INFORMATICS FRAMEWORK FOR EVALUATING MULTIVARIATE PROGNOSIS MODELS: APPLICATION TO HUMAN GLIOBLASTOMA MULTIFORME

by

Stephen Richard Piccolo

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Biomedical Informatics

The University of Utah

May 2011

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Stephen Richard Piccolo**

has been approved by the following supervisory committee members:

Lewis J. Frey	, Chair	**11/18/2010** Date Approved
Philip S. Bernard	, Member	**11/18/2010** Date Approved
Lisa Cannon-Albright	, Member	**11/18/2010** Date Approved
Karen Eilbeck	, Member	**11/18/2010** Date Approved
Peter J. Haug	, Member	**11/18/2010** Date Approved

and by **Joyce A. Mitchell** , Chair of the Department of **Biomedical Informatics**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

For decades, researchers have explored the effects of clinical and biomolecular factors on disease outcomes and have identified several candidate prognostic markers. Now, thanks to technological advances, researchers have at their disposal unprecedented quantities of biomolecular data that may add to existing knowledge about prognosis. However, commensurate challenges accompany these advances. For example, sophisticated informatics techniques are necessary to store, retrieve, and analyze large data sets. Additionally, advanced algorithms may be necessary to account for the joint effects of tens, hundreds, or thousands of variables. Moreover, it is essential that analyses evaluating such algorithms be conducted in a systematic and consistent way to ensure validity, repeatability, and comparability across studies. For this study, a novel informatics framework was developed to address these needs. Within this framework, the user can apply existing, general-purpose algorithms that are designed to make multivariate predictions for large, hetergeneous data sets. The framework also contains logic for aggregating evidence across multiple algorithms and data categories via ensemble-learning approaches. In this study, this informatics framework was applied to developing multivariate prognisis models for human glioblastoma multiforme, a highly aggressive form of brain cancer that results in a median survival of only 12-15 months. Data for this study came from The Cancer Genome Atlas, a publicly available repository containing clinical, treatment, histological, and biomolecular variables for hundreds of patients. A variety of variable-selection approaches and multivariate algorithms were applied in a cross-validated design, and the quality of the resulting models was measured using the error rate, area under the receiver operating characteristic curve, and log-rank statistic. Although performance of the algorithms varied substantially across the data categories, some models performed well for all three metrics—particularly models based on age, treatments, and DNA methylation. Also encouragingly, the performance of ensemble-learning methods

often approximated the best individual results. As multimodal data sets become more prevalent, analytic approaches that account for multiple data categories and algorithms will be increasingly relevant. This study suggests that such approaches hold promise to guide researchers and clinicians in their quest to improve outcomes for devastating diseases like GBM.

To Laurel, Kate, and Joshua

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Upon entering the Biomedical Informatics program at the University of Utah, I had only a vague idea of the field of biomedical informatics. I had information-systems expertise but had taken only one college-level course in biology—12 years prior to that time—and had little understanding of the medical, statistics, and public-health disciplines. Despite that lack of knowledge, I had a strong desire to learn and to develop my skills. Thanks to the guidance and patience of numerous mentors, friends, and family members, I have grown immensely during my time as a doctoral student and feel prepared for the exciting opportunities that lie ahead.

First and foremost, I gratefully acknowledge my wonderful wife, Laurel (Harmon) Piccolo. We started dating as I started this doctoral program, and we were married by the end of my first year. Many times I have been discouraged or flustered, and just as many times she has buoyed me up with heartfelt words of encouragement and expressions of confidence. Additionally, Laurel has cared for our wonderful children, cooked glorious meals, and performed innumerable chores to make up for the long hours required by my studies. What a marvelous companion, support, and friend!

Secondly, I acknowledge our two young children, Kate and Joshua. Their smiling faces, innocent love, and abundant energy have lifted my spirits each day and provided motivation to push through difficult obstacles.

Thirdly, I express gratitude to several family members who have been instrumental in helping me reach this point. From an early age, my mother, Alice (Cook) Piccolo, encouraged me to explore a variety of disciplines. With her encouragement, I attended a computer-skills class in high school and discovered a new aptitude and interest. Mom has always been a comforting influence in my life and helped me cope with challenges. My father, Richard Piccolo, has taught me the value of consistency, hard work, and responsibility. During most of his career, Dad has arisen before sunrise and ridden his bicycle to work. Often he has followed his day job with a second or

third job to support his family. I also thank my aunt, Janice Piccolo, and my pseudo-aunt, Linda Palmer. In addition to offering their cheery demeanors and insightful perspectives on life, they generously allowed us to live in their basement apartment during the final (and most demanding) stages of completing this dissertation. And many thanks be to Roger and Maxine Harmon, Laurel's parents, who brighten my day each time I see them and who continuously offer support beyond any expectation.

Next I heartily thank Dr. Lewis J. Frey, who has guided me through the long yet intellectually stimulating process of defining a project and executing it. Lewis made himself available for weekly meetings, which have been invaluable in moving this project forward. Along the way, he has challenged me to think innovatively and allowed me flexibility in defining this project. Many times his intuition has proven priceless.

Many thanks also go to Drs. Philip Bernard, Lisa Cannon-Albright, Karen Eilbeck, Peter Haug, Joyce Mitchell, Julio Facelli, and John Hurdle, who have provided mentoring and guidance. In addition to offering specific advice, they have exemplified a desire to improve humanity through their research.

Completing this dissertation would not have been feasible without the generous training fellowship (1T15-LM007124) provided by the National Library of Medicine (NLM) and indirectly by the American public. Of high importance to our family, this income enabled Laurel to stay home with Kate and Joshua without her having to seek outside employment. Many thanks also go to the Evan Johnson Lab at Brigham Young University and the Andrea Bild lab at the University of Utah for providing part-time employment and generous funding during the months after my NLM fellowship ended.

Many thanks are in order for several people with whom I did not interact directly but without whom this study would not have been possible. Among these individuals are 1) the sample donors who contributed data to The Cancer Genome Atlas and the research groups who preprocessed the data and hosted it, 2) research groups who donated data to the UCI Machine Learning Repository, 3) open-source software developers who created and refined tools that were used for this project, and 4) referees who provided valuable feedback on a paper that was submitted to the Neuro-

Oncology journal.

An allocation of computer time from the Center for High Performance Computing at the University of Utah is gratefully acknowledged. An allocation of computer time from the Fulton Supercomputing Laboratory at Brigham Young University also is gratefully acknowledged.

I appreciatively acknowledge the many individuals who have made the Department of Biomedical Informatics and the University of Utah what they are today. Several administrative staff in the department have been particularly helpful and have also been friends: Kathy Stoker, JoAnn Thompson, Linda Galbreath, Teri Birch, and Lynn Ford. I express many thanks and much admiration to my fellow students who have provided fellowship, inspiration, and even laughs. Thanks also to Jonathan Balzotti of the University Writing Center for style suggestions in writing parts of this manuscript.

Of special note, I acknowledge Ms. Kim D. Murphy, who was a project coordinator for Dr. Frey through the core part of my dissertation research. She has offered valuable advice, is a friend of the family, and has provided inspiration through her courageous outlook in the face of a recent diagnosis with stage IV colorectal cancer.

Finally, I thank God for giving me life, for strengthening me through trials, and for blessing me with health, guidance, and peace.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Glioblastoma multiforme (GBM) is a highly aggressive form of brain cancer that results in a median survival of 12-15 months, [2] a short duration relative to most other cancers. [3] However, the length of time that GBM patients survive after diagnosis is variable—some patients survive only a few weeks while others survive many years [4]—a pattern that suggests that individual patient and tumor characteristics influence tumor aggressiveness, responses to treatments, and ultimately survival. Unfortunately, despite extensive research efforts to date, GBM survival has not been extended appreciably at the population level. [5] An understanding of factors associated with GBM survival time could help researchers and physicians identify patients less likely to respond to standard treatments [1] and could help researchers identify mechanisms driving disease severity.

The lack of progress in developing reliable methods to prospectively differentiate between longer-term survivors (LTS) and shorter-term survivors (STS) likely results—at least in part—from the complexity of interactions between clinical, demographic, and treatment factors as well as from the biological complexity of tumor initiation and progression. Various studies have identified an array of candidate prognostic factors for GBM, [2,4,6–19] yet it has remained unclear how best to account for the effects of multiple factors working in concert to affect prognosis. In response to this problem, some researchers have proposed multivariate prognosis models. [1, 11, 19, 20] While such studies have shown promise for aiding clinicians with the task of predicting a patient's prognosis, no existing model attempts to account for all candidate prognostic factors for GBM—existing models typically account for a single category of molecular data and one or more clinical variables such as age. Additionally, many algorithms exist for developing multivariate prognosis models, but each GBM study has typically

employed only one such technique. Two reasons for these limitations are apparent: 1) it is economically infeasible for individual research labs to acquire multiple categories of clinical and molecular data for a reasonably sized patient cohort, and 2) no informatics framework has been available for performing analyses that compare and combine the outputs of multiple algorithms in a systematic and consistent way. Fortunately, a recent initiative by the United States National Institutes of Health has resulted in a publicly available data repository—The Cancer Genome Atlas (TCGA) [21]—that contains data for hundreds of GBM patients. This resource is unique in the breadth of data it contains—TCGA contains thousands of data points representing clinical characteristics, treatments administered, and molecular features profiling each patient's tumor. Such a resource enables an unprecedented opportunity to evaluate previously reported prognosis models, to discover new candidate prognosis models, and to make systematic comparisons across data categories and algorithms. Accordingly, the overaching goal of this research was to develop an informatics framework to perform such analyses and to apply the framework to an analysis of GBM survival.

1.2 Main Objectives

Having a vast array of data for each patient presents not only opportunities but also challenges. One important challenge is to filter the data into variable sets that represent the key factors influencing patient outcome and that ignore irrelevant factors. A related challenge is to develop models that account for potentially intricate combinations of heterogeneous factors that jointly affect patient outcome. In response to such challenges, the statistics and machine-learning communities have developed general-purpose algorithms designed to filter variable sets and to make multivariate predictions in many diverse contexts; however, such algorithms have been applied only minimally in the context of predicting GBM survival, so their potential usefulness in this setting is unknown. Thus, one key objective of this research was to apply machine-learning algorithms in this context.

Theory and practical application suggest that no single algorithm is universally optimal for performing classification tasks. [22] One algorithm may be well suited for extracting meaningful patterns from a particular class of patient data whereas a

second algorithm may be less suitable for the task. However, the second algorithm may, in turn, be more effective than the first algorithm on a different class of patient data. The heterogeneous nature of the TCGA data provides further justification to employ multiple algorithms—for example, some variables are naturally discrete (e.g., sex, mutation status) while others are inherently continuous (e.g., age, gene expression) and follow diverse numerical distributions. Thus, another key objective of this research was to explore methods that combine the outputs of multiple algorithms into a unified prediction for each patient.

Having identified a multivariate model that differentiates successfully between two classes (e.g., LTS and STS), a subsequent challenge is to interpret the model. This challenge can be especially difficult when the model contains a large number of variables, a scenario often encountered with biomolecular-profiling data. Invididual variables used in the model can be evaluated one at a time. However, it is often desirable to interpret a model as a whole to better understand the underlying mechanisms driving the model's success. Gene-set enrichment analysis (GSEA) [23] is one approach for interpreting sets of genomic variables. In GSEA, selected genes are evaluated against known functional categories in an attempt to identify functional categories that are associated with the genes more than would be expected by chance. However, when only a subset of known genes have been profiled, a selection bias may impact the results and subsequent intepretation. Consequently, another objective of this research was to explore this bias and develop a method to account for it.

1.3 Hypotheses

In accordance with the main objectives, the following hypotheses were evaluated in this study:

- Multivariate algorithms can be used to derive clinical and biomolecular models that differentiate significantly between GBM patients who survived a relatively long (LTS) or short (STS) time after diagnosis.

- Methods that aggregate evidence across multiple data categories and algorithms can differentiate between LTS and STS better than using methods that use

evidence from a single data category or algorithm.

- Biologically relevant aspects of the models can be identified via comparison with prior knowledge and via accounting for gene-selection bias.

The remaining sections of this dissertation describe in further detail the research that was performed in this study. The Background section describes the clinical need motivating this study, prior research that has been conducted in this area, and how this study aims to improve on prior work. The Methods section describes the algorithms that were used to formulate GBM survival models, explains modifications that have been made to existing approaches, and outlines the study's experimental design; additionally, this section describes the informatics framework that was developed and details at a technical level its advantages over other approaches. The Results section outlines the findings of this study in substantial detail. The Discussion section provides additional interpretation of the results, addresses notable side observations that came about during the analysis, explains limitations of this study, and suggests how the research could be developed further in future studies. Finally, the Conclusion postulates on the potential implications of this study for the biomedical informatics research community.

CHAPTER 2

BACKGROUND

2.1 Univariate Predictors of GBM Survival

Over the past decades, researchers have searched for clinical, histological, and treatment-associated factors that appear to shorten or lengthen overall GBM survival. An association between age at diagnosis and GBM survival has been reported consistently and repeatedly. [4, 6–15] Other factors reported to have some prognostic relevance include Karnofsky Performance Status (KPS) (a measure of that patient's well being at the time of diagnosis), [8–14] extent of tumor resection, [9, 13–15] radiation therapy, [6, 14, 16] and tumor necrosis. [7, 9, 24] In 2005, a phase III clinical trial also suggested that treatment with Temozolomide, an oral alkylating agent, extends survival in many patients. [2]

Due to technological advances, researchers have also searched for prognostic factors at the biomolecular level. Although many types of biomolecular alterations—for example, DNA point mutations, DNA methylation changes, DNA amplifications and deletions, and mRNA expression changes—have been observed regularly in GBM tumors, [17, 25, 26] a prognostic relationship has been demonstrated for only a few alterations, including *IDH1* mutations, [17, 27] *MGMT* hypermethylation, [12, 18] *EGFR* amplification, [13] *CDK4* amplification, [14] *MDM2* amplification, [13] and *FABP7* expression. [19] Unfortunately, for some of these associations, conflicting results have been observed in other studies. [4, 12, 14, 16]

2.2 Multivariate Predictors of GBM Survival

Though it is promising that individual factors with prognostic relevance have been observed for GBM, their value for predicting a given patient's survival may be limited, because individual factors often have a variable impact on survival. This variability may, in part, result from combinations of factors that have cumulative or

interacting effects. [28, 29] For example, in a recent study, simultaneous EGFR and p53 alterations appeared to have a joint effect on GBM survival, whereas no effect was observed for either alteration alone. [14] To account for such combinatory effects, researchers have proposed multivariate models. [1, 11, 19, 20]

For example, in one study, multivariate techniques were applied to mRNA expression data for a group of GBM patients. [19] Liang, et al. examined whether GBM patients—whose tumors by definition have similar histopathological features—could be grouped into survival-associated subpopulations, based on genomewide mRNA expression levels in the tumors. [19] They used agglomerative hierarchical clustering, a popular *unsupervised-learning algorithm*, to divide patients into subpopulations (clusters) that had relatively similar mRNA-expression profiles. Using a subset of genes that were correlated with survival, they observed that the two upper-level clusters had median survivals of 4 months and 25 months, respectively. Although the generalizability of the model was not assessed, this finding suggested that mRNA expression data may have potential to differentiate between LTS and STS.

Unsupervised-learning algorithms, such as the hierarchical-clustering method used by Liang, et al., are designed to extract instrinsically meaningful patterns from data, but they may not identify patterns that are optimally relevant to a particular outcome (i.e., survival)—by nature, unsupervised-learning methods ignore outcomes. While such approaches are useful in many scenarios (including when the outcome is unknown), they may not be ideal when the researcher's intent is to correlate independent variables with a particular outcome. Thus, in complement to unsupervised-learning methods, the statistics and computer-science communities have also developed a large variety of *supervised-learning algorithms* designed to identify multivariate patterns associated with outcomes of interest. Hundreds of supervised-learning algorithms have been published in the literature and have been applied in many fields of study, but such methods have been applied only minimally to GBM data in relation to survival. [1, 11, 30]

In one study, Lamborn, et al. used *recursive partitioning analysis* (RPA), [31] a supervised-learning technique that divides patients into subpopulations according to associations between combinations of independent variables and the outcome. [11]

Lamborn, et al. investigated whether combinations of clinical variables could be used to assign GBM patients to risk groups that had significantly different overall survivals. Derived from a large sample of GBM patients who had been enrolled in clinical trials and who had received radiation treatment, their RPA model was based on age at diagnosis, KPS, and tumor anatomic site. When the model was evaluated on the same data set from which it was derived, two-year survival estimates for the groups were 65%, 35%, 17%, and 4%, respectively. In a follow-up analysis, they also observed that this method could be used to detect survival-associated subpopulations among patients that had received a particular treatment regimen. Although the generalizability of the models was not assessed, the findings suggested that multivariate models based on clinical data have potential to elucidate differences in survival expectations among GBM patients as a whole or among patients receiving a given treatment.

In a recent study, Colman, et al. applied supervised-learning methods to mRNA expression data. [1] They explored the potential to predict an individual GBM patient's survival status at the time of diagnosis. After examining multiple independent data sets, they identified nine genes that consistently were differentially expressed between patients who survived fewer than two years and patients who survived longer. Then they developed a risk score based on the combined expression of the nine genes: genes that were typically over-expressed were assigned a positive weight, while under-expressed genes were assigned a negative weight. Then an optimal cutoff value for distinguishing LTS from STS was identified using RPA. Colman, et al. tested this method on a separate data set and found that the multigene score was an independent, significant predictor of survival.

The results published by Colman, et al. suggest clinical viability for multivarate prognosis models based on mRNA expression data. In fact, based on their findings, the authors have marketed a test designed to distinguish LTS from STS in clinical and research settings. [32] However, further exploration of the utility of such methods is warranted. For example, Colman, et al. used univariate statistical approaches to identify differentially expressed genes, yet multivariate approaches may reveal gene subsets that influence survival jointly but not individually. [14] Additionally, the

multigene scoring approach employed by Colman, et al. is basic and unsuitable for data sets that contain discrete values. Accordingly, two foci of this study were to use multivariate algorithms to filter variables sets and to make survival predictions using multiple categories of clinical and molecular data.

2.3 Ensemble Predictors of GBM Survival

The studies described so far have been limited in at least two ways. Firstly, each study evaluated a single category of GBM patient data (i.e., clinical, mRNA) in isolation, whereas many types of prognostic factors (e.g., clinical, treatments, biomolecular) have been observed for GBM patients and thus may influence GBM tumor progression. [25] Secondly, each study reported applying only a single algorithm to a respective data set, and it is likely that different algorithms perform better for different categories of data or even for different subsets of patients. These limitations appear commonly in other studies of GBM survival. Accordingly, other key foci of this study were to evaluate the potential prognostic value of various categories of patient/tumor data and the effectiveness of multiple algorithms for producing accurate predictions of GBM survival.

Demonstrating the potential to develop prognostic models for a single category of patient/tumor data provides an indication of the priority that should be placed on collecting that data. However, intuition suggests that it should be possible to improve prognostic models by aggregating insights across multiple data categories and algorithms. In one study that implemented such an approach, Nigro, et al. [33] acquired mRNA expression data and DNA copy-number data for a cohort of GBM patients. Having applied filtering and clustering methods for the data categories separately, they found that in both cases, one cluster contained predominantly STS ($<$ 2 years survival), while the other cluster had a combination of STS and LTS ($>$ 24 months). Importantly, individual patients were not grouped identically for the two data categories, an indication that the mRNA expression and DNA copy-number data contained complementary prognostic information. Consequently, the authors concluded that "analyzing genetic signatures on both DNA and RNA levels may result in more robust molecular classification schemes than those derived from either

experimental method alone" (p. 1685). [33] However, few such studies have been conducted using multiple categories of patient/tumor data for the same set of patients. Part of the reason may be that it has been economically infeasible for individual research labs to obtain high-throughput molecular data—let alone multiple categories of molecular data—for large patient cohorts. In fact, the patient cohort for the Nigro, et al. study consisted of only 34 patients. Fortunately, the U.S. National Cancer Institute and National Human Genome Research Institute have teamed to invest in TCGA with the goal to "improve our ability to diagnose, treat, and prevent cancer" by exploring "the entire spectrum of genomic changes involved in human cancers". [34] TCGA contains detailed clinical data (including time to survival) and multiple categories of high-throughput biomolecular data for a cohort of over 300 GBM patients. Thus this vast resource provides an unprecedented opportunity to explore multivariate prognosis models.

Although many supervised-learning algorithms exist, no single algorithm is universally optimal. [22] And generally speaking, the performance of a given algorithm depends on the nature of the data to which it is applied. [22] Thus it may be useful to apply multiple algorithms to potentially shed light on which algorithm(s) are well suited (or not) to a particular category of data. However, lacking a priori knowledge of which algorithm would perform best in a given scenario, *ensemble methods*—which consider the outputs of multiple algorithms—have been suggested as a way to yield better results than any single algorithm. [35] To gain advantage from combining the outputs of multiple algorithms, it is desirable that the algorithms vary substantially in the predictions that they make, a phenomenon known as classifier diversity. [36] When classifier diversity exists (and in the common scenario where algorithms predict imperfectly), the various algorithms commit prediction errors on different patients. Ensemble methods attempt to capitalize on classifier diversity by aggregating the collective insights provided by the individual algorithms—where one algorithm fails, another algorithm may succeed. Many ensemble approaches have been proposed in the literature, ranging in complexity from a simple majority vote [37] to more advanced methods, such as those that assign weights to predictions based on the perceived quality of the predictions. [38] In this study, a variety of such ensemble

approaches have been applied to the problem of GBM survival prediction.

2.4 Assessing Clinical Relevance

Although it is an interesting academic exercise to apply multivariate algorithms to patient data in novel ways, the value of such research can be increased substantially if it can be shown that the methods have clinical utility. In the case of survival analyses, clinical utility means demonstrating that patients who survive a relatively long period of time can be distinguished from patients who survive a relatively short time. Having separated patients into distinct groups, standard survival-analysis techniques (e.g., the log-rank test [39] and Kaplan-Meier curves [40]) can be used to assess the overall differences in survival between the groups. A method that successfully separates patients into survival-associated groups can be invaluble in clinical settings at the time of diagnosis. [41] For example, if a patient knew her survival would likely be relatively short, she might opt for more aggressive treatments and/or to enter a (potentially risky) clinical trial. [42]

One advantage of unsupervised-learning techniques is that they assign patients to discrete groups, so the overall survival of the groups can be compared in a straightforward way using survival-analysis techniques. Conversely, supervised-learning algorithms rely upon an explicit outcome, which they attempt to predict. For the analyses performed in this study, the outcome is survival, which is naturally a continuous variable, and obvious groupings among patient survival times may not arise (as in GBM). Consequently, researchers often use an alternative approach—patients surviving longer than a given number of days or years are placed in one group, and patients surviving shorter than that threshold are placed in another group. By dividing the patients into distinct groups, researchers are able to present findings that are intuitive to clinicians and that can be assessed using survival-analysis techniques that are familiar to clinicians. An additional reason for grouping patients this way is that supervised-learning algorithms have traditionally focused more on problems with discrete outcomes, and thus existing methods are more developed in this area. Thus, in this study, the outcome variable is a discretized form of survival: patients are designated as either LTS or STS based on whether the patient survived a specific

number of days. (Details are outlined in the Methods section regarding the various methods used to perform this discretization.)

2.5 Assessing Biological Relevance

If a model based on biomolecular data can be demonstrated to have high clinical relevance to GBM survival, a logical next goal is to interpret the biological meaning of the model. Identifying specific biological mechanisms that drive GBM tumor aggressiveness could help bench researchers develop targeted treatments, potentially leading to longer survival for patients subject to those biological mechanisms. In addition to exploring biological mechanisms at a granular level, researchers have also attempted to gain insights on the underlying functional categories associated with the genes that appear to have most influence on the outcome. This class of methods, commonly referred to as gene-set enrichment analysis (GSEA), [23] performs a functional analysis on the top-level genes and identifies the known biological processes or functions that are associated with the selected genes. Because some biological processes have more genes associated with them than others, GSEA methods typically correct for such bias. However, GSEA methods are typically designed for analysis of high-throughput molecular data–such as mRNA-expression microarray data—that profile almost every known gene. When custom (not genomewide) molecular assays are used for molecular profiling, GSEA may need further refinement. In developing such assays, researchers must decide a priori which molecular features to measure, due to cost and/or technological restrictions. Because these decisions are at least partially subjective, it is plausible that molecular features are selected in favor of genes already believed to influence the disease(s) being studied. This potential limitation is of particular relevance to researchers investigating multimodal data sets like TCGA because each molecular-profiling platform measures a different set of molecular variables. So having a method that could be applied more generally and that could account for selection bias would be desirable.

CHAPTER 3

METHODS

3.1 Data

GBM data for this study were downloaded from the TCGA data portal [43] on August 26, 2010. For each GBM patient, the portal contains clinical observations, treatments that have been administered, histological observations of tumor samples, number of days that patients survived, and biomolecular data that have been acquired using various high-throughput profiling technologies. For this study, age, sex, and Karnofsky performance status (KPS) were labeled as *clinical* variables. *Treatment* variables were the following: radiotherapy, temozolomide, dexamethasone, lomustine, bevacizumab, or *other drug* (indicating whether a given patient had received any other type of drug treatment); each treatment variable contained binary values indicating whether patients had received any of the given treatment. *Histology* variables included in the analysis were number proliferating cells, percent tumor cells, percent tumor nuclei, percent necrosis, percent stromal cells, percent inflammatory infiltration, percent lymphocyte infiltration, percent monocyte infiltration, percent granulocyte infiltration, percent neutrophil infiltration, percent osinophil infiltration, presence of endothelial proliferation, presence of nuclear pleomorphism, presence of palisading necrosis, and presence of cellularity. When a low value range (e.g., <5% necrosis) was given, the range was rounded to bottom of the range; high-value ranges (e.g., >95% tumor cells) were rounded up.

Molecular-level data representing DNA somatic mutations, DNA copy number changes, DNA methylation states, mRNA expression levels, and miRNA expression levels were also included in the analyses. These data were acquired using the following technologies: Sanger sequencing, Agilent Human Genome CGH Microarray 244A microarrays, Illumina DNA Methylation OMA002 and OMA003 custom pan-

els, Affymetrix HG-133A microarrays, and Agilent 8 x 15K Human miRNA-specific microarrays.

The TCGA data portal contains raw data as well as data that have been pre-processed by the TCGA Consortium. [21] For somatic mutations, TCGA "Level 3" data, which had been summarized at the gene level, were used. Somatic mutations that had been marked as "silent" were excluded, based on an expectation that synonymous mutations would by nature not have prognostic relevance. Additionally, only mutations that had been validated and were considered "somatic" by the TCGA Consortium were included. Finally, any gene that had fewer than two mutations was excluded. For DNA copy number, the Level 3 data were mapped to the UCSC hg19 version of the human genome, [44] and a mean value was calculated for each chromosomal band. [45] For DNA methylation, Level 2 data were used (in the absence of Level 3 data); these data contain "beta values" representing the proportion of methylated molecules on the complementary probes for each locus. These data had not been summarized at the gene level; thus, to get gene-level values, the mean beta value across all probes associated with a given gene was calculated. For mRNA and miRNA expression, Level 3 data, which had already been preprocessed and summarized, were used.

For each patient, the overall survival time was calculated as the difference in days between the date of initial pathologic diagnosis and the date of decease. Patients having no recorded diagnosis date, having a pretreatment history, or missing >50 percent of data for a given data category were excluded. After performing this filtering step, data for 313 GBM patients remained. Of those patients, 307 had data for at least four data categories, and 100 had data for all categories. Table 3.1 lists the number of patients and the number of data points for each category, along with the proportion of missing values for each category. The *All Data* category represents a combined data set containing the union of all other data categories.

3.2 Model Validation Procedure

Even when a multivariate algorithm is capable of fitting a model to a data set with extremely high accuracy, the model may fail to generalize and thus have limited

clinical relevance for the overall population. Cross validation is one technique for assessing how well models generalize. Such an approach is essential, especially when analyzing a data set like TCGA that is drawn from multiple institutions, includes heterogeneously treated patients, and contains thousands of potential predictor variables.

In cross validation, the data instances (patients) are partitioned into sets of equal size (or as close to equal as possible). In turn, each set is held separate for testing, and the remaining instances are used to train a model. The trained models then are used to predict the outcome of the respective test instances. One cross-validation parameter that can be varied is the number of folds used. For example, in ten-fold cross validation, the data set is divided into ten partitions, resulting in ten disjoint sets of test instances. In leave-one-out cross validation, each test set contains only a single instance, and the remaining instances are used for training. In this study, ten-fold cross validation was used. Where possible, stratification was also used; stratification attempts to spread instances from each class (e.g., LTS, STS) evenly across the folds.

In order to estimate optimal parameters for training a model, this study also employed *nested cross validation*; with this approach, instances in each training set are further subdivided into internal training and test sets; the parameters that perform best internally are then used to train models on the outer training sets. Nested cross validation provides performance estimates that approximate what can be obtained on independent test sets. [46]

Having assigned patients to cross-validation folds, variable-selection approaches were applied to the training sets. The purpose of such approaches is to reduce the effect of noise in the data by focusing the models on variables that are most relevant to the outcome. Having ranked the variables for each data category, the top-1, 5, 10, 50, 100, 500, or 1000 variables (if that many were available) were identified for each outer training set. Because the optimal number of variables to include in a given model was unknown, a similar selection procedure was applied to the internal cross-validation folds. The number of top-ranked variables, n, that performed best on each training set—based on lowest average area under the receiver operating characteristic curve (AUC) across the internal folds—was considered optimal. Classification algorithms

then used the best n features to train a model and predict the survival status of patients in the respective outer folds. This procedure was performed independently for each combination of data category, feature-selection approach, and classification algorithm.

3.3 Variable Selection Approaches

In filtering variable sets, four selection approaches were used.

The first approach, *None*, is simply to perform no filtering. Two reasons for using this approach may have merit: 1) it is possible that every variable contributes some information, so excluding any variable may reduce classification performance, and 2) many classification algorithms have intrinsic techniques for filtering variables that are particularly well suited to those algorithms, so pre-filtering with a contradictory approach may negatively impact classification performance.

The second selection approach, *prior knowledge filtering*, requires a manual literature review to identify variables that have been reported to bear prognostic relevance for each of the data categories. The purpose of applying this technique is to identify all GBM prognosis variables for which the scientific community has reached some consensus. Selecting variables based on prior knowledge also sets a baseline against which the quantitative algorithms can be compared. For this study, a variable was considered to constitute prior knowledge if two or more articles, published in peer-reviewed journals, had reported the variable to be prognostic for GBM. Two exceptions to this rule were allowed: 1) if only one article had reported any candidate prognosis variable for a given data category, the variables from that article were used, or 2) if a single study had validated candidate variables across multiple independent data sets, those variables were considered robust and given preference over variables reported in separate studies. Few articles researching the prognostic relevance of miRNA expression have been published to date, so a single article was considered sufficient for that category. The Colman et al. mRNA expression signature was derived from multiple independent studies and thus was used. Table 3.2 lists each data category along with the prior-knowledge variables that were selected for each category.

Another selection approach used in this study was the Support Vector Machines-Recursive Feature Elimination (SVM-RFE) algorithm. [47] SVM-RFE is based on Support Vector Machines, a powerful classification algorithm that assigns a weight to each variable, quantifying its ability to discriminate the classes. SVM-RFE uses a backward search: variables with the lowest weights are removed in an iterative fashion, and variables are ranked according to the order in which they are eliminated. [47] SVM-RFE has been shown to perform well on high-dimensional data sets with complex dependencies among variables. [47] The implementation of this algorithm in the Weka software package [48] was used in this study. The algorithm was configured to eliminate 10% of variables in each iteration; when 10 or fewer variables remained, one variable was eliminated in each iteration. Otherwise, default configuration settings were used.

