A COMPARISON OF LOGISTIC REGRESSION TO RANDOM FORESTS FOR EXPLORING DIFFERENCES IN RISK FACTORS ASSOCIATED WITH STAGE AT DIAGNOSIS BETWEEN BLACK AND WHITE COLON CANCER PATIENTS

by

Ming Geng

BMed, Hunan Medical University, 1992

MMed, Xiang-Ya School of Medicine, Central-South University, China, 2001

Submitted to the Graduate Faculty of

Department of Biostatistics

The Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2006

University of Pittsburgh

Graduate School of Public Health

This thesis was presented

by

Ming Geng

It was defended on

January 31, 2006

and approved by

Sati Mazumdar, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Edmund M. Ricci, PhD, Professor, Department of Behavioral and Community Health Sciences, Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Carol K. Redmond, ScD, Distinguished Service Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

A COMPARISON OF LOGISTIC REGRESSION TO RANDOM FORESTS FOR EXPLORING DIFFERENCES IN RISK FACTORS ASSOCIATED WITH STAGE AT DIAGNOSIS BETWEEN BLACK AND WHITE COLON CANCER PATIENTS

Ming Geng, MS

University of Pittsburgh, 2006

Introduction: Colon cancer is one of the most common malignancies in America. According to the American Cancer Society, blacks have lower survival rate than whites. Many previous studies suggested that it is because blacks were more likely to be diagnosed at a late stage. Hence, understanding the factors associated with colon cancer stage at diagnosis has important public health implication. **Objectives:** The objectives of this study are twofold: 1) To compare logistic regression modeling to Random Forests classification with respect to variables selected and classification accuracy; and 2) To evaluate the factors related to colon cancer stage at diagnosis in a population based study. Many studies have compared Classification and Regression Trees (CART) to logistic regression and found that they have very similar power with respect to the proportion correctly classified and the variables selected. This study extends previous methodological research by comparing the Random Forests classification techniques to logistic regression modeling using a relatively small and incomplete dataset. **Methods and Materials:** The data used in this research were from National Cancer Institute Black/White Cancer Survival Study which had 960 cases of invasive colon cancer. Stage at diagnosis was used as the dependent variable for fitting logistic regression models and Random Forests Classification to multiple potential explanatory variables, which included some missing data. **Results:** Odds ratio (blacks vs. whites) decreased from 1.628 (95%CI: 1.068-2.481) to 1.515 (95% CI: 0.920-2.493) after adjustment was made for patient delay in diagnosis, occupation, histology and grade of tumor. Race became no longer important after these variables were entered in the Random Forests. These four variables were identified as the most important variables associated with racial disparity in colon cancer stage at diagnosis in both logistic regression and Random Forests. The correct classification rate was 47.9% using logistic regression and was 33.9% using Random Forests. **Conclusion:** 1)**.** Logistic regression and Random Forests had very similar power in variable selection. 2). Logistic regression had higher classification accuracy than Random Forests with respect to overall correct classification rate.

TABLE of CONTENTS

LIST of TABLES

LIST of FIGURES

ACKNOWLEDGEMENTS

My foremost thank goes to my adviser Dr. Carol K Redmond. Without her, this thesis would not have been possible. I thank her for her patience, encouragement, and consideration that carried me on through difficult times, and for her insights and suggestions that helped to improve my research skills.

Thanks to Dr. Sati. Mazumdar, who advised me and helped me in various aspects of my thesis, her valuable feedback helped me to improve the thesis in many ways. I also thank Dr. Edmund M. Ricci for careful reading and thoughtful comments.

I also acknowledge the contributions of the National Cancer Institute Surveillance, Epidemiology, and End Results Program and the Black/White Cancer Survival Study Group in the conduct of this important study and thank them for these important data available to me.

Finally, I am forever indebted to my husband, my parents, my brothers and sisters-in-law for their understanding, endless patience and encouragement when it was most required.

1.0 INTRODUCTION

Colon cancer is one of the most common malignancies among men and women in America. The American Cancer Society estimates that there will be about 104,950 new cases in 2005[1]. White patients had a higher survival rate than black colon cancer patients. The survival rate was 64.9% for whites and 54.7% for blacks from 1995 to 2001[2, 3], and this disparity may be widening in recent years since mortality from colon cancer has decreased faster for whites than for blacks[4]. Several studies have found that blacks are more likely to be diagnosed at later stage than their white counterparts and stage at diagnosis is the most important predictor of the racial survival difference [5, 6]. Therefore, determining factors that are associated with stage at diagnosis has great importance in secondary prevention of colon cancer.

In this study, logistic regression and Random Forests are used to explore factors that are associated with colon cancer stage at diagnosis and the extent of this association.

1.1 LOGISTIC REGRESSION

Figure 1 summarizes the advantages and disadvantages of logistic regression and Random Forests. Logistic regression is a very popular analytical tool in epidemiological

Feature	Logistic Regression	Random Forests
Parametric analytic tool	Yes	No
Provides probability outcome , such as odds ratio	Yes	No
Controls confounding	Yes	Yes
Tests interaction	Yes	Yes*
Does not require a special assumption of data set	No	Yes
Robust to outliers	No	Yes
Does not require transformation of continuous variable	No	Yes
Does not encounter numerical problem	No	Yes
Never suffer from over-fitting	No	Yes
Automatically selects important variable	No	Yes
Automatically handles missing values	No	Yes
Can handle imbalance data effectively	No	Yes

Note: * RandomForests™ version 1.0 has not offered this capability yet

Figure 1: Comparison of Logistic Regression and Random Forests

studies. It is widely used both in case-control and cohort studies. The logistic function [$f(n) = 1/(1+e^{-n})$] has several desirable properties. First, its range is always between 0 and 1 when the independent variable varies from -∞ to ∞, so the logistic regression model can be used to model the probability of an individual developing disease. Second, the logistic function has a sigmoid shape which applies to the disease condition. In addition, logistic regression can control confounding and assess interaction very effectively when there are several confounders or the confounder is a continuous variable[7]. Its most attractive advantage may be that the researcher can calculate an odds ratio and its confidence interval directly, so that the results can be interpreted easily. The probability of a given subject developing a disease during a fixed time interval can also be calculated. Logistic regression has also been extended to analyze data in which the dependent variable has more than two levels or the dependent variable is an ordinal variable. Although logistic regression is a useful method, it has four main disadvantages. First, its use is based on a special assumption about the data set [7], namely that the relationship between the mean of dependent variable and a set of independent variables follows a logistic distribution and that the errors are binomially distributed. In reality, however, this assumption may not be true. The assumption of a binomial distribution for the error term may be invalid due to overdispersion in the dependent variable[8]. Second, some data points, termed influential values, may have an undue influence on the overall fit of the model, either on the set of parameter estimates or on a single parameter estimate [8]. Sometimes, an observation may have undue influence so that including a term in the model based on a likelihood ratio test is due solely to that observation. Third, several numerical problems may be encountered when a logistic regression model is fitted. Numerical problems include zero cell count in a contingency table, collinearity between covariates, or a set of covariates that separates the dependent variable completely and the maximum likelihood estimates do not exist[7]. Fourth, if the data set contains a large number of variables, a problem of overfitting may occur. Variable selection is not easy and is time consuming, especially when interactions are taken into account. Although most software packages offer an option for stepwise variable selection method, this method has been criticized by some researchers for several drawbacks. For example, Harrell et al pointed out that a stepwise variable selection method using only a significance level as the criterion for entering a variable does not take into account the multiple comparison problem and may select noise variables that cause a decrease in the predictive power This is true especially in later steps of the variable selection [9,10]. For instance, let α be the type I error. Suppose variable A has no relationship to the outcome, the probability that it is not selected in the model at each step is $(1- \alpha)$, the probability of it remaining outside the model after the nth step is $(1- \alpha)^n$, and the probability of it selected in the model by mistake after the nth step is [1-(1- α)^n] > α, so the type I error increases with the number of comparisons. Furthermore, selecting a noise variable may cause the model to become unstable. That is, small changes in the dataset may result in different variables selected.

1.2 RANDOM FORESTS

An alternative statistical procedure called Classification and Regression Trees (CART) has been developed by Breiman and his colleagues[11]. In CART the root node contains all observations and every node is divided into two children nodes depending on a yes-no answer to a question, such as whether a patient is male or is older than 65 years, until the cases in the same node are homogeneous[12]. CART is easy to use and to interpret, but classification accuracy may be low and the classification may be unstable. Breiman noted "If we change the data a little, the tree picture can change a lot" [13].

Random Forests is an extension of CART to a group of trees. A main advantage of Random Forests as compared with CART is its higher classification accuracy. From a theoretical view Random Forests has several advantages over logistic regression. First, it can select important variables automatically no matter how many variables are used initially[14]. The algorithm which Random Forests uses to select important variables is different from stepwise variable selection in logistic regression. Random Forests estimates a variable's importance based on the margin of cases. Margin is defined as "the proportion of votes for the true class minus the maximum proportion of votes for other classes"[11]. Random Forests always uses 63% of cases to construct each tree and the remaining 37% of cases compose "out-of-bag" (OOB) used to evaluate the performance of each tree. To assess the importance of a variable, one first randomly permutes all values of this variable and runs each tree to the OOB cases with permuted values. One then calculates the difference between the original margin and the new margin for OOB cases. The average decrease of the margin over all OOB cases and all trees is used as the criterion to estimate the importance of the variable[11] whereas in logistic regression the stepwise variable selection method is based on the likelihood ratio test[7]. Second, Random Forests provides methods to handle missing values automatically [14]. Section 3.2.3 presents in detail two methods for handling missing values. Third, it never suffers the problem of overfitting. Breiman states that the test set error rates are monotonically decreasing and converge to a limit as the number of trees increases approaching ∞[13]. Because each tree is constructed using 63% of the dataset selected at random with replacement and each node is split using the best split in a small random sample of available variables (usually a square root of the number of available variables are selected at random as a potential splitter), every tree is constructed at random and is independent from other trees. Therefore, adding trees to the forest does not cause a problem of overfitting[12]. Another reason why Random Forests never suffers this problem is that each tree must be grown to the maximum size and does not need pruning. Because of the algorithm of choosing a splitter for each node mentioned above, with the potential splitter set differently from node to node and the tree grown to full size, an important variable will eventually be in the tree [11]. By combining a large number of these un-pruned trees (usually 500 trees) one can obtain powerful predictors[13]. An experiment has shown that pruning trees hurts the performance of Random Forests[11]. Fourth, it can handle imbalanced data[11, 13, 14]. That is, it is more efficient for a dataset in which the number of cases in the class of interest is small compared with another class. Fifth, it works efficiently in large data sets[11, 13, 14]. The main disadvantage of Random Forests is that one cannot obtain a probability measure, like the odds ratio and its confidence interval. Although variable importance given by Random Forests shows which variables confound, the effect of adjusting for confounding cannot be assessed quantitatively. Another disadvantage of Random Forests is that, unlike decision trees analysis which is a single tree, the model is very complex and the tree structure is in an invisible "black box". Therefore, the relationship between a particular level of a variable and outcome and the extent of the relationship is unknown. Random Forests theoretically offers an approach named "prototypes" which can give information about how a variable is related to the outcome, but Random Forests software (Version 1.0) does not have this ability if the independent variable is a categorical variable. At present the prototype is only available if the predictor is a continuous variable.

