PREDICTION ACCURACY OF SNP EPISTASIS MODELS GENERATED BY MULTIFACTOR DIMENSIONALITY REDUCTION AND STEPWISE PENALIZED LOGISTIC REGRESSION

by

Amy M. Perkins

B.S. in Mathematics, Westminster College, 2007

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Amy M. Perkins

It was defended on

July 29, 2010

and approved by
Stewart Anderson, PhD
Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Lan Kong, PhD
Assistant Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Sachin Yende, MD, MS
Assistant Professor, Department of Critical Care Medicine
School of Medicine, University of Pittsburgh

Thesis Advisor: Stewart Anderson, PhD
Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

PREDICTION ACCURACY OF SNP EPISTASIS MODELS GENERATED BY MULTIFACTOR DIMENSIONALITY REDUCTION AND STEPWISE PENALIZED LOGISTIC REGRESSION

Amy M. Perkins, M.S.

University of Pittsburgh, 2010

Conventional statistical modeling techniques, used to detect high-order interactions between SNPs, lead to issues with high-dimensionality due to the number of interactions which need to be evaluated using sparse data. Statisticians have developed novel methods Multifactor Dimensionality Reduction (MDR), Generalized Multifactor Dimensionality Reduction (GMDR), and stepwise Penalized Logistic Regression (stepPLR) to analyze SNP epistasis associated with the development of or outcomes for genetic disease. Due to inconsistencies in published results regarding the performance of these three methods, this thesis used data from the very large GenIMS study to compare the prediction accuracies of 90-day mortality in SNP epistasis models. Comparisons were made using prediction accuracy, sensitivity, specificity, model consistency, chi-square tests, sign tests, and biological plausibility. Testing accuracies were generally higher for GMDR compared to MDR, and stepPLR yielded substandard performance since the models predicted that all subjects were alive at ninety days. Stepwise PLR, however, determined that IL-1A SNPs IL1A_M889, rs1894399, rs1878319, and rs2856837 were each significant predictors of 90-day mortality when adjusting for the other SNPs in the model. In addition, the model included a borderline significant, second-order interaction between rs28556838 and rs3783520 associated with 90-day mortality in a cohort of patients hospitalized with community-acquired pneumonia (CAP). The public health importance of this thesis is that the relative risk for CAP may be higher for a set of SNPs across different genes. The ability to predict which patients will experience a poor outcome may

lead to more effective prevention strategies or treatments at earlier stages. Furthermore, identification of significant SNP interactions can also expand the scientific knowledge about biological mechanisms affecting disease outcomes. Altogether, the GMDR method yielded higher prediction accuracies than MDR, and MDR performed better than stepPLR when establishing SNP epistasis models associated with 90-day mortality in the GenIMS cohort.

Keywords: biostatistics, single nucleotide polymorphism, community-acquired pneumonia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF TERMS

BIC	Bayesian Information Criterion
CAP	Community-Acquired Pneumonia
CVC	Cross-Validation Consistency
EPCR	Endothelial Protein C Receptor
GenIMS	Genetic and Inflammatory Markers of Sepsis
GMDR	Generalized Multifactor Dimensionality Reduction
HWE	Hardy-Weinberg Equilibrium
ICAM-1	Intercellular Adhesion Molecule 1
IL-1A	Interleukin 1 Alpha
IL-1B	Interleukin 1 Beta
IL-1RN	Interleukin 1 Receptor Antagonist
IL-6	Interleukin 6
IL-10	Interleukin 10
IRRR	Iteratively Reweighted Ridge Regression
LD	Linkage Disequilibrium
LR	Logistic Regression
LTA	Lymphotoxin A
MBL	Mannose-Binding Lectin
MDR	Multifactor Dimensionality Reduction
PA	Prediction Accuracy
PAI-1	Phosphoribosylanthranilate Isomerase 1
SE	Standard Error
SIRS	Systemic Inflammatory Response Syndrome

SN	Sensitivity
SNP	Single Nucleotide Polymorphism
SP	Specificity
stepPLR	Stepwise Penalized Logistic Regression
TA	Testing Accuracy
TLR-4	Toll-Like Receptor 4
TNF	Tumor Necrosis Factor
TRA	Training Accuracy

PREFACE

I would like to take the opportunity to thank my advisor Dr. Stewart Anderson and my thesis committee members Dr. Lan Kong and Dr. Sachin Yende for their time and dedication to help me succeed in this endeavor. I would also like to recognize the CRISMA Center and GenIMS Investigators for their support throughout this project and over the past three years. I also owe a lot of gratitude to my family Dennis, Diane, Tracy, Ben, and Gracyn Perkins and my friends and colleagues Jenny Slobodian, MinJae Lee, and Wenzhu Bi for their unending love and encouragement.

Futhermore, GenIMS was funded by the National Institute of General Medical Sciences R01 GM61992, with additional support from GlaxoSmithKline for enrollment and clinical data collection, and Diagnostic Products Corporation for the cytokine assays.

1.0 INTRODUCTION

Numerous clinical studies have uncovered single nucleotide polymorphism (SNP) associations with genetic diseases, but only recently have statisticians developed adequate methods to analyze high-order interactions among SNPs related to the development of disease. A polymorphism is a common variation in DNA sequence and is present in more than one percent of the population. Gene-gene interaction, or epistasis, is defined in biology as the physical interaction between biomolecules [1]. Alternatively, the statistical concept of epistasis merely describes population variation, and discovering which determinants contribute to population variation has major implications for public health [2]. Conventional statistical modeling techniques, used to detect high-order interactions between SNPs, lead to issues with high-dimensionality due to the number of interactions which need to be evaluated using sparse data. As a result, statisticians developed novel methods such as non-parametric Multifactor Dimensionality Reduction (MDR) and Generalized Multifactor Dimensionality Reduction (GMDR) and parametric forward stepwise Penalized Logistic Regression (stepPLR) to model SNP interactions.

The central aim of this thesis is to compare the performance of MDR, GMDR, and stepPLR using prediction accuracy, sensitivity, specificity, model consistency, chi-square tests, sign tests, and biological plausibility. A secondary objective of this thesis is to identify which SNPs exhibit significant, high-order interactions in models of 90-day mortality for a cohort of patients hospitalized with community-acquired pneumonia (CAP). The Genetic and Inflammatory Markers of Sepsis (GenIMS) study is one of the largest multicenter observational cohort studies (N = 2,320) with genotype data for inflammatory response genes. Hence, this thesis will analyze SNPs on the genes EPCR, ICAM-1, IL-1A, IL-1B, IL-6, IL-10, MBL, PAI-1, TLR-4, and TNF. This thesis has public health significance because the genetic

data analysis results could be used in the future for risk assessment and therapy for CAP. Based on prior studies, analysis of the GenIMS data using the GMDR method is expected to yield higher prediction accuracies and better results for discovering SNP epistasis associated with 90-day mortality when compared to MDR and stepPLR.

Additional introductory material is mentioned in Chapter 1 before discussing the background literature review in Chapter 2, statistical methods and description of the data in Chapter 3, results in Chapter 4, discussion and conclusions in Chapter 5, and the appendices thereafter.

For complex diseases, it is rare for a single gene to determine disease status or its outcome, so it is important to look concurrently at the effect of multiple SNP genotypes at different loci [3]. There are two types of models which may be characterized when looking at two-locus models. A model under which a particular genotype at one locus causes the disease independently of the genotype on the second locus is called a heterogeneity model, and an epistasis model refers to the case when the genotypes at the two loci work dependently to cause disease [3]. It is vital to understand that, occasionally, an individual with the disease-related genotype will never develop the disease. This phenomenon is measured by *penetrance* which is "a genetic term meaning the proportion of individuals with a disease-causing gene that actually shows the symptoms of the disease" [4]. Accordingly, statisticians have employed various statistical methods to characterize the associations between SNPs.

Conventional statistical modeling techniques to detect high-order interactions between SNPs, such as generalized linear regression, are rarely efficient due to the number of interactions which need to be evaluated. Statisticians refer to this problem as the "curse of dimensionality" since the number of high-order interactions can easily outnumber the total sample size [2]. Implementing generalized linear regression to model complex interactions can be highly involved, be difficult when building models, and make it hard to interpret the resulting parameter estimates [3]. There may also be problems with overfitting, correlation among the genetic factors, and empty cells in tables used to define the interactions [4]. Likewise, it is not appropriate to start model building by conditioning on significant main effects (SNPs) because potentially significant interactions between the non-significant and significant main effects could be left out of the final model. Standard additive or multiplicative

interaction models make it difficult to characterize the associations between SNPs because of the inherent complexity of epistasis [3]. Thus, MDR, GMDR, and stepPLR have been developed to uncover these complex interactions, but published articles have given inconsistent results when comparing the performance of these three methods.

Clinical studies have shown that genetics plays a role in the susceptibility to and outcomes of infectious disease. CAP is the most common infectious cause of hospitalization in developed countries and has been studied in several gene-association studies [5]. Increased mortality after CAP has been hypothesized to occur due to dysregulated host immune response to the organism. During hospitalization for CAP, several pathways, including inflammatory and coagulation cascades, are activated. Interactions between different proteins within these pathways have been well described. For example, hypersector genotypes within TNF, IL-6, and IL-1A genes may be associated with higher 90-day mortality. To date, however, few studies have examined whether epistasis across different genes is associated with outcomes of CAP.

Recalling the central aim of this thesis, the prediction accuracies of MDR, GMDR, and stepPLR will be used to compare the methods' performance in discovering which SNPs exhibit significant interactions in models of 90-day mortality in a cohort of CAP patients. Results from each method will give the best one-, two-, three-, and four-locus models for the SNPs within each gene and again across all genes. The prediction accuracy, sensitivity, specificity, model consistency, and the chi-square and sign tests (for significance of the association between the interaction and case-control status) of the final models will be the basis for comparison of the methods. The particular SNPs in the final models will be recorded to assess whether the three methods tend to include interactions of the same SNPs. In addition, MDR will conduct a hierarchical cluster analysis to characterize the relationship of the SNPs in the model as epistatic, independent, or correlated. Consequently, any conclusions about the performance of MDR, GMDR, and stepPLR will not be an absolute resolution but merely a look at how well they function for the GenIMS data set.

2.0 BACKGROUND LITERATURE REVIEW

Many publications have identified the methods that work best for determining SNP epistasis in case-control studies, but they have given inconsistent results which often depend on the composition of the data. Hua He et al. [6] published a study on simulated data and kidney transplant patients; these investigators excluded any SNPs with a minor allele frequency less than five percent or when more than ten percent of the data were missing. In addition, they conducted Fisher's exact test for all the SNPs and selected only the top 100 most significant SNPs for interaction detection purposes. This, however, is a potential limitation to their study since the authors could have overlooked important interactions for which the main effects were not significant. They concluded that stepPLR is more powerful than MDR when the effects of the SNPs are additive, but MDR performs better when complex interactions are observed. Moreover, "MDR was particularly good in detecting the weak effects of a purely epistatic interaction" [6].

An earlier study by Park and Hastie [4] suggests that stepPLR is more powerful than MDR particularly in the presence of multiple SNP interaction factors, for MDR was only able to discover a portion of the significant interactions using simulated data. Comparing the performance of stepPLR to MDR in a real data set, the authors noted that stepPLR achieved a higher sensitivity than MDR if the specificity for stepPLR was fixed equivalent to that of MDR (and likewise for the specificity) [4]. Thus, epistasis models generated by stepPLR yielded better classification of cases and controls than MDR.

Lou et al. [2] proposed the GMDR method as an extension of MDR to adjust for discrete or quantitative baseline covariates and to allow for an unequal number of observations for cases and controls. GMDR produced the same results as MDR when analyzing a simulated data set with a case-control ratio of 1:1 and no adjustment for covariates. After adjusting

for meaningful covariates, GMDR yielded higher prediction accuracy and cross-validation consistency in models of high-order SNP interactions. Thus, accounting for the variation in a covariate that is known to increase disease risk will yield a model that can more accurately predict disease status. The authors applied the two methods to a real data set with 191 smokers and 191 non-smokers and found that, although both indicated the same best four-locus model, GMDR produced a higher prediction accuracy, a higher cross-validation consistency, and a significant p-value [2].

A study recently published in *Respiratory Medicine* [7] investigated polymorphisms in anti-inflammatory and inflammatory genes and their relationship with susceptibility, severity of illness, and outcome in adult CAP patients. The authors conducted this study because variation in the clinical presentation of CAP patients can be affected by "different pathogens, variable virulence in different strains of microorganisms, increasing age, and underlying diseases...but genetic variability affecting the host response may also influence the susceptibility to it and the severity and outcome of infection" [7]. This suggests the possibility of gene-environment interactions as well. The following methods were implemented to determine the significance of these relationships: chi-square tests, odds ratios, binary logistic regression, Kaplan-Meier survival curves, and log-rank tests. No significant associations were found between polymorphisms on the genes TNF, LTA, IL-6, or IL1-RN and disease severity or outcome, nor was there a significant interaction between the genotypes for SNPs on TNF and LTA [7]. Unfortunately, these investigators did not use MDR, GMDR, or stepPLR which could have augmented a comparison of their performance.

Gallagher et al. [8] published a study on associations between inflammatory response genes and severity of systemic inflammatory response syndrome (SIRS) in CAP patients. These researchers found a significant, increasing linear trend between the IL10_M1082 G allele and SIRS score; this G allele was also associated with mortality in CAP patients. No significant associations were observed between disease severity and TNFA_M308 or IL6_M174, and no significant interaction was found between these three SNPs [8].

