npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02783-3

Handling missing modalities in multimodal survival prediction for non-small cell lung cancer

Check for updates

Ruffini Filippo[1,2,12] ✉, Camillo Maria Caruso[1,12], Claudia Tacconi[3], Lorenzo Nibid[4,5], Francesca Miccolis[6], Marta Lovino[6], Carlo Greco[3,7], Edy Ippolito[3,7], Michele Fiore[3,7], Alessio Cortellini[8,9], Bruno Beomonte Zobel[10], Giuseppe Perrone[4,5], Bruno Vincenzi[8,9], Claudio Marrocco[11], Alessandro Bria[11], Elisa Ficarra[6], Sara Ramella[3,7,13], Valerio Guarrasi[1,13] ✉ & Paolo Soda[1,2,13] ✉

Accurate survival prediction in non-small cell lung cancer (NSCLC) requires integrating clinical, radiological, and histopathological data. Multimodal deep learning (MDL) can improve precision prognosis, but small cohorts and missing modalities limit its clinical applicability, as conventional approaches enforce complete-case filtering or imputation. We present a missing-aware multimodal survival framework that combines computed tomography (CT), whole-slide histopathology images (WSI), and structured clinical variables for overall survival modeling in unresectable stage II–III NSCLC. The framework uses foundation models (FMs) for modality-specific feature extraction and a missing-aware encoding strategy that enables intermediate multimodal fusion under naturally incomplete modality profiles. By design, the architecture processes all available data without dropping patients during training or inference. Intermediate fusion outperforms unimodal baselines and both early and late fusion strategies, with the trimodal configuration reaching a C-index of 74.42. Modality-importance analyses show that the fusion model adapts its reliance on each data stream according to representation informativeness, shaped by the alignment between FM pretraining objectives and the survival task. The learned risk scores produce clinically meaningful stratification of disease progression and metastatic risk, with statistically significant log-rank tests across all modality combinations, supporting the translational relevance of the proposed framework.

Lung cancer persists as a leading cause of cancer-related mortality worldwide, with non-small cell lung cancer (NSCLC) accounting for ~85% of all diagnosed cases[1]. The advent of targeted therapy and immunotherapy has markedly reshaped the therapeutic landscape of NSCLC[2]. However, the variability in individual treatment effectiveness makes robust risk stratification a pressing clinical need. Accurate risk estimation is supportive of critical therapeutic choices[3], such as selecting the appropriate treatment modality, adjusting treatment intensity, and defining follow-up strategies, according to each patient's expected clinical trajectory. Therapeutic outcomes for NSCLC might not be fully attributable to a single tumor biomarker or image-based diagnostic technique. This means that understanding a patient's prognostic risk requires information drawn from complementary sources[4,5].

Histopathology, radiology, genomic information, and electronic health records (EHR) provide multi-scale views of tumor morphology, molecular profile, and patient health status[6]. Histopathology, obtained from digitalized whole slide images (WSI) of the biopsy examination, remains the gold standard for diagnosis, providing high-resolution insight into the microscopic morphology of pathological tissue. This examination holds significant prognostic value by revealing the tumor microenvironment and immune infiltration patterns, which are key determinants for survival. Yet, due to the invasive and focal nature of biopsies, they frequently fail to capture the complete spatial variability of the tumor.[7]. Radiological imaging provides a non-invasive, macroscopic view of the entire tumor burden. In NSCLC, this role is primarily fulfilled by computed tomography (CT). Although CT scans overcome the spatially localized and invasive nature of biopsies, they provide a limited survival signal when used in isolation to predict disease outcome[8]. Molecular sequencing provides genomic information that characterizes the tumor's underlying genetic landscape. Although the identification of oncogenic drivers and mutations yields

A full list of affiliations appears at the end of the paper. ✉e-mail: filippo.ruffini@unicampus.it; valerio.guarrasi@unicampus.it; paolo.soda@umu.se

critical predictive biomarkers for targeted therapy, genomics alone does not capture the physical properties of the pathology. Finally, EHRs aggregate the essential patients' clinical context, ranging from physiological parameters and comorbidities to treatment history, into structured tabular data. These clinical variables are indispensable for linking biological findings to actual survival outcomes and monitoring disease progression[9,10]. To enable robust modeling and harmonization across patients, a common practice is to process EHR data, aggregating key physiological variables, comorbidities, and treatment history, uniforming them into structured tabular formats.

Faced with this intricate tangle of multi-source information, Artificial Intelligence (AI) research in medicine has converged toward multimodal deep learning (MDL), where multiple modalities are fused to capture a holistic view of the biological mechanisms underlying the disease[11]. In NSCLC, prognosis is shaped by factors spanning multiple biological scales, from cellular-level tissue alterations to macroscopic tumor characteristics and broader clinical indicators. MDL aims to capture the latent non-linear relationships between these complementary views, i.e., WSIs, CT scans, and clinical data, translating their joint signal into more accurate survival estimation and supporting the transition toward personalized cancer care[12].

The integration of these complementary data streams enables clinicians to move beyond a static snapshot of disease status and more accurately anticipate disease progression and patient outcomes. To translate this multimodal perspective into actionable predictions, prognostic estimation can be formulated as an overall survival (os) prediction task[13], the standard quantitative framework for estimating patient outcomes[14]. Unlike binary classifiers, which collapse prognosis into a static mortality label, OS modeling uses time-to-event analysis to estimate survival probabilities over time. Handling of censored observations, i.e., patients who are lost to follow-up or who remain event-free at analysis, is a key strength of this formulation, thus ensuring that all available longitudinal information contributes to estimate the survival prediction[15].

From these survival functions, OS models generate continuous, patient-specific risk scores that quantify individual treatment trajectories. These scores are then used to stratify patients into clinically meaningful risk groups, typically via median or quartile thresholds, to distinguish individuals with aggressive disease trajectories from those who are likely to experience more indolent courses[16]. Such stratification is central to precision oncology, informing decisions on therapeutic intensity, surveillance frequency, and eligibility for targeted or experimental treatments.

Historically, survival modeling has been largely driven by machine learning (ML) approaches, most notably the cox proportional hazards (CPH) model and random survival forests (RSF)[17]. While effective on structured, low-dimensional clinical variables, these models are inherently limited in their ability to operate directly on high-dimensional data such as CT volumes or WSIs. This limitation has motivated the adoption of MDL as a more suitable framework for accurate and clinically actionable survival estimation in NSCLC[16,18–20]. Despite the clear advantages of MDL, its applicability in real clinical scenarios remains fundamentally constrained by the limited availability of comprehensive multimodal datasets. In oncological settings, two pervasive structural limitations systematically hinder the robustness and generalizability of MDL models: restricted cohort sizes and the occurrence of partially missing or fragmented modalities[5,21]. First, the inherent scarcity of labeled data combined with the heterogeneity of medical datasets creates a high risk of overfitting, particularly when training imaging encoders, such as convolutional neural networks (CNNs), from scratch on high-dimensional inputs. To overcome this bottleneck, the field has increasingly adopted foundation models (FMs), whose large-scale pretraining enables them to capture generalist representations of the underlying data distribution[22]. This allows the extraction of high-level, semantically rich features from each input modality, reducing the dependency on large task-specific labeled datasets.

In healthcare, FMs have demonstrated remarkable robustness in high-variance clinical settings[23,24]. While their utility as feature extractors is well-validated for diagnostic tasks across both CT and WSI images[25–30], their application to prognosis prediction remains underexplored[31]. Unlike static diagnosis, deploying these generalist representations for survival analysis requires validating whether they capture the temporal and prognostic signals hidden within the tumor.

Second, medical data are inherently incomplete, with missingness occurring at the level of both individual variables and entire diagnostic modalities[32,33]. Clinical tabular data may be incomplete due to non-response, data entry errors, patient attrition, data corruption, or systematic omissions linked to specific clinical pathways. At the modality level, CT scans may be unavailable for patients treated at external institutions, and histopathology slides may be absent when biopsies are not performed or are technically unsuccessful. Consequently, real-world cohorts rarely contain a complete and uniform set of measurements across all patients.

On these grounds, this work presents a missing-aware multimodal framework for OS prediction in NSCLC, validated on an internal cohort of unresectable stage II–III patients undergoing radical chemoradiotherapy. A central aspect of this study is the exploration of Foundation Models as feature extractors for both CT volumes and WSIs within a small clinical cohort, investigating whether generalist representations acquired through large-scale pretraining can effectively substitute for task-specific encoder training in data-scarce oncological settings. Building on this premise, the proposed framework comprises three stages: (i) FM-based unimodal feature extraction for CT and WSI modalities, mapping raw imaging data into high-level representations without requiring cohort-specific fine-tuning of vision encoders; (ii) missing-aware representation learning, where each modality is encoded through a transformer-based architecture with an adaptive masking mechanism that enables the network to structurally ignore unavailable inputs; and (iii) an intermediate multimodal fusion strategy, in which the modality-specific latent representations are concatenated and processed by a multimodal head, capturing cross-modal interactions and producing a continuous, patient-specific hazard estimate. The framework is resilient to missing modalities by design, allowing both training and inference on incomplete patient profiles without imputation or sample exclusion. Additionally, we performed a systematic modality-importance analysis by progressively masking individual data streams at test time and monitoring the corresponding impact on survival prediction, providing an interpretable view of modality relevance and clinically actionable insights on model reliability.

Results

This section presents the results of the proposed three-stage, missing-aware multimodal learning framework. We report performance across unimodal, bimodal, and trimodal configurations, with a focus on robustness under incomplete patient profiles.

Cohort characteristics

Our study is based on a retrospective cohort of 179 patients with unresectable Stage IIA, IIB, IIIA, IIIB, and IIIC NSCLC treated with radical Chemo-Radiotherapy. Tables 1 and 2 present a univariate analysis of baseline cohort characteristics in relation to 5-year overall survival status. Statistical significance was evaluated using one-way analysis of variance (ANOVA) for continuous variables and the χ^2 test for categorical variables.

Patients are stratified into Survivors ($n = 80$) and Non-survivors ($n = 99$) based on their 5-year survival status. While the primary endpoint is time-to-event overall survival, this univariate stratification serves as an exploratory characterization of the cohort, offering an interpretable summary of baseline differences and confirming that the observed covariate trends align with the survival outcome.

Across demographic variables, the mean age of the cohort was 69.8 ± 12.0 years, with similar distributions in the Survivor (69.6 ± 11.3) and Non-survivor (70.0 ± 12.5) groups ($p = 0.820$). A significant difference was observed for sex ($p = 0.039$), with males representing 78/99 (78.8%) of non-survivors compared to 51/80 (63.7%) of survivors, indicating a statistically significant association between sex and worse survival outcomes. Stage at diagnosis was similarly distributed ($p = 0.540$), with stage IIIA observed in 48/99 non-survivors and 36/80 survivors, and Stage IIIB in 41/99 and 30/80

Table 1 | Cohort characteristics stratified by survival status

Characteristic	Overall (n = 179)	Non-survivor (n = 99)	Survivor (n = 80)	P value
Age (years)	69.8 ± 12.0	70.0 ± 12.5	69.6 ± 11.3	0.820
Sex				0.039*
Female	50 (27.93%)	21 (21.2%)	29 (36.2%)	
Male	129 (72.07%)	78 (78.8%)	51 (63.7%)	
Weight (kg)	70.3 ± 13.9	72.4 ± 14.0	68.1 ± 13.5	0.160
Height (cm)	168.3 ± 8.9	169.8 ± 9.1	166.7 ± 8.5	0.112
NRS	0.3 ± 1.3	0.4 ± 1.5	0.2 ± 1.0	0.419
Smoking habitus				0.521
Ex-smoker	43 (53.75%)	23 (56.1%)	20 (51.3%)	
Smoker	28 (35.00%)	15 (36.6%)	13 (33.3%)	
Non-smoker	9 (11.25%)	3 (7.3%)	6 (15.4%)	
Cigarettes/day	22.4 ± 18.7	23.8 ± 19.5	21.0 ± 18.2	0.547
Stage at diagnosis				0.540
IIA	2 (1.12%)	1 (1.0%)	1 (1.2%)	
IIB	5 (2.79%)	3 (3.0%)	2 (2.5%)	
IIIA	84 (46.93%)	48 (48.5%)	36 (45.0%)	
IIIB	71 (39.66%)	41 (41.4%)	30 (37.5%)	
IIIC	17 (9.50%)	6 (6.1%)	11 (13.8%)	
Histology				0.331
Adenocarcinoma	90 (50.28%)	48 (48.5%)	42 (52.5%)	
Squamous carcinoma	80 (44.69%)	44 (44.4%)	36 (45.0%)	
NOS	5 (2.79%)	3 (3.0%)	2 (2.5%)	
Other	4 (2.23%)	4 (4.0%)	0 (0.0%)	
Comorbidity 1				0.479
Vascular	48 (64.00%)	25 (65.8%)	23 (62.2%)	
Metabolic	12 (16.00%)	4 (10.5%)	8 (21.6%)	
Renal	2 (2.67%)	2 (5.3%)	0 (0.0%)	
Other	11 (14.67%)	6 (15.8%)	5 (13.5%)	
Pulmonary	2 (2.67%)	1 (2.6%)	1 (2.7%)	
Comorbidity 2				0.756
Vascular	6 (11.76%)	2 (7.7%)	4 (16.0%)	
Metabolic	26 (50.98%)	15 (57.7%)	11 (44.0%)	
Renal	3 (5.88%)	2 (7.7%)	1 (4.0%)	
Pulmonary	5 (9.80%)	2 (7.7%)	3 (12.0%)	
Other	11 (21.57%)	5 (19.2%)	6 (24.0%)	
Comorbidity 3				0.253
Vascular	1 (6.67%)	0 (0.0%)	1 (12.5%)	
Metabolic	1 (6.67%)	1 (14.3%)	0 (0.0%)	
Renal	2 (13.33%)	0 (0.0%)	2 (25.0%)	
Other	10 (66.67%)	6 (85.7%)	4 (50.0%)	
Pulmonary	1 (6.67%)	0 (0.0%)	1 (12.5%)	
ECOG PS				0.101
0	60 (70.59%)	26 (60.5%)	34 (81.0%)	
1	23 (27.06%)	16 (37.2%)	7 (16.7%)	
2	2 (2.35%)	1 (2.3%)	1 (2.4%)	

Continuous variables are presented as mean ± standard deviation, and categorical variables as n (%). P values were calculated using one-way analysis of variance for continuous variables and chi-square test for categorical variables. *RT* Radiotherapy, *Ch.T* Chemotherapy. Statistical significance is indicated as follows: $p < 0.05$ (*), $p < 0.005$ (**), and $p < 0.001$ (***).

patients, respectively. Histological subtype stratification were comparable between the groups (p = 0.331), with the majority of cases identified as adenocarcinoma (48/99 for non survivor compared to 42/80 for survivor) and squamous carcinoma (44/99 and 36/80). Similar patterns were observed for renal, pulmonary, and other comorbidity categories. For example, vascular comorbidities were reported in 25/38 (65.8%) non-survivors and 23/37 (62.2%) survivors (p = 0.479), and metabolic comorbidities in 15/26 (57.7%) and 11/25 (44.0%) patients (p = 0.756). Among molecular biomarkers, EGFR mutation status differed between groups (p = 0.033), with positive cases found exclusively among Survivors (7.5% compared to 0% in Non-survivors). This difference highlights the known favorable impact of EGFR mutations, likely attributable to the distinct biology of this subtype and the availability of effective targeted therapies[34]. Moreover, a highly significant variation was also observed in radiotherapy techniques (p < 0.001), reflecting the survival advantage associated with modern delivery methods. Survivors were more frequently treated with VMAT (60.3%), whereas usage of 3D-CRT was nearly three times higher in the non-survivor group (43.2% vs. 16.2%), suggesting an association between older conformal techniques and poorer prognostic outcomes.