The final variable-selection approach was the RELIEF-F algorithm. [49] For a random subset of instances (i.e., patients in this case), RELIEF-F compares each instance against other instances of the same class (i.e., LTS or STS) and of a different class; it then calculates a score for each variable based on whether values are similar for instances of the same class and dissimilar for instances of a different class. [49] The resulting score is a continuous value with higher numbers signifying a better ability to differentiate and zero/negative values signifying no ability to differentiate. This study used the Weka implementation of this algorithm with default configuration settings.

3.4 Classification Algorithms

In TCGA, some data categories contain hundreds or thousands of independent variables. Additionally, variables vary not only in their semantics but also in their scales of measurement. [50] Consequently, classification algorithms used in this study needed to be capable of handling a large number of variables and multiple variable types. The following algorithms meet these criteria and were employed in this study: C5.0 Decision Trees, Naïve Bayes Classifier (NBC), and Support Vector Machines (SVM). Each of these algorithms has been applied broadly in a variety of contexts and represents a considerably different algorithmic approach.

The C5.0 Decision Trees algorithm [51] is conceptually similar to the RPA al-

gorithm (used in the Lamborn, et al. study [11]) in that it uses combinations of variables to assign patients to subgroups that are homogeneous in relation to the outcome variable; however, unlike RPA, C5.0 Decision Trees is designed to handle both continuous and discrete variables, [52] handle missing values, perform well on large data sets, and account for intervariable dependencies.

NBC is based on Bayes' Theorem of conditional probabilities and calculates the class' posterior probabilities as the product of the conditional probabilities for each variable. [53] For simplicity, NBC assumes independence between variables; yet despite its simplicity, NBC often performs as well as or better than more sophisticated algorithms. [54] For this study, the Weka implementation of this algorithm was used. Instead of the default settings, which assume that continuous variables follow a normal distribition (which is not always the case in the TCGA data), a nonparametric kernel density estimator was used to characterize continuous variables. This method has been shown to reduce errors compared to the normality assumption. [55]

The SVM algorithm [56] uses a mathematically derived hyperplane to separate instances of different classes; the instances lying on the hyperplane's margin constitute support vectors, and the algorithm seeks a maximal margin between the hyperplane and the support vectors. [57] For this study, the Weka wrapper of the LibSVM library [58] and the radial-basis function kernel (default setting) were used.

3.5 Ensemble Learning Approaches

Ensemble-learning methods are designed to create aggregate predictions based on evidence from multiple individual predictions. For this study, a variety of existing ensemble approaches were applied, and modified versions of existing approaches were developed. In performing ensemble learning, predictions from each combination of data category, feature-selection approach, and classification algorithm were considered. (Predictions for the *All Data* category were excluded.)

The first and simplest ensemble approach, *majority vote*, [37] counts the number of predictions a patient received for a given outcome (i.e., LTS or STS) and makes an aggregate prediction in favor of the outcome that received the most votes; in situations where each outcome received the same number of votes, the predicted

outcome is selected at random.

An advantage of *majority vote* is its simplicity; however, because each prediction is given an equal weight, the aggregate prediction may be influenced heavily by incorrect individual predictions. The second approach, *simple weighted vote*, [38] attempts to place most emphasis on the individual predictions it believes to be most informative. The implementation used in this study assigns a weight to each individual prediction based on the AUC attained via nested cross validation for the relevant combination of data category, feature-selection algorithm, and classification algorithm. *Squared weighted vote* squares the weights used in the simple weighted vote in an attempt to place exponentially higher emphasis on predictions that perform best in nested cross validation. Two additional novel approachs—*LTS predictive-value weighted vote* and *STS predictive-value weighted vote*—assign weights based on the percentage of times (learned via nested cross validation) that patients were predicted correctly as being LTS or STS, respectively. These approaches were motivated by the need to improve predictive performance in data sets that have an unbalanced class distribution and thus may favor predictions for either class. For example, *STS predictive-value weighted-vote* places the most emphasis on predictions that it believes are effective for identifying STS correctly and thus may perform relatively well on data sets containing a small proportion of STS.

For each patient, the *Select Best* approach makes an aggregate prediction based on the individual prediction that received the highest weight (AUC).

Two additional ensemble methods use the posterior probabilities assigned by individual classification algorithms. The *mean probability* method averages the probabilities for each outcome across all predictions—the outcome with the highest mean probability is selected. *Weighted mean probability* assigns a weight, based on the inner-fold AUC, to each probability, and then calculates the mean for each outcome. Finally, *Stacked Generalization* uses the posterior probabilities from the original predictions and trains a second-level classification algorithm to make aggregate predictions based on those values. [59] In stacked generalization, any classification algorithm can be used for the second-level predictions; however, this study used C5.0 Decision Trees because it is designed to handle dependencies between attributes (in this case,

predictions)—a condition that would be expected to occur when first-level algorithms make similar predictions—and because its models are easily interpretable.

3.6 Outcome Discretization

In accordance with previous studies of GBM survival, [1, 30, 33] each patient was designated as either a longer-term survivor (LTS) or shorter-term survivor (STS). In the literature, several survival thresholds have been used to distinguish LTS from STS; one of the most common thresholds is two-year survival [1,33]; however, in other studies, different thresholds have been used. [12, 60] In this study, various survival thresholds were used in an exploration of the effects the threshold has on performance. Two-year survival was used in the initial experiments: patients surviving longer than two years were labeled as LTS, while patients surviving shorter were labeled STS. In a subsequent experiment, an empirical method was used to estimate which split point would result in the best classification performance. [61] This method evaluates many split points via an optimization procedure, attempting to split the patient population into two groups that are subject to "different underlying disease mechanisms" (p. 95). [61] In this approach, 1) a series of candidate split points are determined: patients are sorted by their respective survival times, and the median survival separating each set of adjacent patients constitutes a candidate split point; 2) for each split point, patients are designated as either STS or LTS, depending on their actual survival time; 3) cross validation is used to calculate an error rate at each split point; and 4) the split point that results in the lowest error rate (compared to the error rate that would have been achieved if the majority class had been predicted by default) is selected.

In implementing the split-point selection procedure, several modifications differed from the published approach. The original authors used the C4.5 Decision Trees algorithm for the cross-validation step; however, Quinlan has released an updated version of the algorithm called C5.0 Decision Trees, which was used in this analysis. Additionally, the published method selects the split point that performs best on the full data set; however, this approach may result in models more likely to overfit the data and thus that fail to generalize to external data sets. Consequently, in this study, a split point was selected separately, via nested cross validation, for each outer

cross-validation fold. Finally, because of the large number of variables in the TCGA data set and because of the higher computational burden of performing nested cross validation, 10-fold cross validation was used rather than leave-one-out cross validation (as used by the original authors).

3.7 Performance Metrics

Having performed cross validation, a survival prediction existed for each patient. The quality of the predictions was measured using various metrics: error rate, AUC, and the log-rank survival statistic.

The error rate is calculated as the percentage of test instances misclassified across all folds. (Of related interest is the proportion of patients of either class that were predicted accurately.)

The AUC considers the posterior probabilites produced by the classification algorithms; these probabilities represent the confidence with which the predictions were made. In calculating the AUC, various confidence thresholds are used, and the true positive rate is measured against the false positive rate across all thresholds.

In many machine-learning studies, the error rate and AUC are the performance metrics of primary interest; however, in this study, the outcome variable (survival) is naturally continuous. Thus, the performance was also measured using the log-rank statistic. [39] The overall survival times of patients predicted as LTS were compared against the overall survival times of patients predicted as STS. Subsequently, Kaplan-Meier curves [40] were used to create a visual representation of the overall survival differences between the two groups. The R project [62] and its survival package [63] were used for calculating the log-rank statistic and producing the graphs.

The performance of the ensemble approaches was measured using the same metrics by which individual algorithms were measured.

3.8 Custom Software Requirements

The complex nature of the analyses that were performed in this study necessitated a custom software implementation to orchestrate the various analysis steps. The name of this software package is *ML-Flex* because it is designed to perform machine-learning analyses in a flexible and extensible way. ML-Flex is a command-line tool,

written in the Java programming language. Although ML-Flex was developed for the analyses performed in this study, its extensible nature also suits it well for more general application. The following paragraphs explain the requirements that were met by ML-Flex, along with brief explanations of the software's architecture.

A vast array of multivariate algorithms have been published; however, the algorithms are implemented in a variety of programming languages, have inconsistent interfaces, and use a variety of file formats. Additionally, due to the heterogeneous nature of the methods used to acquire and/or preprocess each category of TCGA data, a variety of text-based data formats have been used to represent the data. Thus, ML-Flex's first requirement was to parse arbitrary data formats and store them in local files that use a common data structure. Having collected and consolidated input data, a second requirement was to transform data values based on criteria specific to that type of data—for example, the DNA methylation data in TCGA are recorded at the probe level, but a gene-centric value needed to be computed across all probes for a given gene.

Having processed and stored the transformed data, the next requirement was to apply variable-selection and classification algorithms to the data. Because the TCGA data sets are large, and nested cross validation—a computationally intensive procedure—needed to be performed multiple times, the software also needed to execute in parallel across multiple computing nodes and multiple processing cores within each node, thus decreasing the amount of time required to complete the analyses.

Prior to performing the analyses, third-party software libraries with implementations of the algorithms were identified; however, no single library implemented all the algorithms. Furthermore, Weka is written in the Java programming language, while C5.0 Decision Trees is written in C. Consequently, an additional requirement was that ML-Flex interface extensibly with third-party software written in many programming languages. Accordingly, ML-Flex needed to contain logic for formatting the data in whatever structure was required by a given third-party tool. For third-party software packages that are not accessible via a Java application programming interface, ML-Flex interfaces with them in the following way: 1) a formatted text file is saved to the local file system, 2) the third-party software is invoked via command-line calls

using the `Runtime` class in Java, and 3) after the software returns an exit code, the software's output files are parsed, and the results are recorded. Thus each software package can operate natively, without requiring any language-specific wrappers.

Another important requirement was that ML-Flex output the analysis results to structured text files that could then be imported into statistical packages and graphics libraries to assist in reporting the results.

3.9 Custom Software Features

Using Java's object-oriented nature, several abstract classes were developed within ML-Flex. The abstract classes contain generic functionality for the main programming logic and specify default parameters. For example, to extract and transform the raw data, an abstract class called `AbstractDataProcessor` orchestrates the process of parsing input files, transforming data values, and storing transformed data. Concrete classes contain the specific logic and parameters required to extract and transform each category of raw data. This extensible design makes it possible to add new categories of data to analyses with minimal coding effort.

No existing software package contained implementations of all ensemble-learning approaches that were performed in this study, so this functionality was added to ML-Flex. ML-Flex uses an abstract class to retrieve the outputs of individual algorithms and assign weights to individual predictions. It then uses concrete classes to implement the specific logic for the various ensemble approaches.

Although implementations of nested cross validation exist in third-party software packages, a custom implementation of this logic was incorporated into ML-Flex because it was central to the TCGA analyses and because the custom implementation facilitated executing tasks in parallel.

To process the data in parallel, a simple, coarse-grained architecture was implemented. This approach is designed to operate in a cluster-computing environment with minimal duplication of computations, without deadlocks, and with no need to communicate directly across processing nodes or threads. For example, for variable selection, ML-Flex instantiates a list of Java objects indicating each unique combination of data category, algorithm, and cross-validation fold; then using the `java.util.concurrent` framework, multiple, independent threads access the file

system to check whether a given combination has been processed and the results have been stored; if the combination has not been processed, the thread checks the file system for an empty, correspondingly named *lock file* indicating that the combination is currently being processed by another thread; if no lock file exists, the thread attempts to use an atomic transaction to create the lock file; having successfully created the lock file, the thread performs variable selection, the results are stored on the file system, and the lock file is deleted. Similar logic is used to perform classification tasks.

Even though ML-Flex is designed to process data quickly via parallelism, the design of ML-Flex sacrifices some performance in favor of scalability and flexibility. Although computing nodes operate independently and thus do not communicate directly to optimize performance, any number of computing nodes may be used, and individual computing nodes may run any operating system that supports the various software components. (However, it is essential the clustering environment employ a shared file system that allows atomic file creation.) Because text files track processing status, ML-Flex is restartable, a desirable feature in cluster-computing environments where server reliability may be limited. Before a job is restarted, lock files simply must be deleted to clear the processing queue. Additionally, ML-Flex's design makes it possible to assign additional computing nodes to a job even after the job is already running.

3.10 Custom Software Validation

Any custom software implementation requires a concerted effort to verify that the software functions properly. ML-Flex was evaluated on various data sets for which the expected outcome (or at least an approximation thereof) was known in advance.

For the first validation, the actual survival values of the TCGA patients were used. As a preliminary step, patients who survived longer than the median were designated as LTS, and the remaining patients as STS. Then for each patient, 900 continuous values were generated at random using the standard normal distribution and assigned to the patient. Subsequently, 100 binary values were generated for each patient, and the symbols were shuffled at random. The resulting data set contained 1000 variables,

each of which should be expected to have no ability to discriminate between LTS and STS. The purpose of creating and evaluating this data set was to ensure that the software contained no obvious problem that would give a positive result when none was expected.

The second validation data set was generated using an approach similar to what was used for the first set. However, in this case, three of the continuous random values were increased by 50 points for each STS. Additionally, two of the binary values were generated such that STS were always assigned a value of zero, while LTS were always assigned a value of one. Although this data set mostly contained randomly generated noise, it should be expected that 1) the variable-selection algorithms would place the highly discriminatory variables at the top of their ranked lists, and 2) the classification algorithms would be able to discriminate perfectly (or at least nearly so) between LTS and STS. The purpose of evaluating this data set was to ensure that when an obvious signal did exist, it could be found.

The final validation data sets were downloaded from the UCI Machine Learning Repository, [64] which contains various "real-world" data sets that have been evaluated in many machine-learning studies over the years. Although few of the data sets are cancer related, the purpose of this validation step was to ensure that the classification results obtained on these data sets were similar to results that have been achieved in other studies; achieving similar results would also suggest that the software developed for this study could perform effectively on data sets with varying levels of expected discrimination. The UCI data sets used for validation were Iris, Breast Cancer (Wisconsin), [65] Hepatitis, Horse Colic, Ionosphere, Pima Indians Diabetes, Statlog (Australian Credit Approval), Statlog (Heart), Tic Tac Toe Endgame, Connectionist Bench (Sonar), and Congressional Voting Records. With the exception of Iris, each of these data sets contain two classes; for consistency, Iris was separated into two separate data sets—one containing *setosa* and *versicolor* data (which are linearly separable), and the other containing *versicolor* and *virginica* data (not linearly separable).

3.11 Bias Correction Procedure for Gene Set Enrichment Analysis

The GOstats [66] package was used to perform GSEA. GOstats uses the hypergeometric distribution to correct for functional categories that have a relatively large (or small) number of genes associated with them and thus would be more (or less) likely to reveal associations with the selected genes. Using this package, an association was tested between genes in top-performing models and biological pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) homo sapiens database. [67]

To correct for potential gene-selection bias, the following method was used. For a given data category, a subset of genes of the same size as the top-performing model was randomly selected from the full set of genes profiled. GSEA was performed for the random gene set, and the resulting p-values were recorded for each pathway. When no result was returned for a pathway, the p-value was 1.0 by default. The same process was repeated 1000 times. For each pathway, an empirical p-value was then calculated by comparing the actual p-value with the distribution of p-values for randomly selected gene sets. Empirical p-values below 0.05 were considered to be statistically significant.

Table 3.1. Summary of TCGA patients and variables included in the analyses for each data category after filtering steps were performed.

Data Category	# Patients	# Variables	Proportion Missing Data
Clinical	313	3	0.081
Treatments	313	6	0.004
Histology	313	6	0.006
DNA Methylation	188	2189	0.020
Somatic Mutations	112	154	0.000
DNA Copy Number	305	320	0.000
mRNA Expression	279	12042	0.000
miRNA Expression	276	534	0.000
All Data	313	15254	0.156

Table 3.2. Variables that have been associated with GBM prognosis in the literature and that were used in Experiment 2.

Data Category	Variables
Clinical	Age, KPS
Treatments	Radiation, temozolomide
Histology	Percent necrosis
DNA Methylation	MGMT
Somatic Mutations	IDH1, TP53
DNA Copy Number	7p, 9p, 10q23, 12q, 19p
mRNA Expression	PDPN, AQP1, CHI3L1, RTN1, EMP3 GPNMB, IGFBP2, OLIG2, LGALS3
miRNA Expression	hsa-miR-196a, hsa-miR-196b

CHAPTER 4

RESULTS

4.1 Validation Experiment: Simulated Data With and Without Structure

Table 4.1 lists results of the validation experiment that was performed using random, simulated data for which no significant separation between LTS and STS was expected. Indeed, the algorithms differentiated poorly between LTS and STS. In fact, in most cases, the error rate and AUC values were slightly worse than would be expected by chance (0.498 and 0.500, respectively). None of the log-rank p-values were below 0.05. Figure 4.1 displays an ROC curve for the result obtained with the C5.0 Decision Trees classification algorithm (with no variable selection). As expected, the curve lies close to the x = y line, which represents the result that would be expected by random chance. Figure 4.2 displays the associated Kaplan-Meier curves, which overlap substantially (as expected). Table 4.2 lists results for the ensemble-learning approaches. Again as expected, the results are similar to what would be expected by chance.

Table 4.3 lists results of the validation experiment in which a subset of variables were simulated to separate LTS from STS, whereas the remaining variables were random noise. As expected, the algorithms separated the two classes perfectly, resulting in error rates of 0.0, AUC values of 1.0, and extremely low log-rank p-values; these results approach the upper limits of performance that could be expected from the TCGA GBM data. Figure 4.3 and Figure 4.4 display an ROC curve and Kaplan-Meier curves, respectively, for the results obtained using the C5.0 Decision Trees classification algorithm (and no variable selection). Table 4.4 lists results for the ensemble-learning approaches; these, too, performed extremely well, as expected.

4.2 Validation Experiments: UCI Machine Learning Repository Data

Table 4.5 list results of validation experiments that used data from the UCI Machine Learning Repository. No single algorithm always performed better than the rest. And in many cases, predictive performance varied substantially from one algorithm to another on a given data set. The SVM algorithm performed considerably worse than C5.0 Decision Trees and NBC on the Statlog data sets but considerably better on the Tic Tac Toe data. Such variability illustrates the concept of classifier diversity—some algorithms are better suited than others to particular data sets. Accordingly, the ensemble-learning approaches often performed better than individual algorithms (see Table 4.6), though the results were mixed overall. Notably, the performance of the ensemble-based approaches was highly consistent across the various approaches.

As a means of comparison, these tables also list results that were reported for the UCI data sets on TunedIT.org, a web site that allows researchers to report classification results for a wide variety of algorithms. In these tables, the TunedIT values represent the mean error rate across all Weka classifiers that fall under the *bayes*, *functions*, and *trees* classifier categories. For every data set except Ionosphere, the TunedIT results fall within the range of results that were obtained using ML-Flex. It is reasonable that the TunedIT results would vary moderately from the ML-Flex results because the TunedIT results were obtained using a variety of algorithms and configuration settings.

4.3 TCGA Experiment 1: Full Data Set, Two-Year Survival

In an initial exploration of prognostic models that could be derived with the algorithmic variable-selection approaches, an initial experiment was performed using the full TCGA data set and two-year survival as the survival split point. Results obtained using SVM-RFE variable selection are listed in Table 4.7. (Results for the remaining variable-selection approaches are listed in Tables 4.8 and 4.9.)

A relatively high AUC value (0.683) and a significant log-rank p-value (0.0159) were observed for the NBC models based on age, KPS, and gender data. However, when variable selection was performed—thus limiting the models to a single clinical

variable—the models became less stable and performed relatively poorly. This result suggests that age, KPS, and gender each contributed some information that helped differentiate between patients who survive longer or shorter than two years.

The second-lowest p-value (0.00104) and highest AUC value (0.705) were attained by NBC models trained on treatment variables. This result suggests that patients surviving longer than two years received different overall treatments from patients surviving less than two years. A look at the underlying data reveals a clear trend between patient survival and the overall number of treatments received (see Figure 4.5). Two possible explanations for this trend are that 1) the more treatments a patient receives, the better her survival expectation, or 2) the longer a patient survives, the more treatments she is likely to receive. The latter would likely be a confounding effect.

When SVM-RFE variable selection and the SVM classification algorithm were applied, DNA methylation models attained significance according to the log-rank statistic (see Figure 4.6; $p = 0.427$). Interestingly, the models performed best when the top-1000 ranked genes were included, and predictive performance tended to improve as the number of genes increased (see Figure 4.7). Many studies have demonstrated an association between gene-specific methylation and outcome—including a relationship between MGMT methylation and GBM survival [2]—but *global* methylation patterns have also been suggested to influence tumor initiation and progression. For example, recent research suggests that global hypomethylation contributes to oncogene activation, loss of imprinting, and decrease in genomic instability. [68] On the other side of the spectrum, global hypermethylation can silence transcription of many genes—including tumor suppressors—and is recognized as a common molecular abnormality across various cancers. [69] In fact, some cancer studies have shown a prognostic relationship between methylation patterns across many genes even when no relationship existed for the individual genes. [69] In the TCGA methylation data, 58.4% of profiled genes were more highly methylated in STS than in LTS (based on mean difference, see Figure 4.8), a trend that suggests a slight bias toward hypermethylation in the most aggressive tumors.

The best-performing somatic-mutation models contained 50 genes, but it appears

the presence of such a large number of genes resulted in models being overfit to the training data, which resulted in poor generalization. The top-ranked genes (e.g., PRAME, FRAP1, GRM1, PTCH1, EPHA3) were mutated differentially between LTS and STS; however, because many of the genes were mutated infrequently (often only two to three times across the population), the models failed to generalize. It may be that limiting the analysis to genes that are mutated more frequently would stabilize variable selection and result in better-performing models.

When all data were combined into an aggregate data set, C5.0 Decision Trees models performed well according to the log-rank statistic ($p = 0.00474$) and AUC 0.452). Perhaps most notable about this result is that none of the C5.0 models attained significance for individual data categories; however, when all data were combined, the algorithm was much more successful at separating LTS from STS. The interpretable nature of C5.0 Decision Trees makes it feasible to investigate hypotheses about factors that may interact across data categories to affect survival. For example, when the C5.0 rule learner was applied to the full data set, it suggested that patients would be STS if they received lomustine treatment and had relatively low methylation for either CD86 or IRAK3; conversely, it suggested that patients would be LTS if they received lomustine treatment but had relatively high methylation for both CD86 and IRAK3. These rules held true for approximately 26/32 (82%) of patients who received lomustine. Such rules could guide researchers in developing treatments tailored to a patient's tumor–molecular profile; however, the true clinical and biological relevance of such rules can only be speculated until further validation.

Although the NBC algorithm appears to have performed best overall, classifier performance varied substantially across algorithms in this experiment. This higher classifier diversity may be one reason the ensemble approaches resulted in AUC values that were consistently higher than most individual classifiers (see Table 4.10). However, the best performing ensemble approach was *Select Best* (AUC = 0.676, $p = 0.00762$). By nature, Select Best is influenced strongly by the best individual performers, which in this case were models based on clinical and treatments data.

In many cases, the ensemble approaches predicted all patients as LTS, likely an indication that the ensemble approaches were influenced heavily by class imbalance

between LTS and STS. However, as expected, the *LTS predictive-value weighted vote* method helped counter some of the effects of class imbalance in comparison to the *majority vote* and *simple weighted vote* methods.

One result from this experiment was counterintuitive and helps illustrate that care must be taken to interpret the results. In this experiment, the lowest log-rank p-value was attained when the SVM-RFE and NBC algorithms were applied to the clinical data. Interestingly, Kaplan-Meier curves for the predictions reveal that the subset of patients predicted as LTS survived a significantly *shorter* time compared to the remaining patients (see Figure 4.9; p = 4.46e-05), an unexpected trend. One limitation of the log-rank statistic is that significance can be attained even when predictions are inaccurate; consequently, when evaluating survival-status predictions it is essential also to examine Kaplan-Meier curves visually. A second important observation is that the error rate and AUC sometimes conflict with each other. In this case, the error rate was 0.220, substantially worse than the baseline expected by chance (0.169). Conversely, the AUC was 0.590, substantially better than expected by chance (0.500). The reason for this discrepancy is that NBC was unable to decipher whether 23 of the patients were LTS or STS and thus predicted the two classes with equal probabilities. Then by default, NBC designated these patients as LTS—an apparently arbitrary decision. Of the 23 patients, 22 were actually STS. This explains not only why the error rate was low but also why the LTS-predicted patients had low survival as a group. However, because the AUC is not sensitive to arbitrary thresholds, it indicated reasonably overall good performance, as illustrated by the ROC curve for this result (see Figure 4.10).

4.4 TCGA Experiment 2: Prior Knowledge Variables, Two-Year Survival

The second experiment on TCGA data was performed using the prior-knowledge variables and two-year survival as a split point. Table 4.11 lists results for the individual classification algorithms.

Two results attained statistical significance via the log-rank test: NBC models trained on TP53 and IDH1 somatic mutations (p=0.00479, see Figure 4.11) and NBC models trained on mRNA expression levels (p=0.00599, see Figure 4.12). No

other result was statistically significant.

Interestingly, even though the NBC somatic-mutation models attained significance via the log-rank test, the AUC value (0.520) was only slightly higher than what would be expected by random chance (0.500). Additionally, the error rate (0.196) was slightly worse than would be expected by chance (0.196). Taken together, these results demonstrate again a key observation that arose in this study: no single metric is suitable in isolation to quantify classification performance. In this case, the log-rank statistic highlights a small subset of patients who were predicted as LTS. Of those patients, only three survived longer than two years. However, as a group, the patients predicted as LTS had significantly longer survival than patients predicted as STS. In fact, two of the LTS-predicted patients survived 603 and 691 days—values that were close to the survival threshold. Thus while these patients were misclassified—thus impacting the error rate and AUC—the NBC algorithm still identified a patient subset with potential clinical relevance. Indeed, recent research has suggested that IDH1 mutations offer a survival advantage, [17, 27] which may partially explain this result.

The fact that mRNA-expression models performed well was unsurprising, given that Colman, et al. had derived the gene set from multiple, independent data sources and had used two-year survival as their threshold. [1] That their gene set also generalized to TCGA at a significant level lends further credence to the clinical relevance of that gene set. However, it was somewhat surprising that no model attained significance for any other data category. Even clinical models, which accounted for age and KPS—two well-known GBM prognostic variables—failed to reach significance. Again in this case, it is important to acknowledge that no single performance metric is adequate. Even though the log-rank statistic was not significant for the NBC models, the AUC value was relatively high 0.676 (see ROC curve in Figure 4.13). A look at the Kaplan-Meier survival curves (see Figure 4.14) reveals that a subgroup of LTS were identified accurately; however, the log-rank statistic was heavily influenced by the fact that the two curves intersect at one point. Thus in this case, the AUC was a more telling metric of performance than the log-rank statistic. Because the AUC is derived from posterior probabilities generated by the classification algorithms, the ordering of the probabilities influences the AUC value; thus if, for example, the actual

STS are predicted as STS with greater overall probability than the actual LTS, the AUC can be informative even when the log-rank statistic and error rate are not.

Across all performance metrics in this experiment, NBC performed better than C5.0 Decision Trees and SVM. In many cases, C5.0 Decision Trees and SVM predicted all patients as LTS. NBC may have been less impacted by the class imbalance between LTS and STS, a factor that can have a strong impact on the performance of many classifiers, include C5.0 Decision Trees and SVM. [70]

When combining predictions across all data categories and algorithms, no ensemble-learning approach attained statistical significance (see Table 4.12).

Taken together, these results suggest that multivariate prognosis models trained on prior-knowledge variables have some ability to identify patients surviving longer than two years. In particular, performance for somatic-mutation and mRNA expression models was better than what was obtained using algorithmic variable selection. In TCGA Experiment 1, the SVM-RFE algorithm ranked IDH1 and TP53 mutations 10th and 37th best across all cross-validation folds. In the same experiment, SVM-RFE ranked none of the Colman, et al. mRNA expression genes highly; in fact, none of the nine mRNA expression genes were ranked in the top 1000 across all folds. These results demonstrate that the algorithmic variable-selection approaches do not always coincide with prior knowledge, particularly in high-dimensional data sets where a considerable amount of noise is likely.

4.5 TCGA Experiment 3: Full Data Set, Empirical Survival Discretization

The experiments so far have used two-year survival, an arbitrarily decided threshold, to distinguish LTS from STS. This section describes the results of an experiment in which the survival threshold was determined empirically (for each cross-validation fold).

In preliminary analyses, the empirical split-point method appeared to favor median survival. A closer look revealed two possible explanations for this apparent bias: 1) because C5.0 Decision Trees—the classification algorithm used by this method—performs best when the class distribution is balanced, [70] it should naturally bias toward the median, and 2) the existing method for correcting the error rate against

chance expectation favors thresholds near the median. The following example illustrates the latter point. If the candidate threshold were 100 days, approximately 90% of patients would be LTS; thus if all patients were predicted by default to be LTS, the error rate would be 0.100. Then if the algorithm distinguished perfectly between LTS and STS at this threshold, the error rate would be 0.000, and the improvement over chance expectations would be 0.100. However, if the candidate threshold were 360 days, approximately 50% of patients would be LTS, and the default error rate would be 0.500. Then if the algorithm could distinguish only moderately between LTS and STS—for example, achieving an error rate of 0.390—the improvement over chance would be 0.110. Thus even though the algorithm classified perfectly at 100 days, the 360-day threshold would be selected. In fact, 100 days would be selected only if the improvement over chance exceeded 0.100 for no other threshold.

In evaluating ways to address this limitation, two other metrics were considered: the AUC and log-rank statistic. To compare the effectiveness of these metrics, the following simulations were performed. Using the actual GBM survival values, a continuous independent variable was generated for each patient such that patients below a given, artificially defined threshold had much lower values than patients above that threshold. Thus it would be expected that the algorithm could differentiate perfectly between LTS and STS at these thresholds. Figures 4.15, 4.16, and 4.17 show the results of a simulation where the artificial threshold was 360 days (approximately median survival in this data set). When the error rate (corrected for what would be observed if the majority class were predicted by default) or AUC were used, the simulated threshold was identified correctly and precisely. When the log-rank statistic was used, several "optimal" thresholds were identified in a tie, and the median of these values was near the correct threshold. These findings suggest that if the true threshold for distinguishing LTS from STS is near the median, any of the three metrics could be used. Figures 4.18, 4.19, and 4.20 show the results of a simulation where the artificial threshold was 100 days. When the error rate was used, a large number of "optimal" thresholds (including 100 days) was identified; among these, the median was 235.5 days. A similar result was observed for the log-rank statistic. However, when AUC was used, the threshold was identified precisely at 99.5 days—no ties occurred. Taken

together, these results suggest that the AUC is better than the error rate or log-rank statistic at identifying survival split points. The likely reason for AUC's advantage is that its calculation is insensitive to class imbalances. [71]

When the AUC was used as the survival-threshold metric on the TCGA data, the selected thresholds varied between 69.5 days and 147 days across the cross-validation folds. Thus, in contrast to the previous experiments, in which most patients were labeled STS, the great majority of patients were labeled LTS in this experiment.

