAMP
NF1ALT
NCOR1ALT
E
Altered
Not altered
TP53
TMB-H
9p21DEL
ERBB2MUT
11q13AMP
GATA3
non-Luminal
Luminal
TP53
MYC
9p21DEL
non-Basal
Basal
TMB-H
ERBB2MUT
TP53
11q13AMP
non-Her2
Her2
F
Low risk
Intermediate
High risk
Pos.cor
Neutral
Negative.cor
ZFHX3
ANKRD11
CDH1
AURKA
FAT1
CDKN2B
CDKN2A
PAK1
FGFR1
PIK3CA
CBFB
MYC
PALB2
BRCA1
BRCA2
MDM2
PPM1D
KMT2C
KRAS
PTEN
SPEN
ARID1A
ERBB2
NF1
FOXA1
TBX3
MEN1
AKT1
ESR1
GATA3
CCND1
CDKN1B
MCL1
RUNX1
TP53
MAP2K4
NCOR1
MAP3K1
PIK3R1
RB1
SF3B1
SMAD4
KMT2A
FOXP1
MDM4
G
Hedgehog signaling
Estrogen response early
Wnt beta catenin signaling
Estrogen response late
Kras signaling dn
Uv response dn
Apical surface
Myogenesis
Epithelial mesenchymal transition
Angiogenesis
Pancreas beta cells
Coagulation
G2m checkpoint
E2f targets
Myc targets v1
Mtorc1 signaling
Oxidative phosphorylation
Unfolded protein response
Cholesterol homeostasis
Reactive oxygen species pathway
Xenobiotic metabolism
Interferon alpha response
Interferon gamma response
Protein secretion
Uv response up
Bile acid metabolism
Peroxisome
Fatty acid metabolism
Myc targets v2
Adipogenesis
Glycolysis
Complement
Androgen response
Pi3k akt mtor signaling
Dna repair
Mitotic spindle
Allograft rejection
Il6 jak stat3 signaling
Inflammatory response
Heme metabolism
Tnfa signaling via nfkb
Hypoxia
P53 pathway
Tgf beta signaling
Apical junction
Notch signaling
Spermatogenesis
Kras signaling up
Apoptosis
Il2 stat5 signaling
Low
Intermediate
High
H
Enriched in
Intermediate-risk/estrogen-high
Low-risk/estrogen-high
-log10(q value)
log2(Odds ratio)

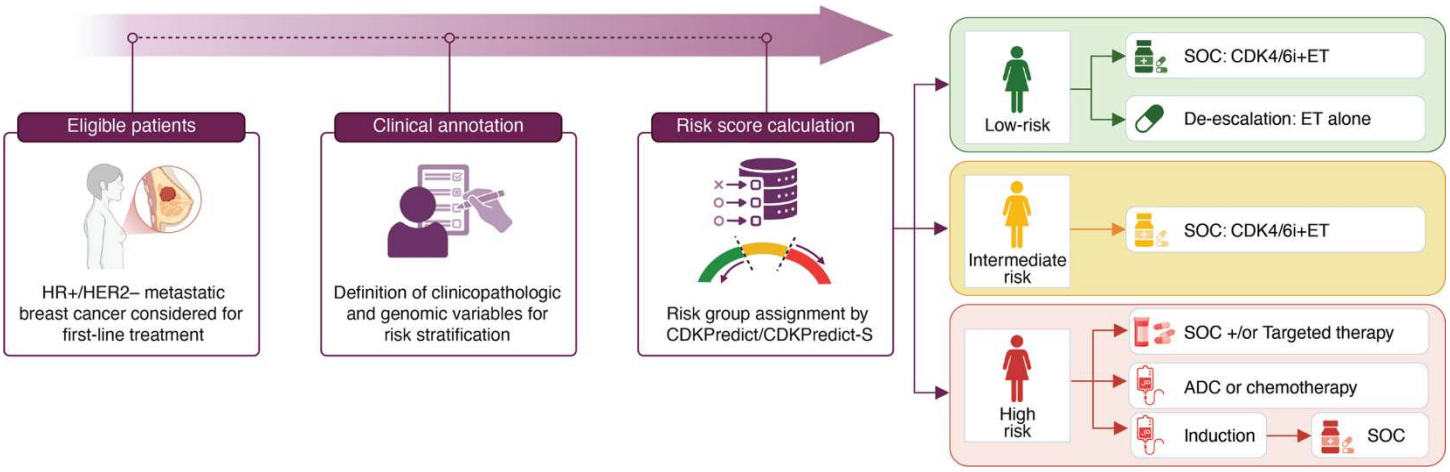
Figure S9. Plausibility of CDKPredict. (A) Distribution of Cox-coefficients across the n=100 base learners included in the ensemble clinicogenomic model. **(B)** Correlation between CDKPredict risk-scores and fraction genome altered (FGA; *left*), showing a positive correlation resulting in a distinct distribution of FGA across risk groups, with a bimodal distribution observed prominently among intermediate- and high-risk tumors (*right*). **(C)** Gene alteration enrichment analysis in HRD-dominant compared with non-HRD dominant tumors, calculated using the Fisher's exact test and significance tested using the Benjamini-Hochberg procedure. **(D)** Same, analyzing gene alterations enrichment in APOBEC-dominant compared with non-APOBEC dominant tumors. **(E)** Co-segregation of genomic features included in CDKPredict and transcriptomic PAM50 intrinsic subtypes, representing genes showing an overall enrichment or mutual exclusivity in the relative subtype as calculated by a Fisher's exact test ($P<0.05$), with genomic variables in bold showing genes with a statistically significant association once corrected for family wise error ($q<0.05$ using the Benjamini-Hochberg procedure). **(F)** Heatmap showing the gene expression profile of features included in CDKPredict across risk-strata and categorized according to the Spearman's correlation (not corrected for multiple testing, $P<0.05$) in genes showing a positive, non-significant, and negative correlation with clinicogenomic risk-scores. **(G)** Representation of pathway-level activity across risk-strata, calculated as the mean expression for each group utilizing the Hallmark gene set collection; pathways represented in bold denote those with significantly different expression between low- and high-risk tumors, calculated using the Mann-Whitney U test and significance using the Benjamini-Hochberg procedure ($q<0.05$). **(H)** Gene enrichment analysis in CDKPredict intermediate-risk and low-risk tumors showing a high Hallmark estrogen response early activity compared with the non-estrogen response early high/CDKPredict risk counterparts.



937 **Figure S10. Benchmarking CDKPredict to an augmented model including higher-order**
938 **genomic features. (A)** Model development schema; in detail, besides features included in
939 CDKPredict, the augmented model included FACETS signatures (fraction genome altered and
940 whole genome doubling events), tumor mutational signatures (dominant status), and
941 hemizygosity in tumor suppressor genes showing a significant prognostic association with PFS
942 on CDK4/6i+ET (**Methods**). Thereafter, the model was fitted on the training set using the same
943 CDKPredict architecture and tested in the independent MSK-validation set. **(B)** Time dependent
944 AUC using a 5-year timeframe between CDKPredict and the comprehensive model, with dots
945 showing the mean AUC across n=1,000 bootstrap iterations in the validation cohort obtained
946 using the same fitted model, using the same reproducible artificial splits between the two models
947 (random seed). **(C)** Same as before using the Harrell's C-index; altogether, the two models
948 demonstrated similar performances, emphasizing the adoption of CDKPredict at the advantage
949 of higher generalizability. **(D)** Representation of feature coefficients across the n=100 base
950 learners included in the comprehensive model.



951 **Figure S11. Deployment and assessment of the clinical and clinicogenomic version of**
952 **CDKPredict-S**. **(A)** Probability density function of risk-scores across risk-strata of the clinical
953 CDKPredict-S, which was refitted on the overall MSK-cohort before model deployment. Arrows
954 and the respective values represent the risk-score cutoff between groups, calculated using the
955 Brent's method. **(B)** Same for the clinicogenomic CDKPredict-S. **(C)** Association between
956 continuous risk-scores of the clinical CDKPredict-S and PFS on first-line CDK4/6i+ET in the
957 overall independent external cohort with available clinical data (n=1,019), showing a clear
958 stepwise association between increasing risk-scores and shorter time to progression. **(D)** Same,
959 using the clinicogenomic model in the external validation cohort of patients with available baseline
960 tumor genomic sequencing (n=660).



961 **Figure S12. CDKPredict as a triage for risk-adapted therapeutic intervention in HR+/HER2–**
962 **MBC.** Following clinicogenomic data annotation and risk score calculation, patients assigned to
963 each risk-strata could be evaluated in clinical trials for de-escalated or escalated treatment
964 approaches. Low-risk patients could be evaluated for de-intensified approaches with ET alone,
965 particularly in resource-limited settings; conversely, CDKPredict identifies high-risk patients
966 showing poor outcomes on standard first-line CDK4/6i+ET, for which escalated trials including
967 targeted therapy +/- CDK4/6i+ET, first-line ADC, or induction with ADC or with chemotherapy-
968 based regimens followed by CDK4/6i+ET maintenance could be prioritized.
969

ONLINE METHODS

Study cohort and prospective sequencing

The Memorial Sloan Kettering (MSK) Breast Translational cohort encompassed 1,264 individuals with HR+/HER2– breast cancer who received routine care at MSK, with prospectively collected and manually curated clinical data. Inclusion criteria consisted of the receipt of first-line CDK4/6i+ET, together with baseline tumor genomic sequencing, defined as sequencing performed on a tumor sample collected either before (89.8% of total samples) or from treatment start up to 28 days (129 of 1,264 of patient-specific samples, 10.2%); the 28-day timeframe from first-line treatment initiation was deliberately selected given the usual 2–3-week gap between prescription and first dose in US practice, resulting in many samples ultimately being collected after treatment start; moreover, unlike transcriptional profiles, tumor DNA profiles are not expected to change within days of starting therapy, therefore reflecting an accurate baseline genomic profile; for the same reason and considering the first-line setting, we assumed this selection criterion would not result in potential immortal time bias. An additional criterion for study eligibility was the receipt of first-line CDK4/6i+ET for at least 28 days. Patients who received investigational agents together with CDK4/6i+ET or induction chemotherapy before CDK4/6i+ET were excluded from the study population. For patients with multiple samples satisfying the temporal sequencing criteria, a manual review of the cases was conducted, prioritizing samples collected at a more recent time relative to the start of CDK4/6i+ET, utilizing the tumor cellularity and sequencing depth as a second criterion for sample selection. The present study was approved by the Memorial Sloan Kettering Cancer Center (MSK) Institutional Review Board (IRB) and all patients provided written informed consent for tumor and paired normal DNA sequencing and review of medical records for clinical annotations.

