



Research Article

Expository techniques and integrative bioinformatics for survival modeling: Screening potential biomarkers in hepatocellular carcinoma

Sumaila ABUBAKARI^{1,2,*} , Filiz KARAMAN² , Radiatu ABDUL-SOMED³

¹Department of Electrical/Electronic Engineering, School of Engineering, University of Engineering and Agricultural Sciences, Bunso, E51774, Ghana

²Department of Statistics, Faculty of Arts and Science, Yildiz Technical University, Davutpaşa Campus, İstanbul, 34220, Türkiye

³Department of Pharmaceutical Technology, University of Technology and Applied Sciences, Navrongo, UK-0215-5322, Ghana

ARTICLE INFO

Article history

Received: 14 August 2024

Revised: 13 January 2025

Accepted: 14 May 2025

Keywords:

Bioinformatics; Biomarkers;
Liver Cancer; Survival Modeling;
Hepatoblastoma

ABSTRACT

Hepatocellular carcinoma (HCC) is the most common malignant form of primary liver cancer and one of the leading causes of liver-related deaths. Risk estimation of HCC depends on a matrix of several risk factors, from clinicopathological to gene expression features. From an etiological perspective, the survival prospects of HCC vary from patient to patient. Therefore, a comprehensive study that captures complex relationships among risk factors is one of the ways to quantify prognosis and/or diagnosis, thereby harnessing personalized medicine. We integrated survival analysis with bioinformatics to identify diagnostic and prognostic biomarkers in HCC. Using RNA-seq data from TCGA-LIHC, we screened differentially expressed genes (DEGs) and applied Cox regression, machine learning (LASSO-penalized modeling), and pathway enrichment analysis to prioritize genes linked to overall survival (OS) and tumor progression. Among 3,623 DEGs (685 downregulated, 2,938 upregulated), we identified three core prognostic genes (AP1M1, CASC9, and CA9) significantly associated with OS (logrank $p < 0.001$, hazard ratio > 2.0). Diagnostic biomarkers (CLEC4M, SLC5A1, SLC22A1, SFRP5, CNDP1, TERT, and ZIC2) were enriched in metabolic signaling pathways. Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis revealed five significantly altered pathways, with the strongest enrichment in 'hsa04976: Bile Secretion', underscoring disrupted metabolic homeostasis as a hallmark of HCC aggressiveness. Our study reveals novel biomarkers with both diagnostic and prognostic potential, offering a framework for targeted HCC management. These findings highlight the potential of integrative omics-survival modeling to advance precision oncology.

Cite this article as: Abubakari S, Karaman F, Abdul-Somed R. Expository techniques and integrative bioinformatics for survival modeling: Screening potential biomarkers in hepatocellular carcinoma. Sigma J Eng Nat Sci 2026;44(4):2714–2734.

*Corresponding author.

*E-mail address: sumaila.abubakari@std.yildiz.edu.tr

This paper was recommended for publication in revised form by
Editor-in-Chief Ahmet Selim Dalkilic



INTRODUCTION

Survival analysis serves as a special modeling technique used to handle time-to-event data, primarily due to the presence of ‘censored’ observations. When we refer to this technique as a special one, we are simply referring to its ability to deal with data where the event times are known for some subjects (individuals) while unknown for others, yet the unknown event times cannot be treated strictly as missing values. Further, the distribution of event times is normally skewed—fewer later events than earlier ones. In context, we use the expression *time-to-event* data to refer to observed data on the time taken for the occurrence of the said event (e.g., cancer relapse), commonly collected along with clinical or omics risk features. Though with varying terminologies, time-to-event data occur in different fields of research, namely; precision medicine [1], graduate education [2], loan delinquency [3], and engineering [4].

The genetic profiling of a patient tends to provide significant markers towards disease structure and development, in what is popularly known as *personalized medicine*. In cancer studies, oftentimes, Progression-free survival (PFS) and overall survival (OS) are some of the endpoints commonly used for modeling the survival prospects of patients in a wide range of diseases. By PFS, we mean the duration from when treatment is initiated to the time when the disease gets worse, while OS refers to the survival time from when an individual is diagnosed with a condition or when treatment is initiated to the actual event time. This is usually regarded as a measure of disease control rate (DCR)—thus, partial remission, complete remission, and stable disease rate. Note that the *event time* is defined as the duration for which a condition (e.g., cancer) is observed to have occurred. Thus far, time-to-event data intrinsically depends on whether the event has occurred or not. When it is not observed in the individual within the study time, we say the subject has been “censored”. Censoring is often

grouped into left-, right-, and interval. The most common one is right censoring. The differences in these groupings lie in the range of the time the event is observed. For instance, in right-censoring, we observe the lower limit; the upper limit in left-censoring, and both in the interval-censoring. A simple illustration of right-censoring and actual survival time distribution of different cancer types are shown in Fig. 1a and Fig. 1b respectively. For a comprehensive discussion on various forms of censoring, interested readers are referred to Leung et al. [5] and Klein and Moeschberger [6]. In this paper, we assume survival data with a right-censored paradigm. Further, the status variable is binarized so that a ‘1’ means the event has occurred and ‘0’ otherwise. A typical survival data with some clinical risk factors is shown in Table 1 for a cohort of Lung cancer patients [7]. Observe that the status variable is binary, where ‘1’ is a code for censored individuals (i.e., 63/228) and ‘2’ is a code for ‘dead’ individuals (i.e., 165/228). Because the distribution of the status variable is skewed, a basic logistic regression where the status variable is the response variable would lead to reliability issues in estimating the regression coefficients [8]. Further, using the event time as the response variable in simple regression affects the predictive strength of the model [9].

Following the rapid progression of technology in DNA microarrays, survival prospects of cancer patients and other forms of diseases have efficiently been improved due to such technologies presenting better pathways to evaluate proteome-wide or gene expression levels (microarray data). Stated differently, *omics* data (i.e., genomics, proteomics, epigenomics, etc.) present us with pieces of information that could serve as important biomarkers for assessing patients’ prognosis, or serve as prior data for future research direction [10]. However, such microarrays are complex and exceedingly larger than sample sizes (i.e., $p \gg n$), invariably due to their high dimensionality. Also, this stream of

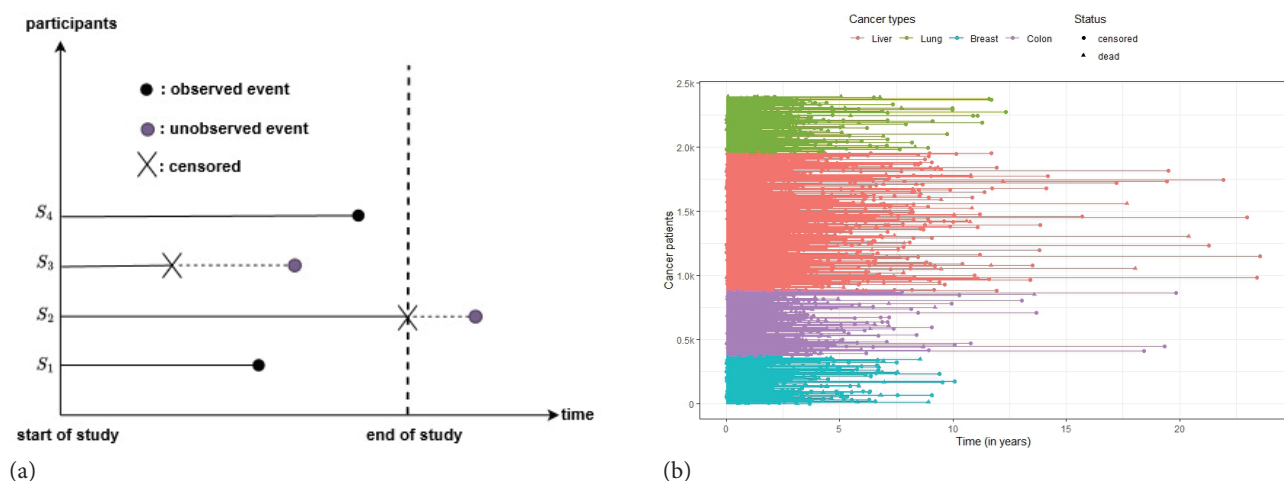


Figure 1. Simple illustration of right censoring (a) Right-censored survival time distribution (a) for selected TCGA cancer datasets (Liver, Lung, Breast, Colon) Overall survival time distribution.

Table 1. An illustration of components of survival data: survival time, censoring, and binarization of the status variable. The data is drawn from 228 patients from the North Central Cancer Treatment Group (NCCTG) [7] with advanced lung cancer. Here, survival time is measured in days, ‘1’ = ‘censored’, ‘2’ = ‘dead’

Inst.	Time	Status	Age	Sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
3	306	2	74	1	1	90	100	1175	NA
3	455	2	68	1	0	90	90	1225	15
3	1010	1	56	1	0	90	90	NA	15
32	105	1	75	2	2	60	70	1025	5
6	174	1	66	1	1	90	100	1075	1
22	177	1	58	2	1	80	90	1060	0

data is collected with potentially massive missing values (NAs), inter-dataset collinearity, and noise-laced covariates [11]. Due to the high-dimensional space of the expression features, analytical techniques for these kinds of data go beyond classical survival modeling. To this end, recent advances in high-throughput genomics have transformed survival analysis by integrating multi-omics data (e.g., transcriptomics, proteomics) to identify prognostic biomarkers. For example, single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics now enable granular profiling of tumor heterogeneity, revealing novel DEGs linked to HCC aggressiveness [12–14]. However, these technologies generate high-dimensional datasets ($p \gg n$) laced with noise, collinearity, and missing values. Traditional Cox models struggle with such dimensionality, prompting the adoption of machine learning (ML) and penalized regression (e.g., LASSO, ridge, or elastic net) for feature selection and predictive accuracy. Recent benchmarks highlight that hybrid models combining clinical and omics data outperform purely genomic or clinical models. For instance, integrating radiomics with transcriptomics has significantly improved HCC survival prediction in AUC metrics [15]. In benchmarking several methods for high-dimensional ‘omics’ data, Vinga [16] provided a comprehensive review of penalized regression approaches focused on survival modeling. Bøvelstad et al. [17] have also assessed the comparative predictive power of several ML models to show improvement in model prediction. In their study, the authors considered using ‘tuning’ methods and demonstrated performance by using both clinical and gene expression data. In a similar study, Bommert et al. [18] also benchmarked various penalized methods for feature selection in high-dimensional cases and suggested using the extended L_2 -regularizer [19] for stability in feature selection, as well as for accuracy. As times go, more researchers continue to provide better insights into various modeling techniques, from parameter tuning to comparative analysis of classical and modern-day survival models, especially for health sciences [20–22].

The novelty of this study lies in its integrative approach, which combines survival analysis with advanced bioinformatics to identify and validate significantly differentially

expressed genes (DEGs) in hepatocellular carcinoma (HCC). We do not focus solely on gene expression patterns but also on clinical survival outcomes, allowing our model to uniquely correlate DEGs with these outcomes. This approach pinpoints biomarkers with direct prognostic and diagnostic relevance. By bridging computational biology with patient-centric data, we not only uncover molecular drivers of HCC aggressiveness but also deliver actionable targets for improving therapeutic strategies—a critical advance in a field where early detection and treatment efficacy remain major challenges.

The remainder of this paper is organized as follows: In section 2, we present the structure of time-to-event analysis of “omics” datasets and their characteristics. In section 3, we both cover classical and modern survival modeling techniques for fitting low- and high-dimensional data. In section 4, we overview methods for assessing and validating prediction models. Section 5 centrally focuses on integrative analyses that use survival techniques and bioinformatics modeling tools to screen potential predictive biomarkers, including the methods, results, and discussion. We conclude our study with concluding remarks in section 6.

Data classification

Assessing the survival prospect of patients is one of the foci in what is usually termed *personalized medicine*. This assessment is done across a collection of different risk factors, clinicopathological and/or omics-related, to minimize bad prognosis. Thus, the endpoint goal is to target decision processes that screen insightful risk factors toward improved treatment targets.

