# Coxmos: Interpretable survival models for high-dimensional and multi-omic data

**Pedro Salguero**,[1] **Anabel Buendía-Galera**[1] **and Sonia Tarazona**[1,*]

[1]Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Valencia, 46022, Spain

*Corresponding author: sotacam@eio.upv.es

## Abstract

**Motivation:** Survival analysis is fundamental in many fields, including medicine. However, the analysis of high-dimensional datasets, such as omic or multi-omic data, presents significant challenges due to data dimensionality, multicollinearity, and the limited interpretability of existing methods. Classical Cox models struggle in these contexts, and machine learning approaches often lack transparency.

**Results:** To overcome these limitations, we present Coxmos, an R package that integrates adapted Cox regression models with variable selection and PLS-based approaches tailored for high-dimensional and multi-block data. Coxmos also provides validation, comparison, interpretation, and visualization tools. Benchmarking across clinical, omic, and multi-omic datasets showed that Coxmos outperformed other state-of-the-art machine learning methods, while offering biological interpretation. We showcased Coxmos' functionalities on an ovarian cancer dataset, highlighting its potential to integrate multiple omic layers, identify relevant predictors, and assess their impact on patient survival.

**Availability:** Coxmos is freely available at CRAN: https://cran.r-project.org/web/packages/Coxmos.

## 1. Introduction

Survival analysis is a fundamental statistical approach in medicine and other fields to model the time until a specific event occurs (such as death or disease onset) as a function of other covariates (age, gender, treatment, etc.), while accounting for censored observations. The *Cox Proportional Hazards (PH) model* (Cox, 1972) remains one of the most widely used methodologies. This semi-parametric model expresses the hazard at time $t$ for an individual with covariate vector $\mathbf{x}$ as in Equation 1.

$$h(t \mid \mathbf{x}) = h_0(t) \cdot e^{\beta^\top \mathbf{x}}, \tag{1}$$

where $h_0(t)$ is the baseline hazard function and $\beta$ the vector of coefficients associated to the covariates. The Cox model assumes that covariate effects are constant over time (PH assumption), requires that the number of observations $n$ exceeds the number of predictors $p$ and cannot properly handle multicollinearity.

Besides survival prediction, an essential aim of survival models is to identify the most influential survival predictors, a task that becomes increasingly complex as the number of predictors grows (Witten and Tibshirani, 2010). This is the case of high-dimensional (HD) datasets, including omic data, where the number of variables $p$ exceeds the number of observations $n$, multicollinearity is frequent, and the number of events per variable (EPV) —that determines the statistical power of the models— is too low (Concato et al., 1995; Peduzzi et al., 1995; Vittinghoff and McCulloch, 2007). To address these limitations, variable selection techniques such as stepwise procedures or shrinkage methods like Elastic Net have been proposed (Guyon and Elisseeff, 2003).

Machine learning (ML) approaches are another alternative to model survival in HD settings (Pölsterl et al., 2016; Leger et al., 2017; Spooner et al., 2020). Techniques such as Support Vector Machines (SVM), Boosted Models, Random Forests, or Neural Networks have been adapted to handle right-censored data (Shivaswamy et al., 2007; Tutz and Binder, 2006; Friedman, 2002; Chen and Guestrin, 2016; Ishwaran et al., 2008; Faraggi and Simon, 1995; Hothorn, 2006; Knottenbelt et al., 2025). ML models often achieve high accuracy, and some of them incorporate feature selection strategies, but their interpretability is still limited.

Conversely, dimensionality reduction techniques such as Partial Least Squares Regression (PLS) (Wold, 1966; Sjöström et al., 1983) provide a more interpretable alternative to deal with HD data. PLS models rely on a small number of latent variables that capture most of the variability in the original dataset and benefit from multicollinearity. Various adaptations of PLS for survival analysis have been developed. Nguyen and Rocke (Nguyen and Rocke, 2002) applied the PLS model on survival time without

considering censoring and used the extracted PLS components as predictors in a Cox PH model. Park et al. (Park et al., 2002) reformulated the PLS algorithm to introduce Poisson regression to handle censoring. Li and Gui (Li and Gui, 2004) integrated Cox models within the PLS algorithm, thus considering censored observations. Bastien (Bastien, 2008) developed the idea of using deviance residuals from a null Cox model as the response variable in the PLS model and implemented it in the *plsRcox* R package (Bastien et al., 2015), which also includes sparse (variable selection) and kernel-based (nonlinear relationship) versions. However, this package lacks tools for optimizing, evaluating and interpreting PLS-based survival models.

An even more complex scenario arises when dealing with multiblock (MB) datasets, such as those obtained from multi-omic experiments, where the relationship between blocks can increase the multicollinearity issue and complicate the identification of relevant survival predictors. Approaches such as MB Random Forest (Hornung and Wright, 2019), IPF Lasso (Boulesteix et al., 2017) or Priority Lasso (Klau et al., 2018) are interesting options for MB datasets. However, these models often provide limited insight into the relative importance of individual omic layers. While MB-PLS regression models have been widely used for MB data (Westerhuis et al., 1998), to our knowledge, they have not yet been adapted for survival analysis.

Despite the wide range of methodologies developed for survival analysis on HD or MB data, there are still some challenges to address, such as overfitting, model interpretability, or the lack of robust tools for model evaluation and comparison. Solving these issues requires integrating efficient cross-validation procedures, predictive performance metrics, and advanced visualization tools to support model selection and results interpretation. To overcome these limitations, we have developed Coxmos, an R package providing a unified and flexible framework for survival analysis in HD and MB settings. Coxmos implements four modeling strategies for HD data: the Cox proportional hazards model with HD-adapted stepwise or Elastic Net variable selection, and three PLS-based approaches: PLS-ICOX, PLS-DRCOX, and PLS-DACOX. It also extends the PLS approaches for survival modeling in MB studies. The package includes automated hyperparameter optimization via cross-validation, variable selection strategies, and different performance metrics for model optimization, evaluation and comparison.

## 2. Materials and methods

### 2.1. Overview of Coxmos package

We present Coxmos, a comprehensive R package that addresses the challenges of survival analysis for HD and MB data. Our package provides functionalities for data exploration and preprocessing, assumption validation, performance evaluation and comparison, model interpretation, and prediction (Figure 1). Survival methods in the package include the Cox PH model with improved stepwise and Elastic Net variable selection, and sparse PLS-based methods for single-matrix and MB datasets. We also implemented model optimization via cross-validation and advanced visualization capabilities to enhance model robustness and interpretability.

#### 2.1.1. Data exploration and preparation

Before survival analysis, in order to assess data quality, Coxmos generates visualizations for the distribution of events and censored observations within a specified time range. As the stability and statistical power of the model depend on the EPV ratio and a value lower than 10 increases the risk of overfitting and unstable estimates, the package also reports the EPV for the analysed dataset.

Finally, prior to model fitting, Coxmos prepares the data by transforming categorical variables into dummy variables and creates a dataset partition into training and test subsets for proper validation that preserves the distribution of event/censored observations.

#### 2.1.2. Survival models in Coxmos

**Cox PH models with variable selection**

Coxmos implements two variable selection strategies for the Cox PH model that have been adapted to HD scenarios: an improved stepwise procedure and the ElasticNet regularization method with a two-hyperparameter optimization.