The external, nationwide multi-institutional cohort encompassed 564 individuals affected by HR+/HER2– breast cancer, who received first-line CDK4/6i+ET in community or academic settings, with available patient-level clinicogenomic data resulting from referral to a tertiary-level sequencing facility for baseline tumor genomic profiling on samples collected before and up to 28 days from starting first-line CDK4/6i+ET. The same exclusion criteria, including a treatment duration below 28 days, investigational agents or induction chemotherapies, among others, were utilized for the study eligibility in the external multi-institutional cohort.

For validating a scalable model for variable-resource settings, in addition to the abovementioned external cohort, we leveraged a cohort of 455 patients, who received routine care in a single-

center institution in Italy (European Institute of Oncology, *IEO-cohort*). Inclusion criteria included HR+/HER2– MBC and receipt of first-line CDK4/6i+ET for at least 28 days. Patients with or without baseline genomic sequencing (performed before or within 3 months from start of first-line CDK4/6i+ET) were included in the study cohort. The temporal criterion for tumor sequencing was expanded relative to the MSK-cohort to facilitate the collection of a larger external independent validation cohort, hypothesizing that secondary resistance mechanisms would be less likely to emerge only after a limited duration of treatment with CDK4/6i+ET.

Genomic sequencing

For patients included in the MSK-cohort, samples were sequenced in a CLIA-certified laboratory using the FDA-authorized, tumor-normal MSK-IMPACT sequencing panel, as described previously[36]. Patients provided consent for tumor profiling under an institutional IRB-approved research prospective protocol, 'Tumor Genomic Profiling in Patients Evaluated for Targeted Cancer Therapy' (NCT01775072). Tumor type and purity were assessed by a dedicated pathologist from hematoxylin and eosin-stained slides of tumor samples. In detail, DNA was extracted from formalin-fixed, paraffin-embedded tumor and matched normal (peripheral blood) samples. Targeted sequencing was conducted with custom DNA probes against exonic and intronic regions according to the MSK-IMPACT version, encompassing 341 genes (n= 4 samples, 0.3%), 410 genes (n= 127, 10%), 468 genes (n= 666, 52.7%) and 505 genes (n= 467, 36.9%). Tumor samples with no detectable somatic alterations likely due to low tumor content or low sequencing depth (<100x) were excluded (n=30). Tumor mutational burden (TMB) was calculated as the number of non-synonymous mutations per megabase sequenced (mutations per megabase, Mut/Mb). TMB high (TMB-H) was defined as ≥10 Mut/Mb, consistent with the FDA-approved indication for the agnostic use of immune-checkpoint inhibitors, as previously defined[68].

Somatic genomic alterations were filtered for 'oncogenic' or 'likely oncogenic' alterations using OncoKB, a precision oncology knowledge-base that tracks the effects of cancer variants and their potential clinical actionability (http://oncokb.org). Genomic alterations were classified as actionable using a level of evidence from 1 to 4, where Level 1-2 alterations indicate a standard therapeutic intervention and levels 3A-4 include investigational therapeutic alterations, which may direct a patient toward a clinical trial relevant to that biomarker. Germline alterations were included only for *BRCA*1/2 and *PALB2*, in line with their previous relevant prognostic role for patients receiving CDK4/6i+ET.

Allele specific copy number, tumor purity and ploidy were inferred using FACETS (v0.5.14)[79]. Copy number quality call was evaluated using a set of consecutive FACETS quality control metrics (https://github.com/taylor-lab/facets-preview/). A total of 1,085 tumors (85.8%) passed these filters and were included in the analyses of fraction genome altered (FGA) and whole-genome doubling events (WGD). FGA was defined as the proportion of the profiled genome with absolute $\log_2$ copy number changes greater than 0.2. WGD was inferred as previously described[80], with tumors classified as WGD-positive when more than 50% of the autosomal genome exhibited a major copy number equal or greater than 2.

Genomic signature analysis was conducted using Signature multivariate analysis (SigMA; https://github.com/parklab/SigMA), whose performance for imputing HRD and APOBEC leveraging the MSK-IMPACT sequencing assay has been previously validated[32,81]. As previously defined, only samples with at least n=5 SNVs were considered for mutational signature prediction (n=946, 74.8%). To increase the sensitivity of the analysis and to gauge a larger spectrum of genome-wide mutagenic processes extending beyond covered regions, both silent mutations and SNVs originating from off-target reads were included, using as a quality filter a sequencing depth of ≥100 reads in the involved genomic region, together with at least 5% reads supporting the alternate variant (variant allele frequency of 5%).

In the external multi-institutional cohort, tumor sequencing was performed using a commercially available genomic sequencing panel interrogating up to 648 cancer-related genes, as previously validated[51]. In detail, genomic profiling was conducted using the solid tumor Tempus platforms xT and the legacy xO assay, while plasma-based profiling was conducted using the liquid biopsy xF platform[53]. Among 564 patients, 417 (73.9%) underwent tissue-based testing and 147 (26.1%) underwent plasma-based testing. The specific assay distribution was xT (n=6), xT.v2 (n=62), xT.v3 (n=81), xT.v4 (n=267), xF.v2 (n=146), xF.v3 (n=1), and xO (n=1).

For patients included in the IEO-cohort, baseline genomic sequencing was available for 21.1% cases (96 of 455). In detail, tumor genomic sequencing was conducted using an analytically validated and commercially available hybrid capture, tumor-only solid-tissue sequencing panel interrogating 523 cancer-related genes[74] for n=58 cases (60.4%). For the remaining 38 cases (39.6%), tumor sequencing was conducted using an analytically validated and commercially available cfDNA-based tumor-only sequencing assay targeting 50 cancer-related genes[75].

Clinical data annotation for the MSK-cohort

Clinical data were gathered from the MSK Breast Translational Program's clinical database, which contains comprehensive, longitudinal information on patient demographics, pathology, treatment regimens, and clinical outcomes. Structured data were systematically and manually curated using a standardized clinical data annotation process by a dedicated team of clinical data annotators and were maintained under rigorous quality-control procedures by experienced data managers. The data lock for the analysis was set on October 29, 2025.

Progression-free survival was defined as i) a clinical or radiographic disease progression prompting change in systemic therapy or recommendation for ablative local therapy directed at site(s), or ii) any cause of death. Patients lost to follow-up, or who were still receiving first-line CDK4/6i+ET at last follow-up, or who discontinued treatment without evidence of either clinical or radiological progression (e.g., toxicity, patient preference or insurance coverage) were right censored. When endocrine therapy or CDK4/6i changed to another compound (endocrine compounds or CDK4/6i, respectively) for reasons other than disease progression (e.g., toxicity, patient preference or insurance coverage), the time on successive regimens was combined to more accurately capture the real-world PFS. When patients discontinued CDK4/6i after 28 days due to toxicity while continuing the endocrine agent alone, the overall PFS achieved before progression or death on the same regimen or up to the right censoring time were combined.

Train and test stratification in the MSK-cohort

For developing the model, the MSK-cohort was divided into a training set (70%, n=884) and an internal validation set (30%, n=380). The train-test split was conducted in a stratified manner, balancing the rate of i) events, ii) liver metastasis, iii) treatment-free interval ≤12 months, and iv) the availability of FACETS (passing quality control criteria, as abovementioned) between the train and test set. The abovementioned factors were selected to ensure a similar median follow-up (median follow-up by reverse Kaplan-Meier method in the train set of 49.7 months, IQR: 26.8–79.9; in test set of 50.6 months, IQR: 24.9–74.3), a balanced representation of relevant prognostic clinical features for preliminary model validation, and to achieve a balanced representation of FACETS-level signatures for assessing the clinicogenomic model augmented with higher-order genomic features.

For evaluating the balanced representation of clinical and genomic variables between the train and test set, univariable tests were conducted using the Fisher's exact test (categorical variables).

Furthermore, for evaluating the genomic representation of features between the train and test cohort, we leveraged an adversarial random forest classifier encompassing n=100 trees, which was trained to distinguish instances belonging to the training set or test set using genomic features present in at least 1.5% of the overall MSK-cohort. The evaluation was performed on hold-out samples using a 10-fold cross-validation. The model discriminative performance was measured with the area under the receiver operating characteristic curve (AUC), with a score equal to 0.5 representing a features distribution not reflecting an imbalanced split of genomic variables in a higher-order dimension between the train and the internal validation set. In addition, for evaluating the multi-dimensional embedding of genomic features, a principal component analysis (PCA) was conducted (extracting the first two principal components). To quantify the contribution of original features to the first two principal components in the representative illustration, PCA loadings were extracted from eigenvectors accessed using the pca.components_ attribute in scikit-learn.

Features engineering

Clinical and genomic data, collected as abovementioned, were utilized for developing a machine learning framework designed to predict the risk of early progression from first-line CDK4/6i+ET.

Specifically, for clinical variables, features selection was conducted using knowledge-based criteria that included clinical variables relevant to outcomes with CDK4/6i+ET. Continuous variables were engineered to result in a categorical representation to maximize the clinical interpretability of the model, broad applicability and to mitigate differences originating from inter-observer or pre-analytical factors (e.g., percentage of nuclei showing estrogen and progesterone receptors staining). The final set of features included in the model comprised: presence or absence of metastasis involving the liver (i) or central nervous system (ii) at the time of CDK4/6i+ET initiation; iii) presence of bone-only disease; iv) treatment-free interval equal to or below 12 months, defined as tumor recurrence occurring during endocrine therapy or within 12 months from the end of adjuvant therapy for either toxicity or treatment plan termination; v) high-grade tumor as per Nottingham Histologic Score[82]; for patients with multiple samples or without available metastatic samples, those collected closer to the start of first-line CDK4/6i+ET and samples from metastatic sites were prioritized; vi) de novo metastatic disease, defined as a newly diagnosed metastatic disease or with locoregional extension considered not amenable for curative-intent locoregional treatments; vii) age at metastatic diagnosis ≤50 years; viii) use of fulvestrant as an endocrine backbone; ix) negative progesterone receptor, defined as 0% of tumor cell nuclei showing receptor staining; for patients with multiple samples, priority was given to the

sample collected at a more recent time relative to first-line treatment start and to those collected from metastatic sites; x) estrogen receptor percentage equal to or below 10%; and xi) above 10% and equal or below 80%, using the >80% as a reference category to reduce collinearity in the variables; xii) number of distant metastatic organs equal to 2 or 3 or xiii) equal to or above to 4, using as reference the presence of at most one distant metastatic organ or locoregional disease to reduce collinearity; xiv) genomic sequencing conducted on a metastatic sample.