Survival Data

Survival data typically comprises at least two random variables—event-status and survival time. The survival time is the time from the onset of a study of a condition of interest (e.g., cancer relapse) to the time of its actual occurrence or to the time of closure of the study. On the other hand, the event status is a Bernoulli variable indicating whether the event has occurred or not (Fig. 1a).

Let δ be a Bernoulli variable so that the observed survival time $T = \min\{T^* \leq C\}$, given that T^* is the recorded

event time while C is the censoring time. Therefore, $\delta = \mathbb{1}\{T^* \leq C\}$, thus;

$$\delta = \begin{cases} 1 & \text{if } T^* = X \\ 0 & \text{if } T^* = C, \end{cases}$$

where $\mathbb{1}\{\cdot\}$ is a binary indicator function of status. The probability of an individual surviving beyond time t is characterized by the survivor function, $S(t) = \mathbb{P}(T^* > t) = 1 - F(t)$ where $F(t)$ is the cumulative distribution function (CDF).

The hazard function $h(t)$ is another estimator that gives insight into the survival prospects of subjects. It is defined as the instantaneous risk of death at time t given that, the individual up till now has survived time t , mathematically expressed as;

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T^* < t + \Delta t | T^* > t)}{\Delta t}, \quad h(t) \geq 0 \quad \forall t.$$

Clinico-Demographic Data

While the overall survival time (OS) may be a sufficient estimate, comprehensive analyses of other associated risk factors can provide better insights into the survival prospects—for example the influence of clinicopathological on survival modeling. Like in Okorie et al. [23], several cancer studies depend on demographic factors such as gender, race, vital status, ethnicity, cause of relapse, etc., to study their impact on prognosis. In personalized medicine, these factors are usually controlled to measure their influence on health outcomes. For instance, clinical risk factors such as hypercholesterolemia, myocardial infarction, left ventricular ejection, etc., are considered to influence patient survival in cardiovascular disease (CVD) [24]. The clinicopathological factors to consider largely depend on what cancer is being considered to model the target outcome, as illustrated in Bączyk and Kozłowska [25]. Here, the authors

Table 2. Example of demographic and clinical factors [24]

Risk factors	Data type
Age	continuous
Sex	binary
Marital status	categorical
Education	categorical
Rheumatoid arthritis duration	continuous
VAS Pain (0 - 10cm)	continuous
Morning stiffness	continuous
Grip strength measurement (right)	continuous
Grip strength measurement (left)	continuous
Ritchie Articular Index	continuous

discuss the influence of clinical and demographic factors in treating rheumatoid arthritis (RA) by considering the factors shown in Table 2.

Omics Data

Owing to the fast advancement of bioinformatic tools, researchers have rushed to employ the knowledge of biology, computer science, medicine, and mathematics to better characterize patients' prognosis. Modern microarray technology invariably produces tonnes of complex data, the so-called 'omics' data—i.e., genomics, transcriptomics, proteomics, metabolomics, phenomics, and metagenomics. Each of these components shown in Fig. 2 serves a unique purpose towards human disease improvement. This is because of the structure of molecular information they process or possess. For example, in multi-omics analysis, genomics carries data of an individual's genome like genetic variation; proteomics carries data on the individual's set of proteins; transcriptomics for the individual's

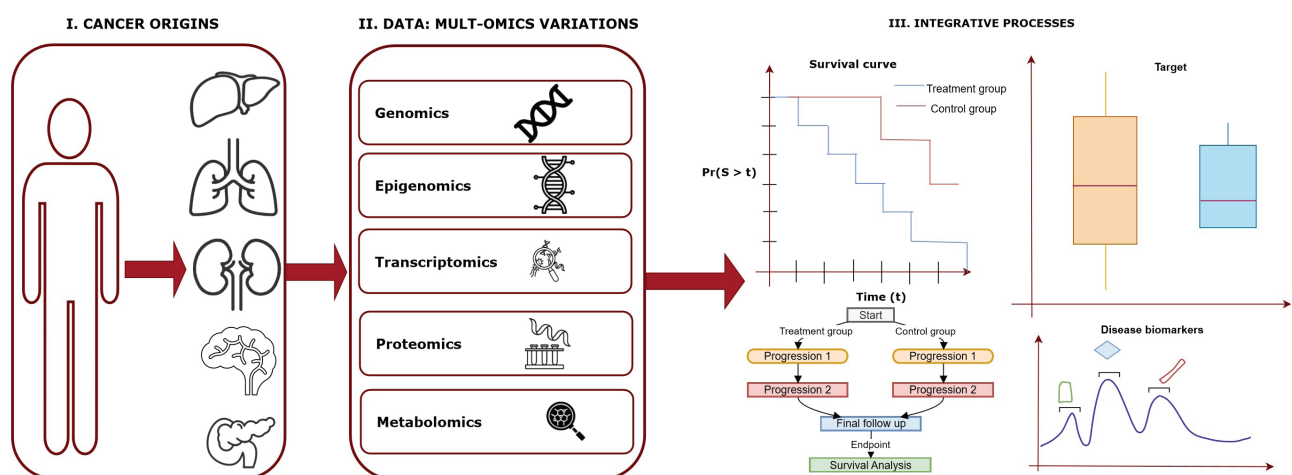


Figure 2. Integrative concept of multi-omics signatures for comprehensive bioinformatics modeling toward cancer detection, biomarker screening, and treatment assessment.

RNA transcripts generated by a genome; metabolomics for metabolites from metabolism. To understand the molecular mechanisms driving disease progression, integrative analysis across multi-omics platforms is pivotal for novel biomarker discovery [26, 27].

Omics data generated from genomic-powered tools traditionally comprises DNA molecules as carriers of genetic traits such as single nucleotide polymorphisms (SNP). As stated differently, a genome can be viewed as a blueprint that carries an individual's cellular construction. DNA molecules are polymers constituting simpler chemical units known as *nucleotides*. There are four basic forms of nucleotides, namely; adenine (A), thymine (T), cytosine (C), and guanine (G). Depending on how many minor alleles (AA, AG, GC, or TT) each individual carries at a particular location, each nucleotide's characteristic can be coded as 0, 1, or 2. Summarily, a genomics dataset consists of DNA molecule sequences found in a single cell or a population of cells.

Omics data related to transcriptomes consists of RNA molecules (e.g., *mRNA*). Transcriptomics comes in different types or forms such as ribosomal RNA (*rRNA*), messenger RNA (*mRNA*), transfer RNA (*tRNA*), and non-coded RNA (*ncRNA*). The *mRNAs* are the most important because they act as intermediary messengers between genomes and proteomes. For this reason, transcriptomics generally focuses on *mRNA* and is traditionally used to evaluate the sequences of expression levels of DNA. At a distributional level, microarray expression levels can fairly be represented by a normal distribution following a $\log(a + 1)$ transformation, often with base 2. The cell count of *mRNA* molecules varies depending on the needs of proteins, and a transcriptomics dataset contains a list of every *mRNA* sequence present at any particular time. Transcriptomics may be viewed as “snapshots” of the portion and regions of the genome that are now being transcribed. In recent years, RNA-seq has eclipsed DNA microarrays due to their ability to mirror wholesome sequencing of transcriptomes rather than focusing on some selected transcripts/probes. Further, microRNAs tend to play crucial roles in repressing *mRNA* targets, especially in humans [28].

A proteome is derived from ‘genome’ and ‘protein.’ Proteomics is a collection of large-scale data on proteins using detection methods through mass spectrometry or immunoassays. To profile proteins, the most current and widespread method uses high-resolution 2D electrophoresis to discriminate proteins from different samples in a parallel fashion. Thereafter, it is followed by staining of differentially expressed proteins (DEPs) by mass spectrometry. Like transcriptomics, we can use a $\log(a + 1)$ transformation to obtain nearly normally distributed protein expression data. Generally, profiling proteins using the mass spectrometry-based method tends to produce thousands of protein features as opposed to using the Reverse Phase Protein Arrays (RPPA)—which produces about 200 protein features [29].

Metabolomics refers to the scientific research of quantitatively studying the chemical processes of cell metabolism (metabolites) believed to carry useful biological fingerprints. Like proteomics, metabolomics shares identical statistical features in the sense that they can be profiled using mass spectrometer-based tools. Comprehensive metabolomics profiling can play a significant role in disease diagnosis or understanding the dynamics of complex biological systems. For example, using the nuclear magnetic resonance-based method (NMR), Emwas et al. [30] discuss the role of NMR-based metabolomics aimed at disease diagnosis and the limitations around such methods. Again, at the distributional level, metabolomics data often fit well with the logarithmic normal distribution. Note that subfields of metabolomics can be employed in different diverse ways. For example, exometabolomics involves the study of extracellular metabolites, which can draw many applications, such as biofuel mining, bioengineering, or how small cells interact to cause cellular transformation.

Issues with Missing Values in Omics Data

The presence of missing values in complex omics data can often be traced to many sources and has the potential to significantly affect model fitting or cause assessment challenges. For large-scale data such as transcriptomics, genomics, etc., missingness could be attributable to inherent limited technology, measurement errors, poor experimental design, and low sample quality [31,32]. Missing values in omics data can be grouped into three kinds, namely, missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR). Considering transcriptomics data, for example, [33] argue that transcriptomics data have approximately no less than 10% missingness, and this affects over 90% of genes. In many cases, this missing value proportion can be attributed to faulty slides or corrupted images. For metabolomics, the proportion of missing values is around 11%–21%, tending to affect 80% of the features. There are available methods to handle missing values (MAR, MNAR, and MCAR), from the simplest imputation techniques to complex methods. Classical approaches that employ multiple imputations may produce biased results. One suggested solution is to identify features in the omics data with high rates of missing values, say above 60%, and drop such features from the data before running imputation on the remaining features in the data (e.g., KNN). Whether we choose listwise, pairwise, or variable deletion, the ultimate aim should be targeted at methods that avoid introducing extreme bias into the data. It is important to note, however, that some methods are completely impute-free. For example, mixture models are impute-free. Thus, they handle missing values when directly used (see Taylor et al. [34]).

Survival Prediction Modeling

In many applications of survival analyses, one of the core goals is to estimate the probability of an event occurring at

time t . This event in many medical research may be death due to a medical condition. However, classical methods such as Cox proportional hazards (CoxPH), despite their huge success in statistical science, are hugely limited in an era of growing complexity of survival data. Thus far, the appropriateness of several survival models has been linked to whether they are feasible with high dimensionality. This section briefly overviews survival modeling in the context of both high- and low-dimensional survival data.

Modeling Low-Dimensional Survival Data

Let X_i be a matrix of p covariates (clinicopathological and omics features), T_i be the recorded survival time, δ_i be the Bernoulli variable of censorship status, and n be the number of patients. Then the observed survival data, D , is represented as $D = \{T_i, \delta_i, X_i, \forall i=1,2,\dots,n\}$. For right-censored survival data, the survivor function, $S(t)$, is estimated by taking into consideration the individuals at risk (n_i), the number of observed events (d_i), and the number of censored subjects at time t_i . For distribution-free estimation, Kaplan and Meier [35] proposed the *product limit estimator* based on multiplying out the conditioned probabilities of the event occurring. Mathematically, it is stated as

$$S_{km}(t) = \prod_{i:t_i < t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i < t} \left(1 - \frac{d_i}{n_i}\right),$$

$$S_{km}(t_k) = \frac{n_{k-1} - d_{k-1}}{n_{k-1}} S_{km}(t_{k-2}), \dots \text{ [by recursion]},$$

where $S_{km}(t)$ is the KM estimator for untied survival times, i.e., $T_1 < T_2 < \dots < T_k$; ($k \leq n$).