There are many studies comparing CART with logistic regression[15, 16]. For example, Rudolfer and Paliouras compared CART with logistic regression using data of carpal tunnel syndrome patients. Delen and Walker compared decision trees with logistic regression for the purpose of predicting breast cancer survivability. Few studies, however, have compared Random Forests with logistic regression in the medical or epidemiological domain. For this reason, the objectives of this study are twofold. The first objective is to compare logistic regression modeling with Random Forests. This objective is evaluated in two ways. The first is by comparison of the variables selected. In addition to three design variables, the dataset contains 18 variables, making variable selection an important consideration. Variable importance can be obtained directly from Random Forests. The extent of change in the odds ratio and in its confidence intervals, as well as improvement of model fit after adjusting for a certain variables, are used as criteria to assess variable importance in a logistic regression model. The second aspect of comparison between the two methods is the classification accuracy. The overall correct classification rate is used as the criterion for this purpose. The second objective of this thesis is to apply logistic regression modeling and Random Forests to evaluate the factors related to stage at diagnosis in a population based study of black/white colon cancer patients[17].

2.0 LITERATURE REVIEW

Lee et al[18] compared 21 analytic tools, including logistic regression, Random Forests, and CART for seven microarray data sets with sample sizes ranging from 918 to 2467. Five data sets were from different types of cancer patients and the other two data sets consisted of cancer patients with non-cancer patient controls. Since the variability of gene expression can be greater in tissue involved with cancer than in normal tissue, these two data sets were heterogeneous. The mean error rate was used as the criterion to compare the performance of these methods. The objective of this study was to provide the most appropriate classification tool in various specific situations. Although three gene selection methods were used in their analysis and they had significant influence on the performance of the statistical tools, Random Forests always performed much better than logistic regression and CART in all seven data sets. CART was better than logistic regression in some data sets and worse in the others. The outcome of the comparison depended as well on the gene selection methods. The seven data sets were complete, so methods of handling missing values offered by these analytic tools were not evaluated or compared.

In 1998, Rudolfer and Paliouras published a paper comparing CART with the proportional odds model, which is the extension of a traditional logistic regression model using data on carpal tunnel syndrome (CTS)[15]. These data were collected at electromyography clinics in a hospital. There were 1710 CTS patients in this study who were randomly divided into a training sample of 850 observations and a test sample of 860 observations. There were three kinds of independent variables: history, clinical examination, and nerve condition studies. The outcome variable was the severity of disease, which had four levels (No disease, mild CTS, moderate CTS, and severe CTS). Their interest focused on two aspects: variable selection and classification accuracy. They found that the important predictors chosen by these two methods overlapped and that the most important variable was selected by both. These two methods also gave very similar results with respect to classification accuracy (the correct classification rate was 79.3% for CART and 78.4% for logistic regression). Therefore, Rudolfer and Paliouras concluded that no important difference existed between logistic regression and CART for their dataset.

Their study is somewhat similar to the work presented in this thesis but differs in three aspects. First, they compared logistic regression with CART, whereas in this study, Random Forests is used instead of CART. Second, their data set was complete, so the algorithm for handling missing values offered by CART was not examined in their study. Third, a stepwise variable selection process was used in constructing the proportional odds model in their study. The variable selection method in this study is described in detail in Section 3.3.

In 2004, Delen and Walker[16] used logistic regression and decision trees to predict breast cancer survivability using data from the SEER Cancer Incidence Public-Use Database for the years 1973-2000. They defined survival as “any incidence of breast cancer where the person is still living after sixty months (5 years) from the date of diagnosis". The dependent variable was a binary variable coded as 0 and 1, where 0 denoted death and 1 denoted survival. Their data set contained 202,932 cases and 16 predictor variables. In this study, CART achieved a correct classification rate of 93.62%, with sensitivity of 96.02% and specificity of 90.66%. The correct classification rate, sensitivity, and specificity obtained from logistic regression was 89.20%, 90.17%, and 87.86%, respectively. These results are unexpectedly good. CART, however, had higher classification power than logistic regression. They used sensitivity analysis to identify important predictor variables, so variable selection capability was not compared between CART and logistic regression.

The original SEER Breast Cancer data contained 433,277 cases and 72 variables. They removed records where the cause of death was something other than breast cancer, and records of patients who were censored. Some variables were considered to be important and had missing values, so they kept these variables and deleted the records which contained the missing data. Although they checked the effect on the distributions of other variables of removing these records and found that the change in the distribution was not considerable, whether their conclusion can be applied to other situations is still questionable.

In addition to the studies mentioned above, there are other studies that have compared logistic regression and Random Forests or CART. For example, Chatellier[19] used CART and logistic regression to predict cardiovascular risk and compared their performance. After deleting records with missing values, the final data set contained

15,444 cases, of which 10,296 cases were used as a training sample and 5,148 as a test sample. The outcome in this study was “6-year incidence of the combined endpoint defined by occurrence of myocardial infarction, stroke or cardiovascular death”, which is a binary variable. The areas under ROC curves and their 95% confidence intervals were used as criteria to evaluate the predictive power of the model. In this study the authors showed that the performance of these two models were very similar. In a related study, Lin and Wu et al[20] used logistic regression and Random Forests to predict protein-protein interactions using the area under the ROC curve as the criterion to compare them. They used a large sample without missing values and concluded that Random Forests performed better than logistic regression. Ture et al[21] compared CART with logistic regression in a data set about hypertension, Garzotto[22] compared CART with logistic regression in prostate cancer detection, and Stark and Pfeiffer[23]compared these two methods in the veterinary epidemiological domain. The samples used in these studies were without missing values or missing values were estimated by certain methods, such as Stark and Pfeiffer, who used stochastic regression imputation, in the data preparation step. They found that CART and logistic regression had similar classification accuracy.

3.0 MATERIALS AND ANALYSIS

3.1 DESCRIPTION of DATA SET

The data used in this thesis are from National Cancer Institute Black/White Cancer Survival Study. Details of the design and conduct are provided in[17]. Study subjects were black and white colon cancer patients aged 20-79 years old who lived in one of three metropolitan areas, Atlanta, New Orleans, or San Francisco/Oakland. All patients were identified through population-based tumor registries and were newly diagnosed with invasive colon cancer between January 1, 1985 and December 31, 1986. All eligible black patients were included in the study, and a probability sample of white patients was selected. White patients were frequency matched with black patients by sex, age group (20-49, 50-64, and 65-79), and metropolitan area of residence. These three variables constitute the design variables. The dataset includes somewhat more white than black colon cancer cases. There were 983 patients; twenty-three patients were excluded because their stage at diagnosis was unknown, so the study sample consisted of 960 patients (441 blacks and 519 whites).

All variables used in this study were categorical variables. Appendix A lists the variables used and their record source. A previous publication has described the variables in detail[5]. Only 70.5% of black patients and 71.1% of white patients were interviewed.

Although there was no difference in the interview response rate between blacks and whites, the ability of patients to be interviewed was highly related to the stage of disease. Patients with metastatic disease were less likely to be interviewed. Stage was used as the dependent variable in all analyses . The Duke's Classification was the basis for staging the colon cancer patients[24].

3.2 METHODOLOGY

3.2.1. POLYTOMOUS LOGISTIC REGRESSION MODEL

Since the dependent variable, stage at diagnosis, had 4 levels, the polytomous logistic regression was used here. It is an extension of traditional logistic regression models and can simultaneously fit 3 logit models using the same reference group[7]. In this analysis, Duke's Stage A and white patients were used as the reference group.

3.2.2. PROPORTIONAL ODDS MODEL

Since the levels of stage at diagnosis were ordered, a proportional odds model was also suitable for the study. It is an extension of binary logistic regression models and is most commonly used when the dependent variable is an ordinal variable. It postulates a linear form for the log cumulative odds: $\ln [P(Y\leq j|X)/P(Y>j|X)] = \alpha_j + \beta X$, where β is the vector of parameters for X[7]. These common parameters β's reflect the proportional odds assumption. Therefore, the only difference in the models is the intercept terms, α_j. For example, in this analysis the dependent variable had 4 ordered levels, Stage A, Stage B, Stage C, and Stage D(j=1, 2, 3, 4); we can consider 3 binary logistic regression models with j=1, 2, 3. This means that the estimates from the three binary models can be pooled to provide just one set of β estimates. By calculating exp(β), we obtain an estimate of the cumulative odds ratio for the value of x differing by one unit.

3.2.3. RANDOM FORESTS

Breiman and Cutler explain Random Forests and discuss the relationship and difference between it and CART on their website[11, 13, 14]. They state that "Random Forests has its root in CART"[11]. It is an ensemble of trees. It combines trees by voting in a classification problem after each tree is grown. When data are input down each tree in the forest, each tree casts a vote at its terminal node for the class, and the forest chooses the classification which has the most votes[11]. Random Forests differs from CART in several aspects. First, unlike CART, which uses all the data to construct the tree, Random Forests uses 63% of the data, which are randomly selected from the original data set with replacement to build each tree and the remaining 37% of data called "out of bag" (OOB), are used to get an unbiased estimate of the classification error rate when the tree is added into the forest [11]. Second, when each node in every tree is split, Random Forests first randomly selects a small subset of the total available variables as the potential splitters and uses the best one to split the node[11]. Usually, the size of the small subset equals the square root of the number of total available variables, whereas in CART, the best of all the available variables is used to split the node. For these two features, trees in Random Forests are different from each other, but combining trees will only be beneficial if the trees are different. Each tree in Random Forests is a weak learner which is defined as "a prediction function that has low bias"[13].Usually, a weak learner has low bias and high variance, so it is not an accurate predictor. The classification power of each tree in Random Forests is lower than that of CART which is a single tree. Combining weak learners, however, can result in estimates with low bias and low variance, which is the main reason why Random Forests does better than CART in terms of classification accuracy. Another advantage of this combination is that every tree can uncover somewhat different aspects of the data structure[12]. Third, each tree is grown to the largest possible extent, so no pruning is needed. Fourth, the algorithm for handling missing values is different from that used in CART. Random Forests offers two ways to handle missing values[11, 14]. The easiest way is to use the mode of the non-missing value to replace the missing value in the same class if the variable is a categorical variable. This method is fast, but is not suitable when the proportion of missing values is large. In the advanced method, missing values are replaced inaccurately first. The forest is run and the proximities are computed. The missing values are refilled with the most frequent non-missing value in the same class where the frequency is weighted by proximity. That is, the case which is close (use proximity as criterion) to the case with missing value has more weight in estimating the missing value. The process is repeated 4-6 times. The advanced method can handle a large proportion of missing values. Breiman and Cutler demonstrated this method using a DNA data set which has 50% of the data deleted at random and showed that the test set error was less than 10%[14]. How large a proportion of missing values can be handled before the method becomes problematic and whether or not this method can handle missing values which are not missing at random are unknown. In general, Random Forests is more accurate than CART. Breiman demonstrated that the error rate obtained from Random Forests is two-thirds of the CART error rate[11].