*Stepwise variable selection (CoxSW)*

The classical stepwise (SW) method (Efroymson, 1960) —based on iteratively adding or removing variables— cannot always be used in HD scenarios since existing algorithms initialize from a full model with all the covariates, which cannot be fitted when $p \geq n$. In addition, the SW algorithm may compromise the survival predictive performance when the EPV criterion is not met (Flom and Cassell, 2007; Thompson, 1995).

In Coxmos, we adapted the SW procedure from the *My.stepwise* R package (Company, 2017) to better handle HD datasets. First, univariate Cox models are fitted for each covariate instead of a full model. For the backward option, the initial model includes the most significant variables according to their univariate p-values while fulfilling the EPV constraint imposed by the user. The forward option adds variables with the most significant univariate p-values until the EPV constraint is no longer met.

While *My.stepwise* only relies on ANOVA tests to iteratively compare models, Coxmos combines them with Akaike's Information Criterion (AIC) (Akaike, 1973) to select variables contributing to a measurable improvement of model performance. This modification enhances algorithmic convergence and ensures a more robust and consistent selection process.

*ElasticNet strategy (CoxEN)*

The ElasticNet (EN) regularization method (Simon et al., 2011) combines Lasso and Ridge approaches into the penalization term in Equation 2:

$$P_\alpha(\beta) = \lambda[\alpha \sum |\beta_i| + (1-\alpha) \sum \beta_i^2], \tag{2}$$

where $\lambda$ is the regularization parameter and $\alpha$ is a value between 0 and 1 which gives the weights for Lasso and Ridge penalizations. Coxmos applies an adaptation of the EN method in the *glmnet* R package (Simon et al., 2011). We included the EPV restriction and a custom cross-validation (CV) to also optimize the $\alpha$ parameter, since only $\lambda$ optimization is performed in *glmnet*.

**PLS-based survival models**

Coxmos implements two types of sparse PLS-based survival models (sPLS): (1) methods for HD datasets (e.g., omic data), comprising sPLS-ICOX, sPLS-DRCOX and sPLS-DACOX; and (2) models for

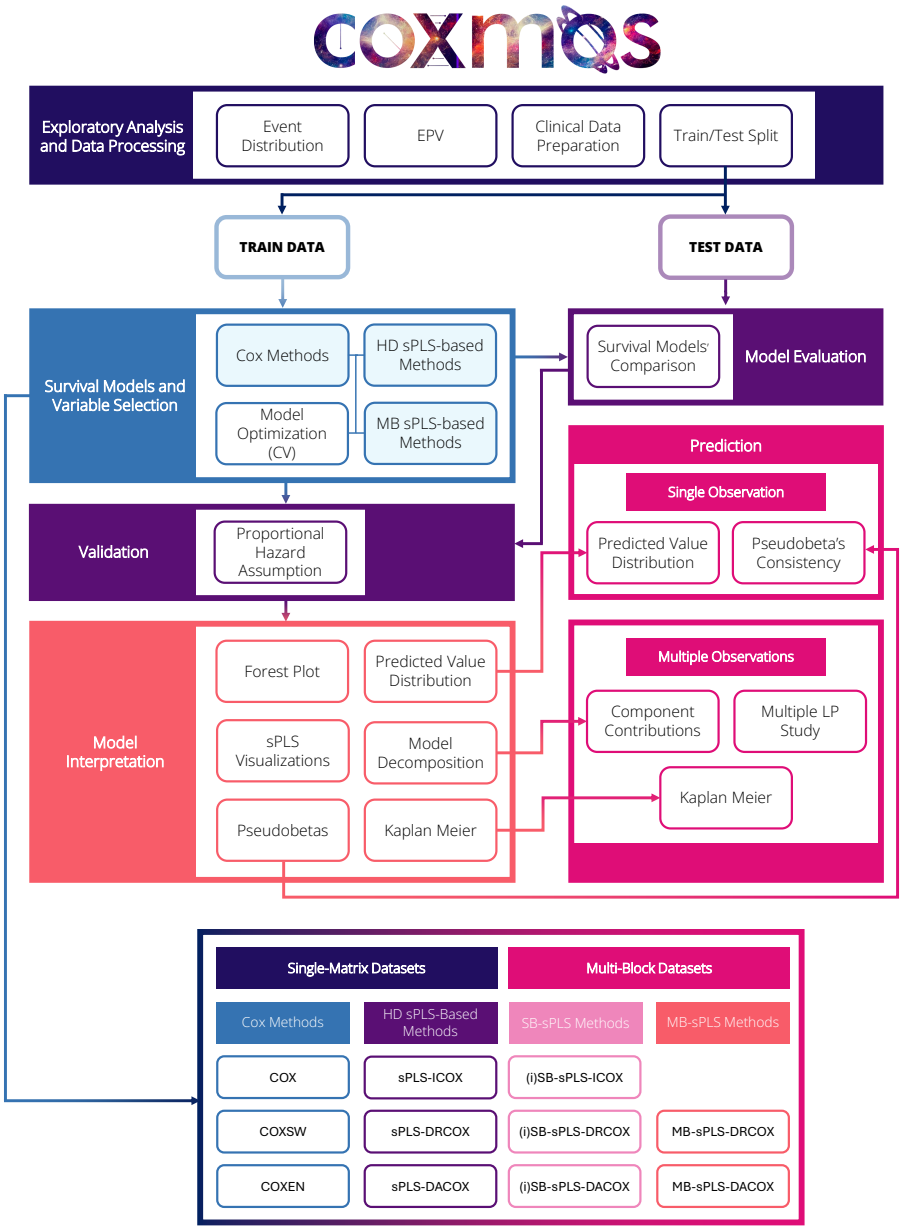


Figure 1: Overview of Coxmos package functionalities for survival analysis on high-dimensional (HD) and multi-block (MB) datasets.

MB datasets (e.g. multi-omic datasets), including two different approaches: single-block —(i)SB-sPLS-ICOX, (i)SB-sPLS-DRCOX and (i)SB-sPLS-DACOX— and multi-block —MB-sPLS-DRCOX and MB-sPLS-DACOX. For all models, hyperparameters are optimized by applying a cross-validation procedure (penalty factors, number of PLS components, number of variables to retain per component, etc.). Moreover, for the sake of transparency and reusability, the package returns all computed matrices (scores, residuals, loadings and weights). These sPLS survival models are briefly described in the following sections, but further details can be found in Suppl. Section 1.1.

PLS regression (Wold, 1966) models the relationship between a predictor matrix $\mathbf{X}$ and a response matrix $\mathbf{Y}$ by projecting them onto a low-dimensional latent space, while aiming to maximize the covariance between $\mathbf{X}$ and $\mathbf{Y}$ latent variables or components. In this work, we used the NIPALS iterative procedure (Höskuldsson, 1988) to estimate PLS models. A key step in the classical NIPALS algorithm is the estimation of the vector of weights $\mathbf{w_a}$ for each component $a = 1, ..., A$ (see step 2 in Suppl. Section 1.1.1), which relates $\mathbf{X}$ and $\mathbf{Y}$ in the PLS model.

*sPLS-ICOX*

The *sparse PLS individual Cox* (sPLS-ICOX) method in Coxmos was adapted from Bastien et al. (2015). In this approach, step 2 in NIPALS algorithm mentioned above is replaced with a univariate Cox regression model for each predictor $j$. Therefore, $w_{aj} = \hat{\beta}_{aj}$,

where $\hat{\beta}_{aj}$ is the estimated Cox regression coefficient of predictor $j$ in component $a$. The retained $A$ PLS components are the predictors in the final multivariate Cox model.

