The Development and Application of a Risk Index to Predict Individualized Chronic Disease Risk

by

Reagan John Kelly

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2009

Doctoral Committee:
Professor Sharon R. Kardia, Chair
Professor Hosagrahar V. Jagadish
Professor Patricia A. Peyser
Associate Professor Kerby A. Shedden

Dedication

To my family, whose help and love made everything possible

Acknowledgements

I would like to thank Drs. Jagadish, Peyser, and Shedden for agreeing to serve on my dissertation committee and for their constructive and insightful comments on this work.

I would also like to thank Dr. Kardia for serving as my advisor and for giving me the chance to explore my research interests while also providing me with more experience in how to conduct research than I could have ever anticipated.

Thanks also to Todd Greene, Kristin Meyers, Joyce Wu, Yan Sun, Koji Yanagisawa, Doug Jacobsen, Linda Feldkamp, Tracy Fuller, Jian Chu, and Lori Carey for all of their help through this process.

Thanks to my family, whose support, encouragement, and understanding allowed me to reach this milestone.

Lastly, thanks to Jennifer Smith, my friend, colleague, and constant companion. Without her help, encouragement, and compassion this work would not exist.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SNP – Single Nucleotide Polymorphism

CHD – Coronary Heart Disease

FHS – Framingham Heart Study

PCA – Principal Components Analysis

CART – Classification and Regression Trees

RF – Random Forests

SVM – Support Vector Machines

ROC Curve – Receiver Operating Characteristics Curve

AUC – Area Under the Curve

GRS – Genetic Risk Score

Chapter 1

Introduction

1.1 Public Health Importance of Selected Chronic Diseases

Chronic diseases, such as hypertension, type 2 diabetes, and heart disease affect 109 million Americans (DeVol, et al, 2007), or more than 1 in every 3 Americans. These diseases are estimated to cost the U.S. $1.3 trillion annually through both direct healthcare spending on treating the diseases and their consequences and indirect costs through lost wages and productivity. Between 1980 and 2007, type 2 diabetes incidence rates have increased from 3.5% to 7.8% (National Center for Health Statistics), and a similar, if less dramatic, rise has been reported in hypertension (Tu, et al, 2008). Heart disease prevalence has increased from 8% of the adult US population in 1981 (Collins, 1986) to 11% of the adult US population in 2006 (Pleis, et al, 2007), and remains the single greatest cause of death among US adults (Rosamond, et al, 2008). The increase in obesity among children and young adults may mean that the rates of these common chronic diseases will continue to rise (Lee, 2008).

By their nature, chronic diseases have long-term cost impacts, and while they can be controlled they cannot be cured. The most cost-effective way to deal with these diseases is to focus on preventing or delaying their onset and, in individuals who have already developed a chronic disease, to prevent the development of complications (Russell, 2009). These preventive efforts, however, will be most cost-effective, and may even be

cost-saving, if targeted at those individuals at highest risk. The identification if individuals at increased risk for developing a chronic disease or at higher risk of a complication from a chronic disease, however, is not routinely done in clinical practice (Emery, et al, 2001). For many chronic diseases, in fact, there are no widely used methods to estimate an individual's risk. An important exception, described in detail below, is the Framingham Coronary Heart Disease (CHD) risk score.

The promise of genomic medicine has been that information about an individual's genome could be used to identify the diseases they are at increased risk for and give insight into how to treat the diseases they have. As the "omics" revolution has progressed it has become possible to measure all of the genes being expressed in a sample of cells from an individual, identify the most abundant proteins in those cells, profile all of the metabolites in those cell, and genotype nearly a million single nucleotide polymorphisms (SNPs). This has provided great hope that risk prediction methods can be developed that integrate "omics" information to improve prediction accuracy as well as allow for the early identification of individuals at increased risk of developing a disease. However, one of the key stumbling blocks on the path toward genomic medicine is the lack of translational studies on how genetic information can be used to better predict, diagnose, and treat diseases. Most complex diseases, like hypertension, diabetes, and heart disease, have multifactorial genetic and environmental etiologies and will need complex, or at least multivariable risk prediction methods to make clinically useful predictions.

This dissertation develops, tests, and applies a risk prediction method designed to predict susceptibility to chronic disease. In order to do so, first the characteristics that make risk prediction methods suited for use in clinical practice, namely analytical validity, clinical utility, and clinical validity (Holtzman, et al, 1997), are examined in Section 1.2. In Section 1.3 a short survey of different types of risk prediction methods and their characteristics is outlined, focusing on methods that exclusively use data easily measured in a clinical setting, methods that exclusively use "omics" data that is collected in a high-throughput manner but is typically not collected in clinical practice, and methods which integrate clinical and high-throughput "omics" data together. In Section 1.4, genetic risk scores, a class of risk prediction methods that use information about SNPs to make predictions about an individual's risk of disease, are reviewed to identify ways that this class of methods can be improved using the insights about risk prediction methods from Section 1.3. Finally, Section 1.5 provides an overview of the risk prediction method that is developed and of the structure of the remainder of this dissertation.

When assessing or developing any risk prediction method there are a number of decisions that must be made and issues that must be considered. Figure 1-1 gives a graphical overview of these. Before a risk prediction method can be used there are three categories of issues that must be examined, data issues, modeling building issues, and model assessment issues. To address the data issues, the outcome being examined, the types of variables being used as input, and any transformations or modifications that must be done to those variables needs to be chosen. To address model building issues, the way the risk prediction method models the available variables as a function of the outcome, the way

variables are selected, and the way the parameters are estimated must be chosen. Lastly, to address model assessment issues, the way the predictive accuracy of the method will be estimated and how to interpret the model must be decided. Chapter 2 describes in detail the choices that have been made for the risk prediction method that forms the core of this dissertation. Section 1.2 discusses the features that are relevant to clinical risk prediction specifically.

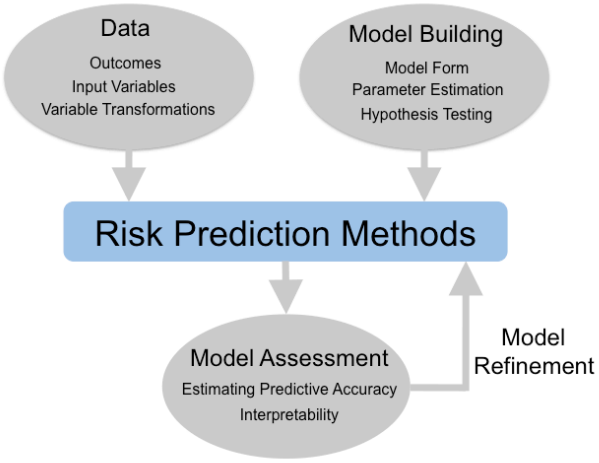


Figure 1-1 Considerations for Risk Prediction Methods

1.2 Features of Clinical Risk Prediction Methods

Creating a system that can make clinical predictions requires more than simply predictive accuracy. There are three important features that these systems must exhibit to come into broad clinical use, and any attempt at creating such a system must be guided by three principles: analytical validity, clinical validity and clinical utility (Haddow, et al, 2004). Figure 1-2 shows the relationship between these elements and the features of risk prediction methods described in Section 1.1

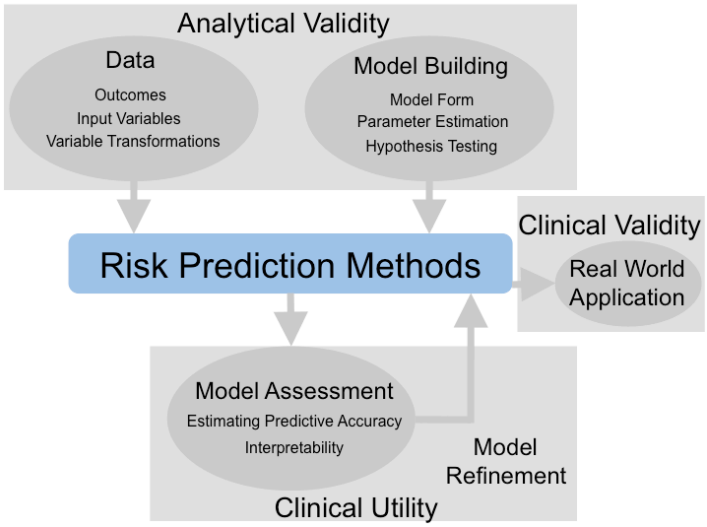


Figure 1-2 Relationship of Analytical Validity, Clinical Utility, and Clinical Validity with Risk Prediction Methods

1.2.1 Analytical Validity

Any algorithm that will be used by physicians must also offer an assessment of its analytical validity, or how well it identifies and classifies at-risk individuals in the population (Haddow, et al, 2004). Minimally, it should provide an estimate of how well it has performed in the past and how well it can be expected to perform in the future. There are numerous ways to assess this, but the most standard is an estimate of prediction accuracy (what proportion of predictions were correct) or of prediction error (what proportion of predictions were incorrect). Typically, a cross-validation scheme is used to measure the algorithm's performance. This allows for a reliable assessment of performance without the added requirement of a second, fully independent test population. It is important to remember, however, that the performance estimate is valid only for individuals similar to those used to build the predictor. Applying a predictor that

was created using people older than sixty with hypertension to, for example, a population of forty-year olds with normal blood pressure will produce a performance estimate that is unlikely to be valid.

1.2.2 Clinical Utility

Next, the prediction that comes out of the system must have clinical utility, which means that a physician must be able to make a treatment decision based on the prediction made (Haddow, et al, 2004). It is unrealistic to expect that in every case any algorithm is capable of extracting enough information to make a useful prediction, but at least one of the potential predictions should offer clinicians information that can be used to tailor a diagnosis or treatment plan. For example, if a given type of cancer has a known, rare subset that responds extremely well to a particular type of treatment, an algorithm that can distinguish that subtype would be of clinical utility even though it says nothing about the majority of patients that the prediction algorithm is applied to. Conversely, an algorithm that can distinguish between two common molecular sub-types of a cancer, whose treatments, therapy responses, and prognoses are essentially identical, would not be of real clinical utility, even though it makes a very accurate prediction about the majority of patients with respect to sub-type.

Clinical utility can be assessed by examining a risk prediction method's impact in four areas: how it affects a physician's understanding of the diagnosis of a patient, how it affects a physician's choice of treatments for a patient, how the use of the information from this risk prediction method affects the clinical outcome of a patient (either in terms

of mortality or in terms of quality of life), and how the use of the information from this risk prediction method has a societal benefit (such as improved cost-effectiveness in treating a particular disease (Tatsioni, et al, 2005). While the assessment of a risk prediction method's impact on patient's outcomes or cost-effectiveness would require separate studies, Grosse suggests that this is not always necessary, and that impact on a doctor's diagnostic thinking or therapeutic choice could be sufficient to determine a risk prediction method's clinical utility even if health outcomes and cost-effectiveness data do not exist (Grosse, et al, 2006).

1.2.3 Clinical Validity

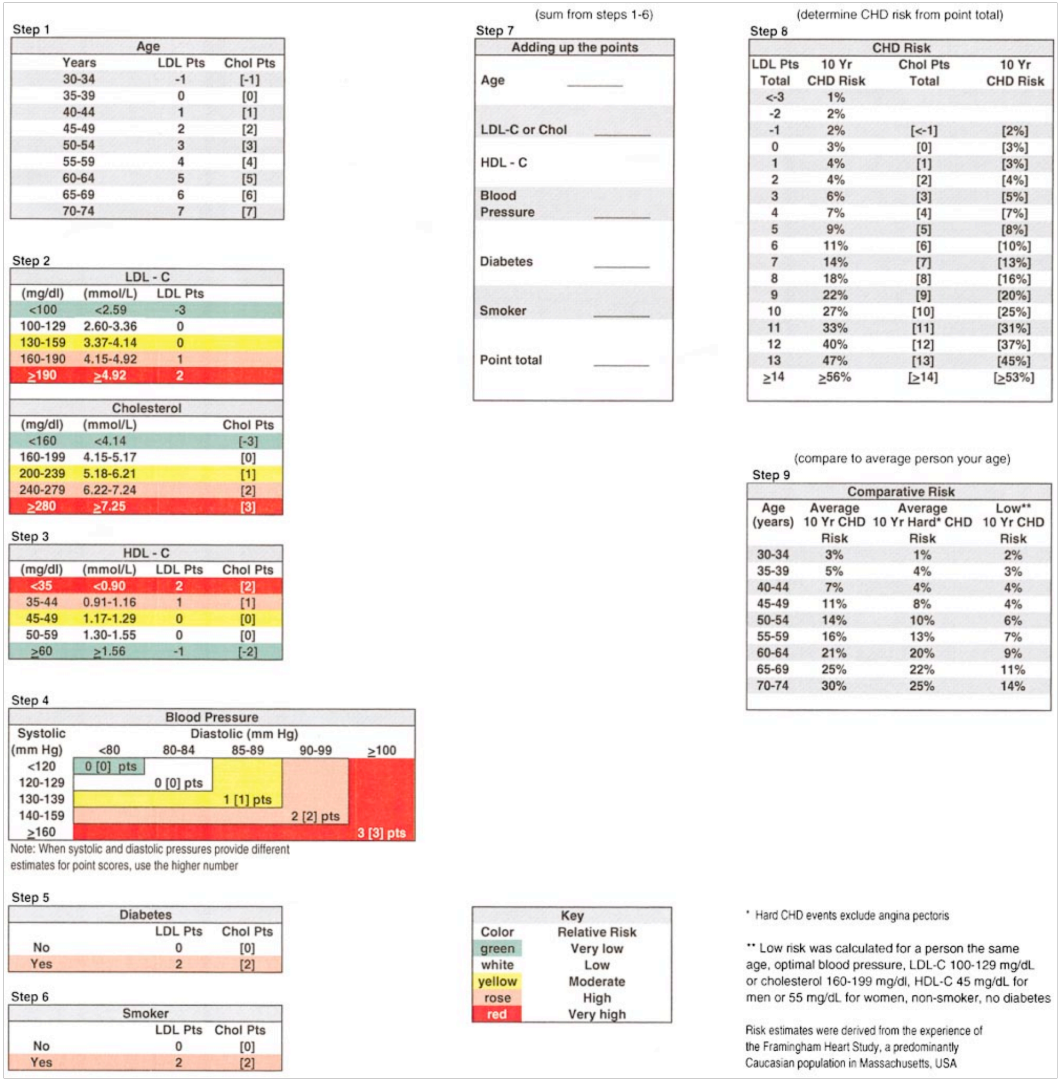
Finally, any prediction system that is intended for wide-spread clinical use should also provide evidence of clinical validity, or how well the system performs when applied to a clinical population (Holtzman, et al, 1997). Clinical validity is assessed using a several criteria: 1) clinical sensitivity, or the probability that a person who develops the disease was identified as high risk, 2) clinical specificity, the probability that a person who does not develop the disease was identified as low risk, 3) positive predictive value, the probability a person who has been predicted as being high risk will develop the disease, and 4) negative predictive value, the probability that a person who has been predicted as low risk will not develop the disease (Holtzman, et al, 1997). Clinical validation of risk prediction methods is an important step; however, it most often accomplished through the collection of a secondary validation set after the method is used in clinical practice. For example, Aaronson, et al. (Aaronson, et al, 1997) describe the heart failure survival score (HFSS), which will be discussed in detail below as an example of clinical nomogram-

based risk prediction method. The study, however, was conducted prior to the wide-spread use of β-blocker treatment in heart failure. In order to test if the HFSS was valid in patients treated with β-blockers, Koelling, et al. (Koelling, et al, 2004) undertook a follow-up study which found that HFSS strata was a significant predictor of survival in both β-blocker treated and untreated groups, and that the area under the receiver operator characteristics (ROC) curve (AUC), a measure of the predictive accuracy of a model, for the HFSS was similar in both β-blocker treated and untreated patients. Because of prospective nature of clinical validation studies, however, a discussion of the clinical validity of the risk index described is beyond the scope of this dissertation.

1.3 Risk Prediction Methods

Risk prediction methods have been a topic of much research for more than two decades. As medical and epidemiological research has advanced and the risk factors underlying chronic diseases and other adverse events have been identified, ways of assessing a patient's risk have followed. One of the most famous risk prediction methods is the Framingham CHD risk score (Wilson, et al, 1998). This is a simple algorithm that assigns a numerical score to the values of several easily obtained clinical variables and translates the patient's final score into their risk of developing CHD within ten years. The Framingham CHD risk score sheet is shown in Figure 1-3. Other risk prediction methods can predict a patient's risk of disease recurrence (Stephenson, et al, 2005), their likelihood of responding to a drug (Thuerigen, et al, 2006), or how aggressive their cancer is (Spurgeon, et al, 2006). These prediction methods run the gamut from simple to complex, and their discriminatory power and predictive utilities vary widely. This review

of risk prediction methods focuses on three main groups: 1) prediction methods that use exclusively clinical measurements, 2) methods that use gene expression or proteomic signatures, and 3) methods that integrate clinical and gene expression measurements.



Step 1

Age		
Years	LDL Pts	Chol Pts
30-34	-1	[-1]
35-39	0	[0]
40-44	1	[1]
45-49	2	[2]
50-54	3	[3]
55-59	4	[4]
60-64	5	[5]
65-69	6	[6]
70-74	7	[7]

Step 2

LDL - C		
(mg/dl)	(mmol/L)	LDL Pts
<100	<2.59	-3
100-129	2.60-3.36	0
130-159	3.37-4.14	0
160-190	4.15-4.92	1
≥190	≥4.92	2

Cholesterol		
(mg/dl)	(mmol/L)	Chol Pts
<160	<4.14	[-3]
160-199	4.15-5.17	[0]
200-239	5.18-6.21	[1]
240-279	6.22-7.24	[2]
≥280	≥7.25	[3]

Step 3

HDL - C			
(mg/dl)	(mmol/L)	LDL Pts	Chol Pts
<35	<0.90	2	[2]
35-44	0.91-1.16	1	[1]
45-49	1.17-1.29	0	[0]
50-59	1.30-1.55	0	[0]
≥60	≥1.56	-1	[-2]

Step 4

Blood Pressure					
Systolic (mm Hg)	Diastolic (mm Hg) <80	80-84	85-89	90-99	≥100
<120	0 [0] pts				
120-129		0 [0] pts			
130-139			1 [1] pts		
140-159				2 [2] pts	
≥160					3 [3] pts

Note: When systolic and diastolic pressures provide different estimates for point scores, use the higher number

Step 5

Diabetes		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

Step 6

Smoker		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

(sum from steps 1-6)

Step 7

Adding up the points	
Age	_____
LDL-C or Chol	_____
HDL - C	
Blood Pressure	_____
Diabetes	_____
Smoker	_____
Point total	_____

(determine CHD risk from point total)

Step 8

CHD Risk			
LDL Pts Total	10 Yr CHD Risk	Chol Pts Total	10 Yr CHD Risk
<-3	1%		
-2	2%		
-1	2%	[<-1]	[2%]
0	3%	[0]	[3%]
1	4%	[1]	[3%]
2	4%	[2]	[4%]
3	6%	[3]	[5%]
4	7%	[4]	[7%]
5	9%	[5]	[8%]
6	11%	[6]	[10%]
7	14%	[7]	[13%]
8	18%	[8]	[16%]
9	22%	[9]	[20%]
10	27%	[10]	[25%]
11	33%	[11]	[31%]
12	40%	[12]	[37%]
13	47%	[13]	[45%]
≥14	≥56%	[≥14]	[≥53%]

(compare to average person your age)

Step 9

Comparative Risk			
Age (years)	Average 10 Yr CHD Risk	Average 10 Yr Hard* CHD Risk	Low** 10 Yr CHD Risk
30-34	3%	1%	2%
35-39	5%	4%	3%
40-44	7%	4%	4%
45-49	11%	8%	4%
50-54	14%	10%	6%
55-59	16%	13%	7%
60-64	21%	20%	9%
65-69	25%	22%	11%
70-74	30%	25%	14%

Key	
Color	Relative Risk
green	Very low
white	Low
yellow	Moderate
rose	High
red	Very high

* Hard CHD events exclude angina pectoris

** Low risk was calculated for a person the same age, optimal blood pressure, LDL-C 100-129 mg/dL or cholesterol 160-199 mg/dl, HDL-C 45 mg/dL for men or 55 mg/dL for women, non-smoker, no diabetes

Risk estimates were derived from the experience of the Framingham Heart Study, a predominantly Caucasian population in Massachusetts, USA

Figure 1-3 The Framingham CHD Risk Score Worksheet

1.3.1 Clinical Nomograms

The oldest and most heavily researched type of risk prediction method is the clinical nomogram. This class of methods combines demographic and clinical measurements,

such as age, gender, cholesterol, and blood pressure that are reasonably easily obtained by a physician and make a prediction about a person's risk of a disease or complication. The Framingham CHD risk score is a well known example, and the score sheet distributed to physicians is shown in Figure 1-3. To use this nomogram, a physician needs a patient's age, LDL or total cholesterol level, HDL level, systolic and diastolic blood pressure, current smoking status, and diabetes status. Using this information the doctor adds up a patient's points from the risk score sheet (each risk table has two sets of point values; the LDL points column is used if the doctor has the patient's LDL cholesterol level and the cholesterol points are used if the doctor has the patient's total cholesterol level. Once points have been assigned for each of these variables they can be added up and an estimate of 10-year CHD risk can be obtained from the table. Lastly, the doctor can then compare the individual's 10-year CHD risk to the average 10-year CHD risk for an individual in the same age group.

The Framingham CHD risk score discussed above is one example, but others exists for predicting whether or not a patient's cancer is likely to be malignant (Lu, et al, 2003), how long a patient will survive after transplant (Thuluvath, et al, 2003), and nearly any other clinically important outcome or event. These predictions are typically based on fairly simple statistical methods, such as logistic regression (Thuluvath, et al, 2003), proportional hazards modeling (Aaronson, et al, 1997,Wilson, et al, 1998), and classification and regression trees (CART) (Spurgeon, et al, 2006). Outcomes that have proven resistant to prediction with traditional statistical methods are modeled using more advanced statistical methods, such as support vector machines (SVM) (Lu, et al, 2003)

and random forests (RF) (Ward, et al, 2006). This section briefly reviews each of these clinical nomograms to illustrate the type of modeling and predictions that are the standard in the field.

Wilson, et al. (1998) used Cox proportional hazards modeling to define the Framingham CHD risk score. This score is simple for physicians to calculate and apply to individual patients, provides a quantitative measure of risk, and had been validated in a number of external studies (Milne, et al, 2003,Ramachandran, et al, 2000,McEwan, et al, 2004). Physicians can then tailor interventions to reduce 10-year CHD risk for patients at very high risk. However, it is not clear how well patients understand and can use absolute measures of risk like those that are provided by this score, meaning that the impact of this risk score on patient's medical decision making is unclear. Additionally, the score's applicability and reliability is poorly understood for African Americans and Asians.

Thuluvath, et al. (2003) developed a logistic regression model to predict survival one month, one year, and five years after organ transplant based on pre-transplant clinical characteristics. This method is extremely straightforward for physicians to understand and implement, and, in the test sample, the proportion of survival at each time point is concordant with the prediction of the models. No assessment of the predictive accuracy is provided, however, and so, while the method is effective at estimating the proportion of individuals that survive it is not clear that the method is effective at predicting the survival of particular individuals.

Spurgeon, et al. (2006) classified prostate cancer samples into high and low aggressiveness groups using classification trees built with both demographic information and results from prostate ultrasounds. This method is simple to create and interpret and has excellent sensitivity, identifying 91% of truly aggressive tumors. However, its specificity and positive predictive values are quite low, meaning it has difficulty identifying less aggressive tumors.

Aaronson, et al. (1997) used Cox modeling to develop a heart failure survival score (HFSS) which is used to stratify patients with end-stage heart failure by high, medium, and low risk of death. This is easy for a physician to implement and interpret and is very easily applied to new patients. Log-rank tests show good separation between high, medium, and low risk. This study was performed in a sample collected between 1993 and 1995, at a time before the wide-spread use of β-blocker to treat heart failure patients. A validation study described in detail above (Koelling, et al, 2004), however, confirmed the HFSS's applicability in patients using β-blockers.

Ward, et al. (2006) developed a risk prediction model using Random Forests to predict short-term mortality from lupus erythematosus. The model's overall misclassification is 11-13%, which is quite good, but the model was difficult to develop, very sensitive to changes in the modeling parameters, such as the number of variables considered at each split and the number of trees constructed, and not validated by any external sample. Lu, et al. (Lu, et al, 2003) describe an even more complex model, developed using a Support Vector Machine (SVM) with a radial basis function as its kernel. It predicts ovarian

cancer mortality using clinical, histological, and ultrasound variables. The model's performance is good and well-balanced, with a sensitivity of 85% and a specificity of 84%, but the method is extremely complex to develop and train, or calibrate to making predictions about a given population, and the resulting model is nearly impossible for physicians to interpret. An overview of these six risk prediction methods based on clinical data is given in Table 1-1 to illustrate the breadth of modeling approaches and highlight their advantages or disadvantages.

Overall, clinical nomograms have a number of advantages that make them attractive choices for risk prediction. They base their prediction on variables that are commonly used in clinical practice and are fairly straightforward to obtain. Using these variables has the additional advantage of giving the predictive models some amount of interpretability. This might offer a physician insight into how to treat the disease or what preventative steps to take. Also, most clinical nomograms are fairly straightforward to calculate. Since they typically depend on a small number of variables, it is simple to develop either score sheets (e.g., the Framingham CHD risk score, Figure 1-3) or even web-based applications (National Cancer Institute) that physicians can easily enter information into and derive a risk estimate. Even more complex clinical nomograms using more sophisticated statistical modeling can be easily converted to computer programs that are relatively simple for physicians to use to estimate a patient's risk.

Table 1-1 Brief Overview of Selected Clinical Nomograms (Alphabetic by First Author)

Author	Year	Description	Pros	Cons
Aaronson (Cox Modeling)	1997	Uses proportional hazards modeling to stratify patients with end-stage congestive heart failure into high and low risk of death. Log rank tests indicate that the model does a good job of separating low, medium, and high risk.	Non-invasive variables. Simple to calculate and interpret for a physician. Quickly applicable to a single new patient.	
Lu (Support Vector Machines)	2003	Uses Support Vector Machines with a Radial Basis Function kernel to identify malignant ovarian tumors using clinical, histological and ultrasound measurements. In a validation set the model gives a sensitivity of 85%, a specificity of 84%, and a positive predictive value of 73%.	Good predictive ability. Validation provides good support for the results.	Extremely complicated to create and train. Complete black box (results not easy to interpret).
Spurgeon (Classification and Regression Trees)	2006	Combines demographic information with data from prostate ultrasound to assess prostate cancer aggressiveness. In a validation sample the model gives a sensitivity of 91%, a specificity of 33.5%, and a positive predictive value (PPV) of 12.7%.	Very sensitive, easy to calculate, and simple for physician to interpret.	Poor specificity and PPV.
Thuluvath (Logistic Regression)	2003	Uses logistic regression modeling that includes clinical characteristics prior to transplant to predict survival at one month, one year, and five years. The proportion of survival observed at each time point is concordant with the prediction.	Simple to create.	Authors provide no realistic assessment of predictive accuracy. Despite extremely large sample size, only use a single validation as opposed to cross-validation. Modeling approach does not accurately reflect the complexity of the underlying phenotype.
Ward (Random Forests)	2006	Predicts short-term mortality in patients with systemic lupus erythematosus using Random Forests with 47 clinical and demographic variables. The mean classification error ranged from 11-13%.	Good performance.	Impossible to determine the relationships between the variables. Difficult to train and tune. Lacks validation to determine broad applicability.

Wilson (Cox Modeling)	1998	Uses Cox proportional hazards modeling from a large cohort (~6000) using 10-year incidence of CHD to develop a risk score. The area under the ROC Curve ranges from 0.69-0.77, indicating that the model predicts CHD risk well.	Easy to apply to a patient. Validated in numerous studies.	Applicability to any group other than Caucasians unclear. Overestimates older patients' risk of CHD.

Clinical nomograms also have disadvantages. Because they are typically developed using a small number of variables and relatively simple statistical techniques, they are not always well suited to predicting extremely complex outcomes, such as which individuals will require adjuvant chemotherapy. The development and validation of these risk prediction methods typically requires very large samples and a large amount of time for follow-up of outcomes, but their broad applicability is uncertain. Even the Framingham CHD score, which has been validated in a number of studies, is not necessarily applicable to groups other than Caucasians between 30 and 74 years of age, and it also overestimates CHD risk in older individuals.

1.3.2 Molecular Signatures

Molecular signature detection methods are one of the newest classes of risk prediction methods, and they have gained increasing popularity as gene expression arrays have dropped in cost and improved in quality. Simply, these methods attempt to identify a subset of gene expression values that allows samples to be classified into risk groups (e.g., low, medium, and high risk) or disease subgroups (e.g., histological subtypes of cancer with known differences in treatment or survival outcomes). The algorithms used to identify the signature and classify the samples vary in their complexity. The most straightforward algorithms use fairly simple methods such as hierarchical clustering or

linear discriminant analysis to group the samples and identify genes that are differentially expressed between the groups to define the signature (Pawitan, et al, 2005). More complex methods can involve using Support Vector Machines (Thuerigen, et al, 2006) or Random Forests (Hoffmann, et al, 2006) to identify the signature and classify the samples. Some molecular signature risk prediction methods may also offer new ways of thinking about the mechanism for the disease or outcome. While the identification of a gene in a molecular signature is not hard evidence of a causal role, the signature genes offer a starting place for deeper mechanistic investigation that cannot be achieved with clinical variables alone.

Briefly, Thuerigen, et al. (2006) developed a Support Vector Machine based method to predict whether a patient will respond to a course of chemotherapy. Using a technique called recursive feature elimination they identify a 512 gene signature that provided 78% sensitivity and 90% specificity in a validation set. The complexity of the method makes interpretation difficult, and the only way to measure the signature is to run a full gene expression array – something not typical in clinical practice. Pawitan, et al. (2005) used Linear Discriminant Analysis to identify a 64 gene signature that stratifies a group of breast cancer patients into high, medium, and low risk of distant relapse or death to identify those patients with the greatest need for adjuvant chemotherapy. The model does well at identifying patients who do well without therapy and who do poorly despite therapy, but it does not identify those who would respond well to treatment or those who would do poorly without treatment, which are more clinically useful classifications. Hoffman, et al. (2006) used Random Forests to identify subgroups of patients in a small

set of childhood acute lymphocytic leukemia patients. While the predictive accuracy is high, little can be said about the resulting model, both because it is difficult to interpret and because the sample used to generate it was so small. Table 1-2 presents an overview of these three molecular signature based risk prediction schemes to illustrate the breadth of modeling approaches used to create molecular signatures and the types of predictions these methods make. While this section focused on gene expression-based molecular signatures, these issues extend to proteomic-based biomarker detection, metabolomic-based metabolite profiles, and genome-wide SNP genotypes.

Table 1-2 Brief Overview of Selected Molecular Signatures (Alphabetic by First Author)

Author	Year	Description	Pros	Cons
Hoffmann (Random Forests)	2006	Uses Random Forests to find a small set of genes for subgroup distinction in childhood ALL. Cross-validation showed a prediction accuracy of 98%.	High accuracy.	Not enough predictive performance metrics given to assess how well it performs.
Pawitan (Linear Discriminant Analysis)	2005	Using Linear Discriminant Analysis, identified 64 genes to determine which breast cancer patients need adjuvant chemotherapy in addition to surgery. Log rank tests show good separation between high, medium, and low risk groups.	Good at identifying people who do well without treatment and people who do poorly despite treatment.	Difficulty identifying people who would do well with treatment and those who would do poorly without treatment (need clinical trial for that).
Thuerigen (Support Vector Machines)	2006	Using Support Vector Machines and Recursive Feature Elimination, found a 512 gene signature that they used to predict whether or not a patient will have a complete response to a specific course of chemotherapy. In a validation set the signature gave a sensitivity of 78%, a specificity of 90%, and a positive predictive value of 64%.	Very high sensitivity and specificity. Validation set provides convincing evidence for applicability.	No simple way to measure the signature; have to run a microarray. Only predicts whether the person has a complete response, not whether they survive.

Molecular signatures developed with traditional statistical methods (e.g., Pawitan, et al. 2005) are potentially useful because they are straightforward to construct and can be interpreted. This makes them easy for physicians to use and easy to develop into a score sheet or application. However, the scale and complexity of the data being examined means that these methods are not well suited for complex outcomes like chronic disease risk, where the causes are not only multifactorial, but there are likely numerous pathways leading to the outcome.

Molecular signatures using more sophisticated machine learning algorithms have the advantage of being able to predict more complex outcomes, as well as deal effectively with large numbers of weak predictors, or variables which explain only a small fraction of the variability in the outcome. They often use sophisticated variable selection procedures which help tune the models very precisely, giving high predictive accuracy, sensitivity, and specificity. The complexity of the methods, though, makes it difficult if not impossible for physicians to interpret the results. While tuning can provide excellent performance in one dataset, it can also lead to over-fitting, which reduces the model's applicability to other populations. Additionally, while the molecular signature methods discussed above have been assessed with external validation sets, no further validation has been published, and none of the methods are near real-world use in clinical practice.

1.3.3 Integrative Approaches

Most risk prediction methods focus on one type of data, such as clinical data or high-throughput "omics" data. Recently, however, more integrative approaches to risk prediction that combine information from high-throughput biological assays with clinical data have been attempted. These approaches are important, because they attempt to more accurately reflect the complexity of the underlying conditions and to utilize the full range of data that is available for prediction. Attempting to predict a person's risk of a disease, then, by only considering gene expression misses an important source of information that may be provided by clinical measurements. At the same time, clinical variables are not uniformly accurate predictors, and a specific clinical phenotype may have different

meanings depending on the genetic background of the individual. Ignoring information about genetics, then, misses a potentially important tool for stratifying disease risk.

Integrative methods are, for the time being, fairly uncommon, but they mark the beginning of a trend towards risk prediction methods that are respectful of the complexity of the outcome being predicted. They range in complexity from logistic regression (Stephenson, et al, 2005) and Cox proportional hazards modeling (Pittman, et al, 2004) to advanced statistical techniques such as partial-sliced inverse regression (Li, 2006).

Three examples of this type of integrative methods are reviewed in Table 1-3. Stephenson, et al. (Stephenson, et al, 2005) develop a logistic regression model to predict prostate cancer recurrence. When prediction was made using only gene expression data, the accuracy was only 53%, but when a clinical nomogram was added to the gene expression data the accuracy rose to 89%. The clinical nomogram + gene expression data improved the accuracy of predictions for those people that the clinical nomogram alone classified as indeterminate, but for those that were well predicted by the nomogram the gene expression data did not offer any additional benefit. Pittman, et al. (Pittman, et al, 2004) describe a "clinico-genomic model" that uses Cox proportional hazards modeling with both clinical data and gene expression data to predict breast cancer recurrence. Combining the data provides better performance than either the clinical data or the gene expression data alone. However, the clinical data seems to underperform in this case, raising some questions about the broad applicability of these results. Li, et al. (Li, 2006) combine clinical data and 40 principal components based on gene expression data using

partial-slice inverse regression to predict recurrence in a dataset of diffuse large B-cell lymphoma. The combined clinical and gene expression data provided better separation between the medium- and low-risk strata than the clinical data alone. However, the method is extremely complex and the use of 40 principal components means that there is no way to interpret the results.

Table 1-3 Overview of Selected Integrative Approaches (Alphabetic by First Author)

Author	Year	Description	Pros	Cons
Li (Partial-sliced Inverse Regression)	2006	Uses partial-sliced inverse regression to predict survival in DLBCL (diffuse large B-cell lymphoma). The first 40 principal components are used for the inverse regression. Combined model shows good separation by the log rank test between the low and medium risk groups.	By adding clinical and genomic information together, this method achieves much better separation between medium and low risk groups.	Because principal components are used, the results have no scientific interpretation. The method is complicated to perform.
Pittman (Cox Modeling)	2004	Incorporates clinical info and gene expression to predict breast cancer recurrence with proportional hazards modeling. The combined data gives 90% sensitivity and 90% specificity. The gene expression only model and the clinical only model can only reach 70-75% sensitivity to achieve 90% specificity.	Combining clinical and genomic data offers better performance than either individually.	Clinical predictors do not perform as well as might be expected, potentially preventing an accurate comparison between the clinical data only model and the clinical data plus genomic data model.
Stevenson (Logistic Regression)	2005	Uses logistic regression to predict prostate cancer recurrence with two models: exclusively gene expression, and gene expression + clinical nomogram. Assessed using leave-one-out cross-validation and compared to the clinical nomogram alone. The clinical + gene expression model has an accuracy of 89% vs. 53% from the gene expression data alone.	Gene expression data significantly improves the prediction for people who the nomogram classifies as indeterminate.	For patients whom the nomogram predicts well, this method does not add any value.

The primary advantage of these integrative approaches is an increase in predictive accuracy. It is often the case that while a simple risk prediction method does a fairly good job at predicting most of the individuals, they often do a poor job classifying those individuals that are at an intermediate level of risk. The integrated systems, however, perform quite well for those individuals predicted at intermediate risk using these more simple approaches approach. The extra information that results from using both types of data yields a prediction method that is superior to ones built from either of the data types alone for those individuals initially assessed as having intermediate risk. However, these risk prediction methods produce models that are often not as easily interpretable to physicians as clinical nomograms, because the introduction of gene expression levels requires a detailed understanding of cellular molecular biology and genetics not typically found among physicians. Despite these disadvantages, integrative risk prediction systems offer significant promise for risk prediction and are likely to become more common as "omics" technologies move into clinical practice. One of the goals of this dissertation is to develop and refine a risk prediction method that integrates clinical data and single nucleotide polymorphism (SNP) genotypes to increase high predictive accuracy for individuals at intermediate risk.

1.4 Genetic Risk Scores

Since the goal of this dissertation is to utilize emerging genetic technologies to improve risk prediction, a more detailed examination of genetic risk scores follows. "Genetic risk score" is the name used to describe a class of risk prediction methods that work by

summing an individual's risk alleles and using the final score as a basis for risk prediction. It is well established that chronic diseases have a strong genetic component, with heritability estimates for hypertension of 30% (Agarwal, et al, 2005), for diabetes from 72-78% (Permutt, et al, 2005), and for coronary heart disease from 49-51%(Fischer, et al, 2005), and polymorphisms have been identified which are known to influence an individual's risk of developing these particular chronic diseases (See Table 1-4). These well-validated associations are important to understanding the etiology of chronic diseases, but they are of limited value by themselves. They represent either a significant risk to a very small number of people (an example of this would be the BRCA1/2 mutations, which greatly increase lifetime risk of breast and ovarian cancer , but affect only a small number of women (Antoniou, et al, 2003)), or, more commonly, a small to moderate risk which is modified by other polymorphisms, environmental factors, and health behaviors (Cho, 2009,Burke, et al, 2007). Rather than focus on single polymorphisms, genetic risk scores combine information from a number of polymorphisms with the explicit purpose of making a prediction about an individual's risk.

Table 1-4 Examples of Selected Genetic Associations with Chronic Diseases

Chronic Disease	Type of Study	Reference	Association	Notes
Hypertension	Candidate Gene	(Rice, et al, 2000,Rigat, et al, 1990,Turner, et al, 1999)	Angiotensin Converting Enzyme (*ACE*)	The insertion/deletion polymorphism in the A*CE* gene is one of the most widely validated genetic contributors to hypertension risk.
	Genome Wide	(Wang, et al, 2009,Org, et al, 2009)	*STK39* *CDH13*	Recent studies have identified *STK39* as a hypertension susceptibility locus in Amish & other Caucasian samples (Wang) and *CDH13* as a hypertension susceptibility locus in two

				European samples (Org).
Diabetes	Candidate Gene	(Grant, et al, 2006)	*TCF7L2*	Polymorphisms in *TCF7L2* has been associated with diabetes risk in a number of studies.
	Genome Wide	(Scott, et al, 2007,Zeggini, et al, 2007,Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, et al, 2007)	*TCF7L2* *CDKAL1* *FTO*	Working together, large consortia have identified polymorphisms in three genes which contribute to diabetes risk in several independent samples.
Heart Disease	Candidate Gene	(Klerk, et al, 2002)	*MTHFR*	A meta-analysis found that a polymorphism in *MTHFR* increased heart disease risk in Europeans, but not in North Americans. The authors suggest this may be due to an interaction between the polymorphism and folate levels.
	Genome Wide	(McPherson, et al, 2007,Helgadottir, et al, 2007,Schunkert, et al, 2008)	9p21	Polymorphisms from 9p21 have been associated with heart disease risk independent of hypertension or diabetes in a number of studies.

The risk index method developed and tested in this dissertation draws from the idea of a genetic risk score and the risk index method utilized by Beer, et al (Beer, et al, 2002). Genetic risk scores (GRS) are particularly appealing because they allow for the compiling of a large amount of potential risk information in a straightforward manner. Ideally, they also allow for a semi-quantitative comparison of the disease risk for two individuals (i.e., while the overall scale of the risk score may differ between populations, the difference in risk score between two individuals from the same population should provide information about one individual's risk relative to the other). Looking closely at a few examples of GRS methods offers insight into their best characteristics and how to leverage these while minimizing the drawbacks of the methods.

1.4.1 A Simple Example of a Genetic Risk Score

Morrison, et al. (2007) propose a straight-forward approach to developing a GRS, focusing on the impact of polymorphisms on a survival phenotype. Using time to developing coronary heart disease as their outcome, the authors use the Atherosclerosis Risk in Communities (ARIC) cohort to develop a GRS using 116 polymorphisms. Working separately in whites and blacks, the authors begin by using each polymorphism as the explanatory variable in a Cox proportional hazards regression, and excluding all polymorphisms with a p-value > 0.10. They identified 11 polymorphisms which passed this criteria in whites and 11 which passed this criteria in blacks. These *n* polymorphisms are then summed for each individual *j*, giving $GRS_j = \sum_{i=1}^{n} SNP_{ij}$, with each SNP genotype coded as 1 for the risk-conferring homozygote, 0 for the heterozygote, and -1 for the risk-lowering homozygote. This set of GRS values for the sample being investigated was then used as an explanatory variable in a Cox proportional Hazards model, and the authors found it to be significantly associated with the time to developing CHD in both whites and blacks.

The authors then take an important next step, and assess the predictive ability of the GRS above and beyond already existing measurements of CHD risk. They added the GRS to the Atherosclerosis Risk in Communities (ARIC) Score (ARCS), a nomogram developed using the ARIC study to assess an individual's risk of developing CHD, and assessed the area under the receiver operating characteristics curve (AUC) for the Cox model containing only the ARCS and for the Cox model containing both the ACRS and the GRS. The authors found that in whites including the GRS in addition to the ARCS

proved to be only marginally statistically significant, while in blacks the AUC for the ARCS + GRS proved to be statistically significantly larger than the AUC for the ARCS alone. Finally, the authors investigated whether the predictive ability of the GRS was equal over the entire range of the ARCS. When the ARCS values were split into tertiles, Cox modeling in each tertile showed the GRS to be a significant predictor in each tertile. Additionally, Cox modeling showed that there was no significant interaction between the GRS and the ARCS tertiles, indicating the impact of the GRS on predictive ability is consistent across the range of ARCS values and not confined to improving prediction in individuals with extreme ARCS values.

The GRS approach described by Morrison, et al. has a number of advantages. It is simple to implement, deals easily with a potentially large number of polymorphisms, and can be used in conjunction with pre-existing risk assessment methods. However, the method assumes an approximately equivalent effect of each polymorphism, while polymorphisms may in fact have a broad range of effects (Knudsen, et al, 2001). Also, as with the method described by Horne, at al. (2005), there is no model selection procedure in place, which, if there are multiple polymorphisms capturing the same genetic variation, could lead to an over- or under-estimate of risk. Lastly, this method as described assumes complete data is available for each individual. The Morrison, et al. (2007) approach is explicitly described as a proof-of-concept to show the utility of aggregating genetic information to predict risk of incident chronic disease in a longitudinal study. Only a fairly small number of polymorphisms were considered, but the early success from this method

suggests that a GRS can be of substantial utility in identifying individuals at risk of developing CHD.

1.4.2 A More Complex Approach to a Genetic Risk Score

The goal of the GRS described by Horne, et al. (2005) is to predict clinically relevant chronic disease endpoints using features of the underlying biology. Specifically, because of the complexity of the biology of common chronic disease endpoints, the authors work under the assumption that it will be easier to connect genetic polymorphisms with measurable intermediate phenotypes known to influence the outcome of interest. The authors choose this set of intermediate biological phenotypes and then collect a set of genetic polymorphisms thought to influence the intermediate phenotypes.

Each polymorphism is modeled as the explanatory variable in a linear regression of the intermediate phenotype, and polymorphisms that achieve univariate significance are retained. For each retained polymorphism s influencing intermediate phenotype i there is a beta coefficient β_{is} estimated from the linear modeling. Each of the intermediate phenotypes is then used as the explanatory variable in a univariate logistic regression of the outcome, and for each intermediate phenotype i there is a beta coefficient β_i. Then, if n is the number of intermediate phenotypes examined and p_i is the number of polymorphisms retained for intermediate phenotype i, then these can be added, giving the genetic risk score for person j: $GRS_j = \sum_{i=1}^{n} \frac{\beta_i}{|\beta_{\max}|} \sum_{s=1}^{p_i} \beta_{is} * genotype_{isj}$, where β_{max} is given by $max\{|\beta_1|,\ldots,|\beta_n|\}$ and $genotype_{isj}$ is the genotype for the s^{th} polymorphism for the i^{th} intermediate phenotype in person j. The genetic risk score variable is then used as the

explanatory variable in a logistic regression of the clinical endpoint and its performance assessed. When used with large numbers of polymorphisms the GRS will approximate a continuous variable, but when used with a small number only a limited number of values will be possible. In this case, the authors suggest dividing the GRS into several groups and using the resulting categorical variable as the independent variable in the predictive model.

Horne, et al. (2005) illustrate their GRS method using the clinical endpoint of coronary artery disease (CAD), with three intermediate phenotypes, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides (TG). One SNP was genotyped in each of three genes – *CETP*, *ABCA1*, and *HL*. None of these three SNPs significantly predicted CAD in a logistic regression model, even after adjustment for age, sex, hypertension, diabetes, hyperlipidemia, family history of early CAD, tobacco use, and C-reactive protein levels. For the GRS, none of the three SNPs were associated with differences in LDL-C or TG. On the other hand, HDL-C variation was significantly associated with all three SNPs. The resulting GRS_{tot} distribution ranged discontinuously from -2.09 to 0, and there were few predictors, so the authors stratified the sample into five groups based on their GRS_{tot} value. When that strata was used as the explanatory variable in a logistic regression (using strata 4, the largest group, as the reference group) the authors found that groups 1, 2, and 3, which had lower mean values of GRS_{tot} had significantly beneficial odds ratios compared to group 4. When this stratification was collapsed into two group (with groups 1, 2, and 3 forming the "low risk" strata and groups 4 and 5 comprising the "high risk" strata) the significance

remained, with the low risk strata having an odds ratio of 0.77. Although this example is quite small in scale, it offers evidence that polymorphisms not directly associated with a clinical outcome can be used to stratify risk using intermediate phenotypes.

This approach to combining genetic information to make predictions about clinical endpoints offers some advantages. Beginning with clinically relevant endpoints ensures that the prediction that will be made has a direct relationship to disease. Moving from the clinical endpoints to intermediate phenotypes reduces the potential search space to a more manageable level and adds an element of interpretability to the eventual results. The polymorphisms considered for building the predictive model are chosen from genes with prior evidence of involvement with one of the intermediate phenotypes. The resulting predictive model is then grounded in functional knowledge, so even if the precise mechanism by which the polymorphisms act is unknown, as is often the case (Rebbeck, et al, 2004), there is still a body of evidence connecting the polymorphism to the clinical endpoint. At the same time, by restraining the search space in this way, many potentially predictive polymorphisms are excluded either because they lack *a priori* evidence associating them with an intermediate phenotype known to influence the clinical endpoint or because they are linked with an intermediate phenotype whose role in affecting the clinical endpoint is either not recognized or poorly understood. Additionally, because there is no selection method implemented beyond simple univariate significance, the presence of multiple polymorphisms marking the same genetic information might lead to an over-estimate of the amount of risk being conferred. Although this is not the case in

the example discussed by the authors, which focuses on only three polymorphisms, if a larger number of polymorphisms were examined this could quickly become an issue.

1.4.3 An Application of the Genetic Risk Score Concept to Other Data Types

Beer, et al. (2002) describe a risk classification method they term a "risk index" that sums information about risk from a number of different variables and uses that sum to make a qualitative predction about an individual's risk. This approach deals with gene expression values, and so is not strictly a genetic risk score, but it shares many features with other genetic risk scores. This method focuses on classifying lung cancer patients as at high or low risk of death based on gene expression profiling. Using gene expression levels as quantified by Affymetrix oligonucletide expression arrays as predictor variables instead of SNP genotypes, Beer, et al. define a risk index, a linear combination of the 50 most significant predictors of survival as assessed by univariate Cox proportional hazards modeling weighted by their beta coefficients from the Cox model. The final number of predictors used, 50, resulted from empirical testing of models comprised of the top 10, 20, 50, and 75 genes. Fifty was chosen because this model had the highest association with the survival outcome. This risk index is given by the equation $RI_j = \sum_{i=1}^{50} \beta_i * E_{ij}$, where RI_j is the risk index for the j^{th} person, β_i is the Cox proportional hazards model for the i^{th} most significant gene expression variable, and E_{ij} is the value of i^{th} most significant gene expression variable for person j. A value for the risk index is then calculated for a new individual, and they are classified as high or low risk depending on whether their risk index value is above or below the 60^{th} percentile value among the sample initially used to generate the risk index.

The rationale for this risk index approach is simple: when dealing with a small dataset with a large number of potential features to examine ($p >> n$), it is impossible to estimate them all simultaneously. So, to take into consideration the greatest amount of information, the features were treated as independent predictors and used univariately to make predictions. Beer, et al. (2002) found that this procedure stratified their sample of 67 stage I tumors into groups with significantly different survival, with a log rank test p-value of 0.0006 for the comparison between the high and low risk classes. The risk index developed with the original sample was also able to stratify an independent sample of 84 tumors into high- and low-risk groups with significantly different survival times (P=0.003).

The risk index has several advantages in its design. It is straightforward to implement, but examines multiple models to determine which performs best. In the example given it also performs well, and the list of gene expression values used to build the risk index offers some measure of interpretability. This method, however, has a very simple variable selection procedure, considers only a fairly small number of variables, and does not provide a comprehensive assessment of its predictive performance at the group and individual level. The risk index can be improved as a risk prediction tool by expanding the variables being examined to include common clinical variables in addition to genetic polymorphisms, implementing a more sophisticated variable selection procedure, and adding measures of analytical validity.

1.5 Developing, Testing, and Applying the Risk Index to Chronic Disease Risk Prediction

Examining the current GRS methodologies offers several lessons for developing new approaches. First, begin with a clinically relevant endpoint And focus on providing a qualitative assessment of risk (e.g., high risk / low risk), and this type of prediction is most useful in a clinical setting. Second, create methods that can be integrated with other risk assessments both to demonstrate the usefulness of the assessment above and beyond other methods and also to improve overall predictive performance. Third, integrate weightings of genetic effects into GRS methods, in contrast to the uniform weighting found in Morrison, et al. (2007). The weights allow for an assessment of the relative importance of particular polymorphisms or gene expression levels, and account for differences when not all variables have an equivalent impact (Ryall, et al, 1992). Last, it is important to have provisions to deal with missing variable values. It will often be the case in real-world data that an individual will be missing one or more variable values, particularly when a large number have been investigated. Careful consideration of this eventuality can prevent potential biases that may unintentionally arise when individuals with missing variable values have differential contributions to a risk index formed by many variables and end up being systematically assigned higher or lower risk values because of their missing values.

Using existing genetic risk score methodologies for inspiration, this dissertation develops, extends, and tests a risk prediction system based on the risk index methodology described by Beer et al. (2002). This system will make predictions about an individual's level of

risk for developing a particular outcome (e.g., developing diabetes) based on both clinical data and genetic factors. Chapter 2 will describe in detail the development of this statistical methodology. Chapter 3 will describe a simulation study that was used to investigate the method's performance with different scales of data. Lastly, chapter 4 will discuss the application of this method to a dataset from the Framingham Heart Study's Offspring Cohort and it's ability to predict risk of three different chronic disease outcomes: ten-year incident hypertension, ten-year incident diabetes, and prevalent hypertension.

By extending and rigorously testing the genetic risk score methodology, the risk index method developed in this dissertation will provide insight into the potential applications that genetic risk score methods are well suited to. If the risk index methodology is shown to be a prediction method that performs well then it will offer a practical way for clinicians to classify an individual's risk of developing a given chronic disease as high or low. This method could then be deployed using a particular clinic's existing population and used to assess risk in any new individuals entering the clinic, and would allow physicians to assess its performance and determine the appropriateness of the predictions for the population being examined. To assess the performance of the clinical + genetics model developed by the risk index, it's predictive performance will be compared both to a Random Forests model built using the clinical and genomic model and to a risk index model developed using only the available clinical data.

Chapter 2

The Development of a Risk Index Prediction Method

The purpose of this dissertation is to develop and evaluate the performance of a risk prediction method called the risk index. The risk index is composed of the linear combination of the values of a set of covariates weighted by their regression coefficients estimated from univariate models of the covariates predicting the particular outcome of interest. Using a particular risk index model, a risk index value can be calculated for each individual in a dataset. In addition, the risk index can then be used to make predictions about a new individual's risk of developing the outcome (e.g., in an independent testing set). Figure 2-1 shows a graphical overview of the risk index procedure, and the mathematical details are laid out in section 2.1. Choice of performance criteria and assessment methods are laid out in section 2.2, SNP selection methodology is described in section 2.3, and section 2.4 describes the comparison between the risk index methodology and Random Forests (RF), a widely-used machine-learning algorithm that has been used for risk prediction with high-dimensional data.

2.1 The Risk Index

The risk index method requires a sample of n individuals with a vector of outcomes $\boldsymbol{Y}$, where y_j represents the outcome of the j^{th} individual. The implementation of the risk index method used for this project requires a dichotomous outcome (coded as 0 for "no" and 1 for "yes"), although it is straightforward to extended the procedure to consider

either continuous or survival outcomes. The risk index method also requires two matrices of covariates, $\boldsymbol{X_{cov}}$ and $\boldsymbol{X_{SNP}}$, with $\boldsymbol{X_{cov}}$ representing a set of typical clinical covariates (e.g., demographic and anthropometric measurements, biochemical measurements, past medical history) and $\boldsymbol{X_{SNP}}$ representing a set of genotypes, from either a candidate gene study or from a genome-wide association study. $\boldsymbol{X_{cov}}$ is an $n \times v_{cov}$ matrix where v_{cov} is the number of variables in the set of typical clinical covariates and $x_{cov\,kj}$ is the value of the k^{th} variable from the set of typical clinical covariates for the j^{th} person. Similarly, $\boldsymbol{X_{SNP}}$ is an $n \times v_{SNP}$ matrix where v_{SNP} is the number of variables in the set of genotype variables and $x_{SNP\,kj}$ is the value of the k^{th} variable from the set of genotype variables for the j^{th} person.

Risk index models will be built for each of the two sets of variables, clinical covariates and genotype variables. The risk index model for the set of genotypes variables begins with the risk index model that has been created for the set of clinical covariates. The risk index model for the clinical covariates takes the form of $RI_{\mathrm{cov}\ j} = \sum_{l=1}^{l_{\mathrm{cov}}} \beta_{\mathrm{cov}\ l} * x_{\mathrm{cov}\ lj}$, and the risk index model for the genotype variables takes the form $RI_{\mathrm{SNP|cov}\ j} = \sum_{l=1}^{l_{\mathrm{cov}}} \beta_{\mathrm{cov}\ l} * x_{\mathrm{cov}\ lj} + \sum_{l=1}^{l_{\mathrm{SNP}}} \beta_{\mathrm{SNP}\ l} * x_{\mathrm{SNP}\ lj}$. For these models, j is the individual that the risk index value is being calculated for, l_{cov} and l_{SNP} are the number of variables in the risk index model for the clinical covariates and the genotype variables respectively. β_{covl} and β_{SNPl} are the beta coefficient for the l^{th} variable in clinical covariates and the genotype variables respectively, and $x_{cov\,lj}$ and $x_{SNP\,lj}$ is the value of the l^{th} variable in clinical covariates and the genotype variables respectively for the j^{th} person. The final risk index models are denoted RI_{cov} and $RI_{SNP|cov}$, where RI_{cov} represents the model built

using $\boldsymbol{X_{cov}}$, and $RI_{SNP|cov}$ representing the models built with $\boldsymbol{X_{SNP}}$ given RI_{cov}. For mathematical precision, in this chapter the risk index models will be referred to as RI_{cov} and $RI_{SNP|cov}$,. In Chapters 3, 4, and 5 these models will be reffered to as the Clinical and Clinical + Genotype risk index models, respectively.

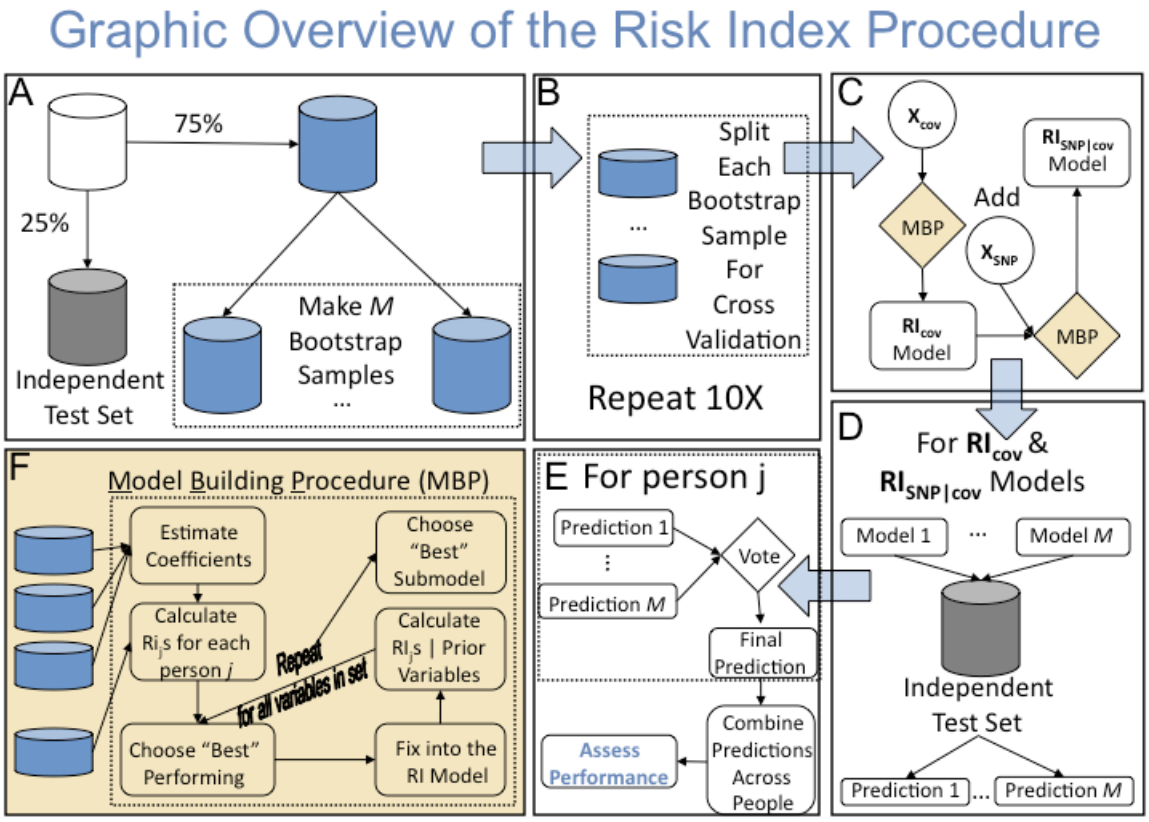


Figure 2-1 A Visual Overview of the Risk Index Method

2.1.1 Logistic Regression

As it is currently implemented, the risk index methodology begins the model building procedure by estimating logistic regression models of the dichotomous outcome being examined with each individual clinical covariate and genotype variable as an explanatory variable. Logistic regression is a form of the generalized linear model (GLM) that is

based on the logit function, $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$ (Hosmer, et al, 2000), where $\frac{p}{1-p}$ represents the odds of the outcome variable being equal to one (ie, the probability that the outcome is equal to one divided by the probability that the outcome is equal to zero). The logit function is then modeled as a linear function of the explanatory variable, $\text{logit}(p) = \alpha + \beta\, x$, or, expressed in terms of probability, $p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$. Estimates of α and β are made by maximizing the likelihood function $l(\beta) = \prod_{j=1}^{n} \zeta(x_j)$, where $\zeta(x_j) = p^{y_j}(1-p)^{1-y_j}$, which takes the value of p when y_j equals 1 and (1-p) when y_j equals 0.