For selecting genomic variables to be included in the model, we considered somatic tumor variants annotated as oncogenic or likely oncogenic by OncoKB[48,49]. For considering differences in the biology of distinct alteration classes, which is particularly relevant in the context of breast cancer, alterations affecting oncogenes were classified as [$Gene_{MUT}$] for genomic mutations including single nucleotide variants and small insertion-deletions (indels), [$Gene_{AMP}$] for amplifications, and [$Gene_{FUS}$] for gene fusions or structural variants. Assuming a common role of pathogenic variants to converge on a loss of function of tumor suppressor genes, oncogenic and likely oncogenic alterations belonging to distinct mutational classes were considered together for tumor suppressor genes as [$Gene_{ALT}$]. Tumor suppressor genes were defined according to OncoKB (https://www.oncokb.org/cancer-genes). Amplifications including *CCND1*, *FGF4*, *FGF3*, and *FGF19* were merged into 11q13 amplicon considering their co-localization within the same genomic amplicon and to avoid unstable model estimates due to collinearity. In the same manner, deletions involving *CDKN2A* and *CDKN2B*, whose co-deletion represents a recurrent genomic event in breast cancer, were classified together as amplicon $9p21_{DEL}$. For selecting genomic features to include in the model, we employed a frequency-based features selection to include genomic alterations present in at least 1.5% of samples to avoid possible statistical noise derived from rare features, resulting in n=42 genomic variables; moreover, we reasoned that a crude frequency-based selection could mitigate i) the risk of a data-driven selection bias if conducted in the overall cohort before model evaluation in the hold-out MSK validation cohort, and ii) statistical noise expected if conducted only on the training set (n=884) given the smaller sample size and the sparsity of genomic biomarkers. Besides them, based on prior biological knowledge, we included *BRCA1* (n=16), *PALB2* (n=11), *FAT1* (n=16), and *PIK3R1* (n=16) alterations, whose prognostic importance has been widely established in breast cancer by our group among others[13,16,58,59]; importantly, this biology-driven selection is based on prior validated studies with established evidence across multiple cohorts, arguing against a potential circularity bias; moreover, we reasoned a circularity bias in features selection would have resulted in model overfitting to the training set and poor external performances in the available external validation

cohorts. To account for previously reported technical differences across sequencing platforms[83], rather than being normalized and engineered as a continuous variable, the TMB score was one-hot encoded considering the absolute number of mutations, including i) 6-9 Mut/Mb (n=153, 12.1%) and ii) ≥10 Mut/Mb (n=95, 7.5%), using the TMB: 0-5 Mut/Mb as the reference (n=1,016, 80.4%) to reduce collinearity. Collectively, together with the two TMB categories, 48 genomic features were included in the genomic and clinicogenomic model.

For developing the augmented model leveraging higher-order genomic variables to benchmark with CDKPredict, besides the abovementioned variables, we included tumor mutational signatures, FGA, and hemizygosity in tumor suppressor genes, the latter defined from the allele-specific copy number analysis (using FACETS) as a total copy number of n=1 together with a lesser copy number of n=0 involving the specific gene. Specifically, tumor mutational signatures derived with SigMA were categorized to represent APOBEC- and HRD-dominant groups (by one-hot encoding), using other signatures as the reference level. For categorizing FGA, we utilized the Youden's J statistic to define the optimal cutoff maximizing the specificity and sensitivity to detect patients experiencing a PFS event within 12 months, while excluding samples that were right censored before 12 months. The 12-month timeframe was selected to emphasize the prediction discrimination in a time-frame useful for employing baseline, risk-informed decision making in patients with breast cancer considered for first-line therapy. The FGA cutoff was determined at a 0.35 threshold, as calculated in the training set and applied for classifying samples in the internal validation set. For selecting genes showing hemizygosity, we performed a univariable Cox model for each tumor suppressor gene correcting the analysis for fraction of the genome carrying loss of heterozygosity events (fraction loss of heterozygosity), as determined with FACETS, and WGD events, with the correction conducted to isolate the prognostic role of hemizygosity events from co-occurring passenger events in the context of large scale genomic instability. Ultimately, tumor suppressor genes showing a false discovery rate <0.1 (calculated with the Benjamini-Hochberg procedure) and with a prior demonstrated role to promote endocrine resistance were selected, ultimately including hemizygosity events in *RB1*, *BRCA1*, *BRCA2*, *TP53*, *LATS2*, *NF1* and *PTEN*.

Rationale for developing ReCAST

For developing ReCAST, we developed a supervised machine learning framework to predict time-to-event outcomes from high-dimensional, sparse clinicogenomic data, available at https://github.com/mskcc/ReCAST.

ReCAST was specifically designed to address a key limitation of conventional feature selection and modeling when applied to sparse clinicogenomic datasets, namely the tendency to discard or under-weight infrequent yet meaningful predictor variables. This is particularly relevant for structured genomic data, which are characteristically sparse: an individual genomic alteration is typically present in a minority of cases, such that biologically informative events may exhibit a low prevalence in any given cohort. The consequences of this limitation are especially pronounced in cancer, where rare alterations conferring resistance to therapy carry large effects in the few patients who harbor them. In this setting, penalized Cox models may shrink the coefficients of rare features toward zero unless their contribution to the objective function is large enough to outweigh the regularization term. This condition is rarely met by low-prevalence variables, particularly when more prevalent covariates already account for much of the gain in the objective function at the population level:

$$\hat{\beta} = \arg\max_{\beta} \left(\ell(\beta) - \lambda \sum_{k=1}^{p} |\beta_k| \right)$$

where $\hat{\beta}$ represents the calculated vector of model coefficients, $\ell(\beta)$ denotes the Cox partial log likelihood to be maximized by the features vector, λ the regularization parameter of the Cox model, and $\sum_{k=1}^{p} |\beta_k|$ the L1 penalty for Least Absolute Shrinkage and Selection Operator (Lasso) models. Altogether, this might limit the ability of the prediction framework to capture low frequency but prognostically relevant variables, which have been extensively demonstrated to have a critical role in determining resistance to therapy.

In addition, to facilitate the potential clinical implementation of the model in routine care, we aimed to establish a model with clear interpretability, leveraging therefore a Cox-model framework.

ReCAST framework

For developing ReCAST, we developed an ensemble machine learning framework encompassing n=100 penalized Cox-models as base learners, building on a previously developed framework at MSK[84,85], to result in a single-patient risk-score prediction. In detail, each base learner consists of an L1-penalized Cox proportional hazard model, wherein each single model was fitted on a distinct subset of the training set defined with a resampling with replacement strategy (bootstrap). Such approach was deliberately employed to increase the heterogeneity in the features representation across base learners, subjecting each single model to a distinct patient-level and

features-level representation of the overall training set. Similarly, the L1-penalty term was determined to emphasize the relative importance of potentially rare yet meaningful features by forcing each base learner to maximize the objective function leveraging a distinct subset of features.

In addition, cognizant of the intrinsic tendency of the L1-penalty to over-shrink the relative importance of infrequent yet potentially important variables, we leveraged a previously developed framework[42] to employ a feature-specific penalty term across each distinct base learner (i.e., adaptive lasso). In detail, each of the n=100 base learners was fitted using feature-specific penalty factors estimated from a prior L2-penalized Cox model fitted on the specific bootstrap subcohort of each base learner. Herein, the regularization strength of the L2-penalized model was selected using the best 3-fold cross validation score, considering a hyperparameter space included in the log-uniform scale between -2 and 1. The score was defined as the mean time-dependent AUC utilizing n=10 time bins from 3 to 36 months. Such time frame was selected to optimize the model for discriminative accuracy in a clinically meaningful timeframe for potential clinical decision-making considered for the particular therapeutic setting. Each feature-specific penalty term for the subsequent adaptive lasso optimization was determined as the inverse of the absolute value of the L2-penalized coefficients. Collectively, variables with larger initial coefficients, therefore more important for increasing the partial log-likelihood of the L2-penalized Cox model, received smaller penalties in the subsequent adaptive lasso step, whereas variables with negligible effects were penalized more strongly.

Finally, each L1-penalized base learner was fitted utilizing variable-specific penalty weights, calculated as above, on the respective bootstrap-resampled cohort. In detail, for finding the optimal regularization strength of the L1 penalty term, the initial set of candidate penalty terms was defined by fitting the L1-penalized model to the bootstrap-resampled cohort, gathering a set of n=100 candidate penalty values ranging from a maximum regularization strength setting all coefficients to 0, to a minimal penalty performing no feature selection (i.e., leaving all coefficients non-zero). Subsequently, the optimal regularization strength was selected according to the best time-dependent AUC score, with the same definition as above and calculated within a 3-fold cross-validation framework.

Calculation of risk-scores

During each base learner training, a risk-score was computed on out-of-bag samples. Importantly, such strategy was employed to gather an objective and reliable assessment of the model performances on the overall cohort employed for training the ensemble model, which represents a critical prerequisite for early model development. Indeed, such strategy allowed us to mitigate the risk of data leakage and overconfident predictions of the model resulting from risk-scores computed on samples utilized for model training, enabling a reliable estimation and calibration of the model during its early development stage. Moreover, we assumed an objective estimation of train scores would further enable the definition of reliable cutoffs which could potentially better generalize once considering a single-patient prediction of independent samples. For each patient, an individualized continuous score was defined from the mean of the predicted risk-scores across base learners. Finally, risk-scores were min-max normalized between 0 and 100 using the 1st and 99th percentiles of raw scores to avoid distributional shifts from relevant outliers.

For predictions on samples not utilized during model fit, risk-scores were defined by leveraging the full set of n=100 base learners, with each computing the continuous scores derived from the vector multiplication encompassing the sample-specific features and the base-learner specific coefficients.

Definition of discrete risk groups

Following the definition of patient-specific risk-scores, we next aimed to gather a risk-class definition to facilitate the clinical implementation of model predictions for informing potential decision making. For this scope, ReCAST employs a Gaussian mixture model (GMM) using the training risk-scores measured as abovementioned to result in a data-driven classification. This strategy was deliberately employed to avoid predefined quantile-based thresholding, which we assumed would not reflect a reliable and data-driven objective stratification of risk classes. Specifically, patient-specific training risk-scores were modelled using pre-defined Gaussian components (n=2 for the clinical and genomic only model, n=3 for the clinicogenomic model, resulting from the concordance analysis as defined in the study results) to fit a GMM. After fitting, the component specific means, variances and weights (i.e., proportion of samples included in each component) were extracted. Finally, risk-group thresholds were defined by calculating the root (i.e., value equal to 0) of the difference between the adjacent components' probability density functions, corresponding to the point at which their densities were equal and thus where components intersected, which was calculated using Brent's method as implemented in scipy.optimize.brentq.

