A COMPREHENSIVE RISK STRATIFICATION MODEL FOR PROGNOSTICATION AND ASSISTING WITH THERAPEUTIC DECISION-MAKING FOR MULTIPLE MYELOMA PATIENTS

A THESIS

Presented to the Department of Mathematics and Statistics

California State University, Long Beach

In Partial Fulfillment

of the Requirements of the Degree

Master of Science in Applied Statistics

Committee Members:

Hojin Moon, Ph.D. (Chair)
Tianni Zhou, Ph.D.
Kagba Suaray, Ph.D.

College Designee:

Tangan Gao, Ph.D.

By Brian Song

B.S., 2014, University of California, Irvine

August 2018

ABSTRACT

A COMPREHENSIVE RISK STRATIFICATION MODEL FOR PROGNOSTICATION AND ASSISTING WITH THERAPEUTIC DECISION-MAKING FOR MULTIPLE MYELOMA PATIENTS

By

Brian Song

August 2018

The goal of the research is to improve current risk stratification models of multiple myeloma by developing a novel statistical decision algorithm. The increase in precision would assist in providing optimal treatments for multiple myeloma cancer patients depending on the risk of progression at the time of diagnosis. If progression of cancer is imminent, then risk-adapted therapy would be a considerable option. Larger amount of data supplied from multiple clinics were gathered to obtain better prognosis. The data are available from the Synapse website under the Multiple Myeloma DREAM Challenge site. Although both genomic variation data and gene expression data were available, the study was done with the latter in conjunction with general patient data. Preliminary research has shown that the microarray data were not standardized among the different clinics, so the study required additional preprocessing before aggregating all data for comprehensive investigation. Accelerated Time Failure model is used to screen insignificant variables for easier processing, reducing 17,308 markers to 4,503. A combination of random forest models and likelihood ratio test is utilized to further reduce potentially significant biomarkers. The remaining biomarkers are used in multiple statistical models to determine the optimal model that best represents the data. The efficacy of the model is checked by using two clinics to train the model to predict the third clinic. The average and standard deviation of the resulting statistics are used to validate the consistency of the model for different clinics. We show that an improvement in current risk stratification models can be obtained.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Multiple myeloma is when multiple plasma cells become cancerous. Because of the nature of plasma cells, this cancer generally occurs within the bone marrow. With the existence of cancer in bones, there will be many symptoms, such as low blood count (anemia), low platelet count (thrombocytopenia), issues with bone stability, infections, kidney problems, and many other issues. This cancer tends to come into effect later in life (as can be seen in Figure 1). To date, there is no known cure for this cancer, and the survival rate tends to be around three years, albeit some patients have lived up to ten years (Alexander et al. 2007).

From 15-25 percent of multiple myeloma patients will progress or die within 18 months of diagnosis regardless of treatment. To approach this aggressive disease, risk-adapted therapy is implemented as a potential standard method of treatment. However, due to the lack of standardization in multiple myeloma patient risk stratification models, along with current models yielding results with limited precision, a search for greater performance is embarked.

There has been progress in the development of patient stratification methodologies, but there is still much opportunity to further improve stratification efforts to optimize treatment strategies and to develop new therapies that address unmet medical needs to multiple myeloma patients. Currently, there is the Myeloma Prognostic Risk Signature (MyPRS), which is a clinical test to monitor, generate prognosis, assess risk, and manage cancer treatments by some genomic markers selected from seventeen peer-reviewed journals. The methodology of this test consists of immunohistochemistry for digital/virtual karyotyping and RNA analysis for gene expression profiling. Researchers at University of Arkansas for Medical Sciences (UAMS) created a seventy-gene prognosis score (range 0-100) to measure the risk for relapse or shorter overall survival. Upon obtaining a score, depending on the severity of the cancer, therapeutic measures can be taken accordingly (Van Laar et al. 2014). Improvements in this treatment can be made, such as higher risk patients (ranging from 15-30 percent of all patients) obtaining alternative treatment regime and/or referral to an appropriate clinical trial. Also, for the larger population of patients diagnosed with a lower risk, a reduction in intensity of treatments can be beneficial. Since the risk-adapted therapy based on gene signatures is our major concern, our goal is to develop a comprehensive risk stratification model for prognostication in order to improve patients' survival and ultimately find a cure for multiple myeloma.

Synapse partnered with the Myeloma Genome Project (MGP) to design the Multiple Myeloma DREAM Challenge. The challenge was created in a collaborative effort to obtain many multiple myeloma patient data to create a generalized model that best stratifies high risk patients. A high risk patient is defined as a patient whose disease progresses (or causes death) within 18 months from time of diagnosis, which is approximated as 18*30.5= 549 days. The challenge is divided into three objectives: (a) identifying high risk patients using genomic variation data (DNA single nucleotide variants, indel, etc.), (b) identifying high risk patients using gene expression data (RNA-seq or microarray), (c) identifying high risk patients using any available data. For this study, we will focus on the second objective and use expression data. The challenge is set up such that three clinical data will be available for analysis and training, and four clinics will be hidden from the contestants in which the model will be tested against. The area under the curve will be averaged out across the four clinics, and the average will be the final score for the competition. The goal of the analysis is to find significant biomarkers that can distinguish whether a person is at risk of progression. As the competition has ended early, this study will mimic the competition by using two clinics at a time to train and test against the third clinic. Then the score will be averaged out on the prediction of the three clinics and will closely represent the scores of others.

CHAPTER 2

STATISTICAL BACKGROUND

(Penalized) Logistic Regression

Logistic regression is a statistical method to classify a dataset to two or more outcomes dependent on one or more independent variables. This is parametric linear regression model, which can be defined as

$$logit(p) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

where

$$odds = \frac{p}{1-p} = \frac{probability\ of\ event\ occuring}{probability\ of\ event\ not\ occuring}$$

and

$$logit(p) = ln(\frac{p}{1-p})$$

b_0 is the intercept of the linear regression equation, and$b_1, \ldots, b_n$ represent the coefficients for the independent variables $x_1, x_2, \ldots, x_n$.

By setting up the equation

$ln(\frac{p}{1-p})$=$b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$

And solving for p, we get

$$\frac{p}{1-p} = e^{b_0+b_1x_1+b_2x_2+\ldots+b_nx_n}$$

$$\Rightarrow \frac{1-p}{p} = e^{-(b_0+b_1x_1+b_2x_2+\ldots+b_nx_n)}$$

$$\Rightarrow \frac{1}{p} - 1 = e^{-(b_0+b_1x_1+b_2x_2+\ldots+b_nx_n)}$$

$$\Rightarrow \frac{1}{p} = 1 + e^{-(b_0+b_1x_1+b_2x_2+\ldots+b_nx_n)}$$

$$\Rightarrow p(x) = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n)}}$$

This shows that the probability of an event occurring takes the form of a sigmoid function for logistic regression.

Assume a univariate equation: $odds = e^{b_0+b_1x_1}$. For a continuous variable x_1, we can look at the odds ratio of one unit increase in the variable:

$$Odds\, ratio = \frac{odds(x_1+1)}{odds(x_1)} = \frac{e^{b_0+b_1(x_1+1)}}{e^{b_0+b_1x_1}} = e^{b_1}$$

We can see that the odds of an event occurring in the univariate equation multiply by $e^{b_1}$for every one unit increase of x_1. For a multivariate equation, the odds of the event occurring multiplies by $e^{b_n}$for every one unit increase of $x_n$when all other variables are constant. For a categorical independent variable, the associated constant can be defined as the presence of the category with respect to the base category.

The constants cannot be estimated via maximum likelihood estimation (MLE), as there is no closed-form expression for the equation. They are instead estimated through an iterative process, such as Newton’s method, to solve the general MLE equation:

$$min_f \sum_{i=1}^{n} V(f(x_i), y_i)$$

Fitting a logistic regression model may yield an ill-posed problem (a solution does not exist or the behavior of the model changes drastically given different conditions) or fall under the issue of overfitting. Both cases can occur when there are many independent variables considered for the model from potential causes such as multicollinearity. Including a regularizer to the loss function helped overcome this detriment, which penalizes the variables and only selects a subset of variables that would yield a well-posed model. The penalized loss function can simply take the form:

$$min_f \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda R(f)$$

where $R(f)$is the regularization term and λ is a parameter that controls the significance of the regularization term.

One popular regularization function is the Tikhonov regularization, also recognized as L_2 regularization and ridge regression. This is done by shrinking large coefficients to prevent overfitting. The regularization term takes the form

$$R(f) = ||\beta||_2^2$$

This regularization utilizes the Euclidean norm to put more weight in a solution with desirable properties. The objective function is to be minimized without the norm getting too big, limiting the amount of weight one parameter can influence the model. Many of the lesser significant variables (which could be deemed insignificant due to multicollinearity) will get coefficients shrunk closer to 0, marginally influencing the model.

Another popular penalty function is the least absolute shrinkage and selection operator (LASSO) method, which utilizes absolute norm for penalization. This regularizer is set to solve the objective function:

$$min_f \sum_{i=1}^{n} V(f(x_i), y_i) = min_f \sum_{i=1}^{n} (y_i - x^T\beta)^2 \text{ such that } \sum_{j=1}^{m} \left|\beta_j\right| \leq t$$

LASSO essentially caps the sum of the absolute value of the coefficients to be less than a fixed value, and drop variables to satisfy the criteria. This differs from ridge in that ridge balances out the coefficients as opposed to removing them. The limitation of this regularizer arises when there is high-dimensional data, with limited sample. LASSO would choose only a selected few and drop the rest. Alongside that, any confounding variables would be dropped if one of the variables were selected.

To balance out the conservative nature of ridge and overcome the limitations of LASSO, elastic net was developed. The regularization function takes the form:

$$[(1-\alpha)\big||\beta|\big|_2^2 + \alpha\big||\beta|\big|_1]$$

where α is a pre-defined constant that weighs the amount of effect LASSO and ridge would proportionally influence the objective function. Elastic net uses a weighted sum of the two regularization techniques to use the coefficient dampening effects of ridge and reduce dropping of coefficients from LASSO. In doing so, the final equation will be parsimonious while the potential issue of multicollinearity is reduced.

Accelerated Failure Time Model

The Accelerated Failure Time (AFT) Model was proposed as an alternative in response to the limitations of classical proportional hazard models. Proportional hazard models define the relationship of events and covariates as an event rate at time t conditional on survival time until time t or later, whereas AFT Model defines the relationship as each covariate accelerating or decelerating the time to event. AFT overcomes the limitations of requiring an assumption for proportional hazards and requiring a parametric statistical distribution for survival time (Zare et al. 2015). For time T, the AFT model is defined as

$$Y_i = {x_i}^T\beta + W_i,$$

where $Y = logT$, and i.i.d. $W_i \sim f$ are the error terms. AFT is considered a semi-parametric model, as the ${x_i}^T\beta$ portion is the linear equation that contains parameters, and W_i is assumed an independently and identically distributed random variable from an unspecified distribution function *f*. T is commonly a distribution such as Weibull, exponential, logistic, or normal, depending on which best fits the trend of the data.

Random Forest

Random forest is an ensemble learning, non-parametric model that can be used for both regression and classification problems. This supervised model functions as a collection of decision trees, which individually generates a prediction, and the results of all trees are aggregated by either mean (for regression) or mode (for classification). Decision tree uses the predictor that best splits the data based on the response, and continues splitting until the samples are well split. Decision tree is considered a weak learner, as it could split the data too much, causing the model to overfit. The collaborative decision of an ensemble of trees averages out the high variance brought by each tree, yielding results that are more stable by "popular" vote. Each tree is grown using a subset of predictors, and a subset of the sample. It is common to use the square root of the total number of variables as the number of predictors each tree will consider. With each decision tree taking in different subsets of predictors, the model becomes robust to common issues such as multicollinearity and interactive effects among predictors.

