Comparative Survival Modeling of Low-Grade Glioma Patients Using Targeted Proteomic Data

Ava Berenji
Harvard-Westlake School, Los Angeles, CA

Abstract

Low-grade gliomas (LGG) have long and highly variable survival times, making individualized survival prediction challenging. In this study, we analyzed targeted proteomic data from 422 LGG patients from The Cancer Genome Atlas (TCGA) to evaluate the association between protein expression and overall survival. After preprocessing, 217 proteins were summarized into 18 latent factors, and one representative protein per factor was selected based on univariate Cox regression. Multivariable survival models were then fit, including both the Cox proportional hazards model and parametric approaches, with proteins violating the proportional hazards assumption excluded. Model performance was evaluated using goodness-of-fit metrics and the concordance index. This study aims to determine which protein markers are most strongly associated with survival, and to compare the performance of semi-parametric and parametric models trained on proteomic data in the presence of heavy right-censoring.

Keywords: Survival analysis, Kaplan-Meier curve, Cox proportional hazards model, Weibull survival model, log-logistic model, generalized gamma model, c-index, proteomics, TCGA, Low-grade glioma

1 Introduction

1.1 Background

Gliomas are a diverse group of brain tumors that arise from glial cells, which provide structural and functional support in the central nervous system [1]. They are classified by the World Health Organization (WHO) into grades I-IV based on their aggressiveness [2]. The most aggressive tumor of the spectrum is glioblastoma (grade IV), which mainly affects older adults. In contrast, low-grade gliomas (LGG), classified as WHO grade II and III tumors, mainly affect younger adults and generally grow more slowly and thus tend to have longer survival times [3]. Nevertheless, LGGs remain incurable, and survival times are highly variable [4, 5]. While survival time is measured in years, it differs substantially across patients, so making accurate estimates is essential for planning treatment, including whether, when, and how aggressively to intervene, as well as for long-term monitoring and follow-up [6].

A major advance in LGG classification came with the 2021 WHO classification update, which shifted diagnosis away from histology (observation) toward molecular markers such as genetics. Two alterations are now central to defining the major subtypes: mutation of the IDH gene and codeletion of chromosomal arms 1p and 19q [2]. Tumors with both IDH mutation and 1p/19q codeletion are classified as oligodendroglioma, associated with longer survival and stronger treatment responses [7]. Tumors with IDH mutation but without 1p/19q codeletion are classified as astrocytoma and exhibit more variable outcomes [2]. However, tumors lacking an IDH mutation, even when histologically low-grade, often behave aggressively and have survival outcomes similar to glioblastoma [8].

Although the molecular framework substantially helps stratify LGG prognosis into major types, survival prediction still varies within each subtype, indicating that molecular classification alone is insufficient for individualized survival prediction [9, 10]. Thus, identifying additional biologically informative markers remains an important objective for improving patient-specific estimates of survival time.

Most previous attempts to individually predict LGG survival time have focused on clinical variables and genomic (DNA) or transcriptomic (RNA) features, with relatively limited use of proteomics data [11]. While DNA and RNA describe what a tumor is capable of producing, they do not directly measure which processes are active at a given time. Moreover, after production, proteins may interact with other proteins, or be modified, activated, or inactivated, which determines their final function in the cell. Since proteins directly drive tumor growth and treatment response, measuring protein expression directly may provide a more functionally relevant view of tumor biology that influences patient survival [12].

By combining survival data with protein expression measurements, survival models can be built to examine and estimate how specific proteins are associated with patient outcomes, providing insight into potential prognostic biomarkers which may be relevant for clinical decision-making and personalized therapy strategies.

1.2 History of Proteomics in Cancer Research

Proteomics, the large-scale study of proteins, emerged in the 1970s with the development of two-dimensional gel electrophoresis, allowing researchers to separate and analyze complex mixtures of proteins. The term ”proteome” was later coined in 1994 by Marc Wilkins, a doctoral student at Macquarie University, to describe the complete set of proteins expressed by a genome [13]. However, early gel-based techniques were slow, poorly reproducible, and limited in scale, which made them impractical for large cancer studies. Major advances in the early 2000s, notably in mass spectrometry (MS) and bioinformatics, addressed many of these issues by enabling high-throughput identification, meaning that thousands of proteins could be measured efficiently in a single study, and relatively improved quantification. The Human Proteome Project, formally organized in 2010, aimed to map the human proteome, and many associated initiatives (such as Cancer-HPP) have identified protein expression patterns associated with various cancer types [14, 15]. Despite these advances, reproducibility and accurate quantification still remained a challenge.

In glioma research, early proteomic studies were largely exploratory, focusing on discovering biomarkers rather than formally modeling survival. Many studies compared protein expression in tumor tissue, cerebrospinal fluid, and plasma between tumor and non-tumor samples to find consistent differences [16, 17]. However, these studies were often limited by small sample sizes and measurement accuracy. The development of targeted proteomic platforms, particularly reverse phase protein arrays (RPPA), marked a turning point. By using antibodies to bind to specific proteins, RPPA allows their abundance to be measured more accurately, reproducibly, and across many samples at once. Importantly, RPPA measurements reflect relative protein expression levels that enable comparison across samples, rather than absolute protein concentrations [18]. The platform contributed important data to large, clinically annotated datasets containing survival outcomes such as those generated by The Cancer Genome Atlas (TCGA), making it possible to formally model patient survival using proteomics data.

1.3 Literature Review

Before proteomics data was available, survival prediction in glioma relied primarily on clinical variables such as age, tumor grade, and performance status. These models demonstrated modest predictive ability, typically achieving C-indices ranging from 0.624 to 0.754 [19]. Subsequently, genomic and transcriptomic data were incorporated, capturing more survival-associated patterns. A study by Yuan et al. demonstrated that the addition of molecular data significantly improved survival prediction compared to clinical models alone [19]. However, genomics and transcriptomics measurements only reflect the potential for protein production, rather than the actual abundance or activity of these proteins.

Proteomic data emerged later due to greater technical and cost requirements, but it offered a more direct measurement of cellular function, capturing true protein abundance and post-translational modifications. Several studies found that proteomic data contain further prognostic signals beyond genomics or transcriptomics. For instance, a 2021 study by Yanovich-Arad et al. demonstrated that proteomics reveals additional and sometimes stronger associations with glioblastoma survival than transcriptomic (RNA) data [20].

Despite this potential, proteomics remains the least explored approach to survival prediction in glioma. Most proteomic survival modeling studies have focused on glioblastoma rather than low-grade glioma. This emphasis was partly due to ease of modeling: glioblastoma exhibits shorter survival times, high event rates, and more limited censoring, which simplifies survival model estimation and validation. For example, the PROTGLIO model developed by Stetson et al. used proteomic markers from The Cancer Genome Atlas (TCGA) to predict overall survival in glioblastoma patients, achieving a C-index of 0.82 in training and 0.70 on the testing set [21]. Also using TCGA data, Patil et al. identified proteomic pathways and individual proteins associated with glioblastoma subtypes and survival time, with several protein markers and cell-cycle–related pathways agreeing with those discussed in this study [22]. Thus, proteomic features relevant to survival in glioblastoma may also be informative in lower-grade glioma.

