Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

Modeling Discrete Survival Time Using Genomic Feature Data

Kyle Ferber and Kellie J. Archer

Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA.

ABSTRACT: Researchers have recently shown that penalized models perform well when applied to high-throughput genomic data. Previous researchers introduced the generalized monotone incremental forward stagewise (GMIFS) method for fitting overparameterized logistic regression models. The GMIFS method was subsequently extended by others for fitting several different logit link ordinal response models to high-throughput genomic data. In this study, we further extended the GMIFS method for ordinal response modeling using a complementary log-log link, which allows one to model discrete survival data. We applied our extension to a publicly available microarray gene expression dataset (GSE53733) with a discrete survival outcome. The dataset included 70 primary glioblastoma samples from patients of the German Glioma Network with long-, intermediate-, and short-term overall survival. We tested the performance of our method by examining the prediction accuracy of the fitted model. The method has been implemented as an addition to the ordinalgmifs package in the R programming environment.

KEYWORDS: classification, ordinal response, gene expression, survival analysis, R

SUPPLEMENT: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

CITATION: Ferber and Archer. Modeling Discrete Survival Time Using Genomic Feature Data. *Cancer Informatics* 2015:14(S2) 37–43 doi: 10.4137/CIN.S17275.

RECEIVED: November 04, 2014. **RESUBMITTED:** December 18, 2014. **ACCEPTED FOR PUBLICATION:** December 25, 2014.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Methodology

FUNDING: Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM011169. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: ferberkl@vcu.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE). Published by LibertasAcademica. Learn more about this journal.

Introduction

Frequently in high-throughput genomic research, we want to fit a statistical model using gene expression data in order to predict a future outcome. This has become a modern challenge for statisticians because there are far more features or variables (p) than samples (n). This is an obstacle in two regards. First, the design matrix will not be full-rank. Thus, there is an infinite number of solutions to the system of equations. Even small perturbations in the data will lead to large fluctuations in the coefficient estimates. Second, given the vast interrelatedness of genes, collinearity is likely to be a problem. Collinear predictors further contribute to the instability of the parameter estimates. Recently, penalization (also referred to as regularization) has stood out as an effective method to combat these two issues. There are several popular penalization methods, but the defining characteristic of them all is that they introduce bias into the parameter estimates in exchange for a reduction in variance. In many cases, penalization improves the model's predictive accuracy and, relatedly, reduces the mean squared error (MSE) of the parameter estimates.[1] In cases where model parsimony and interpretability are important, the least absolute shrinkage and selection operator (LASSO) penalization method is effective as it shrinks many parameter estimates to be exactly zero.[2] The generalized monotone incremental forward stagewise (GMIFS) method is an algorithm that can be used in logistic regression to produce a monotone LASSO solution.[3] The GMIFS method was subsequently extended by Archer et al for fitting several different logit link ordinal response models to high-throughput genomic data[4] including the cumulative logit, forward continuation ratio (CR), backward CR, stereotype logit, and adjacent category models. Herein we describe the GMIFS algorithm for ordinal

response modeling using a complementary log-log (cloglog) link, which is useful for discrete survival Modeling. Therefore, in the Discrete Survival Analysis Section, we describe the model formulation for modeling a discrete survival outcome. In the GMIFS Method for Ordinal Response Modeling section, we present the GMIFS method for the forward CR model using the cloglog link. Next, in the Application section, we discuss the motivating dataset that examined survival in glioblastoma (GBM) patients. The Results section examines the model performance in terms of parsimony, resubstitution error, and cross-validation (CV) error. Finally, in the Conclusion we provide concluding remarks, including limitations of the study.

Discrete Survival Analysis

Survival analysis encompasses methods in which the outcome variable is time to event (eg, time to death, disease relapse, etc.). The particular method used in the analysis will depend on the scale of the survival times collected. Ideally, these will be measured on a continuous scale, but sometimes for a variety of reasons, researchers only collect times on a discrete scale. For instance, for many diseases, it is impossible to record the precise date and time of relapse (ie, a continuous measurement) because the needed data are often only collected at a physician visit. Thus, we are forced to work with discrete times. Furthermore, discrete times are used when the latent scale of the response times is discrete.