Because the selected survival thresholds varied across cross-validation folds (see Figure 4.21), the definition of LTS and STS was different for each fold. For example, in fold 1, the selected threshold was 90.0 days, while in fold 4, the selected threshold was 124.5 days. Because of this variability, it would not have been reasonable to assess overall performance as if the threshold were fixed across all folds (like in previous experiments). One alternative for addressing this problem is to calculate the performance metrics separately for each fold, using only the test instances from that fold. Having obtained a metric for each test set, the overall performance then would be calculated as the mean performance across the folds. Especially on small data sets, this approach is conservative for the log-rank statistic because only 1/k (where k is the number of folds) instances are available for each calculation, so the upper performance bound is reduced. Additionally, whenever the log-rank statistic cannot be computed for a given fold (when all predictions are of the same class), its value must default to zero (equivalent to a p-value of 1.0); consequently, the overall performance estimate may become even more conservative. Tables 4.13, 4.14, and 4.15 list results for this experiment using the mean-across-folds method of measuring performance. Several models performed quite well, despite the conservativeness of this evaluation approach.

In examining the C5.0 models that were used to determine the thresholds, it became apparent that radiation treatment was an extremely strong predictor of survival status at the selected thresholds. In cross validation, predictions based on treatment data (which include radiation) performed exceptionally well by all measures, across all variable-selection approaches and classification algorithms. Figure 4.22 illustrates that patients who received no radiation treatment survived drastically

less time than patients who received radiation treatment. Identifying a relationship between radiation treatment and GBM survival is not novel—multiple studies have previously observed this relationship, even going back many decades. [6,14,16] In fact, radiation treatment is part of the standard of care for GBM. [72] However, the fact that radiation treatment separated STS from LTS so strikingly in this experiment—despite the presence of thousands of other potential predictor variables—suggests that patients who receive no radiation treatment can typically expect a very short survival, regardless of other clinical or biomolecular factors. However, at least for the TCGA GBM patients, such a strong association between radiation-treatment status and survival time likely represents a confounding effect. Some patients may feel they are too old or frail to receive radiation treatment; for these patients, failure to receive radiation may simply be a surrogate indicator of a patient's age or overall health. Indeed, in the TCGA data, patients who chose not to receive radiation treatment were considerably older (see Figures 4.23) and had lower overall KPS (see Figure 4.24) than patients who received radiation. Additionally, only a slight trend exists between patient survival and the number of days before the treatment started for patients who received radiation treatment (see Figure 4.25); this suggests that non-radiation-treated patients did not simply delay treatment and then decease prematurely as a consequence. Taken together, these observations strongly suggest that including radiation treatment in multivariate models introduces a confounding effect.

Aside from radiation treatment, drug treatments also appeared to have an effect on patient survival. However, a closer look at the data revealed a potential confounding effect. Patients who received radiation treatment also received a greater number of overall treatments than patients who received no treatment (see Figure 4.26). Although this finding is expected—radiation-treated patients survive longer and thus may have opportunities to receive more drug treatments—part of the success in using treatments to predict survival may stem from this bias.

The SVM algorithm performed relatively well on the mRNA expression data when RELIEF-F was used for variable selection, attaining a mean AUC of 0.640. Because radiation treatment was such a strong predictor of survival in this experiment, the

possibility existed that the expression of some mRNA genes was a surrogate marker for radiation treatment status. To test this hypothesis, correlation was measured between the top-ranked mRNA expression genes and radiation treatment using Spearman's rank-based rho statistic (excluding missing values). Of the top-500 mRNA genes, 18.0% were correlated with radiation treatment status, while only 8.9% of all mRNA genes were correlated with radiation treatment status. Accordingly, expression of these genes may influence tumor aggressiveness to the point that patient well-being is affected and patients are less likely to receive radiation treatments.

Table 4.16 lists results of the ensemble-learning approaches for this experiment. All approaches had AUC values that were 0.780 or higher, and most approaches attained significant log-rank p-values. As expected, the *Select Best* predictions were influenced heavily by the individual predictions based on treatment data (including radiation treatment). However, *Stacked Generalization* also performed quite well, even though its second-level predictions accounted not only for treatment-related predictions but also for predictions based on various categories of molecular data. And true to its design, the *STS-predictive value weighted vote* method helped counteract the effects of class imbalance and performed better than all other voting methods.

Overall, the results from this experiment suggest that a single independent variable with extremely high prognostic relevance (radiation treatment) can dominate the threshold-selection method. In this experiment, clinical, treatments, and mRNA expression data each contained a strong signal that could be discerned by the classification algorithms. Indeed, the models performed well despite the likely effects of class imbalance at the selected thresholds. When evaluating the performance of these models, one should also consider that the conservative mean-across-folds method was used to calculate the performance metrics.

4.6 TCGA Experiment 4: Radiation-treated Patients, Median Survival

Because radiation treatment was a strong individual predictor in the previous experiment and because radiation treatment is part of the standard of care for GBM, non-radiation-treated patients were excluded for TCGA Experiment 4. Additionally, to remove any effects of class imbalance, median survival was used as the threshold.

After excluding patients who received no radiation treatment, 261 patients remained, and the median survival was 423 days. Table 4.17 lists the results from applying the RELIEF-F variable-selection approach and the various classification algorithms to this subset of patients (see Tables 4.18 and 4.19 for results for the other variable-selection approaches).

Once again, treatment data (not including radiation) distinguished well between LTS and STS. All three classification algorithms performed well. As in previous experiments, this success appears to be attributed at least partly to the relationship between survival and the total number of treatments received. The models that performed best were based on all five drug-treatment variables; however, interestingly, the top-ranked treatment variable was *other drug treatment*—a variable that indicates whether any of a number of infrequently administered treatments were given. In many cases, these "other" treatments were targeted therapies (e.g., tamoxifen, rapamycin, gefitinib, imatinib) that have been used primarily to treat other cancers. In other cases, the treatments were hormonal (e.g., valproic acid, levetiracetam), chemotherapy (e.g., carmustine, cisplatin), or immunotherapy treatments (e.g., dendritic cell vaccine, erlotinib) that each may have a positive effect on some patients but that individually have not been shown to increase GBM survival in phase III clinical trials.

The clinical data also performed quite well in this experiment. In particular, C5.0 and NBC models based on the top-ranked clinical variable–age in every cross-validation fold—attained statistical significance. Figure 4.27 displays Kaplan-Meier curves for the NBC predictions. This result was unsurprising given that age is a well-known GBM prognostic variable—the TCGA data show that as age increases, GBM survival tends to decrease (see Figure 4.28). However, clinical-model performance was considerably better in this experiment—with median survival as the threshold—than in the other experiments. It may be that age influences a GBM patient's prognosis more strongly when other dominant factors (e.g., no radiation treatment) are not at play.

Models based on DNA methylation also reached statistical significance. In fact, nearly every algorithm performed well. Once again, the performance of the models usually increased as the number of genes in the models increased (see Figure 4.29).

One factor that may be relevant to global DNA methylation is patient age. Across all methylation genes profiled, (12.9%) were significantly correlated with age (Wilcoxon rank-sum test; $p < 0.05$). Contrarily, only 5.1% of genes were correlated when the age values were randomly permuted (repeated 100 times). Even though causality should not be inferred from these calculations, it seems likely that global methylation patterns are affected by age. A recent study of gastrointestinal cancer demonstrated just such an effect. [73] As such, survival predictions based on DNA methylation may be confounded by age effects. To test this supposition further, an additional experiment was performed using only age and DNA methylation variables (and limited to patients with data for both categories). Neither data category outperformed the other in all cases (see Table 4.20). When the two data categories were combined (into a single data set or via ensemble-learning methods), predictive performance often improved slightly (see Table 4.21). These results suggest that age and DNA methylation contain different but complementary prognostic information.

In this experiment, every ensemble method attained statistical significance (see Table 4.22). All but one (Stacked Generalization) had an AUC value that approached the maximum individual result for this experiment. In terms of error rate and AUC, *Simple Weighted Vote* performed (slightly) better than *Majority Vote*. Further, *Squared-weighted Vote* performed better than *Simple Weighted Vote*. Additionally, *Weighted Mean Probability* performed better than *Mean Probability*. Similar trends occurred in the other experiments, a finding that attests to the value of using weight-based measures to improve ensemble results.

4.7 TCGA Experiment 5: Radiation-treated Patients, Median Survival, No Treatment Data

As has been demonstrated previously, a clear relationship exists between the number of treatments that a patient receives and the number of days a GBM patient survives (see Figure 4.5). Additionally, at the time of diagnosis, physicians do not always know which treatments will be administered to a given patient, and physicians often alter treatment regimens based on a patient's response to what has already been administered. Thus the prognostic utility of treatment data is limited, and including

treatment data in prediction models may cause a confounding effect. Indeed, the results for TCGA Experiment 4 were consistently good when treatment data were included in models. In an attempt to avoid such confounding effects, an additional experiment was conducted in which treatment data were excluded. As in TCGA Experiment 4, non-radiation-treated patients were excluded from this experiment, and the median was used as the split point between LTS and STS.

Table 4.23 lists the results from applying the various variable-selection approaches and classification algorithms to an aggregate data set containing data for all categories except treatments. And Table 4.24 lists the results from applying the various ensemble-learning approaches. In some cases, the predictions differentiated between LTS and STS at statistically significant levels; this finding suggests that models based solely on clinical, histology, and biomolecular data—which can each be collected at the time of diagnosis— can be used to inform prognosis predictions. However, in most cases, the performance of the algorithms was considerably worse than in TCGA Experiment 4. For example, the best AUC value and log-rank p-value for this experiment were 0.594 and 7.94e-05, respectively, and in many cases the log-rank p-values were not statistically significant; however, in TCGA Experiment 4, the best AUC and log-rank p-value were 0.706 and 9.96e-09, respectively, and all but two log-rank p-values were statistically significant. Interestingly, the best result for this experiment was obtained using no variable selection and the SVM classification algorithm; when these approaches were applied to the data set containing treatment data (TCGA Experiment 4), the performance was substantially worse (AUC = 0.519, log-rank p-value = 0.032). Taken together, these differences suggest further that treatment data can cause a confounding effect and thus should be excluded or at least treated specially.

In an additional attempt to account for the likely confounding effects of treatments, a follow-on experiment was conducted in which only patients who had been treated with radiation and temozolomide were included. For the subset of patients who met these criteria (n = 134), none of the algorithms differentiated between LTS and STS at a statistically significant level, and the AUC values were near 0.500 (see Tables 4.25 and 4.26). These results suggest the difficulty of using TCGA data

to differentiate between patients who will respond relatively well to a given drug treatment and patients who will not respond well. Even though all patients in this experiment received temozolomide, the treatment regimens were not consistent among all patients, and many of the patients received various other drug treatments. It is possible that more effective prediction models could be derived from clinical trials that administer consistent treatment regimens to all patients. However, clinical trials may lack sufficient funding to acquire multiple categories of high-throughput molecular data for an adequately large patient cohort.

4.8 Gene Set Enrichment Analysis of DNA Methylation Genes

Across all TCGA experiments, the best-performing biomolecular models were derived using hundreds of DNA methylation variables. A search of the literature reveals that some of the top-ranked genes have previously been associated with—or at least have a plausible connection with—tumorigenesis. However, with such a large number of variables, it can be challenging to assess an entire model's biological relevance. GSEA is one approach for providing such interpretation.

The KEGG database contains hundreds of pathway diagrams representing existing knowledge about protein interactions that affect particular biological processes. Also included in KEGG are diagrams that represent the proteins involved in particular cancers. Additionally, KEGG contains aggregate pathway diagrams that combine individual pathways according to some theme. One such aggregate diagram is *pathways in cancer*, which attempts to represent all protein interactions involved in any cancer type that is currently represented in KEGG.

Initially, the standard GSEA method was applied to the top-1000 ranked (via RELIEF-F variable selection) methylation genes from TCGA Experiment 4. Table 4.27 lists the most significantly associated pathways. The top pathway, *pathways in cancer*, was assigned a p-value of 2.87e-15, indicating a high likelihood that the selected genes are known to be involved in human cancers in general. Several other top-ranked pathways represent either individual cancers (e.g., *bladder cancer*, *pancreatic cancer*) or other pathways that could plausibly drive aberrant tumor growth (e.g., TGF-beta signaling pathway, focal adhesion). Such findings were expected for

two reasons: 1) it is reasonable to expect that at least some methylation genes that differentiate LTS from STS would be associated with known cancer-related pathways, and 2) the 2189 profiled methylation genes were manually selected by the TCGA Consortium and thus would likely be biased toward genes from known cancer-related pathways. As a way to validate the latter assumption, standard GSEA was applied to the full set of methylation genes that had been profiled (see Table 4.28). Again in this case, *pathways in cancer* and several known cancer-related pathways were identified with extreme significance. These findings confirm that the selected genes were biased strongly in favor of prior knowledge about cancer. (These findings also provide evidence that GSEA behaves as expected.)

In an attempt to account for selection bias, the GSEA permutation approach (described in Methods) was applied to the top-1000 methylation genes. As expected, *pathways in cancer* and most other pathways that had been significant prior to the correction were no longer considered significant (see Table 4.29). However, several pathways remained significant, including *TGF-beta signaling pathway* and *ErbB signaling pathway*, which each are known to play a role in tumorigenesis. [74,75] Also near the top of the list were *NOD-like receptor signaling pathway* and *Toll-like receptor signaling pathway*, which are involved in generating innate immune responses and may be interconnected with each other. [76] The immune system is essential not only for protecting against foreign pathogens but also for attacking tumor cells. [77] Other significant pathways may not have an intuitive connection with tumorigenesis but do affect the nervous system (e.g., *olfactory transduction*, *amyotrophic lateral sclerosis*, *Huntington's disease*). These pathways may have been selected simply because they share genes with pathways that drive GBM tumorigenesis. It is also possible that some pathways (or their subcomponents) that behave aberrantly in other diseases also are deregulated in some GBM tumors, even though their phenotypic manifestations are different.

It is important to note that pathways that were statistically significant before bias correction were not necessarily spurious findings; in fact, these pathways may have strong relevance to GBM survival. However, the bias-correction technique provides a way to generate hypotheses about pathways that may influence GBM survival but that

may have been overlooked in this context. It should also be noted that KEGG is hand curated and contains only a subset of known biological pathways—any enrichment study is limited by the knowledge source upon which it is based.

Table 4.1. Cross-validation results when 900 randomly simulated continuous variables and 100 randomly simulated binary variables were used. Median survival was the split point between longer-term survivors and shorter-term survivors. The purpose of this experiment was to serve as a negative test, ensuring that when no obvious signal existed in a data set, the performance metrics would indicate such.

Variable Selection Approach	Classification Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
None	C5.0	0.188	0.958	0.094	0.526	0.761
None	NBC	0.185	0.981	0.000	0.524	0.126
None	SVM	0.169	1.000	0.000	0.606	N/A
SVM-RFE	C5.0	0.188	0.958	0.094	0.526	0.761
SVM-RFE	NBC	0.185	0.981	0.000	0.524	0.126
SVM-RFE	SVM	0.169	1.000	0.000	0.579	N/A
RELIEF-F	C5.0	0.188	0.958	0.094	0.526	0.761
RELIEF-F	NBC	0.169	1.000	0.000	0.581	N/A
RELIEF-F	SVM	0.169	1.000	0.000	0.553	N/A

Table 4.2. Cross-validation results when 900 randomly simulated continuous variables and 100 randomly simulated binary variables were used. Ensemble-learning approaches were applied, and median survival was the split point between longer-term survivors (LTS) and shorter-term survivors (STS). The purpose of this experiment was to serve as a negative test, ensuring that when no obvious signal existed in a data set, the performance metrics would indicate such.

Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Majority Vote	0.173	0.996	0.000	0.518	2.65e-18*
Simple Weighted Vote	0.173	0.996	0.000	0.518	2.65e-18*
Squared-Weighted Vote	0.173	0.996	0.000	0.518	2.65e-18*
LTS Predictive Value Weighted Vote	0.192	0.958	0.075	0.525	0.939
STS Predictive Value Weighted Vote	0.173	0.996	0.000	0.518	2.65e-18*
Select Best	0.173	0.996	0.000	0.561	0.772
Mean Probability	0.179	0.985	0.019	0.556	0.435
Weighted Mean Probability	0.176	0.985	0.038	0.552	0.925
Stacked Generalization	0.543	0.465	0.449	0.457	0.206

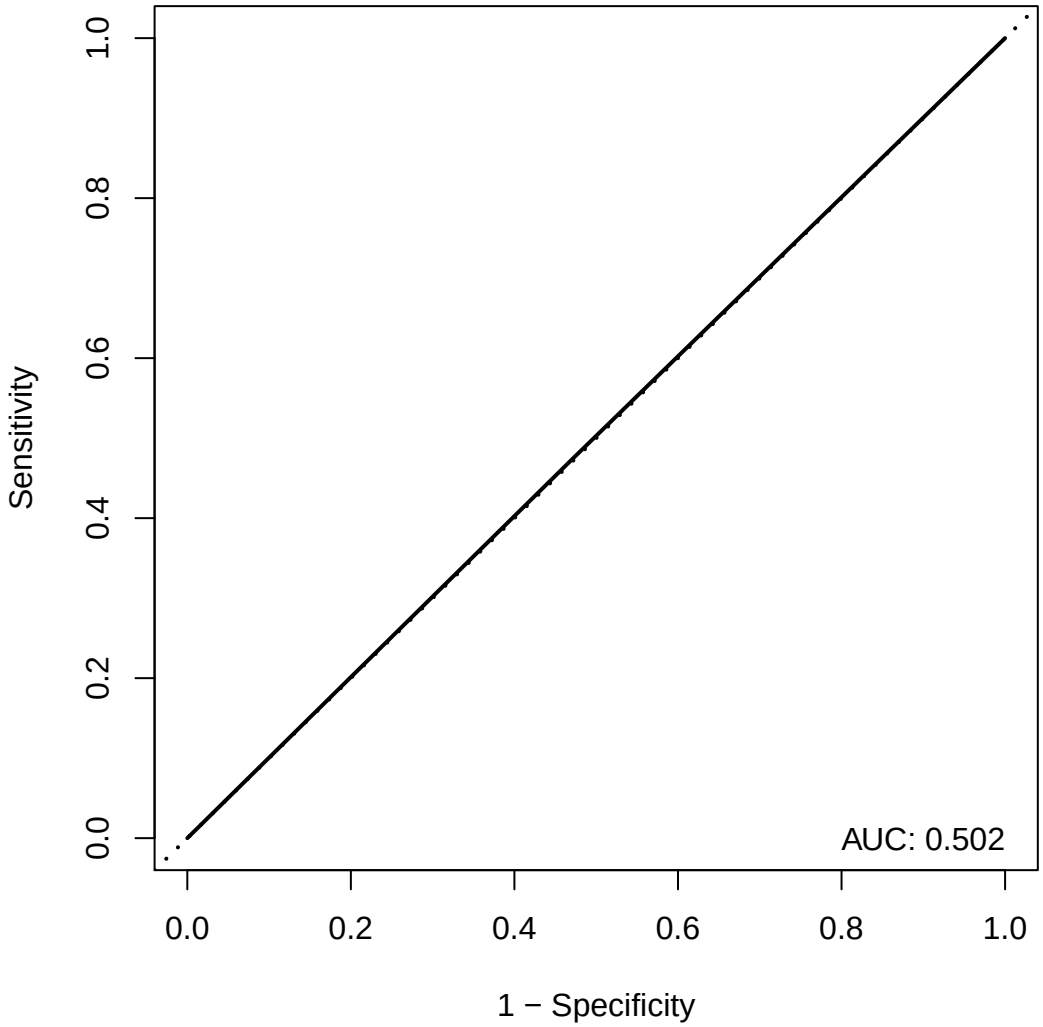

Figure 4.1. Receiver operating characteristic curve for validation experiment in which the C5.0 Decision Trees algorithm attempted to discriminate between longer-term survivors and shorter-term survivors using 900 randomly simulated continuous variables and 100 randomly simulated binary variables. No variable selection was performed.

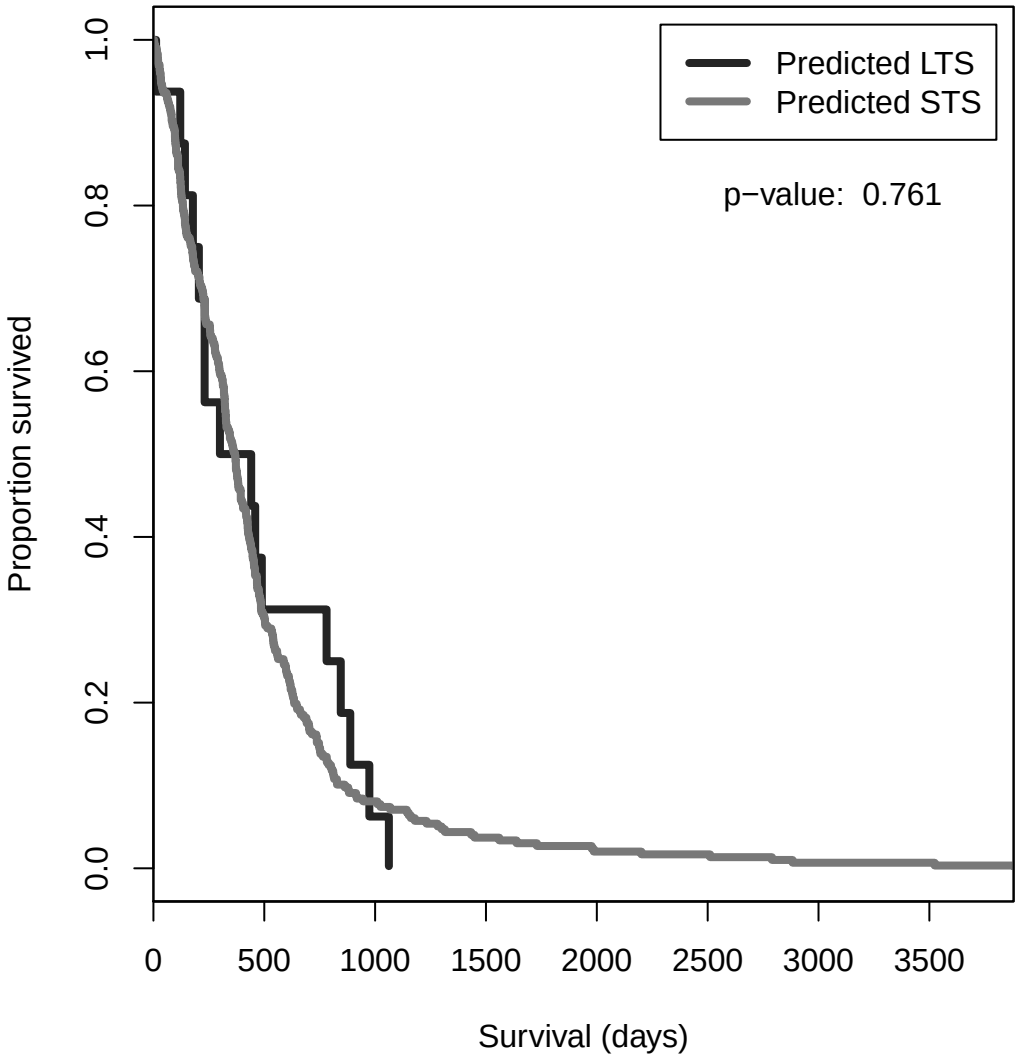

Figure 4.2. Kaplan-Meier curves for validation tests in which the C5.0 Decision Trees algorithm attempted to discriminate between longer-term survivors and shorter-term survivors using 900 randomly simulated continuous variables and 100 randomly simulated binary variables. No variable selection was performed.

Table 4.3. Cross-validation results when 900 randomly generated continuous variables and 100 randomly generated binary variables were used. Three of the continuous variables and two of the binary variables were modified to support perfect discrimination between longer-term survivors (LTS) and shorter-term survivors (STS). Median survival was the split point between LTS and STS. The purpose of this experiment was to serve as a positive test, indicating that when an obvious signal existed in a data set, it could be found.

Variable Selection Approach	Classification Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
None	C5.0	0.000	1.000	1.000	1.000	5.43e-40*
None	NBC	0.000	1.000	1.000	1.000	5.43e-40*
None	SVM	0.000	1.000	1.000	1.000	5.43e-40*
SVM-RFE	C5.0	0.000	1.000	1.000	1.000	5.43e-40*
SVM-RFE	NBC	0.000	1.000	1.000	1.000	5.43e-40*
SVM-RFE	SVM	0.000	1.000	1.000	1.000	5.43e-40*
RELIEF-F	C5.0	0.000	1.000	1.000	1.000	5.43e-40*
RELIEF-F	NBC	0.000	1.000	1.000	1.000	5.43e-40*
RELIEF-F	SVM	0.000	1.000	1.000	1.000	5.43e-40*

Table 4.4. Cross-validation results when 900 randomly generated continuous variables and 100 randomly generated binary variables were used. Three of the continuous variables and two of the binary variables were modified to support perfect discrimination between longer-term survivors (LTS) and shorter-term survivors (STS). In this experiment, ensemble-learning approaches were applied, and median survival was the split point between LTS and STS. The purpose of this experiment was to serve as a positive test, indicating that when an obvious signal existed in a data set, it could be found.

Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Majority Vote	0.000	1.000	1.000	1.000	5.43e-40*
Simple Weighted Vote	0.000	1.000	1.000	1.000	5.43e-40*
Squared-Weighted Vote	0.000	1.000	1.000	1.000	5.43e-40*
LTS Predictive Value Weighted Vote	0.000	1.000	1.000	1.000	5.43e-40*
STS Predictive Value Weighted Vote	0.000	1.000	1.000	1.000	5.43e-40*
Select Best	0.000	1.000	1.000	1.000	5.43e-40*
Mean Probability	0.000	1.000	1.000	1.000	5.43e-40*
Weighted Mean Probability	0.000	1.000	1.000	1.000	5.43e-40*
Stacked Generalization	0.000	1.000	1.000	1.000	5.43e-40*

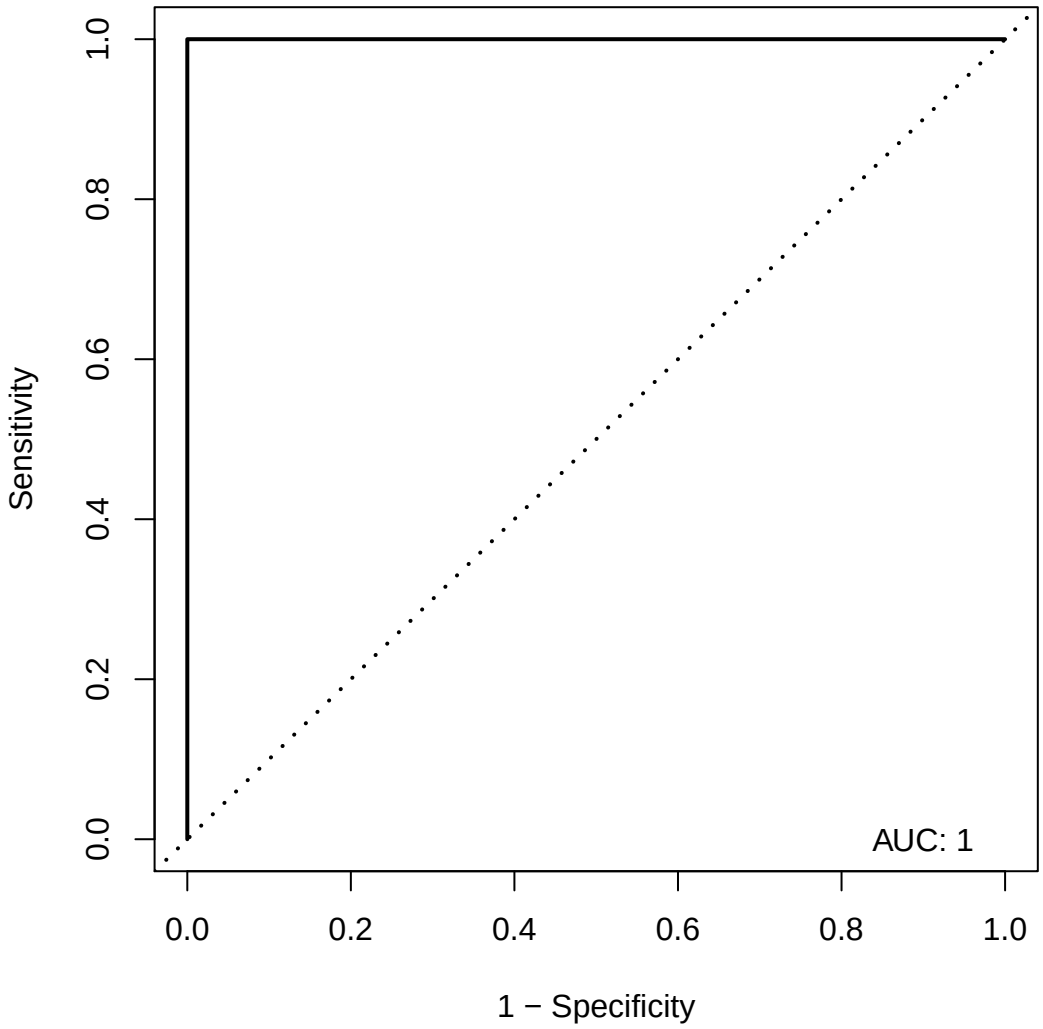

Figure 4.3. Receiver operating characteristic curve for validation tests in which the C5.0 Decision Trees algorithm attempted to discriminate between longer-term survivors (LTS) and shorter-term survivors (STS) using 900 randomly simulated continuous variables and 100 randomly simulated binary variables. Three of the continuous variables and two of the binary variables were modified to support perfect discrimination between LTS and STS. No variable selection was performed.

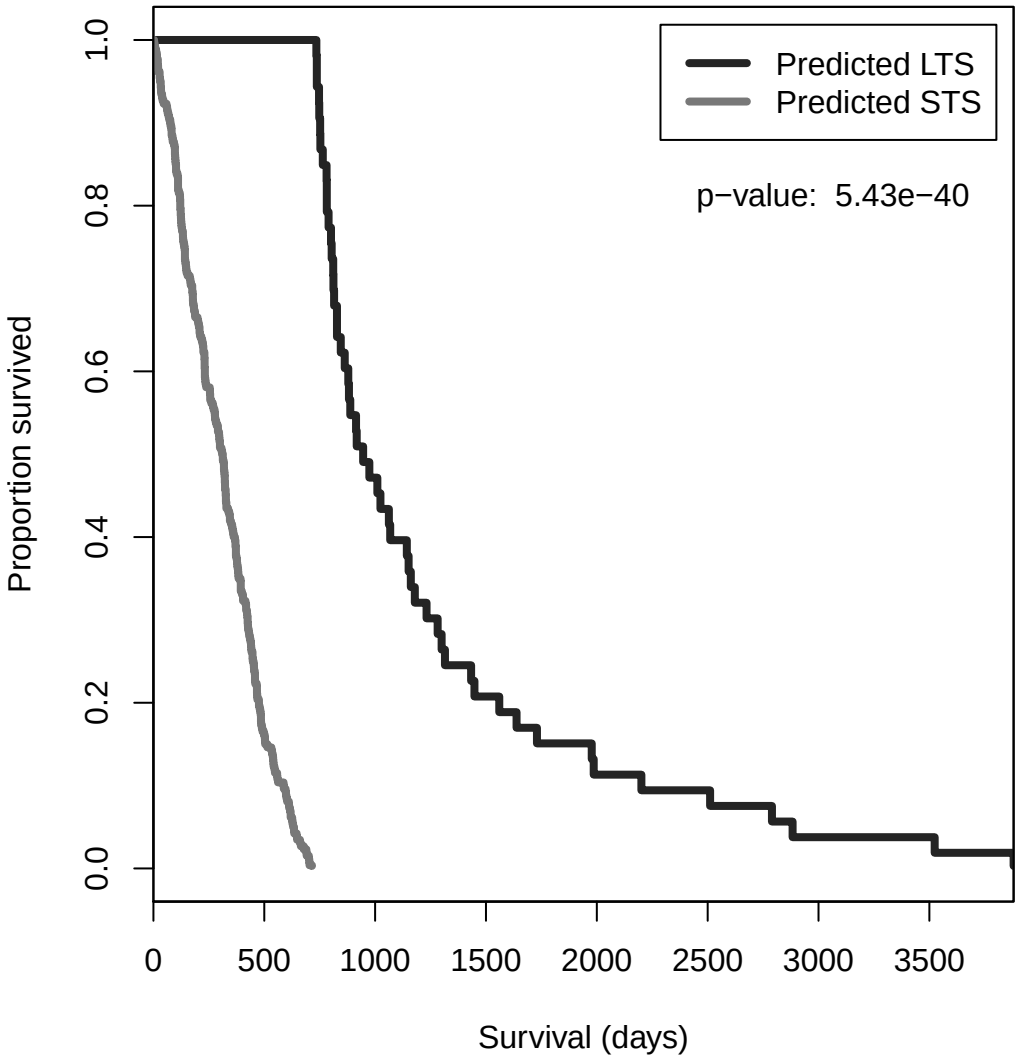

Figure 4.4. Kaplan-Meier curves for validation tests in which the C5.0 Decision Trees algorithm attempted to discriminate between longer-term survivors (LTS) and shorter-term survivors (STS) using 900 randomly simulated continuous variables and 100 randomly simulated binary variables. Three of the continuous variables and two of the binary variables were modified to support perfect discrimination between LTS and STS. No variable selection was performed.