Once estimated in the training cohort, the cutoffs were locked and subsequently applied unchanged for risk-group assignment both in the training set and in the validation sets. Importantly, this approach enabled the definition of risk groups in a data-driven manner, without leveraging arbitrary and possibly not biologically reasonable quantile-based cutoffs.

Machine learning architectures utilized for benchmarking

For benchmarking ReCAST to current state-of-the-art machine learning architectures, we investigated a series of machine-learning models accommodating time-to-event data[86].

First, we aimed to benchmark ReCAST to ensemble trees, which have long been considered gold-standard models for predictions using tabular data[86]. We evaluated a random forest designed for time-to-event data (as implemented in https://github.com/sebp/scikit-survival), and an eXtreme Gradient Boosting (XGBoost) model using a Cox survival objective (utilizing the option objective="survival:Cox" from https://github.com/dmlc/xgboost) for managing right-censored data. Moreover, we evaluated linear models which have represented the most commonly adopted frameworks for cancer prognostication in clinical oncology[1-3,6,87]. In particular, we evaluated a Cox model employing either an Elastic Net penalization (mixing of L1 and L2 regularization in the Cox partial log likelihood optimization) or a Lasso Cox model (using an L1 regularization). Finally, we evaluated a previously developed ensemble model at MSK (i.e., OncoCast), which was previously validated to predict time-to-event outcomes in patients with solid tumors[84,85]. Briefly, OncoCast represents an ensemble model conducting an average risk-estimates from base learners, potentially consisting of distinct ML architectures; for the scope of comparison in the current work, we evaluated OncoCast with the default set of n=100 models, with each consisting of an L1-penalized Cox regression. Each model in the OncoCast implementation was trained using the default 70% set of the training set repartition, wherein the Cox partial log likelihood was optimized using the C-index metric. Further details about OncoCast can be retrieved from the original publication[85].

State-of-the-art transformers architectures for prognostication using tabular data

We further evaluated several state-of-the-art transformer architectures specifically designed for predictions in the setting of tabular data. First, we employed a previously developed transformer model designed for time-to-event prediction (Clinical Transformer)[44]. Briefly, the Clinical Transformer includes an input embedding layer designed to project input features to an

embedding space starting from feature-name and feature-value pairs. Next, it possesses a transformer encoder, which computes self-dot product attention using a multilayer bidirectional transformer encoder similar to BERT[88], comprising multi-head self-attention and feed-forward layers. A special classification [CLS] token is included in each patient representation, and its final representation following the attention mechanism is used as an aggregate patient embedding. For survival prediction, this embedding is passed to a final linear output layer generating a continuous risk-score. For the Clinical Transformer, we evaluated both direct- and gradual-learning fine-tuning strategies, the latter of which has been previously shown to improve clinical outcomes prediction[44]. In direct learning, the model is initialized without pretraining and optimized directly against the survival endpoint. In gradual learning, the same development dataset is utilized for a first step of self-supervised masked-feature pretraining. Using the training partition, 20% of feature names were randomly replaced by a [MASK] token, while their associated feature value was retained to mirror the original pretraining scheme[44]; the model was fine-tuned to reconstruct the masked feature names utilizing a cross-entropy loss and their corresponding values using a mean-squared-error loss. The resulting pretrained parameters were subsequently used to initialize the transformer for supervised fine-tuning using the survival endpoint.

We next evaluated recently developed transformer foundation models specifically designed for prediction using tabular data[45,46]. Unlike conventional models, such transformers are pretrained using collections of millions of synthetic data to learn statistical priors, followed by predictions performed on unseen tabular datasets through in-context learning, where no modification of trained parameters occurs during prediction (zero-shot predictions). Specifically, during in-context learning, labelled observations (training partition) and unlabeled queries (test partition) are provided jointly as context, allowing the transformer to predict the dependent variable using variable-wise and sample-wise attention.

In this context, we applied the pretrained TabPFN version 2.5 (as implemented in the original publication[45]) and version 3[89]. We further employed TabFM (version 1.0), both in its default zero-shot prediction implementation and its official ensemble configuration, as specified in https://github.com/google-research/tabfm.

Model benchmark optimization

For comparing the distinct models, we used a stratified 5-fold cross validation (CV) framework limited to the training set, with the splitting strategy designed to balance the number of PFS events

across folds. In this context, fine-tuning was conducted on the 4-fold training partition with predictions computed on the validation fold to provide an objective measurement of model predictions to unseen data, mitigating the risk of overconfident estimates. Moreover, for ensemble trees (survival random forest and XGBoost) and non-ensemble linear models (Elastic Net and Lasso regularized Cox models) a nested cross validation was employed, in which across each iteration and 4-fold training partitions, hyperparameter tuning was conducted using a first-step of stratified 3-fold CV (using PFS event as the stratification criterion) to derive the best set of hyperparameters optimizing the C-index.

In detail, for Elastic Net penalized Cox model, a grid search evaluated an L1 to L2 ratio ranging from 0.1 to 0.9 using 0.1 increments; moreover, the grid search reiterated across regularization strengths (alpha) including 0.001, 0.005, 0.01 and 0.05. For the Lasso penalized Cox model, the mixing parameter was fixed at 1.0, corresponding to an L1 regularization, and the same four penalty strengths were evaluated. The maximum number of iterations was set to 100,000 for both models.

For the random survival forest, ten hyperparameter configurations were randomly sampled (using randomized search) within each outer training fold. The search space included 100 or 500 trees; maximum tree depths of 3, 6 or 9; minimum numbers of observations required to split an internal node of 2, 5 or 10; minimum terminal-node sizes of 5, 10 or 20; and the number of variables considered at each split, defined as either the square root of the total number of features or 50% of the available features.

XGBoost was implemented using a Cox proportional-hazards loss to accommodate right-censored observations. As with random forest, ten configurations were sampled by randomized search within each inner training partition in the nested CV framework. The evaluated search space comprised learning rates of 0.01, 0.05, 0.10 or 0.20; maximum tree depths of 3, 6 or 9; row-subsampling proportions of 0.7 or 1.0; feature-subsampling proportions of 0.6 or 1.0; L1-regularization strengths of 0 or 0.1; L2-regularization strengths of 0.1 or 1.0; minimum child weights of 1 or 5; maximum delta steps of 1 or 5; and 100, 500 or 1,000 boosting iterations, with the higher ceiling of the latter increased compared to the random forest considering the intrinsic gradient boosting architecture by its base learners.

For training the Clinical Transformer, the feature-value pairs were processed by an encoder comprising two transformer layers, each containing multi-head self-attention. Our implementation used a hidden dimension of 64, four attention heads, an intermediate feed-forward dimension of 128 and dropout probabilities of 0.10 for both the hidden representations and attention probabilities. Such network dimension was deliberately selected to restrict the number of parameters considering the limited sample size of our training set for a transformer architecture. As abovementioned, a [CLS] token was added to each patient representation to result in an aggregated patient embedding following attention, which was finally passed to a linear output layer generating a continuous risk-score. As abovementioned, we conducted both direct and gradual learning. For direct learning, optimization employed an AdamW with a learning rate and weight decay of 1×10^{-4}; the objective function consisted of a differentiable approximation of the C-index, as implemented in the original publication[44]. In gradual learning, we performed 20,000 optimization steps using a batch size of 128, an AdamW optimization with a learning rate of 1×10^{-3} and weight decay of 1×10^{-4}. The resulting encoder parameters were subsequently used to reinitialize the survival model. In each 5-fold iteration, for deriving the optimal number of optimization epochs, the 4-fold training partition was divided into an 80:20 training–validation split (stratified for PFS event), where a model fitted on the inner training partition was trained for a maximum of 300 epochs using as an early stopping criterion an improvement no greater than 10^{-4} in the C-index metric for 40 consecutive epochs. The derived best epoch was then selected for training the clinical transformer on the complete training folds.

For TabPFN and TabFM, which are zero-shot foundation models, no parameter fine-tuning was conducted on the study cohort, as per their intended implementation.

Generation of risk-scores and model comparisons

Using out-of-fold predictions within the 5-fold CV framework, each patient was assigned a model-specific risk-score. To allow easier inter-model comparisons, each predicted risk-score was normalized to a 0-100 score, using a 1st and 99th percentile normalization from risk-scores derived from the 4-fold training set distribution. Considering that tabular zero-shot foundation models have not been specifically designed to compute risk-scores but rather to perform predictions in classification or regression tasks, we adopted a 12-month PFS-probability landmark analysis (excluding from in-context training samples without a PFS event and a follow-up below 12 months), where the probability was afterwards converted to a 0-100 risk-score. The 12-month

timeframe was selected as considered an appropriate time-point to reflect a lack of clinical benefit from first-line CDK4/6i+ET.

For the benchmarking analysis, continuous risk-scores were used to calculate point estimates considering the C-index and AUC[t], with the latter calculated as the integrated mean between 6 months and 60 months using 10 time bins.

Uncertainty in performance estimates was quantified using 1,000 patient-level bootstrap resamples. Within each bootstrap iteration, patients were sampled with replacement and the C-index and the AUC[t] were recalculated for each model. The 2.5^{th} and 97.5^{th} percentiles of the resulting distributions were used to derive bootstrap 95%CI. The metrics obtained in each iteration were afterwards utilized for a contrast score (metric of the model vs. metric of the reference model). To ensure fair comparability, all models utilized the same and reproducible bootstrap resamples to provide paired differences. To generate a descriptive measure of the probability that the comparator model outperformed the reference model under patient resampling, we calculated the proportion of bootstrap iterations in which the difference was greater than zero.

Cross-fitted model-reliance analysis of rare genomic alterations

Cognizant of the existence of rare yet critical genomic alterations promoting clinical resistance to CDK4/6i+ET, we evaluated the capability of the different models to retain sensitivity to rare genomic features for prognostication, which we assumed represents a prerequisite for reliable and plausible single-patient predictions. For this scope, we evaluated rare yet mechanistically and clinically validated drivers of CDK4/6i+ET resistance, namely *BRCA1*, *BRCA2*, *RB1* and *PALB2*. Considering the rare occurrence and to simulate a model later fine-tuned on the overall MSK cohort before external deployment, the development cohort and the previously locked internal-validation cohort were pooled to increase the number of patients carrying the abovementioned genomic alterations. Critically, the pooled cohort was used exclusively for this exploratory model-reliance analysis, assuming such analysis to reflect inherent models' characteristic to retain sensitivity to priori defined rare genomic drivers of CDK4/6i+ET resistance, rather than a context-specific metric determining artificial data-leakage in model selection.