In contrast, low-grade glioma exhibits long and highly variable survival times, leading to heavy right-censoring, where many patients still survive by the end of the study. These characteristics reduce statistical power and complicate model assumptions; thus, low-grade glioma has received less attention in proteomics survival modeling studies. Nevertheless, several studies using TCGA data have identified protein markers associated with low-grade glioma survival. Using univariate and multivariable Cox analysis, Liu et al. associated HIST1H2BK with poor LGG prognosis [23]. Patil et al. identified a set of four proteins (CHK2_pT68, MSH6, ARID1A and PAXILLIN), also using multivariable Cox analysis, that may segregate LGG patients into high and low-risk groups [24].

While the Cox proportional hazards model remains the most common framework for survival analysis due to its flexibility and interpretability, it relies on the assumption that hazard ratios remain constant over time. This proportional hazards assumption is frequently violated in complex cancer datasets, including glioma, particularly in the presence of long survival times [25], harming reliability and interpretability. A recent paper reviewing 39 glioma survival modeling studies reports that only a minority of published nomogram models satisfied the proportional hazards assumption [26], and advocates instead for the use of parametric machine learning in this scenario.

Fully parametric survival models, such as Weibull, log-logistic, and generalized gamma models, offer an alternative by accommodating non-constant hazard ratios and directly estimating baseline hazard. Studies across multiple cancer types have demonstrated that parametric models can achieve predictive performance comparable to, and in some cases exceeding, that of Cox models, even when proportional hazards assumptions hold. For example, Khaksar et al. in non–small cell lung cancer and Teshnizi et al. in leukemia reported improved predictive accuracy using parametric survival models relative to the Cox model [27, 28]. Although parametric models have been explored across many different cancer types, their application to proteomics and glioma datasets remains limited. This study addresses the gaps discussed thus far by comparing both Cox and parametric models trained on TCGA RPPA proteomic data to predict survival in LGG.

1.4 Dataset Description

Proteomic and clinical data for this study were obtained from publicly available datasets generated by The Cancer Genome Atlas (TCGA). Protein expression data were sourced from the TCGA-LGG Level 4 RPPA dataset curated by The Cancer Proteome Atlas (TCPA) [29]. This dataset was accessed via a publicly available GitHub repository associated with the TRGAted platform, which attaches survival endpoints from the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) with protein expression data from the TCPA dataset [12, 30]. The dataset can be found at https://github.com/ncborcherding/TRGAted.

The TCGA-LGG Level 4 RPPA release includes expression measurements for 219 proteins across 426 tumor samples and is provided as fully processed data [12]. Normalization was performed using replicates-based normalization (RBN), which reduces technical variation between batches, allowing protein measurements to be validly compared across patients [31]. Data for each protein were then scaled by TRGAted into z-scores. As discussed in Section 1.2, RPPA measurements reflect relative protein abundance values rather than absolute concentrations[18]. No additional normalization or scaling was applied in this study.

All protein expression measurements were performed on tissue from primary, untreated tumors, as this is a TCGA criterion for studied cancers [32]. Clinical outcome data, including overall survival time and event status, were obtained from the corresponding TCGA clinical files [12]. Overall survival was defined as the time from initial diagnosis to death or last follow-up, with patients who were alive at last follow-up treated as right-censored.

2 Data Preprocessing

The dataset contained expression measurements for 219 proteins across 426 low-grade glioma (LGG) patients. Four patients were excluded due to missing or invalid survival time, leaving 422 patients for analysis. We restricted preprocessing to the protein expression columns. Two proteins, *Alpha-catenin* and *PARP1*, had substantial missing data. Although biologically relevant, their functions were adequately represented by other proteins with complete data, so they were removed. After this step, all 217 remaining proteins had complete measurements. All subsequent analyses (dimension reduction, feature screening, and survival modeling) were performed on this cleaned dataset.

3 Methods

3.1 Feature construction

To reduce dimensionality in the protein expression data while preserving biologically meaningful structure, factor analysis was applied to the 217 proteins using the `fa()` function in the `psych` package. The extraction method was minimum residuals (minres), and factors were rotated using oblimin rotation, which is appropriate when latent factors are expected to be correlated in biological systems. The number of retained factors was determined by parallel analysis, which suggested that 18 factors were significant (Figure 1).

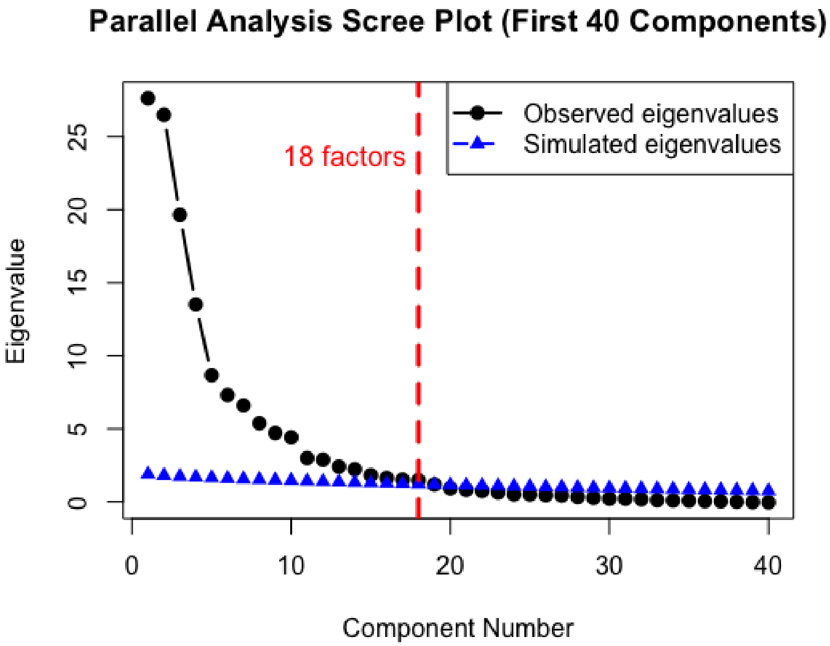

Figure 1: Zoomed parallel analysis scree plot (components 1–40). The observed eigenvalues (solid line) exceed the simulated eigenvalues (dashed line) through the 18th component, indicating that 18 factors should be retained.

From each factor, we selected the five proteins with the largest squared loadings. Each candidate protein was then evaluated for association with overall survival using a univariate Cox proportional hazards model. From each factor's candidate set, the protein with the smallest Cox regression p-value was chosen as the representative predictor. If a protein appeared as the top candidate in multiple factors, it was retained for the factor where it had the strongest loading, and the remaining factor was then assigned the next-best protein by p-value. This ensured that each factor contributed a unique representative protein to subsequent analyses.

Four selected proteins (Annexin A7, YAP, BAK, EIF4G) showed significant violations of the proportional hazards assumption, indicating that their effects on survival were not constant over time, and were therefore excluded from Cox modeling. When excluded across all models, predictive performance was not adversely affected, with negligible changes in C-index. The final predictor set therefore excluded these proteins to ensure Cox model validity and to maintain a consistent feature set across all models for future comparison.

We added two demographic variables, age and gender, as additional predictors, with race excluded because group sizes were too small for reliable modeling. Thus, the final predictor set contained 16 predictors. Age was reported in 10-year ranges in the TCGA clinical data. For modeling purposes, each age range was mapped to its midpoint and treated as a continuous predictor. Gender was coded as a categorical (factor) variable.

