Real-world EHR-derived progression-free survival across successive lines of therapy informs metastatic breast cancer risk stratification

Xuan Zhao[1], Thibault Niederhauser[2], Zsolt Balazs[2,3], Andreas Wicki[3,4], Bowen Fan[†2,3*], Michael Krauthammer[†2*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Klingelbergstrasse 48, Basel, 4056, Basel-Stadt, Switzerland.
[2]Department of Quantitative Biomedicine, University of Zürich, Schmelzbergstrasse 26, Zürich, 8091, Zürich, Switzerland.
[3]Department of Medical Oncology and Hematology, University Hospital Zürich, Rämistrasse 100, Zürich, 8006, Zürich, Switzerland.
[4]Faculty of Medicine, University of Zürich, Pestalozzistrasse 3, Zürich, 8032, Zürich, Switzerland.

*Corresponding author(s). E-mail(s): bowen.fan@uzh.ch; michael.krauthammer@uzh.ch;
Contributing authors: xuazhao@student.ethz.ch; thibault.niederhauser@uzh.ch; zsolt.balazs@uzh.ch; andreas.wicki@usz.ch;

Abstract

Guideline-based recommendations for metastatic lines of therapy (mLoTs), especially second lines and beyond, are comparatively sparse due to challenges in later-line treatment efficacy quantification. Scalable real-world evidence that captures the interaction between treatment and disease progression is therefore especially valuable, as regimens become increasingly individualized, confounding intensifies, and progression is rarely recorded as a structured EHR endpoint. We present a framework to (i) reconstruct clinically coherent mLoTs from longitudinal EHR using radiology-anchored progression evidence and (ii) generate individualized progression-free survival (PFS) estimates from a line-start multimodal snapshot in a highly heterogeneous cohort. In 2,881 patients contributing

[†]These authors jointly supervised this work.

8,791 metastatic mLoTs, the selected model shows good discrimination over a 2-year horizon (Antolini's C = **0.681 ± 0.006**; cumulative/dynamic AUC at 1 year = **0.824 ± 0.008**). Predicted risk strata closely track Kaplan–Meier trends across line number and tumor subtypes, supporting risk stratification with similar calibration across sub-cohorts. Model prediction primarily relies on clinically plausible signals of recent metastatic burden and tumor markers, with limited dependence on surveillance cadence or subtype labels, and is robust to modality ablation. Together, this framework supports scalable evidence generation for metastatic line PFS patterns and interpretable, broadly calibrated prognostication to inform risk assessment and care planning in heterogeneous metastatic practice.

1 Introduction

Metastatic breast cancer (mBC) is an advanced disease typically treated with successive palliative systemic lines of therapy (metastatic lines of therapy; mLoTs) [1, 2]. At line start, treatment benefit estimation can help guide surveillance and subsequent-line planning. In both trials and real-world studies, such benefit is commonly summarized by progression-free survival (PFS), defined as time from treatment initiation to disease progression or death [3]. However, such evidence is strongest for curative or first-line metastatic therapy [4]. Beyond first line (mLoT2+), treatment gains often attenuate while heterogeneity and regimen customization increase [5–7], limiting the availability and generalizability of efficacy quantification. Predictive modeling of treatment response heterogeneity can support individualized decisions beyond average treatment effects [8], and scalable real-world evidence generation is a practical route to closing this gap.

Machine learning (ML) models that integrate multimodal clinical data can strengthen these precision oncology approaches [9]. Prior work on mBC PFS prediction is comparatively limited and often focuses on specific regimens, homogeneous cohorts, or curated feature sets. For example, Varma et al. [10] used natural language processing (NLP) to derive PFS in endocrine therapy plus CDK4/6 inhibitor-treated patients; Merzhevich et al. [11] modeled PFS in a ribociclib-treated cohort; and Cabel et al. [12] studied later-line outcomes in triple-negative breast cancer (TNBC). A natural next step is an interpretable, multimodal model trained on diverse real-world care that leverages population-level progression dynamics to share information across subtypes and heterogeneous regimen mixes.

We present a framework that uses longitudinal EHR to individualize PFS estimation at line level and address two practical challenges. First, clinically coherent mLoT reconstruction and credible PFS labeling are difficult in routine data: line boundaries and segmentation algorithms vary across cancer types, data sources and conventions [13–16]. Furthermore, progression is rarely recorded as a structured endpoint, often inferred from unstructured documentation, increasingly via NLP approaches [10, 17–19]. Second, rich multimodal predictors can induce leakage and shortcut learning; for example, complete within-line treatment records can

Table 1: Cohort size, PFS summary, and first-line (mLoT1) treatment-category combinations overall and by disease subtype. Median PFS is Kaplan–Meier estimated (days); event rate is percent.

	HR/HER2 status				
	±/±	−/−	−/+	+/−	+/+
Patients	2,881	440	147	1,948	346
Total mLoTs	8,791	1,032	437	6,130	1,192
Avg mLoTs	3.05	2.35	2.97	3.15	3.45
Event rate	86.1	91.1	76.0	86.5	83.3
Median PFS	115	85	155	116	145
mLoT1	2,342	389	141	1,565	247
– Event rate	71.1	83.3	63.1	69.3	67.6
– Median PFS	231	103	268	275	240
Top 1 mLoT1		Chemo	Chemo, Biologic	Hormone, Targeted	Chemo, Biologic
– # (%)		289 (74.3%)	81 (57.4%)	703 (44.9%)	82 (33.2%)
Top 2 mLoT1		Chemo, Immuno	Biologic	Hormone	Hormone, Biologic
– # (%)		39 (10.0%)	25 (17.7%)	528 (33.7%)	55 (22.3%)

encode post-initiation adaptations (e.g., switches for toxicity or early progression), and extensive surveillance patterns could proxy for poor prognosis.

Our workflow (Fig. 1a) addresses these challenges by (i) reconstructing mLoTs with clinically informed heuristics and stringent filtering and anchoring PFS to NLP-derived radiology progression assessments; (ii) predicting PFS from a line-start multimodal snapshot and deliberately limiting non-biological proxies. We evaluate discrimination, calibration, and interpretability to demonstrate model robustness and potential clinical utility.

Applied to the mBC subset of MSK-CHORD [19], this pipeline yields a heterogeneous cohort of clinically plausible treatment regimens and outcomes. The selected model shows good agreement between predicted and observed outcomes overall and reasonable calibration across disease subtypes. These properties support clinically interpretable, decision-point risk estimates that can inform surveillance intensity and proactive subsequent-line planning.

2 Results

2.1 Cohort overview

We applied our mLoT reconstruction and PFS labeling pipeline to the mBC subset of MSK-CHORD [19], a public clinicogenomic dataset of patients with metastatic solid tumors treated at Memorial Sloan Kettering Cancer Center (MSK). The analytic cohort includes 2,881 patients contributing 8,791 mLoTs (Table 1). HR+/HER2− disease comprises 67.6% of patients, followed by 15.3% triple-negative breast cancer

Table 2: Global performance across survival models for PFS prediction. C-index: Antolini's C [22]. Mean AUC: weighted mean of the integrated cumulative/dynamic time-dependent area under the ROC curve (cdAUC [23]). AUC @ 1y and AUC @ 2y: cdAUC at 1 and 2 years. IBS: integrated Brier score [24].

Metric	CoxPH	DeepHit	DeepSurv	GBSA	RSF
C-index	0.679 ± 0.007	$\mathbf{0.685 \pm 0.008}$	0.680 ± 0.009	0.681 ± 0.006	0.674 ± 0.006
mean AUC	0.754 ± 0.008	$\mathbf{0.759 \pm 0.007}$	0.755 ± 0.011	0.757 ± 0.006	0.747 ± 0.005
AUC @ 1y	0.818 ± 0.008	0.811 ± 0.008	0.816 ± 0.012	$\mathbf{0.824 \pm 0.008}$	0.814 ± 0.009
AUC @ 2y	0.842 ± 0.013	0.806 ± 0.026	0.841 ± 0.015	$\mathbf{0.847 \pm 0.012}$	0.845 ± 0.012
IBS	0.126 ± 0.004	0.140 ± 0.007	0.125 ± 0.005	$\mathbf{0.124 \pm 0.004}$	0.125 ± 0.004

(TNBC) and 17.1% HER2+. The distribution is broadly consistent with population statistics [21] with modest TNBC and HER2+ enrichment expected in a metastatic cohort. PFS differs across line numbers and disease subtypes. mLoT1 shows lower event rates and longer median PFS than population average, a result of both the predominance of mLoT2+ episodes (73.4%) and diminishing benefits. TNBC mLoTs have higher event rates, shorter median PFS, and fewer lines received than other subtypes, plausibly reflecting fewer effective treatment options and a more severe prognosis. Among patients with valid mLoT1 regimens, the median number of therapy lines received in the first year is 1.

Regimen composition differs substantially across subtypes (Table 1 and Extended Data Tables 4 and 5). In mLoT1, top 2 treatment combination backbones account for a substantial share of episodes. TNBC is dominated by chemotherapy backbones (74.3%), with common agents including capecitabine (19.5%) and paclitaxel (13.6%). HR−/HER2+ regimens are centered on HER2-targeted biologics, most often combined with chemotherapy (together 75.1% of mLoT1 episodes); paclitaxel–pertuzumab–trastuzumab is the single most common drug combination (38.3%). In HR+/HER2− disease, endocrine therapy with and without targeted agents dominates (44.9% and 33.7%, respectively). However, HR+ disease remains heterogeneous in drug combinations: in HR+/HER2−, the top two drug combinations account for only 21.3% of episodes (letrozole + palbociclib: 12.5%; letrozole alone: 8.8%), and similarly in HR+/HER2+ for a combined 21.9%.