The administration of adjuvant immunotherapy was significantly associated with improved survival (p = 0.010). Its higher prevalence among Survivors (32.5% vs. 15.2%) reflects the benefit of consolidation checkpoint inhibition in Stage III NSCLC, reinforcing that access to contemporary systemic therapy remains a major prognostic determinant in this setting[35].

However, a dichotomy emerged when analyzing the intermediate clinical endpoints, highlighting that mortality is fundamentally driven by the disease's trajectory rather than baseline demographics. Progression-free survival (PFS) events were strongly correlated with mortality ($p < 0.001^{***}$): the vast majority of non-survivors (78.8%) experienced a progression event compared to less than half of the survivors (45.0%). This pattern was further reinforced by the significantly higher rate of distant metastasis (DM) in the non-survivor group (64.6% vs. 33.8%, $p < 0.001^{***}$), indicating systemic failure as a hallmark of poor prognosis. These findings confirm that while the two groups were demographically and clinically balanced at baseline, their outcomes were determined by the distinct biological aggressiveness of the tumor, manifesting as therapeutic resistance (PFS events) and systemic spread (DM). We examine in detail the impact of progression events on model performance in the following results sections. Finally, no statistically relevant differences were observed for the remaining baseline characteristics. Variables such as age (p = 0.820), TNM stage (p = 0.540), histology (p = 0.331), and ECOG performance status (p = 0.101) were evenly distributed between survivors and non-survivors groups.

Missing modalities and tabular data sparsity

The cohort in this study is characterised by significant data heterogeneity, presenting variable levels of missingness and feature sparsity. As displayed in Fig. 1, data availability is unevenly distributed across acquisition modalities and within the tabular clinical domain itself. At the modality level (Fig. 1a), WSI constitutes the primary bottleneck, being available only for 33.5% of the cohort (n = 60). In contrast, radiological imaging (CT) and the tabular modality show near-complete coverage, with only 2.8% of cases lacking a scan and 0% lacking the collection of at least the basic information. Nevertheless, the intersection of modalities remains narrow: 22.9% of patients (n = 41) possess the full trimodal profile (CT + WSI + Tabular). By contrast, radiological imaging (CT) and clinical data (Tabular) demonstrate near-complete availability, with only 2.8% of cases missing a CT scan and no cases lacking baseline patient information.

However, full modality-level availability in the Tabular modality does not imply that its constituent clinical features have been fully collected. Figure 1b shows that the structured clinical data, despite being technically present for all patients, exhibits substantial internal sparsity. Several clinically relevant molecular biomarkers, such as EGFR, ALK, and MET mutation status, are absent in 73.7%, 73.2%, and 96.6% of patients, respectively. The absence of these molecular features reflects a structural limitation of standard clinical workflows, where molecular testing is not

Table 2 | (Table 1 Continued) Cohort characteristics stratified by survival status

Characteristic	Overall (*n* = 179)	Non-survivor (*n* = 99)	Survivor (*n* = 80)	*P* value
EGFR mutation				0.033*
Negative	41 (87.23%)	23 (100.0%)	18 (75.0%)	
Positive	6 (12.77%)	0 (0.0%)	6 (25.0%)	
ALK rearrangement				0.233
Negative	45 (93.75%)	24 (100.0%)	21 (87.5%)	
ositive	3 (6.25%)	0 (0.0%)	3 (12.5%)	
MET alteration				1.000
Negative	4 (66.67%)	1 (100.0%)	3 (60.0%)	
Positive	2 (33.33%)	0 (0.0%)	2 (40.0%)	
PD-L1 (%)	0.2 ± 0.3	0.2 ± 0.3	0.2 ± 0.3	0.776
RT dose (Gy)	59.1 ± 8.9	58.3 ± 10.3	60.1 ± 6.8	0.176
Number of fractions	28.6 ± 5.3	27.9 ± 6.3	29.3 ± 3.8	0.211
RT technique				< 0.001***
3DCRT	43 (30.28%)	32 (43.2%)	11 (16.2%)	
VMAT	80 (56.34%)	39 (52.7%)	41 (60.3%)	
Mixed	6 (4.23%)	2 (2.7%)	4 (5.9%)	
RA	13 (9.15%)	1 (1.4%)	12 (17.6%)	
Induction Ch.T				0.465
CC	13 (43.33%)	7 (50.0%)	6 (37.5%)	
gem	1 (3.33%)	0 (0.0%)	1 (6.2%)	
C	15 (50.00%)	6 (42.9%)	9 (56.2%)	
CC-gem	1 (3.33%)	1 (7.1%)	0 (0.0%)	
Concomitant Ch.T				0.351
CC	58 (44.96%)	31 (42.5%)	27 (48.2%)	
Gem.	13 (10.08%)	8 (11.0%)	5 (8.9%)	
C	8 (6.20%)	3 (4.1%)	5 (8.9%)	
CC+Gem	35 (27.13%)	22 (30.1%)	13 (23.2%)	
AI.	2 (1.55%)	0 (0.0%)	2 (3.6%)	
other	13 (10.08%)	9 (12.3%)	4 (7.1%)	
RT (days)	48.7 ± 12.2	48.3 ± 11.5	49.1 ± 13.0	0.695
Suspension days	2.9 ± 7.6	3.8 ± 9.9	2.0 ± 4.2	0.279
Adjuvant immunotherapy				0.010*
No	138 (77.09%)	84 (84.8%)	54 (67.5%)	
Yes	41 (22.91%)	15 (15.2%)	26 (32.5%)	
Esophageal toxicity				0.183
Grade 0	57 (67.86%)	31 (72.1%)	26 (63.4%)	
Grade > 0	27 (32.14%)	12 (27.9%)	15 (36.6%)	
Pulmonary toxicity				0.250
Grade 0	75 (88.24%)	36 (83.7%)	39 (92.9%)	
Grade > 0	10 (11.76%)	7 (16.3%)	3 (7.1%)	
Hemoglobin toxicity				0.276
Grade 0	76 (89.41%)	36 (83.7%)	40 (95.2%)	
Grade > 0	9 (10.59%)	7 (16.3%)	2 (4.8%)	
Neutrophil toxicity				0.709
Grade 0	72 (84.71%)	37 (86.0%)	35 (83.3%)	
Grade > 0	13 (15.29%)	6 (14.0%)	7 (16.7%)	
Platelet toxicity				0.440
Grade 0	68 (80.00%)	36 (83.7%)	32 (76.2%)	
Grade > 0	17 (20.00%)	7 (16.3%)	10 (23.8%)	
PFS event				< 0.001***
Censored	65 (36.31%)	21 (21.2%)	44 (55.0%)	
Uncensored	114 (63.69%)	78 (78.8%)	36 (45.0%)	
DM event				< 0.001***
Censored	88 (49.16%)	35 (35.4%)	53 (66.2%)	
Uncensored	91 (50.84%)	64 (64.6%)	27 (33.8%)	

Continuous variables are presented as mean ± standard deviation, and categorical variables as *n* (%). *P* values were calculated using one-way analysis of variance for continuous variables and chi-square test for categorical variables. *RT* RadioTherapy, *Ch.T* ChemoTherapy, *PFS* progression-free survival, *DM* distant metastasis. Statistical significance is indicated as follows: $p < 0.05$ (*), $p < 0.005$ (**), and $p < 0.001$ (***).

systematically performed, and retrospective retrieval of these results is often challenging or unfeasible. Laboratory and histopathological descriptors (e.g., PD-L1 expression) demonstrate analogous levels of incompleteness for analogous reasons. Beyond genomic and histopathological descriptors, clinical information, such as comorbidities, is often incompletely documented, leading to substantial under-reporting of cardiovascular, metabolic, and respiratory conditions. Cross-modal and intra-modal sparsity have direct consequences for learning. Conventional multimodal architectures enforce a complete-case design: all modalities must be present, or the patient is imputed or excluded.

In our cohort, this constraint would remove 77.1% of cases (Fig. 1a), comprising patients with incomplete diagnostic workups, comorbidities, or limited access to comprehensive staging procedures. Critically, missing modalities may reflect disease severity itself. Patients with the most aggressive trajectories are plausibly those for whom full staging was never completed, meaning their exclusion would systematically deplete the cohort of its hardest and most informative cases. Selecting only complete cases introduces systematic bias and erodes both sample size and model generalizability. Our architecture admits the full cohort by design, handling missing modalities and clinical features without imputation or exclusion.

Missing-aware multimodal survival modeling

To support multimodal survival prediction under realistic clinical missingness, we propose a three-stage framework that decouples FM feature extraction, missing-aware unimodal encoding, and intermediate multimodal fusion. Figure 2 illustrates the complete pipeline, and a comprehensive methodological presentation is provided in "Methods".

Step-1 performs modality-specific pre-processing and feature extraction through functions $\Phi_i(\cdot)$, mapping each raw input x_i to a fixed embedding x_i^*. Φ_1 denotes the tabular pre-processing pipeline, Φ_2 encodes CT volumes via CT-FM[30], and Φ_3 implements a multi-stage WSI pipeline of dedicated foundation models for tissue segmentation, patch encoding, and multi-slide aggregation.

Step-2 transforms each x_i^* through a modality-specific encoder $F_i(\cdot)$ with missing-aware tokenization and masked self-attention, yielding a latent representation h_i conditioned only on observed inputs. Each h_i is then processed by an ODST layer $E_i(\cdot)$ and a fully connected projection to produce a unimodal risk estimate.

Step-3 implements CONCATODST, the proposed intermediate fusion model. The unimodal encoders are loaded from Step 2 and kept frozen. The latent vectors h_i are concatenated ($\oplus$) into a unified representation h_m, passed through the multimodal module $E_m(\cdot)$, composed of a

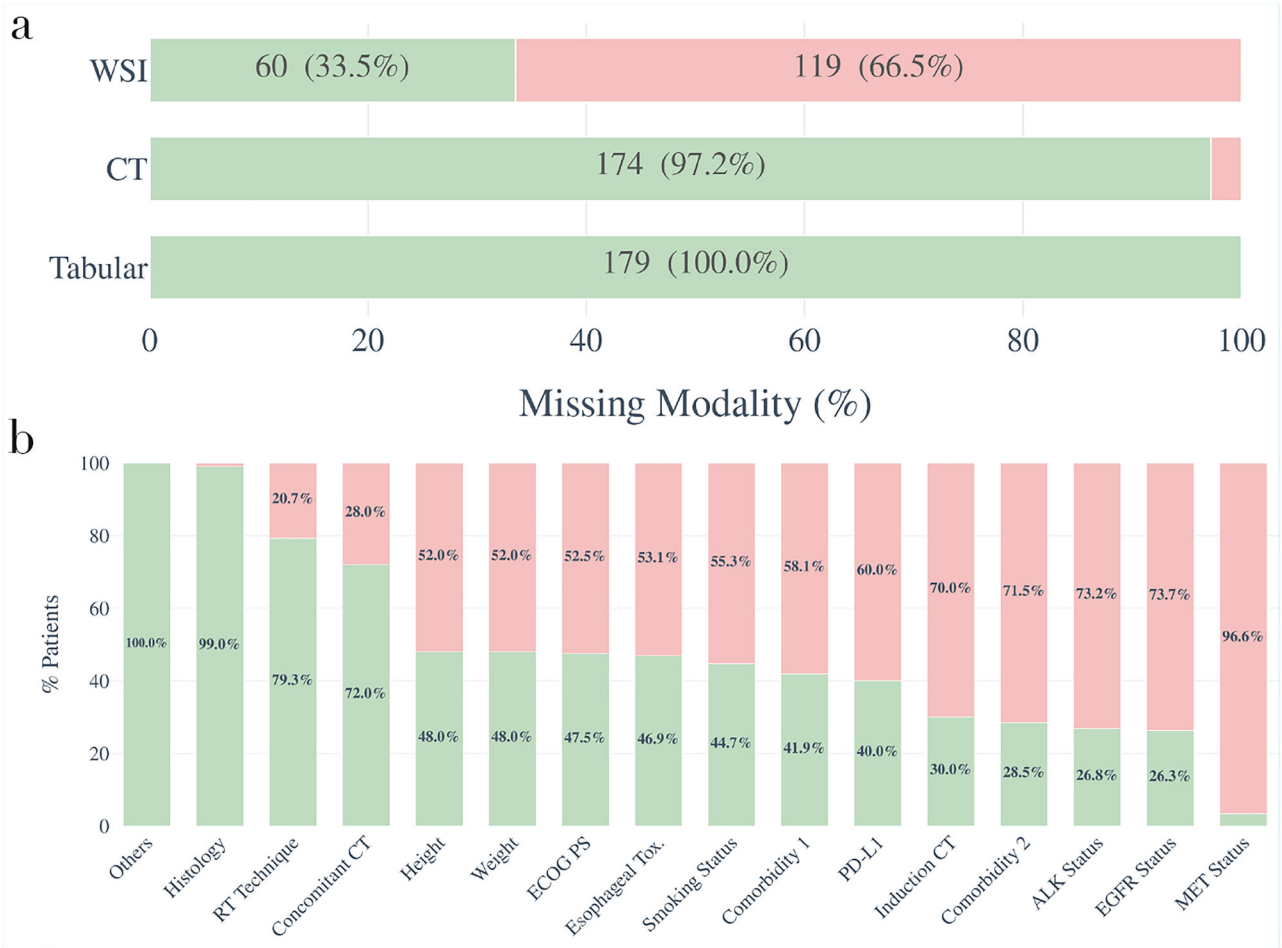


Fig. 1 | Overview of modality availability and feature-level sparsity within the cohort. Green indicates the proportion of available data, whereas red denotes missingness. **a** reports the availability of each modality (WSI, CT, and tabular data) across patients. **b** illustrates feature-wise missingness within the tabular modality, highlighting the proportion of absent values for each clinical variable.