Although many published studies have tried to characterize the relationship between SNP interactions and CAP outcomes, it is difficult to make generalizations from their conflicting results due to methodological problems. Discrepancies often involve the associations

between the SNP genotypes and disease risk, which is often the case for genes IL-10 and TNF. The methodological problems may lead to problematic interpretations and include poor statistical power, variant populations with unknown confounding factors, cases of linkage disequilibrium, or stratified populations [7]. More importantly, the statistical power of the methods can depend on the allele frequencies of the chosen SNPs [6]. Poor results may be obtained when including SNPs which are homogeneous with respect to genotype (i.e. when the minor allele frequencies are too small to identify significant differences in case-control status). Therefore, further evaluation of statistical methods and SNP associations and larger population samples are necessary to reach a valid consensus.

3.0 METHODS

3.1 MULTIFACTOR DIMENSIONALITY REDUCTION

Multifactor dimensionality reduction (MDR) is a data mining procedure which reduces high-dimensional genetic data to a single dimension using constructive induction [1]. This method was proposed by Ritchie et al. in 2001 [9], and the open-source software was developed by Hahn, Ritchie, and Moore [10]. MDR is non-parametric and does not require a predefined model for the gene-gene interactions. After reducing the dimensionality of the SNP data, all possible n-locus SNP interactions are tested for association with a complex genetic disease, for this method can utilize relatively small sample sizes and detect interactions without assessing main effects [1].

MDR requires that the data set has a case-control ratio of 1:1 [9]; in other words, it must include an equal number of cases and controls. The MDR software (version 1.8 beta) requires that the data be formatted as a text (.txt) file with the first few columns containing the genetic data and the last column designating the case-control status (a binary variable coded as zero or one). The SAS code (software version 9.2) used to create the data files is given in Appendix A. The SNPs are categorical variables with three levels: homozygous for the common allele coded as zero, heterozygous coded as one, and homozygous for the less common allele coded as two. Any missing data must be addressed before implementing the MDR software. One way is to delete subjects and/or SNPs until the data set is square, but this is not ideal since the sample size is largely decreased and important information can be lost [11]. A second option is to code the missing data as another level of the categorical variable; for example, a missing value could be represented by a '4' or '9' [11]. MDR then incorporates this fourth level of information into the models. This procedure is appropriate

only when there are relatively few missing genotypes and they are missing at random. Lastly, the missing data may be imputed using frequency-based or multivariate methods, but this approach is also only appropriate with a few missing data points which must be missing at random [11]. Imputation is recommended most often because it yields a complete data set without any loss of information and provides a simpler interpretation of the final model without a fourth level coded for missing data. Once the data set is correctly formatted, the MDR software can be executed.

After importing the data set as a text file, there is an option to filter the genetic data to get a smaller subset which is more likely to exhibit dependencies or interactions between SNPs [11]. One may select the Relief-F, chi-square, or odds ratio filter methods and the subset size which is quantified by a number, percentage, or threshold criteria for the subset of SNPs. For example, the subset may include the top ten SNPs, the top twenty-percent of SNPs, or those SNPs with a p-value less than 0.20. These filters can yield poor results [12], but filtering is still a better option than conditioning on main effects (significant univariate SNPs) because many, potentially significant interactions would be left untested [11]. This thesis implemented the Relief-F algorithm for all fifty-five SNPs combined to obtain the top twenty SNPs for the epistasis models. Filtering the SNPs based on biological plausibility is another option for excluding a number of extraneous SNPs. The advantages of filtering are that it reduces the number of SNPs which need to be exhaustively tested for interactions in MDR and that it is computationally tractible [11]. One of the disadvantages of filtering is that some important SNPs can be excluded from the analysis since the adequacy of the filtered subset fully relies on the performance of the filter [11]. The MDR software configuration is briefly discussed in the next paragraph.

The MDR software allows the user to configure the analysis to obtain specific outputs. One is allowed to specify a random seed which is used to divide the data set for cross-validation [11]. The attribute count range denotes the order of the interactions to be tested, and, by default, MDR selects the best 1-, 2-, 3-, and 4-locus models [11]. The default cross-validation count is set at ten, but the user may request a different number of data subsets for cross-validation. Selecting the compute fitness landscape box produces a list of all-possible SNP interactions and their training accuracies, so determination of the second or third best

models is possible [11]. A model's training accuracy (TRA) ranges from zero to one, is calculated on the training data set and averaged across all ten cross-validation intervals, and is defined by the following formula:

$$TRA = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Here, TP is the number of true positives; TN is the number of true negatives; FP is the number of false positives; and FN is the number of false negatives. The testing accuracy (TA) also refers to the proportion of subjects correctly classified as cases or controls but is calculated using Equation 3.1 on the testing data set. This thesis substitutes the term prediction accuracy (PA) for the testing accuracy of the model. Finally, the options for search type are exhaustive, forced, or random. As stated by Moore, "MDR has traditionally carried out an exhaustive search of all possible 2-way, 3-way, and up to n-way combinations of SNPs" [11]. A forced analysis allows one to type in the names of desired SNPs to obtain estimates of the prediction accuracies for the 2nd or 3rd best models or an unbiased estimate of the prediction accuracy when the cross-validation consistency does not equal ten [11]. Cross-validation consistency (CVC) is defined as the number of times (out of ten) that the particular factors were selected as the best model during cross-validation. After the MDR options are properly configured, the user clicks the run analysis button.

The MDR software implements a series of six steps for each of the cross-validation intervals to generate the best n-locus model. In Step 1, the data set is partitioned into ten subsets (or the specified magnitude of the cross-validation); nine of the subsets are used as the training set, and the remaining subset constitutes the independent testing set [10]. Step 2 selects all possible combinations of n factors from the list of available SNPs. Step 3 utilizes the training set to place the n factors and their categorical genotypes into a contingency table representing the n-dimensional space. For example, two SNPs each with three possible genotypes yields a 3-by-3 contingency table, and the subjects are placed in the table cell corresponding to their haplotype. Step 4 labels each multifactor cell in the contingency table as high-risk or low-risk based on the ratio of cases to controls [10]. If the ratio is greater than or equal to one, the cell is labeled high-risk and shaded a dark gray in the table; the cell is labeled as low-risk and shaded light gray if the ratio does not exceed one. The haplotypes

that are not represented in the data set are left as blank, white cells. This explains how the data are reduced from n-dimensional space to one-dimensional space, for a subject's haplotype (cell) is a categorical variable with two possible levels - high-risk or low-risk. Step 5 estimates the training accuracy of the n-locus model using cross-validation on the training set [10]. The naïve Bayes classifier uses a subject's n genotypes to identify the subject as high-risk (case) or low-risk (control) using the following formula:

$$v_{NB} = \underset{v_j \in V}{arg\,max}\, P(v_j) \prod_{i=1}^{n} P(a_i|v_j) \tag{3.2}$$

Here, v_j represents one of a set of V classes for $j = 1, 2$; a_i is one of n factors (SNPs) [1]. Afterwards, the training accuracy is calculated for each model using the classification results for the training data set. MDR selects the model with the highest training accuracy as the best n-locus model for the first cross-validation, and the CVC is tabulated at this stage. Step 6 utilizes the independent testing set along with the naïve Bayes classifier to estimate the prediction accuracy for the selected n-locus model. Steps 1-6 are repeated for each of the ten cross-validation intervals, and the prediction accuracies are averaged for each n-locus model. Lastly, MDR uses CVC to determine the best n-locus model across the 10-fold cross-validation [6], and the chi-square statistic is calculated for this model to determine whether the factors are significantly associated with the genetic disease.

The final output displayed by the MDR software includes several criteria to summarize the performance of the n-locus model and to select an overall best model. The final best 1- to n-locus models are displayed along with the training accuracy, prediction accuracy, sensitivity (SN), specificity (SP), cross-validation consistency, and chi-square test p-value (χ^2). A CVC less than ten indicates that the estimate for the prediction accuracy is biased, so the model should be rerun with a forced analysis to obtain an unbiased estimate for the PA [11]. It is up to the user to select the overall best epistasis model based on maximum prediction accuracy and CVC. In the case where two different models yield the highest PA and the highest CVC, the more parsimonious model should be selected [10]. It is also plausible that the second or third best n-locus model shown in the fitness landscape could be a better predictor of disease status if a forced analysis of the secondary models produces a higher prediction accuracy [11]. The "rule of thumb" for interpreting the PA reveals that

a number greater than 0.54 is intriguing, a number greater than 0.60 is usually statistically significant, and a number beyond 0.70 is rarely encountered [11]. Moreover, overfitting is a potential problem when the PA continues to rise as additional factors are included in the model [11]. MDR also displays the dendrogram tab which visually depicts a hierarchical cluster analysis to indicate epistasis, correlation, or independence between factors. The length of the connecting lines implies the strength of the relationship; shorter lines represent stronger relationships. The color of the lines reveal the type of relationship as epistatic (red or orange), correlated (blue or green), or independent (tan) [11]. Fortunately, the MDR output is easily saved for future reference, and diagrams may be exported for use in papers or presentations.

MDR was developed to better detect high-order interactions among SNPs associated with genetic diseases and inherently has a number of advantages and disadvantages. The advantages of the MDR method are 1) it has the ability to simultaneously detect and characterize multiple SNPs associated with genetic disease status; 2) it is nonparametric; 3) it is genetic model-free; and 4) its cross-validation strategy limits the number of false-positives from multiple testing [9]. MDR is also able to detect interactions in the absence of any significant main effects. The disadvantages of MDR are a) it lacks power when the real effect is additive instead of multiplicative; b) it cannot label empty cells in high-dimensional contingency tables as high-risk or low-risk; and c) it suffers from an unstable risk assignment when the ratio of cases to controls is approximately one [4]. Furthermore, this method is computationally intensive when more than ten SNPs need to be analyzed, is often hard to interpret, is only applicable to balanced case-control studies, and is limited in its capacity to predict disease status in an independent testing set when the best model exhibits high-dimensionality and the sample size is small [9]. Currently, the MDR software cannot incorporate non-genetic characteristics which could greatly improve the best model's prediction accuracy nor can it easily run data simulations. Hence, MDR has overcome many of the computational hurdles for discerning gene-gene interactions associated with genetic disease, but there are still instances where it could be improved.

3.2 GENERALIZED MULTIFACTOR DIMENSIONALITY REDUCTION

Generalized Multifactor Dimensionality Reduction (GMDR) is another constructive induction approach which extends the MDR method to incorporate unbalanced study designs, meaningful non-genetic covariates, and continuous outcomes [2]. This method is also non-parametric and genetic model free. MDR is a special case of GMDR, and the two methods are equivalent for balanced case-control studies without any covariate adjustment [2]. GMDR attempts to overcome the "curse of dimensionality" by using the same strategy as MDR to reduce the dimensionality of the data from n to one. Another key difference is that the one-dimensional classification of the contingency table cells as high-risk or low-risk is based on the score of a generalized linear model instead of the ratio of cases to controls [2]. Hence, GMDR is able to use more of the available data for building the SNP epistasis models and improving the prediction accuracies.

The generalized linear model underlying the GMDR method defines y_i as the binary disease status for subject i with expectation $E(y_i) = \mu_i$. The model is of the form:

$$l(\mu_i) = \alpha + \mathbf{x}_i^T \beta + \mathbf{z}_i^T \gamma \tag{3.3}$$

where $l(\mu_i)$ is the logit link function, α is the intercept, $\mathbf{x}_i$ is the vector of genetic predictor-variables, $\mathbf{z}_i$ is the vector of non-genetic covariates, and β and γ are the respective parameter vectors. The score-based statistics derived from the model in Equation 3.3 are calculated for each subject i using the following formula:

$$S_i^T = \sum_j \frac{x_{ij}(y_i - \hat{\mu}_i)}{\sqrt{\hat{Var}(y_i)}} \tag{3.4}$$

where S_i^T is a measure of the normalized contributions to the scores of the genetic effects and $\hat{Var}(y_i)$ is the estimated variance of y_i [2].

The GMDR software (version 0.7 beta) was developed in 2007 by Lou, Chen, Yan, and Li to implement the GMDR method [2]. Missing data must be deleted, imputed, or categorically coded as in MDR, and the text data files (.txt) are formatted exactly the same except the GMDR data files may include non-genetic covariates. The SAS code used to create the data files is given in Appendix A. The GMDR method substitutes the average score statistic for

the ratio of cases to controls in each cell, but the remaining steps are identical to those carried out by MDR. The maximum likelihood estimates and scores for each subject are calculated under the null hypothesis $H_0 : \beta = 0$ (ie. there are no significant genetic factors or epistasis) [2]. Step 3 of the GMDR method computes the average score statistic in each cell. In Step 4, the cells are labeled as high-risk or low-risk based on whether the average score statistic meets/exceeds or does not exceed the threshold value, respectively. The threshold equals the ratio of cases to controls in the data set, and a value of zero is equivalent to a one-to-one case-control ratio. Due to the high dimensionality of the data and the comparatively small sample sizes, some of the cells can be empty in the training set and not in the testing set. The GMDR method identifies these as misclassification cells when adding together the scores of the high-risk and low-risk cells. This strategy penalizes the use of many subdivisions in a small sample and is consistent with statistical parsimony [2]. The GMDR software displays the final best 1- to n-locus models along with the training accuracy, prediction accuracy, sensitivity, specificity, cross-validation consistency, and chi-square test and sign test p-values. As in MDR, the best n-locus model is selected based on the highest training accuracy, and the maximum balanced prediction accuracy and CVC are used to determine the overall best epistasis model. The balanced prediction accuracy is calculated using the formula (sensitivity + specificity)/2 to yield an unbiased estimate for unbalanced case-control studies. Lastly, the chi-square and sign tests determine whether the factors are significantly associated with the genetic disease.