In later lines, subtype-specific differentiation is partially attenuated and broadly applicable regimen categories recur across groups. Chemotherapy-only becomes the top category in HR−/HER2− and HR+/HER2−, while biologic-only and chemo+biologic dominate the HER2+ cohorts; HR+ disease still retains substantial endocrine-based regimens.

2.2 Model performance

We compared five model families spanning proportional hazards (PH) and non-PH assumptions and linear, tree-based, and deep learning approaches under nested cross-validation (nested CV). Model output is line-specific PFS probability curve over a 2-year horizon. A gradient-boosted survival model with CoxPH loss and regression

trees as base learner from scikit-survival (GBSA [25]) achieves the near-best global C-index and mean AUC, as well as best global IBS and horizon-specific AUC (Table 2). We therefore select GBSA for downstream analysis.

Time-dependent performance varies more by evaluation horizon than by model class (Fig. 2a). Cumulative/dynamic AUC (cdAUC) increases with horizon, improving by ≈ 0.1 across models over the first year and then flattening in the second year (additional improvement ≈ 0.02). In contrast, the IPCW C-index decreases over time, with a notable decline (≈ 0.01) in the first 180 days followed by a relatively stable plateau. Across both metrics, DeepHit shows an additional decline in the final 180 days of follow-up, suggesting less stable long-term performance and is thus not considered for further analysis despite achieving the best global C-index and mean AUC. Full per-horizon IPCW-C, AUC and Brier scores are reported in Extended Data Table 6; subtype-stratified curves exhibit wider uncertainty in less prevalent groups (Extended Data Fig. 6).

2.3 Accurate risk stratification despite heavily front-loaded event distribution

The population Kaplan–Meier (KM) curve computed from ground-truth PFS labels (red solid line; Fig. 2b) shows frequent early progression and rapid risk-set attrition, with 1,412 (16.1%) mLoTs at risk at 1 year and 552 (6.3%) at 730 d ($\approx$2 years). The mean predicted survival curve (red dashed line; Fig. 2b), obtained by averaging out-of-fold (OOF) GBSA-predicted survival, closely tracks the KM (mean absolute error (MAE) = 0.008), corresponding to an average absolute difference of 0.8% in probability over time. Because follow-up is sparse in the tail, we report restricted mean survival time (RMST) truncated at 365 days (i.e., the area under the survival curve up to 1 year). The population RMST expects an average of 167 progression-free days for an mLoT within the first year, while mean predicted RMST is only 3 days lower. Low MAE and RMST difference together suggest predictions that capture overall progression dynamics at the population level.

Stratifying the population by predicted 365-day PFS probability yields marked prognostic separation. The high-risk tertile has a KM RMST of 102 d and only 2% of mLoTs remain at risk 1 year after line start, compared with 244 d and 35% in the low-risk tertile. Calibration remains broadly consistent within each tertile, with the largest deviations observed in selected low-risk strata. MAEs are 0.030, 0.011, and 0.009 in the low-, mid-, and high-risk groups, respectively.

We also perform tertile stratification using within-group thresholds for disease subtypes and line groups (mLoT1 vs. mLoT2+), as shown in Fig. 3. This yields equally sized strata within each subgroup, highlights within-group heterogeneity in predicted risk, and facilitates more balanced downstream comparisons of baseline characteristics within each subgroup. Kaplan–Meier curves and mean predicted survival curves separate consistently across tertiles, indicating substantial differences in both predicted and observed outcomes within each subgroup defined by model-derived risk scores. This separation is also reflected in the RMST difference between high- and low-risk tertiles, ranging from 100 d in TNBC to 145 d in HR+/HER2−. At the same time, prognosis differs markedly across cohorts overall. For example, the low-risk

tertiles of both HER2+ cohorts have KM-based RMST values above 230 d, compared with 185 d in TNBC; similarly, the low-risk mLoT1 group reaches 285 d versus 205 d in low-risk mLoT2+. Together, these results show that the model output score not only distinguishes high- and low-risk patients at the population level, but also offers clear stratification within sub-cohorts, enabling characterization of within-cohort heterogeneity.

Despite substantial differences in cohort size, with the smallest cohort being HR−/HER2+ (437 mLoTs) and the largest HR+/HER2− (6,130 mLoTs), agreement between predicted and observed outcomes remains broadly similar across subgroup-specific tertiles. Across risk subgroups, MAE ≤ 0.05 except for the low-risk tertile of HR−/HER2+ and the low-risk tertile of mLoT1. Additionally, ΔRMST $\leq 11\,d$ in all non-low-risk cohorts and remains below $20\,d$ in the low-risk tertiles. The somewhat larger deviations in low-risk tertiles likely reflect longer survival tails, sparser follow-up, and greater within-tertile heterogeneity, all of which leave more room for absolute error in RMST and survival probability estimates. Further discussion of factors that lead to the observed miscalibration is provided in Section 3.3.

2.4 Interpretability analyses highlight clinically coherent risk drivers

Shapley additive explanations (SHAP) [26] on GBSA output of progression probability by 365 days (Fig. 4a) identify dominant predictive features. At population level, signals clearly driving higher progression probability include historic and recent radiology evidence of tumor presence, administration of less frequently used chemotherapy agents, tumor-marker kinetics, ECOG, and TP53 genomic mutation. Lower progression probability drivers include a line being the first metastatic line, absence of genomic sequencing data (likely proxying earlier or less heavily characterized disease in this cohort), and patient age. In line-specific explanations, a low-risk example (Fig. 4b) is associated mostly with being the first metastatic line, the absence of disease burden in recent radiology report, line-start targeted therapy, and low CA15–3 kinetics; conversely, an mLoT2+ high-risk example (Fig. 4c) is associated with recent liver metastasis, chemotherapy exposure, high CEA and CA15–3 kinetics, and ECOG status.

To test whether these signals are potential drivers of within-cohort risk stratification, we computed Cohen's d between cohort-centered baseline feature values between the high- and low-risk tertiles (Extended Data Table 7; positive value indicates higher average feature values in the high-risk tertile), linking feature magnitude to stratification within each subtype or line group. Across cohorts, metastatic burden (especially liver involvement) and first-line/genomic-availability markers remain prominent. In HR+ cohorts and mLoT2+, tumor-marker burden (especially CA15–3) remains among the strongest positive contrasts, whereas mLoT1 contrasts are comparatively dominated by tumor-marker availability and TP53 status. Cancer-reported imaging states are more prominent in the HR−/HER2+ disease. Overall, these results connect global attribution to cohort-specific stratification: the same burden and documentation signals that dominate SHAP are those that are systematically elevated in the high-risk

tertiles, while the precise ranking varies across cohorts, consistent with within-cohort heterogeneity.

Additionally, group-level SHAP attribution contrast highlights feature groups that drive stratification (Extended Data Table 8). Across cohorts, higher-risk lines tend to devote relatively more attribution (positive values) to ECOG, tumor-site burden, or genomics/treatment groups, whereas imaging-coverage and measurement-cadence groups more often account for a larger share in low-risk strata.

2.5 Model robustness to missing modalities, leakage, and non-biological signals

Removing any single clinical modality (definitions in Extended Data Section A.3.1) results in only modest performance degradation, suggesting redundancy in information. Notably, ablating surveillance-pattern features has negligible impact on performance, consistent with their relatively small contribution ($\leq 15\%$) to overall SHAP attribution (Extended Data Fig. 7).

Among single-modality models, imaging-derived features capturing disease burden and metastatic sites provide the strongest standalone discrimination, with planned regimen features performing similarly. By contrast, genomics and clinicopathologic features provide weaker standalone signal.

Including the full set of treatments received during an mLoT as feature set introduces an information-leakage pathway. In particular mLoTs with treatment or agents initiated > 28 days after line start exhibit substantially different PFS distributions from those without late-start agents (Fig. 5c, dashed lines). Models trained without late-start agent masking produce mean predicted curves that more closely match the corresponding KM estimates for both subsets and yield consistent performance improvements over treatment-masked models ($\Delta C = 0.010$ and $\Delta IBS = -0.004$ for GBSA; Extended Data Tables 9 and 10). We interpret these gains as a bias demonstration rather than evidence for a preferable deployment model, because they rely on post-initiation information unavailable at the treatment decision point. To avoid this leakage, we mask treatment features to include only those initiated within 28 days of line start in all of our analysis except for Fig. 5c.

3 Discussion

Metastatic lines of therapy planning require individualized clinical decisions, ideally coupled with forecasted line-level benefit and near-term progression risk. Later-line palliative efficacy is difficult to quantify in real-world data because treatment selection and outcomes are tightly coupled to evolving disease burden, prior therapy history, and care pathways, limiting the feasibility of controlled study designs. We address this with an auditable EHR pipeline that reconstructs mLoT and corresponding outcomes in routine practice, and then trains machine learning models to predict PFS using multimodal feature snapshots at line start. In our set, modeling a deliberately heterogeneous real-world cohort—rather than fragmented subtype- or regimen-specific subsets—yields broadly calibrated risk stratification from model output within disease subtypes and lines despite strong cohort imbalance. Risk attribution is dominated by

recent disease burden rather than baseline diagnosis or subtype labels, suggesting that the model leverages clinically meaningful metastatic-stage state and captures shared signals beyond subtype-specific labels.