Referring back to the low-dimensional NCCTG lung cancer data we illustrated in Table 2, the data can be used to obtain some basic survival insights. For example, Fig. 3a shows the KM survival curve for the NCCTG lung cancer

data with a median survival time of about 306 days (i.e., OS ≥ 10 months). The 1- and 2-year survival probabilities are 0.409 and 0.120, respectively. In Fig. 3b, the OS of the patients is stratified by sex (i.e., male-female). The P-value of 0.0013 shows a significant difference in the survival probabilities of the patients when sex is factored in. This comparison of the difference between the two survival curves is evaluated with the *log-rank testing*. From the survival curves by sex, it appears the female patients have relatively better survival prognosis than their male counterparts.

While the Kaplan-Meier approach can be used to estimate the survival probability in a covariate case-by-case framework, it fails to jointly measure the prognosis of patients of several covariates at the same time. Oftentimes too, the survivability of the individual may depend on several different factors—clinical and/or omics features. Under a multivariable setting, the CoxPH [36] can fit the survival data. Generally, the CoxPH is formulated as;

$$h(t|x_i) = h_0(t) \exp \left(\sum_{j=1}^p x_{ij} \beta_j \right), \quad (1)$$

where $h_0(t)$ is the unspecified baseline hazard, β is a vector of regression coefficients, and X is a matrix of risk factors or features together with the survival time and the status variable. As the name suggests, the Cox proportional hazard model is valid when the proportionality assumption is fairly met (extremely hard to hold for most untransformed omics data). Because the baseline hazard need not be specified, CoxPH is widely considered a semi-parametric model. The CoxPH allows for two group hazards to be compared in what is termed as *hazard ratio* (HR).

The “curse of dimensionality” associated with high-dimensional ($p \gg n$) omics data renders the CoxPH model

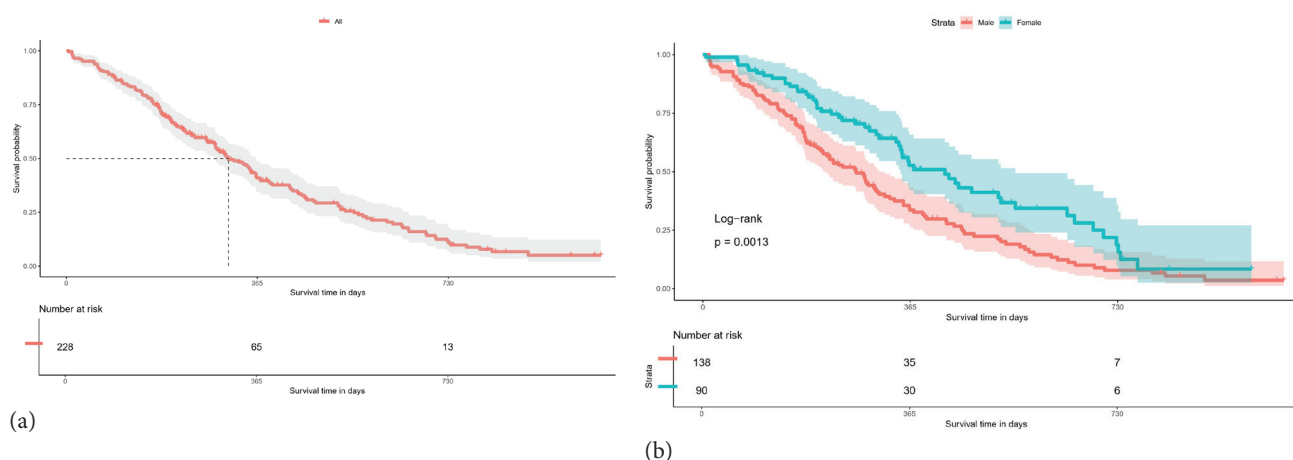


Figure 3. KM curves of the lung cancer patients. Adapted from: C. Loprinzi, et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. 1994. Available at: <https://doi.org/10.1200/jco.1994.12.3.601> **A.** Kaplan-Meier overall survival curve of NCCTG lung cancer patients. **B.** Overall survival estimation stratified by gender/sex.

impracticable in many cases. For low-dimensional data, the modeling method is usually used together with feature/variable selection methods for multivariable survival prediction, and there exists sufficient literature on practical considerations for handling problems with low-dimension survival models [37].

Modeling High-Dimensional Survival Data

As stated earlier, the enormity of high-dimensional features makes the applicability of traditional survival models infeasible. For example, DNA methylation can produce hundreds of thousands of features. This high dimensionality invariably affects multivariable modeling in terms of computational time and interpretability. One hack around this problem is to consider a subset of pre-selected features to be used in univariate case-by-case modeling. Even so, the pre-selected features will usually be selected based on their high statistical significance (e.g., $p < 0.01$). That approach, however, poses serious problems, one of them being that there is the likelihood of missing out on some important variables having significant associations with the outcome variable [38].

Further, it is possible to use total variance explained (TVE) to pre-select features to be used in the model. The reasoning behind this approach is that the higher the variance explained, the higher the predictive signal associated with the features. Zhao and Zucknick [39] argue that features accounting for 50% or more of the total variance can be selected and included in the model. An approach like this helps reduce the dimensionality of omics data having thousands of features and thus improves stability in the selection. In a comprehensive benchmark study, Zhao et al. [40] demonstrated the importance of pre-selection of features towards stability under a Bayesian stochastic selection setting. In a similar study, Bommert et al. [18] revealed that, of all the filter selection techniques, the simple variance approach was the most outstanding method under evaluation metrics such as computational time and stability.

For a univariate CoxPH, it is useful to consider adjusting the P-values of the covariates/features to draw insights from multiple comparative feature analyses. The reason behind this is that it helps contain the influence of false positives, e.g., false discovery rate (FDR). Remember that the issue of false discovery rate is a serious problem when performing simultaneous hypothesis testing, and for this reason, many researchers have devoted valuable time to producing detailed research on methods and solutions to the problem of multiple testing (see [41,42]). While the univariate CoxPH may be used at the pre-analysis stage to pin down influential features, it is important to remember that such an approach fails to capture interactions in the case of multi-omics survival data.

The complexity with which omics data are presented makes the classical survival methods infeasible for capturing nonlinearity and complicated interactions among hundreds of features. These features may give strong predictive

signals toward patients' survival. While pre-selection of subsets of features could help in dimension reduction, such an approach can be bias-prone, rather needlessly. To this end, various state-of-the-art modern methods in literature offer alternative methods that are superior both in computational time and adaptability to nonlinearities of feature effects. For efficient variable selection, one may employ supervised learners such as penalized regression methods with appropriate loss functions such as L_1 -norm or L_2 -norm. In essence, the regularized selection methods tend to unselect sparse features—thus, of the thousands of features, the model selects those features that have significant effects on survival prospects.

Machine Learning Methods for High-Dimensional Data

In recent years, several machine learning (ML) methods have been used to bridge the gap of limitations associated with classical modeling techniques. Generally, ML methods, herein referred to as *learners*, are equipped with the ability to analyze complex data by revealing complex interactions among features that are otherwise hard to discern. This section discusses such learners.

Unsupervised learning

The primary goal of unsupervised learners is to recognize the difficult-to-find patterns in complex data, e.g., transcriptomics data and DNA methylation microarray data. For example, cardiovascular diseases (CVDs) may better be understood by uncovering their underlying molecular signatures so that the complex level of abstractions is representative of the underlying phenotypes for better clinical outcomes [43]. Unsupervised classifiers in this case use the large-scale data to repeatedly 'learn' hidden patterns and transform their complexity into a discernible low-dimensional plane. A popular data reduction technique is the Principal Component Analysis (PCA), where several hundreds of features in the data are represented with fewer features based on the Total Explained Variance. Thus, a high-dimensional hyperplane is projected into a low-dimensional plane.

While the dimension-reduction technique in PCA is tied to extracting maximized variance in the data, other proposed methods can achieve similar goals, including t-SNE, which uses nonlinear methods. In general, t-SNE reduces high-dimensional data into a low dimension using algorithms such that dissimilar data points are kept apart and similar data points are kept together. The t-SNE uses the Kullback-Leibler (KL) divergence to evaluate this difference. Besides t-SNE and PCA, the Uniform Manifold Approximation and Projection (UMAP) is another dimensional-reduction method with the algorithm behind it depending on three core assumptions: i). uniformity of data distribution on the Riemannian fold, ii). the Riemannian metric has a local constant, iii). the connection to the manifold is done locally [44]. In high-dimensional scenarios, one useful advantage of the UMAP techniques lies in the

method's superior visualization capabilities coupled with enhanced shorter computational time.

Because unsupervised modeling strives to discover hidden patterns or structural forms of the unlabeled data, the approach comes with obvious challenges, especially with survival prediction. The resulting model from repeated 'learning' may fail to establish an interpretation between the survival prospects and signals in the hidden layers. This is true with several ML techniques that tend to produce complex models that are difficult to understand—the so-called black box models.

Penalized Cox Modeling

A patient's survival information may be better studied by considering the joint effect of clinicopathological and multi-omics risk features. This is particularly important in personalized medicine. However, with the enormous nature of omics data, feature-selection methods become necessary in selecting the most important factors that potentially contribute to survival prediction. For example, while genome-wide association studies (GWAS) may serve as a pathway to test the statistical significance of thousands of genetic variations across many genomes, the problem of sparsity in high-dimensional microarray data poses serious challenges with non-robust prediction models. One approach to overcoming this problem involves conducting multiple t-tests to identify vital disease-linked phenotypes.

Another approach to identifying relevant features linked to the survival outcome is to employ regularization methods in the traditional CoxPH. Regularization is an ML technique employed to improve and prevent overfitting a probabilistic model. For example, to identify the most significant features among hundreds of thousands of omic features, different penalty functions can be used to shrink to zero the regression coefficients of insignificant features. This is done by utilizing the penalized partial likelihood function;

$$L(\beta|\mathcal{D}) = \prod_{i=1}^n \left\{ \frac{\exp(\beta'x_i)}{\sum_{j \in R(t_i)} \exp(\beta'x_j)} \right\}^{\delta_i} \quad (2)$$

$$\log(L(\beta|\mathcal{D})) = \log \left(\prod_{i=1}^n \left\{ \frac{\exp(\beta'x_i)}{\sum_{j \in R(t_i)} \exp(\beta'x_j)} \right\}^{\delta_i} \right),$$

where, as before, $\mathcal{D} = \{(T_p, \delta_p, X_i) \forall n = 1, 2, \dots, n\}$ is the observed survival data (*i.i.d*) together with the features (p), $R(t_i) = \{j: t_j \geq t_i\}$ is the risk set at time t_i that consist of the censored and uncensored patients before time t_i . We wish to estimate the regression coefficients, β 's and shrink the insignificant features to zero. One extension to the CoxPH is the penalized *lasso* Cox technique as proposed by Tibshirani [45]. This method uses the L_1 -norm loss function such that,

$$\frac{2}{n} \cdot L(\beta|\mathcal{D}) - \lambda \|\beta\|_1,$$

where λ is the so-called 'tuning' parameter used to regulate the penalty signal. As $\lambda \rightarrow \infty$, the coefficients get more regularized. This means a $\lambda = 0$ implies the coefficients are not penalized. Note here that the factor $2/n$ is conveniently chosen for scaling. Two other popular penalized Cox regression methods are ridge regression and elastic net Cox regression. While the lasso-Cox method uses L_1 -norm for penalization, the ridge method uses L_2 -norm, and both 'norms' are used in the case of elastic net to improve the lasso method. The penalty function for the elastic-net Cox regression [19], therefore, is $\lambda \sum_{l=1}^p |\beta_l^2| + (1 - \alpha) \sum_{l=1}^p \beta_l^2$. Binder and Schumacher [46] in a major extension, considered a more tailored penalization of features where it is possible to exceptionally include mandatory variables into the model without penalizing them. To demonstrate the effectiveness of the proposed extension, the authors used microarray data for patients diagnosed with diffuse large B-cell lymphoma. Under different algorithmic schemes, several penalized loss functions have been proposed for feature selection, including Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR-Cox) [47], Smoothly Clipped Absolute Deviation (SCAD) [48], Adaptive-lasso Cox [49], etc.