The GMDR software produces the same output as MDR with the addition of the sign test, and the results are interpreted in the same way. The prediction accuracy indicates the proportion of subjects that were correctly classified as cases or controls. For example, a prediction accuracy of 0.60 implies that 60% of the subjects were correctly classified as cases or controls. Sensitivity is defined as the proportion of subjects, who died within ninety days, that were correctly classified as cases. A sensitivity of 0.85 implies that 85% of subjects, who died within ninety days, were actually classified as cases. Similarly, specificity is defined as the proportion of subjects, who were alive after ninety days, that were correctly classified as controls. For example, a specificity of 0.83 implies that 83% of subjects, who were alive after ninety days, were actually classified as controls. Recall that CVC is measured on a ten

point scale for 10-fold cross-validation, so a CVC of eight indicates that the particular model was selected eight times out of ten as the best model. Additionally, a significant chi-square test reveals that the SNP or interaction between SNPs is associated with 90-day mortality. A significant sign test suggests that the best model with one or more SNPs is significantly better than the null model. This thesis used an $\alpha = 0.05$ level of significance to interpret the test results.

GMDR was developed to overcome some of the limitations present in the MDR method. The advantages of the GMDR method are 1) it can adjust for non-genetic covariates; 2) it can analyze dichotomous and continuous traits; and 3) it can make use of unbalanced case-control studies [2]. When a trait is significantly associated with the covariate(s), GMDR will produce better prediction accuracies. Furthermore, GMDR does not depend on score or likelihood properties to calculate the average score statistics in each cell. Similar to MDR, the disadvantages of GMDR include a) the intensive, high-dimensional computations; b) the issue with empty cells in the training data which are not empty in the testing data set [2]; and c) the inability to easily run data simulations. The authors are planning to extend GMDR to family-based study designs and are working on better optimal algorithms to improve the performance of the method [2].

3.3 STEPWISE PENALIZED LOGISTIC REGRESSION

Stepwise Penalized Logistic Regression (stepPLR) modifies the standard logistic regression (LR) method by adding a quadratic penalization of the L_2-norm of the coefficients in SNP epistasis models. Stepwise PLR was proposed by Park and Hastie [4] in 2007 and has a number of advantages over standard LR. The coefficients of the SNPs and their interactions are estimated by minimizing the following penalized negative log-likelihood:

$$L(\beta_0, \beta, \lambda) = -l(\beta_0, \beta) + \frac{\lambda}{2}||\beta||^2 \tag{3.5}$$

where β_0 is the vector of initial values for the genetic variable coefficients, β is the vector of coefficients which must be estimated, $l(\beta_0, \beta)$ is the binomial log-likelihood, λ is the

regularization parameter, and $||\beta||^2$ is the L_2-penalization. The value for the constant λ is determined using cross-validation and selecting the value that yields the largest log-likelihood [6]; a large λ will fit the data more smoothly. This thesis found $\lambda = 8$ to perform best even though the default value is 10E-4. The model fitting utilizes the iteratively reweighted ridge regression (IRRR) algorithm to estimate the coefficients [4]. The final model is selected according to the Bayesian information criterion (BIC) with forward selection followed by backward deletion [6]. Stepwise PLR incorporates the asymmetric hierarchy principle to allow the SNP interaction terms to enter the model more easily [4]. Under this principle, any factor or interaction of factors in the model can form a new interaction with any other single factor even if the single factor has not been added to the model. The addition or deletion of genetic factors at each step is based on the score defined by the following formula:

$$S = d + cp * df \tag{3.6}$$

where d is the deviance, cp is the complexity parameter equal to the log of the sample size, and df is the effective degrees of freedom. Note this is not the same as the score statistic in the GMDR method. For stepPLR, the final best model is the one with minimum score S [6].

Park and Hastie developed the stepPLR software (version 0.91) as an R software package downloadable for Mac and Unix platforms. For this software package, any missing data must be imputed or deleted before importing the data sets. The SAS code used to create the data files is given in Appendix A, and the R code (version 2.10.1) used to implement stepPLR is given in Appendix B. Unfortunately, the software does not allow the user to specify the largest order (n) of the interactions to be tested; the user can only restrict the number of terms in final model. Setting "max.terms = 4" permits the final model to include up to four SNPs. Another option is to use only forward selection for the model fitting, but this tended to yield the same results. The software output displays the coefficient estimates, standard errors, Z-statistics, Z-test p-values, and the score and prediction error of the model. After subtracting the prediction error from one to obtain the prediction accuracy, the performance of stepPLR can be compared to MDR and GMDR.

Stepwise PLR has many advantages over standard LR, but it is not without its own limitations. In standard LR, there are problems with overfitting due to the large number of

parameters that need to be estimated in relatively small data sets. The SNPs can be correlated which degrades the fit of the model under standard LR, and special parametrization is needed for empty or nearly empty cells in the contingency table [4]. With so few data points in some of the cells, there is large instability in the parameter estimation [6], and these issues are magnified for higher order interactions. Since stepPLR employs quadratic penalization, it has a number of advantages: 1) the large number of coefficients can be fit in a stable fashion amid high-dimensionality; 2) the use of dummy variables to code each SNP genotype yields direct interpretations of the interaction terms and removes the collinearity problem which degrades the fit of the model; 3) sparse data in the contingency tables do not affect the stability of the model fitting; and 4) the sample size does not limit the number of SNPs that can be included in the final model [4]. Stepwise PLR also performs well when the effect of multiple SNPs is additive [6]. The disadvantages of stepPLR are a) the strength of the main effects can be reduced for large values of λ; b) the penalization tends to break up a large main effect coefficient into smaller interaction pieces; c) stepPLR possibly includes interaction terms just to account for the regularized main effects [4]; and d) the method can only handle genetic data and a binary trait and does not incorporate non-genetic covariates. Therefore, stepPLR overcomes many of the standard LR limitations, but additional improvements could be made.

3.4 DESCRIPTION OF DATA

Data obtained from the Genetic and Inflammatory Markers of Sepsis (GenIMS) study were analyzed by each of the three methods to characterize high-order SNP interactions related to 90-day mortality. This cohort study obtained genetic and clinical data from 2,320 patients (aged eighteen or older) who were admitted to the emergency department and subsequently diagnosed with community-acquired pneumonia (CAP). Written consent was obtained from the patients (or proxies) after the study was approved by the institutional review boards of the University of Pittsburgh and all twenty-eight participating hospital sites [13]. The GenIMS investigators sought to uncover patterns between anti-inflammatory and pro-

inflammatory biomarker levels and the primary outcomes defined as 90-day mortality (N = 212, 11%) or severe sepsis (N = 583, 31%). In-hospital mortality was determined by study nurses, and post-discharge mortality was determined by telephone interviews and a National Death Index search [13].

This analysis was limited to 1,704 CAP patients of White race and SNPs on the EPCR, ICAM-1, IL-1A, IL-1B, IL-6, IL-10, MBL, PAI-1, TLR-4, and TNF genes. Each SNP has three possible genotypes which were treated as a categorical variable with the following three categories: homozygous for the common allele (coded as 0), heterozygous (coded as 1), and homozygous for the less common allele (coded as 2). The bases for the alleles are adenine (A), cytosine (C), thymine (T), and guanine (G). Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) were assessed before removing SNPs with only two genotypes and a minor genotype frequency less than 1% or with more than 10% missing data. Incorporating domain-specific knowledge also reduced the number of genes which were anticipated to be significantly associated with 90-day mortality. Out of the 1,704 subjects, 47.1% were females with an average age of 67.9 years, and the males had an average age of 68.1 years. For statistical analysis purposes, cases were defined as subjects who died within ninety days, and subjects still alive at ninety days were classified as controls. There were 191 cases (average age 78.4 years) and 1,513 controls (average age 66.7 years), but these numbers changed for each particular SNP due to missingness or the MDR requirement for an equal number of cases and controls. Prior studies using human subjects included 120 cases and 120 controls, but this sample size was too small to detect interactions. It was expected that 191 cases for this analysis would be sufficient. This data set was selected for analysis because CAP is the leading cause of sepsis, and many studies have been published regarding genetic polymorphisms related to the development of sepsis. Taking this a step further, high-order SNP interactions were analyzed for association with 90-day mortality to evaluate whether MDR, GMDR, or stepPLR yielded the best performance for the GenIMS data set.

4.0 RESULTS

The selected SNPs on each of the ten genes were tested for Hardy-Weinberg equilibrium and linkage disequilibrium before implementing MDR, GMDR, and stepPLR to detect SNP epistasis associated with 90-day mortality. The results of these tests, along with the genotypes and percentages of missing data, are given in Table 1. Some SNPs with more than ten percent missing data were included in the analysis to adequately conduct each method, but this did not largely affect the sample sizes for each gene. Each SNP was found to meet Hardy-Weinberg equilibrium because all of the p-values were greater than $\alpha = 0.05$. Linkage disequilibrium among the SNPs was tested within each gene, and all but two of the genes yielded the necessary significant p-values. The genes TLR-4 and TNF had p-values greater than $\alpha = 0.05$ but were analyzed nonetheless for comparison to other studies.

Table 1: Hardy-Weinberg equilibrium and linkage disequilibrium tests

Gene	**SNP**	**Genotypes**	**Missing (%)**	**HWE**[a]	**LD**[b]
EPCR	rs2069940	CC, CG, GG	28.66	0.99	<0.0001
	rs867186	AA, AG, GG	15.73	0.12	
ICAM-1	rs1799969	AA, AG, GG	0.39	0.20	<0.0001
	rs5030340	CC, CT, TT	9.27	0.99	
	rs281432	CC, CG, GG	5.69	0.77	
	rs281438	GG, GT, TT	10.28	0.49	

Table 1 continued.

Gene	SNP	Genotypes	Missing (%)	HWE[a]	LD[b]
IL-1A	IL1A_M889	CC, CT, TT	5.87	0.95	<0.0001
	rs2856838	CC, CT, TT	4.05	0.28	
	rs1894399	AA, AG, GG	4.21	0.58	
	rs3783520	AA, AG, GG	5.14	0.53	
	rs3783546	CC, CG, GG	9.97	0.44	
	rs1878319	CC, CT, TT	14.17	0.71	
	rs2856837	CC, CT, TT	21.26	0.45	
	rs2856836	CC, CT, TT	7.32	0.36	
IL-1B	IL1B_M511	AA, AG, GG	5.69	0.92	<0.0001
	IL1B_3957	AA, AG, GG	5.58	0.11	
	rs1071676	CC, CG, GG	12.46	0.34	
	rs3917365	CC, CT, TT	8.72	0.99	
	rs1143623	CC, CG, GG	7.32	0.83	
IL-6	IL6_M174	CC, CG, GG	6.87	0.27	<0.0001
	rs2069827	GG, GT, TT	1.25	0.73	
	rs2069861	CC, CT, TT	3.43	0.99	
	rs2069849	CC, CT, TT	1.09	0.38	
	rs1548216	CC, CG, GG	5.37	0.11	
	rs1800795	CC, CG, GG	5.14	0.47	
IL-10	IL10_M1082	CC, CT, TT	6.28	0.69	<0.0001
	IL10_M819	CC, CT, TT	5.58	0.73	
	rs1800872	AA, AC, CC	0.70	0.36	
	rs3024505	CC, CT, TT	3.35	0.91	
	rs1800894	AA, AG, GG	5.37	0.32	
	rs3024496	CC, CT, TT	5.30	0.73	

Table 1 continued.

Gene	SNP	Genotypes	Missing (%)	HWE[a]	LD[b]
MBL	rs11003125	CC, CG, GG	49.14	0.88	<0.0001
	rs1800450	AA, AG, GG	49.14	0.15	
	rs1800451	AA, AG, GG	49.14	0.99	
	rs7096206	CC, CG, GG	49.14	0.78	
	rs5030737	CC, CT, TT	49.14	0.35	
PAI-1	rs13238709	CC, CT, TT	11.60	0.76	<0.0001
	rs2227683	AA, AG, GG	6.15	0.72	
	rs2227665	AA, AG, GG	9.03	0.99	
	rs7242	GG, GT, TT	10.67	0.95	
TLR-4	TLR4_M896	AA, AG, GG	5.98	0.65	>0.05
	rs11536898	AA, AC, CC	8.26	0.69	
	rs11536897	AA, AG, GG	3.66	0.76	
	rs12344353	CC, CT, TT	7.94	0.99	
	rs5030725	GG, GT, TT	3.27	0.99	
	rs2770146	AA, AG, GG	7.87	0.26	
	rs1927912	AA, AG, GG	5.61	0.99	
	rs5030729	AA, AG, GG	5.53	0.99	
	rs1927914	CC, CT, TT	6.07	0.79	
	rs5030717	AA, AG, GG	6.54	0.99	
TNF	TNFA_M308	CC, CT, TT	5.81	0.99	>0.05
	TNF_238	AA, AG, GG	12.09	0.06	
	TNF_857	CC, CT, TT	7.02	0.90	
	rs3093672	CC, CT, TT	1.25	0.99	
	rs3093662	AA, AG, GG	14.88	0.99	

[a] HWE = Hardy-Weinberg Equilibrium p-value, [b] LD = Linkage Disequilibrium p-value

4.1 EPCR RESULTS

The MDR method utilized 103 cases and 103 controls in the data set for the EPCR SNPs, and the results are shown in Table 2. The overall best model was the one-locus model with rs2069940 because it had the highest prediction accuracy 0.5146. The specificity of the model was relatively good at 0.8252, and the CVC was high at nine. This was a very poor model because the sensitivity was only 0.2039, and the non-significant chi-square test indicated that rs2069940 was not significantly associated with 90-day mortality (p = 0.87).