Our main contributions are: (i) a transparent algorithmic module for mLoT mining and PFS labeling; (ii) a multimodal machine learning framework providing interpretable, broadly calibrated risk stratification and prognostication robust to modality ablation. These may support evidence enrichment for metastatic-line PFS patterns and provide a basis for supervisable patient risk stratification and line-level prognostication in routine practice.

3.1 Clinically coherent cohort construction

We prioritize label fidelity and auditability over maximal sample size in constructing the study cohort, using heuristic filters to reduce bias from incomplete documentation and imperfect data capture. Clinical validity of the derived mLoTs is supported by concordance with subtype-specific standards of care. In mLoT1, regimen backbones align with guideline-consistent practice: chemotherapy-dominant treatment in TNBC, HER2-targeted therapy (often combined with chemotherapy) in HER2+ disease, and endocrine therapy with optional targeted agents in HR+/HER2− disease [7]. In mLoT2+, broadly applicable regimen classes (chemo-only, biologic-only, and chemo+biologic) recur across subtypes, consistent with later-line escalation, convergence toward palliative backbones, and increasing use of treatments not restricted to a single biomarker-defined group.

At population level, outcome patterns align with clinical expectations: prognosis is better in earlier lines and poorer in TNBC relative to other subtypes. However, treatment efficacy is not explained by cohort difference alone due to confounders and cohort composition. For example, contrasts in median mLoT1 PFS across HER2+ groups may co-vary with the number of mLoTs received, event rates, sample size, and shifts in regimen mix. Furthermore, two cohort properties motivate joint modeling across subtypes and lines. First, within each subtype—especially HR+ disease—agent-level regimens exhibit a long tail beyond a few dominant combinations, indicating substantial heterogeneity in real-world treatment. Second, regimen classes recur across subtypes, particularly in later lines, creating shared structure that can be leveraged while preserving subtype-specific effects. Taken together, these considerations suggest that efficacy quantification may benefit from multimodal models that integrate patient-, disease-, and treatment-trajectory features to account for heterogeneity and differential outcomes.

3.2 Extensive model performance evaluation

We performed extensive model selection using nested cross-validation with complementary discrimination and calibration metrics, identifying GBSA as a strong and reasonably well-calibrated model for downstream analyses. Under the proportional hazards (PH) assumption [27], individuals share a common baseline hazard and differ through time-invariant covariate effects, so a single risk score computed at line

start can summarize relative risk for patient profiling. This aligns with our predominantly static, tabular covariates and with the strong performance of the regularized Cox model, suggesting that the dominant signal is approximately linear with stable effect directions. The early concentration of progression events further favors models that prioritize features separating patients at highest near-term risk.

Time-dependent performance trends appear driven more by data and feature constraints than by model class. The decline of IPCW C-index is consistent with diminishing prognostic value of baseline features as time from line initiation increases, and with the front-loaded population event distribution where most learnable signal resides. Late-horizon performance is further challenged by fewer events per discrete time interval, increasing IPCW weight penalty for incorrect ranking. cdAUC increases with horizon because longer horizons better separate sustained non-progressors from lines that progress within the evaluation window. At short horizons (e.g., $t = 90$ days), small timing differences among high-risk lines can fall on either side of the case/control cutoff and increase variance; by $t = 365$ days, the cutoff more cleanly separates lines that progress within a year from those that remain progression-free beyond a year, resulting in good performance.

3.3 Risk stratification with good overall calibration

Across the overall cohort and within each subtype, GBSA produces calibrated risk strata with clear within-group prognostic separation and close agreement between predicted survival curves and Kaplan–Meier estimates (low MAE and small differences in 1-year RMST). Stratification remains strong despite substantial subgroup size imbalance, supporting the value of training a single model on a heterogeneous cohort with larger effective sample sizes and extracting shared predictive signals. Because separation is observed across heterogeneous real-world regimens, the score can serve as a baseline prognosis at line start that is comparatively robust to variation in local treatment mix. Clinically, higher-risk patients could be prioritized for earlier imaging, closer follow-up, and proactive planning for subsequent-line options, whereas lower-risk patients may avoid unnecessary escalation when clinically appropriate. The same stratification can support trial enrichment (e.g., identifying patients likely to progress early) and benchmarking within-subtype outcomes against calibrated expectations.

Calibration was not perfect in all risk strata, although the deviations were limited in magnitude and concentrated in a small subset of groups, particularly some low-risk tertiles. We consider this pattern plausible for two reasons. First, calibration is reported here after tertile-based stratification, a common presentation in the clinical risk-stratification literature and one that is also aligned with our downstream comparisons of baseline characteristics across risk groups. Because tertiles discretize a continuous risk score, they can obscure within-bin heterogeneity, especially when the score distribution is skewed, and may therefore accentuate apparent misalignment between average predicted and observed outcomes in some strata. Second, intrinsic real-world care pattern and observational bias may also contribute. In our cohort, low-risk tertiles were more often enriched for mLoT1, less genomic characterization, and greater biomarker missingness. This likely reflects real-world care patterns in which

patients with more indolent disease often undergo less intensive surveillance and accumulate fewer documented events or measurements. As a result, predictions in these groups may be based on sparser information, which can make calibration more challenging. We therefore interpret the observed deviations as arising from the interaction between the modeling framework and the limitations of real-world observational data and care patterns, rather than as evidence of severe model failure. In future work, leakage-safe summaries of pre-metastatic trajectory and prior treatment history may improve calibration in mLoT1 while preserving the decision-point validity of the score.

3.4 Risk driver analysis

The model can also support clinical review by highlighting major contributors to predicted risk. This relies on a feature design principle of *decision-point validity*: covariates are restricted to information credibly available at mLoT initiation. During feature engineering, we coarsely encoded care intensity and surveillance patterns for highly prognostic signals (e.g., radiology and lab measurement) to motivate reliance on biological signals. Consistent with this aim, features designed as surveillance-pattern proxies have minimal impact when ablated and account for only a small fraction of attribution mass.

We also found that including the complete set of drug agents administered within an mLoT introduces a leakage pathway that improves apparent predictive performance and agreement with observed curves. Presumably, these features implicitly encode post-line information such as drug switches due to intolerance or early signs of progression. They also induce immortal-time-related bias, in that longer PFS mechanically permits more drug administrations. To preserve decision-point validity, we apply masking and time-windowing as a conservative design choice.

As a result, line-level attributions yield clinically coherent risk drivers and provide an audit layer for plausibility checks and hypothesis generation. Top attributions concentrate on established prognostic domains (metastatic burden, genomics, and tumor-marker kinetics), and their monotonic directions are biologically sensible (e.g., liver involvement or TP53 alterations should not reduce progression risk). Subtype and initial diagnosis are not among the top contributors, implying that the model relies more on recent disease-burden signals. A plausible explanation is that by mLoT2+, much of the biologic subtype signal has already been "acted upon" through prior regimen selection and earlier outcomes, and the remaining variation in near-term progression risk is increasingly governed by current metastatic burden, marker trajectories, and context- and eligibility-driven treatment choices. Clinically, such line-specific attribution shifts can improve the interpretability of risk scores at the decision point: they suggest which baseline signals are most informative to review at different treatment stages.

Interpretation of some features, however, needs to be put in context. First, the model is not designed to estimate counterfactual treatment effects, and observed associations between treatment features and progression risk are likely confounded by indication and other factors. Treatment-feature attributions in this observational setting should be interpreted in the context of clinician selection, eligibility, and disease context rather than as causal treatment effects. In particular, associations such as

higher risk with rare chemotherapy or lower risk with targeted/biologic therapy likely reflect confounding by indication and eligibility differences rather than counterfactual benefit. Similarly, the feature indicating first metastatic line should be interpreted in context: it captures the clinical pattern that later lines tend to perform worse and likely also reflects the fact that patients in their first metastatic line have not yet experienced progression on prior metastatic therapies, rather than a protective effect of first-line status per se.

Second, missingness in genomic features is associated with lower predicted progression risk, which likely reflects care patterns and disease trajectory: patients with genomic data are more likely to have received prior lines, undergone more intensive evaluation, or had more aggressive disease, whereas absence of such data may proxy for earlier-line disease or less intensive care. We retained this signal to preserve decision-point validity, because genomic data are often unavailable at line start in real-world practice and the model must make predictions from information available at that time. In real-world practice, the biological effect of genomic alterations is difficult to separate from the clinical processes that determine whether and when testing occurs, especially given potential time differences between sample acquisition, sequencing, and treatment initiation. Importantly, our ablation results suggest that this proxy does not materially drive performance: removing genomic features caused only negligible performance degradation, while a genomics-only model performed poorly.