Validating Survival Models

In statistical analysis, model fitting can go wrong on several fronts, including overfitting. For example, regularization techniques strive to prevent the overfitting of models by introducing different penalty functions, thereby improving the statistical inference of the unseen data. To this end, validating a model is crucial to performance assessment and can help reveal the model's underlying violations or specifications. For instance, the proportionality assumption of hazards (PH) in a CoxPH model need not be seriously violated such that we do not see two survival curves cross each other. One way to validate this can be done by examining graphical diagnostics using scaled Schoenfeld residuals [50]. Note that Schoenfeld residuals are not time-dependent. This means that, for a diagnostic plot of the Schoenfeld residuals against time t , if the plot shows non-randomness against time, then the PH assumption is most likely violated. Once one or more assumptions are conspicuously violated, other corrective approaches can be adopted to improve predictive power. In the case of CoxPH, for example, we can adopt various functions to transform the features (e.g., log transformation). Another alternative too can be considering the effect of interactions for time-varying covariates (see Li et al. [51]) or via stratification. In the case of high-dimensional data ($p \gg n$), models' diagnostic checks are more challenging than in the low-dimensional sense. Bin et al. [52] addressed some challenges of validating methods associated with prediction models and suggested the need to modify classical methods to fit the high-dimensional omics data. For other models too, it is important to relax stringent assumptions so

that modeling can be adjusted accordingly without unduly introducing bias.

Survival prediction and calibration

A good statistical model is one that leads to a valid inferential deduction with high predictive accuracy. In the context of survival modeling, this would mean a probabilistic model aimed at improving predictive performance. Further, for a given prediction model, one could assess the adjusted effects on survivability when new risk features are introduced. Assessing the predictive power of a model refers to evaluating the magnitude of the prediction error on a test dataset after the model is ‘trained.’ Various approaches have been proposed to evaluate predictive power, from discriminant metrics to overall model performance. For example, we can simply use robust statistics (e.g., median) to discriminate prognostic estimates, followed by a log-rank test [53,54].

Assessing discriminative power

The performance of a prediction model can be evaluated via discrimination—defined as the model’s capacity, for example, to distinguish upregulated expressed genes from the downregulated ones. Common discriminative metrics for survival outcomes include Harrell’s concordance index [55], Uno’s concordance index [56], and receiver operating characteristic curve (ROC) [57]. As a goodness-of-fit measure, Harrell’s concordance index evaluates risk scores by considering all pairs of samples of the data on comparative grounds and identifies *concordant pairs* and *discordant pairs* through probability estimates. The ratio of the number of concordant pairs to that of the sum of the concordant and discordant pairs is the c-index. Here, the measure is independent of time and gives better global estimates of the prediction model. C-index lies between 0 and 1—an index score of 0 indicates poor discrimination, while an index score of 0.5 means the prediction is no better than tossing a fair coin. If the survival data has a high censoring rate, Harrell’s concordance index is biased, making Uno’s index a better alternative.

Another popular discriminative measure considers the Receiver Operating Characteristic (ROC) with its Area Under the Curve (AUC). It is a diagnostic measure of discrimination that underpins specificity and sensitivity. For example, to predict survivability for a fixed time point (e.g., 10-year survival prediction), a ROC curve can be considered since it averages the value of specificity, taking into consideration all possible values of sensitivity, or vice versa [58]. Like Harrell’s index, it ranges between 1 and 0. A value of 1 indicates perfect predictive risk scoring, while 0 indicates imperfect predictive risk scoring. This way, a value of 0.5 means the discrimination does no better than a fairly tossed coin.

Assessing calibrating power

To improve the predictive power of survival models, calibration may be considered. Calibration evaluates the

closeness of the predicted probability outcome to the actual outcome. Stated differently, we are interested in measuring the degree of closeness between the predicted and the observed. A properly calibrated model is one in which the observed and the predicted agree in the context of well-established statistical methods (e.g., Pearson χ^2 statistic). For instance, [59] suggested the calibration method for hazard models in cases where the PH assumption is violated.

Overall model performance

Quantifying uncertainty in probabilistic models can be considered diagnostic metrics for assessing the goodness of calibration or for evaluating discriminative power [60]. The theory behind this is based on what is termed in statistical modeling as scoring rules. Proper scoring rules generally guide how to evaluate overall model performance. Common examples include the Brier score [61], pinball loss [62], etc.

As a measure of overall performance, the Brier score integrates both calibration and discrimination measures for binary predictive outcomes. For a given survival data, the Brier score (BS) is computed as follows:

$$\text{Brier score (BS)} = \frac{1}{n} \sum_{j=1}^n (p_j - o_j)^2,$$

where n is the number of individuals/subjects, p_j is the event’s probability as predicted by the model, and o_j is the observed outcome for the i th individual. As before, the status variable is coded such that 1 means the event is observed while 0 means a non-observed event. This implies that a score closer to zero means the model’s performance is good.

For different time points, the time-dependent Brier score as proposed by Graf et al. [61] is calculated as;

$$BS(t) = \frac{1}{n} \sum_{j=1}^n \left\{ \frac{\hat{S}(t|t_j)^2 \mathbb{1}\{t_j \leq t \text{ \& } \delta_j = 1\}}{\hat{G}(t_j)} + \frac{\{1 - \hat{S}(t|t_j)\}^2 \mathbb{1}\{t_j > t\}}{\hat{G}(t)} \right\}, \quad (3)$$

where the j th patient has survival time t_j , $\hat{S}(t|t_j)$ is the survival probability predicted by the Cox learner, and $\hat{G}(t)$ is the estimate using the Kaplan-Meier method. The overall model performance, relative to the follow-up times, can be evaluated with a single metric—thus, integrating the multiple Brier scores. It is called the integrated Brier score (ibrier).

Generally, overfitted models tend to exhibit poor generalization to the unseen data. The fitted model may show great performance with the training data but abysmal performance with the new or unseen data. To overcome this undesirable problem, a survival model needs to be validated on more independent data with representativeness mirroring the input data values. A common step in a validation process is to split the dataset into two distinct parts: the training set and the testing set. Though splitting the data could lead to partial loss of information, resampling (e.g., bootstrapping, jackknifing) can be used to improve performance. Resampling methods have been shown to be

robust to unduly observations in the data, besides the fact that they allow one to examine the performance measure and estimate its uncertainty (see Gerds and Kattan [63]).

Real Application: Survival and Bioinformatics Modeling of HCC

In oncology, liver cancer happens to be one of the leading causes of death, only coming third after lung and colorectal cancers [64]. Of all the cases of liver cancer, hepatocellular carcinoma roughly accounts for 80%–90% [65]. With viral hepatitis (i.e., Hep B and Hep C) infections being the central risk indicators of HCC development, there is a need for more novel prognostic and/or diagnostic modeling approaches toward improved personalized medicine. Thus far, the high dimensionality of ‘multi-omics’ data presents the research society with an opportunity to assess the cumulative effect of clinical and omic features on disease and patient management. This section focuses on survival prediction based on differentially expressed genes. The analysis is done using survival modeling concepts that are covered herein. The data used is based on patients diagnosed with Hepatocellular Carcinoma (HCC) and publicly hosted on the TCGA Research Network: <https://www.cancer.gov/tcga>.

In recent years, researchers have proposed various biomarkers for building HCC prediction risk scores and validating these biological features. For instance, Parikh et al. [66] review a class of biomarkers described as ‘promising’ for modeling and detecting hepatocellular carcinoma in its early stages. Here, we explore differential gene expression

analysis (DGEA), screening significant genes for further machine-learning analyses. Our findings suggest a potential research path for exploring new insights into the predictive role of certain genes in the overall survival time of HCC patients or for further prospective studies to validate the screened genes.

Methods

Data collection and pre-processing

Using the TCGAbiolinks package in R, we downloaded expression and clinical data from The Cancer Genome Atlas corresponding to hepatocellular carcinoma (TCGA-LIHC) and processed the data for survival and bioinformatics analyses. The downloaded data consisted of 421 HCC cases, including 364 primary solid tumors (TP) and 50 solid tissue normal (NT) samples. Additionally, the TCGA-LIHC dataset involved 60,660 gene expression features. Individuals with incomplete survival information were censored. Pre-processing also included merging certain cancer stages—Stage III, IIIA, IIIB, and IIIC were combined into Stage III. Due to the relatively low counts of Stage IV cases (2), Stage IVA (1), and Stage IVB (2), they were merged into Stage IV. A summary of the clinicopathological factors is shown in Table 3. Note that patients with missing survival data were excluded from subsequent analyses. Due to the high dimensionality of the data ($p \gg n$), the M3C R package was used to conduct unsupervised exploratory analyses, principally t-SNE, PCA, and UMAP. This approach was used to determine if TP and NT samples were significantly

Table 3. Summary of key clinical features and pathological features of the TCGA-LIHC dataset. Note that NA’s observations were filtered out

Survival status	Alive (n=255)	Dead (n=164)	Total (N=421)
Age in yrs ($\bar{x} \pm SD$)	59.10 $\pm$ 13.40	61.70 $\pm$ 14.00	60.11 $\pm$ 13.66
Gender:			
Female	75	66	141
Male	174	98	272
Definition:			
Primary Solid Tumor	234	130	364
Solid Tissue Normal	15	34	49
AJCC-Pathologic-Stage:			
Stage I	134	52	186
Stage II	61	34	95
Stage III	2	4	6
Stage IIIA	32	29	61
Stage IIIB	4	4	8
Stage IIIC	4	6	10
Stage IV	0	2	2
Stage IVA	1	0	1
Stage IVB	0	2	2

discriminated. As described in the next section, the data was further processed after identifying the significantly expressed genes.

Differential gene expression analysis

Using the DESeq2 package, the downloaded expression data (TCGA-LIHC) was processed in several stages. First, the genes with low counts (< 10) were filtered out. The resulting data was then transformed using the variance stabilizing transformation function (i.e., `DESeq2::vst()`). Using a significance level of $\alpha = 0.05$, we obtained expressed genes via the function `DESeq2::DESeq()`. However, to control the false discovery rate (FDR), the Benjamini-Hochberg procedure was used to obtain the adjusted P-values. Therefore, differentially expressed genes (DEGs) were screened using the criteria: $p < 0.05$ and $|\log_2\text{FoldChange}| > 1.50$. Differentially expressed genes (3,640) were classified into 'Downregulated', 'Upregulated', and 'Not significant' as shown in Fig. 4a.

Prognostic and diagnostic expressed genes

To identify prognostic and diagnostic genes, the DEGs were ranked by their $|\log_2\text{FoldChange}|$ such that the corresponding transcriptome profiling data of the expressed genes were extracted, log-transformed, and filtered using the 5% varied features, i.e., `M3C::featurefilter(percentile = 5, method = "var")`. This resulted in 182 screened genes for prognostic and diagnostic examination.

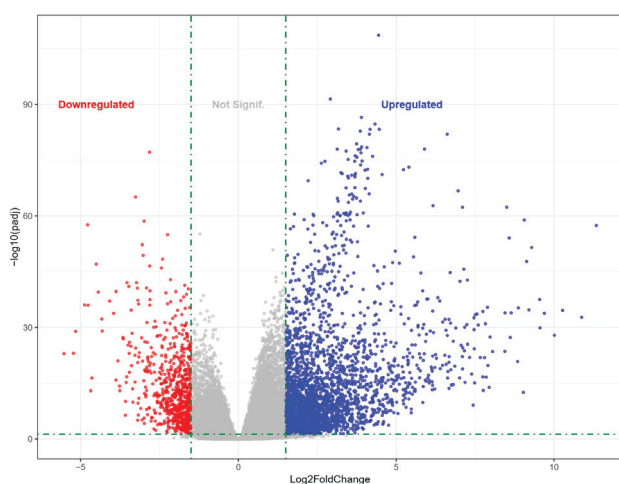
To identify potential diagnostic genes, different machine-learning algorithms were used. However, because the original data is severely unbalanced (i.e., 364 tumors vs. 50 normal tissues), we performed Synthetic Minority Oversampling Technique (SMOTE) analysis to 'create' a balance between the two unbalanced classes. Note that

SMOTE analysis involves running an algorithm to balance uneven datasets—thus creating a balance between oversampled and undersampled groups in datasets. This is done by 'compensating' the unfavored class (in this case, normal tissues). In R, this is done by running the `performance::smote()` function. The resulting data was then subjected to five machine learning classification models (Fig. 5a–b) to determine which of the algorithms best discriminates between primary tumor and normal tissue samples. Hyperparameter tuning was performed for the selected ML model through Recursive Feature Elimination Cross-validated (RFE-CV), thus `caret::rfe()`. The 'best' model was used to extract the most important features. These features were then used in a penalized Cox regression (lasso) to screen the final diagnostic genes (Fig. 5c–d).