fully connected layer downscaling to 1/3 of the input dimension followed by an ODST layer, and a final fully connected layer that generates the multimodal survival risk prediction y. The ablation study on the choice of $E_m(\cdot)$ is detailed in the supplementary material (Section S.1). The comparison study for the selection of the CT-FM foundation model for the CT modality is detailed in the supplementary material (Section S.3)

Prognostic separability in FMs' representation spaces

To enable unimodal representation learning under limited cohort size, we extracted imaging features from pretrained foundation models (FMs). For WSI, the multi-instance nature of WSI images requires a two-stage pipeline: patch-level feature extraction followed by aggregation into a single slide-level embedding. Therefore, we combined CLAM-based patch aggregation with MI-Zero vision encoders and the TITAN framework[25,26,36], producing a 768-dimensional representation. For CT volumes, vision transformer-based FMs pretrained on large-scale volumetric datasets expose a class token that serves directly as a compressed scan-level embedding, reducing extraction to a single forward pass. Selecting among available CT-specific FMs was less immediate, as available FMs for CT differ substantially in pretraining strategy and output dimensionality. We therefore conducted a comparative analysis across CT-CLIP[29], MERLIN[27], and CT-FM[30]; the comparison is detailed in the supplementary material (Section S.3). Based on this comparison, we selected CT-FM[30], which outputs a 512-dimensional token per volume. For completeness, tabular variables were included after imputing missing values using median imputation for continuous features and k-nearest neighbors for categorical features.

Using these embeddings, we ran a UMAP-based representation analysis to probe the prognostic signal encoded by each modality (Fig. 3). This representation analysis serves as a preliminary, mortality-based assessment rather than a direct reflection of the time-to-event task. Patients are stratified into Survivors and Non-Survivors by 5-year status, mirroring the cohort characterization in "Cohort characteristics". WSI embeddings (Fig. 3a) show locally coherent substructures, likely reflecting histological tissue organization, yet without statistically significant outcome-level separation (p = 0.2097). The unsupervised histopathological representations encode meaningful morphological patterns but do not linearly discriminate prognostic groups, plausibly due to high intra-class heterogeneity and the limited number of available WSIs for this study. Differently, CT embeddings (Fig. 3b) exhibit a statistically robust linear separation between Survivors (green) and Non-Survivors (red), confirmed by a significant group-separation test (p = 0.0421). The tabular feature space (Fig. 3c) shows high variance and no evident outcome-group separation (p = 0.2952), confirming that clinical variables alone are insufficient to stratify patients in this cohort. Taken together, these observations motivate a multimodal integration strategy, as none of the three data streams, when considered in isolation, provides a fully reliable prognostic separation, yet they capture complementary signals that may become synergistic when jointly modeled. In the following section, we move beyond exploratory mortality-based analyses and quantitatively

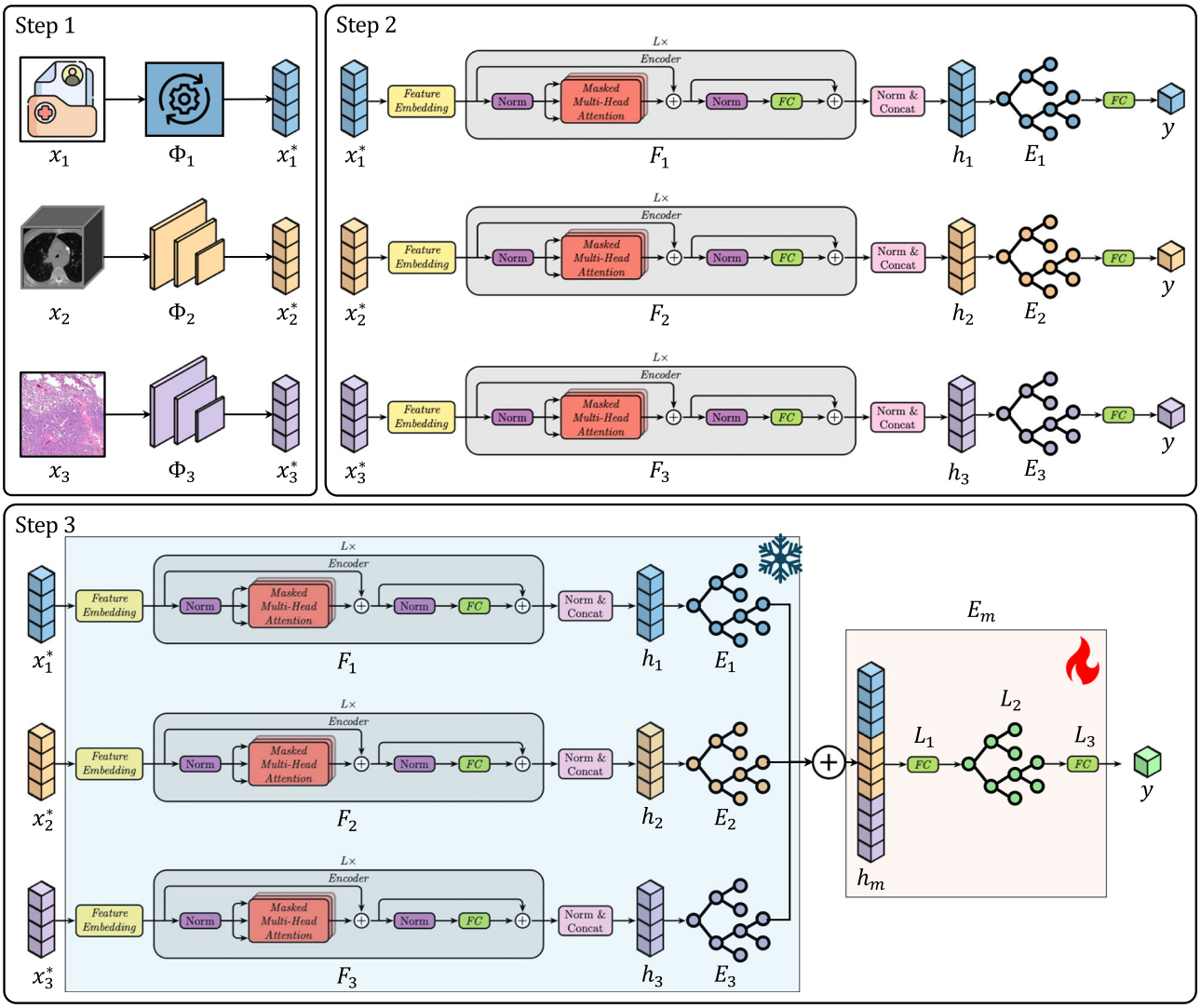


Fig. 2 | Schematic overview of the proposed multimodal survival framework. The framework integrates three clinical data streams: Pathology (WSI), Radiology (CT), and Tabular data. Step-1 extracts fixed embeddings via domain-specific foundation models. Step-2 processes each modality through a dedicated NAIM+ODST encoder with built-in missing modality handling. Step-3 implements ConcatODST, where unimodal latent representations are concatenated and passed through a multimodal module to predict the patient-specific hazard y.

evaluate the predictive performance of each unimodal representation on the full survival modeling task.

FMs enable unimodal survival modeling

Each unimodal representation was evaluated on the time-to-event OS task under three ML baselines: CPH[37], RSF[38], and SGB[39]. Regarding the Tabular modality, the three survival models differ in their treatment of missing values. These methods differ in how they handle missing tabular features. CPH and RSF, by design, handle missing entries during model fitting and are therefore applied directly to the available clinical features. In contrast, SGB requires complete inputs and was trained on KNN-imputed data. For imaging modalities, patients without a CT scan or WSI were excluded from that unimodal analysis only. Cross-validation splits were kept fixed across all experiments; reported performance thus varies in effective sample size across modalities, reflecting each modality's coverage in the cohort ("Missing modalities and tabular data sparsity").

Three DL architectures were considered alongside the ML baselines. We first considered a transformer-based model inspired by the Not-Another-Imputation-Method (NAIM)[32], which addresses missing values through masked self-attention, operating directly on incomplete inputs without imputation. The NN baseline shares NAIM's continuous and categorical embedding scheme but removes the transformer layers, replacing them with a standard MLP with KNN-imputed inputs. The NN model serves as a vanilla baseline to isolate the contribution of NAIM's missing-aware attention mechanism. Lastly, NAIM+ODST retains the full NAIM encoder but replaces the linear survival head with an oblivious differentiable sparsemax trees (ODST)[40], targeting higher-order non-linear interactions between the learned representations. More implementation details for the ODST hyperparameters are reported in the supplementary materials (Section S.2).

Table 3 reports unimodal results across all six models under 5-fold cross-validation, using C-index, Uno C-index, and td-AUC. For WSI, the high dimensionality of the FM embeddings challenged all the selected models. Among them, RSF achieved the strongest overall ranking (C-index 65.47, Uno 62.51), while the NN failed to generalize, yielding poor results across the board (Uno 45.21, td-AUC 41.64). Lastly, NAIM +ODST reaches the highest td-AUC among DL models (55.33) and Uno c-index (56.28). For CT features extracted via CT-FM, SGB led ML models (C-index 70.22, Uno 66.24). Among DL models, NAIM+ODST achieved the highest td-AUC (74.57) across all the models, yielding the

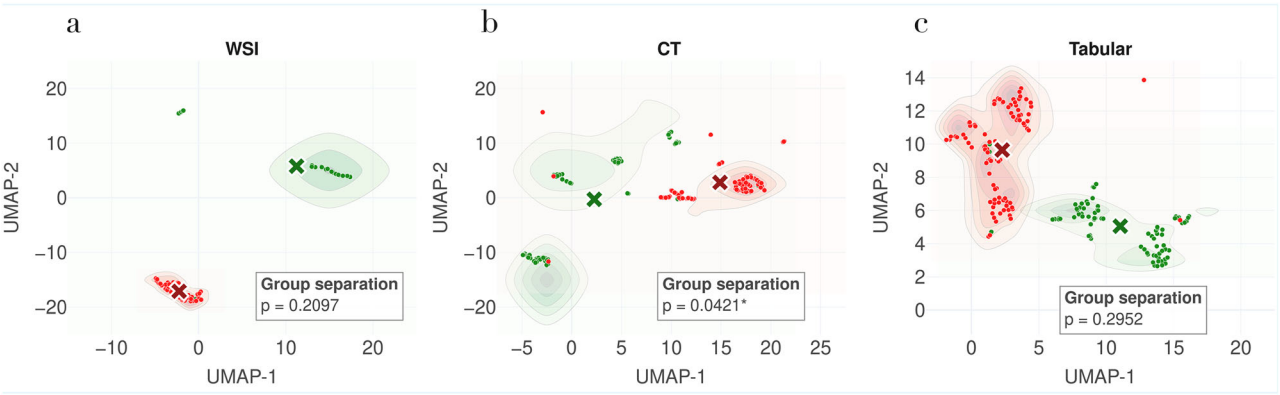


Fig. 3 | SPrognostic separability in foundation model embedding spaces. Two-dimensional UMAP projections of foundation model embeddings for WSI (**a**), CT (**b**), and tabular data (**c**). Each plot shows the distribution of survivors (green) and non-survivors (red), with kernel density contours and class centroids. Reported *p* values quantify group separation in the embedding space.