High dimensional discrete survival data. Assume there are n independent subjects (i = 1, 2, 3,..., n) and p features per subject, where $p >> n$. Because this design matrix will be singular, traditional statistical methods (eg, OLS) are not applicable. The data are often presented as follows:

- Let Y_i represent the discrete survival time response variable that takes on the values ($j = 1, 2..., K$), where K is the largest value of Y observed.
- To facilitate the formation of the likelihood, we define an $n \times K$ response matrix as follows:

$$y_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{otherwise} \end{cases}$$

- A $p \times 1$ vector of covariates, $\boldsymbol{x}_i$, is observed for each subject.

The forward CR model with a complementary log-log link function. With discrete survival data, we are generally interested in modeling the discrete hazard rate defined as

$$\pi_{ij} = \pi_j(\boldsymbol{x}_i) = P(Y_i = j | Y_i \geq j, \boldsymbol{x}_i).$$

This is also the form of a probability modeled by a forward CR model. Furthermore, if it is reasonable to assume that the data were generated by a continuous-time proportional hazards model, then we use the complementary log-log (cloglog) link function,[5]

$$\log[-\log(1 - \pi_{ij})] = \alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}$$

Here α_j represents the intercept, or threshold, for the jth class. Notice that α_j is the only component of the model that depends on time. Thus, the functions for the K time points are parallel, which implies we are assuming proportional hazards.

Likelihood. We define the likelihood as a product of n conditionally independent Bernoulli random variables,[6] where π_{ij} is the discrete hazard rate and $(1 - \pi_{ij})$ is the conditional complement of π_{ij} given by $P(Y_i > j | Y_i \geq j, \boldsymbol{x}_i)$ for the forward CR model.

$$L = \prod_{i=1}^{n} \prod_{j=1}^{K-1} \pi_{ij}^{y_{ij}} \left(1 - \pi_{ij}\right)^{\sum_{k=j}^{K} y_{ik} - y_{ij}}$$

Now define $\boldsymbol{\pi}_j = (\pi_{1j}, \pi_{2j}, \ldots, \pi_{nj})^T$. When using the cloglog link, the derivative of the log-likelihood is then given by

$$\frac{\delta \log L}{\delta \beta_p} = \sum_{j=1}^{K-1} \left[\boldsymbol{x}_p^T \exp\left\{-\exp\left\{\alpha_j + \boldsymbol{X}\boldsymbol{\beta}\right\} + \alpha_j + \boldsymbol{X}\boldsymbol{\beta}\right\} \left[\frac{\boldsymbol{y}_i}{\boldsymbol{\pi}_j} - \frac{\sum_{k=j}^{K} \boldsymbol{y}_k - \boldsymbol{y}_j}{\boldsymbol{1} - \boldsymbol{\pi}_j} \right] \right]$$

We use the generalized monotone incremental forward stagewise algorithm to solve for the penalized solution:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left(\log\left[L(\boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X}) \right] - \lambda \sum_{p=1}^{P} |\beta_p| \right)$$

The tuning parameter, λ, controls the amount of shrinkage. As λ increases, the number of parameter estimates that will be shrunk to zero also increases. Using these coefficient estimates and the estimates for the α's (described later), we can recursively estimate the probability that subject i belongs to class j where

$$P\left(Y_i = j | \boldsymbol{x}_i\right) = \pi_{ij} * P\left(Y_i \geq j | \boldsymbol{x}_i\right) = \begin{cases} \pi_{ij} & \text{for } j = 1 \\ \pi_{ij} * \left[1 - \sum_{i=1}^{j-1} P\left(Y_i = j | \boldsymbol{x}_i\right)\right] & \text{for } 1 < j \leq K \end{cases}$$

Subject i is then classified to the class that corresponds to the maximum class-specific probability.