For the GMDR method, 103 cases and 715 controls were used to generate the models for the EPCR SNPs. The results are shown in Table 2. The prediction accuracies and CVCs were identical for the one- and two-locus models. Thus, rs2069940 was selected as the overall best model because it was more parsimonious. The prediction accuracy was low at 0.5228. A very low sensitivity of 0.2036 was observed, but the specificity of the model was much higher at 0.8420. The CVC was very high at ten. Despite that, the chi-square test concluded that rs2069940 was not significantly associated with 90-day mortality (p = 0.62). The sign test was borderline significant (p = 0.05). Altogether, MDR and GMDR both chose rs2069940 as the overall best model, but GMDR had slightly better prediction accuracy than MDR.

Table 2: EPCR SNP epistasis models generated by MDR and GMDR

Method	**Best Models**	**TRA[a]**	**PA[b]**	**SN[c]**	**SP[d]**	**CVC[e]**	χ^{2}[f]	**Sign[g]**
MDR	rs2069940	0.5146	0.5146	0.2039	0.8252	9	0.87	NA
(N=206)	rs2069940, rs867186	0.5210	0.4660	0.4660	0.4660	10	0.76	NA
GMDR	rs2069940	0.5229	0.5228	0.2036	0.8420	10	0.62	0.05
(N=818)	rs2069940, rs867186	0.5229	0.5228	0.2036	0.8420	10	0.62	0.05

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity, [e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 103 cases and 715 controls to determine the best epistasis model for EPCR. The model shown in Table 3 only included main effects for rs2069940 and rs867186 and no interactions. The Z-test for rs2069940 indicated that this SNP was not significantly associated with 90-day mortality (p = 0.18). The Z-test for rs867186 suggested that this SNP was not significantly associated with 90-day mortality as well (p = 0.54). The standard errors were large compared to the magnitude of the estimates which signified a poorly fit model. The prediction accuracy was high at 0.8741. This estimate, however, was biased because this stepPLR model predicted all subjects were controls.

Table 3: EPCR SNP epistasis model generated by stepPLR

Best Model (N=818)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
rs2069940	0.1074 (0.0793)	0.18	632.31	0.8741
rs867186	-0.0697 (0.1125)	0.54		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrated that EPCR SNPs rs2069940 and rs867186 exhibited weak epistasis as indicated by the long, red line in Figure 1.

Figure 1: MDR interaction dendrogram for EPCR SNPs. The long, red line indicates weak epistasis between rs2069940 and rs867186.

4.2 ICAM-1 RESULTS

The MDR method utilized 118 cases and 118 controls in the data set for the ICAM-1 SNPs, and the results are shown in Table 4. Even though the three-locus model had the highest prediction accuracy 0.5636, the overall best model was the one-locus model with rs281432 because it had a CVC of ten and was the more parsimonious model. The prediction accuracy was fairly low at 0.5297. Unfortunately, this is a very poor model since the specificity was only 0.2712. The sensitivity was also poor at 0.7881. Finally, the chi-square test concluded that rs281432 was not significantly associated with 90-day mortality ($p = 0.74$).

For the GMDR method, 118 cases and 845 controls were used to generate the models for the ICAM-1 SNPs. The results are shown in Table 4. The same one-locus and four-locus models were selected by GMDR and MDR. Again, the three locus model exhibited the highest prediction accuracy 0.5369, but the one-locus model with rs281432 was selected as the overall best model because it had a CVC of ten and was the more parsimonious model. The prediction accuracy was fairly low at 0.5219. This was also a poor model since the specificity was only 0.2589; the sensitivity was also low at 0.7848. In addition, the chi-square test concluded that rs281432 was not significantly associated with 90-day mortality ($p = 0.38$), and the sign test was also non-significant ($p = 0.17$). Thus, MDR had slightly better prediction accuracy than GMDR.

Table 4: ICAM-1 SNP epistasis models generated by MDR and GMDR

Method	**Best Models**	**TRA**[a]	**PA**[b]	**SN**[c]	**SP**[d]	**CVC**[e]	χ^{2}[f]	**Sign**[g]
MDR (N=236)	rs281432	0.5476	0.5297	0.7881	0.2712	10	0.74	NA
	rs281432, rs5030340	0.5730	0.5593	0.8136	0.3051	9	0.50	NA
	rs281432, rs5030340, rs281438	0.5880	0.5636	0.8475	0.2797	7	0.45	NA
	rs281432, rs5030340, rs281438, rs1799969	0.6088	0.5254	0.5763	0.4746	10	0.80	NA
GMDR (N=963)	rs281432	0.5353	0.5219	0.7848	0.2589	10	0.38	0.17
	rs281432, rs281438	0.5511	0.5292	0.6453	0.4130	7	0.48	0.17
	rs281432, rs281438, rs1799969	0.5621	0.5369	0.6327	0.4410	7	0.52	0.17
	rs281432, rs281438, rs1799969, rs5030340	0.5693	0.5165	0.5903	0.4426	10	0.58	0.17

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity,
[e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 118 cases and 845 controls to determine the best epistasis model for ICAM-1. The model shown in Table 5 included main effects for rs1799969, rs5030340, and rs281432 and an interaction between rs1799969 and rs281432. None of the Z-tests for the main effects were significant (all $p > 0.10$). The Z-test for rs1799969*rs281432 also suggested that this interaction was not significantly associated with 90-day mortality ($p = 0.61$). The standard errors were large compared to the magnitude of the estimates which signified a poorly fit model. The prediction accuracy was high at 0.8775. This estimate, however, was biased because this stepPLR model predicted all subjects were controls.

Table 5: ICAM-1 SNP epistasis model generated by stepPLR

Best Model (N=963)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
rs1799969	-0.0777 (0.1242)	0.53	740.66	0.8775
rs5030340	-0.0556 (0.1215)	0.65		
rs281432	0.0527 (0.0565)	0.35		
rs1799969*rs281438	-0.0122 (0.0241)	0.61		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrated that ICAM-1 SNPs rs281432 and rs281438 exhibited relatively strong correlation as indicated by the medium-sized, blue line in Figure 2. The dendrogram also displayed a relatively strong correlation between rs1799969 and rs5030340 as shown by a second medium-sized, blue line in the figure. The long, tan line showed independence between the two clusters of SNPs.

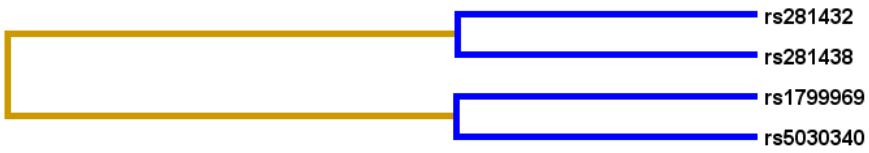

Figure 2: MDR interaction dendrogram for ICAM-1 SNPs. The medium-sized, blue lines indicate relatively strong correlation between rs281432 and rs281438 and between rs1799969 and rs5030340. The long, tan line indicates independence between the two clusters of SNPs.

4.3 IL-1A RESULTS

The MDR method utilized 104 cases and 104 controls in the data set for the IL-1A SNPs, and the results are shown in Table 6. The overall best model was the one-locus model with rs3783546 because it had the highest prediction accuracy (0.5481) and CVC (10) among all four models. The specificity was low at 0.5096, and the sensitivity was relatively low at 0.5865. The chi-square test concluded that rs3783546 was not significantly associated with 90-day mortality ($p = 0.66$).

For the GMDR method, 104 cases and 753 controls were used to generate the SNP epistasis models for IL-1A. The results are shown in Table 6. The one-locus model with rs3783546 was selected as the overall best model because it had the highest prediction accuracy (0.5579) and CVC. The sensitivity was relatively low at 0.6164, and the specificity was even lower at 0.4994. The CVC was very high at ten. The chi-square test, however, concluded that rs3783546 was not significantly associated with 90-day mortality ($p = 0.36$), but the sign test was borderline significant ($p = 0.05$). This was the first overall best model with a prediction accuracy larger than 0.55 and could be investigated more thoroughly in future research. Here, GMDR performed better than MDR.

Table 6: IL-1A SNP epistasis models generated by MDR and GMDR

Method	**Best Models**	**TRA**[a]	**PA**[b]	**SN**[c]	**SP**[d]	**CVC**[e]	χ^{2}[f]	**Sign**[g]
MDR (N=208)	rs3783546	0.5641	0.5481	0.5865	0.5096	8	0.66	NA
	IL1A_M889, rs2856838	0.5721	0.5240	0.5000	0.5481	6	0.83	NA
	IL1A_M889, rs2856838, rs1878319	0.5721	0.5240	0.5000	0.5481	5	0.83	NA
	IL1A_M889, rs2856838, rs1878319, rs1894399	0.5721	0.5240	0.5000	0.5481	5	0.83	NA
GMDR (N=857)	rs3783546	0.5574	0.5579	0.6164	0.4994	9	0.36	0.05
	rs3783546, rs3783520	0.5671	0.5540	0.4664	0.6416	8	0.33	0.17
	rs3783520, IL1A_M889, rs2856838	0.5684	0.5547	0.4664	0.6430	9	0.33	0.17
	rs3783520, IL1A_M889, rs2856838, rs1878319	0.5684	0.5547	0.4664	0.6430	9	0.33	0.17

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity, [e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 104 cases and 753 controls to determine the best epistasis model for IL-1A. The model shown in Table 7 included six main effects and two interactions. The results of the Z-tests indicated that IL1A_M889, rs1894399, rs1878319, and rs2856837 were each significant predictors of 90-day mortality when adjusting for the other SNPs in the model (all $p < 0.05$). In addition, the Z-test for the interaction between rs28556838 and rs3783520 was borderline significant ($p = 0.06$). The SNPs rs2856836 and rs3783520 and the interaction rs2856838*rs1894399 were not significantly associated with 90-day mortality (all $p \geq 0.09$). The standard errors were fairly small with respect to the magnitude of the parameter estimates. The prediction accuracy was high at 0.8787, but this estimate was biased because all subjects were predicted as controls.

Table 7: IL-1A SNP epistasis model generated by stepPLR

Best Model (N=857)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
IL1A_M889	-0.1142 (0.0512)	0.03	646.66	0.8787
rs1894399	0.1071 (0.0483)	0.03		
rs1878319	-0.1071 (0.0483)	0.03		
rs2856837	-0.1005 (0.0484)	0.04		
rs2856836	0.0939 (0.0553)	0.09		
rs3783520	0.0461 (0.0565)	0.42		
rs2856838*rs3783520	-0.1292 (0.0696)	0.06		
rs2856838*rs1894399	0.0627 (0.0689)	0.36		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrated that IL-1A SNPs rs1894399, IL1A_M889, and rs1878319 exhibited strong correlation as indicated by the short, blue lines in Figure 3. The dendrogram also displayed a strong correlation between rs3783546 and this cluster of three SNPs as shown by the other short, blue line in the figure. The long, green line showed weak correlation between rs2856838 and this cluster of four SNPs.

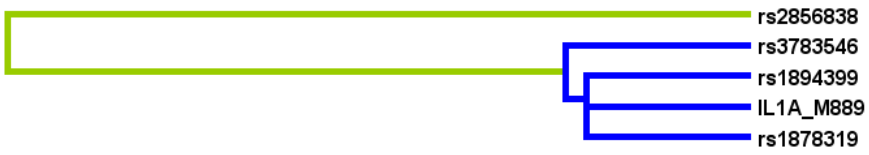

Figure 3: MDR interaction dendrogram for IL-1A SNPs. The short, blue lines indicate strong correlation between rs1894399, IL1A_M889, and rs1878319 and between rs3783546 and this cluster of three SNPs. The long, green line indicates weak correlation between rs2856838 and this cluster of four SNPs.

4.4 IL-1B RESULTS

The MDR method utilized 111 cases and 111 controls in the data set for the IL-6 SNPs, and the results are shown in Table 8. Even though the three-locus model had the highest CVC, the two-locus model with an interaction between IL1B_M511 and rs3917365 was selected as the overall best model because it had the highest prediction accuracy and was the more parsimonious model. The prediction accuracy, however, was still low at 0.5090. The sensitivity was poor at 0.5856, and the specificity was very low at 0.4324. The CVC was also low at five. The chi-square test concluded that IL1B_M511*rs3917365 was not significantly associated with 90-day mortality ($p = 0.93$).

For the GMDR method, 111 cases and 839 controls were used to generate the models for the IL-1B SNPs. The results are shown in Table 8. The two-locus model with an interaction between rs1071676 and rs1143623 exhibited the highest prediction accuracy (0.5237) and CVC, so it was selected as the overall best model. This was a poor model because the sensitivity was low (0.5598) and the specificity was even lower at 0.4875. The CVC, however, was very high at ten. The chi-square test concluded that the interaction between rs1071676 and rs1143623 was not significantly associated with 90-day mortality ($p = 0.49$), and the sign test was also non-significant ($p = 0.38$). Altogether, GMDR had better prediction accuracy than MDR.

