5-2010

SURVIVAL PREDICTION FOR BRAIN TUMOR PATIENTS USING GENE EXPRESSION DATA

Vinicius Bonato

SURVIVAL PREDICTION FOR BRAIN TUMOR PATIENTS

USING GENE EXPRESSION DATA

by

Vinicius Bonato, PhD
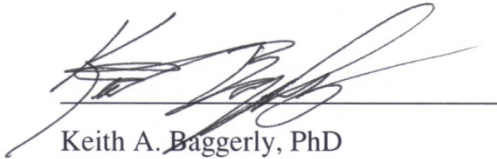
APPROVED:
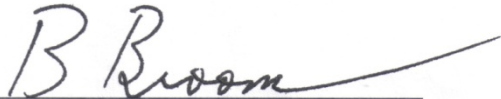
Supervisory Professor: Kim-Anh Do, PhD

Veerabhadran Baladandayuthapani, PhD

Keith A. Baggerly, PhD

Bradley M. Broom, PhD

Andrei S. Rodin, PhD

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

SURVIVAL PREDICTION FOR BRAIN TUMOR PATIENTS

USING GENE EXPRESSION DATA

A

THESIS

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

M. D. Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

by

Vinicius Bonato, PhD

May, 2010

Abstract

SURVIVAL PREDICTION FOR BRAIN TUMOR PATIENTS USING GENE EXPRESSION
DATA

Vinicius Bonato, B.S., Universidade Estadual de Campinas, Campinas, SP, Brazil;

M.Sc., Universidade Estadual de Campinas, Campinas, SP, Brazil;

Ph.D., Universidade Estadual de Campinas, Campinas, SP, Brazil;

Supervisory Professor: Dr. Kim-Anh Do

Brain tumor is one of the most aggressive types of cancer in humans, with an estimated median survival time of 12 months and only 4% of the patients surviving more than 5 years after disease diagnosis. Until recently, brain tumor prognosis has been based only on clinical information such as tumor grade and patient age, but there are reports indicating that molecular profiling of gliomas can reveal subgroups of patients with distinct survival rates. We hypothesize that coupling molecular profiling of brain tumors with clinical information might improve predictions of patient survival time and, consequently, better guide future treatment decisions. In order to evaluate this hypothesis, the general goal of this research is to build models for survival prediction of glioma patients using DNA molecular profiles (U133 Affymetrix gene expression microarrays) along with clinical information. First, a predictive Random Forest model is built for binary outcomes (i.e. short *vs.* long-term survival) and a small subset of genes whose expression values can be used to predict survival time is selected. Following, a new statistical methodology is developed for predicting time-to-death outcomes using *Bayesian ensemble trees*. Due to a large heterogeneity observed within prognostic classes obtained by the Random Forest model, prediction can be improved by relating time-to-death with gene expression profile directly. We propose a Bayesian ensemble model for survival prediction which is appropriate for high-dimensional data such as gene expression data. Our approach is based on the ensemble "sum-of-trees" model which is flexible to incorporate additive and interaction effects between genes. We specify a fully Bayesian hierarchical approach and illustrate our methodology for the CPH, Weibull, and AFT survival models. We overcome the lack

of conjugacy using a latent variable formulation to model the covariate effects which decreases computation time for model fitting. Also, our proposed models provides a model-free way to select important predictive prognostic markers based on controlling false discovery rates. We compare the performance of our methods with baseline reference survival methods and apply our methodology to an unpublished data set of brain tumor survival times and gene expression data, selecting genes potentially related to the development of the disease under study. A closing discussion compares results obtained by Random Forest and Bayesian ensemble methods under the biological/clinical perspectives and highlights the statistical advantages and disadvantages of the new methodology in the context of DNA microarray data analysis.

Acknowledgments

Table of Contents

List of Figures

List of Tables

1. Introduction

Glioma is a type of central nervous system (CNS) cancer affecting the glial cells (Cairncross *et al.*, 1998). Glial cells are responsible for building the neuronal myelin sheath, forming the CNS structural tissue, and providing physical protection to adjacent neuronal cells (Junqueira & Carneiro, 2005). Glioma is the most frequent (~40%) type of primary brain tumor (PBT), with an average of worldwide annual occurrence close to 190,000 cases (Castells *et al.*, 2009) resulting in more than 140,000 deaths each year with 10,000 of them occurring only in the United States (Nutt *et al.*, 2003). Despite major efforts to reduce deaths caused by this disease, the mean survival time of newly diagnosed malignant glioma patients remains at approximately 12 months (Furnari *et al.,* 2007) and after 24 months of surgical resection nearly 90% of patients are dead (Louis *et al.*, 2007; Nutt *et al.*, 2003). In the list of deaths caused by cancer, gliomas are ranked in first for children under the age of 15 years and are ranked in second for individuals ranging from 15 to 34 years of age (Castells *et al*., 2009).

Glioma mainly develops in the brain and is usually classified based on the glial cell type primarily affected by the tumor such as astrocytomas (astrocytes), oligodendrogliomas (oligodendrocytes) or oligoastrocytomas (astrocytes and oligodendrocytes)(Louis *et al.*, 2007). There is a common sense in the literature supporting the idea that neurodevelopment and gliomas are highly related. Supposedly, malignant gliomas originate from uncontrolled neural stem/progenitor cells located in the forebrain which manifest mesenchymal phenotypes and become able to invade surrounding tissues (Bachoo *et al.*, 2002; Carro *et al.*, 2010; Phillips *et al.*, 2006). The typical glioma phenotype, which includes invasion, proliferation, migration, necrosis, and angiogenesis, is triggered by factors present in the adjacent extracellular matrix (ECM) and by the abnormal production of cell surface growth-factor receptors (Giese *et al.*, 2003). High tumor recurrence (95%), in spite of adjuvant therapies such as surgical removal followed by chemotherapy and/or radiotherapy, is attributed to the capacity that putative cancer stem cells have to infiltrate the normal surrounding parenchyma of the main tumor areas, making complete removal rarely possible (Giese *et al.*, 2003).

1.1. Glioma classification and diagnosis

Following the WHO grading system, gliomas can be classified on the basis of histopathological features as (I) slow-growing, circumscribed, benign tumors which can be surgically removed, (II) diffuse slow-growing tumors presenting well-differentiated cells which due to its propensity of infiltration can turn the cancer incurable by surgery, (III) diffuse anaplastic tumors characterized by atypical nuclei and significant proliferative activity with the capacity of infiltrating extensively throughout the brain parenchyma being more fatal than lower-grades, and (IV) neoplasm with more characteristics of malignancy such as predominance of undifferentiated cells, vascular proliferation, atypical nuclei, cellular polymorphism (hence the name *multiforme*), high proliferative activity, and tissue necrosis (Furnari et al., 2007; Louis *et al.*, 2007). The majority of malignant gliomas are specifically subtyped as: *diffuse astrocytoma* (WHO Grade II), *anaplastic astrocytoma* (III), *glioblastoma multiforme* (IV), *oligodendroglioma* (II), *anaplastic oligodendroglioma* (III), *oligoastrocytoma* (II), and *anaplastic oligoastrocytoma* (III).

Diffuse astrocytoma (DA) typically occurs in young adults (30 – 40 years of age) and is commonly located in the supratentoral brain region. DA has an inherent tendency to become anaplastic and, eventually, progress to *glioblastoma multiforme*. DA has an annual incidence rate of 1.4 new cases/1 million population with males being slightly more affected than females (M/F=1.18). The mean survival time after tumor removal ranges between 6 to 8 years but the survival time chiefly depends on its eventual progression to *glioblastoma multiforme*, normally occurring after 4-5 years (Louis *et al.*, 2007).

Anaplastic astrocytoma (AA) typically affects adults (45 – 55 years of age) and is preferentially located in the cerebral hemispheres. AA is considered an intermediate glioma subtype assigned between DA and *glioblastoma multiforme* since after approximately 2 years of its diagnosis it tends to progress to *glioblastoma multiforme* (secondarily). However, AA is also recorded to arise without previous history of a less malignant tumor (primarily). Males are more affected than females (M/F=1.31) and older patients have shorter survival times (Louis *et al.*, 2007).

Glioblastoma multiforme (GBM) is the most frequent glioma subtype (~90%), most commonly affecting people between 45 and 75 years of age (Louis *et al.*, 2007). It preferentially develops in deep white matter regions (Tatard *et al.*, 2010) and is characterized by large histological/molecular heterogeneity (Louis *et al.*, 2007). GBM can arise primarily, without previous history of brain tumor, or secondarily, progressing from lower grade glioma subtypes (DA and AA). Some evidence indicates that *anaplastic oligodendroglioma* might also be a precursor to GBM (DeAngelis, 2009). Approximately 9,000 new cases are estimated to occur in the United States per year with males being predominantly more affected than females (M/F=1.26). GBM is considered a very aggressive cancer and due to its invasive nature cannot be completely removed by surgery. The survival rate for newly diagnosed patients is estimated as 42.4% at 6 months, 17.7% at 1 year, and 3.3% at 2 years, but older patients tend to have even worse prognoses (Louis *et al.*, 2007).

Oligodendroglioma (O) typically affects adults (40 – 45 years of age) and preferentially arises in the cerebral hemispheres. Approximately 5-6% of gliomas are classified as O and tumors of this type frequently present chromosomal deletions in the arms 1p and 19q. Males are slightly more affected than females (M/F=1.1). Estimates show that in the United States close to 50% of patients survive 10 years after diagnosis and surgical resection (Louis *et al.*, 2007).

Anaplastic oligodendroglioma (AO) is reported mainly in adults between 45 and 50 years of age. AO preferentially develops in the frontal lobe and can arise primarily or secondarily (from O) tumors. Approximately 1.2% of the PBT are classified as AO with males being slightly more affected than females (M/F=1.1). Median survival time estimates are quite variable ranging from 1 to 4 years depending on the studied population (Louis *et al.*, 2007).

Oligoastrocytoma (OA) mainly affects adults between 35 to 45 years of age with annual incidence estimated as 1 case per million population. However, reported incidence has increased over the last decade, probably due to improvements in histological recognition techniques. Males are more affected than females (M/F=1.3) and the median survival time is reported to be around 6 years (Louis *et al.*, 2007).

Anaplastic oligoastrocytoma (AOA) accounts for 1-4% of all gliomas and affects patients with median age of 44 years. It is still uncertain if AOA arises only primarily or also secondarily from OA. Males are more affected than females (M/F=1.15) and the median survival time of AOA patients is estimated as 2.8 years (Louis *et al.*, 2007).

Imaging techniques, as magnetic resonance imaging, are noninvasive techniques currently used by clinicians to first diagnoses and assess the glioma and characterize its surrounding tissues (Diehn *et al.*, 2008). Following tumor resection, a biopsy is performed and the diagnosis is confirmed by histological examination (Castells *et al.*, 2009). However, some authors suggest that histopathological classification might be inaccurate sometimes (Castells *et al.*, 2009; Nutt *et al.*, 2003). For example, even though primary and secondary gliomas are considered distinct diseases, they share histological similarities and are distinguishable only if a lower grade lesion is previously recorded. The classification is, therefore, subject to error in situations where secondary gliomas rapidly arise and progress (Nobusawa *et al.*, 2009). Histopathological classification is also strongly susceptible to inter-observer variability and is not easily reproducible (Gravendeel *et al.*, 2009; Nutt *et al.*, 2003). Coons *et al.* (1997) and Giannini *et al.* (2001) show that the concordance in diagnosis based on histological features can range from only 5% to a maximum of 80% among experienced neuropathologists and neurosurgeons. Since the following treatment decisions and therapy response predictions are based on the glioma subtype, a more objective and accurate method of glioma classification is urgently needed (Nutt *et al.*, 2003).

Lately, immunohistochemical markers have been used for classification of gliomas. Examples are the GFAP (glial fibrillary acidic protein) which is always present in astrocytomas but seldom in oligodendrogliomas, and OLIG2 (oligodendrocyte transcription factor 2), a specific oligodendroglioma marker (Furnari *et al.*, 2007). Recent research efforts indicate that molecular profiling of gliomas is also a promising tool for tumor classification and consequently for better outcome prediction (Hayden, 2010). For instance, losses on chromosomes 1p and 19q as well as mutations in CDKN2A, IDH1 and TP53 might be associated with a poor prognosis for oligodendroglioma (O and AO) patients (Cairncross *et al.*, 1998). Also, distinct patterns of gains and losses of chromosomal regions can distinguish primary (EGFR

4

amplification) and secondary (TP53 mutation) GBM lesions (Maher *et al.*, 2006). In addition, Phillips *et al.* (2006) and Verhaak *et al.* (2010) identify three to four GBM subtypes, namely Proneural, Neural, Classical (Proliferative), and Mesenchymal, based on the analysis of microarray expression of signature genes suggesting that high-grade gliomas are associated with the over-expression of "stem cell" genes (proliferative/mesenchymal phenotype) while low-grade gliomas have higher expression of neuronal genes (well-differentiated phenotype).

1.2. Molecular alterations in gliomas

The most common genetic alterations in gliomas occur in biological processes directly involved in cell proliferation, apoptosis, necrosis, angiogenesis, and invasion (Furnari *et al.*, 2007). The RB (retinoblastoma) and p53 pathways are usually found altered in gliomas (Furnari *et al.*, 2007; TCGA, 2008) and are directly associated with cell cycle regulation via the control of mitogenic factors such as MAPK (mitogen-activated protein kinase) and RTK (receptor tyrosine kinase). RB controls cell proliferation by inactivating mitogenic factors of the MAPK cascade which induces the formation of cyclins and their associations with CDK (cyclin-dependent kinase) complexes. Some PI3K (phosphoinositide 3-kinase) family members are important pieces of the RTK pathway and have been reported to have their expression altered in gliomas (Kang *et al.*, 2006). Similarly, the p53 transcription factor regulates the promoters of thousands of genes, among them many CDK complexes, and is best known for its role in activating apoptosis and consequently tumor suppression. Not only TP53 mutations are commonly found in gliomas but also two of its key negative regulators – MDM4 (Mdm2-like p53-binding protein) and CHD5 (Chromodomain-helicase-DNA-binding protein 5) – have their expression frequently altered. In addition, PTEN (phosphatase and tensin homolog), one important tumor suppressor regulated by p53, is also downregulated in 50% of gliomas (Ohgaki *et al.*, 2004). In approximately 40% of GBM samples, the abnormal activation of the cell proliferation pathway might be explained by the overexpression of the EGFR (epidermal growth factor receptor) gene, a membrane receptor.

In addition to the p53 network, another group of molecules called the "death-receptors" are particularly involved in the control of apoptosis. The "death receptors" are membrane receptors such as TNRF1 (tumor necrosis factor receptor 1), TRAIL R1 (TNF-related apoptosis-inducing ligand receptor 1), TRAIL R2 (TNF-related apoptosis-inducing ligand receptor 2), and FAS (TNF receptor superfamily, member 6), which are linked to the activation of caspases and have been reported altered in gliomas (Furnari *et al*., 2007). Likewise, members of the Bcl-2 (B-cell CLL/lymphoma 2) family are frequently altered in gliomas (Furnari *et al*., 2007). Bcl-2 family members are known by their roles in modulating cell death (pro-apoptotic or anti-apoptotic roles), through caspase pathway activation, and by initiating migration and invasion processes in glioma cells (Wick *et al*., 2004). It is suggested that apoptotic/necrotic pathways are, at some degree, interconnected, which is corroborated by the discovery of the protein Bcl2L12 which shares homologies with members of the Bcl-2 family and has the ability to deactivate caspases, hence switching apoptosis to necrosis in gliomas (Nicotera & Melino, 2004).

Angiogenesis, microvascular proliferation of endothelial cells (Stiver *et al*., 2004), is a marked characteristic of both primary and secondary GBM and is one of the many histological features used to identify distinct subtypes of gliomas. The underlying molecular mechanisms of angiogenesis involve many different regulators (especially VEGF, PDGF, IL-8, thrombospondins, endostatin, and interferons) of the hypoxia-induced factor (HIF) pathway (Nyberg *et al*., 2005) and it has been shown that many common mutations found in gliomas — PTEN, EGFR, and CMYC genes — modulate angiogenesis via the HIF pathway (Shchors *et al*., 2006). Likewise, underlying biochemical processes of cell invasion — another hallmark of gliomas — have similarities with the development of SNC (Furnari *et al*., 2007). Many genes involved in cell invasion regulation have already been reported altered in gliomas, including members of the family of metalloproteases (Wang *et al*., 2003a), urokinase-type plasminogen activator (Landau *et al*., 1994), cysteine proteases (McCormick, 1993), IGFBP2 (Wang *et al*., 2003b), ephrins, and P311 (Furnari *et al*., 2007).

1.3. Treatment and prognosis

Applying therapy to glioma patients is still a challenging task since the majority of patients experience undesirable side effects (Furnari *et al.* 2007) and show little improvement after neurosurgery, chemotherapy, and radiotherapy (Freije *et al.*, 2004). Two factors are believed to contribute to the poor prognosis (Cairncross *et al.*, 1998). First, gliomas present remarkable molecular heterogeneity, and consequently therapies targeting a specific pathway attain survival improvement for only a small fraction of the patients. Second, the hemato-encephalic barrier provides tumor cells with the ability to evade chemotherapy and the invasive nature of gliomas makes complete tumor resection almost impossible. In spite of these difficulties, some good results have been obtained in recent years. A third of the oligodendroglioma (O and AO) patients have been shown to be particularly sensitive to combined treatment with procarbazine, lomustine, and vincristine (PCV) due to mutations on chromosomes 1p and 19q (Cairncross *et al.*, 1998). Mixed gliomas (OA and AOA) have also been reported to have a substantial improvement in prognosis after PCV therapy (Cairncross *et al.*, 1998). In addition, glioma patients with unmethylated MGMT promoters become more sensitive to alkylating agents and respond positively to temozolomide treatment (DeAngelis, 2009; Hegi *et al.*, 2005; TCGA, 2008). Freije *et al.* (2004) also report some improvement in a small group of patients treated with irinotecan.

Currently, prognosis is basically based on tumor grading/subtypes and patient age (Furnari *et al.*, 2007). However, the molecular heterogeneity of gliomas along with the variability in response to therapy highlight the importance of cataloging and discovering the intricate genome alterations in glioma tumors which, consequently, will improve prognosis and bring new perspectives of therapy for this disease in the field of small molecule drugs, therapeutic antibodies or RNA interference based factors (Freije *et al.*, 2004). Further, increasing the sample size and using adequate methods to handle peculiarities of molecular data can, in fact, improve survival prediction of glioma patients and better guide disease management.

1.4. Survival prediction using DNA microarray data

During the past few years, a new promising avenue in genomics and molecular biology became available. Microarray data can provide information about thousands of genes simultaneously and, therefore, provide a complete picture of the functioning of whole genomes and, consequently, generate alternatives for cancer treatment (Gentleman *et al.*, 2005). Gene expression profiling using DNA microarray technology has successfully identified molecular subtypes of cancer and revealed associations of gene expression patterns with disease recurrence and survival prognosis of patients (Alizadeh *et al*., 2000; Berchuck *et al*., 2005; Garber *et al*., 2001; Sorlie *et al*., 2001; Yeoh *et al.*, 2002). Survival prediction is often formulated in terms of categorical outcomes (e.g. 'poor' *vs*. 'good' prognosis) which might be useful for guiding decisions about cancer management and treatment (Ross, 2009). However, due to a large heterogeneity observed within prognostic classes, prediction of time to a clinical event/occurrence can be poor. Improvement of survival prediction accuracy can be attained by relating time-to-event to gene expression profiles directly. This requires specific survival analysis methods to account for the presence of censored outcomes, such as the (multivariable) Cox proportional hazards (CPH) model (Cox, 1972) and the accelerated failure-time model (AFT) (Klein & Moeschberger, 1997).