We complemented this methodology with a sparse approach to select the most relevant predictors, which sets $w_{aj} = 0$ when the associated $\hat{\beta_{aj}}$ is not significant for a given $\alpha$, which is optimized by CV strategies.

*sPLS-DRCOX*

The *sparse PLS Deviance Residual Cox* (sPLS-DRCOX) method is an extension of the splsRcox algorithm (Bastien, 2008). The response variable in the PLS model is the vector of deviance residuals computed from a null Cox model with no predictors. The resulting PLS components are the predictors in the final Cox regression model. In Coxmos, we implemented a heuristic algorithm based on the Lasso variable selection in Rohart et al. (2017), where a near-optimal combination of the number of variables selected per component is obtained.

*sPLS-DACOX*

In the *sparse PLS Discriminant Analysis Cox* (sPLSDA-Cox) model, the response variable is the censoring status (event vs. censored). Although time-to-event information is not used in the PLS-DA model, the resulting components are used as predictors in the final Cox model. This method incorporates the same heuristic variable selection strategy as in sPLS-DRCOX.

*Single-block sPLS methods for MB data*

This MB approach applies an sPLS-COX model to each block separately and all the components obtained for all blocks are the predictors in the final multivariate Cox survival model. The sPLS-COX model can be any of sPLS-ICOX, sPLS-DRCOX or sPLS-DACOX. Two strategies for selecting the number of components per block were implemented: the Single-Block (SB) approach (SB-sPLS-ICOX, SB-sPLS-DRCOX and SB-sPLS-DACOX), where the number of components and the sparse penalty, optimized by CV, are the same across all blocks; and the iterative SB (iSB) approach (iSB-sPLS-ICOX, iSB-sPLS-DRCOX and iSB-sPLS-DACOX), which allows a different number of components and penalty values for each block by performing independent CV.

*Multi-block sPLS methods*

In this case, all blocks are simultaneously analyzed in the same sPLS-based survival model (Tenenhaus and Tenenhaus, 2011; Tenenhaus et al., 2014). The two methods implemented, MB-sPLS-DRCOX and MB-sPLS-DACOX, are based on Rohart et al. (2017) and computed using *mixOmics*, where a block-correlation matrix reflecting relationships among blocks is used. Coxmos can automatically estimate this matrix from pairwise PLS models across all omic blocks. The heuristic variable selection procedure previously described is also integrated into these methods.

### 2.1.3. Model optimization and evaluation

**Cross-validation**

Coxmos applies a repeated k-fold CV on the training data to optimize model hyperparameters. Users can choose the number of folds and repetitions, as well as the performance metric or a weighted combination of them.

**Predictive performance metrics**

Coxmos includes different performance metrics for survival: the Akaike Information Criterion (AIC) (Akaike, 1973), the C-index (Harrell et al., 1982), the area under the Receiver Operating Characteristic (ROC) curve (AUC) and the Integrated Brier Score (IBS) (Graf et al., 1999). For AUC, six estimation procedures were implemented following the cumulative/dynamic (C/D) approach or the incident/dynamic (I/D) strategy: *survivalROC* (C/D) (Heagerty et al., 2000), *cenROC* (C/D) (Beyene and Ghouch, 2020), *nsROC* (C/D) (Pérez-Fernández et al., 2018), *smoothROCtime* (C/D and I/D) (Díaz-Coto et al., 2020), and *risksetROC* (I/D) (Heagerty and Zheng, 2005). As AUC can be computed for each time point, Coxmos provides both AUC measurements across time and the average of AUC values for a set of time points.

The package also renders plots and statistical methods (t-test, ANOVA, Wilcoxon rank-sum test, and Kruskal-Wallis test) to evaluate and compare different models and identify the best-performing ones according to these metrics.

### 2.1.4. Model validation

The Schoenfeld residuals are used in Coxmos to check the proportional hazards (PH) assumption in survival models by integrating functionalities from *survival::cox.zph* (Therneau, 2023) and *survminer::ggcoxzph* (Kassambara et al., 2024). Both residual plots and Schoenfeld's statistical tests were implemented.

### 2.1.5. Model interpretation

**Forest Plot**

Coxmos generates forest plots summarizing the estimated effects of covariates or PLS components, which were adapted from the *survminer* R package (Kassambara et al., 2021).

**Predicted value distribution**

Coxmos provides different types of predictions from the training data using (Therneau, 2023): linear predictor (lp), risk, expected number of events, and survival probability. Also, density plots and histograms are offered to visualize the distribution of predicted values grouped by event status to assess the model's discriminative ability.

**Variable or component contributions**

Coxmos plots the contribution to AUC of each original variable, PLS component, or linear predictor, over a range of time points, to identify the most relevant variables for model's performance.

**sPLS visualization**

In Coxmos, PLS-based models can be interpreted through loading plots for variable contribution to components, score plots with projected observations, or biplots, which combine variables and observations.

**Pseudobeta coefficients**

The estimated Cox regression coefficients $\hat{\beta}$ (Equation 1) reflect both the magnitude and direction of each predictor's effect, i.e., whether they act as a protective or risk factor, and are key to the interpretation of survival models. However, in sPLS-Cox models, the predictors included in the final Cox model are PLS components, which limits the interpretability at the original variable level. To better assess the effect of each original predictor, we defined the pseudobeta coefficients by leveraging the relationship between the coefficients ($\hat{\beta}$) of the PLS components $\mathbf{T}$ and the transformed weights ($W^*$) of the original variables, which connect response and predictors in PLS models (see Suppl. Section 1.2).

**Kaplan-Meier curves**

The Kaplan-Meier (KM) curves in the *survminer* R package (Kassambara et al., 2024) combined with the log-rank test were adapted in Coxmos to be displayed for all types of models and can be applied to the model's linear predictor, specific PLS components, or original variables. For numerical variables, an optimal cutoff to divide patients into two groups of lower and higher survival rates in KM curves is estimated by Coxmos and can be used in test data to assess the model's ability to differentiate potential risk groups.

#### 2.1.6. Survival prediction for new observations

Given a new set of patients with their corresponding covariate values, making predictions from a fitted Cox regression model is straightforward. However, for PLS-based survival models, the covariate values must be properly scaled and transformed into scores prior to obtaining such predictions. Coxmos supports this functionality, and most of the previously described tools can also be applied to these new observations (see Suppl. Section 1.3 for further details).

For a single observation, the predicted value can also be added to the training predicted value distribution (see previous section) to visualize its position relative to the training patients. Additionally, the new patient's pseudobeta coefficients can be compared with those from the trained model. In the case of multiple new observations (e.g. test data), the variable or component contribution to the model's performance can be obtained to assess which variables have more impact on survival and to detect possible overfitting. As with single observations, pseudobeta coefficient plots can also be generated simultaneously for multiple patients. Finally, to evaluate model performance on unseen data, cutoff values from training KM curves can be applied to stratify new patients and assess the model's ability to separate them into groups with different survival probabilities, according to the log-rank test p-value.

### 2.2. Model comparison

We compared the predictive performance of survival models in Coxmos and machine learning (ML) methods on different datasets by using the C-Index and AUC metrics. An Analysis of variance (ANOVA) was applied to identify significant differences between methods and datasets (block factor), and least significant difference (LSD) intervals at 5% significance level allowed pairwise comparisons for the identification of the best-performing methods.