GMIFS Method for Ordinal Response Modeling

The incremental forward stagewise (IFS) method is an iterative algorithm that produces a penalized solution for a linear regression model.[3] The GMIFS method is an extension of IFS capable of fitting overparameterized logistic regression models.[3] The GMIFS algorithm was extended by Archer et al (2014) for fitting several different logit link ordinal response models to high-throughput genomic data.[4] We updated this method to allow for the use of a complementary log-log link function. The steps of the GMIFS algorithm for ordinal response modeling are as follows[4]:

1. Enlarge the predictor space as $\tilde{\boldsymbol{X}} = [\boldsymbol{X} : -\boldsymbol{X}]$, where $\boldsymbol{X}$ represents the standardized predictors.
2. Initialize the α's to their empirical values. For the forward CR model with a cloglog link, these are initialized as $\alpha_j = \log\left(-\log\left(1 - \frac{\sum_{i=1}^{n} y_{ij}}{\sum_{i=1}^{n}\sum_{k=j}^{K} y_{ik}}\right)\right)$.
3. For step $s = 0$, initialize the components of $\hat{\boldsymbol{\beta}}^{(s)}$ as $\hat{\beta}_1 = \hat{\beta}_2 = \ldots = \hat{\beta}_P = \hat{\beta}_{P+1} = \ldots = \hat{\beta}_{2P} = 0$.
4. Find $m = \text{argmin} - \delta \log L / \delta \beta_p$ at the current estimate $\hat{\boldsymbol{\beta}}^{(s)}$.
5. Update $\hat{\beta}_m^{(s+1)} = \hat{\beta}_m^{(s)} + \varepsilon$.
6. Estimate the α's by maximum likelihood, treating $\hat{\boldsymbol{\beta}}^{(s)}$ (from step 5) as fixed.
7. Repeat steps 4 to 6 until the difference between two successive log-likelihoods is smaller than a pre-specified tolerance, τ.

The rationale for enlarging the predictor space is that it allows us to avoid taking the second derivative of the log-likelihood. Once the algorithm has converged, we can obtain the penalized solution by $\hat{\beta}_p = \hat{\beta}_p - \hat{\beta}_{p+P}$.[4] Furthermore, in step 5, ε is a small incremental value; we used 0.001 in our analysis.

Application

Glioblastoma. Glioblastomas (GBMs) are highly malignant and aggressive tumors that arise from the supportive tissue of the brain. Among all primary brain and central nervous system (CNS) tumors, they are the second most common after meningiomas, which are predominantly benign, and the five-year survival rate for GBM patients is less than 4%.[7] Aside from the aggressiveness of the tumors, one possible explanation for the low survival rate is that GBMs are rare in young people; the median age at diagnosis is 64, and the age group with the highest incidence rate is 75–84 year olds.[7] Treatment involves surgical removal of as much of the tumor as is safely possible followed by radiotherapy and/or chemotherapy.[8] The Cancer Genome Atlas (TCGA) Research Network revealed a subtype of GBM related to the mRNA expression and methylation of a set of genes that affects young adults and has an increased survival rate. Researchers also discovered four molecular subtypes of GBM that have unique responses to treatment and gene mutations that could lead GBMs to become resistant to therapy after a standard chemotherapy treatment.[9,10] These findings highlight the importance of genomic research in the study and treatment of GBM.

Data. We downloaded the raw CEL files for GSE53733 from Gene Expression Omnibus.[11] The investigators used Affymetrix HG-U133 v2.0 GeneChips to measure gene expression from patients' tumor samples taken from their initial operation. In the dataset, there were $n = 70$ GBM patients, of which 16 had an overall survival (OS) of less than 12 months, 31 patients had an OS between 12 and 36 months, and 23 patients had an OS greater than 36 months.[12] The patients' survival times were reported by the investigator as short-, intermediate-, and long-term OS. There were $p = 54{,}613$ features per subject in the CEL files after excluding control probe sets. However, after processing the data to remove probe sets with MAS5 present calls in <30% of the subjects,[13] 31,744 features remained. Furthermore, a 3′:5′ ratio much different from 1 for the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is associated with poor cDNA and cRNA quality.[14] Thus, we removed one subject with a 3′:5′ GAPDH ratio greater than 3, leaving us with 69 subjects. We then used the RMA method to obtain probe set expression summaries for our statistical analysis.[15] Afterward, we fit a forward CR model using the cloglog link with $\varepsilon = 0.001$ and $\tau = 0.00001$.