3.3 ANALYTIC METHODS

Differences between blacks and whites in colon cancer stage at diagnosis and individual factors of interest were initially evaluated using a chi-square test for contingency tables.

Although it would be optimal to divide the dataset into two subsets, a training sample and a test sample, it cannot be done for the logistic regression model due to small sample size. An overfitting problem was encountered when an attempt was made to fit a logistic model using half or two-thirds of the data as a training sample. Therefore, the same data is used to construct the logistic regression model and to evaluate the performance of the model.

The logistic regression model was fitted in two ways. First, a logistic regression model was fitted including missing values by setting a separate category "unknown", and then the model was refitted excluding missing values. Whites and Duke's Stage A patients were used as the reference group for the logistic regression analysis.

The relationship between each factor and stage was evaluated for blacks and whites separately in a proportional odds model. However, when the proportional odds model was fitted for tumor grade in blacks and other variables (smoking, grade of tumor, total delay, patient delay, occupation, income, insurance type, poverty index, education and body mass index) in whites, the proportional odds assumption was not satisfied, hence a polytomous logistic regression model was also fitted for these variables in whites. Quasi-complete separation problems occurred when a polytomous logistic regression model was fitted for tumor grade in both races and for patient delay in whites. Quasi-complete separation occurs when the distribution of covariates of two groups overlap only at a few tied values[7]. For example, if blood pressure is a predictor in a study and all patients' blood pressures are higher or equal to 120 mmHg, whereas blood pressure in all controls are lower than or equal to 120 mmHg. The values of blood pressure of these two groups overlap only on 120 mmHg. In this situation the maximum likelihood estimators do not exist, which is referred to as quasi-complete separation. Quasi-complete separation is suggested by unreasonably large standard errors and is sensitive to the sample size, as well as the number of covariates included in model[7]. If one uses a backward variable selection method and includes all of the potential predictors simultaneously in the model initially, this problem is more likely to occur.

Individual factors that showed a statistically significant association with stage at the P<0.10 level in either blacks or whites were examined in a multivariate polytomous logistic regression model to assess the association between race and stage controlling simultaneously for the three design variables (sex, age group and metropolitan area of residence) and the factor of interest.

Factors that showed statistically significant associations with stage in both blacks and whites were examined in a multivariate polytomous logistic regression model using a backward selection method. At first, all factors were included in the model simultaneously, but the quasi-complete separation problem occurred. Since the quasi-complete separation problem is sensitive to the number of variables included in the model, variables “education” and “smoking” were excluded from the model to eliminate this problem, and the backward selection method was used. These two variables were excluded because their variances were unreasonably large suggesting that their influence on the model was not estimated well and keeping these two variables in the model resulted in more than two other variables excluded. Because the objectives of this study were to compare logistic regression with Random Forests in regard to variable selection and detecting factors associated with colon cancer stage at diagnosis, a reasonable strategy to overcome this difficulty was to include as many variables as possible in the initial model. After excluding all factors which did not show statistical significance, variables “education” and “smoking” were added into the model to see whether they significantly improved the model fit, but the quasi-complete separation problem arose again.

Model fit was checked using the method recommended by Hosmer and Lemeshow of fitting and calculating logistic regression diagnostics using the individual logistic regression approach[7]. Hosmer and Lemeshow’s test statistic was used to check the global goodness-of-fit of the model. This statistic follows a chi-square distribution. For the model which included Stage A and Stage B patients (Stage A as reference), Chi-Square(8)=14.51 and p=0.07. For the model which included Stage A and Stage C patients (Stage A as reference), Chi-Square (8)=3.62 and p=0.89. For the model which included Stage A and Stage D patients (Stage A as reference), Chi-Square(8)=14.62 and p=0.07. All p-values were larger than 0.05, indicating that the global goodness-of-fit of the model was good. An index plot of deviance residuals was made to see whether there were any outliers. Any point whose deviance residual was beyond -2 to +2 is considered an outlier, indicating that the model does not fit this point well. The index plot of $\Delta\beta$ was used to check whether there was any point which had undue influence on each parameter estimate. Eleven observations were indicated as outliers, but no error was found in the data records. In general, the model fit was good.

There were 430 patients (194 blacks and 236 whites) in the analysis after patients with missing values on any of the covariates were excluded. Neither the polytomous logistic regression model nor the proportional odds model could be fitted for blacks and whites separately even when the model included only the three design variables (and their two and three way interactions) either because the quasi-complete separation problem was met or the proportional odds assumption was not satisfied. The proportional odds assumption was also not satisfied when the proportional odds model was fitted for the whole data set without missing values.

A multivariate polytomous logistic regression model was fitted for the data without missing values to assess the relationship between race and stage adjusted for the three design variables and the factor of interest. Three variables, patient delay, income, and usual health care source, were found to be important. The importance of tumor grade was unknown, since it could not be included in the model with race and three design variables due to quasi-complete separation.

A forward selection method was used to fit a multivariate polytomous logistic regression model, but the variable "patient delay" cannot be in the model simultaneously with any of the other two important variables, so the final model either included patient delay or usual health care source and income in addition to the three design variables (and their interactions) as well as race.

The advanced missing value imputation method in Random Forests was used because of the large proportion of missing values. All class weights were set to 1 regardless of the number of cases each class contained. The number of trees was set to 500 and the number of potential splitters for each node was set to 5 (the approximate square root of the number of independent variables).

Polytomous logistic regression and proportional odds models were fitted using SAS Version 8.0. Random Forests was performed using RandomForests™ Version 1.0. The frequency distribution of the black and white colon cancer patients by stage at diagnosis (Figure 2) was obtained using Stata 8.0.

4.0 RESULTS

4.1 DISEASE STAGE AT DIAGNOSIS

The frequency distribution of the black and white colon cancer patients by stage at diagnosis is shown in Figure 2. Stage of disease at diagnosis is significantly associated with race (Chi-Square2(3)=7.76, P=0.0153), with black patients more likely to be diagnosed at advanced stage than their white counterparts.

4.2 RELATIONSHIP BETWEEN RACE AND FACTORS OF INTEREST

Table 1 presents the racial distribution of all variables used in the analysis. There were no significant differences in grade and histology of tumor, alcohol use, total delay and interview status between black and white patients (P>0.05). The difference between blacks and whites in patient delay was marginally significant (P=0.049).Compared with whites, black patients were less likely to be married or partnered, and more likely to experience co morbidity, to have income less than $10,000 or poverty index lower than 125. Blacks were also less well educated and more likely to have an unskilled occupation as well as being overweight. A greater proportion of blacks than whites had public insurance or had their usual health care provided by a public clinical facility. The proportions of former smokers and those with a professional job were higher for whites than for blacks.

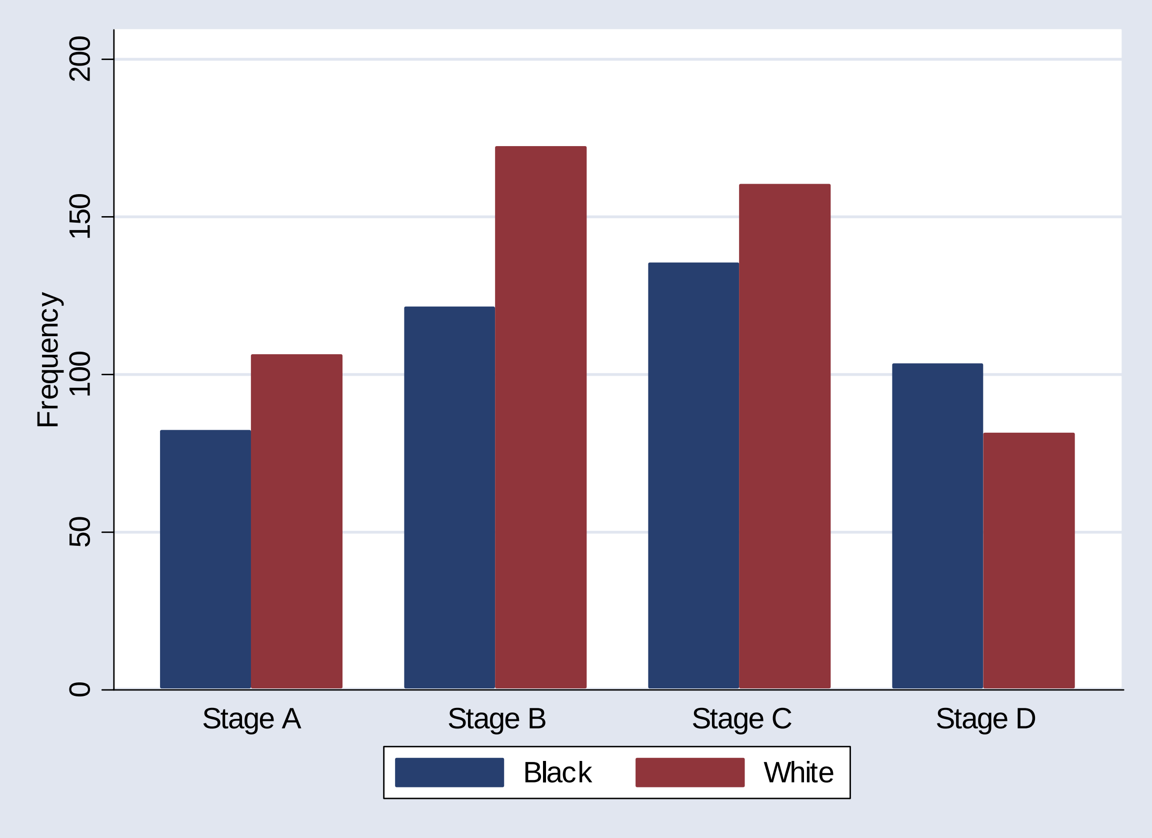

Figure 2: Frequency Distribution of Black and White Colon Cancer Patients by Stage at Diagnosis

