# plasma: Partial LeAst Squares for Multiomics Analysis

Kyoko Yamaguchi[1], Salma Abdelbaky[1], Lianbo Yu[2], Christopher C. Oakes[1], Lynne V. Abruzzo[3], and Kevin R. Coombes[4]

[1]Division of Hematology, Department of Internal Medicine, Ohio State University, Columbus, OH USA
[2]Department of Biomedical Informatics, Ohio State University, Columbus, OH USA
[3]Department of Pathology and Laboratory Medicine, Medical University of South Carolina, Charleston, SC USA
[4]Department of Biostatistics, Data Science, and Epidemiology, School of Public Health, Georgia Cancer Center at Augusta University, Augusta, GA USA

## Abstract

Recent growth in the number and applications of high-throughput "omics" technologies has created a need for better methods to integrate multiomics data. Much progress has been made in developing unsupervised methods, but supervised methods have lagged behind. We present a novel algorithm, `plasma`, to learn models to predict time-to-event outcomes from multiomics data sets. Plasma uses two layers of existing partial least squares algorithms to first select components that covary with the outcome and then construct a joint Cox proportional hazards model. We apply `plasma` to the stomach adenocarcinoma (STAD) data from The Cancer Genome Atlas. We validate the model both by splitting the STAD data into training and test sets and by applying it to the subset of esophageal cancer (ESCA) containing adenocarcinomas. We use the other half of the ESCA data, which contains squamous cell carcinomas dissimilar to STAD, as a negative control. Our model successfully separates both the STAD test set (p $= 2.73 \times 10^{-8}$) and the independent ESCA validation data ($p = 0.025$) into high risk and low risk patients. It does not separate the negative control data set (ESCA squamous cell carcinomas, $p = 0.57$). The performance of the joint multiomics model is superior to that of the individually trained models. It is also superior to the performance of an unsupervised method (Multi Omics Factor Analysis; MOFA) that finds latent factors to be used as putative predictors in a post-hoc survival analysis. Many of the factors that contribute strongly to the `plasma` model can be justified from the biological literature.

### Significance

To fill the unmet need for supervised multiomics methods, we introduce `plasma`, an algorithm based on partial least squares that integrates multiomics features into biologically relevant "components" that can predict patient outcomes.

### Availability and Implementation

The `plasma` R package can be obtained from The Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=plasma. The latest version of the package can always be obtained from R-Forge at https://r-forge.r-project.org/R/?group_id=1746. Source code and data for the analysis presented here can be obtained from GitLab, at https://gitlab.com/krcoombes/plasma.

### Contact

Email: kcoombes@augusta.edu

## Supplementary Information

Supplementary material is available from the journal web site.

# Introduction

Recent years have seen the development of numerous algorithms and computational packages for the analysis of multiomics data sets. As with other applications of machine learning, problems addressed by these algorithms are divided into two broad categories: unsupervised (e.g., clustering or class discovery) or supervised (including class comparison and class prediction) (1). Advances in the area of unsupervised learning have been broader and deeper than advances in supervised learning.

One challenge when trying to integrate multiomics data sets arises from multiple levels of missing data. At a low level, individual features in some data sets may fail to be measured in individual samples. It is usually fairly easy to deal with low-level missing data. In some cases, they can be ignored; if not, appropriate imputation methods may be able to fill in the blanks. At a high level, some patients may have been omitted entirely from one of the omics data sets. A common, highly conservative, response is to work with the subset of patients who have "complete data"; that is, the intersection of samples used in all data sets. This approach may throw away a lot of potentially useful data. Especially for time-to-event ("survival") analyses, where the statistical power depends on the number of events, restricting to the complete data subset may reduce the total sample size to a point where it is no longer possible to learn a useful prognostic model.

As of this writing, we do not know of any supervised multiomics method that can learn to predict outcomes when some samples have only been assayed on a subset of the omics data sets. In this paper, we present a novel algorithm, which we call "*plasma*", to predict time-to-event outcomes in the presence of incomplete data. The core idea of plasma relies on a two-step process, built on ideas that have proved useful for analyzing related problems in single omics data sets.

1. Use partial least squares (PLS) with Cox regression analysis (2) to identify factors that can predict time-to-event outcomes in each individual omics data set.
2. For each pair of omics data sets, use samples common to both data sets to train and extend models of each factor to the union of assayed samples. For this step, we use PLS with linear regression (3).

Our approach is motivated, in part, by the unsupervised method of Multi-Omic Factor Analysis (MOFA) (4). MOFA, like plasma, does not require all omics assays to have been performed on all samples. It is still able to discover class structure across omics data sets. Also like plasma, it starts with a standard method – Latent Factor Analysis – that is known to work well on a single omics data set. It then fits a coherent model that identifies latent factors that are common to, and able to explain the data well in, all the omics data sets under study. Our investigation (unpublished) of the factors found by MOFA suggests that, at least in some cases, it is approximately equivalent to a two-step process analogous to the one we propose in plasma.

To illustrate the plasma method, we use the stomach adenocarcinoma (STAD) and esophageal cancer (ESCA) data sets from The Cancer Genome Atlas (TCGA).

# Methods

Our computational method is implemented and the data are available in version 1.1.3 of the `plasma` package (https://CRAN.R-project.org/package=plasma).

## Data

The results included here are in whole or part based upon data generated by the TCGA Research Network. We downloaded the entire STAD and ESCA Level 3 data sets (5,6) from the FireBrowse web site (http://firebrowse.org/ (7)) on 6 August 2018. We filtered the data sets so that only the most variable, and presumably the most informative, features were retained. To summarize:

1. From TCGA, we obtained 140 columns of clinical, demographic, and laboratory data on 436 patient samples. We removed any columns that always contained the same value, and any columns that were duplicates of other data in the set. We also removed any columns whose values were missing in more than 25% of the patients. We converted categorical variables into sets of binary variables using one-hot-encoding. We then separated the clinical data into three parts:

    1. Outcome (overall survival)
    2. Binary covariates (75 columns)
    3. Continuous covariates (3 columns)
2. Exome sequencing data for 430 patients with STAD were obtained as mutation allele format (MAF) files. We removed any gene that was mutated in fewer than 4% of the samples. The resulting data set contained 1329 mutated genes.
3. Methylation data for 388 STAD patients were obtained as beta values computed by the TCGA from Illumina Methylation 450K arrays. We removed any CpG site for which the standard deviation of the beta values was less than 0.25 or for which the mean was within 0.15 of the boundary values of 0 or 1. The resulting data set contained 2,291 highly variable CpGs.
4. Already normalized sequencing data on 2,588 microRNAs (miRs) was obtained for 382 patients. We removed any miR for which the standard deviation of normalized expression was less than 0.10, which left 1,064 miRs in the final data set.
5. Already normalized sequencing data on 20,531 mRNAs was obtained in 411 patients. We removed any mRNA whose mean normalized expression was less than 5 or whose standard deviation was less than 1.25. The final data set included 1,690 mRNAs.
6. Normalized expression data from reverse phase protein arrays (RPPA) was obtained from antibodies targeting 133 proteins in 350 patients. All data were retained for further analysis.