In spite of their wide use in other settings, these standard multivariable survival methods cannot be directly applied to clinical outcome prediction using gene expression data since the number of covariates (genes) under investigation is considerably larger than the number of samples (patients) -- this is called the "*large p, small n problem*" (West, 2003). Many different strategies have been employed to solve this high dimensionality problem. For example, clustering techniques have been applied for grouping correlated sets of genes; the average expression level of each cluster is then used as a covariate (metagene) in the survival model (D'haeseleer, 2005). Likewise, linear combinations of covariates obtained by partial least squares (Nguyen & Rocke, 2002; Park *et al.*, 2002) or the principal components of the design matrix (Li & Gui, 2004) have been used as explanatory variables in survival regression models. In addition, some authors propose the use of penalized versions of the CPH model, L1-penalized (Lasso regression) and L2-penalized (ridge regression), for estimating parameters while simultaneously

performing variable selection (Gui & Li, 2005; Park *et al.*, 2002; Tibshirani, 1997; Zou & Hastie, 2005). Similarly, Huang *et al.* (2006) and Datta *et al.* (2007) developed penalized variants of the AFT model for fitting models in high-dimensional data settings. Bayesian techniques for variable selection have also been developed for Weibull and CPH models (Lee & Mallick, 2003) as well as the AFT model (Sha *et al.,* 2006).

While these strategies address the high-dimensionality problem with some degree of success, they fail to incorporate complex interactions between genes, since the genes are modeled in an additive and linear manner. Ensemble methods like bagging (Breiman, 1996), boosting (Friedman, 2001), and random forests (Breiman, 2001) are very flexible alternatives for accommodating variable interactions and are notably more stable in high-dimensional settings (Breiman, 1996, 2001). Tree ensemble methods are based on a recursive partitioning strategy, where a tree is built by successively splitting the observations into binary nodes. If X_p denotes the splitting covariate, one node contains all observations with $X_p \leq \rho$, while the other node contains all observations with $X_p > \rho$. The covariate X_p and the threshold ρ are selected based on a splitting criterion and the growing or pruning of trees is based on a stopping criterion (see details in the following sections).

Because ensemble methods use a linear combination of trees to fit data variation where each tree fits part of the data, these methods have been shown to have high predictive accuracy (Lee *et al.*, 2005). These ensemble methods were originally developed for modeling binary or continuous responses. Extensions for modeling survival data, often called survival ensembles (Hothorn *et al.*, 2006), address the censoring problem by growing relative risk forests (Ishwaran *et al.,* 2004), by imputing censored observations (Ishwaran *et al.,* 2008), or by using a Kaplan-Meier curve aggregation procedure to predict survival of a new observation (Hothorn *et al.*, 2004). In general, survival ensemble methods estimate a survival function for each terminal node of the tree, weighing censored observations differently and then, prediction is performed by dropping down the tree a new observation (Hothorn *et al.*, 2006). A different approach proposed by Schmid & Hothorn (2008) estimates both the predictor function of the AFT model

and simultaneously the scale parameter, so that a boosting algorithm can be applied to minimize a pre-defined loss function. Even though Bayesian estimation has been shown to improve the predictive performance of tree models with nominal or continuous responses (Chipman *et al.*, 1998; Denison *et al*., 1998; Pittman *et al*., 2004), Bayesian survival ensembles are still limited to the study of Clarke & West (2008) who proposes tree-based Weibull models for outcome prediction of advanced stage ovarian cancer.

2. Objectives

The hypothesis of this research is that molecular profiling of brain tumors coupled with clinical information might better predict patient survival time. In addition, we hypothesize that the use of survival methods which incorporate interactions and work well in high-dimensional settings can improve prediction accuracy for survival times. We evaluate these hypotheses by building models for survival prediction of glioma patients (section 3.1).

First, a predictive Random Forest model is built for binary outcomes (i.e. short *vs*. long-term survival) and a small subset of genes whose expression values can be used to predict survival time is selected (section 3.2). Random Forest predictive accuracy along with a list of the most important genes used for prediction are shown in section 4.1. Following, a new statistical methodology is developed for predicting time-to-death outcomes using *Bayesian ensemble trees*. Due to the large heterogeneity observed within prognostic classes obtained by the Random Forest model, prediction can be improved by relating time-to-death outcomes to the gene expression profiles directly. The new approach is based on the ensemble "sum-of-trees" model (Chipman *et al.*, 2006) and hence is defined by a likelihood and a prior. A fully Bayesian hierarchical approach is specified with uncertainty in estimation being propagated at each stage of hierarchy to make predictions. The new methodology is illustrated using three popular survival models - CPH (section 3.3.1), Weibull (section 3.3.2), and AFT models (section 3.3.3). The new approach is unique as it overcomes the lack of conjugacy using a latent variable formulation to model the covariate effects and, as a result, model fitting becomes efficient and computationally less expensive (section 3.3.4). The new approach is non-parametric and incorporates additive and interaction effects between genes which results in high predictive accuracy as compared to other methods. In addition, it provides a model-free way to select important predictive prognostic markers based on controlling false discovery rates (section 3.5). The predictive accuracy of the new method is compared (section 4.2) with baseline reference survival methods reviewed by van Wieringen *et al.* (2009) using the breast cancer data set of Van't Veer *et al*. (2002). The results of the new methodology applied to the brain tumor data set are

11

presented in section 4.3. A closing discussion compares results obtained by Random Forest and Bayesian ensemble methods under the biological/clinical perspectives and stresses the statistical advantages and disadvantages of the new methodology in the context of DNA microarray data analysis (section 5).

3. Methods

3.1. The data set

A set of gene expression profiles for brain tumor patients is used here to identify molecular and genetic signatures which could be of prognostic value. The data set contains gene expression measurements, patient age, tumor grade, and survival information for 734 patients obtained from nine different cancer treatment institutions (Table 1).

The survival time after diagnosis ranges from 1 to 698 weeks with 15% of the observations being censored (Figure 1). Survival time was discretized into short-term survival (STS ≤ 24 months — 70% of patients) *vs.* long-term survival (LTS > 24 months — 30% of patients) before using the Random Forest classification method. Gene expression values of 11,911 genes are obtained using three different Affymetrix microarray chips (142 from HT-U133A, 355 from U133A, 237 from U133Plus2 -- Figure 2 and Table 1). The annotation was reformulated using a customized CDF file organized by the BrainArray Group, Department of Psychiatry, University of Michigan (see more details in Dai *et al.*, 2005 and at the group's website: *brainarray.mbni.med.umich.edu/*). The data was pre-processed using Batch normalization available in the JMP Genomics SAS® software and then quantile normalization (*affy* package in R) in order to account for batch effects. The data have not been made publicly available by the time when this thesis was written but, upon request, the data can be obtained from Dr. Erik P. Sulman (Department of Radiation Oncology, MD Anderson Cancer Center).

Instead of the original set of 11,911 genes, we work with a reduced set of 142 *metagenes* which are obtained by applying an unsupervised clustering algorithm, Gene Shaving (Do *et al.,* 2003; Hastie *et al.,* 2001). Gene Shaving identifies the largest principal component, clusters the genes highly correlated with it and shaves out the less correlated ones. Then the following largest principal component is found and the procedure repeats until around 85% of the genes were shaved out. The metagenes are then constructed from the clusters as signed average of its members. As a result, 1,623 genes were selected and grouped in 142 metagenes ranging in sizes from 2 to 87 genes (Figure 3).

Even though the methods treated here are capable of handling this high-dimensional setting (11,911 x 734), we decide to work with a lower dimensional data set (144 x 734) to be able to compare our results with multivariable versions of the linear competing methods using the same set of predictors. In addition to the 142 metagenes, clinical covariates such as patient age and histopathological tumor grade (II, III, or IV) were also added as covariates in the survival model. For the survival ensemble methods, a cross-validation procedure was performed with the data being randomly split 10 times into training and test sets with a 2:1 ratio. We first build the predictor using the training set and then assess and compare the performance of different methods using evaluation measures calculated for the test set.
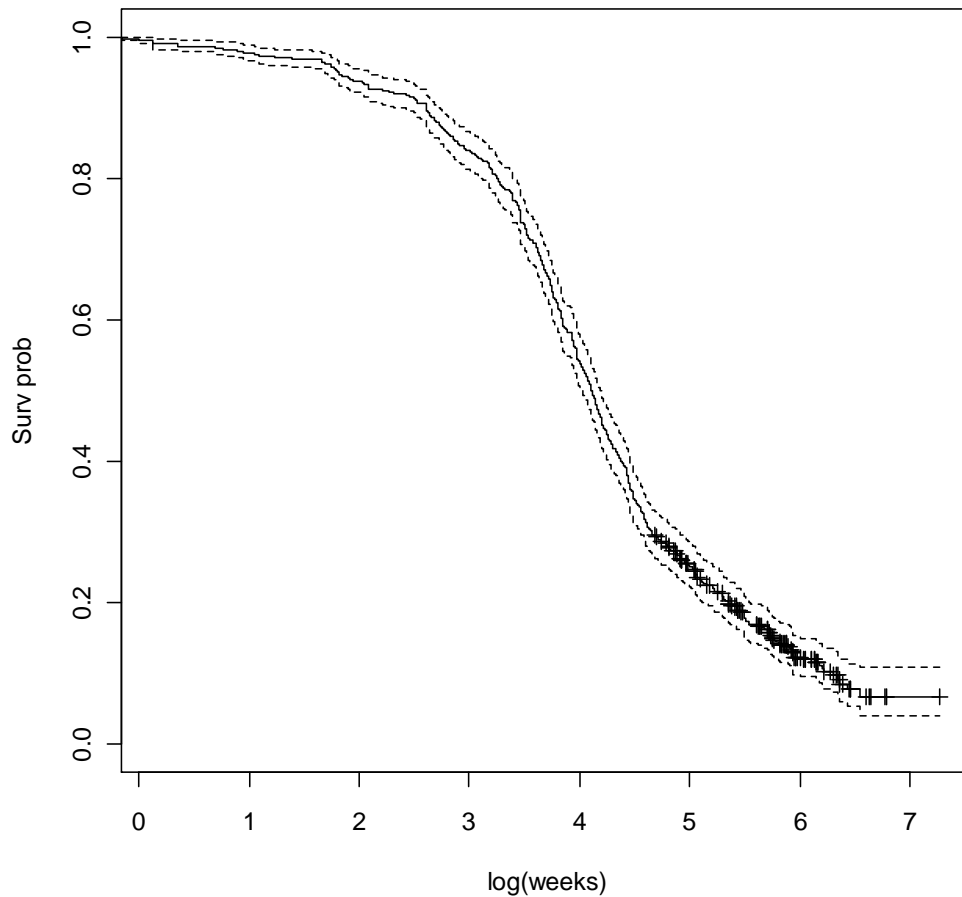
Figure 1: Kaplan-Meier survival curve (solid line) for all 734 glioma patients along with the 95% confidence interval (dashed lines). Plus signs indicate right-censored observations. Horizontal axis shows the natural log of survival times in weeks.
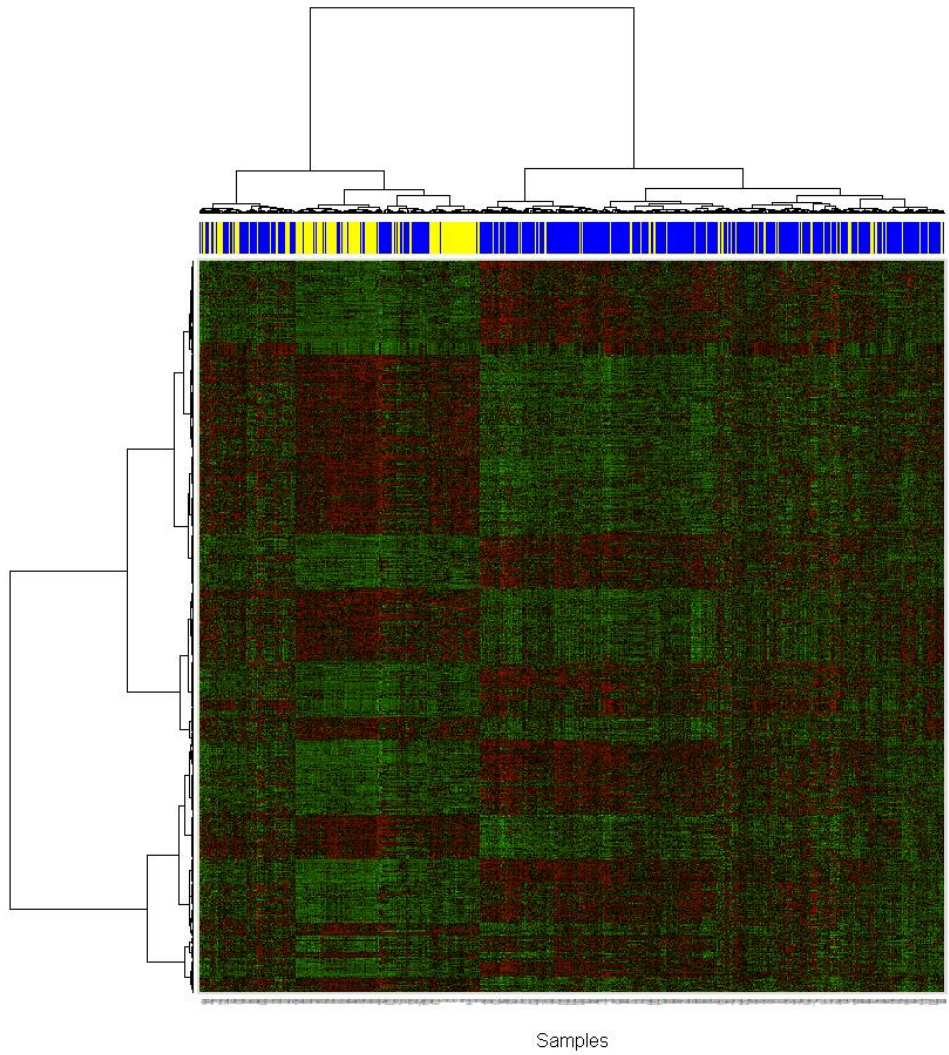
Figure 2: Heatmap of the top 2,000 genes selected by a t-test comparing averages of expression values for STS and LTS groups of glioma patients. Green color spots represent under-expressed values and red color spots represent over-expressed values. Samples are depicted in the horizontal axis and genes in the vertical axis. Dendrograms were obtained by *"Ward"* method. Blue labels at the top of the figure indicate STS patients while yellow labels indicate LTS patients.
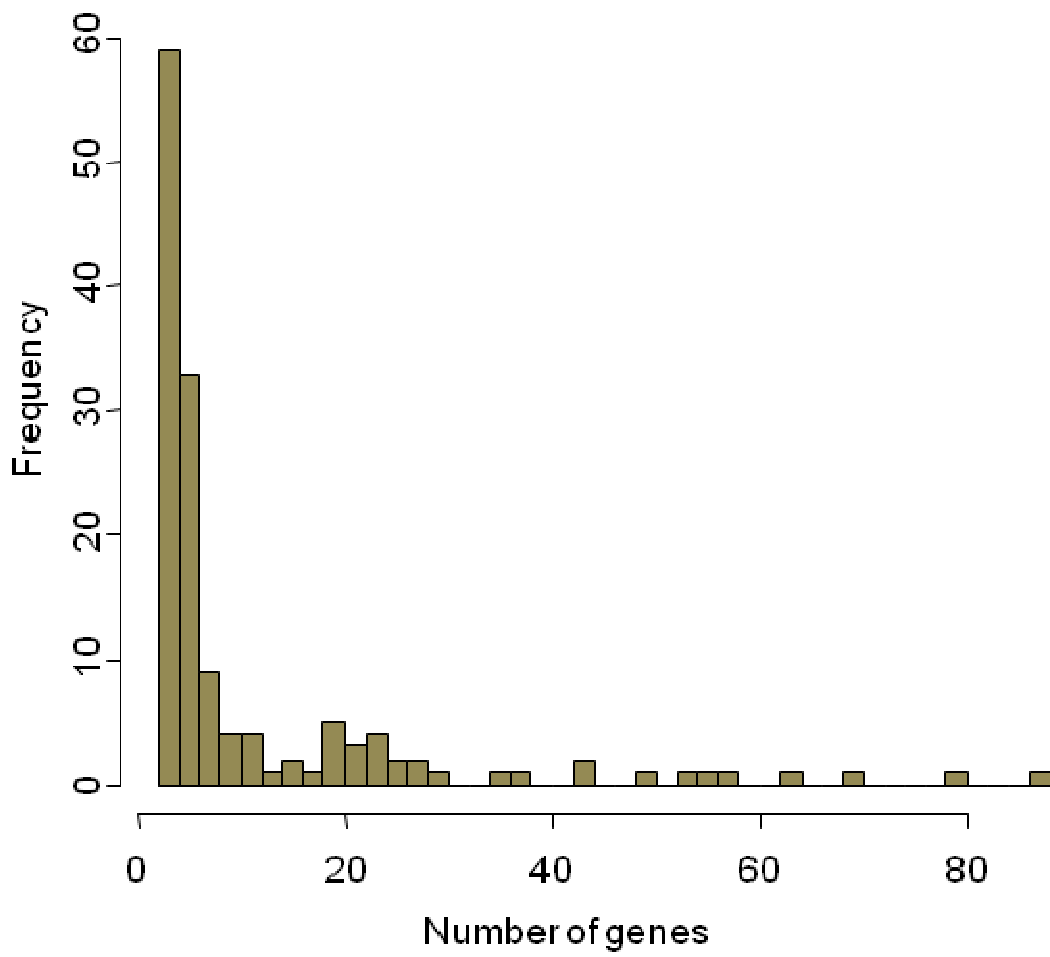
Figure 3: Distribution of sizes of the 142 metagenes obtained by Gene-Shaving.

Table 1. Cancer center providers and distributions of array types and survival times of glioma samples.

Institution	Total of samples	Array Type			Survival	
		HT-U133A	U133A	U133Plus2	STS	LTS
UCLA	166	0	157	9	130	36
MDA	145	32	75	38	74	71
TCGA	110	110	0	0	97	13
Henry Ford	104	0	0	104	58	46
EORTC	65	0	0	65	54	11
Collins	61	0	61	0	46	15
UCSF	31	0	31	0	24	7
Duke	31	0	31	0	28	3
Belgium	21	0	0	21	3	18
Total	**734**	**142**	**355**	**237**	**514**	**220**

3.2. Random Forest

Random Forest (RF) is an ensemble tree-based method which is appropriate for classification of DNA microarray data because it handles high-dimensionality well, it is able to perform model-free variable selection, and it is flexible enough to incorporate interactions between genes, which confers on it a highly predictive accuracy when compared to alternatives (Diaz-Uriarte & Alvarez de Andres, 2006). The Random Forest (RF) model is based on a collection of individual tree-structured predictors $\{l(\mathbf{X}, \square_i)\}$, where $\mathbf{X}$ is the data and $\square_i$ represents a subset of randomly selected covariates chosen with replacement to build the trees (Breiman, 2001). Each tree votes for the most popular class at input $\mathbf{x}$ and the size of i has to be set a priori. If i is set equal to the total number p of covariates the bagging algorithm is obtained (Breiman, 1996). RF is applicable for classification or regression and does not assume any particular stochastic model. The prediction error within sample converges to a limit as the number of trees increases (Breiman, 2001). Trees are grown as large as possible and are not pruned as in other tree-based methods. A bootstrap sample of samples from the original data set is used to build each tree in the forest and the samples not sampled are referred to as out-of-bag (OOB) observations. OOB prediction is obtained by the majority of the votes involving only those trees that did not contain the corresponding sample. The frequency that a covariate is used to build the trees defines a measure of variable importance which is used here for feature selection (Shi & Horvath, 2006).