#### 2.2.1. Coxmos models

To compare models in Coxmos and benchmark them against state-of-the-art survival analysis tools, we fitted them on the training data and optimized them via cross-validation. Specifically, we used a 5-fold cross-validation with 3 repetitions on 80% of each dataset, leaving 20% of the data for test, where performance metrics for model comparison were computed.

The classical Cox model and CoxSW do not require hyperparameter tuning, whereas the two regularization parameters are optimized in the CoxEN model. For PLS-based methods, the tuned hyperparameters depend on the model: for sPLS-ICOX, both the number of PLS components and the penalty factor for variable selection; for sPLS-DRCOX and sPLS-DACOX, the number of PLS components and the number of variables to retain per component; and for (i)SB-sPLS and MB-sPLS models, the same as for their HD counterparts. The block-correlation matrix in MB-sPLS models was computed automatically by Coxmos. HD methods in Coxmos were also applied to MB datasets as an alternative strategy, after concatenating and block-scaling the omic layers (see Suppl. Section 1.4).

The metric used for model optimization was the average AUC at 15 equally spaced time points. Penalty parameters were explored over the range 0.1-0.9 in increments of 0.2. For the number of variables to retain per component in the heuristic search, the minimum number of variables was set to 1, the maximum to the total number of variables, and the initial number of cutpoints to 10. The maximum number of PLS components considered was 8.

#### 2.2.2. Machine Learning models for survival analysis

We compared our PLS-based survival methods to ML survival algorithms. For HD data, we tried CoxBoost (Binder and Schumacher, 2008) from *CoxBoost* R package v1.5 (Binder, 2020); MBoost (Hofner et al., 2014) from the *mboost* R package v2.9.11 (Hothorn et al., 2024); and Random Survival Forests (RSF) (Ishwaran et al., 2008) from the *randomForestSRC* R package v3.4.1 (Ishwaran et al., 2023). For MB data, we considered IPF-LASSO (Boulesteix et al., 2017) from the *ipflasso* R package v1.1 (Boulesteix et al., 2019); Priority-LASSO (Klau et al., 2018) from the *prioritylasso* R package v0.3.1 (Klau et al., 2023); and Block Forest (Hornung and Wright, 2019) from the *blockForest* R package v0.2.6 (Hornung and Wright, 2023). To ensure a fair comparison with Coxmos, we also applied a 5-fold cross-validation with the same training and test partitions (further details in Suppl. Section 1.5). We also fitted RSF and Block Forest models on selected variables (SV) according to their importance in the complete model (denoted as RSF SV and Block Forest SV). The number of selected variables in each of these models was set to the maximum number selected by Coxmos models for the same dataset.

### 2.3. Data

The methods implemented in the Coxmos R package were tested and compared on various HD and multi-omic datasets with different characteristics (Table 1). More details about each study and data preprocessing can be found in the Suppl. Section 1.6.

**Table 1.** Description of datasets used in this work, including the number of patients and variables, the number of events, and the EPV value.

| Dataset | Data Type | N. Patients | N. Var. | N. Event | EPV |
|---|---|---|---|---|---|
| BRCA | RNA-Seq | 571 | 19318 | 68 | 0.004 |
| BRCA | miRNA | 571 | 642 | 68 | 0.106 |
| BRCA | Proteomic | 571 | 518 | 68 | 0.184 |
| BRCA | Methylation | 571 | 17436 | 68 | 0.004 |
| BRCA | Clinical | 571 | 22 | 68 | 6.8 |
| HNSC | RNA-Seq | 443 | 21520 | 152 | 0.007 |
| HNSC | miRNA | 443 | 793 | 152 | 0.192 |
| HNSC | Mutation | 443 | 106 | 152 | 1.43 |
| HNSC | CNV | 443 | 57964 | 152 | 0.003 |
| HNSC | Clinical | 443 | 7 | 152 | 21.714 |
| HGSOC | RNA-Seq | 290 | 12370 | 176 | 0.0142 |
| HGSOC | miRNA | 290 | 523 | 176 | 0.3365 |
| HGSOC | CNV | 290 | 15055 | 176 | 0.0117 |
| HGSOC | Methylation | 290 | 17441 | 176 | 0.0101 |

To showcase Coxmos functionalities, we used the High-Grade Serous Ovarian Cancer (HGSOC) dataset from The Cancer Genome Atlas Ovarian Cancer Cohort (Bell et al.) described in Table 1. The study lasted 5481 days and included 290 patients, 60.69% of whom experienced the event (death). Four omic modalities were measured: gene expression, microRNA expression, Copy Number Variations (CNV), and DNA methylation, and data was preprocessed with the ATHAR pipeline (see Suppl. Section 1.6.2).

## 3. Results and discussion

### 3.1. Comparison of HD methods on clinical datasets

We first compared the performance of Cox and PLS-based models in Coxmos to ML methods on the clinical datasets, where the full Cox regression model with no variable selection could also be applied. The ANOVA tests (Suppl. Tables S1 and S2) revealed a statistically significant effect of the dataset on model performance for both the C-Index (p-val $< 0.0001$) and AUC (p-val $= 0.002$) metrics, showing that the characteristics of the data may influence the choice of the survival model. In contrast, almost statistically significant differences were found between methods for these metrics (C-Index p-val $= 0.1087$ and AUC p-val $= 0.0802$), suggesting a similar performance of the compared models on low HD data, such as clinical data. Still, the consequently overlapping LSD intervals for the C-Index (Fig.2A) showed that PLS-based survival models (C-index mean $\bar{x}_C \simeq 0.7$) slightly outperform the full Cox and CoxSW models ($\bar{x}_C = 0.69$), while CoxEN and ML boost models perform worse ($\bar{x}_C < 0.65$). RF was the best performing ML method on average, close to Coxmos PLS-based models. Among these, sPLS-DRCOX achieved the highest mean C-Index and AUC values (Suppl. Figure S1A).

### 3.2. Comparison of HD methods on omic data

The previous HD survival models were also compared on omic datasets, which present a much higher dimensionality than clinical data. Obviously, we had to exclude the full Cox regression model from this comparison.

Unlike in the clinical data, the ANOVA test for omic data (Suppl. Tables S3 and S4) revealed statistically significant differences in model performance across both datasets and methods for the C-Index and AUC metrics (p-val $< 0.0001$). The C-Index LSD intervals (Fig. 2B) showed that all ML methods (with $\bar{x}_C < 0.65$) and coxEN ($\bar{x}_C \simeq 0.66$) presented a significantly lower performance than the rest of Coxmos methods (from $\bar{x}_C \simeq 0.76$ for sPLS-DACOX to nearly 0.84 for sPLS-DRCOX). AUC results (Suppl. Figure S1B) mostly confirmed these findings. However, while sPLS-DRCOX stands out as the best-performing method for the C-Index, coxSW achieved the highest AUC value ($\bar{x}_{AUC} \simeq 0.65$). The overlapping LSD intervals for most Coxmos models, although indicating non-significant differences between them, suggest that Coxmos PLS-based survival models can properly handle extremely high dimensionality and multicollinearity present in omic data. Although our adapted version of coxSW rendered good results, it must be noted that applying it is not always feasible under a HD scenario due to computational restrictions, as happened with CNVs in the HNSC dataset.