For each alteration, we performed n=50 repetitions of five-fold cross-validation, with outer folds stratified according to alteration-carrier status. Within each repetition, models were fitted using four outer folds with predictions computed on the evaluation fold. For Elastic Net and Lasso

penalized Cox models, random survival forest and XGBoost, the same nested cross validation strategy for hyperparameter tuning was carried out as previously described. Considering the computational limitation of such analysis, the Clinical Transformer was not included in the cross-fitted reliance analysis.

In each outer validation fold, the analysis was restricted to patients carrying the alteration under evaluation. For every carrier, two feature representations were generated: the observed representation, in which the alteration retained its recorded value, and an ablated representation, in which only the corresponding alteration indicator was changed from present to absent, while retaining unchanged all remaining clinical and genomic variables. Both representations were evaluated using the same fitted model. Risk-scores were generated using the same fold-specific scaling procedures as abovementioned to ensure model comparability (score from 0 to 100, normalized using the 1st and 99th percentiles from the 4-fold training partition). For each carrier, the ablated risk-score was calculated as the difference between the risk-score with and without the alteration of interest, hypothesizing such feature ablation analysis could provide an objective and measurable metric of feature contribution for both linear and non-linear models. To avoid treating repeated predictions from the same patient as independent observations, the n=50 out-of-fold differences were first summarized by their median for each patient, alteration and model.

Development of CDKPredict and validation strategy

CDKPredict represents the implementation of ReCAST for classifying patients with HR+/HER2– breast cancer at high risk of early progression to first-line CDK4/6i+ET. In the preliminary development of CDKPredict, the ensemble model was trained using the abovementioned framework on the MSK-cohort training set, encompassing 70% (n=884) of the MSK-cohort cases. For the internal validation of CDKPredict, risk-scores and the corresponding risk classes were defined on independent samples encompassed in the internal validation cohort (30% of the development cohort, n=380).

For preliminarily assessing the specificity of CDKPredict for CDK4/6i+ET and endocrine based regimens, we assembled two independent cohorts of patients with HR+/HER2– MBC treated with first-line ET alone (n=402) or first-line chemotherapy (n=265) at MSK, and evaluated model performances against those receiving CDK4/6i+ET included in the internal hold-out validation set using the original model fine-tuned on the training cohort; in addition, we retrieved patient-level data and clinicogenomic annotation from two independent MSK cohorts of patients with triple-

negative MBC (TNBC) treated with first-line chemotherapy (n=547), and patients with HER2+ MBC who received first-line THP (i.e., taxane-based regimen, trastuzumab and pertuzumab, n=315). The same inclusion criteria were applied as in the original CDK4/6i+ET cohort (e.g., sequencing timing). For assessing model specificity, CDKPredict risk score was evaluated in Cox-regression models stratified per treatment arm to accommodate distinct baseline hazards for each treatment arm given the non-proportional hazards (Schoenfeld's residual $P\text{-value}_{\text{CDK4/6i+ET vs. ET alone vs. chemotherapy}}=0.00039$, $P\text{-value}_{\text{CDK4/6i+ET vs. 1L THP vs. TNBC chemotherapy}}=1.6\text{x}10^{-12}$), and using beyond main terms an interaction term between CDKPredict risk scores and treatment regimens. In addition, we evaluated the cumulative dynamic AUC[t] using a 12-month landmark analysis considering i) the relevance of detecting early-progressors for the potential utility of risk-adapted therapeutic interventions; ii) considering the expected lower PFS for patients not on first-line CDK4/6i+ET or ET alone, and particularly for HR+/HER2– MBC candidate for first-line chemotherapy for visceral crisis or other clinicopathological surrogates of endocrine resistance, and in patients with TNBC (first-line ET cohort: 13.2 months [95%CI 10.4–15.6]; first-line chemotherapy in HR+/HER2–: 5.7 months [95%CI 4.8–6.7]; first-line THP: 17.2 months [95%CI 14.3–23.7]; first-line chemotherapy in TNBC: 4.3 months [95%CI 3.8–4.9]). Differences in the AUC[t] at each timepoint (3, 6, 9, 12 months and mean across timepoints) were measured across each bootstrap iteration and significance was measured by calculating the number of iterations in which the AUC[t] was superior to the comparator arm using a two-tailed distribution. Considering the distinct censoring distribution across cohorts, the inverse probability of censoring weighting (as implemented in sksurv.metrics.cumulative_dynamic_auc) was estimated from the original training cohort for the CDK4/6i+ET arm, while for other cohorts it was estimated separately within each cohort. As a sensitivity analysis to account for less favorable clinical profiles among patients with HR+/HER2– MBC treated with first-line chemotherapy, we performed a propensity score matching between the chemotherapy and CDK4/6i+ET cohorts (using the MatchIt R package). Propensity scores were estimated by logistic regression using baseline clinical variables (ER and PR expression, tumor grade, number of metastatic sites, liver, CNS and bone-only disease, TFI ≤12 months, de novo metastatic disease and age ≤50 years), using a maximum distance of 0.2 standard deviations for pairs of the propensity score, with balance assessed by standardized mean differences (threshold <0.1, as shown in Table S6).

For gathering the external validation of CDKPredict, before deployment, the model was refitted and risk-cutoffs between risk-groups fine-tuned using the full MSK-cohort (n=1,264 cases), as per standard machine learning practice. This strategy was deliberately selected for optimizing the

ensemble model to the full set of variables and observations in the MSK-cohort, wherein rare yet meaningful features could have been allocated in an unbalanced manner between the train and the internal validation set. After model training on the full cohort, the final model coefficients across the n=100 base learners together with the fine-tuned cutoffs were locked and applied without any change to the external cohorts. For each patient, the complete preprocessing and feature-encoding pipeline used during training was replicated, after which the locked coefficient vectors were used to generate continuous risk-scores and the corresponding risk-group assignment.

To assess the clinical advantage of CDKPredict over the TFI ≤12 months criteria, we performed decision curve analysis comparing CDKPredict high-risk patients to those satisfying the TFI ≤12 months criterion, with net benefit calculated at each threshold using the 12-month progression as estimated from a Kaplan-Meier estimator to account for censored observations. For assessing calibration of CDKPredict in the external cohort, a Cox model fitted on out-of-bag risk estimates from the MSK cohort was used to measure the predicted probability of progression at 12 and 24 months; once divided into deciles, this was compared to the observed PFS-rate at 12 and 24 months using a Kaplan-Meier estimator.

Development of CDKPredict-S for baseline risk stratification in variable resource settings

Cognizant of the fact genomic sequencing remains unavailable in many clinical settings or is represented by panels interrogating a restricted number of genes, we intended to develop a scalable model, namely CDKPredict-S. Specifically, CDKPredict-S was designed to provide reliable risk-predictions, regardless of whether baseline somatic genomic sequencing was available, to remain robust in the presence of missing variables, and to rely on a restricted number of variables to facilitate its large-scale implementation.

For patients receiving baseline genomic sequencing, we developed a clinicogenomic version of CDKPredict-S designed to employ clinicogenomic variables for risk predictions. Specifically, CDKPredict-S was developed including the same clinical variables of CDKPredict; for genomic variables, CDKPredict-S was built using a restricted number of genomic biomarkers with established clinical relevance in HR+/HER2– MBC, that is having an established prognostic/predictive role for first-line CDK4/6i+ET, or representing actionable biomarkers for targeted therapies and currently investigated in clinical trials in the first-line setting for HR+/HER2– MBC; the final set of genomic variables for the scalable model included *PIK3CA*, *PTEN*, *AKT1*, *BRCA1*, *BRCA2*, *ESR1*, *TP53*, *RB1*. Thereafter, CDKPredict-S was developed using the same

framework and risk-group definition as CDKPredict. For patients not receiving genomic sequencing, we developed a second version of CDKPredict-S, which was designed to rely on the clinical risk-estimation for computing risk-predictions. In detail, a second CDKPredict-S model was trained and validated using the abovementioned clinical variables, together with clinical-only cutoffs to define cases at high-, intermediate- and low-risk of early progression to CDK4/6i+ET.

To mitigate systematic risk distortions originating from missing variables, CDKPredict-S was further designed to learn the overall prevalence of features from the development set (MSK-cohort) to estimate the marginal risk-contribution; such decision was undertaken hypothesizing the prevalence of the detected clinicogenomic features among 1,264 cases could approximate the prevalence of the variable among the general population of patients with breast cancer candidates for front-line CDK4/6i+ET. For estimating the baseline marginal risk-contribution of the feature, the overall mean coefficients originating from the n=100 base learners of the final model refitted on the overall MSK-cohort before deployment were multiplied by the learned prevalence of the features, resulting in a baseline risk-score for the missing variable and calibration of the model in risk estimation relative to the risk-strata cutoffs. Importantly, CDKPredict-S was designed to rely on the marginal risk-contribution from the development set rather than conducting data imputation considering the lack of biological and clinical plausibility in estimating a missing categorical feature from the specific variables included in CDKPredict-S.

Transcriptomic analysis

The transcriptomic analysis was conducted on a subset of cases included in the external cohort with baseline whole-exome transcriptome sequencing (n=450 of 564, 79.8%). Samples were processed using the Tempus RS RNA sequencing assay. Briefly, exome capture was performed using IDT xGen Exome Research Panel probes on a minimum of 50ng of total RNA extracted from FFPE tissue. Libraries were sequenced on the Illumina HiSeq 4000 platform to a target depth of 30–50 million reads. The bioinformatics pipeline utilized Kallisto (v0.44.0) for transcript-level pseudo-alignment and quantification against the Ensembl GRCh37 (release 75) reference. Gene-level counts were derived by summing the abundance of protein-coding transcripts covered by the RS panel.

For the downstream RNA-seq analysis, first, low-expressed genes across the study cohort were removed to retain genes totaling at least 1 count per million across at least 10% of the cohort (passing from 18,890 to 18,208 genes). Next, raw read counts were normalized using the variance

stabilizing transformation (VST) as implemented in DESeq2 using the PyDESeq2 package (https://github.com/scverse/PyDESeq2).