3.3. Survival ensembles

We denote the observed data for the ith patient ($i = 1, ..., n$) as t_i, the survival time, along with δ_i, the event indicator function, where $\delta_i = 0$ if data is right censored and $\delta_i = 1$ if it is not. In addition to the survival response, the p-dimensional vector of covariates (genes/probes) potentially associated with the ith patient survival time, $\mathbf{X}_i$, is also available. Let $\mathbf{t} = (t_1, ..., t_n)$ denote the vector of survival times and let $\mathbf{X}_{n \times p}$ (samples by metagenes) denote the matrix of gene expression data. In the following, we develop survival distribution models which aids to predict the survival time of a new patient with covariates $\mathbf{X}_{new}$.

Modeling of survival data usually proceeds in two steps: (1) specification of a sampling distribution $p(\mathbf{t}|f(\mathbf{X}))$, conditional on a function of the covariates $f(\mathbf{X})$, such as modeling either the hazard function (as in CPH models) or the survival time directly (as in Weibull and AFT models) and (2) the regression function $f(\mathbf{X})$ which models the covariate effects. Usually, for computational convenience it is assumed that the covariates are linearly and independently related to survival, such that $f(\mathbf{X}) = X'\beta$ where β is a vector of p unknown regression coefficients that captures the covariate effects on survival time or hazard. There are two drawbacks of this approach. First the linear and independent assumption is a restrictive one. Second, and more importantly, in high throughput studies such as gene expression data the problem becomes much more complex when p, the dimension of $\mathbf{X}$, is very large, possibly larger than the sample size n. This makes the estimation of β unstable and the high-dimensionality problem is exacerbated if interactions between covariates are considered. Dimension reduction approaches such as feature selection or partial least squares alleviate the problem to a certain degree. However, these methods are based on linear relationships between the response and the covariate which may not be very realistic. If the actual f is nonlinear, these models may fail to produce reasonable prediction due to lack of flexibility. We propose to model $f(\mathbf{X})$ in a flexible manner using ensemble methods that not only

accommodate nonlinear effects but also incorporates interactions of covariates to estimate the effects on survival time as we describe below.

3.3.1 Ensemble-based Proportional Hazards Regression

The proportional hazards model (Cox, 1972) is one of the most popular survival models in the statistical literature. Rather than modeling time-to-event directly, it models the hazard function $h(t)$, at any time t as

$$h(t|x) = h_o(t)exp(\omega),$$

where $h_o(t)$ is the baseline hazard function and ω is an unknown function modeling the associated latent covariate effect. The joint conditional survival function of **t** in the CPH model can then be written as

$$S(\mathbf{t}|\omega, \Lambda) = exp\left(-\sum_{i=1}^{n} \Lambda(t_i) exp(\omega_i)\right),$$

where Λ represents the cumulative hazard function. The associated complicated form of the likelihood makes it impossible to express conditional distributions of the parameters (ω, Λ) in closed forms (Ibrahim *et al*., 2001). As a result, the drawing of posterior distributions requires the sampling of all model parameters using complex MCMC procedures at each iteration, making the process computationally intensive with potential poor mixing, especially in high-dimensional settings.

We simplify the joint likelihood in two ways. First, for the cumulative hazard function we follow the approach in Kalbfleisch (1978) by specifying a Gamma process prior for Λ, such that

$$\Lambda \sim \mathcal{GP}(a\Lambda^*, a),$$

where Λ^* is the mean process and a is a weight parameter about the mean with $\Lambda(t) \sim Gamma(a\Lambda^*(t), a)$. The utility of using the Gamma process prior is that the Λ vector can be analytically integrated out such that the marginal likelihood conditional on ω can be written as

$$L(\mathbf{t}|\omega) = \exp\left(-\sum_{i=1}^{n} aW_i\Lambda^*(t_i)\right)\prod_{i=1}^{n}(a\lambda^*(t_i)W_i)^{\delta_i},$$

where δ_i is the indicator for event, $V_i = \sum_{l\in R(t_i)}^{n} \exp(\omega_l)$, $i = 1, \dots, n$, $R(t_i)$ is the set of individuals at risk at time t_i, and $W_i = \{1 - \exp(\omega_i)/(a + V_i)\}$.

Second, we modify the model by treating the ω_i's as random latent variables, conditional on which the t_i's are independent of $\mathbf{X}_i$'s by the following factorization: $p(t_i|\omega_i)p(\omega_i|\mathbf{X}_i)$. This latent variable feature allows us to elicit conjugate priors for ω_i's, making the sampling from full conditionals fast and more efficient. Specifically, we assume a Gaussian process on $p(\omega_i|\mathbf{X}_i)$, such that $\omega_i = f(\mathbf{X}_i) + \epsilon_i$, where $f(\mathbf{X}_i)$ is the regression function and ϵ_i are residual random effects assumed to be distributed $Normal(0, \sigma^2)$. The residual random effects ϵ_i accounts for the unexplained sources of variation in the data, most probably due to explanatory variables (genes) not included in the study (Lee & Mallick, 2003).

Therefore, the full conditional of the *Ensemble-based Proportional Hazards regression* is written as

$$p(\omega|\Lambda,\mathbf{X},\sigma^2,\mathbf{t},\delta) \propto \exp\left(-\sum_{i=1}^{n} aW_i\Lambda^*(t_i)\right)\prod_{i=1}^{n}(a\lambda^*(t_i)W_i)^{\delta_i} \times \exp\left(-\frac{1}{2\sigma^2}(\omega - f(\mathbf{X}))'\mathbf{I}(\omega - f(\mathbf{X}))\right)$$

$$\log\left(p(\omega|\Lambda,\mathbf{X},\sigma^2,\mathbf{t},\delta)\right) \propto -\sum_{i=1}^{n} aW_i\Lambda^*(t_i) + d\sum_{i=1}^{n}\log(a\lambda^*(t_i)W_i) + MVN(f(\mathbf{X}),\sigma^2\mathbf{I}),$$

and the joint posterior survival function defined as

$$S(\mathbf{t}|\omega) = \left(\frac{a}{a+\exp(\omega)}\right)^{a\Lambda^*},$$

We approximate $f(\cdot)$ using a tree-based ensemble method to model not only the non-linear effects of the genes but also account for the high-dimensionality of the data. We use the "sum-of-trees" approach of Chipman *et al.* (2006) (CGM, henceforth) called the Bayesian Additive Regression Trees (BART) model as our candidate choice due to its excellent predictive performance on a variety of data sets. Compared to other ensemble methods, BART is also preferable because it is explicitly defined in terms of a full probability model, i.e. with likelihood and priors, and, therefore, a full Bayesian hierarchical approach can be implemented for estimation of all relevant uncertainties. We present a brief review of BART below and refer to CGM for more details.

Let **T** represent a single decision tree containing both internal and terminal nodes. Internal nodes of the tree are grown through recursive partitions of the data using splitting rules. Splitting rules produce binary splits of the data and are defined in terms of splitting variables and cutoff values. Dropping an individual with covariates $\mathbf{x}_i$ down the tree assigns it to a terminal node according to the tree splitting rules. Let each tree be indexed by B terminal nodes and define $\boldsymbol{\mu}=(\mu_1, \dots, \mu_B)$ as the vector of averages μ_b of individuals assigned to the same node b, where b = 1,…,B. So each observation can be mapped by a function f such that $f(\mathbf{x}_i) = g(\mathbf{x}_i, \mathbf{T}, \boldsymbol{\mu})$.

Since BART is a "sum-of-trees" model f can be approximated by

$$f(\mathbf{X}) = \sum_{m=1}^{M} g(\mathbf{X}, \mathbf{T}_m, \boldsymbol{\mu}_m),$$

where M is the total number of trees. Compared to single tree models, BART is more flexible since several trees incorporate the additive effects and, consequently, improve estimation.

We note that the number of regression trees M set for the tree-ensemble methods dictates how often a covariate will be selected to be part of the model. CGM show that setting a relatively small number of trees benefits the variable selection procedure since variables compete with each other to improve fit and therefore, relevant predictors should appear more frequently in the tree model. Because

we are interested in exploring the BART variable selection feature, we performed a previous study examining the trade-off between total number of trees and computational time and found out that setting $M = 40$ is a feasible number for all survival ensemble methods described in this section (Figure 4).

To complete the full Bayesian hierarchical formulation of our ensemble-based proportional hazard regression model we need to specify the following priors: $p(\omega|f(\mathbf{X}), \sigma^2)$, $p(\sigma^2|\boldsymbol{\phi})$, and $p(f|\boldsymbol{\phi})$ where $\boldsymbol{\phi} = (T_1, \mu_1, \ldots, T_m, \mu_m)$ represents the tree-specific parameters. Our prior for $p(f)$ is of the form,

$$p(f) = \prod_{m=1}^{M} p(\mathbf{T}_m, \boldsymbol{\mu}_m) = \prod_{m=1}^{M} p(\mathbf{T}_m) p(\boldsymbol{\mu}_m|\mathbf{T}_m),$$

where the second equality is obtained by recursively conditioning on the terminal nodes.

We follow CGM and define $p(\mathbf{T}_m)$ by three factors: (i) the distribution on the splitting variable assignments at each interior node is a uniform prior over all available variables, (ii) the distribution on the splitting rule assignment in each interior node, conditional on splitting variable is a uniform distribution over the set of available splitting values, and (iii) the probability that a node at depth d is nonterminal is given by $c(1 + d)^{-e}$, where $c \in (0, 1)$ and $e \in (0, \infty)$ are fixed parameters controlling the size of the tree. Following CGM we set $c = 0.95$ and $e = 2$ to give prior probabilities of $(0.05, 0.55, 0.28, 0.09, 0.03)$ for trees to have $(1, 2, 3, 4, \geq 5)$ terminal nodes, respectively. As in CGM we assume i.i.d conjugate normal priors for $p(\boldsymbol{\mu}_m|\mathbf{T}_m)$. Assigning prior distributions for the set of tree parameters $\mathbf{T}$ and $\boldsymbol{\mu}$ constrains the sizes of the trees avoiding the model being populated by non-informative covariates. This imposed variation in tree sizes grants BART flexibility for accommodating main effects as well as their interactions of different orders (more than one splitting rule) which results in a better predictive performance obtained by BART when compared to competing methods such as random forest and boosting algorithms. To complete the prior formulations we assume a conjugate inverse chi-square distribution on σ^2 as $[\sigma^2] \sim \nu\eta/\chi_\nu^2$, with ν being a data-determined fixed hyperparameter.

Concisely, the complete hierarchical Bayesian model for ensemble-based CPH model can be written as,

$$[\mathbf{t}|\omega] \sim L(\mathbf{t}|\omega),$$

$$[\omega_i|f(\mathbf{X}_i), \sigma^2] \sim Normal(f(\mathbf{X}_i), \sigma^2),$$

$$f(\mathbf{X}_i) \sim Tree(\boldsymbol{\phi}),$$

$$\sigma^2 \sim \chi_v^2.$$

3.3.2. Ensemble-based Weibull Regression

The Weibull model is a parametric model used extensively in describing lifetimes and can be reparameterized both as a CPH as well as an AFT model (Klein & Moeschberger, 1997). The Weibull distribution is indexed by a shape parameter τ and scale parameter ψ_i and it models the probability of survival at time t_i for patient i as,

$$f(t_i|\tau, \psi) = \tau\psi_i exp\left(-\psi_i t_i^{\tau}\right)\mathbf{I}_{(t_i>0;\tau>0;\psi_i>0)}.$$

Reparameterizing the scale parameters as $\omega_i = \log\left(\psi_i\right)$ the Weibull likelihood can be written as,

$$f(t_i|\tau, \omega_i) = \tau t_i^{\tau-1} exp\left(\omega_i - exp(\omega_i t_i^{\tau})\right)\mathbf{I}_{(t_i>0;\tau>0)},$$

and the survival function as $S(t_i|\tau, \omega_i) = exp(-exp(\omega_i)t_i^{\tau})$. Letting $\Delta = \sum \delta_i$ represent the number of censored observations, the joint likelihood function for the parameter τ and the vector of parameters $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)$ becomes

$$L(\tau, \boldsymbol{\omega}|\mathbf{X}, \mathbf{t}, \delta) = \prod_{i=1}^{n} f(t_i|\tau, \omega_i)^{\delta_i} S(t_i|\tau, \omega_i)^{1-\delta_i}$$

$$L(\tau, \boldsymbol{\omega}|\mathbf{X}, \mathbf{t}, \delta) = (\tau)^{\sum \delta_i} (t_i)^{\sum \delta_i(\tau-1)} exp\left(\sum_{i=1}^{n}(\delta_i\omega_i) - \sum_{i=1}^{n}\delta_i exp(\omega_i)t_i^{\tau} - \sum_{i=1}^{n}(1-\delta_i)exp(\omega_i)t_i^{\tau}\right)$$

$$L(\tau, \boldsymbol{\omega}|\mathbf{X}, \mathbf{t}, \delta) = \tau^{\Delta}(t_i)^{\sum \delta_i(\tau-1)} exp\left(\sum_{i=1}^{n}(\delta_i\omega_i) - \sum_{i=1}^{n} exp(\omega_i)t_i^{\tau}\right)$$

$$L(\tau, \boldsymbol{\omega}|\mathbf{X}, \mathbf{t}, \delta) = \tau^{\Delta} exp\left(\sum_{i=1}^{n}(\delta_i\omega_i) + \sum_{i=1}^{n}\delta_i(\tau-1)log(t_i) - \sum_{i=1}^{n} exp(\omega_i)t_i^{\tau}\right)$$

$$L(\tau, \boldsymbol{\omega}|\mathbf{X}, \mathbf{t}, \delta) = \tau^{\Delta} exp\left(\sum_{i=1}^{n}(\delta_i\omega_i + \delta_i(\tau-1)log(t_i)) - \sum_{i=1}^{n} exp(\omega_i)t_i^{\tau}\right).$$

For convenience, we let $\theta = log(\tau)$ and write the conditional distribution of the vector $\boldsymbol{\omega}$ as

$$p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{t}, \delta, \theta) \propto exp\left(\theta\Delta + \sum_{i=1}^{n}\left(\delta_i\omega_i + \delta_i(e^{\theta} - 1)log(t_i)\right) - \sum_{i=1}^{n} exp(\omega_i)t_i^{e^{\theta}}\right)$$

$$\times exp\left(-\frac{1}{2\sigma^2}(\omega - f(\mathbf{X}))'\mathbf{I}(\omega - f(\mathbf{X}))\right).$$

Since ω_i's are conditionally independent, we conveniently draw their posterior distributions componentwise from

$$p(\omega_i|\omega_{l\neq i}, \mathbf{X}, \mathbf{t}, \delta, \theta) \propto exp\left(\theta\Delta + \delta_i\omega_i + \delta_i(e^{\theta} - 1)log(t_i) - exp(\omega_i)t_i^{e^{\theta}}\right)$$

$$\times exp\left(-\frac{1}{2\sigma^2}(\omega_i - f(\mathbf{X_i}))^2\right).$$

Following, we also use Metropolis-Hastings to draw the conditional of θ from

$$p(\theta|\boldsymbol{\omega}, \mathbf{X}, \mathbf{t}, \delta) \propto exp\left(\theta\Delta + \sum_{i=1}^{n}\left(\delta_i\omega_i + \delta_i(e^\theta - 1)log(t_i)\right) - \sum_{i=1}^{n}exp(\omega_i)t_i^{e^\theta}\right) \times e^{\theta(\tau_o-1)}e^{-k_o e^\theta}e^\theta.$$

As in the previous section, we model the covariate effects ω_i's as latent variables as $Normal(f(\mathbf{X}_i), \sigma^2)$, with f being modeled using BART. We complete our hierarchical model assigning a conjugate gamma prior on τ as $Gamma(\tau_o, k_o)$, with fixed but vague hyperparameters. Thus our ensemble-based Weibull regression model can be written as,

$$[t_i|\tau, \omega_i] \sim Weibull(\tau, \omega_i),$$

$$[\tau] \sim Gamma(\tau_o, k_o),$$

$$[\omega_i|f(\mathbf{X}_i), \sigma^2] \sim Normal(f(\mathbf{X}_i), \sigma^2),$$

$$f(\mathbf{X}_i) \sim Tree(\boldsymbol{\phi}),$$

$$\sigma^2 \sim \chi_\nu^2.$$

3.3.3. Ensemble-based Accelerated Failure Time Model

The AFT model is a parametric survival model that assumes that the individual survival time t_i depends on the multiplicative effect of an unknown function of covariates $f(\mathbf{X}_i)$ over a baseline survival time α. The AFT model (on log-scale) can be written as,

$$log(t_i) = \alpha + f(\mathbf{X}_i) + \epsilon_i, \qquad i = 1, \dots, n$$

where f captures the covariate effects affecting the (log) survival time directly.

We assume the random errors ϵ_i's are normally distributed, however other distributions such as extreme-value or t distribution (Klein & Moeschberger, 1997) can easily be adopted. Note that under an extreme-value distribution, the AFT model is equivalent to the Weibull model described previously.

As before, let ω be a latent variable such that $\omega_i = f(\mathbf{X}_i) + \epsilon_i$, where ϵ_i's are i.i.d $Normal(0, \sigma^2)$. The AFT model can now be expressed as

$$\begin{cases} \log(t_i^*) = \alpha + f(\mathbf{X}_i), & \text{if} \quad \delta_i = 1, \\ \log(t_i^*) > \alpha + f(\mathbf{X}_i) & \text{if} \quad \delta_i = 0, \end{cases}$$

where α is assigned a conjugate normal prior distribution as $Normal(\alpha_o, \alpha_c)$, where α_o and α_c are fixed hyperparameters.

After estimating ω, we define $r_i = \log(t_i^*) - \omega_i$ where $[r|\alpha, \sigma^2] \sim Normal(\alpha, \sigma^2)$. Now, we specify a conjugate prior for α as $[\alpha|\alpha_o, \alpha_c] \sim Normal(\alpha_o, \alpha_c)$, which makes the posterior distribution of $[\alpha|r, \sigma^2, \alpha_o, \alpha_c]$ be an updated $Normal(\alpha^*, \sigma^*)$, where $\alpha^* = \frac{\sigma^2 \alpha_o + \alpha_c \sum r_i}{\sigma^2 + n\alpha_c}$ and $\sigma^* = \sqrt{\frac{\alpha_c \sigma^2}{\sigma^2 + n\alpha_c}}$.

Therefore, the censored survival times ($\delta_i = 0$) are sampled from univariable normal distributions $Normal(\alpha + \omega_i, \sigma^2)$ truncated at $\log(t_i^*)$.