For each tree, n samples are randomly selected with repeat through bootstrap aggregating, or bagging. More formally, given a training set $X = x_1, x_2, \dots, x_n$ with response variables $Y = y_1, y_2, \dots, y_n$, B samples are randomly selected with replacement. When building each tree in a random forest, a chunk (usually one-third) of the sample is set aside from the pool of sample for bagging. This subset is then utilized to calculate "out-of-bag" error. The out-of-bag error rate is a nice measurement of the performance of the model, as the scores will be generated from a subset of samples that were not used per tree in the random forest. This error rate is very useful when tuning the hyperparameters to best fit the data. Hyperparameters such as number of trees or max number of splits for all trees are tunable parameters that will require grid search cross validation (training model with different combination of hyperparameters and choosing the best model based on a statistic such as "area under curve" or AUC) to best optimize the model. The out-of-bag error can be used to measure the significance of predictors (Breiman 2001). Out-of-bag error is measured by predicting sample x_i using the trees that did not contain that sample. To measure the importance of a predictor, the out-of-bag error of each tree is measured with the predictor being included and not included, and then the difference is recorded. The average of the difference across the trees is recorded, and the value is standardized by the standard deviation. The larger score obtained indicates a higher valued predictor, as it is representative of a large change in error when the predictor is included (or removed) from the model (Zhu, Zheng, and Korosok 2015). This score is recognized as the mean decrease in accuracy. Alternatively, the mean decrease in Gini coefficient is another measurement of variance significance. This is the measurement of how much each variable contributes to the homogeneity of the nodes and leaves in the random forest.

Support Vector Machines

Support vector machines (SVM) is a method of supervised learning models commonly used for classification (although a version of SVM has been proposed to work for regression as well). First invented by Vapnik and Chervonenkis in 1963, this model discriminates two classes by determining a hyperplane that would best separate them. Although there are potentially infinite hyperplanes that discriminate the classes well (imagine two clusters in a 2-D graph with any number of lines that can be drawn between them), the best hyperplane is calculated based on maximizing the distance from the closest points to the hyperplane. These close points are the support vectors. By maximizing the distance, the likelihood of misclassification is then minimized.

Let’s consider a situation where two classes are linearly separable. If we assume two values of y_i to be either 1 or -1 (i.e., two classes in a classification model), then we create a hyperplane defined as (1) $\vec{w} * \vec{x} - b = 1$ and (2) $\vec{w} * \vec{x} - b = -1$, where the samples of one class are on or above the boundary of the first hyperplane, with the samples of the second class below the second hyperplane. Then the hyperplane that satisfies $\vec{w} * \vec{x} - b = 0$ (essentially the hyperplane that can be drawn in the center of the two hyperplanes) would represent the optimal splitting hyperplane. The support vectors would be the vectors $\vec{x}_i$ that are on the margins of the two hyperplanes (1) and (2).

This hyperplane can extend further than a linear classification with implementation of the kernel trick to define the hyperplane in a higher dimension for more complex datasets. The kernel trick is a method of computing the inner product of the training inputs to transform a non-linear separation of points to a linear map, which can then be fit. Popular kernel functions for SVM are linear ($\vec{x}_i * \vec{x}_i$), polynomial $(\vec{x}_{i*}\vec{x}_i)^d$, radial basis function exp(-γ‖$\vec{x}_i - \vec{x}_j||^2$) for γ>0, and sigmoid/hyperbolic tangent $tanh$(κ$\vec{x}_i * \vec{x}_j$+c) for some κ>0and c<0.

CHAPTER 3

STUDY BACKGROUND

The data used consist of patients from three locations. The GSE24080UAMS dataset is a collection of 559 microarray experiments contributed by the Myeloma Institute for Research and Therapy at the University (UAMS) in in Little Rock, Arkansas. The gene expression profiling of highly purified bone marrow plasma cells was performed in newly diagnosed patients with multiple myeloma. The bone marrow aspirates were enriched for plasma cells by anti-CD138 immuno-magnetic bead selection of mononuclear cell fractions of bone marrow aspirates.

HOVON65 consists of 282 microarray experiments on newly diagnosed multiple myeloma patients enrolled in the phase III HOVON65/GMMD-HD4 clinical trials. Bone marrow was used to purify plasma cells to a purity of >80%. The analysis was performed on the Affymetrix HG-U133 Plus 2.0 platform.

E-MTAB-4032 dataset (also recognized as Kryukov) is a collection of 149 microarray experiments which were enriched by magnetic cell sorting and samples with >80% purity as determined by flow cytometry and cytospin was used for RNA isolation and microarray analysis.

The microarray datasets are all affymetrix chips and CEL files have been reprocessed via the R oligo package with with robust multichip average (RMA) in order to provide consistency. Alongside the microarray data, patient clinical information is also available from the Multiple Myeloma Research Foundation (MMRF), such as the age, gender, and International Staging System (ISS). The ISS is a system that categorizes the stage of multiple myeloma cancer based on four factors: the amount of albumin in the blood, the amount of beta-2-microglobulin in the blood, the amount of lactate dehydrogenase (LDH) in the blood, and the specific gene abnormalities (cytogenetics) of the cancer.

CHAPTER 4

PRELIMINARY ANALYSIS

As Table 1 shows, in each clinic, the number of patients vary greatly, having 559, 282, and 149 for UAMS, HOVON, and E-MTAB respectively. Alongside that, the number of high risk patients also differ. In Table 2, it is also noticeable that the death rate, progression rate, and proportion of stages of cancer for each clinic also differ. UAMS has the lowest death and progression rate of 31 percent and 45 percent respectively, and the E-MTAB patient data has the highest death and progression rate of 37 percent and 97 percent respectively. These vastly varying clinics were intentionally chosen so that when merging the three clinics together, the model would be built to be as generalized as possible.

To have an idea of how the immensely large data looks, we implement principal component analysis (PCA) to reduce the dimensions of the patients' microarray data to fifty (85 percent of explained variance retained as seen on Figure 2), and then reducing the dimensions to two with t-distributed stochastic neighbor embedding (t-SNE) to visualize the patients that are similar in microarray information. t-SNE is utilized to reduce the dimensions to two because it discriminates distant neighbors more, so clusters would be better distinguishable. Looking at Figure 3, it is noticeable that three clusters were well separated. These clusters are separated perfectly by the clinics. This is because the variability from the hybridization process performed to obtain the data is greater than the variability within the microarray data. Looking at Table 3, which shows the average expression levels at each biomarker for each class for each clinic, it is clear that inter-clinic variation is high. For the first column of data, for instance, it is noticeable that the variation between the classes in each clinic is much less than the variation of the values as a whole. This makes the collected experimental data across each clinic not entirely homogeneous.

Meta-analysis is a highly valued approach to biostatistical data, as there is generally a lack of patient data for studies. So, when there are multiple sources of the same information, finding a method to adequately combine data is important. For this analysis, two approaches are considered: standardize the microarray data to all have approximately equal mean and standard deviation and analyze the clinical data separately and obtain similar results. The former approach is the better method to determine a generalized model that can accurately predict high risk patients. This is ideal because the potentially significant markers would be shown as expressive regardless of the procedures taken to obtain the expression data. The latter approach would be beneficial to have an idea of what individual biomarkers were considered significant within the clinic. However, there would be limited use of the results, as the significant markers would only be relevant to the specific procedure taken, which is not necessarily the standard method. So we implemented the former approach and standardized the markers before merging together. We can see in Figure 4 that the clusters of the clinics were well mixed.

As mentioned previously, the ISS is a categorization of the stages of multiple myeloma cancer. To validate the differentiation of the stages, we fit the progression-free survival times of each group onto a Kaplan-Meier estimation, then perform Log-Rank test to check if the survival times are statistically different. The results of comparing the three ISS stages for each clinic is shown on Table 4, 5, and 6. The Kaplan-Meier curves are also depicted in Figure 5, 6, and 7. It is noticeable that the progression-free survival times are much less for ISS 3. This makes sense, as it is the representation of the latest of cancer. ISS 1 and 2 were shown statistically different for E-MTAB, HOVON, and UAMS with p values 0.00232, 0.00687, and 0.0211, respectively. For

E-MTAB, the survival times between ISS 2 and 3 were not shown to be statistically different, but was shown otherwise for the other two clinics. This is also in line with what is visually noticeable in Figure 5, as the ISS 2 and ISS 3 survival curve look to overlap. Unsurprisingly, the survival times of ISS 1 and ISS 3 groups were shown to be statistically different.

CHAPTER 5

MODEL DESIGN

Preprocessing

The microarray data from each clinic consists of over 17,000 Entrez IDs, with many non-overlapping markers. The Entrez IDs that did not appear in all clinics have been filtered out from the study, leaving 17,038 IDs. Each expression data within each clinic was standardized to set mean as zero and standard deviation as one so that the data can be merged together without bias from each clinic, and bias across the expression data when selecting significant markers. The 'Clinic' was recorded before the three expression data were merged as to keep track of which clinic the patients originated from. From the MMRF Clinical data, ISS, progression-free survival time, and high risk flag were merged to the expression data.

Variable Screening

Because of the large number of predictors, variable screening is a requirement for efficient analysis. To approach this, Cox proportional hazard was considered to model survival time against predictors. However, checking the proportional hazard assumption by calculating the scaled Schoenfeld residual indicated that 2,212 of the 17,038 variables violated the assumption. The alternative course of action taken was to utilize the univariate Accelerated Failure Time model. The Progression-Free Survival Time with censored flags were used against each Entrez ID, and the p-value of the predictors were recorded. All predictors under p-value of 0.05 were kept, and the remainder were screened. By having the p-value set at 0.05, it has the tendency to output more predictors as possibly significant. Being conservative with the results reduce the likelihood of important predictors being filtered out during this step. This screening process removed 12,535 markers, yielding 4,503 potentially important markers.

To gain a better insight on the data, we also separately screened the variables by the individual clinics. The number of markers remaining for GSE24080 UAMS, HOVON65, and E-MTAB4032 were 4,320, 2,511, and 2,493, respectively. Of the selected markers, 301 were common across the three clinics. What this entails is that some of the markers would not be as expressive from one hybridization process to the next. Considering the screening done with the combined samples considered more markers as potentially significant, it is likely that the markers that yield some significance in each clinic would be considered significant in the overall scheme.

Variable Selection

The predictors can still be reduced, as there are still thousands of markers. To reduce the number of predictors further, Random Forest Classification was utilized. To set up the procedure, high risk flag was used as the response, with using only the markers derived from variable screening. Censored data would not be useful in this step, as predicting requires known classes, so patients with censored information was removed. There are 287 high risk patients in total out of 977 total patients, which indicates an imbalance of classes. To reduce the likelihood of the models generated to be biased towards the non-high risk class, downsampling was performed to match the size of each class. Because there is a significantly large number of predictors, a high number of trees would be required to reduce risk of interaction within the predictors. The rest of the parameters defined for the random forest was set as the default values based on the R package ‘randomForest’. 20 trials of ten fold cross validation was performed, as to reduce correlation occurring within trees. By setting up the procedure as such, bias from the sample would also be neutralized, as it would average out across the many trees in the many random forests.

For each model training in the cross validation (200 total), whether a marker was considered or not was counted. The idea is that if a marker was not expressive or very low discrimination power relative to the subset of markers used in each tree, then it would yield a low number of variables used. The idea is that the more important markers would be used in the trees to help separate the sample into two classes. If the used markers from an individual tree split the data well enough, then the rest of the markers would not be used, as it is not required. Nearly all variables came out used, which is indicative of each individual marker minutely influencing the predictive power of the model, causing the model to utilize any information it can.

The total number of times each variable was used to split the data per tree across all trials was also recorded. One variable can be used multiple times within a tree if it is predictive enough to split the data well. So if the variable was used in the trees frequently, it is easy to notice the relative significance of the variable with respect to other variables. From the results, the total counts were a bit more diverse.

We utilize the random forest cross validation results to select variables for the final model. The counts of occurrence in trees was sorted from highest count to least to view the most potentially significant variable to the least. The variables were included into a logistic regression model one at a time and the likelihood ratio test was performed. Because one variable is added at a time, one model being compared is a subset of the other. This gives the opportunity to compare models to each other when one particular variable is considered for the model. To set up the test, the whole dataset was considered. This is because for the likelihood ratio test, the likelihood ratio statistic is set up as$= \frac{L(\theta_0|x)}{L(\theta_1|x)} = \frac{f(\cap_i \quad x_i|\theta_0)}{f(\cap_i \quad x_i|\theta_1)}$ where $\theta_0$and $\theta_1$represents the different set of parameters used. Because the variables were sorted relative to their importance by random forest, each subsequent marker being included would potentially be less significant as it is being included into the model. The iterative inclusion of variables continues until the function does not find a significant predictor after 100 successive variables were tested. The patience level of 100 was used to conservatively include potentially important variables while still finishing at a short period of time. Testing with patience as 80 and 120 yielded the same markers, so it is likely that the ending results contain sufficient number of markers and is not likely to find more. After obtaining the final results, variable inflation factor (VIF) was tested to remove any multicollinearity. Then, backward stepwise selection was performed to reduce the model. This was to overcome any potential variables included later that seems to be more significant than the initial variables included.