### 3.3. Comparison of MB methods

Finally, we evaluated models for MB data: the multi-omic BRCA, HNSC and HGSOC studies. We considered four groups of models:

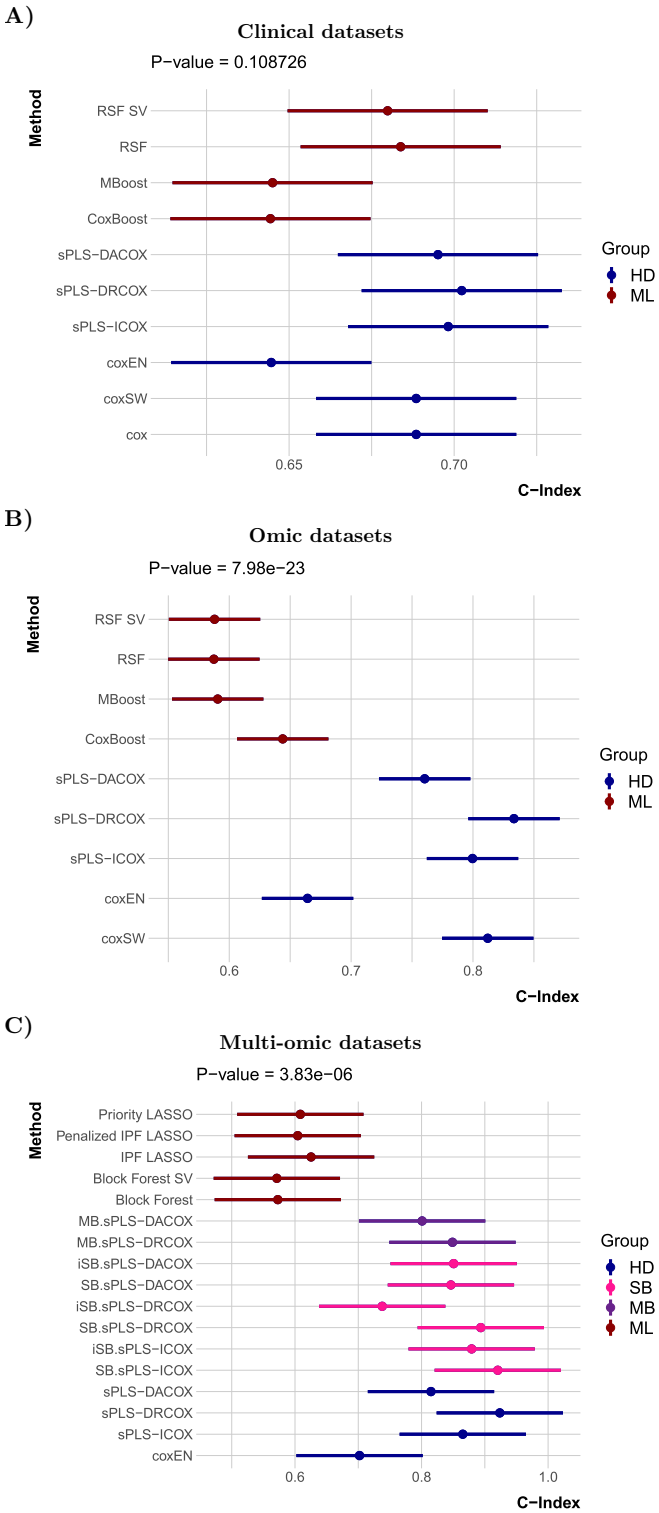


Figure 2: LSD intervals at 5% significance level for the C-Index computed on test data. (A) Results for clinical data from BRCA and HNSC datasets for HD and ML methods. (B) Results for omic data from BRCA, HNSC and HGSOC datasets for HD and ML methods. (C) Results for multi-omic data from BRCA, HNSC and HGSOC datasets for HD, MB and ML methods.

HD methods in Coxmos applied to concatenated and block-scaled datasets; SB and MB models in Coxmos, which directly handle MB data; and ML models for MB datasets. Considering the size of these multi-omic datasets, full Cox and CoxSW models had to be excluded from the comparison. We also excluded non-MB ML methods due to the poor performance obtained in the previous comparisons.

The ANOVA tests (Suppl. Tables S5 and S6) revealed a significant effect of the dataset on the C-Index and AUC (p-val = 0.009 and p-val < 0.0001, respectively) and also for the method (p-val < 0.0001 and p-val = 0.0008, respectively). The C-Index LSD intervals (Fig.2C) showed that all Coxmos methods obtained an average C-Index higher than 0.7 and clearly outperformed ML models ($\bar{x}_C \leq 0.62$). Although LSD intervals overlapped for Coxmos methods, SB.sPLS-ICOX and sPLS-DRCOX presented a slightly better performance on average, while coxEN obtained the worst average performance among Coxmos models. A similar trend was observed for the AUC metric (Suppl. Figure 1C): ML-based methods yielded the lowest values, followed closely by MB approaches, while SB and HD models achieved the highest AUC scores. In particular, SB.sPLS-DRCOX and sPLS-DRCOX showed the best AUC values.

Again, PLS-based survival methods in Coxmos showed their ability to efficiently analyse complex and HD data while properly addressing the multicollinearity issue. Interestingly, we found no clear superiority of MB methods over the concatenation strategy, or for the more flexible iSB models over their SB counterparts. The choice of the PLS-based strategy may depend on the characteristics of the dataset and highlights the importance of having robust tools to optimize and compare different models, as those provided by Coxmos.

### 3.4. Case study on ovarian cancer

This section showcases Coxmos' functionalities by analysing a multi-omic High-Grade Serous Ovarian Cancer (HGSOC) dataset from TGCA (Bell et al.). All figures and results shown were generated with the Coxmos R package.

A first exploration of the number of patients across time intervals and the proportion of them reaching the event of interest (Figure 3) shows a higher concentration of events at earlier time points. This may reduce statistical power and hamper reliable prediction at later time intervals, with a lower number of observations.

For validating survival models, the dataset was split into training (80%) and test observations (20%). The EPV values per omic block in the training set are below the commonly recommended threshold of 5–10 EPVs (Table 2), which reflects the pronounced high dimensionality of this dataset, requiring efficient multivariate approaches that can handle such HD data.

**Table 2.** Event per variable (EPV) values per omic block.

| Block | CNV | Genes | Methylation | miRNA |
|---|---|---|---|---|
| EPV | 0.0093 | 0.0113 | 0.0080 | 0.2676 |

In this complex scenario, the optimization and evaluation of different survival models is essential, and Coxmos provides the tools that facilitate this task. We compared the PLS-based survival methods in Coxmos —SB and MB methods, and HD methods on concatenated and scaled blocks— and coxEN. After hyperparameter tuning via 5-fold CV on training data with 3 repetitions, the C-Index performance metric was calculated for the optimized model on the test set (Fig. 4). The best predictive performance was achieved by sPLS-DRCOX (C-Index = 0.95), followed closely by SB.sPLS-ICOX and SB.sPLS-DRCOX (C-Index = 0.92).

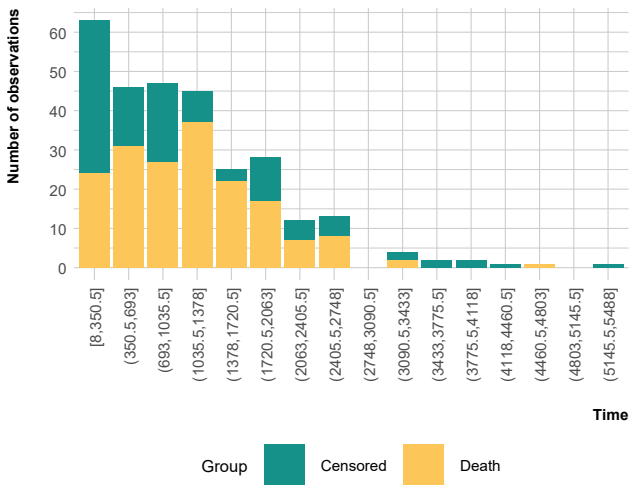


Figure 3: Distribution of patients and survival events per day interval in HGSOC dataset.