Another example of contextual interpretation is age: higher age associates with lower predicted risk (Fig. 4a), which is unlikely to be protective biologically. This pattern plausibly reflects cohort composition (older patients enriched for HR-positive and HER2-negative disease; Extended Data Table 3b) and/or more conservative care patterns that affect derived PFS. This may create a transportability hazard if age, subtype composition, or treatment-selection patterns differ in deployment settings. In our setting, removing age does not lead to a statistically significant degradation in model performance for GBSA using threshold of 0.05 (Extended Data Tables 11, and 12).

With these caveats, model explanations can support decision-making by making risk drivers explicit at the moment treatment lines begin, and surface hypotheses for follow-up (e.g., confirmatory imaging or marker re-evaluation when kinetics dominate). Furthermore, risk-score contribution analysis, in combination with risk stratification, yields a short list of top features that drive heterogeneity both within and across subtypes and suggest clinically meaningful sub-phenotypes. Lastly, reliance distributed across multiple modalities implies that useful predictions can remain obtainable when particular data streams are missing, as commonly occurs with care transitions and variable testing. Such robustness may broaden applicability beyond highly curated subsets and can provide a baseline for assessing how differences in data availability or practice patterns could bias downstream analyses.

3.5 Limitations and outlooks

This study has several limitations. First, PFS labels in the study cohort are derived from routine-care documentation depending on surveillance intensity and care pathways specific to the contributing institution. Prior work has shown that comparability

of real-world PFS to trial-based PFS endpoints is limited [28, 29] due to measurement cadence and progression ascertainment. Accordingly, RMST estimates may not be directly comparable to survival estimates from protocol-defined clinical trials. Evaluation was restricted to nested cross-validation within the single-site MSK-CHORD cohort. This design was chosen to provide an approximately unbiased estimate of performance in the presence of hyperparameter tuning and model selection; however, it remains an internal validation procedure and cannot establish transportability. To our knowledge, no open-access dataset currently matches the combination of longitudinal treatment history, progression labels, and genomic sequencing information required for a like-for-like external validation. An important next step will be to assess out-of-site performance and generalizability. Although the fitted model and its learned weights may not be directly transportable across institutions, the general modeling framework and the feature patterns associated with risk in this study may still inform similar analyses in other settings. Accordingly, the reported score should be interpreted as a baseline prognostic model for real-world PFS in this source cohort, not as a transportable treatment recommender or a surrogate predictor of overall survival.

Second, clinically important details needed to determine treatment-switch rationale and progression status are not available for every line. Heuristic mLoT segmentation from structured EHR records is limited by data granularity, often missing drug-switch rationale, surgical intent, and peri-metastatic context, while NLP-derived radiology labels may miss full-report nuance and introduce noise in PFS timing. These can lead to imperfect mLoT classification and potential informative censoring, a common challenge in real-world oncology PFS analyses. In our setting, informative censoring may occur when an mLoT advances to a subsequent line without a qualifying progression label and is instead censored at the last validated non-progression assessment. If the treatment change was related to undocumented disease worsening, outcome of these lines may differ from those remaining under observation, potentially biasing PFS estimates [30, 31]. In our cohort, this risk is bounded above by 674/8,791 mLoTs (7.7%). Because this mechanism can arise only for lines assigned non-progression censoring, and because administrative censoring at 2 years is non-informative by design, only mLoTs censored before the 2-year cutoff may be affected by this bias. Moreover, some among these lines likely reflect non-progression-related treatment changes, such as toxicity, patient preference, or planned escalation, which would not constitute informative censoring. This limitation could be further addressed by higher resolution in the current routine-care data and calls for future work on richer data capture and modeling approaches.

Third, models summarizing risk at line-start do not benefit from additional within-line dynamics such as response, toxicity, dose changes, or interval imaging and biomarker measurements. This baseline-only design, together with incomplete treatment detail, represents mLoTs that incorporate late-start agents or step-up strategies with noticeably worse calibration. In such cases, clinically meaningful determinants of outcome occur after the modeled time origin, reducing informativeness. Addressing this would require time-updated covariates or line definitions that explicitly represent post-initiation adaptations, and a careful leakage-free design.

Future work could pursue two directions in parallel. First, learned mLoT segmentation methods that integrate structured therapy timing with radiology and clinical signals could reduce reliance on heuristics and improve boundary accuracy. Scalable weak supervision or clinician-in-the-loop audits could further provide training signal and ongoing quality control. More accurate mLoTs would reduce outcome-label noise and increase cohort size eligible for exploration. Leveraging fine-grained information may also promote transportability across sites with different documentation, surveillance, and treatment mixes.

Second, the adaptation of dynamic time-series models that could update risk within a line and across successive mLoTs at predefined, leakage-safe update times would considerably improve model performance. With target trial emulation and sensitivity analyses, the same infrastructure could be extended toward comparative effectiveness while mitigating immortal time and time-varying confounding. Higher-fidelity inputs (raw imaging, longitudinal labs/vitals, free-text notes, and richer molecular profiling) could improve prognostic resolution, especially in later-line heterogeneous disease. These extensions would broaden applicability beyond curated cohorts and support both calibrated stratification and more credible real-world treatment-effect studies across diseases.

4 Methods

4.1 Cohort construction

We included MSK-CHORD patients with breast cancer as the single primary malignancy who had palliative metastatic treatment records and radiology-derived progression assessments. Metastatic onset was approximated as the earliest of (i) the first mention of a distant metastatic site in radiology reports (excluding regional lymph nodes and nonspecific sites) or (ii) the diagnosis date for de novo stage IV disease.

4.2 mLoT segmentation

The segmentation algorithm uses systemic treatment records and NLP-annotated radiology progression assessments. For each patient, mLoT1 begins at the first systemic therapy start date after metastatic onset and advances according to the following rules, reflecting clinical conventions and oncologist input:

- An mLoT advances after a radiology-confirmed progression event that occurs at least 28 days after the current line start date;
- An mLoT advances after a new systemic therapy drug (not present in current mLoT) that starts at least 1 year after the current line start date;
- An mLoT does not advance after local treatment changes (e.g., bone-directed agents, surgery, radiation therapy);
- The next mLoT begins at the start date of the first subsequent systemic therapy.

All treatments administered between the start and end of an mLoT are considered part of the mLoT for downstream filtering. The 28-day threshold is a pragmatic compromise based on clinician input: it allows routine scheduling and assessment latency

around line start while avoiding attribution of an observed progression to a regimen with essentially no exposure. Comparable operational conventions for line-of-therapy analyses have also been used in other oncology real-world studies [15, 16, 32, 33]. An exception applies when a new systemic treatment is initiated within 28 days before the recorded progression that ends the current mLoT. In that case, it is assigned to start a new mLoT to avoid attributing progression to a regimen with insufficient exposure or radiology-assessment latency.

4.3 PFS labeling and censoring

For survival analysis, we label each segmented mLoT from above with a PFS(event, time) pair. The event label is binary, indicating whether the line ended with a progression event or was censored. The PFS time is defined as the number of days from line start to the label-defining event, and is not the time in between the start of two consecutive mLoTs. We use the following rule set to assign PFS labels:

- **Progression event:** an mLoT is labeled as progression if it is terminated by a qualifying radiology-confirmed progression event. The time between line start and the first such progression event is the PFS time;
- **Death event:** for the final mLoT, death without a prior documented progression is also counted as a progression event. The time between line start and death is the PFS time;
- **Right censoring:** if neither of the above occurs, the line is right-censored at the time of the last radiology-confirmed non-progression assessment. PFS time is the number of days from line start to that last non-progression assessment;
- **Exclusion:** lines without any follow-up radiology assessment are not assigned PFS labels and excluded from modeling.

Radiology-confirmed progression is determined by the NLP-annotated labels indicating presence of cancer progression in impression section of radiology reports, taken from the MSK-CHORD dataset. These labels are validated in the original MSK-CHORD study against manually curated labels and achieved AUC of 0.97, precision and recall of 0.9 [19] in identifying progression status. We therefore consider these labels as an upstream validated component and do not attempt to re-estimate its accuracy within the breast-cancer subset.

We additionally apply a pragmatic rule set to further strengthen mLoT regimen and PFS label fidelity. Constructed mLoT and PFS pairs are excluded if:

- No radiology assessment occurred during the mLoT;
- No radiology assessment occurred within 90 days prior to the start of the mLoT;
- Both chemotherapy and endocrine therapy were administered within the mLoT;
- HR-negative patients received endocrine therapy.

Among 2,294 QC exclusions, 2,114 involve concurrent chemotherapy and endocrine therapy within the same mLoT, and 261 involve endocrine therapy in HR-negative disease. Concurrent chemotherapy and endocrine therapy within the same mLoT were excluded based on clinician input, because these treatments are generally not administered together within a single metastatic line in breast cancer, making such overlap

more consistent with transition or segmentation ambiguity than a valid single mLoT. Excluding mLoTs without radiology assessments in the 90 days prior to line start removes a cohort with distributionally different PFS patterns (Extended Data Fig. 8) and enforces a minimal baseline disease-state anchor at initiation. This reduces bias from incomplete capture (e.g., outside-system care) that could distort PFS and induce model reliance on non-biological proxies. Together, these rule-based filters prevent contradictions with standard-of-care patterns and limit spurious correlations between documentation intensity, subtype labels, and hazard.