Table 3 | Unimodal survival performance across WSI, CT, and tabular modalities for all evaluated models: three ML baselines (CPH, RSF, SGB) and three DL architectures (NN, NAIM, NAIM+ODST)

Modality	Model	Type	C-index	td-AUC	Uno
WSI	CPH	ML	49.71 ± 2.27	43.79 ± 3.29	50.62 ± 2.26
	RSF	ML	**65.47 ± 1.15**	**62.68 ± 2.69**	**62.51 ± 1.67**
	SGB	ML	55.57 ± 4.16	53.95 ± 7.61	52.55 ± 5.62
	NN	DL	47.85 ± 6.47	41.64 ± 5.95	45.21 ± 5.49
	NAIM	DL	55.14 ± 6.86	48.67 ± 7.18	52.55 ± 6.09
	NAIM+ODST	DL	58.81 ± 4.04	55.33 ± 4.65	56.28 ± 2.69
CT	CPH	ML	58.13 ± 2.51	60.78 ± 2.28	56.88 ± 2.12
	RSF	ML	66.62 ± 6.09	71.90 ± 7.71	61.38 ± 5.15
	SGB	ML	**70.22 ± 5.05**	74.48 ± 5.76	**66.24 ± 4.37**
	NN	DL	59.98 ± 4.24	59.39 ± 3.84	58.63 ± 4.81
	NAIM	DL	63.39 ± 2.25	61.57 ± 2.16	61.23 ± 3.48
	NAIM+ODST	DL	68.75 ± 5.29	**74.57 ± 6.42**	64.42 ± 4.86
Tabular	CPH	ML	59.72 ± 4.78	65.05 ± 7.39	57.75 ± 3.39
	RSF	ML	**67.47 ± 3.61**	**75.28 ± 3.84**	62.39 ± 2.20
	SGB	ML	57.58 ± 4.12	60.75 ± 5.36	56.72 ± 3.95
	NN	DL	60.67 ± 4.12	66.48 ± 5.82	55.94 ± 2.27
	NAIM	DL	63.85 ± 3.76	69.68 ± 4.60	61.75 ± 3.33
	NAIM+ODST	DL	65.52 ± 4.45	72.73 ± 5.57	**63.56** ± 4.35

Results are reported as mean ± SE over fivefold cross-validation. **Bold** and underline indicate the best and second-best score per modality, respectively.

second highest C-index (68.75), outperforming both NAIM (63.39) and the NN (59.98). For Tabular features, RSF was the strongest model (C-index 67.47, td-AUC 75.28). Similar to the result on the CT, NAIM+ODST obtained the second-best score behind RSF on C-index (65.52) and surpassed it on Uno (63.56 vs 62.39), while improving over the standalone NAIM encoder by 3.0 points in td-AUC (72.73 vs 69.68). Based on this consistent performance across disparate modalities, NAIM+ODST was selected as the primary unimodal DL encoder for this study. This model serves as the unified backbone for mapping both structured clinical features and high-dimensional imaging representations (CT and WSI) into a shared prognostic space.

Intermediate fusion strategies maximize multimodal synergy

Building on the unimodal baselines presented in the previous section, Table 4 summarizes multimodal survival performance under three fusion paradigms, early, late, and intermediate, across all bimodal and trimodal modality combinations. A detailed description of all architectures and training procedures is provided in "Methods".

Early fusion ($\oplus$NAIM+ODST) concatenates the FM-extracted embeddings of all available modalities and feeds the joint vector into NAIM+ODST, the best-performing unimodal DL encoder, trained end-to-end directly on the combined input. Late fusion (NAIM+ODST^{n}) trains one independent NAIM+ODST per modality and combines their outputs at the decision level by rank-based score aggregation, as described in "Methods". This ensemble baseline explicitly avoids any feature-level interaction between modalities. Intermediate fusion (ConcatODST) builds on frozen unimodal encoders pretrained on the same cross-validation fold, with their FC prediction heads removed. The latent representations are concatenated and processed by the MM module: a FC layer down-scaling the fused representation by one-third, followed by an ODST layer modeling non-linear cross-modal interactions. The freezing strategy and MM module configuration are supported by the ablation study in Section S.1.

Table 4 | Survival prediction performance across unimodal and multimodal configurations under different fusion strategies

Fusion	Modalities	Model	C-ix	td-AUC	Uno
Unimodal	WSI	RSF	65.47 ± 1.15	62.68 ± 2.69	62.51 ± 1.67
	WSI	NAIM+ODST	58.81 ± 4.04	55.33 ± 4.65	56.28 ± 2.69
	CT	SGB	70.22 ± 5.05	74.48 ± 5.76	66.24 ± 4.37
	CT	NAIM+ODST	68.75 ± 5.29	74.57 ± 6.42	64.42 ± 4.86
	Tabular	RSF	67.47 ± 3.61	75.28 ± 3.84	62.39 ± 2.20
	Tabular	NAIM+ODST	65.52 ± 4.45	72.73 ± 5.57	63.56 ± 4.35
Early	WSI+CT+Tab	$^{\oplus}$NAIM+ODST	73.26 ± 4.05	77.73 ± 5.04	68.94 ± 5.04
	WSI+CT	$^{\oplus}$NAIM+ODST	71.28 ± 4.09	74.27 ± 5.19	67.97 ± 4.43
	WSI+Tab	$^{\oplus}$NAIM+ODST	66.65 ± 3.08	67.77 ± 3.15	65.00 ± 3.60
	CT+Tab	$^{\oplus}$NAIM+ODST	71.93 ± 4.74	<u>78.94 ± 5.96</u>	68.59 ± 3.80
Late	WSI+CT+Tab	NAIM+ODST^{3}	71.50 ± 4.58	76.60 ± 5.97	67.63 ± 3.63
	WSI+CT	NAIM+ODST^{2}	<u>74.02 ± 3.89</u>	<u>76.52 ± 4.60</u>	**70.35 ± 4.37**
	WSI+Tab	NAIM+ODST^{2}	67.44 ± 3.08	67.19 ± 3.08	**70.35 ± 4.07**
	CT+Tab	NAIM+ODST^{2}	69.00 ± 6.11	76.93 ± 7.80	63.86 ± 4.80
Interm.	WSI+CT+Tab	ConcatODST	**74.42 ± 4.25**	**80.74 ± 5.21**	<u>69.01 ± 3.79</u>
	WSI+CT	ConcatODST	73.13 ± 3.67	77.96 ± 4.73	68.60 ± 3.64
	WSI+Tab	ConcatODST	70.63 ± 1.41	75.34 ± 2.64	66.77 ± 2.56
	CT+Tab	ConcatODST	69.90 ± 4.94	76.32 ± 6.11	66.09 ± 4.51

Models are evaluated using Concordance Index (C-ix), time-dependent AUC (td-AUC), and Uno's C-index, reported as mean ± standard deviation over fivefold cross-validation. Unimodal baselines include both classical machine learning models and the proposed NAIM+ODST architecture. Multimodal results are grouped by fusion paradigm: early fusion ($^{\oplus}$NAIM + ODST), late fusion (NAIM+ODST^{n}), and intermediate fusion (ConcatODST). Best-performing results for each metric are highlighted in bold, while second-best values are <u>underlined.</u>

The trimodal intermediate fusion model (ConcatODST, WSI+CT +Tab) achieves the highest C-index (74.42 ± 4.25) and td-AUC (80.74 ± 5.21) in this study. On C-index, it improves by 4.20 points over the strongest unimodal ML model (CT SGB, 70.22 ± 5.05); on td-AUC, by 5.46 points over the best unimodal result (Tabular RSF, 75.28 ± 3.84). CT features contribute positively across all configurations. The WSI+CT late fusion model achieves a C-index of 74.02 ± 3.89 and the highest Uno C-index among all evaluated models (70.35 ± 4.37), outperforming the corresponding early (71.28 ± 4.09, Uno 67.97) and intermediate (73.13 ± 3.67, Uno 68.60) fusion variants, the only configuration where late fusion exceeds its intermediate counterparts. This result is notable given the 64% WSI missingness rate at test time: CT features compensate for absent histopathology, stabilizing predictions that would otherwise depend on the WSI unimodal signal alone (best DL C-index 58.81 ± 4.04).

Among bimodal configurations, CT+Tab early fusion reaches a td-AUC of 78.94 ± 5.96, the second-highest td-AUC across all models and fusion strategies. The CT+Tab intermediate model (69.90 ± 4.94) yields a lower C-index than its early counterpart (71.93 ± 4.74), suggesting that joint optimization does not uniformly help when both modalities already carry complementary discriminative signals at the feature level. By contrast, WSI +Tab underperforms relative to configurations involving CT. The WSI +Tab intermediate model achieves a C-index of 70.63 ± 1.41, lower than both trimodal intermediate (74.42) and WSI+CT late (74.02), and its td-AUC (75.34) falls behind CT+Tab early (78.94) by 3.60 points. The low variance of the WSI+Tab intermediate configuration (C-index SEM 1.41) reflects stable but ceiling-limited performance, likely because WSI missingness prevents the histopathological branch from contributing a strong independent signal.

Across fusion strategies, intermediate fusion outperforms early fusion in the trimodal setting (74.42 vs 73.26 in C-index; 80.74 vs 77.73 in td-AUC), confirming that end-to-end joint optimization extracts synergies not available to feature-level concatenation. Late fusion reaches competitive performance only for WSI+CT, where the two unimodal branches are individually strong (CT NAIM+ODST C-index 68.75; WSI RSF 65.47); averaging predictions from two well-calibrated models suffices here. For CT +Tab late fusion, C-index drops to 69.00 ± 6.11 despite both unimodal baselines exceeding 65, confirming that unweighted score averaging is sensitive to mismatched prediction scales across modality types.

Robustness to modality dropout and adaptive attention

To assess the robustness of the proposed framework under incomplete modality profiles, we investigate how progressive removal of individual data streams affects survival prediction in the trimodal ConcatODST intermediate fusion model. We conducted a controlled missingness analysis by progressively masking individual modalities in the test set from their natural baseline to complete unavailability (100% missing), while keeping the remaining modalities at their observed rate. This evaluation setting serves as a post-hoc Explainable AI (XAI) mechanism. Specifically, the performance drop induced by suppressing a modality can be interpreted as an estimate of the model's reliance on that data stream.

Figure 4.a decomposes modality contributions in the trimodal setting and reveals a consistent prognostic hierarchy. Masking Tabular data (green curve) leaves performance stable across the full missingness range, suggesting that the WSI and CT streams jointly compensate for absent clinical variables in this trimodal setting. Masking WSI (blue curve) has a limited effect until complete removal, where C-index drops moderately to ≈ 70, consistent with the high natural missingness rate already present during training. CT (red curve) is the most sensitive modality: progressive masking drives a consistent performance decline reaching a minimum near 75% missing, followed by a partial recovery at 100%. Taken together, the trimodal model maintains a C-index above 66 under complete removal of any single modality, demonstrating that the missing-aware fusion design yields a model robust to the loss of any individual data stream.

In WSI+CT (Fig. 4b), masking WSI causes a continuous and sharp degradation reaching ≈ 40 at complete removal; masking CT shows a more moderate decline across the intermediate range before collapsing at 100%. Given the substantially higher availability of CT at test time (95% vs 36% for WSI), the model likely relies more heavily on the radiological stream under natural conditions, with WSI contributing a complementary signal. In WSI +Tabular (Fig. 4c), masking Tabular data drives a steep and monotonic

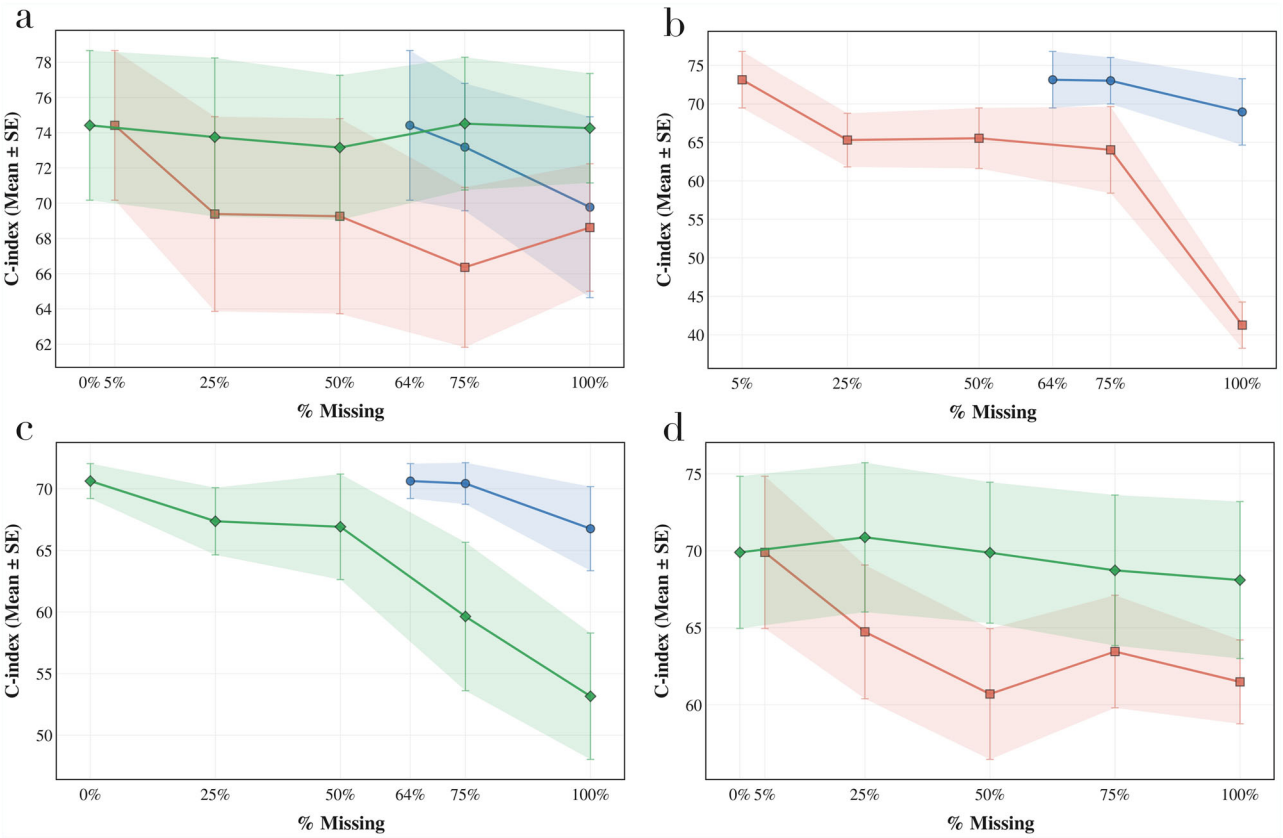


Fig. 4 | Robustness to modality dropout under increasing missingness. Performance of multimodal intermediate fusion models as a function of progressive modality removal. Each panel reports the mean C-index ± standard error over fivefolds. Each line represents the performance when a specific modality is removed at test time, with the *x* axis denoting the fraction of test samples for which that modality is missing. Color coding identifies the modality being removed: green corresponds to Tabular, red to CT, and blue to WSI. Solid lines denote CT-FM encoders, while dashed lines denote Merlin encoders (**a**, **b**, **d**). Panels correspond to different modality combinations: **a** WSI+CT+Tabular, **b** WSI+CT, **c** WSI+Tabular, and **d** CT+Tabular.

performance drop to ≈ 52 at 100% missing, while masking WSI leaves C-index stable across the full range (≈ 70), indicating that clinical variables are the primary driver in this bimodal configuration. In CT+Tabular (Fig. 4d), masking CT produces an early performance drop that persists across all missingness levels, whereas masking Tabular data has a comparatively small effect, with C-index remaining near the bimodal baseline throughout. Notably, when a bimodal model is forced into a unimodal inference regime (i.e., by masking one modality entirely, performance decreases relative to the full bimodal setting, as expected due to information removal, yet in several cases remains competitive with, or even exceeds, the corresponding unimodal baselines. This behavior can be observed in the WSI+Tabular where the removal of the WSI slides yields 67.92 ± 2.14 C-index, which is slightly higher than the best performing unimodal ML model trained on Tabular data (RSF, 67.47 ± 3.61). This suggests that the model does not merely combine independent unimodal predictors, but learns internal representations that remain informative even when only a subset of modalities is available, supporting the practical utility of the proposed missingness-aware fusion approach.