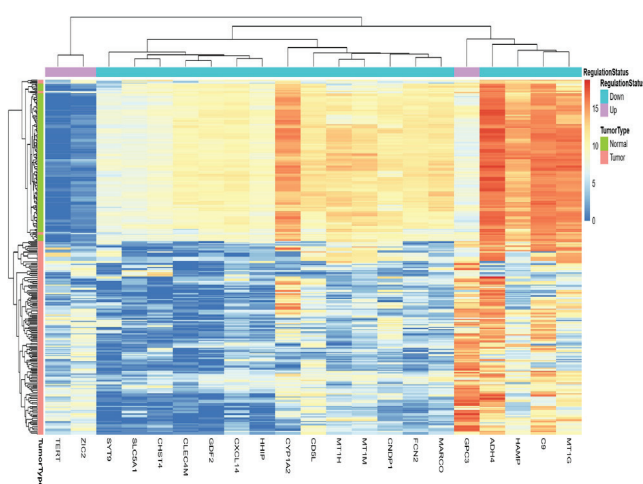
To identify potential prognostic genes, again, the original 182 filtered genes were combined with their corresponding survival data. The data was partitioned into a 60/40 ratio—60% comprising the training dataset ($N = 290$) and 40% representing the testing dataset ($N = 123$). A univariate CoxPH was performed to identify differentially expressed genes (features). Additionally, the features found to be statistically significant were further used in a stepwise multivariable CoxPH modeling—using both backward and forward selection methods and the decision criterion based on the Akaike Information Criterion (AIC). A risk score was estimated from the final estimated model as

$$\text{Risk Score} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{gene}_1 + \hat{\beta}_2 \times \text{gene}_2 + \dots + \hat{\beta}_n \times \text{gene}_n + \text{some clinical effects}, \quad (4)$$

where $\hat{\beta} = \hat{\beta}_0, \dots, \hat{\beta}_n$ are the estimated regression coefficients of regressing the survival time over the most significant extracted expression features; $\text{gene}_1, \text{gene}_2, \dots, \text{gene}_n$.



(a)



(b)

Figure 4. Volcano plot (a) of 3,623 differentially expressed genes screened under the criteria: $|\log_2\text{FC}| > 1.50$, $\text{FDR} < 0.05$. Down- and upregulated genes are colored red and blue, respectively. (b) Hierarchical clustering (heatmap) of potential diagnostic features, consisting of 21 differentially expressed genes between primary tumors (TP) and normal tissues (NT) in TCGA-LIHC. Screening of the biomarkers was based on recursive feature elimination with cross-validation (RFE-CV).

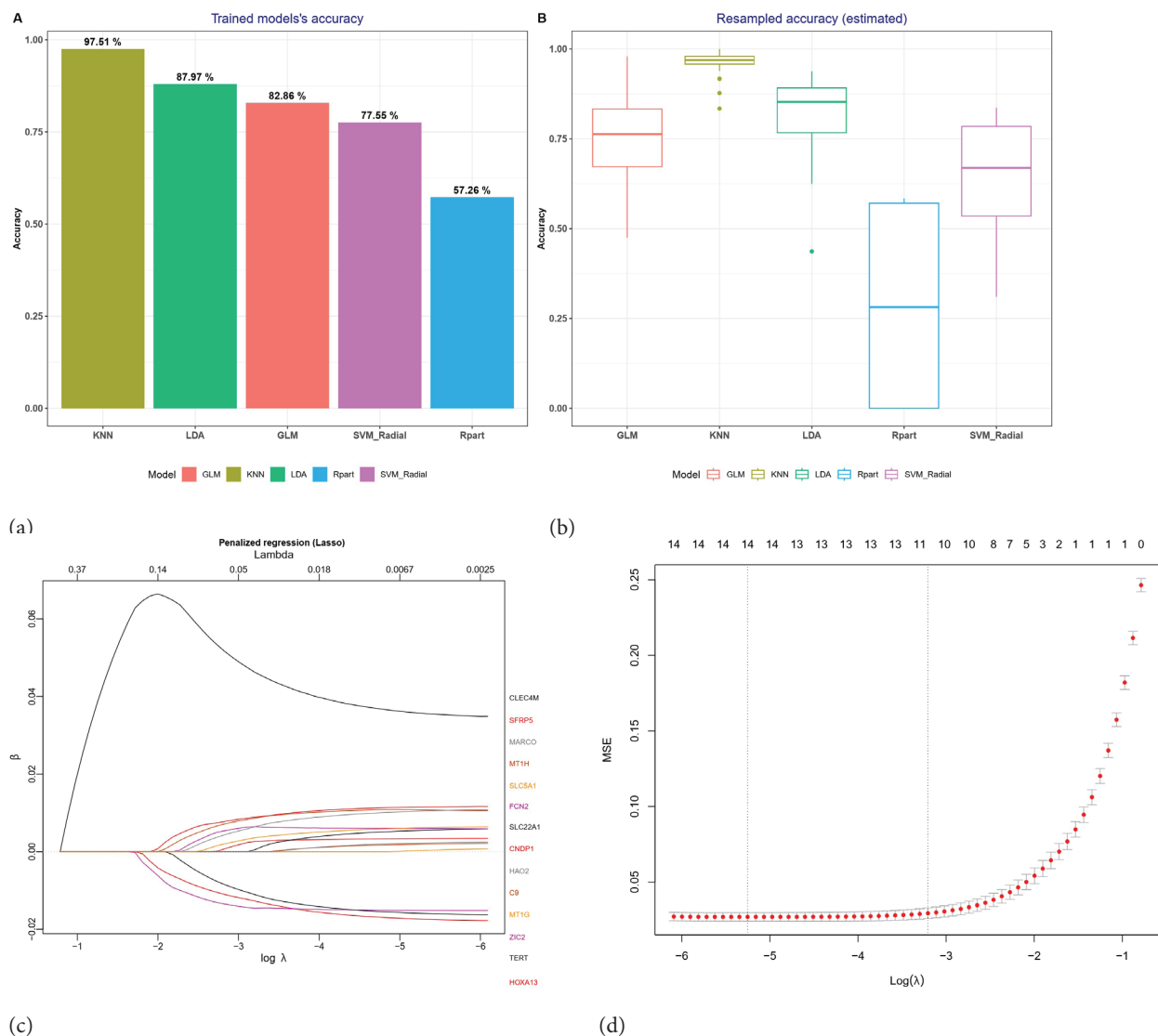


Figure 5. (a) Bar plot of the accuracies of 5 ML classification algorithms (training dataset) used to screen the differentially expressed gene (DEGs) diagnostic biomarkers. Methods with hyperparameters were further tuned for the final screening. (b) Performance assessment of the trained methods through resampling. (c) Penalized lasso regression to screen 21 diagnostic- and LIHC-specific differentially expressed genes. (d) Plot showing the cross-validated mean square errors (MSE).

For each of the train and test datasets, patients were stratified into high- and low-risk groups based on the median risk score. Kaplan-Meier (KM) curves were performed for each screened DEG (Fig. 6a–f). The performance of the potential prognostic genes from the multivariable CoxPH was evaluated using the Receiver Operating Curve analysis (ROC). Performance was assessed using sensitivity and specificity (Fig. 7c–d).

Gene set enrichment analysis (GSEA)

Gene set enrichment analysis (GSEA) was explored for the two risk groups (low- and high-risk). This was done using the clusterprofiler R package to obtain the biological processes (BP) related to the identified expressed genes.

The criteria for obtaining significant GSEA results were based on $p < 0.05$, FDR < 0.05 , and a normalized enriched score greater than one (i.e., NES > 1).

Statistical analysis

All the statistical analyses considered in this study were done using R Software (4.4.1). Five machine learning classification algorithms were considered for diagnostic screening, followed by a lasso CoxPH regression. Potential prognostic screening was done via univariate and multivariable CoxPH, and performance was assessed via ROC analysis. The Benjamini-Hochberg procedure was adopted to deal with false discovery rates.

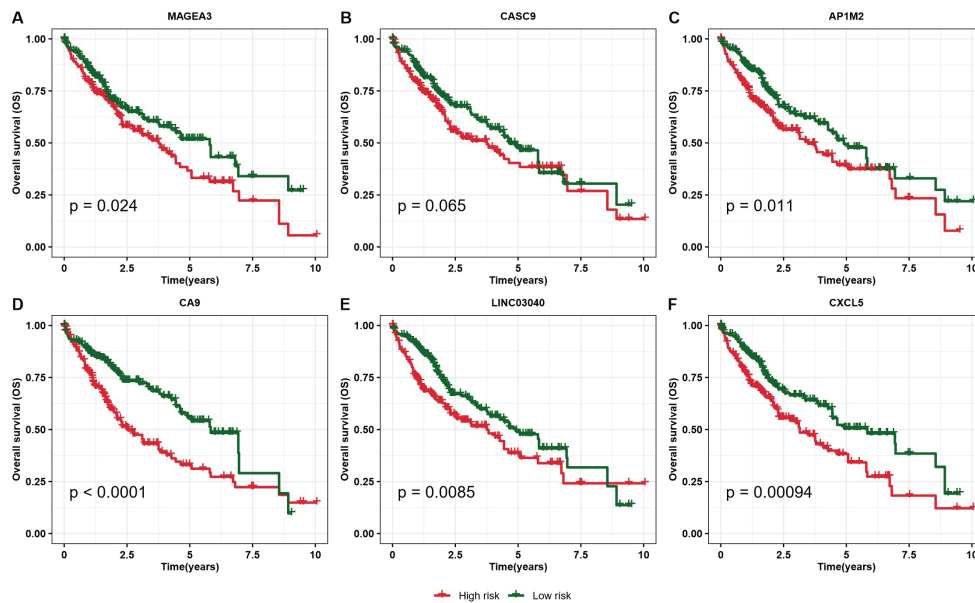


Figure 6. Kaplan-Meier (OS) plots of the six differentially expressed genes screened as prognostic biomarkers in the univariate CoxPH modeling.