After selecting variables through the likelihood ratio test, 43 variables were selected by the time the function finished searching. There were no variables that indicated multicollinearity, but five were dropped during the backward selection step. In the end, 38 variables remained, which can be seen in Table 7.

Model Selection

With the selected 38 variables, different models were tested to select the optimal algorithm. To validate the generalization of the model, two clinics would be used to train the model, and the third clinic would be used as a test set. In total, there will be three combinations and results to determine the efficacy of the model. The average and standard deviation of the metrics across the three tests would be considered for comparing the power of each method. This metric is of importance as it would be the best representation of how well the model is generalized.

CHAPTER 6

RESULTS

Random forest, SVM, and Logistic Regression (penalized and unpenalized) were all considered for potential models. Random forest was built with 500 trees, with default hyperparameters as defined in the R package ‘randomForest’. SVM was built using linear kernel and was performed using the R package ‘e1071’. Base R function ‘glm()’ was used for unpenalized logistic regression, and R package ‘glmnet’ was used for penalized logistic regression. R package ‘glmnet’ has a parameter α that can be predefined to define a proportion of LASSO and ridge regularization that would influence the model. The values of α used were numbers between 0 and 1 (inclusive) in increments of 0.2. Note that when ⍰= *0*, the regularization function is identical to LASSO, and ridge when ⍰= *1*.

As the distribution of classes is still proportionally more in favor of low risk patients, either downsampling or upsampling (resampling the lesser populated classes with replacement to match the higher populated class) was implemented on each method to balance out the model. To make the results more in tune with realistic outcomes, raw unstandardized data were used to fit the model.

All the resulting performance of each model for each clinic is listed in Tables 8, 9, and 10. Based on the results obtained from averaging the metrics across the three combinations, shown in Table 11, logistic regression with downsampling yielded the strongest model. With the highest average area under the curve of 81 percent, it can be seen as a more discriminative model. The summary of the performance of this model is shown in Table 13. It is worth considering support vector machine with linear kernel did have a bit more balanced specificity and sensitivity compared to logistic regression, despite having slightly lower AUC. From Table

12, it can be seen that the AUC from the three clinics vary very slightly for logistic regression, indicating a very robust result. Looking at the individual results of the clinics in Tables 14, 15, and 16, E-MTAB4032 were predicted as nearly entirely high risk, whereas the predictions of the other two clinics were biased towards low-risk.

To put perspective on the results of logistic regression, we can compare with the competition results. Upon the end of the competition, the highest average AUC obtained for challenge 2 was 0.6954. Granted the results of logistic regression were based on different data, the method to obtain the metrics were mimicked. The results of this analysis were also obtained with less data. A large performance difference shows great potential in the methods performed for this analysis

It is worth noting the decrease in performance from regularized logistic regression when compared to the unregularized counterpart. It is possible the variable selection step already reduced the chance of overfitting and resolved effects of multicollinearity through checking for VIF. Since adding a regularization factor has a built-in function of selecting variables, it is worth considering comparing penalized logistic regression without the variable selection step. The results are shown in Table 17. The best results from this method seem to occur from LASSO, with an average AUC of 65 percent. This score, however, is nearly as poor as the results from random forest after variable selection, which performed the worst compared to the other models. This is potentially because 4,503 variables is far too many predictors relative to the small sample size for regularized logistic regression to be able to handle the issues of overfitting and multicollinearity.

CHAPTER 7

CONCLUSION

The microarray data consisting of 17,032 markers has been filtered to 4,503 through variable screening with Accelerated Failure Time model. With random forest, a list of variables ordered by importance was generated, in which the variables were tested one at a time through likelihood ratio test, yielding 43 potentially significant variables. After checking for VIF and performing backward stepwise selection, the number of variables was reduced to just 38. Multiple models were considered to compare amongst each other to gauge the best performing model that best discriminates high risk patients from others. Through testing multiple models, it is shown that the best model to generalize the data is logistic regression balanced with downsampling. Comparing with multiple models gives a general relative power of logistic regression, which is good in verifying the efficacy of the results.

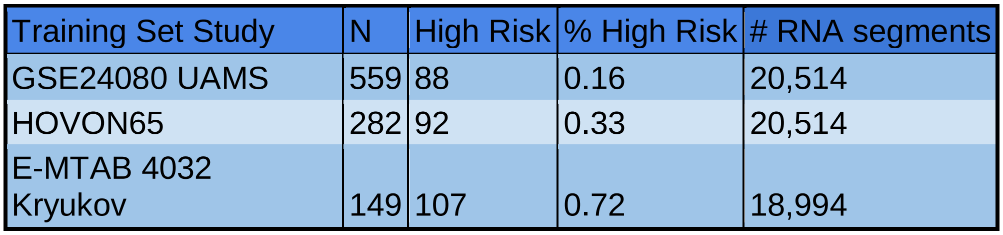
A difficult barrier in having consistent data arises from the nature of how expression data is obtained. Different procedures can be performed in the multistep process in obtaining expression levels of DNA sequences. For future improvements in model performance, alternative methods to combine multiple clinic data can be considered. One proposed method is looking into ways to reduce bias from utilizing different platforms such that just the variability from the expression data is left (Kim et al. 2007). There are currently new methods being tested and developed to reduce variance of expression levels associated to the individual clinics to create an ideal aggregated dataset for model development.

APPENDIX

TABLES AND FIGURES

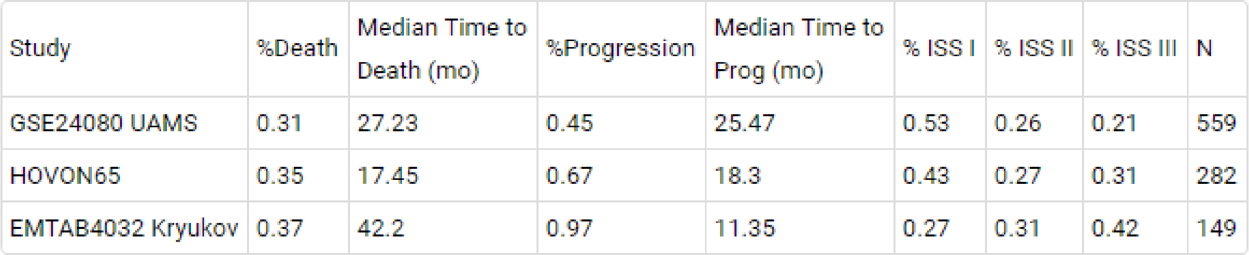
LIST OF TABLES

TABLE 1. Variant Sample Size and High Risk Proportions



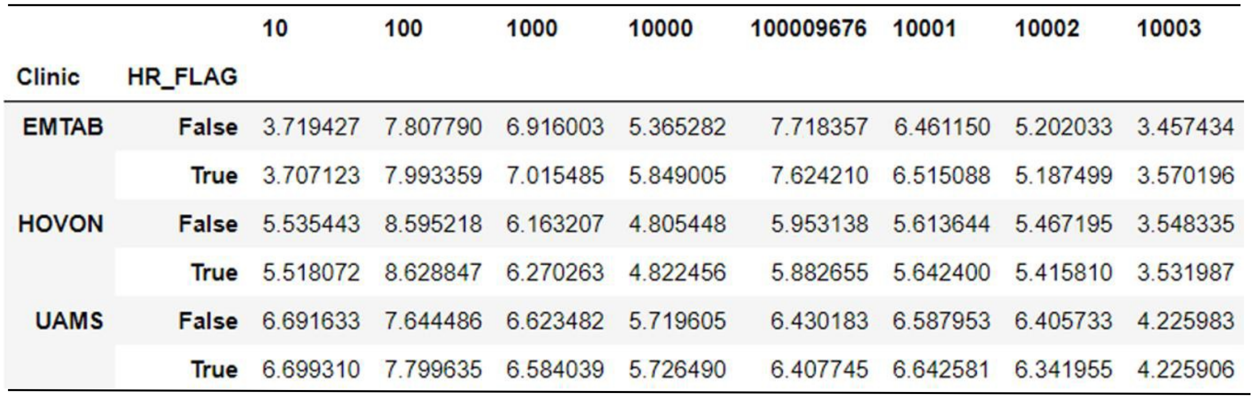
Training Set Study	N	High Risk	% High Risk	# RNA segments
GSE24080 UAMS	559	88	0.16	20,514
HOVON65	282	92	0.33	20,514
E-MTAB 4032 Kryukov	149	107	0.72	18,994

TABLE 2. Variant Death Rate, Progression Rate, and Stages of Cancer



Study	%Death	Median Time to Death (mo)	%Progression	Median Time to Prog (mo)	% ISS I	% ISS II	% ISS III	N
GSE24080 UAMS	0.31	27.23	0.45	25.47	0.53	0.26	0.21	559
HOVON65	0.35	17.45	0.67	18.3	0.43	0.27	0.31	282
EMTAB4032 Kryukov	0.37	42.2	0.97	11.35	0.27	0.31	0.42	149

TABLE 3. Average Expression Value Across Each Clinic, for High/Low Risk Patients



		10	100	1000	10000	100009676	10001	10002	10003
Clinic	HR_FLAG								
EMTAB	False	3.719427	7.807790	6.916003	5.365282	7.718357	6.461150	5.202033	3.457434
	True	3.707123	7.993359	7.015485	5.849005	7.624210	6.515088	5.187499	3.570196
HOVON	False	5.535443	8.595218	6.163207	4.805448	5.953138	5.613644	5.467195	3.548335
	True	5.518072	8.628847	6.270263	4.822456	5.882655	5.642400	5.415810	3.531987
UAMS	False	6.691633	7.644486	6.623482	5.719605	6.430183	6.587953	6.405733	4.225983
	True	6.699310	7.799635	6.584039	5.726490	6.407745	6.642581	6.341955	4.225906

TABLE 4. Final Selected Variables from Variable Selection Step

X8407	X8852	X84333	X57733	X784
X114569	X150967	X10170	X3797	X9764
X26147	X151963	X8349	X146802	X400657
X284837	X2030	X1992	X84898	X939
X112483	X22987	X79576	X7813	X7639
X4060	X64130	X283848	X7291	X80823
X79000	X91445	X84707	X51099	
X58515	X158431	X128826	X871	

TABLE 5. Results on E-MTAB 4032 Using HOVON65 and GSE24080 UAMS to Build Model

EMTAB						
Model	**ACC**	**SEN**	**SPE**	**PPV**	**NPV**	**AUC**
Random Forest + downsample	0.34	0.53	0.27	0.20	0.62	0.66
Random Forest + upsample	0.73	0.00	0.99	0.00	0.74	0.61
Logistic Regression + downsampling (34 markers)	0.74	0.03	1.00	1.00	0.74	0.81
Logistic Regression + upsampling (34 markers)	0.74	0.03	1.00	1.00	0.74	0.77
glmnet (alpha = 0) + upsampling	0.74	1.00	0.03	0.74	1.00	0.75
glmnet (alpha = 0.2) + upsampling	0.76	1.00	0.08	0.75	1.00	0.74
glmnet (alpha = 0.4) + upsampling	0.78	1.00	0.16	0.77	1.00	0.73
glmnet (alpha = 0.6) + upsampling	0.78	1.00	0.16	0.77	1.00	0.72
glmnet (alpha = 0.8) + upsampling	0.78	1.00	0.16	0.77	1.00	0.72
glmnet (alpha = 1) + upsampling	0.77	1.00	0.11	0.76	1.00	0.71
glmnet (alpha = 0) + downsampling	0.74	0.81	0.55	0.84	0.51	0.76
glmnet (alpha = 0.2) + downsampling	0.69	0.72	0.61	0.84	0.43	0.74
glmnet (alpha = 0.4) + downsampling	0.62	0.58	0.74	0.86	0.38	0.73
glmnet (alpha = 0.6) + downsampling	0.58	0.51	0.76	0.86	0.36	0.71
glmnet (alpha = 0.8) + downsampling	0.61	0.57	0.74	0.86	0.38	0.71
glmnet (alpha = 1) + downsampling	0.65	0.64	0.66	0.84	0.40	0.71
SVM with linear kernel	0.75	0.05	1.00	1.00	0.75	0.80