Survival analysis and disease progression stratification

To assess the translational clinical utility of the intermediate fusion models, we performed an aggregated risk stratification analysis. Risk scores predicted on each validation fold were concatenated into a single distribution, from which an optimal threshold was derived via the log-rank test. This threshold was then applied to the corresponding test-fold predictions, partitioning patients into High-Risk and Low-Risk groups.

Kaplan-Meier survival curves were estimated separately for the High-Risk and Low-Risk groups using the observed survival times, and the log-rank test was applied to assess the statistical significance of the separation between the two strata (Fig. 5). All four intermediate fusion configurations achieved significant risk-group discrimination. CT+Tabular (Fig. 5a, p = 0.048), WSI+CT (Fig. 5b, $p = 0.017$), WSI+CT+Tabular (Fig. 5c, $p = 0.023$), and WSI+Tabular (Fig. 5d, p = 0.011) all achieve statistically significant separation between the two risk strata, confirming that intermediate fusion produces clinically discriminative risk scores across all modality combinations tested.

To further validate the clinical relevance of the learned risk scores, we assessed whether the stratification aligns with disease progression patterns beyond overall survival. We focused on two complementary endpoints: progression-free survival (PFS), a composite measure capturing local progression, metastatic spread, and censoring status, and distant metastasis (DM), which reflects systemic tumor dissemination. Figure 6 reports the Kaplan–Meier curves obtained from the trimodal intermediate fusion model (WSI+CT+Tabular). Patient-specific risk scores were aggregated across test folds and used to stratify patients into Low-Risk and High-Risk groups. Within each risk group, survival differences were evaluated separately for PFS (Fig. 6a) and DM (Fig. 6b).

For PFS, the Low-Risk group shows a significant separation between patients who experience progression and those who do not (log-rank p = 0.007), while in the High-Risk group the separation does not reach significance (p = 0.090). A similar pattern holds for DM, where the Low-Risk group achieves significant discrimination between patients who develop distant metastases and those who remain metastasis-free (p = 0.002), whereas the High-Risk group approaches but does not reach significance (p = 0.059). The analysis of the resulting survival curves reveals a structural asymmetry across risk strata. For both PFS and DM, the no-event subgroups

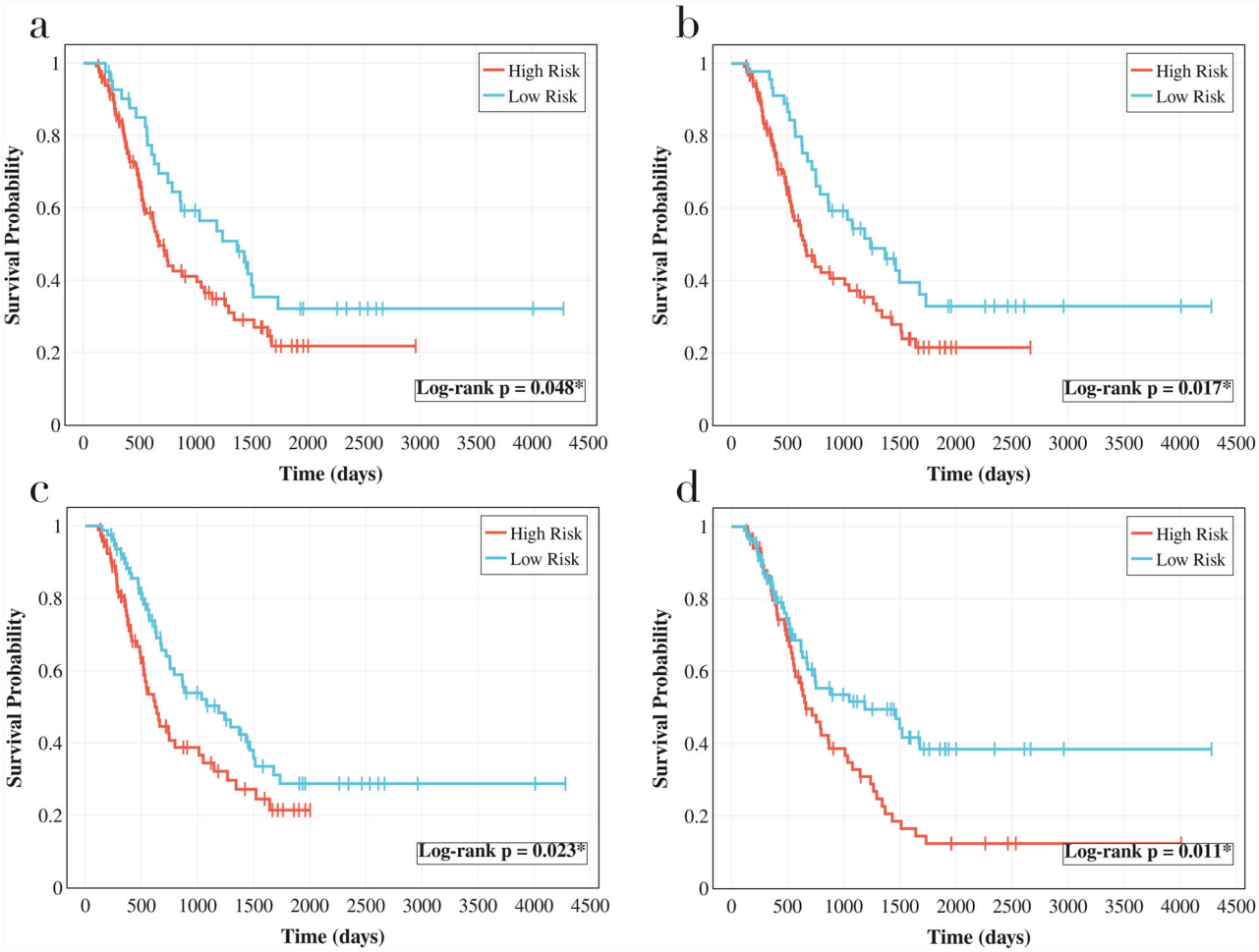


Fig. 5 | Risk-stratified survival analysis across multimodal configurations. Stratified Kaplan–Meier survival curves for intermediate fusion models (`ConcatODST`), with follow-up time (days) on the *x* axis and survival probability on the *y* axis. Patients are stratified into Low-Risk and High-Risk groups according to the predicted 5-year outcome score produced by each model. The blue curve represents the Low-Risk group, while the red curve corresponds to the High-Risk group. Panels correspond to different modality combinations: **a** CT+Tabular, **b** WSI+CT, **c** WSI+CT+Tabular (trimodal), and **d** WSI+Tabular.

differ sharply between Low-Risk and High-Risk patients: event-free Low-Risk patients sustain high survival probabilities throughout follow-up, while event-free High-Risk patients decline at a substantially faster rate. The risk score thus characterizes patients even in the absence of observed progression, capturing latent biological differences that separate indolent from aggressive disease courses. By contrast, the event-positive subgroups across both risk strata are largely overlapping, indicating that patients who experience progression or distant metastasis converge toward a common survival trajectory irrespective of their initial risk classification.

This dual behavior underscores the clinical utility of the proposed stratification. The model identifies a Low-Risk subpopulation with stable disease control that may be managed with reduced surveillance intensity. In the High-Risk group, the clear separation among event-free patients supports the use of the risk score as an early warning signal, flagging patients who carry elevated baseline risk and may benefit from intensified monitoring or preemptive therapeutic escalation before progression manifests.

Discussion

This study addresses a structural gap in multimodal oncology modeling, namely the near-universal assumption of complete modality availability. Biopsies are not always feasible, histopathology slides are not always digitized, and CT scans acquired at referring institutions may be unavailable at the treating center. Treating missingness as a training design constraint rather than a pre-processing problem allowed us to retain 179 patients in the analysis instead of the 41 who would have remained under a complete-case filter. This distinction carries direct clinical consequences. The excluded patients are not a random subset. Patients with the most aggressive disease trajectories are precisely those for whom full staging workups are most likely to be incomplete, and their systematic exclusion biases any prognostic model toward the more favorable end of the risk spectrum ("Missing modalities and tabular data sparsity").

The prognostic hierarchy observed across modalities is consistent with known biological properties of locally advanced NSCLC. Structured clinical variables, which encode treatment decisions, molecular driver status, and performance status, carry prognostic information that imaging alone cannot replicate ("Cohort characteristics"). The significant univariate associations of EGFR mutation status ($p = 0.033$), radiotherapy technique ($p < 0.001$), and adjuvant immunotherapy ($p = 0.010$) with 5-year survival reflect the extent to which treatment-era factors, rather than baseline tumor biology alone, determine outcome in this population. The survival advantage of modern volumetric arc techniques over three-dimensional conformal radiotherapy and the impact of durvalumab consolidation following the PACIFIC regimen are well-established clinically; their reflection in the tabular stream confirms that the model is not learning noise.

From a clinical perspective, the most important question raised by this work is how a continuous risk score can be made actionable within existing oncology workflows. The stratification analysis showed significant separation between High-Risk and Low-Risk groups across all four intermediate fusion configurations (log-rank p ranging from 0.011 to 0.048), providing statistical evidence of discriminative capacity, but statistical significance

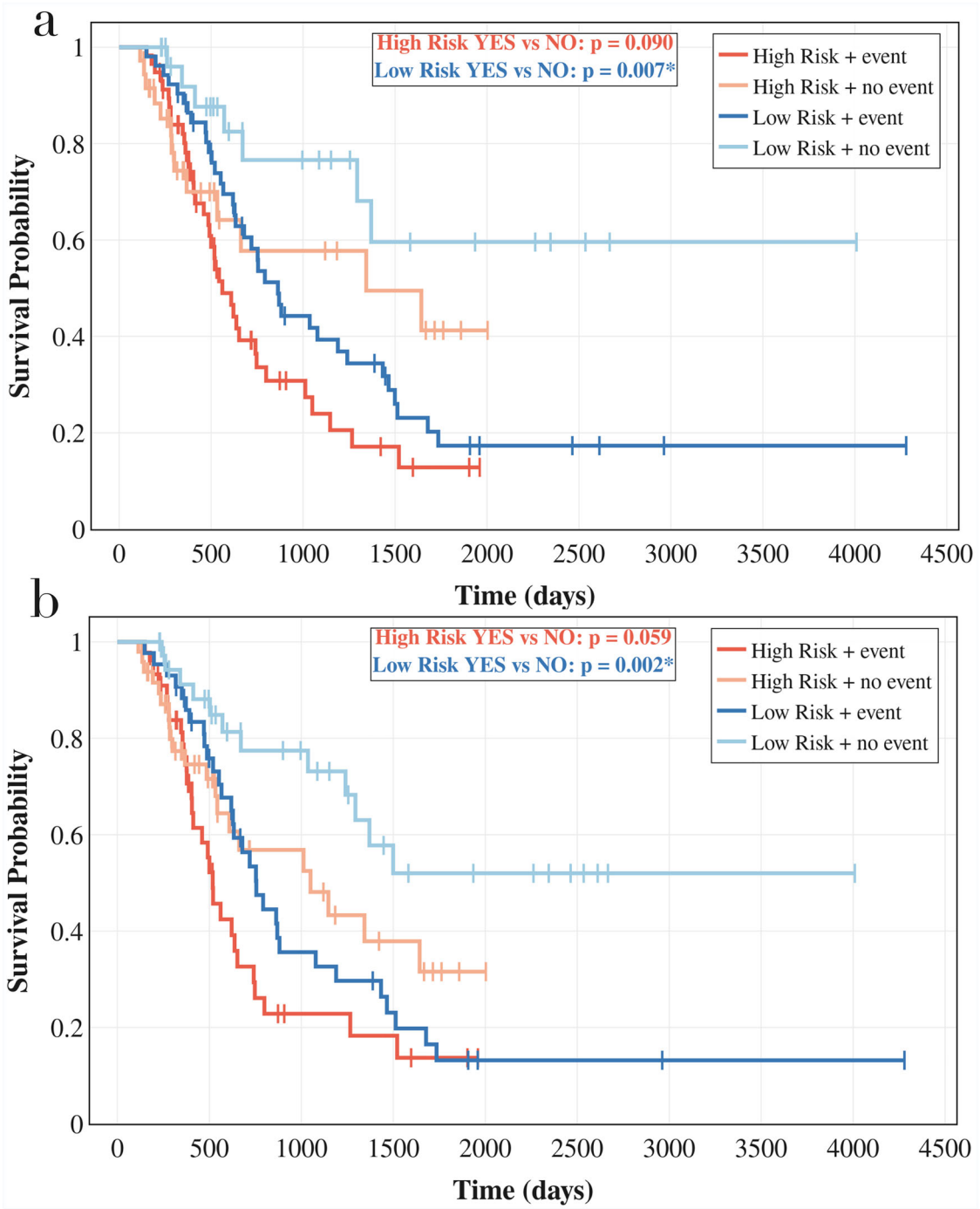


Fig. 6 | Progression-based survival stratification using the trimodal intermediate fusion model. Kaplan–Meier curves for progression-free survival and distant metastasis. Patients are first stratified into Low-Risk and High-Risk groups using a threshold derived from validation folds, and this threshold is subsequently applied to the test set. Within each risk group, patients are further separated according to event status (event vs no-event) for the corresponding endpoint. This stratification highlights differences in survival trajectories conditioned on both predicted risk and observed progression patterns. **a** reports progression-free survival (PFS), while **b** corresponds to distant metastasis.