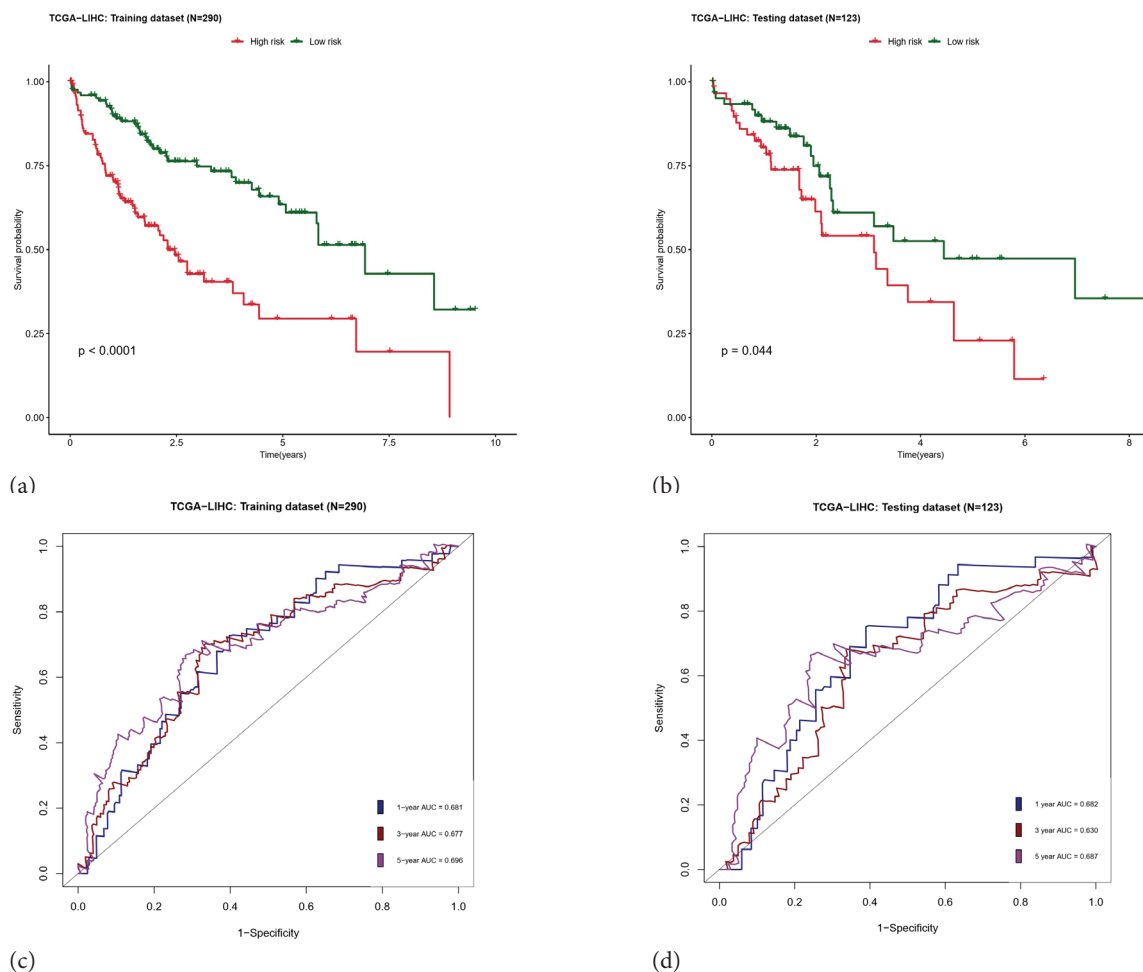


Figure 7. Assessment of the final six screened prognostic biomarkers. Based on the median risk score, KM analysis was performed between low—and high-risk groups (a-b). The six features' ROC evaluation was analyzed using sensitivity and specificity (c-d).

RESULTS AND DISCUSSION

Results

Using the TCGA repository, the data on hepatocellular carcinoma comprised primary tumor samples ($n = 364$) and normal tissue samples ($n = 50$). A SMOTE analysis to get data balanced resulted in 147 normal samples and 196 primary tumor samples. Through differential gene expression analysis, we screened 3,623 significantly expressed genes ($p < 0.05$, $|\log_2\text{FoldChange}| > 1.50$), including 685 downregulated genes and 2,938 upregulated genes, as shown in Fig. 4a. A hierarchical clustering of the diagnostic screened genes (21) showed clear discrimination between the two tumor classes—thus, normal and primary tumor samples, as shown in Fig. 4b.

Feature selection of the expressed genes was performed in stages. Initially, the top 5% (182 genes) accounting for the total explained variance (TVE) were screened for further diagnostic and prognostic evaluation. For diagnosis, the log-transformed RNA data corresponding to the filtered genes was partitioned (70/30) into training and testing datasets and was used in five machine learning classification models, namely decision trees (RPART), linear discriminant analysis (LDA), radial support vector machine (rad_svm), k-nearest neighbors (KNN), and generalized linear model (GLM). A 5-fold cross-validation method with 3 repeats was also applied in training the classification model. Models with hyperparameters were further tuned in the validation stage. Note that the simple LDA model has no hyperparameters and was, thus, not hyperparameter-tuned. The results showed the KNN as the best-performing method (Fig. 5a–b) for discriminating between normal and tumor samples for the DEGs. Variable importance for this method revealed the top 21 features. Therefore, 21 features were selected for further regularized modeling (here, LASSO). Using the `cv.glmnet` function, 10-fold cross-validation was finally performed to identify significant diagnostic features (genes/biomarkers) shown in Fig. 5c–d.

To identify the prognostically expressed genes, we filtered the overall survival data and combined it with the mRNA data of the selected 182 genes. As before, the data was partitioned into training and testing sets. Initially, a univariate Cox proportional hazards (CoxPH) analysis was performed at a significance level of ($p < 0.01$). The results revealed five significant genes: CXCL5, MAGEA3, CASC9, AP1M2, and CA9. Fig. 6a–f shows the Kaplan-Meier (KM) analysis of these six genes concerning overall survivability. The cumulative effect of the univariate-selected genes was further evaluated using a stepwise multivariable CoxPH analysis with both backward and forward selection methods. Using the Akaike Information Criterion (AIC) as the decision criterion, the results identified three significant genes: CASC9, CA9, and AP1M2. These three genes were then used to compute the risk score of the patients as

$$\text{Risk Score} = 0.0469 \times \text{AP1M2} + 0.0644 \times \text{CASC9} + 0.0764 \times \text{CA9},$$

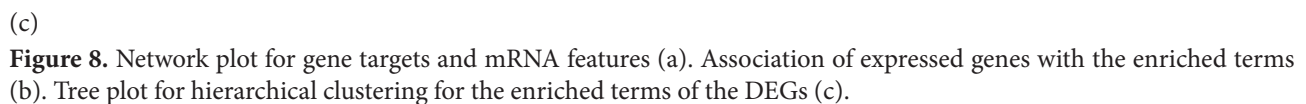
Using this risk score, patients were classified into 2 groups—those whose mean risk score was greater than the median risk score were classed into the high-risk group, whereas those with a mean risk score lower than the median risk score were grouped as the low-risk class. The overall survival curve (KM) for both the training dataset and the validation datasets was obtained (Fig. 7a–d). The performance of the trained dataset was validated through a Receiver Operating Curve (ROC) analysis with performance metrics based on sensitivity and specificity as shown in Fig. 7c–d. Notice that the overall survival time is evaluated at 1-, 3-, and 5-year time points.

To gain further biological (cellular and molecular) insights into the enriched terms of the 182 DEGs, we performed gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. The GO analysis results revealed mostly enriched terms in “Hepatocellular Adenoma,” “Hepatoblastoma,” “Liver Regeneration Disorder,” and “Chronic Hepatitis” (Fig. 8), involving some of the identified diagnostic and prognostic genes screened in this study. The KEGG signaling pathways analysis revealed five pathways, with the highest count in the “hsa04976: Bile Secretion (Digestive System)” pathway. Two of the DEGs identified in our diagnostic analysis (“SLC22A1” and “SLC5A1”) were found in this pathway. These two genes, therefore, point to a crucial area for further research (Appendix Fig. S1 and Fig. S2).

Discussion

Hepatocellular carcinoma is one of the leading causes of liver-related deaths, with Hepatitis B and C known to be some of the risk factors influencing the high incidence. While extensive advances have been made in several treatments and/or management targets, there is still room for utilizing existing and emerging bioinformatic techniques to identify insightful biomarkers—oncogenetic or tumor-suppressors. Combining bioinformatics and survival modeling methods can, therefore, enhance the screening process of diagnostic and prognostic biomarkers toward targeted therapeutics. In this study, we strove to use the expository techniques earlier discussed with differential gene expression analysis to identify a gene or cluster of genes with potential prognostic/diagnostic signatures. In the present study, we performed a detailed bioinformatics analysis of transcriptome (gene expression) data on hepatocellular carcinoma (TCGA-LIHC) on two fronts: those deemed potentially diagnostic and those that may potentially possess prognostic insights.

While 3623 genes were significantly differentially expressed ($\text{FDR} < 0.05$; $|\log_2\text{FoldChange}| > 1.50$), our study focused mainly on the top 5% of them (i.e., 182) based on Total Variance Explained (TVE). Under the scope of selected machine-learning classification algorithms,



With the DEGs screened as potential diagnostic genes, some have already been reported in the literature in their association with the pathogenesis of hepatocellular carcinoma. More particularly, we found solute carrier family 22 member 1 (SLC22A1) and solute carrier family 5 member 1 (SLC5A1) significant in the KEGG pathway enrichment

analysis (hsa04976: Bile secretion). As a polyspecific transporter to vital organs such as the liver, kidney, and intestine, Gao et al. [67] in their study found SLC5A1 to be highly expressed and inhibit the growth of pancreatic cells. Our findings therefore point to SLC5A1 as a potential diagnostic biomarker for HCC. Additionally, while CLEC4M was highly expressed and screened as a diagnostic biomarker in our study, other studies have identified it for its association with poor prognosis of hepatocellular and colon cancer (see [68,69]). Other highly expressed diagnostic HCC biomarkers identified included SFRP5, CNDP1, TERT, and ZIC2.

From a prognostic perspective, our study focused on the overall survival (OS) prospects of HCC related to the screened DEGs. Six hub DEGs were revealed in a univariate CoxPH model, including CXCL5, MAGEA3, CASC9, AP1M2, CA9, and LINC03040, while a further multivariable modeling of these six genes revealed three significant genes (CASC9, AP1M2, and CA9). For instance, the prognostic potential of the C-X-C motif chemokine 5 (CXCL5) for HCC, as shown in the univariate modeling (Fig. 6F) has also been identified in related experimental studies [70]. In the case of CASC9, while it was marginally significant in the univariate case, its prognostic effect, along with AP1M1 and CA9, revealed strong statistical significance (Fig. 7A–B). In HBV infection, the upregulation of AP1M1 (AP-1 complex subunit mu-1) has been found to promote liver cancer cell proliferation [71]. Since HBV is one of the most common risk factors for HCC incidence, AP1M1 appears to be a promising biomarker for HCC prognostic quantification. Furthermore, Yao et al. [72] observed in their study that the long non-coding RNA of CASC9 can invade, promote, and inhibit the apoptosis of hepatocellular carcinoma cells, thus being closely linked to a poor prognosis. While high expression of carbonic anhydrase 9 (CA9) is closely linked to several cancer types, our findings reveal its independent potential as a novel biomarker for patients with HCC. Other studies have shown that CA9 is associated with poor prognosis in patients with resectable HCC [73, 74]. Combining these three DEGs (AP1M1, CASC9, and CA9) potentially provides meaningful prognostic insights into the overall survival of HCC patients. The impact of these genes was also assessed on clinicopathological variables such as age groups (<60 years vs. >60 years) and cancer stage (stage I, II, III+). The KM overall survival analysis ($p < 0.05$) showed that the two age classes were significant, but the cancer stage groups were marginally not significant (Appendix Fig. S3). To the best of our knowledge, no recent study has captured all three DEG biomarkers as a prognostic risk score estimator for HCC patients. Finally, in terms of biologically enriched terms, for example, CXCL5, as shown in Figure 8a–c, is associated with chronic liver disease, “hepatoblastoma,” “hepatocellular adenoma,” “intrahepatic cholangiocarcinoma,” and “malignant neoplasm of the gastrointestinal tract.”

In the present study, we adopted an analytical approach combining bioinformatics analysis and survival risk modeling to identify significantly expressed genes for HCC risk prediction, emphasizing their diagnostic and prognostic insights. However, our study has several limitations:

1. The TCGA-LIHC repository primarily contains samples from Caucasian and Latino patients, resulting in an unbalanced dataset in terms of ethnicity and limiting generalizability. Future studies should use a more balanced dataset.
2. Proper validation of the genes identified in our study would require a multicenter prospective cohort study rather than the retrospective study conducted here. A prospective study for the same research goal is likely a future focus of our work.
3. Another limitation of our study is the lack of comprehensive clinicopathological variables (tumor size, Tumor-Node-Metastasis (TNM) grade, and stage). Future work may consider the impact of such clinicopathological variables, including a detailed evaluation of the biological processes (BP) for the identified DEGs.

