Original Research

Robust Multimodal Fusion for Survival Prediction in Cancer Patients

Cancer Informatics
Volume 24: 1–23

DOI: 10.1177/11769351251376192
journals.sagepub.com/home/cix
Sage

Dominic Flack[1], Aakash Tripathi[2], Asim Waqas[2], Ghulam Rasool[2] and Dimah Dera[1]

Abstract

Objectives: Multimodal deep learning models have the potential to significantly improve survival predictions and treatment planning for cancer patients. These models integrate diverse data modalities using early, intermediate, or late fusion techniques. However, many existing multimodal models either underperform or show only marginal improvements over unimodal models. To establish the true efficacy of multimodal survival prediction models, it is critical to demonstrate consistent and substantial advantages over unimodal counterparts.

Methods: In this paper, we introduce the Robust Multimodal Survival Model (RMSurv), a novel discrete late fusion model that leverages synthetic data generation to compute time-dependent weights for various modalities. RMSurv utilizes up to 6 distinct data modalities from The Cancer Genome Atlas Program (TCGA) non-small cell lung cancer and the TCGA pan-cancer datasets to predict overall survival over a period of 10 years. The key innovations of RMSurv are the calculation of time-dependent late fusion weights using a synthetically generated dataset and a new statistical feature normalization technique to enhance the interpretability and accuracy of discrete survival predictions. We evaluate the performance of the proposed method and several alternatives with cross validation using the concordance index, and vary the number of modalities included. We also create a late fusion simulation to highlight the complex relationships of multimodal fusion.

Results: In our experiments, RMSurv outperforms the best unimodal model's Concordance index (C-Index) by 0.0273 on the 6-modal TCGA Lung Adenocarcinoma (LUAD) dataset. Existing late and early fusion methods improved the C-index by only 0.0143 and 0.0072, respectively. RMSurv also performs best on the combined TCGA non-small-cell lung cancer dataset and the TCGA pan-cancer dataset.

Conclusions: These advancements underscore RMSurv's potential as a powerful approach for survival prediction, establishing robust multimodal benefits, and setting a new benchmark for survival prediction models in pan-cancer settings.

Keywords
multimodal, cancer, survival prediction, fusion

Received: 14 April 2025; accepted: 14 August 2025

Introduction

Background

Multimodal learning in oncology is an emerging research area with great potential to improve cancer research and patient care. Multimodality refers to various types of data, including but not limited to radiological and diagnostic imaging, clinical and demographic data, histopathology slides, or molecular information. Multi-omics analysis uses data from genomics, transcriptomics, and similar fields for medical research. Survival prediction, a critical aspect of cancer research, involves estimating how long a patient is likely to live after diagnosis or treatment, aiding in personalized treatment planning, resource allocation, and clinical trial design. Survival prediction models are best evaluated using the concordance index, or C-index, which measures the fraction of pairs of predicted risk scores that match the ground truth. Several existing fusion methods merge

[1]Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, NY, USA
[2]Department of Machine Learning, Moffitt Cancer Center, Tampa, FL, USA

Corresponding author:
Dominic Flack, Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, 54 Lomb Memorial Dr, Rochester, NY 14623, USA.
Emails: daf6674@rit.edu; dominic416@gmail.com

heterogeneous data modalities such as clinical records, -omics data, and histopathology images for survival prediction.[1,2] Survival models are typically categorized as Cox and discrete models. Cox-based models estimate a single risk score for a Cox proportional hazards model, which will be converted into a survival probability over time.[3] Discrete models directly output hazards across multiple discrete time periods, and the predicted survival is calculated as the cumulative product of the complement of the hazard.[4] The data from various modalities can be fused at different stages using early, intermediate, or late fusion methods. The simplest method is early fusion, in which data from all modalities are combined into a single feature vector used as the model input.[1] Intermediate fusion combines features within the intermediate or hidden layers of the machine learning model and allows for modeling complex interactions between data modalities.[1] Late fusion involves independently training unimodal models for each modality and subsequently creating an ensemble for the final outputs. Given the data-driven nature of fusion methods, determining the optimal approach can be challenging and application dependent. In this paper, we provide a comprehensive comparison of various fusion methods to investigate the differences in their performance.

The key challenge for existing data fusion models is to consistently achieve a multimodal advantage, meaning performance for a given task that surpasses the best unimodal model. Even if the model produces a multimodal advantage for some modalities (eg, imaging, genomics, etc.), it won't be helpful in real-world medical decision-making unless its performance clearly exceeds that of the unimodal model based on readily available data like a patient's age, gender, or cancer stage. Therefore, studies that show a multimodal advantage for carefully selected modalities and exclude clinical data or other outlier high-performing modalities are limited in their potential for future clinical application. Existing intermediate and early multimodal fusion models often demonstrate a multimodal advantage when using the ideal combination of a maximum of 2 to 3 modalities,[5-8] but experience a sharp degradation in performance when additional weaker modalities are added.[4] These models also experience inconsistent performance, where some datasets demonstrate a multimodal disadvantage when the architecture, selected modalities, and number of modalities are not optimal.[8] This problem fundamentally limits the potential of employing machine learning for cancer survival prediction. Just as a physician can consider dozens of data types in their prognoses, the ideal system should be robust enough to extract signal from an unlimited number of modalities without being heavily affected by the noise.

A potential cause of the low robustness to weak modalities is the optimization strategy of early and intermediate fusion, which minimizes a combined loss function based on the training accuracy of the multimodal model. For small and noisy cancer cohort datasets, where the training-set cases and test-set cases will differ considerably, the overfitting of each modality will compound when they are all fused together as a single model.

A promising solution proposed in a recent work[9] is late fusion, in which each modality will be used for training a separate model using a distinct loss function, and the weighting of modalities will be determined based on validation-set C-index instead of the optimization based on training-set C-index. This method showed robustness and a modest multimodal advantage in up to 6 modalities across many datasets, but is limited by its single validation set, ad-hoc analytical weight calculation, lack of time-dependent weighting, and absence of normalization.[9]

In this paper, we propose the Robust Multimodal Survival Model (RMSurv), which uses a synthetically generated dataset to empirically optimize the weighting for discrete late fusion. We further improve the model by using time-dependent weights to represent the performance of each modality over time and normalize the output to correct the distribution. We present multiple variations of late fusion methods and compare them to existing methods on 3 datasets with a varying number of data modalities. We also present a simulation to reveal the underlying correlation-based relationships of late fusion, and introduce a novel pathology report embedding modality, which shows promising results for a new class of text-based survival prediction modalities.

Related Work

Several machine learning methods for survival prediction are based on the Cox proportional hazards model.[3] In the Cox model, covariates such as age, gender, and cancer stage are weighted and linearly combined to calculate an exponential risk score, which scales the baseline hazard function.[3] Ching et al[10] developed Coxnnet, which used a neural network to transform high-dimensional -omics data into lower-dimensional features that were used as covariates for the Cox model. Since the Cox-nnet model compresses the features into a single risk score, variables have proportional, time-invariant effects on the baseline hazard and survival predictions.[4]

Vale-Silva and Rohr[4] addressed the proportionality problem with the MultiSurv model. The MultiSurv model showed improved performance by replacing the single risk score output with multiple discrete risk outputs, allowing the model to directly calculate survival probabilities over different time periods. This approach captured time-varying influences of variables so that factors like cancer stage might be more influential in early years, while age and other features could be more dominant in later years. The paper also highlighted the modality inclusion problem for survival models.[4] Six modalities, including clinical, multi-omics, and whole slide images (WSI), were available, but the model performed best when only using 2 modalities: clinical data and gene expression.[4]

Intermediate fusion methods have the advantage of modeling rich cross-modal interactions.[1] Chen et al[11] integrated the Kronecker product into their "Pathomic Fusion" model to maintain unimodal features while generating features for each cross-modal interaction. They later integrated discrete outputs with the same intermediate fusion method

in the PORPOISE model.[8] The number of features scales exponentially with the number of modalities, making this method infeasible for 6 or more modalities. The concatenation method (early fusion) was shown to outperform the "Pathomic Fusion" model on some datasets such as TCGA Lung Adenocarcinoma (LUAD).[11] For other datasets like TCGA Lung Squamous Cell Carcinoma (LUSC), the best unimodal model outperformed the intermediate multimodal fusion model.[8] Li et al[12] developed another intermediate fusion model, HFBSurv, that utilized attentional factorized bilinear modules to model unimodal and bimodal interactions more efficiently than the Kronecker product.

Several improvements to unimodal architectures have been incorporated into intermediate fusion systems to further increase performance. Gomaa et al[6] incorporated a vision transformer model for MRI images and a fully connected network for clinical data in their multimodal fusion model. The fusion method used cross-attention mechanisms and non-proportional discrete survival predictions, and demonstrated strong multimodal benefits in glioblastoma datasets.[6] Their model showed modest improvements in the bi-modal comparison to existing methods in both unimodal and multimodal setups.[6] Another recent study by Luo et al[5] proposed combining a vision transformer whole-slide-image model with genomic features in an intermediate fusion model. They added an additional layer after feature concatenation, which uses the Dempster–Shafer theory to assess the uncertainty of each modality in the final prediction.[5] This model outperformed several other multiple instance learning methods in a multimodal comparison with 3 cancer types.[5] Yang et al[7] adopted a similar approach in their MMsurv model, which used a novel bilinear pooling and transformer fusion layer and a 2-step multi-instance learning approach. In their experiment, the model outperformed existing unimodal and multimodal methods on 4 out of 6 datasets.[7]

One potential limitation of early and intermediate fusion approaches are the highly variable performance based on the dataset and number of modalities included.[4,8] Late fusion, by contrast, has shown promise in providing a robust multimodal advantage across many datasets and a number of modalities.[9] Furthermore, the available data types vary significantly across cancer types, so an approach that can easily and consistently incorporate any combination of modalities and unimodal architectures will be advantageous.[13] Late fusion methods do not model feature interactions between modalities, but instead train each unimodal model separately, and use an ensemble to fuse independent predictors of survival.[1,9] The ensemble approach can easily exclude missing modalities, prevents outsize influence of high-dimensional modalities, and performs well with heterogenous and weakly correlated modalities.[1,14] Some models, such as MultiSurv, combine features at the final layers, but instead of using an ensemble, they train all sub-models as 1 model with a common loss function.[4] By doing this, the relative weights of modalities are based on the training set accuracy and can cause overfitting.

Nikolaou et al[9] showed a multimodal advantage on 25 of 33 datasets tested using the simple late fusion method "AZ-AI multimodal pipeline." In this fusion model, 1 validation set C-index is calculated for each modality, and the linear combination weight for each modality is set by subtracting 0.5 from each C-index value and normalizing them.[9] Since this method only estimates the complex relationship when combining modalities, and only uses 20% of the training data for validation, there is significant room for improvement. In particular, this strategy ignores the impact of the correlation between predictions and the time-dependent accuracy of predictions. Furthermore, the model simply excludes modalities with validation C-index below 0.52, which is another ad-hoc assumption that can be improved using the proposed method. Despite these limitations, the method provides a very consistent multimodal advantage for up to 6 modalities included, even on small datasets.[9] The authors of this study concluded that different multimodal fusion methods are better for different settings, and that late fusion is best suited for applications where the risk of overfitting is high, such as small sample sizes.[9]

Methods

Late Fusion Simulation

Intuition suggests that the performance of a combination of 2 predictions depends on the accuracy of each prediction and the correlation between the 2. Combining nearly identical, highly correlated predictions will not add signal to the combined prediction. Likewise, linear combinations with low correlation can benefit from the independent signal of each modality. However, even with zero correlation, a survival prediction with very low C-index will just add noise to a highly accurate survival prediction. Therefore, an ad-hoc relationship that uses only the C-index as an input to calculate late fusion weights is not capable of modeling the true empirical relationship. Here we describe a fully synthetic dataset, distinct from RMSurv, to simulate late fusion and ground this intuition. We use the results of this simulation to explain the need for an empirical strategy like RMSurv, and to explain why adding more modalities to a model often decreases performance in multimodal fusion research. These results are shown in the "Late Fusion Simulation Results" Section. We also use this simulated dataset to calculate weights in our "synthetic weights" alternative weight calculation method, which we explain in the "Alternative Weight Calculation Options" Section.

For this simulation, we generate synthetic risk scores and survival times with arbitrarily set C-indices and cross-modality correlations by sampling from a multivariate normal distribution. This method does not directly use C-index as an input but instead requires a positive semi-definite covariance matrix. To get around this, we need to model the non-linear C-index as a linear correlation. We achieve this by converting the C-index of each modality into a Pearson correlation between the modality and the survival times using analytical estimates. We run a binary search algorithm to repeatedly generate distributions to correct for errors in the analytical estimates and match the C-index to its corresponding correlation metric. This approximation of

the C-index as a Pearson correlation results in negligeable error in a 2-dimensional simulation, but results in some small unavoidable error between the desired and actual C-indices and correlations when using 6 modalities within the simulation. After assigning a Pearson correlation to each modality, we add a new row and column to the existing Pearson correlation matrix to combine these into a unified matrix that represents cross-modality correlations between risk scores and correlations between risk scores and survival times. We then find the nearest positive semidefinite matrix and generate our normally distributed samples. We apply an iterative process to reduce the error between the desired and actual C-indices and correlations, then we can test the performance with varying weights given to each modality.

Binary Search Procedure. In the initialization step, we set the lower bound of Spearman's correlation, $\rho_{s_{low}}$, to 0, and the upper bound, $\rho_{s_{high}}$, to .9999. We also define a tolerance level, $\text{tol} = 1\times10^{-4}$, to determine convergence. During each iteration, we calculate the midpoint Spearman's correlation,

$$\rho_s = \frac{\rho_{s_{low}} + \rho_{s_{high}}}{2},$$

and convert ρ_s to an approximate Pearson correlation,

$$\rho_p = 2\sin(\frac{\pi}{6}\rho_s).$$

We then set ρ_p of the other modalities to zero, generate synthetic data using ρ_p, and compute the observed C-index, $c_{observed}$, between the generated risk score and survival times. If $c_{observed} > c_{desired}$, we update $\rho_{s_{high}} = \rho_s$. If $c_{observed} < c_{desired}$, we update $\rho_{s_{low}} = \rho_s$. This process continues until $|c_{observed} - c_{desired}| < \text{tol}$ or the maximum number of iterations is reached.