We also apply administrative censoring at 730 days (2 years) to mitigate the influence of long-term survivors with limited follow-up and to focus on clinically relevant risk stratification. In the final analytic cohort, 1,226 of 8,791 mLoTs are censored, with 674 lines censored after a documented non-progression assessment and 552 administratively censored at 730 days (Fig. 1b). This design, separating mLoT advancement and PFS labeling, ensures derived labels are valid within the modelled duration to the best of documented evidence, and limits the impact of noise in real-world data documentation and capture.

4.4 Feature extraction

Features are extracted and engineered from the MSK-CHORD dataset and are grouped into five modalities: clinicopathologic, radiology, tumor markers, treatment regimen, and genomics (Extended Data Section A.3.1). Clinicopathologic features summarize baseline demographics and disease characteristics. Radiology features are derived from radiology reports within 90-day windows prior to line start, capturing tumor presence in specific organs and cancer presence via imaging different body regions (several regions share the same report outcomes). Tumor-marker features summarize CA15–3 and CEA kinetics using various log-transformed aggregation statistics. Treatment features encode drug agents and local therapies administered within 28 days of line start, taking top 5 plus "other" category for each systemic treatment. Genomic features encode somatic mutations and copy-number aberrations in a curated gene list. To best ensure decision-point validity, all genomic features are masked for pre-sequencing lines, with an additional missingness flag to indicate that genomic profiling was not performed before line start. This enables the model to not mistake missing genomic features as wild-type status. Under this time-restricted encoding, genomic data are unavailable at line start for 3,010 of 8,791 modeled mLoTs, spanning 1,805 of 2,881 patients. Additionally, a flag indicating first metastatic line is included, to account for the established prior that treatment benefit tends to diminish across successive lines and that mLoT1 tends to perform better than later lines. Missing numeric measurements are imputed with 0 after standardization; missing binary presence indicators are imputed with 0, or as tri/quad state variables (e.g., positive/negative/unknown status). Missingness flags, last day of observation, and cancer-presence report status by imaged body region are additionally considered as a "surveillance patterns" modality.

Two design choices specifically mitigate information leakage and non-biological shortcuts. Because mLoTs are constructed retrospectively, naively encoding the full within-line treatment record risks inducing immortal time bias. Longer PFS mechanically permits additional drug administrations and regimen modifications, which may

reflect longer time at risk as well as post-baseline factors such as early progression signals, toxicity-driven switches, or patient preference. As shown in Results, leveraging late-start agents can improve predictive performance but yields predictions that are not interpretable as baseline prognosis. Accordingly, we mask drug agents initiated ≥ 28 days after line start to preserve interpretability as line-start prognosis.

4.4.1 Conservative feature engineering mitigates surveillance-pattern and history short-cutting

Because progression is observed through clinical surveillance, models can learn care-process proxies rather than underlying disease risk. For example, given diminishing benefit across successive mLoTs, the model may learn to associate heavier prior treatment with shorter subsequent PFS. We therefore omit explicit treatment-history features, under the assumption that the current regimen already reflects clinician synthesis of prior response and tolerability. We also suppress fine-grained surveillance by omitting measurement counts and timestamps in biomarker kinetics and radiology screening, except for the most recent day of observation and value. Surveillance pattern variables are evaluated in addition to clinical modalities in ablation experiments.

4.5 Model selection and training

We benchmarked the following survival models that span varying complexity, adherence to the PH assumption, and interpretability:

- (PH, linear) CoxPH [25].
- (PH, tree-based) Gradient-boosted survival model with CoxPH loss (GBSA).
- (Non-PH, tree-based) Random survival forest (RSF) [34].
- (PH, deep learning) DeepSurv [35].
- (Non-PH, deep learning) DeepHit [36].

This model set spans varying levels of complexity, adherence to the PH assumption, and interpretability. Code for model implementations and evaluation metrics is adapted from the framework and Python notebooks in [37].

We performed nested cross-validation (nested CV) with 5 outer folds for test evaluation and 3 inner folds for hyperparameter selection. This provides an unbiased estimate of held-out performance under hyperparameter exploration and model selection. For GBSA, RSF, DeepSurv, and DeepHit, we evaluated five preselected configurations (Extended Data Section A.3.2) spanning low, middle, and high capacity models with different regularization strengths. For CoxPH, we tuned the L2 regularization strength α over 30 log-spaced values between 10^{-1} and 10^{3}. Best-performing configuration candidates in each outer fold were those with average inner-fold C-index within 0.003 of the best-performing configuration. The final configuration is selected in favor of lower-complexity and more strongly regularized configurations. CV splits were stratified by HR/HER2 subtype and whether a patient has at least one mLoT censored at 2 years. All mLoTs from a given patient were assigned to the same split to prevent leakage. Detailed candidate grids are provided in Extended Data Section A.3.2, and

selected outer-fold configurations are summarized in Extended Data Table 16. Ablation experiments use the best configuration selected, restricting configurations to those selected at least once in the full-modality nested CV to reduce computational burden.

4.6 Evaluation

We harmonized each model's output to a survival probability curve on a common time grid over 2 years and evaluated performance on the held-out test set. Global metrics are Antolini's C [22] cumulative/dynamic AUC (cdAUC) [23]. Antolini's C considers all comparable patient pairs over follow-up and reports the fraction for which the patient with the shorter observed PFS receives higher predicted risk. cdAUC is computed at each time t by forming the time-specific risk set, pairing individuals who experience an event prior to time t with those who remain event-free past t, and estimating the probability that the model assigns higher risk to the former; inverse-probability weights correct for censoring. An average over the evaluation time range yields the reported single overall discrimination score. For time-dependent evaluation, we report IPCW C-index [38], cdAUC, and Brier score [24] at 30-day increments. To prevent degenerate inverse-probability-of-censoring weights at the administrative-censoring boundary, any PFS event occurring exactly at 730 d is shifted to 729.999 d (i.e., $730-0.001$). This negligible adjustment avoids censoring-survival estimates reaching 0 at the evaluation horizon and stabilizes IPCW-based metrics. For the same reason, IPCW-based metrics at the 2-year horizon are evaluated at $730-\varepsilon$ rather than exactly 730 d. When a nominal horizon (e.g., 730 d) is not exactly represented on a model's evaluation grid, we evaluate at the closest available time point and report it at the nominal horizon.

4.7 Risk stratification and Kaplan–Meier summaries

Risk stratification and interpretability analyses are performed on out-of-fold predictions from the primary model (GBSA). mLoTs are stratified into low/mid/high risk tertiles using the predicted 1-year progression probability, and we compare the mean predicted survival curve to the Kaplan–Meier estimate within each stratum. For proportional-hazards models, the risk ranking is time-invariant and thus the same tertile cutoffs apply across time horizons. We summarize the discrepancy using the mean absolute error (MAE) between predicted and observed survival curves over time; low MAE means the predicted survival probabilities differ from the corresponding Kaplan–Meier estimate by only a small amount on average. We additionally report RMST at 365 days, which informs the expected time free of progression or death within the first year of treatment.

4.8 SHAP contributions

We interpret GBSA predictions at the feature-group level using SHAP values computed from SHAP's PermutationExplainer on the predicted PFS probability *at* 365 days. Although models trained on different folds may learn different baseline hazards and risk-score calibrations, the predicted PFS probability at a fixed time point is directly comparable across folds. To intuitively correlate higher SHAP value output

with higher progression risk, we report SHAP contributions for $1 - \text{PFS}(365)$ (i.e., progression probability *by* 365 days).

4.8.1 Baseline-feature value contrasts

Let x_{ij} denote the value of baseline feature j at line initiation for line i. To report subtype- (or mLoT-) specific contrasts relative to the cohort baseline, we re-center each baseline feature within cohort G,

$$\tilde{x}_{ij} = x_{ij} - \frac{1}{|G|} \sum_{i \in G} x_{ij}. \tag{1}$$

We then quantify separation between risk strata using Cohen's d computed on $\tilde{x}_{ij}$ within cohort G:

$$d_j^{(x)}(G) = \frac{\bar{\tilde{x}}_{j,\mathrm{H}}(G) - \bar{\tilde{x}}_{j,\mathrm{L}}(G)}{\sigma_{p,j}(G)}. \tag{2}$$

Positive $d_j^{(x)}(G)$ indicates higher mean feature values in the high-risk tertile, while negative values indicate higher mean feature values in the low-risk tertile. We apply this same baseline-feature value contrast within each HR/HER2 subtype and within each mLoT stratum (mLoT1 and mLoT2+).

4.8.2 SHAP-based contrasts at group level

Let ϕ_{ij} denote the SHAP attribution for feature j on line i for the prediction $1 - \text{PFS}_i(365)$. We additionally map baseline features into clinically motivated feature groups (Extended Data Table 17) and summarize each group's contribution using a per-line relative-importance (share) score,

$$s_g(i) = \frac{\left|\sum_{j \in g} \phi_{ij}\right|}{\sum_{g'} \left|\sum_{j \in g'} \phi_{ij}\right|}, \tag{3}$$

where g denotes a feature group and the denominator sums absolute SHAP magnitudes across groups. This normalization yields the fraction of total absolute attribution assigned to group g for line i and avoids within-group sign cancellation. We compare these group shares directly between risk strata within cohort G (again without re-centering),

$$d_g^{(s)}(G) = \frac{\bar{s}_{g,\mathrm{H}}(G) - \bar{s}_{g,\mathrm{L}}(G)}{\sigma_{p,g}^{(s)}(G)}. \tag{4}$$

Positive $d_g^{(s)}(G)$ indicates that group g accounts for a larger share of total absolute attribution in the high-risk tertile; negative values indicate a larger share in the low-risk tertile.