alone does not determine clinical utility. Oncologists managing unresectable Stage III NSCLC operate within a relatively narrow therapeutic decision space. The clinical value of a risk score in this setting, therefore, lies not in treatment selection per se, but in surveillance intensity, early escalation triggers, and patient counseling. The within-group PFS and DM stratification produced by the trimodal model enables this pathway more concretely. Within the Low-Risk group, patients who subsequently progressed were separated from those who remained event-free with high significance (p = 0.007 for PFS, $p = 0.002$ for DM), while the High-Risk group showed a faster overall decline even among patients who did not progress ("Survival analysis and disease progression stratification"). This asymmetry implies that the risk score is not simply summarizing binary outcome status, but captures a latent biological aggressiveness that manifests as a trajectory rather than an event. Clinically, a High-Risk, event-free patient is not the same as a Low-Risk, event-free patient, even when their observed histories are identical at the time of scoring. This distinction is precisely the kind of information that

could justify earlier post-treatment imaging in High-Risk patients, or enrollment into consolidation therapy trials for those with elevated baseline risk scores who nonetheless complete chemo-radiotherapy without acute progression.

Translating these risk scores into clinical practice requires threshold definitions that go beyond the log-rank-optimized binary cutoff used in this retrospective analysis. The optimal cutoff derived from validation-fold score distributions served as a methodological tool for assessing stratification capacity, not as a clinically calibrated decision boundary. For prospective use, thresholds should be defined relative to institutional event rates and treatment protocols. Risk score integration would be most useful as a continuous variable overlaid on existing stratification systems such as TNM staging and ECOG performance status, not as a replacement for either. The practical deployment pathway, we envision is one where the multimodal risk estimate is surfaced within the tumor board workflow, alongside the radiological read and pathological report, as an additional quantitative input that extends the prognostic horizon without displacing established clinical reasoning. Importantly, the model's tolerance for missing modalities means it can produce a score for any patient who has at least tabular data available, which is the universal modality in this cohort. This property is essential for clinical adoption, as a tool that silently fails or refuses inference when a slide is unavailable is not usable in routine care.

Several open questions condition whether the performance reported here translates beyond this cohort. The study population is deliberately narrow, comprising unresectable Stage II–III NSCLC treated with radical chemo-radiotherapy at a single institution. The exclusion criteria that define this population produce a clinically homogeneous group but limit generalization. The prognostic dynamics of resectable Stage II disease, where surgery is the primary treatment, or of metastatically Stage IV disease managed with SBRT and systemic therapy, differ substantially from those modeled here. Multi-center validation within the same clinical indication is the prerequisite for generalizability, and it requires partner institutions that share not only tumor boards and treatment protocols but also data infrastructure that permits harmonization of CT acquisition parameters, WSI digitization standards, and tabular variable definitions across sites.

Beyond cohort scope, two technical directions represent the most consequential next steps. The first is tumor-aware CT feature extraction. All three foundation models evaluated in this study encode the full thoracic volume, treating gross tumor burden, mediastinal involvement, and uninvolved lung parenchyma as an undifferentiated signal (Section S.3). Restricting extraction to the gross tumor volume or clinical target volume, as defined by the treating radiation oncologist, would concentrate the radiological representation on the tissue directly responsible for local failure and metastatic seeding. This is not merely an engineering refinement, given that GTV delineation already exists in the treatment planning workflow for every patient who undergoes radical chemo-radiotherapy, making it a zero-cost additional input from a data collection standpoint. Whether foundation models conditioned on anatomically restricted inputs recover prognostic signal that whole-volume encoders suppress is an empirical question this study cannot answer, but the biological rationale for testing it is strong. The second direction concerns fusion expressiveness. The intermediate fusion architecture used here, a concatenation-based head with frozen unimodal encoders, was chosen for its interpretability and resistance to overfitting on a small cohort. Cross-attention mechanisms that allow modality representations to condition one another during feature refinement, or dynamic gating networks that weight each stream by its observed completeness and estimated informativeness per patient, would in principle extract richer cross-modal interactions. These approaches require larger missing-aware multimodal cohorts to train reliably without collapsing to degenerate solutions, and the field does not yet have well-validated benchmarks for this setting. A clear priority is therefore the construction of multi-institutional datasets with heterogeneous modality profiles collected under standard clinical workflows. Computational feasibility is a practical barrier to clinical translation that has received less attention than model architecture. Foundation model inference for a single patient in this pipeline requires WSI tissue segmentation, patch-level encoding across thousands of image tiles, slide-level aggregation, and volumetric CT embedding, none of which is compatible with real-time clinical use under current hardware constraints. Knowledge distillation of the imaging FMs into lightweight surrogates, or caching of embeddings computed at the time of treatment planning, would reduce inference latency to a degree compatible with tumor board timelines. The tabular and survival head components are computationally trivial; the bottleneck is entirely in the FM extraction stage.

Finally, the incremental value of the multimodal risk score over established prognostic indices remains to be quantified. This study benchmarks modality combinations against one another and against unimodal baselines, but does not include direct comparison against *TNM* stage, *ECOG* performance status, or published nomograms for NSCLC[41]. Demonstrating net reclassification improvement over these existing tools, or equivalently showing that the multimodal score stratifies patients within TNM substages where clinical staging provides insufficient discrimination, is the evidence threshold that would move this class of model from methodological interest to clinical relevance. These comparisons are the natural next step once a prospectively collected, multi-site cohort with complete staging and outcome data is available.

Methods

Materials and cohort selection

Data collection was carried out at the Fondazione Policlinico Universitario Campus Bio-Medico of Rome and encompassed 353 patients with histologically confirmed NSCLC, all diagnosed at stage II–III following the 9th edition TNM classification of lung cancer[42] and enrolled retrospectively or prospectively between 2012 and 2024. From this cohort, we retained only individuals classified as unresectable stage II–III disease[43] (IIA, IIB, IIIA, IIIB, or IIIC) using the 9th edition of the TNM classification for Thoracic Oncology[42]. Accordingly, the study employed the following exclusion criteria to ensure a homogeneous cohort: (i) patients with resectable tumors who underwent surgery in neoadjuvant settings; (ii) patients treated with palliative intent; (iii) cases with incomplete baseline clinical information or missing follow-up; (iv) patients with small cell lung cancer. In addition, we removed patients with non-NSCLC histologies, metastatic or recurrent disease at presentation, prior thoracic surgery or postoperative RT/RTCT, those managed with non-standard therapeutic pathways (e.g., SBRT alone or systemic therapy only), and cases lacking verifiable diagnostic imaging or pathology.

After applying these criteria, the final study population consisted of 179 patients, providing a homogeneous cohort representative of the clinical course of locally advanced, non-resectable NSCLC. For every included patient, a comprehensive set of structured clinical variables was curated to capture the biological profile and treatment history. As detailed in Tables 1 and 2, data collection was stratified into four key domains: (i) general characteristics, including demographic and physiological baselines such as sex, age, weight, height, smoking status, daily cigarette consumption, family history of neoplasms, numeric pain rating scale (NRS), and specific comorbidity profiles (vascular, metabolic, renal, pulmonary); (ii) tumor biology, encompassing histological subtype (adenocarcinoma, squamous cell carcinoma, NOS), TNM classification, molecular biomarkers (EGFR, ALK, MET, PD-L1), and other biopsy details; (iii) treatment parameters, covering chemotherapy details (induction cycles, specific drug schemes), radiotherapy metrics (total dose, fractionation, technique, duration), treatment interruptions, and the administration of adjuvant immunotherapy; and (iv) toxicity and outcomes, recording grade-specific adverse events (esophageal, pulmonary, hematological).

Beyond Overall Survival (OS), the study incorporated a comprehensive assessment of disease progression patterns to characterize failure modes. Secondary clinical endpoints included: (i) distant Metastasis (M), indicating systemic disease spread to non-contiguous organs; and (ii) Progression-Free Survival (PFS), calculated as the time from treatment initiation to the first documented event of local and distant progression or death.

Imaging acquisition. For each patient, multimodal imaging data were systematically acquired prior to the initiation of radical therapy to establish a baseline for prognostic modeling.

- **Radiology**. CT scans were retrospectively retrieved from the institutional picture archiving and communication system (PACS). Specifically, we collected the treatment planning CTs acquired within 30 days prior to the start of Chemoradiotherapy. Unlike standard diagnostic scans, these planning CTs are acquired under rigorous protocols to ensure geometric accuracy for dose calculation, providing a standardized and reproducible representation of the macroscopic tumor burden and thoracic anatomy across the entire cohort.
- **Pathology**. High-resolution WSIs were generated from diagnostic tissue biopsies obtained via bronchoscopy or CT-guided percutaneous needle aspiration. Formalin-fixed paraffin-embedded (FFPE) tissue blocks were sectioned and stained with Hematoxylin and Eosin (H&E) according to standard clinical workflows. Subsequently, the physical glass slides were digitized at ×20 magnification (0.5 µm/pixel) using a high-throughput slide scanner. This digitization process transformed the analog histological samples into gigapixel-resolution digital slides, ensuring the preservation of the intricate microscopic tumor architecture and cellular phenotypes required for computational analysis.

Ethical approval

This study was conducted in accordance with the Declaration of Helsinki. Ethical approval was obtained from the Ethical Committee of Fondazione Policlinico Universitario Campus Bio-Medico. Data collection spans two regulatory phases: a retrospective phase approved on 30 October 2012 and registered on ClinicalTrials.gov (Identifier: NCT03583723), and a prospective phase approved on 16 April 2019 (Identifier: 16/19 OSS). All patients provided written informed consent for the use of their clinical and imaging data for research purposes.

Methodological framework overview

We propose a three-stage multimodal framework designed to predict survival outcomes from incomplete heterogeneous data. As illustrated in Fig. 2, the pipeline consists of: (1) Unimodal feature extraction using pre-trained FM to bypass data scarcity; (2) Missing-aware representation learning using the `NAIM` architecture; and (3) our proposed intermediate fusion strategy.

Step 1: preprocessing & feature extraction

We formalize the three-stage multimodal learning pipeline illustrated in Fig. 2. Given three heterogeneous input modalities, Tabular (x_1); CT (x_2), and WSI (x_3), the goal is to construct a unified representation for downstream survival prediction. We denote by $\Phi_i : \mathcal{X}_i \to \mathbb{R}^{d_i}$ the modality-specific feature extractors producing high-level embeddings $x_i^* \in \mathbb{R}^{d_i}$ leveraging state of the art FMs for each modalities or classical pre-processing pipelines.

Tabular data (x_1).

To ensure consistency during model training, all clinical features underwent a standardized preprocessing pipeline. Categorical variables (e.g., comorbidities, histology) were one-hot encoded, while ordinal features (e.g., disease stage, ECOG status) were integer-encoded to preserve their inherent hierarchy. Numerical features (e.g., age, weight) were standardized via z-score normalization to zero mean and unit variance. The resulting feature vector is denoted as $x_1^* = \Phi_1(x_1)$, where Φ_1 represents the composite preprocessing function.

Radiology imaging (x_2).

Volumetric radiological features are extracted using CT-FM[30], a SegResEncoder-based FM pretrained via intra-sample contrastive learning (SimCLR) on 148,000 CT scans from the Imaging Data Commons. Raw CT volumes undergo the following preprocessing steps, consistent with the CT-FM encoding pipeline: (i) Orientation: all volumes are reoriented to a canonical SPL (Superior-Posterior-Left) coordinate system. (ii) Windowing & Normalization: voxel intensities are clipped to the range $[-1024, 2048]$ HU and linearly scaled to the [0, 1] interval. (iii) Foreground cropping: background regions are removed via foreground detection, retaining only the anatomically informative portion of the volume without imposing a fixed spatial resolution or crop size. The preprocessed volume x_2 is passed through the CT-FM encoder $\mathcal{E}_{CT}$, and the resulting feature maps are reduced to a single embedding via global average pooling, yielding a 512-dimensional representation:

$$x_2^* = \Phi_2(x_2) = \mathrm{AvgPool}\big(\mathcal{E}_{CT}(x_2)\big) \in \mathbb{R}^{512} \tag{1}$$

Pathology imaging (x_3).