3.2 Backward Elimination

Backward elimination was used to obtain reduced, interpretable models for the Cox, Weibull, log-logistic, and generalized gamma models. Random Survival Forests were excluded because they do not rely on statistical significance testing or parametric coefficient estimates.

Starting from the full set of 16 predictors, backward elimination was performed independently for each model. Only protein predictors were eligible for removal, whereas age and gender were retained in all models as baseline demographics.

At each iteration, the predictor with the largest Wald test p-value was removed if its p-value exceeded $\alpha = 0.05$. The model was then refit, and this procedure was repeated until all remaining protein predictors met the significance threshold. Reduced models were fit separately for each model family, resulting in slightly different predictor sets across models.

3.3 Model Fitting and Evaluation

All survival models were fit in R using the survival, flexsurv, and randomForestSRC packages. We compared five modeling approaches: the Cox proportional hazards model, Weibull, log-logistic, generalized gamma, and Random Survival Forests.

The Cox proportional hazards model was estimated using the coxph() function. Parametric survival models (Weibull, log-logistic, and generalized gamma) were fit using flexsurvreg. Random Survival Forests were trained using an ensemble of survival trees.

Model performance was evaluated using both goodness-of-fit and predictive discrimination. For parametric models, goodness-of-fit was assessed using the Akaike Information Criterion (AIC), the corrected AIC (AICc), and the Bayesian Information Criterion (BIC), with lower values indicating better fit. These criteria were not reported for the Cox model, which does not specify a full likelihood, or for Random Survival Forests, which are nonparametric.

For all models, predictive performance was assessed using the concordance index (C-index), which measures how well a model ranks patients by survival time, while accounting for censoring.

4 Theoretical Framework of Survival Analysis

4.1 Kaplan–Meier Estimator

In survival studies, the central object of interest is the survival function,

$$S(t) = \mathbb{P}(T > t),$$

which gives the probability that the time-to-event variable T exceeds t. Unlike many standard statistical settings, survival data are often subject to censoring: for some individuals, we only know that the event has not yet occurred by a certain time, but not the exact failure time. Right-censoring of this kind is most common in practice.

A widely used nonparametric approach for estimating $S(t)$ under censoring is the Kaplan–Meier (KM) estimator. Suppose the ordered distinct event times are $t_1 < t_2 < \cdots < t_k$, where d_j events occur at t_j and n_j individuals remain at risk immediately before t_j. The estimator is defined as

$$\widehat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right).$$

This product form reflects the conditional probability of surviving past each event time, and results in a step function that decreases only at observed event times. The plot of $\widehat{S}(t)$ is known as the Kaplan–Meier curve, and it remains the standard descriptive tool for visualizing survival experiences across groups.

4.2 Random Censoring Model

In survival analysis, observed data often consist of event times that may be censored. To formally handle this, one commonly adopts the random censoring framework. Suppose we observe n independent pairs (t_i, δ_i), where

$$t_i = \min(T_i, C_i), \quad \delta_i = \mathbf{1}\{T_i \le C_i\},$$

with T_i denoting the true survival time, C_i the censoring time, and δ_i an indicator of whether the event is observed. Assuming independence between T_i and C_i, the likelihood contribution of the i-th subject is

$$L_i = \big[f(t_i)\big]^{\delta_i} \big[S(t_i)\big]^{1-\delta_i},$$

where $f(t)$ denotes the density and $S(t) = 1 - F(t)$ the survival function of T. This expression omits terms involving the distribution of the censoring time. Under the assumption of independent censoring, these terms factorize and do not involve the parameters governing the survival distribution, and therefore can be ignored for inference on T. For a sample of n independent observations, the likelihood is

$$L_p = \prod_{i=1}^{n} L_i,$$

and model parameters are estimated by maximizing L_p under a chosen parametric or semiparametric specification of f and S.

4.3 Cox Proportional Hazards Model

The Cox model provides a semiparametric approach that avoids specifying the baseline survival distribution. It assumes proportional hazards, meaning that covariate effects act multiplicatively on the hazard function and that hazard ratios between individuals are constant over time. Violations of this assumption can lead to biased or unstable estimates of coefficients.

For a subject with covariates $\mathbf{X}$, the survival function is

$$S(t \mid \mathbf{X}) = \big[S_0(t)\big]^{\exp(\boldsymbol{\beta}'\mathbf{X})},$$

where $S_0(t)$ is an unspecified baseline survival function. Inference on $\boldsymbol{\beta}$ is based on the partial likelihood that eliminates the unspecified baseline hazard:

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\mathbf{X}_i)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{X}_j)} \right]^{\delta_i},$$

with $R(t_i)$ the risk set at time t_i. This allows estimation of relative covariate effects without requiring a parametric assumption for $S_0(t)$, which can then be estimated in a second step.

4.4 Weibull Model

The Weibull model is a fully parametric approach that allows the hazard rate to vary over time in a flexible manner. With a set of covariates $x_1, \ldots, x_k$, the survival and density functions are given by

$$S(t) = \exp(-\lambda t^{\gamma})\,, \quad f(t) = \gamma \lambda t^{\gamma-1} \exp(-\lambda t^{\gamma})\,,$$

where $\lambda = \exp\{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)\}$ is the scale parameter and $\gamma > 0$ is the shape parameter. Parameter estimation is carried out by maximizing the full likelihood function.

4.5 Log-Logistic Model

Another alternative is the log-logistic distribution, which allows hazards that rise and then fall over time. The density and survival functions are

$$f(t) = \frac{(\gamma/\lambda_i)(t/\lambda)^{\gamma-1}}{\left[1 + (t/\lambda)^{\gamma}\right]^2}, \quad S(t) = \frac{1}{1 + (t/\lambda)^{\gamma}},$$

with $\lambda = \exp\{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)\}$ and $\gamma > 0$. Model parameters are estimated by maximizing the full likelihood function. The Log-logistic model has the proportional odds property. Indeed, the odds for experiencing the event before time t are

$$\frac{1 - S(t)}{S(t)} \;=\; (t/\lambda)^{\gamma} \;=\; t^{\gamma}\,\lambda^{-\gamma} \;=\; t^{\gamma}\,\exp\{\gamma(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)\}.$$

Hence, for two individuals, the ratio of these odds is independent of time, and thus, the odds are proportional. The estimated odds have the form:

$$\frac{1 - \hat{S}(t)}{\hat{S}(t)} \;=\; t^{\hat{\gamma}}\,\exp\{\hat{\gamma}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)\}.$$

The estimated regression slopes yield the following interpretation. If x_1 is a continuous variable, then a one-unit increase in x_1 results in a percent change in the estimated odds computed as

$$\frac{\text{odds}|_{x_1+1} - \text{odds}|_{x_1}}{\text{odds}|_{x_1}} \cdot 100\% \;=\; \left(e^{\hat{\gamma}\hat{\beta}_1} - 1\right) \cdot 100\%.$$

4.6 Generalized Gamma Model

The generalized gamma model encompasses a wide family of distributions, including the exponential, Weibull, and log-normal as special cases. Its density is

$$f(t) = \frac{\lambda p}{\Gamma(d)} \big(\lambda t\big)^{dp-1} \exp\big\{ -(\lambda t)^p \big\}, \quad t > 0,$$

where $\lambda > 0$ is a scale parameter, $p > 0$ a shape parameter, and $d > 0$ a family parameter. Covariates are introduced by modeling λ as a log-linear function of predictors. The survival function is

$$S(t) = 1 - \frac{\gamma\big(d, (\lambda t)^p\big)}{\Gamma(d)},$$

where $\gamma(\cdot, \cdot)$ is the incomplete gamma function. The parameters of this model are estimated from data by the full-likelihood maximization method.