## Imputation

We imputed missing data for any patient sample assayed in an input data set that yielded partial incomplete data (that is, at least some meaningful data for that particular sample). The underlying issue is that the PLS models for individual omics data sets will not make predictions on a sample if even one data point is missing. As a result, if a sample is missing at least one data point in every omics data set, then it will be impossible to use that sample at all.

For a range of available methods and R packages, see the CRAN Task View on Missing Data (https://CRAN.R-project.org/view=MissingData). We also recommend the R-miss-tastic web site on missing data (https://rmisstastic.netlify.app/). Their simulations suggest that, for purposes of producing predictive models from omics data, the imputation method is not particularly important. Because of the latter finding, we have implemented two simple imputation methods in the `plasma` package:

1. `meanModeImputer` will replace any missing data by the mean value of the observed data if there are more than five distinct values; otherwise, it will replace missing data by the mode. This approach works for both continuous data and for binary or small categorical data.
2. `samplingImputer` replaces missing values by sampling randomly from the empirical data distribution.

For both the STAD and ESCA data, we used the sampling imputer.

## Computational Approach

The `plasma` algorithm is based on Partial Least Squares (PLS), which has been shown to be an effective method for finding components that can predict clinically interesting outcomes (3). The workflow of the plasma algorithm is illustrated in **Figure 1** in the case of three omics data sets. First, for each of the omics data sets, we apply the PLS Cox regression algorithm (`plsRcox` Version 1.7.7 (2)) to the time-to-event outcome data to learn separate predictive models (indicated in red, green, and blue, respectively). Each of the individual omics models consists of three kinds of regression models:

- The `plsRcoxmodel` contains the coefficients of the components learned by PLS Cox regression. The number of components is determined automatically as a function of the logarithm of the number of features in the omics data set. The output of this model is a continuous prediction of "risk" for the time-to-event outcome of interest.
- Next, two separate models are constructed using the prediction of risk on the training data.
  - The `riskModel` is a `coxph` model using continuous predicted risk as a single predictor.
  - The `splitModel` is a `coxph` model using a binary split of the risk (at the median) as the predictor.

Each of these models may be incomplete, since they are not defined for patients who have not been assayed (shown in white in **Figure 1**) using that particular omics technology. In the second step, for each pair of omics data sets, we apply the PLS linear regression algorithm (`pls` Version 2.8.3 (8)) to learn how to predict the Cox regression components from one data set using features from the other data set. This step extends (shown in pastel red, green, and blue, resp.) each of the original models, in different ways, from the intersection of samples assayed on both data sets to their union. Third, we average all of the different extended models (ignoring missing data) to get a single coherent model of components across all omics data sets. Assuming that this process has been applied to learn the model from a training data set, we can evaluate the final Cox regression model on both the training set and a new test set of patient samples.

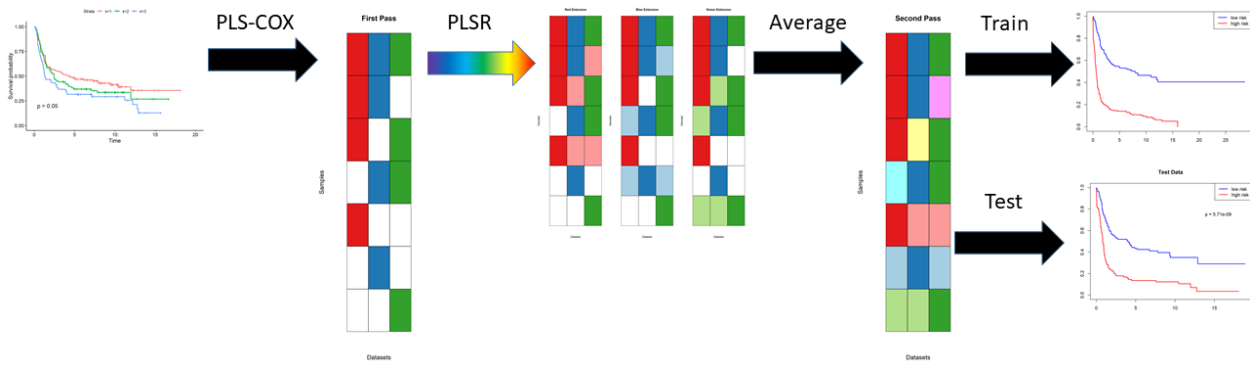


Figure 1: Workflow schematic for plasma algorithm with three omics data sets. See "Computational Approach" in the main text for an explanation.

All computations were performed in R version 4.4.0 (2024-04-24 ucrt) of the R Statistical Software Environment (9). Cox proportional hazards models for survival analysis were fit using version 3.5.8 of the `survival` R package (10).

## Gene Enrichment

In order to interpret the model, we converted feature identifiers into gene names. This task was straightforward for the mutation, mRNASeq, and RPPA data sets. We used annotation files downloaded from the Illumina web site (https://webdata.illumina.com/downloads/productfiles/humanmethylation450/humanmethylation450_15017482_v1-2.csv) to extract associated genes (possibly more than one per locus) from the methylation data set. Gene enrichment (pathway or annotation) analysis was performed by uploading gene lists to ToppGene (https://toppgene.cchmc.org/) (11). Results from ToppGene were limited to 40 entries per ToppGene category.

## Terminology

Because of the layered nature of the plasma algorithm, we intend to use the following terminology to help clarify the later discussions.

1. The input data contains a list of *omics data sets*.
2. Each omics data set contains measurements of multiple *features*.
3. The first step in the algorithm uses PLS Cox regression to find a set of *components*. Each component is a linear combination of features. The components are used as predictors in a Cox proportional hazards model, which predicts the log hazard ratio as a linear combination of components.
4. The second step in the algorithm creates a secondary layer of components. We do not give these components a separate name. They are not an item of particular focus; we view them as a way to extend the first level components to more samples by "re-interpreting" them in other omics data sets.

## Preparing the Data

To be consistent with the `MOFA` R package (4), all of the data sets are arranged so that patient samples are columns and assay features are rows. Each data set includes the same complete set of patients, with columns entirely composed of "NA's" to indicate samples that were not assayed.

## Split Into Training and Test

We randomly assigned patients to separate training and test sets. We used 60% for training and 40% for testing. **Figure 2** presents a graphical overview of the number of samples (`N`) and the number of features (`D`) in each omics component of the training and test sets. Before proceeding, we confirmed that there was no significant difference in survival between the randomly assigned training and test sets (**Supplemental Figure S1**).

