Predicting survival times for neuroblastoma patients using RNA-Seq expression profiles

Tyler Grimes, Alejandro Walker, Susmita Datta, Somnath Datta[1]
Department of Biostatistics, University of Florida, FL 32610, USA

1 Introduction

Neuroblastoma is the most frequently diagnosed cancer in the first year of life and the most common extra-cranial solid tumor in children. It accounts for 5% of all pediatric cancer diagnoses and 10% of all pediatric oncology deaths. These numbers have improved over the past decade, but accurate prognosis for the disease has remained a challenge (Bosse and Maris, 2016). The difficulty is due to the highly heterogeneous nature of neuroblastoma; cases can range from tumors that spontaneously regress on their own, to aggressive tumors that spread unabated by treatment.

In 1980, the MYCN oncogene was identified as a biomarker for clinically aggressive tumors. It has since been one of the most important markers for stratifying patients. Genome-wide association studies have found many other genes having alleles that are associated with an increased risk of neuroblastoma. However, while aberrations of these genes indicate an increased susceptibility to the disease, these markers are less useful in stratification once the patient is diagnosed.

Predicting survival outcomes using gene expression data has been explored with promising results (Formicola et al., 2016; Tan et al., 2008). These studies use gene expression profiles from microarrays with classification methods to stratify patients into risk groups.

In this study, we undertake the CAMDA 2017, Neuroblastoma data integration challenge. In our analysis, clinical data and expression profiles from RNA-Seq data are integrated together to model survival times directly. The effects of using various feature levels of expression profiles (genes, transcripts, and introns) are examined and compared to a model without RNA-Seq data. The inclusion of RNA-Seq profiles is shown to increase the prediction accuracy for both overall survival and event free survival times. These models can also be used as a classifier to accurately identify high-risk groups. We end by discussing our continued research into the use of an ensemble predictor, which will integrate the several models developed here to further improve prediction accuracy.

2 Datasets

The datasets can be accessed from the GEO database with accession number GSE49711 (Su et al., 2014; Zhang et al., 2015). The data are comprised of tumor samples from 498 neuroblastoma patients from seven countries: Belgium (n = 1), Germany (n = 420), Israel (n = 11), Italy (n = 5), Spain (n = 14), United Kingdom (n = 5), and United States (n = 42). Several clinical variables are available for each patient, along with the RNA sequencing information from their tumor sample. Su et al. (2014) randomly separated the data into a training set and testing set; this partition is recorded with the clinical data and is also used here.

[1] To whom correspondence should be addressed (somnath.datta@ufl.edu).

2.1 Clinical data

The clinical data consist of 11 variables. In this study, three of these variables are used as covariates in our models; these include sex, age, and MYCN status (see table 1).

There are two outcomes of interested, namely overall survival and event free survival. Overall survival is calculated as the time from diagnosis to death from disease or the last follow-up date, if the patient survived. Event free survival was calculated from diagnosis to the time of tumor progression, relapse, or death from disease or to the last follow-up, if no event occurred.

The censoring rates are shown in table 1 as death from disease and progression ("No" corresponds to right-censoring).

Table 1: Clinical variables

Variables	Training	Testing
Sex		
Male	146	141
Female	103	108
Age		
< 18 months	156	144
≥ 18 months	93	105
MYCN Status		
Normal	199	202
Amplified	47	45
N/A	3	2
Death from Disease		
Yes	51	54
No	198	195
Progression		
Yes	89	94
No	160	155

2.2 RNA-Seq Dataset

The RNA-Seq data provide annotations at three feature levels, giving datasets comprised of 60,776 genes, 263,544 transcripts, and 340,414 introns, respectively. A hierarchical version of the transcript annotation is also provided but was not used.

Genes and transcripts without an NCBI ID were removed. Any RNA-Seq variables with over 90% of zeroes for counts were also omitted. A database of 3681 important genes related to neuroblastoma was obtained from the GeneCards Suite (Safran et al., 2010). This dataset was used to subset the remaining genes and transcripts, resulting in 3401 genes and 48288 transcripts. For the introns, their predictive ability for survival time was ranked by fitting each intron in a Cox proportional hazards model for the overall survival time of patients in the training set. The top 10,000 introns with the smallest p-values (testing that the coefficient is zero) were used.

3 Accelerated failure time (AFT) models

The AFT model relates the log survival times to a linear combination of the covariates using the regression equation $\log(y) = X\beta + \epsilon$, where $y \in \mathbb{R}^{+n}$ denotes the observed survival times for n observations, X the $n \times p$ matrix with columns containing the predictor variables for each observation, $\beta \in \mathbb{R}^p$ the unknown parameter of interest, and $\epsilon \in \mathbb{R}^n$ an unobservable random error that is assumed to be independent of X. The predictors X are centered and scaled prior to fitting the model.

Since $p > n$, ordinary least squares (OLS) is not appropriate and would over-fit on the observed data. We consider four alternative approaches to fit the AFT model. These involve latent factor and regularization techniques. Both of these require the selection of one or more tuning parameters. These parameters can be determined using k-fold cross validation. In this study, 10-fold cross validation is used and implemented in R using two packages discussed in the following sections.