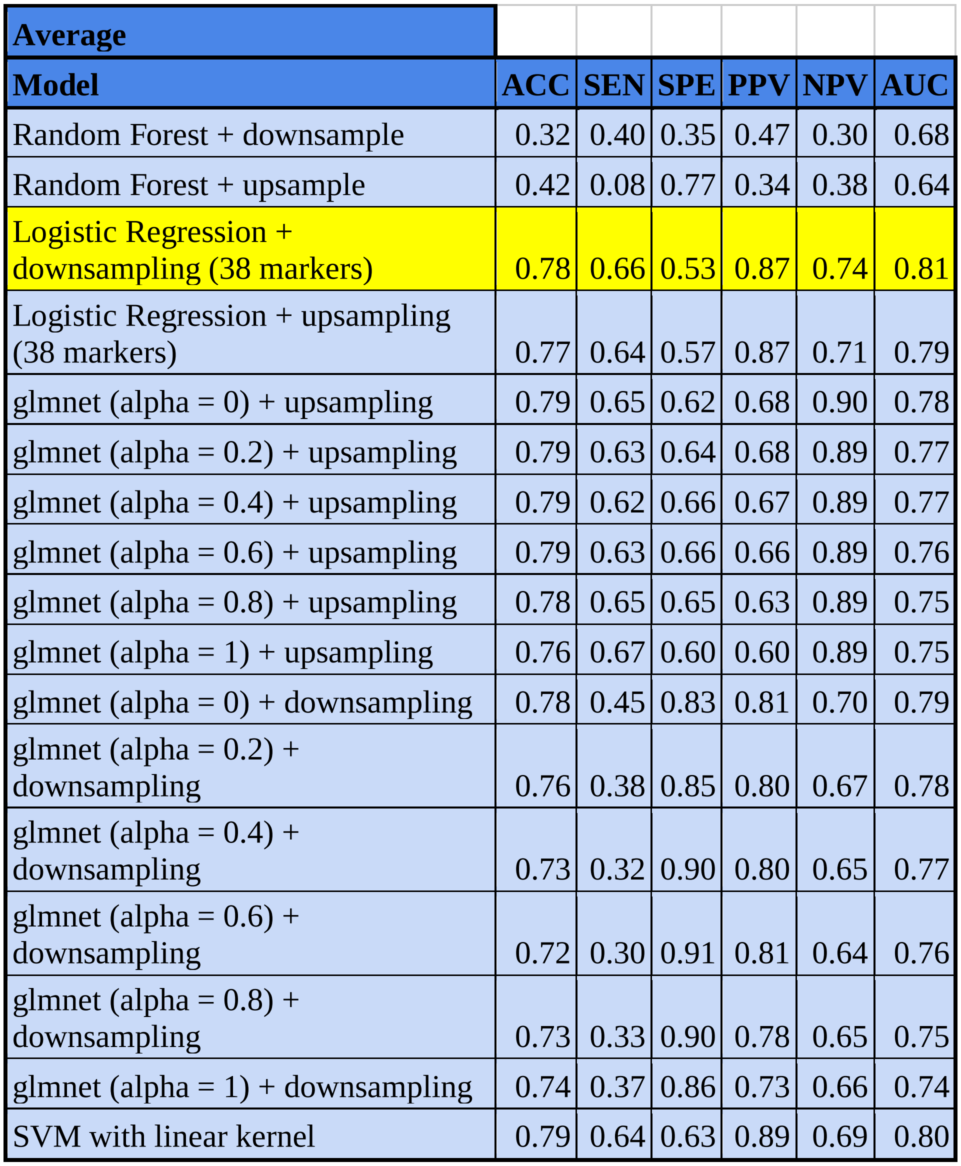
TABLE 6. Results on HOVON65 Using E-MTAB 4032 and GSE24080 UAMS to Build Model

HOVON						
Model	**ACC**	**SEN**	**SPE**	**PPV**	**NPV**	**AUC**
Random Forest + downsample	0.43	0.55	0.20	0.57	0.18	0.68
Random Forest + upsample	0.30	0.06	0.77	0.34	0.29	0.66
Logistic Regression + downsampling (34 markers)	0.73	0.97	0.25	0.72	0.82	0.82
Logistic Regression + upsampling (34 markers)	0.72	0.99	0.18	0.71	0.89	0.82
glmnet (alpha = 0) + upsampling	0.77	0.55	0.88	0.71	0.80	0.82
glmnet (alpha = 0.2) + upsampling	0.77	0.51	0.90	0.71	0.78	0.81
glmnet (alpha = 0.4) + upsampling	0.76	0.49	0.89	0.69	0.78	0.80
glmnet (alpha = 0.6) + upsampling	0.76	0.49	0.90	0.70	0.78	0.80
glmnet (alpha = 0.8) + upsampling	0.76	0.49	0.90	0.71	0.78	0.80
glmnet (alpha = 1) + upsampling	0.75	0.45	0.91	0.71	0.76	0.79
glmnet (alpha = 0) + downsampling	0.73	0.26	0.96	0.77	0.72	0.82
glmnet (alpha = 0.2) + downsampling	0.71	0.22	0.96	0.74	0.71	0.81
glmnet (alpha = 0.4) + downsampling	0.70	0.20	0.96	0.69	0.70	0.80
glmnet (alpha = 0.6) + downsampling	0.72	0.23	0.96	0.75	0.71	0.79
glmnet (alpha = 0.8) + downsampling	0.71	0.22	0.96	0.74	0.71	0.78
glmnet (alpha = 1) + downsampling	0.71	0.21	0.97	0.76	0.71	0.77
SVM with linear kernel	0.77	0.92	0.48	0.78	0.76	0.81

TABLE 7. Results on GSE24080 UAMS Using E-MTAB 4032 and HOVON65 to Build Model

UAMS						
Model	**ACC**	**SEN**	**SPE**	**PPV**	**NPV**	**AUC**
Random Forest + downsample	0.21	0.14	0.58	0.63	0.11	0.69
Random Forest + upsample	0.24	0.18	0.56	0.68	0.11	0.66
Logistic Regression + downsampling (34 markers)	0.87	0.97	0.33	0.89	0.66	0.80
Logistic Regression + upsampling (34 markers)	0.84	0.90	0.52	0.91	0.51	0.79
glmnet (alpha = 0) + upsampling	0.86	0.40	0.94	0.57	0.89	0.78
glmnet (alpha = 0.2) + upsampling	0.86	0.38	0.95	0.58	0.89	0.77
glmnet (alpha = 0.4) + upsampling	0.85	0.36	0.94	0.53	0.89	0.76
glmnet (alpha = 0.6) + upsampling	0.84	0.40	0.93	0.50	0.89	0.76
glmnet (alpha = 0.8) + upsampling	0.81	0.45	0.88	0.41	0.90	0.75
glmnet (alpha = 1) + upsampling	0.76	0.56	0.80	0.34	0.91	0.74
glmnet (alpha = 0) + downsampling	0.87	0.26	0.99	0.82	0.88	0.79
glmnet (alpha = 0.2) + downsampling	0.87	0.19	0.99	0.81	0.87	0.78
glmnet (alpha = 0.4) + downsampling	0.86	0.17	0.99	0.83	0.86	0.77
glmnet (alpha = 0.6) + downsampling	0.86	0.17	0.99	0.83	0.86	0.76
glmnet (alpha = 0.8) + downsampling	0.86	0.19	0.99	0.74	0.87	0.75
glmnet (alpha = 1) + downsampling	0.85	0.26	0.96	0.58	0.87	0.74
SVM with linear kernel	0.86	0.94	0.42	0.90	0.56	0.79

TABLE 8. Average Metrics of Results from Three Clinics



Average						
Model	**ACC**	**SEN**	**SPE**	**PPV**	**NPV**	**AUC**
Random Forest + downsample	0.32	0.40	0.35	0.47	0.30	0.68
Random Forest + upsample	0.42	0.08	0.77	0.34	0.38	0.64
Logistic Regression + downsampling (38 markers)	0.78	0.66	0.53	0.87	0.74	0.81
Logistic Regression + upsampling (38 markers)	0.77	0.64	0.57	0.87	0.71	0.79
glmnet (alpha = 0) + upsampling	0.79	0.65	0.62	0.68	0.90	0.78
glmnet (alpha = 0.2) + upsampling	0.79	0.63	0.64	0.68	0.89	0.77
glmnet (alpha = 0.4) + upsampling	0.79	0.62	0.66	0.67	0.89	0.77
glmnet (alpha = 0.6) + upsampling	0.79	0.63	0.66	0.66	0.89	0.76
glmnet (alpha = 0.8) + upsampling	0.78	0.65	0.65	0.63	0.89	0.75
glmnet (alpha = 1) + upsampling	0.76	0.67	0.60	0.60	0.89	0.75
glmnet (alpha = 0) + downsampling	0.78	0.45	0.83	0.81	0.70	0.79
glmnet (alpha = 0.2) + downsampling	0.76	0.38	0.85	0.80	0.67	0.78
glmnet (alpha = 0.4) + downsampling	0.73	0.32	0.90	0.80	0.65	0.77
glmnet (alpha = 0.6) + downsampling	0.72	0.30	0.91	0.81	0.64	0.76
glmnet (alpha = 0.8) + downsampling	0.73	0.33	0.90	0.78	0.65	0.75
glmnet (alpha = 1) + downsampling	0.74	0.37	0.86	0.73	0.66	0.74
SVM with linear kernel	0.79	0.64	0.63	0.89	0.69	0.80

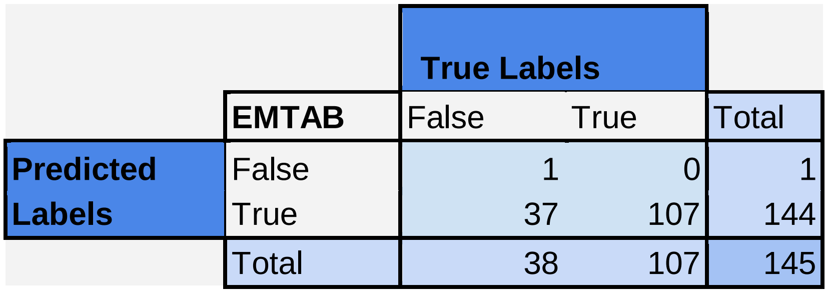
TABLE 9. Standard Deviation of the Metrics from the Results of the Three Clinics

Standard Deviation						
Model	**ACC**	**SEN**	**SPE**	**PPV**	**NPV**	**AUC**
Random Forest + downsample	0.11	0.23	0.20	0.23	0.27	0.02
Random Forest + upsample	0.27	0.09	0.22	0.34	0.32	0.03
Logistic Regression + downsampling (38 markers)	0.08	0.55	0.41	0.14	0.08	0.01
Logistic Regression + upsampling (38 markers)	0.07	0.53	0.41	0.15	0.20	0.02
glmnet (alpha = 0) + upsampling	0.06	0.31	0.51	0.09	0.10	0.03
glmnet (alpha = 0.2) + upsampling	0.06	0.33	0.49	0.09	0.11	0.03
glmnet (alpha = 0.4) + upsampling	0.05	0.34	0.44	0.12	0.11	0.04
glmnet (alpha = 0.6) + upsampling	0.04	0.32	0.43	0.14	0.11	0.04
glmnet (alpha = 0.8) + upsampling	0.02	0.31	0.42	0.19	0.11	0.04
glmnet (alpha = 1) + upsampling	0.01	0.29	0.44	0.23	0.12	0.04
glmnet (alpha = 0) + downsampling	0.08	0.32	0.24	0.03	0.18	0.03
glmnet (alpha = 0.2) + downsampling	0.10	0.30	0.21	0.05	0.22	0.03
glmnet (alpha = 0.4) + downsampling	0.12	0.23	0.14	0.09	0.24	0.04
glmnet (alpha = 0.6) + downsampling	0.14	0.18	0.12	0.06	0.26	0.04
glmnet (alpha = 0.8) + downsampling	0.13	0.21	0.14	0.07	0.25	0.04
glmnet (alpha = 1) + downsampling	0.10	0.24	0.18	0.14	0.24	0.03
SVM with linear kernel	0.06	0.51	0.32	0.11	0.11	0.01