Thus our ensemble-based AFT model can be succinctly written as,

$$[t_i|\alpha, \omega_i] \sim Normal(\alpha + \omega_i, \sigma^2),$$

$$[\alpha|\alpha_o, \alpha_c] \sim Normal(\alpha_o, \alpha_c),$$

$$[\omega_i|f(\mathbf{X}_i), \sigma^2] \sim Normal(f(\mathbf{X}_i), \sigma^2),$$

$$f(\mathbf{X}_i) \sim Tree(\boldsymbol{\phi}),$$

$$\sigma^2 \sim \chi_v^2.$$
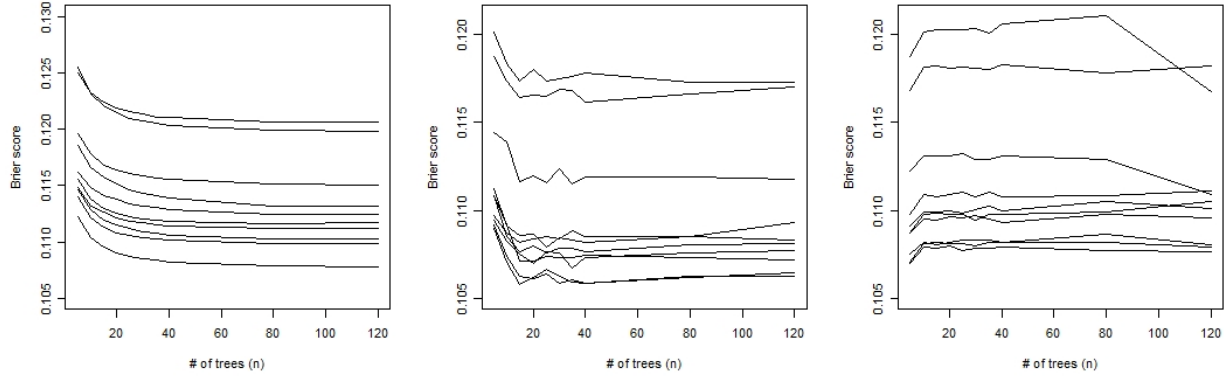
Figure 4: Size of trees. Figure shows the Brier Score for the test data depending on the number of trees set for the BART model. Left plot - AFT-TREE; center - WEI-TREE; Right - CPH-TREE. Each line represents one training/test split of data.

3.3.4. Model fitting via MCMC

We use Markov Chain Monte Carlo (MCMC, Gilks *et al.*, 1996) algorithms to generate samples from our posterior distributions. The specific drawing scheme for the CPH model uses a Gibbs sampler to estimate the set of parameters $(\boldsymbol{\omega}, \boldsymbol{\phi}, \sigma^2)$. The Gibbs sampling method iterates $k = 1, \dots, K$ times through the following steps:

(1) Update $\boldsymbol{\phi}$ using the Bayesian backfitting MCMC algorithm described in CGM;

(2) Update $[\sigma^2 | \boldsymbol{\phi}]$ using a Gibbs sampler;

(3) Update $[\omega_i | \boldsymbol{\phi}, \sigma^2]$, where $i = 1, \dots, n$, using for each ω_i a Metropolis-Hastings procedure with a proposal density $q(\omega_i, \omega_i^*)$ which generates moves from the current state ω_i to a new state ω_i^*. The probability of accepting the change is given by

$$\pi_{\omega_i} = min\left(1, \frac{\mathrm{p}(\omega_i^* | \omega_{l \neq i}, \mathbf{X}, \mathbf{t}) \mathrm{q}(\omega_i^*, \omega_i)}{\mathrm{p}(\omega_i | \omega_{l \neq i}, \mathbf{X}, \mathbf{t}) \mathrm{q}(\omega_i, \omega_i^*)}\right).$$

The posterior distributions of the Weibull model parameters $(\boldsymbol{\omega}, \boldsymbol{\phi}, \tau, \sigma^2)$ are obtained in a similar manner:

(1) Update $\boldsymbol{\phi}$ using the Bayesian backfitting MCMC algorithm described in CGM;

(2) Update $[\sigma^2|\boldsymbol{\phi}]$ using a Gibbs sampler;

(3) Update $[\omega_i|\boldsymbol{\phi}, \tau, \sigma^2]$ componentwise, where $i = 1, \dots, n$, using for each ω_i a similar Metropolis-Hastings procedure with probability of accepting the change given by

$$\pi_{\omega_i} = min\left(1, \frac{\mathrm{p}(\omega_i^*|\omega_{l \neq i}, \mathbf{X}, \mathbf{t}, \delta, \tau) \mathrm{q}(\omega_i^*, \omega_i)}{\mathrm{p}(\omega_i|\omega_{l \neq i}, \mathbf{X}, \mathbf{t}, \delta, \tau) \mathrm{q}(\omega_i, \omega_i^*)}\right).$$

(4) Update $[\tau|\boldsymbol{\omega}, \boldsymbol{\phi}, \sigma^2]$ also using the Metropolis-Hastings procedure with acceptance probability

$$\pi_{\tau} = min\left(1, \frac{\mathrm{p}(\tau^*|\boldsymbol{\omega}, \boldsymbol{\phi}, \mathbf{t}, \delta) \mathrm{q}(\tau^*, \tau)}{\mathrm{p}(\tau|\boldsymbol{\omega}, \boldsymbol{\phi}, \mathbf{t}, \delta) \mathrm{q}(\tau, \tau^*)}\right).$$

The drawing scheme for the AFT model parameters $(\boldsymbol{\omega}, \alpha, \sigma^2)$ follows these steps:

(1) Update $\boldsymbol{\phi}$ using the Bayesian backfitting MCMC algorithm described in CGM;

(2) Update $[\sigma^2|\boldsymbol{\phi}]$ using a Gibbs sampler;

(3) Obtain $[\alpha|\boldsymbol{\phi}, \sigma^2, \mathbf{t}]$;

(4) Update ω_i if $\delta_i = 1$;

(5) Sample from a $Normal(\alpha + \omega_i, \sigma^2)$ truncated at t_i if $\delta_i = 0$.

3.3.5. FDR-based Variable Selection for Ensemble Models

As mentioned before, BART offers a model-free mechanism to do variable selection. Let $\mathrm{p}(\gamma_p)$ denote the posterior probability inclusion of gene γ_j in the model, $j = 1, \dots, p$. We approximate $\mathrm{p}(\gamma_p)$ based on the relative frequency of occurrence o_{ik} of the ith covariate across k MCMC samples as

$$\mathrm{p}(\gamma_p) \equiv \frac{1}{\mathrm{K}} \Sigma_{k=1}^{\mathrm{K}} o_{ik},$$

where o_{ik} is the indicator function $\mathbf{I}(\gamma_j \in \mathbf{X}^k)$, where $\mathbf{X}^k$ is the set of covariates used to the build the tree model in the kth MCMC iteration.

We consider any covariate γ_j with $p(\gamma) < \varphi$, for some threshold φ, as significantly used (true discovery) in the model. The significance threshold φ can be set to control the average Bayesian FDR (Morris *et al.*, 2008) and it is thus a Bayesian q-value (Storey, 2003). In our study, we are interested in finding the value φ_ξ that controls the overall FDR at the level ξ, meaning that we expect only $100\xi\%$ of the covariates declared as significantly used in the model are false discoveries. For all covariates γ_j, $j = 1, \dots, p$, we first sort $p_j = p_0(\gamma_j)$ in descending order to yield $p_{(j)}$, $j = 1, \dots, p$. Then $\varphi_{0.2} = p(\gamma_j)$, where $\gamma_j = max\left\{j^*: j^{*-1}\left(1 - \sum_{j=1}^{j^*} p_{(j)}\right) \le \xi\right\}$. The set of regions $\chi_{\varphi\xi}$ then can be claimed to be significant covariates based on an average Bayesian FDR of ξ.

3.3.6. Performance Assessment

We assess the performance of our method using cross-validation, i.e., we randomly split the data 10 times into mutually exclusive training and test sets in the proportion 2:1, build the predictor using the training set, and then predict survival for the test set and compare it with the observed survival. We use two measures of predictive performance, the Brier score (BS) and the coefficient of determination (R^2). The BS is a specialized measure of goodness-of-fit for survival models (Graf *et al.*, 1999) and is given by

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\left(\hat{S}(t|\mathbf{X}_i)\right)^2 \mathbf{I}(t_i < t \wedge \delta_i = 1)}{\hat{\kappa}(t_i)} + \frac{\left(1 - \hat{S}(t|\mathbf{X}_i)\right)^2 \mathbf{I}(t_i > t)}{\hat{\kappa}(t)}\right],$$

where $\hat{\kappa}(\cdot)$ is the Kaplan-Meier estimate of the survival distribution for the observations $(t_1, \dots, t_n)$ and $\mathbf{I}$ denotes a indicator function. For BS we utilize the training data $\mathbf{t}$ and $\mathbf{X}$ to fit a model $p(\mathbf{t}|\mathbf{X})$ and use it to obtain the survival distribution $\hat{S}(t_*|\mathbf{t}, \mathbf{X}_*)$ for a future patient with covariate $\mathbf{X}_*$. Brier score ranges from 0 to 1 and the smaller the score the better is the fit.

The R^2 measure is the usual coefficient of determination of the fitted model and is estimated as

$$R^2 = 1 - exp\left(-\frac{2}{n}\left(L(\hat{\omega}) - L(0)\right)\right),$$

where $L(\cdot)$ denotes the log-likelihood function. In order to obtain the R^2, we first estimate $\hat{\boldsymbol{\omega}}$, the vector of latent covariate effects, using the median of the posterior distribution and then use it as a predictor in the univariable version of the specific underlying model. For example, the vector $\hat{\boldsymbol{\omega}}$ estimated from the AFT-Tree model is used as the predictor vector in a univariable AFT. R^2 also ranges from 0 to 1 and values close to 1 indicate good fits.

4. Results and Discussion

4.1. Random Forest

Random Forest model was used to fit the binary responses (STS *vs.* LTS) of the patients studied here. We first performed a study about the parametric settings to run Random Forest models in R. The default setting for the *randomForest* function sets the number of trees (*ntree*) to 500 and the number of covariates sampled for each binary split (*mtry*) as the square root of the total number of covariates. In our case, *mtry*=12. The bigger the number of trees used to build the model, the better the fitting is expected to be, however it increases computational time. Similarly, sampling a small number of covariates as candidate for splitting the nodes of the trees can leave important covariates out, hence compromising the fitting. On the other hand, sampling a large number of covariates can populate the model with non informative covariates, resulting in overfitting and increasing of the computation time (Rodin *et al.*, 2009). We run Random Forest 10 times for different settings of the parameters *ntree* (50, 100, 200, 500, 1000) and *mtry* (1, 3, 6, 12, 25, 50, 100, 144) and a summary of the misclassification error for the training data is presented in Table 2. Clearly, the misclassification error is smaller when the number of trees is equal to 1,000 as well as when the parameter *mtry* is equal to the default value already implemented in the R function.

Followoing, a Random Forest model was fitted and the binary responses (STS *vs.* LTS) of the patients studied here were predicted by this model. In order to reduce computation time and still obtain a small misclassification error, we opted by using the default parameters in the R function (*ntree*=500 and *mtry*=12). The overall misclassification rate of the prediction for the training set data (Table 2) was 18.9% (STS - 18.8% and LTS - 19%), and the overall misclassification rate for the testing set (OOB) was 25.3% (STS - 22.6% and LTS - 31.8%) indicating a slight overfitting in training set. Figure 5 allows a visual evaluation of the predictive performance of RF model. For 100% accuracy, one should find only red symbols at the left side of the dashed line and only black symbols at the right. Besides the inherent misclassification error, there is a large heterogeneity observed within classes which shows that splitting the survival response in a binary category might cause some loss of information. The AUC, another

measure of performance, obtained for the training data is 0.83 and for the testing set is 0.77 (Figure 6). A large misclassification error along with large heterogeneity presented within classes do not allow us to make any suggestion in order to change disease management.

The top 20 most important variables used to build the RF model are listed in Table 3. Some gene members of the top metagenes are highly related to the development of cancer/glioma phenotypes based on a search performed at the OMIM database (*http://www.ncbi.nlm.nih.gov/omim/*). The top 3 most important metagenes found by the RF method are the same found by the *survival-ensemble* methods and further investigation about them will be presented in the following sections. In addition, *patient age* is also a factor commonly found important by both classes of methods.
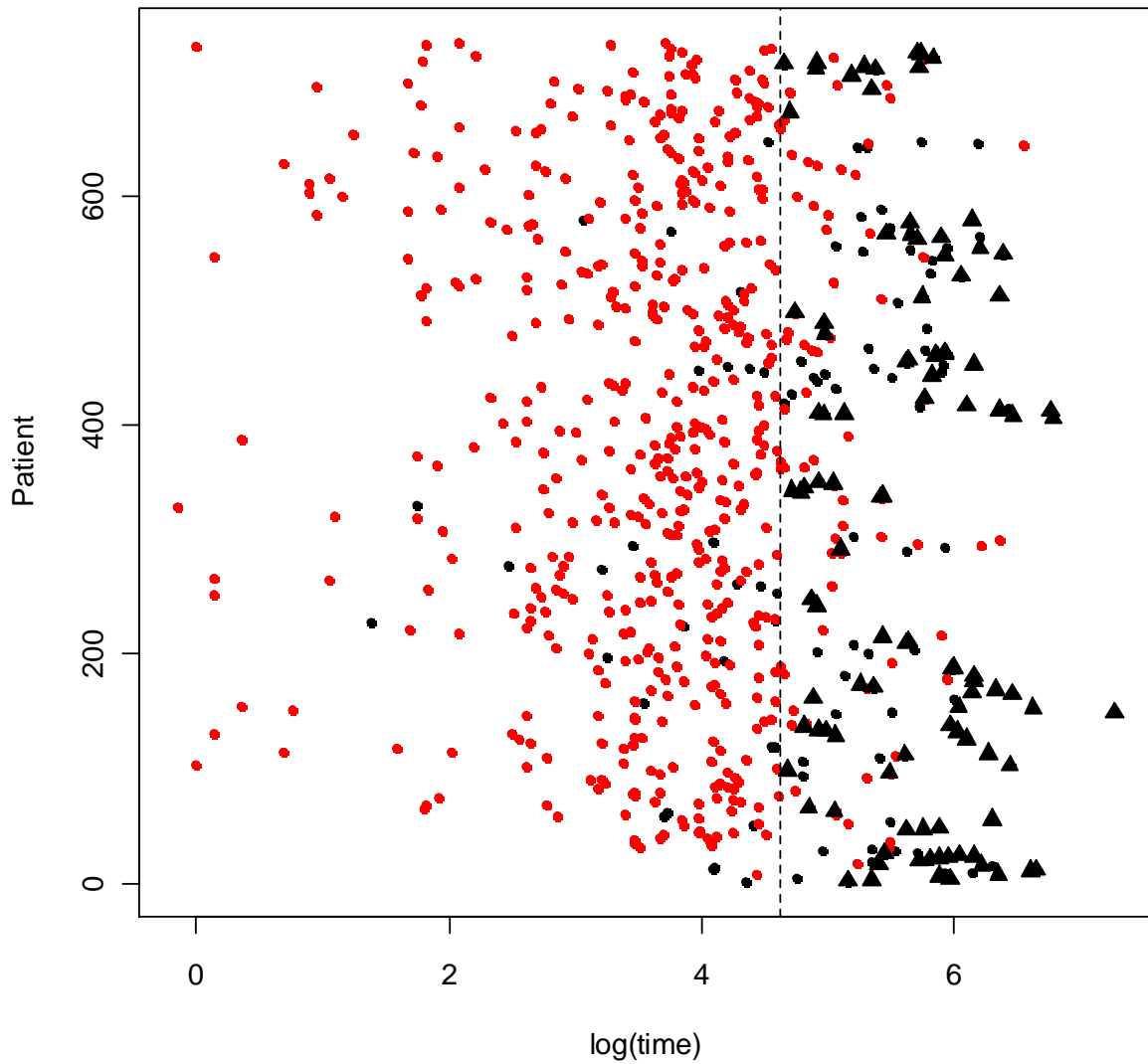
Figure 5: Plot comparing the observed survival times with classes predicted by the Random Forest model. Dots represent time of death and triangles represent censored times. Red symbols are observations predicted as STS and black symbols are observations predicted as LTS. The vertical dashed line indicates the value of 24 months used to discretize the survival time in STS *vs*. LTS.
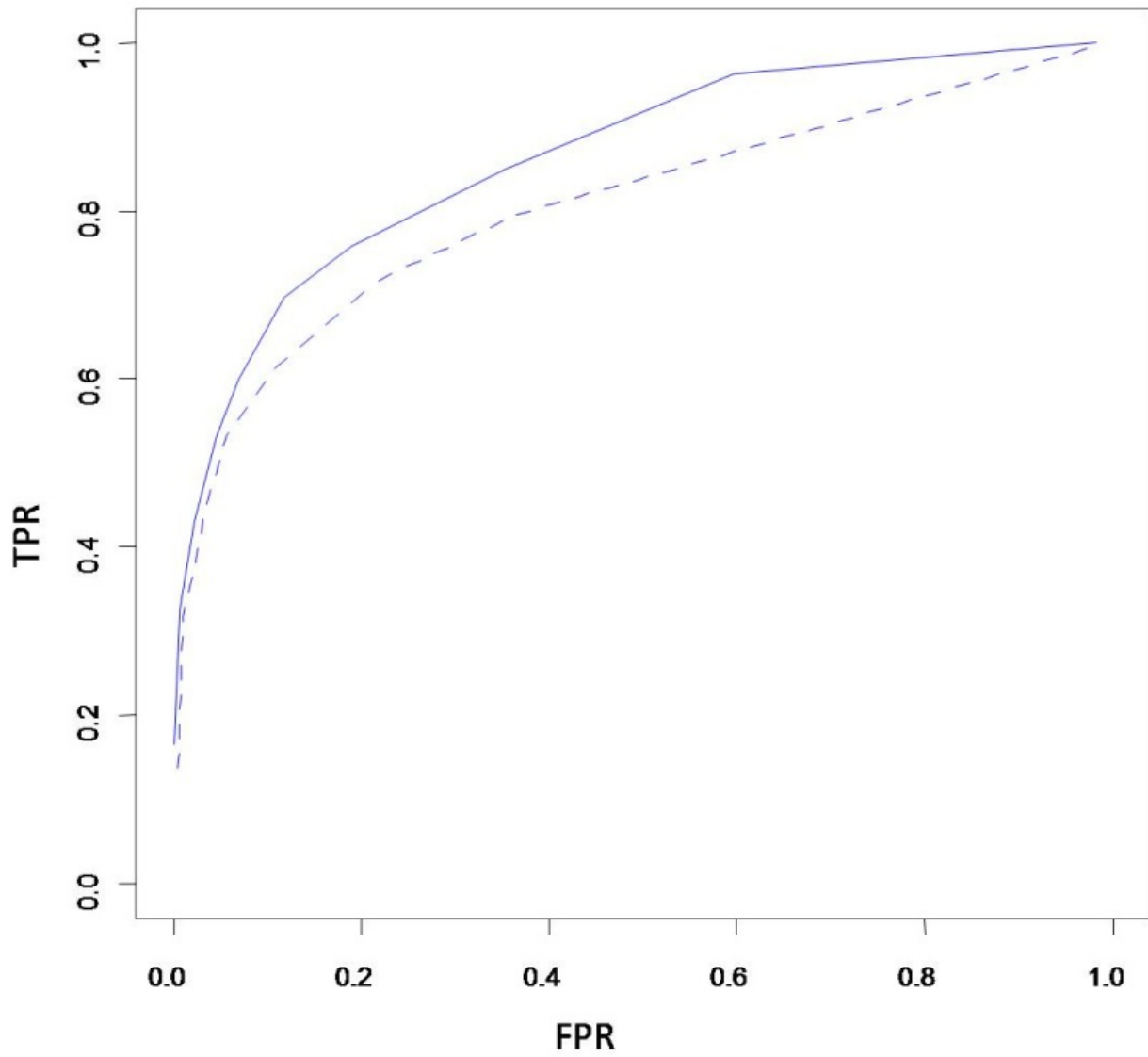
Figure 6: ROC for training set (solid lines) and test set (dashed lines) showing the performance in classification of STS and LTS by the Random Forest model.