Table 1: Distribution of Independent Variables in Black and White Colon Cancer Patients

Variable	White		Black		P-value˘
	No	%	No	%	
Total	519	54.1	441	45.9	
Age category					
20-49	48	5.0	57	5.9	
50-64	176	18.3	148	15.4	
65-79	295	30.7	236	24.6	
Sex					
Female	296	30.8	239	24.9	
Male	223	23.2	202	21.0	
Metropolitan area					
Atlanta	115	12.0	139	14.5	
New Orleans	164	17.1	136	14.2	
San Francisco/Oakland	240	25.0	166	17.3	
Marital status					<0.01
Married/partnered	344	35.8	221	23.0	
Widowed	90	9.4	99	10.3	
Separated/divorced	46	4.8	71	7.4	
Never married	28	2.9	29	3.0	
Unknown	11	1.2	21	2.2	
Smoking					0.02
Never	154	16.0	149	15.5	
Former	171	17.8	107	11.2	
Current	43	4.5	49	5.1	
Unknown	151	15.7	136	14.2	
Alcohol use					0.54
Never use	489	50.9	408	42.5	
Formerly used	13	1.4	13	1.4	
Currently uses	17	1.8	20	2.1	
Total delay category					0.95
<1 month	95	9.9	83	8.7	
1-3 month	91	9.5	80	8.3	
3-6 month	67	7.0	52	5.4	
>= 6 month	93	9.7	73	7.6	
Unknown	173	18.0	153	15.9	
Occupation class					<0.01
Homemaker	30	3.1	14	1.5	
Managerial/professional	128	13.3	45	4.7	
Technical/sale/administrative	105	10.9	48	5.0	
Skilled	66	6.9	90	9.4	
Unskilled	40	4.2	110	11.5	
Unknown	150	15.6	134	14.0	

Table 1 (continued): Distribution of Independent Variables in Black and White Colon Cancer Patients

Variable	White		Black		P-value˜
	No	%	No	%	
Patient delay category					0.05
No sympt/unkn sympt˜	67	7.0	31	4.7	
<1 month	157	16.4	150	15.6	
1-3 month	58	6.0	53	5.6	
3-6 month	37	3.9	24	2.5	
>= 6 month	48	5.0	40	4.2	
Unknown	152	15.8	143	14.9	
Household income category					<0.01
<$10000	50	5.2	113	11.8	
$10000-$19999	93	9.7	67	7.0	
$20000-$34999	78	8.1	40	4.2	
>=$35000	108	11.3	33	3.4	
Unknown	190	19.8	188	19.6	
Insurance group category					<0.01
None	11	1.2	30	3.1	
Public	34	3.5	106	11.0	
Any private	324	33.8	174	18.1	
Unknown	150	15.6	134	13.7	
Usual care group					<0.01
None	49	5.1	61	6.4	
Public	17	1.8	67	7.0	
Private	301	31.4	179	18.7	
Unknown	152	15.8	134	14.0	
Poverty index category**					<0.01
0-125	47	4.9	111	11.6	
126-200	37	3.9	46	4.8	
201-300	55	5.7	27	2.8	
301-400	45	4.7	21	2.2	
>400	144	15.0	48	5.0	
Unknown	191	19.9	188	19.6	
Education category					<0.01
<=High school	92	9.6	174	18.1	
High school	116	12.1	68	7.1	
>High school	161	16.8	69	7.2	
Unknown	150	15.6	130	13.5	
Body mass index quartile (weight(kg)/height(m)^2)***					<0.01
1	116	12.1	52	5.4	
2	117	12.2	82	8.5	
3	103	10.7	97	10.1	
4	86	9.0	116	12.1	
Unknown	97	10.1	94	9.8	

Table 1 (continued): Distribution of Independent Variables in Black and White Colon Cancer Patients

Variable	White		Black		P-value˘
	No	%	No	%	
Interview status					0.84
Interview	369	38.4	311	32.4	
Non-interview	150	15.6	130	13.5	
Tumor histologic type					0.91
Other	47	4.9	39	4.1	
IDC+/-*	472	49.2	402	41.9	
Duke's stage					0.02
Stage A	106	11.0	82	8.5	
Stage B	172	17.9	121	12.6	
Stage C	160	16.7	135	14.1	
Stage D	81	8.4	103	10.7	
Tumor grade					0.13
Grade 1	132	13.8	122	12.7	
Grade 2	295	30.7	265	27.6	
Grade 3	70	7.3	41	4.3	
Unknown	22	2.3	13	1.4	
Comorbidities					<0.01
No	149	15.5	88	9.2	
Yes	286	29.8	299	31.2	
Unknown	84	8.8	54	5.6	

Note: *. IDC+/- means Adenocarcinoma / adenosquamous

** A poverty index was calculated using the information of household income and the number of person supported by this income and divided by the national 1986 poverty-level income for a family of that size.

*** Body mass index was calculated using formula [weight(kg)/height(m)^2]. Participants were categorized by marking cut points at the sex-specific 50th and 85th percentiles for men and women aged 20-29 years in the Second National Health and Nutrition Examination Survey.

˜. No sympt/unkn sympt means no symptom or unknown symptom status

˘. All P-value are from Chi-Square test

4.3 RESULTS FROM LOGISTIC REGRESSION

4.3.1. RELATIONSHIP BETWEEN STAGE AND FACTORS OF INTEREST

Table 2 and Table 3 present the relationship between stage at diagnosis and all factors in a logistic regression analysis for blacks and whites separately.

Marital status had a relationship with stage at diagnosis in black patients only. Co morbidity and alcohol use were not associated with stage at diagnosis in either race. All other factors examined had an association with stage in both blacks and whites. Among the variables smoking, marital status, income, insurance, usual health care source, total delay, occupation, education, and body mass index, the category designated "unknown" which represented the missing values in that variable was significantly related to higher stage at diagnosis in blacks. The "unknown" category of total delay, education, and body mass index was associated with Stage D in whites. Patient delay was positively related to higher stage at diagnosis in blacks and may also be positively associated with higher stage in whites. The higher income or poverty index was negatively related to higher stage at diagnosis in whites. Histology of tumor was significantly associated with stage, with histology other than adenocarcinoma / adenosquamous related to higher stage in whites.

Table 2: Cumulative Odds Ratio and 95% Confidence Interval of All Factors

Characteristics˜	White		Black	
	OR (95% CI**)	P-value*	OR (95% CI**)	P-value*
Marital Status		0.17		0.02
Married/Partnered				
Widowed	0.9 (0.6-1.5)		0.8 (0.5-1.2)	
Separated/Divorced	0.6 (0.3-1.1)		1.0 (0.6-1.6)	
Never Married	0.5 (0.2-1.0)		0.8 (0.4-1.7)	
Unknown	0.8 (0.3-2.4)		0.2 (0.1-0.5)	
Smoking		<0.01ˆ		<0.01
Current				
Never	1.3 (0.7-2.4)		1.0 (0.6-1.9)	
Former	1.6 (0.9-3.0)		1.2 (0.6-2.2)	
Unknown	0.7 (0.4-1.3)		0.4 (0.2-0.8)	
Alcohol use		0.65		0.17
Never				
Formerly used	1.6 (0.6-4.4)		2.2 (0.8-6.1)	
Currently uses	0.9 (0.4-2.3)		1.7 (0.7-4.0)	
Total delay category		0.03ˆ		<0.01
<1 month				
1-3months	0.9 (0.5-1.5)		0.7 (0.4-1.3)	
3-6months	0.9 (0.5-1.6)		1.2 (0.6-2.2)	
≥6months	0.9 (0.5-1.5)		1.0 (0.6-1.8)	
Unknown	0.5 (0.3-0.8)		0.4 (0.2-0.7)	
Occupation class		<0.01ˆ		<0.01
Homemaker				
Managerial/professional	1.2 (0.6-2.6)		0.5 (0.2-1.5)	
Technical/sale/administrative	0.7 (0.4-1.6)		0.3 (0.1-1.0	
Skilled	1.0 (0.4-2.2)		0.4 (0.1-1.0)	
Unskilled	0.8 (0.3-2.0)		0.5 (0.2-1.4)	
Unknown	0.5 (0.2-1.0)		0.2 (0.1-0.4)	
Patient delay category		<0.01ˆ		<0.01
No sympt/unkn sympt˜				
<1month	0.2 (0.1-0.4)		0.3 (0.2-0.7)	
1-3months	0.2 (0.1-0.3)		0.3 (0.1-0.8)	
1-3months	0.2 (0.1-0.4)		0.4 (0.1-1.0)	
≥6months	0.3 (0.2-0.7)		0.4 (0.2-1.1)	
Unknown	0.1 (0.1-0.2)		0.1 (0.1-0.3)	
Household income category		<0.01ˆ		0.01
<$10000				
$10000-$19999	1.3 (0.7-2.5)		0.7 (0.4-1.3)	
$20000-$34999	1.9 (1.0-3.8)		2.0 (1.0-4.0)	
>=$35000	2.3 (1.2-4.3)		1.1 (0.5-2.6)	
Unknown	0.8 (0.5-1.5)		0.6 (0.4-1.0)	
Insurance group category		<0.01ˆ		<0.01
None				
Public	1.3 (0.4-4.5)		0.5 (0.2-1.2)	
Private	2.0 (0.6-6.2)		0.6 (0.3-1.4)	
Unknown	0.9 (0.3-2.8)		0.2 (0.1-0.5)	

Table 2 (continued): Cumulative Odds Ratio and 95% Confidence Interval of All Factors