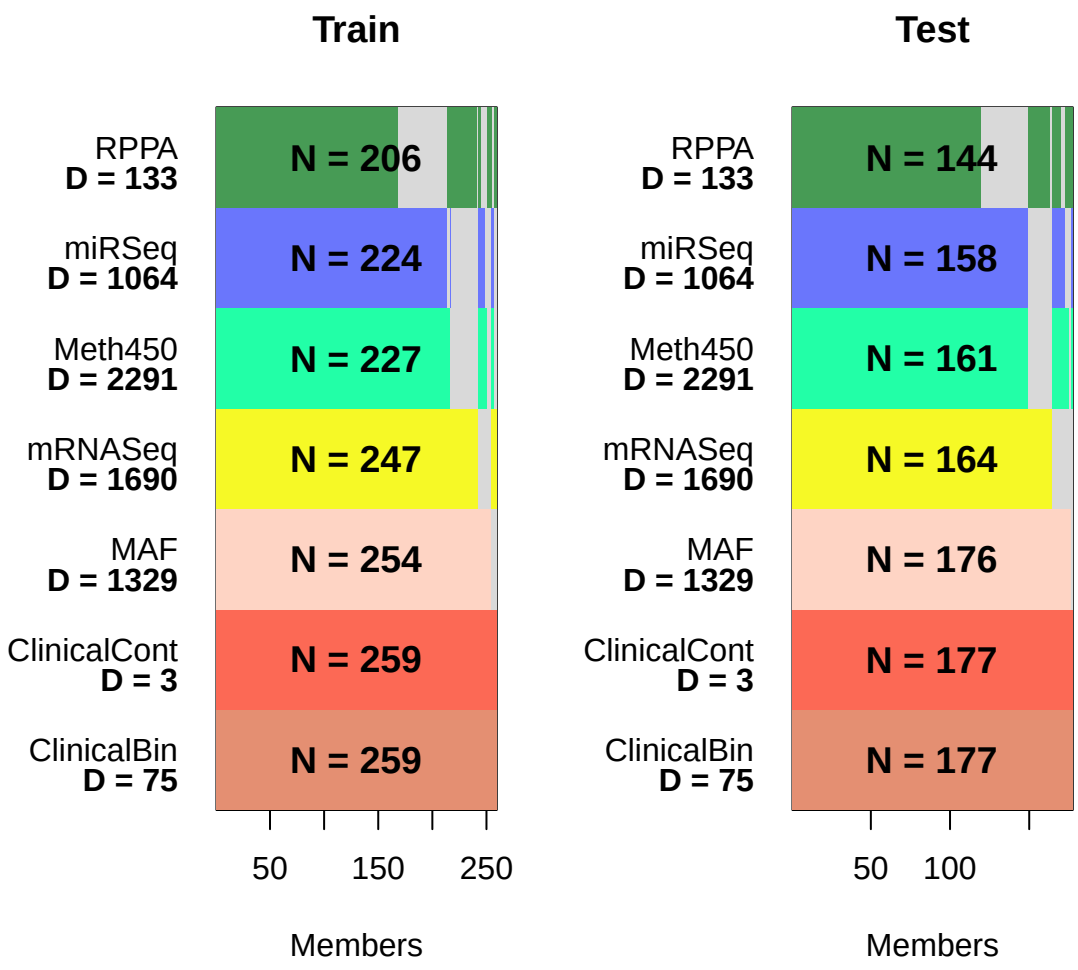


Figure 2: Overview of training and test data. (N is the number of samples in a data set; D is the number of features. RPPA = reverse phase protein arrays; Meth450 = Illumina 450K methylation arrays; MAF = mutation allele format; ClinicalBin = binary clinical features; ClinicalCont = continuous clinical features.)

# Results

## Individual PLS Cox Regression Models

In the first step of the `plasma` algorithm, we fit PLS Cox models on each omics data set separately. On the training set, all seven contributing data sets are able to find a PLS model that can successfully separate high risk from low risk patients in the training data (**Figure 3**).

## Training and Testing a Unified Model of Overall Survival

The second step of the algorithm is to extend the individual omics-based models across other omics data sets. Since this step is performed for all pairs of data sets, in our case there are seven different sets of predictions of each PLS component. These different predictions are averaged. The structure of complete models created is the same as for the separate, individual omics models. **Figure 4** shows the final composite Kaplan-Meier plot using the predicted risk, split at the median value of the training data, on both the training and test data sets. The model yields a statistically significant separation of outcomes between the high and low risk patients in both the training ($p = 3.99 \times 10^{-18}$) and test ($p = 2.73 \times 10^{-8}$) sets.

## Single Omics Predictions on the Test Set

Next, we compared the test results of the joint omics model to the predictions made on the test data from the separate single-omics models. The Kaplan-Meier plots in **Supplemental Figure S2** show that most of the individual models have poor performance on the test data, especially when compared to the results of the combined model learned jointly from all the omics data sets.

## Independent Validation

Because previous unsupervised analyses of the ESCA data suggest that there are two classes of esophageal cancer (one similar to stomach adenocarcinomas and one similar to head-and-neck squamous cell cancers) (6,12), we split the ESCA data into these two types and used them both to evaluate the independent predictive power of the model (**Figure 5**). As we expected, the model performed poorly on the squamous cell cancers ($p = 0.576$) and performed well on the adenocarcinomas $p = 0.025$).

## Interpreting the Model

At this point, our model appears to be a fairly complex black box. We have constructed a matrix of components, based on linear combinations of actual features in different omics data sets. These components are then combined in yet another linear model that predicts the time-to-event outcome through Cox regression. In this section, we want to explore how the individual features from different omics data sets contribute to the models. Our first step in opening the black box is to realize that not all of the components discovered from the individual omics data sets survived into the final composite model. Some components were eliminated (by stepwise feature selection based on the Akaike Information Criterion) because they appeared to be nearly linearly related to components found in other omics data sets. So, we can examine the final composite model more closely.

Table 1: Coefficients of the omics components retained in the final model. (coef = log hazard ratio, exp(coef) = hazard ratio)

| | coef | exp(coef) | se(coef) | z | p |
|---|---|---|---|---|---|
| MAF1 | 1.531 | 4.624 | 0.382 | 4.006 | 0.00006 |
| MAF2 | 1.283 | 3.608 | 0.191 | 6.729 | 0.00000 |
| MAF3 | 6.092 | 442.244 | 1.509 | 4.037 | 0.00005 |
| miRSeq1 | 1.131 | 3.098 | 0.334 | 3.387 | 0.00071 |
| miRSeq3 | 0.263 | 1.301 | 0.104 | 2.532 | 0.01135 |

| | coef | exp(coef) | se(coef) | z | p |
|---|---|---|---|---|---|
| miRSeq4 | -0.985 | 0.373 | 0.449 | -2.193 | 0.02833 |
| ClinicalBin1 | -1.700 | 0.183 | 0.955 | -1.780 | 0.07501 |
| ClinicalBin2 | -6.407 | 0.002 | 3.309 | -1.936 | 0.05285 |
| ClinicalBin3 | -4.426 | 0.012 | 1.961 | -2.258 | 0.02397 |
| mRNASeq1 | -0.180 | 0.835 | 0.128 | -1.405 | 0.16005 |
| mRNASeq2 | 0.949 | 2.582 | 0.292 | 3.248 | 0.00116 |
| mRNASeq4 | 0.779 | 2.178 | 0.436 | 1.788 | 0.07385 |
| Meth4501 | -0.095 | 0.909 | 0.041 | -2.301 | 0.02140 |
| Meth4502 | -1.992 | 0.136 | 0.924 | -2.156 | 0.03105 |

We see that at least one component discovered from four of the five "true" omics data sets survived in the final model; both the continuous clinical data (ClinicalCont) and the protein (RPPA) components failed to make the cut. In addition, three components from the binary clinical data were retained in the final model.