Table 4.5. Cross-validation results when classification algorithms were applied to several UCI Machine Learning data sets. Values indicate the error rate that was attained for respective combinations of algorithm and data set. The TunedIT values represent the mean error rate that was observed on the TunedIT.org web site across all Weka classifiers that fall under the *bayes*, *functions*, and *trees* categories.

Data Set	C5.0	NBC	SVM	TunedIT
Breast Cancer	0.053	0.026	0.036	0.055
Hepatitis	0.247	0.156	0.221	0.193
Horse Colic	0.119	0.198	0.244	0.192
Ionosphere	0.094	0.077	0.060	0.118
Pima Indians	0.272	0.249	0.352	0.270
Statlog (Australian)	0.157	0.183	0.442	0.168
Statlog (Heart)	0.193	0.159	0.444	0.216
Tic Tac Toe	0.374	0.308	0.069	0.157
Sonar	0.250	0.288	0.327	0.268
Voting	0.030	0.095	0.042	0.055

Table 4.6. Cross-validation results when ensemble-learning approaches were applied to several UCI Machine Learning data sets. Values represent the error rate for respective combinations of data set and ensemble method. The TunedIT values represent the mean error rate that was observed on the TunedIT.org web site across all Weka classifiers that fall under the *bayes*, *functions*, and *trees* categories.

Data Set	Majority Vote	Simple Weighted Vote	Squared-Weighted Vote	LTS Predictive Value Weighted Vote	STS Predictive Value Weighted Vote	Select Best	Mean Probability	Weighted Mean Probability	Stacked Generalization	TunedIT
Breast Cancer	0.029	0.029	0.029	0.039	0.029	0.029	0.030	0.030	0.034	0.055
Hepatitis	0.201	0.195	0.195	0.169	0.195	0.201	0.201	0.195	0.195	0.193
Horse Colic	0.122	0.122	0.122	0.159	0.122	0.122	0.122	0.125	0.125	0.192
Ionosphere	0.066	0.066	0.066	0.063	0.066	0.066	0.068	0.066	0.077	0.118
Pima Indians	0.243	0.237	0.240	0.225	0.243	0.238	0.249	0.247	0.253	0.270
Statlog (Australian)	0.132	0.132	0.132	0.172	0.132	0.132	0.141	0.135	0.136	0.168
Statlog (Heart)	0.152	0.156	0.148	0.148	0.152	0.152	0.144	0.152	0.156	0.216
Tic Tac Toe	0.210	0.210	0.201	0.069	0.201	0.210	0.186	0.175	0.023	0.157
Sonar	0.221	0.221	0.226	0.255	0.221	0.221	0.226	0.231	0.231	0.268
Voting	0.044	0.044	0.044	0.042	0.044	0.044	0.044	0.044	0.046	0.055

Table 4.7. Cross-validation results when all patients were included, two-year survival was the split point between longer-term survivors and shorter-term survivors, and the *SVM-RFE* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.220	0.938	0.000	0.590	4.46e-05*
	SVM	0.204	0.954	0.019	0.455	0.238
Treatments	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.182	0.931	0.264	0.707	0.000388*
	SVM	0.169	1.000	0.000	0.488	N/A
Histology	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.173	0.992	0.019	0.497	0.615
	SVM	0.169	1.000	0.000	0.544	N/A
DNA Methylation	C5.0	0.245	0.880	0.100	0.490	0.328
	NBC	0.181	0.962	0.067	0.554	0.427
	SVM	0.181	0.943	0.167	0.557	0.037*
Somatic Mutations	C5.0	0.196	1.000	0.000	0.500	N/A
	NBC	0.277	0.889	0.045	0.559	0.385
	SVM	0.214	0.967	0.045	0.395	0.231
DNA Copy Number	C5.0	0.170	0.988	0.020	0.504	0.43
	NBC	0.272	0.835	0.180	0.516	0.563
	SVM	0.167	0.996	0.000	0.543	0.0797
mRNA Expression	C5.0	0.262	0.861	0.146	0.504	0.892
	NBC	0.211	0.918	0.167	0.570	0.108
	SVM	0.240	0.887	0.146	0.536	0.551
miRNA Expression	C5.0	0.261	0.846	0.229	0.538	0.987
	NBC	0.196	0.952	0.104	0.540	0.657
	SVM	0.174	1.000	0.000	0.589	N/A
All Data	C5.0	0.233	0.877	0.226	0.552	0.00338*
	NBC	0.204	0.942	0.075	0.561	0.383
	SVM	0.169	0.977	0.113	0.643	0.0233*

Table 4.8. Cross-validation results when all patients were included, two-year survival was the split point between longer-term survivors and shorter-term survivors, and the *None* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.166	0.981	0.113	0.683	0.0159*
	SVM	0.173	0.996	0.000	0.553	0.614
Treatments	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.204	0.915	0.208	0.705	0.00104*
	SVM	0.169	1.000	0.000	0.463	N/A
Histology	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.173	0.985	0.057	0.474	0.487
	SVM	0.169	1.000	0.000	0.524	N/A
DNA Methylation	C5.0	0.324	0.772	0.167	0.469	0.21
	NBC	0.170	0.981	0.033	0.584	0.495
	SVM	0.160	1.000	0.000	0.547	N/A
Somatic Mutations	C5.0	0.196	0.989	0.045	0.517	0.0854
	NBC	0.259	0.922	0.000	0.453	0.0578
	SVM	0.205	0.989	0.000	0.397	0.337
DNA Copy Number	C5.0	0.184	0.965	0.060	0.512	0.676
	NBC	0.302	0.784	0.260	0.556	0.303
	SVM	0.164	1.000	0.000	0.496	N/A
mRNA Expression	C5.0	0.262	0.853	0.188	0.520	0.362
	NBC	0.215	0.913	0.167	0.533	0.134
	SVM	0.176	0.996	0.000	0.574	0.866
miRNA Expression	C5.0	0.250	0.860	0.229	0.544	0.122
	NBC	0.261	0.882	0.062	0.432	0.0669
	SVM	0.174	1.000	0.000	0.586	N/A
All Data	C5.0	0.246	0.850	0.283	0.567	0.00474*
	NBC	0.240	0.908	0.038	0.503	0.433
	SVM	0.169	1.000	0.000	0.632	N/A

Table 4.9. Cross-validation results when all patients were included, two-year survival was the split point between longer-term survivors and shorter-term survivors, and the *RELIEF-F* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.176	0.992	0.000	0.694	0.494
	SVM	0.169	1.000	0.000	0.451	N/A
Treatments	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.195	0.958	0.057	0.681	0.116
	SVM	0.179	0.988	0.000	0.498	0.187
Histology	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.173	0.981	0.075	0.463	0.185
	SVM	0.169	1.000	0.000	0.505	N/A
DNA Methylation	C5.0	0.239	0.892	0.067	0.480	0.609
	NBC	0.186	0.943	0.133	0.600	0.0609
	SVM	0.160	1.000	0.000	0.412	N/A
Somatic Mutations	C5.0	0.196	1.000	0.000	0.500	N/A
	NBC	0.214	0.967	0.045	0.468	0.366
	SVM	0.205	0.989	0.000	0.430	0.337
DNA Copy Number	C5.0	0.164	1.000	0.000	0.500	N/A
	NBC	0.308	0.796	0.160	0.512	0.986
	SVM	0.164	1.000	0.000	0.488	N/A
mRNA Expression	C5.0	0.262	0.879	0.062	0.471	0.405
	NBC	0.262	0.866	0.125	0.489	0.721
	SVM	0.186	0.983	0.000	0.465	0.63
miRNA Expression	C5.0	0.217	0.908	0.188	0.548	0.0865
	NBC	0.178	0.991	0.021	0.484	0.198
	SVM	0.174	1.000	0.000	0.615	N/A
All Data	C5.0	0.204	0.942	0.075	0.509	0.0943
	NBC	0.153	0.969	0.245	0.671	1.89e-05*
	SVM	0.169	0.996	0.019	0.610	0.322

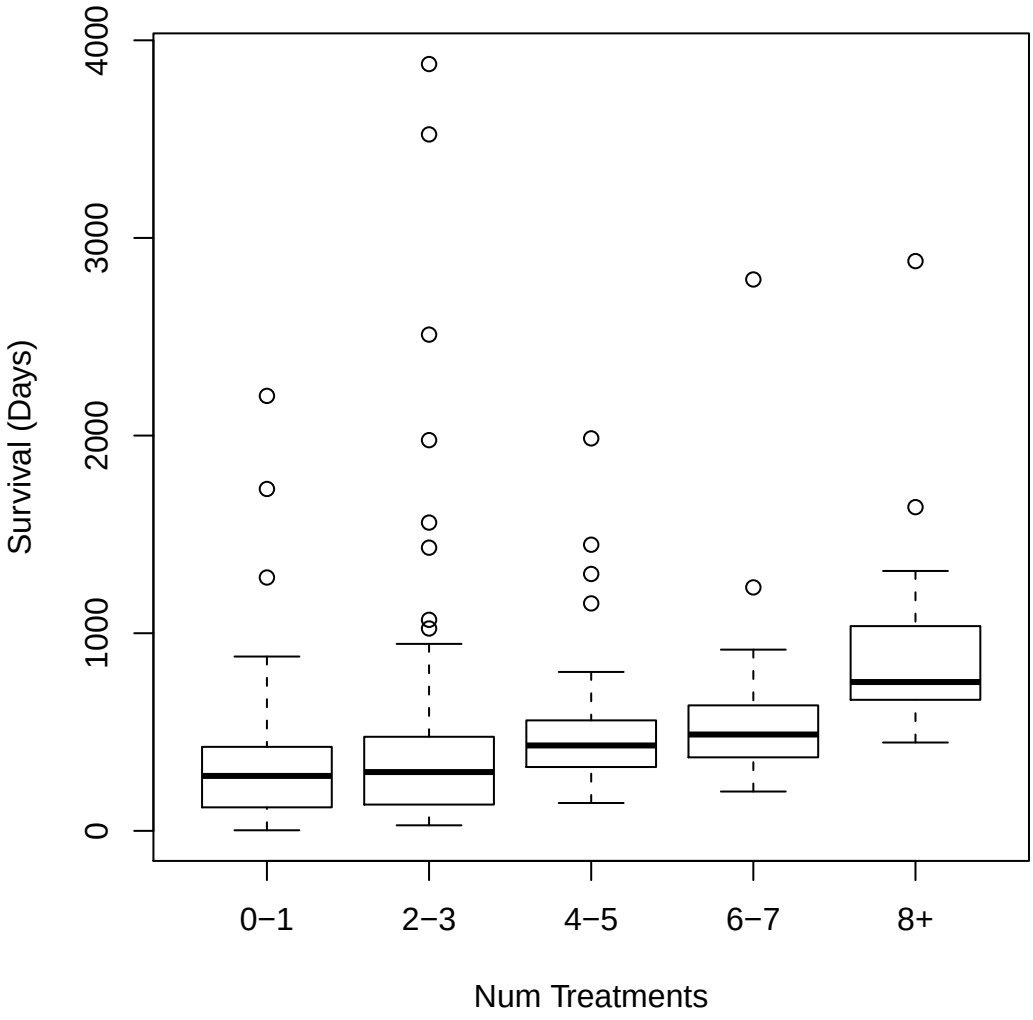

Figure 4.5. Patient overall survival versus the total number of treatments received by each patient.

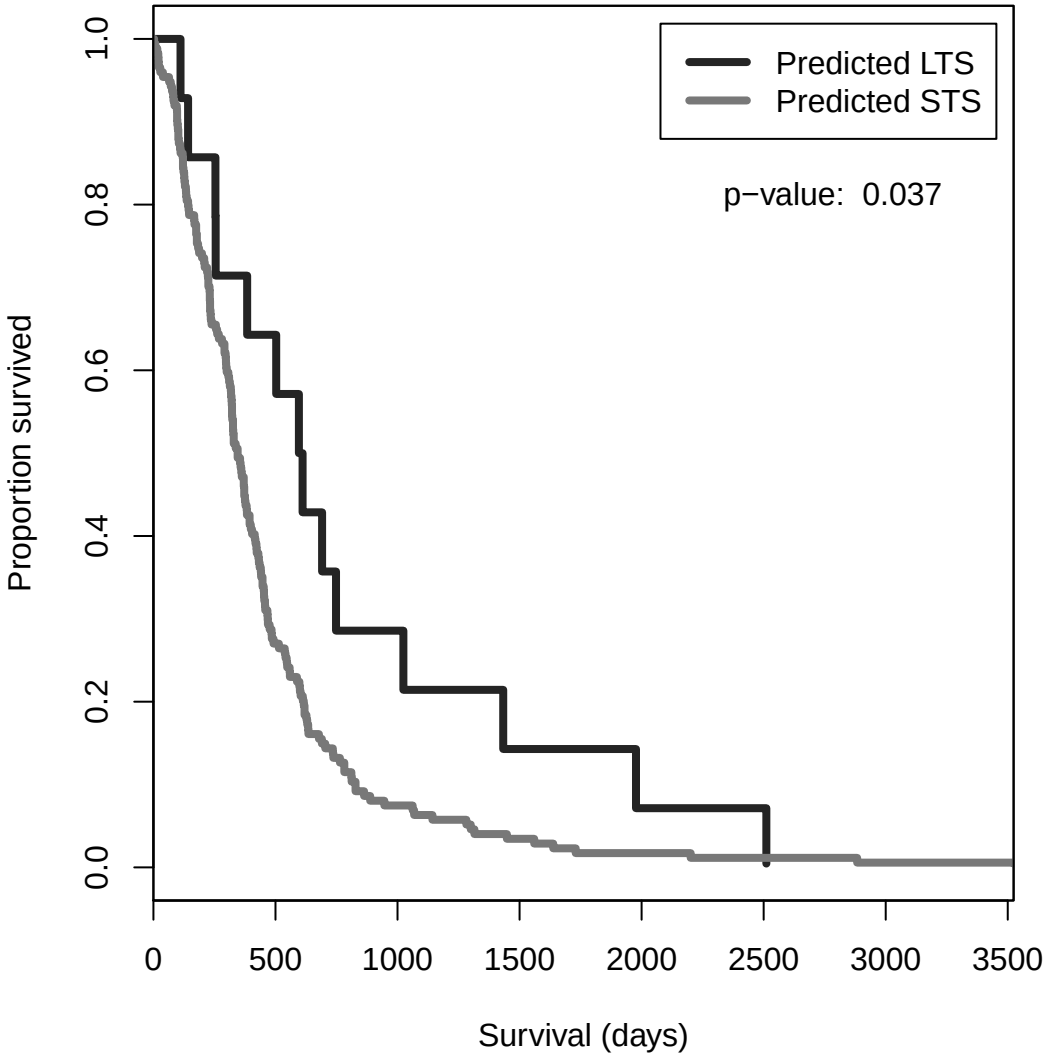

Figure 4.6. Kaplan-Meier curves comparing overall survival of patients predicted as longer-term survivor (LTS) versus patients predicted as shorter-term survivor (STS) for SVM models trained on DNA methylation data. Support Vector Machines-Recursive Feature Elimination was used for variable selection, and two-year survival was the split point between LTS and STS.

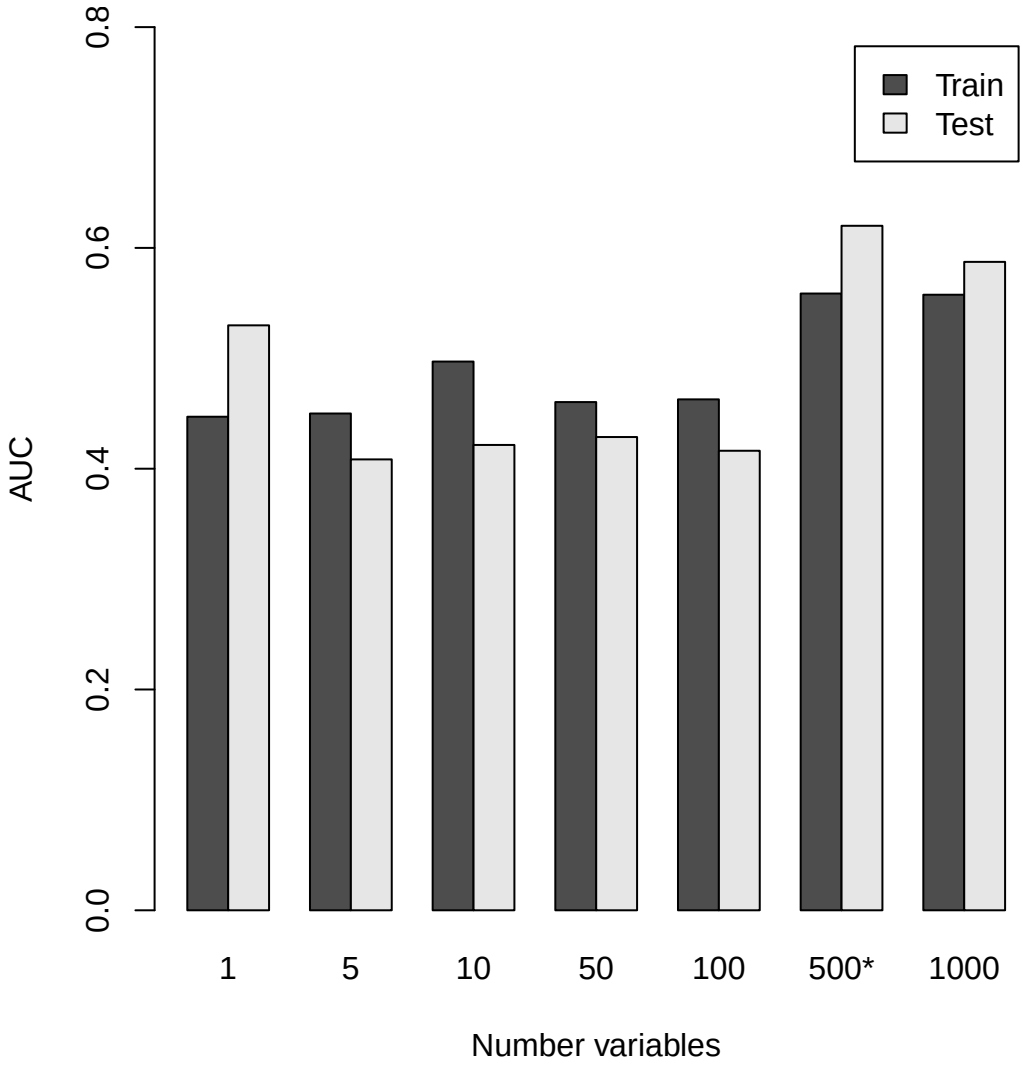

Figure 4.7. Area under receiver operating characteristic curve versus number of DNA methylation genes included in Support Vector Machines models. Support Vector Machines-Recursive Feature Elimination was used for variable selection, and two-year survival was the split point between longer-term survivors and shorter-term survivors.

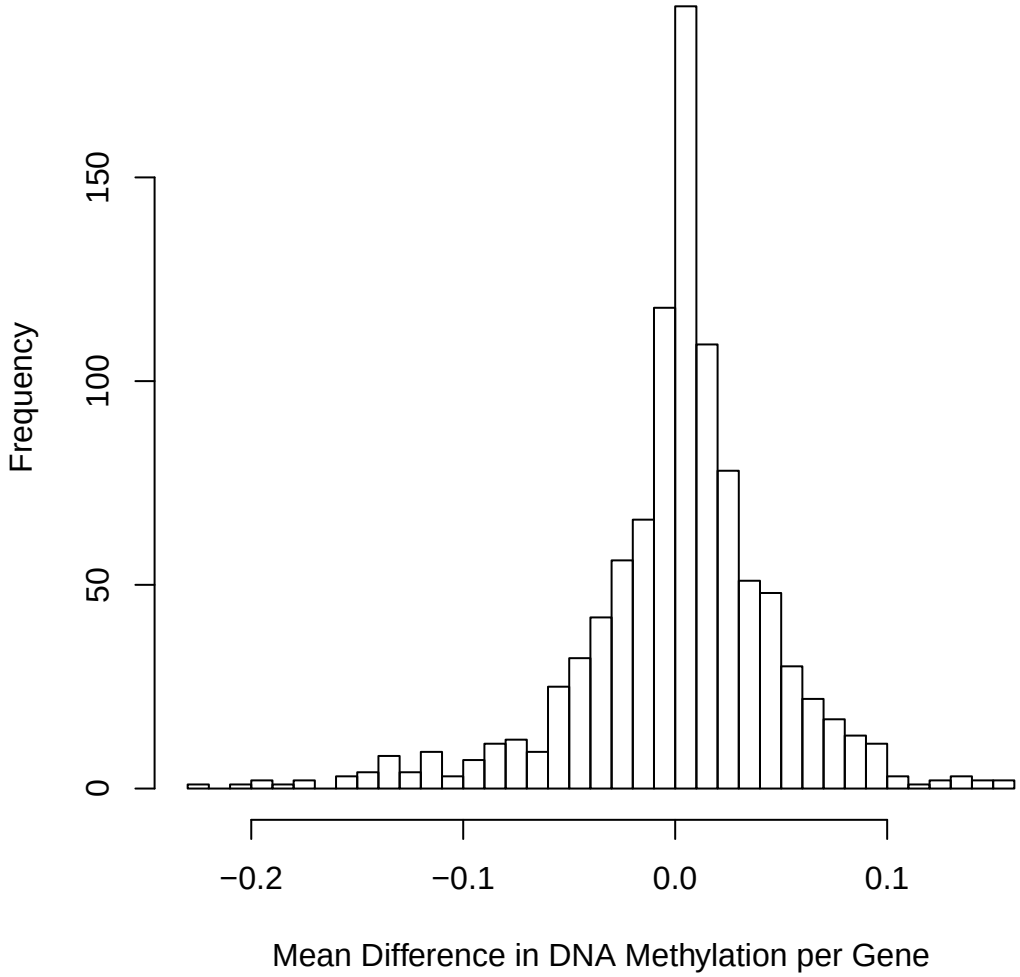

Figure 4.8. Mean difference in global DNA methylation between longer-term survivors (LTS) and shorter-term survivors (STS) for each gene that was profiled. Two-year survival was the split point between LTS and STS.

Table 4.10. Cross-validation results when all patients were included, two-year survival was the survival split between longer-term survivors and shorter-term survivors, and ensemble-learning approaches were applied.

Ensemble Method	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Majority Vote	0.169	1.000	0.000	0.601	N/A
Simple Weighted Vote	0.169	1.000	0.000	0.610	N/A
Squared-Weighted Vote	0.169	1.000	0.000	0.614	N/A
LTS Predictive Value Weighted Vote	0.169	1.000	0.000	0.620	N/A
STS Predictive Value Weighted Vote	0.169	1.000	0.000	0.603	N/A
Select Best	0.192	0.938	0.170	0.676	0.00762*
Mean Probability	0.169	1.000	0.000	0.641	N/A
Weighted Mean Probability	0.169	1.000	0.000	0.651	N/A
Stacked Generalization	0.220	0.900	0.189	0.544	0.0408*

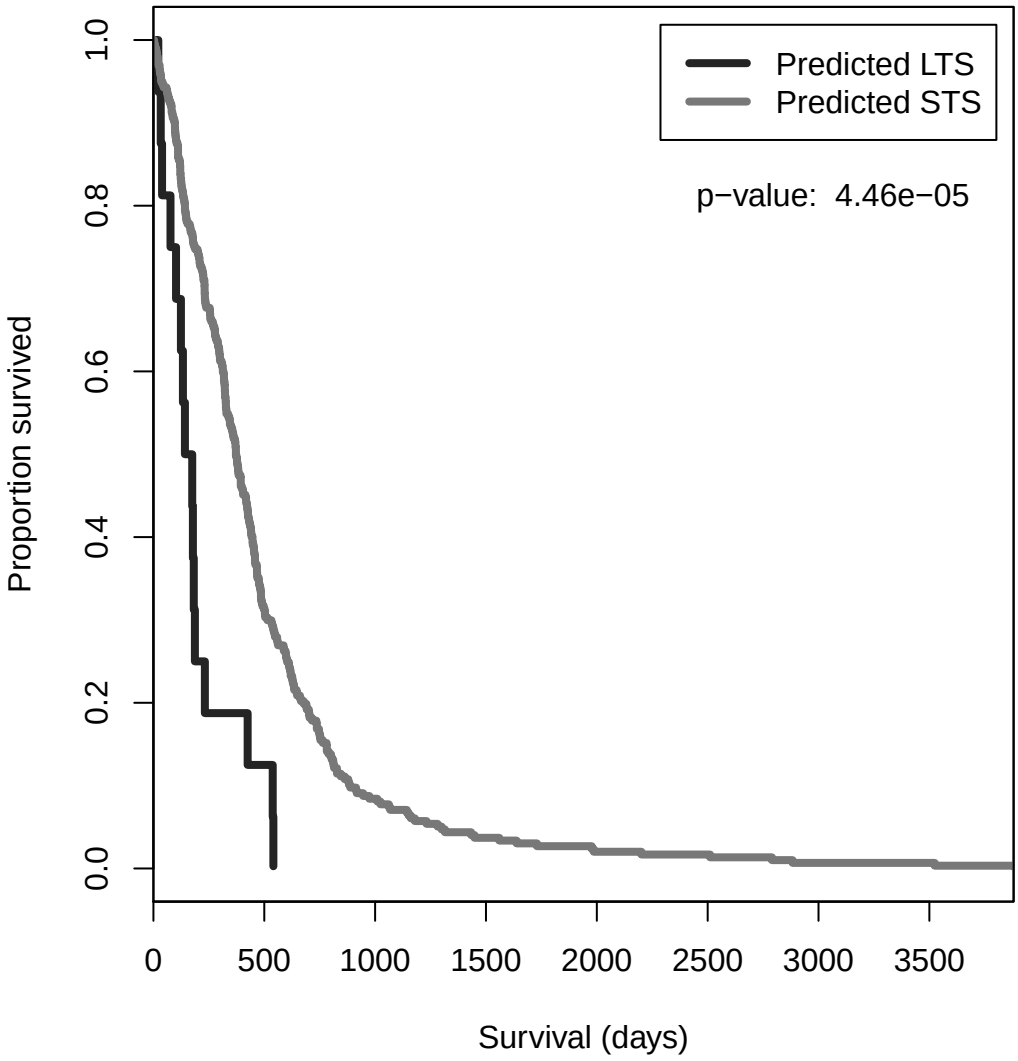

Figure 4.9. Kaplan-Meier curves comparing overall survival of patients predicted as longer-term survivor (LTS) versus patients predicted as shorter-term survivor (STS) for NBC models trained on clinical data. Support Vector Machines-Recursive Feature Elimination was used for variable selection, and two-year survival was the split point between LTS from STS.

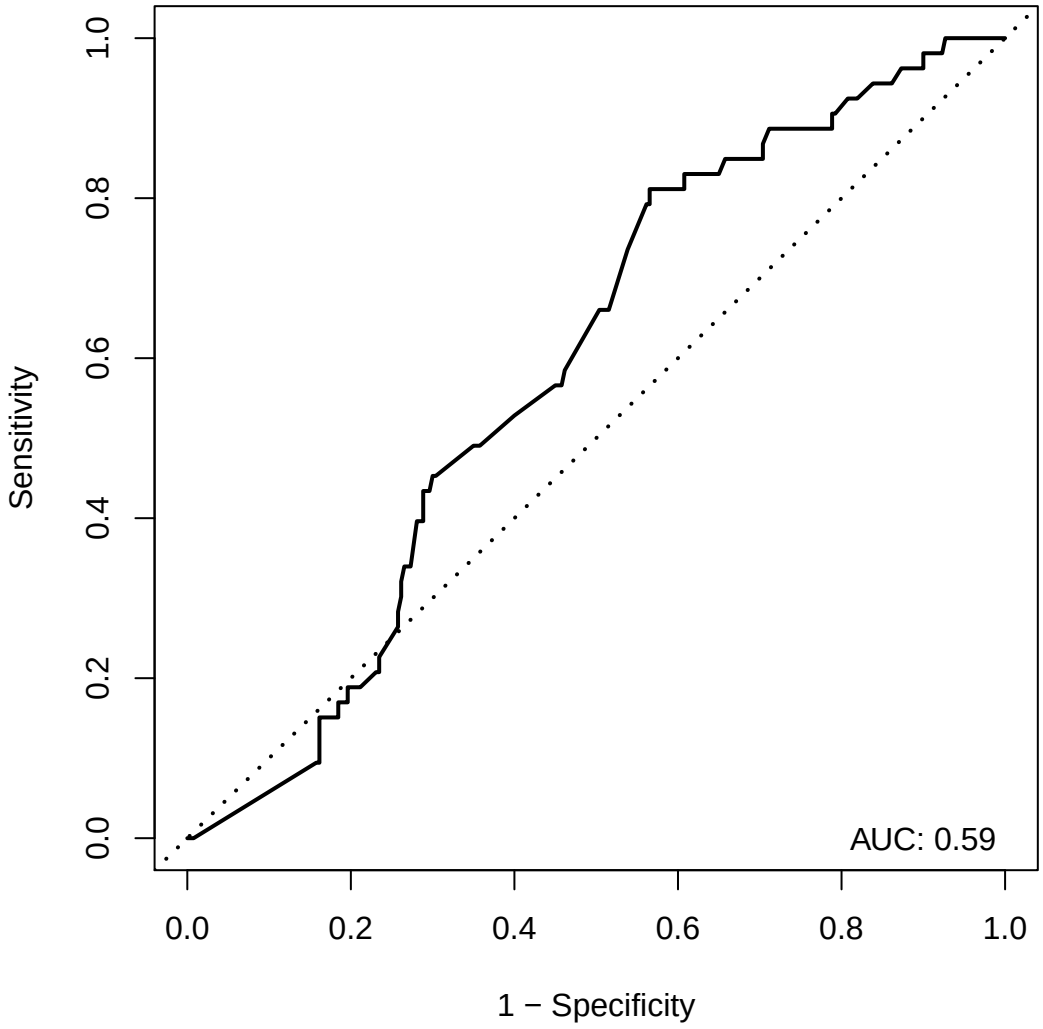

Figure 4.10. Receiver operating characteristic curve for NBC models trained on clinical data. Support Vector Machines-Recursive Feature Elimination was used for variable selection, and two-year survival was the split point between longer-term survivors and shorter-term survivors.