CONCLUSION

Survival analysis is considered a specialized modeling technique in inferential statistics, primarily due to the existence of well-established methods capable of handling various censoring mechanisms for survival outcome estimation. However, with the growing complexity of multi-omics data, especially in bioinformatics, the need for integrative methods has become an increasingly popular predictive modeling approach. Our study considered two parts: expository techniques in ‘omics’ survival modeling and bioinformatics analysis of gene expression data (specifically, TCGA-LIHC), with a primary focus on screening diagnostic and prognostic biomarkers for hepatocellular carcinoma patients.

In recent decades, researchers have devoted considerable resources to identifying novel biomarkers with potential prognostic and/or diagnostic insights to improve personalized medicine for hepatocellular carcinoma (HCC) patients. Therefore, the application part of this study focused on identifying such novel biomarkers. Our study identified 3,623 differentially expressed genes (DEGs), including 685 downregulated and 2,938 upregulated genes. Using a series of statistical and bioinformatics analytical methods, three significant prognostically differentially expressed genes—AP1M1, CASC9, and CA9—were identified in a multivariable Cox regression as potentially having prognostic insights into the tumorigenesis of HCC. Additionally, several strongly expressed genes were identified as diagnostic biomarkers for their ability to significantly differentiate HCC primary tumor samples from normal solid tissues. These included CLEC4M, SLC5A1, SLC22A1, SFRP5, CNDP1, TERT, and ZIC2. While these findings may provide a theoretical basis for targeted management and

evaluation of HCC in personalized medicine, it is worth mentioning that a multicenter prospective study would be required to confirm our key findings.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in The Cancer Genome Atlas database (TCGA-LIHC), with dbGaP Study Accession phs000178. The R code for the complete analyses is available with the corresponding author on a reasonable request.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

- [1] Tong L, Shi W, Isgut M, Zhong Y, Lais P, Gloster L, et al. Integrating multi-omics data with EHR for precision medicine using advanced artificial intelligence. *IEEE Rev Biomed Eng* 2024;17:80–97. [\[CrossRef\]](#)
- [2] Arib MAA. Survival analysis of students not graduated on time using Cox proportional hazard regression method and random survival forest method. *J Stat Data Sci* 2023;13–21. [\[CrossRef\]](#)
- [3] Thackham M. Survival analysis: applications to credit risk default modelling [PhD thesis]. Macquarie University; 2022. [\[CrossRef\]](#)
- [4] Karim MdR, Islam MA. Reliability and survival analysis. Singapore: Springer Singapore; 2019. [\[CrossRef\]](#)
- [5] Leung KM, Elashoff RM, Afifi AA. Censoring issues in survival analysis. *Annu Rev Public Health* 1997;18:83–104. [\[CrossRef\]](#)
- [6] Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. New York: Springer; 2003.
- [7] Loprinzi CL, Laurie JA, Wieand HS, Krook JE, Novotny PJ, Kugler JW, et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. *North Central Cancer Treatment Group. J Clin Oncol* 1994;12:601–607. [\[CrossRef\]](#)
- [8] Green MS, Symons MJ. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *J Chronic Dis* 1983;36:715–723. [\[CrossRef\]](#)
- [9] Henderson R. Problems and prediction in survival-data analysis. *Stat Med* 1995;14:161–184. [\[CrossRef\]](#)
- [10] DeGroat W, Mendhe D, Bhusari A, Abdelhalim H, Zeeshan S, Ahmed Z. IntelliGenes: a novel machine learning pipeline for biomarker discovery and predictive analysis using multi-genomic profiles. *Bioinformatics* 2023;39. [\[CrossRef\]](#)
- [11] Hasin Y, Seldin M, Lusi A. Multi-omics approaches to disease. *Genome Biol* 2017;18:1–15. [\[CrossRef\]](#)
- [12] Tang M, Xu Y, Pan M. Integrative analysis of scRNA-seq and bulk RNA-seq identifies plasma cell related genes and constructs a prognostic model for hepatocellular carcinoma. *J Hepatocell Carcinoma* 2025;427–444. [\[CrossRef\]](#)
- [13] Liu J, Yu Z, Liu Q, Dou C, Cao P, Xie X. A novel 5-differentially expressed gene (DEG) signature predicting the prognosis in patients with metastatic liver malignancies and the prognostic and therapeutic potential of SPP1. *Int J Clin Oncol* 2025;1–18. [\[CrossRef\]](#)
- [14] Cao L, Liu M, Ma X, Rong P, Zhang J, Wang W. Comprehensive scRNA-seq analysis and identification of CD8+ T cell related gene markers for predicting prognosis and drug resistance of hepatocellular carcinoma. *Curr Med Chem* 2024;31:2414–2430. [\[CrossRef\]](#)
- [15] Miranda J, Horvat N, Fonseca GM, Araujo-Filho J de AB, Fernandes MC, Charbel C, et al. Current status and future perspectives of radiomics in hepatocellular carcinoma. *World J Gastroenterol* 2023;29:43. [\[CrossRef\]](#)
- [16] Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. *Brief Bioinform* 2021;22:77–87. [\[CrossRef\]](#)
- [17] Bøvelstad HM, Nygård S, Borgan Ø. Survival prediction from clinico-genomic models—a comparative study. *BMC Bioinformatics* 2009;10:1–9. [\[CrossRef\]](#)
- [18] Bommert A, Welchowski T, Schmid M, Rahnenführer J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Brief Bioinform* 2022;23. [\[CrossRef\]](#)
- [19] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw* 2011;39:1. [\[CrossRef\]](#)
- [20] Madjar K, Zucknick M, Ickstadt K, Rahnenführer J. Combining heterogeneous subgroups with graph-structured variable selection priors for Cox regression. *BMC Bioinformatics* 2021;22:586. [\[CrossRef\]](#)
- [21] Badik ŞT, Akar M. Machine learning classification models for the patients who have heart failure. *Sigma J Eng Nat Sci* 2024;42:235–244. [\[CrossRef\]](#)

- [22] Abubakari S, Karaman F. Neutral benchmarking of survival models in health sciences: comparative study of classical and machine learning techniques. *Gazi Univ J Sci Part A Eng Innov* 2024;11:518–534. [\[CrossRef\]](#)
- [23] Okorie IE, Moyo R, Nadarajah S. The impact of socio-demographic factors on the survival of cancer patients in Zimbabwe. *Sci Rep* 2021;11:12309. [\[CrossRef\]](#)
- [24] Nazli SA, Rosman A, Mohd Kasim NA, Al-Khateeb A, Ul-Saufie AZ, Md Radzi AB, et al. Coronary risk factor profiles according to different age categories in premature coronary artery disease patients who have undergone percutaneous coronary intervention. *Sci Rep* 2024;14:15326. [\[CrossRef\]](#)
- [25] Bączyk G, Kozłowska K. The role of demographic and clinical variables in assessing the quality of life of outpatients with rheumatoid arthritis. *Arch Med Sci* 2018;14:1070–1079. [\[CrossRef\]](#)
- [26] Abubakari S, Dizman YA, Karaman F. Integrated meta-analysis identifies keratin family genes and associated genes as key biomarkers and therapeutic targets in metastatic cutaneous melanoma. *Diagnostics (Basel)* 2025;15. [\[CrossRef\]](#)
- [27] Menyhárt O, Györfy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput Struct Biotechnol J* 2021;19:949–960. [\[CrossRef\]](#)
- [28] Bartel DP. Metazoan microRNAs. *Cell* 2018;173:20–51. [\[CrossRef\]](#)
- [29] Monsivais D, Parks SE, Chandrashekar DS, Varambally S, Creighton CJ. Using cancer proteomics data to identify gene candidates for therapeutic targeting. *Oncotarget* 2023;14:399–412. [\[CrossRef\]](#)
- [30] Emwas A-HM, Salek RM, Griffin JL, Merzaban J. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics* 2013;9:1048–1072. [\[CrossRef\]](#)
- [31] Kong W, Hui HWH, Peng H, Goh WWB. Dealing with missing values in proteomics data. *Proteomics* 2022;22:2200092. [\[CrossRef\]](#)
- [32] Huang L, Song M, Shen H, Hong H, Gong P, Deng H-W, et al. Deep learning methods for omics data imputation. *Biology (Basel)* 2023;12:1313. [\[CrossRef\]](#)
- [33] De Brevern AG, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics* 2004;5:1–12.
- [34] Taylor S, Ponzini M, Wilson M, Kim K. Comparison of imputation and imputation-free methods for statistical analysis of mass spectrometry data with missing data. *Brief Bioinform* 2022;23. [\[CrossRef\]](#)
- [35] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;53:457–481. [\[CrossRef\]](#)
- [36] Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol* 1972;34:187–202. [\[CrossRef\]](#)
- [37] Heinze G, Wallisch C, Dunkler D. Variable selection—a review and recommendations for the practicing statistician. *Biom J* 2018;60:431–449. [\[CrossRef\]](#)
- [38] Lee S, Lim H. Review of statistical methods for survival analysis using genomic data. *Genomics Inform* 2019;17. [\[CrossRef\]](#)
- [39] Zhao Z, Zucknick M. Structured penalized regression for drug sensitivity prediction. *J R Stat Soc Ser C Appl Stat* 2020;69:525–545. [\[CrossRef\]](#)
- [40] Zhao Z, Banterle M, Lewin A, Zucknick M. Multivariate Bayesian structured variable selection for pharmacogenomic studies. *J R Stat Soc Ser C Appl Stat* 2024;73:420–443. [\[CrossRef\]](#)
- [41] Xing X, Zhao Z, Liu JS. Controlling false discovery rate using Gaussian mirrors. *J Am Stat Assoc* 2023;118:222–241. [\[CrossRef\]](#)
- [42] Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, et al. A practical guide to methods controlling false discoveries in computational biology. *Genome Biol* 2019;20:1–21. [\[CrossRef\]](#)
- [43] Chung NC, Mirza B, Choi H, Wang J, Wang D, Ping P, et al. Unsupervised classification of multi-omics data during cardiac remodeling using deep learning. *Methods* 2019;166:66–73. [\[CrossRef\]](#)
- [44] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv [Preprint]*. 2018. [\[CrossRef\]](#)
- [45] Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med* 1997;16:385–395.
- [46] Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* 2008;9:1–10. [\[CrossRef\]](#)
- [47] Yang S, Yuan L, Lai Y-C, Shen X, Wonka P, Ye J. Feature grouping and selection over an undirected graph. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2012. p. 922–930. [\[CrossRef\]](#)
- [48] Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. *Ann Stat* 2002;30:74–99. [\[CrossRef\]](#)
- [49] Zhang HH, Lu W. Adaptive lasso for Cox's proportional hazards model. *Biometrika* 2007;94:691–703. [\[CrossRef\]](#)
- [50] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994;81:515–526. [\[CrossRef\]](#)
- [51] Li Y, Wang F, Li R, Sun Y. Semiparametric integrative interaction analysis for non-small-cell lung cancer. *Stat Methods Med Res* 2020;29:2865–2880. [\[CrossRef\]](#)
- [52] Bin RD, Herold T, Boulesteix A-L. Added predictive value of omics data: specific issues related to validation illustrated by two case studies. *BMC Med Res Methodol* 2014;14:1–23. [\[CrossRef\]](#)