3.1 Dimension Reduction In order to fit the AFT model with $p > n$ predictors, we consider four different dimension reduction techniques. These include partial least square regression (PLS) (Boulesteix

and Strimmer, 2007), sparse partial least square regression (SPLS) (Chun and Keles, 2010), the lasso (Tibshirani, 1996) and the elastic net (Zou et al, 2005). These procedures require the selection of tuning parameters, which is done by 10-fold cross validation using the R packages "spls" and "glmnet."

4 Imputation for right-censoring

Let $\{(y_i, \delta_i, X_i) \mid i = 1, \dots, n\}$ denote the set of observed survival times, indicators for death from disease, and the p-dimensional vector of covariates for the n patients in the dataset. Let T_i denote the true survival times for patient $i = 1, \dots, n$. If the ith patient's survival time is censored (i.e. $\delta_i = 0$) then we only observe $y_i < T_i$. That is, T_i is unobserved.

To deal with this right-censoring, the dataset imputation procedure from Mostajabi et al. (2011) is used. An initial estimate $\hat{\beta}^{(1)}$ is obtained by fitting the AFT model using only the uncensored data. Then, in each of $k = 1, \dots, n_K$ iterations, first calculate the Kaplan-Meier estimate $\hat{S}^{(k)}(e)$ of the distribution of model error using $\{(e_i, \delta_i) \mid i = 1, \dots n\}$ where $e_i = \log(y_i) - X_i^T \hat{\beta}^{(k)}$. Then, n_J new datasets are imputed by replacing each censored $\log(y_i)$ with $X_i^T \hat{\beta}^{(k)} + e_i^*$, where e_i^* is a sampled model residual from the conditional distribution $\hat{S}^{(k)}(\, e \mid e > \, e_i)$. This condition ensures that the imputed observation will be larger than the observed right-censored time. The new datasets are used to compute estimates $\hat{\beta}_j^{(k)}$ for $j = 1, \dots, n_J$. At the end of the iteration, the estimate is updated by $\hat{\beta}^{(k+1)} = \frac{1}{n_J} \sum_{j=1}^{n_D} \hat{\beta}_j^{(k)}$. The process is repeated for n_K iterations and the final estimate $\hat{\beta}^{(n_K)}$ is returned.

To balance between computation time and simulation variability, we chose to run $n_K = 5$ iterations, imputing $n_J = 10$ datasets in each.

5 Results

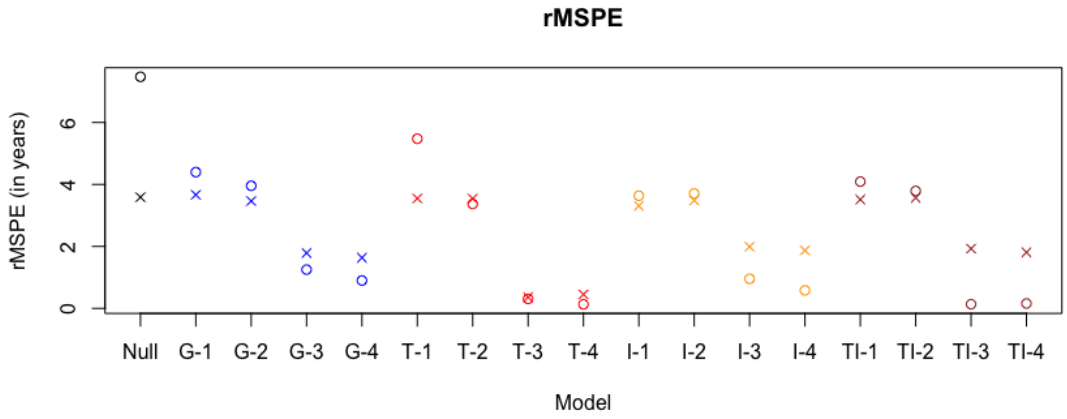
Thirteen models are considered in total, and each is used to estimate overall survival and event free survival. For a baseline of comparison, a "null" model is fit with clinical variables only (no RNA-Seq data). The other twelve models use clinical variables and RNA-Seq data; these models use genes, transcripts, introns, or both transcripts and introns, each fit using the four methods PLS, SPLS, lasso, and elastic net. Performance is measured using a weighted root mean squared prediction error (MSPE), which is defined by

$$\text{rMSPE} = \left(\frac{1}{\sum \delta_i} \sum_{i=1}^{n} \delta_i \left(y_i - X_i^t \hat{\beta} \right)^2 \right)^{1/2}$$

5.1 Prediction of survival times

Overall, the models using RNA-Seq data all perform better than the model with clinical variables only (see figure 1). The lasso and elastic net provide better accuracy than SPLS and PLS when predicting overall survival; they also perform better for event free survival, but the difference is less stark. Interestingly, there do not seem to be substantial differences in the predictive capabilities between the different RNA-Seq datasets, even after integrating both transcripts and introns into the model.