Results

After the GMIFS algorithm converged, we examined two models: (a) the model selected my minimizing the AIC criterion and (b) the model resulting from the convergence of the GMIFS algorithm (Fig. 1). Using the full dataset, the AIC-selected model misclassified 10 of the 69 patients, while the converged model only misclassified one patient (Tables 1 and 2). However, the AIC-selected model was more parsimonious with 25 non-zero coefficients, while the converged model contained 46 non-zero coefficients. The 25 probe sets that had non-zero coefficient estimates in the AIC-selected model are shown in Table 3. Furthermore, for each model, we examined the sensitivity and specificity for diagnosing short-term survival as well as the sensitivity and specificity for diagnosing short- or intermediate-term survival (Tables 4 and 5). Among the probe sets with non-zero coefficient estimates, the one with the largest absolute coefficient estimate in both models (among probe sets with known gene symbols) was designed to interrogate HD Domain Containing 2 (HDDC2). Long-term survivors had higher HDDC2 expression levels than short- and intermediate-term survivors (Fig. 2). This result agrees with another GBM study that showed that HDDC2 was significantly downregulated in short-term survivors compared to long-term survivors.[12] There was also a clear positive relationship between Nucleoside-Triphosphatase, Cancer-Related (NTPCR) expression and survival time (Fig. 3). Interestingly, researchers have shown that NTPCR is overexpressed in neuroblastomas,[16] but no study has associated NTPCR with

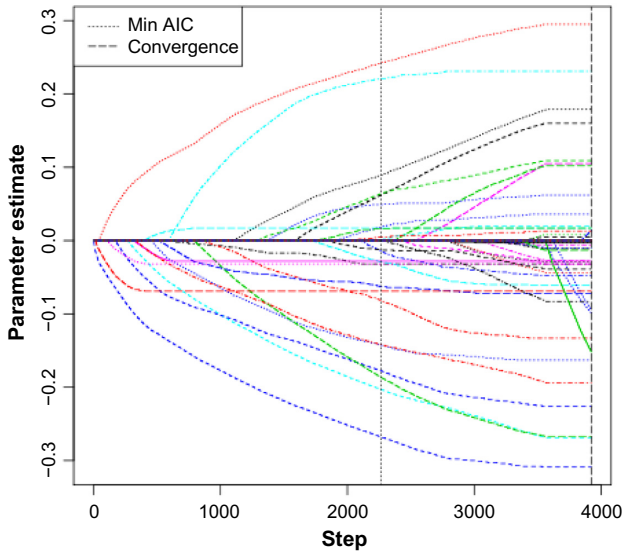

Figure 1. Coefficient paths for our forward CR model using a complementary log-log link.
Notes: The first vertical dashed line signifies the step in the algorithm when the AIC was minimized. The second vertical dashed line marks the step when the algorithm converged.

GBM. Additionally, one of the probe sets that had a non-zero coefficient estimate was designed to interrogate the gene PDZ and LIM domain 4 (PDLIM4), which has previously been studied in association with gliomas. Researchers examined the expression of the gene at the protein level for patients with gliomas and discovered that the median OS for patients with high levels of the protein (PDLI4) was significantly shorter than for patients with low protein levels.[17] We compared the mean $\log_2$ expression levels of PDLIM4 for patients with short-, intermediate-, and long-term OS using Welch's T-test. Patients with short-term OS had significantly higher expression levels than patients with long-term OS at the Bonferroni adjusted $\alpha = \frac{0.05}{3} = 0.017$ significance level ($P = 0.0019$), and patients with intermediate-term OS also had significantly higher expression levels than patients with long-term OS ($P < 0.0001$). The difference in mean expression levels between those with short-term OS and those with intermediate-term OS was not significant. Thus, it appears that both the gene expression levels and the protein levels of the gene are lower for patients who survive longer.

A common critique of a model fitted from high-dimensional data is that the final model, even if selected by minimizing AIC, is not parsimonious. In this example, critics may say that given a sample size of 69 subjects, including 25 coefficients in the model is overfitting, and that the model performance is likely a result of chance. In response, we fit two additional models whose performances will be a result of chance alone. First, we fit a model with the same gene-expression data used in our example, but we randomly permuted the response vector. Next, we fit a model using our original response vector, but instead of using the gene expression data, we used a design matrix filled with $31{,}744 \times 69 = 2{,}190{,}336$ random variables generated from a Gaussian distribution with a mean and standard deviation equal to the corresponding sample statistics of the gene expression data. If we exclude regions of underfitting and overfitting, the model fit with the gene expression data and the original response vector had better performance than the other two models whose performances are a result of chance rather than a relationship between the features and the response (Fig. 4).