Generating Synthetic Data Using a Gaussian Copula. We begin by constructing an initial symmetric Pearson correlation matrix, **P**, for M modalities. For instance, when $M = 3$, **P** is a 3×3 matrix defining the correlations between each pair of modalities. We then add an additional row and column to incorporate survival time correlations, resulting in an $(M+1)\times(M+1)$ matrix, $\mathbf{\Sigma}_p$. This expanded matrix takes the form:

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix} \rightarrow \Sigma_p = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{1T} \\ \rho_{21} & 1 & \rho_{23} & \rho_{2T} \\ \rho_{31} & \rho_{32} & 1 & \rho_{3T} \\ \rho_{T1} & \rho_{T2} & \rho_{T3} & 1 \end{bmatrix},$$

where ρ_{iT} represents the Pearson correlation between modality i and the survival time T.

Next, we verify that Σ_p is semipositive definite, because this is required for generating valid multivariate normal samples. If Σ_p is not semipositive definite, we adjust it to the nearest positive definite matrix using Higham's[15] algorithm. We then use the adjusted Σ_p to generate N samples from a multivariate normal distribution, where each sample corresponds to risk scores for M modalities and a single survival time.

After drawing these samples, we transform each normal variable Z_i into a uniform variable U_i using the standard normal cumulative distribution function, Φ:

$$U_i = \Phi(Z_i),\ i = 1,\ldots,M,T.$$

We then apply the desired marginal distributions. For each modality i, we map U_i back to a standard normal distribution by $R_i = \Phi^{-1}(U_i)$. For the survival times, we map U_T to an exponential distribution with rate parameter λ:

$$T = -\frac{\ln(1-U_T)}{\lambda}.$$

Optimizing Weights with Population-Based Search. We define a combined risk score $R_{combined}$ by summing each modality's risk score R_i, weighted by w_i:

$$R_{combined} = \sum_{i=1}^{M} w_i R_i.$$

Our objective is to maximize the C-index of $R_{combined}$ with respect to the survival time T:

$$\max_{\mathbf{w}} c_{combined} = C - index(R_{combined}, T).$$

In a 2-dimensional simulation, we simply compute the combined C-index at 100 relative weights ranging from 0 to 1. However, in higher dimensions where local minima may appear, we use the differential evolution algorithm. We set a population size of 15, a tolerance of 10^{-6}, and a maximum of 100 iterations to find the ideal weights for combining the risk scores.

Data Pre-processing

In our experiments, we use the non-small-cell lung cancer types LUAD and LUSC from the Cancer Genome Atlas (TCGA) database, a large public database of cancer data collected from 2006 to 2015.[16] We also use the TCGA pan-cancer dataset, which includes all 33 cancer types.[17] We separate this data into 3 sets of varying sizes to assess the impact of dataset size on multimodal performance. LUAD, our first dataset, contains 522 cases, with 188 cases uncensored (deceased), and a median follow-up time of 21.6 months.[17] LUSC contains 504 cases and 219 uncensored cases, and a median follow-up time of 21.9 months.[17] Our second dataset is the combined LUAD + LUSC dataset, with a total of 1026 cases and 407 uncensored cases.[17] The pan-cancer (PAN) dataset contains 11 060 cases, and 3622 uncensored cases.[17]

We use 7 total input modalities to evaluate the performance of the fusion methods: clinical data, pathology reports, gene expression, miRNA, DNA methylation, protein expression, and somatic mutation. Gene expression (measuring mRNA) and miRNA are transcriptomic factors, which regulate the expression of genes in cancer cells.[16] DNA methylation is an epigenomic factor that can change gene activity through the addition of methyl groups to DNA.[18] Protein expression measures the presence of specific proteins in cancer cells,[19] and somatic mutation measures alterations to the DNA sequence in cancer cells.[16] To

Table 1. Discretization of Clinical Data Categories.

Attribute	Category	Numeric value
Age	Integer	Integer
Gender	Male	1
	Female	2
Race	White	1
	Asian	2
	Black or African American	3
	Not reported	4
	American Indian or Alaska Native	5
Stage	Stage 0	1
	Stage I	10
	Stage IA	11
	Stage IB	12
	Stage IC	13
	Stage II	20
	Stage IIA	21
	Stage IIB	22
	Stage IIC	23
	Stage III	30
	Stage IIIA	31
	Stage IIIB	32
	Stage IIIC	33
	Stage IV	40
	Stage IVA	41
	Stage IVB	42
	Stage IVC	43
	Not Reported	50

create 6 modalities with similar C-indices for LUAD and LUSC, we concatenate the features from the 2 lowest performing modalities, protein expression, and somatic mutation, into a single modality, which is shortened to "protein expression" in the Results Section for simplicity. For PAN, protein expression performed well as a standalone modality and somatic mutation is excluded. Another important step in the pre-processing of the TCGA data was the removal of the many duplicate cases in Xena[20] and MINDS,[21] where a single patient's -omic and clinical data are saved at different points in time. This can skew results of studies when many patients appear in both the test set and the training set, so we removed these extra cases. The reporting of this study conforms to the TRIPOD-AI statement[22] (see Supplemental File 1).

For the clinical data modality, we use the Multimodal Integration of Oncology Data System (MINDS) database[21] to download the categorical dataset, then discretize it, as shown in Table 1, to produce a vector with 4 values: age, gender, race, and stage.

The pathology report PDF, also downloaded using MINDS, requires more preprocessing, and we use the HoneyBee framework[23] to convert the information into a usable format. This program extracts the text of the PDF and inputs this into the GatorTron-Large transformer model,[24] which produces an embedding with 3584 values. We also tested the HoneyBee method on the clinical data text generated with additional categories, but this underperformed the discretized 4-category method by a wide margin (C-index = 0.582 vs 0.646 on LUAD).

The remaining modalities are tabular-omics data downloaded from the UCSC Xena website.[20] We do not apply any manual feature selection. Instead, we reduce the number of features by removing duplicate features, features with many constant values, and features with very low variability. The thresholds for feature removal were adjusted for each modality based on the number of features removed and the performance of the unimodal models in cross-validation. The feature selection was performed on the LUAD + LUSC dataset and pan-cancer dataset, and the features for the LUAD dataset are inherited from LUAD + LUSC. Table 2 shows the resulting number of features for each modality. We apply median imputation for missing values, and the system performs comparably when zero-filling missing values as well. Normalization was not applied to the features before input into the model. We do not use any right-censored cases to calculate training loss. For LUAD and LUAD + LUSC, our results are from fivefold cross-validation repeated for 10 seeds. For the pan-cancer dataset, we perform fivefold cross-validation on just 2 seeds due to training time constraints and much lower unimodal variance compared to the other datasets.

Although MRI images, whole slide images, and copy number information were available, we did not include these modalities in the final testing. TCGA LUAD has MRI images for fewer than 10% of patients, so this modality performed poorly. We used embeddings of slide images generated with the UNI pretrained vision transformer model,[25] but the performance on this modality did not exceed 0.53 C-index. Even when fine-tuning on images from LUAD and including multiple-instance learning, the modality frequently performed below 0.5 C-index and was excluded. The unimodal model trained on tabular copy number data performed inconsistently, so this data modality was also excluded.

Unimodal Architecture

For a late fusion model, the ensemble can combine predictions from models with various different architectures. To simplify this study, we use the same architecture in all unimodal models used in late fusion. The unimodal model outputs used in late fusion are exactly the same for each late fusion strategy, but the linear combination weights vary depending on the method. This setup isolates the effect of the late fusion weight calculation method. Our unimodal discrete model uses twenty 6-month time periods and outputs a hazard score for each. This allows the model to account for non-proportional effects of individual features. The survival time of each case is converted into its respective time bin, and survival times exceeding 10 years are set to the final time bin. The negative log-likelihood loss function optimizes the model by increasing the hazard probability at the true time bin of the survival time and decreasing the hazard probability for the preceding time bins. In our preliminary experiments, we modified several existing models into discrete versions, including a self-normalizing

Table 2. Number of Features in Each Modality After Pre-Processing.

Modality	LUAD	LUAD + LUSC	PAN
Clinical data	4	4	4
Pathology report	3584	3584	3584
Gene expression	16829	16829	192958
miRNA	1012	1012	634
DNA methylation	4931	4931	38943
Protein expression			210
Protein expression + somatic mutation	1204	1204	

network, a gradient boosting tree, a simplified fully connected network, and a modified version of the HFBSurv architecture.[12] We achieved the best results with the model based on HFBSurv, which was modified into a discrete unimodal model by removing the cross-modal portion of the architecture and modifying the output layer into 20 discrete hazard outputs instead of the original single risk score output. This results in a simplified model with a series of fully connected input layers with Tanh activation, a modality-specific attentional factorized bilinear module, and a series of fully connected output layers. We use the negative log-likelihood loss function instead of the original Cox partial likelihood loss function, and we also add dropout in the first 2 fully connected layers.

An advantage of the late fusion approach is the ability to tune hyperparameters for unimodal models. We noticed that some unimodal models would experience more overfitting than others when using the same hyperparameters, so the unimodal models were manually tuned with cross-validation by modifying the learning rate (5×10^{-5} to 1.2×10^{-4}), number of epochs (40-100), and number of neurons within the input layers (48-256). This is an overlooked limitation of early and intermediate fusion, which apply the same hyperparameters to all sub-models. These unimodal model hyperparameters were consistent across all late fusion experiments.

Robust Multimodal Survival Model (RMSurv)

We develop a novel, robust multimodal data fusion approach to model the complex relationship between modalities and optimize the weight calculation strategy for discrete late fusion. The process uses nested cross-validation to estimate generalized C-indices, creates a synthetic dataset based on these estimates and the model outputs, and performs a grid search to find the optimal multimodal ensemble weights.

Nested Cross-Validation. One limitation of the existing method is the use of only a single validation set for calculating the C-index of each unimodal model.[9] We propose the use of nested cross validation to use 100% of the training data for validation and achieve C-index predictions with lower variance and lower average difference from the test-set C-indices. The RMSurv method starts by performing a nested fivefold cross-validation on the training set to calculate the average validation C-index for each modality. This involves splitting the training data for each modality into fivefolds and creating 5 new training subsets with onefold held out for validation. For each fold, we train the unimodal models, test them on the designated validation set, and record the C-index for each modality. We then average the C-index across all fivefolds. This only uses ground truth information from the combined training set, but in contrast to the training set C-indices, this represents the generalized accuracy by evaluating the C-indices on the hidden validation sets.

Re-Training with Full Training Set. After the nested cross-validation, the training set is combined without any held-out validation data, and the unimodal models are all trained again. The test set model inputs are passed through the model, and the outputs are recorded. This step is performed in the existing method,[9] and aims to improve performance compared to the nested cross-validation models since 100% of the training data is included.

Sampling Survival Times. Next, we randomly sample ground truth survival times from the full training set, sampling as many survival times as there are test cases. Both censored and uncensored survival times were sampled in our experiments. We align each sampled survival time to a test set case, along with its unimodal model outputs, in a random permutation. This sampling approach perfectly models the actual cross-modality correlations and provides a strong estimate for the distribution of the test set survival times without leaking the actual test-set survival times.

Optimize Survival Time Assignment. Before this step, the distributions and correlations of the synthetic dataset are properly modeled, but the C-indices are randomly initialized and do not match our average validation C-indices calculated with the nested cross-validation. This step will augment the order of the sampled survival times such that the synthetic dataset will inherit the desired validation C-indices. We optimize how survival times are assigned by defining a total loss function based on the squared difference between achieved and desired C-indices:

$$Loss = \sum_{i=1}^{M} \left(c_{achieved,i} - c_{desired,i} \right)^2,$$

where M is the number of modalities, and $c_{desired,i}$ is the average validation C-index of a given modality.

We iteratively improve the survival time assignment by swapping the rank of 2 randomly selected survival times, evaluating the loss after each swap, and accepting the new assignment whenever it reduces the loss. We stop early if the loss drops below a predefined threshold (10^{-6}), and we impose a maximum of 10000 iterations to limit computation. This limit was not reached for the 3 datasets tested. This iterative search results in a simulated dataset that maintains the original cross-correlation between data modalities but replaces the test set ground truth survival times with simulated survival times matching the average validation C-indices.

Optimize Weights via Population-Based Grid Search. Now that the synthetic dataset is complete, we can empirically calculate the ideal weights with a grid search. This avoids the limitations of ad-hoc methods by implicitly considering correlations and number of modalities in addition to C-indices. The predictions for each modality are combined into 1 with a linear combination of model outputs. We optimize the linear combination weights via a population-based grid search. We define an objective function to maximize the C-index $c_{combined}$ for the combined risk score $R_{combined}$ and the survival time T:

$$\max_{\mathbf{w}} c_{combined} = C - index\left(R_{combined}, T\right).$$

Using the differential evolution algorithm,[26] we ensure that all weights are nonnegative and normalized to sum to 1. We set the population size to 15, allowing the mutation factor to vary between 0.5 and 1. We also set the recombination probability to 0.7 and limit the algorithm to a maximum of 100 iterations.

Final Testing. Finally, we test the model with the test set ground truth survival times, which are held out until this step. Figure 1 shows a visual representation of the RMSurv strategy. The detailed steps of the proposed RMSurv method are explained in Algorithm 1.

Time Dependent-RMSurv (TD-RMSurv)

RMSurv and the other late fusion methods described in the "Alternative Weight Calculation Options" section all output a hazard for 20 discrete time bins, each representing a 6-month period over a 10-year time frame. In the baseline RMSurv scheme, the same M late fusion weights are applied to each of the time bins. Because the weights do not change over time, the C-index and correlations of each modality are therefore implicitly modeled as if they were constant over time. This is not necessarily true, and the late fusion weights can in fact be optimized for each individual time bin. Thus, we propose the time-dependent RMSurv model (TD-RMSurv) which takes advantage of the discrete architecture and creates a search space of $M \times 20$ weights, 1 per modality per time bin. The rationale for time-dependent weighting is that just like certain features can be more influential at certain time bins in a discrete unimodal model, certain modalities can also be more influential at certain times. This method cannot be used with an ad-hoc strategy, because C-index is calculated using survival times across all time bins, and C-index within single 6-month periods would be excessively noisy and not calculatable for small datasets.

We define the baseline RMSurv method as the following, where the same weight w_i applies across all time bins:

$$\mathbf{y}_{comb,j} = \sum_{i=1}^{M} w_i \cdot \mathbf{y}_{i,j,norm}.$$

For TD-RMSurv, each modality can have a distinct weight $w_{i,j}$ at each time bin j. We define the TD-RMSurv method as the following:

$$\mathbf{y}_{comb,j} = \sum_{i=1}^{M} w_{i,j} \cdot \mathbf{y}_{i,j,norm}.$$

Here, M is the number of modalities included in the ensemble.

RMSurv Algorithm

Algorithm 1. Pseudocode for RMSurv Algorithm.