Table 2: Summary of the misclassification error obtained for the training set using different settings of the parameters *ntree* and *mtry* in the *randomForest* R function. Ten RF models were built using each combination of parameters and the mean and s.d. of the overall training set misclassification error are reported. Parameter *ntree* represents the number of trees to grow and *mtry* is the number of variables randomly sampled as candidates at each split.

| **mtry** | ntree | | | | |
	50	100	200	500	1000
1	24.6±0.6	23.1±0.9	21.4±0.7	19.4±0.7	19.4±0.5
3	24.1±0.8	22.8±1.1	21.9±0.8	19.1±0.6	18.9±0.5
6	24.0±0.8	22.7±0.8	21.7±0.9	19.2±0.5	18.6±0.4
12	24.0±1.3	23.4±0.6	21.4±0.4	18.9±0.6	18.5±0.5
25	23.4±0.9	22.6±1.0	22.0±1.0	19.6±0.4	19.0±0.3
50	23.7±1.0	22.5±0.7	22.0±0.8	20.1±0.5	19.8±0.2
100	24.0±0.9	23.0±0.4	22.9±0.7	20.4±0.5	20.3±0.5
144	23.9±0.9	23.1±1.2	22.8±0.7	20.6±0.4	20.2±0.4

Table 3: Top 20 covariates used to build the RF model. Starred values (*) indicate metagenes containing genes related to the cancer phenotype while double-starred values (**) indicate metagenes containing genes related to the development of glioma phenotype. The directionality (DIR) information shows the influence of the over-expression of a given metagene in the patient survival: "+" means that survival increases with the metagene over-expression and "□" means that survival decreases with the upregulation. The measure of importance is the Gini index attributed to this covariate.

Rank	Covariate	DIR	Measure of importance	Genes
1	metagene82**	+	15.34	ALDH2, DAAM2, SCG3, MXI1, RAP2A
2	metagene99*	+	14.68	ZCCHC24, GLUD1, ID4,SCN3A
3	metagene52**	+	10.49	C1QL1, OSBPL11, CLASP2, MPRIP, PHLPP, GARNL1, ID1, RBPJ, IMPDH2, LRP4, PIK3R1, RIN2, PID1, BAI3, SALL2, TTC3, C11orf2, ADM, HSPA5, TNC, MAOB, C1S
4	metagene81	□	7.40	
5	metagene141	+	7.07	KCND2, ARHGAP12
6	metagene127*	□	5.86	FSTL1, COL4A1, IGKC, RP11-35N6.1
7	metagene85	+	5.82	ATP9A, WASF3, TCEAL2, NAP1L3
8	metagene92*	□	4.77	MDK, MFAP2, THBS4, TPBG, SERPINH1
9	metagene60	+	4.64	SLC9A6, AKAP11, HSP90AB1, HMP19, GPRC5B, ATP6V1G2, TERF2IP, SH3GL2, USP11
10	metagene83**	□	4.62	GADD45A, LY96, LAMP2, PLSCR1, RBP1, ISG15
11	metagene55**	□	4.45	ARPC1B, IFITM2, CTSC, ADFP, SLC39A14, GUSB, HLA-DMA, IRAK1, SERPINE1, PLP2, RNASE1, SLN, SOD2, TUBB6, CD44
12	metagene51	+	4.42	RTN3, PLEKHB1, RTN1, AGXT2L1, TSPAN7
13	metagene45*	+	4.12	CPE, APOE, PEA15
14	Age		3.97	
15	metagene84	+	3.97	CIRBP, ABAT, LRIG1, ZBTB20, ZMIZ1, RASSF2
16	metagene54	□	3.83	NDRG1, NCAN, F13A1, FN1, MSN, P4HB, PTX3, ACTA2, SERPING1, C1R, ACTN1, AKAP12
17	metagene104	+	3.19	ALDOC, FXYD6
18	metagene135*	+	3.05	RPS23, TCF12
19	metagene69	□	3.01	IGFBP2, IGFBP3, PLS3, RPN2, CAPNS1, SRPX
20	metagene42	□	2.96	HOXC4, HOXC6, HOXC10, HOXC11

4.2. Survival Ensembles

4.2.1. Performance Assessment using Breast Cancer Data

We compare the performance of our method with other survival prediction methods tailored for gene expression data as reviewed in van Wieringen *et al.* (2009). We use the breast cancer data set of Van't Veer *et al*. (2002 -- available at *http://www.rii.com/publications/2002/vantveer.html*), which contains gene expression profiles for 295 breast cancer patients and 5,057 gene expression values along with the patient survival outcomes. We reapply the best methods found by van Wieringen *et al.* (2009) ensuring that we work under the same conditions, i.e., the multivariable linear Cox Proportional hazard model (hereafter CPH) with top 10 genes obtained using a univariable Cox regression, the L1-penalized Cox regression (CPH-L1) of Tibshirani (1997), and the L2-penalized Cox regression (CPH-L2) of Gui & Li (2005). In addition, we ran a multivariable linear Weibull model with the top 10 most significant genes obtained by univariable Weibull models, as well as, multivariable linear AFT model with the top 10 genes pre-selected using a univariable AFT analysis. As in van Wieringen *et al.* (2009), we use the top 200 most significant genes obtained by the univariable underlying model to run our ensemble-model versions of accelerated failure time (AFT-TREE), Weibull (WEI-TREE), and Cox (CPH-TREE). A simple long chain (k=10,000 iterations) for each tree model with burning-in of the first half (5,000 samples) is enough to make inferences since different initial values do not alter chain convergence. The cross-validation procedure was repeated 10 times with the data being randomly split into training and test sets with a 2:1 ratio. We use the training set to build the predictor and then assess the performance of competing methods using the test set.

Figure 7 summarizes our results for all the methods considered here. Based on Brier score (Figure 7 - left), the methodology proposed here substantially outperforms the competing methods. Brier score for the ensemble-models is roughly 30% smaller than those for CPH-L1 and CPH-L2 methods, which were reported the top two performing methods in van Wieringen *et al.* (2009). In terms of $1 - R^2$, the tree-based methods are best performing methods (lowest values) along with CPH-L2, which has a very high variability and its performance is highly dependent on the actual split of the data. Interestingly, other

37

ensemble methods such as bagging and random forest yielded a lower R^2 (i.e. higher $1 - R^2$) in the breast cancer data set compared to set of Bayesian ensemble models studied here (medians of R^2 equal to 0.058 and 0.061, respectively, from van Wieringen *et al.,* 2009). In summary, based on these 2 different evaluation measures, we believe that our proposed methodology indeed improves the survival prediction accuracy, which could be attributed to the added flexibility in accounting for additive and non-linear effects.
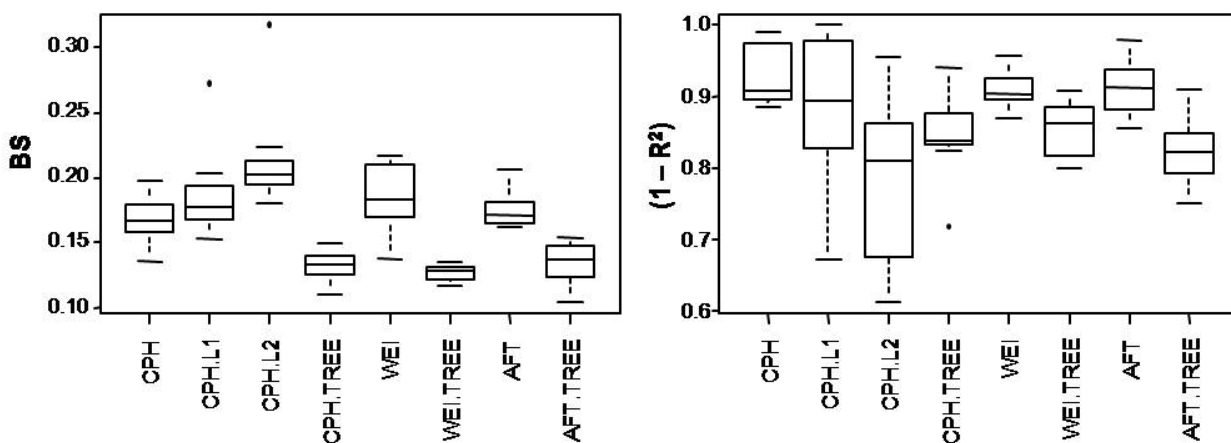

Figure 7: Box plots of performance results for the Brier score (BS -- left) and $(1 - R^2)$ measure (right) applied to the breast cancer data. For both measures, the lower the value the better the performance of the method.

As we mentioned before, one of the advantages of ensemble-models is that it is possible to assess the importance of each covariate for survival prediction using the relative frequency of occurrence as explained in Section 3.3.5. Using a FDR cutoff of 0.2 we found that a total of 15 variables are significant in the AFT-TREE, 11 in the WEI-TREE, and 14 in the CPH-TREE. Three genes (CCT5; BCL2; IL8) are simultaneously listed for AFT-TREE and WEI-TREE, and only one (NDUFS6) for AFT-TREE and CPH-TREE. Genes identified by the models could represent promising targets for further biological investigation. A search at the OMIM database (*http://www.ncbi.nlm.nih.gov/omim/*) confirmed that many of them are reported to be highly related to cancer development. For instance, BCL2 is known as one of

the strongest predictors of shorter overall survival in diffuse large-B-cell lymphoma patients (Lossos *et al*., 2004) and this gene also figures as a prognostic marker for breast cancer in Van't Veer *et al*. (2002) studies. Another example is the STK12 gene which is localized in a region that is frequently deleted in tumors and that contains tumor-related genes such as p53 (Tatsuka *et al*., 1998). Furthermore, BTG2 (Boiko *et al.*, 2006) is known as a major downstream effector of p53-dependent proliferation arrest in human fibroblasts while SESN1 gene expression is known to be modulated by p53 transcripts (Velasco-Miguel *et al.*, 1999).

4.2.2. Application to the brain tumor data

The methods compared here include the multivariable linear versions of AFT, Weibull, CPH, and the proposed ensemble-model versions AFT-TREE, WEI-TREE, and CPH-TREE models. One long chain (k=10,000 iterations) for each tree model is enough for satisfactory convergence and the first half of samples is discarded to make inference.

The Brier score calculated for the AFT-TREE (tree model) is slightly better than the one for its multivariable linear version (AFT) ($\mu_{AFT-TREE} = 0.118 \pm 0.01$ *vs*. $\mu_{AFT} = 0.119 \pm 0.02$) as well as for the Weibull models ($\mu_{WEI-TREE} = 0.116 \pm 0.02$ *vs*. $\mu_{WEI} = 0.112 \pm 0.02$). On the other hand, the BS for the CPH models are essentially the same ($\mu_{CPH-TREE} = 0.110 \pm 0.01$ *vs*. $\mu_{CPH} = 0.110 \pm 0.02$). To further evaluate the predictive ability of our proposed models, we conducted a time-dependent AUC analysis (Figure 8) to compare the prognostic capacity of survival models in different binary splits of the survival response. Time-dependent AUC analysis is frequently used in the clinical literature (Cerhan *et al*., 2007) to help physicians to better categorize patients in terms of survival classes. Here, the proposed ensemble-models do remarkably better than the competing methods showing higher sensitivity.
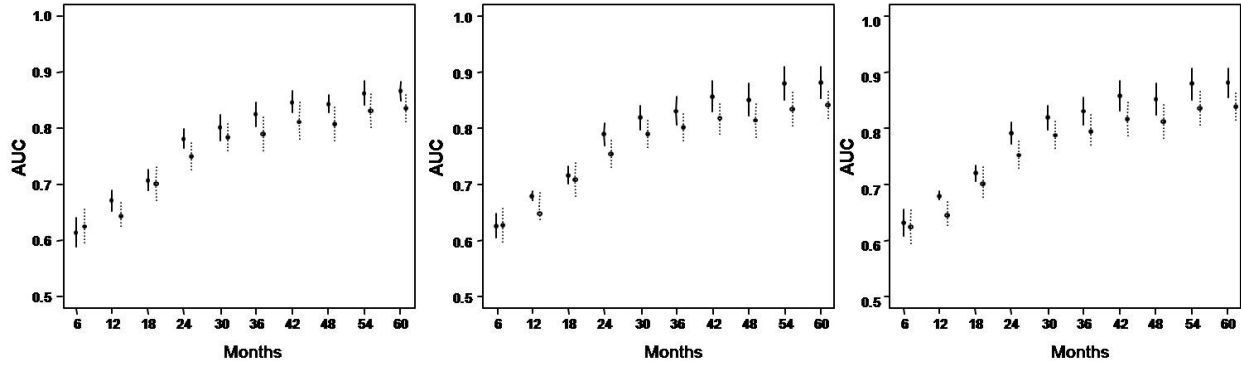
Figure 8: Time-dependent AUC analysis. Figure shows the time-dependent AUC analysis comparing the performance of the proposed ensemble methods with their multivariable linear versions applied to the test data. Dots represent the medians across splits of training/test sets and the lines depict the interquartile limits. Left plot: CPH (dashed lines) and CPH-TREE (solid); Center plot: WEI (dashed) and WEI-TREE (solid); Right plot: AFT (dashed) and AFT-TREE (solid).

A very small number of covariates (Table 4) are significantly used in the AFT-TREE, WEI-TREE, and CPH-TREE models (Figure 9). There is a complete overlapping between the first 3 more important covariates in all models. In addition, the top five mostly used covariates by AFT-TREE and WEI-TREE are the same. Tumor grade is known as one of the most important clinical factors for predicting survival time in brain tumor patients (see section 1) and it was confirmed in our results as one of the covariates more frequently used by all models. Indeed, the level IV of the covariate tumor grade (GBM patients) has a distinct survival curve than the levels II and III corroborating previous findings (Figure 10). In addition, patient age, another important clinical covariate (see section 1), also figures in the top for the AFT-TREE and WEI-TREE models. Glioma patients studied here which are younger than the median age (52 years) present much better prognosis than patients older than 52 years of age (Figure 11).

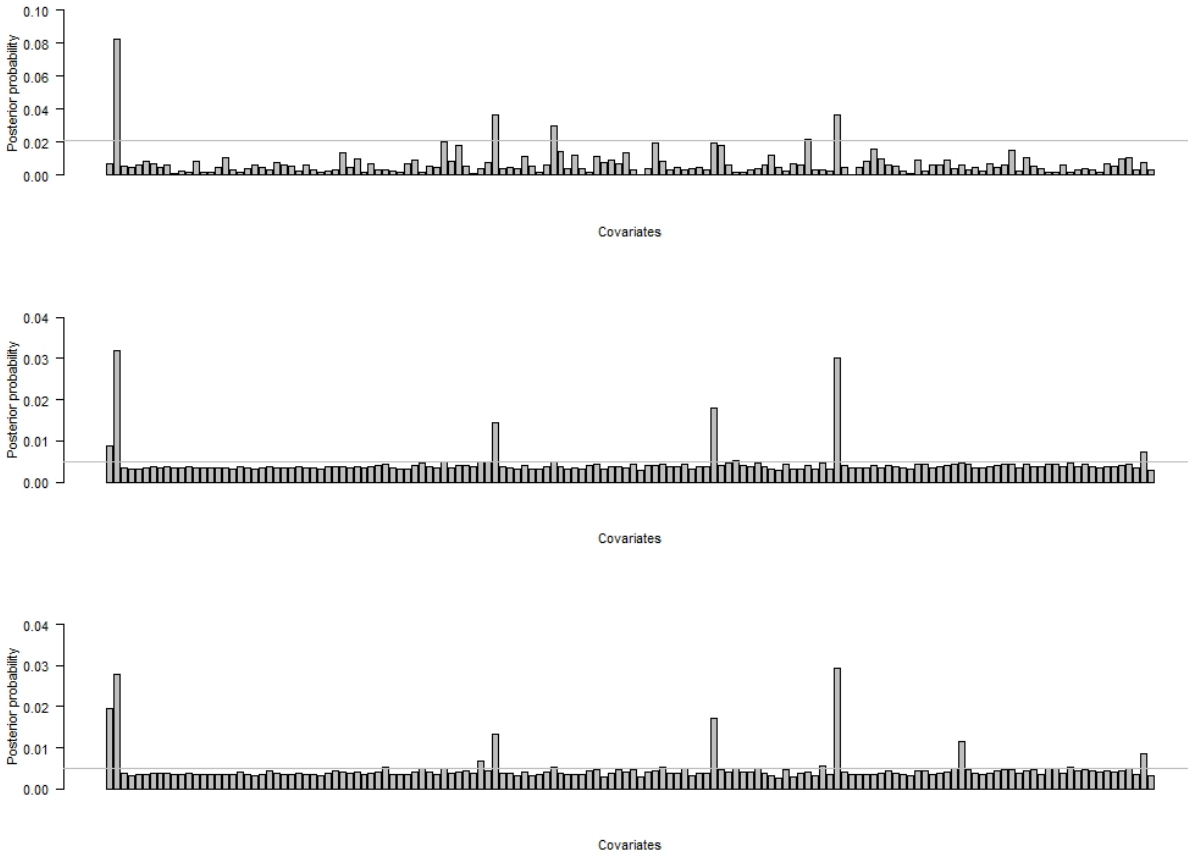**Figure 9:** Posterior probability of a variable appearing in the CPH-TREE (top), WEI-TREE (center), and AFT-TREE (bottom) survival ensemble models for the brain tumor data. Covariates from left to right: patient age, tumor grade, metagene 1, ..., metagene 142. Variables with posterior probability above the horizontal gray line are considered to be significantly used when controlling the FDR at 20%.
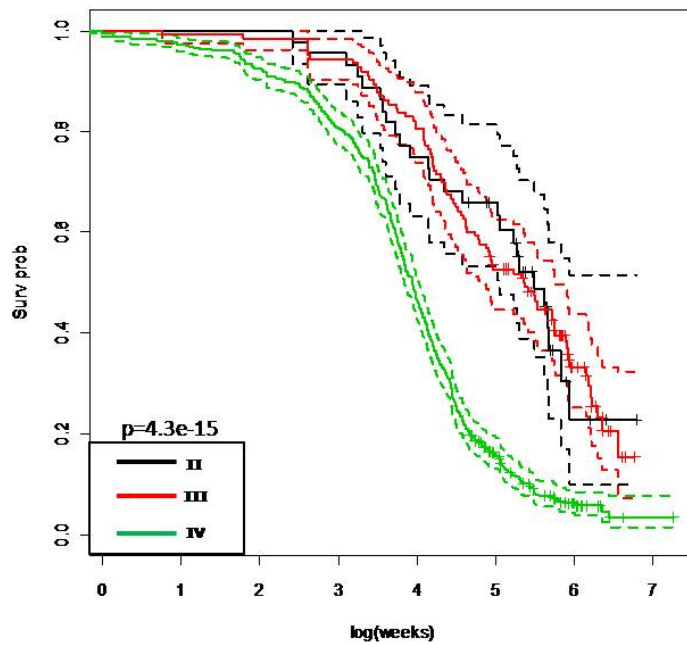
Figure 10: Survival curves (solid) along with the 95% C.I. (dashed) for the gliomas patients studied here. P-value was obtained by a log-rank test. A total of 44 patients are classified as tumor grade II, 120 as grade III, and 570 as grade IV.