4.7 Model Performance Metrics

When comparing statistical models, relying only on fit measures like the log-likelihood can be misleading. Models with more parameters often fit the data better, but this may result in overfitting. To address this, information criteria such as the Akaike Information Criterion (AIC), the corrected AIC (AICc), and the Bayesian Information Criterion (BIC) are commonly used, as they balance model fit with complexity. For each of these criteria, a lower value corresponds to a better-fitting model.

In survival analysis, another widely used tool is the concordance index (C-index), which evaluates a model's ability to discriminate between outcomes. The C-index is especially useful when the likelihood function is not available, as is often the case with machine learning methods. We now describe each of these criteria in more detail.

4.7.1 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) was introduced by Hirotugu Akaike in 1973. It measures the relative quality of a statistical model by approximating the difference between the true data-generating process and the fitted model. The AIC is defined as

$$\text{AIC} = -2\,\ell(\hat{\theta}) + 2k,$$

where $\ell(\hat{\theta})$ is the maximized log-likelihood of the model, and k is the number of estimated parameters. The first term rewards model fit, while the second penalizes complexity.

4.7.2 Corrected Akaike Information Criterion (AICc)

When the sample size n is small relative to the number of parameters k, the AIC can underestimate the risk of overfitting. The AICc introduces a correction for finite samples (Sugiura, 1978; Hurvich & Tsai, 1989):

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1}.$$

As $n \to \infty$, AICc converges to AIC, but for small n, AICc is more reliable.

4.7.3 Bayesian Information Criterion (BIC)

The BIC (Schwarz, 1978) comes from a Bayesian perspective, approximating the marginal likelihood of a model under certain regularity conditions. It is given by:

$$\text{BIC} = -2\ell(\hat{\theta}) + k\ell(n).$$

Compared to AIC, the penalty term grows with $\ell(n)$, making BIC more conservative: it tends to favor simpler models, especially as the sample size increases.

The Concordance Index (C-index)

While AIC, AICc, and BIC are measures of model fit that balance likelihood and complexity, they do not directly assess a model's predictive discrimination. In survival analysis, the **concordance index (C-index)** is widely used. The C-index measures the proportion of all pairs of subjects in which the predictions and outcomes are *concordant*. For survival data, it is defined as:

$$C = \frac{\text{Number of concordant pairs}}{\text{Number of compared pairs}}.$$

Intuitively, it is the probability that, for two randomly selected individuals, the one with the higher predicted risk (or shorter survival time) actually experiences the event earlier. The C-index ranges from 0.5 (no better than chance) to 1 (perfect concordance).

5 Applications and Results

5.1 Descriptive Survival Analysis

After preprocessing, 422 patients remained in the dataset (four patients were excluded due to missing or invalid survival times). A total of 217 proteins were included, following the exclusion of two proteins (α-catenin and PARP1) with excessive missingness. To characterize overall survival in the cohort prior to modeling, a Kaplan–Meier curve was generated.

The Kaplan–Meier curve shows a gradual decline in survival probability over time, with a long right tail extending beyond 6000 days, consistent with the long and variable survival time of LGG. During follow-up, 97 deaths were observed, and 325 (77%) patients were censored, indicating substantial right-censoring. Median overall survival was approximately 2660 days.

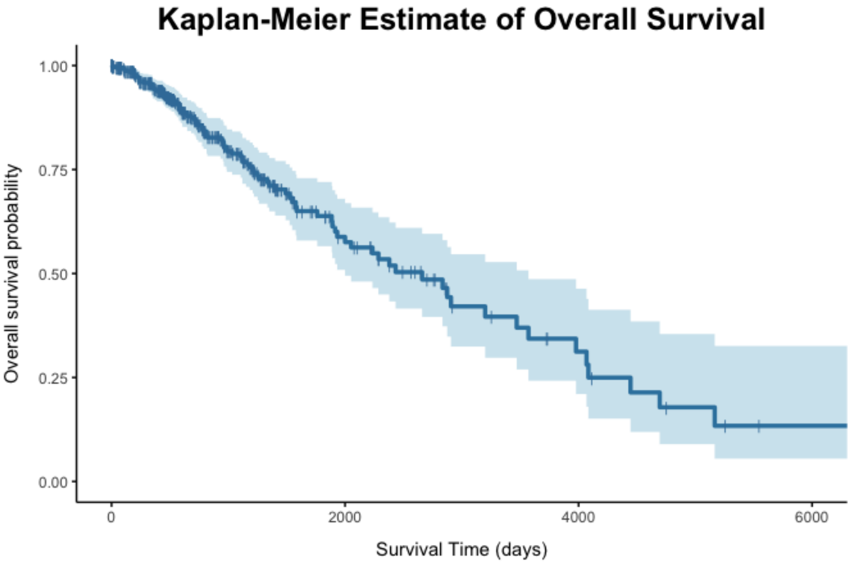

Figure 2: Kaplan–Meier estimate of overall survival for the 422 LGG patients. The solid line represents the estimated survival function, and the ribbons indicate the pointwise confidence intervals.

5.2 Univariate Cox and Full Model Results

Univariate Cox proportional hazards regression was performed for each of the 18 protein predictors and 2 demographic variables, as described in Section 3.1, to assess individual association with overall survival prior to multivariable modeling. In this model, hazard ratios (HR) greater than 1 indicate increased hazard and shorter survival, whereas HR values less than 1 indicate reduced hazard and longer survival.

Several proteins showed statistically significant associations with overall survival, including *ARAF_pS299*, STAT5A, *HER2_pY1248*. Age was the most significant predictor in the univariate model. Full univariate Cox regression results for all 20 covariates are reported in Appendix Table A1.

Although some proteins, such as Annexin A7, were highly significant in univariate analysis, they were excluded from subsequent multivariable Cox models due to violations of the proportional hazards assumption, as discussed in Section 3.1.

In univariate analysis, protein effects were observed in both directions. For example, *PKCPANBETAII_pS660* was associated with reduced hazard (HR = 0.74, $p < 0.0053$), corresponding to longer survival, whereas STAT5A (HR = 1.67, $p < 0.001$) was associated with increased hazard and shorter survival. In the multivariable Cox model, one protein changed hazard ratio direction: *FIBRONECTIN*, which changed HR from 1.23 to 0.80. Statistical significance often differed by several orders of magnitude between models, and even occasionally flipped. Univariate and full multivariable Cox regression results are compared in Appendix Table A2, and full multivariable outputs may be found in this paper's Github repository. Due to interpretability, we will focus on reduced multivariable models obtained via backward elimination.