Characteristics˜	White		Black	
	OR (95% CI**)	P-value*	OR (95% CI**)	P-value*
Usual care group		<0.01ˆ		<0.01
None				
Public	1.3 (0.5-3.7)		1.6 (0.9-3.2)	
Private	1.5 (0.9-2.7)		1.2 (0.7-2.1)	
Unknown	0.7 (0.4-1.3)		0.5 (0.3-0.9)	
Poverty index category		<0.01ˆ		0.07
0-125				
126-200	1.6 (0.7-3.6)		1.2 (0.6-2.4)	
201-300	1.1 (0.5-2.3)		1.3 (0.6-2.9)	
301-400	1.9 (0.9-4.0)		2.0 (0.8-4.7)	
>400	2.3 (1.2-4.2)		1.5 (0.7-3.0)	
Unknown	0.9 (0.5-1.6)		0.7 (0.5-1.1)	
Education category		<0.01ˆ		<0.01
<high school				
High school	0.8 (0.5-1.3)		1.1 (0.7-1.9)	
>high school	1.3 (0.8-2.1)		0.9 (0.5-1.6)	
Unknown	0.5 (0.3-0.8)		0.4 (0.2-0.6)	
Body mass index quartile (weight/height^2)		0.03ˆ		<0.01
1				
2	0.9 (0.5-1.4)		0.5 (0.3-0.9)	
3	1.1 (0.7-1.8)		0.5 (0.3-0.9)	
4	1.2 (0.7-2.1)		0.6 (0.3-1.1)	
Unknown	0.5 (0.3-0.9)		0.3 (0.2-0.6)	
Tumor grade		<0.01ˆ		<0.01ˆ
Grade 1				
Grade 2	0.3 (0.2-0.5)		0.3 (0.2-0.5)	
Grade 3	0.2 (0.1-0.3)		0.1 (0.1-0.3)	
Unknown	1.1 (0.5-2.6)		0.4 (0.1-1.2)	
Comorbidities		0.94		0.18
No				
Yes	1.1 (0.7-1.5)		1.4 (0.9-2.2)	
Unknown	1.1 (0.7-1.8)		1.8 (0.9-3.3)	
Tumor histologic type		0.02		0.08
IDC+/-				
Other	0.5 (0.3-0.9)		0.6 (0.3-1.1)	

Note: * P-value was calculated using likelihood ratio test

** 95% CI represents 95% confidence interval

ˆ Proportional odds assumption is not satisfied

˜ All models include three design variables and their interactions. The first level of each variable was used as reference

˜ No sympt/unkn sympt means no symptom or unknown symptom status

Table 3*.Odds Ratio and 95% Confidence Interval for Some Factors in White Patients

Characteristics	Stage B	Stage C	Stage D	P-value
	OR (95% CI)	OR (95% CI)	OR (95% CI)	
Smoking				<0.01
Current				
Former	1.0 (0.4- 2.8)	0.7(0.2- 1.8)	0.4 (0.1- 1.4)	
Never	1.8 (0.6- 5.0)	1.0 (0.4- 2.8)	0.8 (0.2- 2.9)	
Unknown	1.0 (0.3- 2.9)	1.0 (0.4- 2.8)	1.9 (0.6- 6.3)	
Total delay category				0.06
< 1 month				
1-3 months	0.8 (0.3- 1.7)	1.2 (0.5- 2.8)	1.0 (0.3- 3.2)	
3-6 months	0.8 (0.3- 1.9)	1.0 (0.4- 2.4)	1.3 (0.4- 4.3)	
>=6 months	0.8 (0.4- 1.9)	1.0 (0.4- 2.4)	1.3 (0.4- 3.9)	
Unknown	0.7 (0.3- 1.4)	1.1 (0.5- 2.4)	3.1 (1.2- 8.0)	
Occupation class				<0.01
Homemaker				
Managerial/professional	1.4 (0.5- 4.6)	1.6 (0.5- 5.2)	0.6 (0.1- 2.3)	
Skilled	5.7 (1.6-20.6)	2.5 (0.6-10.1)	1.3 (0.3- 6.8)	
Technical/sale/administrative	2.8 (0.8- 9.0)	3.2 (0.9-10.8)	1.5 (0.4- 6.2)	
Unskilled	4.3 (1.0-17.7)	2.8 (0.6-12.4)	1.7 (0.3-10.1)	
Unknown	1.8 (0.6- 5.6)	2.5 (0.8- 8.2)	3.1 (0.8-11.7)	
Household income category				<0.01
<$10000				
$10000-$19999	0.5 (0.2- 1.7)	0.4 (0.1- 1.4)	0.6 (0.1- 2.4)	
$20000-$34999	0.3 (0.1- 0.8)	0.3 (0.1- 0.9)	0.2 (0.1- 1.0)	
>+$35000	0.3 (0.1- 0.9)	0.3 (0.1- 0.8)	0.1 (0.0- 0.7)	
Unknown	0.3 (0.1- 1.0)	0.5 (0.2- 1.4)	0.9 (0.2- 3.4)	
Poverty index category				<0.01
0-125				
126-200	0.4 (0.1 -1.7)	0.3 (0.3- 1.2)	0.4 (0.1- 2.0)	
201-300	0.4 (0.1- 1.7)	0.5 (0.1- 2.2)	0.5 (0.1- 2.5)	
301-400	0.3 (0.1- 1.2)	0.4 (0.1- 1.7)	0.1 (0.0- 0.7)	
>400	0.2 (0.1- 0.7)	0.2 (0.1- 0.8)	0.1 (0.0- 0.5)	
Unknown	0.3 (0.1- 0.9)	0.4 (0.1- 1.5)	0.6 (0.2- 2.5)	
Education category				<0.01
< high school				
high school	1.3 (0.6- 3.0)	1.2 (0.5- 2.8)	2.1 (0.7- 6.1)	
>high school	0.5 (0.3- 1.1)	0.7 (0.3- 1.5)	0.5 (0.2- 1.5)	
Unknown	0.6 (0.3- 1.4)	1.1 (0.5- 2.4)	3.0 (1.2- 8.0)	
Body mass index quartile				
(weight/height^2)				<0.01
1				
2	0.7 (0.3- 1.6)	0.8 (0.4- 1.6)	1.6 (0.6- 4.2)	
3	1.0 (0.5- 2.2)	0.8 (0.4- 1.8)	1.0 (0.3- 2.9)	
4	1.0 (0.5- 2.3)	0.5 (0.2- 1.2)	1.3 (0.4- 3.6)	
Unknown	0.7 (0.3- 1.7)	0.9 (0.4- 2.0)	3.4 (1.3- 9.1)	

Note: * This table is complementary to Table 2; the relationship between each variable here and stage cannot be assessed using the proportional odds model because the proportional odds assumption is not satisfied. Variable grade is not in this table, because of quasi-complete separation. Stage A and the first level in each variable were used as reference

4.3.2. ASSOCIATION BETWEEN RACE AND STAGE AFTER CONTROLLING FOR OTHER FACTORS

The relationship between stage and race adjusted for other factors of interest is presented in Table 4. Although interview status was associated with stage, it did not significantly affect the relationship between stage and race.

For patients with Stage D disease, the 95% confidence intervals of the odds ratio did not include one when adjusted for smoking, total delay, usual health care source, body mass index, and tumor grade as well as histology of tumor. However, they included one when adjusted for marital status, occupation, patient delay, income, poverty index, insurance type, and education. The change in the odds ratio was 0.079 adjusted for education, 0.342 adjusted for poverty index, and between 0.079 and 0.342 adjusted for marital status, occupation, patient delay, or insurance type, which suggests that these factors may explain part of the race-stage relationship.

For patients with Stage C disease, the 95% confidence intervals of odds ratio included one when controlled for sex, age group, and metropolitan area of residence. The odds ratio and its 95% confidence intervals changed very little when adjusted for the design variables and marital status, smoking, total delay, occupation, insurance type, usual health care source, education, body mass index, or grade and histology of tumor. The odds ratios decreased from above one to below one when adjusted for patient delay, income, and poverty index.

For patients with Stage B disease, the odds ratio and its 95% confidence intervals changed very little when adjusted for all factors associated with stage.

Table 4*. Odds Ratio and 95% Confidence Interval in the Entire Sample for Black Colon Cancer Patients versus White Colon Cancer Patients

Characteristics	Stage B	Stage C	Stage D
	OR (95% CI)	OR (95% CI)	OR (95% CI)
Race Black Vs. White	0.9 (0.6- 1.3)	1.1 (0.7- 1.6)	1.6 (1.1- 2.5)
Martial Status	0.9 (0.6- 1.3)	1.0 (0.7- 1.5)	1.5 (1.0- 2.3)
Smoking	0.8 (0.6- 1.2)	1.0 (0.7- 1.5)	1.6 (1.0- 2.4)
Total delay category	0.9 (0.6- 1.3)	1.1 (0.7- 1.6)	1.6 (1.1- 2.5)
Occupation class	0.8 (0.5- 1.1)	1.0 (0.7- 1.5)	1.5 (1.0- 2.4)
Patient delay category	0.8 (0.5- 1.2)	1.0 (0.6- 1.4)	1.4 (0.9- 2.2)
Household income category	0.8 (0.6- 1.2)	1.0 (0.7- 1.5)	1.3 (0.9- 2.1)
Insurance group category	0.8 (0.5- 1.2)	1.0 (0.7- 1.5)	1.5 (1.0- 2.4)
Usual care group	0.8 (0.6- 1.2)	1.0 (0.7- 1.6)	1.6 (1.1- 2.5)
Poverty index category	0.8 (0.5- 1.2)	1.0 (0.7- 1.5)	1.3 (0.8- 2.0)
Education category	0.8 (0.5- 1.2)	1.1 (0.7- 1.6)	1.5 (1.0- 2.4)
Body mass index quartile (weight/height^2)	0.9 (0.6- 1.3)	1.1 (0.8- 1.6)	1.6 (1.0- 2.4)
Interview status	0.9 (0.6- 1.3)	1.1 (0.7- 1.6)	1.7 (1.1- 2.5)
Histology category	0.9 (0.6- 1.3)	1.1 (0.7- 1.6)	1.6 (1.1- 2.5)
Tumor grade	0.9 (0.6- 1.3)	1.2 (0.8- 1.8)	1.8 (1.2- 2.8)

Note: * All models include three design variables and their interactions. White, Stage A and the first level of each variable shown in Table 1 were used as reference.

Four variables, patient delay, occupation, histology and grade of tumor, remained in the multivariate logistic regression model using a backward variable selection method. The model which included these four variables yielded adjusted odds ratios for race of 1.515 among Stage D, 1.021 among Stage C, and 0.712 among Stage B using whites and Stage A as the reference. These four factors explained part of the excess risk for Stage D versus Stage A disease among black patients compared with whites (odds ratio decreased from 1.628 to 1.515). Retaining histology and grade of tumor in the model significantly improved the model fit, but the odds ratio and its 95% confidence intervals decreased very little when adjusted for histology of tumor, and they became larger when adjusted for grade of tumor. Although these four variables explained some of the race-stage relationship and the 95% confidence intervals for the odds ratio overlapped one when Stage D cases were adjusted for the four variables, the odds ratio was far from 1, indicating that inclusion of these variables did not adequately explain the excess risk for advanced stage disease among blacks (Table 5). In the multivariate polytomous logistic regression model, the variable “occupation” could be replaced by the variable “education”, as seen in Table 6.