For computing PAM50 intrinsic subtypes, we utilized the genefu package (https://github.com/bhklab/genefu), as previously validated[3], to classify samples into five categories using the molecular.subtyping function, and particularly luminal A, luminal B, HER2-enriched, basal-like and normal-like.

Using VST-normalized counts, sample-specific pathway scores were measured using gene set variation analysis (GSVA) as implemented in the GSEApy package (https://github.com/zqfang/GSEApy). The pathway-level annotation was defined according to the Hallmark gene set (n=50 pathways) from the Molecular Signatures Database (MSigDB). For finding pathways significantly differently expressed between CDKPredict low- and high-risk cases, a Mann-Whitney U test corrected for multiple testing (using the Benjamini-Hochberg procedure with a false discovery rate < 0.05) was employed. The correlation between CDKPredict risk-scores and the pathway-level activity was carried out using a Spearman's correlation analysis with significance tested using the Benjamini-Hochberg procedure.

For defining estrogen low, medium and high tumors, we utilized the HALLMARK_ESTROGEN_RESPONSE_EARLY gene set from the Hallmark collection, representing a biological transcriptional output in response to the activation of the estrogen-receptor signaling pathway[62]. Utilizing sample-specific GSVA, we observed a trimodal probability density function, which was modelled using a three-component GMM to define three distinct groups using Brent's method, as described above for finding CDKPredict risk-group cutoffs.
For assessing the relative enrichment of the estrogen response early pathway activity in low-risk compared with high-risk tumors, we conducted a gene set enrichment analysis (GSEA) employing the GSEApy package, with genes ranked according to the metric: $\mathrm{sign}(\log2[\mathrm{Fold\ change}]) \times -\log10(\mathrm{qvalue})$ as calculated by a differential gene expression analysis conducted with DESeq2 (https://github.com/scverse/PyDESeq2) using VST normalization starting from raw counts, using the CDKPredict low-risk group as a reference.

DATA AVAILABILITY

Deidentified data needed to replicate our study findings are provided as Supplementary Tables. In accordance with Tempus and institutional policies of IEO, all requests for data collected or

curated in-house will be evaluated on a case-by-case basis to determine whether the data requested and the use case comply with intellectual property and patient privacy obligations.

CODE AVAILABILITY

The custom code used for developing CDKPredict and CDKPredict-S and the code used for conducting the study analysis are available at https://github.com/mskcc/ReCAST and https://github.com/lboscolobielo/CDKPredict.

REFERENCES

1. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med *351*, 2817–2826. 10.1056/NEJMoa041588.
2. Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F.L., Watson, D., Bryant, J., et al. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol *24*, 3726–3734. 10.1200/JCO.2005.04.7985.
3. Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol *27*, 1160–1167. 10.1200/JCO.2008.18.1370.
4. Kalinsky, K., Barlow, W.E., Gralow, J.R., Meric-Bernstam, F., Albain, K.S., Hayes, D.F., Lin, N.U., Perez, E.A., Goldstein, L.J., Chia, S.K.L., et al. (2021). 21-Gene Assay to Inform Chemotherapy Benefit in Node-Positive Breast Cancer. N Engl J Med *385*, 2336–2347. 10.1056/NEJMoa2108873.
5. Sparano, J.A., Gray, R.J., Makower, D.F., Pritchard, K.I., Albain, K.S., Hayes, D.F., Geyer, C.E., Jr., Dees, E.C., Goetz, M.P., Olson, J.A., Jr., et al. (2018). Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. N Engl J Med *379*, 111–121. 10.1056/NEJMoa1804710.
6. Cardoso, F., van't Veer, L.J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Pierga, J.Y., Brain, E., Causeret, S., DeLorenzi, M., et al. (2016). 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med *375*, 717–729. 10.1056/NEJMoa1602253.
7. Sparano, J.A., Crager, M.R., Tang, G., Gray, R.J., Stemmer, S.M., and Shak, S. (2021). Development and Validation of a Tool Integrating the 21-Gene Recurrence Score and Clinical-Pathological Features to Individualize Prognosis and Prediction of Chemotherapy Benefit in Early Breast Cancer. J Clin Oncol *39*, 557–564. 10.1200/JCO.20.03007.
8. Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Dunbier, A.K., Cowens, J.W., Ferree, S., Storhoff, J., Schaper, C., and Cuzick, J. (2013). Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. J Clin Oncol *31*, 2783–2790. 10.1200/JCO.2012.46.1558.
9. Finn, R.S., Martin, M., Rugo, H.S., Jones, S., Im, S.A., Gelmon, K., Harbeck, N., Lipatov, O.N., Walshe, J.M., Moulder, S., et al. (2016). Palbociclib and Letrozole in Advanced Breast Cancer. N Engl J Med *375*, 1925–1936. 10.1056/NEJMoa1607303.
10. Goetz, M.P., Toi, M., Huober, J., Sohn, J., Tredan, O., Park, I.H., Campone, M., Chen, S.C., Manso, L.M., Paluch-Shimon, S., et al. (2024). Abemaciclib plus a nonsteroidal aromatase inhibitor as initial therapy for HR+, HER2- advanced breast cancer: final overall survival results of MONARCH 3. Ann Oncol *35*, 718–727. 10.1016/j.annonc.2024.04.013.
11. Hortobagyi, G.N., Stemmer, S.M., Burris, H.A., Yap, Y.S., Sonke, G.S., Hart, L., Campone, M., Petrakova, K., Winer, E.P., Janni, W., et al. (2022). Overall Survival with Ribociclib plus Letrozole in Advanced Breast Cancer. N Engl J Med *386*, 942–950. 10.1056/NEJMoa2114663.
12. Im, S.A., Lu, Y.S., Bardia, A., Harbeck, N., Colleoni, M., Franke, F., Chow, L., Sohn, J., Lee, K.S., Campos-Gomez, S., et al. (2019). Overall Survival with Ribociclib plus Endocrine Therapy in Breast Cancer. N Engl J Med *381*, 307–316. 10.1056/NEJMoa1903765.

13. Safonov, A., Lee, M., Brown, D.N., Boscolo Bielo, L., Mehine, M., Bandlamudi, C., O'Leary, B., Shao, H., Vicente, L., Muldoon, D., et al. (2026). Homologous recombination deficiency and hemizygosity drive resistance in breast cancer. Nature *652*, 752–762. 10.1038/s41586-026-10197-0.
14. Turner, N.C., Liu, Y., Zhu, Z., Loi, S., Colleoni, M., Loibl, S., DeMichele, A., Harbeck, N., Andre, F., Bayar, M.A., et al. (2019). Cyclin E1 Expression and Palbociclib Efficacy in Previously Treated Hormone Receptor-Positive Metastatic Breast Cancer. J Clin Oncol *37*, 1169–1178. 10.1200/JCO.18.00925.
15. Herrera-Abreu, M.T., Palafox, M., Asghar, U., Rivas, M.A., Cutts, R.J., Garcia-Murillas, I., Pearson, A., Guzman, M., Rodriguez, O., Grueso, J., et al. (2016). Early Adaptation and Acquired Resistance to CDK4/6 Inhibition in Estrogen Receptor-Positive Breast Cancer. Cancer Res *76*, 2301–2313. 10.1158/0008-5472.CAN-15-0728.
16. Li, Z., Razavi, P., Li, Q., Toy, W., Liu, B., Ping, C., Hsieh, W., Sanchez-Vega, F., Brown, D.N., Da Cruz Paula, A.F., et al. (2018). Loss of the FAT1 Tumor Suppressor Promotes Resistance to CDK4/6 Inhibitors via the Hippo Pathway. Cancer Cell *34*, 893–905 e898. 10.1016/j.ccell.2018.11.006.
17. Li, Q., Jiang, B., Guo, J., Shao, H., Del Priore, I.S., Chang, Q., Kudo, R., Li, Z., Razavi, P., Liu, B., et al. (2022). INK4 Tumor Suppressor Proteins Mediate Resistance to CDK4/6 Kinase Inhibitors. Cancer Discov *12*, 356–371. 10.1158/2159-8290.CD-20-1726.
18. Kudo, R., Safonov, A., Jones, C., Moiso, E., Dry, J.R., Shao, H., Nag, S., da Silva, E.M., Yildirim, S.Y., Li, Q., et al. (2024). Long-term breast cancer response to CDK4/6 inhibition defined by TP53-mediated geroconversion. Cancer Cell *42*, 1919–1935 e1919. 10.1016/j.ccell.2024.09.009.
19. Wander, S.A., Cohen, O., Gong, X., Johnson, G.N., Buendia-Buendia, J.E., Lloyd, M.R., Kim, D., Luo, F., Mao, P., Helvie, K., et al. (2020). The Genomic Landscape of Intrinsic and Acquired Resistance to Cyclin-Dependent Kinase 4/6 Inhibitors in Patients with Hormone Receptor-Positive Metastatic Breast Cancer. Cancer Discov *10*, 1174–1193. 10.1158/2159-8290.CD-19-1390.
20. Kim, S., Armand, J., Safonov, A., Zhang, M., Soni, R.K., Schwartz, G., McGuinness, J.E., Hibshoosh, H., Razavi, P., Kim, M., et al. (2023). Sequential activation of E2F via Rb degradation and c-Myc drives resistance to CDK4/6 inhibitors in breast cancer. Cell Rep *42*, 113198. 10.1016/j.celrep.2023.113198.
21. Cai, Y., Zhao, P., Wu, F., Zhao, H., Shao, H., Marra, A., Patel, P., O'Connell, E., Fink, E., Miele, M.M., et al. (2025). Inhibition of NR2F2 restores hormone therapy response to endocrine refractory breast cancers. Sci Transl Med *17*, eadk7786. 10.1126/scitranslmed.adk7786.
22. Razavi, P., Chang, M.T., Xu, G., Bandlamudi, C., Ross, D.S., Vasan, N., Cai, Y., Bielski, C.M., Donoghue, M.T.A., Jonsson, P., et al. (2018). The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers. Cancer Cell *34*, 427–438 e426. 10.1016/j.ccell.2018.08.008.
23. Toy, W., Shen, Y., Won, H., Green, B., Sakr, R.A., Will, M., Li, Z., Gala, K., Fanning, S., King, T.A., et al. (2013). ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. Nat Genet *45*, 1439–1445. 10.1038/ng.2822.
24. Arruabarrena-Aristorena, A., Maag, J.L.V., Kittane, S., Cai, Y., Karthaus, W.R., Ladewig, E., Park, J., Kannan, S., Ferrando, L., Cocco, E., et al. (2020). FOXA1 Mutations Reveal Distinct Chromatin Profiles and Influence Therapeutic Response in Breast Cancer. Cancer Cell *38*, 534–550 e539. 10.1016/j.ccell.2020.08.003.
25. Costa, C., Wang, Y., Ly, A., Hosono, Y., Murchie, E., Walmsley, C.S., Huynh, T., Healy, C., Peterson, R., Yanase, S., et al. (2020). PTEN Loss Mediates Clinical Cross-

Resistance to CDK4/6 and PI3Kalpha Inhibitors in Breast Cancer. Cancer Discov *10*, 72–85. 10.1158/2159-8290.CD-18-0830.