Require: number of seeds S, maximum iteration K, threshold ϵ
for seed = 1: S **do**
  for outer_fold = 1:5 **do**
    Partition the dataset into training set $\mathcal{D}$ (80%) and test set $\mathcal{T}$ (20%).
    Partition $\mathcal{D}$ into 5 inner folds.
    for nested_fold = 1:5 **do**
      Train unimodal models on 4 folds and validate on the remaining fold.
      Compute the validation C-index for each modality.
    end for
    Average the C-index results for each modality to compute $c_{desired,m}$
    Retrain all unimodal models on the entire $\mathcal{D}$ and record their outputs for $\mathcal{T}$.
    Randomly sample survival times from $\mathcal{D}$ (size = $|\mathcal{T}|$) and assign them to the test outputs in a random permutation.
    for iteration = 1: K **do**
      Define total loss: $Loss = \sum_m (c_{achieved,m} - c_{desired,m})^2$.
      Randomly swap the rank of two assigned survival times.
      Accept the swap if it reduces Loss stop early if Loss $< \epsilon$.
    end for
    Combine modality outputs linearly with weights $\mathbf{w}$, subject to $\sum_i w_i = 1$,
    and use a population-based grid search to maximize the combined C-index.
    Reveal $\mathcal{T}$'s true survival times and compute final performance.
  end for
  Compute the average performance across all outer folds for this seed.
end for
Compute overall average performance across all seeds.

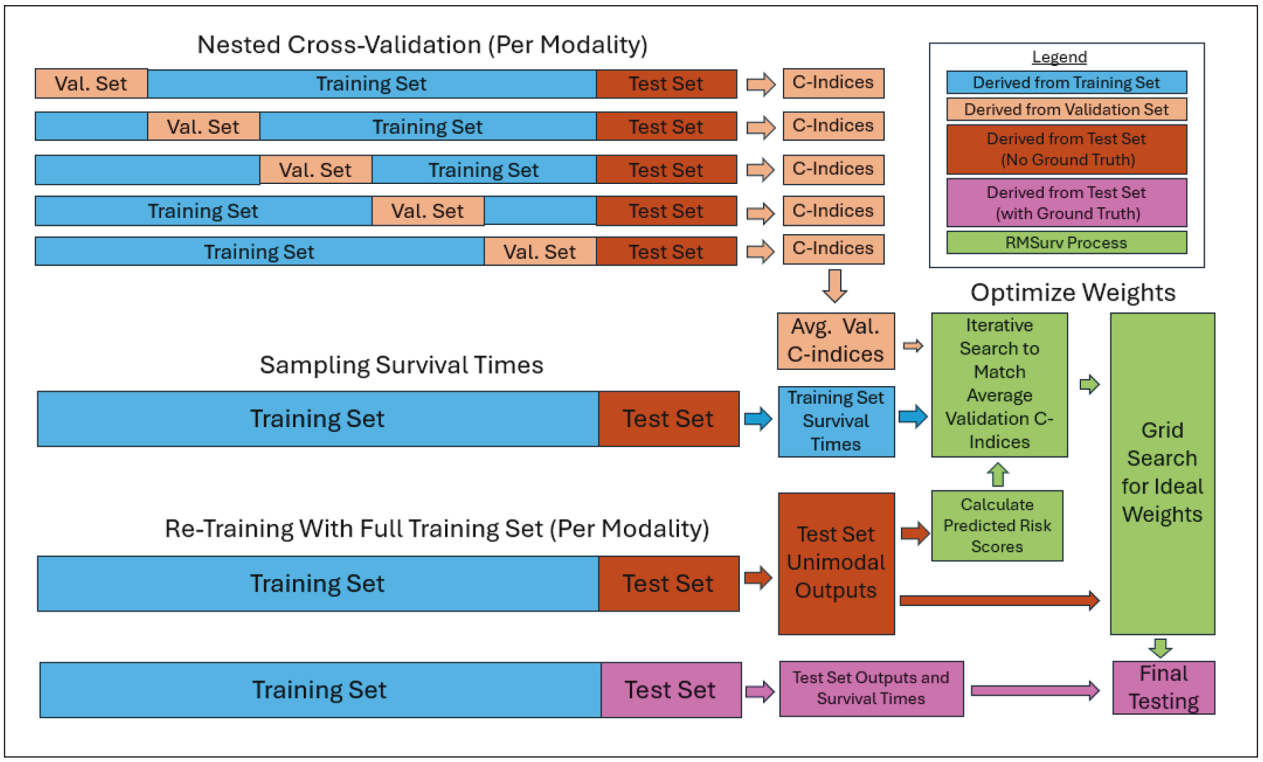

Figure 1. Overview of RMSurv weight calculation scheme. The process starts with a nested cross-validation on the training set (shown in blue) to calculate the average validation accuracy. This provides a representative accuracy for each modality, without introducing overfitting effects. Next, the unimodal models are re-trained on the full training set, and the model outputs are recorded. The survival times are randomly sampled from the training data. The synthetic dataset (shown in green) is generated using a binary search to match the average validation C-indices. Finally, ideal weights for the synthetic dataset are calculated with a grid search, and these weights are applied to the test set with the actual survival times (shown in pink).

Normalization Strategy

In our late fusion experiments we normalize both the unimodal outputs and the unified ensemble outputs, which gives 2 key benefits. First, unimodal models with higher training set C-indices will have greater variance in outputs, so they will have an outsize influence on the linearly combined result. The normalization before the linear combination ensures all unimodal model variances are equal such that there are no nonlinear effects, and the weights are interpretable as their true relative weight on the final output. Second, a linear combination of predictions will reduce the variance as compared to the unimodal outputs. By normalizing after the linear combination, we increase the variance, which avoids all survival predictions being very similar to the mean.

One limitation of the C-index metric is that it only measures the accuracy of the rank, so outputs with biased means and standard deviations will not show any decrease in C-index. This normalization strategy does not significantly change the rank (C-index) of the output risk scores, but it does improve the error as measured by the Integrated Brier Score (IBS). IBS measures squared differences between observed outcomes and predicted survival probabilities over time.[27] The normalization method matches the mean and standard deviation of the test distribution to the ideal training set distribution. Figure 2 shows how we apply this normalization strategy, which is used for each of the late fusion strategies described in "Alternative Weight Calculation Options" for the most fair comparison.

We begin by computing the mean and standard deviation of the training set model outputs for each of the 20-time bins:

$$\mu_{train,j} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_{train,i,j},$$

$$\sigma_{train,j} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{Y}_{train,i,j} - \mu_{train,j}\right)^2}.$$

These statistics are computed for each time bin independently. Next, we define a range of multipliers, for instance, $w_{std} \in \{0.1, 0.2, \ldots, 10.0\}$. For each w_{std}, we normalize the training predictions, perform the weighted linear combination, calculate survival, and then compute the Integrated Brier Score. We select the multiplier that yields the lowest IBS on the training set.

For test data, we normalize the predictions of each modality using the training set statistics and the optimized standard deviation multiplier w_{std}:

$$\mathbf{y}_{norm,i,j} = \left(\frac{\mathbf{y}_{i,j} - \mu_{test,j}}{\sigma_{test,j}}\right)\sigma_{train,j}\, w_{std} + \mu_{train,j}.$$

We then combine the normalized risk scores from all modalities using their respective weights w_i:

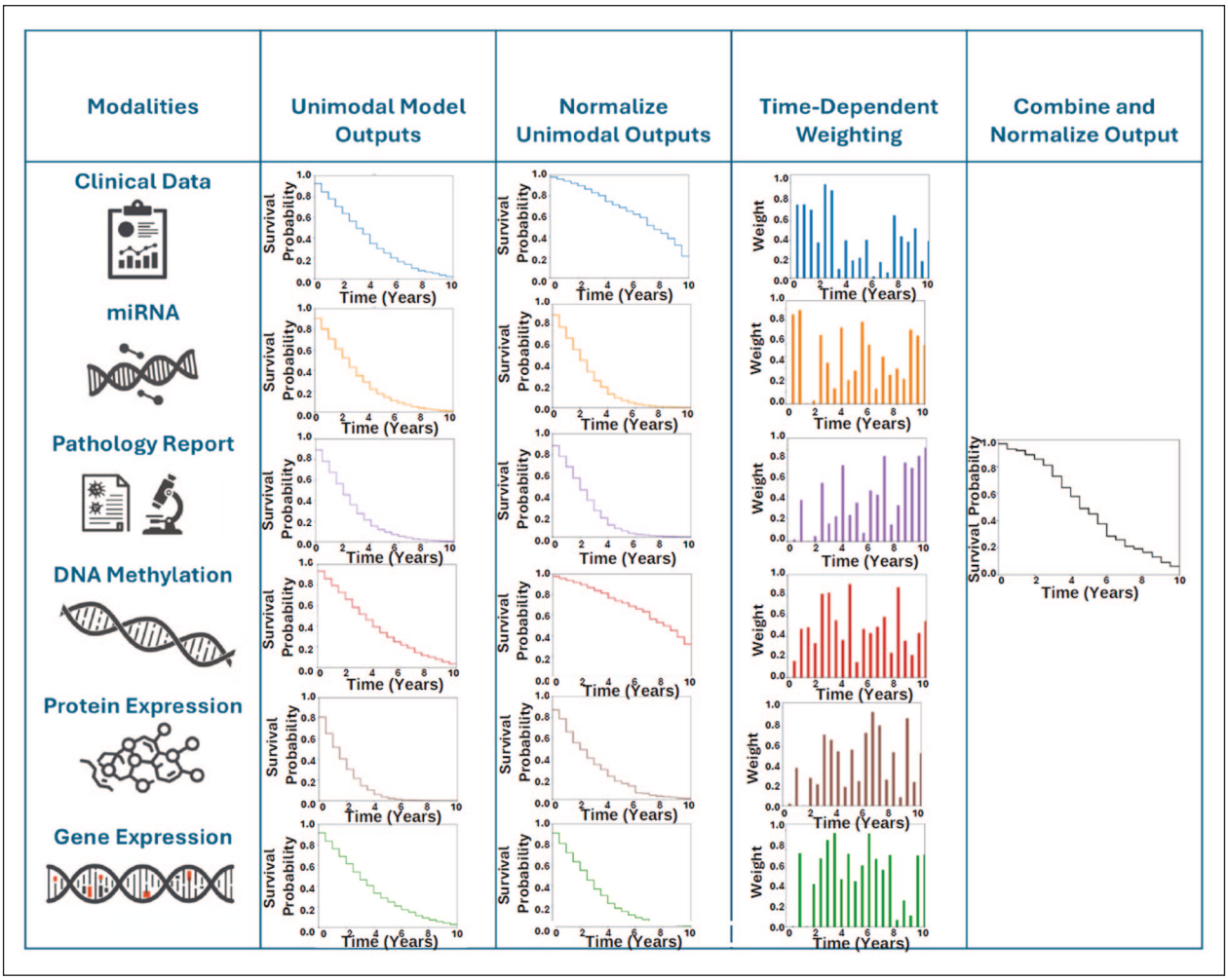

Figure 2. Schematic layout of the proposed late fusion method is presented. Up to 6 unimodal models are separately trained and output unique survival predictions. Before the combination, the outputs are normalized to avoid non-linear influences from modalities with higher variance outputs. The modalities are combined with linear weighting, with the option for time-dependent weighting. The combined output is normalized again due to the inherent decrease in variance at the combination step. The statistics for the normalization are calculated by maximizing the Integrated Brier Score on the training set.

$$\mathbf{y}_{comb,j} = \sum_{i=1}^{M} w_i \, \mathbf{y}_{norm,i,j},$$

where M is the number of modalities in the ensemble. We normalize the combined score once more, again using the training set statistics and w_{std}:

$$\mathbf{y}_{norm,i,j} = \left(\frac{\mathbf{y}_{i,j} - \mu_{test,j}}{\sigma_{test,j}} \right) \sigma_{train,j} \, w_{std} + \mu_{train,j}.$$

Once the combined score is normalized, we compute hazards and survival probabilities via the sigmoid function:

$$H_j = Sigmoid\left(\mathbf{y}_{comb,j}\right), \; S_j = \prod_{k=1}^{j} \left(1 - H_k\right).$$

We use a single risk score per case for the C-index by defining:

$$R_{risk} = -\sum_{j=1}^{T=20} S_j,$$

where S_j is the survival probability at time bin j. Finally, we compute the C-index by pairing R_{risk} with the survival times and censorship indicators.

Alternative Weight Calculation Options

In this section, we describe several alternative late fusion weight calculation methods which are used as a comparison to our proposed method in the "Late Fusion Experiment" Section. There are 5 main weight calculation strategies, and we test each with 1 validation set and with 5 nested cross-validation sets.

The first strategy we test is the existing method proposed by Nikolaou et al,[9] in which the weight for each modality is set by the validation C-index minus 0.5. We describe this as the baseline ad-hoc weighting method. The original model uses only 1 validation set, but we improve on this by using 5 nested validation sets and averaging the C-indices before calculating the weights. This strategy is referred to as the "improved ad-hoc" or "5-val ad-hoc" method. By using 5 validation sets, we aim to decrease the variance in the difference between the actual test set C-indices and the validation set C-indices used in calculating the weights. Figure 3 shows the existing method, and Figure 4 shows the improved method.

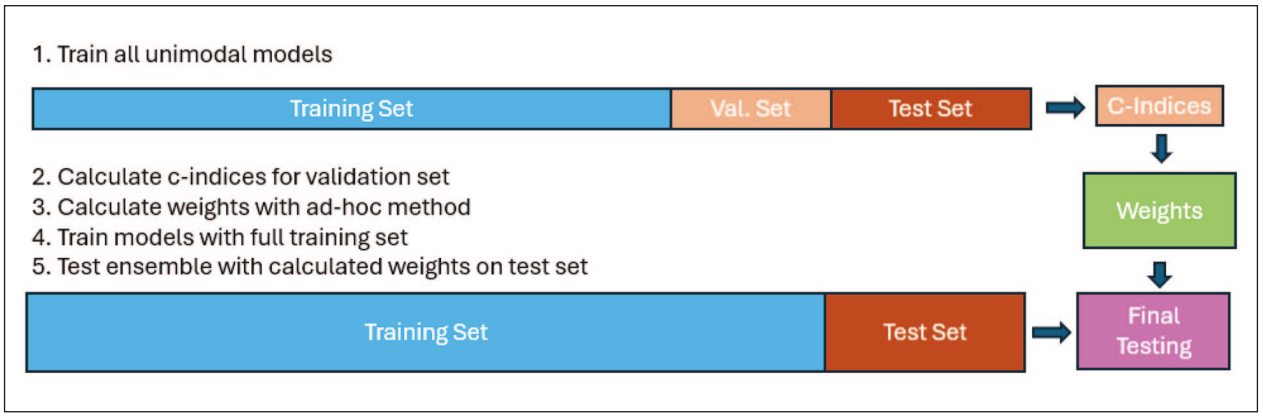

Figure 3. Existing ad-hoc method with 1 validation set. This method was introduced in a previous work and uses simplified C-index estimation and weight calculation methods. By calculating the weight as the difference between the C-index and 0.5, the cross-correlation is not accounted for in either the model outputs or validation set outputs.