### Final (Composite) Weights

A key point to note is that the core of the `plasma` model consists of a composition of two levels of linear models. One model starts with individual omics data sets and predicts the "*components*" that we have learned across all data sets. The second model uses the components as predictors in a Cox proportional hazards model of time-to-event (i.e., overall survival) data. The composite of these two steps is yet another linear model. In particular, we can compute the matrix that links the features from the omics-level data sets directly to the final survival outcome. Features from the individual omics data sets with the highest weight for that omics data set in the final model are shown in **Figure 6**. (We omitted continuous clinical features from this figure since there are only three features, none of which were significant.)

### Gene Enrichment Using ToppGene

For each component, we determined the list of significant genes (arbitrarily defined as those in the top 5% by absolute value) contributing to that component. We uploaded those gene lists to the ToppGene web site (11) to perform gene enrichment analyses. We performed an additional gene enrichment analysis using the union of all significant genes from all components. Finally, we performed a similar analysis using the genes with the largest (top 5% by standardized weight) contribution to the final composite model. We clustered both the ToppGene annotation categories and the model components based on correlation between the vectors of the negative log10 p-values of the enrichment scores. Clustering determined the order of the rows and columns of the pathway annotations presented in **Figure 7**. A more conventional heatmap display of the pathway annotations is also included as **Supplemental Figure S3**. Corresponding bubble plots for other ToppGene database categories (biological process, cellular location, disease, and drug) are shown in **Supplemental Figures S4-S7**.

## Discussion

One of the most interesting aspects of this study is that the `plasma` algorithm successfully discovers a model on the training set that generalizes to produce statistically significant results on the test set (**Figure 4**). However, it accomplishes this goal even though most of the individual omics models do not generalize well to the data set (**Supplemental Figure S2**). We can think of three possible reasons for this performance.

1. "*Wisdom of the crowd.*" There has long been an idea in the machine learning field that combining ensembles of weak models can give rise to a strong model. Well-established examples of this idea are bagging and boosting (13)
2. "*Out of phase.*'' Each omics data set may overfit the model in a different way. Instead of reinforcing each other, the extent of overfitting may cancel out.

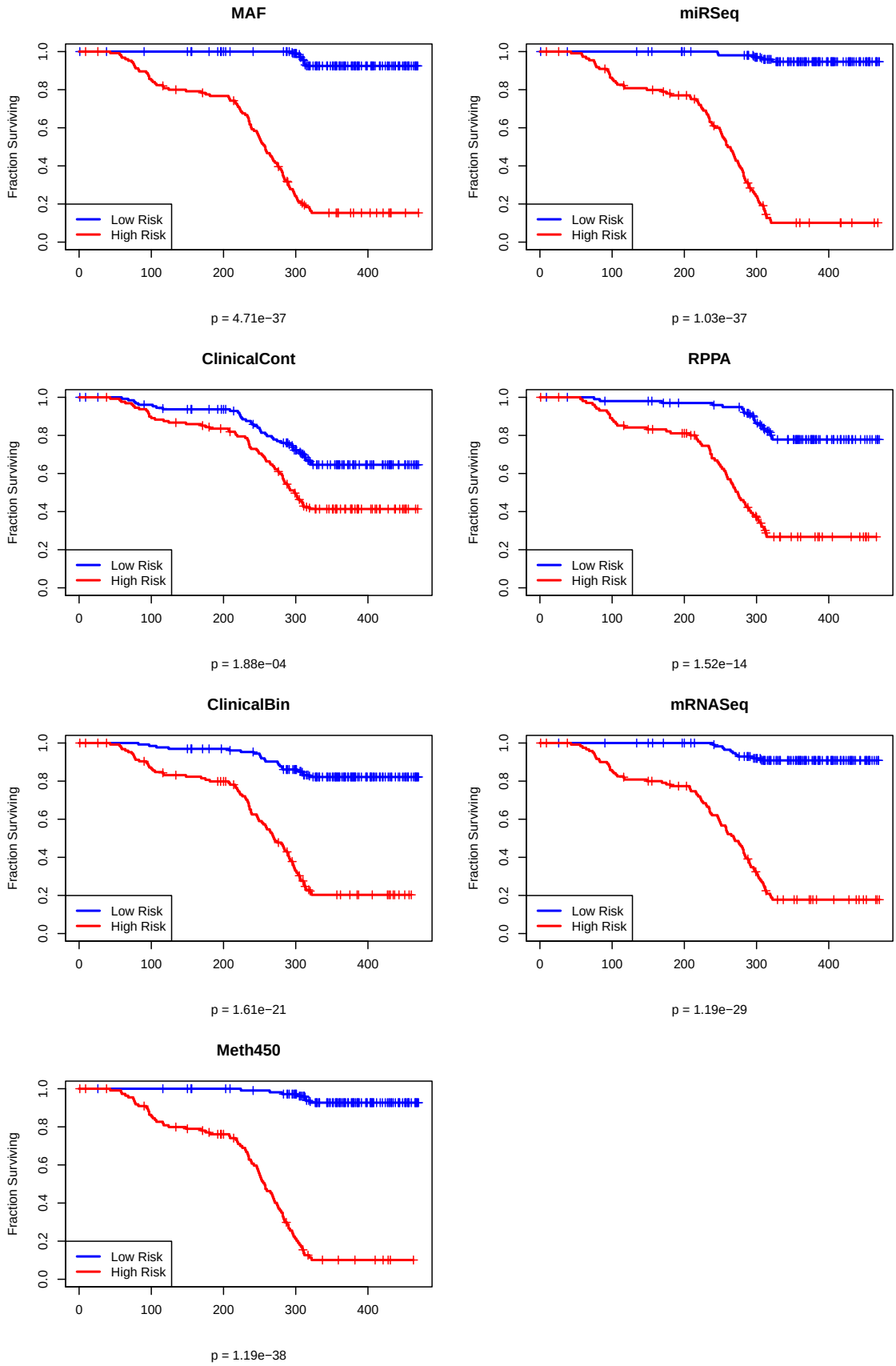


Figure 3: Kaplan-Meier plots of overall survival on the STAD training set from separate PLS Cox omics models.