For categorical explanatory variables, one category is chosen as the reference group and the term e^{β_l} is interpreted as the odds ratio of the two groups (the odds of a person in that category having the outcome divided by the odds of a person in the reference category having the outcome). For continuous explanatory variables, the reference group can be thought of as a person who has a value of zero for the variable, and a person's odds of developing the outcome given the variable is the beta coefficient multiplied by the person's value for that variable, $e^{\beta_l x_{lj}}$.

2.1.2 Dividing the Data

Developing and applying the risk index requires the ability to reliably track the performance of risk index models during the model building process and to estimate the final model's performance on a real-world dataset. To accomplish this, the original

sample of n individuals will be split into two parts, as shown in part A of Figure 2-1. One part will consist of a random twenty-five percent of the sample that is set aside as an "independent test set". This data will be used after the final steps of model building to assess the performance of the risk index models. The remaining seventy-five percent of the data will comprise the "optimization set" and be used to construct the risk index models. This portion of the data will be further subdivided into four equal parts for four-fold cross-validation. A visual representation of this is given in part B of Figure 2-1. Division into cross-validation sets will be repeated ten times to account for random fluctuations in performance due to the stochastic process of data division (Molinaro, et al, 2005). Each part will be used as a "testing set" once, while the remaining three parts will be used as the "training set". The training set will be used to estimate the logistic regression coefficients, and a prediction will be made about whether each person in the testing set will develop the outcome or not. These predictions will then be used to calculate sensitivity, specificity, and other performance characteristics by comparing them to the actual individual outcomes that were observed, as described below.

2.1.3 Building the Risk Index

Building the risk index begins with the set of typical clinical covariates, $\boldsymbol{X_{cov}}$. Part F of Figure 2-1 displays the model building procedure described here in a graphical manner. Each variable $x_{cov\,k}$ is used as the explanatory variable in a univariate logistic regression on Y, and regression coefficients are estimated using the training set. These coefficient estimates are stored and will be used throughout the building of the risk index model for $\boldsymbol{X_{cov}}$. Each of the v_{cov} coefficients are then used to form a one-variable risk index model,

$RI_{\text{cov } j} = \beta_{\text{cov } k} * x_{\text{cov } kj}$, with the associated covariate and calculate the risk index values for people in the training set and testing set. The single best performing (as described in Section 2.2) of the v_{cov} one-variable models is selected, and that variable (denoted variable "1"), is fixed into the risk index model. A new set of v_{cov}-*1* risk index models are then constructed using variable "1" and one of the remaining variables: $RI_{\text{cov } j} = \beta_{\text{cov } 1} * x_{\text{cov } 1j} + \beta_{\text{cov } k} * x_{\text{cov } kj}$. The best performing is chosen, the new variable, "2", is fixed into the model, and the process is repeated until either all variables of the covariate set $\boldsymbol{X_{cov}}$ have been incorporated into the model or some maximum limit (e.g., ten variables) has been reached. This yields a set of successively nested models, with each risk index model containing one more variable than the previous risk index model. The best performing (again, as described in Section 2.2) of these nested models is then selected, giving $RI_{\text{cov } j} = \sum_{l=1}^{l_{\text{cov}}} \beta_{\text{cov } l} * x_{\text{cov } lj}$,, where each variable $x_{cov\,l}$ is in the final, best-performing model, and l_{cov} is the total number of variables in that model. The risk index building procedure is then repeated with the set of covariates $\boldsymbol{X_{SNP}}$, using the risk index model developed for covariate set $\boldsymbol{X_{cov}}$ as a base and successively adding the variable that improves the performance of the model by the greatest amount. This yields

$$RI_{SNP\,j} = \sum_{l=1}^{l_{\text{cov}}} \beta_{\text{cov } l} * x_{\text{cov } lj} + \sum_{l=1}^{l_{SNP}} \beta_{SNP\,l} * x_{SNP\,lj}\,.$$

2.1.4 Accounting for Missing Values

To account for the fact that some individuals may be missing values for covariates that are used in risk index models, an additional term is used to adjust individual risk index values by the amount of available covariate variables. For $\boldsymbol{X_{cov}}$, the value $RI_{cov\,j}$ for the j^{th}

person is divided by $n_{cov\,j}$, the number of clinical covariates for which person j has non-missing values, and for $\boldsymbol{X_{SNP}}$ the value $RI_{SNP\,j}$ for the j^{th} person is divided by $n_{SNP\,j}$, the number of genotype variables for which person j has non-missing values,. This makes the risk index model for clinical covariates $RI_{\text{cov}\,j} = \dfrac{\sum_{l=1}^{l_{\text{cov}}} \beta_{\text{cov}\,l} * x_{\text{cov}\,lj}}{n_{\text{cov}\,j}}$, and the risk index model for genotype variables $RI_{SNP\,j} = \dfrac{\sum_{l=1}^{l_{\text{cov}}} \beta_{\text{cov}\,l} * x_{\text{cov}\,lj}}{n_{\text{cov}\,j}} + \dfrac{\sum_{l=1}^{l_{SNP}} \beta_{SNP\,l} * x_{SNP\,lj}}{n_{SNP\,j}}$.

2.1.5 Making Predictions

The purpose of the risk index methodology is to make predictions about the disease risk of individuals. In order to turn the risk index values into a discrete (yes/no) prediction, a cut-point is selected during the model building procedure (shown in part F of Figure 2-1) using the following algorithm:

1) One of the ten cross-validation iterations is selected
2) For one of the four training set / testing set pairs in that iteration is selected
3) For a given risk index model, risk index values are calculated for each person in the training and testing set
4) The value at the p^{th} percentile of the training set risk index values distribution, c^*, is obtained
5) All individuals in the testing set with an risk index value less than c^* are assigned a prediction of 0 or “low risk of developing the outcome”
6) All individuals in the testing set with an risk index value greater than c^* are assigned a prediction of 1 or “high risk of developing the outcome”

7) Steps 3-6 are performed for the remaining three training set / testing set pairs in the selected cross-validation iteration
8) The performance of the predictions made on the individuals in the testing sets is assessed (as described in Section 2.2)
9) Steps 1-8 are performed for the remaining nine iterations of cross-validation samples
10) Performance of the ten iterations are averaged
11) Steps 1-10 are performed for a range of values of *p*
12) The percentile with the highest performance (as described in section 2.2) is chosen as the cutpoint

This process is encapsulated in part F of Figure 2-1 as "Choosing the 'Best' Performing Model".

2.1.6 Ensemble Prediction

The predictive performance of the risk index building procedure described above depends heavily on the optimization set's split into cross-validation sets. To provide a more stable prediction an additional step inspired by ensemble prediction methods will be performed (Optiz and Maclin, 1999). To begin, a bootstrap sample *S* of the optimization set will be generated. The risk index procedure will be used to generate a risk index model using the clinical covariates and a risk index model using both the clinical covariates and genotype variables. These models will be used to make a high risk/low risk prediction about the individuals in the independent test set that was set aside . The bootstrap procedure will be repeated *W* times, and each person in the independent testing set will be assigned a final high risk prediction if more than (*W*/2)+1 of the models predicted the individual to be

high risk. If more than (W/2)+1 of the models predicted the individual to be low risk that person will be assigned a final prediction of low risk.

2.2 Assessing Performance

Once predictions have been made about individuals in the testing set, they can be compared to the observed values of the outcome for each individual. Table 1 shows the appropriate way to classify individual predictions. From these predictions, performance metrics can be calculated as shown in Table 2.

Table 2-1 Description of a 2x2 Table

	True "Yes"	True "No"
Predicted "Yes"	**T**rue **P**ositive	**F**alse **P**ositive
Predicted "No"	**F**alse **N**egative	**T**rue **N**egative

Table 2-2 Calculation of Performance Metrics

Name	Abbrev.	Formula
Sensitivity	Sen	TP/(TP + FN)
Specificity	Spe.	TN/(TN + FP)
Accuracy	Acc.	(TP + TN)/(TP + TN + FP + FN)
Positive Predictive Value	PPV	TP/(TP + FP)

These performance metrics will be calculated once for each of the ten four-fold cross-validations. The four-fold cross-validation will yield one prediction for each individual, and the values for all individuals will be used to calculate sensitivity, specificity, misclassification, and positive predictive value. These performance metrics will then be averaged across all ten iterations of four-fold cross-validation. Next, sensitivity, specificity, misclassification, and positive predictive value will be calculated for the

independent testing set. Finally a receiver operating characteristics (ROC) curve will be generated and the area under the ROC curve (AUC) will be estimated.

2.2.1 Model Building Optimization Function

Part F of Figure 2-1 provides a visual overview of the model building procedure and indicates that an optimal or "best" new variable is chosen during the actual model building, and an optimal submodel is chosen as the final step in the model building procedure. An additional metric is needed to make these comparisons between sets of predictive models during the model building procedure. Although a number of potential scores have been developed for this purpose, no single score is appropriate for all contexts. The risk index as it will be applied in this context will focus on future disease-state prediction, and so an appropriate performance metric should compare the predicted disease probabilities that are assigned to each person to their eventual outcome. The Brier score, or quadratic score, is a model score that does this and will be used for model comparison and optimization. The Brier score begins with maximum time in which to have the outcome, t^*, an indicator function $I(T_i < t^*)$, and an individual's predicted probability of developing that outcome, $\hat{\pi}(t^* \mid \tilde{X})$. The indicator function $I(T_i<t^*)$ takes the value of 0 if individual i's time to th outcome (T_i) is greater than t* and takes the value of 1 if individual i's time to develop the outcome is less than t*. $\hat{\pi}(t^* \mid \tilde{X})$ is referred to at the "strata-specific outcome probability", which is the probability that a any person in the given risk strata (high or low) that individual i is assigned to will develop the outcome by t^*. This is calculated for a given risk strata by adding the number of individuals individual in the strata who develop the event by t^* and dividing by the total

number of individuals in that strata. The score itself is given by

$$Brier\ Score = \frac{1}{n}\sum_{i=1}^{n}\left(I(T_i < t^*) - \hat{\pi}(t^* \mid \tilde{X})\right)^2$$ and is the average of the squared difference between an individual's outcome status and the probability of a person in the same risk strata developing the outcome by t^*.

2.2.2 Calculating Strata-Specific Outcome Probabilities

An important component of the Brier score is the strata-specific outcome probabilities. For the model building process, performed in the optimization set, the Brier score will be estimated on the testing sets of a given cross-validation iteration and the averaged across all ten iterations. For a particular testing set in a given cross-validation iteration, the strata-specific outcome probabilities will be estimated in the associated training set. This process is illustrated in Figure 2-2. For the cutpoint being investigated, each individual in the training set will be assigned to either the high or low risk strata. For each of the strata, then, the number of individuals who develop the outcome by t^* will be divided by the total number of individuals assigned to that strata. This proportion will be the probability of a person in that risk strata developing the outcome by t^*. For the final model assessment process, performed in the independent test set, the strata-specific outcome probabilities will be estimated in the same manner, but will be done with the optimization set.

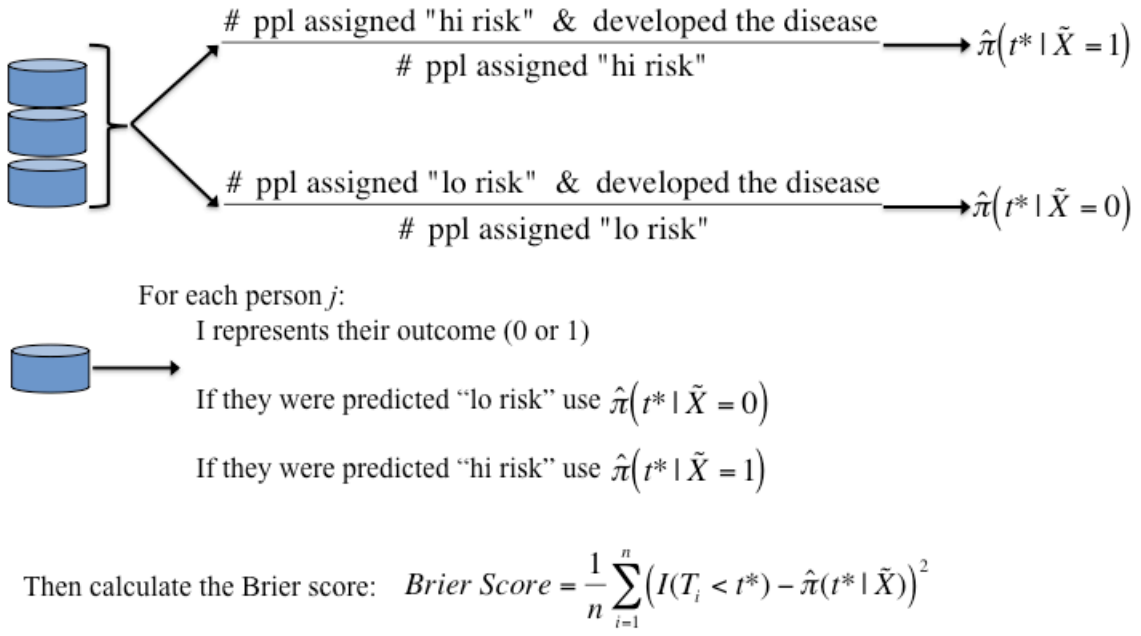


Figure 2-2 An Overview of the Assignment of Strata-Specific Probabilities for Use in the Brier Score

2.2.3 Estimating Individual Predicted Probabilities of Disease

The bootstrapping procedure used to generate the final predictions for each individual provides a simple measure of that individual's predicted probability of disease. The final prediction for an individual is taken to be the prediction which was most often given by the ensemble procedure. Analogously, the predicted probability of disease for the individual is simply the number of bootstrap samples which predicted an individuals would develop the disease being examined divided by the total number of bootstrap samples. Using the binomial distribution a 95% confidence interval we might be able to

construct an estimate of the predicted probability of disease with the Wilson score interval (Wilson, 1927). This interval, given by

$$95\%\ \text{CI} = \frac{\hat{p} + \frac{(1.96)^2}{2n} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{(1.96)^2}{4n^2}}}{1 + \frac{(1.96)^2}{n}}$$, where $\hat{p}$ is the predicted probability of disease and n is the number of bootstrap samples, might be useful physicians to gauge the prediction. We should note that the n bootstraps are not independent as this equation assumes.

2.2.4 Bootstrap Estimate of Model Performance

Beyond individual prediction confidence, it is also important to be able to compare the performance of two different risk index models created with the same data. The performance of the two risk index models is estimated on the independent test set; however, this point estimate is based on a random division of the data, and a bootstrap procedure will provide a confidence interval that can be used for comparisons. To begin, 1000 bootstrap samples will be drawn from the independent test set, denoted $n^*_{ind1},\ldots,n^*_{ind1000}$. Both the risk index models R_{cov} and R_{SNP} will then be applied to each of the 1000 bootstrapped independent test samples, and sensitivity, specificity, misclassification, positive predictive value, and model score will be calculated. For each risk index model the 1000 values for each of these measurements will be sorted, and the lower bound of the 95% confidence interval will be the value at the 2.5th percentile and the upper bound will be the value at the 97.5th percentile.

For each of these performance measures (sensitivity, specificity, misclassification, and positive predictive value), the standard error and the bias can be estimated. For each n_i, there will be 1000 estimates of each of the performance measures. The standard error of the estimates is given by $\hat{\sigma}_B = \sqrt{\frac{\sum_{b=1}^{B}(s_b - \bar{s}_B)^2}{B-1}}$, where B is the total number of bootstrap samples generated (1000 in this case), $\bar{s}_B$ is the value of a particular performance measure averaged over all of the B bootstraps, and s_b is the particular performance measure from the b^{th} bootstrap of the independent test set. The bias, or average deviation of the performance measure in a bootstrap sample from the estimate in the independent test set, is given by $Bias = \frac{\sum_{b=1}^{B}(s_b - s)}{B}$, where s is a particular performance estimate for the independent test set, s_b is the estimate for the same performance measure for the b^{th} bootstrap sample of the independent test set, and B is the total number of bootstrap samples generated.

2.3 SNP Selection

The search through the set of SNPs is exhaustive, and if no limit is placed on the number of SNPs that may be fixed into the risk index model, for g SNPs, the risk index procedure must evaluate $\sum_{h=1}^{g} g - (h-1)$ models, which can be alternatively expressed $g^2 - \frac{g(g-1)}{2}$. If the number of SNPs that may be added to the risk index model is limited to some smaller number, q, then the risk index model building procedure must evaluate $q * g - \frac{q(q-1)}{2}$

models. For a set of 500,000 SNPs and a limit of at most twenty SNPs allowed into the risk index model, there are $20 * 500{,}000 - \frac{20(19)}{2}$, or 9,999,810 models to be examined by the risk index model building procedure. In order to keep the number of models to be examined reasonable, two different SNP selection approaches will be tested. First, each available SNP will be tested for association with the outcome of interest in the entire optimization set. The 1% of SNPs most highly associated with the outcome, as measured by p-value, will be then be used for the risk index procedure because these SNPs are most likely to be strong predictors of the outcome. Second, a principal components analysis (PCA) approach will be used to reduce the dimensionality of the entire set of genotype data.

2.3.1 Principal Components Analysis

Principal Components Analysis (PCA) is a dimensionality reduction tool that is commonly used in many fields of study (Jackson, 2003). PCA computes the eigenvectors and eigenvalues of a p x p covariance or correlation matrix where p is the number of variables being examined. Each of the principal components is orthogonal to all others (i.e., there is no overlap in the variance they explain) and explains a proportion of the total variance of the sample. In this dissertation the PCA will be performed exclusively on SNP genotypes. This means that "variance" here refers to population level genomic variability in the DNA sequence. Principal components are ordered, with the first component explaining the most variance and each successive component explaining successively less of the variance. Because of these features, new variables (i.e., the principal components) can be constructed such that a significant portion of the variance

can be explained with relatively small number of variables, as opposed to the potentially tens or hundreds of thousands of variables used to construct the set of principal components, and each of the new variables explains a discrete subset of the variance in the sample. PCA has been used to correct for population structure in genome-wide association studies with programs such as SMARTPCA (part of EIGENSTRAT) (Patterson, et al, 2006), to precisely determine an individual's population of origin (Price, et al, 2008) and to identify the relationships between subpopulations (Seldin, et al, 2006). The same benefits that make PCA well-suited to population structure analysis, particularly its ability to reduce the number of variables being examined while representing a significant amount of the genetic variation contained in the original set of genotypes, also make PCA a good choice for reducing the dimensionality of SNP data in this project.

The principal components procedure begins with an $n \times p$ set of data, where n is the number of individuals in the dataset and p is the number of variables. A covariance matrix $\mathbf{S}$ is then calculated, where s_{ij} is the covariance of the i^{th} and j^{th} variable if $i \neq j$ and the variance of the i^{th} if $i = j$ (Jackson, 2003). The set of p eigenvalues, λ, for the matrix $\mathbf{S}$ are found by taking the determinant $det(\mathbf{S} - \lambda\mathbf{I})$, where $\mathbf{I}$ is the identity matrix, and solving the resulting p^{th} order polynomial. The eigenvectors u_i are calculated by solving $[\mathbf{S} - \lambda\mathbf{I}]\mathbf{t}_i = 0$ for $\mathbf{t}_i$ and normalizing $\mathbf{t}_i$, $u_i = \frac{t_i}{\sqrt{t_i' t_i}}$. The set of eigenvalues and eigenvectors is then used to create a new set of uncorrelated variables, the principal components. This is done for each of the n subjects by multiplying each of the p eigenvectors by the standardized value of each variable, $z_i = u_i'[x - \bar{x}]$. These new variables, or some subset

of them, can then be used in place of the original variables. For this project, the complete set of SNP genotypes have been decomposed into its principal components (PCs). The PCs that account for 90% of the variation in the sample are then used as the genotype variables for the risk index procedure. Additionally, the set of SNPs most highly associated with the outcome being investigated have been decomposed into its PCs, and the PCs that account for 90% of the variation in the sample are used as the genotype variables for the risk index procedure.

2.4 Performance Comparison

The development of a novel risk prediction algorithm requires a standard against which to judge its performance. An appropriate comparison method must have three essential characteristics: 1) the goals and underlying assumptions of the methods are similar, 2) the methods are capable of dealing with the same types of data, and 3) the performance metrics produced are comparable. This project will compare the performance of the risk index methodology to that of random forests (RF). Developed by Leo Brieman, one of the developers of the classification and regression trees (CART) methodology (Breiman, et al, 1984), RF is a modification of a decision tree algorithm that is capable of dealing with very high-dimensional data (Breiman, 1996).

2.4.1 Classification and Regression Trees

CART is a specific implementation of a decision tree algorithm. Given a dataset of samples with class labels, CART creates a binary tree by splitting the samples at each node in such a way that makes the two new nodes as pure as possible (i.e., they contain as

close to exactly one class as possible) (Breiman, et al, 1984). Once the specified stopping criterion has been reached, class labels can be assigned to each of the terminal nodes based on the classes of the samples in the node. New samples, with unknown class labels, can be placed in the tree, and by following the splitting rules, these samples can be assigned the class label associated with the terminal node they reach. In the following sections the statistical and analytical strategies underlying CART are outlined in detail using the same nomenclature and terminology as is used in Brieman, et al. (1984).

2.4.2 Splitting Nodes

Four things are needed to create the tree: a set of binary questions used to split the nodes, a measure to assess the goodness of a specific split, a stop-splitting rule, and a method to assign class labels to the terminal nodes (Breiman, et al, 1984). The most straight-forward of these requirements is the set of binary questions with which to split the nodes. Breiman, et al. propose a set of standard questions that encompasses all possible splits of a node using a single variable (Breiman, et al, 1984). For each continuous variable x, the questions are of the form "Is $x \leq c$?" for all c ranging over $(-\infty,\infty)$, and for each categorical variable x that takes class labels $(c_1, \ldots,c_n)$, the questions are of the form "Is $x \in S$?", as S ranges over all subsets of $(c_1, \ldots,c_n)$.

2.4.3 Measurement of Impurity

The goal of tree growing is to create a tree in which the terminal nodes hold samples of exactly one class. Due to limitations in CART implementations this result is not always achievable in practice, however it suggests a logical choice for a measure of the goodness

of a split *s* of node *t*: the impurity of the resulting left and right nodes t_L and t_R, respectively (Breiman, et al, 1984). A good measure of impurity *i(t)* should be maximal when each class is equally represented in a single node and equal to zero when exactly one class is represented. The optimal split *s* derived from the set of questions described above is the split that maximizes the decrease in impurity,

$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$, where p_L is the proportion of people in node *t* that are in node t_L and p_R is the proportion of people in node *t* that are in node t_R. Breiman, et al. (1984) suggest two possible impurity measures, the Gini criterion and the Twoing criterion. The Gini criterion is straightforward, and defines impurity as

$i(t) = \sum_{j \neq i} p(j \mid t) p(i \mid t)$. In the case of a two class problem, this reduces to

$i(t) = p(1 \mid t) p(2 \mid t)$, which the authors describe as the appropriate impurity function for two class problems. The Twoing criterion is more complex and begins with grouping all class labels $C = (1,...,J)$ into two super classes $C_1 = (j_1,\ldots,j_n)$ and $C_2 = C - C_1$. Splitting then proceeds as if it were a two class problem. The optimal split, then, depends on the choice of C_1 and C_2, and so all groupings of C_1 and C_2 and all potential splits *s* within those groupings are considered. This is a much more computationally intensive intense procedure than the Gini criterion; however, it has the advantage of grouping similar class labels together. The final tree does not depend heavily on the choice of criteria, but the Gini criterion is the preferred method due to better performance in some instances.

2.4.4 Stop Splitting Criteria

Choosing when to stop growing the tree has proven to be a difficult problem. Although common sense suggests growing the tree until the decrease in impurity falls below some

threshold, in practice that threshold is often either too high, resulting in trees which are too small, or too low, resulting in very large trees that have poor predictive ability (Breiman, et al, 1984). Minimizing the misclassification error rate is one potential solution; however, this measure continually decreases as tree size increases and leads to over-estimation of a large tree's predictive ability. A procedure for both finding the optimal size of a tree and for giving an accurate estimate of the real-world misclassification rate is therefore necessary. Breiman, et al. propose pruning as the best way for reducing tree size and using the estimated misclassification error rate as the criterion for choosing the optimally sized tree. Pruning proceeds recursively, beginning with the largest tree (i.e., the tree that was grown to completion) and identifying a weakest-link subtree that, for some cost penalty α, has a cost-complexity

$R_\alpha(T_T) = R(T_T) + \alpha\left|\tilde{T}_T\right|$ (where $\left|\tilde{T}_T\right|$ is the number of terminal nodes in the tree and $R(T_T)$ is the misclassification error rate) that is equal for the subtree and its root node. The appropriate α is found by finding the node that minimizes $g_1(t) = \begin{cases} \dfrac{R(t) - R(T_t)}{\left|\tilde{T}_t\right| - 1}, & t \notin \tilde{T}_1 \\ +\infty, & t \in \tilde{T}_1 \end{cases}$,

where $\tilde{T}_1$ is the set of terminal nodes in the full tree and $\left|\tilde{T}_t\right|$ is the number of terminal nodes in subtree *t*. Once the weakest-link subtree has been removed, the procedure is repeated until only the root node of tree *T* remains. This yields a nested set of subtrees ranging from a single root node to the full tree created during the growing phase. An estimate of the true misclassification error can then be used to identify the optimal pruned tree. This estimate can be obtained either through the use of a test sample (if the original

sample size is sufficiently large) or by cross-validation. For the test sample method, some proportion of the original samples is left out, a tree is constructed and pruned, and classes in the test sample are obtained with each of the trees. Misclassification rates can then be calculated for each of the trees, and the tree with the lowest rate is selected as the optimal tree. Cross-validation is more complex but also more commonly used, owing to sample size constraints. Briefly, the whole sample is divided into *n* approximately evenly sized groups. As the tree is grown with the full sample, *n* auxiliary trees are grown with each of the possible sets of *n-1* groups (so each group is used to construct *n-1* trees and is left out of 1 tree). Each of the auxiliary trees are pruned using the α that was used for the full tree, and the misclassification error can be estimated by putting each of the *n* groups down the auxiliary tree that it was not used to create. If *n* is sufficiently large, this approximates the misclassification error obtained by putting an independent test set down the full tree, and because there is a misclassification error estimate associated with each of the pruned subtrees, it can be used to identify the optimal subtree.

2.4.5 Assigning Class Labels

There are multiple approaches for assigning class labels to each of the terminal nodes of the tree. The approach preferred by Brieman et al. is to have each sample in a terminal node "vote" on the class label, with the class that is most highly represented being set as the label for that node(Breiman, et al, 1984). If two or more classes are equally represented in a terminal node, they suggest choosing a class label at random from among the most represented classes.

CART has become a very widely used classification method because it flexible, computationally efficient, and creates a tree that is easily interpretable. CART performs well in cases where linear methods such as logistic regression do poorly (Brieman, et al. 1984). However, large numbers of variables can cause a significant increase in the time required to create the tree. For this reason, CART is best suited to data types with fewer variables, such as demographic, clinical, environmental and social, and biochemical variables, or to data types where feature selection has been applied, reducing the number of variables to be considered.

2.4.6 Random Forests

Random Forests (RF), however, an ensemble learning algorithm based on CART, is not limited by the number of variables contained in the dataset. In its most basic form, a large number of trees are grown (e.g.,100 or 1000). Each tree is grown with a bootstrap sample from the original data (i.e., a sample of the same size as the original is chosen, with replacement, from the original sample) and the best split at each node is chosen from among some number of randomly selected input variables. An unknown sample can then be run down each of the trees, and the class label that is selected by the most classifiers can be applied (Breiman, 1996).

With the RF method, trees are grown as in CART but are not pruned. Breiman states that the ensemble nature of the classifier places a limiting value on the generalization error (i.e., the error expected when the classifier is presented with a novel set of samples), removing the need for tree pruning (Breiman, 1996). At each node, some number F of

input variables are selected, and the best possible split among those variables is chosen. Breiman describes two different values for F, $F = 1$ and $F = floor(\log_2(M + 1))$ (where M is the total number of input variables available), but the performance of the algorithm does not heavily depend on F. Error estimation, a problem that is solved only with cross-validation or an independent test set in CART, can be performed using an "out-of-bag" method. When each bootstrap sample is created, approximately one-third of the cases are held out. Then, once all of the trees have been constructed, each case can be run down the one-third of trees it was not used to create. The proportion of incorrect classifications can then be averaged over all cases, giving an estimate of the misclassification error rate. Because only one-third of all trees were used for the classification, this is actually an overestimate of the misclassification error rate, but it is unbiased; thus, it will approach the true misclassification error rate as sample size increases. Unlike CART, RF does not lend itself to simple interpretation. In its raw form it is a "black box", but assessing variable importance can allow insight into the relationships that the RF is modeling. To measure importance, trees are constructed as described above. Then, for all out-of-bag samples for a given tree, the n^{th} variable is permuted. Misclassification error is estimated, and the percent difference between the misclassification error rate estimated with and without the permutation of the n^{th} variable is that variable's measure of importance for that tree. Importance measurements for each variable can then be averaged across all trees, and by standardizing with the standard error of the importance, a z-score for the importance can be obtained and its significance assessed.

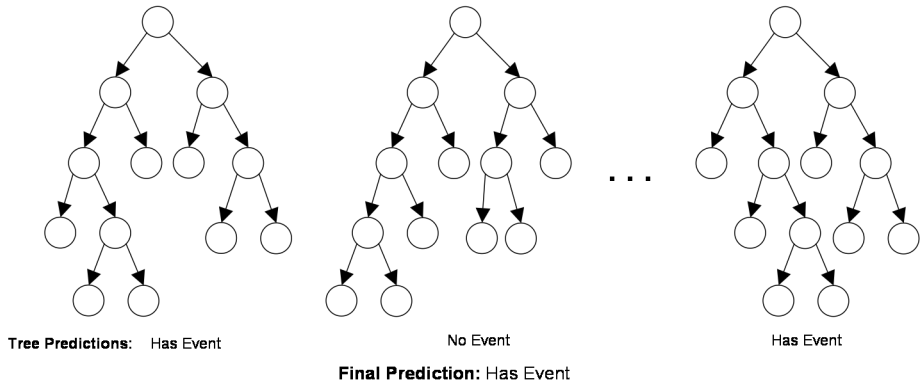


Figure 2-3 An Example of a Random Forest and Its Prediction About a New Individual

RF excels at classification in situations in which a large number of variables and a large number of cases are available (Breiman, 1996). Its performance is excellent even when all available variables have very weak predictive power, a condition that is common when considering genomic features such as SNP genotypes but that prevents other classification algorithms from performing well.

2.4.7 Comparison Methodology

RF's ability to deal with a large number of variables makes it an excellent choice for comparison to the risk index, although the comparison procedure must be done so as to ensure that the comparison is as valid as possible. First, the optimization set used in the construction of the risk index will be used to build a forest of 500 trees. An optimization procedure will be used to identify the number of variables to be examined at each split, *k*. Initially the forest will be generated with $k = \sqrt{v}$, where *v* is the total number of variables in the dataset, and the out-of-bag (OOB) error will be estimated. Then *k* will be increased from this initial value by factors of 2 until the OOB improves by less than 5%. Finally, *k*

will be decreased from its initial value by factors of 2 until the OOB no longer improves by 5% or more. The value of k which provides the lowest OOB will be used.

The true comparison of the methods is their predictive performance on the independent test set and, as with the risk index, a bootstrapping procedure will be used to provide a confidence interval for the forest's performance. As before, 1000 bootstrap samples of the independent test set will be taken, and the forest will be applied to each.

As a default, the final prediction from the random forest is given by a simple voting procedure. If a majority of trees predict an individual is at high risk, then that individual is assigned a prediction of "high risk". If not, the individual is assigned a prediction of "low risk". To fully examine the potential performance of the random forests, predictions will be made using a number of voting cutoffs. To begin, all individual for whom 5% or more of the trees in the forest predict to be at high risk will be assigned a prediction of "high risk", and this cutoff will be increased until only individuals who 95% or more of trees in the forest predict as high risk will be assigned a prediction of "high risk".

Once predictions have been made for each individual, the sensitivity, specificity, misclassification rate, and positive predictive value will be calculated for each of the bootstrap samples. These values will then be sorted and the 95% confidence interval for each of these measurements will be the values at the 2.5^{th} percentile and the 97.5^{th} percentile. The bootstrapped performance estimates of the two techniques will then be compared.

Chapter 3

Simulation Study to Characterize the Performance of the Risk Index

While the final assessment of the utility of any risk prediction algorithm should be rooted in its ability to accurately make predictions about real datasets, the use of simulation datasets is an important step in the characterization of performance. Datasets obtained from real-world studies have complex structures that make it impossible to say definitively that a particular variable is or is not related to the outcome being studied. In a simulation study, however, the correlation structure of the data can be directly specified, giving a dataset where the relationship between each variable and the outcome is known.

Applying the risk index methodology to simulated datasets allows for the systematic investigation of its functioning and performance. By having fine-grained control over the precise structure of the dataset being tested, it is possible to address important questions about the risk index methodology's sensitivity to noise variables, ability to account for correlation among predictor variables, and performance with varying numbers of predictors. The simulation study carried out here examines two different scenarios that the risk index might be expected to handle. First, a small-scale study was simulated consisting of one thousand people with a small number of standard covariates (eight) and a moderate number of polymorphisms (five hundred). This small-scale simulation is patterned after a candidate gene study, where the polymorphisms investigated are not evenly spread throughout the genome, but rather in selected regions thought to be

involved in the disease process being studied. Second, a large-scale study was simulated, with ten thousand individuals, thirty covariates, and approximately forty thousand polymorphisms. This large-scale simulation is patterned after a genome-wide association study, with polymorphisms chosen to reflect the frequencies observed on Chromosome one in the Affymetrix 500K genome-wide genotyping system.

3.1 Small-scale Simulation Study Methodology

The small-scale simulation study consists of one thousand people, eight covariates, and five hundred polymorphisms. The outcome is a dichotomous variable with a 30% prevalence and is simulated as a continuously distributed normal random variable with a mean of 120 and a standard deviation of 40. Approximately 30% of the individuals will have values for the outcome of >140, and this cutoff is used to convert the continuous outcome into the dichotomous outcome.

The covariate and outcome simulation are generated with a multivariate normal random number generator. The correlation matrix between the covariates and the continuous outcome is specified, with the first variable having a 0.56 correlation with the outcome, the second a 0.41 correlation with the outcome, the third a 0.50 correlation with the outcome and an 0.80 correlation with variable one, and the fourth a 0.28 correlation with the outcome. The remaining four variables will be noise variables, with no correlation with the outcome or any of the explanatory variables. Though this simulation is simplistic, it allows the examination of the behavior of the risk index procedure in the presence of highly correlated explanatory variables. In particular, by including correlation

between the first and third variable, it is possible to see the effect of that correlation on the variable selection procedure.

Genotype simulation was performed using the genomeSIMLA program (Edwards, et al, 2008). Genotypes for five hundred SNPs were generated for each individual in the simulation dataset. All of the SNPs were considered independent from every other SNP and had no pairwise linkage disequilibrium. Four polymorphisms were set as associated with the outcome, each with a beta coefficient of between 0.4 and 0.8, corresponding to an odds ratio for a given locus of between 1.5 and 2.2. Genotypes were coded additively, with a value of 0 representing an individual homozygous for the major allele, a value of 1 representing a heterozygous individual, and a value of 2 representing an individual homozygous for the minor allele.

3.2 Small-scale Simulation Study Complete SNP Set Results

3.2.1 Variable Selection

Using the methodology outlined above, 100 small-scale simulation datasets were generated, each with 1000 individuals, eight covariates, and 500 polymorphisms. The risk index procedure was then applied to each of the 100 simulation datasets. The datasets were divided into an independent testing set of 250 individuals and an optimization set of 750 individuals. One hundred bootstrap samples of the optimization set were generated, and the risk index procedure was used to generate Clinical and Clinical + Genotype risk index models for each of the bootstrap samples. Each of these models was then used to make a prediction (high risk or low risk for developing hypertension) about each of the

250 individuals in the independent testing set. For both the Clinical risk index model and the Clinical + Genotype risk index model the predictions from each of the 100 bootstrap samples were used as votes, and the prediction most frequently assigned was designated as the consensus prediction. For each of the 250 individuals in the independent testing set there was one prediction for Clinical risk index model and one prediction for the Clinical + Genotype prediction.

Table 3-1 shows the summary of the variable selection procedure from the Clinical risk index model averaged across the 100 simulation datasets. Variables one, two, and four are most frequently selected; on average, they each appear in more than half of the 100 trimmed Clinical risk index models (a "trimmed model" here refers to a risk index model that has been grown to its maximum size and had the optimal submodel chosen). Variable three, because of its high correlation with variable one, is typically chosen as one of the last variables (on average, variable three is chosen as the sixth, seventh, or eighth variable in 71.05 of the 100 untrimmed Clinical risk index models for a given simulation dataset). However, each trimmed Clinical risk index models contained, on average, 3.83 variables, so variable 3, because it is typically selected into position six, seven, or eight, appears in only an average of 26.58 of the 100 trimmed Clinical risk index models.

Table 3-2 shows the summary of the variable selection procedure from the Clinical + Genotype risk index model averaged across the 100 simulation datasets. No SNP was in more than 5.76 out of 100 Clinical + Genotype risk index models on average. The SNPs most commonly observed in trimmed Clinical + Genotype risk index models were SNP

180 (5.76 out of 100 trimmed Clinical + Genotype risk index models, on average), SNP 168 (5.28 out of 100 trimmed Clinical + Genotype risk index models, on average), SNP 425 (5.27 out of 100 trimmed Clinical + Genotype risk index models, on average), SNP 411 (5.06 out of 100 trimmed Clinical Genotype risk index models, on average), and SNP 73 (5.05 out of 100 trimmed Clinical + Genotype risk index models, on average. The genotype simulation specified 4 SNPs as being associated with the outcome, (i.e., SNP 1, SNP 10, SNP 50, and SNP 100). Although these polymorphisms were, when selected, often chosen as the first variable in the Clinical + Genotype risk index model, the Small-scale Simulation only included these variables in trimmed Clinical + Genotype risk index models in between 1.25 and 1.76 out of 100 times, on average.

Table 3-1 Summary of the Number of Times Each Variable is Selected into a Specific Model Position for the Small-scale Simulation Clinical Risk Index Models

	Variable Position								Total # of Times in Trimmed
Variable	1	2	3	4	5	6	7	8	Model
v1	42.59	6.17	2.6	3.11	3.35	6.28	14.76	21.14	56.15
v2	8.51	39	9.64	4.24	7.79	16.08	4.41	10.33	62.7
v3	12.75	3.76	3.18	3.92	5.34	11.57	22.66	36.82	26.58
v4	1.57	11.35	28.89	11.33	23.18	10.13	7.52	6.03	54.05
v5	8.88	9.12	13.67	20.34	15.24	13.48	12.8	6.47	45.48
v6	8.76	9.79	14.32	19.44	15.12	14.41	12.48	5.68	46.82
v7	8.61	10.43	13.64	18.98	15.3	14.26	12.41	6.37	46.1
v8	8.33	10.38	14.06	18.64	14.68	13.79	12.96	7.16	45.28

Table 3-2 Summary of the Number of Times Selected Genotype Variables are Selected into a Specific Model Position for the Small-scale Simulation Clinical + Genotype Risk Index Models

	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
SNP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
s39	0.27	0.23	0.3	0.4	0.28	0.37	0.32	0.35	0.35	0.3	0.29	0.28	0.31	0.36	0.36	0.3	0.27	0.29	0.34	0.21	6.18	4.91
s51	0.19	0.21	0.3	0.3	0.34	0.22	0.24	0.22	0.4	0.3	0.33	0.25	0.27	0.29	0.29	0.25	0.29	0.32	0.27	0.24	5.52	4.6
s73	0.13	0.44	0.3	0.31	0.35	0.42	0.41	0.29	0.24	0.35	0.46	0.2	0.3	0.29	0.27	0.25	0.47	0.36	0.33	0.32	6.49	5.05
s135	0.16	0.25	0.36	0.3	0.28	0.3	0.22	0.27	0.33	0.31	0.22	0.32	0.28	0.25	0.35	0.31	0.33	0.31	0.22	0.42	5.79	4.55
s168	0.18	0.23	0.34	0.4	0.37	0.43	0.37	0.38	0.37	0.28	0.33	0.32	0.25	0.4	0.35	0.32	0.37	0.23	0.33	0.36	6.61	5.28
s180	0.25	0.29	0.41	0.29	0.33	0.45	0.3	0.43	0.36	0.43	0.52	0.39	0.36	0.32	0.35	0.33	0.26	0.34	0.27	0.33	7.01	5.76
s240	0.2	0.36	0.17	0.33	0.33	0.42	0.23	0.36	0.3	0.32	0.34	0.27	0.21	0.24	0.35	0.38	0.35	0.3	0.28	0.38	6.12	4.81
s273	0.16	0.27	0.34	0.34	0.29	0.35	0.28	0.27	0.27	0.28	0.25	0.28	0.34	0.31	0.28	0.32	0.21	0.35	0.24	0.37	5.8	4.6
s288	0.16	0.23	0.29	0.35	0.26	0.32	0.42	0.23	0.29	0.32	0.28	0.28	0.27	0.28	0.36	0.34	0.26	0.26	0.39	0.34	5.93	4.64
s302	0.12	0.24	0.38	0.41	0.26	0.35	0.23	0.28	0.32	0.4	0.4	0.26	0.34	0.29	0.27	0.25	0.29	0.26	0.33	0.27	5.95	4.79
s329	0.17	0.34	0.32	0.38	0.2	0.4	0.39	0.2	0.25	0.33	0.27	0.29	0.29	0.28	0.29	0.36	0.41	0.32	0.24	0.24	5.97	4.72
s387	0.12	0.27	0.31	0.19	0.33	0.34	0.37	0.26	0.37	0.34	0.24	0.24	0.28	0.33	0.35	0.27	0.3	0.28	0.31	0.33	5.83	4.57
s411	0.15	0.22	0.35	0.33	0.36	0.38	0.24	0.49	0.32	0.41	0.32	0.29	0.29	0.38	0.23	0.38	0.31	0.32	0.26	0.35	6.38	5.06
s413	0.2	0.24	0.31	0.35	0.33	0.34	0.25	0.36	0.24	0.32	0.38	0.33	0.4	0.29	0.24	0.34	0.33	0.38	0.29	0.32	6.24	4.96
s417	0.29	0.38	0.24	0.31	0.22	0.33	0.21	0.42	0.19	0.29	0.38	0.28	0.27	0.22	0.28	0.36	0.24	0.29	0.27	0.26	5.73	4.6
s425	0.17	0.35	0.39	0.28	0.37	0.18	0.37	0.38	0.4	0.41	0.31	0.37	0.33	0.41	0.35	0.3	0.34	0.39	0.4	0.45	6.95	5.27
s462	0.26	0.18	0.25	0.29	0.28	0.24	0.34	0.4	0.27	0.38	0.27	0.34	0.3	0.32	0.18	0.2	0.27	0.35	0.34	0.34	5.8	4.57
s477	0.15	0.32	0.3	0.28	0.26	0.31	0.32	0.31	0.38	0.33	0.29	0.34	0.31	0.21	0.28	0.31	0.3	0.38	0.21	0.29	5.88	4.61
s486	0.16	0.21	0.31	0.46	0.34	0.34	0.38	0.34	0.26	0.23	0.37	0.39	0.3	0.35	0.24	0.33	0.36	0.19	0.38	0.37	6.31	4.81
s494	0.17	0.21	0.25	0.35	0.35	0.38	0.29	0.42	0.34	0.24	0.41	0.48	0.25	0.3	0.27	0.36	0.27	0.33	0.37	0.18	6.22	4.93
s1	0.46	0.14	0.05	0.11	0.05	0.03	0.04	0.04	0.02	0.05	0.02	0.04	0.08	0.05	0.03	0.02	0.03	0.01	0.03	0.06	1.36	1.25
s10	0.3	0.14	0.08	0.11	0.12	0.1	0.12	0.11	0.17	0.06	0.11	0.04	0.05	0.08	0.11	0.06	0.06	0.07	0.1	0.05	2.04	1.76
s50	0.44	0.33	0.14	0.06	0.07	0.05	0.06	0.04	0.05	0.03	0.08	0.08	0.01	0.05	0.04	0.04	0.07	0.08	0.05	0.05	1.82	1.68
s100	0.43	0.2	0.1	0.1	0.1	0.11	0.1	0.05	0.07	0.05	0.05	0.06	0.02	0.07	0.03	0.06	0.05	0.08	0.12	0.04	1.89	1.7

3.2.2 Models

Once the variable selection procedure is finished each of the 100 small-scale simulation datasets have 100 trimmed Clinical and Clinical + Genotype risk index models. Tables 3-3, 3-4, and 3-5 each show five trimmed Clinical risk index models randomly selected from one of three randomly chosen small-scale simulation datasets (datasets #5, #12, and #25). Figures 3-1, 3-2, and 3-3 show the full distribution of risk index values in the optimization set for a randomly selected trimmed Clinical risk index model from each of the three small-scale simulation datasets. Figures 3-4, 3-5, and 3-6 show the full distribution of risk index values in the independent testing set for a randomly selected trimmed Clinical risk index model from each of the three small-scale simulation datasets. In all six of these figures a red line indicates the cut-off point. All individuals with a risk index value greater than or equal to this cut-off point are predicted as "high risk" and all individuals with a value less than this cut-off point are predicted as "low risk". Tables 3-6, 3-7, and 3-8 show the risk index values and predictions from the same set of five Clinical risk index models from the same three small-scale simulation datasets for a set of 25 individuals randomly selected from the optimization set.

Table 3-3 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #5

Bootstrap Sample	Trimmed Clinical Risk Index Model
2	0.1601*v1 + 0.3749*v2 + 0.1046*v4
5	0.1596*v1 + 0.3284*v2 + 0.0047*v6 + 0.0239*v8
44	0.143*v1 + 0.3871*v2 + 0.0864*v4
83	0.1539*v1 + 0.3756*v2 + 0.0766*v3 + 0.055*v4 + 0.0531*v5 + 0.1189*v8
85	0.1472*v1 + 0.417*v2 - 0.2374*v5 + 0.0189*v8

Table 3-4 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #12

Bootstrap Sample	Trimmed Clinical Risk Index Model
24	0.3645*v2 + 0.0979*v3 + 0.0945*v4 - 0.2109*v8
27	-0.0133*v5
37	-0.2249*v5 + 0.0167*v6 + 0.0116*v7 + 0.0195*v8
49	0.1481*v1 + 0.4174*v2 + 0.0845*v3 + 0.0918*v4 - 0.0016*v6
83	0.1352*v1 + 0.3632*v2 + 0.0904*v3 + 0.0856*v4 + 0.1224*v5 + 0.0173*v6 - 0.0492*v8

Table 3-5 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #15

Bootstrap Sample	Trimmed Clinical Risk Index Model
25	0.1271*v1 + 0.3612*v2 + 0.0798*v3 + 0.0501*v6
76	0.3139*v2 - 0.5463*v5 - 0.0426*v8
79	0.1516*v1 + 0.0964*v3 + 0.0909*v4 + 0.0158*v6 - 6e-04*v7 + 0.1363*v8
88	0.1149*v5 + 0.0023*v7
92	0.1537*v1 + 0.2992*v2 + 0.0974*v3 + 0.0072*v6 - 0.1636*v8

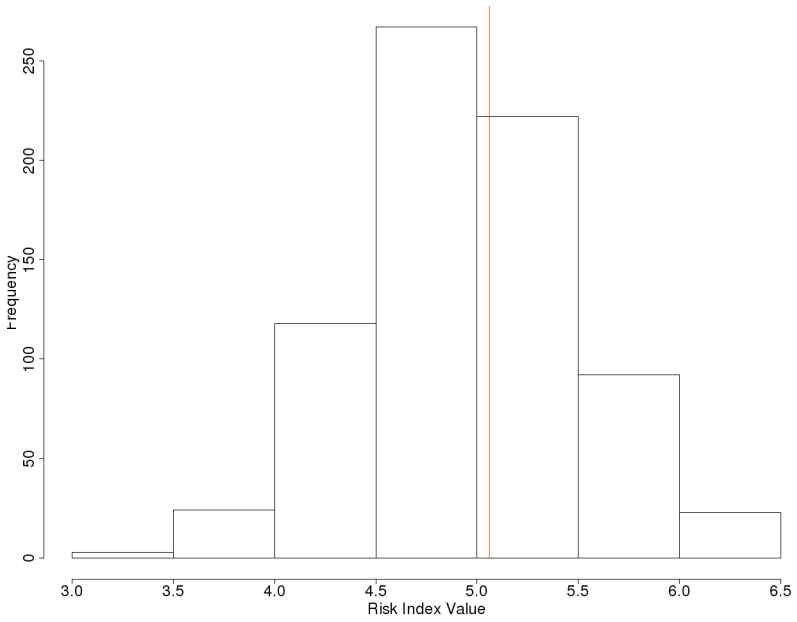


Figure 3-1 Clinical Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #5, Bootstrap Sample #2

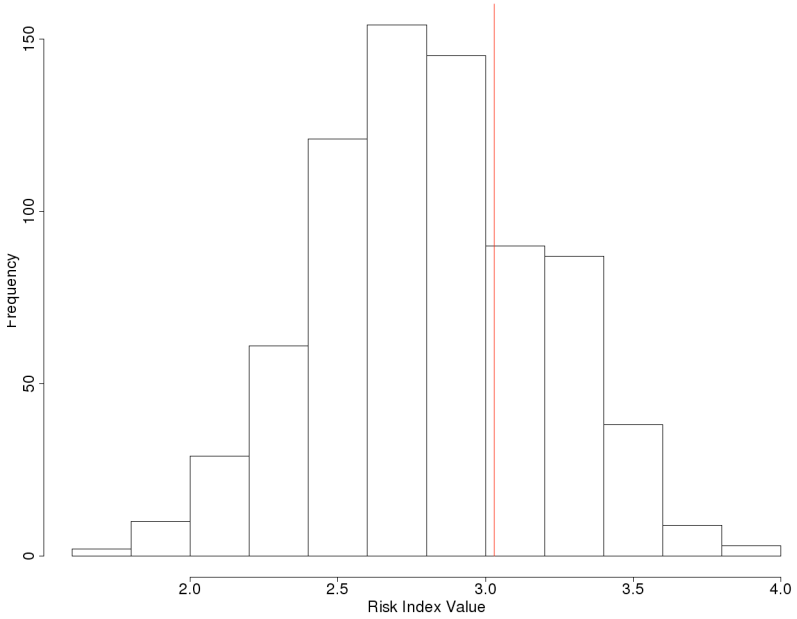


Figure 3-2 Clinical Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #12, Bootstrap Sample #24

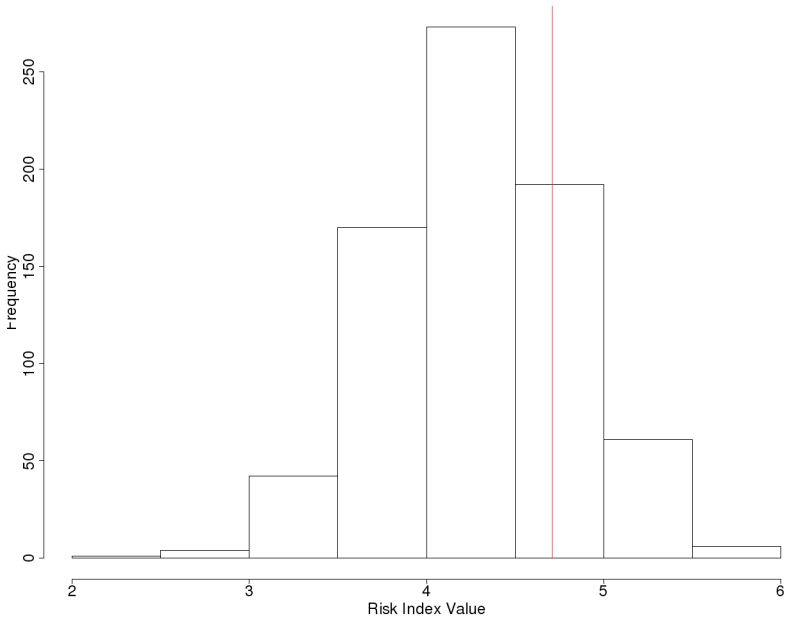


Figure 3-3 Clinical Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #15, Bootstrap Sample #25

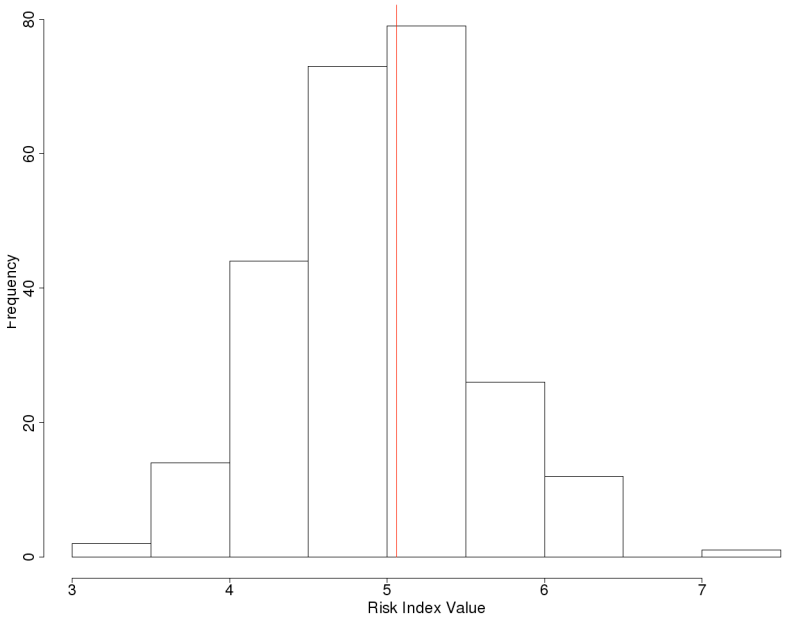


Figure 3-4 Clinical Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #5, Bootstrap Sample #2

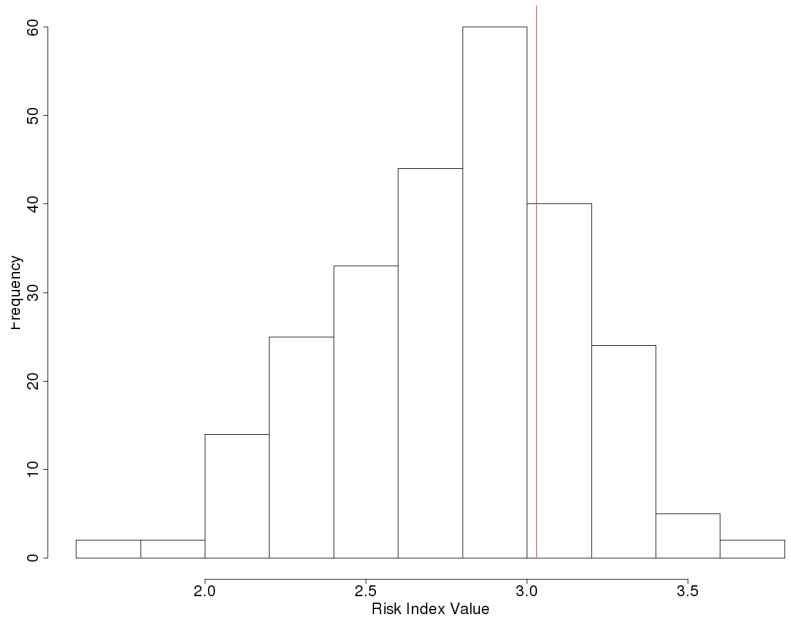


Figure 3-5 Clinical Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #12, Bootstrap Sample #24

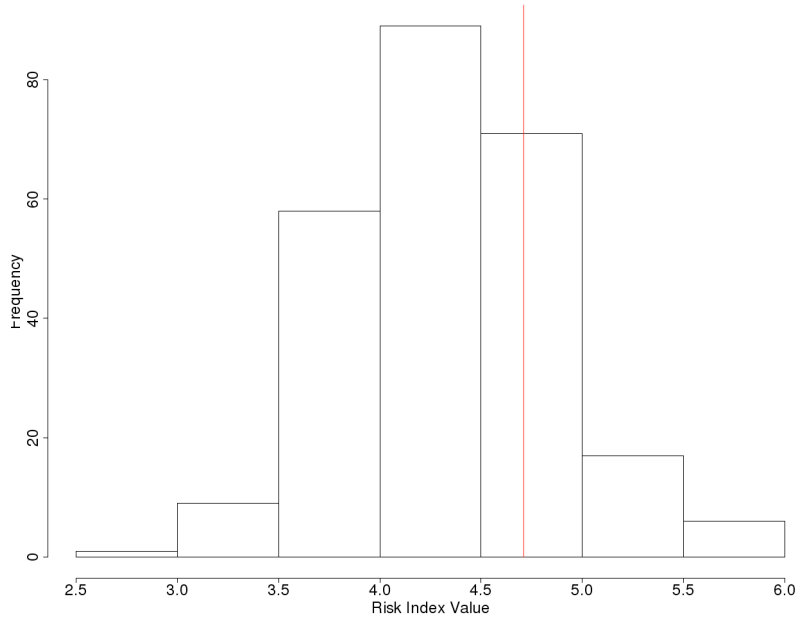


Figure 3-6 Clinical Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #15, Bootstrap Sample #25

Table 3-6 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Small-scale Simulation Dataset #5

		Bootstrap Sample #2 Cutoff Value = 5.061		Bootstrap Sample #5 Cutoff Value = 3.182		Bootstrap Sample #44 Cutoff Value = 4.897		Bootstrap Sample #83 Cutoff Value = 3.675		Bootstrap Sample #85 Cutoff Value = 2.538	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	5.371	1	3.245	1	4.933	1	3.906	1	2.674	1
2	1	4.422	0	2.392	0	3.984	0	2.881	0	1.789	0
3	0	4.650	0	2.735	0	4.335	0	3.060	0	2.265	0
4	0	4.541	0	2.444	0	4.102	0	3.123	0	1.851	0
5	0	3.817	0	2.197	0	3.453	0	2.806	0	1.556	0
6	0	4.879	0	2.713	0	4.434	0	3.282	0	2.126	0
7	0	4.330	0	2.597	0	3.976	0	3.002	0	2.055	0
8	0	4.976	0	3.021	0	4.573	0	3.381	0	2.445	0
9	0	5.104	1	3.049	0	4.648	0	3.634	0	2.432	0
10	1	4.747	0	2.816	0	4.352	0	3.347	0	2.235	0
11	1	5.214	1	3.302	1	4.760	0	3.786	1	2.682	1
12	1	5.183	1	3.251	1	4.701	0	3.757	1	2.555	1
13	0	4.280	0	2.454	0	3.896	0	3.204	0	1.884	0
14	0	4.282	0	2.539	0	3.946	0	3.047	0	2.003	0
15	1	6.074	1	3.694	1	5.559	1	4.125	1	3.115	1
16	0	4.537	0	2.972	0	4.182	0	3.490	0	2.399	0
17	1	5.526	1	3.219	1	5.082	1	3.501	0	2.647	1
18	0	4.225	0	2.668	0	3.855	0	3.145	0	2.044	0
19	1	6.463	1	3.838	1	5.911	1	4.497	1	3.281	1
20	0	5.295	1	2.944	0	4.821	0	3.650	0	2.420	0
21	0	4.835	0	2.825	0	4.448	0	3.190	0	2.322	0
22	0	4.586	0	2.347	0	4.209	0	3.035	0	1.860	0
23	0	4.517	0	2.567	0	4.089	0	3.128	0	1.927	0
24	0	4.250	0	2.414	0	3.883	0	2.928	0	1.838	0
25	1	5.715	1	3.522	1	5.255	1	3.993	1	2.939	1

Table 3-7 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Small-scale Simulation Dataset #12

		Bootstrap Sample #24 Cutoff Value = 3.029		Bootstrap Sample #27 Cutoff Value = -0.131		Bootstrap Sample #37 Cutoff Value = -0.257		Bootstrap Sample #49 Cutoff Value = 4.228		Bootstrap Sample #83 Cutoff Value = 3.277	
		Risk Index		Risk Index		Risk Index		Risk Index		Risk Index	
Individual	Outcome	Value	Prediction	Value	Prediction	Value	Prediction	Value	Prediction	Value	Prediction
1	1	3.302	1	-0.131	1	-0.233	1	4.837	1	3.464	1
2	0	2.483	0	-0.141	0	-0.268	0	3.637	0	2.622	0
3	0	2.721	0	-0.142	0	-0.328	0	4.099	0	2.994	0
4	1	3.168	1	-0.129	1	-0.261	0	4.216	0	3.022	0
5	0	2.464	0	-0.129	1	-0.240	1	3.996	0	2.894	0
6	0	2.457	0	-0.134	0	-0.281	0	3.803	0	2.717	0
7	0	3.198	1	-0.135	0	-0.266	0	4.353	1	3.128	0
8	0	2.862	0	-0.127	1	-0.210	1	3.988	0	2.852	0
9	1	2.678	0	-0.131	1	-0.249	1	3.906	0	2.802	0
10	1	3.096	1	-0.132	0	-0.226	1	4.253	1	3.113	0
11	0	2.938	0	-0.132	0	-0.287	0	4.441	1	3.186	0
12	0	2.598	0	-0.131	1	-0.214	1	4.031	0	2.911	0
13	1	1.978	0	-0.131	1	-0.232	1	2.821	0	2.115	0
14	1	3.388	1	-0.135	0	-0.231	1	4.647	1	3.358	1
15	0	2.590	0	-0.135	0	-0.234	1	3.791	0	2.761	0
16	0	1.937	0	-0.128	1	-0.261	0	3.257	0	2.363	0
17	0	2.400	0	-0.129	1	-0.267	0	3.526	0	2.539	0
18	0	3.305	1	-0.134	0	-0.261	0	4.070	0	2.921	0
19	1	3.143	1	-0.129	1	-0.267	0	4.621	1	3.337	1
20	1	2.989	0	-0.131	1	-0.254	1	4.568	1	3.277	0
21	0	2.767	0	-0.131	0	-0.228	1	3.759	0	2.744	0
22	1	3.439	1	-0.138	0	-0.329	0	4.849	1	3.403	1
23	0	2.300	0	-0.138	0	-0.290	0	3.427	0	2.483	0
24	1	3.291	1	-0.128	1	-0.292	0	4.662	1	3.310	1
25	0	2.725	0	-0.134	0	-0.292	0	3.634	0	2.644	0

Table 3-8 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Small-scale Simulation Dataset #15

		Bootstrap Sample #25 Cutoff Value = 4.710		Bootstrap Sample #76 Cutoff Value = -0.838		Bootstrap Sample #79 Cutoff Value = 3.586		Bootstrap Sample #88 Cutoff Value = 0.643		Bootstrap Sample #92 Cutoff Value = 3.352	
		Risk Index		Risk Index		Risk Index		Risk Index		Risk Index	
Individual	Ouctome	Value	Prediction	Value	Prediction	Value	Prediction	Value	Prediction	Value	Prediction
1	0	3.839	0	-1.095	0	3.038	0	0.621	0	2.752	0
2	0	4.035	0	-1.265	0	3.126	0	0.621	0	3.120	0
3	1	5.608	1	-1.005	0	4.310	1	0.673	1	4.646	1
4	0	3.363	0	-1.107	0	2.713	0	0.642	0	2.340	0
5	1	4.982	1	-1.337	0	4.121	1	0.637	0	3.928	1
6	0	3.799	0	-1.016	0	2.911	0	0.635	0	2.898	0
7	0	4.076	0	-0.964	0	3.242	0	0.630	0	3.119	0
8	0	3.729	0	-1.059	0	2.844	0	0.613	0	2.676	0
9	1	4.375	0	-1.166	0	3.590	1	0.621	0	3.357	1
10	0	4.545	0	-0.827	1	3.325	0	0.603	0	3.465	1
11	0	3.866	0	-0.696	1	2.794	0	0.612	0	2.730	0
12	0	3.336	0	-1.029	0	2.531	0	0.642	0	2.370	0
13	1	4.476	0	-0.930	0	3.431	0	0.645	1	3.485	1
14	0	2.972	0	-0.753	1	2.108	0	0.614	0	1.921	0
15	0	2.329	0	-1.043	0	1.816	0	0.651	1	1.407	0
16	0	3.994	0	-0.948	0	2.921	0	0.626	0	2.964	0
17	1	4.645	0	-0.708	1	3.427	0	0.618	0	3.434	1
18	0	4.291	0	-0.730	1	3.120	0	0.620	0	3.137	0
19	1	4.553	0	-0.949	0	3.376	0	0.646	1	3.445	1
20	1	4.296	0	-0.912	0	3.186	0	0.610	0	3.196	0
21	1	5.095	1	-0.637	1	3.820	1	0.640	0	3.982	1
22	0	3.506	0	-1.229	0	2.766	0	0.634	0	2.576	0
23	1	5.052	1	-0.538	1	3.657	1	0.641	0	3.917	1
24	1	4.552	0	-0.748	1	3.082	0	0.644	1	3.393	1
25	0	4.038	0	-0.874	0	3.003	0	0.622	0	3.010	0

Tables 3-9, 3-10, and 3-11 each show the five trimmed Clinical + Genotype risk index models corresponding with the models shown in Tables 3-3, 3-4, and 3-5. Tables 3-12, 3-13, and 3-14 show the risk index values and predictions from the same set of five Clinical + Genotype risk index models. Figures 3-7, 3-8, and 3-9 show the full distribution of risk index values in the optimization set for a randomly selected trimmed Clinical + Genotype risk index model from each of the three small-scale simulation datasets described above. Figures 3-10, 3-11, and 3-12 show the full distribution of risk index values in the independent testing set for a randomly selected trimmed Clinical + Genotype risk index model from each of the three small-scale simulation datasets. As in the previous set of figures, in all six of these figures a red line indicates the cut-off point. All individuals with a risk index value greater than or equal to this cut-off point are predicted as "high risk" and all individuals with a value less than this cut-off point are predicted as "low risk".