26. Formisano, L., Lu, Y., Servetto, A., Hanker, A.B., Jansen, V.M., Bauer, J.A., Sudhan, D.R., Guerrero-Zotano, A.L., Croessmann, S., Guo, Y., et al. (2019). Aberrant FGFR signaling mediates resistance to CDK4/6 inhibitors in ER+ breast cancer. Nat Commun *10*, 1373. 10.1038/s41467-019-09068-2.
27. Morrison, L., Loibl, S., and Turner, N.C. (2024). The CDK4/6 inhibitor revolution - a game-changing era for breast cancer treatment. Nat Rev Clin Oncol *21*, 89–105. 10.1038/s41571-023-00840-4.
28. Nayar, U., Cohen, O., Kapstad, C., Cuoco, M.S., Waks, A.G., Wander, S.A., Painter, C., Freeman, S., Persky, N.S., Marini, L., et al. (2019). Acquired HER2 mutations in ER(+) metastatic breast cancer confer resistance to estrogen receptor-directed therapies. Nat Genet *51*, 207–216. 10.1038/s41588-018-0287-5.
29. Bertucci, F., Ng, C.K.Y., Patsouris, A., Droin, N., Piscuoglio, S., Carbuccia, N., Soria, J.C., Dien, A.T., Adnani, Y., Kamal, M., et al. (2019). Genomic characterization of metastatic breast cancers. Nature *569*, 560–564. 10.1038/s41586-019-1056-z.
30. Hanker, A.B., Sudhan, D.R., and Arteaga, C.L. (2020). Overcoming Endocrine Resistance in Breast Cancer. Cancer Cell *37*, 496–513. 10.1016/j.ccell.2020.03.009.
31. Kim, R., Yu, J., Lim, J., Oh, B.B., Nam, S.J., Kim, S.W., Lee, J.E., Chae, B.J., Kim, J.Y., Park, G.E., et al. (2026). Whole-genome landscapes of 1,364 breast cancers. Nature *649*, 1282–1291. 10.1038/s41586-025-09812-3.
32. Gupta, A., Gazzo, A., Selenica, P., Safonov, A., Pareja, F., da Silva, E.M., Brown, D.N., Shao, H., Zhu, Y., Patel, J., et al. (2025). APOBEC3 mutagenesis drives therapy resistance in breast cancer. Nat Genet *57*, 1452–1462. 10.1038/s41588-025-02187-1.
33. Andre, F., Su, F., Solovieff, N., Hortobagyi, G., Chia, S., Neven, P., Bardia, A., Tripathy, D., Lu, Y.S., Lteif, A., et al. (2023). Pooled ctDNA analysis of MONALEESA phase III advanced breast cancer trials. Ann Oncol *34*, 1003–1014. 10.1016/j.annonc.2023.08.011.
34. Finn, R.S., Dering, J., Conklin, D., Kalous, O., Cohen, D.J., Desai, A.J., Ginther, C., Atefi, M., Chen, I., Fowst, C., et al. (2009). PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. Breast Cancer Res *11*, R77. 10.1186/bcr2419.
35. Turner, N.C., Im, S.A., Saura, C., Juric, D., Loibl, S., Kalinsky, K., Schmid, P., Loi, S., Sunpaweravong, P., Musolino, A., et al. (2024). Inavolisib-Based Therapy in PIK3CA-Mutated Advanced Breast Cancer. N Engl J Med *391*, 1584–1596. 10.1056/NEJMoa2404625.
36. Cheng, D.T., Mitchell, T.N., Zehir, A., Shah, R.H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z.Y., Won, H.H., Scott, S.N., et al. (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. J Mol Diagn *17*, 251–264. 10.1016/j.jmoldx.2014.12.006.
37. Rugo, H.S., Layman, R.M., Lynce, F., Liu, X., Li, B., McRoy, L., Cohen, A.B., Estevez, M., Curigliano, G., and Brufsky, A. (2025). Real-world progression-free survival of CDK4/6 inhibitors plus an aromatase inhibitor in HR-positive/HER2-negative metastatic breast cancer in United States routine clinical practice. ESMO Open *10*, 105570. 10.1016/j.esmoop.2025.105570.
38. Rugo, H.S., Im, S.A., Joy, A.A., Shparyk, Y., Walshe, J.M., Sleckman, B., Loi, S., Theall, K.P., Kim, S., Huang, X., et al. (2022). Effect of palbociclib plus endocrine therapy on time to chemotherapy across subgroups of patients with hormone receptor–positive/human epidermal growth factor receptor 2–negative advanced breast cancer:

Post hoc analyses from PALOMA-2 and PALOMA-3. Breast *66*, 324–331. 10.1016/j.breast.2022.11.005.

39. Di Leo, A., O'Shaughnessy, J., Sledge, G.W., Jr., Martin, M., Lin, Y., Frenzel, M., Hardebeck, M.C., Smith, I.C., Llombart-Cussac, A., Goetz, M.P., and Johnston, S. (2018). Prognostic characteristics in hormone receptor-positive advanced breast cancer and characterization of abemaciclib efficacy. NPJ Breast Cancer *4*, 41. 10.1038/s41523-018-0094-2.
40. Breiman, L. (2001). Random Forests. Machine Learning *45*, 5–32. 10.1023/A:1010933404324.
41. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754. 10.48550/arXiv.1603.02754.
42. Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association *101*, 1418–1429. 10.1198/016214506000000735.
43. Breiman, L. (1996). Bagging predictors. Machine Learning *24*, 123–140. 10.1007/BF00058655.
44. Arango-Argoty, G., Kipkogei, E., Stewart, R., Sun, G.J., Patra, A., Kagiampakis, I., and Jacob, E. (2025). Pretrained transformers applied to clinical studies improve predictions of treatment efficacy and associated biomarkers. Nat Commun *16*, 2101. 10.1038/s41467-025-57181-2.
45. Hollmann, N., Muller, S., Purucker, L., Krishnakumar, A., Korfer, M., Hoo, S.B., Schirrmeister, R.T., and Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. Nature *637*, 319–326. 10.1038/s41586-024-08328-6.
46. Introducing tabfm: A Zero-shot foundation model for Tabular Data. Google Research - Explore Our Latest Research in Science and AI. Accessed on September 1, 2026 at https://research.google/blog/introducing-tabfm-a-zero-shot-foundation-model-for-tabular-data/
47. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165. 10.48550/arXiv.2005.14165.
48. Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol *2017*. 10.1200/PO.17.00011.
49. Suehnholz, S.P., Nissan, M.H., Zhang, H., Kundra, R., Nandakumar, S., Lu, C., Carrero, S., Dhaneshwar, A., Fernandez, N., Xu, B.W., et al. (2024). Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. Cancer Discov *14*, 49–65. 10.1158/2159-8290.CD-23-0467.
50. Vergouwe, Y., Moons, K.G., and Steyerberg, E.W. (2010). External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol *172*, 971–980. 10.1093/aje/kwq223.
51. Beaubier, N., Tell, R., Huether, R., Bontrager, M., Bush, S., Parsons, J., Shah, K., Baker, T., Selkov, G., Taxter, T., et al. (2018). Clinical validation of the Tempus xO assay. Oncotarget *9*, 25826–25832. 10.18632/oncotarget.25381.
52. Boulos, H., Lo, C., Zhu, W., Driessen, T.M., Yamada-Hanff, J., Harding, T., Lozac'hmeur, A., Pereira, T., Sonnenschein, A., Och, J., et al. (2025). Analytical Validation of Next-Generation Sequencing-Based Comprehensive Liquid Biopsy Assay for Therapy Selection. J Mol Diagn *27*, 383–394. 10.1016/j.jmoldx.2025.02.005.
53. Finkle, J.D., Boulos, H., Driessen, T.M., Lo, C., Blidner, R.A., Hafez, A., Khan, A.A., Lozac'hmeur, A., McKinnon, K.E., Perera, J., et al. (2021). Validation of a liquid biopsy assay with molecular and clinical profiling of circulating tumor DNA. NPJ Precis Oncol *5*, 63. 10.1038/s41698-021-00202-2.