5.3 Backward Elimination and Reduced Models

Backward elimination was performed separately for each model, using a Wald test threshold of $p \leq 0.05$. Each model was initialized with the same set of 16 predictors (14 proteins, along with age and gender). The demographic variables were retained in all models, while protein predictors could be eliminated.

Performance metrics for the reduced models are summarized in Table 1. To assess goodness-of-fit, we report AIC, AICc, and BIC for parametric models only, as semiparametric models such as Cox do not yield comparable metrics. To assess discriminative ability, we report C-index for all models.

Model	AIC	AICc	BIC	C-index
Cox PH	—	—	—	0.844
Weibull	1694.83	1695.36	1743.37	0.844
Log-logistic	1694.34	1694.87	1742.88	0.852
Generalized gamma	1698.51	1698.95	1747.05	0.844

Table 1: Comparison of reduced survival models after backward elimination. Lower AIC, AICc, and BIC indicate better model fit, while higher C-index indicates better discriminative ability.

Across the reduced models, model fit and concordance were strong and very similar, with C-index values clustered within a narrow range (0.844–0.852). Among parametric approaches, the log-logistic model achieved the highest C-index and marginally lower information criteria, although these differences were small and should not be over-interpreted. In terms of AIC, the Weibull and log-logistic models were essentially identical, whereas the generalized gamma model showed slightly higher AIC (by about 4 units), indicating slightly worse fit. Given that no single model family clearly outperformed the others, we discuss results comparatively across all models. Full coefficient tables for all reduced models, along with model-fitting code, are provided in the accompanying GitHub repository.

Variable	**Cox**	**Weibull**	**Log-logistic**	**Gen. gamma**
BRD4	1.41^{**}	0.83^{**}	0.85^{*}	-0.19^{**}
CHK2_pT68	0.71^{*}	1.21^{*}	—	0.20^{*}
CYCLINB1	1.33^{**}	0.87^{**}	0.79^{***}	-0.16^{**}
CYCLINE1	0.74^{*}	1.18^{*}	1.24^{***}	0.19^{**}
FIBRONECTIN	—	—	1.28^{**}	—
HER2_pY1248	1.17^{*}	0.92^{*}	—	—
PAI1	1.27^{**}	0.87^{**}	0.83^{**}	-0.15^{**}
PKCPANBETAII_pS660	—	—	1.26^{*}	—
STAT5ALPHA	1.35^{**}	0.85^{**}	0.81^{***}	-0.18^{**}
X1433ZETA	1.45^{**}	0.82^{**}	0.81^{**}	-0.19^{**}
age	1.05^{***}	0.97^{***}	0.97^{***}	-0.03^{***}
gender (male)	0.93	1.02	0.98	-0.01

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 2: Comparison of effect estimates across reduced survival models. HR is reported for Cox model; HR > 1 indicates shorter survival (Cox). ETR is reported for Weibull and Log-logistic (AFT models); ETR > 1 indicates longer survival (AFT). Coefficients are reported for the generalized gamma model; positive coefficients indicate longer survival.

Across all reduced models, protein predictors exhibited consistent effect direction, with no protein reversing direction between model families. With few exceptions, predictors showed similar statistical significance across models. Most proteins (e.g. BRD4, Cyclin B1, PAI-1, STAT5A, and 14-3-3ζ) were statistically significant in all model families. Some proteins, such as FIBRONECTIN and PKCPANBETAII_pS660, were retained in only a subset of models. Age was highly significant in every model considered.

6 Discussion

In this study, we investigated whether targeted proteomic data can be used to model overall survival in patients with low-grade glioma, and whether performance and key predictors vary across model types. Across all reduced models, performance was consistently strong, with concordance indices clustered within a narrow range (0.84 - 0.86), and similarly close AICs. This suggests that the prognostic signals captured were driven by predictor effects, rather than distribution assumptions or heavy right-censoring. Due to similar performance, we discuss all reduced models comparatively rather than focusing on a single one.

6.1 Model Comparison

The similarity in performance across models likely reflects data properties rather than similarities between models. The LGG dataset contained long and highly variable survival times with substantial right-censoring, which makes it unlikely for a single parametric distribution to fit the data closely.

Among parametric models, the log-logistic model achieved slightly better metrics than the Weibull and generalized gamma models. Although these differences were small and should not be over-interpreted, they are consistent with the biological heterogeneity of low–grade glioma. The log-logistic distribution permits non-monotonic hazard behavior, which may plausibly reflect long periods of disease stability in some patients followed by eventual progression in others–a pattern commonly observed in LGG clinical courses. However, given the minimal performance differences, these results do not support a definitive preference for one classical model over another. Rather than indicating a particular baseline hazard shape, the similar performance across models may be explained by strong covariate effects.

Overall, these findings suggest that, among classical survival models, model choice is less critical than feature construction and biological signal quality. When predictive performance is comparable, simpler, more interpretable models may be selected for further analysis and clinical application.

6.2 Interpretation of Model Output

Across all reduced models in which a predictor was retained, its estimated effects on survival time consistently pointed in the same direction. Proteins associated with shorter survival in one model were never associated with longer survival in another, and vice versa, reinforcing that the underlying biological signal captured by these predictors is more important than choice of model or survival time distribution, as discussed earlier.

However, effect sizes and statistical significance varied by model, and between univariate and multivariable Cox models. FIBRONECTIN reversed direction between univariate and multivariable Cox models, which may reflect correlation with other proteins with related function in the training set. Such reversals highlight the importance of multivariable modeling in high-dimensional proteomic studies; since proteins often work together in the cell, univariate results alone may be misleading.

Age remained the strongest predictor in every model, with older age consistently associated with shorter survival, aligning with its established associations in literature [33, 34]. Gender was not a significant predictor in any model, consistent with prior studies [35].

While the sets of predictors retained after backward elimination were not identical across models, these differences likely reflect the choice of significance thresholds or collinearity among data, rather than contradictory conclusions about protein significance. Importantly, in models where a protein was included, its estimated effect direction on survival time stayed consistent. Thus, while certain models may yield more uncertain estimates for the effect of a protein, the general biological signal appears consistently estimated.

Different models may identify different statistically significant predictors, but several proteins were retained in all four models and thus are analyzed in further detail: BRD4, Cyclin B1, Cyclin E1, PAI-1, STAT5A, and 14-3-3-ζ. Predictors that are consistently directionally aligned but intermittently significant may still represent meaningful biological contributors whose effects are partially masked by collinearity, limited events, or shared pathway structure captured by factor analysis.

6.3 Biological Interpretation of Key Proteins

Several proteins identified as strong predictors of survival have established roles in pathways relevant to Low-Grade Glioma, especially cell-cycle regulation, transcription, and signaling.

Across Cox and parametric models, BRD4 was associated with poorer survival. By binding to acetylated chromatin, BRD4 facilitates transcription. In locations oversaturated with BRD4, deemed "Super-Enhancers", the protein drives rapid transcription of oncogenes, which drive tumors to arise and grow

[36].