Table 3-9 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #5

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
2	0.1601*v1 + 0.3749*v2 + 0.1046*v4 + 0.0011*s16 - 0.0074*s138 + 0.0488*s166 - 0.0444*s172 + 0.0488*s181 - 0.0088*s186 + 0.0268*s204 - 0.0241*s252 - 0.0691*s295 - 0.0267*s348 + 0.2672*s356
5	0.1596*v1 + 0.3284*v2 + 0.0047*v6 + 0.0239*v8 - 0.0142*s110 + 0.1455*s127 - 0.0438*s134 + 0.0473*s152 - 0.0685*s220 + 0.0281*s267 - 0.048*s273 + 0.0209*s326 - 0.0131*s385 - 0.0665*s417
44	0.143*v1 + 0.3871*v2 + 0.0864*v4 - 0.0517*s22 + 0.0501*s58 - 0.0154*s203 + 0.0276*s231 - 0.017*s341 + 0.2082*s396
83	0.1539*v1 + 0.3756*v2 + 0.0766*v3 + 0.055*v4 + 0.0531*v5 + 0.1189*v8 + 0.0076*s58 - 0.0746*s66 - 0.1219*s74 - 0.048*s136 - 0.0812*s192 - 0.0173*s216 - 0.0506*s229 + 0.0561*s236 + 0.1519*s245 - 0.0691*s256 - 0.0374*s264 - 0.0397*s272 - 0.0056*s294 + 0.0758*s315 - 0.0796*s325 + 0.2099*s443 + 0.0547*s479
85	0.1472*v1 + 0.417*v2 - 0.2374*v5 + 0.0189*v8 + 0.0379*s3 - 0.0563*s22 + 0.0814*s137 - 0.0868*s207 + 0.0149*s208 + 0.1528*s230 + 0.0863*s273 - 0.0042*s322 - 0.1984*s361 - 0.3127*s370 - 0.0524*s402 + 0.0735*s411 + 0.0042*s430

Table 3-10 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #12

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
24	0.3645*v2 + 0.0979*v3 + 0.0945*v4 - 0.2109*v8 + 0.0228*s141 + 0.2769*s175 - 0.047*s208 + 0.0244*s237 - 0.0554*s262 + 0.0216*s380 + 0.2255*s482
27	-0.0133*v5 + 0.0094*s5 + 0.012*s28 - 0.0264*s51 + 0.0161*s61 - 0.0541*s125 - 0.0602*s136 - 0.0169*s137 - 0.0381*s167 - 0.0375*s176 - 0.0464*s231 - 0.0178*s314 + 1e-04*s321 + 0.0251*s324 - 0.0459*s328 - 0.0078*s356 - 0.0711*s357 - 0.0019*s378 + 0.0201*s408
37	-0.2249*v5 + 0.0167*v6 + 0.0116*v7 + 0.0195*v8 + 0.0201*s37 + 0.0418*s52 - 0.0061*s67 - 0.0491*s108 + 0.0699*s112 + 0.0688*s125 + 8e-04*s161 + 0.1483*s201 - 0.0363*s208 + 4e-04*s218 + 0.0133*s241 - 0.0445*s276 - 0.0223*s295 - 0.0644*s296 + 0.1699*s328 + 0.044*s365 + 0.0515*s389 + 0.0466*s398 + 0.0049*s452
49	0.1481*v1 + 0.4174*v2 + 0.0845*v3 + 0.0918*v4 - 0.0016*v6 + 0.0303*s42 + 0.0316*s80 + 0.0064*s87 - 0.0536*s133 - 0.0201*s139 + 0.1096*s150 + 0.0791*s197 + 0.0791*s259 - 0.1046*s349 - 0.1563*s369 + 0.1707*s373 + 0.0667*s375 + 0.1017*s406 + 0.0174*s416 + 0.4357*s459 - 0.0326*s465
83	0.1352*v1 + 0.3632*v2 + 0.0904*v3 + 0.0856*v4 + 0.1224*v5 + 0.0173*v6 - 0.0492*v8 + 0.0392*s4 + 0.0742*s7 - 0.0074*s32 + 0.0685*s94 - 0.1162*s141 + 0.2069*s307 + 0.052*s387

Table 3-11 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #15

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
25	0.1271*v1 + 0.3612*v2 + 0.0798*v3 + 0.0501*v6 - 0.0156*s14 + 0.0383*s22 + 0.0094*s49 - 0.0231*s82 + 0.0235*s101 + 0.0356*s113 - 0.0424*s119 + 0.04*s168 - 0.0988*s201 - 0.2504*s255 - 0.0315*s273 - 0.0067*s375 + 0.0185*s385 + 0.0288*s420 - 0.0376*s424 - 0.0148*s458 + 0.0167*s479 + 0.0043*s491 + 0.2736*s496
76	0.3139*v2 - 0.5463*v5 - 0.0426*v8 + 0.4962*s12 + 0.0094*s18 + 0.0132*s56 + 0.0368*s65 + 0.0054*s106 - 0.0112*s124 - 0.0961*s141 - 0.0199*s143 - 0.0262*s214 - 0.1017*s218 - 0.0064*s240 + 0.1556*s257 - 0.0235*s289 + 0.0977*s310 + 0.0272*s311 + 0.1166*s405 + 0.0248*s436 + 0.0533*s485
79	0.1516*v1 + 0.0964*v3 + 0.0909*v4 + 0.0158*v6 - 6e-04*v7 + 0.1363*v8 - 0.1052*s46 - 0.0304*s86 + 0.1201*s158 - 0.096*s162 + 0.0155*s173 + 0.0827*s254 - 0.1274*s265 + 0.4236*s329 - 0.0118*s336 - 0.0026*s337 + 0.0312*s359 - 0.1349*s377 + 0.0325*s405 - 0.1539*s426 + 0.0686*s439 - 0.0048*s444 - 0.0332*s469
88	0.1149*v5 + 0.0023*v7 + 0.0135*s59 - 0.122*s166 - 7e-04*s179 + 0.0221*s195 + 0.0075*s251 - 0.0147*s265 - 0.0067*s352 - 0.0137*s360 + 0.0162*s374 - 0.0627*s384 + 0.0103*s388 - 0.0242*s391 + 0.0109*s406 - 0.0854*s413 + 0.0619*s440 + 0.0079*s448 + 0.0531*s461 + 0.0122*s464 + 0.0347*s495
92	0.1537*v1 + 0.2992*v2 + 0.0974*v3 + 0.0072*v6 - 0.1636*v8 + 0.0106*s51 - 0.2487*s75 - 0.0166*s84 - 0.1151*s90 + 0.0016*s123 - 0.0618*s133 - 0.0326*s136 + 0.0098*s166 - 0.0228*s186 - 0.0633*s193 + 0.0691*s207 + 0.0473*s212 + 0.0169*s215 + 0.0756*s269 + 0.1881*s275 + 0.0381*s366 - 0.0177*s422 + 0.0664*s435 + 0.1145*s475

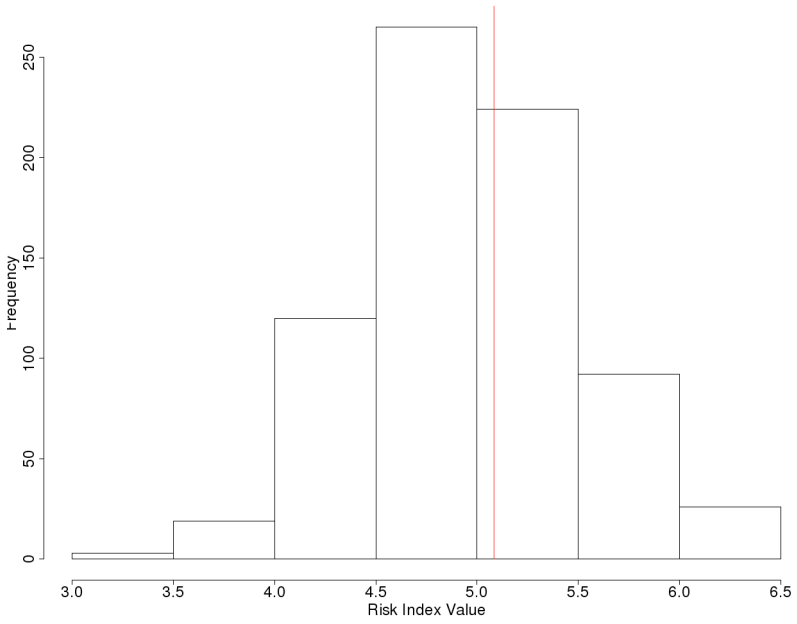


Figure 3-7 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #5, Bootstrap Sample #2

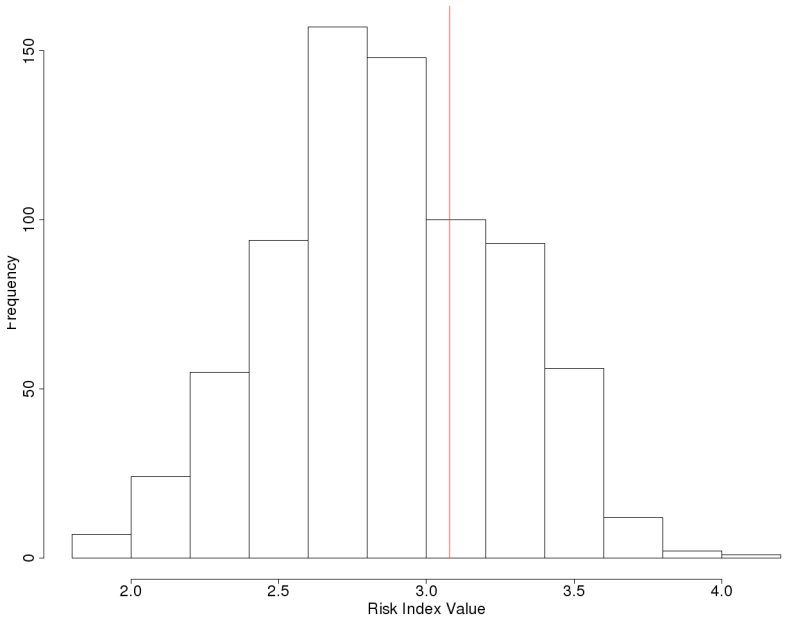


Figure 3-8 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #12, Bootstrap Sample #24

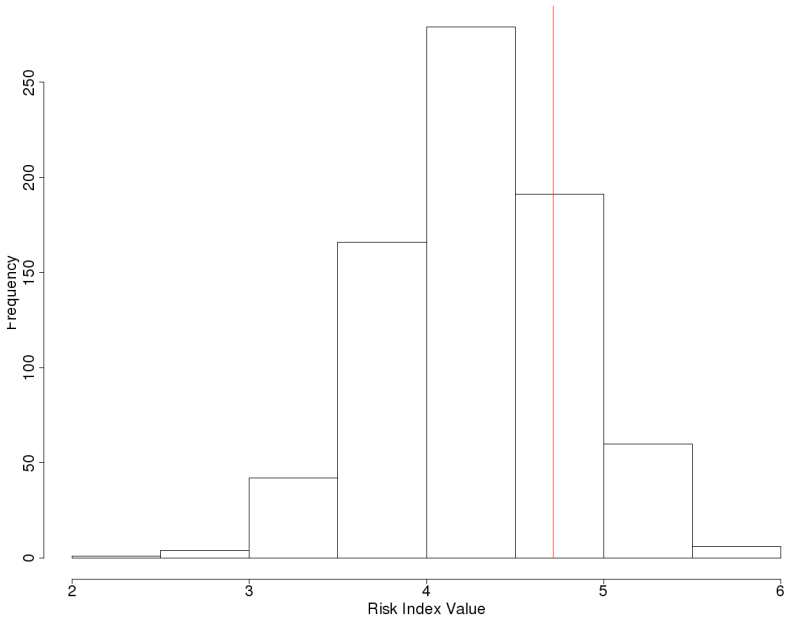


Figure 3-9 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #15, Bootstrap Sample #25

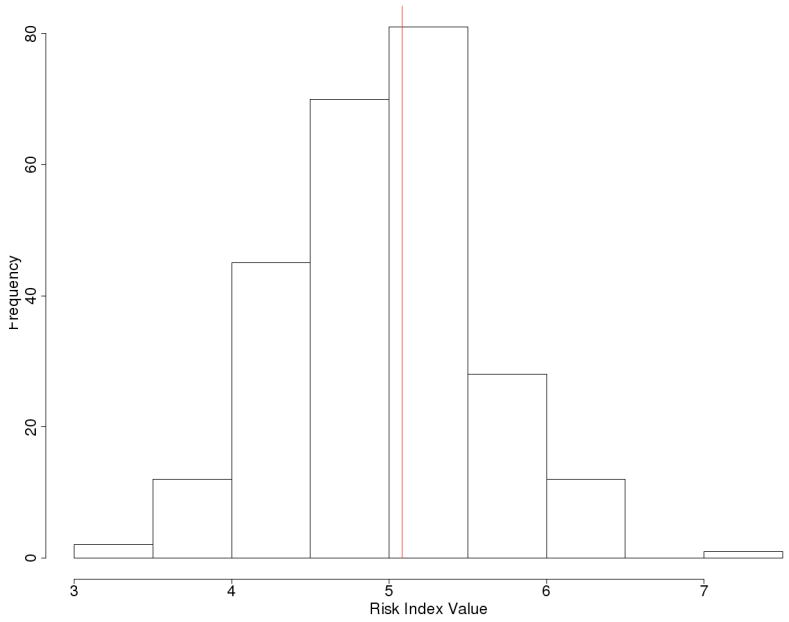


Figure 3-10 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #5, Bootstrap Sample #2

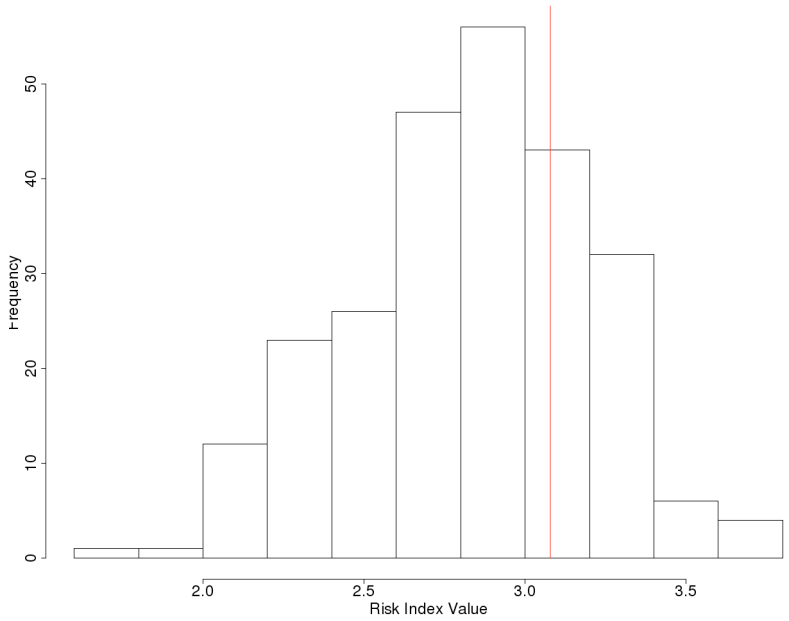


Figure 3-11 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #12, Bootstrap Sample #24

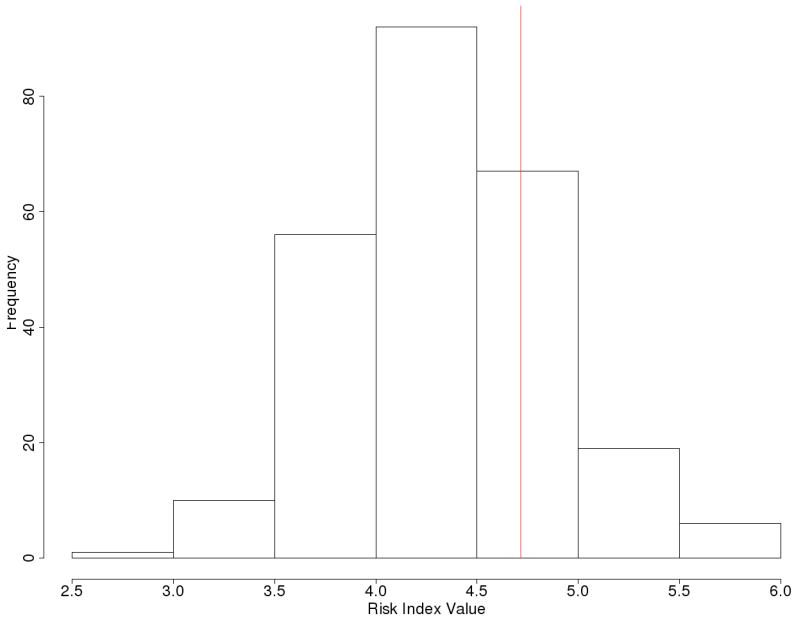


Figure 3-12 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #15, Bootstrap Sample #25

Table 3-12 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Small-scale Simulation Dataset #5

		Bootstrap Sample #2 Cutoff Value = 5.083		Bootstrap Sample #5 Cutoff Value = 3.196		Bootstrap Sample #44 Cutoff Value = 4.923		Bootstrap Sample #83 Cutoff Value = 3.680		Bootstrap Sample #85 Cutoff Value = 2.525	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	5.396	1	3.270	1	4.990	1	3.902	1	2.672	1
2	1	4.448	0	2.409	0	3.984	0	2.885	0	1.811	0
3	0	4.652	0	2.753	0	4.337	0	3.070	0	2.250	0
4	0	4.559	0	2.448	0	4.100	0	3.116	0	1.809	0
5	0	3.843	0	2.237	0	3.528	0	2.806	0	1.549	0
6	0	4.909	0	2.697	0	4.436	0	3.278	0	2.098	0
7	0	4.359	0	2.609	0	3.969	0	3.006	0	2.057	0
8	0	4.967	0	3.020	0	4.570	0	3.388	0	2.440	0
9	0	5.107	1	3.045	0	4.646	0	3.647	0	2.450	0
10	1	4.772	0	2.827	0	4.352	0	3.370	0	2.196	0
11	1	5.211	1	3.321	1	4.823	0	3.797	1	2.686	1
12	1	5.211	1	3.250	1	4.733	0	3.740	1	2.495	0
13	0	4.318	0	2.465	0	3.896	0	3.197	0	1.902	0
14	0	4.317	0	2.542	0	3.943	0	3.034	0	1.981	0
15	1	6.095	1	3.717	1	5.545	1	4.112	1	3.065	1
16	0	4.566	0	2.988	0	4.249	0	3.481	0	2.359	0
17	1	5.531	1	3.230	1	5.118	1	3.508	0	2.669	1
18	0	4.248	0	2.671	0	3.879	0	3.147	0	2.051	0
19	1	6.489	1	3.837	1	5.905	1	4.486	1	3.290	1
20	0	5.282	1	2.958	0	4.846	0	3.653	0	2.416	0
21	0	4.833	0	2.841	0	4.474	0	3.181	0	2.337	0
22	0	4.643	0	2.351	0	4.235	0	3.036	0	1.842	0
23	0	4.520	0	2.565	0	4.124	0	3.133	0	1.900	0
24	0	4.267	0	2.439	0	3.946	0	2.907	0	1.825	0
25	1	5.718	1	3.532	1	5.274	1	3.985	1	2.952	1

Table 3-13 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Small-scale Simulation Dataset #12

		Bootstrap Sample #24 Cutoff Value = 3.078		Bootstrap Sample #27 Cutoff Value = -0.140		Bootstrap Sample #37 Cutoff Value = -0.233		Bootstrap Sample #49 Cutoff Value = 4.236		Bootstrap Sample #83 Cutoff Value = 3.296	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	3.403	1	-0.142	0	-0.230	1	4.838	1	3.483	1
2	0	2.558	0	-0.145	0	-0.257	0	3.667	0	2.671	0
3	0	2.748	0	-0.144	0	-0.320	0	4.080	0	2.985	0
4	1	3.204	1	-0.138	1	-0.249	0	4.201	0	3.040	0
5	0	2.459	0	-0.143	0	-0.211	1	4.019	0	2.899	0
6	0	2.491	0	-0.143	0	-0.270	0	3.815	0	2.749	0
7	0	3.292	1	-0.145	0	-0.243	0	4.361	1	3.138	0
8	0	2.901	0	-0.141	0	-0.189	1	3.987	0	2.874	0
9	1	2.716	0	-0.142	0	-0.233	0	3.920	0	2.833	0
10	1	3.163	1	-0.136	1	-0.210	1	4.251	1	3.109	0
11	0	3.038	0	-0.133	1	-0.275	0	4.468	1	3.209	0
12	0	2.626	0	-0.137	1	-0.211	1	4.036	0	2.925	0
13	1	2.046	0	-0.142	0	-0.204	1	2.828	0	2.109	0
14	1	3.423	1	-0.149	0	-0.204	1	4.680	1	3.421	1
15	0	2.625	0	-0.152	0	-0.220	1	3.819	0	2.773	0
16	0	2.083	0	-0.137	1	-0.246	0	3.248	0	2.389	0
17	0	2.467	0	-0.133	1	-0.250	0	3.547	0	2.547	0
18	0	3.344	1	-0.141	0	-0.241	0	4.072	0	2.971	0
19	1	3.210	1	-0.144	0	-0.246	0	4.639	1	3.350	1
20	1	3.046	0	-0.134	1	-0.256	0	4.563	1	3.321	1
21	0	2.885	0	-0.132	1	-0.223	1	3.779	0	2.790	0
22	1	3.531	1	-0.147	0	-0.298	0	4.893	1	3.421	1
23	0	2.356	0	-0.146	0	-0.273	0	3.431	0	2.531	0
24	1	3.358	1	-0.132	1	-0.281	0	4.668	1	3.362	1
25	0	2.765	0	-0.142	0	-0.266	0	3.622	0	2.681	0

Table 3-14 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Small-scale Simulation Dataset #15

		Bootstrap Sample #25 Cutoff Value = 4.716		Bootstrap Sample #7 Cutoff Value = -0.803		Bootstrap Sample #79 Cutoff Value = 3.577		Bootstrap Sample #88 Cutoff Value = 0.646		Bootstrap Sample #92 Cutoff Value = 3.355	
Individual	Ouctome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	3.854	0	-1.068	0	3.016	0	0.617	0	2.777	0
2	0	4.035	0	-1.191	0	3.106	0	0.617	0	3.139	0
3	1	5.596	1	-0.997	0	4.289	1	0.676	1	4.644	1
4	0	3.358	0	-1.067	0	2.689	0	0.645	0	2.363	0
5	1	4.989	1	-1.313	0	4.125	1	0.641	0	3.960	1
6	0	3.798	0	-0.997	0	2.882	0	0.635	0	2.894	0
7	0	4.088	0	-0.935	0	3.221	0	0.640	0	3.136	0
8	0	3.727	0	-1.048	0	2.827	0	0.611	0	2.687	0
9	1	4.351	0	-1.129	0	3.573	0	0.621	0	3.370	1
10	0	4.540	0	-0.798	1	3.291	0	0.617	0	3.481	1
11	0	3.882	0	-0.647	1	2.805	0	0.603	0	2.716	0
12	0	3.335	0	-0.992	0	2.547	0	0.636	0	2.378	0
13	1	4.494	0	-0.888	0	3.438	0	0.651	1	3.480	1
14	0	2.963	0	-0.680	1	2.087	0	0.610	0	1.927	0
15	0	2.346	0	-0.976	0	1.817	0	0.657	1	1.412	0
16	0	4.022	0	-0.863	0	2.903	0	0.627	0	2.952	0
17	1	4.637	0	-0.676	1	3.429	0	0.622	0	3.458	1
18	0	4.269	0	-0.667	1	3.154	0	0.625	0	3.186	0
19	1	4.569	0	-0.934	0	3.388	0	0.657	1	3.446	1
20	1	4.319	0	-0.904	0	3.185	0	0.606	0	3.202	0
21	1	5.080	1	-0.566	1	3.797	1	0.633	0	4.004	1
22	0	3.491	0	-1.202	0	2.768	0	0.643	0	2.587	0
23	1	5.075	1	-0.488	1	3.655	1	0.643	0	3.930	1
24	1	4.554	0	-0.671	1	3.087	0	0.647	1	3.405	1
25	0	4.048	0	-0.857	0	3.008	0	0.631	0	3.001	0

3.2.3 Predictive Performance

After the variable selection procedure is completed and the models are applied to each individual in the independent testing set then the sensitivity, specificity, misclassification, and positive predictive value are estimated for both the Clinical and Clinical + Genotype risk index models for each of the 100 small-scale simulation datasets. Table 3-15 shows the means and standard deviations of these measurements. To provide a 95% confidence for these measurements of sensitivity, specificity, misclassification, and positive predictive value for each independent testing set, 1000 bootstrap samples were generated. By making predictions about each individual in these bootstrap samples and calculating the sensitivity, specificity, misclassification, and positive predictive value for each bootstrap sample, 95% confidence intervals were estimated for these measurements in each of the 100 small-scale simulation datasets. The mean and standard deviation of the spread (i.e., range) of these confidence intervals for both the Clinical and Clinical + Genotype risk index model is shown in Table 3-16. This provides a view to the variability of the sensitivity, specificity, misclassification, and positive predictive value estimates.Table 3-17 shows the predictive performance and confidence intervals for the three small-scale simulation datasets discussed in Section 3.2.2.

Table 3-15 Means and Standard Deviations of Predictive Performance Estimates for the 100 Small-scale Simulation Datasets

Model	Sensitivity (SD)	Specificity (SD)	Misclassification (SD)	PPV (SD)	AUC (SD)
Clinical	0.606 (0.067)	0.886 (0.031)	0.199 (0.021)	0.704 (0.064)	0.832 (0.027)
Clinical + Genotype	0.589 (0.065)	0.896 (0.030)	0.197 (0.022)	0.717 (0.062)	0.846 (0.024)

Table 3-16 Means and Standard Deviations of Predictive Performance 95% Confidence Intervals for the 100 Small-scale Simulation Datasets

Model	Mean Range of 95% Confidence Interval (SD)			
	Sensitivity	Specificity	Misclassification	PPV
Clinical	0.219 (0.017)	0.093 (0.013)	0.099 (0.006)	0.220 (0.021)
Clinical + Genotype	0.0220 (0.016)	0.090 (0.013)	0.098 (0.006)	0.223 (0.023)

Table 3-17 Predictive Performance Estimates for Three Small-scale Simulation Datasets

Dataset	Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
5	Clinical	0.73 (0.625-0.821)	0.864 (0.811-0.914)	0.175 (0.127-0.223)	0.692 (0.592-0.793)
	Clinical + Genotype	0.703 (0.6-0.797)	0.876 (0.829-0.921)	0.175 (0.131-0.223)	0.703 (0.595-0.8)
12	Clinical	0.528 (0.423-0.658)	0.899 (0.858-0.94)	0.207 (0.159-0.255)	0.679 (0.548-0.8)
	Clinical + Genotype	0.528 (0.418-0.658)	0.916 (0.878-0.955)	0.195 (0.147-0.243)	0.717 (0.593-0.833)
15	Clinical	0.658 (0.545-0.76)	0.886 (0.836-0.93)	0.183 (0.139-0.235)	0.714 (0.603-0.815)
	Clinical + Genotype	0.658 (0.528-0.75)	0.903 (0.853-0.945)	0.175 (0.131-0.227)	0.742 (0.634-0.846)

Using the number of models predicting an individual in the independent testing set as "high risk", receiver operating characteristic (ROC) curves were generated for the Clinical and Clinical + Genotype risk index model for each of the 100 small-scale simulation datasets, and the area under the ROC curve (AUC) was estimated. The average AUC of the Clinical risk index models was 0.832 (SD = 0.027), and the average AUC of the Clinical + Genotype risk index models was 0.846 (SD = 0.024). Figure 3-13, 3-14, and 3-15 show the ROC curves for the Clinical risk index model the three small-scale simulation datasets discussed in sections 3.2.2, and Figure 3-16, 3-17, and 3-18 show the ROC curve for the Clinical + Genotype risk index model from those three selected small-scale simulation datasets.

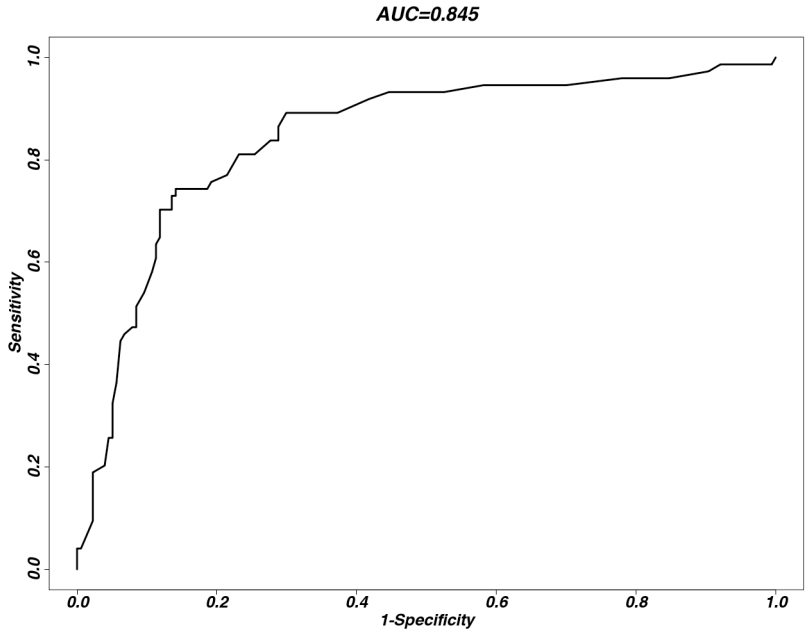


Figure 3-13 ROC Curve of the Clinical Risk Index Model for Small-scale Simulation Dataset #5

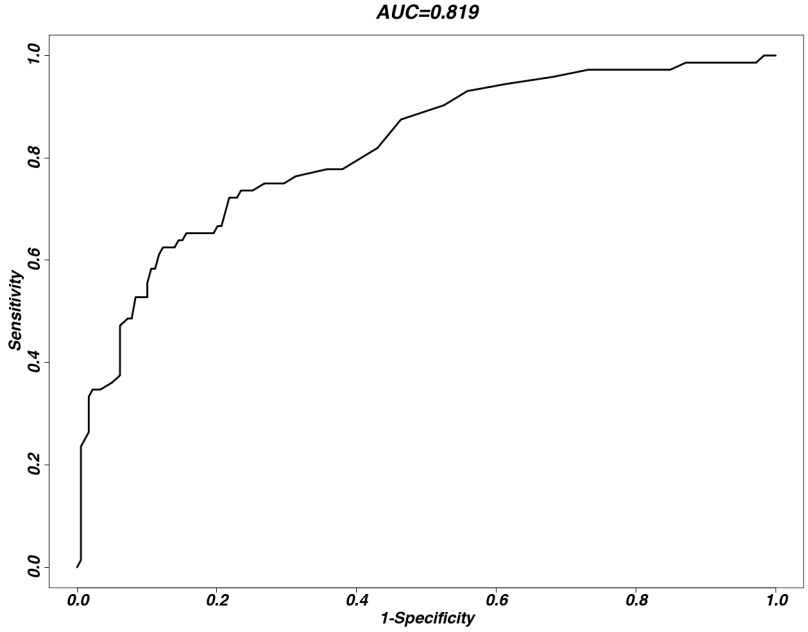


Figure 3-14 ROC Curve of the Clinical Risk Index Model for Small-scale Simulation Dataset #12

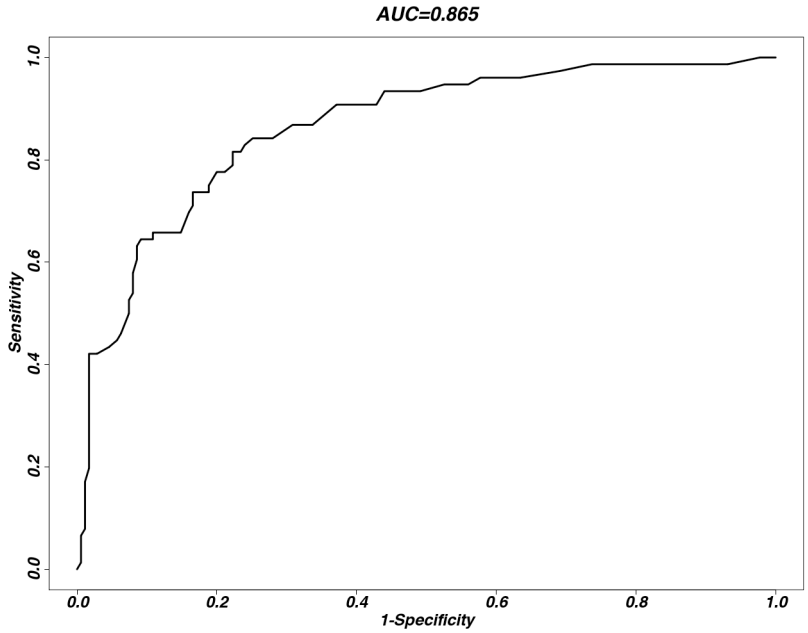


Figure 3-15 ROC Curve of the Clinical Risk Index Model for Small-scale Simulation Dataset #15

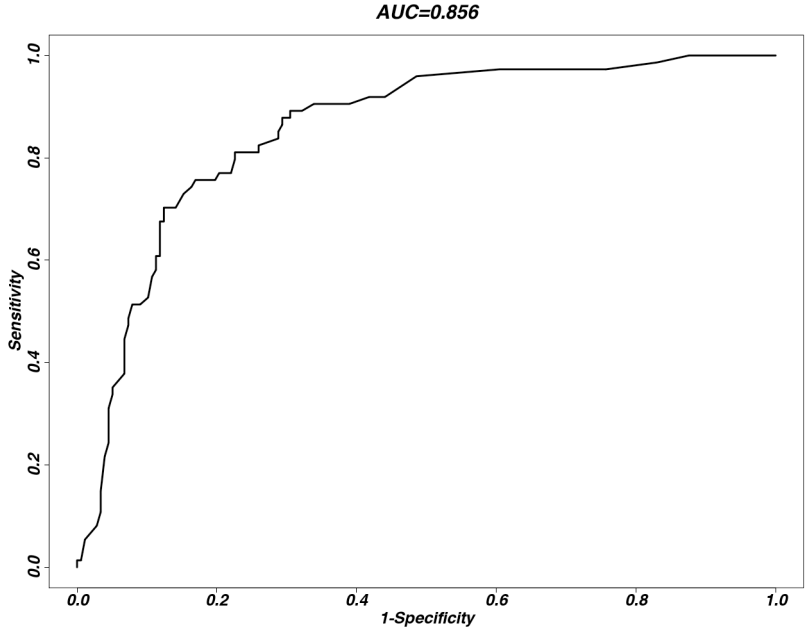


Figure 3-16 ROC Curve of the Clinical + Genotype Risk Index Model for Small-scale Simulation Dataset #5

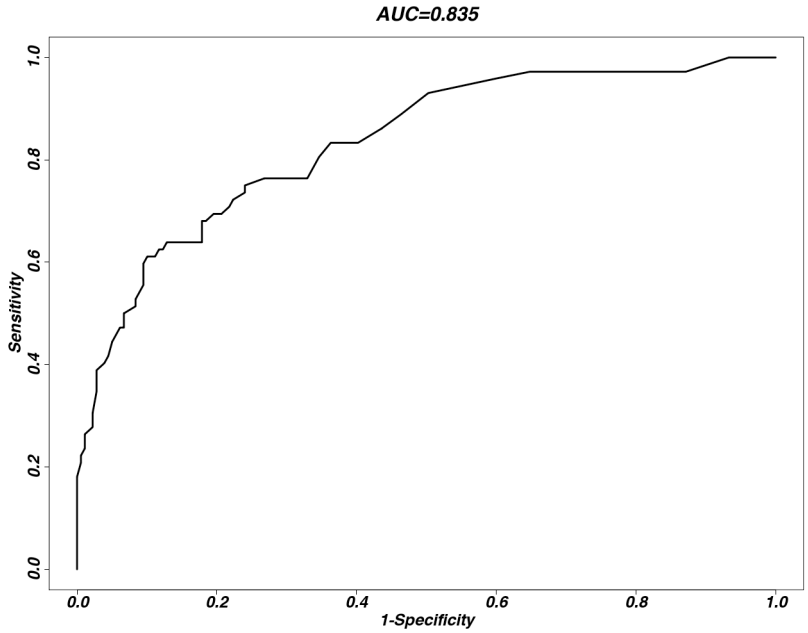


Figure 3-17 ROC Curve of the Clinical + Genotype Risk Index Model for Small-scale Simulation Dataset #12

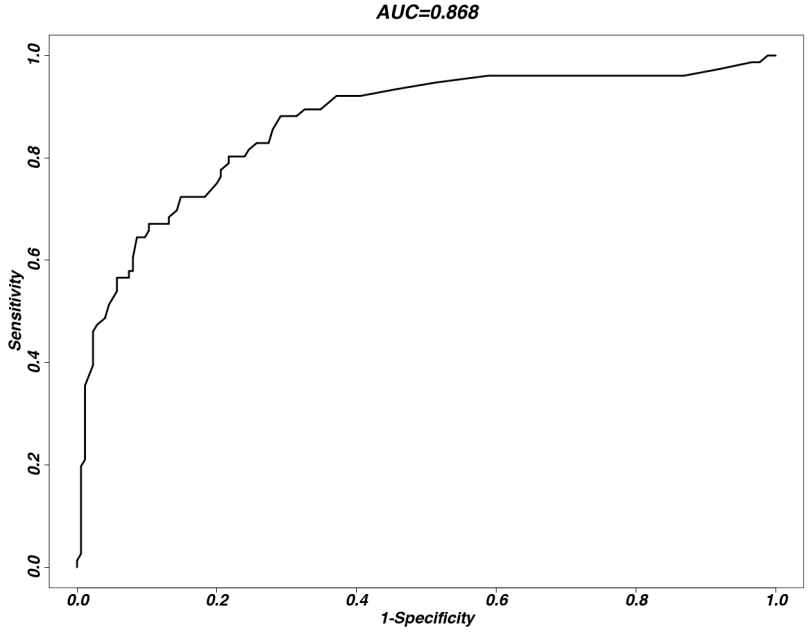


Figure 3-18 ROC Curve of the Clinical + Genotype Risk Index Model for Small-scale Simulation Dataset #15

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. It is currently thought that the proportion of models that predict that an individual as high risk, then, represents the predicted probability of an individual developing the outcome. More theoretical work would be needed to determine this relationship. Using the binomial distribution a 95% confidence interval can be constructed for this estimated probability with the Wilson score interval (Wilson, 1927). Physicians can then use this interval, given by

$$95\%\ \text{CI} = \frac{\hat{p} + \frac{(1.96)^2}{2n} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{(1.96)^2}{4n^2}}}{1 + \frac{(1.96)^2}{n}}$$, to gauge the prediction. As an example, one individual in the independent testing set for the 5th small-scale simulation dataset has a predicted probability of developing the outcome of 0.14. The lower bound for this individual's 95% confidence interval is $\frac{0.14 + \frac{(1.96)^2}{2*100} - 1.96\sqrt{\frac{0.14(1-0.14)}{100} + \frac{(1.96)^2}{4*(100)^2}}}{1 + \frac{(1.96)^2}{100}}$

or 0.085 and the upper bound is $\frac{0.14 + \frac{(1.96)^2}{2*100} + 1.96\sqrt{\frac{0.14(1-0.14)}{100} + \frac{(1.96)^2}{4*(100)^2}}}{1 + \frac{(1.96)^2}{100}}$ or

0.221.

3.2.4 Random Forest Comparison

For each of the 100 small-scale simulation datasets a random forest was generated using the optimization set created by the risk index procedure. Each forest had 500 individual trees, and a tuning procedure was used to find the number of variables *k* considered at

each split that provided the lowest out-of-bag error estimate. Beginning with $k = \sqrt{v}$, where *v* is the total number of predictor variables, the forest was grown and out-of-bag error was measured. Then, the number of variables considered at each split was progressively increased by a factor of two (i.e., $k = 2 * \sqrt{v}$, $k = 4 * \sqrt{v}$, etc.) until the out-of-bag error decreased by less than 5% from the out-of-bag error for the previous value of *k*. Next, returning to $k = \sqrt{v}$, the number of variables considered at each split was progressively decreased by a factor of two (i.e., $k = \frac{1}{2} * \sqrt{v}$, $k = \frac{1}{4} * \sqrt{v}$, etc.) until the out-of-bag error decreased by less than 5% from the out-of-bag error for the previous value of *k*.

For each of the random forests an ROC curve was generated and the AUC was estimated. The mean AUC of the random forest models was 0.987 (SD = 0.006). Figures 3-19, 3-20, and 3-21 show the ROC curve for the random forest generated from the three small-scale simulation datasets described in Section 3.2.2. When working with a dataset that has two possible classes, the standard procedure for a random forest is to assign a prediction to an individual based on a simple majority of votes. When the prevalence of the outcome is less than 50% changing the proportion of votes needed to classify an individual can significantly impact the estimates of performance. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set using a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of 5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the

individual to be “high risk”. Table 3-28 shows the mean and standard deviation of the sensitivity, specificity, misclassification, and PPV for a range of different proportions.

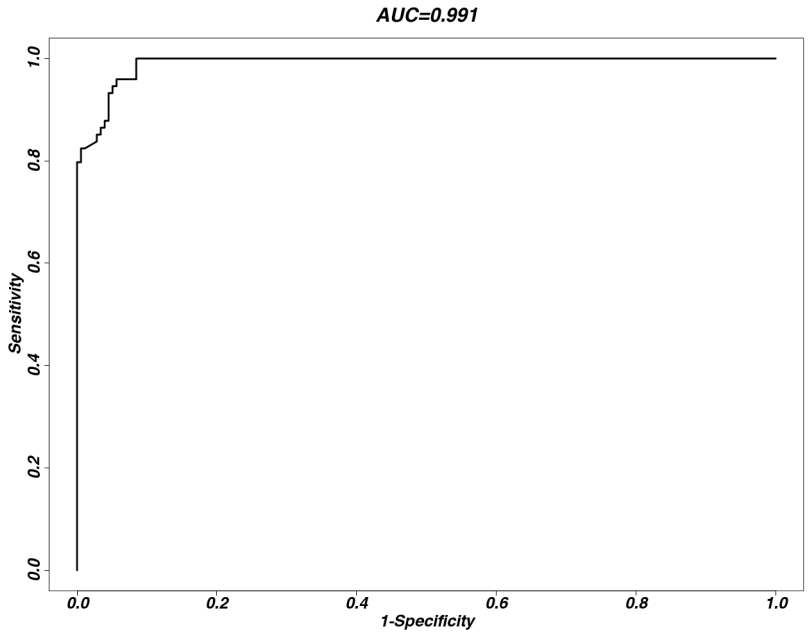


Figure 3-19 ROC Curve of the Random Forest Generated for Small-scale Simulation Dataset #5

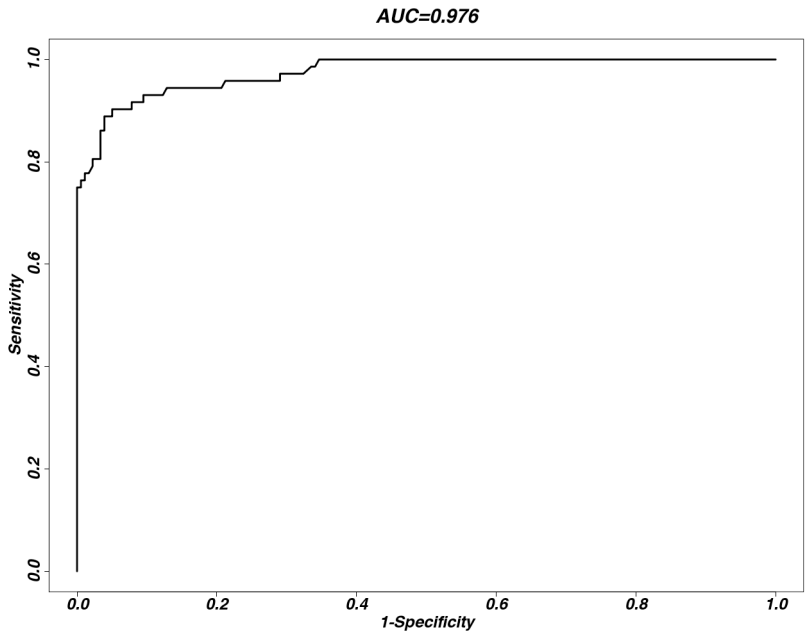


Figure 3-20 ROC Curve of the Random Forest Generated for Small-scale Simulation Dataset #12

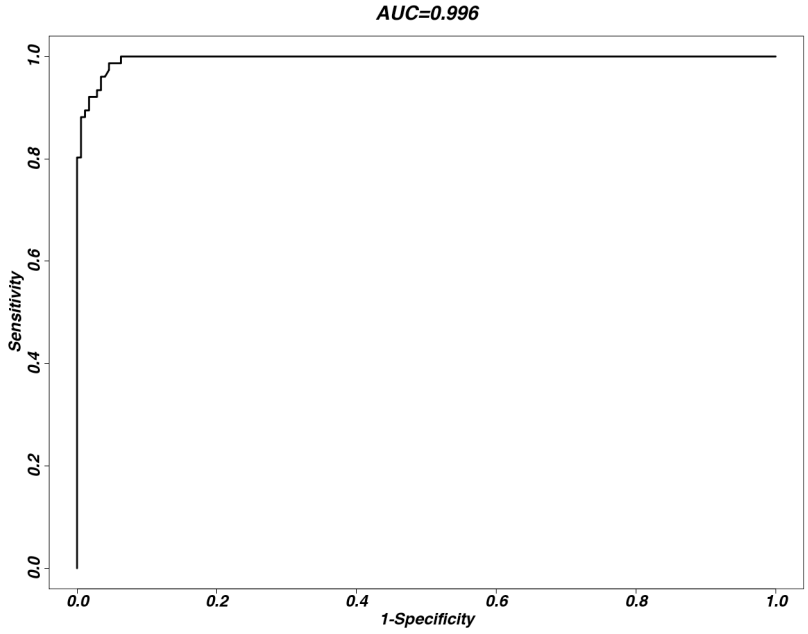


Figure 3-21 ROC Curve of the Random Forest Generated for Small-scale Simulation Dataset #15

Table 3-18 Means and Standard Deviations of Performance Estimates of the Random Forest Models Generated from the 100 Small-scale Simulation Datasets

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	1 (0.002)	0.184 (0.079)	0.565 (0.057)	0.353 (0.037)
0.1	0.998 (0.005)	0.499 (0.069)	0.347 (0.048)	0.47 (0.045)
0.15	0.992 (0.011)	0.705 (0.05)	0.207 (0.034)	0.6 (0.049)
0.2	0.983 (0.015)	0.82 (0.035)	0.13 (0.024)	0.708 (0.047)
0.25	0.972 (0.018)	0.888 (0.028)	0.086 (0.019)	0.794 (0.046)
0.3	0.954 (0.025)	0.922 (0.022)	0.068 (0.015)	0.844 (0.042)
0.35	0.933 (0.029)	0.941 (0.017)	0.061 (0.014)	0.876 (0.037)
0.4	0.915 (0.032)	0.958 (0.015)	0.055 (0.012)	0.905 (0.033)
0.45	0.893 (0.036)	0.969 (0.013)	0.054 (0.013)	0.928 (0.03)
0.5	0.868 (0.041)	0.979 (0.011)	0.055 (0.014)	0.949 (0.027)
0.55	0.839 (0.046)	0.987 (0.01)	0.058 (0.015)	0.966 (0.024)
0.6	0.81 (0.048)	0.993 (0.007)	0.063 (0.015)	0.981 (0.019)
0.65	0.781 (0.049)	0.996 (0.006)	0.07 (0.016)	0.989 (0.015)
0.7	0.717 (0.055)	0.998 (0.003)	0.088 (0.018)	0.994 (0.01)
0.75	0.593 (0.06)	0.999 (0.002)	0.125 (0.021)	0.998 (0.007)
0.8	0.419 (0.074)	1 (0.001)	0.178 (0.026)	0.999 (0.005)
0.85	0.209 (0.069)	1 (0)	0.242 (0.029)	1 (0)
0.9	0.049 (0.043)	1 (0)	0.292 (0.031)	1 (0)
0.95	0.001 (0.003)	1 (0)	0.307 (0.03)	1 (0)

3.2.5 Conclusion

In the small-scale simulation study the risk index procedure quite often identified and selected the covariates that were associated with the outcome. However, the SNPs that were designated as being associated with the outcome were not selected by the risk index procedure more frequently than those SNPs that were designated as having no association with the outcome. Although the “true positive” SNPs have logistic regression coefficients similar to, or even higher than, the “true positive” covariates, as shown in Table 3-29, the median p-values for the “true positive” SNPs are markedly higher than the median p-values for the “true positive” covariates (Table 3-29). This suggests that the standard errors of the logistic regression coefficients are larger for the SNPs than for the covariates, which would lead the SNPs to have a smaller impact on predicting the

outcome. When a risk index was built using only the variables that were simulated to be associated with the outcome (i.e., v1, v2, v3, v4, s1, s10, s50, s100), the performance is slightly better than the best model built using the standard risk index model (Table 3-19)

Table 3-19 Performance Characteristics of the Risk Index Models Including Only Variables Associated with the Outcome

Model	Sensitivity (SD)	Specificity (SD)	Misclassification (SD)	PPV (SD)
Clinical	0.627 (0.119)	0.853 (0.063)	0.216 (0.025)	0.763 (0.005)
Clinical + Genotypes	0.610 (0.123)	0.862 (0.064)	0.216 (0.026)	0.767 (0.028)

Table 3-20 Mean Logistic Regression Coefficients and Median Logistic Regression Coefficient P-values for "True Positive" Variables

Variable	Mean Logistic Regression Coefficient	Median Logistic Regression Coefficient P-value
v1	0.145	1.21E-34
v2	0.38	4.42E-21
v3	0.082	2.08E-29
v4	0.082	7.78E-12
s1	-0.32	3.03E-03
s10	-0.218	6.36E-02
s50	-0.335	9.91E-04
s100	-0.278	1.16E-02

Although the predictive performance of the risk index procedure is quite good, with a mean AUC of the Clinical + Genotype risk index models that is significantly greater than the mean AUC of the Clinical risk index models ($p=0.0002$), the random forest models had a mean AUC of 0.987 (SD = 0.006), which is significantly greater than the mean AUC of the Clinical + Genotype risk index models ($p<2.2e-16$). Tuning the class assignment procedure for the risk index can produce sensitivity, specificity, and positive

predictive values greater than 0.9, much greater than the predictive performance estimates of the risk index models.

3.3 Small-scale Simulation Study Top Principal Components Results

3.3.1 Variable Selection

Using the same procedure as in Section 3.2.1, Clinical and Clinical + Genotype risk index models were constructed for each of the 100 small-scale simulation datasets. In place of the 500 SNPs, a principal components analysis was performed using SMARTPCA (Patterson, et al, 2006), and the principal components that accounted for 90% of the variance among the SNPs were used to build the risk index models. On average, 310 principal components were needed to account for 90% of the variance.

Table 3-20 shows the summary of the variable selection procedure from the Clinical risk index model averaged across the 100 simulation datasets. V1, v2, and v4 are most frequently selected; on average, they each appear in more than half of the 100 trimmed Clinical risk index models. V3, because of its high correlation with v1, is typically chosen as one of the last variables (on average, variable three is chosen as the sixth, seventh, or eighth variable in 70.93 of the 100 untrimmed Clinical risk index models for a given simulation dataset).

Table 3-21 shows the summary of the variable selection procedure from the Clinical + Genotype risk index model averaged across the 100 simulation datasets. No principal component was in more than 5.92 out of 100 Clinical + Genotype risk index models on

average. The principal components most commonly observed in trimmed Clinical + Genotype risk index models were PC 93 (5.92 out of 100 trimmed Clinical + Genotype risk index models, on average), PC 110 (5.89 out of 100 trimmed Clinical + Genotype risk index models, on average), PC 10 (5.85 out of 100 trimmed Clinical + Genotype risk index models, on average), PC 138 (5.8 out of 100 trimmed Clinical Genotype risk index models, on average), and PC 188 (5.75 out of 100 trimmed Clinical + Genotype risk index models, on average.

Table 3-21 Summary of the Number of Times Each Variable is Selected into a Specific Model Position for the Small-scale Simulation Clinical Risk Index Models

	Variable Position								Total # of Times in Trimmed
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Model**
v1	43.96	5.62	3.04	2.98	3.66	6.52	13.74	20.48	57.3
v2	7.57	39.66	10.01	4.24	7.99	15.89	4.34	10.3	62.91
v3	12.66	3.77	2.76	4.03	5.85	11.16	21.89	37.88	25.98
v4	1.83	11.92	29.85	10.55	22.9	9.85	7.39	5.71	55.09
v5	8.44	9.85	13.38	20.18	14.52	13.55	13.57	6.51	46.21
v6	8.47	9.89	13.98	19.64	14.95	14.73	12.11	6.23	46.64
v7	8.49	9.71	13.47	19.46	14.79	14.23	13.4	6.45	45.37
v8	8.58	9.58	13.51	18.92	15.34	14.07	13.56	6.44	45.32

Table 3-22 Summary of the Number of Times Selected Principal Component Variable is Selected into a Specific Model Position for the Small-scale Simulation Clinical + Genotype Risk Index Models

	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
SNP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
PC93	0.35	0.33	0.39	0.39	0.49	0.41	0.36	0.42	0.31	0.37	0.26	0.34	0.33	0.41	0.26	0.3	0.38	0.36	0.41	0.32	7.19	5.92
PC110	0.31	0.33	0.36	0.3	0.37	0.38	0.31	0.42	0.37	0.41	0.32	0.36	0.39	0.44	0.35	0.35	0.36	0.46	0.4	0.33	7.32	5.89
PC10	0.35	0.34	0.4	0.31	0.3	0.38	0.34	0.3	0.31	0.34	0.36	0.32	0.44	0.34	0.31	0.4	0.5	0.38	0.3	0.35	7.07	5.85
PC138	0.31	0.37	0.49	0.35	0.39	0.37	0.33	0.44	0.38	0.36	0.43	0.33	0.31	0.26	0.33	0.29	0.34	0.37	0.29	0.27	7.01	5.8
PC188	0.35	0.26	0.35	0.49	0.35	0.29	0.36	0.38	0.38	0.38	0.35	0.38	0.39	0.3	0.32	0.38	0.36	0.39	0.45	0.31	7.22	5.75
PC223	0.33	0.41	0.33	0.34	0.34	0.34	0.36	0.37	0.36	0.3	0.3	0.39	0.41	0.4	0.23	0.38	0.36	0.34	0.47	0.38	7.14	5.75
PC296	0.4	0.33	0.41	0.28	0.32	0.4	0.35	0.47	0.28	0.3	0.37	0.34	0.28	0.34	0.42	0.33	0.31	0.35	0.28	0.31	6.87	5.74
PC211	0.29	0.34	0.42	0.37	0.29	0.38	0.28	0.3	0.41	0.39	0.38	0.38	0.28	0.28	0.24	0.36	0.44	0.37	0.37	0.25	6.82	5.74
PC268	0.41	0.33	0.33	0.3	0.36	0.41	0.3	0.42	0.33	0.32	0.27	0.3	0.34	0.31	0.42	0.33	0.38	0.37	0.35	0.45	7.03	5.71
PC290	0.33	0.29	0.37	0.42	0.34	0.41	0.29	0.42	0.47	0.28	0.34	0.38	0.4	0.33	0.27	0.31	0.31	0.31	0.43	0.3	7	5.69
PC170	0.44	0.29	0.31	0.4	0.43	0.36	0.38	0.3	0.45	0.29	0.41	0.32	0.38	0.35	0.34	0.31	0.36	0.31	0.34	0.37	7.14	5.68
PC237	0.29	0.49	0.32	0.44	0.27	0.35	0.37	0.31	0.34	0.44	0.36	0.24	0.36	0.25	0.27	0.38	0.23	0.34	0.42	0.29	6.76	5.66
PC102	0.42	0.42	0.32	0.37	0.36	0.3	0.49	0.4	0.27	0.33	0.4	0.31	0.25	0.3	0.32	0.31	0.33	0.23	0.22	0.38	6.73	5.64
PC74	0.43	0.31	0.37	0.28	0.25	0.49	0.42	0.29	0.36	0.35	0.29	0.29	0.37	0.34	0.35	0.37	0.27	0.4	0.4	0.39	7.02	5.63
PC32	0.32	0.34	0.41	0.25	0.29	0.33	0.27	0.33	0.33	0.39	0.41	0.29	0.44	0.25	0.4	0.33	0.49	0.35	0.29	0.39	6.9	5.61
PC291	0.42	0.33	0.33	0.35	0.38	0.28	0.45	0.39	0.42	0.37	0.25	0.16	0.37	0.32	0.4	0.35	0.3	0.38	0.37	0.34	6.96	5.59
PC248	0.35	0.31	0.29	0.42	0.4	0.27	0.29	0.35	0.39	0.36	0.32	0.33	0.23	0.38	0.42	0.38	0.36	0.31	0.29	0.36	6.81	5.59
PC227	0.35	0.37	0.37	0.41	0.43	0.49	0.34	0.18	0.34	0.31	0.42	0.32	0.32	0.36	0.36	0.34	0.3	0.31	0.24	0.31	6.87	5.58
PC39	0.3	0.23	0.38	0.4	0.43	0.31	0.19	0.28	0.44	0.31	0.46	0.34	0.35	0.41	0.44	0.3	0.32	0.28	0.37	0.26	6.8	5.58
PC295	0.35	0.41	0.25	0.3	0.39	0.33	0.47	0.36	0.4	0.29	0.3	0.32	0.32	0.4	0.36	0.32	0.27	0.48	0.2	0.39	6.91	5.56
PC244	0.27	0.37	0.35	0.37	0.33	0.37	0.35	0.27	0.31	0.35	0.3	0.34	0.44	0.41	0.32	0.32	0.19	0.35	0.3	0.3	6.61	5.56
PC65	0.43	0.43	0.35	0.29	0.3	0.33	0.36	0.26	0.32	0.42	0.37	0.33	0.26	0.34	0.44	0.35	0.29	0.43	0.28	0.36	6.94	5.55
PC280	0.37	0.49	0.31	0.34	0.32	0.37	0.33	0.31	0.35	0.31	0.31	0.45	0.34	0.33	0.33	0.26	0.3	0.35	0.33	0.29	6.79	5.55
PC9	0.33	0.3	0.41	0.32	0.35	0.33	0.29	0.33	0.24	0.37	0.39	0.38	0.25	0.39	0.39	0.42	0.36	0.38	0.41	0.3	6.94	5.54

3.3.2 Models

Once the variable selection procedure is finished each of the 100 small-scale simulation datasets have 100 trimmed Clinical and Clinical + Genotype risk index models. Tables 3-22, 3-23, and 3-24 each show five trimmed Clinical risk index models randomly selected from one of three randomly chosen small-scale simulation datasets (datasets #5, #12, and #15). These Clinical risk index models are directly comparable to those described in Section 3.2.2 (Tables 3-3, 3-4, and 3-5). Although we will not directly compare these sets of models here they represent another entire set of results for the Clinical risk index models. Figures 3-22, 3-23, and 3-24 show the full distribution of risk index values in the optimization set for a randomly selected trimmed Clinical risk index models from each of the three small-scale simulation datasets. Figures 3-25, 3-26, and 3-27 show the full distribution of risk index values in the independent testing set for a randomly selected trimmed Clinical risk index model from each of the three small-scale simulation datasets. In all six of these figures a red line indicates the cut-off point. All individuals with a risk index value greater than or equal to this cut-off point are predicted as "high risk" and all individuals with a value less than this cut-off point are predicted as "low risk". Tables 3-25, 3-26, and 3-27 show the risk index values and predictions from the same set of five Clinical risk index models from the same three small-scale simulation datasets for a set of 25 individuals randomly selected from the independent test set.