TABLE 10. Best Results Obtained by Logistic Regression + Downsampling

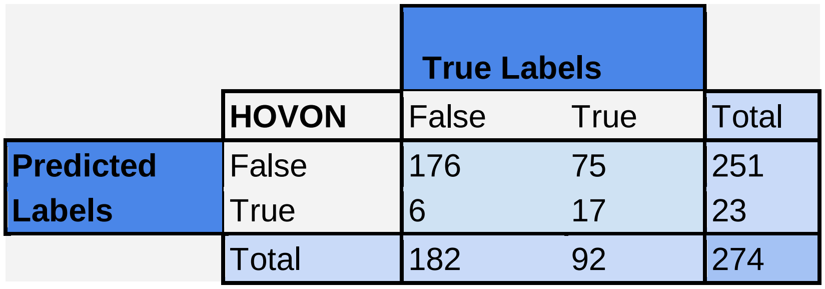
Metrics	EMTAB	HOVON	UAMS	Average	St. D.
ACC	0.74	0.87	0.87	0.78	0.08
SEN	0.03	0.97	0.97	0.66	0.55
SPE	1.00	0.33	0.33	0.53	0.41
PPV	1.00	0.89	0.89	0.87	0.14
NPV	0.74	0.66	0.66	0.74	0.08
AUC	0.81	0.80	0.80	0.81	0.01

TABLE 11. Contingency Table of E-MTAB4032 Predictions from Logistic Regression + Downsampling



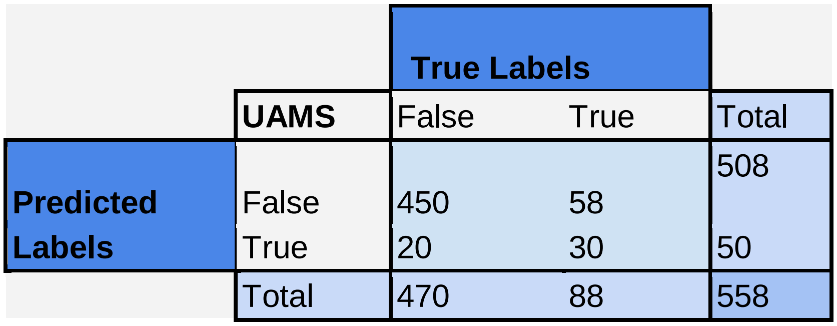
		True Labels		
	EMTAB	False	True	Total
Predicted Labels	False	1	0	1
	True	37	107	144
	Total	38	107	145

TABLE 12. Contingency Table of HOVON65 Predictions from Logistic Regression + Downsampling



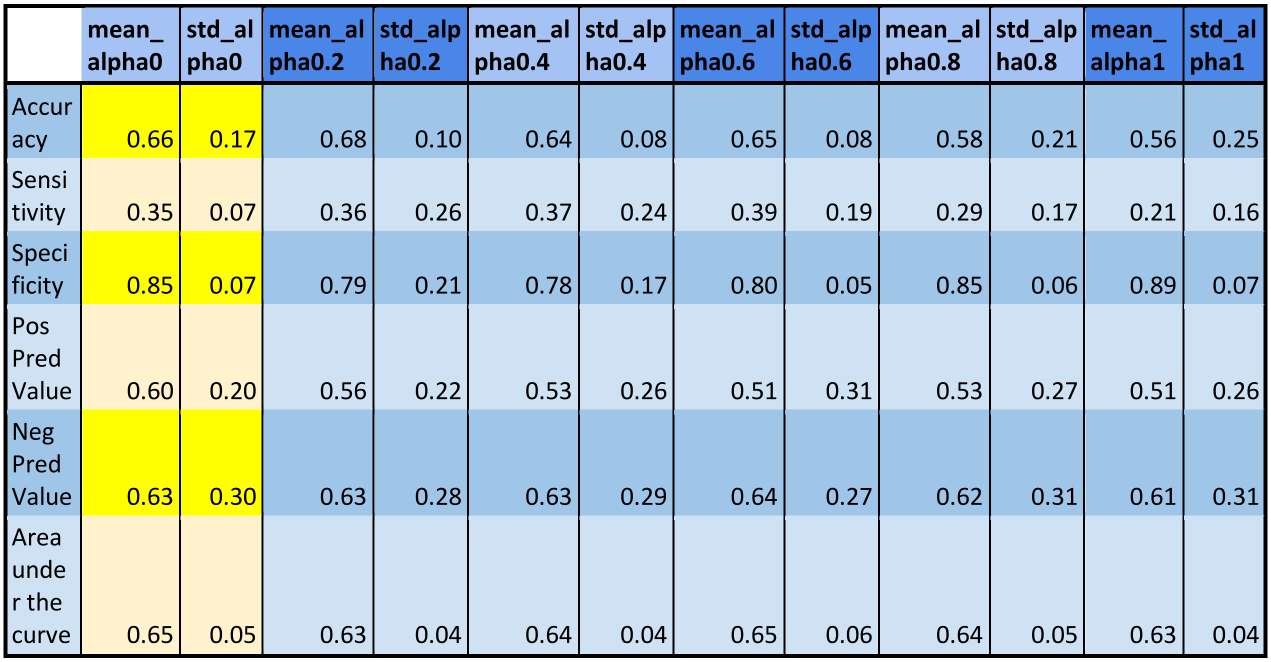
		True Labels		
	HOVON	False	True	Total
Predicted Labels	False	176	75	251
	True	6	17	23
	Total	182	92	274

TABLE 13. Contingency Table of GSE64080 UAMS Predictions from Logistic Regression + Downsampling



		True Labels		
	UAMS	False	True	Total
Predicted Labels	False	450	58	508
	True	20	30	50
	Total	470	88	558

TABLE 14. Results from Performing Regularized Logistic Regression after Variable Screening



	mean_alpha0	std_alpha0	mean_alpha0.2	std_alpha0.2	mean_alpha0.4	std_alpha0.4	mean_alpha0.6	std_alpha0.6	mean_alpha0.8	std_alpha0.8	mean_alpha1	std_alpha1
Accuracy	0.66	0.17	0.68	0.10	0.64	0.08	0.65	0.08	0.58	0.21	0.56	0.25
Sensitivity	0.35	0.07	0.36	0.26	0.37	0.24	0.39	0.19	0.29	0.17	0.21	0.16
Specificity	0.85	0.07	0.79	0.21	0.78	0.17	0.80	0.05	0.85	0.06	0.89	0.07
Pos Pred Value	0.60	0.20	0.56	0.22	0.53	0.26	0.51	0.31	0.53	0.27	0.51	0.26
Neg Pred Value	0.63	0.30	0.63	0.28	0.63	0.29	0.64	0.27	0.62	0.31	0.61	0.31
Area under the curve	0.65	0.05	0.63	0.04	0.64	0.04	0.65	0.06	0.64	0.05	0.63	0.04

LIST OF FIGURES

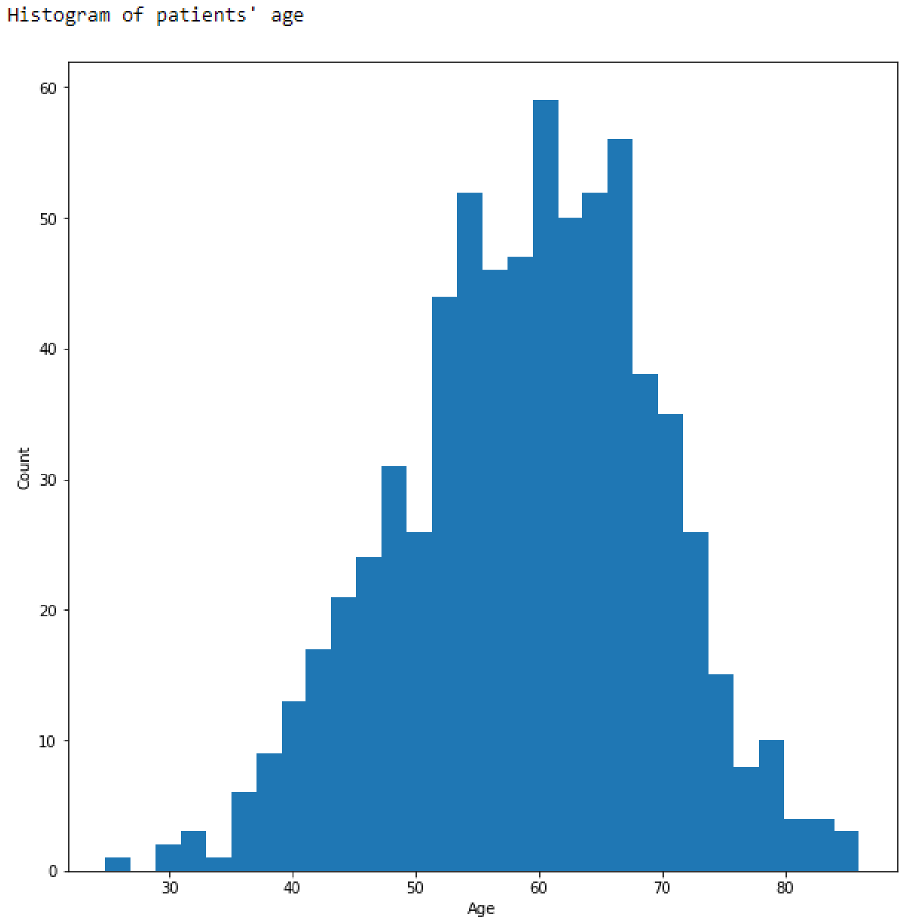
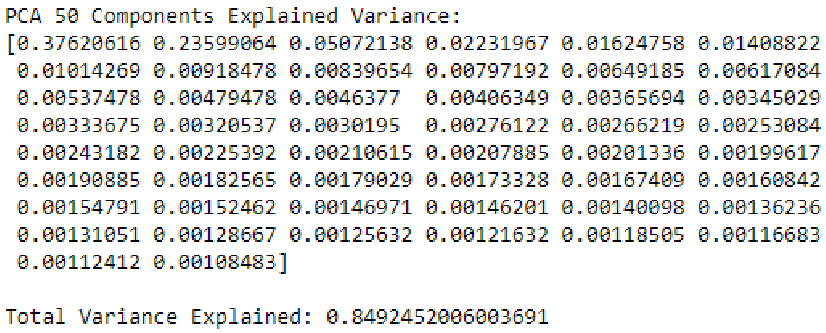
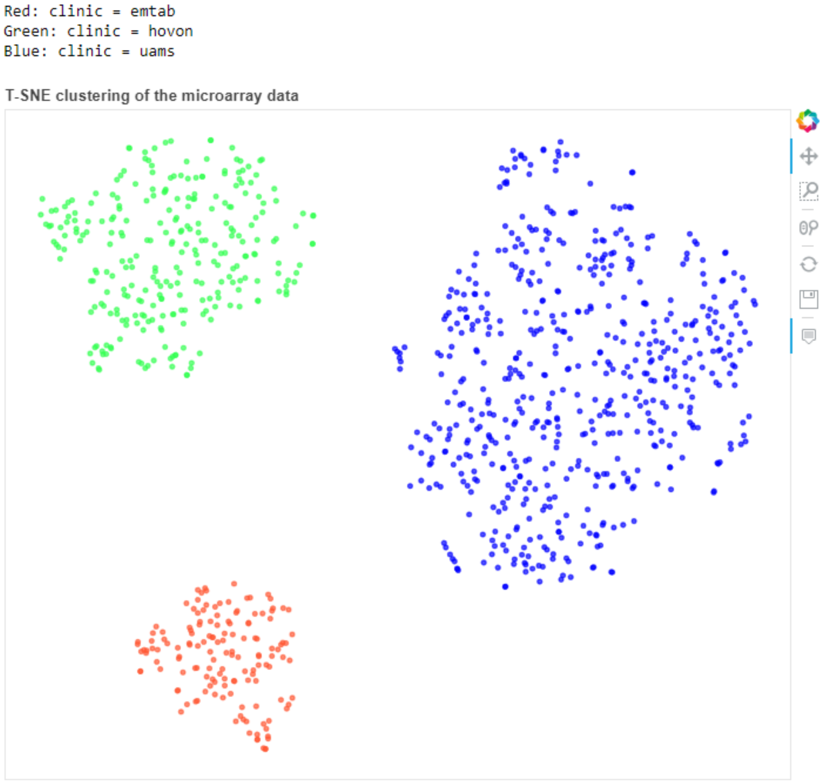


FIGURE 1. Histogram of patients' age across all clinics.