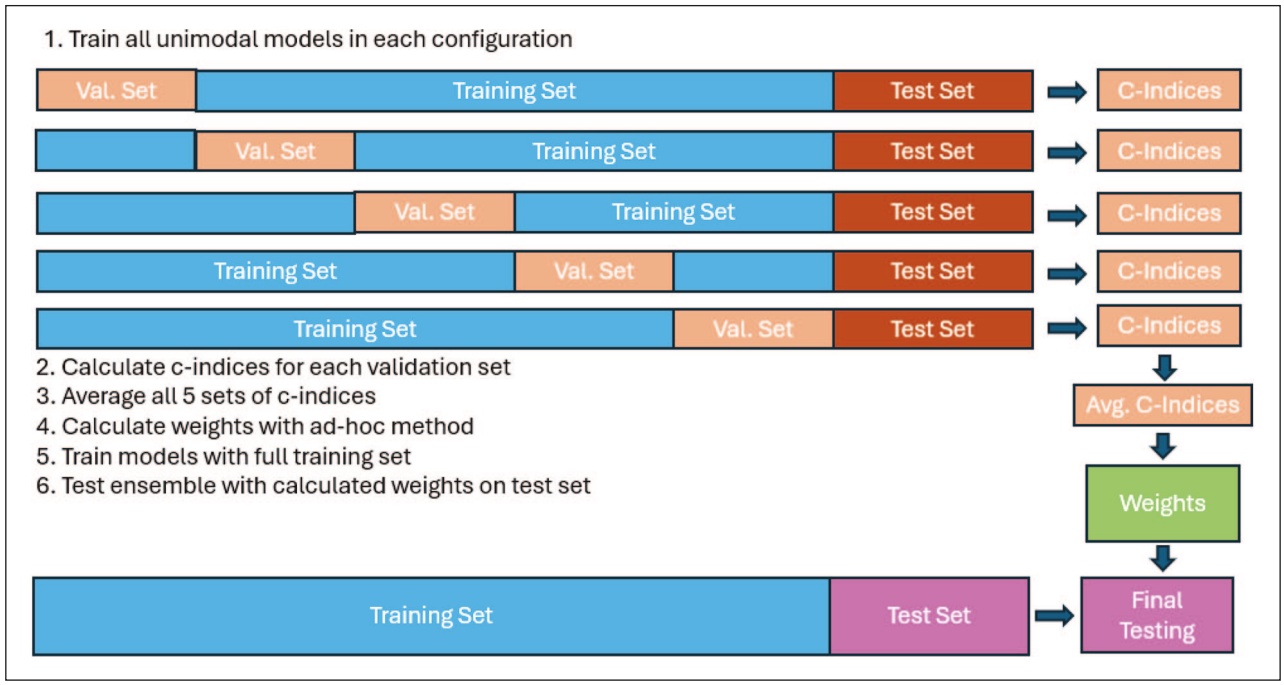

Figure 4. Improved ad-hoc method. This method is identical to the previous method except for the 5 nested validation sets. The averaging of the 5 validation C-indices decreases variance, which reduces the average difference between test set C-indices and validation set C-indices.

The second strategy empirically searches for the best weights for the validation set using a population-based grid search and then applies these weights to the test set outputs to generate the final ensemble output. This is described as the averaged weights method. We also test this method with 5 nested validation sets, and we average the calculated weights to give a more generalized estimate of the ideal test set weights.

Figure 5 shows this strategy used with nested cross-validation.

The third strategy is the synthetic weights method. We calculate the average C-indices just as is done in the ad-hoc method, and then once the models are trained on the full training set, we calculate the Pearson correlation matrix based on the model outputs for the test set. These are used as inputs for the late fusion simulation described above, which generates a fully synthetic dataset and calculates the ideal weights to combine the synthetic risk scores. Figure 6 shows this strategy with nested cross-validation.

The fourth strategy is the RMSurv method. This data flow is similar to the synthetic weights method, but instead of simulating risk scores, the actual unimodal outputs are used with sampled and adjusted training set survival times. The weight search method is similar to the synthetic weights method, but for RMSurv, we use the full discrete survival calculation instead of a single simulated risk score.

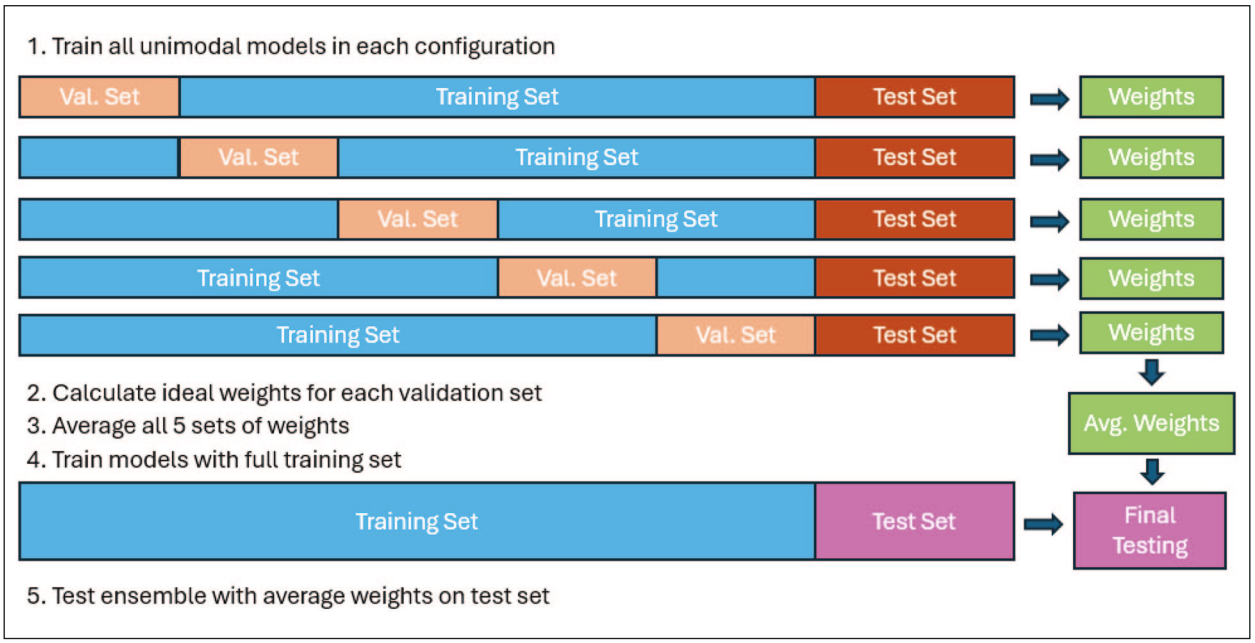

Figure 5. Averaged weights method with 5 nested validation sets. This method leverages a grid search for each validation set and then averages the calculated weights to apply on the test set. While this method does take advantage of the complex combination relationship, it suffers from the high variance in validation sets and does not account for the cross-correlation of the model outputs.

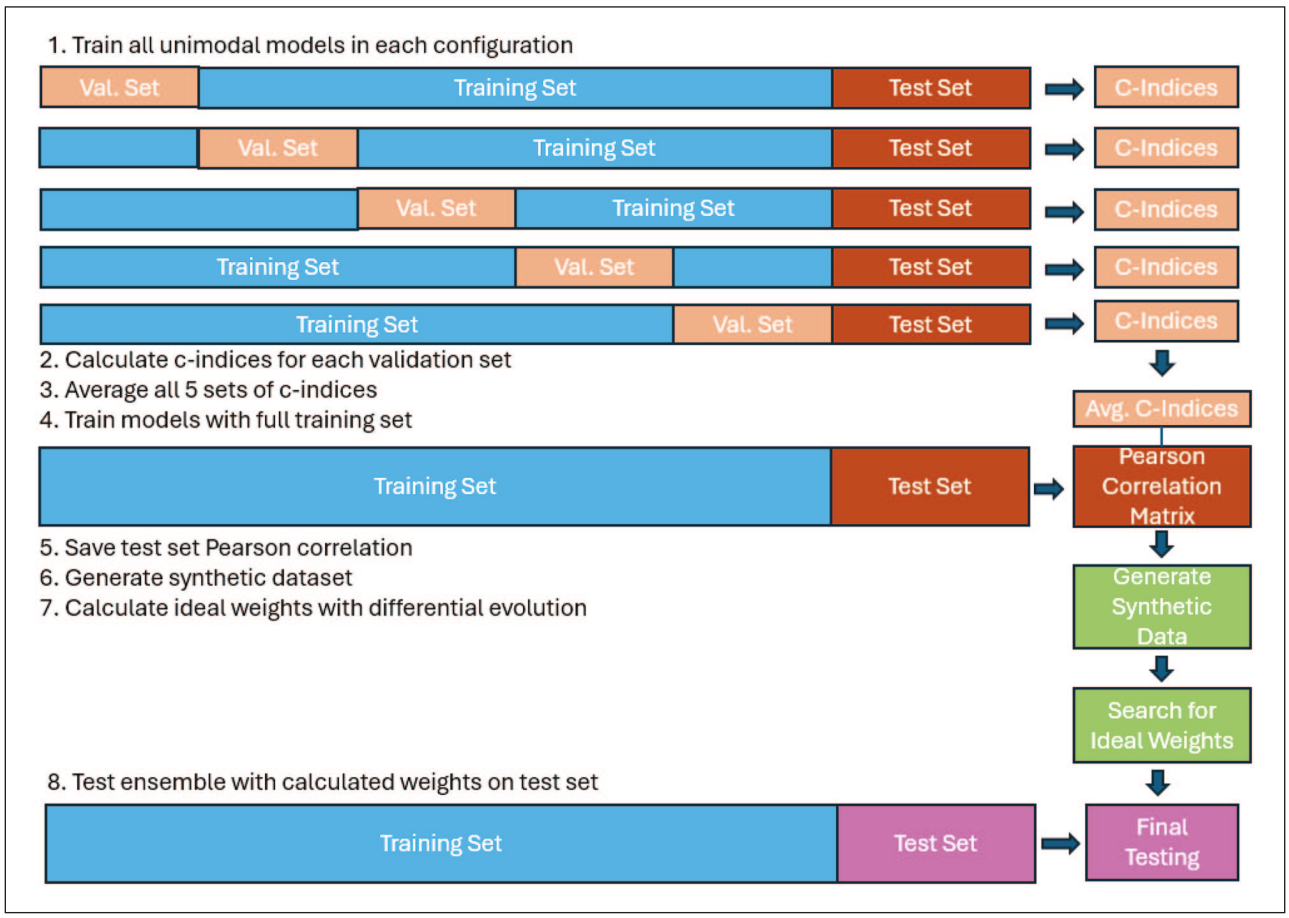

Figure 6. Synthetic weights method with 5 nested validation sets. This method is similar to RMSurv but uses a fully simulated dataset, just like the 1 used in the 2-D linear combination simulation shown above. This simulation uses only the test set correlations and average validation C-indices as inputs. This contrasts with RMSurv, which uses sampled training set survival times, test set model outputs, and average validation C-indices. The grid search is performed on the synthetic dataset to calculate the weights.

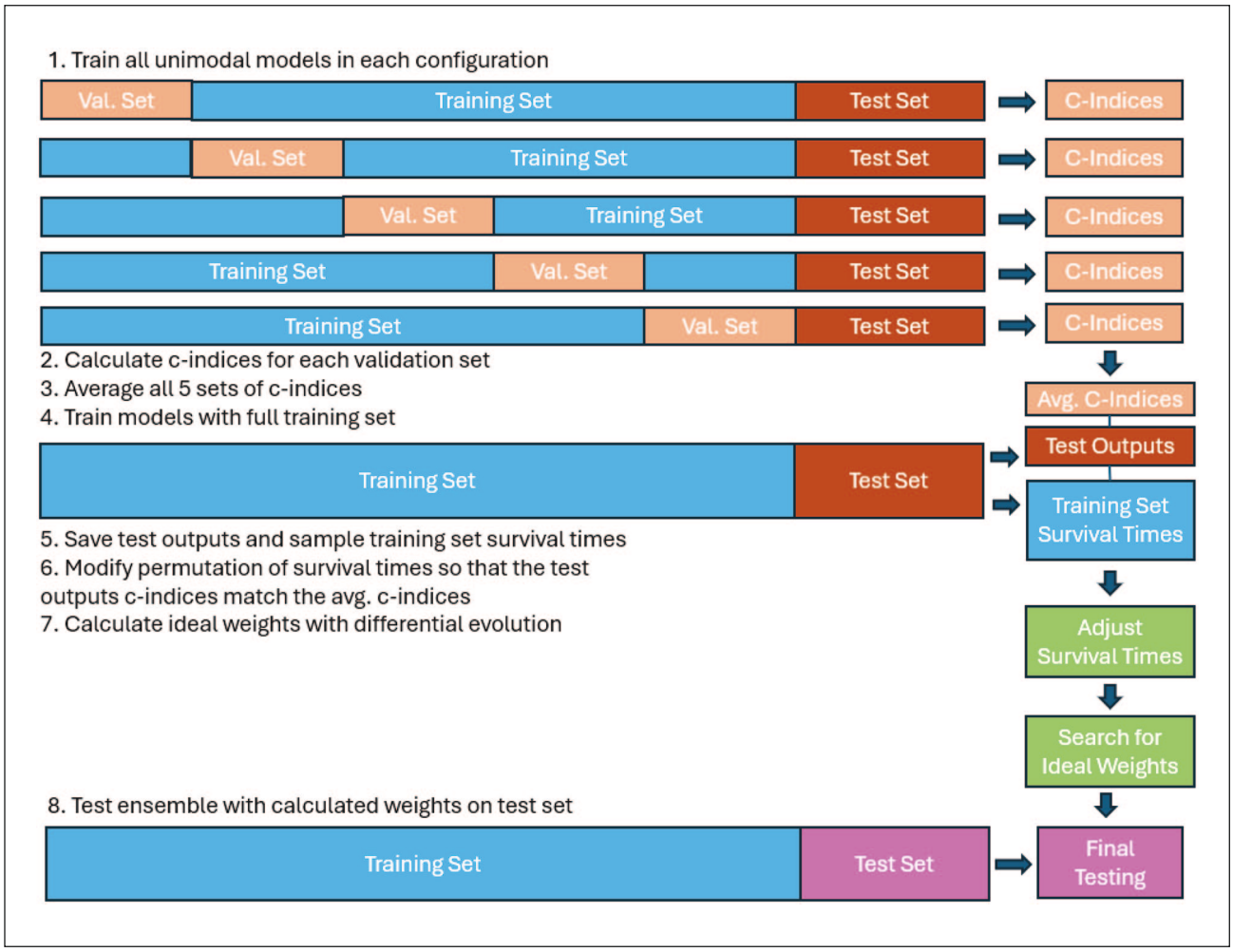

Figure 7. RMSurv method with 5 nested validation sets. This method involves the same nested crossvalidation used in the other methods. Next, the models are trained with the full training set, and both test set outputs and training set survival times are recorded. The synthetic dataset is then generated using a binary search to give the randomly initialized dataset the average validation C-indices of the training set while maintaining the true cross-correlations without leaking the test set labels. A grid search is performed in the last step, and using the original model outputs allows for calculating independent weights for each time bin.