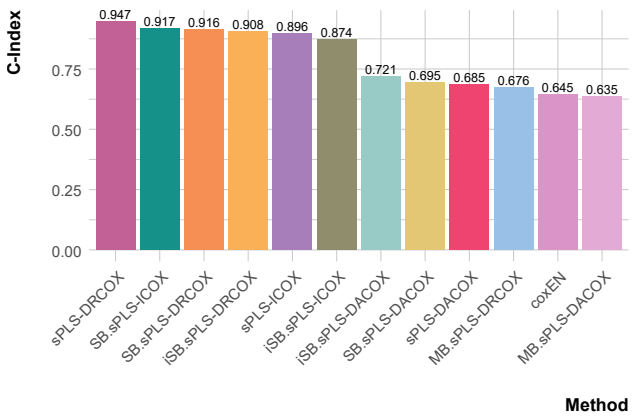


Figure 4: C-Index for survival models in Coxmos on the test set of HGSOC dataset.

Figure 5A shows that the PH assumption is met (Schoenfeld's test p-val = 0.102) for the selected model (sPLS-DRCOX), thus supporting its statistical validity for survival prediction. Although component 4 is statistically significant (p-val = 0.0053), Schoenfeld's residuals are randomly distributed around zero without any clear temporal trend, suggesting no relevant PH assumption violation.

Model optimization yielded six statistically significant PLS components (Table 3 and Fig. 5B), which act as risk factors (Hazard Ratio > 1). To assess each component's predictive performance, Fig. 5C shows their contribution to the AUC in the

**Table 3.** Results for the sPLS-DRCOX model on HGSOC dataset.

| Variable | Coef | exp(Coef) | se(Coef) | Robust se | z | Pr(>\|z\|) |
|---|---|---|---|---|---|---|
| comp_1 | 0.4977 | 1.6449 | 0.0409 | 0.0313 | 15.89 | $7.37 \times 10^{-57}$ |
| comp_2 | 1.3321 | 3.7892 | 0.1020 | 0.0892 | 14.94 | $1.89 \times 10^{-50}$ |
| comp_3 | 0.8525 | 2.3455 | 0.0733 | 0.0547 | 15.57 | $1.11 \times 10^{-54}$ |
| comp_4 | 0.5912 | 1.8062 | 0.0533 | 0.0481 | 12.30 | $9.21 \times 10^{-35}$ |
| comp_5 | 0.7498 | 2.1166 | 0.0670 | 0.0569 | 13.17 | $1.33 \times 10^{-39}$ |
| comp_6 | 0.7456 | 2.1077 | 0.0696 | 0.0707 | 10.55 | $5.27 \times 10^{-26}$ |

training and test datasets. In the training set, the LP consistently achieved higher AUC values across time than any single component. In the test set, the first component outperforms the LP at earlier or later times, probably due to the lower number of observations at these time intervals, rather than a sign of overfitting. Despite this, the generally higher AUC of the LP suggests that multi-layer integration captures more biological signal than any individual layer. The discriminative ability of the model is also supported by the distributions of LP values for censored or event patients (Fig. 5D), as the event ones present, in general, higher LP values (lower survival probability) than the rest.

As the interpretation of PLS components is not straightforward, a key feature of Coxmos is the calculation of pseudobeta coefficients for original values. Fig. 5E shows the top 10 variables with the highest pseudobetas. Interestingly, all of them are miRNAs, seven of which act as risk factors. While Resnick et al. (2009) reported higher serum levels of miR-92 in ovarian cancer patients versus controls, suggesting that miR-92 expression might be associated with poorer survival, the protective effect of miR-148a in the pseudobetas is supported by Zhu et al. (2019) and Zhao et al. (2016), who stated that overexpression of miR-148a reduces cell viability, migration, and invasion of ovarian cancer.

Finally, KM curves (Fig. 5F) showed that the LP clearly separates patients into low- and high-risk groups in the training set (log-rank p-val $< 0.0001$), while the separation is not so clear in the test set (p-val $= 0.1$) but maintains the direction of the effect. Notably, the test KM curves converge at late time points, where the number of observations declines, reducing precision and power, which may contribute to the non-significant result. These results highlight the usefulness of Coxmos' functionalities for conducting a rigorous survival analysis, enabling the identification of both the strengths and weaknesses of the models.

## 4. Conclusion

In this work, we introduced Coxmos, an R package that provides a unified framework for survival analysis in HD and MB settings. Coxmos integrates adapted classical methods with variable selection, such as CoxSW and CoxEN, together with PLS-based approaches specifically designed for HD and MB data. In addition, Coxmos incorporates tools for model optimization, validation, comparison, interpretation, and visualization, supporting reliable analyses and facilitating biological interpretation.

The comparison of methods across clinical, omic, and multiomic datasets revealed several key insights. First, in clinical datasets, model performance was more strongly influenced by the dataset itself than by the modeling strategy. Second, results on omic datasets confirmed the advantages of our PLS-based strategies in HD scenarios, where traditional approaches often struggle. Third, although MB methods were specifically designed for multiomic data and showed strong predictive potential, concatenation with appropriate block scaling often achieved comparable results. Finally and importantly, benchmarking Coxmos against state-of-the-art machine learning approaches demonstrated that the PLS-based methods achieve competitive, and in many cases superior, performance. Combined with their enhanced interpretability, this makes them particularly suitable for survival analysis.

To demonstrate its practical application, Coxmos was applied to an ovarian cancer dataset, showcasing its ability to integrate multiple omic layers, identify relevant predictors, and assess their impact on patient survival. Moreover, Coxmos tools allow the detection of model limitations such as the lack of prediction generalizability due to reduced sample size or the need for careful validation.

In conclusion, Coxmos offers a robust and interpretable framework for survival analysis in complex data contexts, while highlighting the importance of model choice, optimization and validation.

## Competing interests

No competing interest is declared.

## Author contributions statement

S.T. and P.S. conceived the Coxmos idea. P.S. implemented the current version of Coxmos and performed the analyses. A.B.G. helped with the analyses. S.T. supervised the work. S.T., P.S. and A.B.G. wrote and reviewed the manuscript.

## Funding

This work was supported by the Scientific Foundation of the Spanish Association Against Cancer through the project PERME224336TARA and by Instituto de Salud Carlos III through the project AC22/00058 (co-funded by the European Union as part of the Next Generation EU programme and the Recovery and Resilience Mechanism, MRR), both projects under the frame of ERA PerMed (ERAPERMED2022-141 - OVA-PDM). This work was also funded by Generalitat Valenciana (ACIF/2019/081), Spanish MICIN (PID2020-119537RB-100), and Spanish MICIU (PID2023-152976NB-I00).

## Data availability

The databases used in this work are public and were downloaded from The Cancer Genome Atlas Program (TCGA) `https://www.cancer.gov/ccg/research/genome-sequencing/tcga` (2025). The TCGA-BRCA dataset was preprocessed by our team and is available at GitHub `https://github.com/pilargmarch/multiomics2.0`. The code used for the analyses is available at `https://github.com/BiostatOmics/Coxmos_scripts`.