Table 8: IL-1B SNP epistasis models generated by MDR and GMDR

Method	Best Models	TRA[a]	PA[b]	SN[c]	SP[d]	CVC[e]	χ^{2}[f]	Sign[g]
MDR (N=222)	IL1B_M511	0.5335	0.4820	0.5315	0.4324	6	0.86	NA
	IL1B_M511, rs3917365	0.5480	0.5090	0.5856	0.4324	5	0.93	NA
	IL1B_M511, rs3917365, IL1B_3957	0.5636	0.4550	0.4955	0.4144	8	0.67	NA
	IL1B_M511, rs3917365, IL1B_3957, rs1143623	0.5676	0.4550	0.4955	0.4144	7	0.67	NA
GMDR (N=950)	rs1071676	0.5140	0.4858	0.4424	0.5293	5	0.59	0.95
	rs1071676, rs1143623	0.5522	0.5237	0.5598	0.4875	10	0.49	0.38
	rs1071676, rs1143623, IL1B_M511	0.5622	0.5174	0.5409	0.4940	7	0.59	0.38
	rs1071676, rs1143623, IL1B_M511, rs3917365	0.5739	0.5169	0.5674	0.4664	10	0.53	0.17

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity, [e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 111 cases and 839 controls to determine the best epistasis model for IL-1B. The model shown in Table 9 included main effects for IL1B_3957, rs3917365, rs1071676, and IL1B_M511 and an interaction between IL1B_3957 and rs1143623. The Z-tests implied that none of the main effects or the interaction term were significantly associated with 90-day mortality (all $p > 0.10$). The standard errors were large compared to the magnitude of the parameter estimates which signified a poorly fit model. The prediction accuracy was high at 0.8832, but this estimate was biased because this stepPLR model predicted all subjects were controls.

Table 9: IL-1B SNP epistasis model generated by stepPLR

Best Model (N=950)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
IL1B_3957	0.0624 (0.1143)	0.59	713.48	0.8832
rs3917365	0.0037 (0.1217)	0.98		
rs1071676	0.0636 (0.0609)	0.30		
IL1B_M511	0.0252 (0.1172)	0.83		
IL1B_3957*rs1143623	0.0280 (0.0311)	0.37		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrated that the IL-1B SNPs IL1B_3957 and rs3917365 exhibited strong correlation as indicated by the short, blue line in Figure 4. The dendrogram also displayed a relatively strong correlation between IL1B_M511 and rs1143623 as shown by the medium-sized, green line in the figure. The long, green line illustrated weak correlation between these two clusters of SNPs.

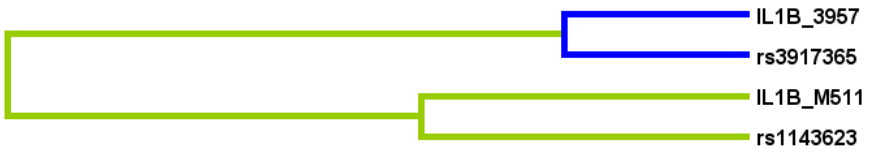

Figure 4: MDR interaction dendrogram for IL-1B SNPs. The short, blue line indicates strong correlation between IL1B_3957 and rs3917365. The medium-sized, green line indicates relatively strong correlation between IL1B_M511 and rs1143623. The long, green line indicates weak correlation between these two clusters of SNPs.

4.5 IL-6 RESULTS

The MDR method utilized 120 cases and 120 controls in the data set for the IL-1B SNPs, and the results are shown in Table 10. The overall best model was the two-locus model with an interaction between rs2069827 and rs1548216 because it had the highest prediction accuracy (0.5375) and the largest CVC. The sensitivity was high at 0.8333, but the specificity was very low (0.2417). The CVC of nine was very high. The chi-square test concluded that the interaction between rs2069827 and rs1548216 was not significantly associated with 90-day mortality ($p = 0.65$).

For the GMDR method, 120 cases and 833 controls were used to generate the models for the IL-6 SNPs. The results are shown in Table 10. The one-locus model with IL6_M174 was selected as the overall best model because it yielded the highest prediction accuracy (0.5341) and the largest CVC. This was also a poor model because a low sensitivity was observed (0.5667); the specificity was even lower at 0.5016. The CVC, however, was high at nine. The chi-square test concluded that IL6_M174 was not significantly associated with 90-day mortality ($p = 0.35$), and the sign test was also non-significant ($p = 0.62$). Hence, MDR had slightly better prediction accuracy than GMDR.

Table 10: IL-6 SNP epistasis models generated by MDR and GMDR

Method	**Best Models**	**TRA**[a]	**PA**[b]	**SN**[c]	**SP**[d]	**CVC**[e]	χ^{2}[f]	**Sign**[g]
MDR (N=240)	rs2069827	0.5250	0.5250	0.8583	0.1917	5	0.74	NA
	rs2069827, rs1548216	0.5417	0.5375	0.8333	0.2417	9	0.65	NA
	rs2069827, rs1548216, IL6_M174	0.5426	0.4875	0.7667	0.2083	8	0.88	NA
	rs2069827, rs1548216, IL6_M174, rs1800795	0.5463	0.5000	0.8000	0.2000	7	0.99	NA
GMDR (N=953)	IL6_M174	0.5342	0.5341	0.5667	0.5016	9	0.35	0.62
	IL6_M174, rs2069861	0.5391	0.5029	0.5667	0.4391	7	0.37	0.62
	IL6_M174, rs2069861, rs1548216	0.5436	0.4985	0.5417	0.4554	6	0.41	0.62
	IL6_M174, rs2069861, rs1548216, rs2069827	0.5473	0.4845	0.5250	0.4440	9	0.41	0.83

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity,
[e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 120 cases and 833 controls to determine the best epistasis model for IL-6. The model shown in Table 11 included four main effects and two interactions between SNPs. The results of the Z-tests indicated that IL6_M174, rs2069849, rs2069816, and rs1800795 were not significant predictors of 90-day mortality (all $p > 0.10$). Furthermore, the interactions rs1800795*IL6_M174 and rs1800795*IL6_M174*rs1548216 were not significant (all $p > 0.10$). The standard errors were relatively large compared to the magnitude of the parameter estimates and signified a poor model. The prediction accuracy was high at 0.8740. This estimate, however, was biased because this stepPLR model predicted all subjects were controls.

Table 11: IL-6 SNP epistasis model generated by stepPLR

Best Model (N=953)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
IL6_M174	0.0069 (0.0879)	0.94	744.47	0.8740
rs2069849	0.0220 (0.0669)	0.74		
rs2069861	0.1285 (0.1228)	0.30		
rs1800795*IL6_M174	0.0320 (0.0333)	0.34		
rs1800795	-0.0191 (0.0848)	0.82		
rs1800795*IL6_M174*rs1548216	-0.0198 (0.0168)	0.24		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrated that IL-6 SNPs rs1548216 and rs2069827 exhibited strong correlation as indicated by the short, blue line in Figure 5. The dendrogram also displayed a relatively strong correlation between IL6_M174 and this cluster of two SNPs depicted by the medium-sized, blue line in the figure. The long, green line showed weak correlation between rs1800795 and this cluster of three SNPs.

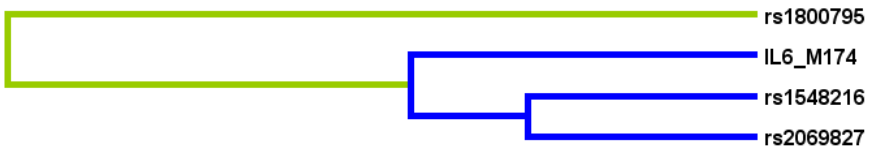

Figure 5: MDR interaction dendrogram for IL-6 SNPs. The short, blue line indicates strong correlation between rs1548216 and rs2069827. The medium-sized, blue line indicates relatively strong correlation between IL6_M174 and this cluster of two SNPs. The long, green line indicates weak correlation between rs1800795 and this cluster of three SNPs.

4.6 IL-10 RESULTS

The MDR method utilized 121 cases and 121 controls in the data set for the IL-10 SNPs, and the results are shown in Table 12. The overall best model was the two-locus model with an interaction between IL10_M819 and IL10_M1082 because it had the highest prediction accuracy and CVC. The prediction accuracy, however, was relatively low at 0.5248. The sensitivity was fairly high at 0.6694, but the specificity was very poor (0.3802). The CVC was very high at ten. The chi-square test concluded that the interaction between IL10_M819 and IL10_M1082 was not significantly associated with 90-day mortality ($p = 0.80$).

For the GMDR method, 121 cases and 880 controls were used to generate the SNP epistasis models for IL-10. The results are shown in Table 12. GMDR selected the one-locus model with IL10_M819 as the overall best model because it had the largest CVC and was more parsimonious than the three-locus model which displayed the highest prediction accuracy (0.5610). The prediction accuracy for IL10_M819 was relatively low at 0.5524. The sensitivity (0.4718) was low while the specificity was fairly high at 0.6330. The CVC was relatively high at eight. The chi-square test, however, concluded that IL10_M819 was not significantly associated with 90-day mortality ($p = 0.46$), and the sign test was not significant ($p = 0.17$). This was the second overall best model with a prediction accuracy larger than 0.55. Here, GMDR performed better than MDR.

Table 12: IL-10 SNP epistasis models generated by MDR and GMDR

Method	**Best Models**	**TRA**[a]	**PA**[b]	**SN**[c]	**SP**[d]	**CVC**[e]	χ^{2}[f]	**Sign**[g]
MDR (N=242)	IL10_M819	0.5459	0.5165	0.5537	0.4793	7	0.87	NA
	IL10_M819, IL10_M1082	0.5771	0.5248	0.6694	0.3802	10	0.80	NA
	IL10_M819, IL10_M1082, rs3024505	0.5886	0.5248	0.6446	0.4050	3	0.80	NA
	IL10_M819, IL10_M1082, rs3024505, rs1800894	0.5992	0.5083	0.6364	0.3802	4	0.93	NA
GMDR (N=1001)	IL10_M819	0.5520	0.5524	0.4718	0.6330	8	0.46	0.17
	IL10_M819, rs3024496	0.5712	0.5580	0.7429	0.3731	7	0.31	0.17
	IL10_M819, rs1800894, rs3024505	0.5807	0.5610	0.4788	0.6432	5	0.42	0.01
	IL10_M819, rs1800894, rs3024505, rs3024496	0.5894	0.5165	0.4712	0.5619	5	0.57	0.17

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity,
[e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 121 cases and 880 controls to determine the best epistasis model for IL-10. The model shown in Table 13 included four main effects and two second-order interactions. The results from the Z-tests indicated that IL10_M1082, rs3024496, rs1800894, and rs1800872 were not significantly associated with 90-day mortality (all $p > 0.10$). On the other hand, the interaction rs3024496*rs1800872 was a significant predictor of 90-day mortality when adjusting for the other SNPs in the model ($p = 0.02$). The interaction between IL10_M1082 and rs1800872 was borderline significant ($p = 0.07$). Overall, the standard errors were fairly small with respect to the magnitude of the parameter estimates. The prediction accuracy was high at 0.8792, but this estimate was biased because all subjects were predicted as controls.

Table 13: IL-10 SNP epistasis model generated by stepPLR

Best Model (N=1001)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
IL10_M1082	-0.1187 (0.0748)	0.11	755.52	0.8792
IL10_M1082*rs1800872	-0.1430 (0.0779)	0.07		
rs3024496	-0.1204 (0.0745)	0.11		
rs3024496*rs1800872	0.1887 (0.0777)	0.02		
rs1800894	-0.0351 (0.1124)	0.76		
rs1800872	-0.1192 (0.1081)	0.27		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrated that the IL-10 SNPs IL10_M819 and rs3024505 exhibited strong correlation as indicated by the short, blue line in Figure 6. The dendrogram also displayed a relatively strong correlation between rs1800894 and this cluster of two SNPs as shown by the medium-sized, green line in the figure. The long, tan line highlighted the independence between IL10_M1082 and this cluster of three SNPs.

Figure 6: MDR interaction dendrogram for IL-10 SNPs. The short, blue line indicates strong correlation between IL10_M819 and rs3024505. The medium-sized, green line indicates relatively strong correlation between rs1800894 and this cluster of two SNPs. The long, tan line indicates independence between IL10_M1082 and this cluster of three SNPs.

4.7 MBL RESULTS

The MDR method utilized 108 cases and 108 controls in the data set for the MBL SNPs, and the results are shown in Table 14. Even though the two-locus model had the highest prediction accuracy 0.5556, the overall best model was the one-locus model with rs1800450 because it had a CVC of eight and was the more parsimonious model. The prediction accuracy, however, was low at 0.5324. This was a very poor model since the specificity was only 0.2593. The sensitivity was relatively high at 0.8056. The chi-square test concluded rs1800450 was not significantly associated with 90-day mortality ($p = 0.72$).

For the GMDR method, 108 cases and 776 controls were used to generate the models for the MBL SNPs. The results are shown in Table 14. The three-locus model exhibited the highest prediction accuracy (0.5429) and was chosen as the overall best model. This model was more parsimonious than the four-locus model with the largest CVC. A very low sensitivity was observed (0.4427), but the specificity was fairly high at 0.6431. The CVC was low at six. The chi-square test concluded that this model was not significantly associated with 90-day mortality ($p = 0.45$), and the sign test was non-significant ($p = 0.17$). Altogether, GMDR had slightly better prediction accuracy than MDR.

Table 14: MBL SNP epistasis models generated by MDR and GMDR

Method	**Best Models**	**TRA**[a]	**PA**[b]	**SN**[c]	**SP**[d]	**CVC**[e]	χ^{2}[f]	**Sign**[g]
MDR (N=216)	rs1800450	0.5324	0.5324	0.8056	0.2593	8	0.72	NA
	rs11003125, rs5030737	0.5556	0.5556	0.2685	0.8426	5	0.53	NA
	rs11003125, rs5030737, rs1800450	0.5792	0.5324	0.6204	0.4444	8	0.76	NA
	rs11003125, rs5030737, rs1800450, rs7096206	0.5921	0.4583	0.5556	0.3611	6	0.69	NA
GMDR (N=884)	rs1800450	0.5298	0.5258	0.7964	0.2551	7	0.48	0.17
	rs11003125, rs5030737	0.5452	0.5139	0.4555	0.5724	5	0.53	0.38
	rs11003125, rs5030737, rs7096206	0.5658	0.5429	0.4427	0.6431	6	0.45	0.17
	rs11003125, rs5030737, rs7096206, rs1800450	0.5821	0.5010	0.4897	0.5124	10	0.45	0.38

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity, [e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 108 cases and 776 controls to determine the best epistasis model for MBL. The model shown in Table 15 included four main effects and one second-order interaction. The results of the Z-tests indicated that none of the main effects (rs11003125, rs1800451, rs5030737, and rs1800450) were significantly associated with 90-day mortality (all $p > 0.10$). The second-order interaction between rs1800450 and rs7096206 was non-significant ($p = 0.27$). The standard errors were large compared to the magnitude of the parameter estimates and signified a poor model. The prediction accuracy was high at 0.8778. This estimate, however, was biased because the stepPLR model predicted all subjects were controls.