Figure 1: Performance measures for each model. The circles and ×'s correspond to overall survival and event free survival, respectively. The blue, red, orange, and brown points correspond to models using genes, transcripts, introns, and introns with transcripts, respectively. The numbers 1-4 represent the PLS, SPLS, lasso, and elastic net methods.



5.2 Pathway analysis

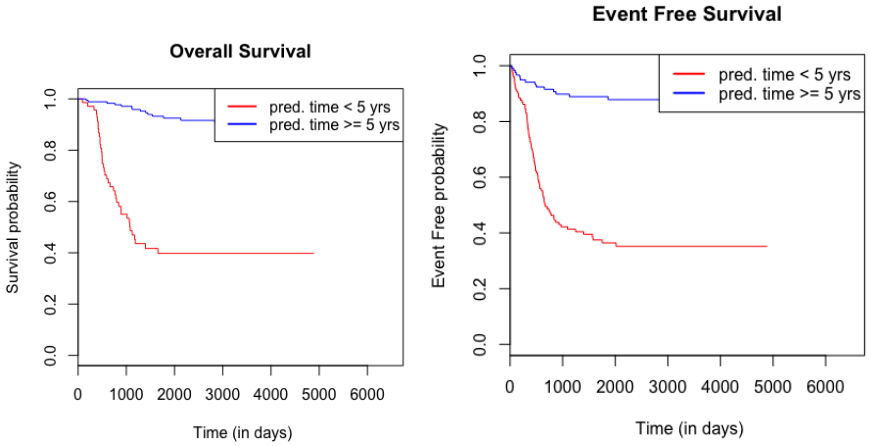
A pathway analysis is performed to evaluate the potential function relevance of the genes selected by the models. We consider the two fitted models G-4 and T-4, which used the elastic net on genes and transcripts, respectively. When predicting overall survival, 185 genes were given positive coefficients by the G-4 model. The Ras, PI3k-Akt, TGF-β, and cell cycle pathways were all expressed; these are related to the cell survival process and are known to be major signaling pathways in cancer (Vogelstein et al., 2013). For event free survival, the model selected only 35 genes, but the same five cell survival pathways were being expressed.

For the T-4 model, 404 transcripts were selected when estimating overall survival. The pathway analysis is done on the genes containing these transcripts. As before, several cell survival pathways were discovered: Ras, PI3k-Akt, Jak-STAT, MAPK, and cell cycle. When predicting event free survival, 198 transcripts were selected. In this case, the MAPK and cell cycle pathways did not show up.

5.3 Kaplan-Meier analysis

These models can also be used to classify patients into high-risk and low-risk groups by setting a threshold for the survival time. Here, we show Kaplan-Meier survival curves using this approach; a patient is classified as high-risk if their predicted survival time is less than 5 years. While the SPLS model has lower prediction accuracy for individual survival times, it produces a wider range of predicted values than the lasso and elastic net, which makes it better suited for a classification scheme. The G-2 model using genes with SPLS is used here; the plots in figure 2 show the resulting survival curves on the validation dataset.

Figure 2: Kaplan-Meier analysis of survival times when using the G-2 model (genes with SPLS) to classify patients. If the predicted survival time is less than 5 years, the patient is classified as high-risk. The difference between the survival curves for both overall survival and event free survival are statistically significant (p-value < 1E-16).



6 Discussion

In this study, we used the AFT model with four dimension reduction techniques and a dataset imputation scheme to predict overall survival and event free survival times of neuroblastoma patients. Three feature levels of an RNA-Seq dataset were considered. Models were fit using the features independently and integrated together (with introns and transcripts). The predictive performances are similar among these four scenarios. The performance depends most on the dimension reduction method used; the lasso and elastic net provide the best accuracy overall. A pathway analysis revealed that the elastic net selected genes and transcripts that are involved in several major signaling pathways in cancer; this attests to the functional relevance of the genes selected by the model.

The predictive powers of these models are all individually better than the baseline model using only clinical data. However, we suspect that by integrating these 16 models together using an ensemble procedure, even better prediction accuracy can be achieved. An ensemble approach also allows for several performance measures to be taken into account, rather than just the rMSPE. We are actively pursuing this line of research and will have results in the near future.

References

Bosse, K.R. et al., 2016. *Cancer*, *122*(1), pp.20-33.
Formicola, D. et al., 2016. *Journal of Translational Medicine*, *14*(1), p.142.
Tan, Q. et al., 2008. *Advances in Computer and Information Sciences and Engineering*, pp.405-409.
Su, Z. et al., 2014. *Genome Biology*, *15*(12), p.523.
Zhang, W. et al., 2015. *Genome Biology*, *16*(1), p.133.
Safran, M. et al., 2010. *Database*, *2010*, p.baq020.
Boulesteix, A.L. et al., 2010. *Briefings in Bioinformatics*, *8*(1), pp.32-44.
Chun, H. et al., 2010. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(1), pp.3-25.
Tibshirani, R., 1996. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.267-288.
Friedman, J. et al., 2010. *Journal of statistical software*, *33*(1), p.1.
Zou, H. et al., 2005. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), pp.301-320.
Vogelstein, B. et al., 2013. *Science*, *339*(6127), pp.1546-1558.