The fifth strategy, TD-RMSurv, uses the same data flow as RMSurv, but the search space is extended to calculate $Mx20$ weights, where M is the number of modalities. Figure 7 shows the flow of data for both of these strategies when using nested cross-validation.

Early and Intermediate Fusion Models

To create an early fusion model most analogous to our proposed late fusion model, we use an identical architecture to what is used in the unimodal models for TD-RMSurv. This architecture is the same as described in the "Unimodal Architecture" section, but all modalities are concatenated into 1 input vector which is used as the model input. We also create a Cox-based early fusion model for our high-level comparison. This model is also based on the same unimodal architecture, but the output layer consists of a single risk score, and we utilize the Cox partial likelihood loss function for optimization. In the 3-modal high-level comparison, we use the 3 best-performing modalities for each dataset. For LUAD, we use clinical data, gene expression, and miRNA. For LUAD + LUSC, we use clinical data, miRNA, and pathology reports. For pan-cancer, we use DNA methylation, miRNA, and protein expression.

We train intermediate fusion models based on the HFBSurv architecture[12] for the high-level fusion method comparison. This architecture employs 3 modalities, incorporating self-attention and bimodal cross-attention mechanisms, along with several fully connected layers. We utilize the 3 best-performing modalities, as described above, for the 3-modal comparison. We modify the HFBSurv architecture to accept 6 modalities for our 6-modal comparison. For each additional modality, we add cross-modality attentional factorized bilinear modules between each pairwise modality to model all bimodal feature interactions. The original HFBSurv has 1 Cox output layer, so we also create modified 3-modal and 6-modal architectures with 20 discrete outputs for a better comparison to the discrete late fusion model.

In both early and intermediate fusion setups, hyperparameters were manually tuned to prevent overfitting. The lack of an automated hyperparameter search is a limitation

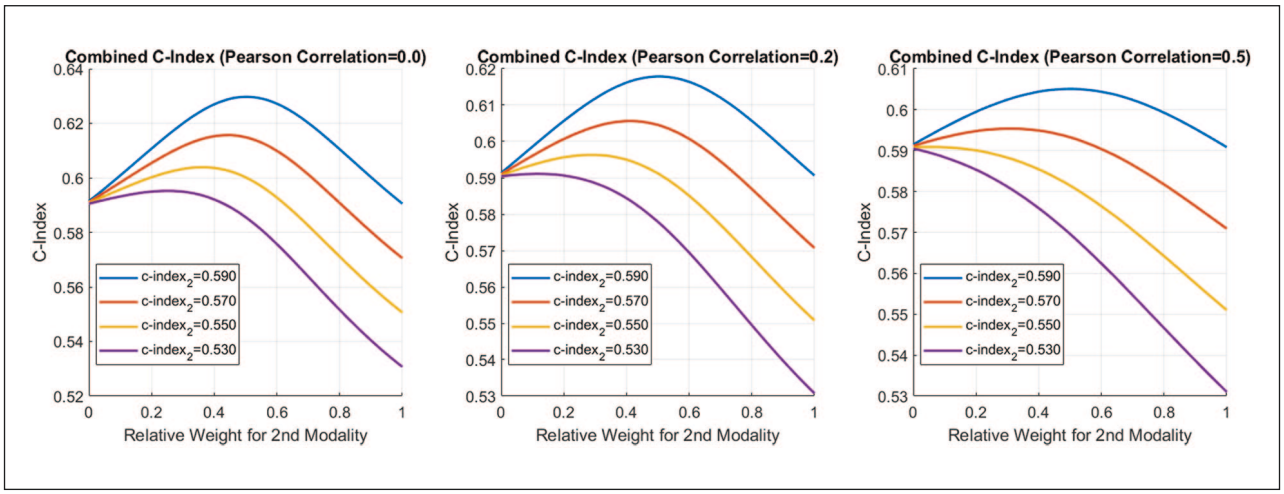

Figure 8. 2-D linear combination simulation with varying Pearson correlation is presented. In these simulations, the first modality is held at 0.59 C-index, and 4 datasets are generated with the second C-index ranging from 0.53 to 0.59. The relative weights are calculated at 100 points to show the ideal weights at the peak of each curve. The figure on the left, with zero correlation, shows the greatest multimodal advantage. The central figure, with a correlation coefficient of .2, represents the average scenario for the LUAD + LUSC dataset. The figure on the right, with a 0.5 correlation, shows a much smaller multimodal advantage, and no advantage with second modalities below 0.55 C-index.

of this comparison, since early and intermediate multimodal fusion models can be highly sensitive to hyperparameters.[2]

Results

Late Fusion Simulation Results

To show the relationship between cross-modality correlation, unimodal C-index, and combined C-index, we create simulated datasets with 2 sets of risk scores with defined C-indices and Pearson correlations. To represent LUAD + LUSC, we set the first modality C-index to 0.59 and show curves with a varying second modality C-index. In our experiments, the cross-modality correlations usually varied between 0 and .5, with an average of about .2. Figure 8 highlights how the relationship changes as a function of correlation. While adding the modality with 0.55 C-index with 0 correlation results in a +0.015 advantage and ideal weight near .4, the same modality cannot improve performance at any weight when the correlation is set to .5. Setting correlation to .2 models the average scenario for the LUAD + LUSC fusion model, and the simulation suggests that modalities below 0.53 C-index will not improve the combined prediction at any weight. Even in this ideally simulated scenario, some modalities are not capable of improving the C-index of another, so it follows that in real datasets with noise and unknown test-set C-indices, the inclusion of certain modalities will decrease performance even with an empirical weight search. This theoretical relationship is validated in our results for our proposed method on LUAD + LUSC, shown in the "Comparison of TD-RMSurv and Ad-hoc Methods" section in which the performance was increased by each modality other than gene expression (0.52 C-index). This simulation shows how the empirical ideal weighting relationship cannot be defined by a simple formula like C-index-0.5, since the result will also depend on correlation and number of modalities included. These results validate the need for a weight calculation method like RMSurv which leverages the true empirical relationships in late fusion.

Unimodal Performance

Figure 9 shows the average unimodal performance on each dataset. For both LUAD and LUAD + LUSC, the unimodal model trained on clinical data outperformed models trained on other data modalities by a wide margin. For the pan-cancer dataset, the best-performing models use the miRNA and DNA methylation modalities. The outputs of these unimodal models are used in the late fusion models of our late fusion experiment.

Late Fusion Experiment

Figure 10 shows a comparison between the 5 late fusion weight calculation methods described above when using all 6 available modalities on the LUAD dataset. The most significant trend in the late fusion experiments was the consistent improvement by using the fivefold nested validation sets instead of a single validation set. Since LUAD is such a small dataset, the decreased variance in the validation C-indices was dramatic and improved performance with every method. Table 3 shows a statistical comparison between the single validation set and 5 validation set nested cross-validation options for each of the 5 strategies. There was a large and significant increase in performance for every strategy.

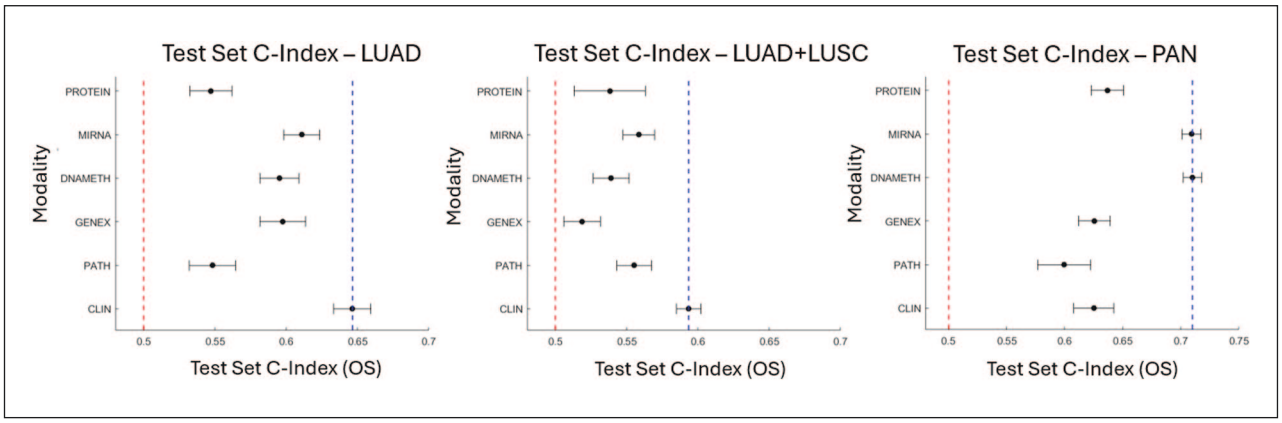

Figure 9. Unimodal performance with 95% confidence interval (CI) is shown for all 3 datasets tested. We train unimodal models using protein expression, miRNA, DNA methylation, gene expression, pathology report PDF embeddings, and clinical data (age, race, gender, stage). The dashed red line marks the minimum predictive performance at C-index = 0.50. The dashed blue line represents the best unimodal model performance. For LUAD and LUAD + LUSC, the strongest unimodal model is clinical data. For the pancancer dataset, which includes 33 cancer types, the strongest unimodal model is DNA methylation.

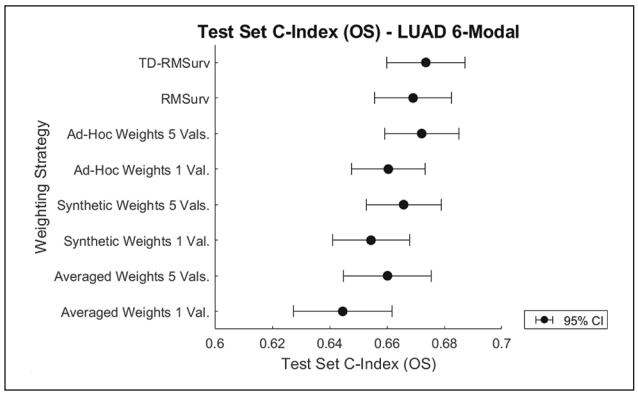

Figure 10. Comparison of the late fusion methods described above on the TCGA LUAD dataset using all 6 modalities. We test several models with both a single validation set and a full nested cross-validation for weight calculation. The use of 5 validation sets in the nested cross-validation consistently improved C-index by over 0.01. TD-RMSurv and Ad-Hoc (5-val) performed the best and are used in subsequent comparisons.

Another surprising result was the relatively strong performance of the Ad-hoc method on LUAD. For this dataset and 6-modal configuration, it outperformed the "averaged weights," "synthetic weights," and baseline RMSurv strategies, and performed nearly as well as TD-RMSurv. This is unintuitive since these methods use empirical grid searches for weight calculation that will take correlation and number of modalities into account. One explanation for this is that the Ad-hoc weighting method can be more beneficial on a very small dataset like LUAD because of a lower sensitivity to small differences in C-index and correlation. LUAD can have as few as 30 uncensored test cases, so the C-indices can be highly variable and result in large differences between the validation and test C-indices. Additionally, the inclusion of more informative cases in either the training set or the testing set can result in an inverse relationship between the testing C-index and the average validation C-index. When these conditions are met, the reduced sensitivity of the ad-hoc approach could be more beneficial than the optimized weighting relationship. The baseline time-independent RMSurv method outperformed the ad-hoc method on the 2 larger datasets, which supports this conclusion. This small dataset phenomenon could also explain the poor performance of the "averaged weights" method. Since this approach uses a highly sensitive grid search that uses the model outputs and cross-correlations within each validation fold, instead of only using an average of the validation C-indices, more of the noise from the small dataset is likely preserved, decreasing performance compared to the other strategies.

The RMSurv method outperformed the synthetic weights method by a wide margin despite the similar approach. There are a few likely explanations for this difference. First, the synthetic weights method models a single risk score per case using a normal distribution, instead of the 20 discrete outputs of the actual distribution, which maintain their exact time-dependent correlations in RMSurv. Second, the simulation assumes an exponential distribution for survival times, instead of sampling from the training distribution. Finally, the 6-dimensional simulation has small errors between the desired and actual C-indices and correlations. These were not possible to remove completely, likely because C-indices are non-linear parameters, which needed to be modeled as linear correlations and formed into a positive semi-definite covariance matrix.

TD-RMSurv (5-val) and the Ad-hoc method (5-val) performed best for LUAD 6-modal, so these 2 strategies are compared in more detail in the "Comparison of TD-RMSurv and Ad-hoc Methods" Section below.

Comparison of TD-RMSurv and Ad-Hoc *Methods.* Figure 11 shows a comparison of TD-RMSurv and the improved ad-hoc late fusion method on 3 datasets of varying sizes. Tables 4 to 6 show a statistical analysis of the performance

Table 3. Statistical Comparison of Nested and Single Validation for 6-Modal LUAD.