Table 3-23 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #5

Bootstrap Sample	Trimmed Clinical Risk Index Model
2	0.004*v6
5	0.1644*v1 + 0.3554*v2
44	0.14*v1 + 0.2395*v5 - 0.0079*v6 - 0.0236*v7
83	0.1366*v1 + 0.3025*v2 + 0.1001*v4 - 0.0116*v6 - 0.0382*v7 - 0.0838*v8
85	-0.079*v5

Table 3-24 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #12

Bootstrap Sample	Trimmed Clinical Risk Index Model
24	0.156*v1 + 0.3719*v2 + 0.1017*v4 + 0.1242*v5 - 0.006*v7 - 0.0213*v8
27	0.1348*v1 + 0.4034*v2 + 0.8165*v5 + 3e-04*v7
37	-0.0194*v8
49	0.1114*v1 + 0.3982*v2 + 0.0657*v3 + 0.0895*v4 - 0.3266*v5 + 0.0495*v6 + 0.0056*v7 - 0.0054*v8
83	0.1531*v1 + 0.2988*v2 + 0.0804*v4 + 0.0115*v7

Table 3-25 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #15

Bootstrap Sample	Trimmed Clinical Risk Index Model
25	0.1571*v1 + 0.405*v2 + 0.0867*v3 + 0.0806*v4 + 0.0141*v6 - 0.0154*v7 + 0.0629*v8
76	0.1696*v1 + 0.2453*v2 + 0.0816*v4 + 0.0044*v6 + 0.0502*v8
79	0.1418*v1 + 0.4494*v2 + 0.0165*v6 - 0.2143*v8
88	0.0115*v6 - 0.0408*v8
92	0.141*v1 + 0.4952*v2 + 0.1181*v4 + 0.0152*v6 - 0.1511*v8

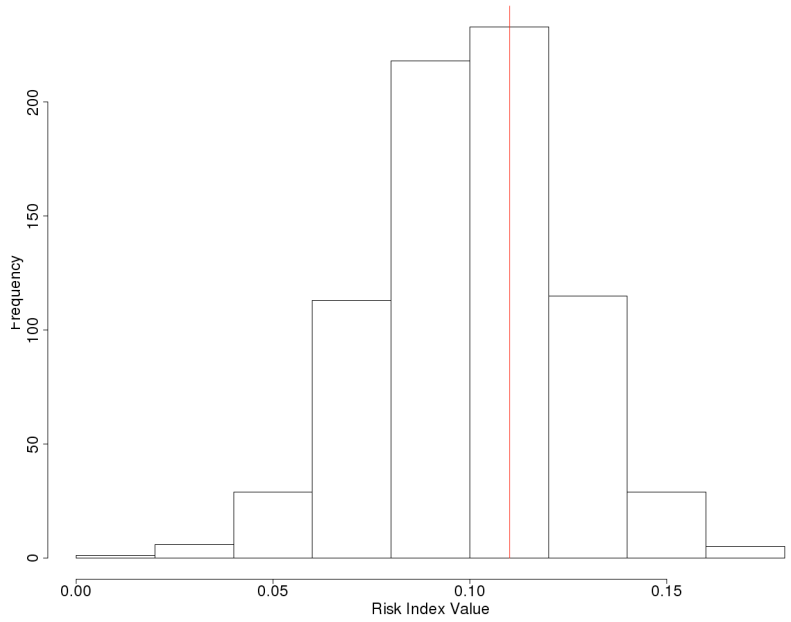


Figure 3-22 Clinical Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #5, Bootstrap Sample #2

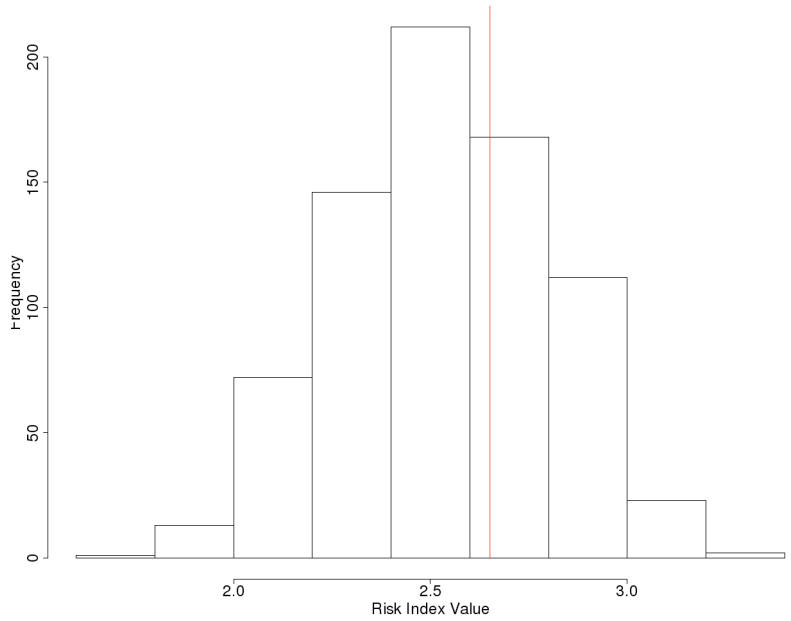


Figure 3-23 Clinical Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #12, Bootstrap Sample #24

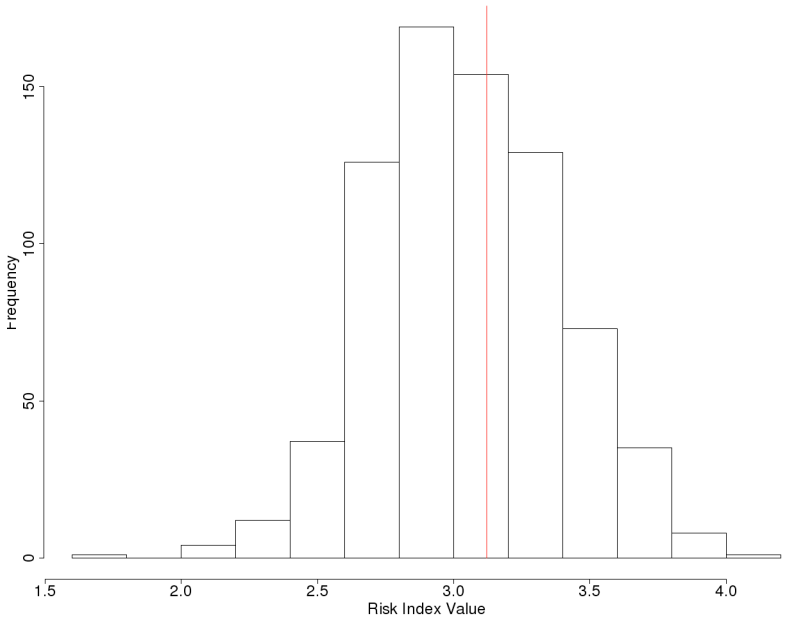


Figure 3-24 Clinical Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #15, Bootstrap Sample #25

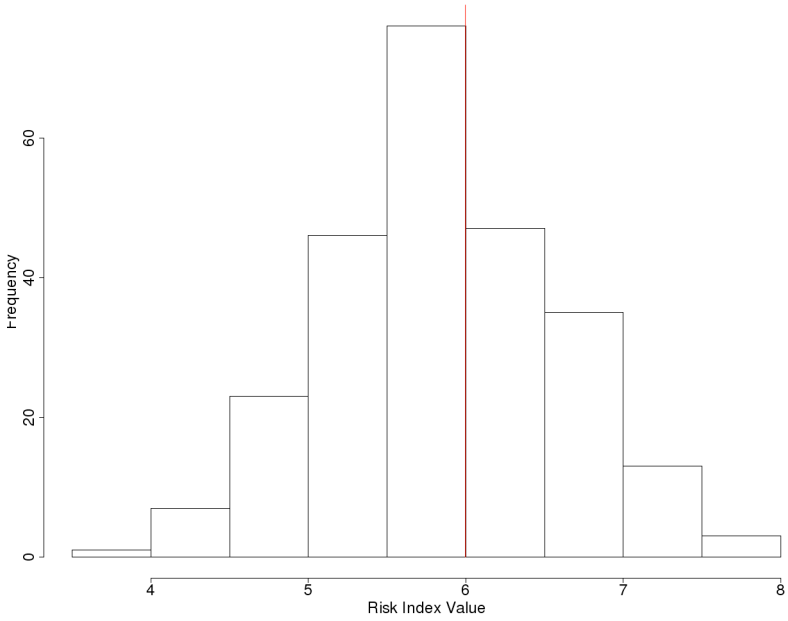


Figure 3-25 Clinical Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #5, Bootstrap Sample #2

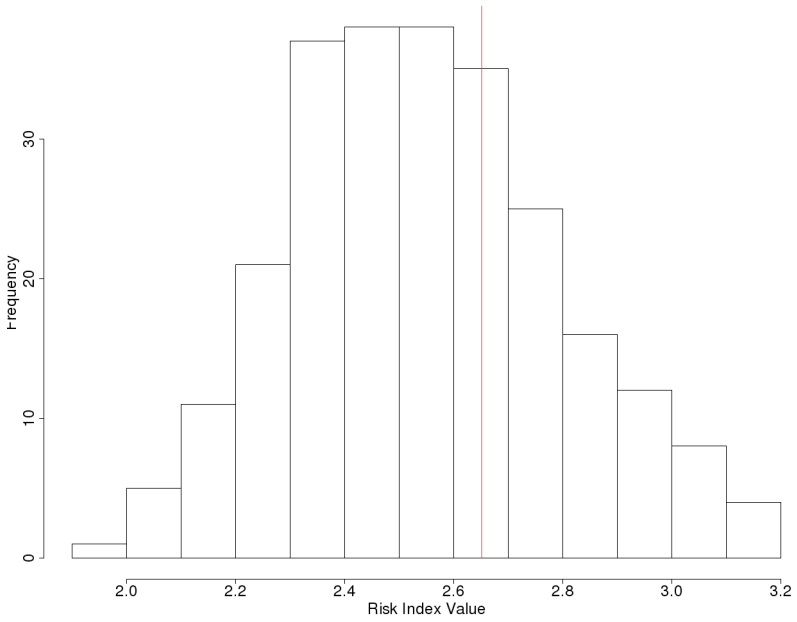


Figure 3-26 Clinical Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #12, Bootstrap Sample #24

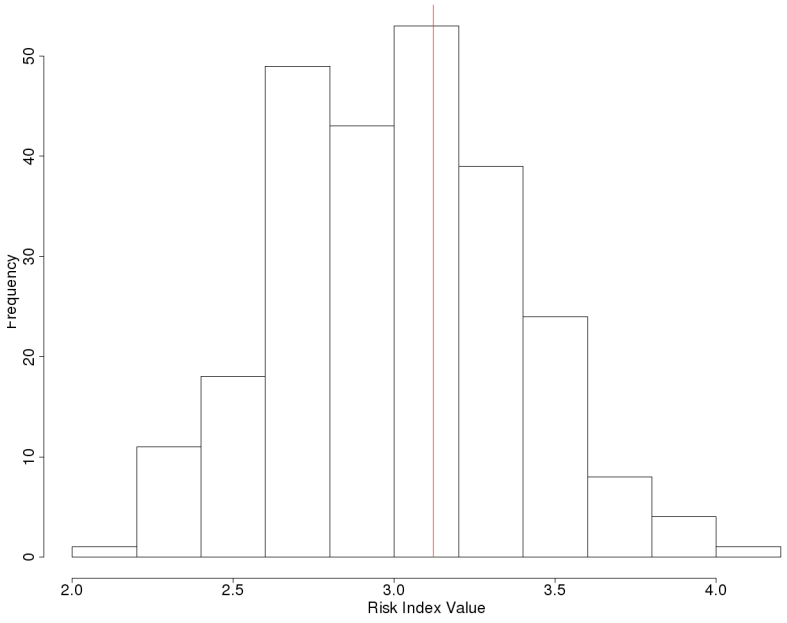


Figure 3-27 Clinical Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #15, Bootstrap Sample #25

Table 3-26 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Small-scale Simulation Dataset #5

		Bootstrap Sample #2 Cutoff Value = 0.110		Bootstrap Sample #5 Cutoff Value = 5.997		Bootstrap Sample #44 Cutoff Value = 2.207		Bootstrap Sample #83 Cutoff Value=1.690		Bootstrap Sample #85 Cutoff Value=-0.772	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	0.081	0	6.614	1	2.553	1	1.882	1	-0.782	0
2	1	0.141	1	6.229	1	2.440	1	1.830	1	-0.783	0
3	1	0.120	1	5.652	0	1.965	0	1.490	0	-0.767	1
4	0	0.131	1	5.033	0	1.867	0	1.544	0	-0.785	0
5	0	0.118	1	5.872	0	1.880	0	1.629	0	-0.795	0
6	0	0.082	0	4.882	0	1.640	0	1.245	0	-0.812	0
7	1	0.109	0	7.009	1	2.111	0	1.929	1	-0.748	1
8	1	0.118	1	5.809	0	1.868	0	1.795	1	-0.835	0
9	0	0.110	0	5.714	0	1.942	0	1.625	0	-0.729	1
10	0	0.088	0	5.434	0	1.972	0	1.624	0	-0.772	0
11	0	0.098	0	5.967	0	1.740	0	1.583	0	-0.803	0
12	1	0.084	0	6.573	1	2.312	1	1.854	1	-0.813	0
13	1	0.118	1	6.255	1	2.339	1	1.784	1	-0.809	0
14	0	0.102	0	5.977	0	2.111	0	1.596	0	-0.813	0
15	1	0.153	1	7.404	1	2.182	0	2.182	1	-0.791	0
16	0	0.105	0	5.502	0	2.080	0	1.360	0	-0.813	0
17	0	0.084	0	5.762	0	2.084	0	1.779	1	-0.823	0
18	0	0.120	1	4.583	0	1.813	0	1.220	0	-0.837	0
19	1	0.062	0	7.008	1	2.383	1	1.879	1	-0.769	1
20	1	0.079	0	6.963	1	2.271	1	1.879	1	-0.791	0
21	0	0.099	0	6.347	1	2.028	0	1.672	0	-0.814	0
22	1	0.121	1	6.243	1	2.112	0	1.651	0	-0.775	0
23	1	0.135	1	6.932	1	2.447	1	1.946	1	-0.766	1
24	0	0.104	0	5.435	0	1.882	0	1.586	0	-0.797	0
25	0	0.125	1	6.279	1	2.196	0	1.655	0	-0.825	0

Table 3-27 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Small-scale Simulation Dataset #12

		Bootstrap Sample #24 Cutoff Value = 2.651		Bootstrap Sample #27 Cutoff Value = 5.049		Bootstrap Sample #37 Cutoff Value = -0.226		Bootstrap Sample #49 Cutoff Value=2.050		Bootstrap Sample #83 Cutoff Value=3.740	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	2.505	0	4.747	0	-0.274	0	1.924	0	3.364	0
2	1	2.836	1	4.999	0	-0.221	1	2.182	1	3.744	1
3	0	2.525	0	4.619	0	-0.213	1	1.927	0	3.323	0
4	0	2.306	0	4.572	0	-0.213	1	1.741	0	3.091	0
5	1	2.691	1	4.717	0	-0.223	1	1.999	0	3.591	0
6	1	2.815	1	5.211	1	-0.256	0	2.277	1	3.763	1
7	0	2.642	0	4.935	0	-0.269	0	1.976	0	3.487	0
8	1	2.894	1	5.076	1	-0.250	0	2.239	1	3.909	1
9	1	2.857	1	5.201	1	-0.200	1	2.212	1	3.821	1
10	1	2.897	1	5.201	1	-0.246	0	2.249	1	3.861	1
11	0	2.251	0	4.232	0	-0.243	0	1.633	0	2.982	0
12	0	2.200	0	4.240	0	-0.233	0	1.589	0	2.928	0
13	1	3.149	1	5.635	1	-0.268	0	2.288	1	4.094	1
14	0	2.315	0	4.267	0	-0.229	0	1.719	0	3.055	0
15	1	2.859	1	5.183	1	-0.207	1	2.183	1	3.871	1
16	0	2.487	0	4.481	0	-0.211	1	1.874	0	3.271	0
17	0	2.487	0	4.810	0	-0.238	0	1.858	0	3.426	0
18	0	2.506	0	4.716	0	-0.203	1	1.978	0	3.356	0
19	0	2.691	1	5.049	0	-0.249	0	2.021	0	3.667	0
20	1	2.270	0	4.404	0	-0.229	0	1.567	0	2.932	0
21	0	2.070	0	3.858	0	-0.282	0	1.556	0	2.742	0
22	1	2.949	1	5.260	1	-0.243	0	2.363	1	3.957	1
23	1	2.849	1	5.204	1	-0.244	0	2.244	1	3.810	1
24	0	2.797	1	4.813	0	-0.246	0	2.218	1	3.739	0
25	0	2.293	0	4.461	0	-0.185	1	1.725	0	2.990	0

Table 3-28 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Small-scale Simulation Dataset #15

		Bootstrap Sample #25 Cutoff Value = 3.121		Bootstrap Sample #76 Cutoff Value = 2.901		Bootstrap Sample #79 Cutoff Value = 2.492		Bootstrap Sample #88 Cutoff Value=-0.068		Bootstrap Sample #92 Cutoff Value=2.921	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	3.180	1	3.072	1	2.574	1	-0.057	1	3.116	1
2	1	3.314	1	2.972	1	2.784	1	-0.073	0	2.958	1
3	1	2.956	0	2.742	0	2.352	0	-0.074	0	3.054	1
4	1	3.350	1	3.213	1	2.660	1	-0.146	0	3.042	1
5	1	3.305	1	2.982	1	2.815	1	-0.095	0	3.197	1
6	0	2.524	0	2.348	0	2.040	0	-0.083	0	2.470	0
7	1	3.472	1	3.194	1	2.661	1	-0.120	0	3.326	1
8	0	2.784	0	2.687	0	2.240	0	-0.144	0	2.571	0
9	1	3.651	1	3.283	1	3.119	1	-0.068	0	3.441	1
10	0	3.021	0	2.800	0	1.990	0	-0.062	1	2.799	0
11	1	2.992	0	2.992	1	2.477	0	-0.056	1	3.299	1
12	0	2.924	0	2.631	0	2.104	0	-0.066	1	2.488	0
13	0	3.082	0	2.704	0	2.491	0	0.012	1	2.746	0
14	1	3.338	1	3.226	1	2.412	0	-0.158	0	2.991	1
15	1	3.703	1	3.546	1	2.930	1	-0.115	0	3.417	1
16	1	3.562	1	3.213	1	2.779	1	-0.135	0	3.225	1
17	1	3.300	1	3.161	1	2.831	1	-0.105	0	3.292	1
18	1	3.367	1	3.215	1	2.705	1	-0.148	0	3.144	1
19	0	2.628	0	2.368	0	1.974	0	-0.039	1	2.177	0
20	0	3.008	0	2.802	0	2.188	0	-0.155	0	2.725	0
21	0	2.498	0	2.375	0	1.955	0	-0.030	1	2.329	0
22	1	3.578	1	3.285	1	3.000	1	-0.159	0	3.425	1
23	1	3.321	1	3.063	1	2.503	1	-0.020	1	3.207	1
24	1	3.411	1	3.141	1	2.570	1	-0.041	1	3.104	1
25	0	3.208	1	3.016	1	2.341	0	-0.175	0	2.897	0

Tables 3-28, 3-29, and 3-30 each show five trimmed Clinical + Genotype risk index models corresponding to the Clinical risk index models shown in Tables 3-22, 3-23, and 3-24. Tables 3-31, 3-32, and 3-33 show the risk index values and predictions from the same set of five Clinical + Genotype risk index models from the same three small-scale simulation datasets for a set of 25 individuals randomly selected from the optimization set. Figures 3-28, 3-29, and 3-30 show the full distribution of risk index values in the optimization set for a randomly selected trimmed Clinical + Genotype risk index model from each of the three small-scale simulation datasets. Figures 3-31, 3-32, and 3-33 show the full distribution of risk index values in the independent testing set for a randomly selected trimmed Clinical + Genotype risk index model from each of the three small-scale simulation datasets. As in the previous set of figures, in all six of these figures a red line indicates the cut-off point. All individuals with a risk index value greater than or equal to this cut-off point are predicted as “high risk” and all individuals with a value less than this cut-off point are predicted as “low risk”.

Table 3-29 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #5

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
2	0.004*v6 + 0.0336*PC7 + 0.0337*PC23 - 0.0028*PC30 + 0.0507*PC58 + 0.0357*PC66 - 0.0138*PC89 + 0.0299*PC94 + 0.0282*PC100 - 0.0062*PC106 - 0.016*PC113 + 0.0046*PC134 + 0.03*PC145 - 0.0269*PC198 - 0.0356*PC211 - 0.0345*PC226 - 0.0495*PC244 + 0.0469*PC258 + 0.0558*PC262 - 0.0123*PC274 + 0.0613*PC302
5	0.1644*v1 + 0.3554*v2 + 0.0146*PC34 + 0.0335*PC45 + 0.0491*PC49 - 0.0321*PC62 - 0.0484*PC77 - 0.0078*PC98 - 0.2477*PC143 - 0.0485*PC144 - 0.3438*PC169 - 0.0411*PC204 + 0.082*PC220 + 0.0639*PC228 - 0.0958*PC235 + 0.2721*PC261
44	0.14*v1 + 0.2395*v5 - 0.0079*v6 - 0.0236*v7 - 0.0163*PC16 + 0.0154*PC21 + 0.0104*PC33 + 0.0654*PC81 - 0.0028*PC99 - 6e-04*PC118 - 0.0194*PC119 + 0.1481*PC125 + 0.0388*PC163 + 0.1901*PC188 - 0.0313*PC217 + 0.0369*PC223 - 0.0824*PC232 + 0.0326*PC258 - 0.0472*PC265 + 0.3294*PC271 - 0.028*PC278 - 0.1468*PC289 - 0.0032*PC296 + 0.1262*PC301
83	0.1366*v1 + 0.3025*v2 + 0.1001*v4 - 0.0116*v6 - 0.0382*v7 - 0.0838*v8 + 0.009*PC25 + 0.3087*PC55 - 0.2883*PC80 + 0.1097*PC133 - 0.1625*PC172 - 0.2277*PC238 + 0.1097*PC260 + 0.1392*PC279
85	-0.079*v5 + 0.0414*PC1 + 0.0456*PC36 - 0.0147*PC48 - 0.0249*PC74 - 0.0231*PC81 + 0.0097*PC110 + 0.0211*PC118 + 0.0284*PC119 - 0.0349*PC159 + 0.014*PC167 + 0.0091*PC179 - 0.017*PC232 + 0.0437*PC234 + 0.044*PC260 + 0.0137*PC262 + 0.0083*PC278 - 0.1161*PC283 - 0.1134*PC293

Table 3-30 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #12

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
24	0.156*v1 + 0.3719*v2 + 0.1017*v4 + 0.1242*v5 - 0.006*v7 - 0.0213*v8 - 0.05*PC20 + 0.0208*PC31 + 0.0604*PC70 + 0.309*PC74 + 0.0226*PC90 - 0.0657*PC91 - 0.1191*PC92 - 0.0217*PC107 + 0.1019*PC152 + 0.0944*PC161 + 0.0321*PC177 - 0.055*PC215 + 0.0841*PC231 + 0.0799*PC240 + 0.0115*PC249 - 0.0425*PC262 + 0.018*PC272 - 0.2959*PC297
27	0.1348*v1 + 0.4034*v2 + 0.8165*v5 + 3e-04*v7 + 0.028*PC50 + 0.0627*PC57 - 0.0183*PC62 - 0.0188*PC111 - 0.1222*PC117 - 0.0167*PC157 - 0.0525*PC163 + 0.0596*PC181 + 0.1473*PC195 - 0.0069*PC210 + 0.1309*PC227 + 0.0691*PC233 + 0.0142*PC235 + 0.3777*PC288 + 0.0419*PC292 - 0.5166*PC298 + 0.0255*PC302 + 0.1219*PC304
37	-0.0194*v8 + 0.033*PC45 + 0.0713*PC93 + 0.1551*PC102 - 0.0501*PC104 + 0.036*PC122 - 0.0147*PC144 + 0.009*PC145 + 0.0787*PC159 - 0.0742*PC170 + 0.009*PC190 + 0.0354*PC193 - 0.0657*PC198 - 0.0797*PC215 + 0.0165*PC223 - 0.0071*PC260 + 0.0564*PC263
49	0.1114*v1 + 0.3982*v2 + 0.0657*v3 + 0.0895*v4 - 0.3266*v5 + 0.0495*v6 + 0.0056*v7 - 0.0054*v8 - 0.1036*PC26 + 0.1341*PC52 - 0.0534*PC75 - 0.0626*PC107 - 0.304*PC121 - 0.0061*PC128 - 0.1893*PC150 + 0.0491*PC151 - 0.0646*PC159 - 0.185*PC183 + 0.0439*PC191 - 0.4759*PC200 + 0.0296*PC217 + 0.1851*PC224 + 0.0286*PC231 + 0.2401*PC274 + 0.0341*PC304 + 0.0939*PC308
83	0.1531*v1 + 0.2988*v2 + 0.0804*v4 + 0.0115*v7 + 0.0084*PC7 - 0.0779*PC25 - 0.0164*PC31 - 0.0246*PC63 - 0.0237*PC67 + 0.0642*PC125 - 0.0172*PC148 + 0.1322*PC151 + 0.1042*PC157 + 0.0662*PC176 + 0.0523*PC177 - 0.2565*PC178 + 0.2854*PC195 - 0.0216*PC218 + 0.0512*PC236 - 0.0718*PC252 - 0.0471*PC262 + 0.0373*PC306

Table 3-31 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Small-scale Simulation Dataset #15

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
25	0.1271*v1 + 0.3612*v2 + 0.0798*v3 + 0.0501*v6 - 0.0156*s14 + 0.0383*s22 + 0.0094*s49 - 0.0231*s82 + 0.0235*s101 + 0.0356*s113 - 0.0424*s119 + 0.04*s168 - 0.0988*s201 - 0.2504*s255 - 0.0315*s273 - 0.0067*s375 + 0.0185*s385 + 0.0288*s420 - 0.0376*s424 - 0.0148*s458 + 0.0167*s479 + 0.0043*s491 + 0.2736*s496
76	0.3139*v2 - 0.5463*v5 - 0.0426*v8 + 0.4962*s12 + 0.0094*s18 + 0.0132*s56 + 0.0368*s65 + 0.0054*s106 - 0.0112*s124 - 0.0961*s141 - 0.0199*s143 - 0.0262*s214 - 0.1017*s218 - 0.0064*s240 + 0.1556*s257 - 0.0235*s289 + 0.0977*s310 + 0.0272*s311 + 0.1166*s405 + 0.0248*s436 + 0.0533*s485
79	0.1516*v1 + 0.0964*v3 + 0.0909*v4 + 0.0158*v6 - 6e-04*v7 + 0.1363*v8 - 0.1052*s46 - 0.0304*s86 + 0.1201*s158 - 0.096*s162 + 0.0155*s173 + 0.0827*s254 - 0.1274*s265 + 0.4236*s329 - 0.0118*s336 - 0.0026*s337 + 0.0312*s359 - 0.1349*s377 + 0.0325*s405 - 0.1539*s426 + 0.0686*s439 - 0.0048*s444 - 0.0332*s469
88	0.1149*v5 + 0.0023*v7 + 0.0135*s59 - 0.122*s166 - 7e-04*s179 + 0.0221*s195 + 0.0075*s251 - 0.0147*s265 - 0.0067*s352 - 0.0137*s360 + 0.0162*s374 - 0.0627*s384 + 0.0103*s388 - 0.0242*s391 + 0.0109*s406 - 0.0854*s413 + 0.0619*s440 + 0.0079*s448 + 0.0531*s461 + 0.0122*s464 + 0.0347*s495
92	0.1537*v1 + 0.2992*v2 + 0.0974*v3 + 0.0072*v6 - 0.1636*v8 + 0.0106*s51 - 0.2487*s75 - 0.0166*s84 - 0.1151*s90 + 0.0016*s123 - 0.0618*s133 - 0.0326*s136 + 0.0098*s166 - 0.0228*s186 - 0.0633*s193 + 0.0691*s207 + 0.0473*s212 + 0.0169*s215 + 0.0756*s269 + 0.1881*s275 + 0.0381*s366 - 0.0177*s422 + 0.0664*s435 + 0.1145*s475

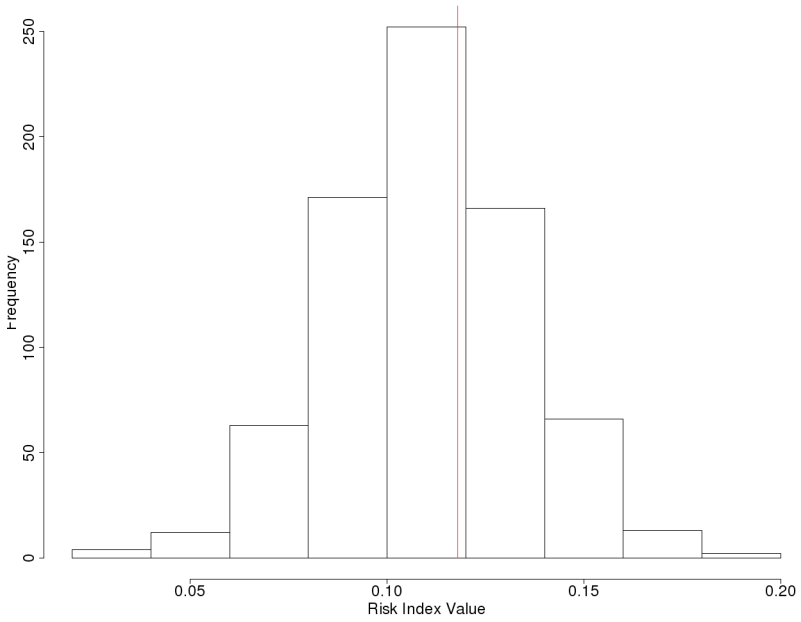


Figure 3-28 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #5, Bootstrap Sample #2

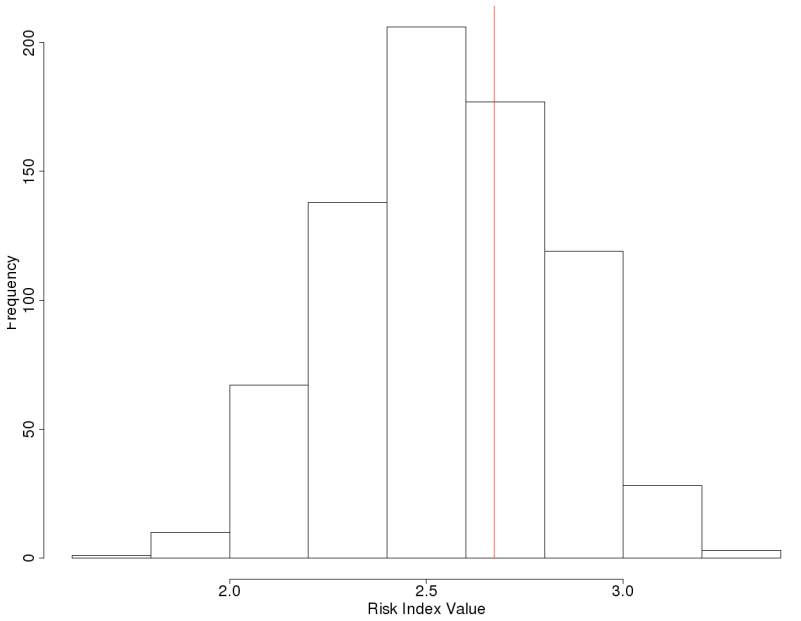


Figure 3-29 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #12, Bootstrap Sample #24

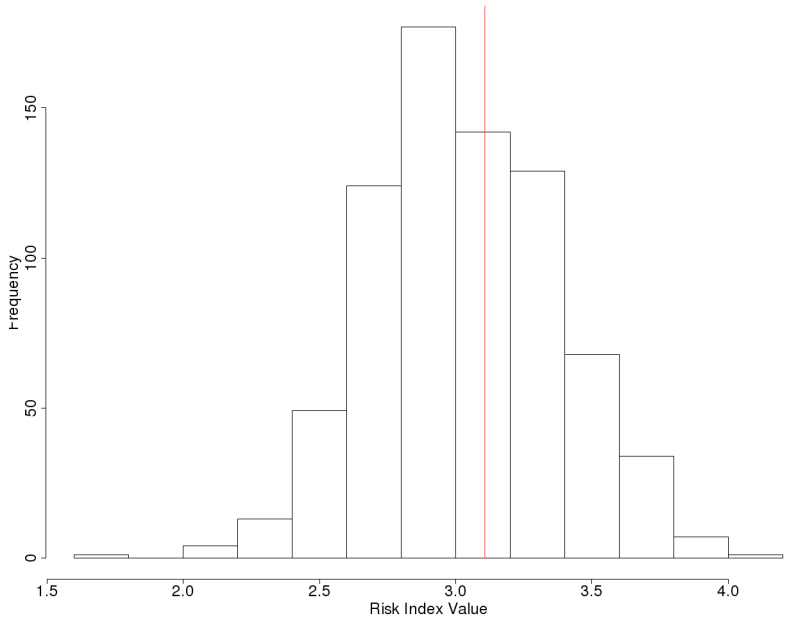


Figure 3-30 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Small-scale Dataset #15, Bootstrap Sample #25

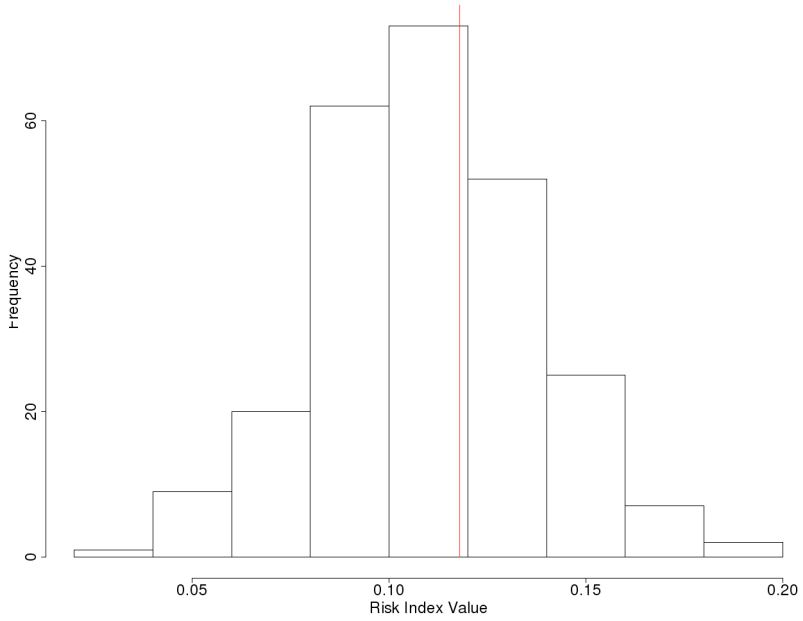


Figure 3-31 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #5, Bootstrap Sample #2

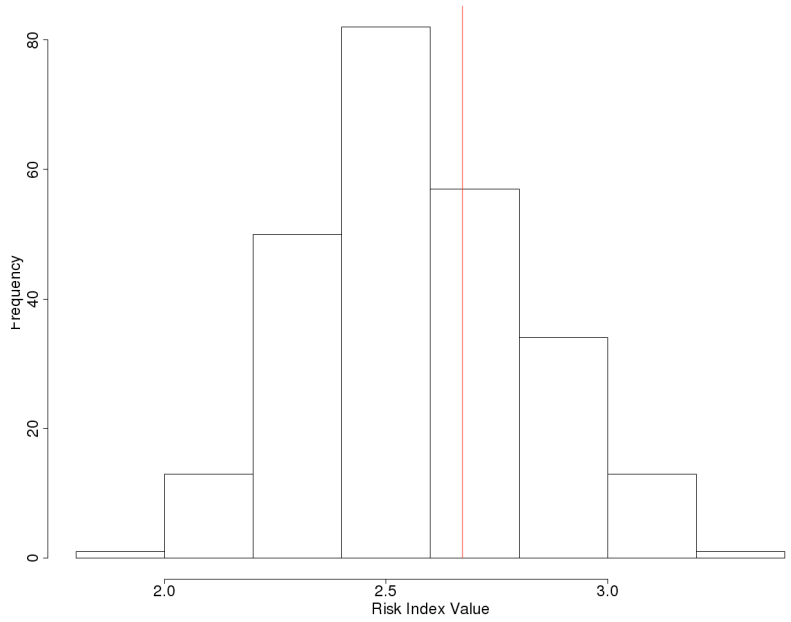


Figure 3-32 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #12, Bootstrap Sample #24

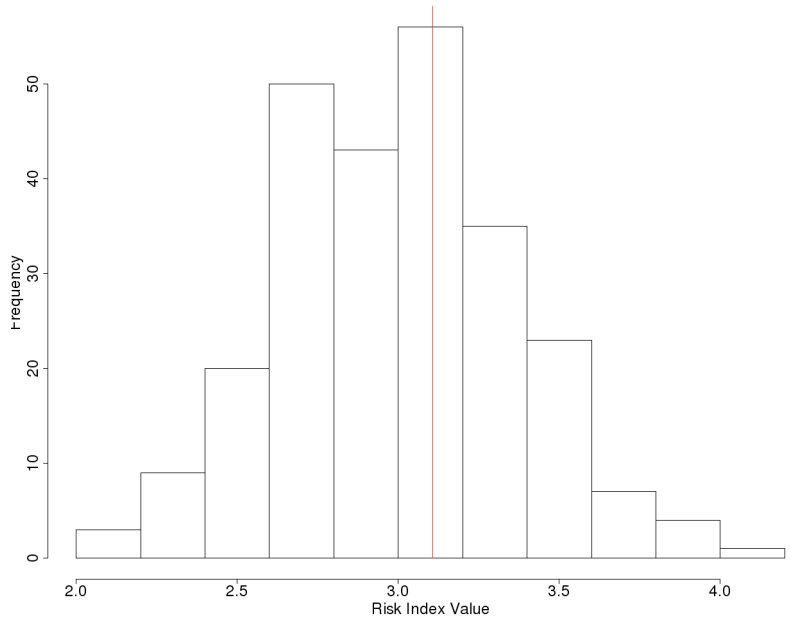


Figure 3-33 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Small-scale Dataset #15, Bootstrap Sample #25

Table 3-32 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Small-scale Simulation Dataset #5

		Bootstrap Sample #2 Cutoff Value = 0.118		Bootstrap Sample #5 Cutoff Value = 6.003		Bootstrap Sample #44 Cutoff Value = 2.225		Bootstrap Sample #83 Cutoff Value=1.712		Bootstrap Sample #85 Cutoff Value=-0.786	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	0.083	0	6.621	1	2.577	1	1.974	1	-0.792	0
2	1	0.155	1	6.230	1	2.445	1	1.872	1	-0.786	0
3	1	0.137	1	5.594	0	1.986	0	1.506	0	-0.771	1
4	0	0.141	1	5.027	0	1.888	0	1.534	0	-0.781	1
5	0	0.135	1	5.898	0	1.896	0	1.624	0	-0.816	0
6	0	0.091	0	4.876	0	1.655	0	1.244	0	-0.814	0
7	1	0.125	1	7.053	1	2.135	0	1.928	1	-0.741	1
8	1	0.125	1	5.823	0	1.885	0	1.917	1	-0.839	0
9	0	0.120	1	5.733	0	1.946	0	1.639	0	-0.735	1
10	0	0.093	0	5.426	0	1.983	0	1.602	0	-0.783	1
11	0	0.103	0	5.976	0	1.754	0	1.598	0	-0.815	0
12	1	0.086	0	6.563	1	2.295	1	1.852	1	-0.821	0
13	1	0.134	1	6.304	1	2.355	1	1.775	1	-0.810	0
14	0	0.114	0	5.962	0	2.133	0	1.663	0	-0.821	0
15	1	0.160	1	7.427	1	2.196	0	2.177	1	-0.795	0
16	0	0.110	0	5.490	0	2.093	0	1.404	0	-0.818	0
17	0	0.095	0	5.767	0	2.082	0	1.768	1	-0.814	0
18	0	0.141	1	4.540	0	1.808	0	1.256	0	-0.844	0
19	1	0.066	0	7.020	1	2.419	1	1.880	1	-0.767	1
20	1	0.077	0	6.980	1	2.268	1	1.933	1	-0.796	0
21	0	0.113	0	6.344	1	2.036	0	1.749	1	-0.823	0
22	1	0.126	1	6.254	1	2.127	0	1.664	0	-0.785	1
23	1	0.150	1	6.946	1	2.470	1	1.958	1	-0.765	1
24	0	0.117	0	5.452	0	1.889	0	1.628	0	-0.792	0
25	0	0.138	1	6.298	1	2.227	1	1.599	0	-0.827	0

Table 3-33 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Small-scale Simulation Dataset #12

		Bootstrap Sample #24 Cutoff Value = 2.672		Bootstrap Sample #27 Cutoff Value = 5.046		Bootstrap Sample #37 Cutoff Value = -0.216		Bootstrap Sample #49 Cutoff Value=1.992		Bootstrap Sample #83 Cutoff Value=3.737	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	2.510	0	4.750	0	-0.258	0	1.890	0	3.370	0
2	1	2.828	1	4.972	0	-0.220	0	2.123	1	3.726	0
3	0	2.551	0	4.572	0	-0.195	1	1.896	0	3.339	0
4	0	2.345	0	4.589	0	-0.192	1	1.672	0	3.076	0
5	1	2.699	1	4.705	0	-0.217	0	1.980	0	3.622	0
6	1	2.838	1	5.199	1	-0.257	0	2.212	1	3.779	1
7	0	2.713	1	4.910	0	-0.249	0	1.946	0	3.479	0
8	1	2.927	1	5.059	1	-0.238	0	2.175	1	3.900	1
9	1	2.853	1	5.206	1	-0.202	1	2.150	1	3.834	1
10	1	2.912	1	5.206	1	-0.235	0	2.205	1	3.890	1
11	0	2.300	0	4.243	0	-0.228	0	1.599	0	2.999	0
12	0	2.239	0	4.200	0	-0.222	0	1.463	0	2.941	0
13	1	3.166	1	5.641	1	-0.257	0	2.233	1	4.090	1
14	0	2.388	0	4.252	0	-0.214	1	1.649	0	3.036	0
15	1	2.882	1	5.219	1	-0.187	1	2.156	1	3.875	1
16	0	2.509	0	4.479	0	-0.216	1	1.812	0	3.258	0
17	0	2.528	0	4.851	0	-0.227	0	1.761	0	3.415	0
18	0	2.526	0	4.672	0	-0.195	1	1.920	0	3.329	0
19	0	2.690	1	5.033	0	-0.245	0	1.940	0	3.688	0
20	1	2.283	0	4.392	0	-0.225	0	1.500	0	2.939	0
21	0	2.083	0	3.829	0	-0.258	0	1.547	0	2.743	0
22	1	2.985	1	5.247	1	-0.249	0	2.322	1	3.985	1
23	1	2.894	1	5.206	1	-0.240	0	2.237	1	3.826	1
24	0	2.812	1	4.807	0	-0.252	0	2.138	1	3.751	1
25	0	2.290	0	4.442	0	-0.169	1	1.625	0	2.990	0

Table 3-34 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Small-scale Simulation Dataset #15

		Bootstrap Sample #245 Cutoff Value = 3.107		Bootstrap Sample #76 Cutoff Value = 2.783		Bootstrap Sample #79 Cutoff Value = 2.466		Bootstrap Sample #88 Cutoff Value=-0.094		Bootstrap Sample #92 Cutoff Value=2.902	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	3.178	1	2.832	1	2.541	1	-0.059	1	3.174	1
2	1	3.305	1	2.771	0	2.702	1	-0.084	1	2.929	1
3	1	2.940	0	2.667	0	2.325	0	-0.083	1	3.022	1
4	1	3.302	1	3.045	1	2.679	1	-0.159	0	3.002	1
5	1	3.301	1	2.844	1	2.784	1	-0.109	0	3.218	1
6	0	2.467	0	2.263	0	2.020	0	-0.091	1	2.488	0
7	1	3.429	1	3.066	1	2.620	1	-0.122	0	3.306	1
8	0	2.722	0	2.574	0	2.219	0	-0.154	0	2.541	0
9	1	3.617	1	3.279	1	3.072	1	-0.070	1	3.411	1
10	0	3.042	0	2.742	0	1.961	0	-0.066	1	2.893	0
11	1	2.924	0	2.865	1	2.458	0	-0.070	1	3.278	1
12	0	2.910	0	2.432	0	2.076	0	-0.073	1	2.477	0
13	0	3.057	0	2.531	0	2.481	1	-0.001	1	2.677	0
14	1	3.289	1	3.170	1	2.362	0	-0.172	0	3.004	1
15	1	3.690	1	3.511	1	2.893	1	-0.124	0	3.419	1
16	1	3.533	1	3.162	1	2.758	1	-0.148	0	3.194	1
17	1	3.309	1	3.037	1	2.820	1	-0.109	0	3.235	1
18	1	3.340	1	3.144	1	2.684	1	-0.156	0	3.072	1
19	0	2.566	0	2.171	0	1.924	0	-0.053	1	2.238	0
20	0	3.022	0	2.565	0	2.172	0	-0.167	0	2.704	0
21	0	2.445	0	2.248	0	1.919	0	-0.035	1	2.300	0
22	1	3.570	1	3.204	1	2.954	1	-0.167	0	3.403	1
23	1	3.324	1	2.938	1	2.457	0	-0.033	1	3.214	1
24	1	3.402	1	3.086	1	2.514	1	-0.050	1	3.109	1
25	0	3.196	1	2.951	1	2.287	0	-0.181	0	2.898	0

3.3.3 Predictive Performance

After the variable selection procedure is completed and the models are applied to each individual in the independent testing set then the sensitivity, specificity, misclassification, and positive predictive value are estimated for both the Clinical and Clinical + Genotype risk index models for each of the 100 small-scale simulation datasets. Table 3-34 shows the means and standard deviations of these measurements. To provide a 95% confidence for these measurements of sensitivity, specificity, misclassification, and positive predictive value for each independent testing set, 1000 bootstrap samples were generated. By making predictions about each individual in these bootstrap samples and calculating the sensitivity, specificity, misclassification, and positive predictive value for each bootstrap sample, 95% confidence intervals were estimated for these measurements in each of the 100 small-scale simulation datasets. The mean and standard deviation of the spread (i.e., range) of these confidence intervals for both the Clinical and Clinical + Genotype risk index model is shown in Table 3-35. This provides a view into the variability of the sensitivity, specificity, misclassification, and positive predictive value estimates. Table 3-36 shows the predictive performance and confidence intervals for the three small-scale simulation datasets discussed in Section 3.2.2.

Table 3-35 Means and Standard Deviations of Predictive Performance Estimates for the 100 Small-scale Simulation Datasets

Model	Sensitivity (SD)	Specificity (SD)	Misclassification (SD)	PPV (SD)	AUC (SD)
Clinical	0.607 (0.066)	0.883 (0.032)	0.201 (0.027)	0.607 (0.067)	0.826 (0.033)
Clinical + Genotype	0.590 (0.063)	0.891 (0.031)	0.200 (0.026)	0.708 (0.067)	0.839 (0.032)

Table 3-36 Means and Standard Deviations of Predictive Performance 95% Confidence Intervals for the 100 Small-scale Simulation Datasets

Model	Mean Range of the 95% Confidence Interval (SD)			
	Sensitivity	Specificity	Misclassification	PPV
Clinical	0.218 (0.017)	0.094 (0.013)	0.098 (0.007)	0.220 (0.022)
Clinical + Genotype	0.220 (0.016)	0.091 (0.013)	0.099 (0.006)	0.222 (0.022)

Table 3-37 Predictive Performance Estimates for Three Small-scale Simulation Datasets

Dataset	Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
5	Clinical	0.553 (0.44-0.658)	0.891 (0.841-0.935)	0.211 (0.163-0.259)	0.689 (0.565-0.803)
	Clinical + Genotype	0.539 (0.429-0.644)	0.897 (0.849-0.938)	0.211 (0.163-0.259)	0.695 (0.574-0.809)
12	Clinical	0.556 (0.437-0.676)	0.911 (0.866-0.95)	0.191 (0.143-0.243)	0.714 (0.581-0.83)
	Clinical + Genotype	0.569 (0.451-0.69)	0.922 (0.88-0.96)	0.179 (0.135-0.227)	0.745 (0.612-0.86)
15	Clinical	0.526 (0.412-0.643)	0.919 (0.878-0.956)	0.203 (0.151-0.255)	0.745 (0.635-0.848)
	Clinical + Genotype	0.526 (0.413-0.635)	0.925 (0.882-0.96)	0.199 (0.155-0.251)	0.759 (0.648-0.862)

Using the number of models predicting an individual in the independent testing set as "high risk", receiver operator characteristic (ROC) curves were generated for the Clinical and Clinical + Genotype risk index model for each of the 100 small-scale simulation datasets, and the AUC for the ROC curve was estimated. The average AUC of the Clinical risk index models was 0.826 (SD = 0.033), and the average AUC of the Clinical + Genotype risk index models was 0.839 (SD = 0.032). Figure 3-32, 3-33, and 3-34 show the ROC curves of the Clinical risk index model the three small-scale simulation datasets discussed in sections 3.2.2, and Figure 3-35, 3-36, and 3-37 show the ROC curve for the Clinical + Genotype risk index model from those three selected small-scale simulation datasets.

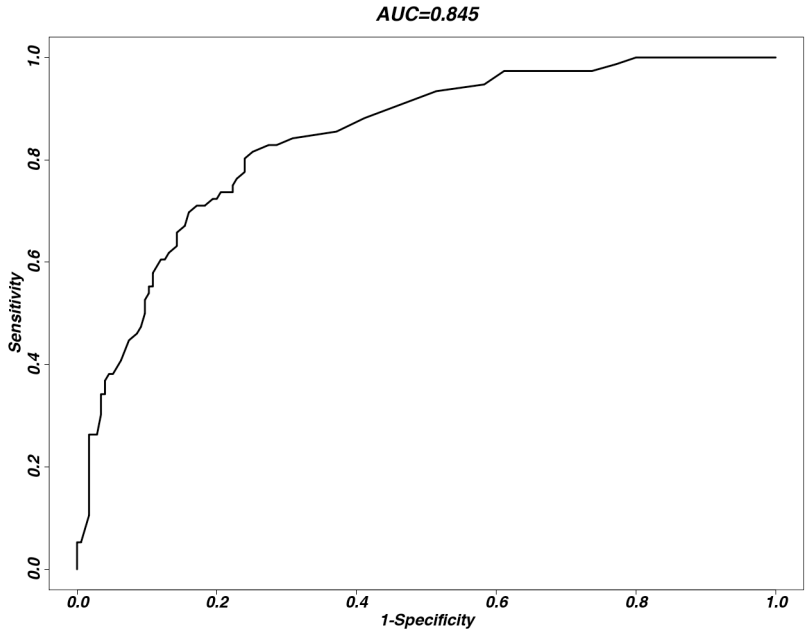


Figure 3-34 ROC Curve of the Clinical Risk Index Model for Small-scale Simulation Dataset #5

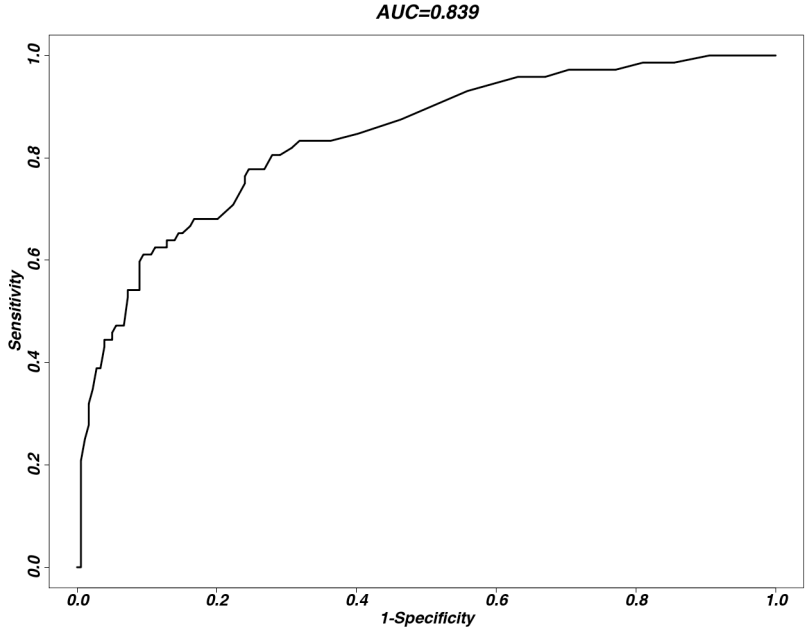


Figure 3-35 ROC Curve of the Clinical Risk Index Model for Small-scale Simulation Dataset #12

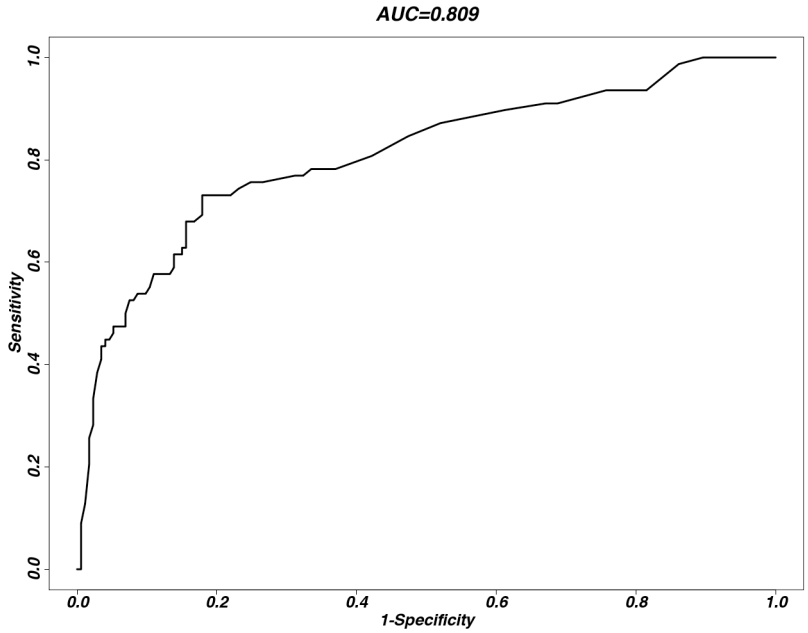


Figure 3-36 ROC Curve of the Clinical Risk Index Model for Small-scale Simulation Dataset #15

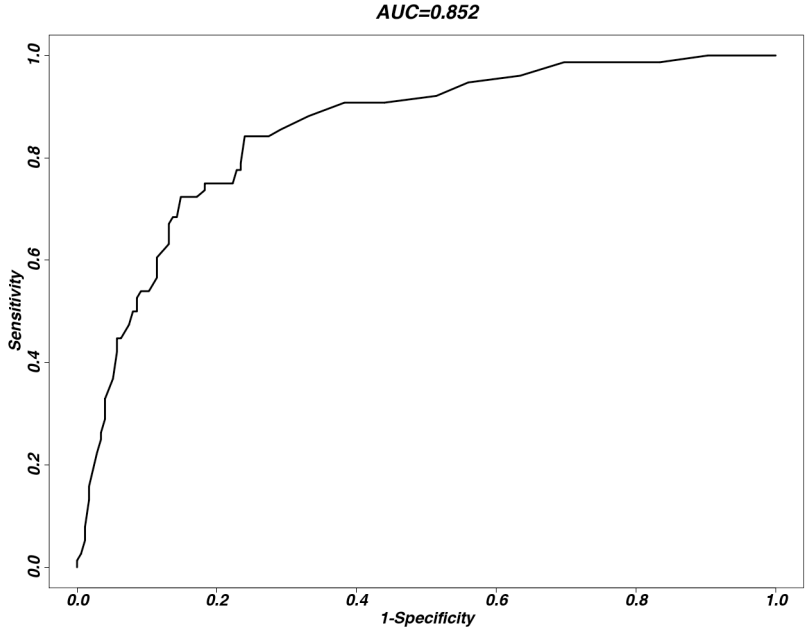


Figure 3-37 ROC Curve of the Clinical + Genotype Risk Index Model for Small-scale Simulation Dataset #5

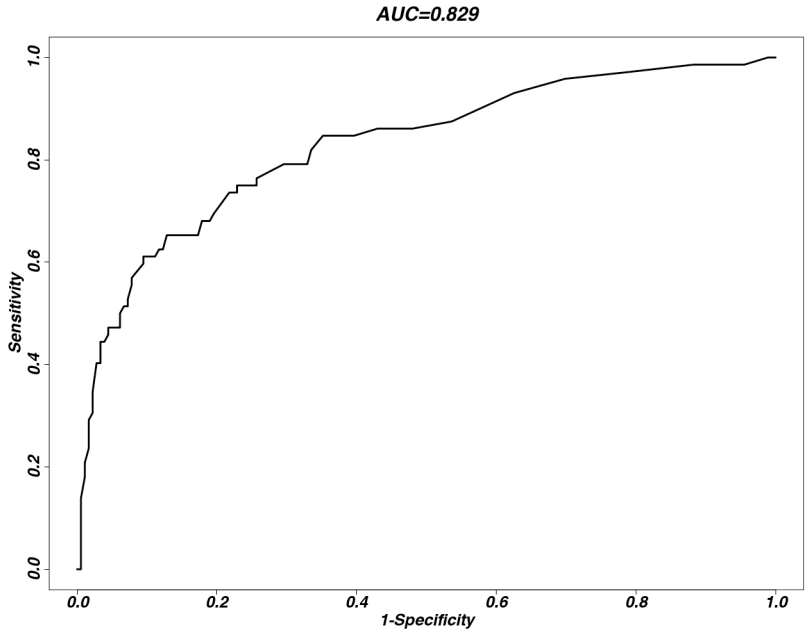


Figure 3-38 ROC Curve of the Clinical + Genotype Risk Index Model for Small-scale Simulation Dataset #12

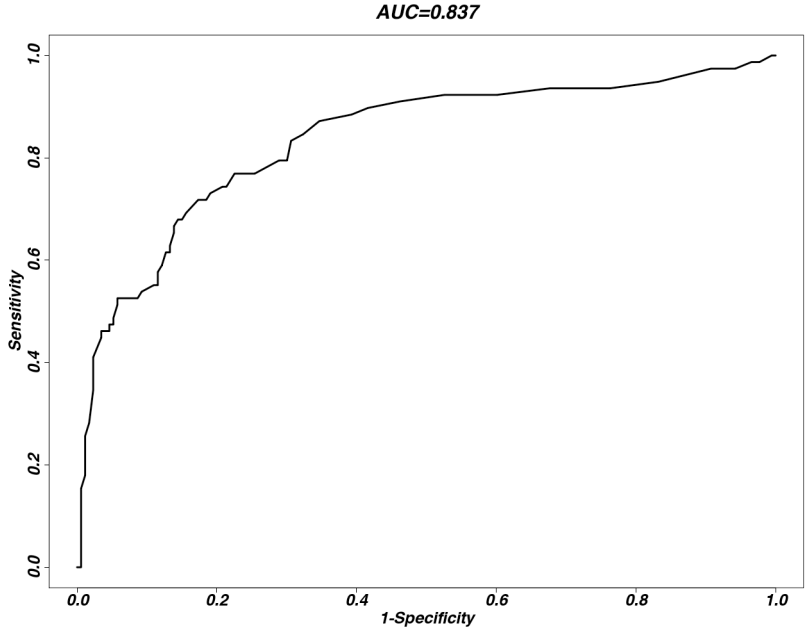


Figure 3-39 ROC Curve of the Clinical + Genotype Risk Index Model for Small-scale Simulation Dataset #15

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. The proportion of models that predict that an individual as high risk, then, represents the predicted probability of an individual developing the outcome. For each individual a 95% confidence interval can be constructed as described in Section 3.2.3

3.3.4 Random Forest Comparison

For each of the 100 small-scale simulation datasets a random forest was generated using the optimization set created by the risk index procedure. The forests were generated using the methodology given in Section 3.2.4. For each of the random forests an ROC curve was generated and the AUC was estimated. The mean AUC of the random forest models was 0.821 (SD = 0.031). Figures 3-40, 3-41, and 3-42 show the ROC curve of the random forest generated from the three small-scale simulation datasets described in Section 3.2.2. When working with a dataset that has two possible classes, the standard procedure for a random forest is to assign a prediction to an individual based on a simple majority of votes. When the prevalence of the outcome is less than 50% changing the proportion of votes needed to classify an individual can significantly impact the estimates of performance. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set using a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of 5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the individual to be "high risk". Table 3-37 shows the

mean and standard deviation of the sensitivity, specificity, misclassification, and PPV for a range of different proportions.