The Coxmos R package and two step-by-step tutorials for its application are freely available at the CRAN repository webpage `https://cran.r-project.org/web/packages/Coxmos/index.html` and at our GitHub `https://github.com/BiostatOmics/Coxmos`.

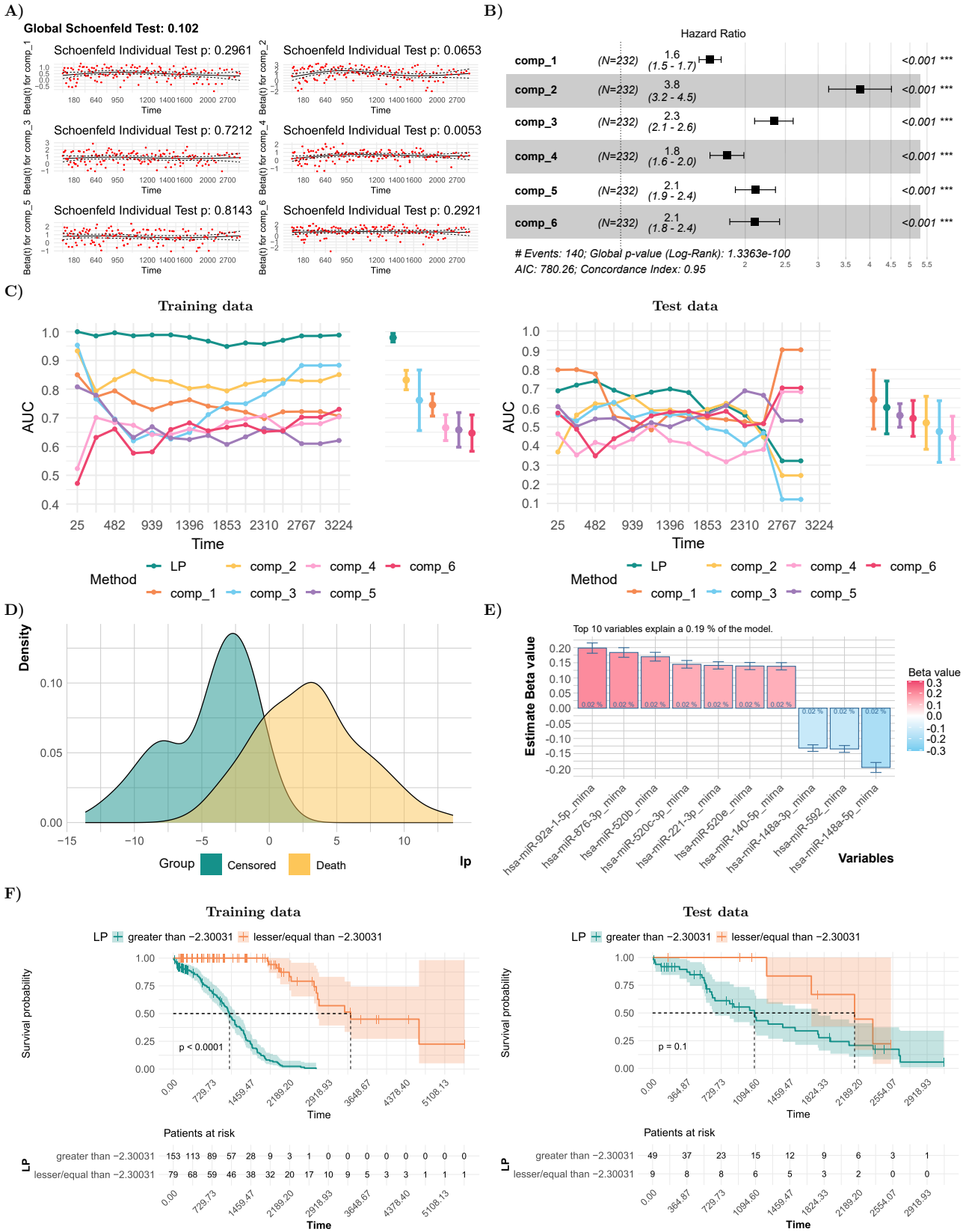


Figure 5: Coxmos results on the HGSOC dataset for the sPLS-DRCOX model. A) Schoenfeld residual test and plots for the model and PLS components. B) Forest plot showing the Hazards Ratio and 95% confidence intervals. C) AUC decomposition into individual components using the training (left) and test (right) data. D) Density plots for the LP per group. E) Estimated pseudobetas for the top 10 variables. F) Kaplan-Meier curve for the LP on the training (left) and test (right) data.

## References


H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*. Springer New York, New York, NY, 1973.

P. Bastien. Deviance residuals based pls regression for censored data in high dimensional setting. *Chemometrics and Intelligent Laboratory Systems*, 2008. doi: 10.1016/j.chemolab.2007.09.009.

P. Bastien, F. Bertrand, N. Meyer, et al. Deviance residuals-based sparse pls and sparse kernel pls regression for censored data. *Bioinformatics*, 2015. doi: 10.1093/bioinformatics/btu660.

D. Bell, A. Berchuck, M. Birrer, et al.

K. M. Beyene and A. E. Ghouch. Smoothed time-dependent receiver operating characteristic curve for right censored survival data. *Statistics in Medicine*, 39:3373–3396, 10 2020. ISSN 10970258. doi: 10.1002/sim.8671.

H. Binder. *Cox Models by Likelihood Based Boosting for a Single Survival Endpoint or Competing Risks*, 2020. URL `https://github.com/binderh/CoxBoost`.

H. Binder and M. Schumacher. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 2008. doi: 10.1186/1471-2105-9-14.

A.-L. Boulesteix, R. D. Bin, X. Jiang, et al. Ipf-lasso: Integrative l1-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine*, 2017. doi: 10.1155/2017/7691937.

A.-L. Boulesteix, M. Fuchs, and G. Schulze. *ipflasso: Integrative Lasso with Penalty Factors*, 2019. URL `https://cran.r-project.org/web/packages/ipflasso`.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939785.

I.-H. S. C. Company. My.stepwise: Stepwise variable selection procedures for regression analysis. `https://cran.r-project.org/web/packages/My.stepwise/index.html`, 06 2017.

J. Concato, P. Peduzzi, T. R. Holford, et al. Importance of events per independent variable in proportional hazards analysis i. background, goals, and general strategy. *Journal of Clinical Epidemiology*, 1995. doi: 10.1016/0895-4356(95)00510-2.

D. Cox. Regression models and life tables (with discussion. *Royal Statistical Society*, 1972. doi: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

S. Díaz-Coto, P. Martínez-Camblor, and S. Pérez-Fernández. smoothroctime: an r package for time-dependent roc curve estimation. *Computational Statistics*, 35:1231–1251, 9 2020. ISSN 16139658. doi: 10.1007/s00180-020-00955-7.

M. A. Efroymson. Multiple regression analysis. *Mathematical Methods for Digital Computers*, 1960.

D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995. doi: 10.1002/sim.4780140108.

P. L. Flom and D. L. Cassell. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In *NorthEast SAS Users Group Inc 20th Annual Conference*, volume 11, 2007.

J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 2002. doi: 10.1016/s0167-9473(01)00065-2.

E. Graf, C. Schmoor, C. Schmoor, et al. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 1999. doi: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5.