4.8.3 Cohen's d and pooled variance.

For any scalar quantity $z(i)$ compared between H and L within cohort G (e.g., $\tilde{x}_{ij}$ or $s_g(i)$), we compute Cohen's d as

$$d = \frac{\bar{z}_{\mathrm{H}} - \bar{z}_{\mathrm{L}}}{\sigma_p}, \quad \sigma_p = \sqrt{\frac{(n_{\mathrm{H}} - 1)s_{\mathrm{H}}^2 + (n_{\mathrm{L}} - 1)s_{\mathrm{L}}^2}{n_{\mathrm{H}} + n_{\mathrm{L}} - 2}}, \tag{5}$$

where $\bar{z}_{\mathrm{H}}$ and $\bar{z}_{\mathrm{L}}$ are means in the high- and low-risk tertiles, s_{H}^2 and s_{L}^2 are their sample variances, and n_{H}, n_{L} are tertile sizes. Throughout, these attributions and contrasts are descriptive and non-causal in this observational setting.

5 Data Availability

The dataset analysed during the current study is the MSK-CHORD dataset [19], released by Memorial Sloan Kettering Cancer Center. It is available via cBioPortal under study identifier `msk_chord_2024`: https://www.cbioportal.org/study/summary?id=msk_chord_2024.

6 Code Availability

Code for data generation, modeling and analysis is available under https://github.com/uzh-dqbm-cmi/mbc-pfs.

7 Acknowledgements

Activities that contributed to this publication were conducted within the scope of the following grants:

- **The LOOP Zurich**: *AI Tumor Board: Personalized Decision Support Project in Precision Oncology.* The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.
- **Swiss National Science Foundation** (Grant No. 10003518): *Medical, Multilingual and Privacy-Preserving Natural Language Processing in the Clinical Domain.* The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

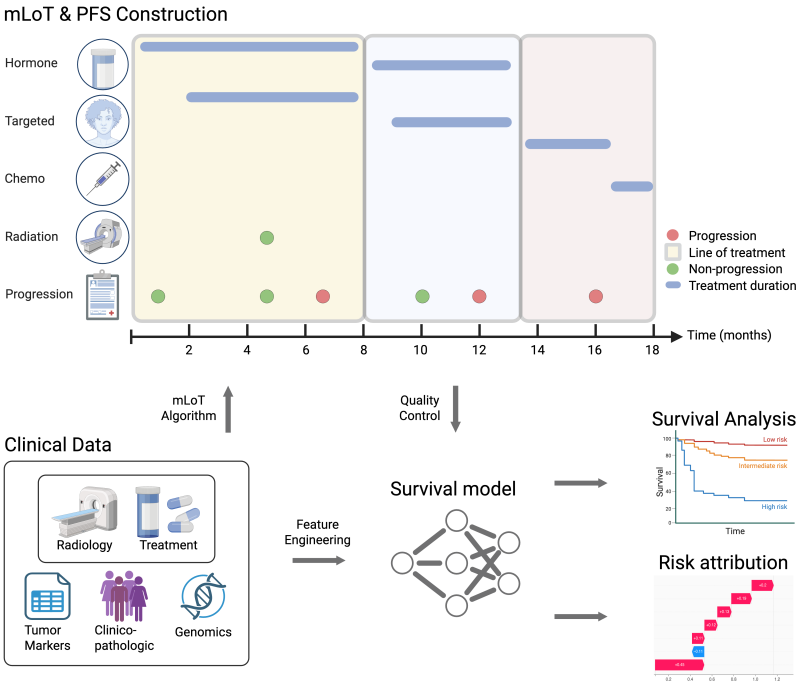
8 Competing interests

All authors declare no financial or non-financial competing interests.

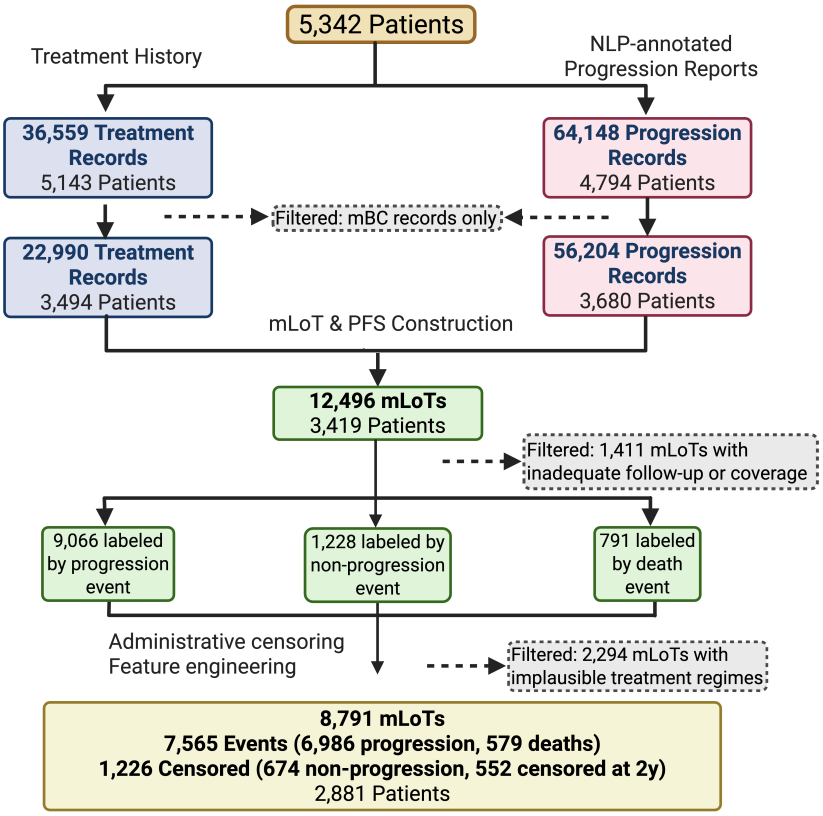
9 Author Contributions

X.Z., B.F., Z.B. and M.K. conceived the study. B.F. and M.K. supervised the study. B.F. advised on the machine learning modeling and feature set design. Z.B. and A.W.

advised on algorithm design, medical content, and clinical interpretation. X.Z. implemented the mLoT extraction algorithm, machine learning modeling, and analysis pipeline. T.N. ran modeling experiments on optimized model parameters, exploratory data analysis, and refined manuscript figures. X.Z. wrote the manuscript with feedback and guidance from all co-authors, who all contributed to the interpretation of results and refinement of the manuscript.

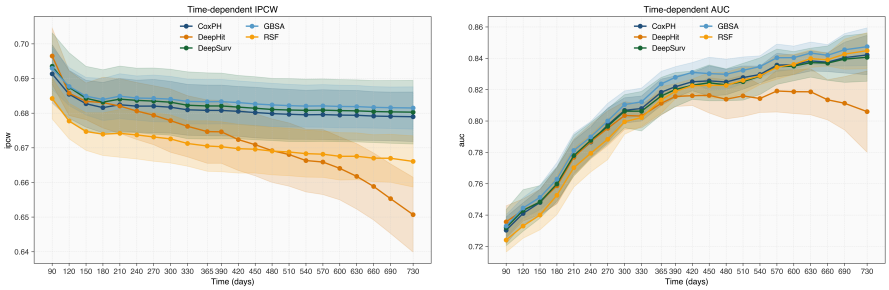


(a) Label extraction and survival modeling workflow.

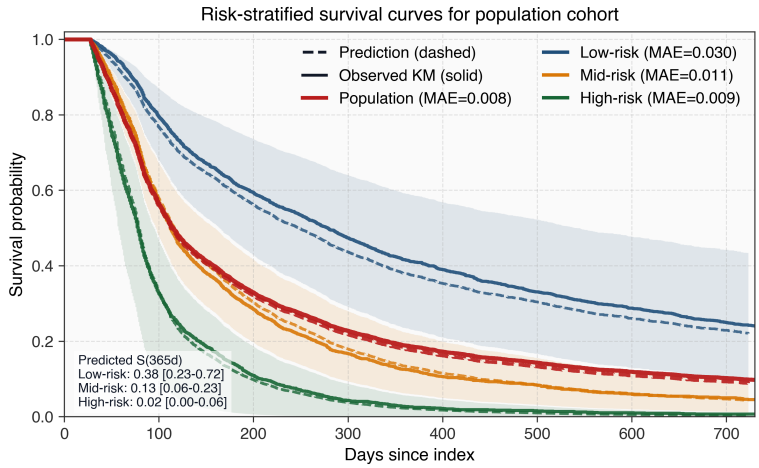


(b) Cohort construction and inclusion criteria.

Fig. 1: Workflow overview. **a**: The workflow combines NLP-annotated radiology reports and treatment history to segment mLoTs and assign PFS outcomes. Survival models are trained on five clinical modalities, with outputs evaluated for discrimination, calibration, interpretability, and robustness. **b**: Derivation of the analytic mLoT cohort and inclusion criteria. Both created with BioRender [20].