Cell cycle regulators Cyclin E1 and Cyclin B1 were also retained in all four models. Cyclin E1, which regulates the G1/S phase transition of the cell cycle, was consistently associated with longer survival in our study. However, in literature, Cyclin E1 is an oncogene whose elevation leads to increased DNA replication and tumor growth. It is unlikely this discrepancy was due to collinearity in the predictor set, as the univariate Cox model confirmed the multivariable direction. However, Cyclin E1 is an important therapeutic target, especially to overcome temozolomide (TMZ) resistance [37]. Since the patients in the TCGA-LGG cohort received treatment, elevated Cyclin E1 levels may make tumor cells more vulnerable to treatment due to replication stress. A study by Yu et al. (2025) found that overexpression of Cyclin E1 was associated with better response to immunotherapy in breast cancer, because it causes genomic instability through replication stress [38]. In advanced-stage ovarian cancer, a 2013 study found that high CCNE1 gene expression (which produces Cyclin E1) is actually associated with prolonged survival time, for a similar reason [39]. Cyclin B1 regulates the G2/M transition and overexpression can cause rapid mitosis and metastasis in cancer [40]. Consistent with the literature, Cyclin B1 was associated with shorter survival across all models in our study.

STAT5A was associated with poorer survival across all models. In general, it has a complex role, being associated with favorable survival in certain cancers while acting as an oncogene in others. In glioma, it is linked to poorer prognosis, agreeing with our findings. Its function is to inhibit apoptosis, delaying programmed cell death and thus driving aggressive cancer growth and progression [41]. A study by Ji et al. (2021) used Cox regression analysis on TCGA and CGGA databases, finding that STAT5A was associated with poorer prognosis in glioma [42].

CHK2_pT68 was associated with longer survival in the Cox, Weibull, and generalized gamma models, whereas it was not included in the reduced log-logistic model. In cancer, it detects damaged DNA and activates Checkpoint kinase 2 (Chk2), stopping the cell cycle. Thus, it acts as a tumor suppressor [43]. Patil et al. (2018) confirmed these results, finding that CHK2_pT68 was protective using Cox models trained on TCGA-LGG data [24].

PAI-1 was associated with poorer survival across all models, agreeing with literature. High levels are linked to increased angiogenesis and tumor detachment, both driving tumor growth and metastasis [44]. The gene SERPINE1, which produces the protein PAI-1, was shown to be associated with poor prognosis in LGG via Cox regression analysis on the TCGA dataset [45]. However, proteomic studies on PAI-1 directly are more rare, especially for Low-Grade Glioma.

14-3-3ζ was associated with poorer survival across all models, agreeing with literature. Its primary role in cancer is to bind to pro-apoptotic proteins, inhibiting their function and thus promoting tumor growth and chemotherapy resistance. Additionally, 14-3-3ζ protects proteins which drive Epithelial-Mesenchymal Transition (EMT), which is a key process leading to metastasis [46].

6.4 Limitations

Several limitations of this study should be acknowledged. First, the TCGA LGG cohort exhibits heavy right-censoring, which increases uncertainty in long-term survival estimates and limits inference in the tail of the survival distribution. Although this does reflect the real clinical course of LGG, it may reduce the performance of classical survival models at longer follow-up times, especially given an already small sample size.

Second, four proteins were excluded from all multivariable models due to violations of the proportional hazards assumption. Although this decision ensured model validity and comparability, it may have removed biologically meaningful predictors whose effects vary over time.

Third, although factor analysis reduces dimensionality, it does not guarantee the predictor set will completely eliminate collinearity. Residual correlation may remain, particularly in biologically interconnected pathways.

Fourth, external validation was not performed. All models were trained and evaluated on a single TCGA dataset; in clinical applications, independent validation on external LGG datasets is necessary. Still, the models provide useful comparative and biological insights.

Finally, RPPA measures a targeted subset of proteins rather than the full proteome. This limits the biological signal captured by factor analysis, as proteins relevant to a survival-related pathway may be missing in the dataset, and prevents discovery of new associations.

6.5 Future Research

As data availability improves, combining proteomics with genomics and/or transcriptomics in a multi-omics analysis can provide deeper insights. However, collinearity and redundancy would need to be carefully handled in these approaches due to the high-dimensional data.

More flexible modeling strategies such as machine learning or deep learning approaches, may also be explored. These models may better account for the long and highly variable survival times, as well as nonlinear associations between proteins. Nevertheless, their use raises challenges in interpretability and overfitting, which must be balanced with model complexity.

Supplemental Materials

The complete dataset and all R code used in this study are available in the GitHub repository at https://github.com/ab2028/lgg-proteomic-survival. Readers are encouraged to access the repository to review and reproduce the analyses.

Acknowledgments

I would like to thank Dr. Olga Korosteleva, Professor of Statistics at CSULB, for her exceptional mentorship and support throughout this research, and Dr. Oleg Gleizer at UCLA Math Circle for introducing me to this research opportunity. My gratitude also extends to Mr. Kevin Weis, my Linear Algebra teacher, Ms. Heather Audesirk, my Honors Chemistry teacher, and Mr. Juan Lopez, my Computer Science teacher, whose engaging instruction and dedication helped me build the essential foundations for my research, as well as my parents, for their constant support, love, and encouragement.

References

[1] Youssef, Gilbert, and Miller, Julie J. "Lower Grade Gliomas". In: *Current Neurology and Neuroscience Reports* 20.7 (2020), p. 21. DOI: 10.1007/s11910-020-01040-8.

[2] Louis, David N. et al. "The 2021 WHO Classification of Tumors of the Central Nervous System: a summary". In: *Neuro-Oncology* 23.8 (2021), pp. 1231–1251. DOI: 10.1093/neuonc/noab106.

[3] Ostrom, Quinn T. et al. "CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013–2017". In: *Neuro-Oncology* 22.Supplement 1 (2020), pp. iv1–iv96. DOI: 10.1093/neuonc/noaa200.

[4] Claus, Elizabeth B. et al. "Survival and low grade glioma: the emergence of genetic information". In: *Neurosurgical Focus* 38.1 (2015), E6. DOI: 10.3171/2014.10.FOCUS12367.

[5] Karabacak, Mert et al. "Prognosis Individualized: Survival predictions for WHO grade II and III gliomas with a machine learning-based web application". In: *npj Digital Medicine* 6 (2023), p. 200. DOI: 10.1038/s41746-023-00932-7.

[6] Chen, Shaohua et al. "Prediction of Survival Outcome in Lower-Grade Glioma Using a Prognostic Signature with 33 Immune-Related Gene Pairs". In: *International Journal of General Medicine* 14 (2021), pp. 8149–8160. DOI: 10.2147/IJGM.S338135.

[7] Antonelli, Manila, and Poliani, Pietro Luigi. "Adult type diffuse gliomas in the new 2021 WHO Classification". In: *Pathologica* 114.6 (2022), pp. 397–409. DOI: 10.32074/1591-951X-823.

[8] Torp, Sverre Helge, Solheim, Ole, and Skjulsvik, Anne Jarstein. "The WHO 2021 Classification of Central Nervous System tumours: a practical update on what neurosurgeons need to know—a minireview". In: *Acta Neurochirurgica* 164.9 (2022), pp. 2453–2464. DOI: 10.1007/s00701-022-05301-y.

[9] Gittleman, Haley, Sloan, Andrew E., and Barnholtz-Sloan, Jill S. "An independently validated survival nomogram for lower-grade glioma". In: *Neuro-Oncology* 22.5 (2019), pp. 665–674. DOI: 10.1093/neuonc/noz191.