Table 5: Black-White Odds Ratio and 95% Confidence Interval for Colon Cancer Stage at Diagnosis

Model	Variable	-2log L	Df*	P**	Odds ratio (95% Confidence interval) blacks vs. whites		
					Stage B	Stage C	Stage D
1	***	2248.1	96		0.7 (0.5-1.1)	1.0 (0.6-1.6)	1.5 (0.9-2.5)
2	Model 1-ptdlygp	2307.8	81	<0.01	0.8 (0.5-1.2)	1.2 (0.7-1.8)	1.7 (1.1-2.8)
3	Model 2-occup	2378.0	66	<0.01	0.9 (0.6-1.4)	1.2 (0.8-1.8)	1.9 (1.2-2.9)
4	Model 3-histcat	2398.7	63	<0.01	0.9 (0.6-1.3)	1.2 (0.8-1.8)	1.8 (1.2-2.8)
5	Model-grade	2525.9	54	<0.01	0.9 (0.6-1.3)	1.1 (0.7-1.6)	1.6 (1.1-2.5)

Note : * Df represents degree of freedom.

** P represents p-value of variable removed; it was calculated based on log-likelihood ratio test.

*** Model 1 included variables: metropolitan area of residence, sex, age category (their 2-way and 3-way interactions), race, occupation class, patient delay category, tumor histologic type, and tumor grade.

Table 6: Black-White Odds Ratio and 95% Confidence Interval for Colon Cancer Stage at Diagnosis (With Occupation Replaced by Education)

Model	Variable	-2log L	Df*	P**	Odds ratio (95% confidence interval) blacks vs whites		
					Stage B	Stage C	Stage D
1	***	2268.7	90		0.8 (0.5-1.2)	1.1 (0.7-1.7)	1.6 (1.0-2.6)
2	Model 1-ptdlygp	2328.0	75	<0.01	0.8 (0.6-1.3)	1.2 (0.8-1.8)	1.8 (1.1-2.9)
3	Model 2-educ	2378.0	66	<0.01	0.9 (0.6-1.4)	1.2 (0.8-1.8)	1.9 (1.2-2.9)
4	Model 3-histcat	2398.7	63	<0.01	0.9 (0.6-1.3)	1.2 (0.8-1.8)	1.8 (1.2-2.8)
5	Model-grade	2525.9	54	<0.01	0.9 (0.6-1.3)	1.1 (0.7-1.6)	1.6 (1.1-2.5)

Note : * Df represents degree of freedom.

** P represents p-value of variable removed; it was calculated based on log-likelihood ratio test.

*** Model 1 included variables: metropolitan area of residence, sex, age category (their 2-way and 3-way interactions), race, education category, patient delay category, tumor histologic type, and tumor grade.

4.3.3. IMPORTANT VARIABLES

Tumor grade, patient delay, occupation or education, and histology of tumor were found to be the most important variables when missing values were included in the analysis. It can be seen that the racial difference in colon cancer stage at diagnosis became non-significant after adjustment was made for these variables. Patient delay, income and usual health care source were also found to be important in the model without missing values.

4.3.4. CLASSIFICATION ACCURACY

Logistic regression gives a predicted probability of each stage for every case. The stage with the highest probability was assigned to every case to achieve classification. Table 7 presents the cross-tabulation of the actual stages and predicted stages which were obtained from the model including all four important variables and missing values. The correct classification rate was not high. However, the ability of the model to evaluate the order of stage levels was good.

Table 7: Classification Accuracy Obtained from Logistic Regression using Entire Sample

	Predicted Duke's stage					
Duke's stage	Stage A	Stage B	Stage C	Stage D	%Correct	Total
Stage A	105	42	29	12	55.9	188
Stage B	37	139	93	24	47.4	293
Stage C	23	86	145	41	49.2	295
Stage D	19	35	59	71	38.6	184
Total	184	302	326	148	47.9	960

The logistic regression model was also fitted without missing values. Unfortunately, the model could not include all important variables simultaneously due to the small sample size, so two models were fitted; one included patient delay and another included income and usual health care source. The results of classification accuracy from these two models are presented in Table 8 and Table 9.

The correct classification rates were close to those obtained in the model including all four important variables and missing values.

Table 8*: Classification Accuracy Obtained from Logistic Regression with Missing Values Excluded

	Predicted Duke's stage					
Duke's stage	Stage A	Stage B	Stage C	Stage D	%Correct	Total
Stage A	22	33	27	2	26.2	84
Stage B	17	75	46	4	52.8	142
Stage C	9	50	83	2	57.6	144
Stage D	3	22	27	8	13.3	60
Total	51	180	183	16	43.7	430

Note: *The model only included patient delay, three design variables and their interactions.

Table 9*: Classification Accuracy Obtained from Logistic Regression with Missing Values Excluded

	Predicted Duke's stage					
Duke's stage	Stage A	Stage B	Stage C	Stage D	%Correct	Total
Stage A	29	32	23	0	34.5	84
Stage B	18	70	54	0	49.3	142
Stage C	10	51	71	12	49.3	144
Stage D	5	21	23	11	18.3	60
Total	62	174	171	23	42.1	430

Note: *The model only included income and usual health source, three design variables and their interactions.

4.4 RESULTS FROM RANDOM FORESTS

4.4.1 VARIABLE IMPORTANCE

Variable importance in the model that included only race, three design variables, and their 2-way and 3-way interactions, as predictors is shown in Figure 3. It can be seen that race is an important variable in predicting colon cancer stage at diagnosis in this model.

Variable	Score
LOCNAGE	100.00
SEXLOCN	55.85
SEXAGE	42.54
RACE	30.92
LOCN	29.18
AGEGP	14.14
SEX	6.42

Note: locn is metropolitan area of residence, sex is gender, and agegp is age category. Locnage, sexlocn, and sexage are their 2-way interactions.

Figure 3: Variable Importance in the Random Forests Model Which Included Only Race, Three Design Variables and Their 2-Way and 3-Way Interactions as Predictor Variables.

Variable importance in the full model is shown in Figure 4. Tumor grade was the most important variable, patient delay second, poverty index third, followed by occupation, education, and poverty index. Race was no longer an important variable when adjustment was made for grade, patient delay, poverty index, and occupation.

Variable	score
GRADE	100.00
PTDLYGP	64.28
POVGP	49.43
OCCUP	40.94
LOCNAGE	28.97
EDUC	27.18
SEXLOCN	20.25
TOTDLYGP	19.38
SMKHX	15.47
SEXAGE	13.86
BMIQ	13.37

Note: locn is metropolitan area of residence, sex is gender, and agegp is age category. Locnage, sexlocn, and sexage are their 2-way interactions. Grade is tumor grade, ptdlygp is patient delay category, povgp is poverty index category, occup is occupation class, educ is education category, totdlygp is total delay category, smkhx is smoking, and bmiq is body mass index quartiles.

Figure 4: Variable Importance from the Full Random Forests Model

4.4.2. CLASSIFICATION ACCURACY

Table 10 shows the classification accuracy obtained from a model without missing values. The total correct classification rate is 30.70%, which is lower than that obtained from the logistic regression model. It can be seen that logistic regression is better than Random Forests in terms of the total correct classification rate, since the logistic regression model which only includes part of the important variables has a higher correct classification rate than Random Forests. However, the correct classification rate for Stage A and Stage D from logistic regression is much lower than that from Random Forests.

The classification accuracy obtained from a model that included missing values is shown in Table 11. The total correct classification rate was 33.85%, which is also lower than that obtained from logistic regression. However, Random Forests performs better than logistic regression model in stage A and Stage D classification.

Table 10: Classification Accuracy Obtained from Random Forests with Missing Values Excluded

	Predicted Duke's stage					
Duke's stage	Stage A	Stage B	Stage C	Stage D	%Correct	Total
Stage A	53	8	8	15	63.1	84
Stage B	53	17	22	50	12.0	142
Stage C	29	26	33	56	22.9	144
Stage D	10	11	10	29	48.3	60
Total	145	62	73	150	30.7	430

Table 11: Classification Accuracy Obtained from Random Forests using Entire Sample

	Predicted Duke's stage					
Duke's stage	Stage A	Stage B	Stage C	Stage D	%Correct percent	Total
Stage A	124	16	18	30	66.0	188
Stage B	96	35	66	96	12.0	293
Stage C	60	58	58	119	19.7	295
Stage D	39	13	24	108	58.7	184
Total	319	122	166	353	33.9	960

4.4.3. POTENTIAL CLUSTER

"Random Forests" has an ability called "scaling" to measure the proximity of data as the distance between cases in geometric space. The closer the proximity, the closer the cases spatially. If prox(n, k) is the proximity between cases n and k, prox(-,k) the average of prox(n,k) over 1st coordinate, prox(n,-) the average of prox(n,k) over 2nd coordinate, and prox(-,-) the average over both, then a new matrix which is the matrix of inner products of the distances is : m(n, k)=.5*(prox(n,k)-prox(n,-)-prox(-,k)+prox(-,-)). If the eigenvalues and eigenvectors of m(n,k) are $\lambda(j)$ and $\upsilon(j)$ respectively, then $\sqrt{\ }\lambda(j)\ \upsilon(j)$ is the jth scaling coordinate[14]. “Random Forests” Version 1.0 uses the largest three eigenvalues and their corresponding eigenvectors of matrix m(n,k) as scaling coordinates to construct a 3-D plot which shows the potential cluster of data.

Figure 5 shows the potential cluster for all observations. It can be seen that there are several regions of overlap. The potential cluster for observations which were misclassified is shown in Figure 6. From this plot it can be seen that the points overlap severely, implying that these observations cannot be separated using the given data. Figure 7 shows the potential cluster for observations that were correctly classified. This plot indicates that points which belong to a different class occupy distinct locations, with only two small regions of overlap. It also can be seen that the points of Stage D are more diverse than others.