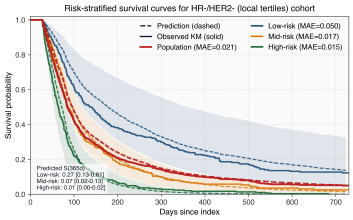
(a) Time-dependent discrimination.



Stratum	Count	180 d	365 d	730 d	KM RMST	ΔRMST
Population	8,791	3,025	1,412	552	167	-3
Low-risk	2,931	1,723	1,038	461	244	-10
Mid-risk	2,930	905	310	79	156	4
High-risk	2,930	397	64	12	102	-2

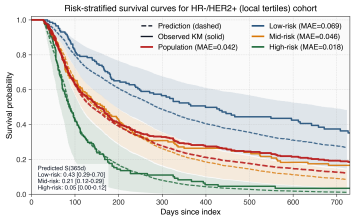
(b) Risk stratification and population calibration.

Fig. 2: Time-dependent model performance and population risk stratification. **a**: IPCW C-index (left) and cdAUC (right) across evaluation horizons (90 d–2 y) for all models. Curves are mean with shaded ±1 SD across nested CV outer fold performances. **b**: Disease subtype population (red) and tertile-stratified (blue, orange, and green) Kaplan–Meier (solid) and mean predicted survival (dashed). Shaded area indicates 95% predicted curves within the tertile. Table shows at-risk counts and RMST summaries. RMST@365: KM RMST with cutoff at 365 days. Δ RMST: predicted RMST - KM RMST. Left legend: range of survival probability at 365 days. Reported 2-year (730 d) summaries are evaluated at the closest available time point on the evaluation grid and labeled as 730 d.



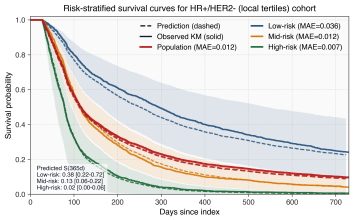
Stratum	Count	180 d	365 d	730 d	KM RMST	ΔRMST
Population	1,032	231	90	31	131	9
Low-risk	344	129	65	27	185	19
Mid-risk	344	73	19	4	125	7
High-risk	344	29	6	0	85	1

(a) HR−/HER2−



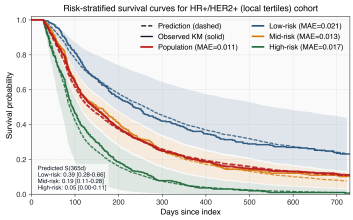
Stratum	Count	180 d	365 d	730 d	KM RMST	ΔRMST
Population	437	180	105	52	194	-9
Low-risk	146	92	61	30	266	-19
Mid-risk	146	65	35	19	193	-7
High-risk	145	23	9	3	124	-2

(b) HR−/HER2+



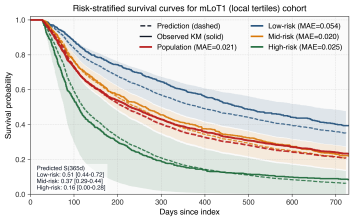
Stratum	Count	180 d	365 d	730 d	KM RMST	ΔRMST
Population	6,130	2,143	1,014	387	168	-4
Low-risk	2,044	1,247	769	331	247	-13
Mid-risk	2,043	626	202	48	156	3
High-risk	2,043	270	43	8	102	-3

(c) HR+/HER2−



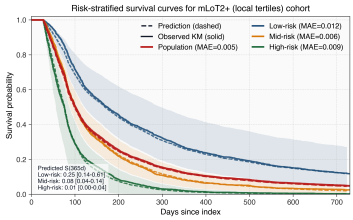
Stratum	Count	180 d	365 d	730 d	KM RMST	ΔRMST
Population	1,192	471	203	82	183	-3
Low-risk	398	213	112	54	236	2
Mid-risk	397	173	74	26	187	-5
High-risk	397	85	17	2	127	-8

(d) HR+/HER2+



Stratum	Count	180 d	365 d	730 d	KM RMST	ΔRMST
Population	2,342	1,239	766	361	226	-4
Low-risk	781	551	392	200	285	-18
Mid-risk	781	443	275	115	237	-5
High-risk	780	245	99	46	157	11

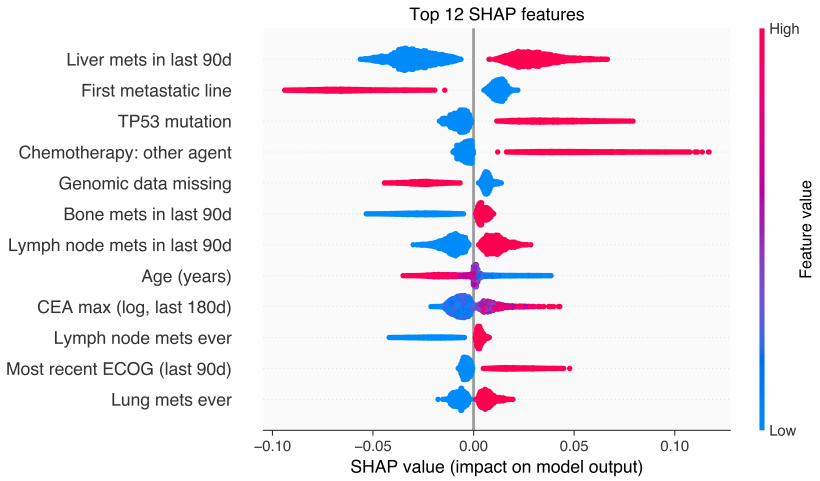
(e) mLoT1



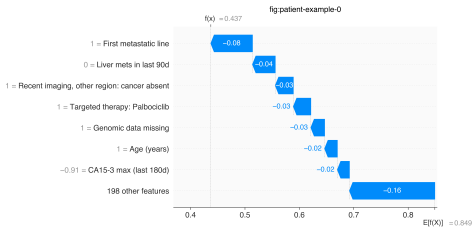
Stratum	Count	180 d	365 d	730 d	KM RMST	ΔRMST
Population	6,449	1,786	646	191	146	-3
Low-risk	2,150	1,009	472	154	205	-6
Mid-risk	2,150	538	140	33	138	1
High-risk	2,149	239	34	4	95	-3

(f) mLoT2+

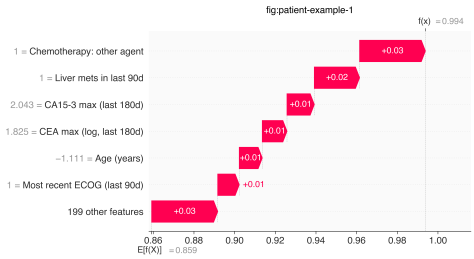
Fig. 3: Disease subtype (top/middle) and line-group (bottom) population (red) and tertile-stratified Kaplan–Meier (solid) and mean predicted survival (dashed). Shaded area indicates 95% predicted curves within the tertile. Table shows at-risk counts and RMST summaries. RMST@365: KM RMST with cutoff at 365 days. Δ RMST: predicted RMST - KM RMST. Left legend: range of survival probability at 365 days. Reported 2-year (730 d) summaries are evaluated at the closest available time point on the evaluation grid and labeled as 730 d.



(a) Top individual features.

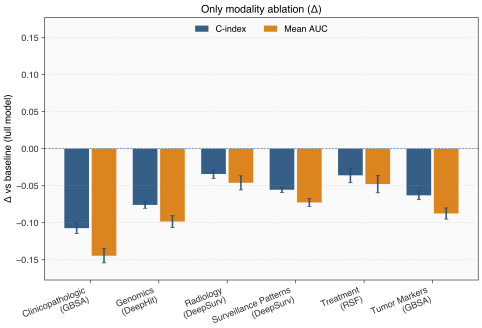


(b) Example mLoT1 top risk drivers.

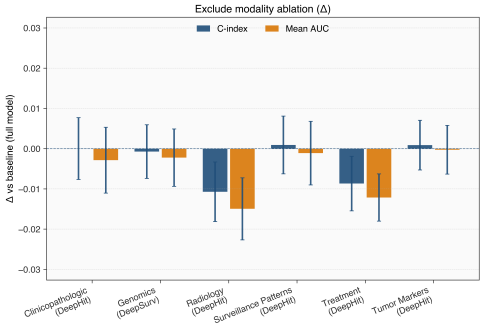


(c) Example mLoT2+ top risk drivers.

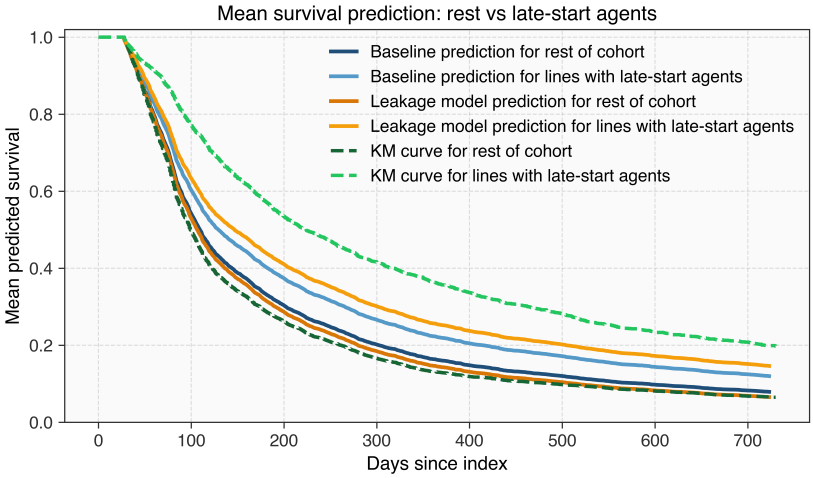
Fig. 4: Interpretability analysis for GBSA risk scores. **a** SHAP summary plot for baseline features showing the top 12 contributors to progression probability by 365 days; red indicates higher feature values and positive SHAP values increase progression probability. **b** and **c**: Waterfall plots illustrating line-level explanations for lines from low- and high-risk population tertile, respectively. $E[f(x)]$ denotes the cohort-mean progression probability and $f(x)$ denotes the line-specific model output.