Figure 11: *Left*: Age distribution; *Right*: K-M survival curves (solid) along with 95% C.I. (dashed) for patients younger and older than the sample median age (52 years). P-value was obtained by a log-rank test.

42

Table 4: Significant covariates (overall FDR of 20%) used to build the *survival-ensemble* models along with its probabilities p(γ) of being included in the final model. Starred values (*) indicate me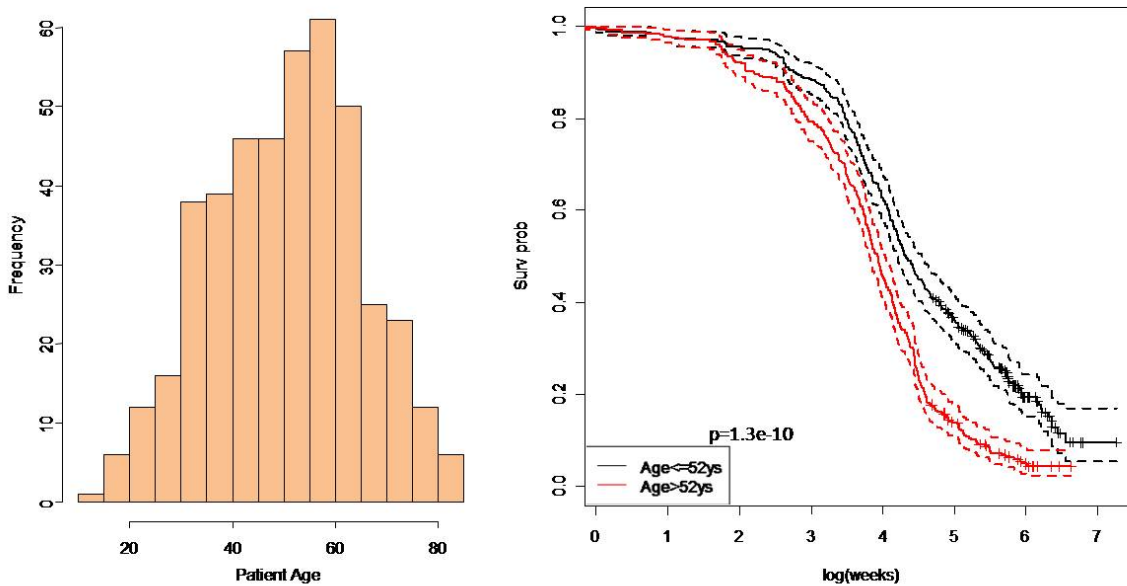tagenes containing genes related to the cancer phenotype while double-starred values (**) indicate metagenes containing genes related to the development of glioma phenotype. (+) means that survival increases with the metagene over-expression and (-) means that survival decreases with the upregulation.

AFT		p(γ)	Weibull		p(γ)	CPH		p(γ)
metagene99*	(+)	0.029	Grade		0.032	Grade		0.082
Grade		0.028	metagene99*	(+)	0.030	metagene52**	(+)	0.037
Age		0.020	metagene82**	(+)	0.018	metagene99*	(+)	0.036
metagene82**	(+)	0.017	metagene52**	(+)	0.014	metagene139*	(□)	0.030
metagene52**	(+)	0.013	Age		0.009	metagene70**	(+)	0.021
metagene116	(□)	0.011	metagene141	(+)	0.007			
metagene141	(+)	0.008	metagene85	(+)	0.005			
metagene50	(□)	0.007	metagene45*	(+)	0.005			
metagene97*	(+)	0.005						

Metagenes 52 and 99 were also found to have one of the highest posterior probabilities of being used in all survival-ensemble models as well as the RF model. Metagene 82 appeared frequently used in the AFT-TREE, WEI-TREE, and RF models while metagene 70 was more frequently used than other variables to build the CPH-TREE model. A following search at the OMIM database (*http://www.ncbi.nlm.nih.gov/omim/*) revealed that these metagenes contain genes known to be associated with the development and progression of many tumors including many associated to brain tumor development as the ones discussed in the section 1. For example, patients having values of expression of metagene 52 above the median show significantly better prognostic curves (Figure 12A). A detailed search shows that the metagene 52 has six genes associated to cancer phenotype: PHLPP, GARNL1, ID1, RBPJ, PIK3R1, and BAI3 (Figure 12B). PHLPP is known for its capacity to dephosphorylate AKT, triggering apoptosis and suppressing tumor growth via the p53 and RTK mitogenic pathways. PHLPP appears downregulated in several colon cancers and glioblastoma cell lines (Gao *et al*., 2005). In our study, PHLPP also appears downregulated in patients with shorter survival (Figure 12B). Mouse embryocarcinoma cells upregulated GARNL1 expression following induction of neuronal differentiation (Heng & Tan, 2002). We show that the gene GARNL1 is upregulated in long survival patients (Figure 12B) and hypothesize that the role of this gene in promoting neuronal differentiation prevents cells going through migration or invasion processes, hence improving prognosis. The protein ID1 is a negative transcriptional regulator of CDKN2A, which is associated with the development of malignant melanoma (Ohtani *et al*., 2001). CDKs, as discussed in section 1, are important members of the mitogenic pathway which control cell proliferation. We show (Figure 12B) that ID1 is overexpressed in long survival patients and hence we hypothesize that ID1 also negatively regulates CDKs and controls cell proliferation. RBPJ is known to be part of the Notch pathway. Activation of the Notch cascade is known to maintain the undifferentiated state in cells (*van* Es *et al*., 2005) and as a result, it might be directly related with mesenchymal phenotypes in gliomas. In addition, the Notch pathway plays a role in neuronal function and development and stabilization of angiogenesis. Hence, the underexpression of RBPJ in short survival patients (Figure 12B) suggests alterations in the Notch pathway causing the formation of mesenchymal

and angiogenesis phenotypes. In addition, it is known that close to 90% of GBM present the PIK3R1 signaling pathway altered (TCGA, 2008) and BAI3 (Brain-specific angiogenesis inhibitor 3) is considered to play a role in suppression of glioblastoma since its expression is absent or significantly reduced in this tumor type (Shiratsuchi *et al*., 1997).

Also, patients having values of expression of metagene 70 above the median show significantly better prognostic curves (Figure 13A). Metagene 70 contains the genes EGFR, FAIM2, SASH1, and PINK1. The EGFR gene is involved in cell signaling, cell proliferation, differentiation, motility, and in tissue development (Wang *et al*., 2004) which makes it one of the most important genes related to the development of gliomas (see section 1). In our study we show that patients with upregulated EGFR (Figure 13B) tend to survive less than other patients. In addition, the gene FAIM2 which is a FAS receptor for a tumor necrosis factor (TNF) (Somia *et al*., 1999) is found altered (Figure 13B). FAS is a "death-receptor", as discussed in section 1, which is involved in pro-apoptotic and anti-apoptotic roles. The gene SASH1 which is downregulated in breast tumor tissues (Zeller *et al*., 2003) is also downregulated in short survival glioma patients (Figure 13B), and the gene PINK1 which is a PTEN-induced putative kinase is also downregulated in short survival glioma patients (Figure 13B) suggesting alterations in the mitogenic signaling and apoptotic pathways.

Likewise, patients presenting downregulation of metagenes 82 (Figure 14A) and 99 (Figure 15A) show significantly better prognostic curves than other patients. Metagene 82 contains the gene MXI1 which negatively regulates MYC oncoprotein, an important glioblastoma tumor inductor (Albarosa *et al*., 1995). MYC plays an important role in regulating cell proliferation, apoptosis (controls the death-receptor Bcl-2), and cell differentiation (Albarosa *et al*., 1995). We show that glioma patients with short survival present downregulation of MXI1 (Figure 14B) suggesting that the MYC oncoprotein is overexpressed and, hence, causing cell proliferation, apoptotic, and mesenchymal phenotypes. To conclude, metagene 99 contains the gene ID4 which is believed to be a putative leukemia suppressor (Yu *et al*., 2005) and here appears overexpressed in long term survival patients (Figure 15B) suggesting an equivalent role in brain tumor suppression as well.
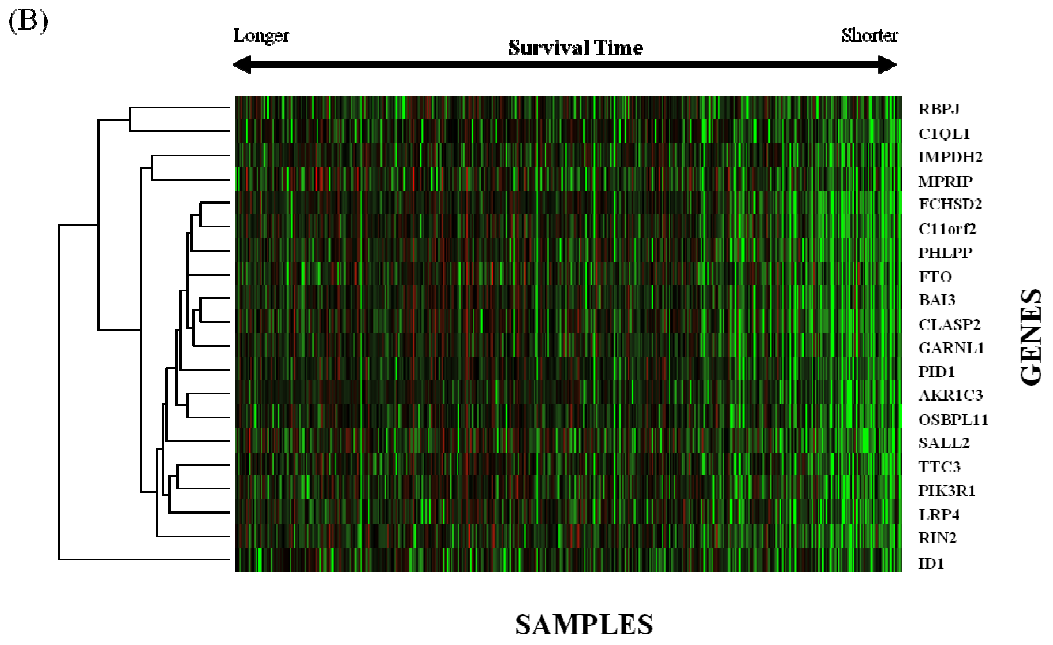
(A)

(B)

Figure 12: *Metagene 52 —* (A) K-M survival curves (solid) along with 95% C.I. (dashed). P-value was obtained by log-rank test. (B) Heatmap of the expression values of genes grouped within metagene 52. Red color spots represent over-expressed values and green color spots represent under-expressed values. Samples are depicted in the horizontal axis and are sorted by survival time (from left to right: longer to shorter survival times). Genes are depicted in the vertical axis. Dendrogram was obtained by *"Ward"* method.
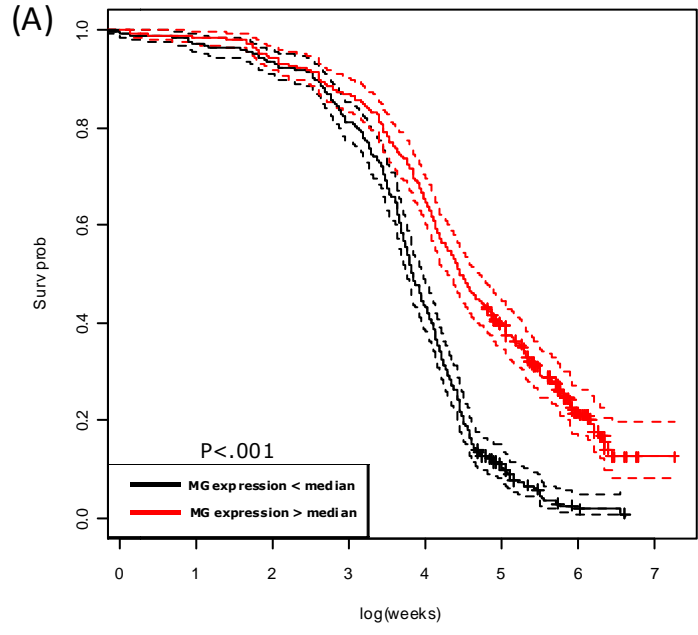
Figure 13: *Metagene 70* — (A) K-M survival curves (solid) along with 95% C.I. (dashed). P-value was obtained by log-rank test. (B) Heatmap of the expression values of genes grouped within metagene 70. Red color spots represent over-expressed values and green color spots represent under-expressed values. Samples are depicted in the horizontal axis and are sorted by survival time (from left to right: longer to shorter survival times). Genes are depicted in the vertical axis. Dendrogram was obtained by *"Ward"* method.
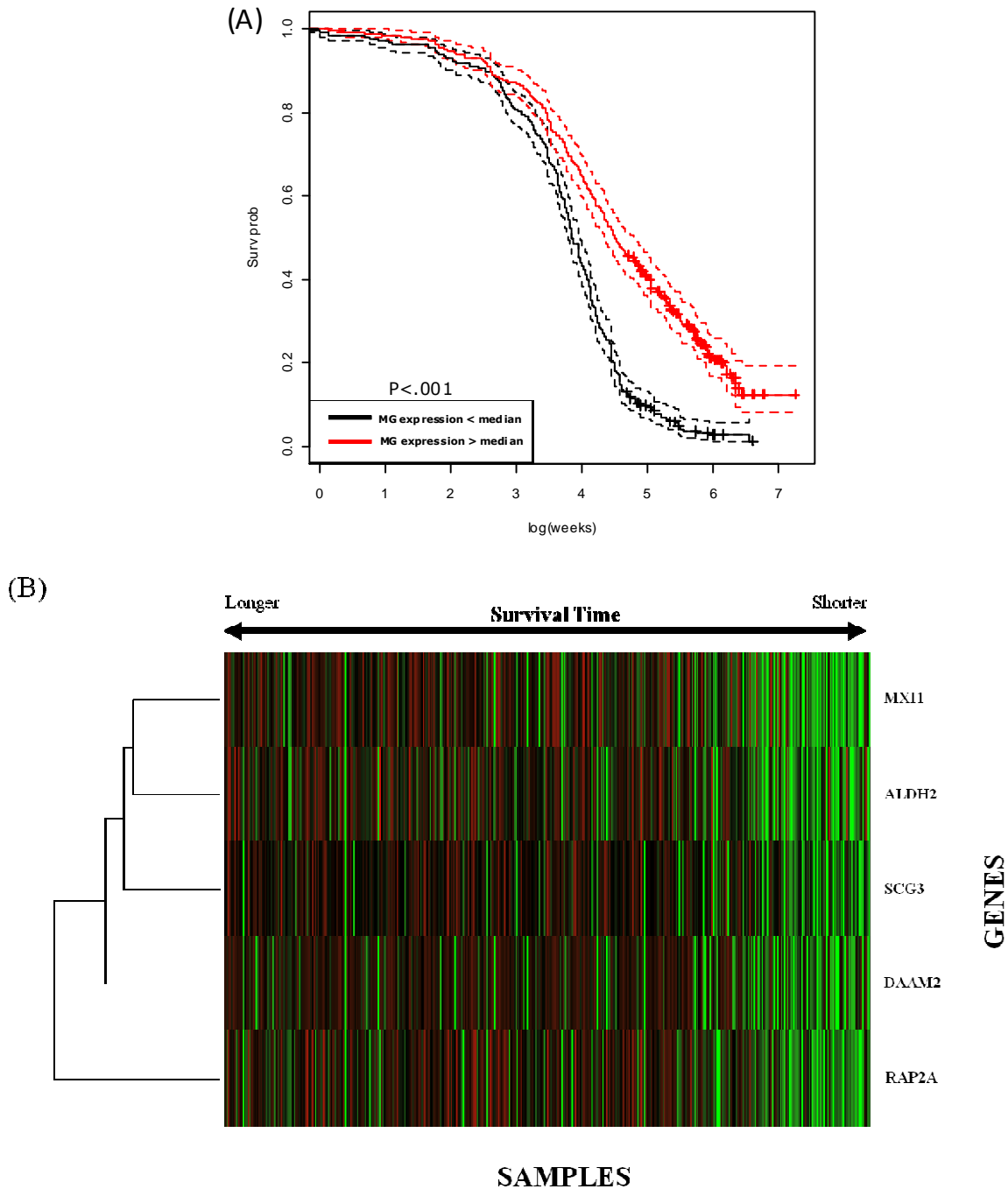
Figure 14: *Metagene 82* — (A) K-M survival curves (solid) along with 95% C.I. (dashed). P-value was obtained by log-rank test. (B) Heatmap of the expression values of genes grouped within metagene 82. Red color spots represent over-expressed values and green color spots represent under-expressed values. Samples are depicted in the horizontal axis and are sorted by survival time (from left to right: longer to shorter survival times). Genes are depicted in the vertical axis. Dendrogram was obtained by *"Ward"* method.
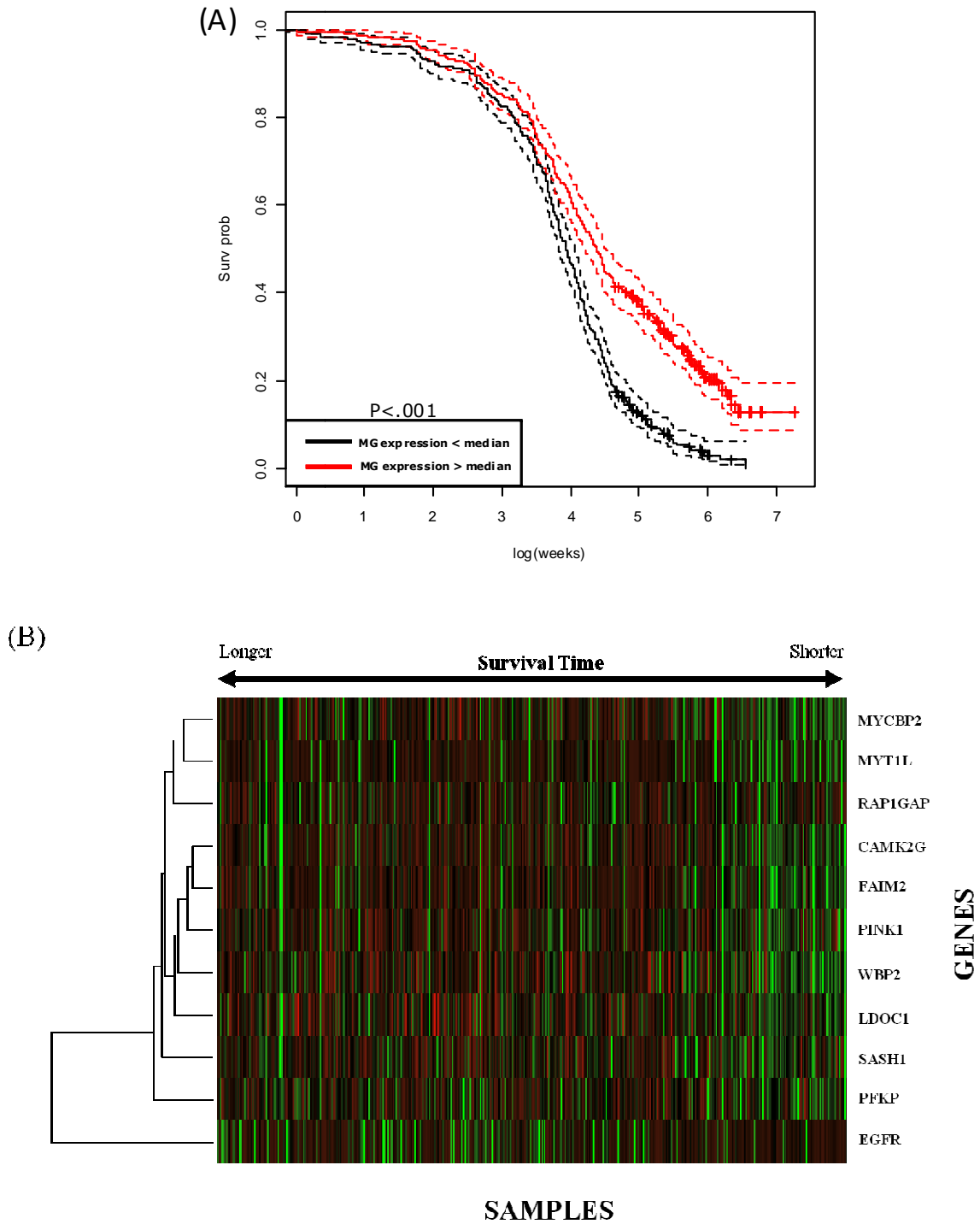
48