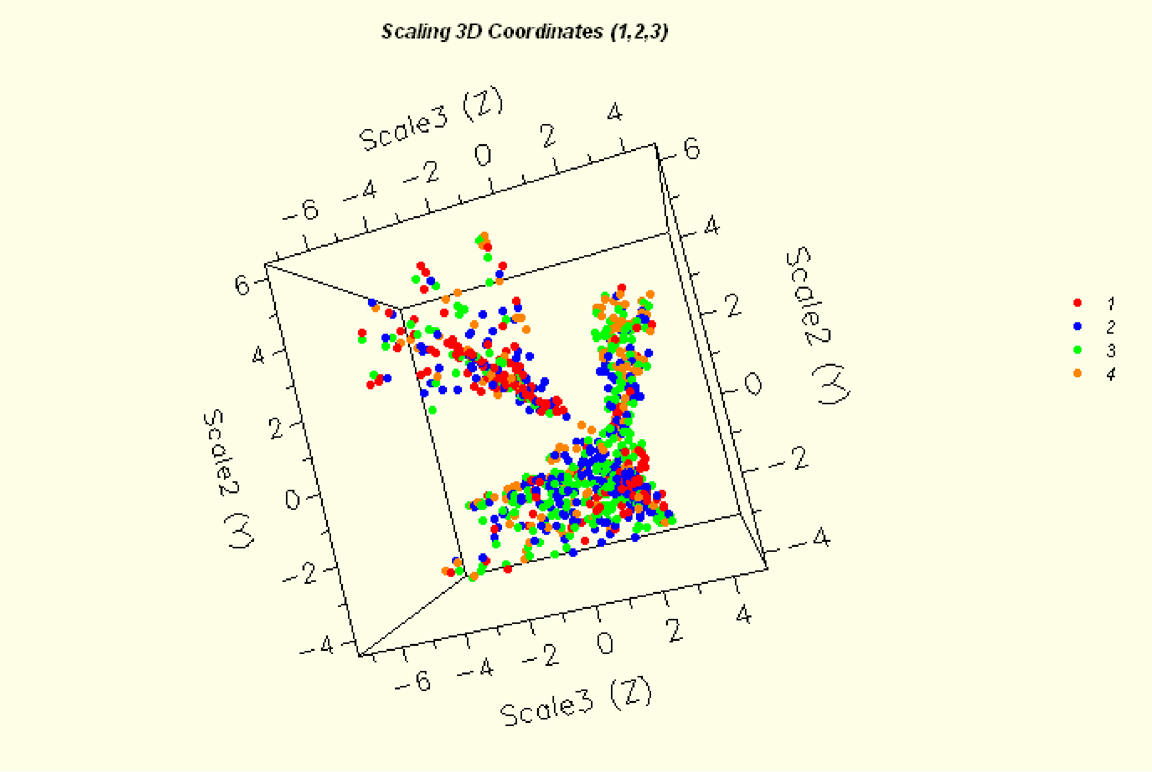

Figure 5: Potential Clusters of All Observations

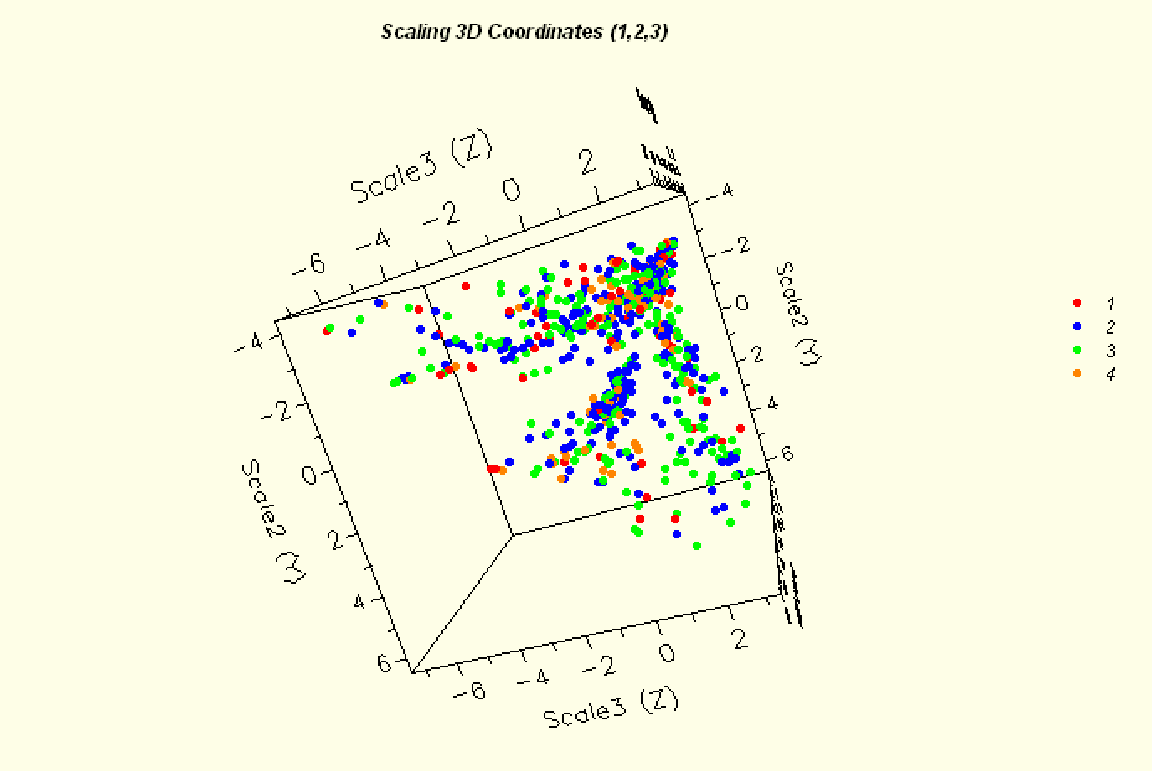

Figure 6: Potential Clusters of Observations That Were Misclassified

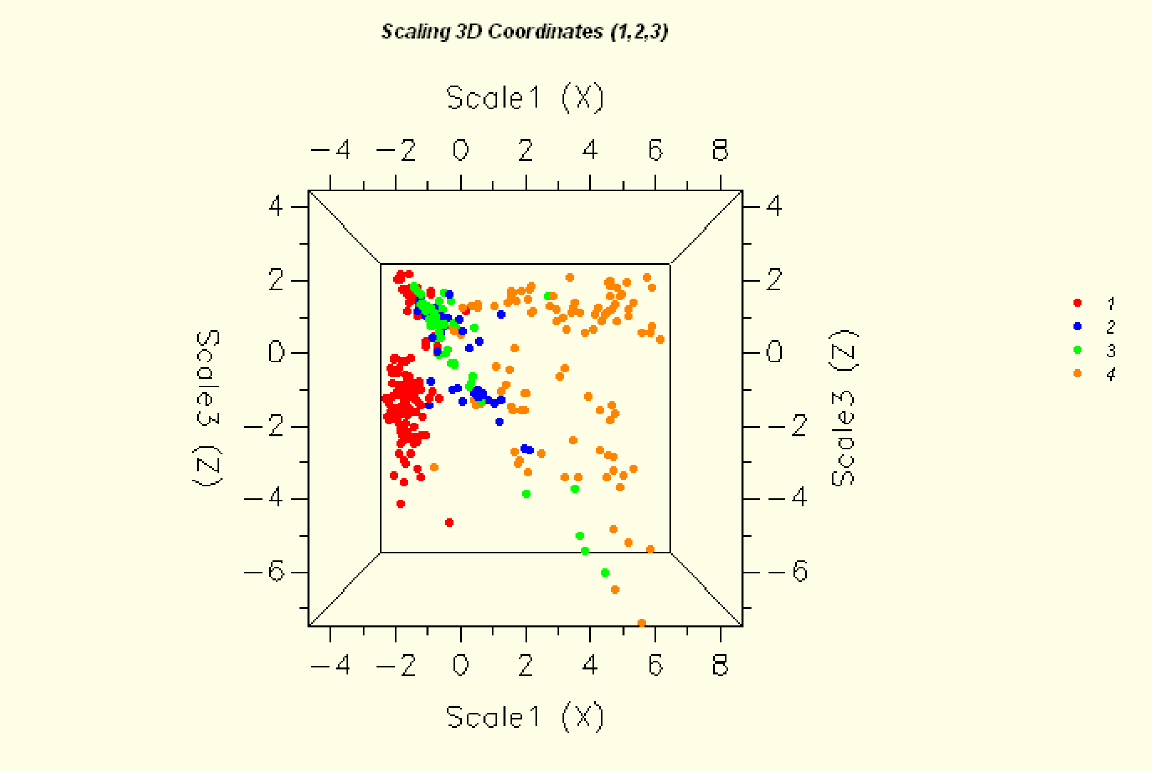

Figure 7: Potential Clusters of Observations That Were Classified Correctly.

5.0 DISCUSSION

Race is shown to be an important variable in both the logistic regression model and Random Forests, which only included sex, age, metropolitan area of residence, and their 2-way and 3-way interactions, as well as race. This result agrees with previous studies indicating a racial disparity in colon cancer stage at diagnosis exists between blacks and whites[5, 25]. Logistic regression, however, gives more details than does Random Forests. Researchers can examine what level of a specific important variable is associated with outcome and the extent of this association through the coefficient of that variable, but Random Forests does not have this feature if the predictor is a categorical variable.

Both logistic regression and Random Forests found that tumor grade, patient delay, and variables related to socioeconomic status (SES) are important variables and race became no longer important when adjustment was made for them. This result is consistent with other similar studies showing that low SES was an important predictor of advanced stage and may account for the disparity in stage at diagnosis between blacks and whites[26-28]. Robinson and Mohilever found a relationship between delay in diagnosis and more advanced stage of colon cancer, with the greater delay attributed to patient delay[29]. Logistic regression performed better than Random Forests again.

Researchers can assess how much of the excess risk for late stage at diagnosis of colon cancer can be explained by each predictor in the logistic regression by evaluting the extent of change in odds ratio and its 95% confidence interval, which cannot be achieved in Random Forests.

The correct classification rates of both logistic regression and Random Forests in this study are much lower than those from similar studies, in which the correct classification rates were at least about 80%[15, 16, 18-23]. Missing values may be the major reason why the correct classification rates in this study are so low. There were no missing values in the similar studies mentioned above. A separate category "unknown" for missing values was set in each variable in logistic regression, but this strategy may not be adequate if the proportion of missing value is large, because creating a separate category "unknown" for missing values in each variable is equivalent to replacing them with the same value. If these missing values had been known, this "unknown" category would be a mixture of different values. Most of the missing values were related to SES (socioeconomic status). SES was found to be an important predictor for colon cancer stage at diagnosis in many similar studies[25-28]. Therefore, replacing unknown SES with the same value may reduce the predictive power of the model. In the analysis with missing values excluded, although the logistic regression models did not include all important variables due to small sample size, the correct classification rates were close to those obtained from the analysis in the entire sample. This may demonstrate indirectly that a better method for handling missing values is needed. Random Forests offers an advanced method to handle missing values. This method was shown effective even when 50% of the values were missing completely at random. In this study, the proportions of missing values in some variables were about 30%, but they were not missing at random.

These missing values were caused by the fact that some patients could not be interviewed, and the ability to be interviewed was highly related to the severity of disease, so the advanced method offered by Random Forests is not effective in this study. This advanced method replaces the categorical missing values by the most frequent non-missing value in the same class where the frequency is weighted by proximity. Since missing values in this study, however, were significantly associated with severity of disease, the value of a variable in a case which has high proximity with the case of interest tends to be missing also. Table 12 lists the correct classification rate when different methods were used to handle missing values. The advanced approach is almost equivalent to the simple approach with respect to overall classification accuracy.