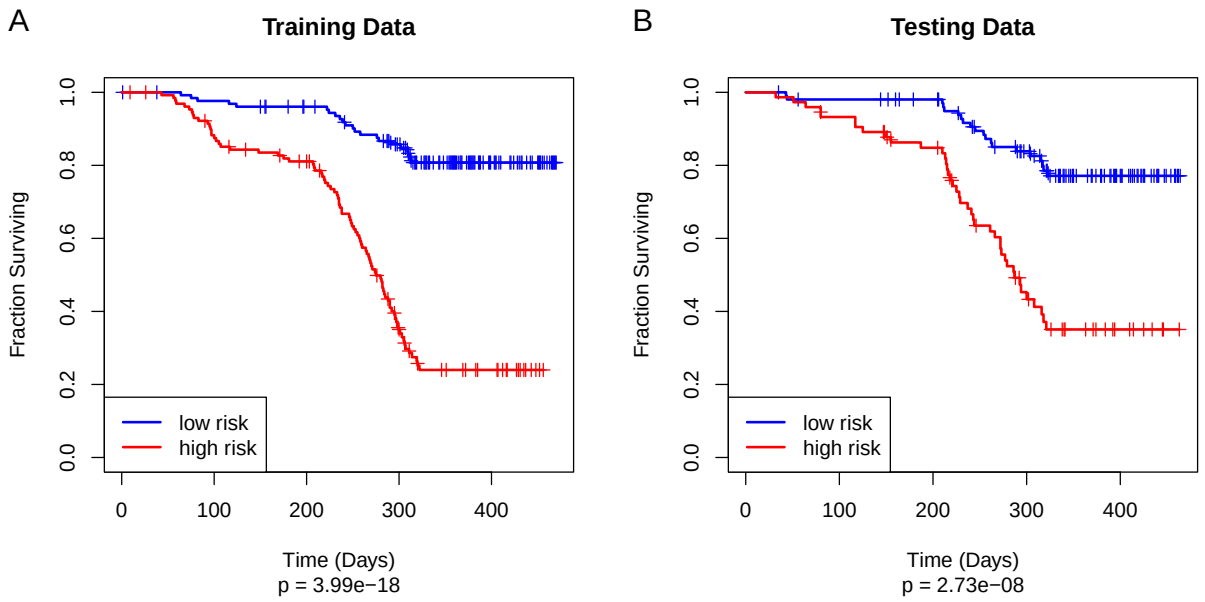


Figure 4: Kaplan-Meier plot of overall survival on (A) the STAD training set and (B) the STAD test set using the unified `plasma` Cox model.

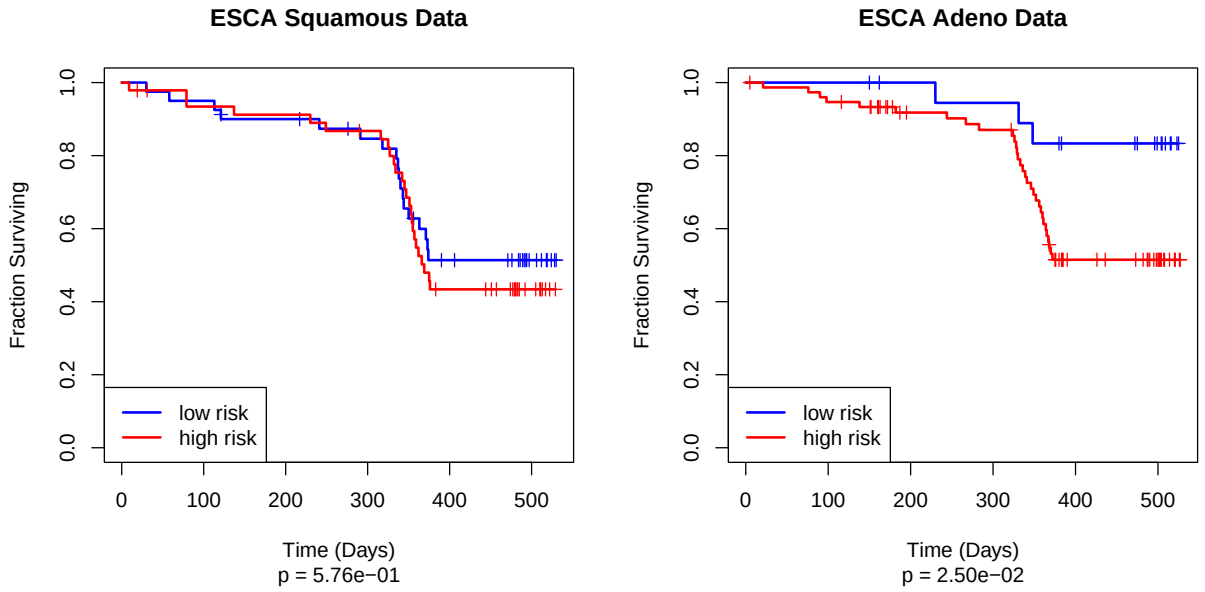


Figure 5: Kaplan-Meier plot of overall survival on (A) the negative control ESCA squamous cell cancers and (B) the ESCA adenocarcinoma validation set using the unified `plasma` Cox model.