Figure 15: *Metagene 99* — (A) K-M survival curves (solid) along with 95% C.I. (dashed). P-value was obtained by log-rank test. (B) Heatmap of the expression values of genes grouped within metagene 99. Red color spots represent over-expressed values and green color spots represent under-expressed values. Samples are depicted in the horizontal axis and are sorted by survival time (from left to right: longer to shorter survival times). Genes are depicted in the vertical axis. Dendrogram was obtained by *"Ward"* method.
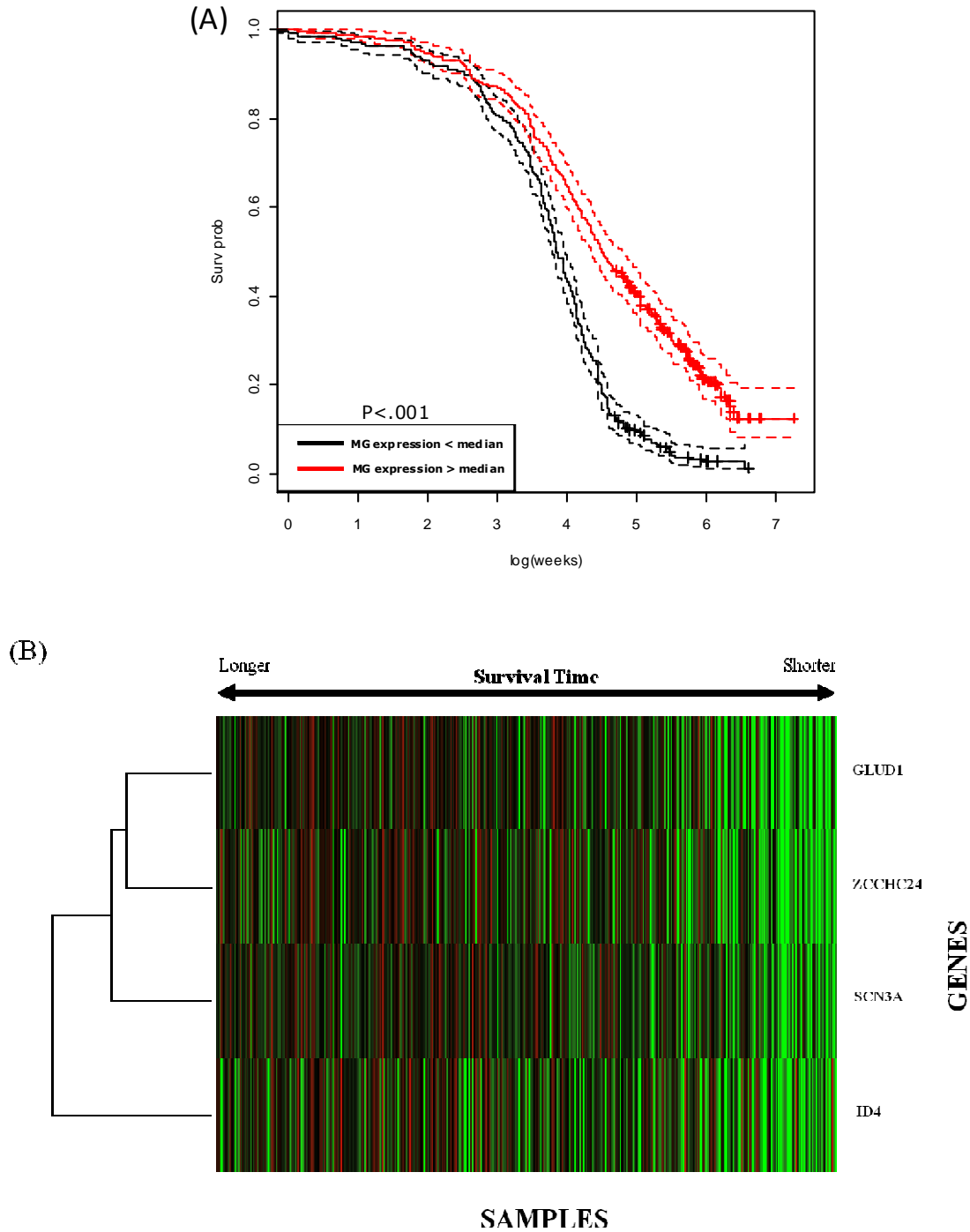
Additionally, partial dependence (PD) functions (Friedman, 2001) obtained by the marginalization of the posterior distribution of $f(\mathbf{X})$ with respect to the covariate(s) of interest are particularly useful to illustrate the marginal effect of one relevant covariate(s) directly on the survival outcome as in the AFT-TREE model. PD function plots (Figure 16) shows that marginal effects of the metagenes 52, 82, and 99 on the survival time estimated by the AFT-TREE model. It shows that the relative upregulation of metagene 52 increases the brain tumor patient survival time in roughly 170 weeks (3.3 years), the relative upregulation of metagene 82 increases glioma patients survival in 130 weeks (2.5 years), and the regulation of the metagene 99 between values 2 and 4 increases the survival of glioma patients in almost 200 weeks (around 4 years). However, just a few patients have these metagenes upregulated (> 4) which makes the confidence intervals bigger at this region, indicating that the interpretation of the PD plots at this region must be done with caution. Another important tool used to identify individual contributions of the covariates on the patient survival time are nomograms. In Figure 17 we show a nomogram of the most important variables in the AFT-TREE model. The interpretation is performed as following. Identify the patient age and draw a vertical line to the "points" axis on the top of the nomogram. Repeat this process for the remaining variables. Sum the points for each individual variable and locate this on the "Total Points" axis at the bottom of the page. The width of a variable axis represents how much it affects the overall survival time. To calculate the log of survival time in weeks, draw a vertical line from the "Total Points" spot on the linear predictor axis. Indeed our results show that expression values of metagenes drastically impact the overall survival in brain tumor patients in addition to clinical covariates. For example, a patient indexed by the median values of the most important covariates found by the AFT-TREE model, i.e., a patient 52 years of age, presenting tumor grade IV, metagene 52 expression equal to -0.36, metagene 82 expression equal to -0.95, and metagene 99 expression equal to -0.75 will receive a "Total points" score around 93 and its expected survival time will be of 75-80 weeks.
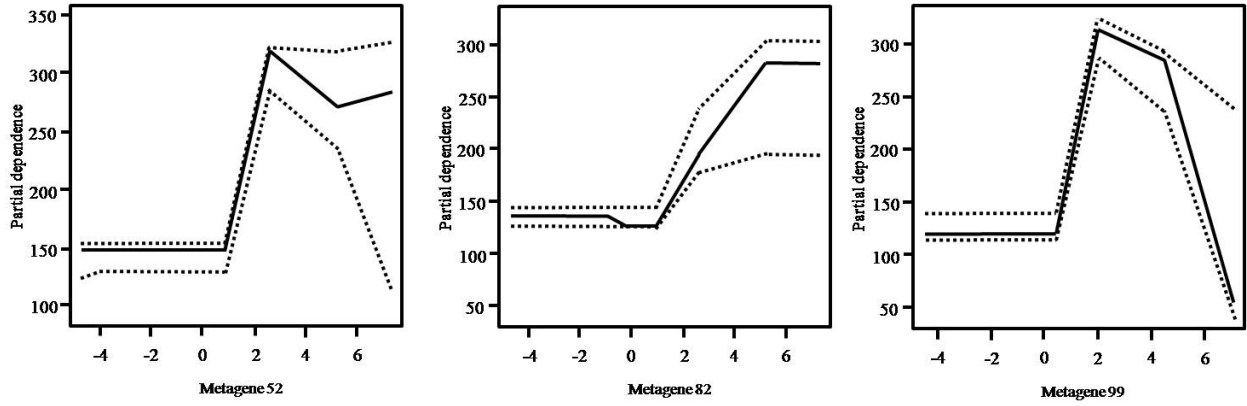
Figure 16: Partial dependence (PD) plots for metagene 52, 82, and 99 in the AFT-TREE model. Solid lines represent the marginal contribution to the survival time and dashed lines are 95% C.I.
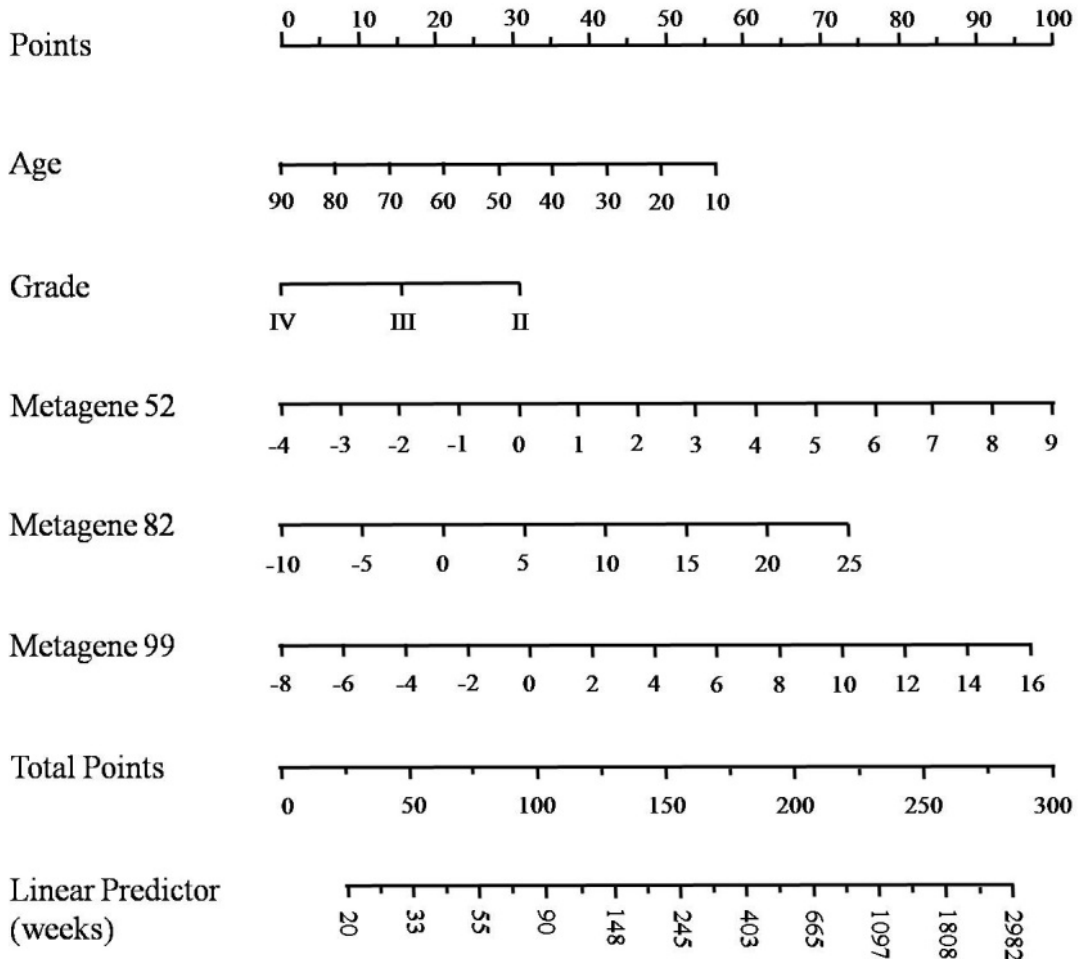

Figure 17: Nomogram of the most important variables in the AFT-TREE model.

5. Conclusion

We built a model for binary response using the Random Forest model. Random Forest is a tree-based ensemble method with high predictive accuracy commonly seen in the Biostatistics literature. We obtained a misclassification error of 19% for training data and 25% for testing data. We also present and discuss a list of the variables most frequently used to build the trees in the RF method. Besides the unavoidable misclassification error, we show that splitting the response in categories (STS *vs.* LTS) might cause loss of information since the heterogeneity within classes is relatively large.

We have proposed Bayesian ensemble methods for survival prediction in the high-dimensional context as in DNA microarray data analysis. We relied on a powerful Bayesian predictive tool (BART) to estimate the covariate effects by means of a latent variable assumed to be normally distributed. BART was chosen because it is flexible to accommodate a high number of covariates and their interactions and properly accounts for the uncertainty of parameter estimation inherent to the Bayesian approach. Nonetheless, the proposed method can be extended to allow the use of any other ensemble method instead.

We incorporated the latent variable estimates in three widely used survival models, named AFT, Weibull, and CPH, and performed Bayesian estimation of additional parameters simultaneously. Our prior choices require less complex MCMC sampling techniques and sometimes, undesirable parameters could be integrated out, as the baseline hazard function in the CPH-TREE model. In addition, our method provides prediction of the survivor function which directly contributes to an adequate personalized management of patients.

The application of our methodology to two different data sets showed that our model outperforms prediction accuracy of many available models.

The screening ability of BART identifies important predictors across trees and training-test splits of data, which allowed us to reveal the impact of many important genes and clinical covariates on the survival of glioma patients. In addition to the predictive ability, the variable selection procedure along with PD functions and nomogram techniques grants high interpretability to the final model.

52

6. References

Albarosa, R.; S. DiDonato; G. Finocchiaro. 1995. Redefinition of the coding sequence of the MXI1 gene and identification of a polymorphic repeat in the 3-prime non-coding region that allows the detection of loss of heterozygosity of chromosome 10q25 in glioblastomas. *Hum. Genet.*, 95:709-11.

Alizadeh, A.A.; M.B. Eisen; R.E. Davis; C. Ma; I.S. Lossos; A. Rosenwald; J.C. Boldrick; H. Sabet; T. Tran; X. Yu; J.I. Powell; L. Yang; G.E. Marti; T. Moore; J.J. Hudson; L. Lu; D.B. Lewis; R. Tibshirani; G. Sherlock; W.C. Chan; T.C. Greiner; D.D. Weisenburger; J.O. Armitage; R. Warnke; R. Levy; W. Wilson; M.R. Grever; J.C. Byrd; D. Botstein; P.O. Brown; L.M. Staudt. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503-11.

Bachoo, R.M.; E.A. Maher; K.L. Ligon; N.E. Sharpless; S.S. Chan; M.J. You; Y. Tang; J. DeFrances; E. Stover; R. Weissleder; D.H. Rowitch; D.N. Louis; R.A. DePinho. 2002. Epidermal growth factor receptor and Ink4a/Arf: convergent mechanisms governing terminal differentiation and transformation along the neural stem cell to astrocyte axis. *Cancer Cell,* 1:269–77.

Berchuck, A.; E.S. Iversen; J.M. Lancaster; J. Pittman; J. Luo; P. Lee; S. Murphy; H.K. Dressman; P.G. Febbo; M. West; J.R. Nevins; J.R. Marks 2005. Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clin. Cancer Res.*, 11:3686-96.

Boiko, A.D.; S. Porteous; O.V. Razorenova; V.I. Krivokrysenko; B.R. Williams; A.V. Gudkov. 2006. A systematic search for downstream mediators of tumor suppressor function of p53 reveals a major role of BTG2 in suppression of Ras-induced transformation. *Genes & Dev.*, 20:236-52.

Breiman, L. 1996. Bagging predictors. *Machine Learning*, 26:123-40.

Breiman, L. 2001. Random forests. *Machine Learning*, 45:5-32.

Cairncross, J.G.; K. Ueki; M.C. Zlatescu; D.K. Lisle; D.M. Finkelstein; R.R. Hammond; J.S. Silver; P.C. Stark; D.R. Macdonald; Y. Ino; D.A. Ramsay; D.N. Louis. 1998. Specific genetic predictors of

chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *J. Natl. Cancer Inst.*, 90:1473-9.

Carro, M.S.; W.K. Lim; M.J. Alvarez; R.J. Bollo; X. Zhao; E.Y. Snyder; E.P. Sulman; S.L. Anne; F. Doetsch; H. Colman; A. Lasorella; K.D. Aldape; A. Califano; A. Iavarone. 2010. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463:318-25.

Castells, X.; J.M. García-Gómez; A. Navarro; J.J. Acebes; O. Godino; S. Boluda; A. Barceló; M. Robles; J. Ariño; C. Arús. 2009. Automated brain tumor biopsy prediction using single-labeling cDNA microarrays-based gene expression profiling. *Diagn. Mol. Pathol.*, 18:206–18.

Cerhan, J.R; S. Wang; M.J. Maurer; S.M. Ansell; S.M. Geyer; W. Cozen; L.M. Morton; S. Davis; R.K. Severson; N. Rothman; C.F. Lynch; S. Wacholder; S.J. Chanock; T.M. Habermann; P. Hartge. 2007. Prognostic significance of host immune gene polymorphisms in follicular lymphoma survival. *Blood*, 12:5439-46.

Chipman, H.A.; E.I. George; R.E. McCulloch. 1998. Bayesian CART model search (with discussion). *JASA*, 93:935–60.

Chipman, H.A.; E.I. George; R.E. McCulloch. 2006. BART: Bayesian Additive Regression Trees, Technical Report, Graduate School of Business, University of Chicago.

Clarke, J. & M. West. 2008. Bayesian Weibull tree models for survival analysis of clinico-genomic data. *Stat. Methodol.*, 5:238-62.

Coons, S.W.; P.C. Johnson; B.W. Scheithauer; A.J. Yates; D.K. Pearl. 1997. Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas. *Cancer*, 79:1381–93.

Cox, D. 1972. Regression models and life tables. *J. R. Stat. Soc. B*, 34:187-220.

Dai, M.; P. Wang; A.D. Boyd; G. Kostov; B. Athey; E.G. Jones; W.E. Bunney; R.M. Myers; T.P. Speed; H. Akil; S.J. Watson; F. Meng. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, 33:e175.

Datta, S.; J. Le-Rademacher; S. Datta. 2007. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics*, 63:259-71.

DeAngelis, L.M. 2009. Anaplastic gliomas. *J. Clin. Oncol.,* 27:5861-7.

Denison, D.G.T.; B.K. Mallick; A.F.M. Smith. 1998. A Bayesian CART algorithm. *Biometrika*, 85:363–77.

D'haeseleer, P. 2005. How does gene expression clustering work? *Nature Biotech.*, 23:1499-1501.

Diaz-Uriarte, R. & S. Alvarez de Andres. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3.