We also performed N-fold (or leave-one-out) CV to assess the generalizability of our models (where $N = 69$). Both the AIC-selected model and the converged model had an N-fold CV error rate of about 44.9% (Tables 6 and 7). Thus, it appears that the AIC-selected model and the model that satisfied the GMIFS convergence criterion predict discrete survival time equally well. We chose the AIC-selected model as our final model as it is more parsimonious and therefore more interpretable.

Table 1. AIC-selected model cross-tabulation of the observed versus the predicted class using the full dataset.

		Observed		
		Short	**Intermediate**	**Long**
	Short	10	0	0
Predicted	**Intermediate**	6	31	4
	Long	0	0	18

Table 2. Converged model cross-tabulation of the observed versus the predicted class using the full dataset.

		Observed		
		Short	**Intermediate**	**Long**
	Short	16	0	0
Predicted	**Intermediate**	0	31	1
	Long	0	0	21

Conclusion

GBM is a particularly dangerous tumor with a low survival rate. A specific and accurate prognosis would be very useful to both the patient and the oncologist. Thus, we were interested in predicting survival time based on a patient's genomic feature data. We used discrete times because the investigators of this particular GBM study reported discrete times. Another case when discrete survival times would be used is when the outcome of interest (eg, disease relapse) can only be assessed at physician visits. The GMIFS algorithm is an effective method for building a classifier for an ordinal response outcome given a high-dimensional covariate space. In this case, we fit a forward CR model with a complementary log-log link function to model discrete survival time. The model resulting from the convergence of the algorithm had only a 1.4% resubstitution error. Using N-fold CV, the model had a

Table 3. Probe sets with non-zero coefficient estimates in the AIC and converged models.

PROBE SET	ENTREZ ID	GENE SYMBOL	CHROMOSOME	$\hat{\beta}_{AIC}$	$\hat{\beta}_{CONVERGED}$	CANCER ASSOCIATIONS
203260_at	51020	HDDC2	6	−0.268	−0.309	Glioblastoma[12]
1557883_a_at	<NA>	<NA>	<NA>	−0.203	−0.269	
206565_x_at	11039	SMA4	5	−0.186	−0.267	
1558723_at	284014	LOC284014	17	−0.178	−0.226	
202447_at	1666	DECR1	8	−0.142	−0.163	Breast cancer[18]
226813_at	84284	NTPCR	1	−0.142	−0.194	Neuroblastomas[16]
209078_s_at	25828	TXN2	22	−0.081	−0.133	Breast cancer[19]
230581_at	<NA>	<NA>	<NA>	−0.069	−0.069	
215962_at	<NA>	<NA>	<NA>	−0.063	−0.072	
1557100_s_at	25831	HECTD1	14	−0.032	−0.032	Breast cancer[20]
242333_at	<NA>	<NA>	<NA>	−0.032	−0.032	
206697_s_at	3240	HP	16	−0.029	−0.061	Non-small cell lung cancer,[21] Hepatocellular carcinoma[22]
222992_s_at	4715	NDUFB9	8	−0.028	−0.028	
219221_at	253461	ZBTB38	3	−0.014	−0.048	Involved in DNA replication and stability[23]
230353_at	284112	LOC284112	17	−0.013	−0.039	
243957_at	400464	LOC400464	15	0.005	0.013	Diffuse large cell B lymphoma[24]
231773_at	9068	ANGPTL1	1	0.016	0.017	Prostate cancer[25]
211564_s_at	8572	PDLIM4	5	0.017	0.017	Glioma,[17] acute myeloid leukemia,[26] Prostate cancer,[27] breast cancer[28]
218669_at	57826	RAP2C	X	0.019	0.036	Acute lymphoblastic leukemia[29]
1561759_at	645513	LOC645513	4	0.049	0.062	
1559283_a_at	285888	CNPY1	7	0.062	0.160	
221900_at	1296	COL8A2	1	0.064	0.109	
203184_at	2201	FBN2	5 9	0.089	0.179	Colorectal cancer[30]
234547_at	<NA>	<NA>	<NA>	0.221	0.231	
229146_at	136895	C7orf31	7	0.242	0.295	