Late fusion strategy	Nested cross-validation	1 Validation set	Mean difference (bootstrap 95% CI)	Paired *t*-test *p*-value
TD-RMSurv	0.6735	0.6647	+0.0088 (+0.0031, +0.0148)	.0054
RMSurv	0.6691	0.6597	+0.0093 (+0.0016, +0.0177)	.0289
Ad-Hoc	0.6721	0.6601	+0.0120 (+0.0069, +0.0171)	.0000
Synthetic weights	0.6658	0.6544	+0.0114 (+0.0032, +0.0200)	.0119
Averaged weights	0.6602	0.6446	+0.0156 (+0.0075, +0.0237)	.0005

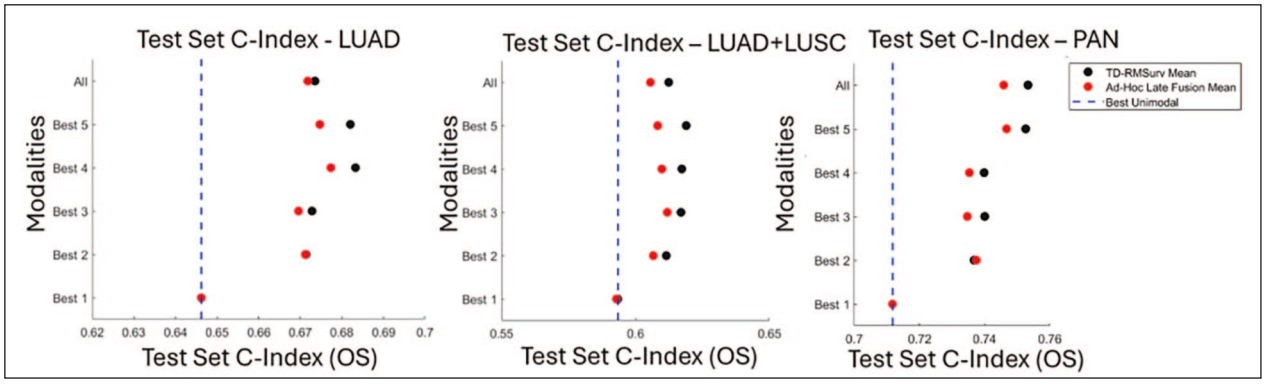

Figure 11. Comparison of the TD-RMSurv and 5 val. ad-hoc weight calculation methods on three datasets.
This chart compares performance when using a varying number of modalities. TD-RMSurv outperformed the ad-hoc method overall.

Table 4. Statistical Comparison of TD-RMSurv and Ad-Hoc Methods on LUAD.

# of Modalities	TD-RMSurv C-index	Ad-Hoc mean C-index	Mean Difference (bootstrap 95% CI)	Paired *t*-test *p*-value
2	0.6714	0.6712	+0.0002 (−0.0033, +0.0036)	.9174
3	0.6728	0.6696	+0.0033 (+0.0007, +0.0059)	.0204
4	0.6833	0.6773	+0.0059 (+0.0016, +0.0101)	.0091
5	0.6820	0.6747	+0.0073 (+0.0031, +0.0118)	.0019
6	0.6735	0.6718	+0.0017 (−0.0029, +0.0063)	.4733

Table 5. Statistical Comparison of TD-RMSurv and Ad-Hoc Methods on LUAD + LUSC.

# of Modalities	TD-RMSurv C-index	Ad-Hoc mean C-index	Mean Difference (bootstrap 95% CI)	Paired *t*-test *p*-value
2	0.6115	0.6066	+0.0048 (+0.0, +0.0)	.0002
3	0.6169	0.6118	+0.0051 (+0.0026, +0.0072)	.0002
4	0.6172	0.6098	+0.0074 (+0.0026, +0.0077)	.0001
5	0.6189	0.6082	+0.0107 (+0.0041, +0.0108)	.0000
6	0.6124	0.6056	+0.0068 (+0.0070, +0.0107)	.0010

Table 6. Statistical Comparison of TD-RMSurv and Ad-Hoc Methods on PAN.

# of Modalities	TD-RMSurv C-index	Ad-Hoc mean C-index	Mean difference (bootstrap 95% CI)	Paired *t*-test *p*-value
2	0.7367	0.7374	−0.0007 (−0.0014, +0.0001)	.1263
3	0.7400	0.7347	+0.0053 (+0.0034, +0.0072)	.0005
4	0.7398	0.7353	+0.0045 (+0.0018, +0.0073)	.0141
5	0.7526	0.7467	+0.0059 (+0.0036, +0.0080)	.0007
6	0.7533	0.7458	+0.0074 (+0.0047, +0.00102)	.0007

on the 3 datasets. TD-RMSurv demonstrated a statistically significant advantage for all configurations on LUAD + LUSC, but did not show a consistent advantage for the 2-modal and 6-modal configurations on LUAD. This suggests that an ad-hoc method will vary in performance between datasets, depending on the size of

Table 7. Statistical Comparison of TD-RMSurv and Baseline RMSurv Methods on LUAD.

# of Modalities	TD-RMSurv C-index	RMSurv mean C-index	Mean difference (bootstrap 95% CI)	Paired *t*-test *p*-value
2	0.6714	0.6672	+0.0042 (−0.0033, +0.0092)	.0952
3	0.6728	0.6689	+0.0039 (+0.0002, +0.0077)	.0486
4	0.6833	0.6732	+0.0100 (+0.0037, +0.0166)	.0042
5	0.6820	0.6712	+0.0109 (+0.0055, +0.0163)	.0003
6	0.6735	0.6691	+0.0044 (−0.0019, +0.0110)	.1894

Table 8. Statistical Comparison of TD-RMSurv and Baseline RMSurv Methods on LUAD + LUSC.

# of Modalities	TD-RMSurv C-index	RMSurv mean C-index	Mean difference (bootstrap 95% CI)	Paired *t*-test *p*-value
2	0.6115	0.6059	+0.0056 (+0.0021, +0.0095)	.0048
3	0.6169	0.6116	+0.0053 (+0.0018, +0.0089)	.0058
4	0.6172	0.6124	+0.0049 (+0.0003, +0.0098)	.0530
5	0.6189	0.6109	+0.0080 (+0.0042, +0.0118)	.0002
6	0.6124	0.6057	+0.0067 (+0.0018, +0.0116)	.0108

Table 9. Statistical Comparison of TD-RMSurv and Baseline RMSurv Methods on PAN.

# of Modalities	TD-RMSurv C-index	RMSurv mean C-index	Mean difference (bootstrap 95% CI)	Paired *t*-test *p*-value
2	0.7367	0.7372	−0.0005 (−0.0013, +0.0003)	.3231
3	0.7400	0.7391	+0.0009 (+0.0026, +0.0014)	.0092
4	0.7398	0.7456	−0.0058 (−0.0112, −0.0011)	.0645
5	0.7526	0.7525	+0.0001 (−0.0014, +0.0012)	.9415
6	0.7533	0.7528	+0.0005 (−0.0003, +0.0013)	.2762

the dataset, and whether the C-indices, correlations, and number of modalities are such that the empirical ideal weights are similar to the calculated ad-hoc weights. The lack of a significant advantage with 2 modalities could be a case where 2 modalities with low correlation create a simple empirical relationship that closely matches the analytical relationship, and the increased sensitivity of the TD-RMSurv method to the noisy dataset becomes disadvantageous. The inconclusive results of the 6-modal configuration could also be explained by increased sensitivity to the weakest modality that is highly variable in performance for this dataset. In the pan-cancer dataset, TD-RMSurv showed a statistically significant advantage in every configuration except for the configuration using the best 2 modalities. For this dataset, the 2 strongest modalities have unimodal outputs with nearly identical C-indices (0.7091, 0.7117), so in the 2-modal configuration the average ideal weight for each is a trivial 0.5 which is what the ad-hoc formula predicts. This explains the close results, and the slight disadvantage of TD-RMSurv may be due to sensitivity to the small amounts of noise on this dataset.

It is important to note that this is a comparison to an improved version of the existing method which uses five-fold nested cross validation, so the increased performance results solely from the novel weight calculation method. The strong results of TD-RMSurv in this experiment validate the theoretical arguments for correlation-sensitive weight calculation and time-dependent weighting. Both late fusion models in this comparison had a consistent multimodal advantage, meaning none of the multimodal models underperformed the best-performing unimodal models.

Comparison of Baseline RMSurv and TD-RMSurv Methods. To isolate the effect of time-dependent weighting, we compare the performance of the RMSurv method with and without time-dependent weighting enabled. Tables 7 to 9 show a statistical comparison of these results for the 3 datasets. For LUAD and LUAD + LUSC, the time-dependent model significantly outperforms the baseline in all but 2 configurations. We found no conclusive advantage for the time-dependent method on the pan-cancer dataset. This finding will be explained in the following section through the differences in the weights calculated by baseline RMSurv and TD-RMSurv on the different datasets.

Interpretation of Calculated Weights. A comparison of the calculated weights of these models can help explain some of the differences in performance. The increased sensitivity of RMSurv is visible in the calculated weights shown in Figure 12. The nearly equal weights from the Ad-hoc method, shown on the left sub-figure, highlight the limitation of the method, which assigns a weight to each modality above 0.5 C-index, regardless of cross-correlations. The RMSurv method, shown in the middle sub-figure, clearly excludes the protein expression based on the cross-correlations in this fold.

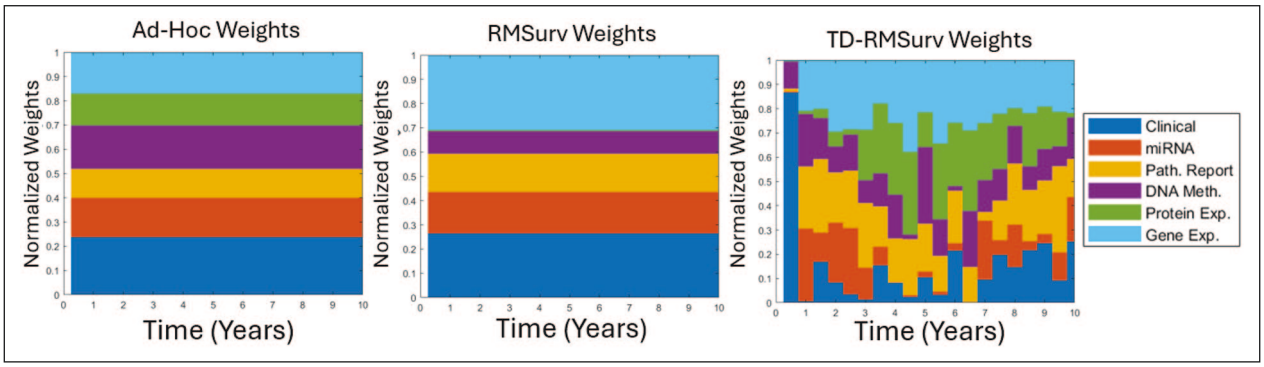

Figure 12. Comparison of calculated weights is shown for LUAD + LUSC seed 10 fold 5.

The weights of the TD-RMSurv model allow the user to interpret the influence of each modality over time. The model assigned a 90% weight to the clinical data for the first time bin and disregarded most of the other modalities. The clinical weighting then decreased to near zero until years 7 to 10. Since the survival prediction is calculated as the cumulative sum of 1-hazard, each subsequent survival prediction will depend on the prediction of the first time bin, which represents the likelihood of survival in the first 6 months. A likely explanation for this is the cancer stage feature, which will be highly predictive for short-term survival. Once the patient survives the first 6 months, other modalities and features become more predictive of the conditional survival. Another interesting change is that the protein modality is now included at certain time bins, where it can improve performance without introducing noise in years 0 to 3.

The time-variant relationships were much more subtle on the larger pan-cancer dataset. The calculated weights for this dataset are shown below in Figure 13. These weights are less variable over time compared to the smaller dataset, which aligns with the smaller performance improvement over the baseline for this dataset. The unimodal models trained on DNA methylation and miRNA outperform the other unimodal models by a wide margin; however, the high cross-correlation between their outputs (~0.7) results in lower weights than what an Ad-hoc method based on the C-index would produce. This allows for significant influence of the other modalities when using the TD-RMSurv method. The clinical modality still holds an outsize weight in the first year, despite having the fourth-highest C-index. These details in the calculated weights of TD-RMSurv highlight information that would be disregarded by a late fusion strategy that does not incorporate correlation-dependent weighting and time-dependent weighting.

Early Fusion Comparison

In this section, we compare our TD-RMSurv model to an early fusion model with a range of modalities. For our early fusion model, the inputs for each modality are concatenated into 1 vector, and the model uses the same discrete architecture as the unimodal models used in our late fusion models. TD-RMSurv late fusion shows a dramatic improvement in performance and robustness compared to the discrete early fusion model.

Figure 14 shows a comparison between the early fusion model and TD-RMSurv on the LUAD dataset. Here, the early fusion model performed inconsistently, sometimes under-performing the clinical-only modality, and never achieving a significant multimodal advantage. There was an unusual drop-off in performance when using 2 modalities for early fusion, which highlights a reliability problem in early fusion. TD-RMSurv, by contrast, shows a strong and consistent multimodal advantage, peaking with the inclusion of 4 modalities.

Figure 15 shows a comparison of early fusion and TD-RMSurv for LUAD + LUSC. The early fusion model performs better on this larger dataset, but shows a modest multimodal advantage, which peaked at 3 modalities. TD-RMSurv shows a stronger multimodal advantage, peaking with the inclusion of 5 modalities.

Figure 16 shows a comparison between the early fusion model and TD-RMSurv for the largest pan-cancer dataset. The performance of the early fusion model decreases significantly with the addition of modality 4, with minimal recovery that does not surpass the best unimodal performance as more modalities are added. For this dataset, modality 4 has many more input features than modalities 1, 2, and 3 combined. This difference in feature length likely caused the dramatic drop in performance, highlighting a limitation of early fusion with concatenation. For late fusion, only the model outputs are combined, ensuring robust model performance in circumstances where the feature length varies or other data characteristics cause overfitting to a single modality. As a result, TD-RMSurv performs strongly on the pan-cancer dataset, achieving its best performance when utilizing all 6 modalities.