Table 4.11. Cross-validation results when all patients were included and two-year survival was used as the split point between longer-term survivors and shorter-term survivors. The *prior-knowledge* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.173	0.981	0.075	0.676	0.0655
	SVM	0.169	1.000	0.000	0.508	N/A
Treatments	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.169	1.000	0.000	0.596	N/A
	SVM	0.169	1.000	0.000	0.497	N/A
Histology	C5.0	0.169	1.000	0.000	0.500	N/A
	NBC	0.169	1.000	0.000	0.431	N/A
	SVM	0.169	1.000	0.000	0.457	N/A
DNA Methylation	C5.0	0.160	1.000	0.000	0.500	N/A
	NBC	0.170	0.987	0.000	0.575	0.302
	SVM	0.165	0.994	0.000	0.510	0.916
Somatic Mutations	C5.0	0.196	1.000	0.000	0.500	N/A
	NBC	0.196	0.967	0.136	0.520	0.00479*
	SVM	0.205	0.989	0.000	0.353	0.628
DNA Copy Number	C5.0	0.164	1.000	0.000	0.500	N/A
	NBC	0.266	0.839	0.200	0.504	0.419
	SVM	0.164	1.000	0.000	0.428	N/A
mRNA Expression	C5.0	0.172	1.000	0.000	0.500	N/A
	NBC	0.211	0.913	0.188	0.538	0.00599*
	SVM	0.172	1.000	0.000	0.459	N/A
miRNA Expression	C5.0	0.174	1.000	0.000	0.500	N/A
	NBC	0.181	0.978	0.062	0.461	0.171
	SVM	0.174	1.000	0.000	0.449	N/A
All Data	C5.0	0.188	0.942	0.170	0.556	0.0299*
	NBC	0.243	0.888	0.113	0.567	0.598
	SVM	0.169	1.000	0.000	0.416	N/A

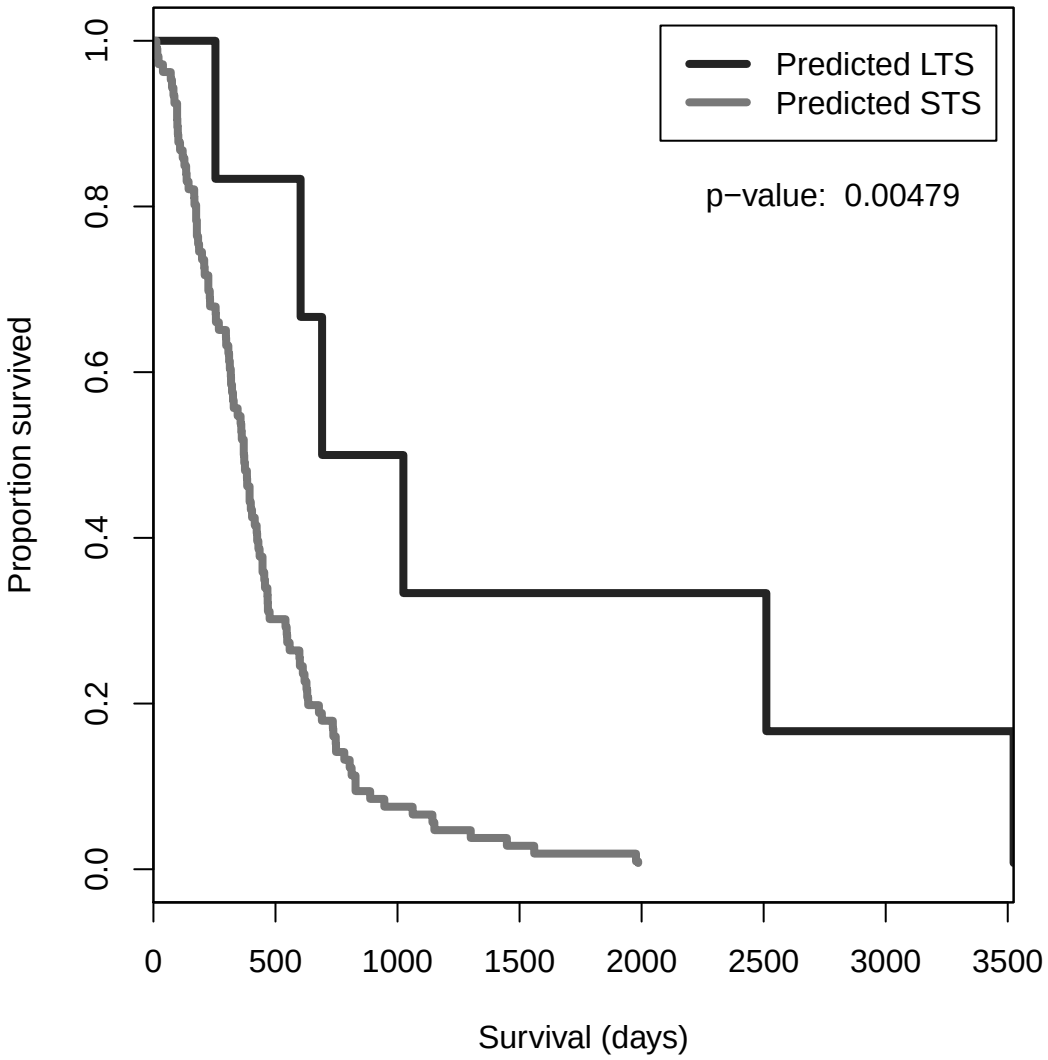

Figure 4.11. Kaplan-Meier curves comparing overall survival of patients predicted as longer-term survivor versus patients predicted as shorter-term survivor for Naïve Bayes Classifier models trained on IDH1 and TP53 somatic mutations.

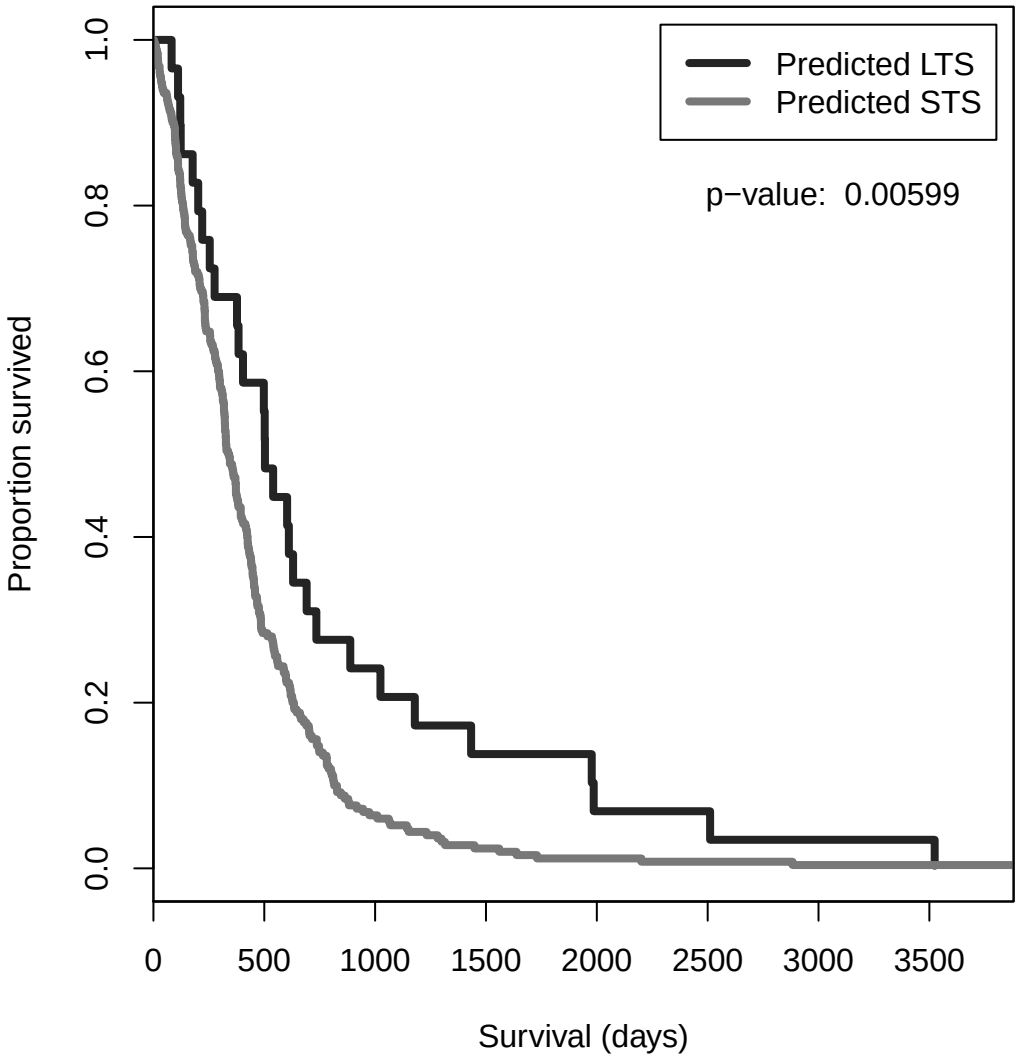

Figure 4.12. Kaplan-Meier curves comparing overall survival of patients predicted as longer-term survivor versus patients predicted as longer-term survivor for Naïve Bayes Classifier models trained on the Colman, et al. mRNA expression profile. [1]

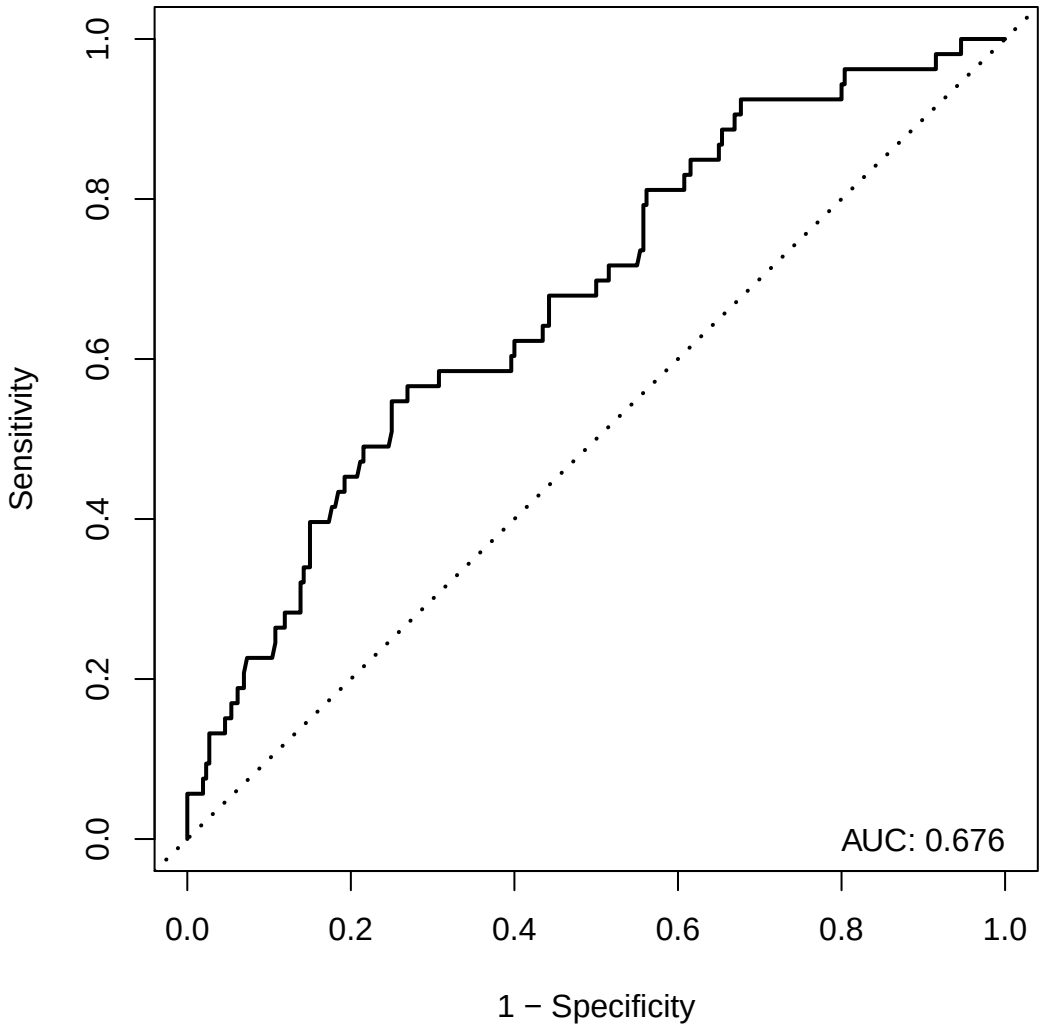

Figure 4.13. Receiver operating characteristic curve for Naïve Bayes Classifier models trained on clinical variables that have been reported in the literature to have prognostic relevance for glioblastoma multiforme.

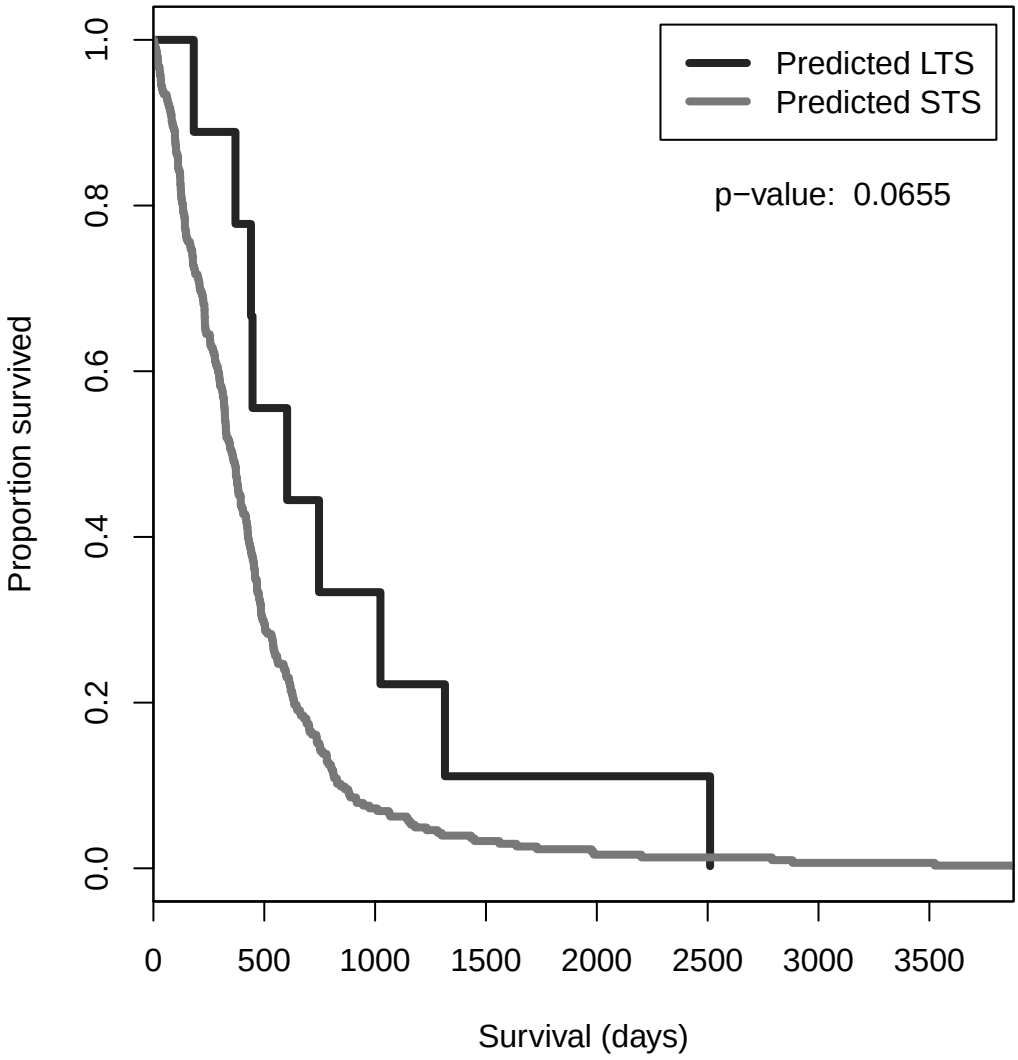

Figure 4.14. Kaplan-Meier curves comparing overall survival of patients predicted as longer-term survivor versus patients predicted as shorter-term survivor for Naïve Bayes Classifier models trained on clinical variables that have been reported in the literature to have prognostic relevance for glioblastoma multiforme.

Table 4.12. Cross-validation results when all patients were included, two-year survival was the split point between longer-term survivors and shorter-term survivors, *prior-knowledge* variables were used, and ensemble-learning approaches were applied.

Ensemble Method	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Majority Vote	0.169	1.000	0.000	0.587	N/A
Simple Weighted Vote	0.169	1.000	0.000	0.583	N/A
Squared-Weighted Vote	0.169	1.000	0.000	0.583	N/A
LTS Predictive Value Weighted Vote	0.173	0.992	0.019	0.599	0.1
STS Predictive Value Weighted Vote	0.169	1.000	0.000	0.584	N/A
Select Best	0.173	0.981	0.075	0.676	0.0655
Mean Probability	0.169	1.000	0.000	0.637	N/A
Weighted Mean Probability	0.169	1.000	0.000	0.646	N/A
Stacked Generalization	0.176	0.988	0.019	0.504	0.743

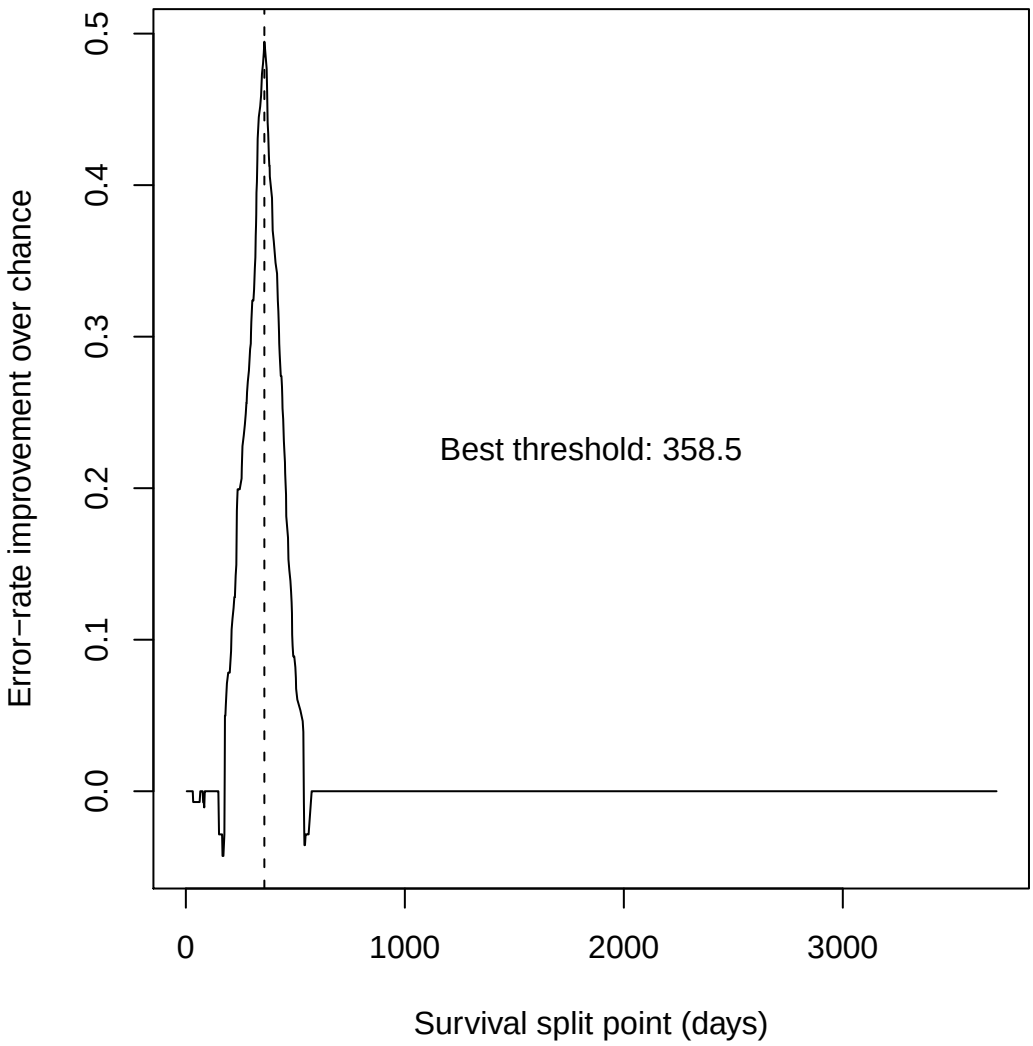

Figure 4.15. Results of empirical split-point selection on data simulated to support perfect separation between longer-term survivors and shorter-term survivors at 360-days survival. The error rate (corrected for what would be observed if the majority class were predicted by default) was used as the evaluation criterion at each split point. When a tie occurred, the median value was selected.

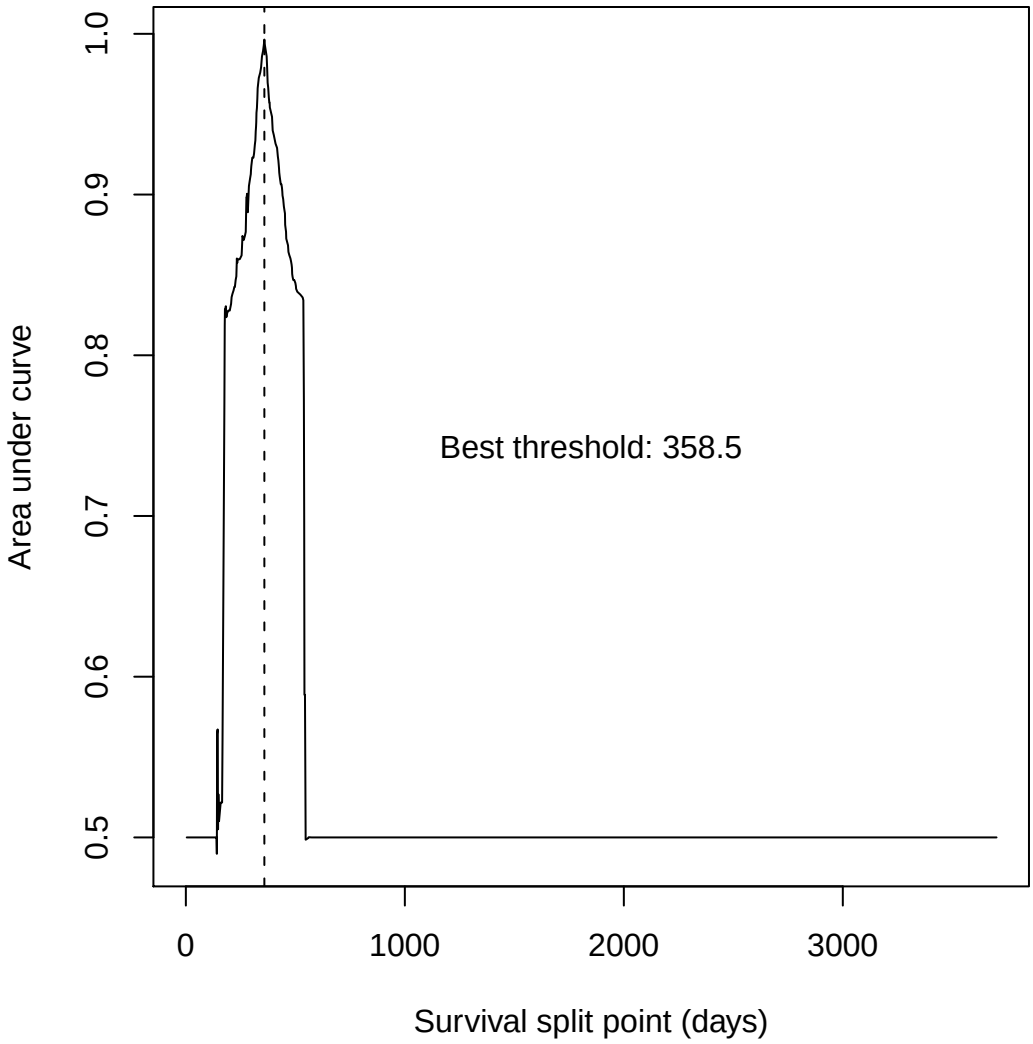

Figure 4.16. Results of empirical split-point selection on data simulated to support perfect separation between longer-term survivors and shorter-term survivors at 360-days survival. The AUC was used as the evaluation criterion at each split point.

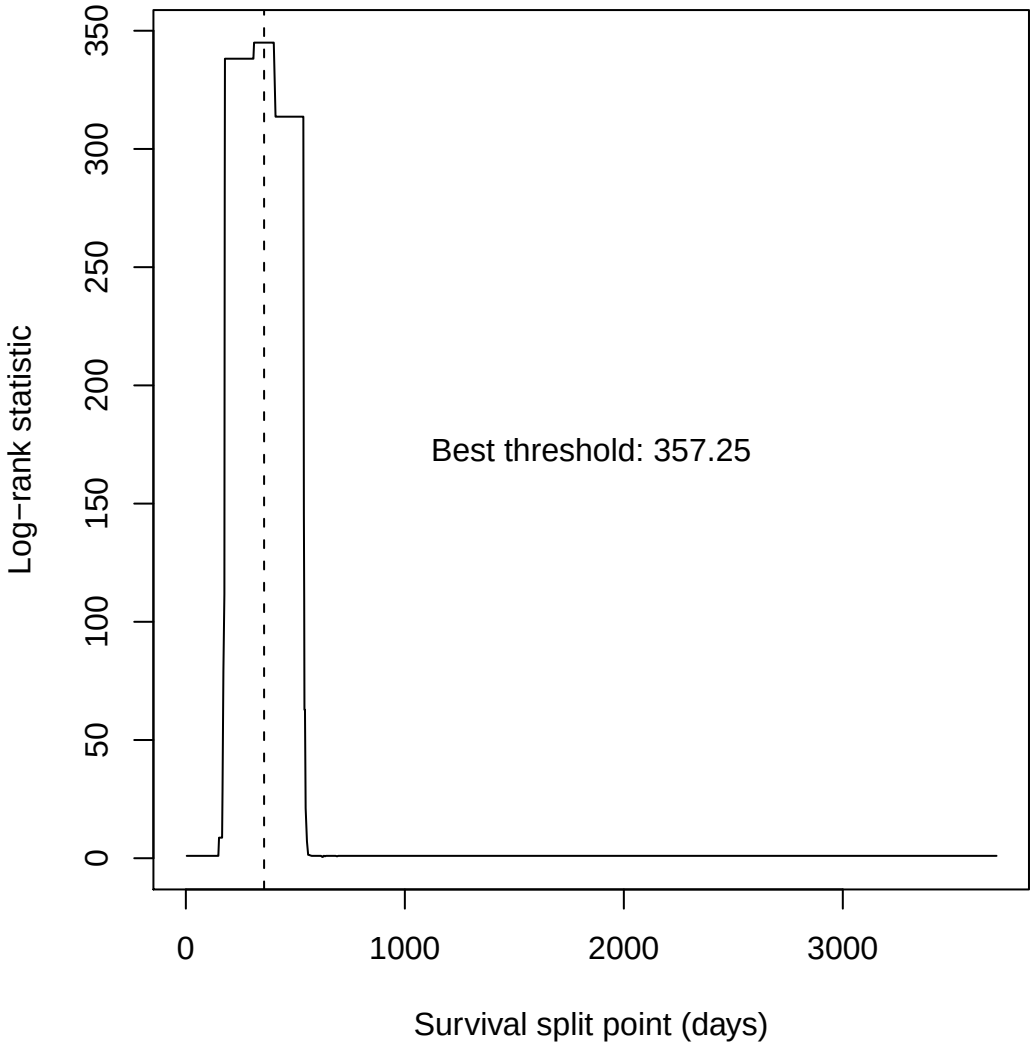

Figure 4.17. Results of empirical split-point selection on data simulated to support perfect separation between longer-term survivors and shorter-term survivors at 360-days survival. The log-rank statistic was used as the evaluation criterion at each split point. When a tie occurred, the median value was selected.

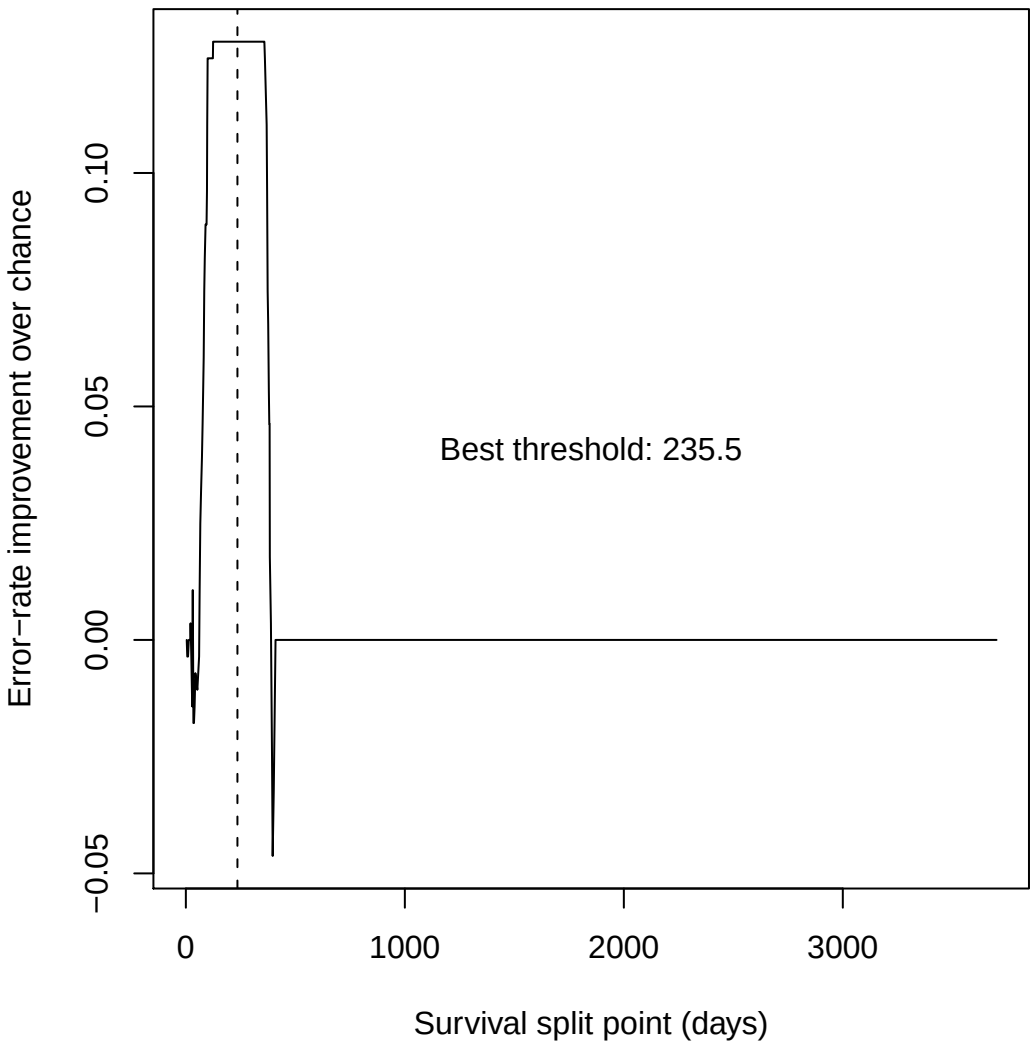

Figure 4.18. Results of empirical split-point selection on data simulated to support perfect separation between longer-term survivors and shorter-term survivors at 100-days survival. The error rate (corrected for what would be observed if the majority class were predicted by default) was used as the evaluation criterion at each split point. When a tie occurred, the median value was selected.