To capture the microscopic tumor characterization, we utilize a hierarchical encoding pipeline to extract a unique representation vector for each WSI. First, WSIs are processed using CLAM[36] (CLustering-constrained Attention Multiple instance learning) with the objective to segment tissue from background artifacts. From each WSI (x_3), we extract a slide-specific set of tissue-containing image tiles. After tissue segmentation, all background and artefactual regions are removed, and the remaining tissue area is decomposed into a non-uniform number N_w of non-overlapping patches of size 256 × 256 pixels at ×20 magnification. For each x_3, we denote this patch set as $P_w(x_3) = \{p_j\}_{j=1}^{N_w}$, where each element p_j corresponds to a spatially localized region that retains diagnostically meaningful microscopic information. This patch-based decomposition allows the model to focus on localized morphological patterns that would otherwise be diluted at whole-slide scale.

Each patch $p_j \in P_w$ is then projected into a descriptive 512-dimensions embedding using the vision encoder of MI-Zero[25] ($\mathcal{E}_{patch}$), a vision-language foundation model pre-trained to align visual features with semantic pathology descriptions. Eq. (2) formalizes the projection step:

$$\mathbf{z}_j = \mathcal{E}_{\mathrm{patch}}(p_j) \in \mathbb{R}^{512},\ p_j \in P_w \tag{2}$$

Finally, TITAN[26](Transformer-based pathology Image and Text Alignment Network), $\mathcal{E}_{wsi}$, aggregates the sequence of patch embeddings into a unified latent representation with 768 dimensions. TITAN leverages a self-supervised masked prediction task to coherently synthesize local cellular phenotypes into a global descriptor. We indicate the final slide-level feature vector in the following Eq. (3):

$$x_3^* = \mathcal{E}_{wsi}\left(\left\{\mathbf{z}_j\right\}_{j=1}^{N_w}\right) \in \mathbb{R}^{768}. \tag{3}$$

By structuring the WSI into interpretable units and encoding them consistently, this pipeline preserves fine-grained microscopic details while ensuring that the representation is compatible with the subsequent multimodal fusion stages. The complete preprocessing pipeline Φ_3, representing the composition of patching, encoding, and aggregation, is formalized as:

$$x_3^* = \Phi_3(x_3) = \mathcal{E}_{\mathrm{wsi}}\Big(\mathcal{E}_{\mathrm{patch}}(p_j)\Big) \in \mathbb{R}^{768}; p_j \in P_w(x_3). \tag{4}$$

Step 2: missing-aware unimodal encoding (`NAIM`)

To robustly accommodate structural missingness, we employ the not another imputation method (`NAIM`)[32] model. Standard deep learning models require complete input vectors; traditionally, this forces researchers to use imputation (replacing missing values with means, medians, or synthetic guesses), which inevitably introduces noise and bias. `NAIM` offers a fundamental paradigm shift: rather than synthesizing fake data, the network is architected to dynamically ignore absent information. In our framework, we instantiate a separate `NAIM` encoder F_i for each modality. Each encoder takes as input the raw feature vector x_i^* (from Step 1) and a binary mask $M_i \in \{0, 1\}^K$ (where 1 indicates missing), projecting them into a robust latent representation h_i.

Dual-stream feature embedding. The first challenge is to convert heterogeneous data (continuous numbers, categorical classes, and missing values) into a unified high-dimensional space where they can interact. NAIM achieves this via a dual embedding strategy that treats missingness not as a null value, but as a specific structural state.

- **Categorical features** (x^{cat}): Discrete variables (e.g., histology, sex) are mapped using a learnable lookup table E^{cat}. If a feature value is present, the lookup table selects the corresponding trainable embedding vector; if the value is missing, the encoder instead retrieves the frozen padding vector. Formally, for the j-th categorical feature, the embedding is defined in Eq. (5).

$$e_j^{cat} = b_j + E_j^{cat}(x_j^{cat}), \tag{5}$$

  where b_j is a learnable bias term specific to that feature index.
- **Numerical features** (x^{num}): Continuous variables (e.g., age, tumor volume) require a more nuanced approach to preserve their magnitude. We employ a specialized lookup table E^{num} containing only two vectors: $\mathbf{v}_{\text{present}}$ (trainable) and $\mathbf{v}_{\text{missing}}$ (frozen zeros). If the feature is missing, the embedding is simply the bias term plus the zero vector. If the feature is present, we multiply the trainable $\mathbf{v}_{\text{present}}$ vector by the scalar value of the feature x_j^{num}. We formulate the numerical embedding in the following equation:

$$e_j^{num} = b_j + x_j^{num} \cdot E_j^{num}(\mathbb{I}_{present}), \tag{6}$$

  where $\mathbb{I}_{present}$ is the indicator index. This mechanism allows the network to learn the significance of the feature via $\mathbf{v}_{\text{present}}$ while scaling it by the actual patient data.

Double-sided masked attention. Once the data is tokenized into embeddings, it is processed by the NAIM Transformer encoder. However, standard self-attention mechanisms would fail because missing tokens would still participate in the softmax normalization, corrupting the attention distribution. NAIM solves this via a double-sided masking strategy. The goal is to ensure that a missing feature neither casts attention (influencing others) nor receives attention (being updated by others).

Let Q, K, V be the query, key, and value matrices derived from the input embeddings. We define a mask matrix $\mathbf{M}$, where $\mathbf{M}_{ij} = -\infty$ if either feature i or feature j is missing, and 0 otherwise. The resulting formulation, reported in Eq. (7), applies the mask before and after the softmax operation, thereby enforcing a complete suppression of missing-feature interactions:

$$\mathrm{A}(Q, K, V) = \mathrm{ReLU}\left(\mathcal{S}\left(\frac{QK^{\top}}{\sqrt{d_h}} + \mathbf{M}\right) + \mathbf{M}^{\top}\right)V. \tag{7}$$

Where $\mathcal{S}$ is the softmax operator. This bidirectional exclusion guarantees that the attention mechanism operates solely on observed features, preserving the functionality of the learned representations even under substantial missingness. Here is how this mechanism works step-by-step:

1. $\frac{QK^{T}}{\sqrt{d_h}} + \mathbf{M}$: Adding $-\infty$ to the attention scores forces the corresponding softmax, values to exactly zero.
2. $+\mathbf{M}^{\top}$: we add the mask again (transposed) after the softmax. This is a crucial reinforcement step that explicitly zeros out any residual connection from a missing feature to others.
3. ReLU: Ensures non-negativity and sparsity in the final attention map.

This mechanism ensures that gradients propagate only through observed data paths, effectively creating an architecture that changes structure based on which data are available for a specific patient.

NAIM**Encoders for Unimodal Survival Prediction**. Each modality is processed independently using a dedicated NAIM encoder. Unlike standard architectures that treat the input as a monolithic vector, NAIM tokenizes the input modality x_i into a sequence of distinct feature embeddings (as defined in Eqs. (5)–(6)). These tokens are processed via the masked self-attention mechanism, which captures intra-modality correlations while strictly preventing information leakage from missing values. The resulting sequence of contextualized embeddings is then flattened to form the comprehensive unimodal representation $h_i \in \mathbb{R}^{d_{model}}$.

To translate the robust latent representation h_i into a continuous survival prediction, the embedding is passed through a hybrid head composed of an ODST[40], noted as E_i, followed by a fully connected (FC) layer. The ODST architecture bridges the gap between the interpretability of decision trees and the gradient-based optimization of DL. Unlike standard decision trees, which rely on greedy, hierarchical splits, an oblivious tree applies a uniform splitting criterion across all nodes at a given depth. In this differentiable implementation, this hard splitting logic is relaxed using a soft gating function (e.g., sigmoid), enabling the entire ensemble to be optimized end-to-end via backpropagation. Formally, the ODST block maps the unimodal embedding h_i to a structured intermediate representation, capturing non-linear interactions between features. This transformed vector is then projected by the FC layer to produce the scalar risk score y_i, as defined in Eq. (8):

$$y_i = \mathrm{FC}\big(\mathrm{E_i}(h_i)\big). \tag{8}$$

Optimization objective. Based on the Cox proportional hazards model, the entire pipeline is trained end-to-end by minimizing the negative log partial likelihood loss[44]. Let N_{ob} be the total number of observed events in a batch. The loss function $\mathcal{L}_{sur}$ is defined as:

$$\mathcal{L}_{sur} = -\frac{1}{N_{ob}} \sum_{i:C_i=1} \left(y_i - \log \sum_{j:T_j \geq T_i} e^{y_j} \right), \tag{9}$$

where:

- $C_i = 1$ indicates the occurrence of the event for patient i (uncensored).
- y_i is the hazard risk score predicted by the network for patient i.
- T_i and T_j denote the observed survival times for patient i and j, respectively.
- The inner summation $\sum_{j:T_j \geq T_i}$ aggregates the risk over the set of all patients j still at risk at time T_i.

This objective enforces the correct relative ordering of survival times by penalizing the model when a patient who experienced an event earlier is assigned a lower risk score than those who survived longer.

Step 3: intermediate fusion and multimodal survival prediction

The framework subsequently transitions to the proposed intermediate fusion strategy. Each modality is processed by its dedicated NAIM encoder and ODST layer (as detailed in "Step 2: missing-aware unimodal encoding (NAIM)"), producing three latent representations: h_1 (Tabular), h_2 (CT), and h_3 (WSI). These embeddings capture modality-specific abstractions while being explicitly conditioned to handle missingness via the double-sided masking mechanism (Eq. (7)). The unimodal encoders are loaded from Step 2 with their weights frozen throughout multimodal training, preserving the modality-specific representations learned during unimodal pretraining and reducing the risk of overfitting on the limited cohort. To construct the unified multimodal patient profile, the frozen latent vectors are fused via concatenation ($\oplus$):

$$h_m = \oplus(h_1,\ h_2,\ h_3) \in \mathbb{R}^{d_1+d_2+d_3}. \tag{10}$$

This fused vector h_m retains complementary prognostic signals from each data source. In the proposed multimodal configuration, referred to as ConcatODST, h_m is passed to the multimodal module E_m, which is the only trainable component in this stage and maps the concatenated representation to the final risk score y_m. E_m comprises two sequential layers L_1 and L_2

followed by a third fully connected layer (L_3), mirroring the unimodal head structure. L_1 is a FC bottleneck that projects h_m to one-third of its input dimensionality, while L_2 is an ODST block that captures non-linear interactions across the fused unimodal embeddings. The full ablation over E_m architecture choices is reported in Section S.1. The E_m mathematical formulation follows:

$$y_m = \mathrm{FC}\left(L_2\left(L_1\left(h_m\right)\right)\right) = E_m(h_m) \tag{11}$$

E_m is optimized end-to-end using the $\mathcal{L}_{surv}$ objective defined in Eq. (9), with gradients propagating exclusively through the multimodal head while the upstream unimodal encoders remain fixed.

Model comparison and experimental setup

To evaluate the proposed intermediate fusion framework, we conducted a structured benchmarking analysis comparing different architectural paradigms, integration strategies, and established ML baselines.

Unimodal baselines and ablation study. Before assessing multimodal fusion, we examined the performance of each modality independently through a unimodal ablation study. For Tabular, CT, and WSI data, we compared the following configurations:

- Machine learning baselines: Cox Proportional Hazards (`CPH`), Random Survival Forests (RSF), and `SGB`, trained on the extracted features.
- Standard neural network (`NN`): a baseline model using the feature embedding layer, as described in "Step 2: missing-aware unimodal encoding (NAIM)", followed by a multi-layer perceptron with ReLU activations.
- `NAIM` (encoder-only): the `NAIM` encoder paired with a linear survival head, isolating the contribution of the masked attention mechanism.
- `NAIM+ODST`: the complete unimodal model combining the `NAIM` encoder with the `ODST` survival head.

Multimodal fusion strategies. To isolate the contribution of the intermediate fusion architecture, we benchmarked it against two alternative integration paradigms, adopting a consistent integration framework (concatenation + ODST head) but varying the optimization scope:

- **Early fusion** ($^{\oplus}$`NAIM+ODST`) Integration is performed directly at the feature level. The FM-extracted embeddings from Step 1 are concatenated into a single unified vector, $x_{\text{early}} = \oplus(x_1^*, x_2^*, x_3^*)$, which is fed into a single `NAIM+ODST` encoder trained end-to-end on the combined input. This baseline tests whether a single missing-aware encoder trained jointly on the concatenated feature space can match the performance of architectures that process each modality independently before fusion.
- **Intermediate fusion (ConcatODST):** In this proposed approach, the architecture is identical to early fusion, but optimization is performed end-to-end. The unimodal encoders remain trainable, allowing the backpropagation of gradients from the multimodal ODST head through to the modality-specific branches. This enables the encoders to adapt their feature representations to maximize cross-modal synergy.
- **Late fusion** (`NAIM + ODST`^{n}): Integration at the prediction level. Independent unimodal models (`NAIM+ODST`) are trained separately for each available modality. At inference time, we employ a ranking-based ensemble: risk scores from each unimodal model are converted into ranks (ordering the cohort from lowest to highest risk). The final multimodal risk score is the sum of these ranks:

$$\mathrm{Rank}_{\text{final}} = \sum_{m \in \mathcal{M}} \mathrm{Rank}\left(y_m\right),$$

where $\mathcal{M}$ is the set of available modalities. This non-parametric aggregation assumes no synergistic interaction between modalities.

Experimental setup. To mitigate overfitting and ensure robust performance estimation, we employed a stratified fivefold cross-validation strategy. The dataset was partitioned into five non-overlapping folds, preserving the ratio of events to censored patients. In each iteration, three folds were used for training, one for validation and one for testing. This procedure was repeated five times, and final metrics are reported as the mean ± standard deviation. Optimization utilized the *AdamW* optimizer (weight decay 1×10^{-5}) with a batch size of 32 for all the training approaches. Training ran for a maximum of 500 epochs with early stopping (patience equal to 50 epochs) based on the loss computed on the validation set. To ensure a fair comparison across fusion paradigms, we adopted a unified two-stage training protocol. In the early fusion setting (ConcatODST*), the `NAIM` encoders were initialized using weights obtained during unimodal pre-training and kept frozen throughout training. In the intermediate fusion setting (ConcatODST), the encoders were initialized identically but remained trainable, enabling end-to-end optimization in which gradients from the survival loss refine modality-specific representations.