High-Level Fusion Method Comparison

In this section, we provide a broader comparison to demonstrate the differences in performance between Cox and discrete models, and early, intermediate, and late fusion models. TD-RMSurv outperforms these alternative methods by a wide margin. The discrete early and

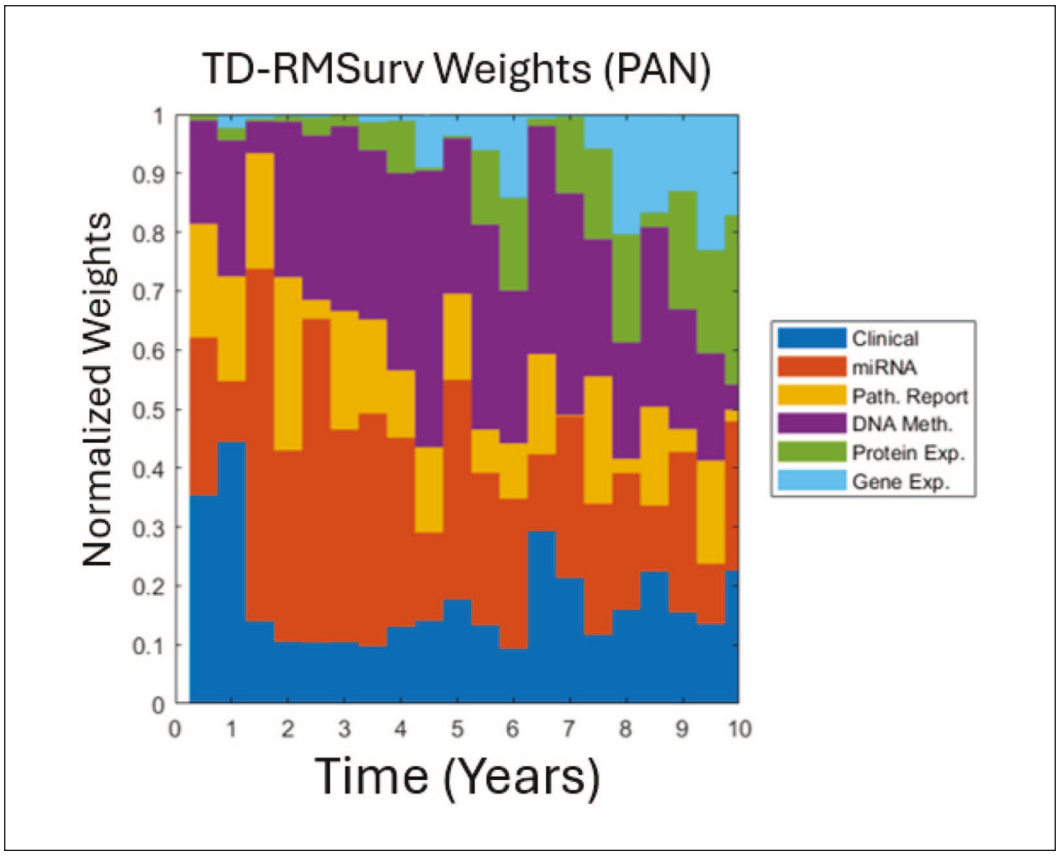

Figure 13. Calculated Weights (TD-RMSurv) for the TCGA pan-cancer dataset for seed 1 fold 5.

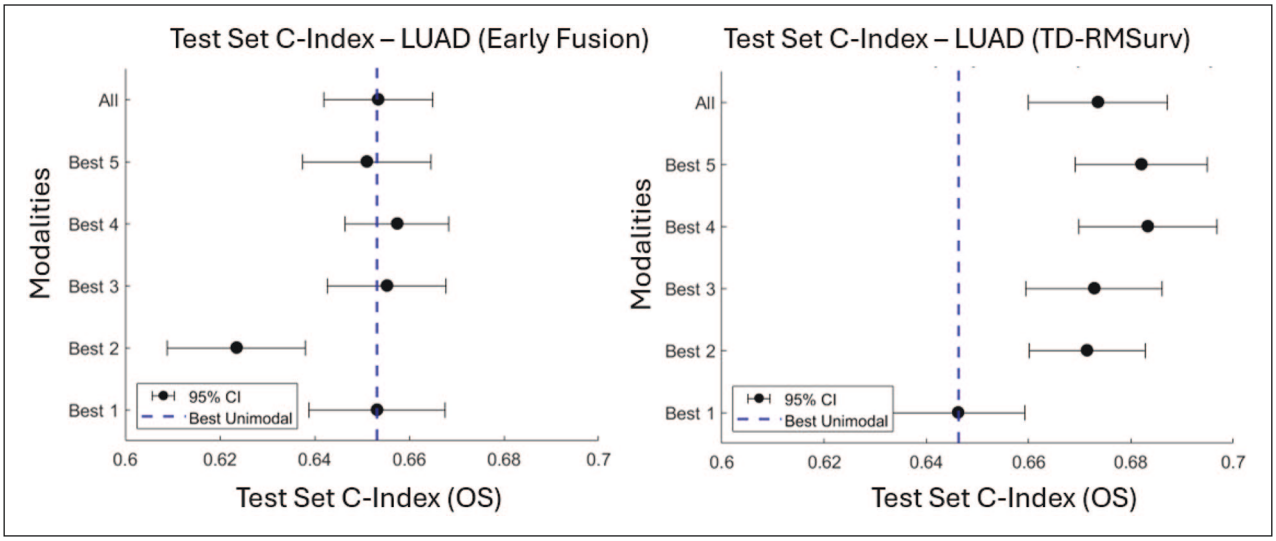

Figure 14. Comparison between early fusion (left) and TD-RMSurv late fusion (right) results for the LUAD dataset.

intermediate fusion models were competitive with late fusion when using the best 3 modalities; however, they sometimes underperformed the unimodal clinical model when using all 6 modalities. The Cox-based models underperformed the discrete models significantly across all 3 datasets.

Figure 17 shows the results for the LUAD dataset. The Cox-based models underperformed the discrete models by a wide margin for this dataset. The discrete early fusion model achieved a small multimodal advantage with both configurations. The discrete intermediate fusion model performed relatively well with the top 3 modalities but experienced a

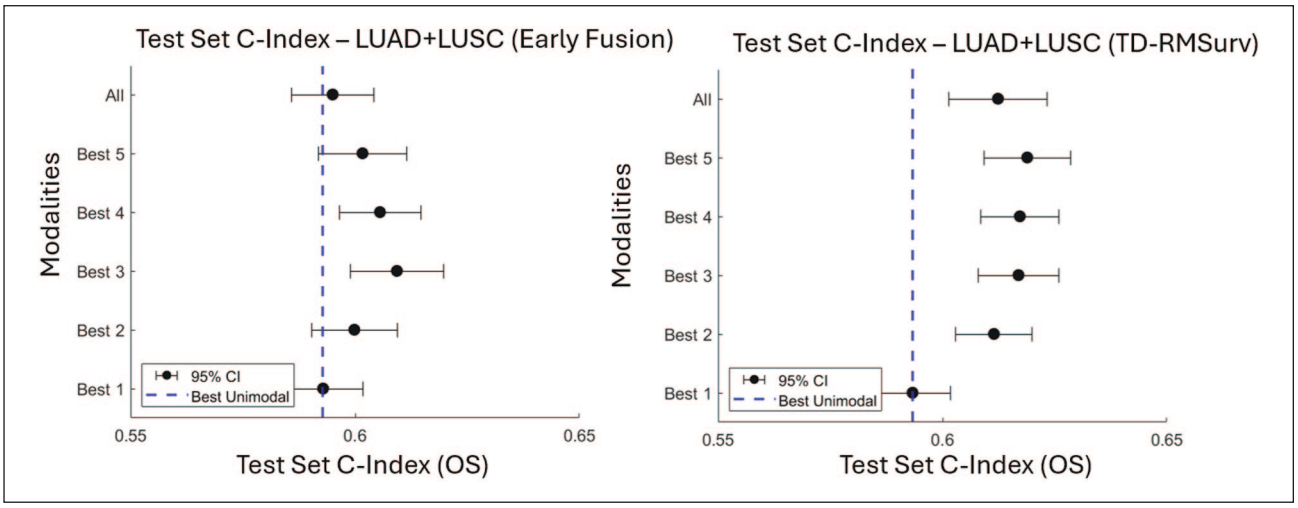

Figure 15. Comparison between early fusion (left) and TD-RMSurv late fusion (right) results for the LUAD + LUSC dataset.

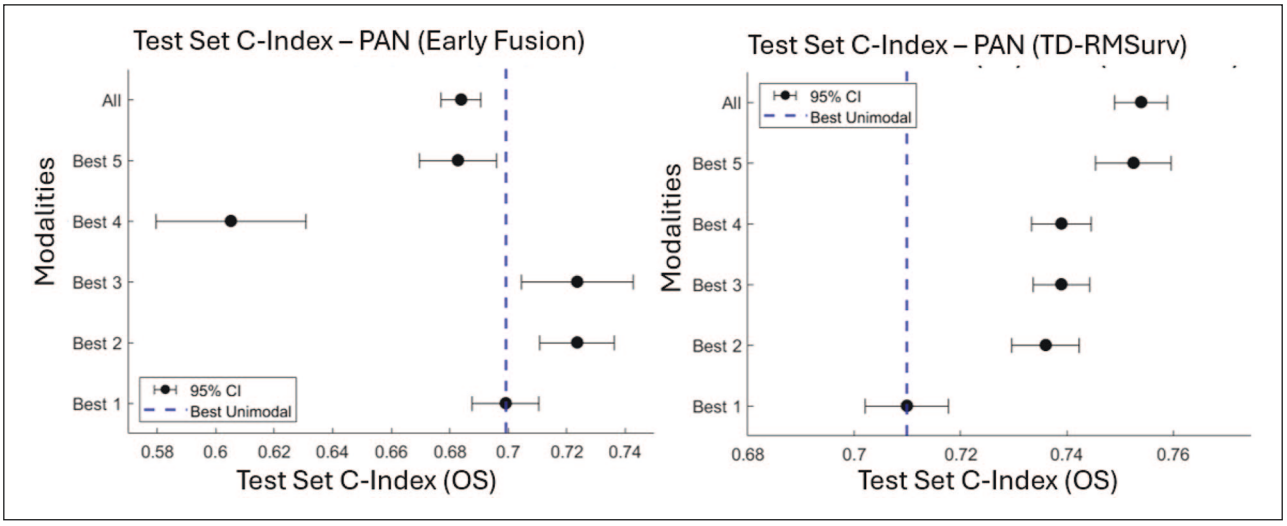

Figure 16. Comparison between early fusion (left) and TD-RMSurv late fusion (right) results for the pan-cancer dataset.

significant performance decline when including all 6 modalities. The TDRMSurv method demonstrates a robust multimodal advantage, even when utilizing all 6 modalities.

For LUAD + LUSC, as shown in Figure 18, the results are similar to LUAD. The Cox-based models also underperformed for this dataset, and TD-RMSurv outperformed early and intermediate fusion by a wide margin in both the 3-modal and 6-modal settings. The discrete early fusion model performed better with 3 modalities included, and the discrete intermediate fusion model again showed a decreased performance with 6 modalities.

Finally, Figure 19 shows the results for the pan-cancer dataset. The Cox models continued to underperform for this dataset, and early fusion remained inconsistent when using 6 modalities. The discrete intermediate fusion model performed slightly below TD-RMSurv for 3 modalities but matched its performance for 6 modalities. This was a surprising result for the intermediate fusion model and may indicate greater potential for intermediate fusion on larger datasets.

Kaplan-Meier Analysis

We validate the robustness of this system through the Kaplan Meier analysis shown in Figure 20. Here we stratify the test set into 3 groups based on the calculated risk score of the TD-RMSurv model for a representative split on the pan-cancer dataset, using all 6 modalities. The risk scores are calculated from the cumulative sum of the discrete survival predictions, just as in the C-index calculations. The effective stratification highlights the qualitative

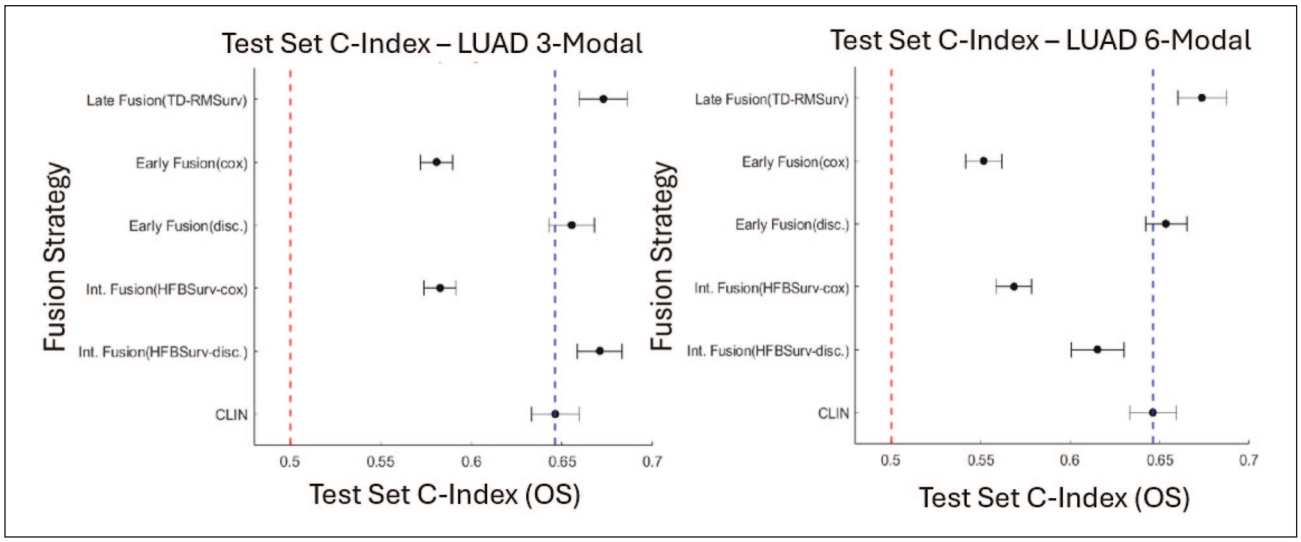

Figure 17. Comparison of multimodal fusion strategies on LUAD. Left: Best 3 modalities. Right: all 6 modalities. The dashed red line indicates the minimum predictive performance at a C-index of 0.50, while the dashed blue line represents the optimal unimodal model performance. The C-index is calculated with 10 cross-validation runs and shown with a 95% confidence interval.