Table 15: MBL SNP epistasis model generated by stepPLR

Best Model (N=884)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
rs11003125	-0.0454 (0.0595)	0.45	681.33	0.8778
rs1800451	-0.0219 (0.0917)	0.81		
rs5030737	0.0890 (0.1238)	0.47		
rs1800450	0.0499 (0.1241)	0.69		
rs1800450*rs7096206	0.0295 (0.0269)	0.27		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrates that the MBL SNPs rs11003125 and rs5030737 exhibited strong epistasis as indicated by the short, red line in Figure 7. The long, tan line illustrated independence between rs1800450 and this cluster of two SNPs. In addition, the longest, tan line showed indpendence between rs7096206 and this clusters of three SNPs.

Figure 7: MDR interaction dendrogram for MBL SNPs. The short, red line indicates strong epistasis between rs11003125 and rs5030737. The long, tan line indicates independence between rs1800450 and this cluster of two SNPs. The longest, tan line indicates independence between rs7096206 and this cluster of three SNPs.

4.8 PAI-1 RESULTS

The MDR method utilized 132 cases and 132 controls in the data set for the PAI-1 SNPs, and the results are shown in Table 16. The overall best model was the one-locus model with rs2227683 because it had the same prediction accuracy as the two-locus model (0.5530) but was more parsimonious. This was the third overall best model with a prediction accuracy larger than 0.55 and could be investigated more thoroughly in future research. The sensitivity was relatively high at 0.7652, but the specificity was low (0.3409). The CVC was high at nine. The chi-square test for rs2227683 was non-significant ($p = 0.55$).

For the GMDR method, 132 cases and 949 controls were used to generate the models for the PAI-1 SNPs. The results are shown in Table 16. The one-locus model with rs2227683 was selected as the overall best model because it yielded the highest prediction accuracy (0.5476) and the largest CVC (10). A fairly high sensitivity was observed (0.7264); the specificity, however, was very low at 0.3688. The chi-square test concluded that rs2227683 was not significantly associated with 90-day mortality ($p = 0.40$), but the sign test was borderline significant ($p = 0.05$). Thus, MDR had slightly better prediction accuracy than GMDR.

Table 16: PAI-1 SNP epistasis models generated by MDR and GMDR

Method	**Best Models**	**TRA**[a]	**PA**[b]	**SN**[c]	**SP**[d]	**CVC**[e]	χ^{2}[f]	**Sign**[g]
MDR (N=264)	rs2227683	0.5530	0.5530	0.7652	0.3409	9	0.55	NA
	rs2227683, rs13238709	0.5530	0.5530	0.7652	0.3409	9	0.55	NA
	rs2227683, rs2227665, rs7242	0.5669	0.4962	0.4394	0.5530	6	0.97	NA
	rs2227683, rs2227665, rs7242, rs13238709	0.5673	0.4924	0.4394	0.5455	10	0.94	NA
GMDR (N=1081)	rs2227683	0.5480	0.5476	0.7264	0.3688	10	0.40	0.05
	rs2227683, rs13238709	0.5480	0.5442	0.7049	0.3835	8	0.42	0.05
	rs2227683, rs2227665, rs7242	0.5561	0.5263	0.4297	0.6228	6	0.54	0.17
	rs2227683, rs2227665, rs7242, rs13238709	0.5603	0.5216	0.4418	0.6014	10	0.54	0.17

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity,
[e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 132 cases and 949 controls to determine the best epistasis model for PAI-1. The model shown in Table 17 included a main effect and a second-order interaction. The Z-test indicated that rs2227665 was associated with 90-day mortality (p = 0.03). The interaction rs2227665*rs2227683 was even more statistically signficant (p = 0.002). The standard errors were relatively small compared to the magnitude of the parameter estimates. The prediction accuracy was high at 0.8779, but this estimate was biased because the stepPLR model predicted all subjects were controls.

Table 17: PAI-1 SNP epistasis model generated by stepPLR

Best Model (N=1081)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
rs2227665	0.1337 (0.0615)	0.03	806.81	0.8779
rs2227665*rs2227683	0.3341 (0.1095)	0.002		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrated that the PAI-1 SNPs rs2227665 and rs2227683 exhibited strong correlation as indicated by the short, blue line in Figure 8. The dendrogram also displayed a relatively strong correlation between rs7242 and this cluster of two SNPs as shown by the medium-sized, green line in the figure. The long, green line showed weak correlation between rs13238709 and this cluster of three SNPs.

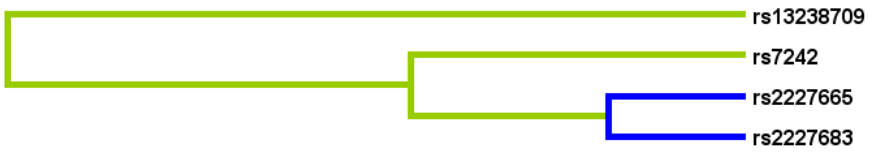

Figure 8: MDR interaction dendrogram for PAI-1 SNPs. The short, blue line indicates strong correlation between rs2227665 and rs2227683. The medium-sized, green line indicates relatively strong correlation between rs7242 and this cluster of two SNPs. The long, green line indicates weak correlation between rs13238709 and this cluster of three SNPs.

4.9 TLR-4 RESULTS

The MDR method utilized 107 cases and 107 controls in the data set for the TLR-4 SNPs, and the results are shown in Table 18. Even though the three-locus model had the highest prediction accuracy 0.5981, the one-locus model with rs2770146 was selected as the overall best model because it had a CVC of ten and was more parsimonious. The prediction accuracy was fairly low at 0.5047. The specificity of the model was relatively high at 0.7009. This was a very poor model because the sensitivity was very low (0.3084) and the chi-square test was non-significant ($p = 0.96$).

For the GMDR method, 107 cases and 771 controls were used to generate the models for the TLR-4 SNPs. The results are shown in Table 18. The prediction accuracy was highest for the three-locus model, but the one-locus model had a CVC of ten and was more parsimonious. Thus, the one-locus model with rs2770146 was selected as the overall best model. A fairly low prediction accuracy was observed (0.5209). The sensitivity was very low at 0.1782, but the specificity of the model was much higher (0.8637). The chi-square test concluded that rs2770146 was not significantly associated with 90-day mortality ($p = 0.47$), and the sign test was also non-significant ($p = 0.17$). As a result, GMDR had slightly better prediction accuracy than MDR.

Table 18: TLR-4 SNP epistasis models generated by MDR and GMDR

Method	**Best Models**	**TRA**[a]	**PA**[b]	**SN**[c]	**SP**[d]	**CVC**[e]	χ^{2}[f]	**Sign**[g]
MDR (N=214)	rs2770146	0.5566	0.5047	0.3084	0.7009	10	0.96	NA
	rs2770146, rs11536897	0.5898	0.5748	0.2617	0.8879	4	0.38	NA
	rs2770146, rs11536897, rs1927914	0.6173	0.5981	0.3458	0.8505	9	0.29	NA
	rs2770146, rs1927914, rs11536898, rs5030717	0.6225	0.5841	0.4112	0.7570	3	0.41	NA
GMDR (N=878)	rs2770146	0.5338	0.5209	0.1782	0.8637	10	0.47	0.17
	rs2770146, rs11536897	0.5561	0.5412	0.2591	0.8232	7	0.36	0.17
	rs2770146, rs5030717, rs11536898	0.5772	0.5674	0.3618	0.7730	7	0.33	0.05
	rs2770146, rs5030717, rs1927914, rs11536897	0.5910	0.5035	0.4264	0.5807	7	0.62	0.38

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity, [e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 107 cases and 771 controls to determine the best epistasis model for TLR-4. The model shown in Table 19 only included a main effect for TLR4_M896 and no interactions. No other SNPs on the TLR-4 gene yielded valid estimates for the model. The Z-test indicated TLR4_M896 was not significantly associated with 90-day mortality ($p = 0.94$). The standard error was very large compared to the magnitude of the estimate which signified a poorly fit model. The prediction accuracy was high (0.8782), but this estimate was biased because this stepPLR model predicted all subjects were controls.

Table 19: TLR-4 SNP epistasis model generated by stepPLR

Best Model (N=878)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
TLR4_M896	-0.0085 (0.1200)	0.94	660.04	0.8782

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrated that TLR-4 SNPs rs2770146 and rs5030717 exhibited strong correlation as indicated by the short, blue line in Figure 9. The dendrogram also displayed strong correlation between rs11536898 and rs1927914 and between rs11536897 and this cluster of two SNPs as indicated by the short, green lines in the figure. The long, tan line depicted independence between these two clusters of SNPs.

Figure 9: MDR interaction dendrogram for TLR-4 SNPs. The short, blue line indicates strong correlation between rs2770146 and rs5030717. The short, green lines indicate strong correlation between rs11536898 and rs1927914 and between rs11536897 and this cluster of two SNPs. The long, tan line indicates independence between these two clusters of SNPs.

4.10 TNF RESULTS

The MDR method utilized 112 cases and 112 controls in the data set for the TNF SNPs, and the results are shown in Table 20. The overall best model was the two-locus model with an interaction between rs3093662 and TNFA_M308 because it had the highest prediction accuracy (0.5536) and the largest CVC. The sensitivity was relatively high at 0.6429, but the specificity was low (0.4643). The CVC was very high at ten. The chi-square test concluded rs3093662*TNFA_M308 was not significantly associated with 90-day mortality (p = 0.61). This was the fourth, overall best model with a prediction accuracy larger than 0.55 and could be investigated more thoroughly in future research.

For the GMDR method, 112 cases and 787 controls were used to generate the models for the TNF SNPs. The results are shown in Table 20. The one-locus model with TNFA_M308 was selected as the overall best model. It yielded the highest prediction accuracy (0.5126) and was more parsimonious than the four-locus model with the largest CVC. The sensitivity was fairly high at 0.7697, but this was a poor model because the specificity was so low (0.2554). The CVC was relatively high at seven. The chi-square test concluded that TNFA_M308 was not significantly associated with 90-day mortality (p = 0.35), and the sign test was also non-significant (p = 0.17). Altogether, MDR had slightly better prediction accuracy than GMDR.

Table 20: TNF SNP epistasis models generated by MDR and GMDR

Method	Best Models	TRA[a]	PA[b]	SN[c]	SP[d]	CVC[e]	χ^{2}[f]	Sign[g]
MDR (N=224)	rs3093662	0.5536	0.5536	0.8482	0.2589	8	0.53	NA
	rs3093662, TNFA_M308	0.5818	0.5536	0.6429	0.4643	10	0.61	NA
	rs3093662, TNFA_M308, TNF_857	0.5833	0.5223	0.6250	0.4196	6	0.83	NA
	rs3093662, TNFA_M308, TNF_857, TNF_238	0.5928	0.5268	0.6518	0.4018	9	0.79	NA
GMDR (N=899)	TNFA_M308	0.5190	0.5126	0.7697	0.2554	7	0.35	0.17
	TNFA_M308, TNF_857	0.5296	0.4845	0.2508	0.7183	6	0.56	0.83
	TNFA_M308, TNF_857, rs3093662	0.5352	0.4458	0.2667	0.6250	3	0.48	0.99
	TNFA_M308, TNF_857, rs3093662, TNF_238	0.5415	0.4633	0.2924	0.6342	8	0.50	0.95

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity,
[e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 112 cases and 787 controls to determine the best epistasis model for TNF. The model shown in Table 21 included two main effects and three interactions. The results of the Z-tests indicated that TNF_857 and rs3093672 were not significantly associated with 90-day mortality (all $p > 0.10$). Similarly, the second-order interactions TNF_857*rs3093662 and TNF_857*TNF_238 were not significant predictors of 90-day mortality (all $p > 0.10$). The Z-test for the third-order interaction TNF_857*rs3093662*TNF_238 was non-significant as well ($p = 0.79$). Overall, the standard errors were very large compared to the magnitude of the parameter estimates, so the model itself was poor. The prediction accuracy was high at 0.8754. This estimate, however, was biased because the stepPLR model predicted all subjects were controls.

Table 21: TNF SNP epistasis model generated by stepPLR

Best Model (N=899)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
TNF_857	0.0285 (0.0465)	0.54	691.06	0.8754
rs3093672	-0.0308 (0.0806)	0.70		
TNF_857*rs3093662	0.0160 (0.0676)	0.81		
TNF_857*rs3093662*TNF_238	0.0247 (0.0931)	0.79		
TNF_857*TNF_238	0.0496 (0.0782)	0.53		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy

The interaction dendrogram produced by MDR illustrated that TNF SNPs rs3093662 and TNF_238 exhibited strong correlation as indicated by the short, blue line in Figure 10. The dendrogram displayed independence between TNF_857 and TNFA_M308 as indicated by the medium-sized, tan line in the figure. The long, tan line depicted independence between these two clusters of SNPs.