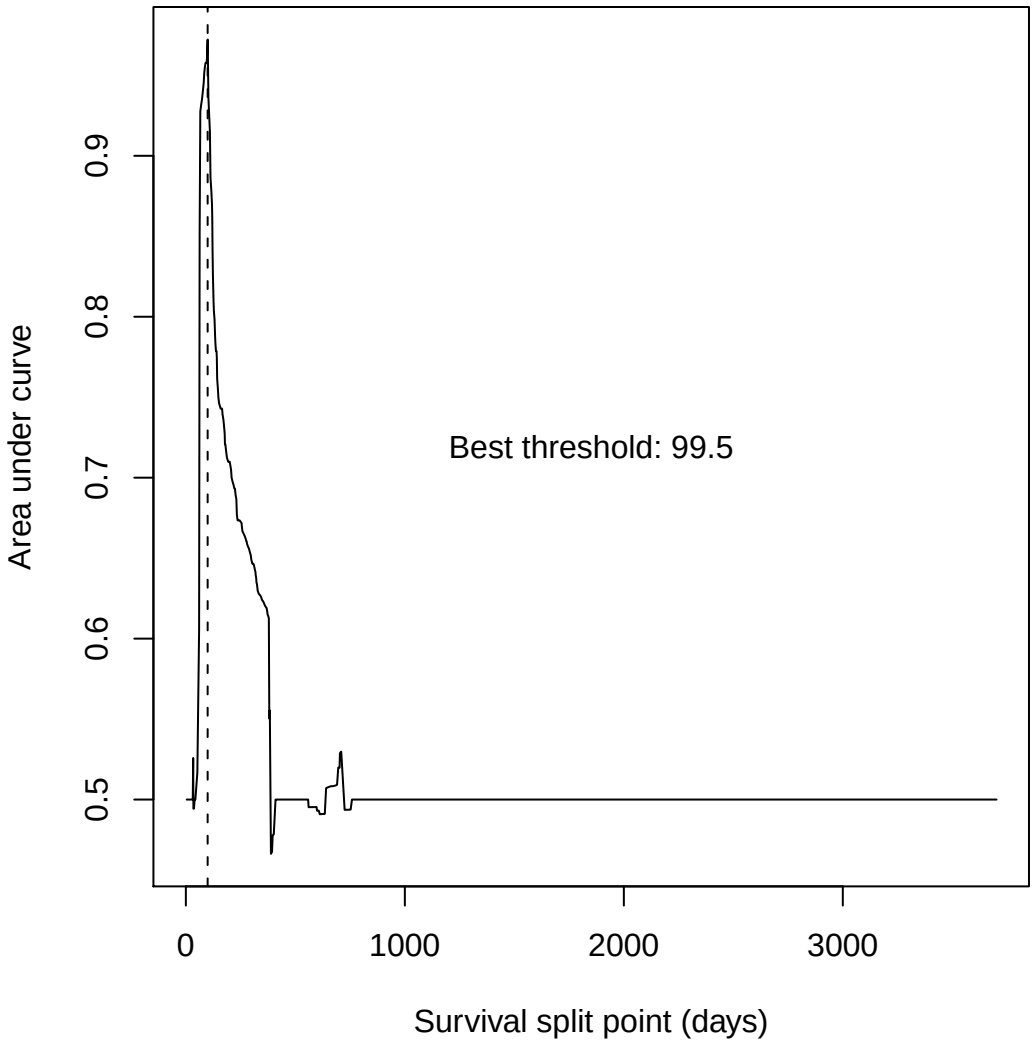

Figure 4.19. Results of empirical split-point selection on data simulated to support perfect separation between longer-term survivors and shorter-term survivors at 100-days survival. The AUC was used as the evaluation criterion at each split point.

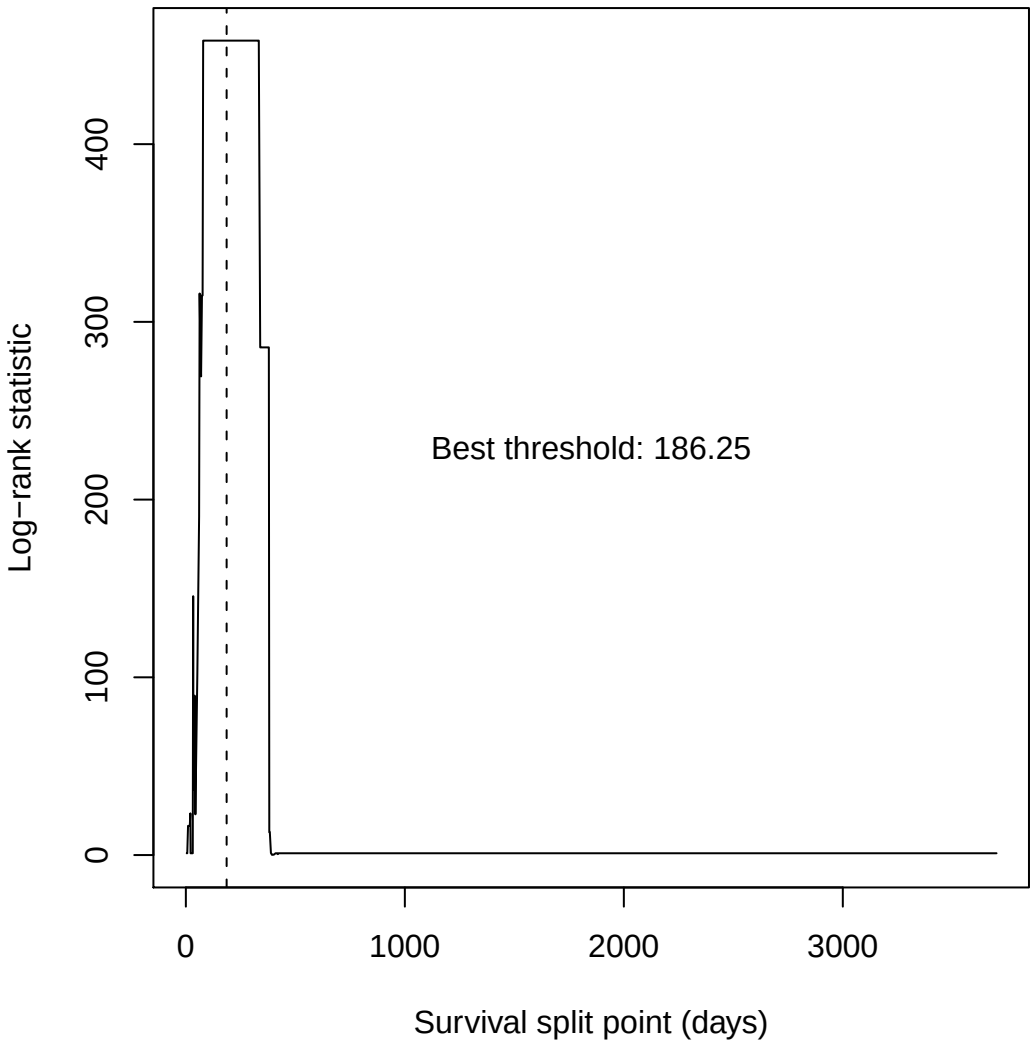

Figure 4.20. Results of empirical split-point selection on data simulated to support perfect separation between longer-term survivors and shorter-term survivors at 100-days survival. The log-rank statistic was used as the evaluation criterion at each split point. When a tie occurred, the median value was selected.

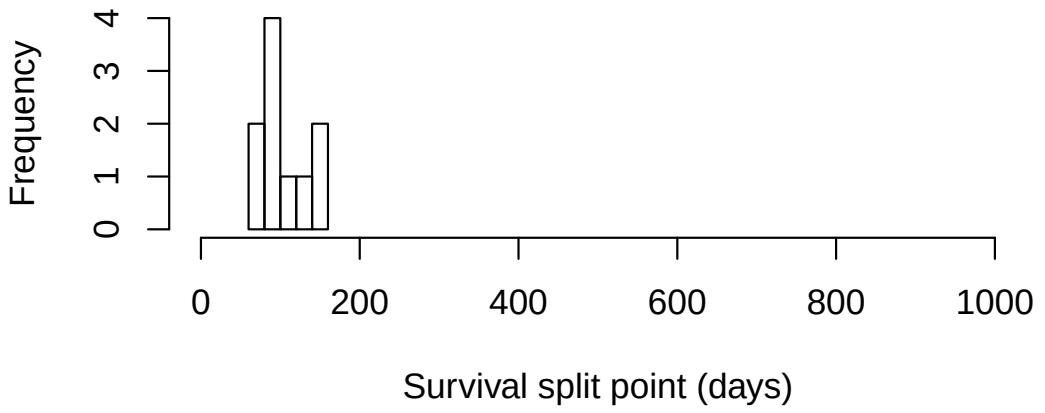

Figure 4.21. Survival split points selected for each cross-validation fold when the empirical split-point method was applied to the full data set.

Table 4.13. Cross-validation results when all patients were included and the empirical split-point method was used to distinguish longer-term survivors from shorter-term survivors in each cross-validation fold. The *SVM-RFE* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.137	0.000	1.000	0.500	N/A
	NBC	0.154	0.000	0.983	0.611	0.782
	SVM	0.150	0.000	0.985	0.491	0.91
Treatments	C5.0	0.118	0.475	0.945	0.710	3.78e-05*
	NBC	0.102	0.642	0.953	0.857	1.87e-06*
	SVM	0.102	0.400	0.970	0.854	0.000104*
Histology	C5.0	0.137	0.000	1.000	0.500	N/A
	NBC	0.153	0.000	0.982	0.539	0.648
	SVM	0.137	0.000	1.000	0.558	N/A
DNA Methylation	C5.0	0.169	0.067	0.922	0.494	0.246
	NBC	0.133	0.025	0.972	0.497	0.574
	SVM	0.161	0.083	0.923	0.530	0.109
Somatic Mutations	C5.0	0.101	0.000	1.000	0.500	N/A
	NBC	0.116	0.000	0.983	0.483	0.949
	SVM	0.109	0.000	0.992	0.516	0.854
DNA Copy Number	C5.0	0.138	0.000	1.000	0.500	N/A
	NBC	0.216	0.062	0.883	0.502	0.481
	SVM	0.141	0.013	0.992	0.507	0.0848
mRNA Expression	C5.0	0.233	0.104	0.866	0.485	0.327
	NBC	0.183	0.083	0.930	0.521	0.0819
	SVM	0.171	0.237	0.897	0.560	0.149
miRNA Expression	C5.0	0.197	0.278	0.877	0.578	0.475
	NBC	0.152	0.070	0.965	0.518	0.302
	SVM	0.129	0.000	1.000	0.589	N/A
All Data	C5.0	0.144	0.350	0.923	0.637	7.4e-05*
	NBC	0.131	0.192	0.958	0.789	0.000548*
	SVM	0.099	0.408	0.967	0.845	4.98e-06*

Table 4.14. Cross-validation results when all patients were included and the empirical split-point method was used to distinguish longer-term survivors from shorter-term survivors in each cross-validation fold. The *None* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.131	0.071	0.988	0.529	0.148
	NBC	0.144	0.067	0.974	0.717	0.024*
	SVM	0.147	0.000	0.989	0.389	0.842
Treatments	C5.0	0.118	0.475	0.945	0.710	3.78e-05*
	NBC	0.102	0.658	0.949	0.864	7.23e-06*
	SVM	0.102	0.400	0.970	0.845	0.000104*
Histology	C5.0	0.137	0.000	1.000	0.500	N/A
	NBC	0.163	0.000	0.970	0.540	0.437
	SVM	0.137	0.000	1.000	0.429	N/A
DNA Methylation	C5.0	0.164	0.158	0.922	0.540	0.114
	NBC	0.120	0.000	0.994	0.497	0.929
	SVM	0.120	0.000	0.994	0.419	0.712
Somatic Mutations	C5.0	0.101	0.000	1.000	0.500	N/A
	NBC	0.116	0.000	0.983	0.506	0.949
	SVM	0.101	0.000	1.000	0.680	N/A
DNA Copy Number	C5.0	0.151	0.050	0.981	0.516	0.509
	NBC	0.295	0.119	0.786	0.451	0.315
	SVM	0.141	0.000	0.996	0.485	0.621
mRNA Expression	C5.0	0.263	0.157	0.817	0.487	0.304
	NBC	0.157	0.083	0.957	0.516	0.018*
	SVM	0.136	0.000	0.995	0.682	0.966
miRNA Expression	C5.0	0.199	0.259	0.876	0.567	0.498
	NBC	0.276	0.071	0.807	0.446	0.306
	SVM	0.129	0.000	1.000	0.608	N/A
All Data	C5.0	0.134	0.438	0.931	0.684	0.000118*
	NBC	0.160	0.058	0.960	0.501	0.056
	SVM	0.141	0.000	0.996	0.701	0.862

Table 4.15. Cross-validation results when all patients were included and the empirical split-point method was used to distinguish longer-term survivors from shorter-term survivors in each cross-validation fold. The *RELIEF-F* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.137	0.000	1.000	0.500	N/A
	NBC	0.205	0.154	0.885	0.707	0.146
	SVM	0.186	0.142	0.913	0.613	0.129
Treatments	C5.0	0.118	0.475	0.945	0.710	3.78e-05*
	NBC	0.102	0.642	0.953	0.856	1.87e-06*
	SVM	0.106	0.367	0.974	0.864	0.000201*
Histology	C5.0	0.137	0.000	1.000	0.500	N/A
	NBC	0.163	0.000	0.970	0.536	0.54
	SVM	0.141	0.000	0.996	0.434	0.88
DNA Methylation	C5.0	0.191	0.000	0.909	0.454	0.511
	NBC	0.146	0.033	0.960	0.501	0.439
	SVM	0.146	0.000	0.964	0.452	0.512
Somatic Mutations	C5.0	0.101	0.000	1.000	0.500	N/A
	NBC	0.126	0.000	0.971	0.439	0.699
	SVM	0.109	0.000	0.992	0.399	0.854
DNA Copy Number	C5.0	0.138	0.000	1.000	0.500	N/A
	NBC	0.401	0.447	0.620	0.589	0.246
	SVM	0.138	0.000	1.000	0.500	N/A
mRNA Expression	C5.0	0.157	0.068	0.956	0.512	0.393
	NBC	0.203	0.073	0.901	0.452	0.187
	SVM	0.146	0.037	0.975	0.640	0.433
miRNA Expression	C5.0	0.182	0.084	0.920	0.502	0.335
	NBC	0.256	0.151	0.791	0.476	0.386
	SVM	0.129	0.000	1.000	0.599	N/A
All Data	C5.0	0.134	0.333	0.950	0.642	0.000458*
	NBC	0.131	0.333	0.945	0.814	0.000262*
	SVM	0.106	0.283	0.983	0.806	0.000686*

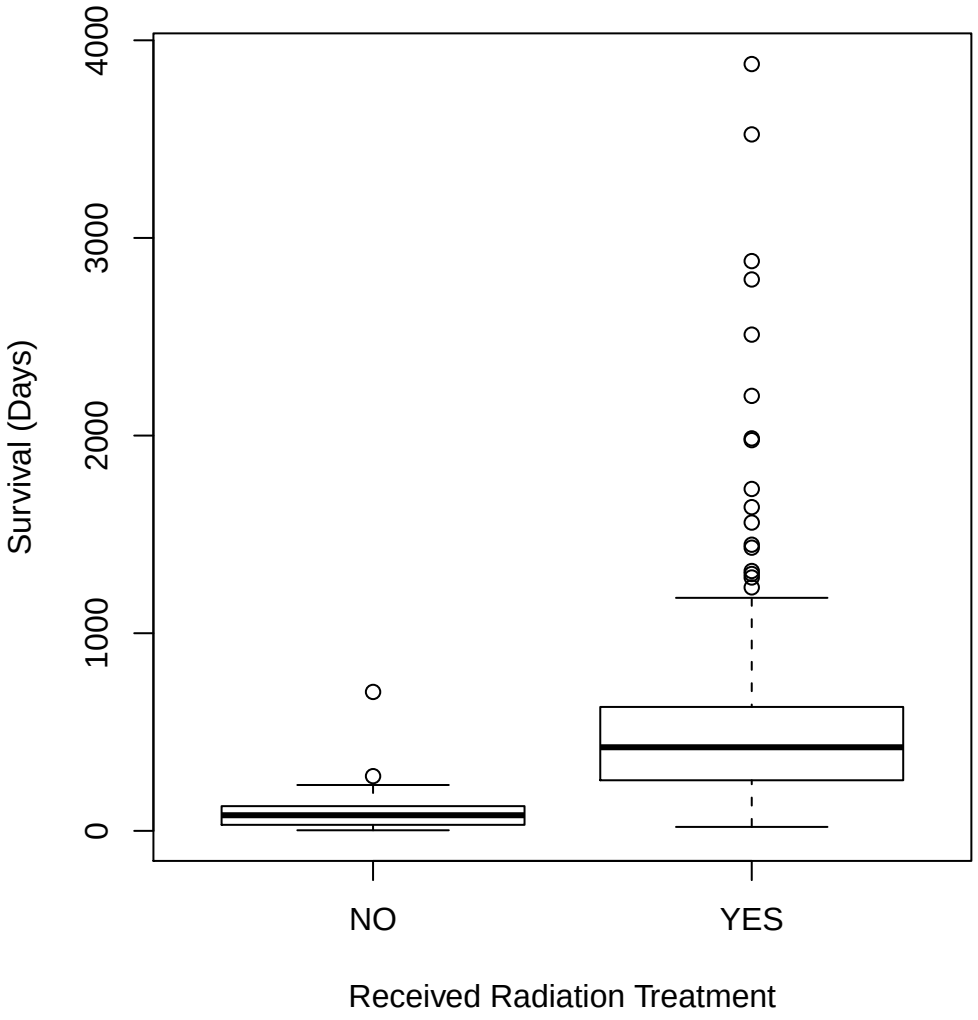

Figure 4.22. Overall survival for patients receiving radiation treatment versus patients not receiving radiation treatment.

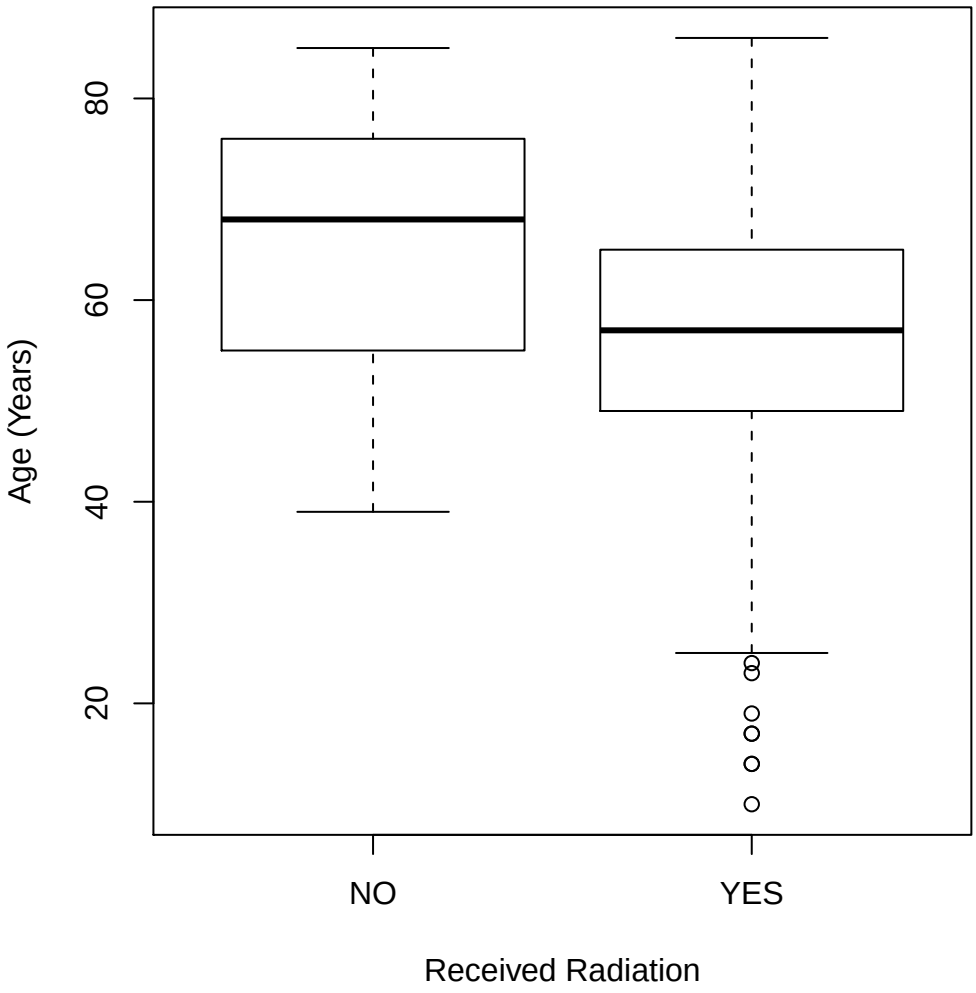

Figure 4.23. Radiation treatment status versus age at diagnosis.

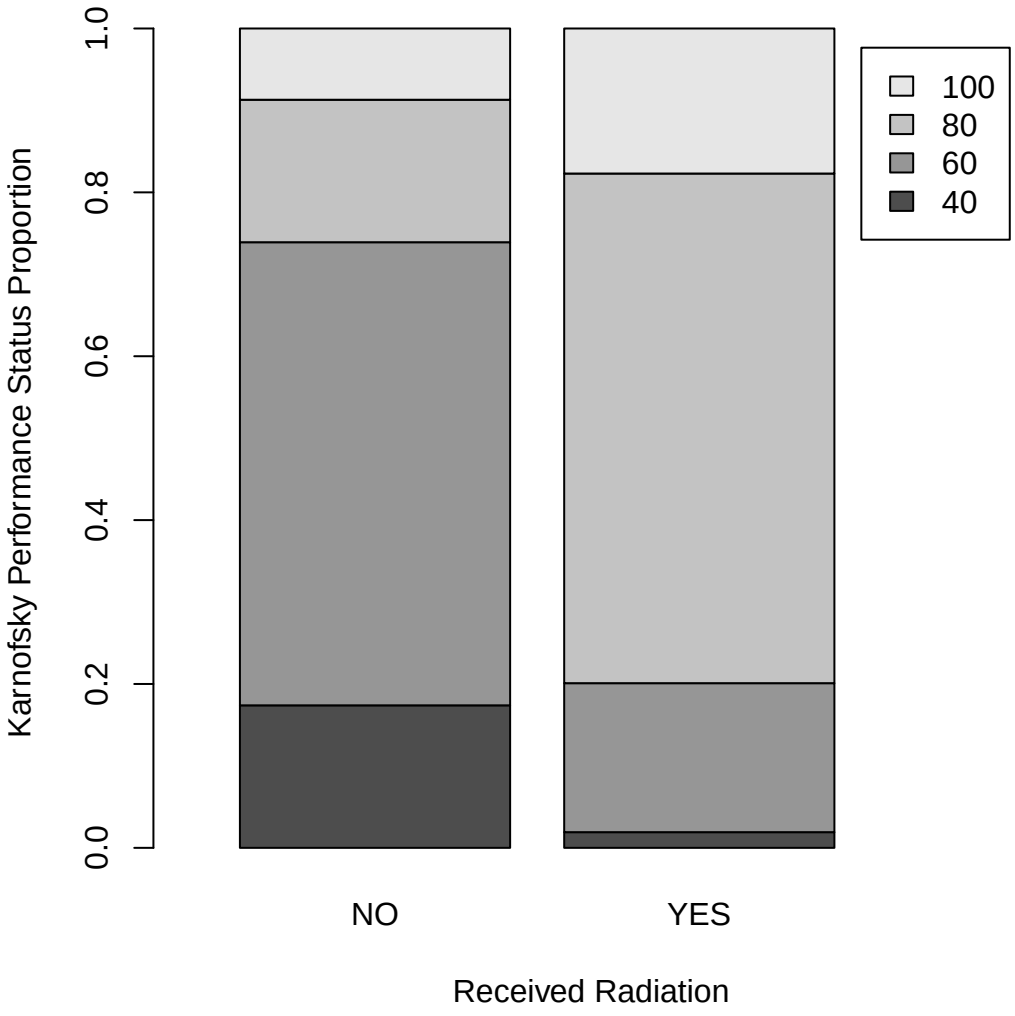

Figure 4.24. Radiation treatment status versus Karnofsky performance status (KPS).

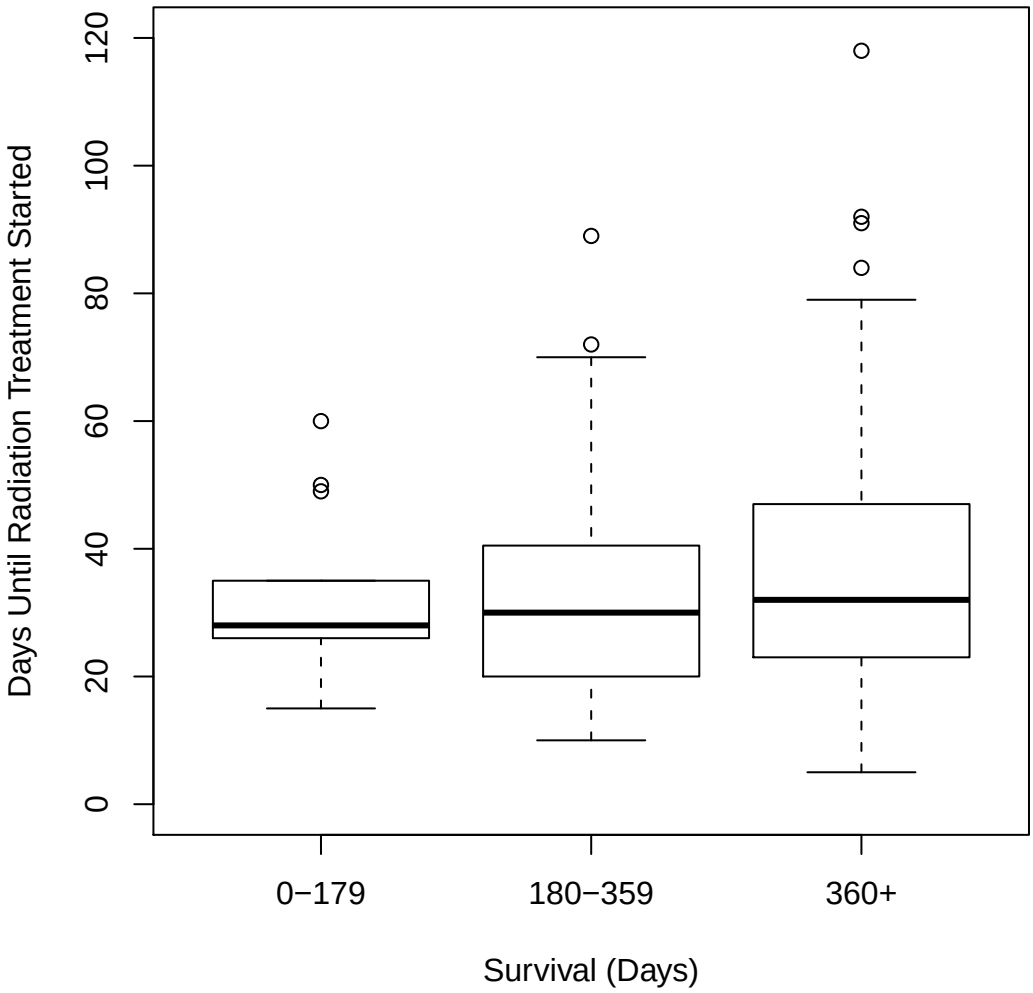

Figure 4.25. Number of days to radiation treatment versus patient overall survival.

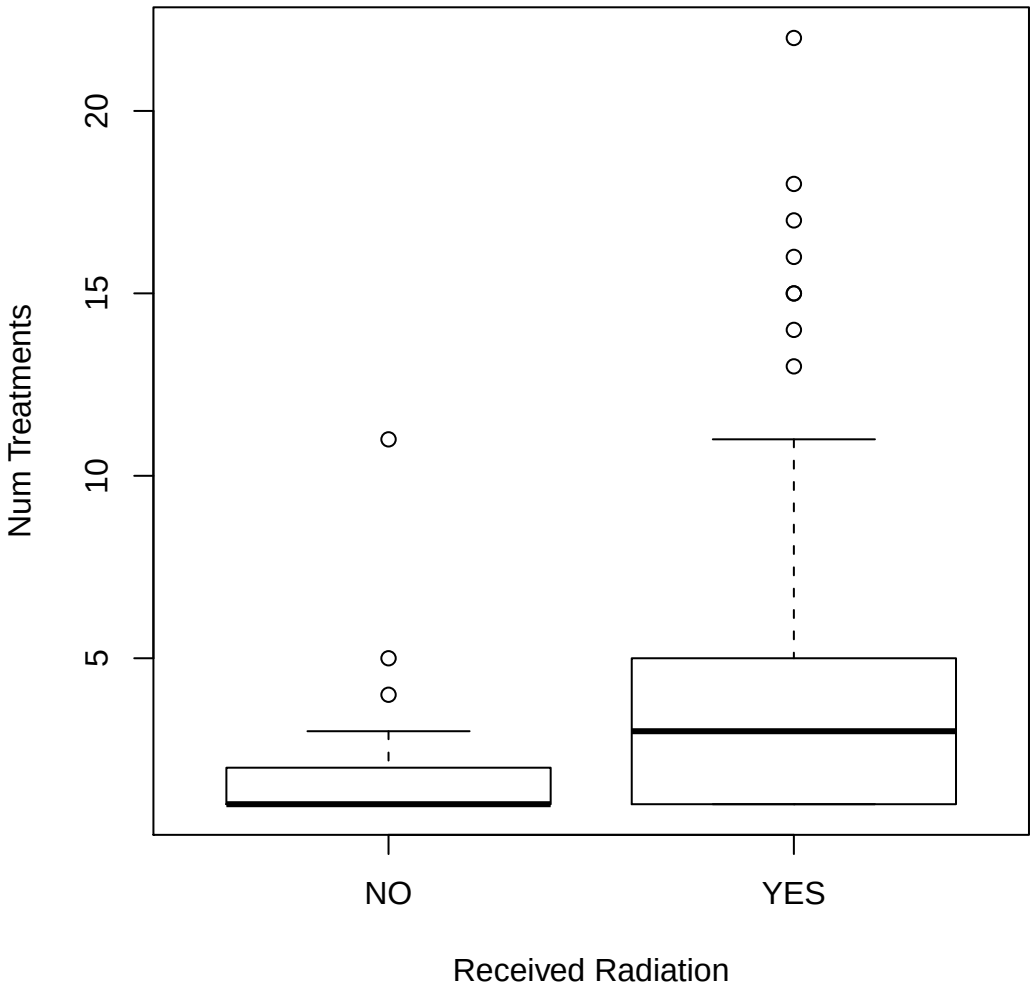

Figure 4.26. Overall number of treatments versus radiation treatment status.

Table 4.16. Cross-valiation results when all patients were included, the empirical survival split-point method was used to distinguish longer-term survivors from shorter-term survivors in each cross-validation fold, and ensemble-learning approaches were applied.