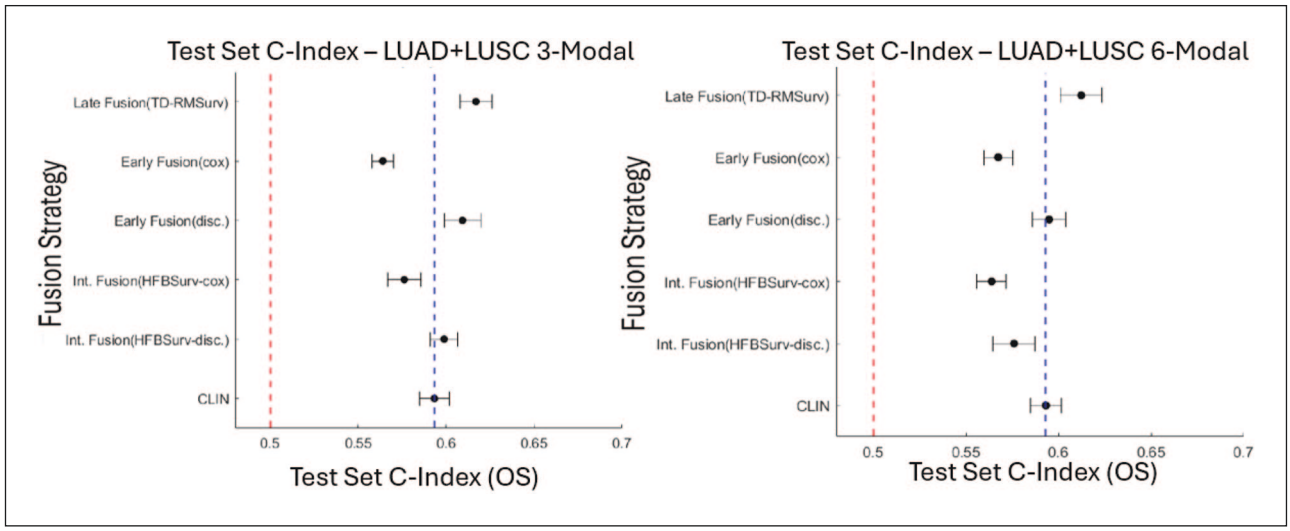

Figure 18. Comparison of multimodal fusion strategies on LUAD + LUSC. Left: Best 3 modalities. Right: all 6 modalities. The dashed red line marks the minimum predictive performance at C-index = 0.50, while the dashed blue line represents the best unimodal model performance. The C-index is calculated with 10 crossvalidation runs and shown with a 95% confidence interval.

reliability of the model when including all available modalities.

Discussion

We found that the TD-RMSurv late fusion method consistently outperformed all unimodal models and multimodal fusion alternatives. We also noticed a consistent improvement by using 5 nested validation sets instead of just 1 validation set for late fusion. Furthermore, other late fusion methods like synthetic weights, ad-hoc, and baseline RMSurv also outperformed the early and intermediate fusion models. Early and intermediate fusion methods necessarily weight modalities based on the combined training-set accuracy, and therefore often overfit to weaker modalities. Late fusion can correct for this, because the errors from individual models tend to be uncorrelated in an ensemble.[1] This advantage was best demonstrated when using all 6 modalities, where the late fusion models maintained or even improved their performance while the other models' performance decreased. These results suggest that for multimodal cancer survival prediction, the benefit of modeling cross-modality feature interactions in early and intermediate fusion is less significant compared to

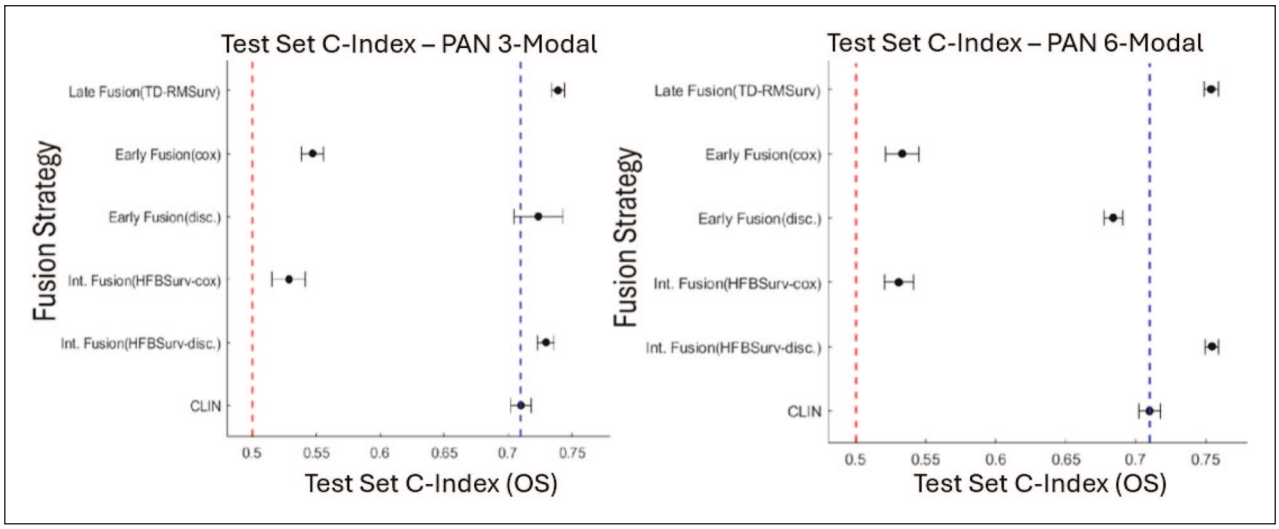

Figure 19. Comparison of multimodal fusion strategies on the pan-cancer dataset. Left: Best 3 modalities. Right: all 6 modalities. The dashed red line indicates the minimum predictive performance at a C-index of 0.50, and the dashed blue line represents the optimal unimodal model performance. The C-index is calculated with a single cross-validation and shown with a 95% confidence interval.

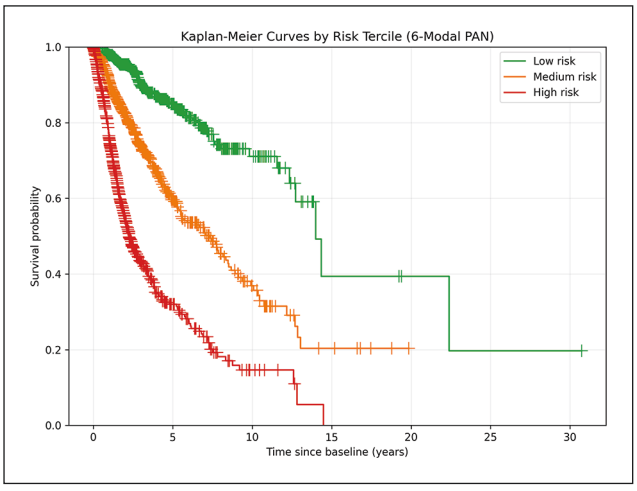

Figure 20. Kaplan Meier Curves showing risk terciles for the 6-modal configuration of TD-RMSurv on the pan-cancer dataset for the test set of seed 1, fold 1 (C-index = 0.7459).

the benefit of treating modalities as dependent random variables with late fusion. Several studies also argue that late fusion is best suited for settings with heterogenous and low correlation modalities, and with small sample sizes.[14,19] The TCGA dataset meets this criteria, so our results support this conclusion.[16]

The increased performance of time-dependent modeling is demonstrated in both the discrete multimodal advantage and the time-dependent weighting advantage. The discrete survival models outperformed the Cox Proportional Hazards models across the datasets, and TD-RMSurv also outperformed baseline RMSurv across all datasets. The weights for TD-RMSurv shown in Figure 12 perfectly demonstrate the limitation of the proportional hazards assumption by showing how the ideal weighting of modalities changes over time.

Our late fusion simulation proves that the modality combination relationship is much more complicated than simply comparing modalities based on their C-index. The true relationship is dependent on the C-index, correlation, and number of modalities. The RMSurv and TD-RMSurv strategies take advantage of this complex relationship, which is likely why they perform better than the ad-hoc method overall.

The pathology report modality was one of the weaker modalities, but its inclusion modestly increased performance on the LUAD + LUSC and pan-cancer datasets when using TD-RMSurv. It is a promising area of research, especially as language models continue to improve.

RMSurv provides a robust multimodal advantage over unimodal models, which is an important step toward the clinical relevance of multimodal survival models. The late fusion approach also allows for easy integration of 6+ modalities, which may require different architectures. For the pan-cancer dataset, TD-RMSurv performed best when using all 6 modalities, which is a step toward including even more modalities on larger datasets. For the LUAD and LUAD + LUSC datasets, performance peaked when including 4 and 5 modalities, respectively. This lack of perfect robustness can be explained by our simulation results, which show that very low performing modalities fundamentally cannot add any signal to a combined prediction for datasets with non-zero correlations. In a very small dataset with high variance, these modalities will occasionally receive some weight when they should be assigned no weight, and the combined performance will decrease. Despite this, TD-RMSurv showed a robust multimodal advantage in all configurations, contrasting the unpredictable performance of the early and intermediate fusion methods.

There is significant potential to improve on this work in the future. This study provides a strong proof of concept with 3 datasets of various sizes, but a broader comparison with many more cancer types is needed to better understand the limitations, especially on very small datasets. Future

implementations could also add histopathology slides, MRI scans, and treatment regimes to combine with existing modalities in late fusion. The pathology report modality introduced in this study also presents an opportunity to add several new text-based modalities such as clinical notes, laboratory tests, and more clinical data features. The architecture of TD-RMSurv could also be improved by varying or increasing the resolution of time bins. By adding 1-month time bins for the first 6 months, for example, we could replace the large jumps in modality weighting between time bins with a more gradual and accurate representation. The multimodal performance of this method, and the clinical utility, is still limited by unimodal performance, so much larger datasets and optimal unimodal architectures for each modality will be needed to outperform traditional prognoses. Future applications should also consider cases where interactions between features across modalities are important. Late fusion cannot model these interactions, so combining certain modalities into early or intermediate fusion sub-models could be beneficial.

Beyond model performance, several other challenges remain for prospective clinical application. Improved datasets or corrective adjustments to survival time sampling will be necessary to account for censoring, which can bias the model toward lower or higher survival probability depending on if the censored cases are included in the training set. Model interpretability for a clinical setting is significantly improved with the proposed method, which can show the normalized survival predictions and relative weights assigned to each modality, but each unimodal model is still a black box in this setup. A future architecture with interpretability at both the unimodal feature level and late fusion output level could be a valuable clinical tool. Finally, cohorts of cases, which would be used as the training set in prospective studies, will change significantly over time as the laboratory tests, recording procedures, environment, treatment methods, and mean overall survival shift. The RMSurv approach will allow for the modeling of the true cross-modality correlations and feature distributions of the prospective test-set, but the underlying C-indices of each modality in the training set would be based on potentially outdated cohorts with stronger or weaker modalities. A system to interpret the changes and uncertainty of modalities could improve the robustness significantly.

Conclusions

In summary, this study highlights the complex relationships within multimodal cancer survival prediction, and introduces the RMSurv model, which uses synthetic data generation, time-dependent weighting, and a novel normalization process. This robust and interpretable system advances the progress toward clinical use of machine learning based survival prediction, even with small datasets.

ORCID iDs

Dominic Flack https://orcid.org/0009-0006-4083-5709
Dimah Dera https://orcid.org/0000-0002-7168-5858

Ethical Considerations

This study used publicly available, de-identified data from The Cancer Genome Atlas (TCGA) database. The original data was collected by the TCGA research network in compliance with all relevant ethical regulations. All patients provided informed consent for their data to be used for research purposes. No patients were involved in the design or reporting of this study.

Consent to Participate

Informed consent to participate was obtained from all individual participants by the TCGA research network at the time of tissue collection.

Consent for Publication

Not applicable.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Dominic Flack. The first draft of the manuscript was written by Dominic Flack, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The author, Dimah Dera, received support from the National Science Foundation, Award number CRII-2401828.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense, the United States Air Force, or the United States Space Force.

Data Availability Statement

This study used publicly available data from the TCGA project which was accessed through Xena: https://xenabrowser.net/datapages/ and MINDS: https://github.com/lab-rasool/MINDS. All code used in this project, including pre-processing code, is freely available on github.

Supplemental Material

Supplemental material for this article is available online.

References

1. Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. 2022;40:1095-1110.
2. Waqas A, Tripathi A, Ahmed S. SeNMo: a self-normalizing deep learning model for enhanced multi-omics data analysis in oncology. *arXiv: 2405.08226*, 2024.
3. Cox D. Regression models and life-tables. *J R Stat Soc Ser B*. 1972;34:187-220.
4. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. *Sci Rep*. 2021;11:13505.

5. Luo H, Huang J, Ju H, Zhou T, Ding W. Multimodal multi-instance evidence fusion neural networks for cancer survival prediction. *Sci Rep*. 2025;15:10470.
6. Gomaa A, Huang Y, Hagag A, et al. Comprehensive multimodal deep learning survival prediction enabled by a transformer architecture: a multicenter study in glioblastoma. *Neurooncol Adv*. 2024;6:vdae122.
7. Yang H, Wang J, Wang W, et al. MMsurv: a multimodal multi-instance multi-cancer survival prediction model integrating pathological images, clinical information, and sequencing data. *Brief Bioinform*. 2025;26:bbaf209.
8. Chen RJ, Lu MY, Williamson DF, et al. Pan-cancer integrative histology-genomic analysis via interpretable multimodal deep learning. *Cancer Cell*. 2022;40:865-878.e6.
9. Nikolaou N, Salazar D, RaviPrakash H, et al. A machine learning approach for multimodal data fusion for survival prediction in cancer patients. *NPJ Precis Oncol*. 2025;9:128.
10. Ching T, Zhu X, Garmire LX. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput Biol*. 2018;14:e1006076.
11. Chen RJ, Lu MY, Wang J, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging*. 2022;41:757-770.
12. Li R, Wu X, Li A, Wang M. HFBSurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics*. 2022;38:2587-2594.
13. Steyaert S, Pizurica M, Nagaraj D, et al. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat Mach Intell*. 2023;5(4):351-362.
14. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform*. 2022;23(2):3-5, 11-12.
15. Higham NJ. Computing the nearest correlation matrix–a problem from finance. *IMA J Numer Anal*. 2002;22:329-343.
16. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45:1113-1120.
17. Liu J, Lichtenberg T, Hoadley KA, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cells*. 2018;173:400-416.e11.
18. Moore L, Le T, Fan G. DNA methylation and its basic function. *Neuropsychopharmacology*. 2013;38:23-38.
19. Akbani R, Ng PK, Werner HM, et al. A pan-cancer proteomic perspective on the Cancer Genome Atlas. *Nat Commun*. 2014;5:3887.
20. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol*. 2020;38:675-678.
21. Tripathi A, Waqas A, Venkatesan K, Yilmaz Y, Rasool G. Building flexible, scalable, and Machine Learning-ready multimodal oncology datasets. *Sensors*. 2024;24:1634.
22. Moons K, Collins G, Reitsma J. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
23. Tripathi A, Waqas A, Yilmaz Y, Rasool G. HoneyBee: a scalable modular framework for creating multimodal oncology datasets with foundational embedding models. *arXiv: 2405.07460*, 2024.
24. Yang X, Chen A, PourNejatian N. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. *arXiv: 2203.03540*, 2022.
25. Chen R, Ding T, Lu M. A general-purpose self-supervised model for computational pathology. *arXiv: 2308.15474*. 2023.
26. Storn R, Price K. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J Glob Optim*. 1997;11:341-359.
27. Park S, Park J, Park S, Kim H. Review of statistical methods for evaluating the performance of survival prediction models. *Korean J Radiol*. 2021;22:213-224.