Table 12: Classification Accuracy Obtained From Random Forests Using Different Methods for Handling Missing Values

Method	Stage A	Stage B	Stage C	Stage D	Overall
Advanced	70.0	12.0	19.7	58.7	33.9
Simple	61.7	15.7	17.3	58.2	33.3
Category*	55.9	23.9	25.8	58.2	37.3

Note: Category means creating a separate category named “unknown” for missing values, and then applying Random Forests to the dataset including the “unknown” category.

Many studies comparing logistic regression with Classification and Regression Trees (CART) found that these two models had very similar classification accuracy[15, 16, 19, 21-23]. Random Forests is a collection of CART-like trees and is more accurate than CART. Some researchers also found that Random Forests was better than logistic regression using error rate as the criterion[18, 20]. The study presented here contradicts their conclusions. The overall correct classification rate is 47.92% for logistic regression and 33.85% for Random Forests. In addition to having higher overall correct classification rate, logistic regression captured the ordering of dependent variables better than did Random Forests. For example, logistic regression classifies 42 Stage A cases as Stage B, but only 29 as Stage C and 12 as Stage D. Random Forests, however, classifies 16 Stage A cases as Stage B and 18 as Stage C, but 30 as Stage D.

Several facts may account for the disparity between this study and previous studies. First, the sample cannot be divided into training sample and test sample due to small sample size, so the same data was used to construct the logistic regression model and to evaluate its performance, whereas Random Forests always uses a random sample which includes 63% of the original data to construct every tree leaving 37% data out of bag as a test sample. Therefore, this comparison is biased somewhat in favor of logistic regression. Second, the sample is somewhat unbalanced; Stage B and Stage C had more cases than Stage A and Stage D. Random Forests can automatically reweigh each target class to achieve equal size. This is equivalent to increasing the weight of Stage A and Stage D, and the classification accuracy of Stage A and Stage D is also increased at the cost of the accuracy for Stage B and Stage C. The ability of the model to capture ordering of the dependent variables may be damaged also since more cases were classified as Stage A and Stage D. Third, Random Forests may work more effectively in large data sets.

Although logistic regression is better than Random Forests with respect to the overall correct classification rate, Random Forests classified Stage A and Stage D cases better than did logistic regression. This demonstrates that Random Forests has good ability to handle imbalanced data. If the class of interest is relatively small or the researcher is interested in a relatively rare disease, Random Forests may work better than logistic regression by increasing the class weight.

6.0 SUMMARY AND CONCLUSIONS

In this study, two statistical methods, logistic regression and Random Forests were used to evaluate the factors which were associated with advanced colon cancer stage at diagnosis. The multi-group classification was used since the dataset had four ordered classes: Stage A, Stage B, Stage C, and Stage D. Because the quasi-complete separation problem was encountered when a polytomous logistic regression model was fitted for some variables and the proportion odds assumption was not satisfied when the proportional odds model was fitted for some other variables, both of these models were fitted to complement one another. All logistic regression models were fitted with and without missing values. Random Forests was run with equal weight in every class regardless of how many cases each class had.

Logistic regression and Random Forests largely agree with each other in variable selection. The important variables selected by these two methods were almost the same. In addition to providing a probability measure, logistic regression performs better than Random Forests with respect to the overall correct classification rate. Different strategies were used for handling missing values in logistic regression and in Random Forests. In logistic regression, missing values in each variable were grouped in a separate category labeled “unknown”, whereas in Random Forests, the advanced approach was adapted to handle missing values automatically. It seems that this advanced approach is not effective when data were not missing at random since the overall classification rate obtained from this advanced approach is very close to that obtained from the simple method.

Although the variables selected by these two methods were almost the same, the variable selection process was much simpler in Random Forests than in logistic regression. This is true especially when a large dataset is used. As mentioned above, some studies have compared logistic regression with Random Forests. Since the stepwise variable selection is open to criticism, whether their conclusion that Random Forests is better than logistic regression depends to some extent on the difference between variable selection methods is unknown. It would be of interest to use Random Forests as an automatic variable selection method and to construct the logistic regression model using selected variables, and then compare the correct classification rate obtained from the resulting logistic model with that obtained from Random Forests.

This study has two major limitations. First, the small sample size leads to partial comparison between Random Forests and logistic regression. Second, a better technique for handling missing values is needed, but is beyond the scope of the present research.

APPENDIX A

THE EXPLANATION OF VARIABLES WHICH WERE USED IN THE ANALYSIS

Variable name	Label	Source	Note
Agegp*	Age at diagnosis		Design variable
Sex*	Patients' gender		Design variable
Locn*	Metropolitan area of residence		Design variable
Race˜	Patients' race	Personal interview	Primary independent variable
Marstatr	Marital status	Personal interview	
Grade	Tumor grade	Pathological review of biopsy and surgical specimens	
Comorbs	Co morbidity	Hospital record abstract	
Smkhx	Smoking	Personal interview	
Alcohol	Alcohol use	Personal interview	
Totdlygp	Total delay category		Defined as the length of time from the patient's recognition a symptom to cancer was diagnosed
Occup	Patients' occupation	Personal interview	Defined according to standard occupational groupings as professional, technical, skilled, and unskilled
Ptdlygp	Patient delay category		Defined as the length of time from the patient's recognition a symptom to her/him first visit to a physician
Incgp	Family income category	Personal interview	
Insgp	Insurance type	Personal interview	
Ucgp	Usual health care source	Personal interview	
Povgp	Poverty index category		Calculated based on income and the family size
Educ	Education category	Personal interview	
BMIQ	Sex-adjusted body mass index quartile		Weight (kg)/Height (m) ^{2}
Inques	Interview status		
Histcat	Histology	Pathological review of biopsy and surgical specimens	
Stageˆ	Stage at diagnosis	Hospital record abstract	Outcome variable
Idnum	Identification number		

Note: 1. * denote design variables
2. ˜ denotes primary independent variable
3. ˆ denotes the dependent variable

BIBLIOGRAPHY

1. [cited; Available from: http://www.cancer.org/docroot/CRI/content/CRI_2_2_1X_How_Many_People_Get_Colorectal_Cancer.asp?sitearea>=

2. [cited; Available from: http://seer.cancer.gov/csr/1975_2002/results_single/sect_01_table.05_2pgs.pdf

3. [cited; Available from: http://seer.cancer.gov/csr/1975_2002/results_single/sect_01_table.06_2pgs.pdf

4. [cited; Available from: http://seer.cancer.gov/csr/1975_2002/results_merged/topic_mor_trends.pdf

5. Mayberry, R.M., Coates, R.J., Hill, L.A., et al, *Determinants of Black/White Difference in Colon Cancer Survival.* Journal of the National Cancer Institute, 1995. **87**: p. 1686-1693.

6. *SEER Cancer Statistics Review*. 1975-2001, National Cancer institute.

7. Hosmer, D.W., Lemeshow, S, *Applied Logistic Regression*. Second ed. 2000, New York: Wiley.

8. Collett, D., *Modelling Binary Data*. Second ed. 2003, Boca Raton: Chapman &Hall/CRC.

9. Harrell, F.E., Lee, K.L., Mark, D.B., et al, *Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat. Med, 1996. **15**: p. 361-387.

10. Harrell, F.E., Lee, K.L., Califf, R.F., et al, *Regression modeling strategies for improved prognostic prediction.* Stat. Med, 1984. **3**: p. 143-152.

11. [cited; Available from: http://nymetro.chapter.informs.org/prac_cor_pubs/RandomForest_SteinbergD.pdf

12. *RandomForests™ version 1.0 Manual*, Salford Institute.

13. [cited; Available from: http://www.math.usu.edu/~adele/forests/ENAR_files/frame.htm

14. [cited; Available from: http://www.math.usu.edu/~adele/forests/cc_home.htm

15. Rudolfer, S.M., Paliouras, G., Peers, J.S, *A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome.* Computers and biomedical research, 1999. **32**: p. 391-414.

16. Delen, D., Walker, G., Kadam, A., (2005) *Predicting breast cancer survivability: a comparison of three data mining methods.* Artificial intelligence in medicine **34**, 113-127 DIO: 10.1016/j.artmed.2004.07.002

17. Howard, J., Hankey, B.F., Greenberg, R.S., et al, *A collaborative study of differences in the survival rates of black patients and white patients with cancer.* Cancer, 1992. **69**: p. 2349-2358.

18. Lee, J. W., Lee, J. B., Park, M., et al (2005) *A extensive comparison of recent classification tools applied to microarray data*. Computational statistics & data analysis **48**, 869-885 DIO: 10.1016/j.csda.2004.03.017

19. Colombet, I., Ruelland, A.,Chatellier, G., et al (2000) *Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression*. Journal of the American medical informatics association **Volume**, 156-160

20. [cited; Available from: http://www.biomedcentral.com/1471-2105/5/154

21. Ture, M., Kurt, I., Kurum, T., et al (2005) *Comparing classification techniques for predicting essential hypertension*. Expert systems with applications **29**, 583-588 DOI: 10.1016/j.eswa.2005.04.014

22. Garzotto, M., Beer, T. M., Hudson, R.G., et al (2005) *Improved detection of prostate cancer using classification and regression tree analysis*. Journal of clinical oncology **23**, 4322-4329 DOI: 10.1200/JCO.2005.11.136

23. Stark, K.C., Pfeiffer, D.U (1999) *The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology-an example*. Intelligence data analysis **3**, 23-35 DOI: 10.1016/S1088-467X(99)00003-7

24. [cited; Available from: http://www.oncologychannel.com/coloncancer/staging.shtml

25. Mostafa, G., Matthews, B. D., Norton, H. J., et al, *Influence of demographics on colorectal cancer.* The American surgeon, 2004. **70**: p. 259-264.

26. Roetzheim, R.G., Pal, N., Tennant, C., et al, *Effects of health insurance and race on early detection of cancer.* J Natl Cancer Inst, 1999. **91**: p. 1409-15.

27. Mandelblatt, J., Andrews, H., Kao, R., et al, *The late-stage diagnosis of colorectal cancer: demographic and socioeconomic factors.* Am J Public Health, 1999. **86**: p. 1794-1797.

28. Schwartz, K.L., Crossley-May, H., Vigneau, F.D., et al, *Race, socioeconomic status and stage at diagnosis for five common malignancies.* Cancer Causes Control, 2003. **14**: p. 761-766.

29. Robinson, E., Mohilever, J., Zidan, J., et al, *Colorectal cancer: incidence, delay in diagnosis and stage of disease.* Eur J Cancer Clin Oncol, 1986. **22**: p. 157-161.