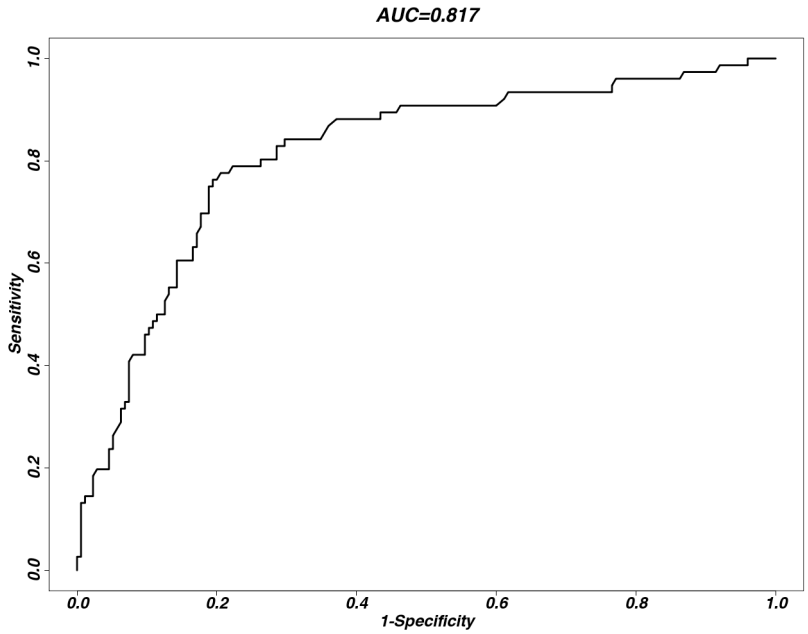


Figure 3-40 ROC Curve of the Random Forest Generated for Small-scale Simulation Dataset #5

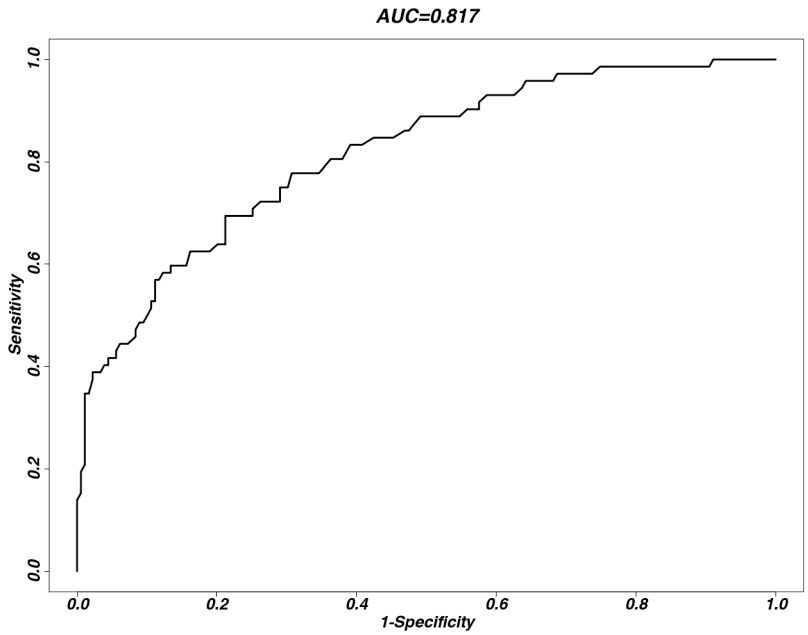


Figure 3-41 ROC Curve of the Random Forest Generated for Small-scale Simulation Dataset #12

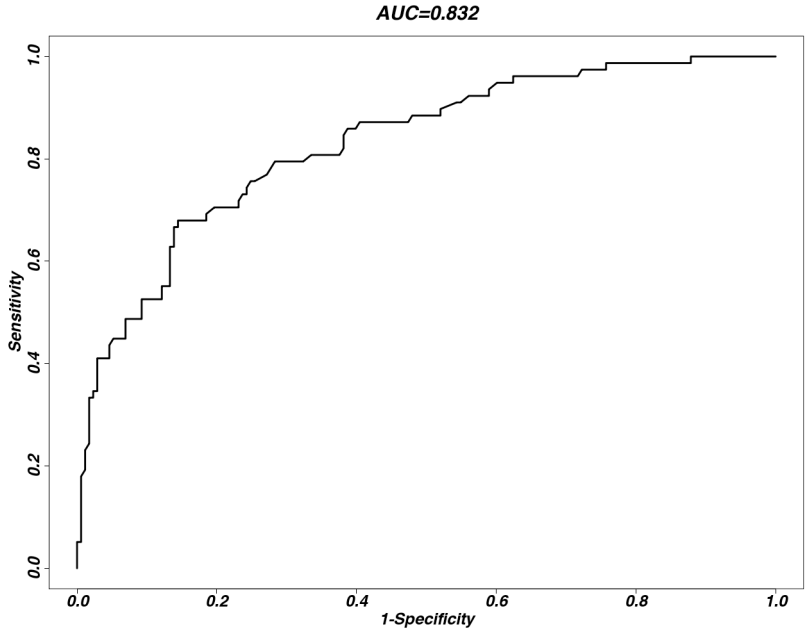


Figure 3-42 ROC Curve of the Random Forest Generated for Small-scale Simulation Dataset #15

Table 3-38 Means and Standard Deviations of Performance Estimates of the Random Forest Models Generated from the 100 Small-scale Simulation Datasets

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	0.991 (0.014)	0.098 (0.073)	0.629 (0.056)	0.327 (0.038)
0.1	0.963 (0.027)	0.264 (0.1)	0.522 (0.069)	0.368 (0.046)
0.15	0.921 (0.039)	0.433 (0.086)	0.417 (0.058)	0.42 (0.051)
0.2	0.859 (0.047)	0.577 (0.064)	0.337 (0.043)	0.474 (0.053)
0.25	0.792 (0.053)	0.693 (0.053)	0.276 (0.035)	0.535 (0.059)
0.3	0.716 (0.064)	0.768 (0.049)	0.247 (0.031)	0.579 (0.064)
0.35	0.631 (0.076)	0.832 (0.045)	0.229 (0.028)	0.627 (0.069)
0.4	0.534 (0.088)	0.884 (0.043)	0.223 (0.027)	0.678 (0.083)
0.45	0.447 (0.097)	0.923 (0.034)	0.223 (0.026)	0.729 (0.086)
0.5	0.345 (0.101)	0.953 (0.027)	0.233 (0.029)	0.779 (0.092)
0.55	0.238 (0.095)	0.975 (0.019)	0.251 (0.033)	0.82 (0.095)
0.6	0.134 (0.081)	0.99 (0.013)	0.273 (0.035)	0.869 (0.12)
0.65	0.048 (0.053)	0.997 (0.007)	0.294 (0.035)	0.903 (0.185)
0.7	0.011 (0.025)	0.999 (0.002)	0.304 (0.034)	0.879 (0.184)

3.3.5 Conclusion

The performance of the small-scale simulation tests using principal components that explain 90% of the variance performed comparably to the small-scale tests using the full set of 100 SNPs. This is not surprising, as the principal components analysis was performed using the set of 500 SNPs examined earlier, and so the principal components are just effectively a compression of the information contained in the SNPs into a smaller number of uncorrelated variables. The specification of no linkage disequilibrium between the simulated SNPs also helps explain why a fairly large number of components are needed to reach 90% of the variance. The performance of the random forests, however, is reduced when using the top principal components. This is likely because the tree-based nature of the random forest method helps identify context-dependent relationships among variables that can be used for classification. Because the principal components are

uncorrelated with each other, however, a random forest is not as effective as in a dataset with significant correlation among variables.

The predictive performance of the risk index procedure is quite good, with a mean AUC of the Clinical + Genotype risk index models that is significantly greater than the mean AUC of the Clinical risk index model ($p=0.008$). Unlike the small-scale simulation tests using the full set of SNPs, the random forest models had a mean AUC that was significantly lower than that of the Clinical + Genotype risk index models ($p=8.4e-5$), and tuning the class assignment procedure can produce sensitivity, specificity, and positive predictive value that is comparable to the risk index methods but does not exceed it.

3.4 Large-scale Simulation Study Methodology

The large-scale simulation study is made up of ten thousand individuals, twenty-nine covariates, and 38,835 polymorphisms. The outcome is generated in the same manner as the small-scale simulation study, and as before a multivariate normal random number generator is used. For the large-scale simulation study, however, rather than specifying the precise correlation matrix for the variables, they were split into blocks, with each block having a range of possible correlations with the outcome and with the other variables. The precise correlation between each variable and the outcome was modeled as a uniformly distributed random variable that takes a value within the range specified for the particular variable. The first and second variables have between a 0.45 and 0.65 correlation with the outcome, the third through fifth variables have a correlation with the outcome of between 0.40 and 0.25, the sixth through sixteenth variables have between a

0.10 and 0.25 correlation with the outcome, and variables seventeen through twenty-nine have between a 0.01 and 0.09 correlation with the outcome. The correlation between variables is simulated as a normal random variable with a mean of 0, and a standard deviation of 0.2, which gives a 99% probability the correlations will be between -0.52 and 0.52, and a 99.9999% probability that the correlation will be between -1.0 and 1.0.

Genotype variable simulation was performed using genomeSIMLA (Edwards, et al, 2008) as with the small-scale simulation study. Using configuration files provided by the authors of genomeSIMLA, 38,835 genotypes were generated so that the final dataset is similar to data that would be obtained from Chromosome one using the Affymetrix 500K Genome-wide genotyping assay (Affymetrix, 2007). Six SNPs were selected as associated with the outcome, with beta coefficients ranging from 0.4 to 0.8, corresponding to an odds ratio at a given locus of between 1.5 and 2.2. As with the small-scale simulation genotypes were encoded additively.

3.5 Large-scale Simulation Study Top 500 SNPs Results

3.5.1 Variable Selection

Using the same procedure as in Section 3.2.1, Clinical and Clinical + Genotype risk index models were constructed for each of the 25 large-scale simulation datasets. The association between the dichotomous outcome and each of the 38,835 SNPs was estimated, and the 500 SNPs with the smallest p-values from this logistic regression analysis were used to construct the risk index.

Table 3-38 shows the summary of the variable selection procedure from the Clinical risk index model averaged across the 25 simulation datasets. Variables v1 through v5 are most frequently selected; on average, they each appear in more than half of the 50 trimmed Clinical risk index models. Variables v12, v15, and v16 are also frequently selected, appearing in 17.7, 17.1, and 16.7 trimmed Clinical risk index models on average. Table 3-39 shows the summary of the variable selection procedure from the Clinical + Genotype risk index model averaged across the 25 simulation datasets. No SNP was in more than 7.29 out of 50 Clinical + Genotype risk index models on average.

Table 3-39 Summary of the Number of Times Each Variable is Selected into a Specific Model Position for the Large-scale Simulation Clinical Risk Index Models

Variable	Variable Position 1	2	3	4	5	6	7	8	Total # of Times in Trimmed Model
v1	31.286	0.429	0.286	0.000	0.286	0.000	0.000	0.000	32.429
v2	0.143	8.000	5.571	3.571	3.714	2.714	2.714	1.571	29.286
v3	0.286	7.286	3.857	3.286	2.714	2.000	1.143	0.857	22.143
v4	0.286	7.000	9.857	3.571	1.857	1.714	2.143	1.714	30.000
v5	0.143	7.286	5.857	7.143	2.571	1.714	1.429	1.143	29.000
v6	0.143	0.714	1.571	2.571	2.714	3.000	2.000	1.571	16.429
v7	0.000	0.286	1.000	0.429	1.571	2.000	1.143	1.857	12.857
v8	0.000	0.286	1.000	1.429	2.857	2.143	1.286	1.286	13.429
v9	0.000	0.571	0.143	1.286	2.571	1.429	2.714	2.857	16.571
v10	0.000	0.143	0.143	0.571	1.000	1.286	0.571	2.429	9.286
v11	0.143	0.286	0.143	0.857	1.000	1.571	1.714	1.429	10.571
v12	0.000	0.286	0.714	2.429	2.714	3.857	3.571	2.429	17.714
v13	0.714	0.143	1.571	1.571	1.429	1.286	1.000	1.000	8.714
v14	0.000	0.143	0.143	0.429	0.429	1.429	1.000	1.429	9.143
v15	0.000	0.000	0.714	1.714	1.286	2.714	3.286	3.286	17.143
v16	0.000	0.000	0.143	1.286	2.000	2.143	3.286	3.143	16.714
v17	2.286	1.857	2.143	1.286	2.143	2.143	1.286	1.286	6.714
v18	0.571	1.143	0.429	1.143	2.143	1.429	2.000	2.000	5.000
v19	0.714	1.143	1.571	2.429	2.857	1.571	1.286	1.857	6.571
v20	1.143	1.429	1.000	1.286	1.571	1.857	2.286	1.714	5.000
v21	1.143	1.286	1.000	1.429	0.429	1.857	1.571	1.857	6.714
v22	1.857	1.571	1.429	1.286	1.857	1.143	1.429	0.714	5.286
v23	1.286	1.429	1.000	1.857	1.286	0.857	1.429	1.000	4.857
v24	2.143	2.286	1.286	1.429	1.143	1.286	1.571	2.143	6.286
v25	1.286	0.714	0.714	0.714	1.714	0.714	1.857	1.714	5.143
v26	1.000	0.571	2.000	1.000	1.000	0.857	1.571	1.286	5.714
v27	2.000	2.571	1.857	1.286	1.143	1.429	1.571	1.571	6.857
v28	0.857	0.571	1.429	1.429	1.000	2.143	1.857	2.857	6.429
v29	0.571	0.571	1.429	1.286	1.000	1.714	1.286	2.000	4.000

Table 3-40 Summary of the Number of Times Selected Genotype Variables are Selected into a Specific Model Position for the Large-scale Simulation Clinical + Genotype Risk Index Models

SNP	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
rs1552124	0.00	1.00	0.14	0.57	0.57	1.00	1.29	0.86	0.71	0.43	0.43	1.14	0.57	0.86	0.71	0.43	0.14	0.29	0.14	0.29	11.57	7.29
rs942904	0.14	0.43	0.57	1.14	0.29	0.29	0.57	0.29	0.57	1.00	0.86	0.86	0.43	0.57	0.71	0.57	0.43	0.43	0.43	0.29	10.86	7.00
rs10788668	0.14	0.43	0.43	0.43	0.86	0.29	0.86	0.71	0.57	0.57	0.14	0.71	0.57	0.43	0.29	0.43	0.71	0.86	0.29	0.14	9.86	6.14
rs12141159	0.00	0.14	0.29	0.29	0.43	1.29	0.57	1.00	0.29	0.86	1.00	0.57	0.57	1.00	0.57	0.86	0.57	0.29	0.14	0.43	11.14	5.57
rs3795479	0.14	0.14	0.14	0.71	0.29	0.71	0.57	0.57	0.71	0.86	0.57	1.29	0.43	0.86	0.43	0.43	0.71	0.00	0.43	0.57	10.57	5.57
rs2241863	0.00	0.57	0.43	0.71	0.43	0.71	1.00	0.29	0.43	0.14	0.14	0.14	0.71	0.29	0.14	0.43	0.43	0.29	0.29	0.00	7.57	5.29
rs2157381	0.00	0.29	0.29	0.43	0.29	0.29	0.14	0.14	0.43	0.86	0.86	0.57	0.14	0.71	0.43	0.71	0.71	0.43	0.71	0.43	8.86	4.86
rs2566753	0.14	0.29	0.29	0.14	0.57	0.43	1.00	0.57	0.00	0.43	0.57	0.14	0.71	0.43	0.57	0.14	0.29	0.29	0.43	0.43	7.86	4.71
rs1373259	0.00	0.00	0.57	0.29	0.29	0.29	0.43	0.57	0.71	0.71	0.57	0.57	0.86	0.43	0.86	0.43	0.71	0.29	0.43	0.71	9.71	4.43
rs7514435	0.14	0.86	0.43	0.14	0.71	0.57	0.29	0.14	0.14	0.29	0.71	0.43	0.14	0.43	0.14	0.29	0.00	0.57	0.00	0.43	6.86	4.43
rs10157886	0.00	0.00	0.00	0.29	0.57	0.57	0.29	0.29	0.29	0.86	0.14	0.43	0.29	0.86	0.43	0.43	1.14	0.00	0.57	0.29	7.71	4.29
rs12141268	0.00	0.00	0.71	0.29	0.43	0.43	0.29	0.29	0.43	0.29	0.29	0.57	0.43	0.29	0.57	0.14	0.14	0.29	0.14	0.29	6.29	4.00
rs2494454	0.29	0.14	0.57	0.14	0.43	0.00	0.14	0.00	0.71	0.29	0.14	0.00	0.57	0.57	0.14	0.29	0.29	0.29	0.14	0.29	5.43	4.00
rs4916041	0.00	0.29	0.29	0.14	0.29	0.43	0.29	0.43	0.43	0.43	0.14	0.86	0.00	0.43	0.14	0.14	0.14	0.71	0.14	0.14	5.86	4.00
rs6682150	0.29	0.43	0.29	0.29	0.00	0.14	0.29	0.14	0.14	0.14	0.00	0.43	0.14	0.29	0.57	0.29	0.43	0.00	0.43	0.29	5.00	3.57
rs10913043	0.14	0.00	0.14	0.43	0.00	0.29	0.29	0.43	1.14	0.29	0.57	0.29	0.57	0.86	0.00	0.29	0.14	0.14	0.57	0.43	7.00	3.57
rs3009947	0.14	0.29	0.14	0.29	0.14	0.14	0.00	0.14	0.43	0.29	0.43	0.29	0.00	0.29	0.29	0.14	0.14	0.14	0.71	0.43	4.86	3.43
rs1797052	0.00	0.00	0.29	0.29	0.14	0.43	0.14	0.71	0.14	0.14	0.29	0.29	0.29	0.14	0.14	0.14	0.43	0.14	0.00	0.43	4.57	3.43
rs12047608	0.00	0.14	0.14	0.14	0.29	0.00	0.29	0.14	0.43	0.14	0.57	0.29	0.29	0.86	0.57	0.29	0.00	0.71	1.00	0.29	6.57	3.14
rs6427160	0.00	0.14	1.00	0.14	0.00	0.43	0.14	0.14	0.29	0.00	0.29	0.14	0.14	0.29	0.00	0.14	0.43	0.14	0.00	0.14	4.00	3.14
rs6693453	0.14	0.43	0.14	0.14	0.14	0.00	0.14	0.57	0.29	0.43	0.00	0.29	0.14	0.29	0.00	0.00	0.14	0.00	0.00	0.14	3.43	3.14

3.5.2 Models

Once the variable selection procedure is finished each of the 25 large-scale simulation datasets have 50 trimmed Clinical and Clinical + Genotype risk index models. Tables 3-40, 3-41, and 3-42 each show five trimmed Clinical risk index models randomly selected from one of three randomly chosen large-scale simulation datasets (datasets #9, #22, and #25). Figures 3-43, 3-44, and 3-45 show the full distribution of risk index values in the optimization set for a randomly selected trimmed Clinical risk index models from each of the three large-scale simulation datasets. Figures 3-46, 3-47, and 3-48 show the full distribution of risk index values in the independent testing set for a randomly selected trimmed Clinical risk index model from each of the three large-scale simulation datasets. In all six of these figures a red line indicates the cut-off point. All individuals with a risk index value greater than or equal to this cut-off point are predicted as "high risk" and all individuals with a value less than this cut-off point are predicted as "low risk". Tables 3-43, 3-44, and 3-45 show the risk index values and predictions from the same set of five Clinical risk index models from the same three large-scale simulation datasets for a set of 25 individuals randomly selected from the optimization set.

Table 3-41 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #9

Bootstrap Sample	Trimmed Clinical Risk Index Model
5	0.0285*v1 + 0.0289*v2 + 0.0133*v4 + 0.0042*v5 + 0.0025*v6 - 0.03*v8 + 0.0029*v12 + 0.0025*v16 + 0.0347*v26 + 0.001*v27
12	0.0299*v1 + 0.0254*v2 + 0.0122*v4 + 0.0045*v5 + 0.0026*v6 - 0.0313*v8 + 0.0035*v11 + 0.0027*v12 + 0.0021*v16 + 0.0011*v27
14	0.0291*v1 + 0.0279*v2 + 0.023*v3 + 0.0129*v4 + 0.0041*v5 + 0.0025*v6 - 0.0265*v8 + 0.0058*v9 + 0.0029*v11 + 0.0032*v12
20	0.0292*v1 + 0.0294*v2 + 0.0136*v4 + 0.0038*v5 + 0.0024*v6 + 0.0038*v7 - 0.0292*v8 + 0.0056*v9 + 0.0027*v11 + 0.0027*v12
23	0.0307*v1 + 0.0259*v2 + 0.0129*v4 + 0.0043*v5 + 0.0026*v6 - 0.0276*v8 + 0.0025*v11 + 0.0033*v12 + 0.0021*v16 - 0.0022*v29

Table 3-42 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #22

Bootstrap Sample	Trimmed Clinical Risk Index Model
11	0.0325*v1 + 0.0109*v2 + 0.0024*v3 + 0.0047*v4 + 0.0029*v5 + 0.0052*v6 + 0.0036*v7 + 0.0024*v9 + 0.0048*v14 - 0.0197*v28
15	0.003*v5 + 0.002*v8 + 0.0177*v15 + 0.0013*v17 - 0.0041*v19 - 3e-04*v20 - 0.0027*v21 - 9e-04*v24 - 0.0058*v25 - 1e-04*v27
23	0.0313*v1 + 0.0112*v2 + 0.0028*v3 + 0.0056*v4 + 0.003*v5 + 0.0055*v6 + 0.0027*v7 - 0.0115*v11 + 0.018*v15 - 0.0181*v28
24	-0.0001*v23
25	-0.0002*v22

Table 3-43 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #25

Bootstrap Sample	Trimmed Clinical Risk Index Model
4	-0.0422*v1 + 0.0359*v2 + 0.0096*v3 - 0.0208*v4 - 0.0136*v5 + 0.0044*v8 - 0.0223*v12 + 0.0052*v13 + 0.0069*v15 + 6e-04*v21
5	-0.0024*v17 + 4e-04*v26 + 3e-04*v27
9	-0.0425*v1 + 0.0334*v2 + 0.0088*v3 - 0.0213*v4 - 0.0137*v5 + 0.004*v8 - 0.0185*v12 + 0.0052*v13 - 0.0129*v14 + 0.0082*v15
18	-0.0386*v1 + 0.0378*v2 + 0.0098*v3 - 0.0206*v4 - 0.013*v5 + 0.0047*v8 - 0.012*v10 - 0.023*v12 - 0.0118*v14 + 0.008*v15
21	-0.0417*v1 + 0.0336*v2 + 0.0091*v3 + 0.0043*v6 + 0.0026*v7 + 0.0378*v9 - 0.0207*v12 + 0.0056*v13 - 0.0016*v19 + 0.0064*v23

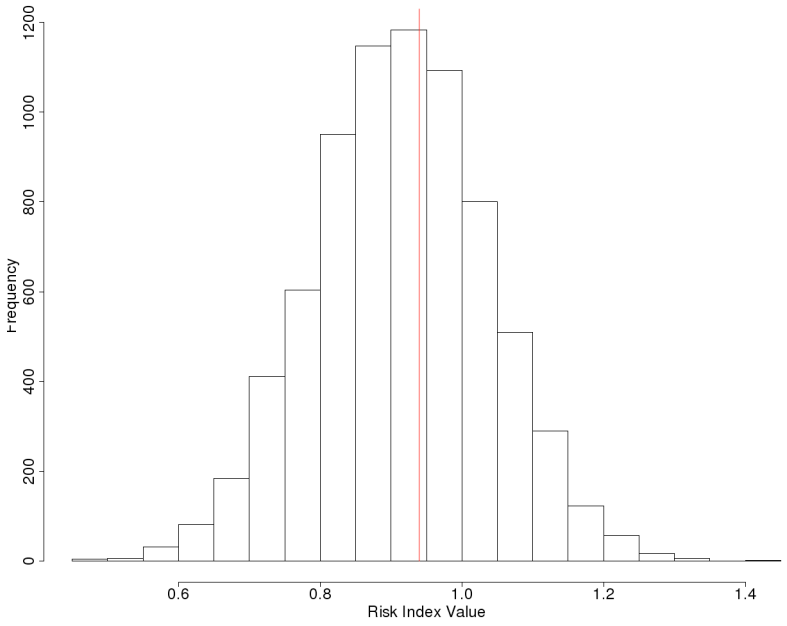


Figure 3-43 Clinical Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #9, Bootstrap Sample #5

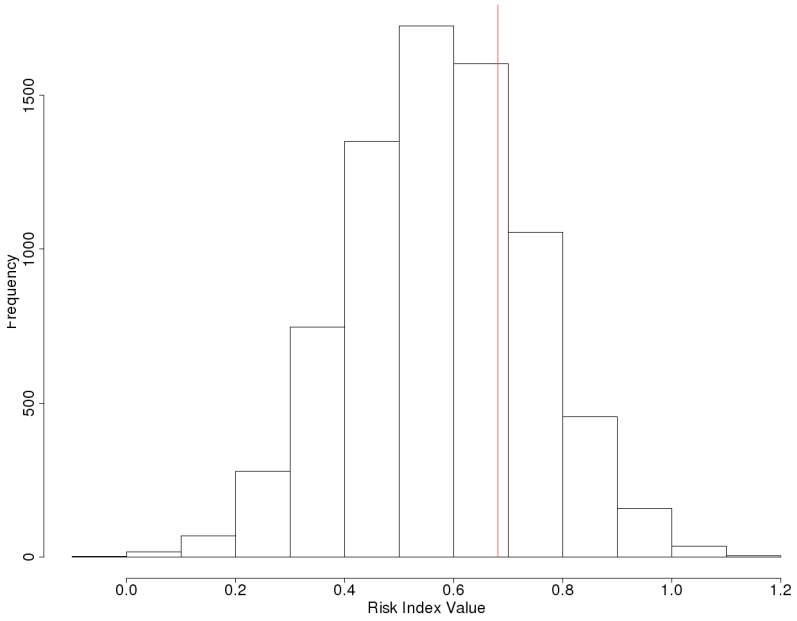


Figure 3-44 Clinical Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #22, Bootstrap Sample #11

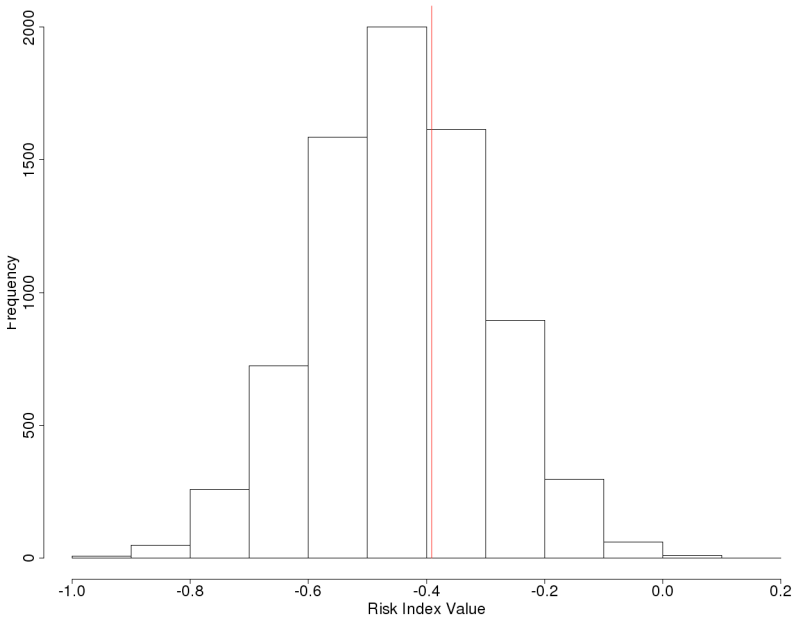


Figure 3-45 Clinical Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #25, Bootstrap Sample #4

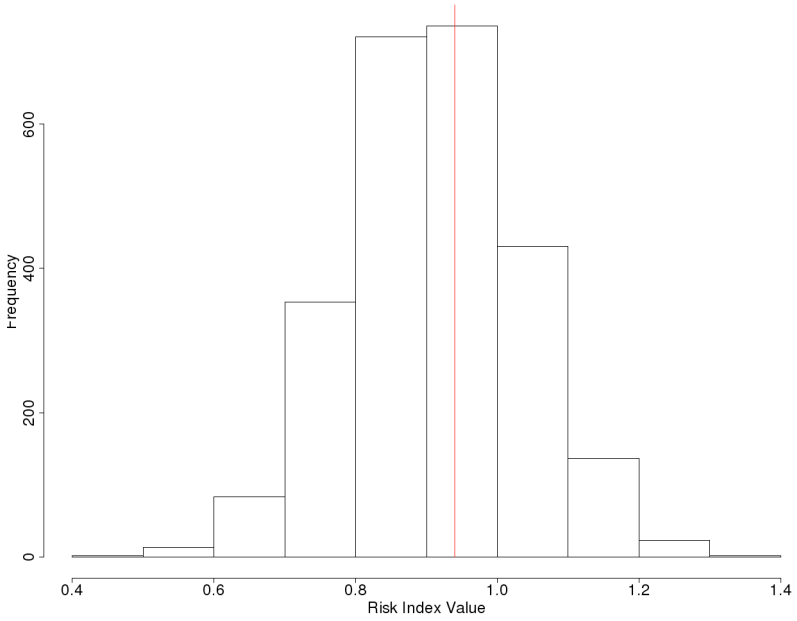


Figure 3-46 Clinical Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #9, Bootstrap Sample #5

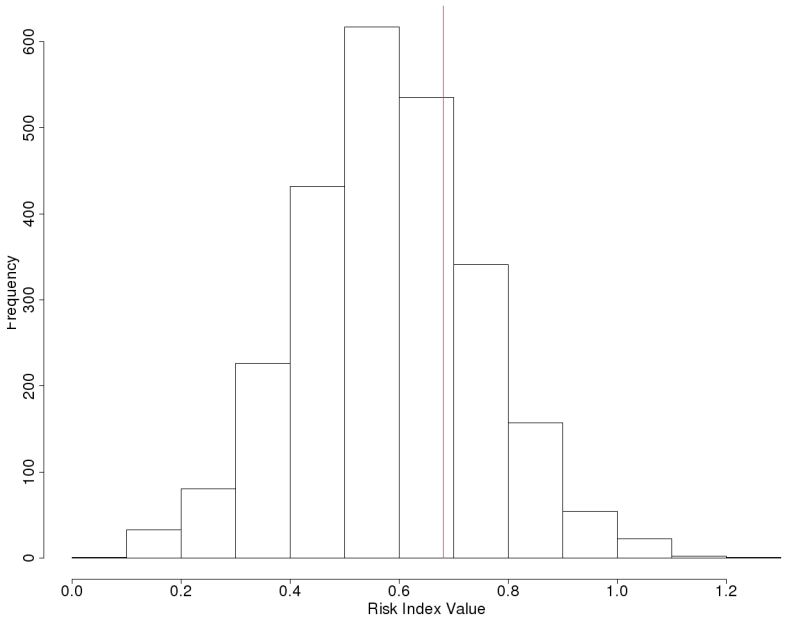


Figure 3-47 Clinical Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #22, Bootstrap Sample #11

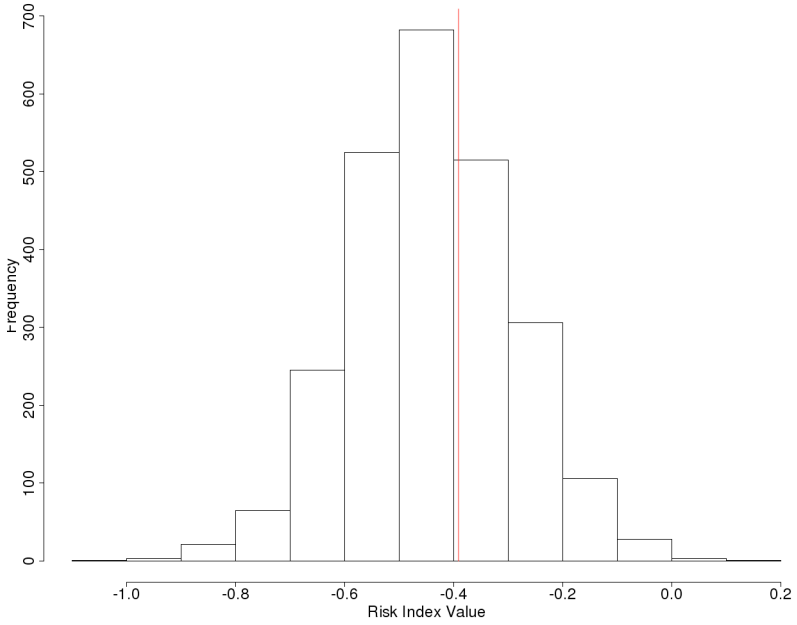


Figure 3-48 Clinical Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #25, Bootstrap Sample #4

Table 3-44 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Large-scale Simulation Dataset #9

		Bootstrap Sample #5 Cutoff Value = 0.940		Bootstrap Sample #12 Cutoff Value = 0.482		Bootstrap Sample #14 Cutoff Value = 0.700		Bootstrap Sample #20 Cutoff Value=0.649		Bootstrap Sample #23 Cutoff Value=-0.482	
Individual	Outcomes	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	1.006	1	0.451	0	0.841	1	0.683	1	0.497	1
2	0	0.859	0	0.384	0	0.607	0	0.497	0	0.406	0
3	1	0.985	1	0.489	1	0.817	1	0.665	1	0.515	1
4	1	1.177	1	0.651	1	0.947	1	0.829	1	0.685	1
5	0	0.721	0	0.254	0	0.405	0	0.330	0	0.269	0
6	0	0.870	0	0.390	0	0.716	1	0.626	0	0.426	0
7	0	0.726	0	0.244	0	0.440	0	0.452	0	0.265	0
8	0	0.824	0	0.290	0	0.586	0	0.499	0	0.336	0
9	1	0.993	1	0.485	1	0.790	1	0.615	0	0.532	1
10	1	0.871	0	0.429	0	0.651	0	0.485	0	0.440	0
11	0	0.731	0	0.189	0	0.482	0	0.418	0	0.229	0
12	1	1.082	1	0.538	1	0.769	1	0.658	1	0.572	1
13	0	0.889	0	0.310	0	0.569	0	0.539	0	0.357	0
14	0	0.899	0	0.350	0	0.583	0	0.566	0	0.401	0
15	0	0.872	0	0.316	0	0.588	0	0.426	0	0.362	0
16	0	0.986	1	0.501	1	0.656	0	0.590	0	0.541	1
17	1	1.005	1	0.474	0	0.780	1	0.709	1	0.522	1
18	1	1.200	1	0.706	1	0.974	1	0.847	1	0.730	1
19	1	1.165	1	0.658	1	0.898	1	0.708	1	0.697	1
20	0	0.906	0	0.364	0	0.714	1	0.572	0	0.415	0
21	0	0.872	0	0.335	0	0.685	0	0.468	0	0.383	0
22	0	0.876	0	0.353	0	0.732	1	0.560	0	0.383	0
23	0	0.934	0	0.377	0	0.551	0	0.551	0	0.418	0
24	1	0.979	1	0.493	1	0.800	1	0.610	0	0.516	1
25	0	0.777	0	0.264	0	0.483	0	0.287	0	0.268	0

Table 3-45 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Large-scale Simulation Dataset #22

		Bootstrap Sample #11 Cutoff Value = 0.680		Bootstrap Sample #15 Cutoff Value = 0.314		Bootstrap Sample #23 Cutoff Value = 0.736		Bootstrap Sample #23 Cutoff Value= -0.005		Bootstrap Sample #25 Cutoff Value = -0.014	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	0.950	1	0.171	0	1.053	1	-0.004	1	-0.014	0
2	0	0.583	0	0.308	0	0.619	0	0.000	1	-0.010	1
3	0	0.435	0	0.329	1	0.523	0	-0.012	0	-0.019	0
4	0	0.591	0	0.217	0	0.716	0	-0.041	0	-0.011	1
5	1	0.743	1	0.428	1	0.828	1	-0.020	0	-0.018	0
6	0	0.529	0	0.303	0	0.627	0	0.009	1	-0.016	0
7	0	0.668	0	0.319	1	0.820	1	-0.024	0	-0.019	0
8	1	0.523	0	0.430	1	0.665	0	0.000	1	-0.014	0
9	0	0.367	0	0.325	1	0.472	0	-0.024	0	-0.016	0
10	0	0.650	0	0.168	0	0.776	1	-0.031	0	-0.021	0
11	0	0.301	0	0.210	0	0.435	0	-0.012	0	-0.021	0
12	0	0.416	0	0.159	0	0.429	0	-0.006	0	-0.013	1
13	1	0.613	0	0.327	1	0.826	1	-0.012	0	-0.015	0
14	1	0.413	0	0.341	1	0.485	0	-0.026	0	-0.019	0
15	0	0.270	0	0.298	0	0.392	0	-0.010	0	-0.013	1
16	0	0.466	0	0.297	0	0.608	0	-0.025	0	-0.013	1
17	0	0.353	0	0.262	0	0.603	0	-0.010	0	-0.014	1
18	1	0.498	0	0.314	1	0.639	0	-0.016	0	-0.008	1
19	0	0.562	0	0.338	1	0.636	0	-0.028	0	-0.018	0
20	0	0.616	0	0.211	0	0.664	0	-0.012	0	-0.019	0
21	0	0.387	0	0.276	0	0.494	0	0.005	1	-0.012	1
22	0	0.393	0	0.122	0	0.573	0	-0.022	0	-0.014	1
23	1	0.826	1	0.257	0	0.916	1	-0.015	0	-0.011	1
24	0	0.397	0	0.319	1	0.636	0	-0.009	0	-0.019	0
25	1	0.756	1	0.392	1	0.930	1	-0.012	0	-0.015	0

Table 3-46 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Large-scale Simulation Dataset #25

		Bootstrap Sample #4 Cutoff Value = -0.391		Bootstrap Sample #5 Cutoff Value = -0.023		Bootstrap Sample #9 Cutoff Value = -0.601		Bootstrap Sample #18 Cutoff Value= -0.612		Bootstrap Sample #21 Cutoff Value = 0.214	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	-0.503	0	-0.087	0	-0.734	0	-0.747	0	0.139	0
2	1	-0.460	0	0.004	1	-0.720	0	-0.843	0	0.118	0
3	1	-0.382	1	0.019	1	-0.616	0	-0.629	0	0.161	0
4	0	-0.453	0	-0.021	1	-0.609	0	-0.669	0	0.073	0
5	1	-0.269	1	0.014	1	-0.399	1	-0.503	1	0.206	0
6	1	-0.251	1	-0.069	0	-0.387	1	-0.446	1	0.346	1
7	0	-0.438	0	0.017	1	-0.597	1	-0.644	0	0.032	0
8	0	-0.253	1	-0.053	0	-0.498	1	-0.485	1	0.305	1
9	0	-0.396	0	-0.050	0	-0.593	1	-0.735	0	0.021	0
10	0	-0.836	0	-0.077	0	-0.990	0	-1.030	0	-0.026	0
11	1	-0.529	0	-0.060	0	-0.675	0	-0.742	0	0.049	0
12	0	-0.524	0	-0.058	0	-0.633	0	-0.705	0	0.017	0
13	0	-0.333	1	-0.070	0	-0.497	1	-0.623	0	0.239	1
14	0	-0.587	0	-0.041	0	-0.853	0	-0.987	0	-0.207	0
15	0	-0.339	1	-0.067	0	-0.529	1	-0.702	0	0.205	0
16	1	-0.306	1	-0.052	0	-0.427	1	-0.537	1	0.244	1
17	1	-0.272	1	-0.029	0	-0.454	1	-0.538	1	0.343	1
18	0	-0.343	1	-0.026	0	-0.496	1	-0.588	1	0.157	0
19	0	-0.480	0	-0.046	0	-0.720	0	-0.764	0	0.016	0
20	0	-0.615	0	-0.011	1	-0.831	0	-0.839	0	-0.034	0
21	1	-0.234	1	-0.053	0	-0.456	1	-0.583	1	0.218	1
22	1	-0.235	1	-0.020	1	-0.378	1	-0.480	1	0.206	0
23	0	-0.583	0	-0.027	0	-0.777	0	-0.798	0	-0.133	0
24	0	-0.459	0	-0.020	1	-0.657	0	-0.678	0	0.242	1
25	1	-0.063	1	-0.062	0	-0.259	1	-0.426	1	0.522	1

Tables 3-46, 3-47, and 3-48 each show five trimmed Clinical + Genotype risk index models corresponding to the Clinical risk index models shown in Table 3-40, 3-41, and 3-42. Tables 3-49, 3-50, and 3-51 show the risk index values and predictions from the same set of five Clinical + Genotype risk index models from the same three small-scale simulation datasets for a set of 25 individuals randomly selected from the optimization set. Figures 3-49, 3-50, and 3-51 show the full distribution of risk index values in the optimization set for a randomly selected trimmed Clinical + Genotype risk index model from each of the three small-scale simulation datasets. Figures 3-52, 3-53, and 3-54 show the full distribution of risk index values in the independent testing set for a randomly selected trimmed Clinical + Genotype risk index model from each of the three small-scale simulation datasets. As in the previous set of figures, in all six of these figures a red line indicates the cut-off point. All individuals with a risk index value greater than or equal to this cut-off point are predicted as “high risk” and all individuals with a value less than this cut-off point are predicted as “low risk”.

Table 3-47 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #9

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
5	0.0285*v1 + 0.0289*v2 + 0.0133*v4 + 0.0042*v5 + 0.0025*v6 - 0.03*v8 + 0.0029*v12 + 0.0025*v16 + 0.0347*v26 + 0.001*v27 + 0.0738*rs859390 + 0.0644*rs12081663 + 0.5065*rs753209 + 0.0725*rs6692452 + 0.0828*rs4652591 - 0.0617*rs4908596 + 0.1819*rs10911598 - 0.1144*rs10783127 - 0.0401*rs184853 - 0.1772*rs17131544 + 0.0034*rs2154367 - 0.0781*rs7528766 - 0.1336*rs1881029 - 0.1489*rs12403147 + 0.003*rs215814 - 0.0421*rs1323126 - 0.056*rs2768761
12	0.0299*v1 + 0.0254*v2 + 0.0122*v4 + 0.0045*v5 + 0.0026*v6 - 0.0313*v8 + 0.0035*v11 + 0.0027*v12 + 0.0021*v16 + 0.0011*v27 + 0.687*rs7520551 - 0.1465*rs7554934 - 0.162*rs12139740 + 0.1052*rs2996655 - 0.0581*rs12058254 - 0.0672*rs6658349 + 0.0114*rs7556384
14	0.0291*v1 + 0.0279*v2 + 0.023*v3 + 0.0129*v4 + 0.0041*v5 + 0.0025*v6 - 0.0265*v8 + 0.0058*v9 + 0.0029*v11 + 0.0032*v12 + 0.1488*rs12060150 + 0.1475*rs2861277 - 0.0839*rs10489322 + 0.1374*rs859390 - 0.1091*rs443386 + 0.0525*rs6694817 + 0.0906*rs2157381 - 0.1301*rs6689228 + 0.0587*rs1339876
20	0.0292*v1 + 0.0294*v2 + 0.0136*v4 + 0.0038*v5 + 0.0024*v6 + 0.0038*v7 - 0.0292*v8 + 0.0056*v9 + 0.0027*v11 + 0.0027*v12 - 0.1466*rs443386 + 0.0317*rs10908327 - 0.1028*rs6675190 + 0.1414*rs7519717 - 0.0702*rs12565849 - 0.1187*rs4908596 + 0.024*rs10495276 + 0.031*rs12124394 + 0.0533*rs10914678 - 0.09*rs11102735 - 0.0076*rs2494884 + 0.0479*rs445633 - 0.0817*rs1980445 + 0.0853*rs2039942
23	0.0307*v1 + 0.0259*v2 + 0.0129*v4 + 0.0043*v5 + 0.0026*v6 - 0.0276*v8 + 0.0025*v11 + 0.0033*v12 + 0.0021*v16 - 0.0022*v29 + 0.8946*rs7520551 + 0.0923*rs7515728 + 0.0912*rs859452 + 0.009*rs4949516 - 0.121*rs10157799 - 0.1499*rs11209805 + 0.5144*rs6659228 + 0.0968*rs12124394 + 0.1581*rs10911598 + 0.1247*rs16840450 - 0.0616*rs7540604 + 0.0546*rs386654

Table 3-48 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #22

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
11	0.0325*v1 + 0.0109*v2 + 0.0024*v3 + 0.0047*v4 + 0.0029*v5 + 0.0052*v6 + 0.0036*v7 + 0.0024*v9 + 0.0048*v14 - 0.0197*v28 + 0.1786*rs3892225 + 0.223*rs16828534 - 0.1237*rs4950371 + 0.0777*rs11210675 + 0.0319*rs16849075 - 0.1109*rs4845222 + 0.033*rs10801446 - 0.0799*rs12406369 + 0.2671*rs10912988 + 0.1421*rs1475766 + 0.0496*rs2039942 - 0.0686*rs4434872
15	0.003*v5 + 0.002*v8 + 0.0177*v15 + 0.0013*v17 - 0.0041*v19 - 3e-04*v20 - 0.0027*v21 - 9e-04*v24 - 0.0058*v25 - 1e-04*v27 + 0.078*rs828505 + 1.0241*rs7520551 + 0.0061*rs6425826 + 0.1881*rs10863400 - 0.5156*rs16860461 + 0.0518*rs2800686 - 0.0202*rs11205175 - 0.0427*rs647924 + 0.0233*rs10493414 + 0.0213*rs1389559 + 0.0275*rs4987299 + 0.0118*rs170261 + 0.0245*rs17032950 - 0.7694*rs11576886 + 0.0033*rs12145484 + 0.0453*rs7542386 + 0.0468*rs474189 + 0.0049*rs1016815
23	0.0313*v1 + 0.0112*v2 + 0.0028*v3 + 0.0056*v4 + 0.003*v5 + 0.0055*v6 + 0.0027*v7 - 0.0115*v11 + 0.018*v15 - 0.0181*v28 + 0.1453*rs16829834 + 0.1044*rs12141159 + 0.2619*rs873525 + 0.0586*rs7554714 + 0.1723*rs2786608 - 0.0127*rs645142 + 0.3927*rs4660345 + 0.0894*rs12089508 + 0.0115*rs10776742 + 0.0712*rs12732088 + 0.0304*rs9657961 + 0.0026*rs11210904 + 0.5356*rs1258022 - 0.0687*rs649352 - 0.024*rs1881029 + 0.03*rs4908817 - 0.7821*rs11576886 - 0.1193*rs652052 - 0.0161*rs1793319
24	-0.0001*v23 - 0.0195*rs12048137 + 0.4799*rs1411400 - 0.3852*rs17131544 + 0.1638*rs894216 - 0.2086*rs647924 - 0.3855*rs284175 - 0.0386*rs4253963 - 0.0223*rs16823912 + 0.0133*rs1475766 - 0.0075*rs6661048 + 0.0222*rs12028179 + 0.003*rs4839312
25	-0.0002*v22 + 0.2237*rs2861311 - 0.1155*rs2050674 - 0.1018*rs12032522 - 0.0407*rs6664830 - 0.1056*rs284175 - 0.1031*rs1202579 + 0.1228*rs17032950 - 0.0285*rs4311892

Table 3-49 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #25

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
4	-0.0422*v1 + 0.0359*v2 + 0.0096*v3 - 0.0208*v4 - 0.0136*v5 + 0.0044*v8 - 0.0223*v12 + 0.0052*v13 + 0.0069*v15 + 6e-04*v21 + 0.111*rs834996 - 0.1008*rs10754164 + 0.1212*rs10890378 - 0.0802*rs16824455
5	-0.0024*v17 + 4e-04*v26 + 3e-04*v27 + 0.0502*rs10914739 + 0.0815*rs873525 + 0.122*rs1411400 + 0.056*rs170261 - 0.5574*rs16860461 - 0.5574*rs563026 + 0.0169*rs696722 - 0.1308*rs1176534 + 0.0135*rs11207408 - 0.0295*rs7553155 + 0.0358*rs10874427 - 0.0342*rs4839312 + 0.6177*rs10890378
9	-0.0425*v1 + 0.0334*v2 + 0.0088*v3 - 0.0213*v4 - 0.0137*v5 + 0.004*v8 - 0.0185*v12 + 0.0052*v13 - 0.0129*v14 + 0.0082*v15 + 0.094*rs16848600 + 0.0591*rs945179 + 0.0418*rs7515728 + 0.065*rs10916131 + 0.0643*rs10914678 + 0.0256*rs12084264 - 0.0874*rs6700777 - 0.8294*rs17131544 + 0.7284*rs1411400 - 0.343*rs3813639 + 0.0165*rs619193 + 0.1626*rs1475766 + 0.0178*rs6657754 - 0.1221*rs10518299 + 0.0597*rs10911065 + 0.2562*rs4311892 - 0.1235*rs12028179 - 0.1014*rs6682150 - 0.0578*rs16830020
18	-0.0386*v1 + 0.0378*v2 + 0.0098*v3 - 0.0206*v4 - 0.013*v5 + 0.0047*v8 - 0.012*v10 - 0.023*v12 - 0.0118*v14 + 0.008*v15 - 0.1935*rs4040617 - 0.0965*rs16848734 + 0.164*rs952023 + 0.0775*rs16864515 + 0.0197*rs16826049 - 0.0388*rs9659765 + 0.0397*rs11121007 + 0.1452*rs11121472 + 0.0481*rs10926660 + 0.1024*rs4474198 + 0.0052*rs4532864 - 0.2001*rs1176534 - 0.2026*rs11589986 + 0.0071*rs16823983 - 0.0796*rs12076197 + 0.0513*rs2386548 - 0.0048*rs4653279
21	-0.0417*v1 + 0.0336*v2 + 0.0091*v3 + 0.0043*v6 + 0.0026*v7 + 0.0378*v9 - 0.0207*v12 + 0.0056*v13 - 0.0016*v19 + 0.0064*v23 - 0.0378*rs1469919 + 0.0896*rs16864515 + 0.0154*rs11264034 + 0.0612*rs542405 - 0.3383*rs17131544 + 0.0461*rs378557 + 0.2682*rs6679643 - 0.0537*rs6699417 + 0.0827*rs2280635 - 0.4996*rs16860461 - 0.1979*rs6667451 + 0.0037*rs6427261 + 0.0222*rs2811620 - 0.1377*rs11589986 - 0.0612*rs10518299 - 0.0101*rs17103767 + 0.0071*rs593861 - 0.0322*rs12068588

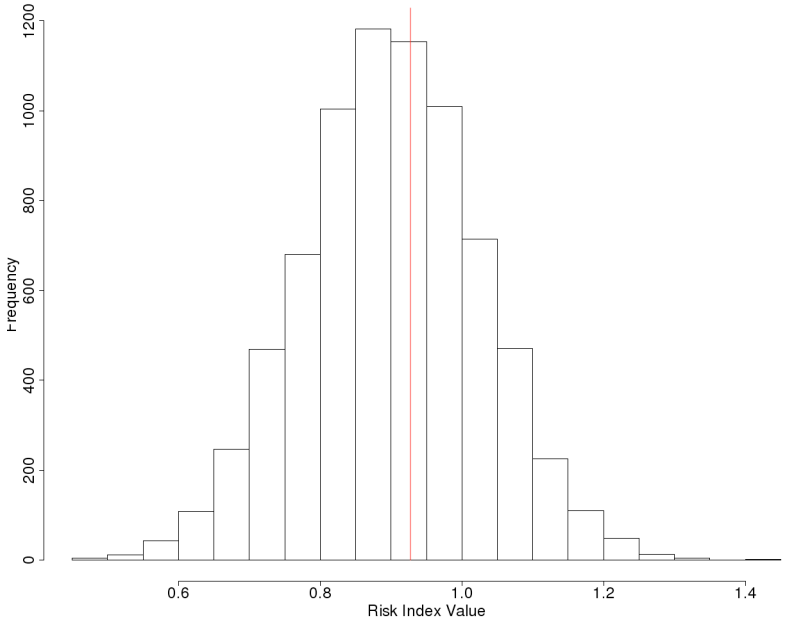


Figure 3-49 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #9, Bootstrap Sample #5

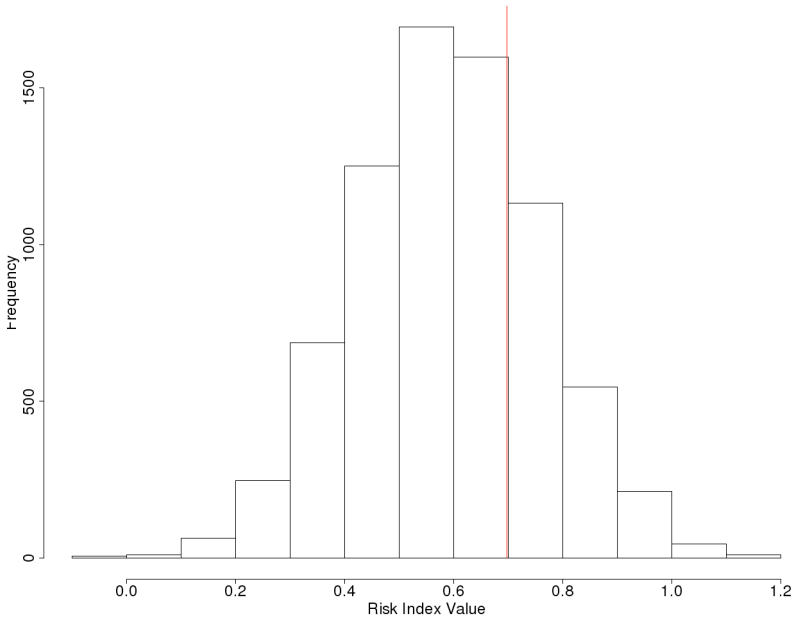


Figure 3-50 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #22, Bootstrap Sample #11

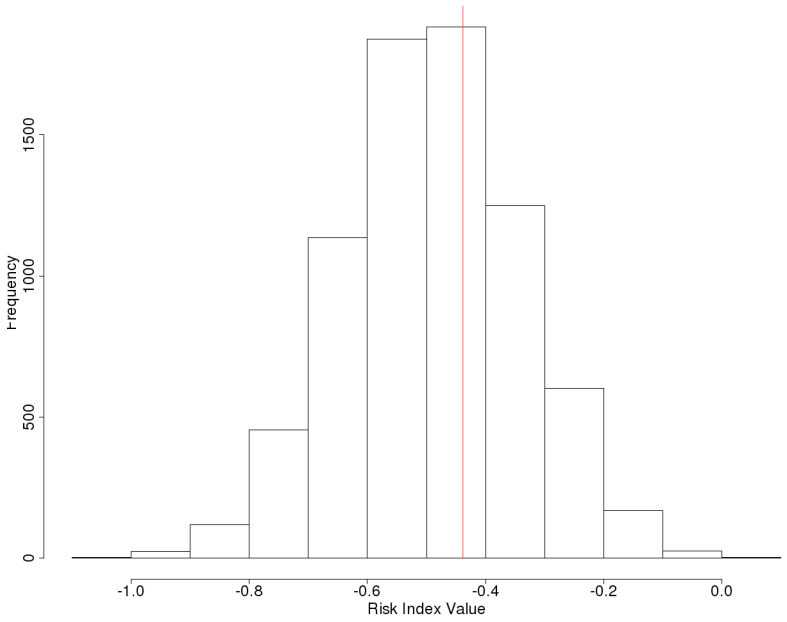


Figure 3-51 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #25, Bootstrap Sample #4

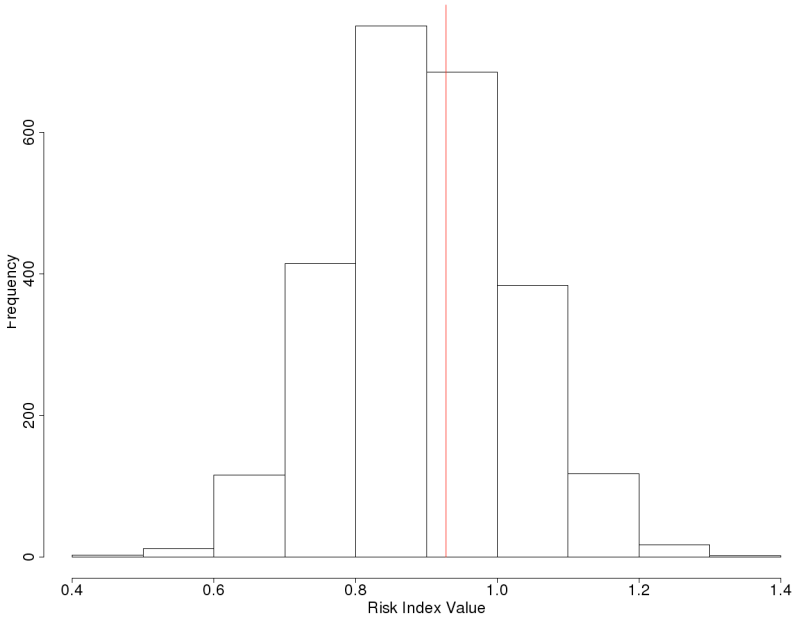


Figure 3-52 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #9, Bootstrap Sample #5