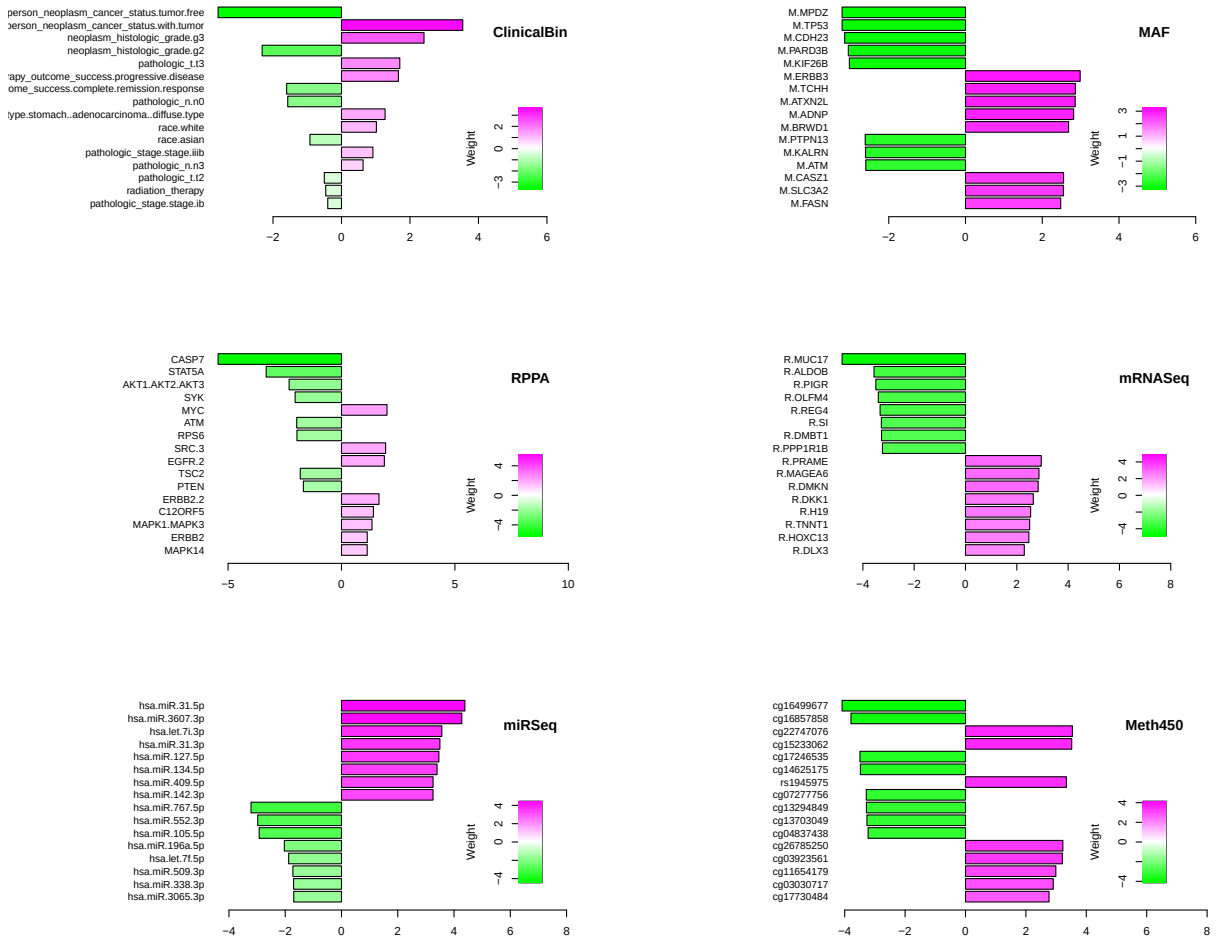


Figure 6: Highly weighted features in the final Cox model. All weights have been standardized within omics data set. Positive weights (magenta) indicate increaded hazard ratio; negative weights (green) indicate decreased hazard ratio. Prefix M = MAF; prefix R = mRNASeq.

Figure 7: Clustering components based on pathway annotation p-values.

3. "*Feature Elimination.*" The combined method successfully identifies useful predictive factors. So, we are still able to fit a generalizable Cox model on the final components.

More research will need to be performed on a variety of data sets to determine whether this phenomenon is more general.

In the original analysis of the STAD dataset by the TCGA Research Network (5), the authors identified four molecularly distinct subtypes. These subtypes, albeit possessing distinct molecular characteristics, were not significantly different between each other in terms of survival outcomes. We also performed an unsupervised analysis using MOFA, and found that the latent factors it uncovered were not related to survival (**Supplemental Table T1-T3, Supplemental Figure S8-S13**) Here, we take a different approach, in that we train a multi-omic model supervised on overall survival, and then dissect the feature weights of omics assays.

First, however, we confirmed that the plasma model for STAD could be validated in an independent cohort. The TCGA-ESCA data were divided into esophageal adenocarcinoma (EAC) and esophageal squamous-cell carcinoma (ESCC). Due to known biological similarities between EAC and STAD (6,12), we hypothesized that the external validation of the plasma-STAD-model with EAC would yield sensible results, and validation with ESCC would serve as a negative control. Indeed, we observed that the Kaplan-Meier plot of OS shows little separation between the low and high-risk patient groups in the validation with ESCC (p=0.576), while there is significant separation (p=0.025) in the validation with EAC (**Figure 5**).

Now, we recognize that a prognostic or predictive model that depends on multiple genome-wide omics technologies is unlikely to move into clinical use any time soon. The costs to perform the assays – both in dollars and in the delays imposed by waiting for all the results – limit its clinical applicability. We expect that it will be necessary to refine the model to use a much smaller set of molecular predictors before it can be clinically useful. Toward that eventual goal, we have performed several analyses to help determine where to focus our attention.

One approach involved performing gene enrichment analysis to better understand the biological role of the complicated factors that plasma found across omics data sets (**Figure 7**). Somewhat surprisingly, the association of the final gene set had weaker associations with any of the biological pathways than did the individual components derived from single omics datasets. However, a heatmap of the same values (**Supplemental Figure S3**) indicates that annotation categories for the final gene list are strongly correlated with those for the union of all the component gene sets. It is possible that the associations are weaker because the final model retains fewer related/redundant genes than the individual PLS models. Future research in other disease contexts will be required to understand this phenomenon.

Nevertheless, the gene enrichment analyses may provide a useful way to focus on important factors for overall survival in gastric cancer. The left and right ends of the bubble plot (and the heatmap) mainly contain pathways that distinguish between the factors. But the central portion of the plots, from the `KEGG_ERBB_SIGNALING_PATHWAY` on the left to the `WP_GLIOBLASTOMA_SIGNAING_PATHWAYS` on the right, primarily consists of signaling pathways that have been identified as important across all the multi-omic factors we found. Moreover, many of these signaling pathways have already been identified in the literature as important in gastric cancer. These include signaling pathways involving ERBB (EGFR) (14), Gastrin (15), Leptin (16), KIT (17), Oncostatin M (18), CKAP4, and the PI3K-AKT-MTOR pathway (19). There is also a strong association with Prolactin signaling, which has previously been highlighted for its role in breast cancer (20) and in colon cancer (21).

One limitation of gene enrichment analysis is that ignores the direction of association between pathways (or genes) and survival. Genes with both positive and negative weights, either for individual components or in the final model, were combined when performing gene enrichment analysis. In order to better understand the contributions of individual gene features, we looked more closely at their weights in the final model of survival (**Figure 6**). We found that some of the binary clinical variables (ClinicalBin) did a good job of separating high and low risk patients in the STAD test set. Furthermore, the weights given to these features in the final composite model make sense. Being tumor free has a large negative weight, indicating decreased risk, while having residual tumor has a positive weight, indicating increased risk. Similarly, having grade 2 or

stage IB tumors or complete response to therapy are markers of good prognosis, while having grade 3 or stage IIIB tumors are markers of poor prognosis.

From the MAF dataset, we found that the most positive weight is from ERBB3, a member of the EGFR family that can heterodimerize with other family members, notably ERBB2 (22). Overexpression of Erbb3 and Erbb1 at the protein and mRNA level in histological specimens of gastric cancer has been significantly associated with tumor relapse and poorly differentiated morphology (23). From the RPPA data set, we observed that ERBB2 possesses a positive weight, and is therefore positively associated with worse outcome. This result, again, matches what is currently known in ERBB2-mutant gastric cancers and other cancers such as breast and lung cancers: overactivity of ERBB2 promotes tumor growth and survival through downstream signaling of the PI3K-AKT and RAS-MAPK pathways. Many of the results from the mRNAseq data set can also be validated from the literature. These include the facts that increased expression of SYK (24), DMBT1 (25), CASP7 (26), OLFM4 (27), and PIGR (28) have been reported to be associated with prolonged survival. High expression of MAGEA6 (29), DKK1 (30), and the MYC-induced long noncoding RNA H19 (31) have been associated with shorter survival.

Some features with large weights are less well-supported by the existing literature. For example, from the MAF data set, we found that the presence of a TP53 mutation was associated with a negative weight, and therefore better OS. This result is supported by a univariate analysis using the STAD TCGA data set which reported that wild type TP53 is associated with poor prognosis in patients with gastric cancer (32), confirming independent findings reported previously (33–35). However, several other studies found no connection between TP53 and survival in stomach cancer (36–39), and at least one study reported that p53-positive tumors were a sign of poor prognosis (40). Earlier studies were characterized by a variety of methods to assess the presence of p53 mutations, including immunohistochemical antibody-based methods and limited direct sequencing of specific mutations. Since our approach is based on whole exome sequencing of the entire gene, as well as the indirect contribution of genes such as TP53 through various components, it is still possible that TP53 has a small, non-independent effect that merely contributes to the effects of other mutations. Similarly, for ATM expression and ATM mutations appearing to be associated with better outcome in our analysis, the literature contains conflicting results, with reports of low ATM expression associated with both poor (41) and good (42) prognosis.