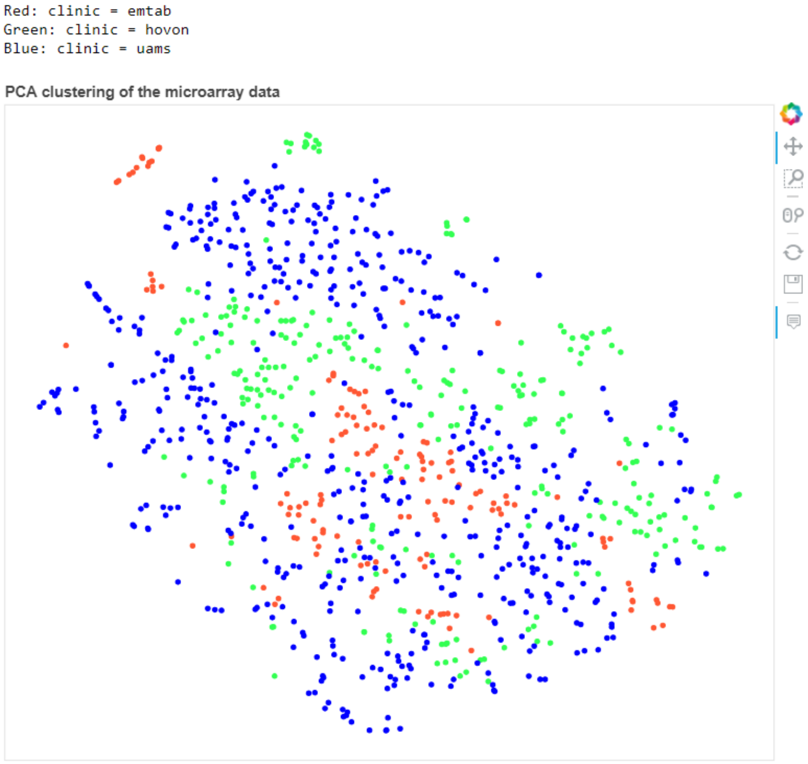
```
PCA 50 Components Explained Variance:
[0.37620616 0.23599064 0.05072138 0.02231967 0.01624758 0.01408822
 0.01014269 0.00918478 0.00839654 0.00797192 0.00649185 0.00617084
 0.00537478 0.00479478 0.0046377  0.00406349 0.00365694 0.00345029
 0.00333675 0.00320537 0.0030195  0.00276122 0.00266219 0.00253084
 0.00243182 0.00225392 0.00210615 0.00207885 0.00201336 0.00199617
 0.00190885 0.00182565 0.00179029 0.00173328 0.00167409 0.00160842
 0.00154791 0.00152462 0.00146971 0.00146201 0.00140098 0.00136236
 0.00131051 0.00128667 0.00125632 0.00121632 0.00118505 0.00116683
 0.00112412 0.00108483]

Total Variance Explained: 0.8492452006003691
```

FIGURE 2. Total variance explained on first 50 components from PCA.



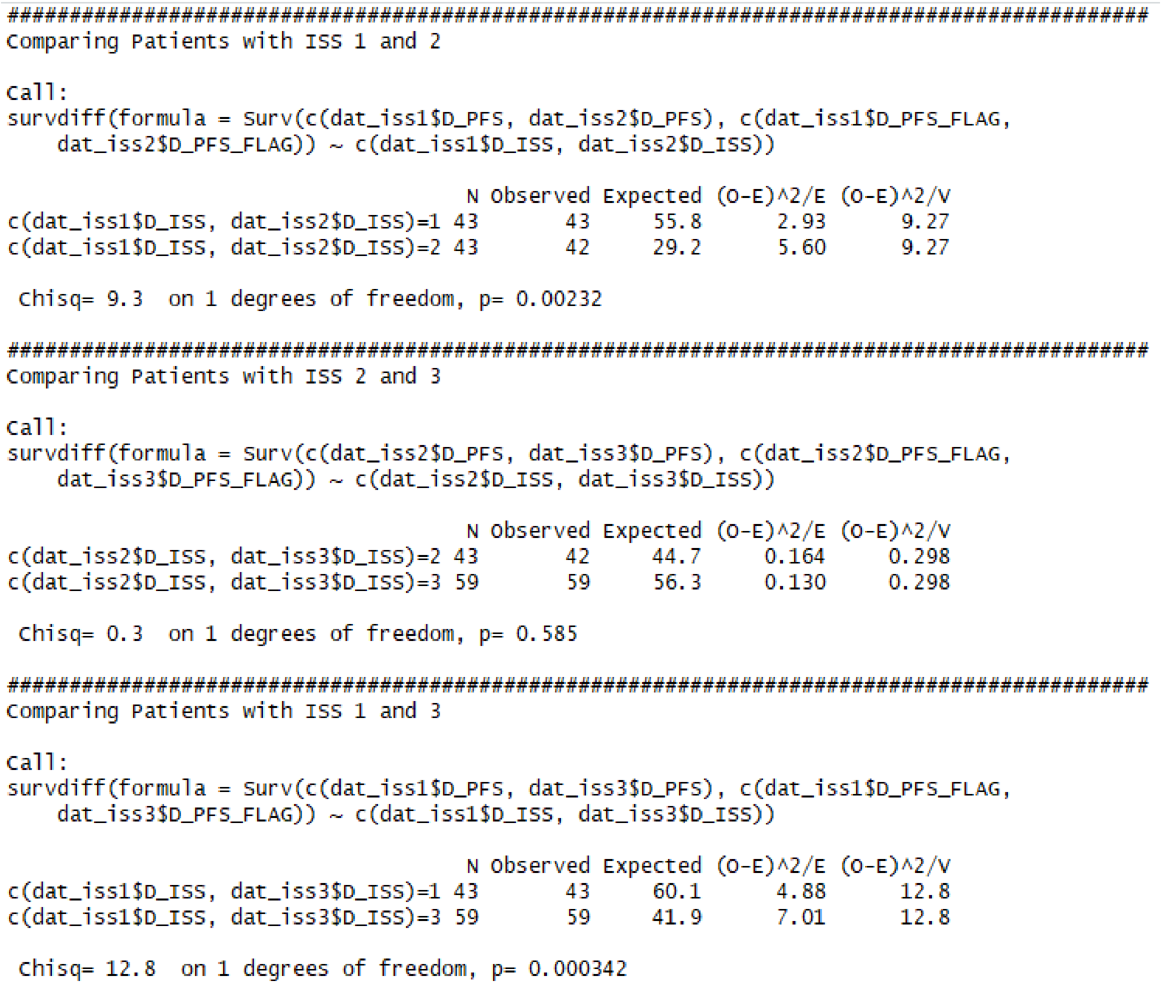
```
Red: clinic = emtab
Green: clinic = hovon
Blue: clinic = uams
```

FIGURE 3. t-SNE two components mapped onto two dimensional graph.



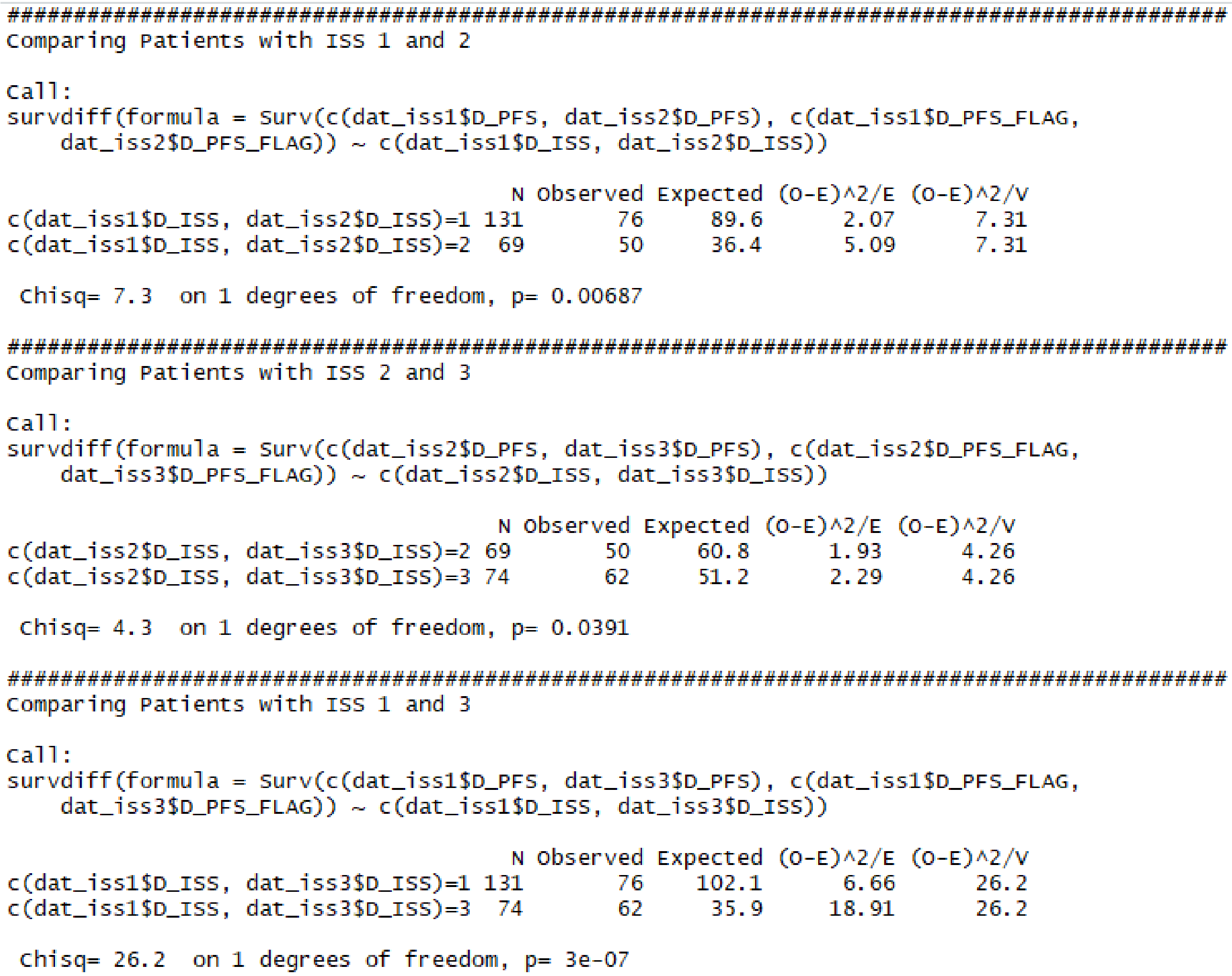
```
Red: clinic = emtab
Green: clinic = hovon
Blue: clinic = uams
```

FIGURE 4. t-SNE graph with values standardized to mean=0 and standard deviation=1.