Ensemble Method	Error Rate	% STS Correct	% LTS Correct	AUC	Log-rank p-value
Majority Vote	0.137	0.000	1.000	0.847	N/A
Simple Weighted Vote	0.134	0.013	1.000	0.855	0.241
Squared-Weighted Vote	0.128	0.042	1.000	0.858	0.0377*
LTS Predictive Value Weighted Vote	0.137	0.000	1.000	0.851	N/A
STS Predictive Value Weighted Vote	0.115	0.246	0.975	0.856	0.000125*
Select Best	0.099	0.675	0.949	0.860	9.77e-07*
Mean Probability	0.137	0.000	1.000	0.849	N/A
Weighted Mean Probability	0.134	0.013	1.000	0.860	0.241
Stacked Generalization	0.121	0.567	0.931	0.749	2.33e-06*

Table 4.17. Cross-validation results when non-radiation-treated patients were excluded and median survival (423 days) was used as the split point between longer-term survivors and shorter-term survivors. The *RELIEF-F* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.398	0.542	0.662	0.602	0.000483*
	NBC	0.418	0.466	0.700	0.590	0.000581*
	SVM	0.467	0.473	0.592	0.541	0.157
Treatments	C5.0	0.333	0.771	0.562	0.666	4.27e-07*
	NBC	0.333	0.702	0.631	0.656	3.5e-06*
	SVM	0.318	0.756	0.608	0.640	3.97e-07*
Histology	C5.0	0.521	0.573	0.385	0.479	0.145
	NBC	0.475	0.374	0.677	0.511	0.411
	SVM	0.533	0.351	0.585	0.421	0.0891
DNA Methylation	C5.0	0.429	0.629	0.500	0.565	0.103
	NBC	0.325	0.888	0.419	0.657	1.32e-05*
	SVM	0.429	0.809	0.284	0.566	0.0785
Somatic Mutations	C5.0	0.459	0.880	0.188	0.534	0.278
	NBC	0.480	0.880	0.146	0.479	0.157
	SVM	0.449	0.560	0.542	0.534	0.126
DNA Copy Number	C5.0	0.492	1.000	0.000	0.500	N/A
	NBC	0.449	0.868	0.224	0.562	0.33
	SVM	0.516	0.209	0.768	0.480	0.759
mRNA Expression	C5.0	0.487	0.479	0.548	0.513	0.808
	NBC	0.457	0.546	0.539	0.553	0.0383*
	SVM	0.491	0.529	0.487	0.535	0.277
miRNA Expression	C5.0	0.536	0.475	0.452	0.463	0.819
	NBC	0.498	0.051	0.965	0.519	0.674
	SVM	0.494	0.373	0.643	0.519	0.637
All Data	C5.0	0.310	0.756	0.623	0.689	9.96e-09*
	NBC	0.333	0.817	0.515	0.687	4.66e-07*
	SVM	0.318	0.725	0.638	0.702	2.56e-07*

Table 4.18. Cross-validation results when non-radiation-treated patients were excluded and median survival (423 days) was used as the split point between longer-term survivors and shorter-term survivors. The *None* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.452	0.489	0.608	0.548	0.0039*
	NBC	0.410	0.580	0.600	0.626	4.25e-05*
	SVM	0.571	0.305	0.554	0.432	0.0145*
Treatments	C5.0	0.318	0.756	0.608	0.682	3.97e-07*
	NBC	0.333	0.702	0.631	0.656	3.5e-06*
	SVM	0.318	0.756	0.608	0.642	3.97e-07*
Histology	C5.0	0.533	0.641	0.292	0.467	0.0192*
	NBC	0.487	0.382	0.646	0.510	0.455
	SVM	0.525	0.359	0.592	0.445	0.86
DNA Methylation	C5.0	0.436	0.573	0.554	0.564	0.135
	NBC	0.380	0.843	0.351	0.628	0.00119*
	SVM	0.399	0.921	0.216	0.586	0.0022*
Somatic Mutations	C5.0	0.541	0.740	0.167	0.453	0.0708
	NBC	0.490	0.680	0.333	0.409	0.893
	SVM	0.459	0.300	0.792	0.543	0.404
DNA Copy Number	C5.0	0.504	0.915	0.064	0.489	0.689
	NBC	0.508	0.651	0.328	0.472	0.443
	SVM	0.508	0.116	0.880	0.500	0.705
mRNA Expression	C5.0	0.462	0.605	0.470	0.537	0.843
	NBC	0.449	0.639	0.461	0.537	0.084
	SVM	0.530	0.630	0.304	0.457	0.6
miRNA Expression	C5.0	0.511	0.466	0.513	0.490	0.395
	NBC	0.515	0.568	0.400	0.480	0.722
	SVM	0.528	0.492	0.452	0.475	0.0395*
All Data	C5.0	0.368	0.588	0.677	0.632	1.46e-05*
	NBC	0.475	0.641	0.408	0.547	0.0646
	SVM	0.475	0.412	0.638	0.519	0.032*

Table 4.19. Cross-validation results when non-radiation-treated patients were excluded and median survival (423 days) was used as the split point between longer-term survivors and shorter-term survivors. The *SVM-RFE* variable-selection approach was used.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Clinical	C5.0	0.456	0.565	0.523	0.544	0.00267*
	NBC	0.395	0.702	0.508	0.614	9.54e-06*
	SVM	0.433	0.519	0.615	0.559	0.00218*
Treatments	C5.0	0.318	0.756	0.608	0.682	3.97e-07*
	NBC	0.333	0.702	0.631	0.656	3.5e-06*
	SVM	0.318	0.756	0.608	0.641	3.97e-07*
Histology	C5.0	0.502	0.763	0.231	0.497	0.832
	NBC	0.494	0.344	0.669	0.492	0.726
	SVM	0.525	0.412	0.538	0.463	0.347
DNA Methylation	C5.0	0.485	0.629	0.378	0.504	0.96
	NBC	0.387	0.798	0.392	0.598	0.00311*
	SVM	0.417	0.764	0.365	0.590	0.0701
Somatic Mutations	C5.0	0.551	0.680	0.208	0.444	0.227
	NBC	0.490	0.800	0.208	0.424	0.53
	SVM	0.480	0.220	0.833	0.462	0.682
DNA Copy Number	C5.0	0.500	0.977	0.008	0.492	0.319
	NBC	0.484	0.550	0.480	0.493	0.358
	SVM	0.476	0.233	0.824	0.483	0.79
mRNA Expression	C5.0	0.513	0.555	0.417	0.486	0.932
	NBC	0.483	0.504	0.530	0.518	0.579
	SVM	0.517	0.454	0.513	0.502	0.551
miRNA Expression	C5.0	0.511	0.568	0.409	0.488	0.709
	NBC	0.481	0.898	0.130	0.510	0.363
	SVM	0.558	0.398	0.487	0.403	0.248
All Data	C5.0	0.349	0.618	0.685	0.651	0.000187*
	NBC	0.398	0.641	0.562	0.618	0.0142*
	SVM	0.352	0.702	0.592	0.656	7.31e-06*

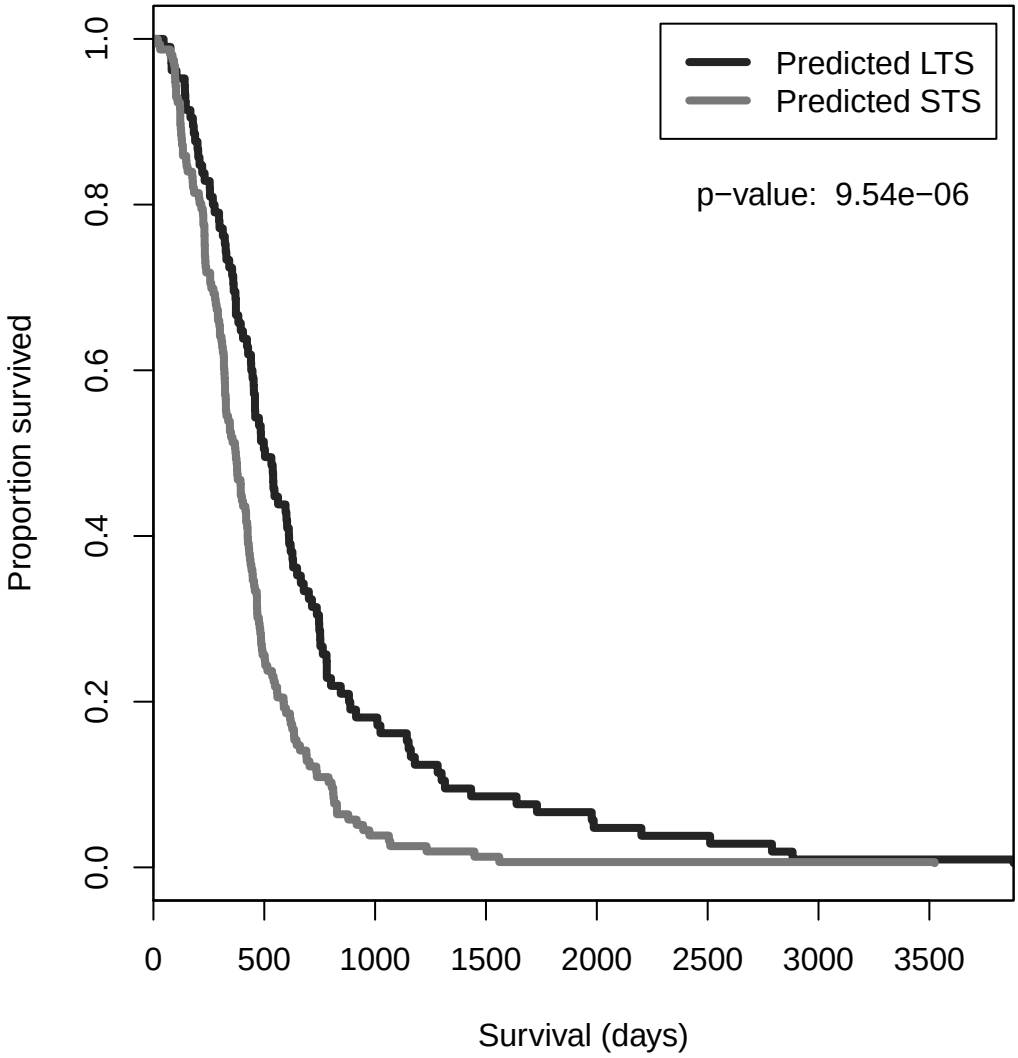

Figure 4.27. Kaplan-Meier curves comparing overall survival of patients predicted as longer-term survivor (LTS) versus shorter-term survivor (STS) when the NBC algorithm was applied to clinical data. Support Vector Machines-Recursive Feature Elimination was used for variable selection, non-radiation-treated patients were excluded, and median survival was the split point between LTS and STS.

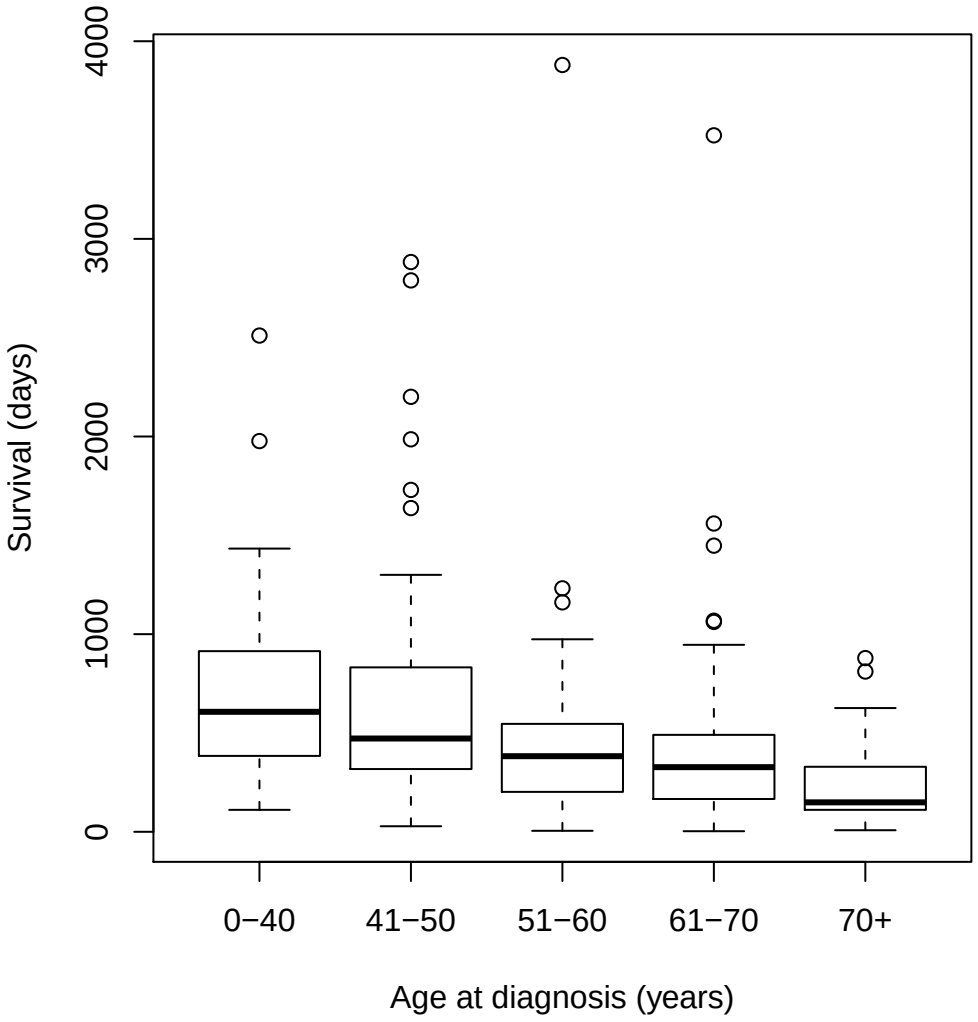

Figure 4.28. Patient overall survival versus age at pathologic diagnosis.

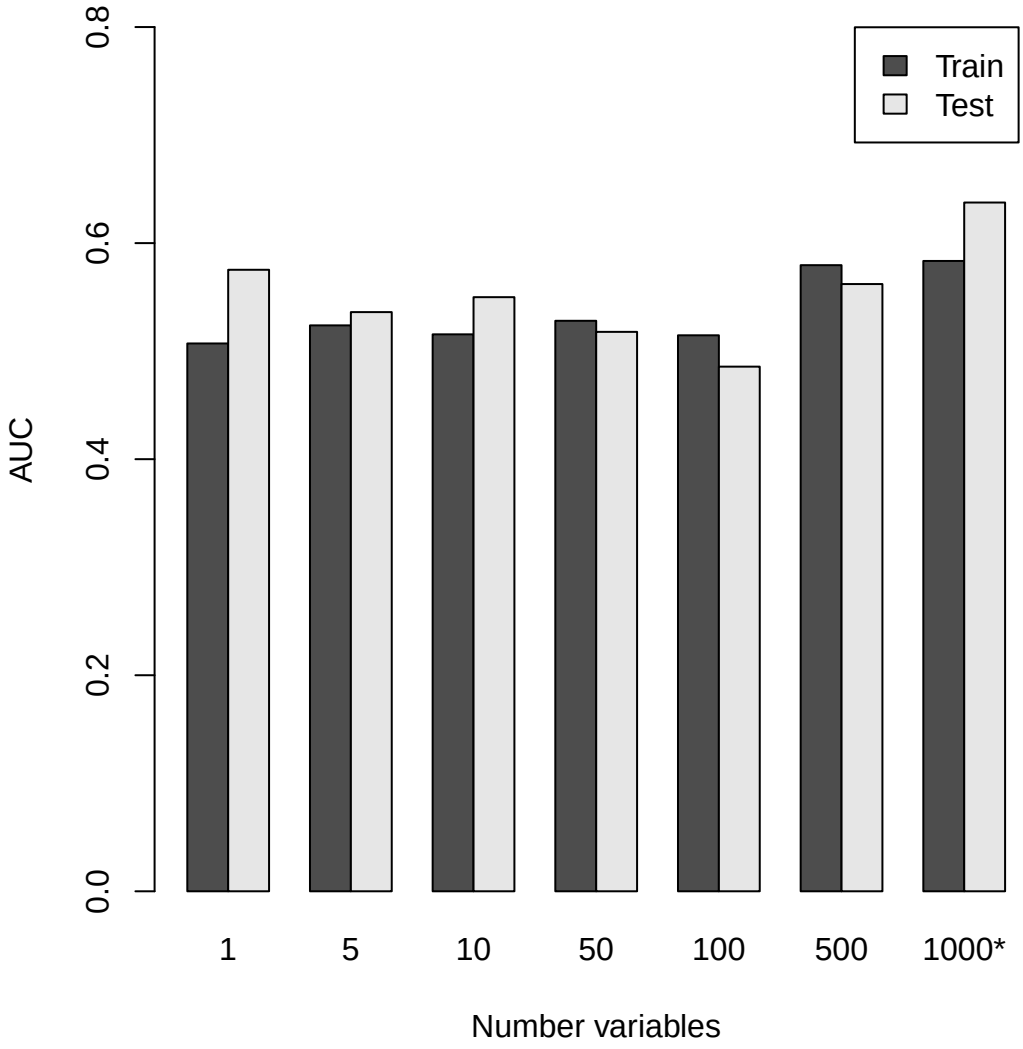

Figure 4.29. Area under receiver operating characteristic curve versus number of DNA methylation genes included in Naïve Bayes Classifier models. Median survival was the split point between longer-term survivors and shorter-term survivors, and variables were ranked using Support Vector Machines-Recursive Feature Elimination.

Table 4.20. Cross-validation results when age at diagnosis and DNA methylation data were used as data categories. No variable selection was applied, non-radiation-treated patients were excluded, and median survival was the split point between longer-term survivors and shorter-term survivors. Only patients with data for both data categories were included.

Data Category	Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Age	C5.0	0.442	0.416	0.730	0.573	0.0145*
	NBC	0.411	0.843	0.284	0.634	0.0436*
	SVM	0.460	0.910	0.095	0.492	0.609
DNA Methylation	C5.0	0.485	0.618	0.392	0.505	0.754
	NBC	0.399	0.843	0.311	0.576	0.000354*
	SVM	0.448	0.876	0.162	0.523	0.0208*
All Data	C5.0	0.485	0.618	0.392	0.505	0.754
	NBC	0.399	0.843	0.311	0.577	0.000354*
	SVM	0.405	0.820	0.324	0.624	0.00657*

Table 4.21. Cross-validation results when ensemble-learning approaches were applied to age data and DNA methylation data. Non-radiation-treated patients were excluded, and median survival was the split point between longer-term survivors and shorter-term survivors. Only patients with data for both data categories were included.

Ensemble Method	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Majority Vote	0.436	0.876	0.189	0.593	0.11
Simple Weighted Vote	0.436	0.843	0.230	0.576	0.03*
Squared-Weighted Vote	0.436	0.843	0.230	0.575	0.03*
LTS Predictive Value Weighted Vote	0.429	0.843	0.243	0.585	0.0144*
STS Predictive Value Weighted Vote	0.423	0.843	0.257	0.566	0.0244*
Select Best	0.460	0.798	0.230	0.530	0.304
Mean Probability	0.436	0.685	0.419	0.611	0.0174*
Weighted Mean Probability	0.436	0.685	0.419	0.592	0.0174*
Stacked Generalization	0.479	0.764	0.230	0.497	0.755

Table 4.22. Cross-validation results non-radiation-treated patients were excluded, median survival (423 days) was the split point between longer-term survivors and shorter-term survivors, and ensemble-learning approaches were applied.

Ensemble Method	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Majority Vote	0.364	0.725	0.546	0.696	1.69e-05*
Simple Weighted Vote	0.349	0.733	0.569	0.703	1.03e-06*
Squared-Weighted Vote	0.326	0.740	0.608	0.706	7.29e-08*
LTS Predictive Value Weighted Vote	0.337	0.725	0.600	0.704	2.63e-07*
STS Predictive Value Weighted Vote	0.352	0.718	0.577	0.703	1.29e-06*
Select Best	0.318	0.733	0.631	0.682	3.65e-07*
Mean Probability	0.352	0.863	0.431	0.689	9.29e-07*
Weighted Mean Probability	0.356	0.840	0.446	0.697	3.9e-06*
Stacked Generalization	0.460	0.588	0.492	0.540	0.0978

Table 4.23. Cross-validation results when non-radiation-treated patients were excluded and median survival (423 days) was used as the split point between longer-term survivors and shorter-term survivors. In this experiment, all data categories except *Treatments* were combined into a single data set.

Variable Selection Approach	Classification Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
None	C5.0	0.475	0.550	0.500	0.525	0.665
None	NBC	0.479	0.641	0.400	0.546	0.0638
None	SVM	0.418	0.481	0.685	0.594	7.94e-05*
SVM-RFE	C5.0	0.475	0.573	0.477	0.525	0.0988
SVM-RFE	NBC	0.429	0.679	0.462	0.580	0.0119*
SVM-RFE	SVM	0.448	0.542	0.562	0.565	0.0106*
RELIEF-F	C5.0	0.498	0.679	0.323	0.501	0.379
RELIEF-F	NBC	0.441	0.702	0.415	0.586	0.0957
RELIEF-F	SVM	0.452	0.527	0.569	0.556	0.139

Table 4.24. Cross-validation results when non-radiation-treated patients were excluded, median survival (423 days) was used as the split point between longer-term survivors and shorter-term survivors, and ensemble-learning approaches were applied. All data categories except *Treatments* were used.

Ensemble Method	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Majority Vote	0.437	0.687	0.438	0.584	0.00222*
Simple Weighted Vote	0.433	0.679	0.454	0.586	0.00238*
Squared-Weighted Vote	0.437	0.679	0.446	0.590	0.00111*
LTS Predictive Value Weighted Vote	0.444	0.649	0.462	0.586	0.0113*
STS Predictive Value Weighted Vote	0.444	0.672	0.438	0.587	0.00372*
Select Best	0.437	0.649	0.477	0.578	0.00411*
Mean Probability	0.444	0.855	0.254	0.583	0.0224*
Weighted Mean Probability	0.433	0.840	0.292	0.584	0.00572*
Stacked Generalization	0.494	0.473	0.538	0.506	0.298

Table 4.25. Cross-validation results when only patients who received radiation and temozolomide treatment were included. Median survival (423 days) was used as the split point between longer-term survivors and shorter-term survivors. All data categories except *Treatments* were combined into a single data set.

Variable Selection Approach	Classification Algorithm	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
None	C5.0	0.530	0.448	0.493	0.470	0.352
None	NBC	0.515	0.284	0.687	0.503	0.987
None	SVM	0.530	0.403	0.537	0.419	0.159
SVM-RFE	C5.0	0.470	0.537	0.522	0.530	0.297
SVM-RFE	NBC	0.507	0.522	0.463	0.515	0.704
SVM-RFE	SVM	0.575	0.418	0.433	0.430	0.18
RELIEF-F	C5.0	0.507	0.627	0.358	0.493	0.991
RELIEF-F	NBC	0.552	0.463	0.433	0.460	0.267
RELIEF-F	SVM	0.507	0.448	0.537	0.457	0.235

Table 4.26. Cross-validation results when only patients who received radiation and temozolomide treatment were included and ensemble-learning approaches were applied. Median survival (423 days) was the split point between longer-term survivors and shorter-term survivors. All data categories except *Treatments* were used.

Ensemble Method	Error Rate	STS Correct	LTS Correct	AUC	Log-rank p-value
Majority Vote	0.515	0.284	0.687	0.485	0.454
Simple Weighted Vote	0.530	0.299	0.642	0.479	0.241
Squared-Weighted Vote	0.537	0.299	0.627	0.477	0.246
LTS Predictive Value Weighted Vote	0.537	0.284	0.642	0.478	0.19
STS Predictive Value Weighted Vote	0.537	0.299	0.627	0.485	0.199
Select Best	0.493	0.552	0.463	0.537	0.727
Mean Probability	0.515	0.239	0.731	0.495	0.637
Weighted Mean Probability	0.515	0.269	0.701	0.490	0.366
Stacked Generalization	0.515	0.448	0.522	0.485	0.141

Table 4.27. Results of standard (noncorrected) GSEA analysis applied to top-1000 ranked DNA methylation genes by the RELIEF-F algorithm. Patients receiving no radiation treatment were excluded, and median survival was used as the split point. The KEGG pathways most highly enriched for the genes are displayed.

KEGG Pathway	p-value
Pathways in cancer	2.87e-15
TGF-beta signaling pathway	3.63e-09
Bladder cancer	2.65e-07
Focal adhesion	3.14e-07
Pancreatic cancer	2.2e-06
Chagas disease	3.61e-06
Chronic myeloid leukemia	4.1e-06
Prostate cancer	4.46e-06
MAPK signaling pathway	5.35e-06
Amyotrophic lateral sclerosis (ALS)	6.11e-06
Toll-like receptor signaling pathway	9.85e-06
Melanoma	1.24e-05
ErbB signaling pathway	1.27e-05
Amoebiasis	1.55e-05
Malaria	2.01e-05
Colorectal cancer	4.25e-05
Fc epsilon RI signaling pathway	5.09e-05
Cytokine-cytokine receptor interaction	7e-05
Leukocyte transendothelial migration	8.94e-05
Axon guidance	0.000109

Table 4.28. Results of standard (noncorrected) GSEA analysis applied to all methylation genes (when patients receiving no radiation treatment were excluded and median survival was used as the split point).

KEGG Pathway	p-value
Pathways in cancer	2.05e-39
Melanoma	4.24e-13
Focal adhesion	5.2e-12
Leishmaniasis	4.07e-11
Prostate cancer	3.33e-10
Bladder cancer	1.08e-09
p53 signaling pathway	4.53e-09
Cytokine-cytokine receptor interaction	5.25e-09
TGF-beta signaling pathway	5.87e-09
Chagas disease	1.47e-08
Axon guidance	3.88e-08
Malaria	1.3e-07
Amoebiasis	2.29e-07
Chronic myeloid leukemia	3.73e-07
MAPK signaling pathway	4.18e-07
Basal cell carcinoma	6.88e-07
Leukocyte transendothelial migration	1.92e-06
Pancreatic cancer	2e-06
Colorectal cancer	2.11e-06
Hematopoietic cell lineage	2.68e-06

Table 4.29. Results of permutation-corrected GSEA analysis applied to top-1000 ranked DNA methylation genes by the RELIEF-F algorithm (when patients receiving no radiation treatment were excluded and median survival was used as the split point).

KEGG Pathway	p-value
Olfactory transduction	0.001
Amyotrophic lateral sclerosis (ALS)	0.004
TGF-beta signaling pathway	0.005
Toll-like receptor signaling pathway	0.008
NOD-like receptor signaling pathway	0.011
Tyrosine metabolism	0.018
Huntington's disease	0.019
ErbB signaling pathway	0.024
Pancreatic cancer	0.027
Fc epsilon RI signaling pathway	0.035
Progesterone-mediated oocyte maturation	0.036
Glycine, serine and threonine metabolism	0.038
Chronic myeloid leukemia	0.039
PPAR signaling pathway	0.04
Glycosaminoglycan biosynthesis - heparan sulfate	0.049

CHAPTER 5

DISCUSSION

5.1 General Observations

Because GBM is a complex disease for which patient survival time is influenced by a variety of heterogeneous factors, one objective of this study was to assess how well multivariate prediction algorithms could differentiate between GBM patients who would survive a relatively long or short time after diagnosis. Using ML-Flex, the custom software package developed for this study, various algorithms were applied to retrospective GBM data from TCGA. In many cases, the algorithms identified subsets of patients that experienced significantly longer or shorter survival than the remaining patients.

A desirable outcome of this study might have been that one particular algorithm or data category performed well in all cases and thus could have been favored for further development of GBM prognosis models. However, in this study, classification performance varied substantially across algorithms and data categories, even though some data categories appeared to be more informative than others. Across the various experiments, better performance was typically observed when variable selection was performed than when all variables were included in the models; this finding coincides with the expectation that many variables have little or no ability to differentiate between LTS and STS. Among classification algorithms, NBC performed quite well and often (at least slightly) better than the other algorithms. However, in many cases, C5.0 Decision Trees and SVM performed (at least slightly) better than NBC. Such variability is not a surprise considering the diversity of the algorithmic approaches—a given algorithm may be suited well to a particular type of data and not to others. Across the experiments, the data categories that appeared to contain most prognostic relevance were clinical, treatments, and DNA methylation; in some experiments, models based on the remaining data categories—histology, somatic mutations, DNA

copy number, mRNA expression, and miRNA expression—performed moderately well, but the algorithms had only marginal success with these categories overall. (Potential reasons for the relatively poor performance of these categories are described in the Limitations section below.)

Rather than compare performance extensively across algorithms or data categories, this study employed ensemble-learning approaches in an attempt to aggregate evidence across multiple algorithms and data categories. In a few cases, the ensemble methods outperformed all individual algorithms. But in general, the ensemble methods' main advantage was their consistency. While performance varied considerably across algorithms and data categories, the ensemble methods often approximated the best individual performers. In some cases, the ensemble methods performed well because they identified and favored one individual algorithm/category combination that generalized well; in other cases, the ensemble methods appear to have extracted complementary evidence from several algorithms/categories. These findings suggest that in the absence of a priori knowledge about which algorithm or data category would be most useful for making survival predictions, ensemble methods may be the most effective choice.

Limited predictive performance was observed for multivariate models based on previously reported GBM prognostic variables when the survival threshold was two years (the threshold used most frequently in the literature). Subsequent experiments revealed that better performance could be attained when the survival threshold was different from two years (and when the variables are not limited by prior knowledge). At least two factors likely influenced these performance differences: 1) prior knowledge about factors influencing GBM prognosis is incomplete, 2) the choice of survival threshold influences predictive performance strongly due to class imbalances.

In evaluating the biological relevance of methylation genes that helped differentiate between LTS and STS, GSEA analyses revealed a strong bias in favor of genes from pathways already believed to affect tumorigenesis. The permutation method attempted to correct for this bias and generated hypotheses about pathways that may drive GBM tumor aggressiveness but may have been overlooked in this context. For the best-performing DNA methylation models, some hypotheses pointed to the body's

innate immune response as a prognostically relevant biological mechanism for GBM. Researchers are actively studying the immune system's role in suppressing tumor growth, and immunotherapies are being investigated as a means to help the immune system target tumors. In fact, at least one such therapy has already shown promise for GBM. In a series of clinical trials conducted at Duke University, patients with an in-frame deletion in the EGFR gene (EGFRvIII) were vaccinated with peptide-pulsed dendritic cells; after treatment, the patients experienced antitumor immune responses with no serious adverse events, and some have had remarkable recoveries. [78]

In this study, three metrics—error rate, AUC, and log-rank statistic—were used to assess model performance. The error rate is frequently reported in machine-learning studies due to its simplicity of calculation and interpretation. However, care must be taken to interpret the error rate when the class distribution is imbalanced—an important consideration in this study. Although the AUC's method of calculation is relatively complex—and thus less intuitive—the AUC's interpretation remains consistent for any class distribution. Additionally, the AUC accounts not only for discrete classes predicted by an algorithm but also for the confidence with which those predictions are made. (In this study, the AUC estimates the likelihood that a classifier assigned a randomly chosen STS a higher probability of being STS than a randomly chosen LTS.) This property also suits the AUC well for assessing model performance in inner cross-validation folds and for serving as a weight in ensemble-learning methods. On the other hand, a disadvantage of the AUC is that it depends on the quality of the algorithms' posterior probabilities; some algorithms, such as C5.0 Decision Trees, are designed only to give discrete probabilities. An important advantage of the log-rank statistic is its familiarity to clinicians. The log-rank p-value provides an intuitive—though arbitrarily determined—assessment of a given model's quality via "statistical significance" thresholds. However, a potential disadvantage of the log-rank statistic is that it weighs outliers heavily. In sum, none of these metrics is adequate in isolation, and sometimes they conflict with each other. Perhaps the best way to assess model performance is to focus on models that perform well according to all three metrics yet to keep an eye open for special cases that otherwise demonstrate clinical relevance (such as the IDH1/TP53 somatic-mutation models).