(a) Single modality model performance.



(b) Ablated modality model performance.



(c) Late-start treatment leakage analysis.

Fig. 5: Modality ablations, leakage checks, and feature-group contributions. **a** and **b** Single-modality models and leave-one-modality-out ablations. Bars show the change in C-index and mean AUC relative to the full multimodal model (ablation minus full; negative indicates worse performance), with error bars denoting ±1 SD across outer folds of nested CV; the best-performing model class for each setting is shown in parentheses. **c** Observed and predicted survival for lines in which all therapies are initiated within 28 days of line start versus lines that include late-start therapies. Dashed lines indicate Kaplan–Meier estimates of PFS; solid lines show mean predicted survival from a standard treatment-masked model (blue) and a model trained without treatment masking (orange), included here as a leakage demonstration rather than a clinical candidate model.

References

[1] Cardoso, F. *et al.* 4th ESO–ESMO international consensus guidelines for advanced breast cancer (ABC 4). *Annals of Oncology* **29**, 1634–1657 (2018).

[2] Fietz, T., Tesch, H., Rauh, J. *et al.* Palliative systemic therapy and overall survival of 1,395 patients with advanced breast cancer - results from the prospective german tmk cohort study. *The Breast* **34**, 122–130 (2017).

[3] U.S. Food and Drug Administration. Clinical trial endpoints for the approval of cancer drugs and biologics: Guidance for industry. Guidance for Industry (2018). Final guidance.

[4] Cardoso, F. *et al.* International guidelines for management of metastatic breast cancer: Combination vs sequential single-agent chemotherapy. *Journal of the National Cancer Institute* **101**, 1174–1181 (2009).

[5] Cottu, P. *et al.* Attrition between lines of therapy and real-world outcomes of patients with HER2-positive metastatic breast cancer in Europe: a cohort study leveraging electronic medical records. *Breast Cancer Research and Treatment* **209**, 419–430 (2025). Epub 2024 Oct 20.

[6] Basmadjian, R. B. *et al.* Retrospective database analysis to assess treatment patterns, clinical outcomes, and healthcare resource utilization in metastatic or recurrent her2-positive breast cancer in alberta. *Cancer Treatment and Research Communications* **45**, 101021 (2025).

[7] Gennari, A., André, F., Barrios, C. H. *et al.* Esmo clinical practice guideline for the diagnosis, staging and treatment of patients with metastatic breast cancer. *Annals of Oncology* **32**, 1475–1495 (2021).

[8] Kent, D. M. *et al.* The predictive approaches to treatment effect heterogeneity (path) statement. *Annals of Internal Medicine* **172**, 35–45 (2020). Epub 2019 Nov 12.

[9] Fountzilas, E., Pearce, T., Baysal, M. A., Chakraborty, A. & Tsimberidou, A. M. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. *npj Digital Medicine* **8**, 75 (2025).

[10] Varma, G., Yenukoti, R. K., Kumar, P. & et al. A deep learning-enabled workflow to estimate real-world progression-free survival in patients with metastatic breast cancer: study using deidentified electronic health records. *JMIR Cancer* **11**, e64697 (2025).

[11] Merzhevich, T. *et al.* Bellazzi, R., José Manuel Juarez Herrero, Sacchi, L. & Blaž Zupan (eds) *Machine learning predictions of overall and progression-free survival in advanced breast cancer.* (eds Bellazzi, R., José Manuel Juarez Herrero, Sacchi,

L. & Blaž Zupan) *Artificial Intelligence in Medicine: 23rd International Conference, AIME 2025, Pavia, Italy, June 23–26, 2025, Proceedings, Part II*, Vol. 15735 of *Lecture Notes in Computer Science*, 267–271 (Springer, Cham, 2025).

[12] Cabel, L. *et al.* Outcome beyond third-line chemotherapy for metastatic triple-negative breast cancer in the french esme program. *The Breast* **56**, 18–25 (2021).

[13] Saini, K. S. & Twelves, C. Determining lines of therapy in patients with solid cancers: a proposed new systematic and comprehensive framework. *British Journal of Cancer* **125**, 155–163 (2021).

[14] Meng, W., Ou, W., Black, W. *et al.* Temporal phenotyping by mining healthcare data to derive lines of therapy for cancer. *Journal of Biomedical Informatics* **100**, 103335 (2019).

[15] Meng, W., Mosesso, K. M., Lane, K. A. *et al.* An automated line-of-therapy algorithm for adults with metastatic non-small cell lung cancer: Validation study using blinded manual chart review. *JMIR Medical Informatics* **9**, e29017 (2021).

[16] Grady, C. B., Hwang, W.-T., Reuss, J. E. *et al.* Determining line of therapy from real-world data in non-small cell lung cancer. *Pharmacoepidemiology and Drug Safety* **33**, e70049 (2024).

[17] Kehl, K. L., Elmarakeby, H., Nishino, M. *et al.* Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA Oncology* **5**, 1421–1429 (2019).

[18] Kehl, K. L., Xu, W., Lepisto, E. *et al.* Natural language processing to ascertain cancer outcomes from medical oncologist notes. *JCO Clinical Cancer Informatics* **4**, 680–690 (2020).

[19] Jee, J. *et al.* Automated real-world data integration improves cancer outcome prediction. *Nature* **636**, 728–736 (2024).

[20] BioRender.com. Figures created with biorender. https://BioRender.com/56dcnu2 (2026). Accessed 11 Feb 2026.

[21] National Cancer Institute (NCI), Surveillance, Epidemiology, and End Results (SEER) Program. Cancer stat facts: Female breast cancer subtypes. https://seer.cancer.gov/statfacts/html/breast-subtypes.html. Accessed 11 Jan 2026.

[22] Antolini, L., Boracchi, P. & Biganzoli, E. A time-dependent discrimination index for survival data. *Statistics in Medicine* **24**, 3927–3944 (2005).

[23] Hung, H. & Chiang, C.-T. Estimation methods for time-dependent auc models with survival data. *Canadian Journal of Statistics* **38**, 8–26 (2010).

[24] Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545 (1999).

[25] Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research* **21**, 1–6 (2020).

[26] Lundberg, S. M. & Lee, S.-I. Guyon, I. *et al.* (eds) *A unified approach to interpreting model predictions.* (eds Guyon, I. *et al.*) *Advances in Neural Information Processing Systems*, Vol. 30, 4765–4774 (Curran Associates, Inc., Red Hook, NY, 2017).

[27] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–220 (1972).

[28] Panageas, K. S., Ben-Porat, L., Dickler, M. N., Chapman, P. B. & Schrag, D. When you look matters: The effect of assessment schedule on progression-free survival. *JNCI: Journal of the National Cancer Institute* **99**, 428–432 (2007).

[29] Ackerman, B. *et al.* Measurement error and bias in real-world oncology endpoints when constructing external control arms. *Frontiers in Drug Safety and Regulation* **Volume 4 - 2024** (2024).

[30] Fleming, T. R., Rothmann, M. D. & Lu, H. L. Issues in using progression-free survival when evaluating oncology products. *Journal of Clinical Oncology* **27**, 2874–2880 (2009).

[31] Campigotto, F. & Weller, E. Impact of informative censoring on the kaplan-meier estimate of progression-free survival in phase ii clinical trials. *Journal of Clinical Oncology* **32**, 3068–3074 (2014).

[32] Bains, S., Kalsekar, A., Amiri, K. I. & Weiss, J. Outcomes and line of therapy patterns among patients with advanced non-small cell lung cancer (ansclc) treated with programmed cell death protein 1/programmed cell death ligand 1 (pd-(l)1) inhibitors and chemotherapy in second-line or later. *JTO Clinical and Research Reports* **3**, 100275 (2022).

[33] Barzi, A. *et al.* Real-world dosing patterns and outcomes of patients with metastatic pancreatic cancer treated with a liposomal irinotecan regimen in the united states. *Pancreas* **49**, 193–200 (2020).

[34] Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals of Applied Statistics* **2**, 841–860 (2008).

[35] Katzman, J. L. *et al.* Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology* **18**, 24 (2018).

[36] Lee, C., Zame, W. R., Yoon, J. & van der Schaar, M. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence* **32**, 2326–2333 (2018).

[37] Chen, G. H. An introduction to deep survival analysis models for predicting time-to-event outcomes (2024). arXiv:2410.01086.

[38] Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L.-J. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**, 1105–1117 (2011). Epub 2011 Jan 13.