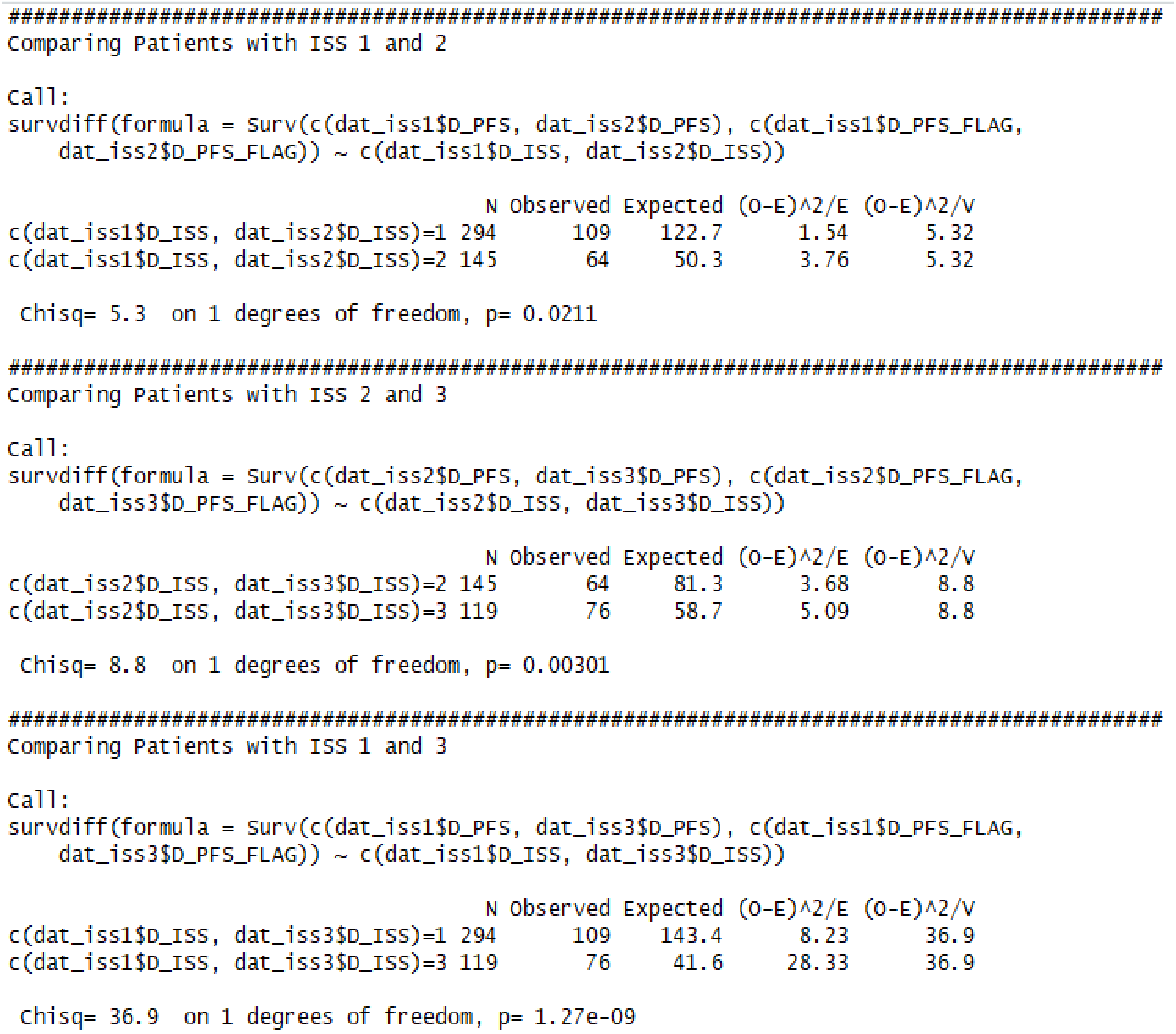
```
###########################################################################################
Comparing Patients with ISS 1 and 2

Call:
survdiff(formula = Surv(c(dat_iss1$D_PFS, dat_iss2$D_PFS), c(dat_iss1$D_PFS_FLAG, 
    dat_iss2$D_PFS_FLAG)) ~ c(dat_iss1$D_ISS, dat_iss2$D_ISS))

                                     N Observed Expected (O-E)^2/E (O-E)^2/V
c(dat_iss1$D_ISS, dat_iss2$D_ISS)=1 43       43     55.8      2.93      9.27
c(dat_iss1$D_ISS, dat_iss2$D_ISS)=2 43       42     29.2      5.60      9.27

 Chisq= 9.3  on 1 degrees of freedom, p= 0.00232 

###########################################################################################
Comparing Patients with ISS 2 and 3

Call:
survdiff(formula = Surv(c(dat_iss2$D_PFS, dat_iss3$D_PFS), c(dat_iss2$D_PFS_FLAG, 
    dat_iss3$D_PFS_FLAG)) ~ c(dat_iss2$D_ISS, dat_iss3$D_ISS))

                                     N Observed Expected (O-E)^2/E (O-E)^2/V
c(dat_iss2$D_ISS, dat_iss3$D_ISS)=2 43       42     44.7     0.164     0.298
c(dat_iss2$D_ISS, dat_iss3$D_ISS)=3 59       59     56.3     0.130     0.298

 Chisq= 0.3  on 1 degrees of freedom, p= 0.585 

###########################################################################################
Comparing Patients with ISS 1 and 3

Call:
survdiff(formula = Surv(c(dat_iss1$D_PFS, dat_iss3$D_PFS), c(dat_iss1$D_PFS_FLAG, 
    dat_iss3$D_PFS_FLAG)) ~ c(dat_iss1$D_ISS, dat_iss3$D_ISS))

                                     N Observed Expected (O-E)^2/E (O-E)^2/V
c(dat_iss1$D_ISS, dat_iss3$D_ISS)=1 43       43     60.1      4.88      12.8
c(dat_iss1$D_ISS, dat_iss3$D_ISS)=3 59       59     41.9      7.01      12.8

 Chisq= 12.8  on 1 degrees of freedom, p= 0.000342 
```

FIGURE 5. Results of E-MTAB Log-Rank test on each ISS stage.



```
###########################################################################################
Comparing Patients with ISS 1 and 2

Call:
survdiff(formula = Surv(c(dat_iss1$D_PFS, dat_iss2$D_PFS), c(dat_iss1$D_PFS_FLAG, 
    dat_iss2$D_PFS_FLAG)) ~ c(dat_iss1$D_ISS, dat_iss2$D_ISS))

                                      N Observed Expected (O-E)^2/E (O-E)^2/V
c(dat_iss1$D_ISS, dat_iss2$D_ISS)=1 131       76     89.6      2.07      7.31
c(dat_iss1$D_ISS, dat_iss2$D_ISS)=2  69       50     36.4      5.09      7.31

 Chisq= 7.3  on 1 degrees of freedom, p= 0.00687 

###########################################################################################
Comparing Patients with ISS 2 and 3

Call:
survdiff(formula = Surv(c(dat_iss2$D_PFS, dat_iss3$D_PFS), c(dat_iss2$D_PFS_FLAG, 
    dat_iss3$D_PFS_FLAG)) ~ c(dat_iss2$D_ISS, dat_iss3$D_ISS))

                                     N Observed Expected (O-E)^2/E (O-E)^2/V
c(dat_iss2$D_ISS, dat_iss3$D_ISS)=2 69       50     60.8      1.93      4.26
c(dat_iss2$D_ISS, dat_iss3$D_ISS)=3 74       62     51.2      2.29      4.26

 Chisq= 4.3  on 1 degrees of freedom, p= 0.0391 

###########################################################################################
Comparing Patients with ISS 1 and 3

Call:
survdiff(formula = Surv(c(dat_iss1$D_PFS, dat_iss3$D_PFS), c(dat_iss1$D_PFS_FLAG, 
    dat_iss3$D_PFS_FLAG)) ~ c(dat_iss1$D_ISS, dat_iss3$D_ISS))

                                      N Observed Expected (O-E)^2/E (O-E)^2/V
c(dat_iss1$D_ISS, dat_iss3$D_ISS)=1 131       76    102.1      6.66      26.2
c(dat_iss1$D_ISS, dat_iss3$D_ISS)=3  74       62     35.9     18.91      26.2

 Chisq= 26.2  on 1 degrees of freedom, p= 3e-07 
```

FIGURE 6. Results of HOVON Log-Rank test on each ISS stage.



```
############################################################################################
Comparing Patients with ISS 1 and 2

Call:
survdiff(formula = Surv(c(dat_iss1$D_PFS, dat_iss2$D_PFS), c(dat_iss1$D_PFS_FLAG,
    dat_iss2$D_PFS_FLAG)) ~ c(dat_iss1$D_ISS, dat_iss2$D_ISS))

                                      N Observed Expected (O-E)^2/E (O-E)^2/V
c(dat_iss1$D_ISS, dat_iss2$D_ISS)=1 294      109    122.7      1.54      5.32
c(dat_iss1$D_ISS, dat_iss2$D_ISS)=2 145       64     50.3      3.76      5.32

 Chisq= 5.3  on 1 degrees of freedom, p= 0.0211

############################################################################################
Comparing Patients with ISS 2 and 3

Call:
survdiff(formula = Surv(c(dat_iss2$D_PFS, dat_iss3$D_PFS), c(dat_iss2$D_PFS_FLAG,
    dat_iss3$D_PFS_FLAG)) ~ c(dat_iss2$D_ISS, dat_iss3$D_ISS))

                                      N Observed Expected (O-E)^2/E (O-E)^2/V
c(dat_iss2$D_ISS, dat_iss3$D_ISS)=2 145       64     81.3      3.68       8.8
c(dat_iss2$D_ISS, dat_iss3$D_ISS)=3 119       76     58.7      5.09       8.8

 Chisq= 8.8  on 1 degrees of freedom, p= 0.00301

############################################################################################
Comparing Patients with ISS 1 and 3

Call:
survdiff(formula = Surv(c(dat_iss1$D_PFS, dat_iss3$D_PFS), c(dat_iss1$D_PFS_FLAG,
    dat_iss3$D_PFS_FLAG)) ~ c(dat_iss1$D_ISS, dat_iss3$D_ISS))

                                      N Observed Expected (O-E)^2/E (O-E)^2/V
c(dat_iss1$D_ISS, dat_iss3$D_ISS)=1 294      109    143.4      8.23      36.9
c(dat_iss1$D_ISS, dat_iss3$D_ISS)=3 119       76     41.6     28.33      36.9

 Chisq= 36.9  on 1 degrees of freedom, p= 1.27e-09
```

FIGURE 7. Results of UAMS Log-Rank Test on each ISS stage.

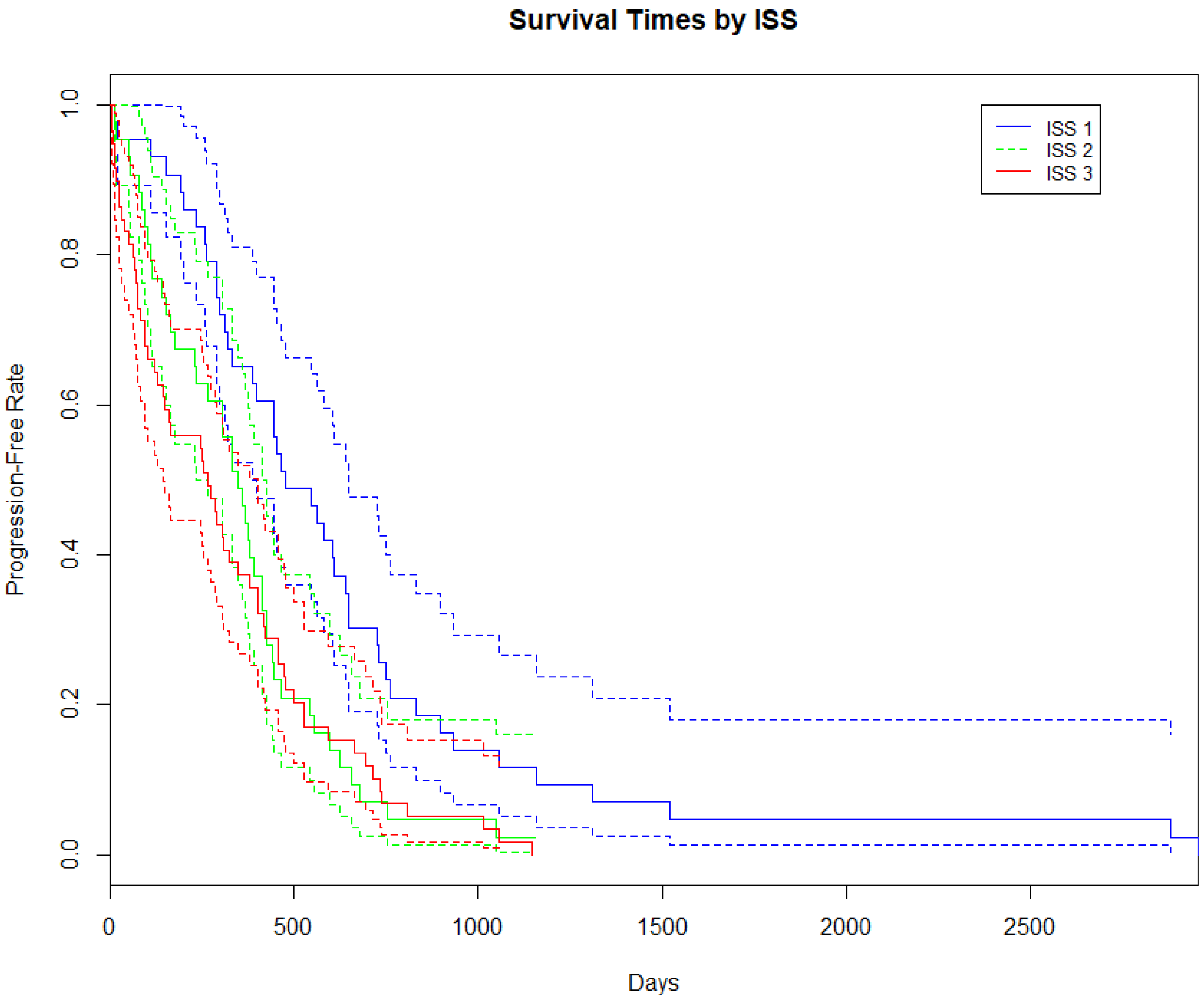


FIGURE 8. E-MTAB 4032 patient progression-free survival rate.