I. Guyon and A. Elisseeff. An introduction of variable and feature selection. *J. Machine Learning Research Special Issue on Variable and Feature Selection*, 3:1157 – 1182, 01 2003. doi: 10.1162/153244303322753616.

J. Harrell, Frank E., R. M. Califf, D. B. Pryor, et al. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030. URL `https://doi.org/10.1001/jama.1982.03320430047030`.

P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61:92–105, 2005. doi: https://doi.org/10.1111/j.0006-341x.2005.030814.x.

P. J. Heagerty, T. Lumley, and M. S. Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 2000.

B. Hofner, A. Mayr, N. Robinzonov, et al. Model-based boosting in r: a hands-on tutorial using the r package mboost. *Computational Statistics*, 2014. doi: 10.1007/s00180-012-0382-5.

R. Hornung and M. N. Wright. Block forests: random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics*, 2019. doi: 10.1186/s12859-019-2942-y.

R. Hornung and M. N. Wright. *blockForest: Block Forests: Random Forests for Blocks of Clinical and Omics Covariate Data*, 2023. URL `https://cran.r-project.org/web/packages/blockForest`.

T. Hothorn. Survival ensembles. *Biostatistics*, 2006. doi: 10.1093/biostatistics/kxj011.

T. Hothorn, P. Buehlmann, T. Kneib, et al. *mboost: Model-Based Boosting*, 2024. URL `https://cran.r-project.org/web/packages/mboost`.

A. Höskuldsson. Pls regression methods. *Journal of Chemometrics*, 1988. doi: 10.1002/cem.1180020306.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, et al. Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860, 2008. doi: 10.1214/08-AOAS169. URL `https://doi.org/10.1214/08-AOAS169`.

H. Ishwaran, M. S. Lauer, E. H. Blackstone, et al. *Random Survival Forests*, 2023. URL `https://www.randomforestsrc.org/articles/survival.html`.

A. Kassambara, M. Kosinski, and P. Biecek. *survminer: Drawing Survival Curves using 'ggplot2'*, 2021. URL `https://CRAN.R-project.org/package=survminer`. R package version 0.4.9.

A. Kassambara, M. Kosinski, and P. Biecek. *survminer: Drawing Survival Curves using 'ggplot2'*, 2024. URL `https://github.com/kassambara/survminer`. R package version 0.5.0.999.

S. Klau, V. Jurinovic, R. Hornung, et al. Priority-lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*, 2018. doi: 10.1186/s12859-018-2344-6.

S. Klau, R. Hornung, A. Bauer, et al. *prioritylasso: Analyzing Multiple Omics Data with an Offset Approach*, 2023. URL `https://cran.r-project.org/web/packages/prioritylasso`.

W. Knottenbelt, W. McGough, R. Wray, et al. Coxkan: Kolmogorov-arnold networks for interpretable, high-performance survival analysis. *Bioinformatics*, 41(8):btaf413, 07 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf413. URL `https://doi.org/10.1093/bioinformatics/btaf413`.

S. Leger, A. Zwanenburg, K. Pilz, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Scientific Reports*, 7(1):13206, 2017. doi: 10.1038/s41598-017-13448-3.

H. Li and J. Gui. Partial cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20 Suppl 1:i208–i215, 2004. doi: 10.1093/bioinformatics/bth900.

D. V. Nguyen and D. M. Rocke. Partial least squares proportional hazard regression for application to dna microarray survival data. *Bioinformatics*, 2002. doi: 10.1093/bioinformatics/18.12.1625.

P. J. Park, L. Tian, and I. S. Kohane. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18 Suppl 1:S120–S127, 2002. doi: 10.1093/bioinformatics/18.suppl_1.s120.

P. Peduzzi, J. Concato, A. R. Feinstein, et al. Importance of events per independent variable in proportional hazards regression analysis. ii. accuracy and precision of regression estimates. *Journal of Clinical Epidemiology*, 1995. doi: 10.1016/0895-4356(95)00048-8.

S. Pérez-Fernández, P. Martínez-Camblor, P. Filzmoser, et al. nsroc: An r package for non-standard roc curve analysis. *The R Journal*, 2018. doi: 10.1007/s00180-020-00955-7.

S. Pölsterl, N. Navab, and A. Katouzian. An efficient training algorithm for kernel survival support vector machines. 11 2016. doi: 10.48550/arXiv.1611.07054.

K. E. Resnick, H. Alder, J. P. Hagan, et al. The detection of differentially expressed micrornas from the serum of ovarian cancer patients using a novel real-time pcr platform. *Gynecologic Oncology*, 112(1):55–59, 2009. ISSN 0090-8258. doi: https://doi.org/10.1016/j.ygyno.2008.08.036. URL `https://www.sciencedirect.com/science/article/pii/S0090825808006835`.

F. Rohart, B. Gautier, A. Singh, et al. mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13, 11 2017. ISSN 15537358. doi: 10.1371/journal.pcbi.1005752.

P. K. Shivaswamy, W. Chu, and M. Jansche. A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 655–660, Oct 2007. doi: 10.1109/ICDM.2007.93.

N. Simon, J. H. Friedman, J. H. Friedman, et al. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 2011. doi: 10.18637/jss.v039.i05.

M. Sjöström, S. Wold, W. Lindberg, et al. A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables. *Analytica Chimica Acta*, 1983. doi: 10.1016/s0003-2670(00)85460-4.

A. Spooner, E. Chen, A. Sowmya, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10(1): 20410, 2020. doi: 10.1038/s41598-020-77220-w.

A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 2011. doi: 10.1007/s11336-011-9206-8.

A. Tenenhaus, C. Philippe, V. Guillemot, et al. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 2014. doi: 10.1093/biostatistics/kxu001.

T. M. Therneau. *A Package for Survival Analysis in R*, 2023. URL `https://CRAN.R-project.org/package=survival`. R package version 3.5-5.

B. Thompson. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55(4):525–534, 1995. doi: 10.1177/0013164495055004001. URL `https://doi.org/10.1177/0013164495055004001`.

G. Tutz and H. Binder. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 2006. doi: 10.1111/j.1541-0420.2006.00578.x.

E. Vittinghoff and C. E. McCulloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, 2007. doi: 10.1093/aje/kwk052.

J. A. Westerhuis, T. Kourti, and J. F. Macgregor. Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics*, 12, 1998. URL `https://api.semanticscholar.org/CorpusID:122887851`.

D. Witten and R. Tibshirani. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 2010. doi: 10.1177/0962280209105024.

H. Wold. Estimation of principal components and related models by iterative least squares. In P. R. Krishnajah, editor, *Multivariate Analysis*, pages 391–420. Academic Press, New York, 1966.

M. Zhao, Z. Su, S. Zhang, et al. Suppressive role of microrna-148a in cell proliferation and invasion in ovarian cancer through targeting transforming growth factor-beta-induced 2. *Oncology Research*, 24(5):353–360, 2016. ISSN 0965-0407. doi: 10.3727/096504016X14685034103275. URL `https://doi.org/10.3727/096504016X14685034103275`.

D. Zhu, D. Yuan, R. Guo, et al. Overexpression of mir-148a inhibits viability and invasion of ovarian cancer ovcar3 cells by targeting foxo3. *Oncology Letters*, 18(1):402–410, July 2019. ISSN 1792-1074. doi: 10.3892/ol.2019.10321. URL `https://doi.org/10.3892/ol.2019.10321`.