Figure 10: MDR interaction dendrogram for TNF SNPs. The short, blue line indicates strong correlation between rs3093662 and TNF_238. The medium-sized, tan line indicates independence between TNF_857 and TNFA_M308. The long, tan line indicates independence between these two clusters of SNPs.

4.11 RESULTS FOR ALL SNPs COMBINED

The MDR method utilized 61 cases and 61 controls in the data set for fifty-five SNPs across all genes, and the results are shown in Table 22. The overall best model was the three-locus model with rs2770146, IL10_M819, and IL6_M174. It had the highest prediction accuracy (0.6803) and was more parsimonious than the nine-locus model with the largest CVC. This was the first overall best model with a prediction accuracy larger than 0.60 and could be investigated more thoroughly in future research. The sensitivity was fairly low at 0.6230, but the specificity was a little better (0.7377). The CVC was surprisingly low at three. The chi-square test concluded rs2770146*IL10_M819*IL6_M174 was not significantly associated with 90-day mortality ($p = 0.20$).

Table 22: All SNPs combined epistasis models generated by MDR

Best Models (N=122)	**TRA[a]**	**PA[b]**	**SN[c]**	**SP[d]**	**CVC[e]**	χ^{2}[f]
rs2770146	0.6066	0.6066	0.6230	0.5902	8	0.46
rs3024505, rs7242	0.6667	0.6311	0.6557	0.6066	3	0.36
rs2770146, IL10_M819, IL6_M174	0.7377	0.6803	0.6230	0.7377	3	0.20
rs2770146, IL10_M819, IL6_M174, rs2856838	0.8224	0.6393	0.7541	0.5246	3	0.32
rs2770146, IL10_M819, IL6_M174, rs2856838, IL1B_M511	0.9035	0.6721	0.8525	0.4918	6	0.20
IL6_M174, rs11003125, IL1B_3957, rs13238709, rs3024505, rs3783546	0.9536	0.6230	0.9180	0.3279	2	0.29
rs2770146, IL10_M819, IL6_M174, IL1B_3957, IL1B_M511, rs1800450, rs13238709	0.9854	0.5574	0.9672	0.1475	2	0.48
rs2770146, IL10_M819, IL6_M174, IL1B_3957, IL1B_M511, rs1800450, rs13238709, rs11003125	0.9927	0.5082	0.9836	0.0328	3	0.85
IL10_M819, IL6_M174, IL1B_3957, IL1B_M511, rs1800450, rs13238709, rs11003125, rs1799969, TLR4_M896	0.9927	0.5410	0.9836	0.0984	10	0.54
IL10_M819, IL6_M174, IL1B_3957, IL1B_M511, rs1800450, rs13238709, rs11003125, rs1799969, TLR4_M896, rs1071676	0.9927	0.5410	0.9836	0.0984	10	0.54

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity,
[e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value

For the GMDR method, 61 cases and 397 controls were used to generate the epistasis models for fifty-five SNPs across all genes. The results are shown in Table 23. The six-locus model including IL6_M174, rs2856838, IL10_M1082, IL1B_M511, rs13238709, and rs2770146 was selected as the overall best model because it yielded the highest prediction accuracy (0.5935). It was also more parsimonious than the ten-locus model with the largest CVC. The sensitivity was very low at 0.3800, but the specificity was relatively high (0.8070). The CVC was fairly high at six. The chi-square test concluded that this model was not significantly associated with 90-day mortality ($p = 0.43$). The sign test was borderline significant ($p = 0.05$). A six-locus statistical epistasis model, however, may not be plausible for biological epistasis. Here, MDR had slightly better prediction accuracy than GMDR.

Table 23: All SNPs combined epistasis models generated by GMDR

Best Models (N=458)	**TRA[a]**	**PA[b]**	**SN[c]**	**SP[d]**	**CVC[e]**	χ^{2}[f]	**Sign[g]**
IL10_M819	0.5557	0.5527	0.4857	0.6197	4	0.37	0.62
IL6_M174, IL1B_M511	0.6084	0.5521	0.4786	0.6256	3	0.61	0.05
IL6_M174, rs7242, rs1143623	0.6779	0.5628	0.6905	0.4351	3	0.40	0.17
IL6_M174, rs7242, IL1B_M511, rs11003125	0.7589	0.5738	0.5267	0.6209	4	0.35	0.38
IL6_M174, rs7242, rs11003125, rs2856838, IL10_M1082	0.8567	0.5267	0.3238	0.7296	3	0.33	0.38
IL6_M174, rs2856838, IL10_M1082, IL1B_M511, rs13238709, rs2770146	0.9351	0.5935	0.3800	0.8070	6	0.43	0.05
IL6_M174, rs2856838, IL10_M1082, IL1B_M511, rs2770146, rs7242, rs11003125	0.9710	0.5368	0.1917	0.8819	6	NA	0.17
IL6_M174, rs2856838, IL10_M1082, IL1B_M511, rs2770146, rs11003125, rs13238709, IL10_M819	0.9850	NA	NA	0.8957	6	NA	0.01
IL6_M174, rs2856838, IL10_M1082, IL1B_M511, rs2770146, rs11003125, rs13238709, IL10_M819, IL1B_3957	0.9910	NA	NA	0.8903	8	NA	0.38
IL6_M174, rs2856838, IL10_M1082, IL1B_M511, rs2770146, rs11003125, rs13238709, IL10_M819, IL1B_3957, IL1A_M889	0.9910	NA	NA	0.8500	9	NA	0.62

[a] TRA = training accuracy, [b] PA = prediction accuracy, [c] SN = sensitivity, [d] SP = specificity,
[e] CVC = cross-validation consistency, [f] χ^2 = chi-square test p-value, [g] Sign = sign test p-value

Stepwise PLR utilized 61 cases and 397 controls to determine the best epistasis model for all fifty-five SNPs combined. The model shown in Table 24 included one main effect and nine interactions. The main effect rs2770146 was not significantly associated with 90-day mortality ($p = 0.31$). Similar results were obtained for the second- through seventh-order interactions because none of the interactions were significant predictors of 90-day mortality (all $p > 0.10$). Overall, the standard errors were large compared to the magnitude of the parameter estimates, so the model itself was poor. This was the first stepPLR model which did not predict all subjects were controls. The software output indicated that the prediction accuracy was 0.8624, but hand calculations yielded a prediction accuracy of 0.8755, a sensitivity of 0.0984, and a specificity of 0.9950. Clearly, there were some discrepancies with how the stepPLR software calculated prediction accuracy. Nevertheless, the model performed very poorly in predicting 90-day mortality since fifty-five out of the sixty-one cases were predicted to be alive after ninety days. This epistasis model overwhelmingly predicted that 450 of the 458 subjects were controls.

Table 24: All SNPs combined epistasis model generated by stepPLR

Best Model (N=458)	**Estimate (SE)[a]**	**Z-test[b]**	**Score**	**PA[c]**
rs2770146	0.1259 (0.1249)	0.31	354.09	0.8624
rs2770146*TNF_857	-0.0056 (0.0333)	0.87		
rs2770146*TNF_857*rs1548216	-0.0111 (0.0666)	0.87		
rs2770146*TNF_857*rs1548216*IL10_M1082	-0.0004 (0.0116)	0.97		
rs2770146*TNF_857*rs1548216*IL10_M1082*TNFA_M308	-0.0065 (0.0215)	0.76		
rs2770146*TNF_857*rs1548216*IL10_M1082*rs12344353	0.0111 (0.0201)	0.58		
rs2770146*TNF_857*rs1548216*IL10_M1082*TNFA_M308*rs11536898	-0.0194 (0.0644)	0.76		
rs2770146*TNF_857*rs1548216*IL10_M1082*rs12344353*rs1800451	0.0229 (0.0207)	0.27		
rs2770146*TNF_857*rs1548216*IL10_M1082*rs12344353*rs1800451 *rs2069849	0.0687 (0.0622)	0.27		
rs2770146*TNF_857*rs1548216*IL10_M1082*TNFA_M308*rs11536898 *rs3093672	-0.0749 (0.0743)	0.31		

[a] SE = standard error, [b] Z-test = Z-test p-value, [c] PA = prediction accuracy
Note: prediction accuracy = 0.8755, sensitivity = 0.0984, and specificity = 0.9950

MDR produced the interaction dendrogram for the SNP epistasis models across all genes. Figure 11 illustrated how various clusters of SNPs exhibited epistasis (red or orange lines), correlation (blue or green lines), or independence (tan lines). The length of the connecting lines implied the strength of the relationship; shorter lines represented stronger relationships. Overall, there were five independent clusters of SNPs. Most notably, IL10_M819 was independent of all other SNPs.

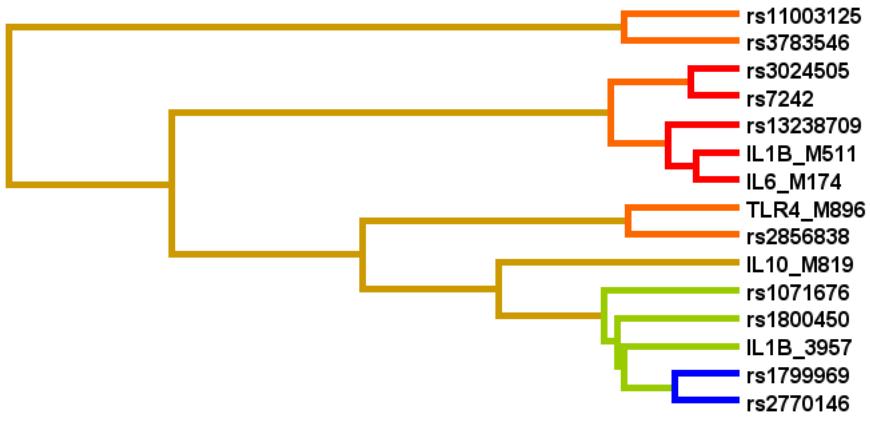

Figure 11: MDR interaction dendrogram for all SNPs combined. The length of the connecting lines implies the strength of the relationship; shorter lines represent stronger relationships. The color of the lines reveal the type of relationship as epistasis (red or orange), correlation (blue or green), or independence (tan).

MDR produced a contingency table for the SNP epistasis models across all genes. Again, the overall best epistasis model included rs2770146, IL10_M819, and IL6_M174, and the table is shown in Figure 12. The high-risk cells were shaded in dark gray; the low-risk cells were shaded in light gray. The blank, white cells were not represented in the sample data. There appeared to be a protective factor for rs2770146 genotype AA when in combination with IL10_M819 genotype CC. In addition, the figure implied an increased risk for rs2770146 genotype GG when in combination with IL10_M819 genotype CC. These SNPs are on genes TLR-4 and IL-10, respectively. One can see the problems with empty cells in higher dimensions for the MDR method, for no clear conclusions could be drawn about the genotypes for the IL6_M174 SNP.

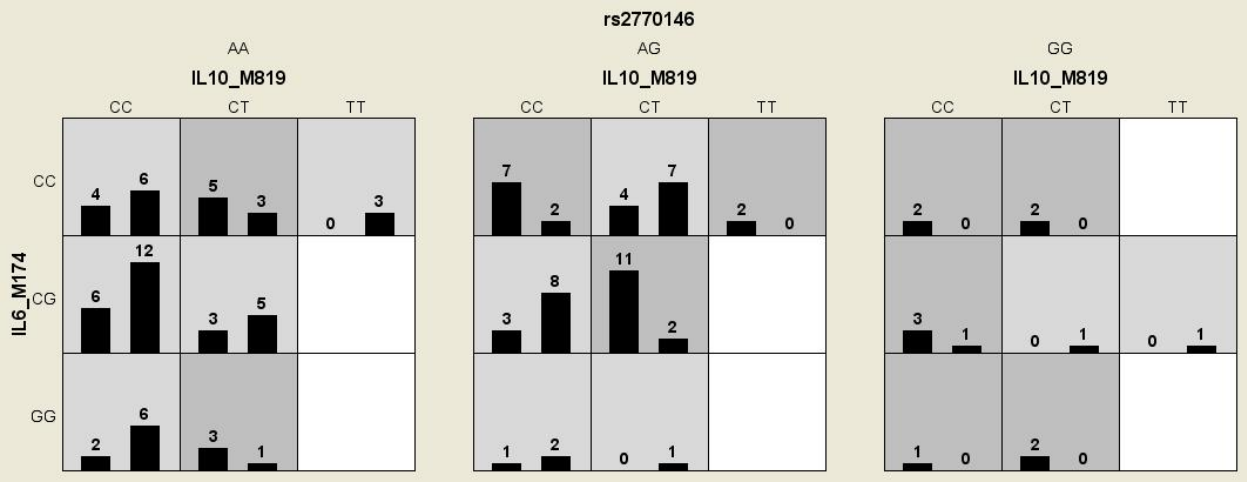

Figure 12: MDR contingency table for all SNPs combined. Dark gray cells are high-risk; light gray cells are low-risk; and white cells are empty in the sample data. There appears to be a protective factor for rs2770146 genotype AA when in combination with IL10_M819 genotype CC and an increased risk for rs2770146 genotype GG when in combination with IL10_M819 genotype CC. These SNPs are on genes TLR-4 and IL-10, respectively.