54. Jeselsohn, R., Buchwalter, G., De Angelis, C., Brown, M., and Schiff, R. (2015). ESR1 mutations-a mechanism for acquired endocrine resistance in breast cancer. Nat Rev Clin Oncol *12*, 573–583. 10.1038/nrclinonc.2015.117.
55. Ma, J., Li, L., Ma, B., Liu, T., Wang, Z., Ye, Q., Peng, Y., Wang, B., Chen, Y., Xu, S., et al. (2024). MYC induces CDK4/6 inhibitors resistance by promoting pRB1 degradation. Nat Commun *15*, 1871. 10.1038/s41467-024-45796-w.
56. Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70. 10.1038/nature11412.
57. Kouros-Mehr, H., Slorach, E.M., Sternlicht, M.D., and Werb, Z. (2006). GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. Cell *127*, 1041–1055. 10.1016/j.cell.2006.09.048.
58. Azim, H.A., Elghazawy, H., Shohdy, K.S., Lasheen, S., Elghazawy, M., Ghazy, R.M., Abdelnasser, D., and Kassem, L. (2025). Impact of BRCA Mutation on Treatment Outcomes of Endocrine Therapy +/- CDK4/6 Inhibitors in Hormone Receptor-Positive/Human Epidermal Growth Factor Receptor-2-Negative Breast Cancer: A Systematic Review and Meta-Analysis. JCO Precis Oncol *9*, e2400841. 10.1200/PO-24-00841.
59. Gorski, J.J., Kennedy, R.D., Hosey, A.M., and Harkin, D.P. (2009). The complex relationship between BRCA1 and ERalpha in hereditary breast cancer. Clin Cancer Res *15*, 1514–1518. 10.1158/1078-0432.CCR-08-0640.
60. Hu, J., Parsons, J., Mineo, B., Bell, J.S.K., Malinauskas, J., Drews, J., Michuda, J., McCarter, C., Dhond, R., Tyndall, J.C., et al. (2021). Abstract 2239: Comprehensive validation of RNA sequencing for clinical NGS fusion genes and RNA expression reporting. Cancer Research *81*, 2239–2239. 10.1158/1538-7445.AM2021-2239.
61. Sinn, B.V., Fu, C., Lau, R., Litton, J., Tsai, T.H., Murthy, R., Tam, A., Andreopoulou, E., Gong, Y., Murthy, R., et al. (2019). SET(ER/PR): a robust 18-gene predictor for sensitivity to endocrine therapy for metastatic breast cancer. NPJ Breast Cancer *5*, 16. 10.1038/s41523-019-0111-0.
62. Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst *1*, 417–425. 10.1016/j.cels.2015.12.004.
63. Turner, N.C., Oliveira, M., Howell, S.J., Dalenc, F., Cortes, J., Gomez Moreno, H.L., Hu, X., Jhaveri, K., Krivorotko, P., Loibl, S., et al. (2023). Capivasertib in Hormone Receptor-Positive Advanced Breast Cancer. N Engl J Med *388*, 2058–2070. 10.1056/NEJMoa2214131.
64. Bidard, F.C., Kaklamani, V.G., Neven, P., Streich, G., Montero, A.J., Forget, F., Mouret-Reynier, M.A., Sohn, J.H., Taylor, D., Harnden, K.K., et al. (2022). Elacestrant (oral selective estrogen receptor degrader) Versus Standard Endocrine Therapy for Estrogen Receptor-Positive, Human Epidermal Growth Factor Receptor 2-Negative Advanced Breast Cancer: Results From the Randomized Phase III EMERALD Trial. J Clin Oncol *40*, 3246–3256. 10.1200/JCO.22.00338.
65. Jhaveri, K.L., Neven, P., Casalnuovo, M.L., Kim, S.B., Tokunaga, E., Aftimos, P., Saura, C., O'Shaughnessy, J., Harbeck, N., Carey, L.A., et al. (2025). Imlunestrant with or without Abemaciclib in Advanced Breast Cancer. N Engl J Med *392*, 1189–1202. 10.1056/NEJMoa2410858.
66. Jhaveri, K., Eli, L.D., Wildiers, H., Hurvitz, S.A., Guerrero-Zotano, A., Unni, N., Brufsky, A., Park, H., Waisman, J., Yang, E.S., et al. (2023). Neratinib + fulvestrant + trastuzumab for HR-positive, HER2-negative, HER2-mutant metastatic breast cancer: outcomes and biomarker analysis from the SUMMIT trial. Ann Oncol *34*, 885–898. 10.1016/j.annonc.2023.08.003.

67. Hyman, D.M., Piha-Paul, S.A., Won, H., Rodon, J., Saura, C., Shapiro, G.I., Juric, D., Quinn, D.I., Moreno, V., Doger, B., et al. (2018). HER kinase inhibition in patients with HER2- and HER3-mutant cancers. Nature *554*, 189–194. 10.1038/nature25475.
68. Marabelle, A., Fakih, M., Lopez, J., Shah, M., Shapira-Frommer, R., Nakagawa, K., Chung, H.C., Kindler, H.L., Lopez-Martin, J.A., Miller, W.H., Jr., et al. (2020). Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. Lancet Oncol *21*, 1353–1365. 10.1016/S1470-2045(20)30445-9.
69. Litchfield, K., Reading, J.L., Puttick, C., Thakkar, K., Abbosh, C., Bentham, R., Watkins, T.B.K., Rosenthal, R., Biswas, D., Rowan, A., et al. (2021). Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. Cell *184*, 596–614 e514. 10.1016/j.cell.2021.01.002.
70. Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J.M., Desrichard, A., Walsh, L.A., Postow, M.A., Wong, P., Ho, T.S., et al. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med *371*, 2189–2199. 10.1056/NEJMoa1406498.
71. Robson, M.E., Tung, N., Conte, P., Im, S.A., Senkus, E., Xu, B., Masuda, N., Delaloge, S., Li, W., Armstrong, A., et al. (2019). OlympiAD final overall survival and tolerability results: Olaparib versus chemotherapy treatment of physician's choice in patients with a germline BRCA mutation and HER2-negative metastatic breast cancer. Ann Oncol *30*, 558–566. 10.1093/annonc/mdz012.
72. Tung, N.M., Robson, M.E., Li, T., Nanda, R., Shah, P.D., Khoury, K., Kimmick, G., Santa-Maria, C., Brufsky, A., DeMeo, M., et al. (2026). TBCRC 048 (Olaparib Expanded) Expansion Cohorts: Phase II Study of Olaparib Monotherapy for Patients With Metastatic Breast Cancer With Germline Mutations in PALB2 or Somatic Mutations in BRCA1 or BRCA2. J Clin Oncol *44*, 653–661. 10.1200/JCO-25-02075.
73. Litton, J.K., Hurvitz, S.A., Mina, L.A., Rugo, H.S., Lee, K.H., Goncalves, A., Diab, S., Woodward, N., Goodwin, A., Yerushalmi, R., et al. (2020). Talazoparib versus chemotherapy in patients with germline BRCA1/2-mutated HER2-negative advanced breast cancer: final overall survival results from the EMBRACA trial. Ann Oncol *31*, 1526–1535. 10.1016/j.annonc.2020.08.2098.
74. Al-Kateb, H., Knight, S.M., Sivasankaran, G., Voss, J.S., Pitel, B.A., Blommel, J.H., Jerde, C.R., Rumilla, K.M., Lee, J.L., Mattson, N.R., et al. (2025). Clinical Validation of the TruSight Oncology 500 Assay for the Detection and Reporting of Pan-Cancer Biomarkers. J Mol Diagn *27*, 292–305. 10.1016/j.jmoldx.2025.01.002.
75. Normanno, N., Machado, J.C., Pescarmona, E., Buglioni, S., Navarro, L., Esposito Abate, R., Ferro, A., Mensink, R., Lambiase, M., Lespinet-Fabre, V., et al. (2023). European Real-World Assessment of the Clinical Validity of a CE-IVD Panel for Ultra-Fast Next-Generation Sequencing in Solid Tumors. Int J Mol Sci *24*. 10.3390/ijms241813788.
76. Youssef, A., Pencina, M., Thakur, A., Zhu, T., Clifton, D., and Shah, N.H. (2023). External validation of AI models in health should be replaced with recurring local validation. Nat Med *29*, 2686–2687. 10.1038/s41591-023-02540-z.
77. Aldea, M., Salto-Tellez, M., Marra, A., Umeton, R., Stenzinger, A., Koopman, M., Prelaj, A., Kehl, K.L., Gilbert, S., Lessmann, M.E., et al. (2026). ESMO basic requirements for AI-based biomarkers in oncology (EBAI). Ann Oncol *37*, 414–430. 10.1016/j.annonc.2025.11.009.
78. Wortelboer, N., van Ommen-Nijhof, A., Konings, I.R., van der Noort, V., van den Pol, E., Guerrero Paez, C., Bakker, S.D., Hovenga, S., Imholz, A.L.T., Kamm, Y.J.L., et al. (2026). Overall Survival With First-Line vs Second-Line CDK4/6 Inhibitor Use in

Advanced Breast Cancer: A Randomized Clinical Trial. JAMA Oncol *12*, 375–383. 10.1001/jamaoncol.2025.6585.

79. Shen, R., and Seshan, V.E. (2016). FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res *44*, e131. 10.1093/nar/gkw520.

80. Bielski, C.M., Zehir, A., Penson, A.V., Donoghue, M.T.A., Chatila, W., Armenia, J., Chang, M.T., Schram, A.M., Jonsson, P., Bandlamudi, C., et al. (2018). Genome doubling shapes the evolution and prognosis of advanced cancers. Nat Genet *50*, 1189–1195. 10.1038/s41588-018-0165-1.

81. Batalini, F., Gulhan, D.C., Mao, V., Tran, A., Polak, M., Xiong, N., Tayob, N., Tung, N.M., Winer, E.P., Mayer, E.L., et al. (2022). Mutational Signature 3 Detected from Clinical Panel Sequencing is Associated with Responses to Olaparib in Breast and Ovarian Cancers. Clin Cancer Res *28*, 4714–4723. 10.1158/1078-0432.CCR-22-0749.

82. Rakha, E.A., El-Sayed, M.E., Lee, A.H., Elston, C.W., Grainge, M.J., Hodi, Z., Blamey, R.W., and Ellis, I.O. (2008). Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol *26*, 3153–3158. 10.1200/JCO.2007.15.5986.

83. Wang, X., Lamberti, G., Di Federico, A., Alessi, J., Ferrara, R., Sholl, M.L., Awad, M.M., Vokes, N., and Ricciuti, B. (2024). Tumor mutational burden for the prediction of PD-(L)1 blockade efficacy in cancer: challenges and opportunities. Ann Oncol *35*, 508–522. 10.1016/j.annonc.2024.03.007.

84. Zauderer, M.G., Martin, A., Egger, J., Rizvi, H., Offin, M., Rimner, A., Adusumilli, P.S., Rusch, V.W., Kris, M.G., Sauter, J.L., et al. (2021). The use of a next-generation sequencing-derived machine-learning risk-prediction model (OncoCast-MPM) for malignant pleural mesothelioma: a retrospective study. Lancet Digit Health *3*, e565–e576. 10.1016/S2589-7500(21)00104-7.

85. Shen, R., Martin, A., Ni, A., Hellmann, M., Arbour, K.C., Jordan, E., Arora, A., Ptashkin, R., Zehir, A., Kris, M.G., et al. (2019). Harnessing Clinical Sequencing Data for Survival Stratification of Patients with Metastatic Lung Adenocarcinomas. JCO Precis Oncol *3*. 10.1200/PO.18.00307.

86. Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? , arXiv:2207.08815. 10.48550/arXiv.2207.08815.

87. Biswas, D., Birkbak, N.J., Rosenthal, R., Hiley, C.T., Lim, E.L., Papp, K., Boeing, S., Krzystanek, M., Djureinovic, D., La Fleur, L., et al. (2019). A clonal expression biomarker associates with lung cancer mortality. Nat Med *25*, 1540–1548. 10.1038/s41591-019-0595-z.

88. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805. 10.48550/arXiv.1810.04805.

89. Grinsztajn, L., Flöge, K., Key, O., Birkel, F., Jund, P., Roof, B., Manium, M., Hoo, S.B., Bühler, M., Garg, A., et al. (2026). TabPFN-3: Technical Report. arXiv:2605.13986. 10.48550/arXiv.2605.13986.