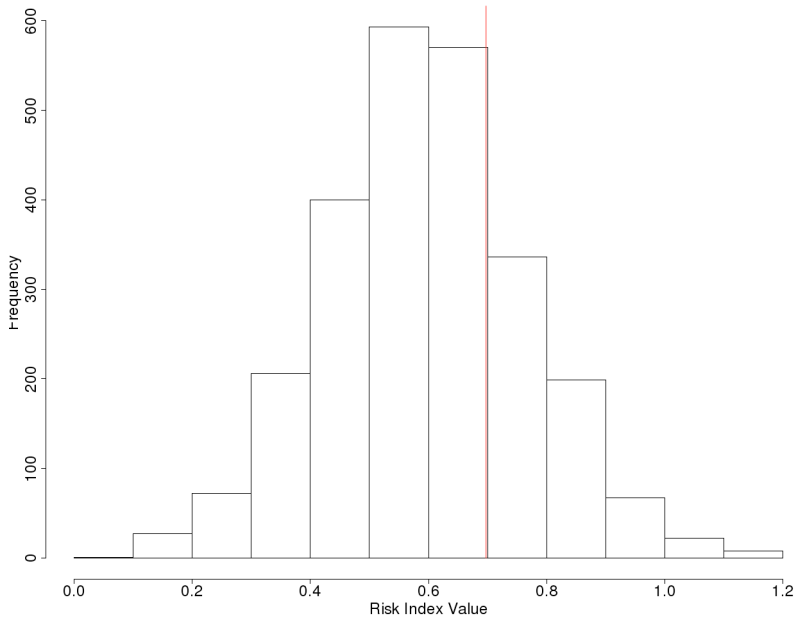


Figure 3-53 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #22, Bootstrap Sample #11

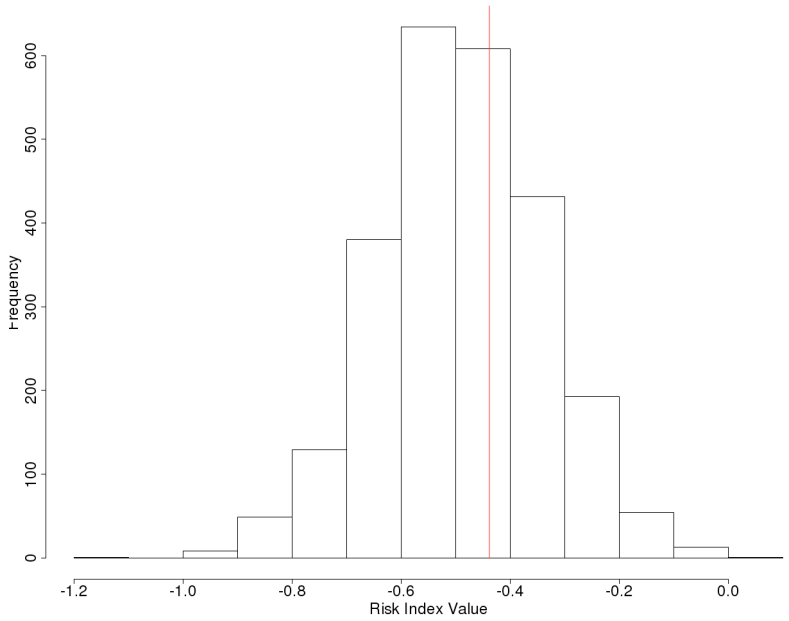


Figure 3-54 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #25, Bootstrap Sample #4

Table 3-50 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Large-scale Simulation Dataset #9

		Bootstrap Sample #5 Cutoff Value = 0.927		Bootstrap Sample #12 Cutoff Value = 0.469		Bootstrap Sample #14 Cutoff Value = 0.696		Bootstrap Sample #20 Cutoff Value=0.647		Bootstrap Sample #23 Cutoff Value=0.509	
Individual	Outcomes	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	0.990	1	0.463	0	0.838	1	0.683	1	0.546	1
2	0	0.845	0	0.372	0	0.599	0	0.473	0	0.417	0
3	1	0.976	1	0.496	1	0.771	1	0.647	0	0.516	1
4	1	1.162	1	0.641	1	0.987	1	0.820	1	0.731	1
5	0	0.693	0	0.223	0	0.398	0	0.335	0	0.261	0
6	0	0.852	0	0.353	0	0.767	1	0.618	0	0.480	0
7	0	0.704	0	0.252	0	0.428	0	0.471	0	0.317	0
8	0	0.812	0	0.289	0	0.595	0	0.481	0	0.368	0
9	1	0.991	1	0.501	1	0.806	1	0.628	0	0.545	1
10	1	0.842	0	0.417	0	0.626	0	0.466	0	0.452	0
11	0	0.715	0	0.168	0	0.438	0	0.419	0	0.247	0
12	1	1.097	1	0.535	1	0.784	1	0.670	1	0.606	1
13	0	0.873	0	0.270	0	0.575	0	0.532	0	0.373	0
14	0	0.876	0	0.320	0	0.630	0	0.582	0	0.429	0
15	0	0.852	0	0.300	0	0.630	0	0.439	0	0.401	0
16	0	0.976	1	0.502	1	0.706	1	0.575	0	0.569	1
17	1	0.993	1	0.423	0	0.811	1	0.680	1	0.555	1
18	1	1.192	1	0.692	1	0.998	1	0.855	1	0.749	1
19	1	1.147	1	0.645	1	0.939	1	0.702	1	0.702	1
20	0	0.898	0	0.363	0	0.675	0	0.579	0	0.425	0
21	0	0.865	0	0.330	0	0.694	0	0.456	0	0.382	0
22	0	0.854	0	0.327	0	0.771	1	0.569	0	0.422	0
23	0	0.931	1	0.311	0	0.498	0	0.551	0	0.461	0
24	1	0.973	1	0.481	1	0.784	1	0.591	0	0.529	1
25	0	0.788	0	0.205	0	0.540	0	0.304	0	0.292	0

Table 3-51 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Large-scale Simulation Dataset #22

		Bootstrap Sample #11 Cutoff Value = 0.698		Bootstrap Sample #15 Cutoff Value = 0.330		Bootstrap Sample #23 Cutoff Value = 0.747		Bootstrap Sample #23 Cutoff Value=0.021		Bootstrap Sample #25 Cutoff Value = -0.024	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	0.956	1	0.192	0	1.052	1	0.026	1	-0.025	0
2	0	0.580	0	0.326	0	0.641	0	0.031	1	-0.020	1
3	0	0.443	0	0.350	1	0.530	0	0.016	0	-0.024	1
4	0	0.645	0	0.242	0	0.763	1	0.028	1	-0.009	1
5	1	0.751	1	0.447	1	0.852	1	-0.007	0	-0.028	0
6	0	0.520	0	0.318	0	0.641	0	0.035	1	-0.021	1
7	0	0.640	0	0.342	1	0.820	1	0.005	0	-0.029	0
8	1	0.568	0	0.444	1	0.700	0	0.027	1	-0.024	1
9	0	0.352	0	0.330	0	0.476	0	-0.013	0	0.001	1
10	0	0.660	0	0.181	0	0.775	1	-0.003	0	-0.031	0
11	0	0.320	0	0.239	0	0.459	0	0.003	0	-0.031	0
12	0	0.400	0	0.178	0	0.447	0	0.004	0	-0.007	1
13	1	0.595	0	0.336	1	0.855	1	0.001	0	-0.025	0
14	1	0.470	0	0.362	1	0.506	0	0.005	0	-0.030	0
15	0	0.290	0	0.317	0	0.438	0	0.006	0	-0.050	0
16	0	0.474	0	0.310	0	0.630	0	0.006	0	-0.050	0
17	0	0.356	0	0.280	0	0.614	0	0.020	0	-0.024	1
18	1	0.499	0	0.338	1	0.643	0	0.009	0	-0.046	0
19	0	0.592	0	0.352	1	0.632	0	-0.013	0	-0.029	0
20	0	0.596	0	0.231	0	0.672	0	0.018	0	-0.029	0
21	0	0.400	0	0.288	0	0.499	0	0.030	1	-0.022	1
22	0	0.447	0	0.116	0	0.592	0	0.004	0	-0.024	1
23	1	0.865	1	0.289	0	0.941	1	0.001	0	-0.016	1
24	0	0.427	0	0.335	1	0.646	0	0.003	0	-0.042	0
25	1	0.791	1	0.414	1	0.941	1	0.002	0	-0.026	0

Table 3-52 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Large-scale Simulation Dataset #25

Individual	Outcome	Bootstrap Sample #4 Cutoff Value = -0.438		Bootstrap Sample #5 Cutoff Value = -0.009		Bootstrap Sample #9 Cutoff Value = -0.612		Bootstrap Sample #18 Cutoff Value= -0.606		Bootstrap Sample #21 Cutoff Value = 0.201	
		Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	-0.568	0	-0.090	0	-0.756	0	-0.766	0	0.115	0
2	1	-0.449	0	0.012	1	-0.719	0	-0.861	0	0.118	0
3	1	-0.354	1	0.030	1	-0.629	0	-0.637	0	0.157	0
4	0	-0.478	0	-0.006	1	-0.629	0	-0.655	0	0.057	0
5	1	-0.315	1	0.019	1	-0.433	1	-0.512	1	0.184	0
6	1	-0.322	1	-0.062	0	-0.442	1	-0.440	1	0.311	1
7	0	-0.488	0	0.013	1	-0.598	1	-0.650	0	0.015	0
8	0	-0.271	1	-0.048	0	-0.508	1	-0.488	1	0.311	1
9	0	-0.441	0	-0.043	0	-0.594	1	-0.727	0	0.021	0
10	0	-0.876	0	-0.066	0	-1.004	0	-0.994	0	-0.018	0
11	1	-0.567	0	-0.042	0	-0.664	0	-0.729	0	0.035	0
12	0	-0.544	0	-0.044	0	-0.633	0	-0.720	0	0.001	0
13	0	-0.424	1	-0.049	0	-0.518	1	-0.617	0	0.225	1
14	0	-0.652	0	-0.033	0	-0.862	0	-0.961	0	-0.207	0
15	0	-0.384	1	-0.064	0	-0.552	1	-0.704	0	0.201	0
16	1	-0.376	1	-0.040	0	-0.442	1	-0.522	1	0.240	1
17	1	-0.272	1	-0.017	0	-0.458	1	-0.529	1	0.333	1
18	0	-0.413	1	-0.016	0	-0.516	1	-0.577	1	0.126	0
19	0	-0.526	0	-0.039	0	-0.701	0	-0.749	0	0.003	0
20	0	-0.660	0	-0.006	1	-0.840	0	-0.853	0	-0.024	0
21	1	-0.254	1	-0.029	0	-0.459	1	-0.581	1	0.231	1
22	1	-0.255	1	-0.025	0	-0.386	1	-0.492	1	0.198	0
23	0	-0.653	0	-0.019	0	-0.783	0	-0.784	0	-0.139	0
24	0	-0.530	0	-0.006	1	-0.661	0	-0.673	0	0.222	1
25	1	-0.035	1	-0.057	0	-0.247	1	-0.441	1	0.496	1

3.5.3 Predictive Performance

After the variable selection procedure is completed and the models are applied to each individual in the independent testing set then the sensitivity, specificity, misclassification, and positive predictive value are measured for both the Clinical and Clinical + Genotype risk index models for each of the 25 large-scale simulation datasets. Table 3-52 shows the means and standard deviations of these measurements. To provide a 95% confidence for these measurements of sensitivity, specificity, misclassification, and positive predictive value, for each independent testing set 1000 bootstrap samples were generated. By making predictions about each individual in these bootstrap samples and calculating the sensitivity, specificity, misclassification, and positive predictive value for each bootstrap sample, 95% confidence intervals were estimated for these measurements in each of the 25 large-scale simulation datasets. The mean and standard deviation of the spread (i.e., range) of these confidence intervals for both the Clinical and Clinical + Genotype risk index model is shown in Table 3-53. Table 3-54 shows the predictive performance and confidence intervals for the three large-scale simulation datasets discussed in Section 3.5.2.

Table 3-53 Means and Standard Deviations of Predictive Performance Estimates for the 25 Large-scale Simulation Datasets

Model	Sensitivity (SD)	Specificity (SD)	Misclassification (SD)	PPV (SD)	AUC (SD)
Clinical	0.725 (0.038)	0.933 (0.016)	0.132 (0.015)	0.831 (0.035)	0.926 (0.015)
Clinical + Genotype	0.734 (0.034)	0.940 (0.015)	0.124 (0.013)	0.849 (0.033)	0.939 (0.012)

Table 3-54 Means and Standard Deviations of Predictive Performance 95% Confidence Intervals for the 25 Large-scale Simulation Datasets

Model	Mean Range of 95% Confidence Interval (SD)			
	Sensitivity	Specificity	Misclassification	PPV
Clinical	0.062 (0.003)	0.024 (0.003)	0.026 (0.002)	0.056 (0.005)
Clinical + Genotype	0.061 (0.003)	0.022 (0.002)	0.026 (0.001)	0.054 (0.004)

Table 3-55 Predictive Performance Estimates for Three Large-scale Simulation Datasets

Dataset	Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
9	Clinical	0.758 (0.726-0.789)	0.878 (0.862-0.893)	0.158 (0.144-0.173)	0.727 (0.696-0.758)
	Clinical + Genotype	0.76 (0.729-0.789)	0.891 (0.876-0.905)	0.148 (0.135-0.162)	0.75 (0.719-0.78)
22	Clinical	0.731 (0.699-0.763)	0.95 (0.94-0.96)	0.115 (0.102-0.127)	0.862 (0.834-0.89)
	Clinical + Genotype	0.746 (0.713-0.776)	0.95 (0.939-0.96)	0.111 (0.098-0.124)	0.863 (0.837-0.89)
25	Clinical	0.78 (0.752-0.809)	0.936 (0.925-0.947)	0.113 (0.101-0.124)	0.849 (0.825-0.876)
	Clinical + Genotype	0.774 (0.746-0.804)	0.945 (0.933-0.955)	0.109 (0.097-0.122)	0.865 (0.838-0.892)

Using the number of models predicting an individual in the independent testing set as "high risk", receiver operator characteristic (ROC) curves were generated for the Clinical and Clinical + Genotype risk index model for each of the 25 large-scale simulation datasets, and the AUC for the ROC curve was estimated. The average AUC for the Clinical risk index models was 0.926 (SD = 0.015), and the average AUC for the Clinical + Genotype risk index models was 0.939 (SD = 0.012). Figure 3-53, 3-54, and 3-55 show the ROC curves for the Clinical risk index model the three large-scale simulation datasets discussed in sections 3.5.2, and Figure 3-56, 3-57, and 3-58 show the ROC curve for the Clinical + Genotype risk index model from those three selected small-scale simulation datasets.

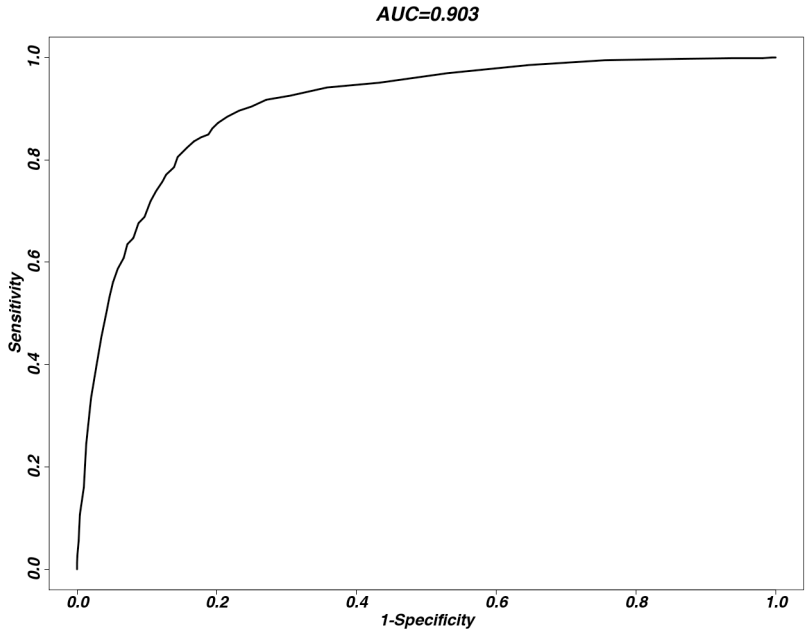


Figure 3-55 ROC Curve of the Clinical Risk Index Model for Large-scale Simulation Dataset #9

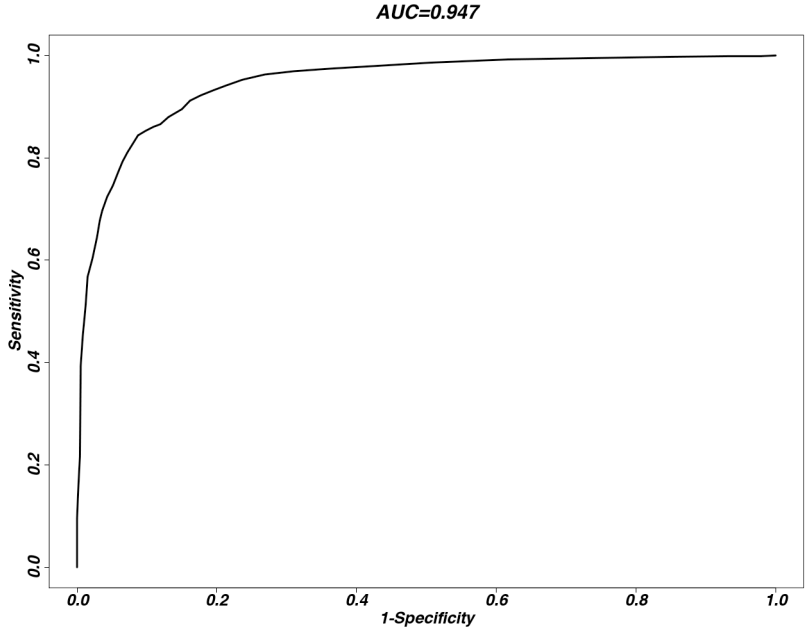


Figure 3-56 ROC Curve of the Clinical Risk Index Model for Large-scale Simulation Dataset #22

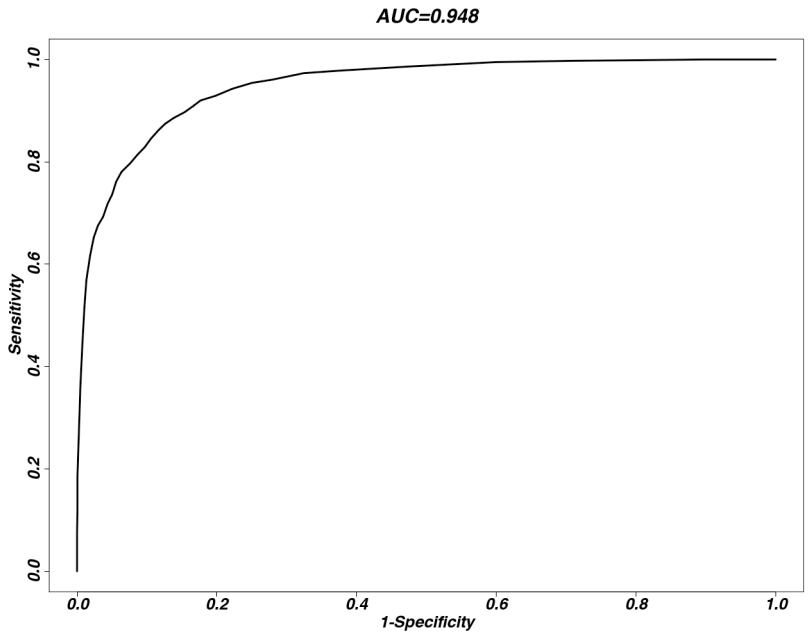


Figure 3-57 ROC Curve of the Clinical Risk Index Model for Large-scale Simulation Dataset #25

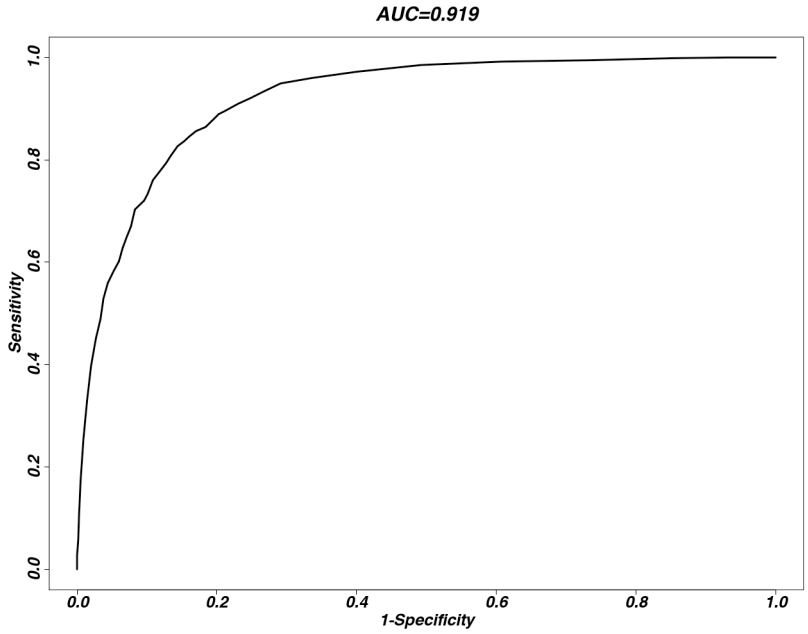


Figure 3-58 ROC Curve of the Clinical + Genotype Risk Index Model for Large-scale Simulation Dataset #9

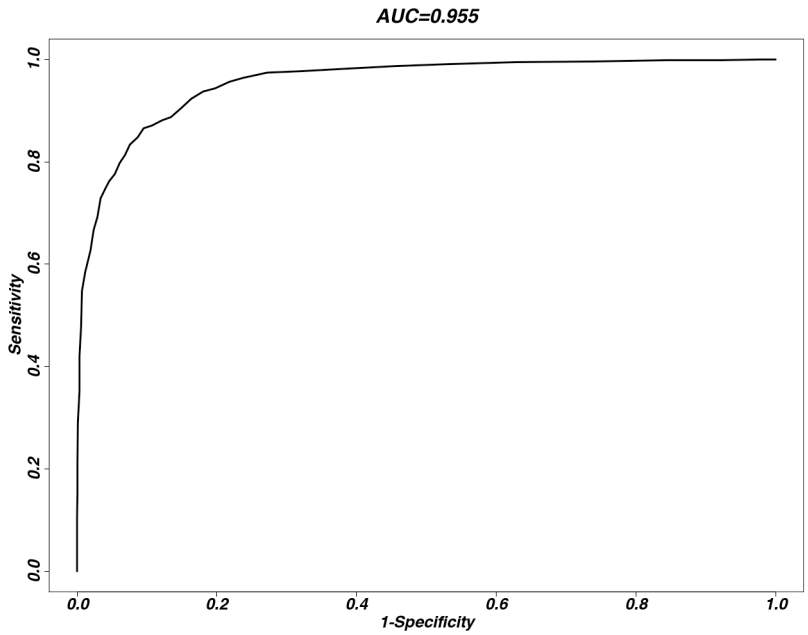


Figure 3-59 ROC Curve of the Clinical + Genotype Risk Index Model for Large-scale Simulation Dataset #22

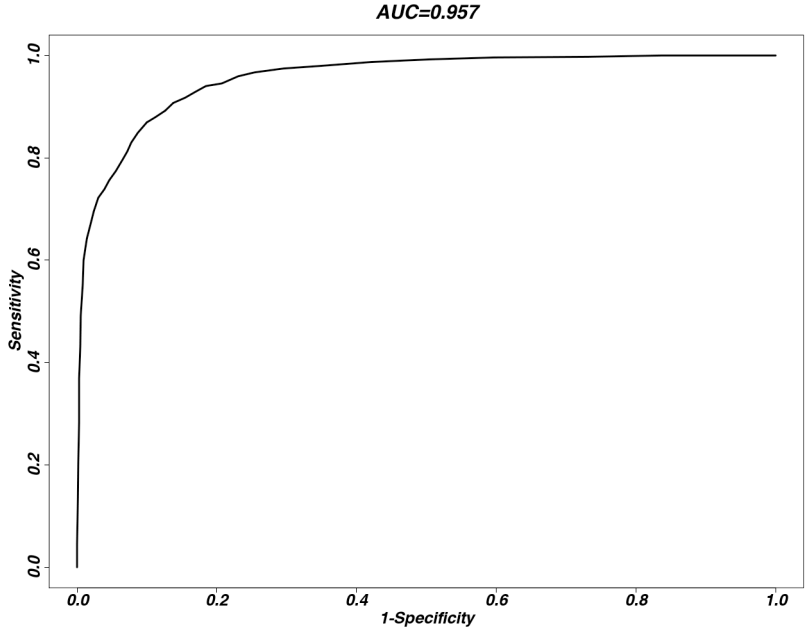


Figure 3-60 ROC Curve of the Clinical + Genotype Risk Index Model for Large-scale Simulation Dataset #25

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. The proportion of models that predict that an individual as high risk, then, represents the predicted probability of an individual developing the outcome. For each individual a 95% confidence interval can be constructed as described in Section 3.2.3

3.5.4 Random Forest Comparison

For each of the 25 large-scale simulation datasets a random forest was generated using the optimization set created by the risk index procedure. The forests were generated using the methodology given in Section 3.2.4. For each of the random forests an ROC curve was generated and the AUC was estimated. The mean AUC for the random forest models was 0.915 (SD = 0.013). Figures 3-61, 3-62, and 3-63 show the ROC curve for the random forest generated from the three small-scale simulation datasets described in Section 3.5.2. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set using a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of 5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the individual to be "high risk". Table 3-55 shows the mean and standard deviation of the sensitivity, specificity, misclassification, and PPV for a range of different proportions.

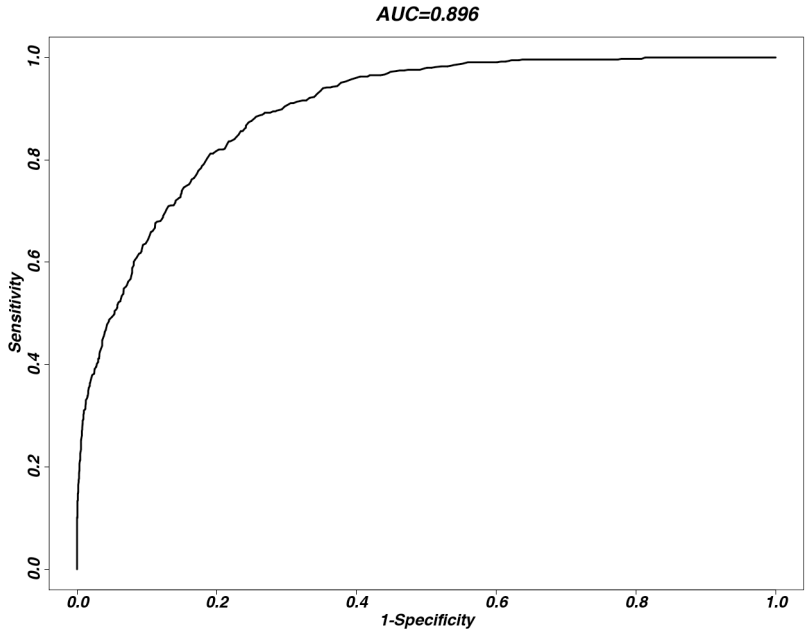


Figure 3-61 ROC Curve of the Random Forest Generated for Large-scale Simulation Dataset #9

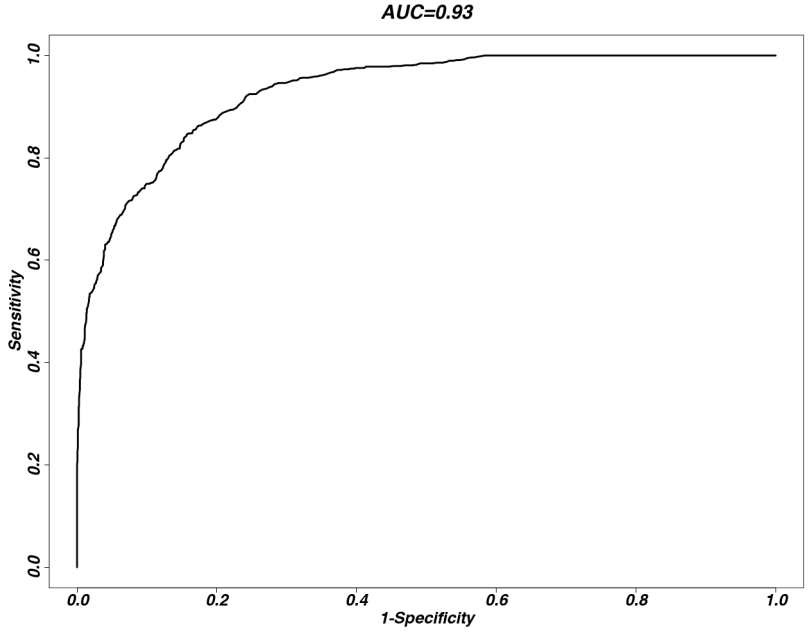


Figure 3-62 ROC Curve of the Random Forest Generated for Large-scale Simulation Dataset #22

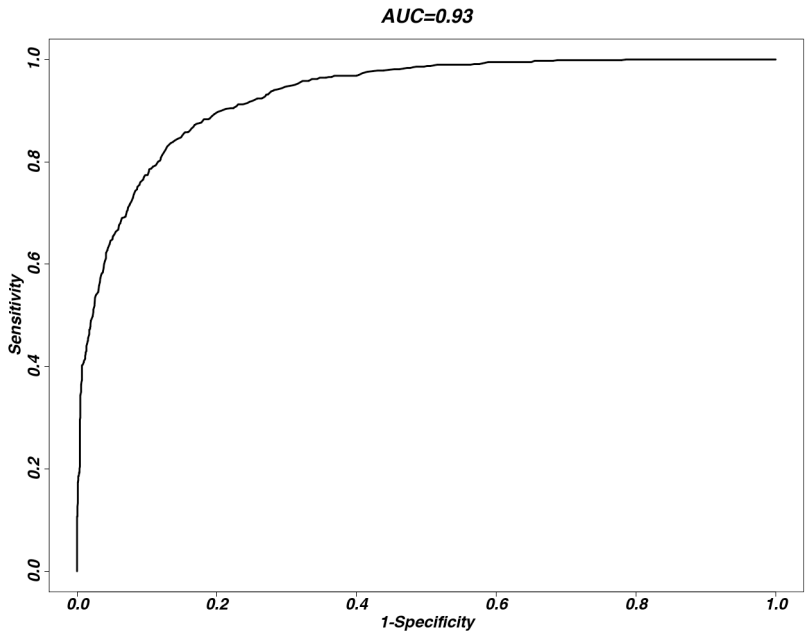


Figure 3-63 ROC Curve of the Random Forest Generated for Large-scale Simulation Dataset #25

Table 3-56 Means and Standard Deviations of Performance Estimates of the Random Forest Models Generated from the 25 Large-scale Simulation Datasets

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	1 (0.001)	0.044 (0.019)	0.657 (0.017)	0.323 (0.012)
0.1	0.996 (0.004)	0.121 (0.028)	0.605 (0.022)	0.341 (0.014)
0.15	0.983 (0.011)	0.199 (0.028)	0.556 (0.02)	0.359 (0.015)
0.2	0.956 (0.021)	0.264 (0.024)	0.519 (0.016)	0.372 (0.014)
0.25	0.914 (0.028)	0.321 (0.018)	0.494 (0.012)	0.38 (0.014)
0.3	0.854 (0.031)	0.369 (0.012)	0.479 (0.01)	0.381 (0.015)
0.35	0.777 (0.032)	0.411 (0.007)	0.475 (0.011)	0.375 (0.016)
0.4	0.694 (0.027)	0.442 (0.006)	0.479 (0.01)	0.361 (0.015)
0.45	0.6 (0.017)	0.471 (0.004)	0.489 (0.006)	0.341 (0.014)
0.5	0.502 (0.001)	0.499 (0)	0.5 (0)	0.314 (0.011)
0.55	0.404 (0.017)	0.527 (0.004)	0.511 (0.006)	0.28 (0.013)
0.6	0.306 (0.027)	0.558 (0.006)	0.521 (0.01)	0.24 (0.019)
0.65	0.22 (0.032)	0.591 (0.007)	0.525 (0.011)	0.196 (0.023)
0.7	0.146 (0.031)	0.631 (0.012)	0.521 (0.01)	0.152 (0.025)
0.75	0.086 (0.028)	0.679 (0.018)	0.506 (0.012)	0.108 (0.027)
0.8	0.046 (0.021)	0.734 (0.024)	0.482 (0.016)	0.07 (0.024)
0.85	0.016 (0.01)	0.805 (0.028)	0.442 (0.02)	0.035 (0.016)
0.9	0.004 (0.004)	0.879 (0.028)	0.395 (0.022)	0.013 (0.009)
0.95	0 (0.001)	0.953 (0.019)	0.345 (0.018)	0.002 (0.005)

3.5.5 Conclusion

The results of the large-scale simulation study using the 500 most highly associated SNPs are extremely promising, and demonstrate robust predictive performance. Both the predictive performance estimates and the AUC for the ROC curves for the 25 large-scale simulation datasets are noticeably higher than the small-scale simulation studies. The average misclassification and PPV are higher than any average misclassification or PPV yielded by the random forest model. As with both small-scale simulation studies the average AUC is significantly higher for the Clinical + Genotype risk index model than for the Clinical risk index model ($p=0.001$), and the average AUC for the Clinical + Genotype risk index model is also significantly greater than the average AUC of the random forest model ($p=1.5e-8$). A risk index model built using only the five most highly

associated covariates and the six associated SNPs had performance characteristics that were comparable to, but somewhat lower than, the best Clinical and Clinical + Genotype risk index models built using the standard variable selection procedure (Table 3-57)

Table 3-57 Performance Characteristics of a Risk Index Model Built Using the Five Most Highly Associated Covariates and the Six Associated SNPs

Model	Sensitivity (SD)	Specificity (SD)	Misclassification (SD)	PPV (SD)
Clinical	0.639 (0.122)	0.853 (0.072)	0.214 (0.026)	0.761 (0.039)
Clinical + Genotype	0.615 (0.102)	0.876 (0.051)	0.204 (0.023)	0.722 (0.085)

3.6 Large-scale Simulation Study Top Principal Components Results

3.6.1 Variable Selection

Using the same procedure as in Section 3.2.1, Clinical and Clinical + Genotype risk index models were constructed for each of the 25 large-scale simulation datasets. In place of the 500 SNPs most highly associated with the outcome, a principal components analysis was performed on the full set of 38,835 SNPs using SMARTPCA (Patterson, et al, 2006), and the top 500 principal components were used to build the risk index models.

Table 3-56 shows the summary of the variable selection procedure from the Clinical risk index model averaged across the 25 simulation datasets. Variables v1 through v5 are the most frequently selected; on average, they each appear in more 14 of the 25 trimmed Clinical risk index models. Variable v14, v13, and v6 are also frequently selected, appearing in 12, 11.8, and 11.7 out of 25 trimmed Clinical risk index models, on average.

Table 3-57 shows the summary of the variable selection procedure from the Clinical + Genotype risk index model averaged across the 25 simulation datasets. No principal component was in more than 1.32 out of 25 trimmed Clinical + Genotype risk index models on average.

Table 3-58 Summary of the Number of Times Each Variable is Selected into a Specific Model Position for the Large-scale Simulation Clinical Risk Index Models

Variable	Variable Position								Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	
v1	14.4	0.92	0.32	0.12	0.12	0.04	0	0	16.64
v2	0.76	2.08	2.32	2.04	0.92	1.08	0.72	0.6	14.2
v3	0.4	4	2.56	1.6	0.96	0.84	0.92	0.52	14.64
v4	0.48	3.64	2.24	1.76	1.12	0.72	0.76	0.6	14.08
v5	0.28	3	3.04	1.56	1.4	1	0.8	0.84	14.56
v6	0.08	0.56	0.48	0.6	0.96	0.72	0.64	1.04	11.68
v7	0.24	0.24	0.28	0.6	0.72	0.88	1.08	0.8	10.32
v8	0.04	0.24	0.76	1.16	0.72	0.8	0.92	0.84	11.32
v9	0.12	0.16	0.4	0.76	0.64	0.68	1.04	1.08	11.6
v10	0.12	0.44	0.6	0.56	0.8	1.08	1.08	0.88	11.44
v11	0.16	0.56	0.64	1.16	0.8	1.16	0.68	0.88	11.16
v12	0.04	0.36	0.72	0.64	1.16	0.6	0.96	1.12	11.64
v13	0.04	0.24	0.44	0.52	1.32	1	0.8	1.24	11.8
v14	0.04	0.28	0.76	0.56	1.16	0.64	0.8	0.84	12
v15	0.04	0.36	0.64	0.64	0.92	1.08	1.08	0.56	10.96
v16	0.08	0.08	0.4	0.64	0.68	1.12	0.64	1.04	11.16
v17	0.64	0.52	0.68	0.88	0.64	0.88	1	1	10.8
v18	0.56	0.6	0.48	1.12	0.64	0.92	0.84	1.04	11.08
v19	0.48	0.56	0.52	0.92	0.76	0.8	1.16	0.76	9.68
v20	0.52	0.72	0.72	0.8	0.84	1.04	0.68	1.12	10.16
v21	0.48	0.68	0.68	1.04	1.4	0.84	1.08	0.8	10.48
v22	0.36	0.72	0.64	0.84	0.76	0.8	0.84	0.76	10.4
v23	0.8	0.6	0.88	0.72	0.76	0.72	0.6	1.08	11.08
v24	0.6	0.6	0.8	0.52	0.72	0.84	1.16	0.68	10.56
v25	0.84	0.52	0.52	0.8	0.8	0.96	0.96	1.08	10.16
v26	0.72	0.72	0.84	0.52	0.64	0.68	1.08	0.64	11.16
v27	0.68	0.64	0.48	0.6	0.84	0.88	0.76	0.96	10.48
v28	0.32	0.52	0.52	0.52	0.84	1.12	0.8	1.24	8.92
v29	0.68	0.44	0.64	0.8	0.96	1.08	1.12	0.96	11.36

* Averaged across 25 bootstrap samples of each optimization set

Table 3-59 Summary of the Number of Times Selected Principal Component Variables are Selected into a Specific Model Position for the Large-scale Simulation Clinical + Genotype Risk Index Models

	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
SNP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
pc407	0.16	0.04	0.12	0.08	0.12	0.08	0.08	0.08	0.2	0.04	0	0.08	0.04	0	0.04	0.04	0.16	0.12	0.12	0.04	1.64	1.32
pc329	0.12	0	0.08	0.12	0.12	0.04	0.12	0.04	0.08	0.12	0	0.08	0	0.04	0.12	0.04	0.08	0.2	0.12	0.04	1.56	1.28
pc359	0	0.12	0.16	0.04	0.08	0.12	0.04	0.08	0	0.08	0.04	0.08	0.16	0.12	0.08	0.12	0.12	0.04	0.04	0.08	1.6	1.28
pc125	0.08	0.08	0.12	0.16	0.08	0	0.08	0	0.04	0.08	0.04	0.04	0.12	0.2	0	0.08	0.08	0.04	0.08	0.16	1.56	1.2
pc232	0	0.08	0.08	0	0.16	0.12	0.04	0.08	0.12	0.04	0.12	0.08	0.04	0.08	0.04	0	0	0.16	0.12	0.2	1.56	1.2
pc277	0.08	0.08	0.04	0.08	0.12	0.16	0.08	0.04	0	0.16	0.04	0.16	0.04	0.04	0	0.04	0.04	0.04	0.16	0.08	1.48	1.2
pc442	0.12	0.04	0.08	0.04	0.16	0.04	0.08	0.12	0	0	0.04	0	0.12	0.08	0.08	0	0.12	0.04	0.12	0.08	1.36	1.2
pc53	0.28	0.08	0.08	0.12	0.04	0.04	0.08	0.12	0.04	0.08	0.08	0.04	0.2	0.04	0	0	0	0.04	0	0	1.36	1.16
pc129	0.12	0	0.08	0.12	0	0.12	0.08	0.12	0.04	0.08	0.08	0.12	0.04	0	0	0.08	0	0	0.12	0.08	1.28	1.16
pc239	0.04	0.12	0.04	0.12	0.08	0	0.04	0	0.24	0.08	0.24	0.12	0	0	0.16	0	0.04	0.12	0.04	0.04	1.52	1.16
pc268	0.04	0.08	0.08	0.16	0.12	0	0.16	0.16	0.08	0.04	0	0	0.04	0.12	0.04	0.08	0.08	0.04	0.04	0.24	1.6	1.16
pc315	0.08	0.28	0.04	0.04	0.04	0.08	0.08	0	0	0.08	0.16	0.04	0.04	0.04	0.04	0.12	0.04	0.04	0.08	0.04	1.36	1.16
pc126	0.08	0.08	0.04	0.04	0.08	0.04	0.12	0.04	0.12	0.04	0.04	0.2	0.04	0.04	0.12	0.04	0	0	0.08	0.04	1.28	1.12
pc141	0	0.08	0.08	0.04	0.04	0.08	0.12	0.12	0.04	0.12	0.12	0	0	0.04	0.12	0.08	0.04	0.12	0.08	0	1.32	1.12
pc144	0.04	0	0.08	0.08	0.04	0.08	0.04	0.12	0.08	0.12	0.04	0.08	0.04	0.04	0.08	0.08	0.04	0.16	0.08	0.04	1.36	1.12
pc172	0.08	0.08	0.08	0.16	0.08	0.08	0.04	0.08	0.16	0.12	0.12	0.04	0.04	0	0	0.12	0.04	0	0.12	0.04	1.48	1.08
pc191	0.04	0.08	0.08	0.04	0.08	0.16	0.04	0.08	0.04	0.08	0.08	0	0.04	0.08	0.04	0.04	0.12	0.08	0.04	0.08	1.32	1.08
pc220	0.08	0.08	0.12	0.08	0.12	0	0.04	0.08	0.2	0	0.04	0.08	0.04	0	0.04	0	0.04	0.04	0.12	0.12	1.32	1.08

3.6.2 Models

Once the variable selection procedure is finished each of the 25 large-scale simulation datasets have 25 trimmed Clinical and Clinical + Genotype risk index models. Tables 3-58, 3-59, and 3-60 each show five trimmed Clinical risk index models randomly selected from one of three randomly chosen large-scale simulation datasets (datasets #9, #22, and #25). Figures 3-64, 3-65, and 3-66 show the full distribution of risk index values in the optimization set for a randomly selected trimmed Clinical risk index models from each of the three small-scale simulation datasets. Figures 3-67, 3-68, and 3-69 show the full distribution of risk index values in the independent testing set for a randomly selected trimmed Clinical risk index model from each of the three large-scale simulation datasets. In all six of these figures a red line indicates the cut-off point. All individuals with a risk index value greater than or equal to this cut-off point are predicted as "high risk" and all individuals with a value less than this cut-off point are predicted as "low risk". Tables 3-61, 3-62, and 3-63 show the risk index values and predictions from the same set of five Clinical risk index models from the same three large-scale simulation datasets for a set of 25 individuals randomly selected from the independent test set.

Table 3-60 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #9

Bootstrap Sample	Trimmed Clinical Risk Index Model
5	0.0304*v1 + 0.0253*v2 - 0.012*v3 + 0.005*v4 + 0.0096*v5 - 0.0029*v7 - 0.0059*v8 + 0.0032*v13 + 0.005*v14 + 0.0029*v15 + 0.0039*v16 - 0.0028*v18 + 0.0043*v19 - 9e-04*v20 + 0.0042*v23 + 0.0185*v25 + 0.0021*v26 + 0*v27 + 5e-04*v29
12	0.0033*v19 - 2e-04*v27
14	0.0282*v1 + 0.0337*v2 - 0.0111*v3 + 0.0065*v4 + 0.0084*v5 + 0.004*v6 - 0.0027*v7 - 0.0067*v8 - 0.0012*v12 + 0.0028*v13 + 0.0036*v14 + 0.0023*v15 + 0.007*v16 + 0.0018*v22 + 0.0044*v23 + 0.0198*v24 + 0.0401*v25 - 0.001*v27 + 0.0012*v28 - 1e-04*v29

20	1e-04*v17 - 7e-04*v28
23	0.0309*v1 + 0.0268*v2 - 0.0127*v3 + 0.009*v5 - 0.0022*v7 - 0.0068*v8 - 0.0022*v12 + 0.0041*v15 + 0.0043*v16 + 3e-04*v17 - 0.0028*v18 + 0.0014*v19 + 9e-04*v20 + 0.001*v21 + 0.0013*v22 + 0.0136*v24 + 0.0029*v26 - 2e-04*v27 + 0.0034*v28

Table 3-61 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #22

Bootstrap Sample	Trimmed Clinical Risk Index Model
11	-0.0816*v1 + 0.1228*v2 + 0.0031*v3 + 0.0047*v4 + 0.0048*v5 + 0.0337*v6 + 0.0023*v8 + 0.0129*v9 - 0.0017*v11 + 0.0036*v12 - 0.002*v14 - 0.0053*v15 + 0.0039*v17 + 0.0019*v19 - 0.0016*v21 - 0.0026*v22 - 1.2314*v25 + 0.0026*v28
15	-0.0929*v1 + 0.0031*v3 + 0.0049*v4 + 0.0059*v5 + 0.0236*v6 + 0.0027*v8 + 0.0096*v9 + 0.0015*v10 - 0.0043*v11 + 0.0026*v12 - 8e-04*v13 - 0.0043*v15 + 0.0038*v17 - 0.0013*v18 + 2e-04*v20 + 0.0022*v22 - 0.2143*v25 + 5e-04*v26 + 0.0013*v28 - 8e-04*v29
23	4e-04*v13 - 0.001*v18 + 1e-04*v26
24	-1e-04*v13 - 1e-04*v21 + 1e-04*v23 - 0.0014*v29
25	-1.9043*v7 + 0.001*v22 - 0.0023*v29

Table 3-62 Clinical Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #25

Bootstrap Sample	Trimmed Clinical Risk Index Model
4	0.009*v1 + 0.0101*v2 + 0.0341*v3 + 0.0134*v4 + 0.0024*v5 - 0.0086*v7 + 0.005*v8 + 0.0178*v9 + 0.0067*v11 - 0.0011*v12 + 0.0058*v13 - 0.0034*v14 + 5e-04*v18 + 5e-04*v20 + 6e-04*v22
5	0.0106*v1 + 0.0099*v2 + 0.0116*v4 + 0.0024*v5 + 0.0128*v6 + 0.0052*v8 + 0.001*v10 + 0.0061*v11 + 1e-04*v20 + 4e-04*v23 - 0.0031*v27 + 7e-04*v29
9	0.0093*v1 + 0.01*v2 + 0.0336*v3 + 0.0127*v4 + 0.0022*v5 + 0.0137*v6 - 0.0075*v7 + 0.0054*v8 + 0.0265*v9 + 0.0069*v11 + 0.0059*v13 + 0.0086*v16 + 5e-04*v18 + 3e-04*v20 + 0.0013*v23 - 2e-04*v24 + 4e-04*v26 + 3e-04*v29
18	0.0098*v1 + 0.01*v2 + 0.034*v3 + 0.0138*v4 + 0.0023*v5 + 0.0136*v6 - 0.0052*v7 + 0.0035*v8 + 0.0121*v9 + 0.0054*v11 - 0.0012*v12 + 0.0052*v13 - 0.0035*v14 + 0.0016*v17 + 7e-04*v18 + 6e-04*v19 + 0.0016*v23 + 0.0013*v26
21	0.0104*v1 + 0.0094*v2 + 0.0288*v3 + 0.0117*v4 + 0.0021*v5 + 0.0144*v6 - 0.0065*v7 + 0.0047*v8 + 0.0122*v9 + 0.0085*v11 - 0.0113*v15 + 3e-04*v20 + 8e-04*v21 - 5e-04*v24

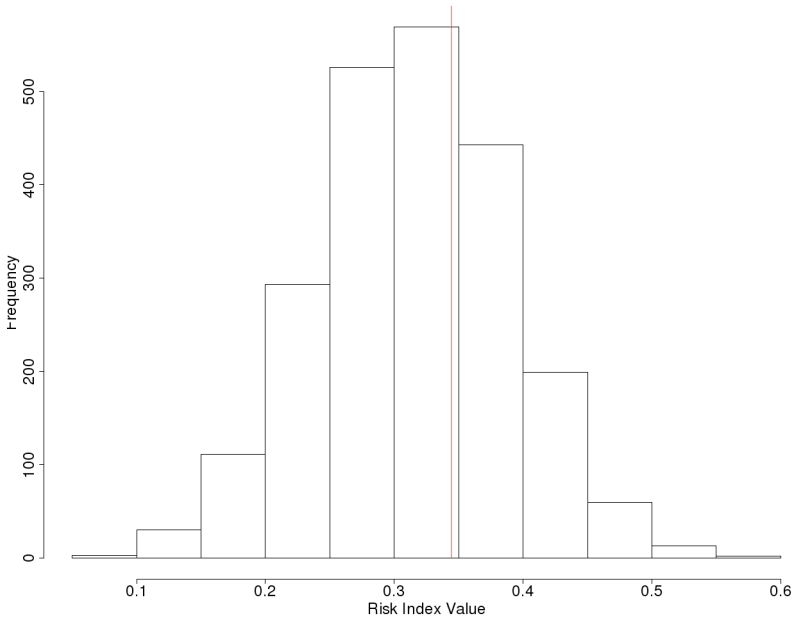


Figure 3-64 Clinical Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #9, Bootstrap Sample #5

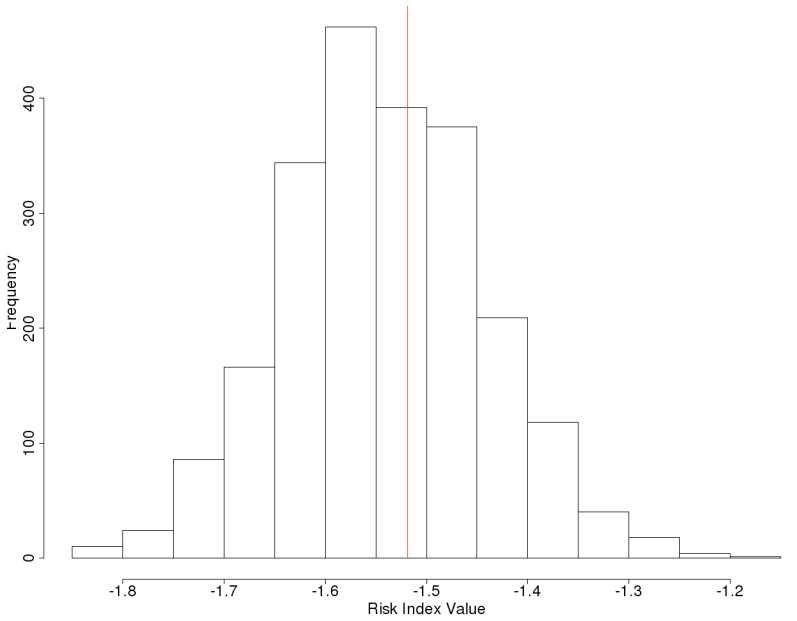


Figure 3-65 Clinical Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #22, Bootstrap Sample #11

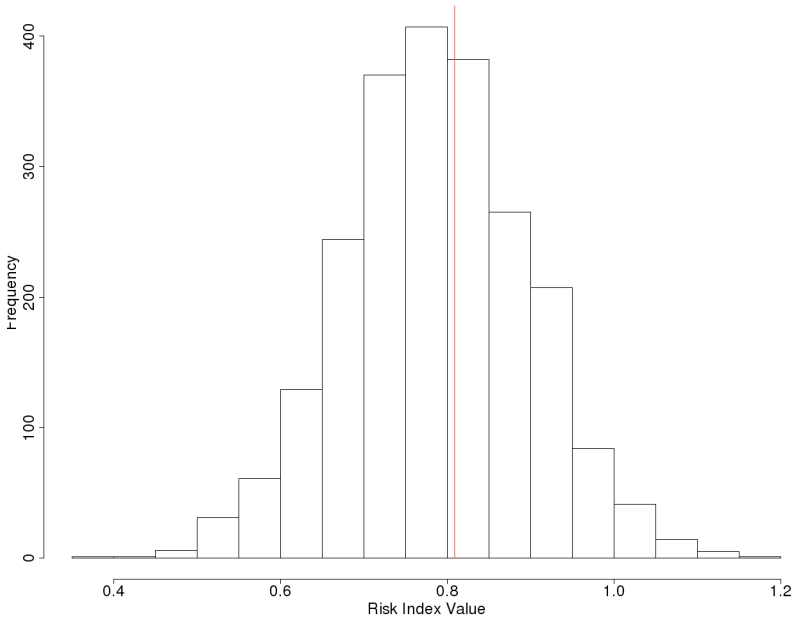


Figure 3-66 Clinical Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #25, Bootstrap Sample #4

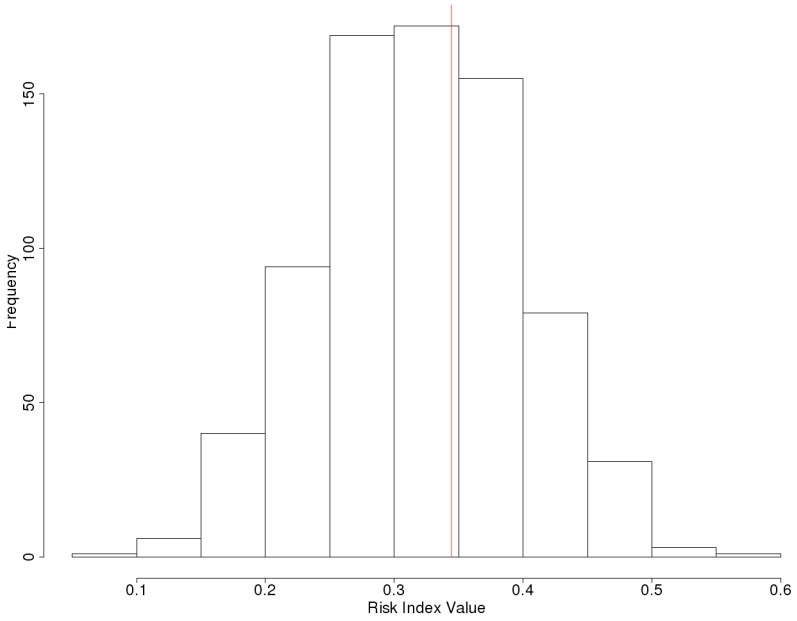


Figure 3-67 Clinical Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #9, Bootstrap Sample #5

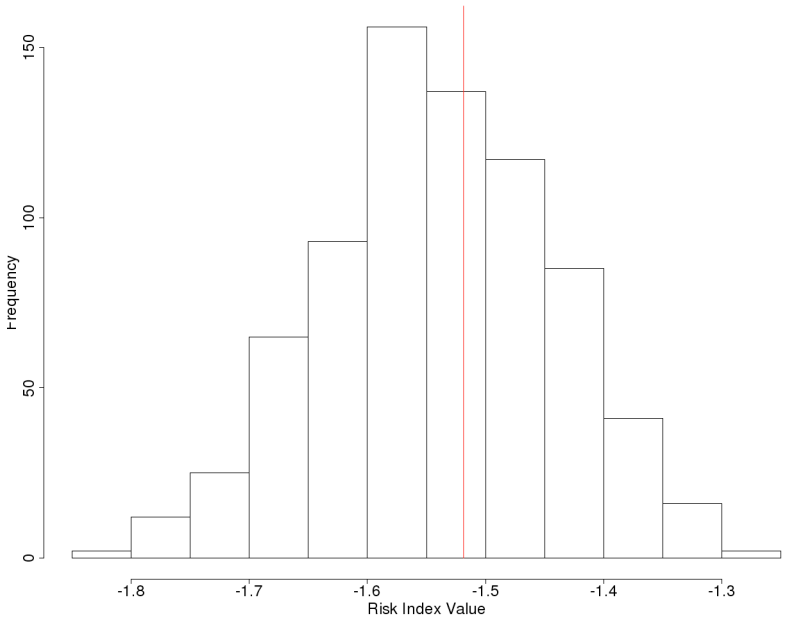


Figure 3-68 Clinical Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #22, Bootstrap Sample #11

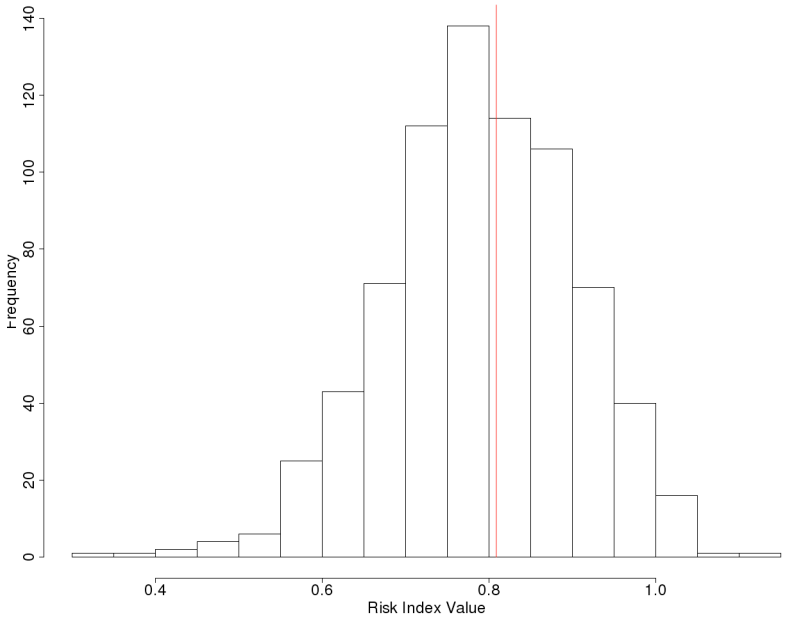


Figure 3-69 Clinical Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #25, Bootstrap Sample #4

Table 3-63 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Large-scale Simulation Dataset #9

		Bootstrap Sample #5 Cutoff Value = 0.940		Bootstrap Sample #12 Cutoff Value = 0.482		Bootstrap Sample #14 Cutoff Value = 0.700		Bootstrap Sample #20 Cutoff Value=0.649		Bootstrap Sample #23 Cutoff Value=-0.482	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	1.006	1	0.451	0	0.841	1	0.683	1	0.497	1
2	0	0.859	0	0.384	0	0.607	0	0.497	0	0.406	0
3	1	0.985	1	0.489	1	0.817	1	0.665	1	0.515	1
4	1	1.177	1	0.651	1	0.947	1	0.829	1	0.685	1
5	0	0.721	0	0.254	0	0.405	0	0.330	0	0.269	0
6	0	0.870	0	0.390	0	0.716	1	0.626	0	0.426	0
7	0	0.726	0	0.244	0	0.440	0	0.452	0	0.265	0
8	0	0.824	0	0.290	0	0.586	0	0.499	0	0.336	0
9	1	0.993	1	0.485	1	0.790	1	0.615	0	0.532	1
10	0	0.871	0	0.429	0	0.651	0	0.485	0	0.440	0
11	0	0.731	0	0.189	0	0.482	0	0.418	0	0.229	0
12	1	1.082	1	0.538	1	0.769	1	0.658	1	0.572	1
13	0	0.889	0	0.310	0	0.569	0	0.539	0	0.357	0
14	0	0.899	0	0.350	0	0.583	0	0.566	0	0.401	0
15	0	0.872	0	0.316	0	0.588	0	0.426	0	0.362	0
16	1	0.986	1	0.501	1	0.656	0	0.590	0	0.541	1
17	1	1.005	1	0.474	0	0.780	1	0.709	1	0.522	1
18	1	1.200	1	0.706	1	0.974	1	0.847	1	0.730	1
19	1	1.165	1	0.658	1	0.898	1	0.708	1	0.697	1
20	0	0.906	0	0.364	0	0.714	1	0.572	0	0.415	0
21	0	0.872	0	0.335	0	0.685	0	0.468	0	0.383	0
22	0	0.876	0	0.353	0	0.732	1	0.560	0	0.383	0
23	0	0.934	0	0.377	0	0.551	0	0.551	0	0.418	0
24	1	0.979	1	0.493	1	0.800	1	0.610	0	0.516	1
25	0	0.777	0	0.264	0	0.483	0	0.287	0	0.268	0

Table 3-64 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Large-scale Simulation Dataset #22

		Bootstrap Sample #11 Cutoff Value = 0.680		Bootstrap Sample #15 Cutoff Value = 0.314		Bootstrap Sample #23 Cutoff Value = 0.736		Bootstrap Sample #23 Cutoff Value= -0.005		Bootstrap Sample #25 Cutoff Value = -0.014	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	0.950	1	0.171	0	1.053	1	-0.004	1	-0.014	0
2	0	0.583	0	0.308	0	0.619	0	0.000	1	-0.010	1
3	0	0.435	0	0.329	1	0.523	0	-0.012	0	-0.019	0
4	0	0.591	0	0.217	0	0.716	0	-0.041	0	-0.011	1
5	1	0.743	1	0.428	1	0.828	1	-0.020	0	-0.018	0
6	0	0.529	0	0.303	0	0.627	0	0.009	1	-0.016	0
7	1	0.668	0	0.319	1	0.820	1	-0.024	0	-0.019	0
8	0	0.523	0	0.430	1	0.665	0	0.000	1	-0.014	0
9	0	0.367	0	0.325	1	0.472	0	-0.024	0	-0.016	0
10	0	0.650	0	0.168	0	0.776	1	-0.031	0	-0.021	0
11	0	0.301	0	0.210	0	0.435	0	-0.012	0	-0.021	0
12	0	0.416	0	0.159	0	0.429	0	-0.006	0	-0.013	1
13	1	0.613	0	0.327	1	0.826	1	-0.012	0	-0.015	0
14	0	0.413	0	0.341	1	0.485	0	-0.026	0	-0.019	0
15	0	0.270	0	0.298	0	0.392	0	-0.010	0	-0.013	1
16	0	0.466	0	0.297	0	0.608	0	-0.025	0	-0.013	1
17	1	0.353	0	0.262	0	0.603	0	-0.010	0	-0.014	1
18	0	0.498	0	0.314	1	0.639	0	-0.016	0	-0.008	1
19	1	0.562	0	0.338	1	0.636	0	-0.028	0	-0.018	0
20	0	0.616	0	0.211	0	0.664	0	-0.012	0	-0.019	0
21	0	0.387	0	0.276	0	0.494	0	0.005	1	-0.012	1
22	0	0.393	0	0.122	0	0.573	0	-0.022	0	-0.014	1
23	1	0.826	1	0.257	0	0.916	1	-0.015	0	-0.011	1
24	1	0.397	0	0.319	1	0.636	0	-0.009	0	-0.019	0
25	1	0.756	1	0.392	1	0.930	1	-0.012	0	-0.015	0

Table 3-65 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical Risk Index Models from Large-scale Simulation Dataset #25

		Bootstrap Sample #4 Cutoff Value = -0.391		Bootstrap Sample #5 Cutoff Value = -0.023		Bootstrap Sample #9 Cutoff Value = -0.601		Bootstrap Sample #18 Cutoff Value= -0.612		Bootstrap Sample #21 Cutoff Value = 0.214	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	-0.503	0	-0.087	0	-0.734	0	-0.747	0	0.139	0
2	0	-0.460	0	0.004	1	-0.720	0	-0.843	0	0.118	0
3	1	-0.382	1	0.019	1	-0.616	0	-0.629	0	0.161	0
4	0	-0.453	0	-0.021	1	-0.609	0	-0.669	0	0.073	0
5	1	-0.269	1	0.014	1	-0.399	1	-0.503	1	0.206	0
6	1	-0.251	1	-0.069	0	-0.387	1	-0.446	1	0.346	1
7	0	-0.438	0	0.017	1	-0.597	1	-0.644	0	0.032	0
8	0	-0.253	1	-0.053	0	-0.498	1	-0.485	1	0.305	1
9	0	-0.396	0	-0.050	0	-0.593	1	-0.735	0	0.021	0
10	0	-0.836	0	-0.077	0	-0.990	0	-1.030	0	-0.026	0
11	0	-0.529	0	-0.060	0	-0.675	0	-0.742	0	0.049	0
12	0	-0.524	0	-0.058	0	-0.633	0	-0.705	0	0.017	0
13	1	-0.333	1	-0.070	0	-0.497	1	-0.623	0	0.239	1
14	0	-0.587	0	-0.041	0	-0.853	0	-0.987	0	-0.207	0
15	1	-0.339	1	-0.067	0	-0.529	1	-0.702	0	0.205	0
16	1	-0.306	1	-0.052	0	-0.427	1	-0.537	1	0.244	1
17	1	-0.272	1	-0.029	0	-0.454	1	-0.538	1	0.343	1
18	1	-0.343	1	-0.026	0	-0.496	1	-0.588	1	0.157	0
19	0	-0.480	0	-0.046	0	-0.720	0	-0.764	0	0.016	0
20	0	-0.615	0	-0.011	1	-0.831	0	-0.839	0	-0.034	0
21	1	-0.234	1	-0.053	0	-0.456	1	-0.583	1	0.218	1
22	1	-0.235	1	-0.020	1	-0.378	1	-0.480	1	0.206	0
23	0	-0.583	0	-0.027	0	-0.777	0	-0.798	0	-0.133	0
24	1	-0.459	0	-0.020	1	-0.657	0	-0.678	0	0.242	1
25	1	-0.063	1	-0.062	0	-0.259	1	-0.426	1	0.522	1

Tables 3-64, 3-65, and 3-66 each show five trimmed Clinical + Genotype risk index models corresponding to the Clinical risk index models shown in Table 3-58, 3-59, and 3-60. Tables 3-67, 3-68, and 3-69 show the risk index values and predictions from the same set of five Clinical + Genotype risk index models from the same three small-scale simulation datasets for a set of 25 individuals randomly selected from the optimization set. Figures 3-70, 3-71, and 3-72 show the full distribution of risk index values in the optimization set for a randomly selected trimmed Clinical + Genotype risk index model from each of the three small-scale simulation datasets. Figures 3-73, 3-74, and 3-74 show the full distribution of risk index values in the independent testing set for a randomly selected trimmed Clinical + Genotype risk index model from each of the three small-scale simulation datasets. As in the previous set of figures, in all six of these figures a red line indicates the cut-off point. All individuals with a risk index value greater than or equal to this cut-off point are predicted as “high risk” and all individuals with a value less than this cut-off point are predicted as “low risk”.