44.9% misclassification rate, significantly better than chance (66% misclassification rate for a three-class outcome), but there is room for improvement. For example, although our method performs automatic variable selection, improvement gains in classification accuracy may be achieved by reducing the dimensionality of the feature set in a meaningful way prior to model fitting. We plan to explore this topic in a follow-up paper. Furthermore, a more accurate classifier could be built with more information. For instance, the five-year survival rate for patients diagnosed between the ages of 0 and 19 is around 19%, while the five-year survival rate for patients diagnosed between the ages of 45 and 54 is only about 3.3%.[7] Additionally, age was significantly different across the three outcome classes in this study[12] but was not made available in the data. Thus, including age as an unpenalized predictor in our model would likely improve its predictive accuracy (the ordinalgmifs R package allows the user to select a subset of predictors that will not be penalized in the GMIFS algorithm[31]). Also, Karnofsky performance status and extent of surgical resection are known prognostic factors for GBM,[32] so

Table 4. AIC-selected model sensitivity and specificity for predicting short-term survival and for predicting short- or intermediate-term survival.

OUTCOME	SENSITIVITY	SPECIFICITY
Short-term survival	63	100
Short- or intermediate-term survival	100	82

Table 5. Converged model sensitivity and specificity for predicting short-term survival and for predicting short- or intermediate-term survival.

OUTCOME	SENSITIVITY	SPECIFICITY
Short-term survival	100	100
Short- or intermediate-term survival	100	95

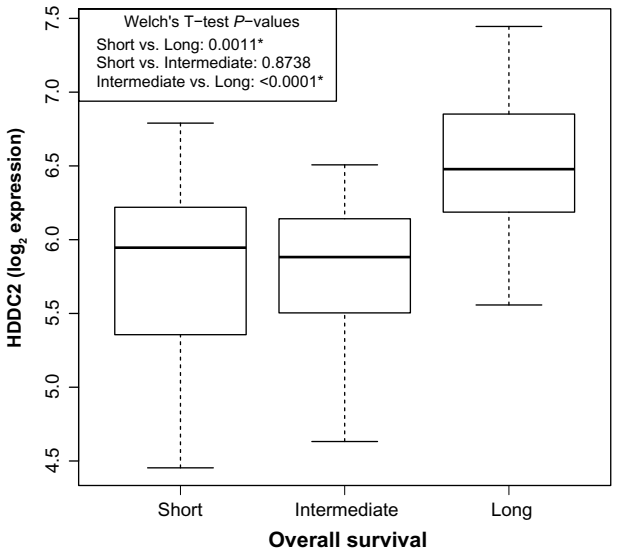

Figure 2. Boxplot of 203260_at (HDDC2) $\log_2$ expression by discrete OS outcome (short-term, intermediate, long-term survival).
Note: *Significant at the Bonferroni adjusted $\alpha = \frac{0.05}{3} = 0.017$ significance level.

they could have been effective unpenalized predictors as well (despite the fact that these two variables were not significantly different across the three classes in this study[12]). Additionally, a specific month of death would have provided more information than a range of months. However, for each discrete time value, we would need enough subjects with that response to fit a reliable model. As the price of microarray experiments decreases, we will have greater access to datasets with a larger number of subjects, making this a reasonable expectation for the future.

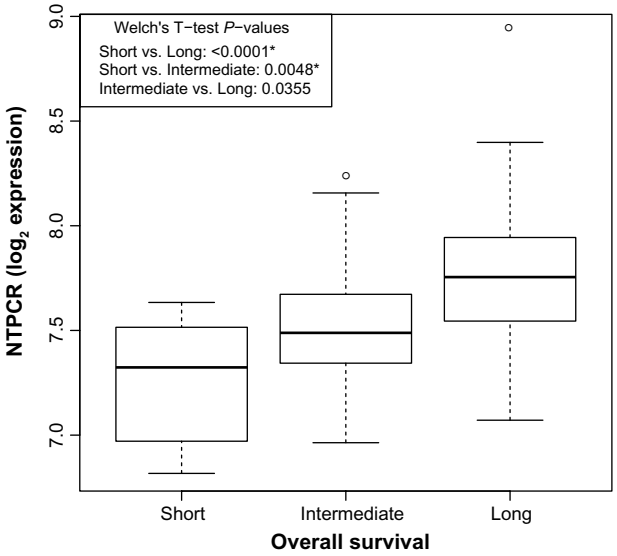

Figure 3. Boxplot of 226813_at (NTPCR) $\log_2$ expression by discrete OS outcome (short-term, intermediate, long-term survival).
Note: *Significant at the Bonferroni adjusted $\alpha = \frac{0.05}{3} = 0.017$ significance level.