The weights from the miRNA-seq data set have mixed agreement with what is known about the role of miRs in STAD. We found that miR-509 and miR-338 have negative weights and thus lower levels should be associated with worse prognosis, which agrees with the literature (43–46). But we also found that miR-552, miR-105, and miR-196a have negative weights, yet the literature suggests that higher levels of these miRs are associated with tumor severity (47–49). Similarly, downregulation of miR-31 has bee reported to be associated with worse overall survival of gastric cancer patients (50), which, again, is the opposite of our results. We are unable to fully explain the discrepancies between our findings and the literature. The differences may be related to the fact that many earlier studies used different technologies (for example, microarrays rather than sequencing). There may also be differences depending on whether assays are targeting primary miRNAs, precursor miRNAs, or 3' or 5' mature miRNAs.

In summary, we have identified a multi-omic machine learning model that learns composites of the variables, i.e. "components", supervised with respect to patient outcome. We found 23 components, 14 of which were retained in the final PLASMA model. We validated the model with 1/3 of the STAD dataset and the adenocarcinoma subset of the ESCA dataset.

To draw comparisons with existing multi-omic modeling procedures, we trained a MOFA model and found 10 orthogonal factors, none of which were significantly associated with survival. Because the `plasma` components are learned in a way that maximizes the covariance in the predictors and the response, these components will be automatically associated with the outcome. Another major difference between MOFA and PLASMA is that while MOFA learns near-orthogonal latent factors, PLASMA may learn highly correlated components. As seen in the gene enrichment analyses with ToppGene, we observe that the correlation of these components serve to highlight signaling pathways highly relevant to gastric cancers and related solid tumors.

The dissection of the weight matrix matrix yields a list of variables from different omics datasets with the greatest contributions to patient outcome. Again, we would like to recognize that PLASMA is not to be used

directly as a multi-omic clinical model, and that implementation of a real clinical assay will require further investigation. However, seeing as PLASMA has found highly weighted features with support from previously published findings, PLASMA can be used as a discovery tool to highlight potentially novel features associated with outcomes of interest. For solid tumors like STAD, in which anti-ErBb2 treatments have yielded mixed results and surgery remains the standard of care, this process may be advantageous in providing hints at novel associations and pathway usages.


## Funding

This work was supported by a collaboration grant from the Department of Biomedical Informatics at The Ohio State University and by startup funds from the Ohio State University Comprehensive Cancer Center and from the Georgia Cancer Center at Augusta University.

*Conflict of Interest:* None declared.


# References


1. Simon RM, Dobbin K. Experimental design of DNA microarray experiments. Biotechniques. 2003;Suppl:16–21.
2. Bertrand F, Maumy-Bertrand M. Fitting and Cross-Validating Cox Models to Censored Big Data With Missing Values Using Extensions of Partial Least Squares Regression Models. Front Big Data. 2021;4:684794.
3. Bastien P, Bertrand F, Meyer N, Maumy-Bertrand M. Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. Bioinformatics. 2015;31:397–404.
4. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14:e8124.
5. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513:202–9.
6. Kim J, Bowlby R, Mungall AJ, Robertson AG, Odze RD, Cherniack AD, et al. Integrated genomic characterization of oesophageal carcinoma. Nature. Nature Publishing Group; 2017;541:169–75.
7. Deng M, Bragelmann J, Kryukov I, Saraiva-Agostinho N, Perner S. FirebrowseR: An R client to the Broad Institute's Firehose Pipeline. Database (Oxford). 2017;2017.
8. Mishra P, Liland KH. Swiss knife partial least squares (SKPLS): One tool for modelling single block, multiblock, multiway, multiway multiblock including multi-responses and meta information under the ROSA framework. Anal Chim Acta. 2022;1206:339786.
9. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
10. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Dietz K, Gail M, Krickeberg K, Samet J, Tsiatis A, editors. New York, NY: Springer; 2000.
11. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37:W305–11.
12. Abrams ZB, Zucker M, Wang M, Asiaee Taheri A, Abruzzo LV, Coombes KR. Thirty biologically interpretable clusters of transcription factors distinguish cancer type. BMC Genomics. 2018;19:738.
13. Sutton CD. Classification and Regression Trees, Bagging, and Boosting. Handbook of Statistics. Elsevier; 2005;24:303–29.
14. Arienti C, Pignatta S, Tesei A. Epidermal Growth Factor Receptor Family and its Role in Gastric Cancer. Frontiers in Oncology. 2019;9:1308.
15. Zhao Y, Wang Q, Zeng Y, Xie Y, Zhou J. Gastrin/CCK-B Receptor Signaling Promotes Cell Invasion and Metastasis by Upregulating MMP-2 and VEGF Expression in Gastric Cancer. Journal of Cancer. 2022;13:134–45.

16. Lin T-C, Hsiao M. Leptin and Cancer: Updated Functional Roles in Carcinogenesis, Therapeutic Niches, and Developments. International Journal of Molecular Sciences. 2021;22:2870.

17. Pan S, Tan J, Deng Y, Wan B-H, Zhang X-Y, Guan B-G. KIT performed as a driver gene candidate affecting the survival status of patients with stomach adenocarcinoma. Oncotarget. 2017;8:70183–91.

18. Su L, Wu X, Li J, Yu Z, Liu B. Interaction of oncostatin M and its receptor OSMR promotes gastric cancer progression via STAT3/FAK/Src signaling. Annals of Oncology. Elsevier; 2018;29:viii678.

19. Matsuoka T, Yashiro M. The Role of PI3K/Akt/mTOR Signaling in Gastric Carcinoma. Cancers. 2014;6:1441–63.

20. Standing D, Dandawate P, Anant S. Prolactin receptor signaling: A novel target for cancer treatment - Exploring anti-PRLR signaling strategies. Frontiers in Endocrinology. 2022;13:1112987.

21. Neradugomma NK, Subramaniam D, Tawfik OW, Goffin V, Kumar TR, Jensen RA, et al. Prolactin signaling enhances colon cancer stemness by modulating Notch signaling in a Jak2-STAT3/ERK manner. Carcinogenesis. 2014;35:795–806.

22. Jaiswal BS, Kljavin NM, Stawiski EW, Chan E, Parikh C, Durinck S, et al. Oncogenic ERBB3 Mutations in Human Cancers. Cancer Cell. Elsevier; 2013;23:603–17.

23. Moghbeli M, Makhdoumi Y, Soltani Delgosha M, Aarabi A, Dadkhah E, Memar B, et al. ErbB1 and ErbB3 co-over expression as a prognostic factor in gastric cancer. Biological Research. 2019;52:2.

24. Nakashima H, Natsugoe S, Ishigami S, Okumura H, Matsumoto M, Hokita S, et al. Clinical significance of nuclear expression of spleen tyrosine kinase (Syk) in gastric cancer. Cancer Letters. 2006;236:89–94.