Diehn, M.; C. Nardini; D.S. Wang; S. McGovern; M. Jayaraman; Y. Liang; K.D. Aldape; S. Cha; M.D. Kuo. 2008. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *PNAS*, 105:5213-18.

Do, K-A; B.M. Broom; S. Wen. 2003. Geneclust. **In**: Parmigiani, G.; E.S. Garrett; R.A. Irizarry; S.L. Zeger. *The analysis of gene expression data: Methods and software.* Springer, New York.

Freije, W.A.; F.E. Castro-Vargas; Z. Fang; S. Horvath; T. Cloughesy; L.M. Liau; P.S. Mischel; S.F. Nelson. 2004. Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.*, 64:6503–10.

Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. *Ann. Stats.*, 29:189-232.

Furnari, F.B.; T. Fenton; R.M. Bachoo; A. Mukasa; J.M. Stommel; A. Stegh; W.C. Hahn; K.L. Ligon; D.N. Louis; C. Brennan; L. Chin; R.A. DePinho; W.K. Cavenee. 2007. Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes & Dev.*, 21:2683–2710.

Gao, T.; F. Furnari; A.C. Newton. 2005. PHLPP: a phosphatase that directly dephosphorylates Akt, promotes apoptosis, and suppresses tumor growth. *Molec. Cell*, 18:13-24.

Garber M.E.; O.G. Troyanskaya; K. Schluens; S. Petersen; Z. Thaesler; M. Pacyna-Gengelbach; M. van de Rijn; G.D. Rosen; C.M. Perou; R.I. Whyte ; R.B. Altman; P.O. Brown; D. Botstein; I. Petersen. 2001. Diversity of gene expression in adenocarcinoma of the lung. *PNAS*, 98:13784-89.

Gentleman, R.; V. Carey; W. Huber; R. Irizarry. 2005. **Bioinformatics and Computational Biology Solutions Using R and Bioconductor.** Springer Inc., New York.

Giannini, C.; B.W. Scheithauer; A.L. Weaver; P.C. Burger; J.M. Kros; S. Mork; M.B. Graeber; S. Bauserman; J.C. Buckner; J. Burton; R. Riepe; H.D. Tazelaar; A.G. Nascimento; T. Crotty; G.L. Keeney; P. Pernicone; H. Altermatt. 2001. Oligodendrogliomas: reproducibility and prognostic value of histologic diagnosis and grading. *J. Neuropathol. Exp. Neurol.*, 60:248–62.

Giese, A.; R. Bjerkvig; M.E. Berens; M. Westphal. 2003. Cost of migration: invasion of malignant gliomas and implications for treatment. *J Clin. Oncol.*, 21:1624–36.

Gilks, W.R.; S. Richardson; D. Spiegelhalter. 1996. **Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics.** Chapman & Hall, New York.

Graf, E.; C. Schmoor; W. Sauerbrei; M. Schumacher. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statis. Med.*, 18:2529-45.

Gravendeel, L.A.M.; M.C.M. Kouwenhoven; O. Gevaert; J.J. de Rooi; A.P. Stubbs; J.E. Duijm; A. Daemen; F.E. Bleeker; L.B.C. Bralten; N.K. Kloosterhof; B. de Moor; P.H.C. Eilers; P.J. van der Spek; J.M. Kros; P.A.E.S. Smitt; M.J. van den Bent; P.J. French. 2009. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res*., 69:9065–72.

Gui, J. & H. Li. 2005. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21:3001-8.

Hastie, T.; R. Tibshirani; D. Botstein; P. Brown. 2001. Supervised harvesting of expression trees. *Genome Biol*., 2:0003.1-0003.12.

Hayden, E.C. 2010. Genomics boosts brain-cancer work. *Nature*, 463:278.

Hegi, M.E.: A.C. Diserens; T. Gorlia; M.F. Hamou; N. de Tribolet; M. Weller; J.M. Kros; J.A. Hainfellner; W. Mason; L. Mariani; J.E.C. Bromberg; P. Hau; R.O. Mirimanoff; J.G. Cairncross; R.C. Janzer; R. Stupp. 2005. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.*, 352:997–1003.

Heng, J.I.T. & S.-S. Tan. 2002. Cloning and characterization of GRIPE, a novel interacting partner of the transcription factor E12 in developing mouse forebrain. *J. Biol. Chem.*, 277:43152-59.

Hothorn, T.; B. Lausen; A. Benner; M. Radespiel-Tröger. 2004. Bagging survival trees. *Statis. Med.*, 23:77-91.

Hothorn, T.; P. Bühlmann; S. Dudoit; A. Molinaro; M.J. van der Laan. 2006. Survival ensembles. *Biostatistics*, 7:355-373.

Huang, J.; S. Ma; H. Xie. 2006. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62:813-20.

Hwang, S.; W.L. Kuo; J.F. Cochran; R.C. Guzman; T. Tsukamoto; G. Bandyopadhyay; K. Myambo; C.C. Collins. 1997. Assignment of HMAT1, the human homolog of the murine mammary transforming gene (MAT1) associated with tumorigenesis, to 1q21.1, a region frequently gained in human breast cancers. *Genomics*, 42:540-42.

Ibrahim, J.G.; M.-H. Chen; D. Sinha. 2001. **Bayesian Survival Analysis**. Springer, New York.

Ishwaran, H.; E.H. Blackstone; C.E. Pothier; M.S. Lauer. 2004. Relative risk forests for exercise heart rate recovery as a predictor of mortality. *JASA*, 99:591-600.

Ishwaran, H.; U.B. Kogalur; E.H. Blackstone; M.S. Lauer. 2008. Random survival forest. *Ann. Appl. Stats.*, 2:841-60.

Junqueira, L. & J. Carneiro. 2005. **Basic Histology: Text & Atlas**. 11th edn. McGraw-Hill Medical, New York.

Kang, S.: A. Denley; B. Vanhaesebroeck; P.K. Vogt. 2006. Oncogenic transformation induced by the p110β, -γ, and -δ isoforms of class I phosphoinositide 3-kinase. *PNAS,* 103:1289–94.

Kalbfleisch, J.D. 1978. Non-parametric Bayesian analysis of survival time data. *J. R. Stats. Soc. B*, 40:214-21.

Klein, J.P. & M.L. Moeschberger. 1997. **Survival Analysis - Techniques for Censored and Truncated Data**. Springer, New York.

Landau, B.J.; H.C. Kwaan; E.N. Verrusio; S.S. Brem. 1994. Elevated levels of urokinase-type plasminogen activator and plasminogen activator inhibitor type-1 in malignant human brain tumors. *Cancer Res.*, 54:1105–08.

Lee, J.W.; J.B. Lee; M. S. Park; H. Seuck. 2005. An extensive comparison of recent classification tools applied to microarray data. *Comp. Stat. Data Anal.*, 48:869-85.

Lee, K.E. & B.K. Mallick. 2003. Bayesian methods for variable selection in survival models with application to DNA microarray data. *Sankhya*, 4:756-78.

Li, H. & J. Gui. 2004. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20:i208-i215.

Lossos, I.S.; D.K. Czerwinski; A.A. Alizadeh; M.A. Wechser; R. Tibshirani; D. Botstein; R. Levy. 2004. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New Eng. J. Med.*, 350:1828-37.

Louis, D.N.; H. Ohgaki; O.D. Wiestler; W.K. Cavenee; P.C. Burger; A. Jouvet; B.W. Scheithauer; P. Kleihues. 2007. **The 2007 WHO Classification of Tumours of the Central Nervous System**. IARC Press, Lyon, France.

Maher, E.A.; C. Brennan; P.Y. Wen; L. Durso; K.L. Ligon; A. Richardson; D. Khatry; B. Feng; R. Sinha; D.N. Louis; J. Quackenbush; P.M. Black; L. Chin; R.A. DePinho. 2006. Marked genomic differences characterize primary and secondary glioblastoma subtypes and identify two distinct molecular and clinical secondary glioblastoma entities. *Cancer Res.*, 66:11502-13.

McCormick, D. 1993. Secretion of cathepsin B by human gliomas in vitro. *Neuropathol. Appl. Neurobiol.*, 19:146–51.

Morris, J.S.; P.J. Brown; R.C. Herrick; K.A. Baggerly; K.R. Coombes. 2008. Bayesian analysis of mass spectrometry data using wavelet-based functional mixed models. *Biometrics*, 64:479-489.

Nguyen, D.V. & D.M. Rocke. 2002. Partial least squares proportional hazard regression for applications to DNA microarray survival data. *Bioinformatics*, 18:1625-32.

Nicotera, P. & G. Melino. 2004. Regulation of the apoptosis-necrosis switch. *Oncogene*, 23:2757–65.

Nobusawa, S.; T. Watanabe; P. Kleihues; H. Ohgaki. 2009. IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas. *Clin. Cancer Res.*, 15:6002–7.

Nutt, C.L.; D.R. Mani; R.A. Betensky; P. Tamayo; J.G. Cairncross; C. Ladd; U. Pohl; C. Hartmann; M.E. McLaughlin; T.T. Batchelor; P.M. Black; A. von Deimling; S.L. Pomeroy; T.R. Golub; D.N. Louis. 2003. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.*, 63:1602–7.

Nyberg, P.; L. Xie; R. Kalluri. 2005. Endogenous inhibitors of angiogenesis. *Cancer Res.*, 65:3967–79.

Ohgaki, H.; P. Dessen; B. Jourde; S. Horstmann; T. Nishikawa; P.L. Di Patre; C. Burkhard; D. Schuler; N.M. Probst-Hensch; P.C. Maiorka; N. Baeza; P. Pisani; Y. Yonekawa; M.G. Yasargil; U.M. Lütolf; P. Kleihues. 2004. Genetic pathways to glioblastoma: A population-based study. *Cancer Res.,* 64:6892–99.

Ohtani, N.; Z. Zebedee1; T.J.G. Huot; J.A. Stinson; M. Sugimoto; Y. Ohashi; A.D. Sharrocks; G. Peters; E. Hara. 2001. Opposing effects of Ets and Id proteins on p16(INK4A) expression during cellular senescence. *Nature*, 409:1067-70.

Park, P.J.; L. Tian; I.S. Kohane. 2002. Linking expression data with patient survival times using partial least squares. *Bioinformatics*, 18:S120-S127.

Phillips, H.S.; S. Kharbanda; R. Chen; W.F. Forrest; R.H. Soriano; T.D. Wu; A. Misra; J.M. Nigro; H. Colman; L. Soroceanu; P.M. Williams; Z. Modrusan; B.G. Feuerstein; K.D. Aldape. 2006. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell,* **9:**157–73.

Pittman, J.; E. Huang; J. Nevins; Q. Wang; M. West. 2004. Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes. *Biostats*., 5:587–601.

Rodin, A.S.; A. Litvinenko; K. Klos; A.C. Morrison; T. Woodage; J. Coresh; E. Boerwinkle. 2009. Use of Wrapper Algorithms Coupled with a Random Forests Classifier for Variable Selection in Large-Scale Genomic Association Studies. *J. Comp. Biol.*, 16:1705–18.

Ross, J.S. 2009. Multigene classifiers, prognostic factors, and predictors of breast cancer clinical outcome. *Adv. Anat. Pathol.*, 16:204-15.

Schmid, M. & T. Hothorn. 2008. Flexible boosting of accelerated failure time models. *BMC Bioinformatics*, 9:1-13.

Sha, N.; M.G. Tadesse; M. Vannucci. 2006. Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, 22:2262-68.

Shchors, K.; E. Shchors; F. Rostker; E.R. Lawlor; L. Brown-Swigart; G.I. Evan. 2006. The Myc-dependent angiogenic switch in tumors is mediated by interleukin 1β. *Genes & Dev.*, 20:2527–38.

Shi,T. & S. Horvath. 2006. Unsupervised learning with random forest predictors. *J. Comp. & Grap. Stats*, 15:118-38.

Shiratsuchi, T.; H. Nishimori; H. Ichise; Y. Nakamura; T. Tokino. 1997. Cloning and characterization of BAI2 and BAI3, novel genes homologous to brain-specific angiogenesis inhibitor 1 (BAI1). *Cytogenet. Cell Genet.*, 79:103-8.

Somia, N.V.; M.J. Schmitt; D.E. Vetter; D. van Antwerp; S.F. Heinemann; I.M. Verma. 1999. LFG: an anti-apoptotic gene that provides protection from Fas-mediated cell death. *PNAS*, 96:12667-72.

Sørlie, T.; C.M. Perou; R. Tibshirani; T. Aas; S. Geisler; H. Johnsen; T. Hastie; M.B. Eisen; M. van de Rijn; S.S. Jeffrey; T. Thorsen; H. Quist; J.C. Matese; P.O. Brown; D. Botstein; P.E. Lønning; A. Børresen-Dale. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, 98:10869-74.

Stiver, S.I.: X. Tan; L.F. Brown; E.T. Hedley-Whyte; H.F. Dvorak. 2004. VEGF-A angiogenesis induces a stable neovasculature in adult murine brain. *J. Neuropathol. Exp. Neurol.*, 63:841–55.

Storey, J.D. 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Stats.*, 31:2013-35.

Tatard, V.M.; C. Xiang; J.A. Biegel; N. Dahmane. 2010. ZNF238 is expressed in postmitotic brain cells and inhibits brain tumor growth. *Cancer Res*. 70:1236–46.

Tatsuka, M.; H. Katayama; T. Ota; T. Tanaka; S. Odashima; F. Suzuki; Y. Terada. 1998. Multinuclearity and increased ploidy caused by overexpression of the aurora- and Ipl1-like midbody-associated protein mitotic kinase in human cancer cells. *Cancer Res.*, 58:4811-16.

The Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061-8.

Tibshirani, R. 1997. The Lasso method for variable selection in the Cox model. *Statis. Med.*, 16:385-95.

van Es, J.H.; M.E. van Gijn; O. Riccio; M. van den Born; M. Vooijs; H. Begthel; M. Cozijnsen; S. Robine; D.J. Winton; F. Radtke; H. Clevers. 2005. Notch/gamma-secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells. *Nature*, 435:959-63.

van't Veer, L.J.; H. Dai H; M.J. van de Vijver; Y.D. He; A.A. Hart; M. Mao; H.L. Peterse; K. van der Kooy; M.J. Marton; A.T. Witteveen; G.J. Schreiber; R.M. Kerkhoven; C. Roberts; P.S. Linsley; R. Bernards; S.H. Friend. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530-536.

van Wieringen, W.N.; D. Kun; R. Hampel; A. Boulesteix. 2009. Survival prediction using gene expression data: A review and comparison. *Comp. Stat. Data Anal.*, 53:1590-1603.

Velasco-Miguel, S.; L. Buckbinder; P. Jean; L. Gelbert; R. Talbott; J. Laidlaw; B. Seizinger; N. Kley. 1999. PA26, a novel target of the p53 tumor suppressor and member of the GADD family of DNA damage and growth arrest inducible genes. *Oncogene*, 18:127-37.

Verhaak, R.G.W.; K.A. Hoadley; E. Purdom; V. Wang; Y. Qi; M.D. Wilkerson; C.R. Miller; L. Ding; T. Golub; J.P. Mesirov; G. Alexe; M. Lawrence; M. O'Kelly; P. Tamayo; B.A. Weir; S. Gabriel; W. Winckler; S. Gupta; L. Jakkula; H.S. Feiler; J.G. Hodgson; C.D. James; J.N. Sarkaria; C. Brennan; A. Kahn; P.T. Spellman; R.K. Wilson; T.P. Speed; J.W. Gray; M. Meyerson; G. Getz; C.M. Perou; D.N. Hayes; TCGARN. 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell,* 17:98–110.

Wang, K.; H. Yamamoto; J.R. Chin; Z. Werb; T.H. Vu. 2004. Epidermal growth factor receptor-deficient mice have delayed primary endochondral ossification because of defective osteoclast recruitment. *J. Biol. Chem.*, 279:53848-56.

Wang, M.; T. Wang; S. Liu; D. Yoshida; A. Teramoto. 2003a. The expression of matrix metalloproteinase-2 and -9 in human gliomas of different pathological grades. *Brain Tumor Pathol.*, 20:65–72.

Wang, H.; H. Wang; W. Shen; H. Huang; L. Hu; L. Ramdas; Y.H. Zhou; W.S. Liao; G.N. Fuller; W. Zhang. 2003b. Insulin-like growth factor binding protein 2 enhances glioblastoma invasion by activating invasion-enhancing genes. *Cancer Res.*, 63:4315–21.

West, M. 2003. Bayesian factor regression models in the "large *p*, small *n*" paradigm. **In**: Bernardo, J.M.; M.J. Bayarri; J.O. Berger; A.P. Dawid; D. Heckerman; A.F.M. Smith; M. West. *Bayesian Statistics 7*. Oxford Univ. Press, Oxford.

Wick, W.: C. Wild-Bode; B. Frank; M. Weller. 2004. BCL-2-induced glioma cell invasiveness depends on furin-like proteases. *J. Neurochem.*, 91:1275–83.

Yeoh, E.J.; M.E. Ross; S.A. Shurtleff; W.K. Williams; D. Patel; R. Mahfouz; F.G. Behm; S.C. Raimondi; M.V. Relling; A. Patel; C. Cheng; D. Campana; D. Wilkins; X. Zhou; J. Li; H. Liu; C.H. Pui; W.E. Evans; C. Naeve; L. Wong; J.R. Downing. 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133-43.

Yu, L.; C. Liu; J. Vandeusen; B. Becknell; Z. Dai; Y.-Z. Wu; A. Raval; T.-H. Liu; W. Ding; C. Mao; S. Liu; L.T. Smith; S. Lee; L. Rassenti; G. Marcucci; J. Byrd; M.A. Caligiuri; C. Plass. 2005. Global assessment of promoter methylation in a mouse model of cancer identifies ID4 as a putative tumor-suppressor gene in human leukemia. *Nature Genet.*, 37:265-74.

Zeller, C.; B. Hinzmann; S. Seitz; H. Prokoph; E. Burkhard-Goettges; J. Fischer; B. Jandrig; L.E. Schwarz; A. Rosenthal; S. Scherneck. 2003. SASH1: a candidate tumor suppressor gene on chromosome 6q24.3 is downregulated in breast cancer. *Oncogene*, 22:2972-83.

Zou, H. & T. Hastie. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, 67:301-20.

Vita

Vinícius Bonato was born in Brasília, Distrito Federal, Brazil on November 18, 1977, the son of Traudi Helena Osterkamp and Carlos Roberto Bonato. After completing his professional high-school in Food Sciences at Colégio Técnico da Unicamp (Cotuca), Campinas, São Paulo state, Brazil in 1995, he entered at Universidade Estadual de Campinas (Unicamp), Campinas, São Paulo state, Brazil. He received the degree of Bachelor of Sciences with a major in Biological Sciences from Unicamp in December, 1999. Then he received a M.Sc. degree in Ecology and Evolutionary Biology in 2002 and a Ph.D. in Ecology and Evolutionary Biology in 2004 both from Unicamp. He reenrolled in the undergraduate program of Statistics in 2005 attending classes until July 2007 at Unicamp. He worked from June 2002 to June 2007 as a Biology instructor at two colleges in Brazil: Universidade Metodista de Piracicaba, Piracicaba, São Paulo state, and Universidade do Espírito Santo do Pinhal, Espírito Santo do Pinhal, São Paulo state. In July of 2007 he entered The University of Texas Health Science Center at Houston Graduate School of Biomedical Sciences.

Permanent address:

Rua do Parque, 153 Apto. 22, Condomínio Bromélias

Parque Villa Flores

Sumaré, SP, Brazil, 13175-660