5.0 DISCUSSION AND CONCLUSIONS

5.1 DISCUSSION

The highest prediction accuracy was obtained from the MDR method for the three-locus model across all SNPs combined. The prediction accuracy for the interaction between rs2770146, IL10_M819, and IL6_M174 indicated that 68.03% of the subjects were correctly classified as cases or controls. This figure does not take into account other baseline characteristics such age, gender, smoking status, or severity of the illness. Moreover, the influence that each genotype of SNP rs2770146 has on 90-day mortality is dependent upon the genotypes of the other two SNPs IL10_M819 and IL6_M174, and so forth. From the contingency table in Figure 12, a subject with genotype CC for SNP IL10_M819 and AA for rs2770146 appeared to have a lower risk of 90-day mortality, but a substitution of genotype GG for rs2770146 implied a higher risk of death. The empty cells in the contingency table made interpretation difficult for the genotypes of IL6_M174 and highlighted the need for even larger sample sizes. Unfortunately, the chi-square test concluded that the third-order interaction itself was not a significant predictor of 90-day mortality, but the results may improve by adjusting for the baseline covariates or including the main effects or additional interactions. It is noteworthy, however, that the interaction dendrogram showed that these three SNPs were independent in the hierarchical cluster analysis.

The results of this analysis clearly showed that stepPLR had the poorest performance when compared to MDR and GMDR. The biggest problem was that stepPLR overwhelmingly predicted that all subjects would still be alive after ninety days, so all of the prediction accuracy estimates were biased. Using an equal number of cases and controls in future analysis may help to control this issue and produce more realistic prediction accuracies.

In this analysis, MDR and GMDR performed almost equally; GMDR produced the highest prediction accuracy for six of the genes. MDR had a much higher prediction accuracy (0.6803) than GMDR (0.5935) when all fifty-five SNPs were combined across all genes. In all, GMDR yielded higher prediction accuracies than MDR, and MDR performed better than stepPLR when identifying SNP interactions associated with 90-day mortality in a cohort of CAP patients. This assessment coincides with a study published by Hua He et al. [6], for the authors stated that MDR performs better when complex interactions are observed. Park and Hastie [4], who published the stepPLR method, claimed that their method yielded better classification of cases and controls than MDR, but this thesis was unable to reach the same conclusion because of the biased estimates obtained using stepPLR. In contrast, the results of this thesis agree with the findings made by Lou et al. [2]. These authors extended the MDR method into GMDR and determined that GMDR tended to produce higher prediction accuracies, higher cross-validation consistencies, and significant p-values (even though identical n-locus models were often obtained). If this analysis had taken into account additional covariates, the findings would have overwhelmingly favored GMDR.

Although many genome-wide association studies of second- and third-order SNP interactions have yielded inconsistent results, the conclusions of this analysis conferred with the results published in *Respiratory Medicine*. No significant associations were found between the SNPs on genes TNF or IL-6 and the CAP outcomes. Similarly, there were no significant interactions found between the TNF SNPs and any of the SNPs on the other nine genes. It was anticipated that TNF, IL-6, and IL-1A genes may be associated with higher 90-day mortality. Stepwise PLR determined that IL-1A SNPs IL1A_M889, rs1894399, rs1878319, and rs2856837 were each significant predictors of 90-day mortality when adjusting for the other SNPs in the model (all $p < 0.05$). In addition, the interaction rs28556838*rs3783520 was borderline significantly associated with 90-day mortality ($p = 0.06$). Interestingly, PAI-1 "supports thrombus formation and cardiovascular events by inhibiting fibrinolysis (thrombus breakdown) and by promoting endothelial dysfunction directly" [3]. Hence, these results complemented previous published studies regarding SNP associations with CAP outcomes.

The public health significance of this thesis involves many advancements related to risk for complex genetic diseases. Simply identifying SNPs associated with worse outcomes are

unlikely to be useful to improve performance of clinical risk prediction models, but the relative risk for CAP may be higher for a set of SNPs across different genes. One of the main goals in human genetics is to discover how DNA sequence variations are related to disease susceptibility, for this could lead to improvements in diagnosis, prevention, and treatment [1]. The ability to predict which patients will experience a poor outcome may lead to more effective prevention strategies or treatments at earlier stages. Identification of significant SNP interactions can also expand scientific knowledge about biological mechanisms affecting disease outcomes. Novel methods for determining risk for known, complex genetic diseases may more frequently involve determination of a patient's genotypes for a combination of particular SNPs. Advances in statistical methodology related to identification of high-order interactions among SNPs associated with disease will most likely play a role in reducing disease and disability for future generations.

Additional work on these methods and the GenIMS data set could reduce the limitations of this analysis. First, removing SNPs because of missing data was one of the biggest problems, for interactions with these SNPs could not be tested. Additional data or imputation could alleviate this issue. Second, this analysis did not take into account any non-genetic covariates since MDR cannot handle this additional data. Including these covariates in GMDR and stepPLR could greatly improve the models' prediction accuracies. Third, MDR and GMDR models did not include main effects nor multiple interactions between SNPs, so forcing these methods to include such covariates could also improve prediction accuracy. Fourth, this analysis did not control for different confounding variables known to affect CAP outcome. Since many subjects who have CAP also develop sepsis, it would be advantageous to adjust for sepsis in the epistasis models. Finally, due to the nature of the three methods, there was no way to validate the conclusions derived from the SNP epistasis models. It was difficult to compare these results to published studies because they often yield conflicting results for method performance and SNP associations with 90-day mortality or simply do not have enough statistical power to facilitate significant findings. Consequently, this thesis provided some useful insights into high-order epistasis models associated with 90-day mortality, but further analysis needs to be done.

Future work involving the GenIMS data set and these three methods could tease out which genotypes are protective against or associated with 90-day mortality in CAP patients. The contingency tables created by MDR could be used as a starting point to determine exactly which genotypes for interacting SNPs are associated with higher risk for 90-day mortality. Specific tests could be carried out using logistic regression and adjusting for age, sex, smoking status, severity of illness, and/or sepsis. Chi-square tests of association between each SNP and 90-day mortality could be completed but were not necessary for this analysis; MDR and GMDR were able to run exhaustive searches for interactions regardless of the significance of main effects. Currently, Ritchie et al. [9] are researching strategies to improve MDR's performance. These strategies include more robust machine learning approaches such as parallel genetic algorithms, a nearest-neighbor method to label empty cells, classifying empty cells in lower dimensions, and applications to unbalanced case-control studies. Other researchers have looked at associations between haplotype clades (C/C/C etc.) using chi-square tests to determine differences in mortality outcomes [14]. Statisticians have already extended the MDR software to conduct this type of analysis in Hap-MDR.

5.2 CONCLUSIONS

This thesis used data from the very large GenIMS study to compare the performance of SNP epistasis models generated by MDR, GMDR, and stepPLR. Prediction accuracies were generally higher for GMDR compared to MDR, and stepPLR yielded substandard performance because the models predicted that all subjects were controls. Stepwise PLR, however, determined that IL-1A SNPs IL1A_M889, rs1894399, rs1878319, and rs2856837 were each significant predictors of 90-day mortality when adjusting for the other SNPs in the model. In addition, the model included a borderline significant second-order interaction between rs28556838 and rs3783520 associated with 90-day mortality in a cohort of CAP patients. The public health importance of this thesis is that the relative risk for CAP may be higher for a set of SNPs across different genes. The ability to predict which patients will experience a poor outcome may lead to more effective prevention strategies or treatments at earlier

stages. Furthermore, identification of significant SNP interactions can also expand scientific knowledge about biological mechanisms affecting disease outcomes. Future analysis using these three methods on the GenIMS data set could improve prediction accuracies by imputing missing data, adding non-genetic covariates, including main effects or multiple interactions between SNPs, and adjusting for potential confounding variables such as sepsis. In all, the GMDR method yielded higher prediction accuracies than MDR, and MDR performed better than stepPLR when generating SNP epistasis models associated with 90-day mortality in the GenIMS cohort.

APPENDIX A

SAMPLE SAS CODE FOR IL-10

```
*Import data sets;
data polygene; set perkins.polygene;
stnum=Study_Number; run;

data clinical; set perkins.clinical;
where race=1 and truecap=1;
keep stnum race truecap everss day90_status; run;

proc sort data=polygene; by stnum; run;
proc sort data=clinical; by stnum; run;

data whiteCAP;
merge polygene (in=ina) clinical (in=inb);
by stnum;
if inb and inc; run;

*Create data sets for MDR and GMDR;
data IL10case; set whiteCAP;
where (day90_status=1) and (IL10_M819 ne "") and (IL10_M1082 ne "")
   and (rs1800872 ne "") and (rs3024505 ne "") and (rs1800894 ne "")
   and (rs3204496 ne ""); run;

data IL10control; set whiteCAP;
where (day90_status=0) and (IL10_M819 ne "") and (IL10_M1082 ne "")
   and (rs1800872 ne "") and (rs3024505 ne "") and (rs1800894 ne "")
   and (rs3204496 ne ""); run;

proc surveyselect data=IL10control out=samp1control method=srs
   sampsize=121 seed=8416; run;

proc sort data=samp1control; by stnum; run;
proc sort data=IL10case; by stnum; run;
```

```
data IL10_MDR_final;
set samp1control (in=inc) IL10case (in=ind);
by stnum;
keep IL10_M819 IL10_M1082 rs1800872 rs3024505 rs1800894 rs3024496
   day90_status; run;

data IL10_GMDR_final; set WhiteCAP;
where (IL10_M819 ne "") and (IL10_M1082 ne "") and (rs1800872 ne "")
   and (rs3024505 ne "") and (rs1800894 ne "") and (rs3024496 ne "");
keep IL10_M819 IL10_M1082 rs1800872 rs3024505 rs1800894 rs3024496
   day90_status; run;

*Frequencies for missing data;
proc freq data=whiteCAP;
table IL10_M819 IL10_M1082 rs1800872 rs3024505 rs1800894 rs3024496
   day90_status; run;

*Tests for HWE and LD;
proc allele data=whiteCAP outstat=ld prefix=Marker perms=10000
   boot=1000 seed=5688;
var IL10_M819_1 IL10_M819_2 IL10_M1082_1 IL10_M1082_2 rs1800872_1 rs1800872_2
   rs3024505_1 rs3024505_2 rs1800894_1 rs1800894_2 rs3024496_1 rs3024496_2; run;


proc print data=ld; run;
```

APPENDIX B

SAMPLE R CODE FOR IL-10

```
library("stepPlr")

## IL10 Data
IL10data<-read.table("IL10_final_GMDR_Day90.txt", sep="\t", header=TRUE)
names(IL10data)
class(IL10data)
attach(IL10data)
dim(IL10data)
x6<-as.matrix(IL10data[,1:6])
y6<-as.matrix(IL10data[,7])

## step.plr
fit6<-step.plr(x6, y6, fix.subset=c(1,1,1,1,1,1), lambda=8,
   cp="bic", max.terms=9, type="both")
summary(fit6)

## predict.stepplr
pred6<-predict.stepplr(fit6, x6, type="class")
summary(pred6)

## cv.step.plr
cvfit6<-cv.step.plr(x6, y6, nfold=10, folds=NULL, lambda=c(1e-4,5,8),
   cp="bic", cv.type="class", trace=TRUE)
cvfit6["error"]

cvfit6["se.error"]
```

BIBLIOGRAPHY

[1] Moore, J. H., Gilbert, J. C., Tsai, C., Chiang, F., Holden, T., Barney, N., & White, B. C. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, 241, 252-261.

[2] Lou, X., Chen, G., Yan, L., Ma, J., Zhu, J., Elston, R. C., & Li, M. D. (2007). A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *The American Journal of Human Genetics*, 80, 1125-1137.

[3] Bastone, L., Reilly, M., Rader, D. J., & Foulkes, A. S. (2004). MDR and PRP: A comparison of methods for high-order genotype-phenotype associations. *Human Heredity*, 58, 82-92.

[4] Park, M. Y. & Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1), 30-50.

[5] Wood, K. A. & Marik, P. E. (2005). Community-acquired pneumonia and outcome: The importance of genetics. *Current Respiratory Medicine Reviews*, 1(2), 159-163.

[6] He, H., Oetting, W. S., Brott, M. J., & Basu, S. (2009). Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study. *BMC Medical Genetics 2009*, 10, I27.

[7] Solé-Violán, J., Rodríguez de Castro, F., García-Laorden, M. I., Blanquer, J., Aspa, J., Borderías, L., Briones, M. L., Rajas, O., Carrondo, I. M., Marcos-Ramos, J. A., Agüero, J. M. F., Garcia-Saavedra, A., Fiuza, M. D., Caballero-Hidalgo, A., & Rodriquez-Gallego, C. (2009). Genetic variability in the severity and outcome of community-acquired pneumonia. *Respiratory Medicine*, 104, 440-447.

[8] Gallagher, P. M., Lowe, G., Fitzgerald, T., Bella, A., Greene, C. M., McElvaney, N. G., & O'Neill, S. J. (2002). Association of IL-10 polymorphism with severity of illness in community acquired pneumonia. *Thorax*, 58(2), 154-156.

[9] Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., & Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69, 138-147.

[10] Hahn, L. W., Ritchie, M. D., & Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19(3), 376-382.

[11] Moore, J. H. "MDR 101." Epistasis Blog. 12 Nov. 2006. 16 Feb. 2009 (http://compgen.blogspot.com/2001_11_01_archive.html).

[12] Han, B., Park, M., & Chen, X. (2009). A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinformatics 2010*, 11(Suppl 3), S5.

[13] Kellum, J. A., Kong, L., Fink, M. P., Weissfeld, L. A., Yealy, D. M., Pinsky, M. R., Fine, J., Krichevsky, A., Delude, R. L., & Angus, D. C. (2007). Understanding the inflammatory cytokine response in pneumonia and sepsis. *Archives of Internal Medicine*, 167(15), 1655-1663.

[14] Sutherland, A. M., Walley, K. R., Manocha, S., & Russell, J. A. (2005). The association of Interleukin 6 haplotype clades with mortality in critically ill adults. *Archives of Internal Medicine*, 165, 75-82.