Finally, it should be noted that even though multivariate models in this study predicted survival status poorly for many data categories, it should not be inferred that effective models are impossible—or even unlikely—to be developed for those data categories. Such inferences could become Type II errors (false negatives) as methods of developing such models are refined. An important goal of this study was to apply commonly used, general-purpose algorithms to the TCGA data for GBM and gain a sense for the performance levels that can be attained when such algorithms are tested in a rigorous and consistent fashion. Other algorithms that are more sophisticated or better suited to the poor-performing data categories may prove valuable in future studies.

5.2 Major Contributions

In 2007, Dupuy and Simon surveyed the literature for studies that had examined a relationship between high-throughput molecular data and cancer outcomes. [79] For 21 of 42 studies that were examined in detail, the authors expressed concern over methodological flaws in the studies' experimental designs. For example, some cross-validated studies performed variable selection on the entire data set rather than within each training set, an approach that can bias results tremendously. Methodological concerns and differences between studies may arise partly because independent labs each develop custom software to perform their analyses, yet considerable time and software-engineering effort is required to ensure validity. ML-Flex was developed as a means to address these gaps by performing cross-validation studies in a rigorous, consistent, and repeatable fashion. Additionally, because repositories like TCGA contain gigabytes of data, the time required to construct multivariate models is an important consideration. Despite the growing availability of high-performance computational resources, most software is not designed to take advantage of multiple central processing units (CPUs) within each server or multiple processing cores within each CPU. Thus in the development of ML-Flex, an additional effort was made to ensure experiments could be executed in a time-efficient manner. ML-Flex uses the threading capability within the Java virtual machine to capitalize on such resources. No software package with this architecture is freely available to researchers at this

time. Thus plans are in place to share ML-Flex with the broader research community in hopes that it will become a standard tool for performing large-scale, cross-validated studies. The extensible nature of ML-Flex should enable developers to customize its use to a broad range of research purposes.

Previous studies of GBM prognosis typically have focused on one or two categories of patient data. Contrarily, in this study, eight categories of patient data have been tested side by side in a comparative assessment of prognostic relevance. Although care must be taken to consider the limitations of such an analysis (especially the possibility of false-negative results), comparative assessments across data categories may help guide investment of financial resources for future analyses. In particular, researchers and funding agencies must weigh the incremental costs of acquiring multiple categories of high-throughput molecular data for each patient.

Although other research groups are analyzing the GBM data in TCGA, no other study has reported the use of supervised-learning algorithms to analyze this data set in relation to prognosis. Thus this study has the potential to pave the way for future supervised-learning studies in TCGA (and elsewhere).

Because various types of molecular aberrations can influence tumor growth, a key goal of the TCGA Consortium is to facilitate development of methods that integrate data across modalities. This study has demonstrated two approaches for performing integrative analyses: 1) combining all data into an aggregate data set and allowing multivariate algorithms to model intercategory relationships at a granular level, and 2) making multivariate predictions for each data category separately and then combining predictions at a coarse level using ensemble-learning approaches. Although neither approach resulted in predictive performance that was consistently better than the best single-category models, the integrative methods used in this study demonstrate alternatives for researchers to consider as multimodal data sets become more common. It may be that such methods perform considerably better when applied to other diseases or when the number of component algorithms is increased.

In this study, algorithmic refinements have also been explored. Firstly, the Edgerton, et al. method of selecting survival split points was modified to use the AUC rather than the chance-corrected error rate. As described in the Results section, a simulation

study demonstrates that this modification should result in better precision when the split point is away from the median. Secondly, three novel, weight-based ensemble methods were developed and applied. While the performance of these methods did not always exceed previously developed methods, they consistently performed better than *majority vote* and *simple weighted vote* and thus show promise for further refinement.

5.3 Limitations

Because biomedical informatics is a relatively young field—especially in the realm of genetics/genomics research—detailed protocols do not exist for performing experiments such as those in this study. General guidelines must be followed to ensure statistical and methodological rigor; however, many seemingly minor decisions—each of which could have a drastic impact on the results—must be made. In this study, an attempt has been made to default such decisions toward the simplest approach. However, arbitrary decisions sometimes had to be made. This section describes such issues as well as extraneous factors that may have impacted performance. It is hoped that this study will serve as a starting point for future research by elucidating methodological factors that must be considered in performing such a study.

The TCGA data are being provided by several research centers throughout the United States. While special efforts are being made to ensure consistency in specimen handling, tumor-sample quality, and clinical data definitions, the TCGA data by nature are heterogeneous and likely to contain noise that could impact predictive performance. Acknowledging this limitation, an added measure of confidence can be placed in models that do perform well on this data set. Additionally, because the data come from different institutions and are handled by a variety of people, data quality can be affected. Thus manual examination of the data and some subjective interpretation were necessary for this study. For example, the drug-treatment data often contained multiple spellings (including mispellings) for a given drug, and sometimes a drug's commercial name was listed instead of its generic name. These inconsistencies were manually corrected before executing the experiments in this study.

Ideally, TCGA would contain data only for GBM patients who had been treated uniformly (as in clinical trials). Such a design would enable assessments of the

effectiveness of specific treatments and of cofactors that may influence treatment responses. However, until recently, no drug treatment had been shown to improve GBM survival in a phase III clinical trial [2]; thus the standard of care has not included specific drug treatments until recently. Consequently, a wide variety of drug treatments (and treatment regimens) have been administered to TCGA patients. The various treatments likely affected patient survival to differing degrees—as already noted, the total number of drug treatments received is associated with patient survival—so treatment status may have a confounding effect on other factors that affect survival. Thus even though this study attempted to address a clinician's need to prospectively estimate patient prognosis at the time of diagnosis—treatment regimens are not always known at the time of diagnosis—treatment data were included in this study to estimate the effects that treatments have on survival. Various alternatives could be employed to deal with the treatment-data confounding effects; for example, 1) analyses may be limited to patients who received specific treatments, 2) treatment data may be excluded from analyses, and 3) predictions based on treatment data may be used as covariates in multivariate survival analyses. The first two of these approaches were employed in TCGA Experiment 5, and the results were mixed. Filtering patients by treatments may result in subpopulations that are inadequately sized to derive generalizable models or attain statistical significance—especially considering the fact that sample sizes in TCGA are limited to begin with. Including treatment-based predictions as covariates in multivariate analyses would provide an estimate of the prognostic value offered by a given data category and algorithm, independent of treatment status; however, this estimate would provide little additional insight due to the heterogeneous nature of the treatment data.

Two variables that are not recorded in TCGA but that may have offered valuable insights are tissue anatomic site and year of diagnosis. In the Lamborn, et al. study, [11] tissue anatomic site showed promise as a prognostic factor; it is plausible that GBM tumors differ in their aggressiveness, operability, and in the effects they have on cognitive function, depending on the location of the brain from which the tumor arises. An evaluation of the relationship between year of diagnosis and patient survival may have offered insights into confounding effects that may result from changes in

treatments, surgical techniques, methods of tumor-sample preservation, etc. that have occurred across the span of years during which the TCGA patients were diagnosed and treated.

One decision that may have impacted the results considerably is the choice of data normalization and summarization techniques. For example, when multiple histology values were provided (e.g., for the top and bottom of slides or for multiple samples), the mean value was used; however, it may be that the maximum or minimum value would better represent these features. Additionally, treatment data were transformed into binary values, even though patients who received multiple doses of a given drug may have gained more benefit than patients who received a single dose. Additionally, all biomolecular data were summarized according to higher-level functional categories. For example, somatic-mutation, DNA methylation, and mRNA expression data were summarized at the gene level; DNA copy-number data were summarized by chromosomal band; and miRNA data were summarized by known gene targets. Although summarizing raw data may sometimes reduce signal, summarization methods enable easier interpretation, reduce computational demands, and may represent underlying biological mechanisms better than raw data. An interesting area for future research would be to assess the effects of various data-summarization approaches on downstream classification performance.

Transforming time-to-event data into a binary outcome (e.g., STS, LTS) can result in a loss of information [79]; however, to be consistent with prior studies that have attempted to separate patients into discrete groups, [1, 19, 20] survival was also discretized in this study. Discretizing survival can also introduce bias if the two groups have different censorship structures [79]; however, such bias was avoided in this study by excluding patients who were still alive (n = 74) when the analysis was performed. Although this exclusion affected sample size only moderately in this study, such an approach may not be acceptable when studying other cancer types for which survival is generally longer. Consequently, multivariate prediction algorithms that retain survival as a continuous variable and account for censorship status (thus allowing living patients to be included) may be more suitable for general application. Alternatively, living patients who have survived longer than the discretization thresh-

old can be included in analyses and labeled LTS; however, to maintain consistency across experiments, that approach was not used in this study.

Instead of discretizing survival using either an abritrary threshold or the empirical split-point method, an alternative approach would be to examine the survival distribution visually and seek to identify natural groupings that have occurred. Figure 5.1 shows the overall distribution of survival times for GBM patients used in this study. No clear bimodal distribution exists in the data; however, noticeable irregularities in the distribution exist around 100 days, 500 days, and 800 days. The irregularity at 100-days survival likely represents (at least in part) the differences in survival between patients who received radiation treatment or not. The irregularities near 500 days and 800 days have no straightforward explanation and may represent thresholds dividing groups of patients whose tumors have different biological underpinnings. One way to formalize this approach would be to use a technique like k-means clustering to identify groups within the distribution quantitatively. One challenge with using this approach is that sample sizes in TCGA are limited. Thus using a population-level database of GBM patient survival—such as what may be found in public cancer registries—could be useful for deriving general-purpose thresholds.

Across the experiments performed in this study, the same data set was evaluated multiple times via cross validation. Although cross validation maintained statistical rigor for each experiment, the possibility exists that enough attempts to attain statistical significance would eventually result in log-rank p-values that fall below the alpha threshold (i.e., $p < 0.05$) by random chance. Researchers have developed many approaches to account for this so-called "multiple-testing bias"; however, no approach is standardly applied in this setting. Perhaps the most conservative correction for multiple tests is the Bonferroni approach, which divides the alpha threshold by the total number of tests; p-values lower than the corrected threshold then are considered significant. Accounting for all combinations of variable-selection approach, classification algorithm, and data category that were tested, plus the various ensemble approaches that were tested, the total number of tests across all experiments was 342. Thus according to the Bonferroni correction, log-rank p-values would need to be lower than 0.0001462 to be considered significant. Several results from this study

attained this level of significance; however, the Bonferroni correction, which assumes independence between tests, is overly conservative in this setting because many of the tests were highly interrelated. A reasonable, though less rigorous, approach to accounting for multiple tests across experiments is simply to place most emphasis on the data categories that perform consistently well. For example, models based on clinical, treatment, and DNA methylation data performed well for multiple algorithms and in multiple experiments, despite differences in survival threshold. Models based on mRNA expression also performed well, though their performance was less robust across algorithms and experiments.

One other way to address multiple-testing concerns is to evaluate prognostic models on an (or preferably multiple) external data set. If a model performs well on independent sets of patients, confidence in the model increases substantially. However, TCGA is unique in its breadth of data, making it infeasible at this time for individual labs to perform external validation. Consequently, this study has simulated external validation with a cross-validation approach. However, as the TCGA Consortium continues to collect data for additional GBM patients, the results of this study can be validated on a separate set of GBM patients.

5.4 Opportunities for Future Work

The multivariate algorithms employed in this study can be configured using various parameters. For simplicity, default parameters were selected in most cases. However, performance can vary dramatically as such parameters are modified. Thus it is possible that parameters other than those used in this study could result in improve performance. One way to estimate optimal parameters for a given classification task is to experiment with various settings in internal cross-validation folds and select settings that perform best internally. A slightly different but related technique, which could be employed using ML-Flex's current design, is to treat different parameter configurations as separate algorithms and use ensemble-learning approaches to combine evidence across the various configurations. If a single configuration performed exceptionally better than other configurations in internal folds, *Select Best* should account for it; otherwise, the collective wisdom produced by the various configurations

may prove beneficial.

TCGA contains various categories of biomolecular data. For some data categories (e.g., DNA copy number, DNA methylation, mRNA expression), data have been profiled using multiple high-throughput technology platforms. For simplicity of computation and interpretation, only a single platform was examined in this study for each data category. However, ML-Flex's extensible design makes it possible to process data from any platform that produces text-based output. Two interesting avenues for future research would be 1) to compare predictive performance across platforms and 2) to combine evidence across platforms via ensemble approaches. For the latter, an additional refinement could be to develop a hierarchical model in which ensemble methods derive an aggregate prediction for each data category, and then aggregate predictions are combined into an overall prediction.

As already noted, class imbalance can affect classification performance because many algorithms are designed for scenarios where the classes are balanced. A recent study observed that class imbalance especially affects high-dimensional data sets, unless a very strong signal exists in the data. [80] In this study, class imbalance often appeared to affect ensemble-learning approaches even more markedly than individual algorithms. For this study, two novel ensemble-learning methods—*STS predictive-value weighted vote* and *LTS predictive-value weighted vote*—were developed in an attempt to counter class-imbalance effects. Although these methods did not always outperform other ensemble approaches, they showed promise as a means to place emphasis on the minority class. In future research, these methods will be explored further and compared with existing methods for dealing with class imbalance, including cost matrices.

When survival values are discretized, the performance of classification algorithms may suffer, particularly for patients whose survival times are near the discretization threshold. In lieu of discretizing survival, an alternative approach would be to retain survival as a continuous variable and use regression algorithms to obtain continuous predictions for each patient. The log-rank statistic and Kaplan-Meier curves—familiar and well-accepted standards for assessing clinical relevance—require that patients be assigned to discrete groups. One way to meet this requirement would be to apply a

clustering algorithm to the continuous survival predictions and compare the actual survival values of patients in each cluster. Such an approach may outperform methods that either 1) discretize survival or 2) perform unsupervised clustering on the entire data set. Additionally, this approach could also be used as an ensemble-learning method because it could account for predictions from multiple data categories and algorithms.

Although this study has focused on GBM, the first cancer type with a substantial amount of data in TCGA, the U.S. government recently announced that an additional $275 million would be invested in TCGA and that it will eventually cover more than 20 cancer types [81]; those efforts are now ongoing. As new data sets become available in TCGA, the tools and methods of this study can be applied to other cancer types. Although prognosis has been the outcome of interest in this study, the ML-Flex package can be used to perform multivariate analyses for any outcome relevant to a given cancer type. Because ML-Flex will be made available publicly, other researchers will be able to explore relationships in TCGA that otherwise would have required a considerable software-engineering effort.

5.5 Relevance to Biomedical Informatics

Shortliffe and Blois have defined *biomedical informatics* as "the scientific field that deals with biomedical information, data, and knowledge—their storage, retrieval, and optimal use for problem solving and decision making." [82] Although other definitions exist, this definition has been used commonly in the field.

Accordingly, the purpose of this study was to help solve an important biomedical problem: short patient survival after GBM diagnosis. In pursuit of this goal, large quantities of biomedical data were retrieved, stored, and processed. Scientific tools and techniques were developed and applied in an attempt to convert data to information, which could then be interpreted and potentially be considered knowledge. It is hoped that such knowledge will serve as building blocks for future research and ultimately have a positive impact on human health.

Biomedical informatics is a highly interdisciplinary field. In isolation, the fields of biology, medicine, computer science, information systems, or statistics lack the

perspective and tools necessary for studies such as this one to be accomplished. However, as these fields continue to become connected through biomedical informatics approaches, scientific progress is sure to accelerate dramatically.

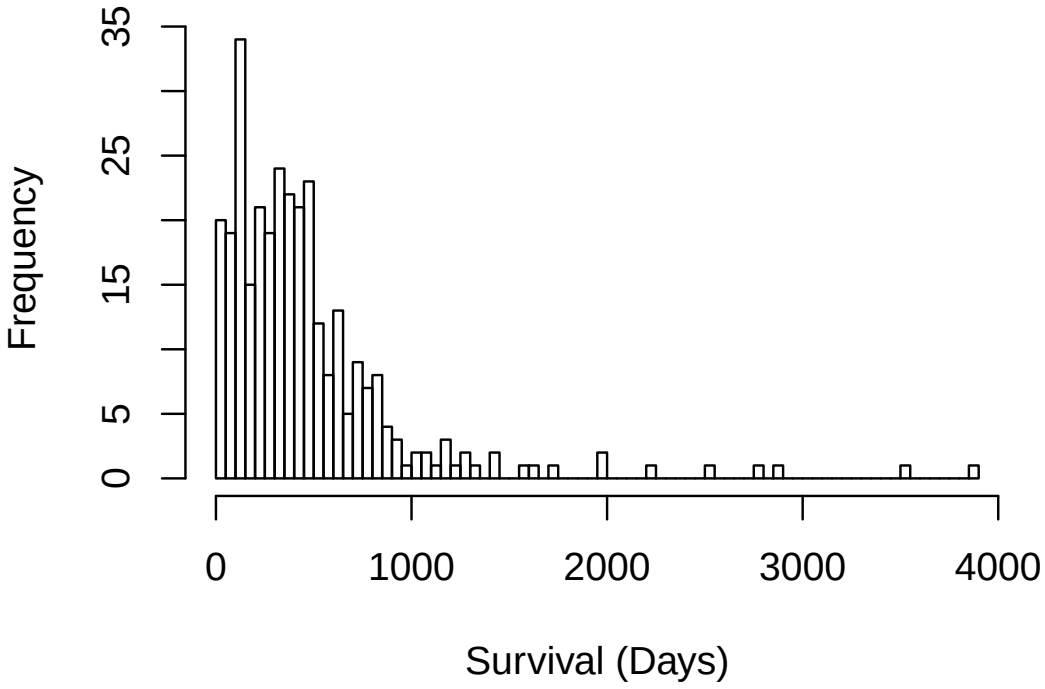

Figure 5.1. Distribution of survival times across all GBM patients that were used in this study.

CHAPTER 6

CONCLUSION

The aggressive nature of GBM leaves clinicians with a relatively short time span to determine optimal treatments for each patient. Although radiation treatment, surgical resection, and temozolomide treatment have shown promise for lengthening GBM survival times, few patients survive longer than five years after diagnosis. A better understanding of factors that are associated with GBM survival—and thus that may indicate a lack of response to standard treatments—could help clinicians prioritize patients for clinical trials and help patients make decisions about entering such trials. An increased understanding of the biological mechanisms that drive tumor aggressiveness—and that may differentiate the most (or least) lethal tumors from the remaining tumors—may also lead researchers to molecularly targeted treatments that improve patient outcomes. [83]

Some prognostic insight may be gained from considering a patient's age or KPS or from examining a patient's tumor under a microscope. However, these data have limited ability to distinguish between LTS and STS. Even though biomolecular aberrations are at the root of tumor initiation and progression, [84] no prognosis model based on biomolecular data is in widespread use by clinicians who treat GBM patients. [1] Until recent years, bench researchers studying GBM prognosis have been limited to small-scale efforts that evaluated one or a few biomolecular variables at a time. Fortunately, technological advances are making it possible for researchers to examine the biomolecular characteristics of cancer cells with increasing granularity and at decreasing costs. (The magnitude of such data sets will only increase as "next-generation" sequencing technologies become more commonplace.) The resulting data deluge necessitates the use of sophisticated informatics techniques to store, retrieve, and analyze the data sets; however, implementation of such techniques often lies

outside the expertise of bench researchers and clinicians.

Although some prognostic factors may determine a GBM patient's fate in isolation, it is likely that multiple factors work in concert to influence tumor aggressiveness and ultimately survival for many patients. In some cases, tens, hundreds, or even thousands of factors may each have a subtle impact on tumor activity. Complicating the situation further, prognostic factors may interact with each other synergistically or antagonistically. For reasons such as these, multivariate approaches to developing prognosis models are warranted—traditionally used statistical techniques for predicting survival may not be suitable in the face of thousands of independent variables, strong dependencies between variables, and multiple scales of measurement. Moreover, it is essential that analyses be conducted in a systematic and consistent way to ensure validity, repeatability, and comparability across studies.

In this study, a variety of multivariate approaches were applied to various categories of data for a cohort of GBM patients, and predictive performance was evaluated in a robust, cross-validated design. Although performance of the algorithms varied substantially across the data categories, some models performed well for all three metrics—particularly models based on age, treatments, and DNA methylation.

Even though a long road may still lie ahead for researchers working to eradicate devastating diseases like GBM, informatics tools and techniques such as those presented in this study promise to guide researchers in their efforts to improve outcomes and explain the biological underpinnings of disease.

REFERENCES

[1] Colman H, Zhang L, Sulman EP, et al. A multigene predictor of outcome in glioblastoma. *Neuro Oncol.* 2010;12(1):49–57.

[2] Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med.* 2005;352(10):987–96.

[3] Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. *CA Cancer J Clin.* 2009;59(4):225–249.

[4] Ohgaki H, Dessen P, Jourde B, et al. Genetic pathways to glioblastoma: A population-based study. *Cancer Res.* 2004;64(19):6892–6899.

[5] Tait MJ, Petrik V, Loosemore A, Bell BA, Papadopoulos MC. Survival of patients with glioblastoma multiforme has not improved between 1993 and 2004: Analysis of 625 cases. *Br J Neurosurg.* 2007;21(5):496–500.

[6] Davis L, Martin J. A study of 211 patients with verified glioblastoma multiforme. *J Neurosurg.* 1949;6(1):33–44.

[7] Burger PC, Green SB. Patient age, histologic features, and length of survival in patients with glioblastoma multiforme. *Cancer.* 1987;59(9):1617–1625.

[8] Gundersen S, Lote K, Hannisdal E. Prognostic factors for glioblastoma multiforme: Development of a prognostic index. *Acta Oncol (Madr).* 1996;35(S8):123–127.

[9] Lacroix M, Abi-Said D, Fourney DR, et al. A multivariate analysis of 416 patients with glioblastoma multiforme: Prognosis, extent of resection, and survival. *J Neurosurg.* 2001;95(2):190–198.

[10] Chandler KL, Prados MD, Malec M, Wilson CB. Long-term survival in patients with glioblastoma multiforme. *Neurosurgery.* 1993;32(5):716–720.

[11] Lamborn KR, Chang SM, Prados MD. Prognostic factors for survival of patients with glioblastoma: Recursive partitioning analysis. *Neuro Oncol.* 2004;6(3):227–235.

[12] Krex D, Klink B, Hartmann C, et al. Long-term survival with glioblastoma multiforme. *Brain.* 2007;130(10):2596–2606.

[13] Houillier C, Lejeune J, Benouaich-Amiel A, et al. Prognostic impact of molecular markers in a series of 220 primary glioblastomas. *Cancer.* 2006;106(10):2218–2223.

[14] Ruano Y, Ribalta T, de Lope AR, et al. Worse outcome in primary glioblastoma multiforme with concurrent epidermal growth factor receptor and p53 alteration. *Am J Clin Pathol.* 2009;131(2):257–263.

[15] Weller M, Felsberg J, Hartmann C, et al. Molecular predictors of progression-free and overall survival in patients with newly diagnosed glioblastoma: A prospective translational study of the german glioma network. *J Clin Oncol.* 2009;27(34):5743–50.

[16] Rich JN, Hans C, Jones B, et al. Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Res.* 2005;65(10):4051–4058.

[17] Parsons WW, Lin JCC, Jones S, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science.* 2008;321(5897):1807–1812.

[18] Hegi ME, Diserens A, Gorlia T, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med.* 2005;352(10):997–1003.

[19] Liang Y, Diehn M, Watson N, et al. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc Natl Acad Sci U S A.* 2005;102(16):5814–5819.

[20] Nutt CL, Mani DR, Betensky RA, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.* 2003;63(7):1602–1607.

[21] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* 2008;455(7216):1061–1068.

[22] Ho YC, Pepyne DL. Simple explanation of the no-free-lunch theorem and its implications. *Journal of Optimization Theory and Applications.* 2002;115(3):549–570.

[23] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102(43):15545–15550.

[24] Homma T, Fukushima T, Vaccarella S, et al. Correlation among pathology, genotype, and patient outcomes in glioblastoma. *J Neuropathol Exp Neurol.* 2006;65(9):846–54.

[25] Ohgaki H, Kleihues P. Genetic pathways to primary and secondary glioblastoma. *Am J Pathol.* 2007;170(5):1445–1453.

[26] Forbes S, Clements J, Dawson E, et al. COSMIC 2005. *Br J Cancer.* 2006;94(2):318–22.

[27] Bujko M, Kober P, Matyja E, et al. Prognostic value of IDH1 mutations identified with PCR-RFLP assay in glioblastoma patients. *Molecular Diagnosis & Therapy.* 2010;14(3):163–169.

[28] Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell.* 1990;61(5):759–67.

[29] Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med.* 2004;10(8):789–99.

[30] Marko NF, Toms SA, Barnett GH, Weil R. Genomic expression patterns distinguish long-term from short-term glioblastoma survivors: A preliminary feasibility study. *Genomics.* 2008;91(5):395–406.

[31] Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and regression trees.* Boca Raton: Chapman and Hall/CRC; 1984.

[32] DecisionDx-GBM. Castle Biosciences Web Site. http://www.castlebiosciences.com/decisiondx-gbm. Published January 1, 2010. Accessed January 28, 2011.

[33] Nigro JM, Misra A, Zhang L, et al. Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. *Cancer Res.* 2005;65(5):1678–86.

[34] The Cancer Genome Atlas Research Consortium. Mission and goal. The Cancer Genome Atlas Web Site. http://tcga.cancer.gov/about/mission.asp. Published January 1, 2007. Accessed October 25, 2010.

[35] Dietterich TG. Ensemble methods in machine learning. *Lect Notes Comput Sci.* 2000;1857(1):1–15.

[36] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn.* 2003;51(2):181–207.

[37] Boland PJ. Majority systems and the condorcet jury theorem. *The Statistician.* 1989;38(3):181–189.

[38] Littlestone N, Warmuth MK. The weighted majority algorithm. *Information and Computation.* 1994;108(2):212–261.

[39] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep.* 1966;50(3):163–170.

[40] Kaplan EL, Meier P. Nonparametric estimation from incomplete observtions. *J Am Stat Assoc.* 1958;53(282):457–481.

[41] Ohno-Machado L. Modeling medical prognosis: Survival analysis techniques. *J Biomed Inform.* 2001;34(6):428–439.

[42] Colman H, Aldape K. Molecular predictors in glioblastoma: Toward personalized therapy. *Arch Neurol.* 2008;65(7):877–83.

[43] The Cancer Genome Atlas Research Consortium. Data portal. The Cancer Genome Atlas Web Site. http://cancergenome.nih.gov/dataportal. Published January 1, 2007. Accessed October 25, 2010.

[44] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860–921.

[45] Cheung VG, Nowak N, Jang W, et al. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature.* 2001;409(6822):953–8.

[46] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics.* 2006;7(1):91.

[47] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn.* 2002;46(1):389–422.

[48] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter.* 2009;11(1):10.

[49] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In: Bergadano F, De Raedt L, eds. Machine Learning ECML94. Berlin / Heidelberg: Springer; 1994:171–182.

[50] Stevens SS. On the theory of scales of measurement. *Science.* 1946;103(2684):677–680.

[51] Quinlan JR. *C4.5: Programs for machine learning.* San Mateo: Morgan Kaufmann; 1993.

[52] Quinlan JR. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research.* 1996;4:77–90.

[53] Langley P, Iba W, Thompson K. An analysis of bayesian classifiers. In: Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose: AAAI Press; 1992:223–228.

[54] Domingos P, Pazzani M. On the optimality of the simple bayesian classifier under zero-one loss. *Mach Learn.* 1997;29(2):103–130.

[55] John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufmann; 1995:338–345.

[56] Vapnik VN. *Statistical learning theory.* New York: Wiley; 1998.

[57] Noble WS. What is a support vector machine? *Nat Biotechnol.* 2006;24(12):1565–7.

[58] Chang C, Lin C. LIBSVM: A library for support vector machines. LIBSVM Web Site. http://www.csie.ntu.edu.tw/∼cjlin/libsvm. Published January 1, 2001. Accessed January 31, 2011.

[59] Wolpert D. Stacked generalization. *Neural Networks.* 1992;5(2):241–259.

[60] Burton EC, Lamborn KR, Feuerstein BG, et al. Genetic aberrations defined by

comparative genomic hybridization distinguish long-term from typical survivors of glioblastoma. *Cancer Res.* 2002;62(21):6205–6210.

[61] Edgerton ME, Fisher DH, Tang L, Frey LJ, Chen Z. Data mining for gene networks relevant to poor prognosis in lung cancer via backward-chaining rule induction. *Cancer Inform.* 2007;3:93–114.

[62] *R: A language and environment for statistical computing* [computer program]. Version 2.11.0. Vienna, Austria: R Foundation for Statistical Computing; 2008.

[63] Therneau, T and Lumley, T. survival: Survival analysis, including penalised likelihood. The R Project for statistical computing Web Site. http://cran.r-project.org/package=survival. Published November 22, 2010. Accessed December 29, 2010.

[64] Frank A, Asuncion A. Machine learning repository. University of California, Irvine, School of Information and Computer Sciences Web Site. http://archive.ics.uci.edu/ml. Published March 1, 2009. Accessed November 1, 2010.

[65] Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc Natl Acad Sci U S A.* 1990;87(23):9193–6.

[66] Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. *Bioinformatics.* 2007;23(2):257–8.

[67] Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* 2006;34(Database issue):D354–7.

[68] Cadieux B, Ching T, VandenBerg SR, Costello JF. Genome-wide hypomethylation in human glioblastomas associated with specific copy number alteration, methylenetetrahydrofolate reductase allele status, and increased proliferation. *Cancer Res.* 2006;66(17):8469–76.

[69] Issa JJ. Methylation and prognosis: Of molecular clocks and hypermethylator phenotypes. *Clin. Cancer Res..* 2003;9(8):2879–2881.

[70] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. *Intelligent Data Analysis.* 2002;6(5):429–449.

[71] Bradley A. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 1997;30(7):1145–1159.

[72] Kleihues P, Cavenee WK, eds. *Pathology and genetics of tumours of the nervous system.* Lyon: IARC Press; 2000.

[73] Suzuki K, Suzuki I, Leodolter A, et al. Global DNA demethylation in gastrointestinal cancer is age dependent and precedes genomic damage. *Cancer Cell.* 2006;9(3):199–207.

[74] Kaminska B, Wesolowska A, Danilkiewicz M. TGF beta signalling and its role in tumour pathogenesis. *Acta Biochim Pol.* 2005;52(2):329–37.

[75] Normanno N, De Luca A, Bianco C, et al. Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene.* 2006;366(1):2–16.

[76] Fritz JH, Girardin SE. How toll-like receptors and nod-like receptors contribute to innate immunity in mammals. *J Endotoxin Res.* 2005;11(6):390–4.

[77] de Visser KE, Eichten A, Coussens LM. Paradoxical roles of the immune system during cancer development. *Nat Rev Cancer.* 2006;6(1):24–37.

[78] Sampson JH, Archer GE, Mitchell DA, Heimberger AB, Bigner DD. Tumor-specific immunotherapy targeting the EGFRvIII mutation in patients with malignant glioma. *Semin Immunol.* 2008;20(5):267–75.

[79] Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst.* 2007;99(2):147–57.

[80] Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics.* 2010;11(1):523.

[81] The Cancer Genome Atlas Research Consortium. The cancer genome atlas project to map 20 tumor types. The Cancer Genome Atlas Web Site. http://tcga.cancer.gov/wwd/program. Published January 1, 2010. Accessed October 25, 2010.

[82] Shortliffe EH, Cimino JJ, eds. *Biomedical informatics: Computer applications in health care and biomedicine.* 3rd ed. New York: Springer; 2006.

[83] Bild AH, Potti A, Nevins JR. Linking oncogenic pathways with therapeutic opportunities. *Nat Rev Cancer.* 2006;6(9):735–741.

[84] Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med.* 2004;10(8):789–799.