The learning rate schedule (η) was defined within the range $\eta_{\min} = 10^{-8}$ to $\eta_{\max} = 10^{-5}$ and consisted of three components. First, a linear warm-up over the initial 50 epochs increased η from $\eta_{\min}$ to $\eta_{\max}$ to avoid early training instability. After warm-up, a plateau detector monitored the validation loss; if no improvement was observed for 20 epochs, a decay schedule was initiated to anneal η back to $\eta_{\min}$ across 12 steps. Training was halted if the validation metric failed to improve for 50 consecutive epochs.

During end-to-end optimization of the intermediate fusion model, we adopted differential learning rates to balance stability and plasticity. The multimodal ODST and fully connected layers were updated using the full learning-rate range (up to $\eta_{\max} = 10^{-5}$), while the pretrained unimodal encoders used a tenfold lower rate (effective $\eta_{\max} = 10^{-6}$), preserving established feature representations while allowing controlled adaptation during cross-modal alignment. All models were implemented in PyTorch and trained on a single *NVIDIA A40 GPU* (48GB VRAM).

Evaluation metrics. Model discrimination was quantified using the Concordance Index (C-index), which measures the probability that, for a random pair of subjects, the subject who experienced the event earlier has a higher predicted risk score. To rigorously account for potential censoring bias, we also reported Uno's C-index[45]. The latter relies on Inverse Probability of Censoring Weights (IPCW) to correct for the unequal probability that individuals remain uncensored over time. By weighting each observation by the inverse of its estimated probability of being uncensored, IPCW compensates for informative censoring and yields a consistent measure of discriminative performance even when the censoring mechanism depends on patient covariates. Lastly, to assess predictive performance over the follow-up period, we computed the time-dependent Area Under the Curve (td-AUC). Specifically, we evaluated the mean cumulative dynamic AUC at the 25th, 50th, and 75th percentiles of the observed event times in the test set.

Statistical analysis. Continuous variables are reported as mean ± standard deviation, categorical variables as frequencies (n, %). Baseline comparisons utilized the Chi-square test (categorical) and one-way ANOVA (continuous). Survival curves were estimated via the Kaplan–Meier method and compared using the log-rank test. Statistical significance was set at a two-sided $p < 0.05$. All analyses were conducted in Python (sciPy, lifelines and scikit-learn).

Data availability

The dataset analyzed in this study was collected and used under institutional ethical approval and informed patient consent. The data are available upon reasonable request for non-commercial research purposes only, addressed to the corresponding authors.

Code availability

The code implementing the proposed framework is publicly available at the following link: https://github.com/arco-group/Handling-Missing-Modalities-in-Multimodal-Survival-Prediction-for-Non-Small-Cell-Lung-Cancer.

Received: 31 December 2025; Accepted: 13 May 2026;
Published online: 22 June 2026

References

1. Huang, Q. et al. Advances in molecular pathology and therapy of non-small cell lung cancer. *Signal Transduct. Target. Ther.* **10**, 186 (2025).
2. Vincent, T. et al. Mechanistic insights into molecular targeted therapy and immunotherapy for lung cancer. *Biochem. Biophys. Res. Commun.* **776**, 152204 (2025).
3. Yuan, X. et al. Systematic review and meta-analysis of artificial intelligence for image-based lung cancer classification and prognostic evaluation. *NPJ Precis. Oncol.* **9**, 300 (2025).
4. Guarrasi, V. et al. Beyond unimodal analysis: multimodal ensemble learning for enhanced assessment of atherosclerotic disease progression. *Comput. Med. Imaging Graph.* **124**, 102617 (2025).
5. Guarrasi, V. et al. Multimodal explainability via latent shift applied to COVID-19 stratification. *Pattern Recognit.* **156**, 110825 (2024).
6. Captier, N. et al. Integration of clinical, pathological, radiological, and transcriptomic data improves prediction for first-line immunotherapy outcome in metastatic non-small cell lung cancer. *Nat. Commun.* **16**, 10 (2025)
7. Guedes, J. et al. A perspective on integrating digital pathology, proteomics, clinical data and AI analytics in cancer research. *J. Proteomics* **320**, 105493 (2025).
8. Paverd, H., Zormpas-Petridis, K., Clayton, H., Burge, S. & Crispin-Ortuzar, M. Radiology and multi-scale data integration for precision oncology. *NPJ Precis. Oncol.* **8**, 158 (2024).
9. Lobato-Delgado, B., Priego-Torres, B. & Sanchez-Morillo, D. Combining molecular, imaging, and clinical data analysis for predicting cancer prognosis. *Cancers* **14**, 3215 (2022).
10. Peng, J., Pond, G. R., Donovan, E. K., Ellis, P. M. & Swaminath, A. A comparison of radiation techniques in patients treated with concurrent chemoradiation for stage III non-small cell lung cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **106**, 985–992 (2020)
11. Doan, L. M. T. et al. Bridging modalities with AI: a review of ai advances in multimodal biomedical imaging. *Commun. Eng.* **5**, 30 (2026).
12. Niu, Y. et al. Multimodal deep learning integrating tumor radiomics and mediastinal adiposity improves survival prediction in non-small cell lung cancer: A prognostic modeling study. *Cancer Med.* **14**, e70125 (2025).
13. Machin, D., Cheung, Y. B. & Parmar, M. *Survival Analysis: A Practical Approach* 2nd edn (Wiley, 2006).
14. Soenksen, L. et al. Integrated multimodal artificial intelligence framework for healthcare applications. *Nat. Comput. Sci.* **5**, 149 (2022)
15. Caruso, C. M., Guarrasi, V., Ramella, S. & Soda, P. A deep learning approach for overall survival prediction in lung cancer with missing values. *Comput. Methods Prog. Biomed.* **254**, 108308 (2024).
16. Lee, J. et al. Gliosurv: interpretable transformer for multimodal, individualized survival prediction in diffuse glioma. *npj Digit. Med.* **8**, 660 (2025).
17. Graf, R., Todd, S. & Baksh, M. Comparison of the Cox proportional hazards model and Random Survival Forest algorithm for predicting patient-specific survival probabilities in clinical trial data. Preprint at https://arxiv.org/abs/2502.03119 (2025).
18. Yang, H. et al. Mmsurv: a multimodal multi-instance multi-cancer survival prediction model integrating pathological images, clinical information, and sequencing data. *Brief. Bioinform.* **26**, bbaf209 (2025).
19. Wu, Y. et al. Multimodal learning for non-small cell lung cancer prognosis. *Biomed. Signal Process. Control* **106**, 107663 (2025).
20. Oh, S., Kang, S.-R., Oh, I.-J. & Kim, M.-S. Deep learning model integrating positron emission tomography and clinical data for prognosis prediction in non-small cell lung cancer patients. *BMC Bioinformatics* **24**, 39 (2023).
21. Yang, H. et al. Multimodal deep learning approaches for precision oncology: a comprehensive review. *Brief. Bioinform.* **26**, bbae699 (2024).
22. Bommasani, R. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258 (2021).
23. Zhang, S. & Metaxas, D. N. On the challenges and perspectives of foundation models for medical image analysis. *Med. Image Anal.* **91**, 102996 (2023).
24. Li, X. et al. Artificial general intelligence for medical imaging analysis. *IEEE Rev. Biomed. Eng.* **18**, 113–129 (2025).
25. Lu, M. Y. et al. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 19764–19775 (IEEE, 2023).
26. Ding, T. et al. A multimodal whole-slide foundation model for pathology. *Nat. Med.* **31**, 3749–3761 (2025).
27. Blankemeier, L. et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature* **652**, 1318–1328 (2026).
28. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181 – 188 (2024).
29. Hamamci, I. E. et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nat. Biomed. Eng.* 1–19 (2026).
30. Pai, S. et al. Vision foundation models for computed tomography. Preprint at https://arxiv.org/abs/2501.09001 (2025)
31. Ruffini, F., Ayllon, E. M., Shen, L., Soda, P. & Guarrasi, V. Benchmarking foundation models and parameter-efficient fine-tuning for prognosis prediction in medical imaging. *Comput. Methods Programs Biomed.* **275**, 109196 (2025).
32. Caruso, C. M., Soda, P. & Guarrasi, V. Not another imputation method: a transformer-based model for missing values in tabular datasets. *AI Open* **7**, 96–122 (2026).
33. Lipkova, J. et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* **40**, 1095–1110 (2022).
34. Lee, C. et al. Impact of egfr inhibitor in non-small cell lung cancer on progression-free and overall survival: a meta-analysis. *J. Natl. Cancer Inst.* **105 9**, 595–605 (2013).
35. Spigel, D. et al. Five-year survival outcomes from the pacific trial: durvalumab after chemoradiotherapy in stage iii non-small-cell lung cancer. *J. Clin. Oncol.* **40**, 1301 – 1311 (2022).
36. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
37. K, M. S. et al. Survival analysis with cox proportional hazards model in predicting patient outcomes. In *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)* 1155–1161 (IEEE, 2024).
38. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).
39. Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. M. & van der Laan, M. J. Survival ensembles. *Biostatistics* **7**, 355–373 (2006).
40. Popov, S., Morozov, S. & Babenko, A. Neural oblivious decision ensembles for deep learning on tabular data. Preprint at https://arxiv.org/abs/1909.06312 (2019).
41. Zhao, Q. et al. Clinical multi-dimensional prognostic nomogram for predicting the efficacy of immunotherapy in nsclc. *Sci. Rep.* **14**, 21380 (2024).
42. Detterbeck, F. C. et al. The proposed ninth edition tnm classification of lung cancer. *Chest* **166**, 882–895 (2024).
43. Daly, M. et al. Management of stage III non-small-cell lung cancer: ASCO guideline. *J. Clin. Oncol.* **40**, 1356 – 1384 (2021).

44. Tong, L., Mitchel, J., Chatlin, K. & Wang, M. D. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med. Inform. Decis. Mak.* **20**, 225 (2020).
45. young Park, S., Park, J. E., Kim, H. & Park, S. H. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). *Korean J. Radiol.* **22**, 1697 – 1707 (2021).

Acknowledgements

Filippo Ruffini is a PhD student enrolled in the National PhD in AI, XXXVIII cycle, course on Health and Life Sciences, organized by the University Campus Bio-Medico di Roma. This work was partially supported by: i) PRIN 2022 MUR 20228MZFAA- AIDA (CUP C53D23003620008, CUP H53D23003480006), ii) PNRR MUR project PE0000013-FAIR, iii) PNRR – DM 118/2023, iv) Kempe Foundation under grant no. JCSMK24-0094, v) Strategic University Projects "IDEA: AI-powered Digital Twin for next-generation lung cancEr cAre" financed by Università Campus Bio- Medico di Roma 2023 (GEN0469). Resources are provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Alvis @ C3SE, partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Conceptualization, F.R., C.M.C., V.G., P.S.; methodology, F.R., C.M.C., V.G., P.S.; software, F.R., C.M.C., V.G., P.S.; validation, F.R., C.M.C., V.G., P.S.; formal analysis, F.R., V.G., P.S.; investigation, F.R., C.M.C., C.T., B.B.Z., B.V., C.M., A.B., E.F., S.R., V.G., P.S.; resources, F.R., C.M.C., V.G., B.B.Z., P.S.; data curation, F.R., C.M.C., C.T., C.G., F.M., M.L., E.I., B.B.Z., M.F., B.B.Z., A.C., G.P., B.V., S.R.; writing-original draft preparation, F.R., S.R., V.G., P.S.; writing-review and editing, F.R., S.R., V.G., P.S.; visualization, C.M.C., C.T., F.M., M.L., A.B., C.M., S.R., V.G., P.S.; funding, S.R., P.S.; supervision, S.R., V.G., P.S. All authors reviewed the manuscript.

Funding

Open access funding provided by Umea University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-02783-3.

Correspondence and requests for materials should be addressed to Filippo Ruffini, Valerio Guarrasi or Paolo Soda.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

[1]Unit of Artificial Intelligence and Computer Systems, Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy. [2]Department of Diagnostics and Intervention, Radiation Physics, Biomedical Engineering, Umeå University, Umeå, Sweden. [3]Operative Research Unit of Radiation Oncology, Fondazione Policlinico Universitario Campus Bio-Medico di Roma, Rome, Italy. [4]Anatomical Pathology Operative Research Unit, Fondazione Policlinico Universitario Campus Bio-Medico di Roma, Rome, Italy. [5]Research Unit of Anatomical Pathology, Department of Medicine and Surgery, Università Campus Bio-Medico di Roma, Rome, Italy. [6]Department of Engineering 'Enzo Ferrari', University of Modena and Reggio Emilia, Modena, Italy. [7]Research Unit of Radiation Oncology, Department of Medicine and Surgery, Università Campus Bio-Medico di Roma, Rome, Italy. [8]Department of Medicine and Surgery, Università Campus Bio-Medico di Roma, Rome, Italy. [9]Operative Research Unit of Medical Oncology, Fondazione Policlinico Universitario Campus Bio-Medico di Roma, Rome, Italy. [10]Operative Research Unit of Radiology and Interventional Radiology, Fondazione Policlinico Universitario Campus Bio-Medico di Roma, Rome, Italy. [11]University of Cassino and Southern Lazio, Cassino, Italy. [12]These authors contributed equally: Filippo Ruffini, Camillo Maria Caruso. [13]These authors jointly supervised this work: Sara Ramella, Valerio Guarrasi, Paolo Soda.
✉e-mail: filippo.ruffini@unicampus.it; valerio.guarrasi@unicampus.it; paolo.soda@umu.se