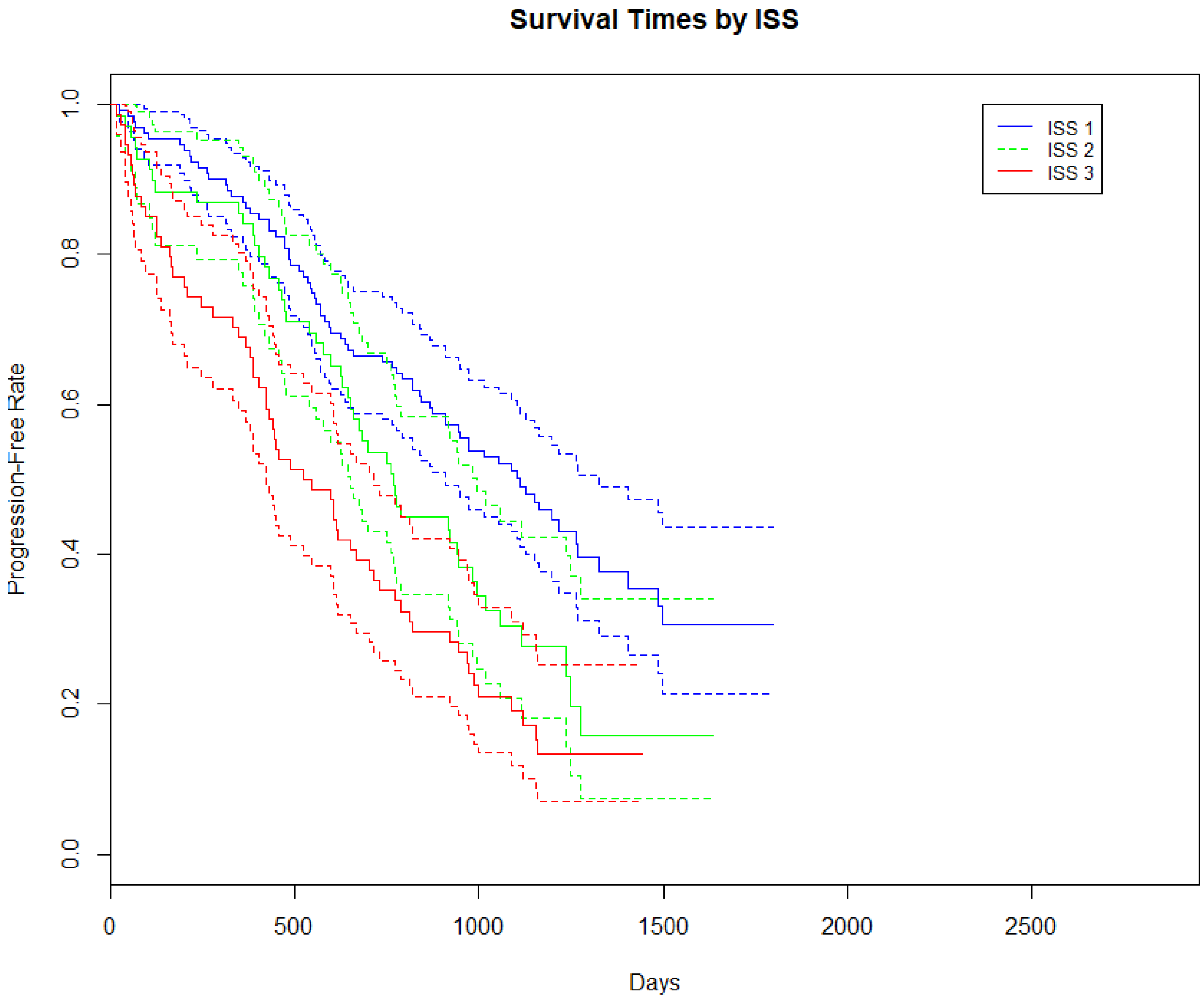


FIGURE 9. HOVON65 patient progression-free survival rate.

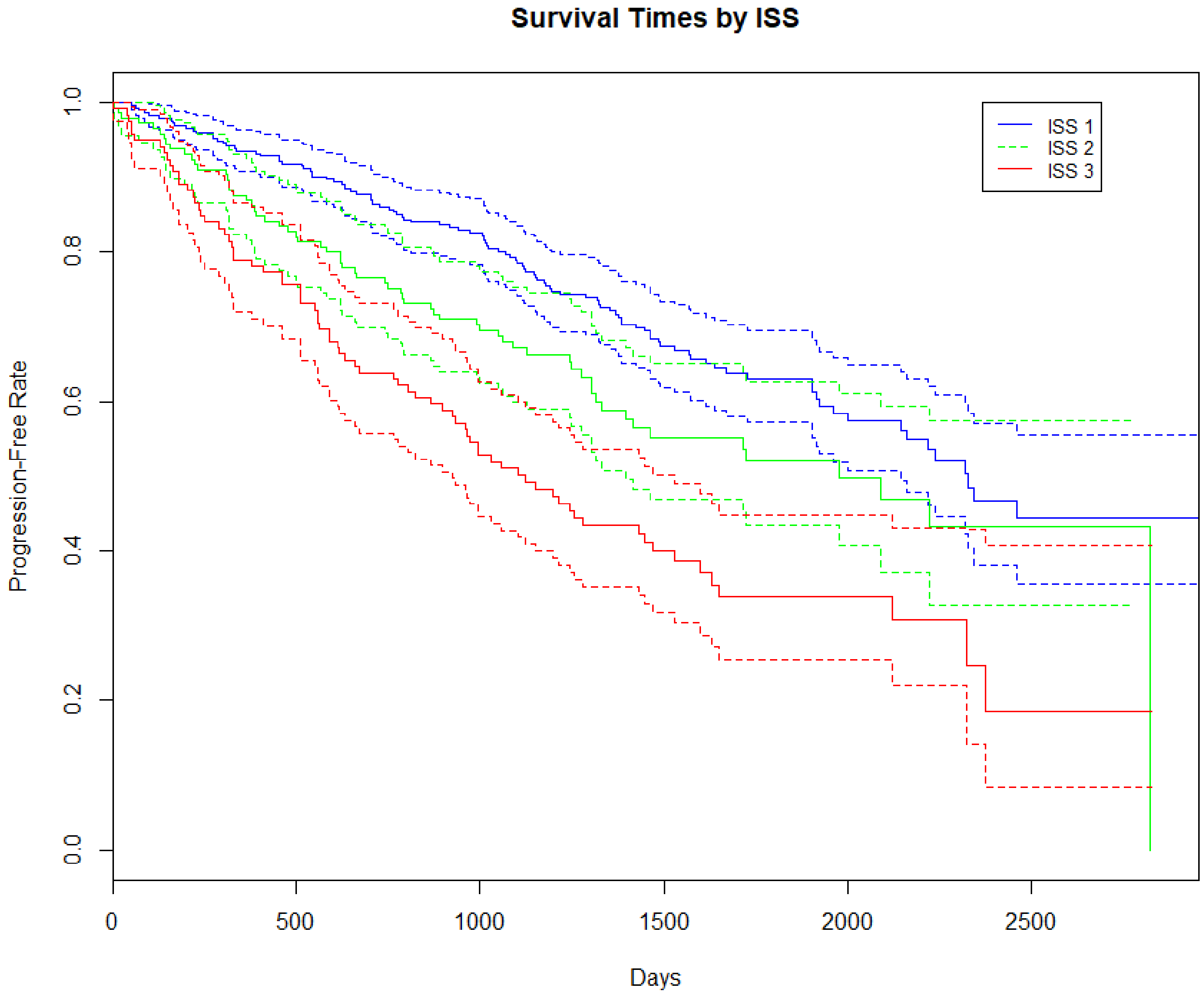


FIGURE 10. GSE24080 UAMS patient progression-free survival rate.

REFERENCES

REFERENCES

Alexander, Dominik D., Pamela J. Mink, Hans-Olov Adami, Philip Cole, Jack S. Mandel, Martin M. Oken, and Dimitrios Trichopoulos. 2007. “Multiple Myeloma: A Review of the Epidemiologic Literature.” *International Journal of Cancer* 120: 40-61. Accessed July 20, 2018. https://onlinelibrary.wiley.com/doi/epdf/10.1002/ijc.22718.

Boser, Bernhard E.. Guyon, Isabelle M., and Vapnik, Vladimir N. 1992. "A Training Algorithm optimal margin classifiers." In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory – COLT '92*, 144. Accessed July 20, 2018. doi: 10.1145 /130385.130401. ISBN 089791497X.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45, no. 1: 5–32. Accessed July 20, 2018. doi: https://10.1023/A:1010933404324.

Kim, Ki-Yeol, Dong Huk Ki, Ha Jin Jeong, Hei-Cheul Jeung, Hyun Cheol Chung, and Sun Young Rha. 2007. “Novel and Simple Transformation Algorithm for Combining Microarray Data Sets.” *BMC Bioinformatics* 8: 218. Accessed July 20, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1914088/pdf/1471-2105-8-218.pdf.

Van Laar, Ryan, Rachel Flinchum, Nathan Brown, Joseph Ramsey, Sam Riccitelli, Christoph Heuck, Bart Barlogie, and John D. Shaughnessy, Jr. 2014. “Translating a Gene Expression Signature for Multiple Myeloma Prognosis into a Robust High-Throughput Assay for Clinical Use.” *BMC Medical Genomics.* Accessed July 20, 2018. https://bmcmedgenomics.biomedcentral.com/track/pdf/10.1186/1755-8794-7-25?site=bmcmedgenomics.biomedcentral.com.

Wei, L. J. 1992. “The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis.” *Statistics in Medicine*, Accessed July 20, 2018. https://onlinelibrary.wiley.com/doi/epdf/10.1002/sim.4780111409.

Zare, Ali, Mostafa Hosseini, Mahmood Mahmoodi, Kazem Mohammad, Hojjat Zeraati, and Kourosh Holakoie Naieni. 2015. “A Comparison Between Accelerated Failure-Time and Cox Proportional Hazard Models in Analyzing the Survival of Gastric Cancer Patients.” *Iranian Journal of Public Health* 44, no. 8: 1095-1102. Accessed July 20, 2018. www.ncbi.nlm.nih.gov/pmc/articles/PMC4645729/.

Zhu, R., D. Zeng, and M. R. Kosorok. 2015. "Reinforcement Learning Trees." *Journal of the American Statistical Association* 110, no. 512: 1770-1784. Accessed July 20, 2018. doi:10.1080/01621459.2015.1036994. PMC 4760114. PMID 26903687.