Table 3-66 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #9

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
5	0.0304*v1 + 0.0253*v2 - 0.012*v3 + 0.005*v4 + 0.0096*v5 - 0.0029*v7 - 0.0059*v8 + 0.0032*v13 + 0.005*v14 + 0.0029*v15 + 0.0039*v16 - 0.0028*v18 + 0.0043*v19 - 9e-04*v20 + 0.0042*v23 + 0.0185*v25 + 0.0021*v26 + 0*v27 + 5e-04*v29 + 10.4268*pc15 + 1.5052*pc17 - 2.5219*pc47 + 1.723*pc48 + 1.2779*pc50 - 4.9046*pc55 + 6.5569*pc91 + 16.9002*pc96 - 1.208*pc119 + 1.549*pc137 + 2.9244*pc188 - 1.4788*pc214 + 2.7539*pc277 - 5.6822*pc342 - 0.9887*pc415 - 4.7684*pc425 + 5.6079*pc475 - 3.1944*pc488
12	0.0033*v19 - 2e-04*v27 + 2.1123*pc54 + 0.504*pc58 - 0.2667*pc69 - 2.4151*pc82 - 0.9506*pc83 + 0.0606*pc136 + 2.0214*pc147 + 0.2685*pc191 - 1.6497*pc194 - 1.1582*pc214 - 0.8613*pc264 + 3.5242*pc283 - 0.0481*pc302 + 1.3812*pc312 - 2.1866*pc331 - 3.633*pc347 + 0.4781*pc394 + 2.0683*pc424 + 1.2825*pc436 - 0.7198*pc454
14	0.0282*v1 + 0.0337*v2 - 0.0111*v3 + 0.0065*v4 + 0.0084*v5 + 0.004*v6 - 0.0027*v7 - 0.0067*v8 - 0.0012*v12 + 0.0028*v13 + 0.0036*v14 + 0.0023*v15 + 0.007*v16 + 0.0018*v22 + 0.0044*v23 + 0.0198*v24 + 0.0401*v25 - 0.001*v27 + 0.0012*v28 - 1e-04*v29 - 4.6747*pc110 + 1.8926*pc175 - 3.5362*pc294 - 1.5334*pc363 + 3.8903*pc487
20	1e-04*v17 - 7e-04*v28 - 1.0924*pc21 - 1.852*pc61 + 0.9555*pc66 - 0.3569*pc70 - 0.6179*pc80 + 0.6884*pc105 - 0.0886*pc114 + 1.3183*pc149 - 1.1099*pc158 - 1.1446*pc197 + 0.1409*pc284 - 1.345*pc291 - 0.5679*pc306 + 2.6538*pc319 - 0.1269*pc325 - 0.3187*pc338 + 1.4281*pc362 - 1.4539*pc439 - 0.8631*pc451 + 0.6004*pc454
23	0.0309*v1 + 0.0268*v2 - 0.0127*v3 + 0.009*v5 - 0.0022*v7 - 0.0068*v8 - 0.0022*v12 + 0.0041*v15 + 0.0043*v16 + 3e-04*v17 - 0.0028*v18 + 0.0014*v19 + 9e-04*v20 + 0.001*v21 + 0.0013*v22 + 0.0136*v24 + 0.0029*v26 - 2e-04*v27 + 0.0034*v28 + 1.9072*pc35 + 4.8736*pc57 - 7.2265*pc64 + 5.0892*pc81 - 1.3005*pc108 - 2.942*pc124 + 4.2528*pc142 - 1.0769*pc182 - 2.057*pc186 + 8.186*pc213 - 0.0804*pc275 + 1.4593*pc279 + 3.5455*pc364 + 10.5203*pc372 + 0.6332*pc391 + 0.3351*pc415 + 3.5862*pc454 + 4.7705*pc456 - 0.0665*pc458 + 0.9987*pc477

Table 3-67 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #22

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
11	0.0816*v1 + 0.1228*v2 + 0.0031*v3 + 0.0047*v4 + 0.0048*v5 + 0.0337*v6 + 0.0023*v8 + 0.0129*v9 - 0.0017*v11 + 0.0036*v12 - 0.002*v14 - 0.0053*v15 + 0.0039*v17 + 0.0019*v19 - 0.0016*v21 - 0.0026*v22 - 1.2314*v25 + 0.0026*v28 - 2.9607*pc9 + 10.6006*pc54 - 1.5149*pc80 + 1.5525*pc111 + 5.052*pc159 + 0.2661*pc166 - 1.2869*pc176 + 2.1985*pc217 + 7.781*pc232 - 3.4101*pc248 + 2.994*pc300 - 3.4148*pc314 + 1.1547*pc315 - 4.3981*pc381 + 0.7789*pc404 - 1.8525*pc433 + 9.1892*pc438 - 0.028*pc448 + 1.1909*pc452
15	0.0929*v1 + 0.0031*v3 + 0.0049*v4 + 0.0059*v5 + 0.0236*v6 + 0.0027*v8 + 0.0096*v9 + 0.0015*v10 - 0.0043*v11 + 0.0026*v12 - 8e-04*v13 - 0.0043*v15 + 0.0038*v17 - 0.0013*v18 + 2e-04*v20 + 0.0022*v22 - 0.2143*v25 + 5e-04*v26 + 0.0013*v28 - 8e-04*v29 - 9.3102*pc20 + 2.253*pc54 - 0.1406*pc74 - 0.3688*pc105 - 2.0127*pc237 - 3.3434*pc289 + 2.3869*pc290 - 9.4705*pc329 - 1.7357*pc341 + 4.1052*pc351 + 1.5242*pc357 - 0.4548*pc392 + 5.4061*pc405 + 4.1047*pc458 - 1.8367*pc473
23	4e-04*v13 - 0.001*v18 + 1e-04*v26 + 0.0157*pc21 + 0.6324*pc35 - 0.6543*pc51 + 1.4247*pc54 - 0.6302*pc64 + 0.0406*pc68 + 1.1264*pc74 - 0.2137*pc82 - 0.1152*pc99 - 0.2328*pc192 - 2.6897*pc203 - 1.3432*pc209 + 1.2853*pc240 + 0.4899*pc260 + 0.8595*pc313 + 0.484*pc338 - 0.2808*pc373 - 1.3172*pc429 - 0.2584*pc453 + 1.046*pc485
24	1e-04*v13 - 1e-04*v21 + 1e-04*v23 - 0.0014*v29 + 0.7593*pc26 - 1.0653*pc46 - 2.1887*pc54 - 1.3976*pc128 - 0.2316*pc207 + 1.4773*pc221 - 0.25*pc228 + 0.1307*pc230 - 2.6488*pc231 + 0.7317*pc279 - 1.2732*pc290 + 3.5582*pc316 - 0.1958*pc373 + 0.8422*pc386 + 0.0368*pc404 - 2.4583*pc418 + 0.9927*pc420 - 0.6906*pc427 + 3.5142*pc494 + 0.159*pc498
25	1.9043*v7 + 0.001*v22 - 0.0023*v29 + 0.6937*pc22 - 0.7539*pc47 - 0.4838*pc53 - 1.8419*pc136 + 0.6845*pc178 + 2.5352*pc179 + 0.7473*pc219 - 0.6718*pc223 - 0.2302*pc228 - 2.613*pc234 - 1.1156*pc265 - 1.4993*pc272 - 4.4805*pc279 + 0.1639*pc331 + 1.9301*pc336 - 1.4362*pc365 + 1.1986*pc400 + 1.1249*pc465 + 1.1365*pc483 + 0.718*pc488

Table 3-68 Clinical + Genotype Risk Index Models for Five Randomly Selected Bootstrap Samples from Large-scale Simulation Dataset #25

Bootstrap Sample	Trimmed Clinical + Genotype Risk Index Model
4	0.009*v1 + 0.0101*v2 + 0.0341*v3 + 0.0134*v4 + 0.0024*v5 - 0.0086*v7 + 0.005*v8 + 0.0178*v9 + 0.0067*v11 - 0.0011*v12 + 0.0058*v13 - 0.0034*v14 + 5e-04*v18 + 5e-04*v20 + 6e-04*v22 - 2.1793*pc46 + 2.4691*pc136 - 3.8813*pc217 + 2.5162*pc259 + 3.6647*pc268 + 0.6671*pc351 + 3.5406*pc356 - 6.6699*pc361 - 1.5907*pc467 + 2.8208*pc476
5	0.0106*v1 + 0.0099*v2 + 0.0116*v4 + 0.0024*v5 + 0.0128*v6 + 0.0052*v8 + 0.001*v10 + 0.0061*v11 + 1e-04*v20 + 4e-04*v23 - 0.0031*v27 + 7e-04*v29 - 4.3389*pc37 + 3.4038*pc38 - 3.4022*pc43 - 2.4419*pc47 - 3.8996*pc58 + 0.9099*pc87 + 2.1992*pc89 + 4.4688*pc173 - 1.5497*pc184 - 0.4699*pc219 + 3.6666*pc254 + 0.1264*pc326 - 2.4101*pc328 - 6.1056*pc330 - 0.4309*pc335 - 2.0316*pc339 - 0.731*pc499
9	0.0093*v1 + 0.01*v2 + 0.0336*v3 + 0.0127*v4 + 0.0022*v5 + 0.0137*v6 - 0.0075*v7 + 0.0054*v8 + 0.0265*v9 + 0.0069*v11 + 0.0059*v13 + 0.0086*v16 + 5e-04*v18 + 3e-04*v20 + 0.0013*v23 - 2e-04*v24 + 4e-04*v26 + 3e-04*v29 - 0.7968*pc69 - 0.6128*pc91 - 13.0616*pc106 - 0.286*pc116 - 1.2397*pc135 + 2.1316*pc136 - 1.8439*pc148 - 6.0649*pc186 - 3.7377*pc205 - 2.3404*pc221 - 7.2134*pc273 - 5.5888*pc301 - 5.2375*pc324 + 1.1473*pc365 + 2.3426*pc404 + 1.6862*pc415 - 12.9176*pc492
18	0.0098*v1 + 0.01*v2 + 0.034*v3 + 0.0138*v4 + 0.0023*v5 + 0.0136*v6 - 0.0052*v7 + 0.0035*v8 + 0.0121*v9 + 0.0054*v11 - 0.0012*v12 + 0.0052*v13 - 0.0035*v14 + 0.0016*v17 + 7e-04*v18 + 6e-04*v19 + 0.0016*v23 + 0.0013*v26 - 1.3138*pc360 + 0.7202*pc388 - 2.0318*pc493
21	0.0104*v1 + 0.0094*v2 + 0.0288*v3 + 0.0117*v4 + 0.0021*v5 + 0.0144*v6 - 0.0065*v7 + 0.0047*v8 + 0.0122*v9 + 0.0085*v11 - 0.0113*v15 + 3e-04*v20 + 8e-04*v21 - 5e-04*v24 + 3.201*pc31 + 1.0042*pc264 - 13.3839*pc295

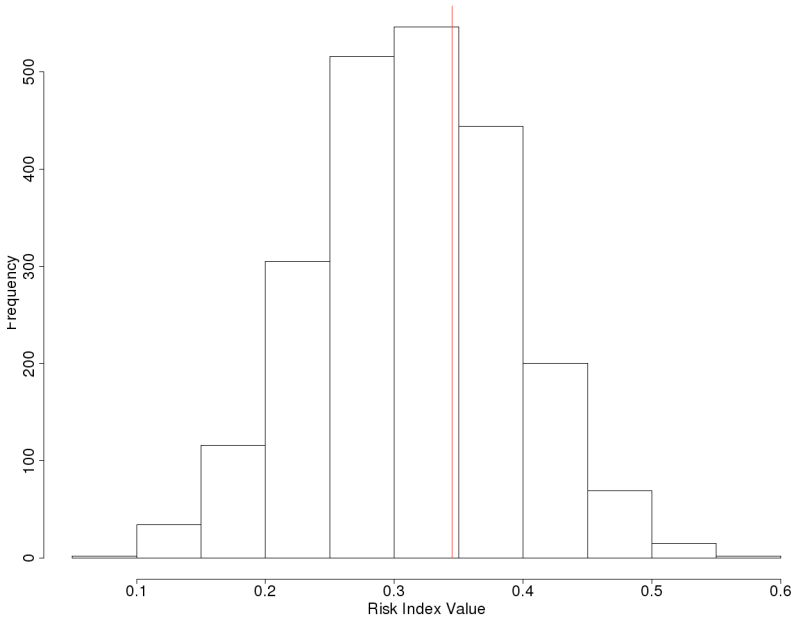


Figure 3-70 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #9, Bootstrap Sample #5

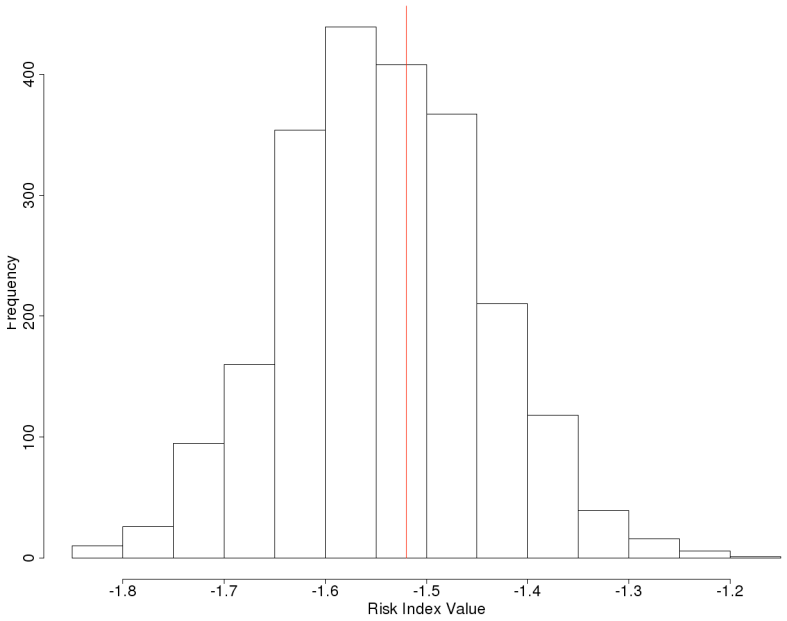


Figure 3-71 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #22, Bootstrap Sample #11

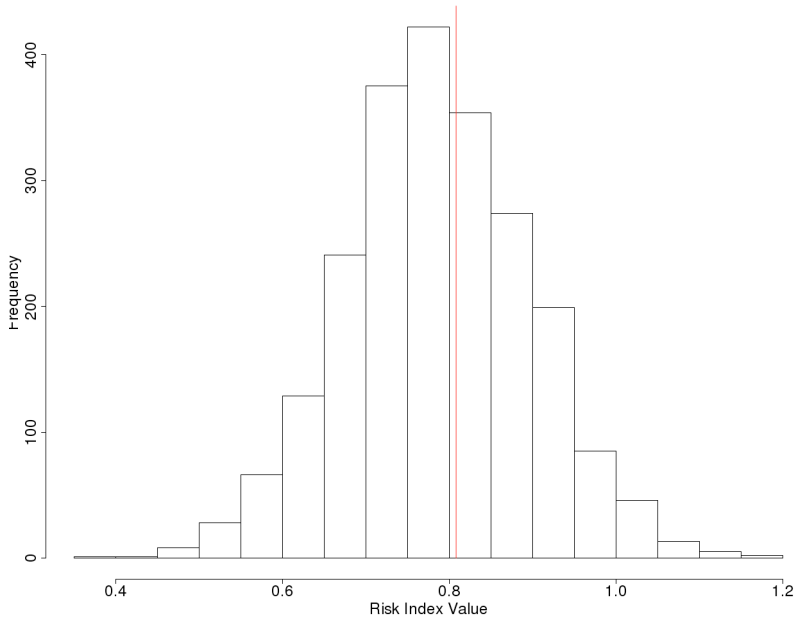


Figure 3-72 Clinical + Genotype Risk Index Value Distribution in the Optimization Set for Large-scale Dataset #25, Bootstrap Sample #4

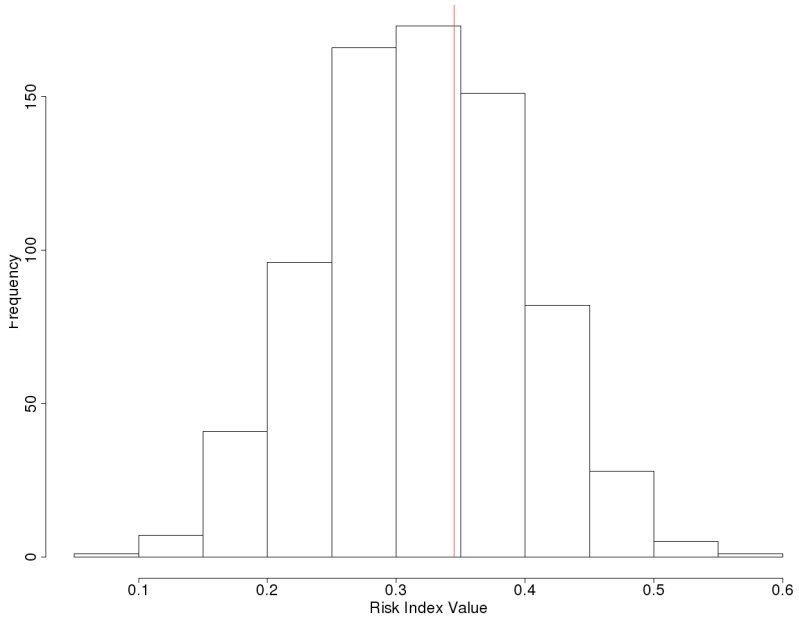


Figure 3-73 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #9, Bootstrap Sample #5

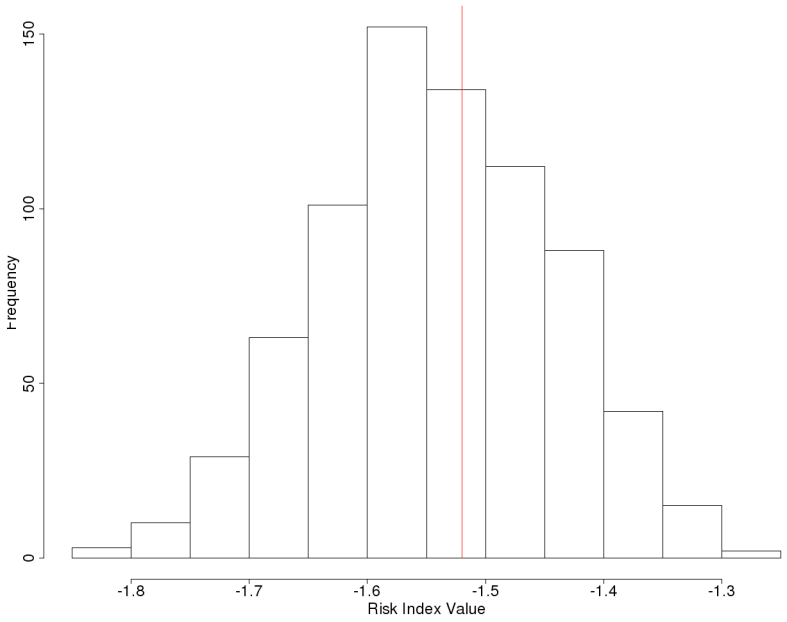


Figure 3-74 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #22, Bootstrap Sample #11

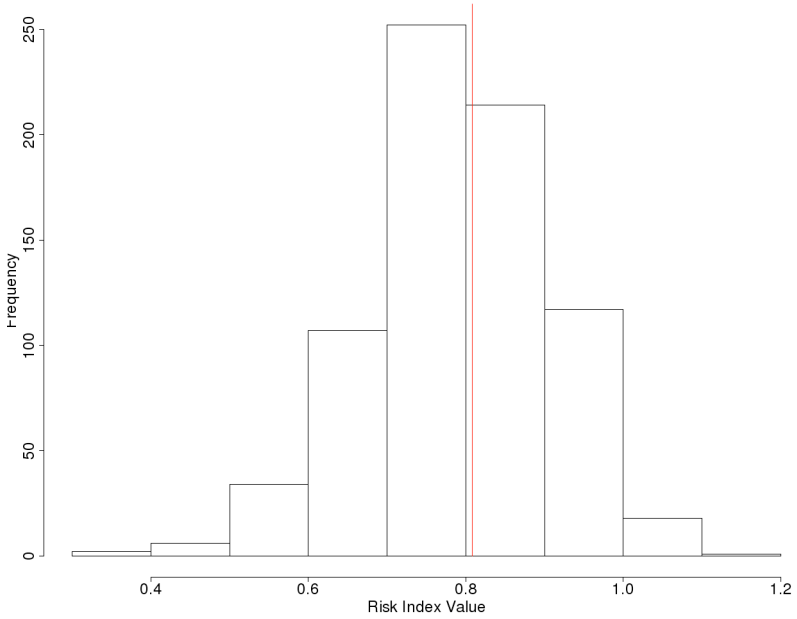


Figure 3-75 Clinical + Genotype Risk Index Value Distribution in the Independent Testing Set for Large-scale Dataset #25, Bootstrap Sample #4

Table 3-69 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Large-scale Simulation Dataset #9

		Bootstrap Sample #5 Cutoff Value = 0.927		Bootstrap Sample #12 Cutoff Value = 0.469		Bootstrap Sample #14 Cutoff Value = 0.696		Bootstrap Sample #20 Cutoff Value=0.647		Bootstrap Sample #23 Cutoff Value=0.509	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	0.990	1	0.463	0	0.838	1	0.683	1	0.546	1
2	0	0.845	0	0.372	0	0.599	0	0.473	0	0.417	0
3	1	0.976	1	0.496	1	0.771	1	0.647	0	0.516	1
4	1	1.162	1	0.641	1	0.987	1	0.820	1	0.731	1
5	0	0.693	0	0.223	0	0.398	0	0.335	0	0.261	0
6	0	0.852	0	0.353	0	0.767	1	0.618	0	0.480	0
7	0	0.704	0	0.252	0	0.428	0	0.471	0	0.317	0
8	0	0.812	0	0.289	0	0.595	0	0.481	0	0.368	0
9	1	0.991	1	0.501	1	0.806	1	0.628	0	0.545	1
10	0	0.842	0	0.417	0	0.626	0	0.466	0	0.452	0
11	0	0.715	0	0.168	0	0.438	0	0.419	0	0.247	0
12	1	1.097	1	0.535	1	0.784	1	0.670	1	0.606	1
13	0	0.873	0	0.270	0	0.575	0	0.532	0	0.373	0
14	0	0.876	0	0.320	0	0.630	0	0.582	0	0.429	0
15	0	0.852	0	0.300	0	0.630	0	0.439	0	0.401	0
16	1	0.976	1	0.502	1	0.706	1	0.575	0	0.569	1
17	1	0.993	1	0.423	0	0.811	1	0.680	1	0.555	1
18	1	1.192	1	0.692	1	0.998	1	0.855	1	0.749	1
19	1	1.147	1	0.645	1	0.939	1	0.702	1	0.702	1
20	0	0.898	0	0.363	0	0.675	0	0.579	0	0.425	0
21	0	0.865	0	0.330	0	0.694	0	0.456	0	0.382	0
22	0	0.854	0	0.327	0	0.771	1	0.569	0	0.422	0
23	0	0.931	1	0.311	0	0.498	0	0.551	0	0.461	0
24	1	0.973	1	0.481	1	0.784	1	0.591	0	0.529	1
25	0	0.788	0	0.205	0	0.540	0	0.304	0	0.292	0

Table 3-70 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Large-scale Simulation Dataset #22

		Bootstrap Sample #11 Cutoff Value = 0.698		Bootstrap Sample #15 Cutoff Value = 0.330		Bootstrap Sample #23 Cutoff Value = 0.747		Bootstrap Sample #23 Cutoff Value=0.021		Bootstrap Sample #25 Cutoff Value = -0.024	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	0.956	1	0.192	0	1.052	1	0.026	1	-0.025	0
2	0	0.580	0	0.326	0	0.641	0	0.031	1	-0.020	1
3	0	0.443	0	0.350	1	0.530	0	0.016	0	-0.024	1
4	0	0.645	0	0.242	0	0.763	1	0.028	1	-0.009	1
5	1	0.751	1	0.447	1	0.852	1	-0.007	0	-0.028	0
6	0	0.520	0	0.318	0	0.641	0	0.035	1	-0.021	1
7	1	0.640	0	0.342	1	0.820	1	0.005	0	-0.029	0
8	0	0.568	0	0.444	1	0.700	0	0.027	1	-0.024	1
9	0	0.352	0	0.330	0	0.476	0	-0.013	0	0.001	1
10	0	0.660	0	0.181	0	0.775	1	-0.003	0	-0.031	0
11	0	0.320	0	0.239	0	0.459	0	0.003	0	-0.031	0
12	0	0.400	0	0.178	0	0.447	0	0.004	0	-0.007	1
13	1	0.595	0	0.336	1	0.855	1	0.001	0	-0.025	0
14	0	0.470	0	0.362	1	0.506	0	0.005	0	-0.030	0
15	0	0.290	0	0.317	0	0.438	0	0.006	0	-0.050	0
16	0	0.474	0	0.310	0	0.630	0	0.006	0	-0.050	0
17	1	0.356	0	0.280	0	0.614	0	0.020	0	-0.024	1
18	0	0.499	0	0.338	1	0.643	0	0.009	0	-0.046	0
19	1	0.592	0	0.352	1	0.632	0	-0.013	0	-0.029	0
20	0	0.596	0	0.231	0	0.672	0	0.018	0	-0.029	0
21	0	0.400	0	0.288	0	0.499	0	0.030	1	-0.022	1
22	0	0.447	0	0.116	0	0.592	0	0.004	0	-0.024	1
23	1	0.865	1	0.289	0	0.941	1	0.001	0	-0.016	1
24	1	0.427	0	0.335	1	0.646	0	0.003	0	-0.042	0
25	1	0.791	1	0.414	1	0.941	1	0.002	0	-0.026	0

Table 3-71 Risk Index Values for 25 Randomly Selected Individuals from the Optimization Set of Five Clinical + Genotype Risk Index Models from Large-scale Simulation Dataset #25

Individual	Outcome	Bootstrap Sample #4 Cutoff Value = -0.438 Risk Index Value	Prediction	Bootstrap Sample #5 Cutoff Value = -0.009 Risk Index Value	Prediction	Bootstrap Sample #9 Cutoff Value = -0.612 Risk Index Value	Prediction	Bootstrap Sample #18 Cutoff Value= -0.606 Risk Index Value	Prediction	Bootstrap Sample #21 Cutoff Value = 0.201 Risk Index Value	Prediction
1	0	-0.568	0	-0.090	0	-0.756	0	-0.766	0	0.115	0
2	0	-0.449	0	0.012	1	-0.719	0	-0.861	0	0.118	0
3	1	-0.354	1	0.030	1	-0.629	0	-0.637	0	0.157	0
4	0	-0.478	0	-0.006	1	-0.629	0	-0.655	0	0.057	0
5	1	-0.315	1	0.019	1	-0.433	1	-0.512	1	0.184	0
6	1	-0.322	1	-0.062	0	-0.442	1	-0.440	1	0.311	1
7	0	-0.488	0	0.013	1	-0.598	1	-0.650	0	0.015	0
8	0	-0.271	1	-0.048	0	-0.508	1	-0.488	1	0.311	1
9	0	-0.441	0	-0.043	0	-0.594	1	-0.727	0	0.021	0
10	0	-0.876	0	-0.066	0	-1.004	0	-0.994	0	-0.018	0
11	0	-0.567	0	-0.042	0	-0.664	0	-0.729	0	0.035	0
12	0	-0.544	0	-0.044	0	-0.633	0	-0.720	0	0.001	0
13	1	-0.424	1	-0.049	0	-0.518	1	-0.617	0	0.225	1
14	0	-0.652	0	-0.033	0	-0.862	0	-0.961	0	-0.207	0
15	1	-0.384	1	-0.064	0	-0.552	1	-0.704	0	0.201	0
16	1	-0.376	1	-0.040	0	-0.442	1	-0.522	1	0.240	1
17	1	-0.272	1	-0.017	0	-0.458	1	-0.529	1	0.333	1
18	1	-0.413	1	-0.016	0	-0.516	1	-0.577	1	0.126	0
19	0	-0.526	0	-0.039	0	-0.701	0	-0.749	0	0.003	0
20	0	-0.660	0	-0.006	1	-0.840	0	-0.853	0	-0.024	0
21	1	-0.254	1	-0.029	0	-0.459	1	-0.581	1	0.231	1
22	1	-0.255	1	-0.025	0	-0.386	1	-0.492	1	0.198	0
23	0	-0.653	0	-0.019	0	-0.783	0	-0.784	0	-0.139	0
24	1	-0.530	0	-0.006	1	-0.661	0	-0.673	0	0.222	1
25	1	-0.035	1	-0.057	0	-0.247	1	-0.441	1	0.496	1

3.6.3 Predictive Performance

After the variable selection procedure is completed and the models are applied to each individual in the independent testing set then the sensitivity, specificity, misclassification, and positive predictive value are measured for both the Clinical and Clinical + Genotype risk index models for each of the 25 large-scale simulation datasets. Table 3-70 shows the means and standard deviations of these measurements. To provide a 95% confidence for these measurements of sensitivity, specificity, misclassification, and positive predictive value, for each independent testing set 1000 bootstrap samples were generated. By making predictions about each individual in these bootstrap samples and calculating the sensitivity, specificity, misclassification, and positive predictive value for each bootstrap sample, 95% confidence intervals were estimated for these measurements in each of the 25 large-scale simulation datasets. The mean and standard deviation of the spread (i.e., range) of these confidence intervals for both the Clinical and Clinical + Genotype risk index model is shown in Table 3-71. Table 3-72 shows the predictive performance and confidence intervals for the three small-scale simulation datasets discussed in Section 3.5.2.

Table 3-72 Means and Standard Deviations of Predictive Performance Estimates for the 25 Large-scale Simulation Datasets

Model	Sensitivity (SD)	Specificity (SD)	Misclassification (SD)	PPV (SD)	AUC (SD)
Clinical	0.758 (0.052)	0.927 (0.024)	0.125 (0.021)	0.827 (0.043)	0.931 (0.011)
Clinical + Genotype	0.749 (0.052)	0.930 (0.024)	0.126 (0.021)	0.832 (0.044)	0.931 (0.011)

Table 3-73 Means and Standard Deviations of Predictive Performance 95% Confidence Intervals for the 25 Large-scale Simulation Datasets

Model	Mean Range of the 95% Confidence Interval (SD)			
	Sensitivity	Specificity	Misclassification	PPV
Clinical	0.108 (0.009)	0.044 (0.007)	0.047 (0.004)	0.101 (0.008)
Clinical + Genotype	0.110 (0.009)	0.043 (0.008)	0.046 (0.004)	0.100 (0.010)

Table 3-74 Predictive Performance Estimates for Three Large-scale Simulation Datasets

Dataset	Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
9	Clinical	0.72 (0.665-0.777)	0.887 (0.86-0.914)	0.166 (0.138-0.194)	0.748 (0.695-0.805)
	Clinical + Genotype	0.703 (0.648-0.76)	0.904 (0.878-0.93)	0.16 (0.134-0.186)	0.774 (0.719-0.829)
22	Clinical	0.833 (0.784-0.88)	0.937 (0.916-0.957)	0.096 (0.075-0.117)	0.862 (0.816-0.902)
	Clinical + Genotype	0.838 (0.791-0.886)	0.937 (0.917-0.958)	0.095 (0.073-0.113)	0.863 (0.82-0.905)
25	Clinical	0.809 (0.758-0.856)	0.908 (0.883-0.931)	0.124 (0.103-0.148)	0.806 (0.756-0.854)
	Clinical + Genotype	0.797 (0.746-0.847)	0.916 (0.892-0.94)	0.123 (0.101-0.145)	0.817 (0.767-0.864)

Using the number of models predicting an individual in the independent testing set as "high risk", receiver operator characteristic (ROC) curves were generated for the Clinical and Clinical + Genotype risk index model for each of the 25 large-scale simulation datasets, and the AUC for the ROC curve was estimated. The average AUC for the Clinical risk index models was 0.931 (SD = 0.021), and the average AUC for the Clinical + Genotype risk index models was 0.931 (SD = 0.022). Figure 3-76, 3-77, and 3-78 show the ROC curves for the Clinical risk index model the three small-scale simulation datasets discussed in sections 3.5.2, and Figure 3-79, 3-80, and 3-81 show the ROC curve for the Clinical + Genotype risk index model from those three selected small-scale simulation datasets.

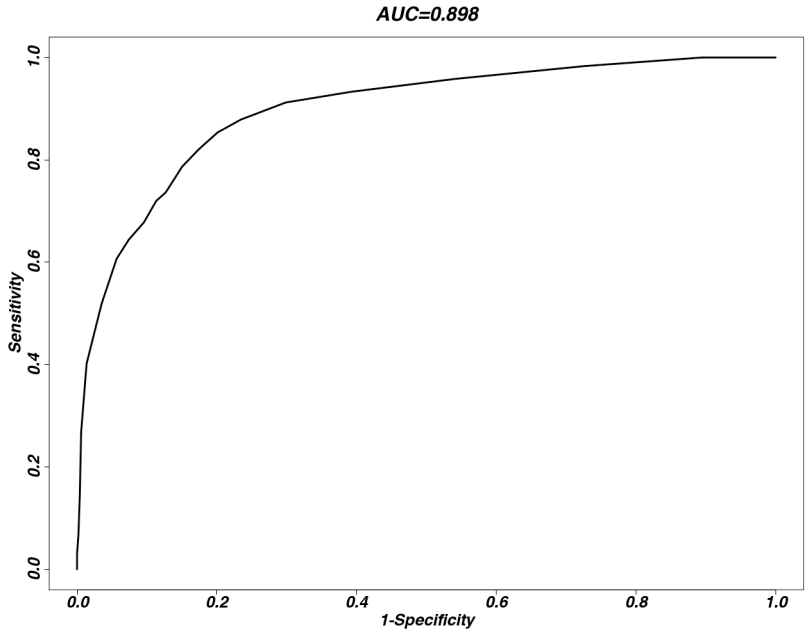


Figure 3-76 ROC Curve of the Clinical Risk Index Model for Large-scale Simulation Dataset #9

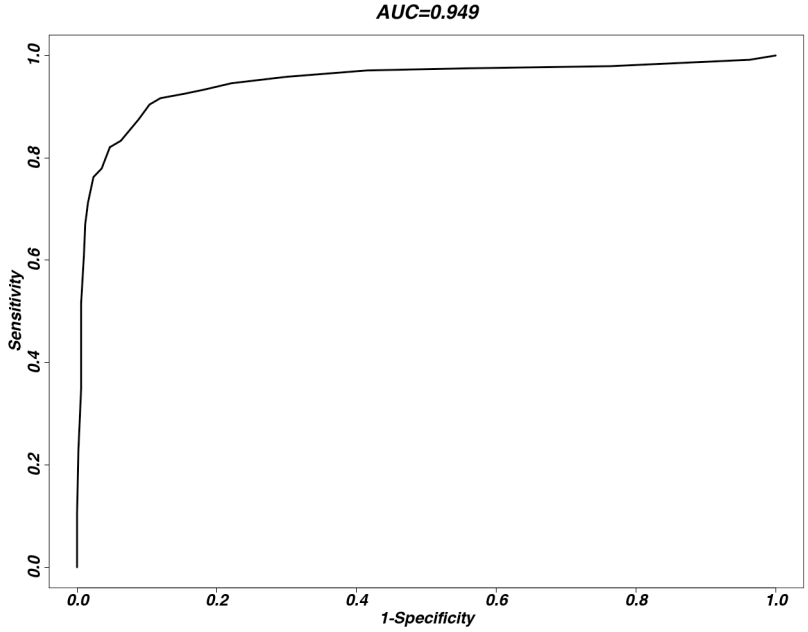


Figure 3-77 ROC Curve of the Clinical Risk Index Model for Large-scale Simulation Dataset #22

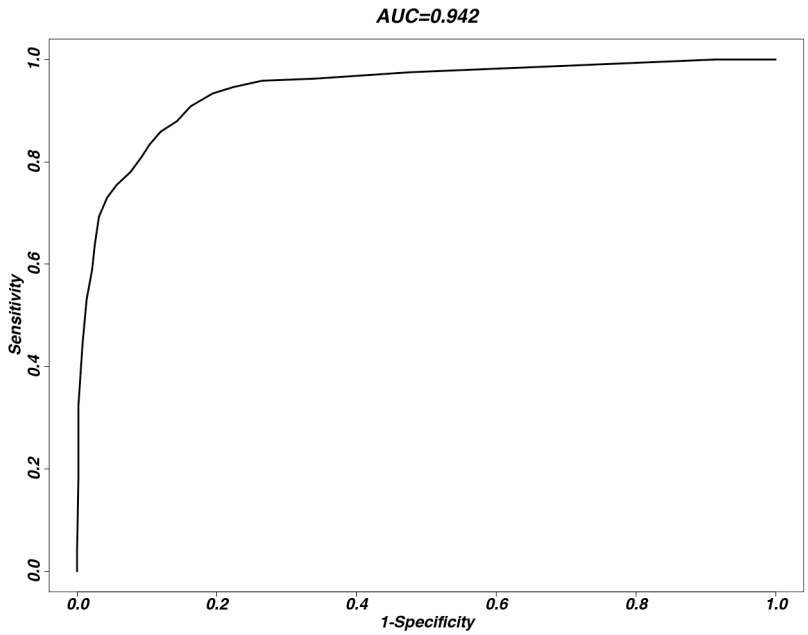


Figure 3-78 ROC Curve of the Clinical Risk Index Model for Large-scale Simulation Dataset #25

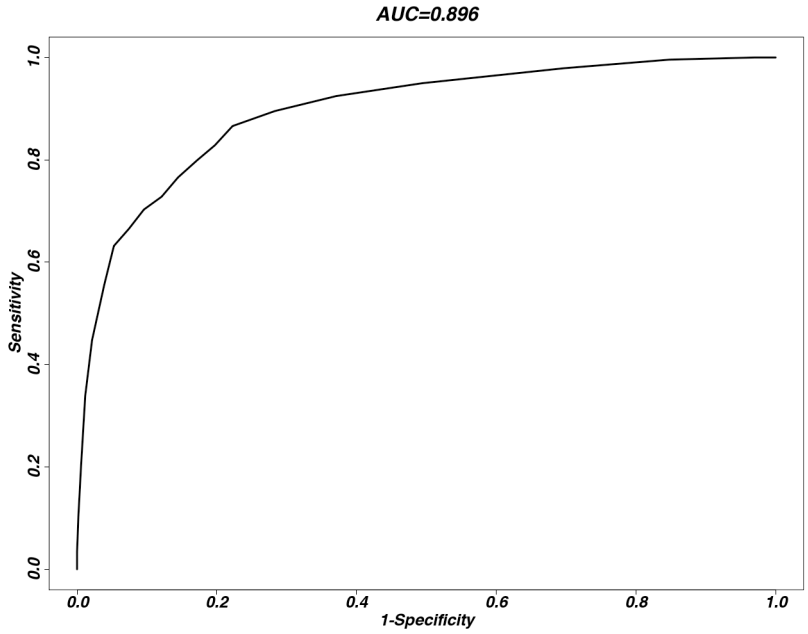


Figure 3-79 ROC Curve of the Clinical + Genotype Risk Index Model for Large-scale Simulation Dataset #9

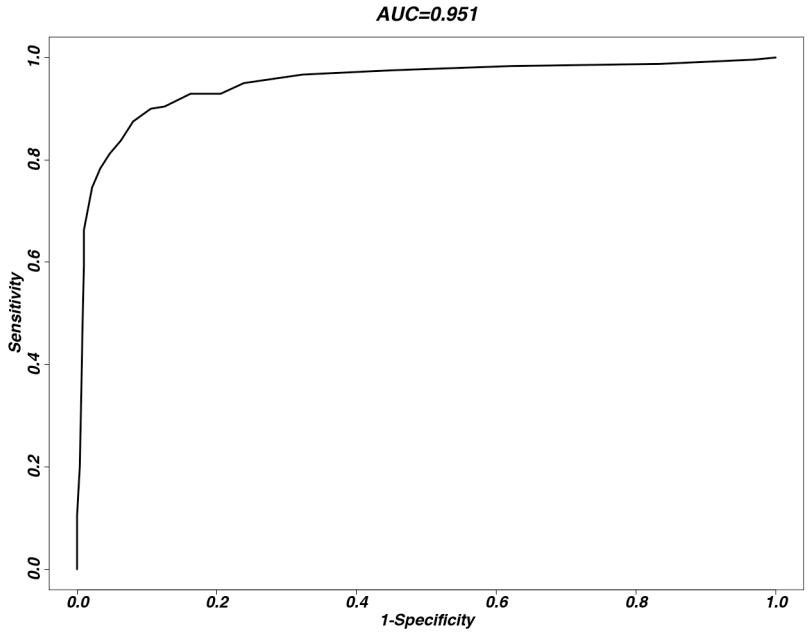


Figure 3-80 ROC Curve of the Clinical + Genotype Risk Index Model for Large-scale Simulation Dataset #22

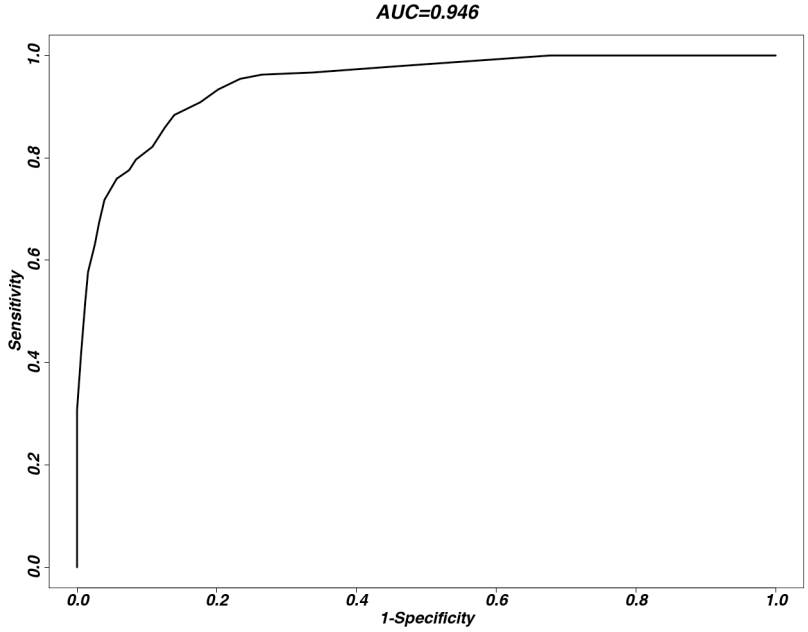


Figure 3-81 ROC Curve of the Clinical + Genotype Risk Index Model for Large-scale Simulation Dataset #25

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. The proportion of models that predict that an individual as high risk, then, represents the predicted probability of an individual developing the outcome. For each individual a 95% confidence interval can be constructed as described in Section 3.2.3

3.6.4 Random Forest Comparison

For each of the 25 large-scale simulation datasets a random forest was generated using the optimization set created by the risk index procedure. The forests were generated using the methodology given in Section 3.2.4. For each of the random forests an ROC curve was generated and the AUC was estimated. The mean AUC for the random forest models was 0.856 (SD = 0.022). Figures 3-82, 3-83, and 3-84 show the ROC curve for the random forest generated from the three small-scale simulation datasets described in Section 3.5.2. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set using a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of 5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the individual to be "high risk". Table 3-73 shows the mean and standard deviation of the sensitivity, specificity, misclassification, and PPV for a range of different proportions.

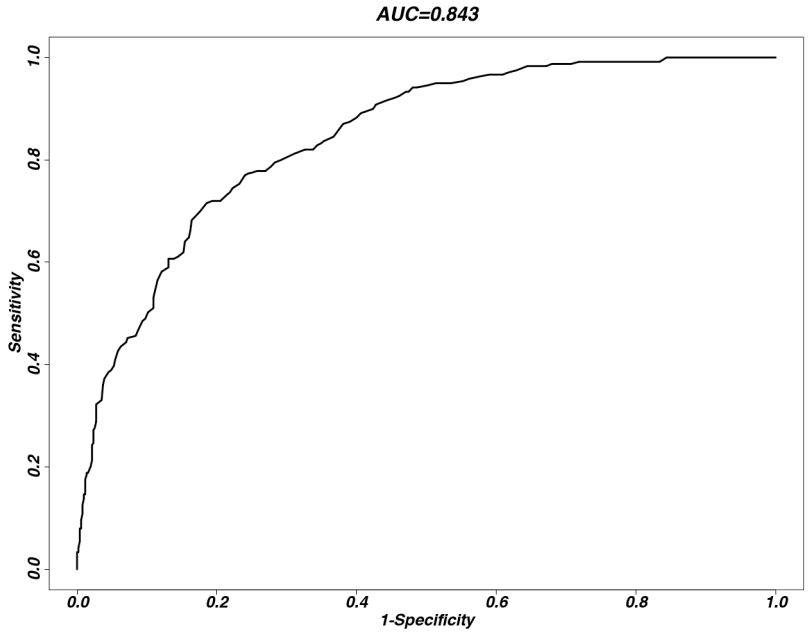


Figure 3-82 ROC Curve of the Random Forest Generated for Large-scale Simulation Dataset #9

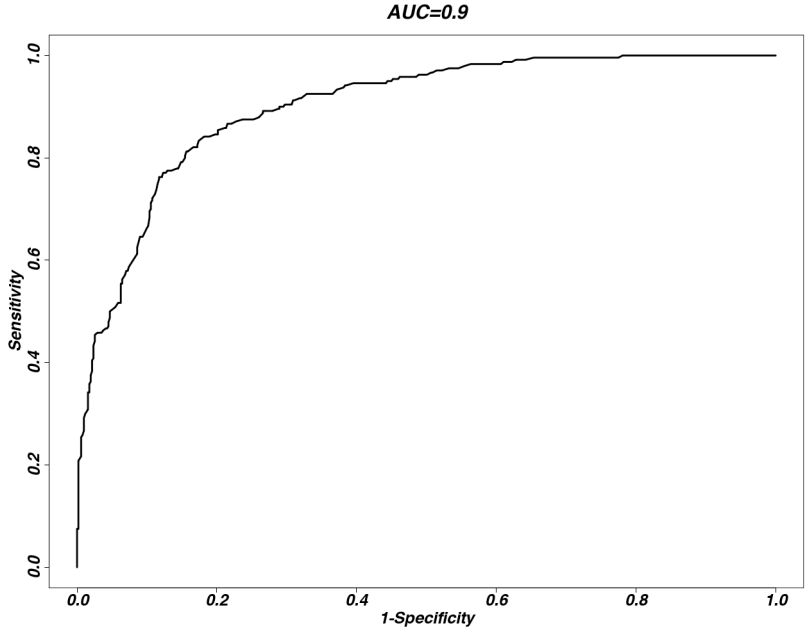


Figure 3-83 ROC Curve of the Random Forest Generated for Large-scale Simulation Dataset #22

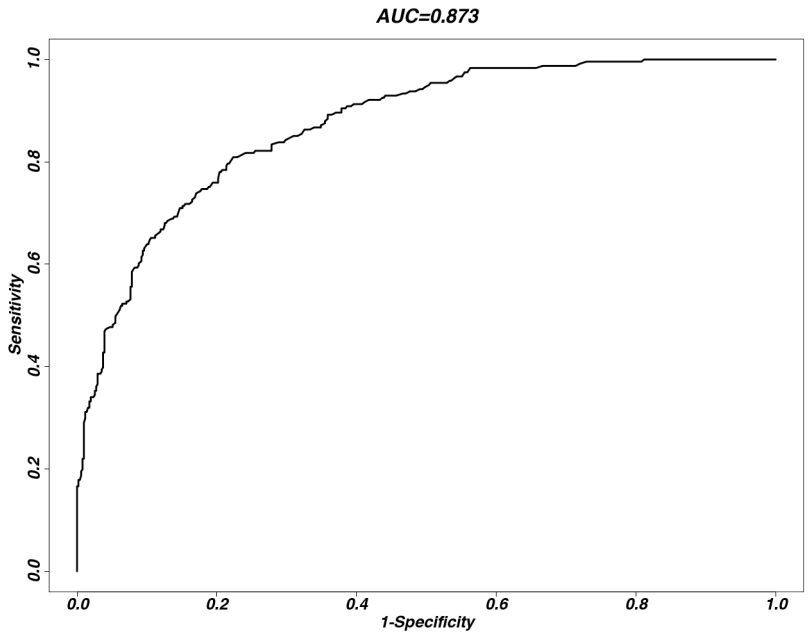


Figure 3-84 ROC Curve of the Random Forest Generated for Large-scale Simulation Dataset #25

Table 3-75 Means and Standard Deviations of Performance Estimates of the Random Forest Models Generated from the 25 Large-scale Simulation Datasets

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	1 (0.001)	0.009 (0.012)	0.684 (0.017)	0.312 (0.016)
0.1	0.999 (0.003)	0.077 (0.052)	0.637 (0.036)	0.327 (0.019)
0.15	0.99 (0.012)	0.228 (0.087)	0.536 (0.056)	0.367 (0.027)
0.2	0.964 (0.025)	0.395 (0.094)	0.428 (0.059)	0.42 (0.034)
0.25	0.912 (0.034)	0.556 (0.077)	0.333 (0.046)	0.483 (0.038)
0.3	0.841 (0.034)	0.685 (0.052)	0.266 (0.032)	0.548 (0.039)
0.35	0.747 (0.04)	0.785 (0.035)	0.227 (0.025)	0.612 (0.04)
0.4	0.653 (0.069)	0.854 (0.026)	0.208 (0.022)	0.67 (0.042)
0.45	0.546 (0.093)	0.909 (0.022)	0.204 (0.023)	0.732 (0.046)
0.5	0.439 (0.103)	0.95 (0.015)	0.209 (0.028)	0.8 (0.04)
0.55	0.326 (0.096)	0.977 (0.011)	0.225 (0.029)	0.865 (0.043)
0.6	0.198 (0.076)	0.992 (0.008)	0.255 (0.025)	0.923 (0.056)
0.65	0.104 (0.06)	0.997 (0.004)	0.28 (0.023)	0.96 (0.05)
0.7	0.042 (0.041)	0.999 (0.001)	0.298 (0.02)	0.978 (0.04)
0.75	0.012 (0.024)	1 (0)	0.306 (0.017)	0.99 (0.037)

3.6.5 Conclusion

The results of the large-scale simulation study using the top 500 principal components are extremely promising, and demonstrate robust predictive performance. As with the large-scale simulation study using the 500 most highly associated SNPs, both the predictive performance estimates and the AUC for the ROC curves for the 25 large-scale simulation datasets are noticeably higher than the small-scale simulation studies. The average misclassification and PPV are higher than any average misclassification or PPV yielded by the random forest model, and the random forests models were not able to provide the same levels of high sensitivity and high specificity simultaneously. Unlike the previous results, however, the Clinical + Genotype risk index model does not have a significantly higher average AUC than the Clinical risk index model ($p=0.98$), however the average AUC for the Clinical + Genotype risk index model is significantly greater than the average AUC of the random forest model ($p=4.4e-16$).

Chapter 4

The Application of the Risk Index Methodology to the Framingham Heart Study

4.1 The Framingham Heart Study

The Framingham Heart Study (FHS) is a large, multi-generational study that has collected health data from residents of Framingham, MA for nearly sixty years. The study was originally developed to allow the study of the epidemiology of cardiovascular disease (Dawber, et al, 1951). The investigators' initial assumption was that cardiovascular disease, unlike infectious diseases, which had received the bulk of epidemiological attention up to that time, has multiple causes and develops over a fairly long span of time. Working from this perspective, the investigators developed a study design calling for approximately 6,000 people who, for practical purposes, would be drawn from a single town of between 25,000 and 50,000 residents. The National Institutes of Health, working with the Massachusetts Health Commission, chose Framingham, MA as the site for this study.

In 1949 the initial cohort of 5,209 subjects were recruited, and by 1952 the initial examination, consisting of a detailed medical history, a comprehensive physical exam including anthropometric measurements, x-rays, electrocardiography, and blood analysis was completed (Dawber, et al, 1951). The ages of the subjects in the initial cohort was 28-62 years, with a mean age of 44.14 years, and 54.7% of the subjects were female (Dawber, et al, 1957). In 1971, 5,124 offspring of the original FHS cohort, along with the

offspring's spouses, were recruited (Feinleib, et al, 1975). The ages of these participants at enrollment ranged from 12-60 years, with a mean of 39.16 years, and 51.5% of the offspring cohort were female. In 2002, 4,095 grandchildren of members of the original cohort were recruited into the third generation cohort (Splansky, et al, 2007). These subjects ranged in age from 19-71 years, with a mean age of 40.16 years, and 53.3% of the subjects were female.

The 14,158 subjects of this study have had biennial exams that assess clinical measures of health, such as blood pressure and biochemical assays, as well as medical history, demographic factors, and psychological, social, and economic measures. Recently, genome-wide genotyping was performed on 6,575 subjects using the Affymetrix Genome-wide Human SNP Array 5.0, which genotypes ~500K SNPs. Other genotyping data, including a 100K SNP array and a 50K SNP array, both from Affymetrix, have also been collected.

4.1.1 Sample Selection for Risk Index Evaluation

This analysis of the FHS data examines the ability of the risk index procedure to combine clinical and genotypic data to make prognostic predictions about an individual's risk of developing a disease. It requires that subjects have both sufficient follow-up time for assessment of disease onset and available DNA samples for genotyping. This analysis focuses on the FHS offspring cohort, which has sufficient follow-up time, but excludes the third-generation cohort, whose follow-up time is insufficient, and the original cohort, which has a small number of subjects who have had DNA collected for genotyping. Of

the 5,214 subjects in the FHS offspring cohort, 2,817 individuals have had DNA collected and been genotyped. These 2,817 subjects were used as the sample for all of the analyses that have been performed on the FHS data..

4.2 Definition of Outcomes

This portion of the dissertation will focus on three outcomes: ten-year incident hypertension, ten-year incident type 2 diabetes (referred to after this as simply "ten-year incident diabetes"), and prevalent hypertension. These diseases represent a range of prevalence (approximately 10% to approximately 35%) representative of common chronic diseases and are known to be strongly influenced by genetic factors. For both incident hypertension and incident diabetes the risk index procedure will be used to identify individuals at high risk of developing the outcome within ten years. Clinical, biochemical, and other predictive variables obtained at the beginning of a ten-year interval will be used to predict the outcome, which is assessed at the end of that ten-year interval. For the offspring cohort the available follow-up time covers only 12 years, and the data available from their first exam is significantly less than what is available from their second exam. Therefore, the ten-year window spanning offspring exam two and offspring exam seven will be used as the source data for both outcomes. For prevalent hypertension the risk index procedure will be used to identify individuals that are hypertensive at exam two.

4.2.1 Ten-Year Incident Hypertension

Using the American Heart Association's diagnostic criteria for hypertension, a subject will be considered to have hypertension in exam two if their average systolic blood pressure is greater than 140 mm Hg, their average diastolic blood pressure is greater than 90 mm Hg, or they are currently taking anti-hypertensive medication (Chobanian, et al, 2003). The 2,283 subjects who began exam two without hypertension will be used for this analysis. Using the same criteria, a subject will be considered to have hypertension in exam seven if their average systolic blood pressure is greater than 140 mm Hg, their average diastolic blood pressure is greater than 90 mm Hg, or they are currently taking anti-hypertensive medication (Chobanian, et al, 2003). Of the 2,283 subjects who did not have hypertension in exam two, 777 (34.0%) developed hypertension by exam seven.

4.2.2 Ten-Year Incident Diabetes

Neither exam two nor exam seven contains a pre-defined variable indicating whether a subject has diabetes. For both exams, however, information on both fasting blood glucose levels and anti-diabetic medications is available. Using the National Diabetes Data Group criteria, a subject will be considered to have diabetes if their fasting blood glucose level is greater than 126 mg/dL or they are taking anti-diabetic medication (National Diabetes Data Group, 1979). At exam two, 2,746 individuals do not have diabetes, and 253 (9.2%) of them develop diabetes by exam seven.

4.2.3 Prevalent Hypertension

Using the American Heart Association's diagnostic criteria for hypertension, a subject will be considered to have hypertension in exam two if their average systolic blood pressure is greater than 140 mm Hg, their average diastolic blood pressure is greater than 90 mm Hg, or they are currently taking anti-hypertensive medication (Chobanian, et al, 2003). Of the 2,817 individuals for whom genotypes are available, 534 (19.0%) are hypertensive at exam two. While hypertension is a straight-forward disease to diagnose, requiring only blood pressure measurements, it is also an excellent model of a complex, multi-factorial disease. For this reason it makes an excellent test of the risk index procedure's ability to identify individuals that have a complex disease, a potentially important application, especially for diseases which are harder to diagnose than hypertension and diabetes, such as auto-immune disorders. Identifying individuals who are at high risk of currently having these diseases could allow physicians to target potentially invasive diagnostic procedures to those most likely to require them.