[10] Poulen, Gaëtan et al. "Huge heterogeneity in survival in a subset of adult patients with resected, wild-type isocitrate dehydrogenase status, WHO grade II astrocytomas". In: *Journal of Neurosurgery* 130.4 (2019), pp. 1289–1298. DOI: 10.3171/2017.10.JNS171825.

[11] Wong, Derek et al. "Integrated proteomic analysis of low-grade gliomas reveals contributions of 1p–19q co-deletion to oligodendroglioma". In: *Acta Neuropathologica Communications* 10 (2022), p. 70. DOI: 10.1186/s40478-022-01372-1.

[12] Borcherding, Nicholas et al. "TRGAted: A web tool for survival analysis using protein data in The Cancer Genome Atlas". In: *F1000Research* 7 (2018), p. 1235. DOI: 10.12688/f1000research.15789.2.

[13] Suran, M. "After the Genome—A Brief History of Proteomics". In: *JAMA* 328.12 (2022), pp. 1168–1169. DOI: 10.1001/jama.2022.7448.

[14] Legrain, Pierre et al. "The Human Proteome Project: Current State and Future Direction". In: *Molecular & Cellular Proteomics* 10.7 (2011), p. M111.009993. DOI: 10.1074/mcp.M111.009993.

[15] Jimenez, Connie R. et al. “The Cancer Proteomic Landscape and the HUPO Cancer Proteome Project”. In: *Clinical Proteomics* 15 (2018), p. 4. DOI: `10.1186/s12014-018-9180-6`.

[16] Gautam, Poonam et al. “Proteins with Altered Levels in Plasma from Glioblastoma Patients as Revealed by iTRAQ-Based Quantitative Proteomic Analysis”. In: *PLoS ONE* 7.9 (2012), e46153. DOI: `10.1371/journal.pone.0046153`.

[17] Kalinina, J. et al. “Proteomics of gliomas: Initial biomarker discovery and evolution of technology”. In: *Neuro-Oncology* 13.9 (2011), pp. 926–942. DOI: `10.1093/neuonc/nor078`.

[18] The RPPA Society et al. “Realizing the Promise of Reverse Phase Protein Arrays for Clinical, Translational, and Basic Research: A Workshop Report”. In: *Molecular & Cellular Proteomics* 13.7 (2014), pp. 1625–1643. DOI: `10.1074/mcp.O113.034918`.

[19] Yuan, Yuan et al. “Assessing the clinical utility of cancer genomic and proteomic data across tumor types”. In: *Nature Biotechnology* 32.7 (2014), pp. 644–652. DOI: `10.1038/nbt.2940`.

[20] Yanovich-Arad, Gali et al. “Proteogenomics of glioblastoma associates molecular patterns with survival”. In: *Cell Reports* 34.9 (2021), p. 108787. DOI: `10.1016/j.celrep.2021.108787`.

[21] Stetson, L. C., Dazard, J.-E., and Barnholtz-Sloan, J. S. “Protein markers predict survival in glioma patients”. In: *Molecular and Cellular Proteomics* 15.7 (2016), pp. 2356–2365. DOI: `10.1074/mcp.M116.060657`.

[22] Patil, Vikas, and Mahalingam, Kulandaivelu. “Comprehensive analysis of Reverse Phase Protein Array data reveals characteristic unique proteomic signatures for glioblastoma subtypes”. In: *Gene* 685 (2019), pp. 85–95. DOI: `10.1016/j.gene.2018.10.069`.

[23] Liu, Weidong et al. “High Levels of HIST1H2BK in Low-Grade Glioma Predicts Poor Prognosis: A Study Using CGGA and TCGA Data”. In: *Frontiers in Oncology* 10 (2020), p. 627. DOI: `10.3389/fonc.2020.00627`.

[24] Patil, Vikas, and Mahalingam, Kulandaivelu. “A four-protein expression prognostic signature predicts clinical outcome of lower-grade glioma”. In: *Gene* 679 (2018), pp. 57–64. DOI: `10.1016/j.gene.2018.08.001`.

[25] Putter, Hein et al. “Long-term survival with non-proportional hazards: results from the Dutch Gastric Cancer Trial”. In: *Statistics in Medicine* 24 (2005), pp. 2807–2821. DOI: `10.1002/sim.2143`.

[26] Xue, Jihao et al. “Limitations of nomogram models in predicting survival outcomes for glioma patients”. In: *Frontiers in Immunology* 16 (2025), p. 1547506. DOI: `10.3389/fimmu.2025.1547506`.

[27] Khaksar, Elahe et al. “Cox Regression and Parametric Models: Comparison of How They Determine Factors Influencing Survival of Patients with Non-Small Cell Lung Carcinoma”. In: *Asian Pacific Journal of Cancer Prevention* 18.12 (2017), pp. 3389–3393. DOI: `10.22034/APJCP.2017.18.12.3389`.

[28] Hosseini Teshnizi, Saeed, and Ayatollahi, Seyyed Mohammad Taghi. “Comparison of Cox Regression and Parametric Models: Application for Assessment of Survival of Pediatric Cases of Acute Leukemia in Southern Iran”. In: *Asian Pacific Journal of Cancer Prevention* 18.4 (2017), pp. 981–985. DOI: `10.22034/APJCP.2017.18.4.981`.

[29] Li, Jun et al. “TCPA: a resource for cancer functional proteomics data”. In: *Nature Methods* 10.11 (2013), pp. 1046–1047. DOI: `10.1038/nmeth.2650`.

[30] Liu, Jianfang et al. “An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics”. In: *Cell* 173.2 (2018), 400–416.e11. DOI: 10.1016/j.cell.2018.02.052.

[31] Akbani, Rehan et al. “A pan-cancer proteomic perspective on The Cancer Genome Atlas”. In: *Nature Communications* 5 (2014), p. 3887. DOI: 10.1038/ncomms4887.

[32] National Cancer Institute. *Cancers Studied in The Cancer Genome Atlas (TCGA)*. Accessed 29 December 2025. 2024. URL: https://www.cancer.gov/ccg/research/genome-sequencing/tcga/studied-cancers.

[33] Yan, Zhiqiang et al. “Predictors of tumor progression of low-grade glioma in adult patients within 5 years follow-up after surgery”. In: *Frontiers in Surgery* 9 (2022), p. 937556. DOI: 10.3389/fsurg.2022.937556.

[34] Kumthekar, Priya et al. “Prognosis of older patients with low-grade glioma: A retrospective study”. In: *Integrative Cancer Science and Therapy* 4.5 (2017), p. 1000255. DOI: 10.15761/icst.1000255.

[35] Gittleman, Haley et al. “Sex is an important prognostic factor for glioblastoma but not for nonglioblastoma”. In: *Neuro-Oncology Practice* 6.6 (2019), pp. 451–462. DOI: 10.1093/nop/npz019.

[36] Qian, Haihong et al. “Super-enhancers and the super-enhancer reader BRD4: tumorigenic factors and therapeutic targets”. In: *Cell Death Discovery* 9 (2023), p. 470. DOI: 10.1038/s41420-023-01775-6.