25. Sousa JF, Ham A-JL, Whitwell C, Nam KT, Lee H-J, Yang H-K, et al. Proteomic profiling of paraffin-embedded samples identifies metaplasia-specific and early-stage gastric cancer biomarkers. The American Journal of Pathology. 2012;181:1560–72.

26. Liang R, Chen W, Chen X-Y, Fan H-N, Zhang J, Zhu J-S. Dihydroartemisinin inhibits the tumorigenesis and invasion of gastric cancer by regulating STAT1/KDR/MMP9 and P53/BCL2L1/CASP3/7 pathways. Pathology, Research and Practice. 2021;218:153318.

27. Luo Z, Zhang Q, Zhao Z, Li B, Chen J, Wang Y. OLFM4 is associated with lymph node metastasis and poor prognosis in patients with gastric cancer. Journal of Cancer Research and Clinical Oncology. 2011;137:1713–20.

28. Fristedt R, Gaber A, Hedner C, Nodin B, Uhlén M, Eberhard J, et al. Expression and prognostic significance of the polymeric immunoglobulin receptor in esophageal and gastric adenocarcinoma. Journal of Translational Medicine. 2014;12:83.

29. Endo M, Kanda M, Sawaki K, Shimizu D, Tanaka C, Kobayashi D, et al. Tissue Expression of Melanoma-associated Antigen A6 and Clinical Characteristics of Gastric Cancer. Anticancer Research. 2019;39:5903–10.

30. Gao C, Xie R, Ren C, Yang X. Dickkopf-1 expression is a novel prognostic marker for gastric cancer. Journal of Biomedicine & Biotechnology. 2012;2012:804592.

31. Zhang E-B, Han L, Yin D-D, Kong R, De W, Chen J. C-Myc-induced, long, noncoding H19 affects cell proliferation and predicts a poor prognosis in patients with gastric cancer. Medical Oncology (Northwood, London, England). 2014;31:914.

32. Deng W, Hao Q, Vadgama J, Wu Y. Wild-Type TP53 Predicts Poor Prognosis in Patients with Gastric Cancer. Journal of Cancer Science and Clinical Therapeutics. 2021;5:134–53.

33. Victorzon M, Nordling S, Haglund C, Lundin J, Roberts PJ. Expression of P53 protein as a prognostic factor in patients with gastric cancer. European Journal of Cancer (Oxford, England: 1990). 1996;32A:215–20.

34. Bazas VM, Lukyanova NY, Lisovenko GS, Rozumiy DO, Potebnya GP. Postoperative autovaccinotherapy for patients with gastric cancer and expression of some proteins in tumor tissue. Experimental Oncology. 2009;31:182–4.

35. Gonçalves AR, Carneiro AJV, Martins I, de Faria PAS, Ferreira MA, de Mello ELR, et al. Prognostic significance of P53 protein expression in early gastric cancer. Pathology oncology research: POR. 2011;17:349–55.

36. Ismail HM, Moneer M, El-Baradie M, Khorshid O, Touny A. Clinicopathologic and prognostic significance of overexpression of her-2/neu and P53 oncoproteins in gastric carcinoma using tissue microarray. Journal of the Egyptian National Cancer Institute. 2007;19:147–57.

37. Sezer C, Yildirim M, Yildiz M, Sezgin A, Donem Dilli U, Goktas S, et al. Prognostic significance of biological apoptosis factors in gastric cancer. Journal of BUON: official journal of the Balkan Union of Oncology. 2013;18:138–46.

38. Liu X, Yu H, Cai H, Wang Y. Expression of CD24, P21, P53, and c-myc in alpha-fetoprotein-producing gastric cancer: Correlation with clinicopathologic characteristics and survival. Journal of Surgical Oncology. 2014;109:859–64.

39. Ahadi M, Moradi A, Musavinejad L, Movafagh A, Moradi A. The Expression of P53, CD44, Ki-67, and HER-2/neu Markers in Gastric Cancer and Its Association with Histopathological Indicators: A Retrospective Study. Asian Pacific journal of cancer prevention: APJCP. 2020;21:1607–14.

40. Lazăr D, Tăban S, Sporea I, Dema A, Cornianu M, Lazăr E, et al. The immunohistochemical expression of the P53-protein in gastric carcinomas. Correlation with clinicopathological factors and survival of patients. Romanian Journal of Morphology and Embryology = Revue Roumaine De Morphologie Et Embryologie. 2010;51:249–57.

41. Abdel-Fatah T, Arora A, Gorguc I, Abbotts R, Beebeejaun S, Storr S, et al. Are DNA repair factors promising biomarkers for personalized therapy in gastric cancer? Antioxidants & Redox Signaling. 2013;18:2392–8.

42. Kim JW, Im S-A, Kim MA, Cho HJ, Lee DW, Lee K-H, et al. Ataxia-telangiectasia-mutated protein expression with microsatellite instability in gastric cancer as prognostic marker. International Journal of Cancer. 2014;134:72–80.

43. Zhang J, Zhu Z, Sheng J, Yu Z, Yao B, Huang K, et al. miR-509-3-5P inhibits the invasion and lymphatic metastasis by targeting PODXL and serves as a novel prognostic indicator for gastric cancer. Oncotarget. 2017;8:34867–83.

44. Sun J, Li J, Zhang W, Zhang J, Sun S, Li G, et al. MicroRNA-509-3p Inhibits Cancer Cell Proliferation and Migration via Upregulation of XIAP in Gastric Cancer Cells. Oncology Research. 2017;25:455–61.

45. Sun F, Yu M, Yu J, Liu Z, Zhou X, Liu Y, et al. miR-338-3p functions as a tumor suppressor in gastric cancer by targeting PTP1B. Cell Death & Disease. 2018;9:522.

46. Liu S, Suo J, Wang C, Sun X, Wang D, He L, et al. Downregulation of tissue miR-338-3p predicts unfavorable prognosis of gastric cancer. Cancer Biomarkers: Section A of Disease Markers. 2017;21:117–22.

47. Feng X, Zhu M, Liao B, Tian T, Li M, Wang Z, et al. Upregulation of miR-552 Predicts Unfavorable Prognosis of Gastric Cancer and Promotes the Proliferation, Migration, and Invasion of Gastric Cancer Cells. Oncology Research and Treatment. 2020;43:103–11.

48. Lin B, Liu C, Shi E, Jin Q, Zhao W, Wang J, et al. MiR-105-3p acts as an oncogene to promote the proliferation and metastasis of breast cancer cells by targeting GOLIM4. BMC cancer. 2021;21:275.

49. Zheng G, Xiong Y, Xu W, Wang Y, Chen F, Wang Z, et al. A two-microRNA signature as a potential biomarker for early gastric cancer. Oncology Letters. 2014;7:679–84.

50. Wang H, Zhang X, Liu Y, Ni Z, Lin Y, Duan Z, et al. Downregulated miR-31 level associates with poor prognosis of gastric cancer and its restoration suppresses tumor cell malignant phenotypes by inhibiting E2F2. Oncotarget. 2016;7:36577–89.