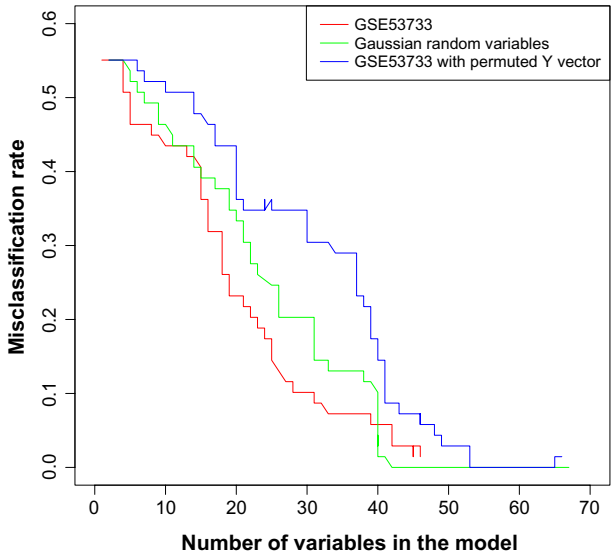

Figure 4. Plot of model misclassification rate by number of variables included in the model for the original GSE53733 data (red line), GSE53733 data with a permuted response (blue line), and Gaussian random variables (green line).

Table 6. N-fold CV: AIC-selected model cross-tabulation of the observed versus the predicted class.

		Observed		
		Short	**Intermediate**	**Long**
	Short	1	1	0
Predicted	**Intermediate**	14	30	15
	Long	1	0	7

Table 7. N-fold CV: converged model cross-tabulation of the observed versus the predicted class.

		Observed		
		Short	**Intermediate**	**Long**
	Short	4	3	1
Predicted	**Intermediate**	12	27	14
	Long	0	1	7

Author Contributions

Conceived and designed the methods: KF, KJA. Analyzed the data: KF. Wrote the first draft of the manuscript: KF. Contributed to the writing of the manuscript: KJA. Agree with manuscript results and conclusions: KF, KJA. Jointly developed the structure and arguments for the paper: KF, KJA. Made critical revisions and approved final version: KF, KJA. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc; 2001.

2. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B*. 1996;58(1):267–88.
3. Hastie T, Taylor J, Tibshirani R, Walther G. Forward stagewise regression and the monotone lasso. *Electron J Statist*. 2007;1:1–29.
4. Archer KJ, Hou J, Zhou Q, Ferber K, Layne JG, Gentry AE. ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer Inform*. 2014;13:187–95.
5. Allison PD. Discrete-time methods for the analysis of event histories. *Sociol Methodol*. 1982;13:61–98.
6. Archer KJ, Williams AAA. L 1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Stat Med*. 2012;31:1464–74.
7. Dolecek TA, Propp JM, Stroup NE, Kruchko C. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2005–2009. *Neuro-Oncol*. 2012;14:v1–49.
8. Stupp R, Mason WP. Molecular characterization of long-term survivors of glioblastoma using genome-and transcriptome-wide profiling. *New Engl J Med*. 2005;352:987–96.
9. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008;455:1061–8.
10. Brennan CW, Verhaak RG, McKenna A, et al; TCGA Research Network. The somatic genomic landscape of glioblastoma. *Cell*. 2013;155:462–77.
11. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30:207–10.
12. Reifenberger G, Weber RG, Riehmer V, et al; German Glioma Network. Molecular characterization of long-term survivors of glioblastoma using genome-and transcriptome-wide profiling. *Int J Cancer*. 2014;135(8):1822–31.
13. McClintick JN, Edenberg HJ. Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinformatics*. 2006;7:49.
14. Dumur CI, Nasim S, Best AM, et al. Evaluation of quality-control criteria for microarray gene expression analysis. *Clin Chem*. 2004;50:1994–2002.
15. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*. 2003;31:e15–e15.
16. Pasdziernik JM, Kaltschmidt B, Kaltschmidt C, Klinger C, Kaufmann M. On the cytotoxicity of HCR-NTPase in the neuroblastoma cell line SH-SY5Y. *BMC Res Notes*. 2009;2:102.
17. de Tayrac M, Saikali S, Aubry M, et al. Prognostic significance of EDN/RB, HJURP, p60/CAF-1 and PDLI4, four new markers in high-grade gliomas. *PLoS One*. 2013;8:e73332.
18. Ursini-Siegel J, Rajput AB, Lu H, et al. Elevated expression of DecR1 impairs ErbB2/Neu-induced mammary tumor development. *Mol Cell Biol*. 2007;27:6361–71.
19. Seibold P, Hein R, Schmezer P, et al. Polymorphisms in oxidative stress-related genes and postmenopausal breast cancer risk. *Int J Cancer*. 2011;129:1467–76.
20. Del Valle PR, Milani C, Brentani MM, et al. Transcriptional profile of fibroblasts obtained from the primary site, lymph node and bone marrow of breast cancer patients. *Genet Mol Biol*. 2014;37:480–9.
21. Park J, Yang JS, Jung G, et al. Subunit-specific mass spectrometry method identifies haptoglobin subunit alpha as a diagnostic marker in non-small cell lung cancer. *J Proteomics*. 2013;94:302–10.
22. Pompach P, Brnakova Z, Sanda M, Wu J, Edwards N, Goldman R. Site-specific glycoforms of haptoglobin in liver cirrhosis and hepatocellular carcinoma. *Mol Cell Proteomics*. 2013;12:1281–93.
23. Miotto B, Chibi M, Xie P, et al. The RBBP6/ZBTB38/MCM10 Axis Regulates DNA Replication and Common Fragile Site Stability. *Cell Rep*. 2014;7:575–87.
24. Kim SJ, Sohn I, Do IG, et al. Gene expression profiles for the prediction of progression-free survival in diffuse large B cell lymphoma: results of a DASL assay. *Ann Hematol*. 2014;93:437–47.
25. Sato R, Yamasaki M, Hirai K, et al. Angiopoietin-like protein 2 induces androgen-independent and malignant behavior in human prostate cancer cells. *Oncol Rep*. 2015;33:58–66.
26. Li Y, Qian J, Lin J, et al. Reduced expression of PDLIM4 gene correlates with good prognosis in acute myeloid leukemia. *Zhongguo Shi Yan Xue Ye Xue Za Zhi*. 2013;21:1111–5.
27. Vanaja DK, Grossmann ME, Cheville JC, et al. PDLIM4, an actin binding protein, suppresses prostate cancer cell growth. *Cancer Invest*. 2009;27:264–72.
28. Xu J, Shetty PB, Feng W, et al. Methylation of HIN-1, RASSF1A, RIL and CDH13 in breast cancer is associated with clinical characteristics, but only RASSF1A methylation is associated with outcome. *BMC Cancer*. 2012;12:243.
29. Lilljebjörn H, Heidenblad M, Nilsson B, et al. Combined high-resolution array-based comparative genomic hybridization and expression profiling of ETV6/RUNX1-positive acute lymphoblastic leukemias reveal a high incidence of cryptic Xq duplications and identify several putative target genes within the commonly gained region. *Leukemia*. 2007;21:2137–44.
30. Hibi K, Mizukami H, Saito M, Kigawa G, Nemoto H, Sanada Y. FBN2 methylation is detected in the serum of colorectal cancer patients with hepatic metastasis. *Anticancer Res*. 2012;32:4371–4.
31. Archer KJ, Hou J, Zhou Q, Ferber K, Layne JG, Gentry A. *ordinalgmifs: Ordinal Regression for High-dimensional Data* 2014. R package version 1.0.3, 2014.
32. Liu Y, Shete S, Etzel CJ, et al. Polymorphisms of LIG4, BTBD2, HMGA2, and RTEL1 genes involved in the double-strand break repair pathway predict glioblastoma survival. *J Clin Oncol*. 2010;28:2467–74.