[37] Liang, Huaxin, Chen, Zhuo, and Sun, Libo. “Inhibition of cyclin E1 overcomes temozolomide resistance in glioblastoma by Mcl-1 degradation”. In: *Molecular Carcinogenesis* 58.8 (2019), pp. 1502–1511. DOI: 10.1002/mc.23034.

[38] Yu, Shibo et al. “Cyclin E1 overexpression triggers interferon signaling and is associated with antitumor immunity in breast cancer”. In: *Journal for Immunotherapy of Cancer* 13.3 (2025), e009239. DOI: 10.1136/jitc-2024-009239.

[39] Pils, Dietmar et al. “Cyclin E1 (CCNE1) as independent positive prognostic factor in advanced stage serous ovarian cancer patients: A study of the OVCAD consortium”. In: *European Journal of Cancer* 50.1 (2014), pp. 99–110. DOI: 10.1016/j.ejca.2013.09.011.

[40] Chen, Hua et al. “Overexpression of CDC2/CyclinB1 in gliomas, and CDC2 depletion inhibits proliferation of human glioma cells in vitro and in vivo”. In: *BMC Cancer* 8 (2008), p. 29. DOI: 10.1186/1471-2407-8-29.

[41] Maninang, Christine, Li, Jinghong, and Li, Willis X. “Expression and prognostic role of STAT5a across cancer types”. In: *Bioscience Reports* 43.8 (2023), BSR20230612. DOI: 10.1042/BSR20230612.

[42] Ji, Wei et al. “Bioinformatics analysis of expression profiles and prognostic values of the signal transducer and activator of transcription family genes in glioma”. In: *Frontiers in Genetics* 12 (2021), p. 625234. DOI: 10.3389/fgene.2021.625234.

[43] Buscemi, Giacomo et al. “DNA damage-induced cell cycle regulation and function of novel Chk2 phosphoresidues”. In: *Molecular and Cellular Biology* 26.21 (2006), pp. 7832–7845. DOI: 10.1128/MCB.00534-06.

[44] Kubala, Marta Helena, and DeClerck, Yves Albert. “The plasminogen activator inhibitor-1 paradox in cancer: A mechanistic understanding”. In: *Cancer Metastasis Reviews* 38.3 (2019), pp. 483–492. DOI: 10.1007/s10555-019-09806-4.

[45] Huang, Xiaoming et al. “Immune-related gene SERPINE1 is a novel biomarker for diffuse lower-grade gliomas via large-scale analysis”. In: *Frontiers in Oncology* 11 (2021), p. 646060. DOI: 10.3389/fonc.2021.646060.

[46] Neal, Christopher L., and Yu, Dihua. “14-3-3ζ as a prognostic marker and therapeutic target for cancer”. In: *Expert Opinion on Therapeutic Targets* 14.12 (2010), pp. 1343–1354. DOI: 10.1517/14728222.2010.531011.

A Appendix.

Factor	Protein	Loading	HR (95% CI)	p-value
Factor1	ARAF_pS299	0.70	0.53 (0.42–0.67)	¡0.001
Factor2	X1433ZETA	0.76	1.15 (0.94–1.41)	0.162
Factor3	AKT_pT308	0.67	1.30 (1.06–1.60)	0.013
Factor4	FIBRONECTIN	0.52	1.23 (1.02–1.49)	0.030
Factor5	PKCPANBETAII_pS660	0.63	0.74 (0.60–0.91)	0.005
Factor6	EIF4G	0.58	1.63 (1.28–2.08)	¡0.001
Factor7	CYCLINE1	0.49	0.72 (0.59–0.88)	0.001
Factor8	CYCLINB1	0.43	1.43 (1.19–1.71)	¡0.001
Factor9	YAP	0.75	1.24 (1.00–1.53)	0.049
Factor10	CHK2_pT68	0.50	0.66 (0.51–0.85)	0.001
Factor11	BRD4	0.44	1.85 (1.52–2.24)	¡0.001
Factor12	BCLXL	0.64	1.27 (1.03–1.55)	0.025
Factor13	CMET	0.72	0.67 (0.54–0.83)	¡0.001
Factor14	HER2_pY1248	0.99	1.52 (1.33–1.73)	¡0.001
Factor15	STAT5ALPHA	-0.56	1.67 (1.38–2.02)	¡0.001
Factor16	BAK	-0.43	0.62 (0.50–0.77)	¡0.001
Factor17	ANNEXINVII	0.50	0.60 (0.52–0.70)	¡0.001
Factor18	PAI1	0.47	1.37 (1.21–1.56)	¡0.001

Table A1: Top protein per factor. Loadings are from factor analysis. Hazard ratios (HR, 95% CI) and p-values are from univariate Cox regression.

Variable	Univariate HR (95% CI), p	Full multivariable HR (95% CI), p
ARAF_pS299	0.53 (0.42–0.67), ¡0.001	0.99 (0.70–1.41), 0.965
X1433ZETA	1.15 (0.94–1.41), 0.162	1.42 (1.06–1.90), 0.018
AKT_pT308	1.30 (1.06–1.60), 0.013	1.15 (0.91–1.46), 0.252
FIBRONECTIN	1.23 (1.02–1.49), 0.030	0.80 (0.59–1.08), 0.147
PKCPANBETAII_pS660	0.74 (0.60–0.91), 0.005	0.76 (0.54–1.08), 0.132
EIF4G	1.63 (1.28–2.08), ¡0.001	–
CYCLINE1	0.72 (0.59–0.88), 0.001	0.75 (0.57–1.00), 0.051
CYCLINB1	1.43 (1.19–1.71), ¡0.001	1.38 (1.09–1.74), 0.006
YAP	1.24 (1.00–1.53), 0.049	–
CHK2_pT68	0.66 (0.51–0.85), 0.001	0.78 (0.57–1.07), 0.128
BRD4	1.85 (1.52–2.24), ¡0.001	1.36 (1.02–1.80), 0.033
BCLXL	1.27 (1.03–1.55), 0.025	1.14 (0.87–1.50), 0.345
CMET	0.67 (0.54–0.83), ¡0.001	0.96 (0.75–1.22), 0.731
HER2_pY1248	1.52 (1.33–1.73), ¡0.001	1.11 (0.94–1.31), 0.202
STAT5ALPHA	1.67 (1.38–2.02), ¡0.001	1.32 (1.04–1.69), 0.025
BAK	0.62 (0.50–0.77), ¡0.001	–
ANNEXINVII	0.60 (0.52–0.70), ¡0.001	–
PAI1	1.37 (1.21–1.56), ¡0.001	1.30 (1.03–1.64), 0.025
age	1.06 (1.04–1.08), ¡0.001	1.05 (1.03–1.07), ¡0.001
gender	0.84 (0.56–1.25), 0.382	0.90 (0.58–1.38), 0.627

Full multivariable model fit statistics: Concordance = 0.852 (SE = 0.018). Likelihood ratio test = 132.1 on 16 df ($p < 0.001$); Wald test = 127.5 on 16 df ($p < 0.001$); Score (log-rank) test = 182.0 on 16 df ($p < 0.001$). Note: YAP, BAK, EIF4G, and Annexin A7 were excluded from multivariable modeling due to violations of the proportional hazards assumption (see Section 3.1 of paper)

Table A2: Univariate and full multivariable Cox proportional hazards regression results.