- [53] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol* 2013;13:1–15. [\[CrossRef\]](#)
- [54] Rahman MS, Ambler G, Choodari-Oskooei B, Omar RZ. Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Med Res Methodol* 2017;17:1–15. [\[CrossRef\]](#)
- [55] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;247:2543–2546. [\[CrossRef\]](#)
- [56] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei L-J. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;30:1105–1117. [\[CrossRef\]](#)
- [57] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000;56:337–344. [\[CrossRef\]](#)
- [58] Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol* 2010;5:1315–1316. [\[CrossRef\]](#)
- [59] Van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med* 2000;19:3401–3415.
- [60] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 2007;102:359–378. [\[CrossRef\]](#)
- [61] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999;18:2529–2545.
- [62] Portnoy S. Censored regression quantiles. *J Am Stat Assoc* 2003;98:1001–1012. [\[CrossRef\]](#)
- [63] Gerds TA, Kattan MW. Medical risk prediction models: with ties to machine learning. Chapman & Hall/CRC; 2021. [\[CrossRef\]](#)
- [64] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024;74:229–263. [\[CrossRef\]](#)
- [65] Llovet JM, Kelley RK, Villanueva A, Singal AG, Pikarsky E, Roayaie S, et al. Hepatocellular carcinoma. *Nat Rev Dis Primers* 2021;7:1–28. [\[CrossRef\]](#)
- [66] Parikh ND, Mehta AS, Singal AG, Block T, Marrero JA, Lok AS. Biomarkers for the early detection of hepatocellular carcinoma. *Cancer Epidemiol Biomarkers Prev* 2020;29:2495–2503. [\[CrossRef\]](#)
- [67] Gao H-F, Chen L-Y, Cheng C-S, Chen H, Meng Z-Q, Chen Z. SLC5A1 promotes growth and proliferation of pancreatic carcinoma via glucose-dependent AMPK/mTOR signaling. *Cancer Manag Res* 2019;3:171–3185. [\[CrossRef\]](#)
- [68] Yu Q, Gao K. CLEC4M overexpression inhibits progression and is associated with a favorable prognosis in hepatocellular carcinoma. *Mol Med Rep* 2020;22:2245–2252. [\[CrossRef\]](#)
- [69] Luo L, Chen L, Ke K, Zhao B, Wang L, Zhang C, et al. High expression levels of CLEC4M indicate poor prognosis in patients with hepatocellular carcinoma. *Oncol Lett* 2020;19:1711–1720. [\[CrossRef\]](#)
- [70] Nie Y, Jiang M, Liu C, Liu Q, Zhu X. CXCL5 has potential to be a marker for hepatocellular carcinoma prognosis and was correlating with immune infiltrates. *Front Oncol* 2021;11:637023. [\[CrossRef\]](#)
- [71] Kou Y, Yan X, Liu Q, Wei X, Zhang B, Li X, et al. HBV upregulates AP-1 complex subunit mu-1 expression via the JNK pathway to promote proliferation of liver cancer cells. *Oncol Lett* 2019;18:456–464. [\[CrossRef\]](#)
- [72] Yao J, Fu J, Liu Y, Qu W, Wang G, Yan Z. LncRNA CASC9 promotes proliferation, migration and inhibits apoptosis of hepatocellular carcinoma cells by down-regulating miR-424-5p. *Ann Hepatol* 2021;23:100297. [\[CrossRef\]](#)
- [73] Kang HJ, Kim IH, Sung CO, Yu E. Expression of CA9 as a novel prognostic marker in resectable hepatocellular carcinoma. *Pathology* 2014;46. [\[CrossRef\]](#)
- [74] Hyuga S, Wada H, Eguchi H, Otsuru T, Iwagami Y, Yamada D, et al. Expression of carbonic anhydrase IX is associated with poor prognosis through regulation of the epithelial-mesenchymal transition in hepatocellular carcinoma. *Int J Oncol* 2017;51:1179–1190. [\[CrossRef\]](#)

APPENDIX

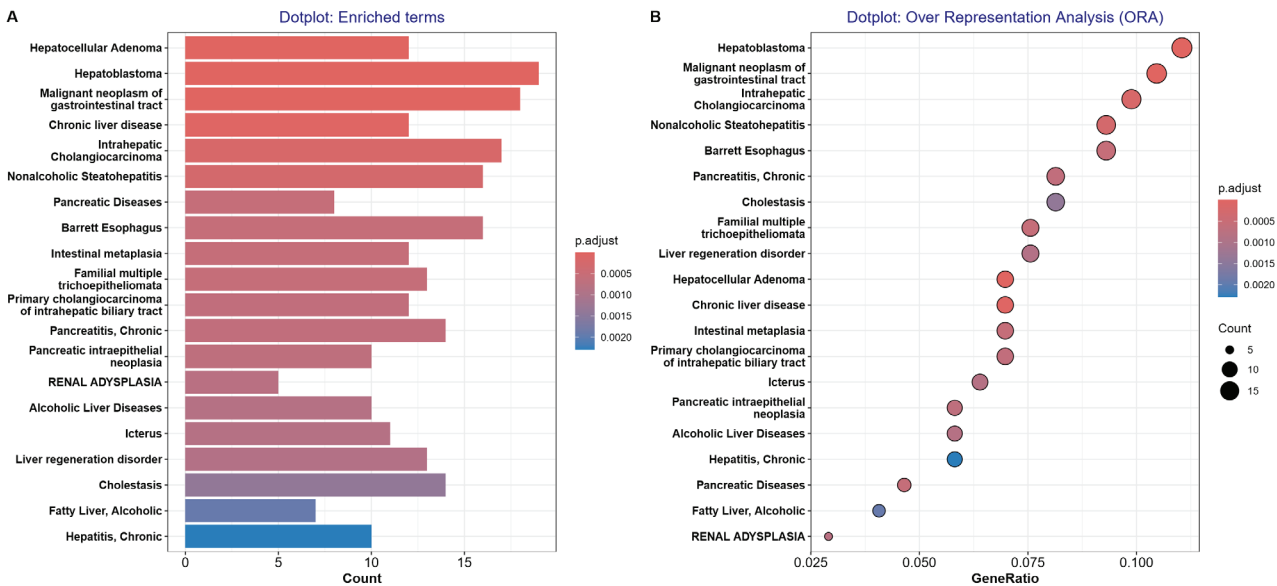


Figure S1. Gene Ontology (GO) revealing enriched terms.

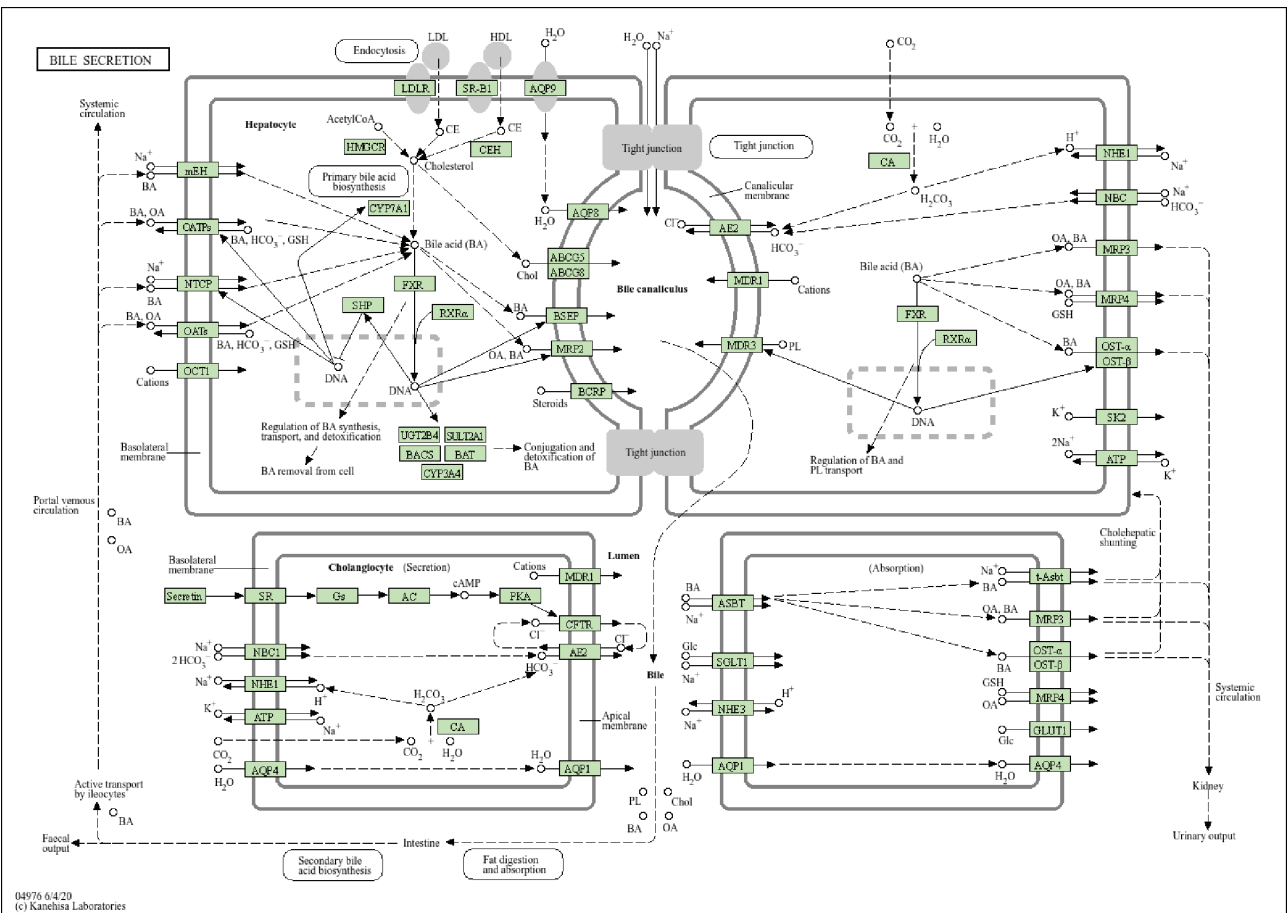


Figure S2. KEGG signaling pathway analysis for hsa04976: Bile Secretion.

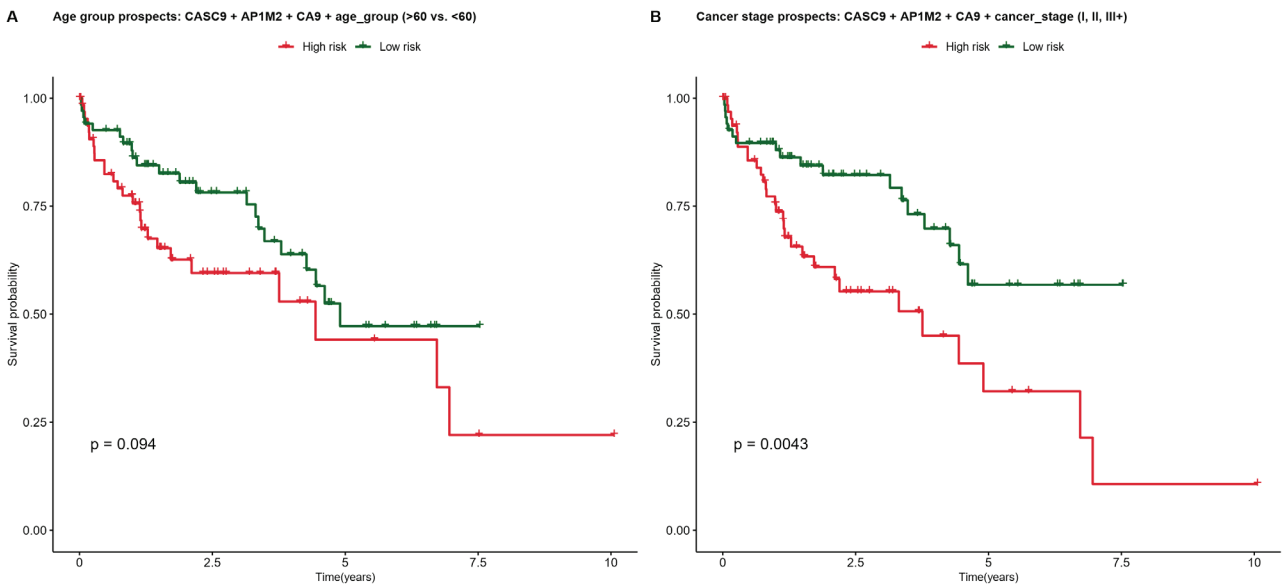


Figure S3. Survival prospects of patients based on stratification by age group and cancer stage, relative to the genes APIMI, CASC9, and CA9.