4.3 Predictor Variable Selection

The FHS data is rich with variables, with each biennial exam comprised of several hundred questions, measurements, and laboratory assessments. This analysis will focus on a fairly small number of clinical variables that are relatively easy to obtain and would be routinely collected in a patient care setting. These variables are age (yrs), sex, weight (lbs), height (in), systolic blood pressure (mm Hg), diastolic blood pressure (mm Hg), total cholesterol (mg/dL), high-density lipoprotein level (mg/dL), low-density lipoprotein level (mg/dL), triglycerides (mg/dL), ever smoked, current smoking status, weekly

alcohol consumption, marital status, left ventricular mass (g), left ventricular ejection fraction (%), blood glucose (mg/dL), blood urea nitrogen (mg/dL), total serum protein level (mg/dL), serum albumin level (mg/dL), serum bilirubin level (mg/dL), serum alkaline phosphatase level (mg/dL), and serum creatine level (mg/dL). Descriptive statistics for these variables in the 2,817 individuals for which genotypes are available are given in Table 4-1.

Table 4-1 Descriptive Statistics of Predictor Variables

Variable	Mean	SD	Range
Age (yrs)	43.2	9.68	17-70
Weight (in)	160.3	34.1	79-326
Height (lbs)	66	3.8	56-79
Cholesterol (mg/dl)	201.6	38.07	52-511
High Density Lipoprotein Cholesterol (mg/dl)	48.98	13.31	14-111
Low Density Lipoprotein Cholesterol (mg/dl)	129.2	34.21	7-311
Triglycerides (mg/dl)	345	287.44	37-6,539
Weekly Alcohol Consumption	3.5	4.88	0-57
Left Ventricular Mass (g)	186.1	55.67	65-476
Left Ventricular Ejection Fraction (%)	74.6	4.36	44-90
Blood Glucose (mg/dl)	97.5	16.1	50-339
Blood Urea Nitrogen (mg/dl)	15.6	4.09	4-49
Serum Protein (mg/dl)	72.2	4.28	54-92
Albumin (mg/dl)	44.5	3.01	34-56
Bilirubin (mg/dl)	72.8	34.48	10-340
Alkaline Phosphotase (mg/dl)	26.6	9.74	4-106
Creatine (mg/dl)	11.5	2.39	4-26
Bilirubin/Creatine Ratio	139.9	42.78	44-467
White Blood Cell Count	63.3	17.57	29-243
Red Blood Cell Count	477.6	42.49	335-685
Hemoglobin	145.3	13.37	79-186
Average Sytolic Blood Pressure (mm Hg)	120.7	15.69	82-203
Average Diastolic Blood Pressure (mm Hg)	77.51	9.52	45-113

4.4 Genotype Variable Selection

As described above, the FHS has extensively genotyped subjects, using Affymetrix platforms measuring 50K, 100K, and 500K genome-wide genotypes. For these analyses, because of constraints of scale and computing resources, the 50K genome-wide genotypes will be used. Focusing solely on autosomal polymorphisms, genotypes were available for 48,071 SNPs. Figure 4-1 shows the calling rates for the 48,071 SNPs in the 2,817 individuals being examined in this project. This graph shows that the majority of SNPs have calling rates of >99% (i.e., fewer than 1% of individuals could not be assigned a genotype for that SNP). To improve the quality of the genotypes and reduce the number of missing genotypes, the HelixTree SNP Variation Suite (Golden Helix) was used to perform SNP genotype imputation. Imputed genotypes with a probability of 75% or greater were retained, and those with a probability lower than 75% were left as missing data. 1,935,361 missing genotypes out of 3,394,984 missing genotypes (57.0%) were imputed with this approach. One observation that potentially accounts for the relatively low proportion of missing genotypes recovered by imputation is that while many subjects were missing a very small number of genotypes, a small subset of individuals were missing the majority of genotypes. Figure 4-1 shows that after imputation the distribution of minor allele frequencies for this collection of SNPs. 4,788 SNPs had a minor allele frequency of less than 5%, and 2,244 SNPs were monomorphic. Monomorphic SNPs were excluded from the analysis.

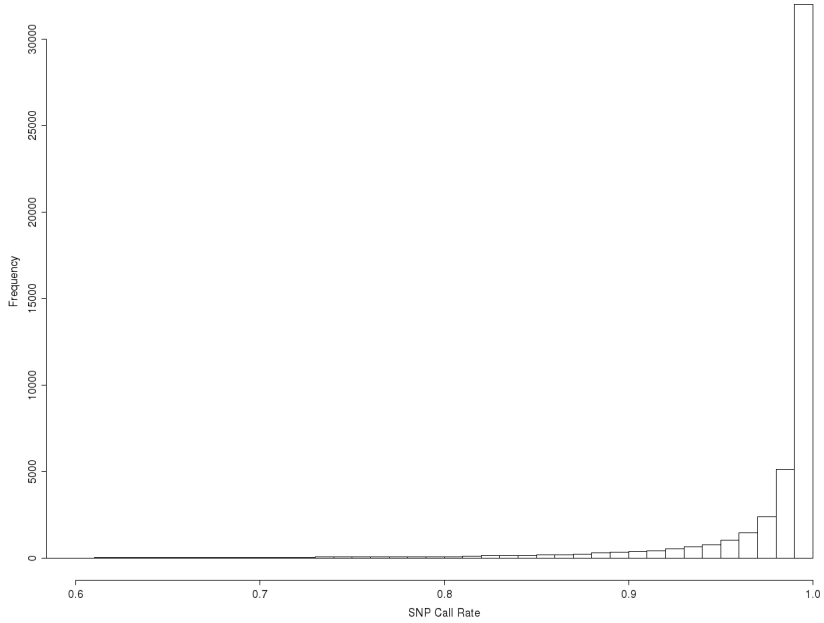


Figure 4-1 Histogram of SNP Call Rates

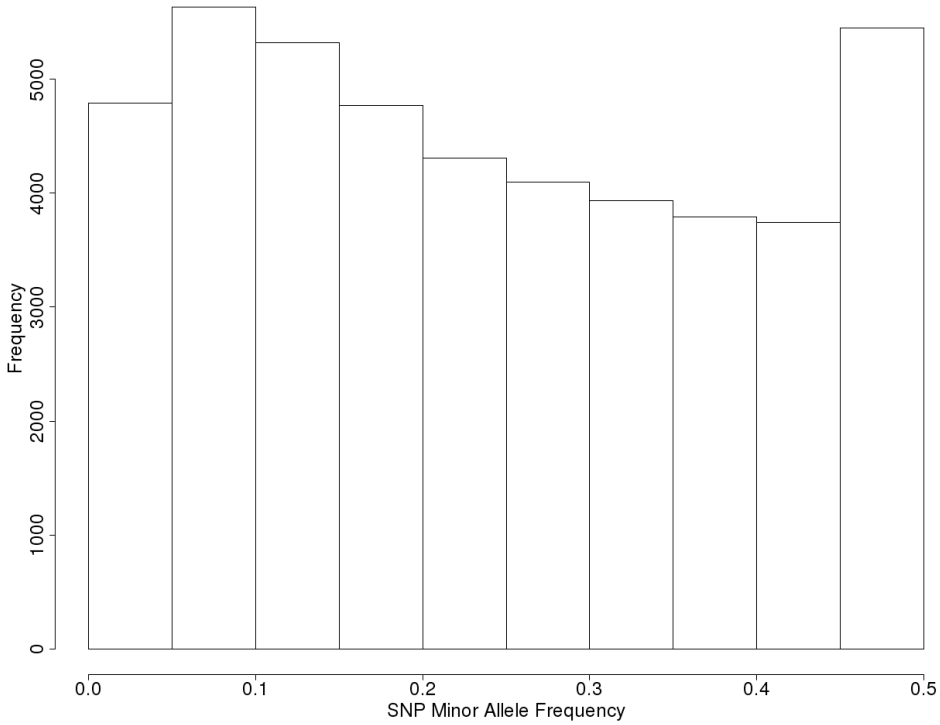


Figure 4-2 Histogram of SNP Minor Allele Frequencies

4.4.1 SNP Selection

To reduce the potential search space and to identify SNPs most likely to be predictive of the outcome being examined, a simple SNP selection procedure was used. The association between each SNP and each of the outcomes was assessed using logistic regression, and the 500 SNPs with the smallest p-values for a particular outcome were selected. Each SNP was encoded using a genotypic model, with the major homozygous genotype being marked as the reference group, and individual coefficients estimated for the heterozygous genotype and the minor homozygous genotype. These SNPs were then used as the genotype predictors for the risk index procedure for that outcome.

4.4.2 Principal Components Analysis

Principal Components Analysis (PCA) will be used as a second procedure to reduce the number of genotype variables being considered (See Chapter 2 for a discussion of PCA). The SMARTPCA program (a part of EIGENSTRAT (Patterson, et al, 2006)) that was used to generate the principal components for this analysis requires a dataset with no missing values. To ensure that only the smallest number of polymorphisms and individuals were excluded, a staged removal procedure was used. First, polymorphisms with a genotyping success rate of less than 99.5%were removed from the data. Next, individuals who were missing more than 0.5% of potential genotypes are removed. Finally, SNPs with less than complete genotyping success were removed. After this data cleaning, 17,268 SNPs in 2,652 people were available for PCA. Using the smartpca program, 500 principal components were estimated. Eigenvectors were computed for each individual for each of the 500 components. The percentage of variance explained by

each component is shown in Figure 4-3. The first 500 components (indicated in red in Figure 4-3) account for 44.6% of the total variance.

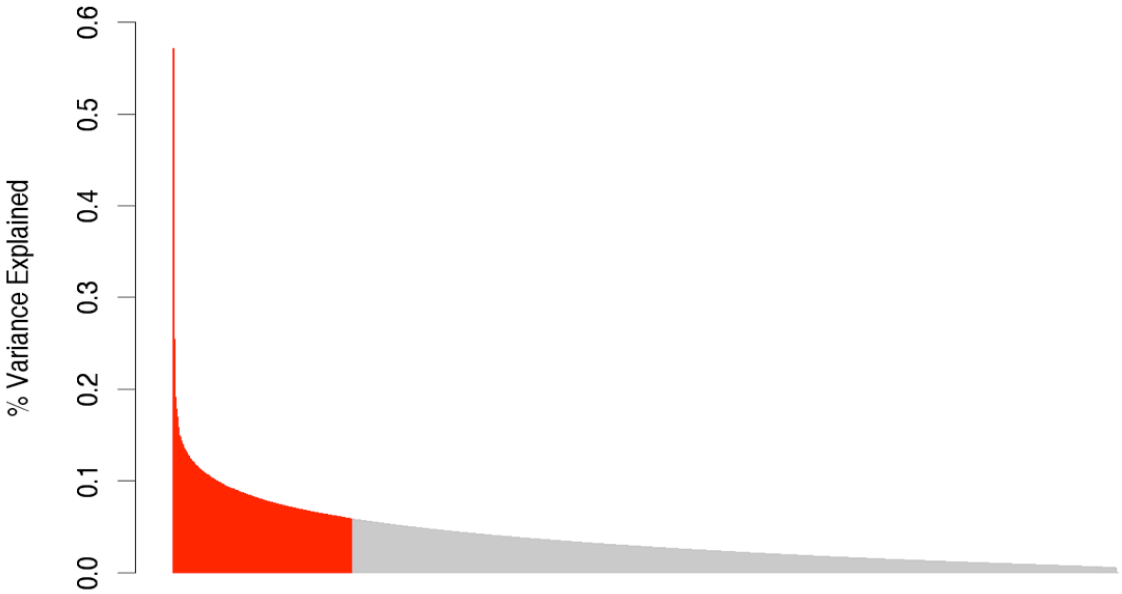


Figure 4-3 Percentage of Variance Explained by Each Principal Component

4.5 Ten-Year Incident Hypertension Results Using 500 Most Highly Associated SNPs

4.5.1 Variable Selection

The risk index procedure was applied to the FHS Offspring cohort data described above. A logistic regression analysis was performed to estimate the association between each of the 48,127 SNPs and the development of hypertension by exam seven. The 500 SNPs with the smallest association p-values were selected for this analysis. Of the 2,283 subjects who did not have hypertension at exam 2, 777, or 34.0%, developed hypertension by exam seven. The data was divided into an independent testing set, consisting of 571 individuals, and an optimization set, consisting of 1712 individuals.

One hundred bootstrap samples of the optimization set were generated, and the risk index procedure was used to generate Clinical and Clinical + Genotype risk index models for each of the bootstrap samples. Each of these models was then used to make a prediction (high risk or low risk for developing hypertension) about each of the 571 individuals in the independent testing set. For both the Clinical risk index model and the Clinical + Genotype risk index model the predictions from each of the 100 bootstrap samples were used as votes, and the prediction most frequently assigned was designated as the consensus prediction.

Table 4-2 gives a summary of the order in which variables were selected for the 100 bootstrap samples of the optimization set used to build the Clinical risk index model for incident hypertension. The relatively small number of clinical variables leads to each of the variables being selected into at least 72 of the 100 untrimmed Clinical risk index models, with height, current smoking status, and marital status selected into all 100 untrimmed Clinical risk index models and weekly alcohol consumption, serum albumin level, serum bilirubin level, and serum creatine level selected into 99 of the 100 untrimmed Clinical risk index models. Fifty out of the 100 trimmed Clinical risk index models contained Marital Status as a variable. Weekly alcohol consumption was included in 47 out of the 100 trimmed Clinical risk index models. Height, systolic blood pressure, and current smoking status, were also frequently included in the set of 100 trimmed Clinical risk index models, appearing in 46, 36, and 31 trimmed Clinical risk index models, respectively.

Table 4-3 gives a summary of the order in which variables were selected for the 100 bootstrap samples of the Optimization Set used to build the Clinical + Genotype risk index model. Because of the much larger number of available genotype variables, no single polymorphism was selected as the first variable in the Clinical + Genotype risk index model by more than 3 of the 100 bootstrap samples. Likewise, no individual polymorphisms was selected as either the second or third variable in the Clinical + Genotype risk index model by more than 4 out of 100 bootstrap samples. Table 4-3 shows a summary of variable selection process for the 15 SNPs that were selected into 20 or more untrimmed Clinical + Genotype risk index models. All of these variables also appear in at least 15 trimmed Clinical + Genotype risk index model.

Table 4-2 Summary of Number of Times Each Variable is Selected into a Specific Model Position for the Clinical Risk Index Models for Incident Hypertension

Variable Name	Variable Position																				Total # of Models	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
Age (yrs)	3	4	4	0	2	3	4	0	2	5	2	3	4	3	6	5	0	4	10	8	72	14
Sex	0	6	5	3	3	3	4	3	6	1	4	7	8	4	9	4	3	3	6	4	86	14
Weight (lbs)	0	1	0	0	1	2	1	3	1	6	5	8	5	4	3	4	10	6	4	12	76	7
Height (in)	4	12	13	10	12	12	9	7	6	2	1	3	3	1	1	1	0	1	0	2	100	46
Systolic Blood Pressure (mm Hg)	27	2	1	0	1	2	0	4	4	2	3	4	2	4	7	2	3	3	5	7	83	36
Diastolic Blood Pressure (mm Hg)	15	4	2	3	1	0	4	3	3	0	2	1	1	8	4	3	7	2	7	5	75	23
Total Cholesterol (mg/dL)	2	0	3	1	0	0	1	1	4	2	5	3	5	4	3	7	11	9	11	6	78	7
High-density Lipoprotein Level (mg/dL)	1	3	2	1	4	4	0	5	4	5	7	6	3	8	5	5	9	8	2	8	90	15
Low-density Lipoprotein Level (mg/dL)	1	2	0	2	3	1	1	4	2	6	2	5	6	6	4	12	8	5	2	13	85	10
Triglycerides (mg/dL)	3	1	1	3	2	1	3	4	3	3	7	2	7	5	10	4	4	5	7	4	79	11
Ever Smoked	0	10	8	7	10	4	12	9	4	6	7	4	6	3	2	3	1	0	0	0	96	29
Currently Smokes	0	11	11	11	4	10	12	8	6	4	1	5	6	3	0	4	1	3	0	0	100	31
Weekly Alcohol Consumption	18	9	6	5	14	9	10	4	6	5	5	1	1	2	1	1	0	1	0	1	99	47
Marital Status	6	16	22	19	12	3	5	2	2	3	3	0	2	0	2	1	0	1	1	0	100	50
Left Ventricular Mass (g)	1	0	2	3	0	2	2	2	5	4	2	4	1	6	5	5	8	4	4	3	63	7
Left Ventricular Ejection Fraction (%)	4	6	6	3	5	3	2	8	6	7	4	7	5	2	2	2	1	3	2	3	81	18
Blood Glucose (mg/dL)	2	0	1	2	3	2	2	4	3	4	4	4	7	4	4	6	8	8	10	7	85	9
Blood Urea Nitrogen (mg/dL)	0	2	1	1	2	6	1	1	6	4	3	3	8	8	7	6	5	6	9	4	83	10
Total Serum Protein Level (mg/dL)	5	5	0	6	5	2	7	3	3	9	5	8	4	7	3	4	5	6	3	3	93	16

Serum Albumin Level (mg/dL)	1	3	1	2	4	12	9	9	10	6	10	8	3	5	5	4	1	3	2	1	99	18
Serum Bilirubin Level (mg/dL)	3	0	2	11	4	7	5	11	5	6	9	6	7	3	5	4	0	4	6	1	99	19
Serum Alkaline Phosphatase Level (mg/dL)	1	1	2	2	0	1	1	2	3	1	4	1	4	5	9	12	10	9	6	5	79	6
Serum Creatine Level (mg/dL)	3	2	7	5	8	11	5	3	6	9	5	7	2	5	3	1	5	6	3	3	99	18

Table 4-3 Summary of Number of Times Selected Genotype Variables are Selected into a Specific Model Position for Clinical + Genotype Risk Index Models for Incident Hypertension

	Variable Position																				Total # of Models	Total # of Times in Trimmed Model
SNP	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**		
rs16995309	1	0	4	0	0	3	0	3	2	3	2	2	1	1	0	0	1	2	3	0	28	23
rs10082778	0	3	0	0	2	1	5	2	0	2	2	2	3	1	3	3	5	4	2	1	41	29
rs241419	0	0	1	1	1	3	2	0	1	3	1	0	0	1	1	1	0	0	1	3	20	18
rs3733920	0	0	1	0	0	1	4	1	0	2	4	3	0	3	1	2	1	1	3	1	28	24
rs6137081	0	0	2	0	2	0	0	2	3	1	3	0	2	0	2	3	4	2	4	0	30	21
rs33965313	0	4	0	1	2	1	3	0	1	0	2	0	1	3	0	3	1	0	0	0	22	16
rs12812222	1	1	1	2	2	1	0	0	1	2	1	1	3	0	0	1	1	0	0	3	21	15
rs6878329	1	1	3	2	3	3	1	6	0	3	2	1	5	1	1	1	2	0	1	2	39	34
rs11693983	2	0	1	1	1	4	1	1	1	1	2	1	1	3	2	1	1	1	0	1	26	21
rs6059153	2	2	1	4	2	6	6	7	3	3	4	6	2	2	2	0	1	0	3	1	57	49
rs17047347	0	0	3	0	1	3	3	1	3	2	1	1	2	2	2	1	3	3	3	1	35	25
rs6004901	2	3	2	1	1	1	2	0	0	1	0	1	2	2	1	0	1	0	0	0	20	18
rs10492357	1	0	0	5	2	0	1	1	1	1	3	0	1	1	2	1	3	6	1	0	30	23
rs1555498	2	2	1	1	0	1	0	3	0	1	3	1	2	4	3	1	3	1	1	1	31	23
rs17128116	1	1	3	5	4	1	3	3	4	2	4	0	3	3	3	2	1	5	3	2	53	39

4.5.2 Models

Table 4-4 shows the trimmed Clinical risk index models for a selection of five random bootstrap samples. Figure 4-4 shows the distribution of risk index values in the optimization set for the Clinical risk index model from one randomly selected bootstrap sample (Bootstrap Sample #9), and Figure 4-5 shows the distribution of risk index values in the independent testing set for the Clinical risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #9). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as "high risk of developing hypertension" and those with a risk index value lower are predicted as "low risk of developing hypertension". Table 4-5 shows the risk index values for 25 randomly chosen individuals from the Independent Testing Set from these five Clinical risk index models along with that risk index model's prediction about each individual, where 0 indicates low risk of developing hypertension and 1 indicates high risk of developing hypertension.

Table 4-4 Five Randomly Selected Trimmed Clinical Risk Index Models for Incident Hypertension

Bootstrap	Trimmed Clinical Risk Index Model
9	0.0697*Systolic Blood Pressure + 0.1913*Marital Status
51	-0.0212*Height + 0.0019*Triglycerides + 0.1182*Current Smoking Status + 0.0207*Weekly Alcohol Consumption + 0.2024*Marital Status + 0.0039*Left Ventricular Mass + 0.0326*Blood Urea Nitrogen + 0.0413*Total Serum Protein - 0.036*Albumin - 0.0076*Bilirubin
59	-0.0017*Height + 0.0747*Current Smoking Status + 0.0109*Weekly Alcohol Consumption + 0.0962*Marital Status - 0.0051*Albumin
72	0.1116*Diastolic Blood Pressure + 0.01*Total Cholesterol + 0.0924*Ever Smoked + 0.0119*Weekly Alcohol Consumption
84	0.0553*Age - 0.1312*Sex + 0.0085*Weight - 0.0152*HDL + 0.3433*Ever Smoked + 0.1008*Current Smoking Status + 0.0066*Weekly Alcohol Consumption + 0.1688*Marital Status + 0.0473*Blood Urea Nitrogen - 0.0013*Bilirubin + 0.0293*Creatine

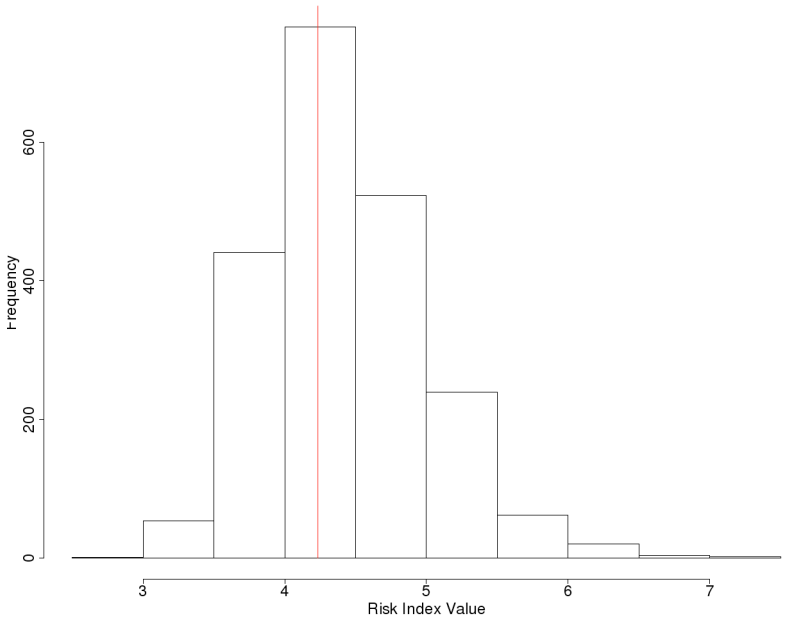


Figure 4-4 Clinical Risk Index Model Risk Index Values Distribution in the Optimization Set, Bootstrap Sample #9

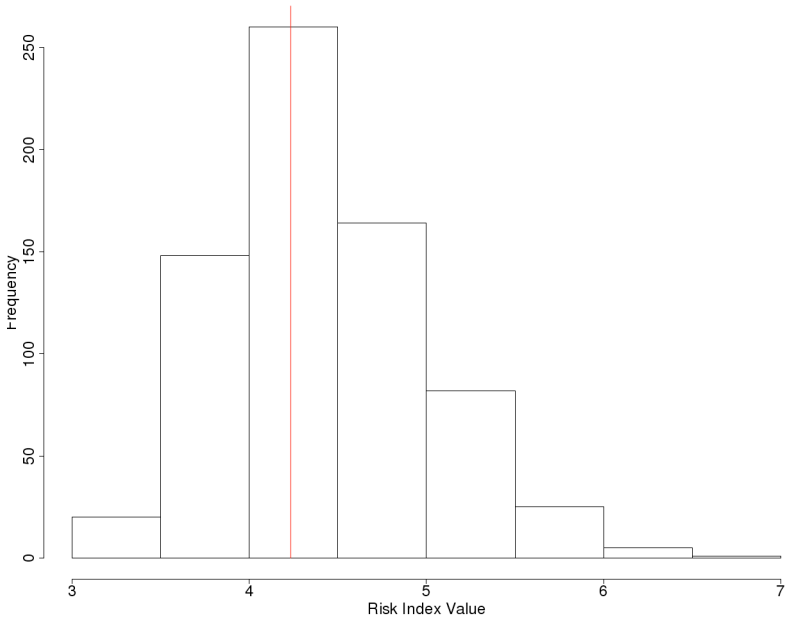


Figure 4-5 Clinical Risk Index Model Risk Index Values Distribution in the Independent Testing Set, Bootstrap Sample #9

Table 4-5 Risk Index Values for 25 Individuals in the Optimization Set for the Clinical Risk Index Models for Incident Hypertension from Five Randomly Selected Bootstrap Samples

Individual	Outcome	Bootstrap Sample #9 Cutoff Value = 4.224		Bootstrap Sample #51 Cutoff Value = 0.168		Bootstrap Sample #59 Cutoff Value = -0.015		Bootstrap Sample #72 Cutoff Value = 2.796		Bootstrap Sample #84 Cutoff Value = 0.394	
		Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	1	5.375	1	0.187	1	0.008	1	2.909	1	0.482	1
2	0	4.025	0	0.127	0	-0.015	0	2.291	0	0.308	0
3	1	5.140	1	0.201	1	-0.006	1	2.982	1	0.453	1
4	0	4.513	1	0.113	0	-0.008	1	2.751	0	0.417	1
5	0	3.511	0	0.050	0	-0.052	0	2.274	0	0.202	0
6	0	3.441	0	0.207	1	0.002	1	2.788	0	0.341	0
7	1	4.425	1	0.225	1	0.006	1	2.970	1	0.407	1
8	1	6.116	1	0.167	0	-0.018	0	3.688	1	0.420	1
9	1	4.513	1	0.182	1	-0.032	0	2.885	1	0.471	1
10	0	4.164	0	0.118	0	-0.031	0	2.669	0	0.365	0
11	1	3.955	0	0.107	0	-0.001	1	2.454	0	0.338	0
12	0	4.381	1	0.217	1	0.068	1	2.688	0	0.415	1
13	1	3.816	0	0.189	1	0.001	1	2.444	0	0.310	0
14	0	4.234	0	0.210	1	-0.022	0	2.575	0	0.410	1
15	1	3.441	0	0.178	1	-0.024	0	2.142	0	0.335	0
16	0	4.792	1	0.112	0	-0.025	0	2.529	0	0.436	1
17	1	4.774	1	0.227	1	0.023	1	2.959	1	0.499	1
18	0	4.025	0	-0.070	0	0.020	1	2.635	0	0.518	1
19	0	3.720	0	0.137	0	-0.025	0	2.458	0	0.325	0
20	1	4.582	1	0.091	0	-0.026	0	2.647	0	0.279	0
21	0	3.816	0	0.094	0	-0.013	1	2.560	0	0.503	1
22	1	5.698	1	0.213	1	-0.011	1	2.828	1	0.439	1
23	0	3.607	0	0.155	0	-0.030	0	2.161	0	0.290	0
24	1	4.286	1	0.142	0	0.114	1	2.697	0	0.588	1
25	1	4.373	1	0.168	0	-0.020	0	2.876	1	0.342	0

Table 4-6 shows the trimmed Clinical + Genotype risk index models for the same set of five bootstraps shown in Table 4-4, and Table 4-7 shows the risk index values and predictions of the same 25 randomly chosen individuals for the 5 bootstrap samples shown in Table 4-5. Figure 4-6 shows the distribution of risk index values in the optimization set for the Clinical + Genotype risk index model from one randomly selected bootstrap sample (Bootstrap Sample #9). Figure 4-7 shows the distribution of risk index values in the independent testing set for the Clinical + Genotype risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #9). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as “high risk of developing hypertension” and those with a risk index value lower are predicted as “low risk of developing hypertension”.

Table 4-6 Five Randomly Selected Clinical + Genotype Risk Index Models for Incident Hypertension

Bootstrap	Trimmed Clinical + Genetics Risk Index Model
9	0.0697*Systolic Blood Pressure + 0.1913*Marital Status + SNP_M-583177(C_G=0.3172, G_G=0.3868) + SNP_M-310902(C_G=0.4776, G_G=0.7179)
51	-0.0212*Height + 0.0019*Triglycerides + 0.1182*Current Smoking Status + 0.0207*Weekly Alcohol Consumption + 0.2024*Marital Status + 0.0039*Left Ventricular Mass + 0.0326*Blood Urea Nitrogen + 0.0413*Total Serum Protein - 0.036*Albumin - 0.0076*Bilirubin + SNP_M-600701(C_T=1.3306, T_T=1.3627) + SNP_M-594285(A_C=0.1237, C_C=15.1625) + SNP_M-319308(C_G=-0.1289, G_G=-13.5691) + SNP_M-180302(C_T=-0.3018, T_T=-0.2546) + SNP_M-592316(C_T=-1.0852, NA) + SNP_M-317848(T_T=1.3633, NA) + SNP_M-324013(A_G=-0.9846, G_G=-1.0584) + SNP_M-179284(C_T=0.0504, T_T=-0.0504) + SNP_M-319058(A_G=-0.3462, G_G=16.2787)
59	-0.0017*Height + 0.0747*Current Smoking Status + 0.0109*Weekly Alcohol Consumption + 0.0962*Marital Status - 0.0051*Albumin + SNP_M-607627(C_T=0.2081, T_T=1.7121) + SNP_M-599757(C_T=-0.6624, T_T=-0.9666) + SNP_M-180069(C_T=0.1609, T_T=0.3887) + SNP_M-306806(C_T=-1.5941, NA) + SNP_M-589310(C_T=0.6218, T_T=0.8646) + SNP_M-185374(A_T=1.0779, NA) + SNP_M-591588(G_T=0.2153, T_T=1.0864) + SNP_M-580964(A_G=0.665, G_G=0.8924) + SNP_M-597220(G_T=0.2554, T_T=0.8698) + SNP_M-603278(C_T=0.0995, T_T=0.9101) + SNP_M-599155(G_T=-1.1167, T_T=-0.4468) + SNP_M-590943(A_C=0.5146, C_C=0.6543) + SNP_M-611577(A_T=-15.0476, T_T=-16.2242) + SNP_M-612475(C_T=-0.424, T_T=-2.0049) + SNP_M-582008(C_T=0.0845, T_T=2.1746) + SNP_M-317848(T_T=2.1421, NA) + SNP_M-597857(A_G=-0.5716, G_G=-0.8797) + SNP_M-594033(C_G=0.2302, G_G=0.4309) + SNP_M-603817(A_G=0.1525, G_G=0.279) + SNP_M-587757(C_T=-17.8997, T_T=-18.0137)
72	0.1116*Diastolic Blood Pressure + 0.01*Total Cholesterol + 0.0924*Ever Smoked + 0.0119*Weekly Alcohol Consumption + SNP_M-580646(C_T=0.2179 ,T_T=0.4832) + SNP_M-177199(C_G=0.3118 ,G_G=0.4673) + SNP_M-597655(A_G=-0.1203 ,G_G=-0.0561) + SNP_M-598920(A_G=-0.2478 ,G_G=-0.6533) + SNP_M-592316(C_T=-0.9288 ,NA) + SNP_M-600051(A_G=-0.2553 ,G_G=-1.0712) + SNP_M-611577(A_T=0.2607 ,T_T=-0.2607) + SNP_M-610223(A_T=0.0567 ,T_T=0.5653) + SNP_M-598811(A_G=0.6177 ,G_G=0.6401) + SNP_M-602035(A_G=-0.129 ,G_G=-0.1316) + SNP_M-327317(C_T=-0.4836 ,T_T=-0.3172) + SNP_M-310902(C_G=14.3295 ,G_G=14.3208) + SNP_M-588824(A_G=0.2089 ,G_G=0.2652) + SNP_M-182866(C_T=-0.1895 ,T_T=0.1413) + SNP_M-594503(A_G=-0.046 ,G_G=0.0596)
84	0.0553*Age - 0.1312*Sex + 0.0085*Weight - 0.0152*HDL + 0.3433*Ever Smoked + 0.1008*Current Smoking Status + 0.0066*Weekly Alcohol Consumption + 0.1688*Marital Status + 0.0473*Blood Urea Nitrogen - 0.0013*Bilirubin + 0.0293*Creatine + SNP_M-580646(C_T=0.5211, T_T=0.3669) + SNP_M-609794(A_T=-0.1066, T_T=-0.617) + SNP_M-589026(A_T=-0.0312, T_T=0.0312) + SNP_M-608322(C_G=0.1269, G_G=0.4254) + SNP_M-580415(A_C=0.0215, C_C=-0.0642) + SNP_M-599120(G_T=0.1178, T_T=0.1993) + SNP_M-185297(C_G=-9.385, G_G=-10.7958) + SNP_M-603138(C_T=-0.1022, T_T=-0.258) + SNP_M-602581(A_T=17.0589, T_T=16.9035) + SNP_M-181501(C_G=-0.1468, G_G=0.9559) + SNP_M-323893(A_G=-0.2881, G_G=-0.4792) + SNP_M-322333(C_T=-0.9651, T_T=-1.2191) + SNP_M-609320(A_G=-0.0513, G_G=0.7844) + SNP_M-319195(C_T=-0.4342, NA) + SNP_M-324013(A_G=0.0311, G_G=-0.0578) + SNP_M-179284(C_T=18.3499, T_T=18.3908) + SNP_M-185106(C_T=-0.4374, T_T=0.4374)

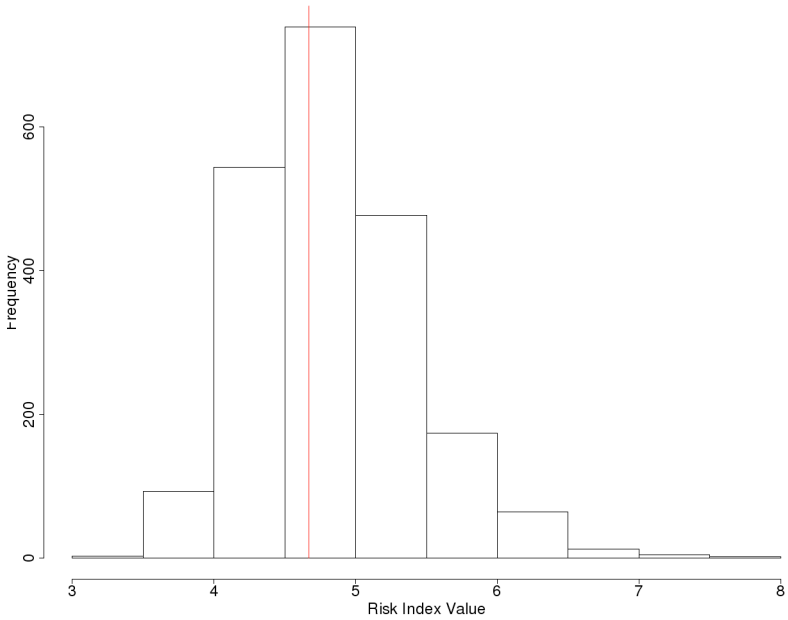


Figure 4-6 Clinical + Genotype Risk Index Model Risk Index Values Distribution in the Optimization Set, Bootstrap Sample #9

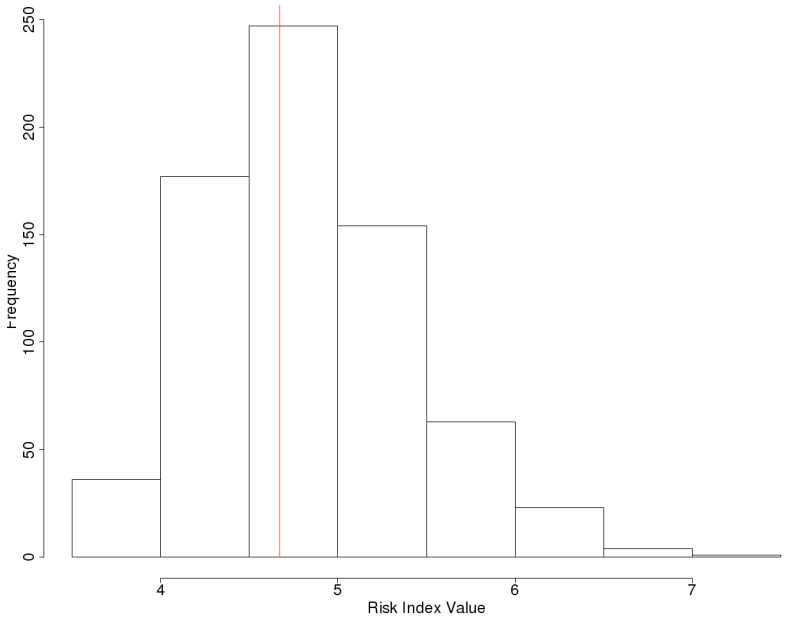


Figure 4-7 Clinical + Genotype Risk Index Model Risk Index Values Distribution in the Independent Testing Set, Bootstrap Sample #9

Table 4-7 Risk Index Values for 25 Individuals in the Optimization Set for the Clinical + Genotype Risk Index Models for Incident Hypertension from Five Randomly Selected Bootstrap Samples

		Bootstrap Sample #9 Cutoff Value = 4.671		Bootstrap Sample #51 Cutoff Value = 0.344		Bootstrap Sample #59 Cutoff Value = -1.538		Bootstrap Sample #72 Cutoff Value = 3.783		Bootstrap Sample #84 Cutoff Value = 1.820	
		Risk Index		Risk Index		Risk Index		Risk Index		Risk Index	
Individual	Outcome	Value	Prediction	Value	Prediction	Value	Prediction	Value	Prediction	Value	Prediction
1	1	5.734	1	0.381	1	-1.530	1	3.947	1	1.897	1
2	0	4.457	0	0.438	1	-1.453	1	3.277	0	1.717	0
3	1	5.499	1	0.344	0	-1.506	1	4.024	1	1.847	1
4	0	5.030	1	0.306	0	-1.617	0	3.794	1	1.858	1
5	0	3.870	0	0.229	0	-1.586	0	3.234	0	1.660	0
6	0	3.959	0	0.387	1	-1.517	1	3.755	0	1.760	0
7	1	4.943	1	-1.103	0	-1.548	0	3.944	1	1.838	1
8	1	6.475	1	0.347	1	-1.505	1	4.673	1	1.850	1
9	1	5.030	1	2.046	1	-1.549	0	3.866	1	1.901	1
10	0	4.523	0	0.264	0	-1.487	1	3.670	0	1.801	0
11	1	4.473	0	0.287	0	-1.550	0	3.450	0	1.780	0
12	0	4.740	1	0.396	1	-1.391	1	3.616	0	1.824	1
13	1	4.213	0	0.349	1	-1.471	1	3.429	0	1.764	0
14	0	4.751	1	0.390	1	-1.674	0	3.539	0	1.840	1
15	1	3.959	0	0.358	1	-1.586	0	3.119	0	1.771	0
16	0	5.150	1	0.259	0	-1.495	1	3.607	0	1.849	1
17	1	5.291	1	0.406	1	-1.496	1	3.907	1	1.926	1
18	0	4.384	0	0.110	0	-1.437	1	3.678	0	1.935	1
19	0	4.079	0	0.317	0	-1.546	0	3.411	0	1.732	0
20	1	4.941	1	0.270	0	-1.414	1	3.635	0	1.727	0
21	0	4.175	0	0.273	0	-1.476	1	3.527	0	1.939	1
22	1	6.215	1	0.360	1	-1.556	0	3.834	1	1.874	1
23	0	4.124	0	0.297	0	-1.542	0	3.109	0	1.698	0
24	1	4.645	0	0.330	0	-1.441	1	3.682	0	2.012	1
25	1	4.732	1	0.348	1	-1.565	0	3.872	1	1.736	0

4.5.3 Predictive Performance

Once predictions were made for each individual in the independent testing set the sensitivity, specificity, misclassification, and positive predictive value were calculated for the Clinical risk index model and the Clinical + Genotype risk index model. One thousand bootstrap samples of the independent testing set were generated, and the 100 trimmed Clinical risk index models and the 100 trimmed Clinical + Genotype risk index models were applied to each individual in each of the 1000 bootstrap samples. The sensitivity, specificity, misclassification, and positive predictive value of the risk index models were calculated for each of the 1000 bootstrap samples, and 95% confidence intervals for each of these measurements were estimated from this data. The estimates and confidence intervals for sensitivity, specificity, misclassification, and positive predictive value are given in Table 4-2. Lastly, using the individual predictions from each of the 100 trimmed Clinical risk index models and 100 trimmed Clinical + Genotype risk index models for the individuals in the independent testing set, receiver operating characteristic (ROC) curves were generated, and the area under the ROC curve (AUC) was estimated for both the Clinical and Clinical + Genotype risk index models (Figure 4-8, Figure 4-9). For the Clinical risk index model the AUC for the ROC curve was 0.567, and for the Clinical + Genotype risk index model the AUC for the ROC curve was 0.475.

Table 4-8 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for Ten-year Incident Hypertension

Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
Clinical	0.667 (0.594 - 0.736)	0.486 (0.444 - 0.53)	0.468 (0.433 - 0.504)	0.308 (0.26 - 0.352)
Clinical + Genotype	0.539 (0.464 - 0.609)	0.457 (0.413 - 0.497)	0.522 (0.487 - 0.559)	0.254 (0.21 - 0.298)

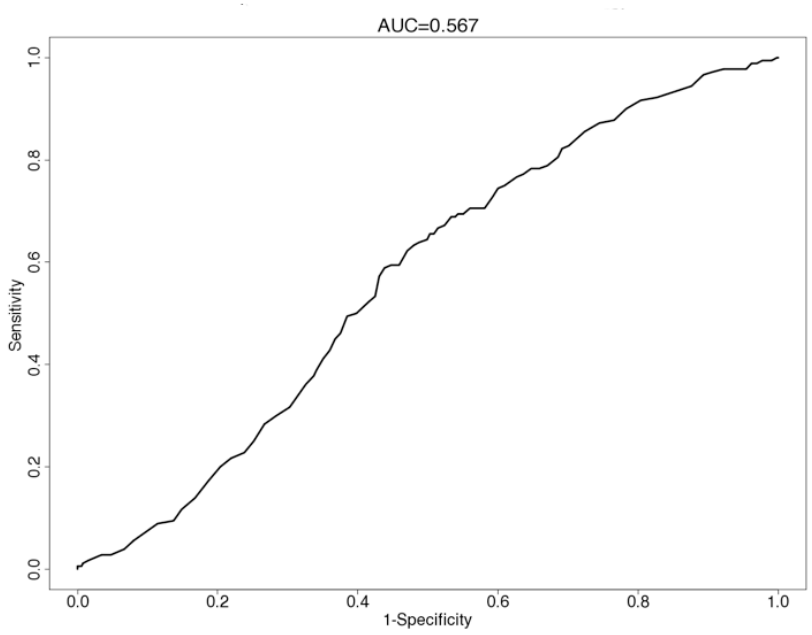


Figure 4-8 ROC Curve and AUC for the Incident Hypertension Clinical Risk Index Model

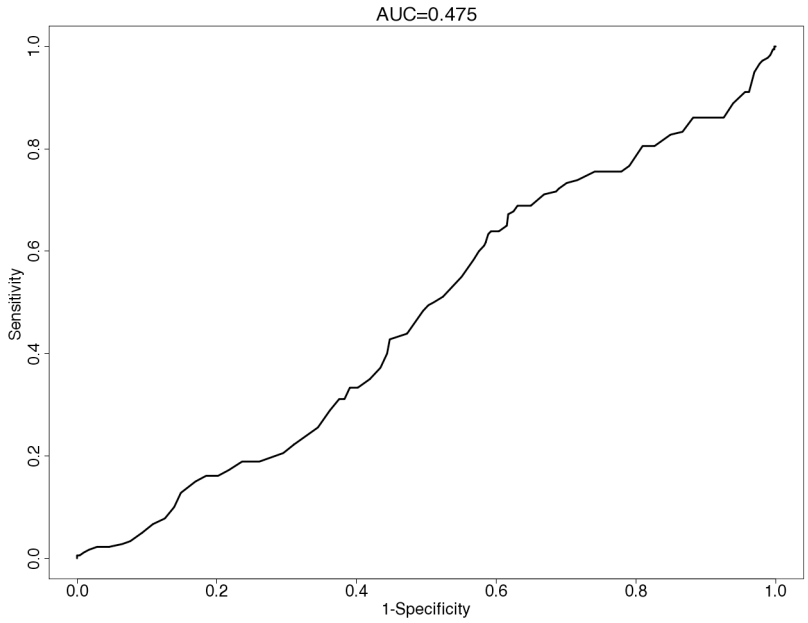


Figure 4-9 ROC Curve and AUC for the Incident Hypertension Clinical + Genotype Risk Index Model

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. The proportion of models that predict that an individual is at high risk of developing hypertension, then, represents the predicted probability of an individual developing hypertension. Using the binomial distribution a 95% confidence interval can be constructed for this estimated probability with the Wilson score interval (Wilson, 1927). This interval, given by

$$95\%\ \text{CI} = \frac{\hat{p} + \frac{(1.96)^2}{2n} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{(1.96)^2}{4n^2}}}{1 + \frac{(1.96)^2}{n}}$$

, can then be used by physicians to gauge the prediction. As an example, one individual in the independent testing set had a predicted probability of developing hypertension from the Clinical risk index model of

0.82. The lower bound for this individual's 95% confidence interval is

$$\frac{0.82+\frac{(1.96)^2}{2*100}-1.96\sqrt{\frac{0.82(1-0.82)}{100}+\frac{(1.96)^2}{4*(100)^2}}}{1+\frac{(1.96)^2}{100}}$$ or 0.733 and the upper bound is

$$\frac{0.82+\frac{(1.96)^2}{2*100}+1.96\sqrt{\frac{0.82(1-0.82)}{100}+\frac{(1.96)^2}{4*(100)^2}}}{1+\frac{(1.96)^2}{100}}$$ or 0.883.

Figure 4-10 shows the distribution of the predicted probability of developing hypertension for the Clinical risk index model in the independent testing set, and Figure 4-11 shows the distribution of the predicted probability of developing hypertension for the Clinical + Genotype risk index model in the independent testing set. In both Figures, a density line is shown on the graph to indicate the density of a normal distribution with the mean and standard deviation matching that of the confidence score distribution.

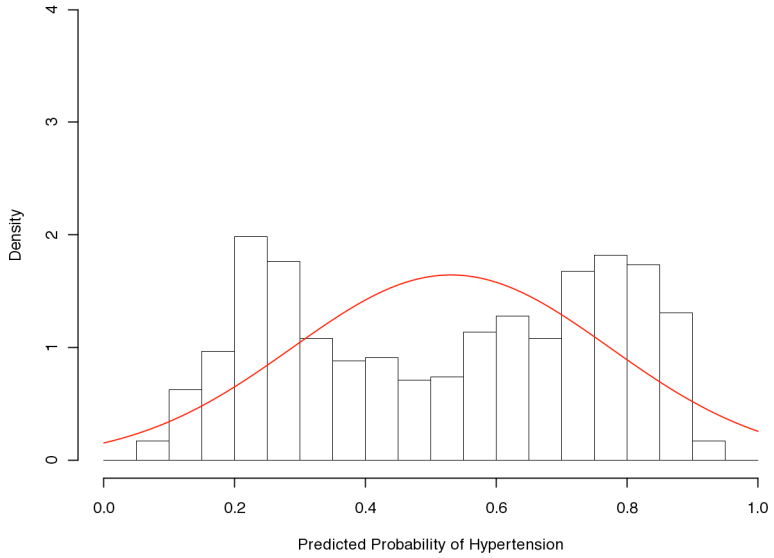


Figure 4-10 Histogram of the Predicted Probability of Developing Hypertension for the Clinical Risk Index Model

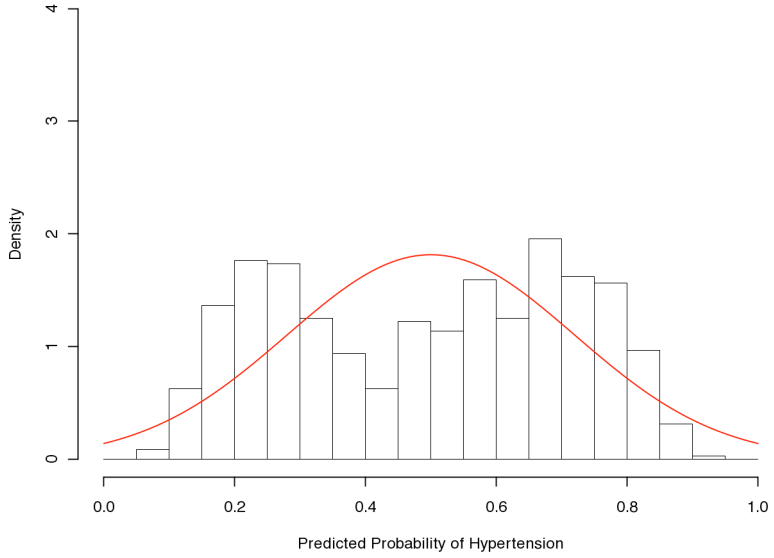


Figure 4-11 Histogram of the Predicted Probability of Developing Hypertension for the Clinical + Genotype Risk Index Model

4.5.4 Random Forests Comparison

A random forest was generated using the optimization set created by the risk index procedure. The forest had 500 individual trees, and a tuning procedure was used to find the number of variables *k* considered at each split that provided the lowest out-of-bag error estimate. Beginning with $k = \sqrt{v}$, where *v* is the total number of predictor variables, the forest was grown and out-of-bag error was measured. Then, the number of variables considered at each split was progressively increased by a factor of two (i.e., $k = 2 * \sqrt{v}$, $k = 4 * \sqrt{v}$, etc.) until the out-of-bag error decreased by less than 5% from the out-of-bag error for the previous value of *k*. Next, returning to $k = \sqrt{v}$, the number of variables considered at each split was progressively decreased by a factor of two (i.e., $k = \frac{1}{2} * \sqrt{v}$, $k = \frac{1}{4} * \sqrt{v}$, etc.) until the out-of-bag error decreased by less than 5% from the out-of-bag error for the previous value of *k*. The optimized *k* chosen was 78, which gave an out-of-bag error estimate of 24.9%.

When working with a dataset that has two possible classes, the standard procedure for a random forest is to assign a prediction to an individual based on a simple majority of votes, when the prevalence of the outcome is less than 50% changing the proportion of votes needed to classify an individual can significantly impact the estimates of performance. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set on a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of

5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the individual to be "high risk". Table 4-9 shows the results of this investigation.

Predictions were then made about each individual in the independent testing set created by the risk index procedure, and the sensitivity, specificity, misclassification, and positive predictive value of the predictions was assessed. One thousand bootstrap samples of the independent testing set were generated, and predictions were made about each individual in each of the bootstrap samples. This data was used to create 95% confidence intervals for the sensitivity, specificity, misclassification, and positive predictive value estimates. Table 4-10 shows the sensitivity, specificity, misclassification, and positive predictive value estimates for the random forest as well as the 95% confidence interval for each estimate. Lastly, using the class votes for the individuals in the independent testing set, an ROC curve was created, and the AUC for the ROC curve was estimated (Figure 4-12).

Table 4-9 Performance Estimates of the Random Forest

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	1.000	0.000	0.659	0.341
0.1	1.000	0.038	0.634	0.349
0.15	0.978	0.166	0.557	0.377
0.2	0.956	0.340	0.450	0.428
0.25	0.898	0.498	0.366	0.480
0.3	0.839	0.615	0.308	0.530
0.35	0.737	0.728	0.269	0.584
0.4	0.657	0.815	0.239	0.647
0.45	0.526	0.887	0.236	0.706
0.5	0.380	0.940	0.251	0.765
0.55	0.219	0.985	0.276	0.882
0.6	0.051	0.992	0.328	0.778
0.65	0.007	1.000	0.338	1.000
0.7	0.000	1.000	0.341	-
0.75	0.000	1.000	0.341	-
0.8	0.000	1.000	0.341	-
0.85	0.000	1.000	0.341	-
0.9	0.000	1.000	0.341	-
0.95	0.000	1.000	0.341	-

Table 4-10 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for the Random Forest Model

Proportion of Votes for "High Risk" Class	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
0.25	0.898 (0.841 - 0.944)	0.498 (0.435 - 0.556)	0.366 (0.318 - 0.413)	0.480 (0.423 - 0.543)
0.3	0.839 (0.775 - 0.901)	0.615 (0.552 - 0.674)	0.308 (0.264 - 0.358)	0.530 (0.462 - 0.596)
0.35	0.737 (0.664 - 0.811)	0.728 (0.673 - 0.781)	0.269 (0.226 - 0.313)	0.584 (0.510 - 0.660)
0.4	0.657 (0.583 - 0.736)	0.815 (0.769 - 0.859)	0.239 (0.199 - 0.279)	0.647 (0.567 - 0.719)
0.45	0.526 (0.448 - 0.607)	0.887 (0.845 - 0.923)	0.236 (0.197 - 0.279)	0.706 (0.615 - 0.793)
0.5	0.380 (0.297 - 0.466)	0.940 (0.908 - 0.967)	0.251 (0.206 - 0.296)	0.765 (0.6579 - 0.868)

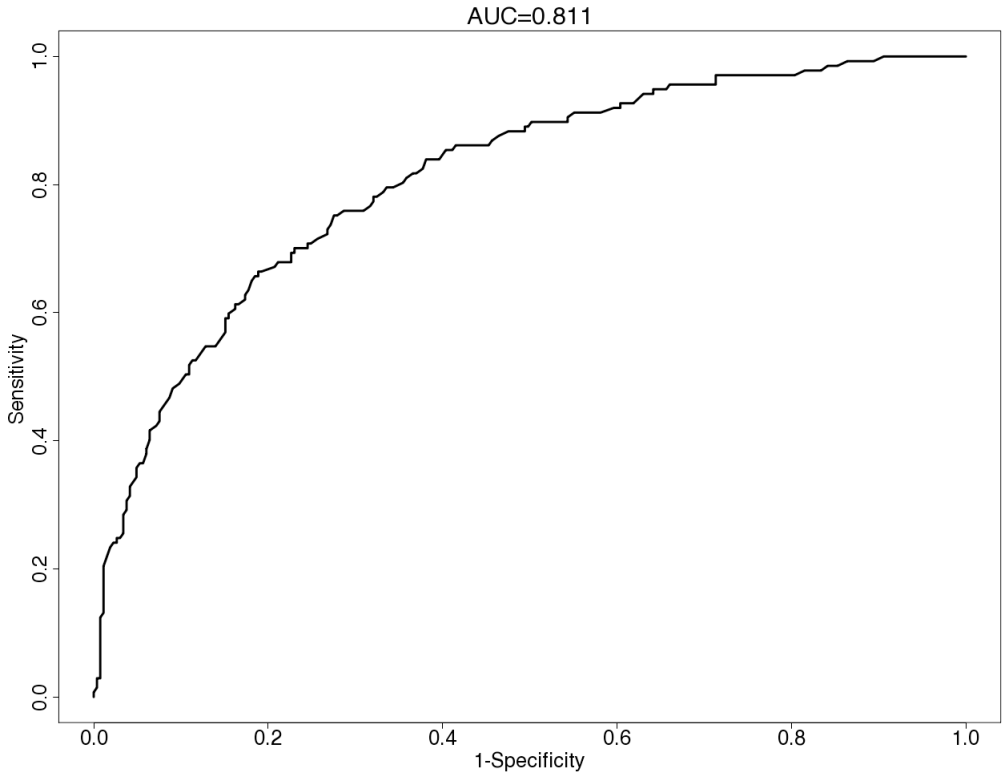


Figure 4-12 ROC Curve and AUC for the Incident Hypertension Random Forest Model

4.5.5 Conclusion

The performance of the risk index procedure in predicting the development of hypertension with a ten-year period is fairly poor, despite a large number of Clinical variables and 500 polymorphisms selected because of their association in logistic regression models with the development of hypertension. The random forest, however, performed much better than the risk index model in overall classification accuracy, prediction specificity, and positive predictive value. However, the Clinical risk index model produced a sensitivity that was significantly higher than the random forest (the 95% confidence interval of the risk index model does not overlap with the 95%

confidence interval for the sensitivity estimate of the random forest) if a simple majority of trees in the random forest is used to assign the predictions. Likewise, the sensitivity estimate for the Clinical + Genotype risk index model has a 95% confidence interval that is higher than the random forest, and it only just overlaps with the 95% confidence interval of the sensitivity estimate of the random forest. However, by lowering the proportion of trees necessary to assign an individual a "high risk" classification, the random forest can yield performance that is noticeably better than the Clinical or Clinical + Genotype risk index models. As Table 4-10 shows, by setting the proportion of trees voting for a classification of "high risk" to 0.35 a nearly balanced sensitivity and specificity, with a misclassification much lower than that of the risk index models. The relatively linear AUC curves for the Clinical and Clinical + Genotype risk index models indicates that varying the proportion of votes need to make a prediction of "high risk" will not yield a marked improvement in sensitivity without a corresponding drop in specificity.

4.6 Ten-Year Incident Hypertension Results Using Top 500 Principal Components

4.6.1 Variable Selection

Using the procedure described in Section 4.5.1, the risk index procedure was repeated, replacing the 500 SNPs most highly associated with 10-year incident hypertension with the top 500 principal components from a principal components analysis of the full set of available SNPs.

Table 4-11 gives a summary of the order in which variables were selected for the 100 bootstrap samples of the optimization set used to build the Clinical risk index model. Fifty-three out of the 100 trimmed Clinical risk index models contained height as a variable, and 51 contained marital status. Diastolic blood pressure was included in 43 out of the 100 trimmed Clinical risk index models. Weekly alcohol consumption, current smoking status, and age were also frequently included in the set of 100 trimmed Clinical risk index models, appearing in 41, 37, and 33 trimmed Clinical risk index models, respectively.

Table 4-12 gives a summary of the order in which variables were selected for the 100 bootstrap samples of the Optimization Set used to build the Clinical + Genotype risk index model. Table 4-12 shows a summary of variable selection process for the 19 PCs that were selected into 8 or more untrimmed Clinical + Genotype risk index models. All of these variables also appear in at least 8 trimmed Clinical + Genotype risk index model.

Table 4-11 Summary of Number of Times Each Variable is Selected into a Specific Model Position for the Clinical Risk Index Model for Incident Hypertension

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
Age (yrs)	4	13	6	0	5	1	2	3	4	4	3	2	3	3	4	3	6	6	1	7	80	33
Sex	0	8	2	7	3	3	7	4	2	7	5	7	3	6	4	6	2	5	1	5	87	22
Weight (lbs)	0	1	0	2	1	0	1	3	1	5	2	2	0	4	4	7	5	8	8	4	58	7
Height (in)	6	8	10	8	10	11	20	5	9	2	3	1	1	0	1	0	2	2	1	0	100	53
Systolic Blood Pressure (mm Hg)	17	1	4	3	0	3	3	0	1	1	7	3	5	4	6	6	4	6	4	5	83	29
Diastolic Blood Pressure (mm Hg)	26	4	4	1	3	3	1	2	1	1	1	1	1	1	2	2	3	1	3	6	67	43
Total Cholesterol (mg/dL)	1	3	0	1	2	2	1	1	3	4	4	2	7	8	5	4	5	7	8	10	78	14
High-density Lipoprotein Level (mg/dL)	4	3	1	3	5	4	1	2	4	4	2	5	5	5	5	3	9	5	9	7	86	21
Low-density Lipoprotein Level (mg/dL)	0	4	5	5	2	2	3	3	3	6	4	3	3	6	4	3	13	5	7	3	84	19
Triglycerides (mg/dL)	2	2	2	3	2	6	2	5	1	1	1	7	5	2	7	9	2	2	9	9	79	19
Ever Smoked	3	4	5	14	4	0	5	5	8	6	4	7	9	4	6	6	3	3	0	2	98	31
Currently Smokes	0	11	16	14	6	7	7	7	2	6	3	4	5	4	3	2	1	0	0	1	99	37
Weekly Alcohol Consumption	13	3	6	3	11	3	7	7	6	5	7	7	3	3	7	2	3	1	0	1	98	41
Marital Status	9	18	10	11	6	12	5	4	5	5	4	1	0	3	2	0	0	2	2	1	100	51
Left Ventricular Mass (g)	3	1	1	1	2	3	1	3	6	2	3	8	4	5	3	5	6	7	8	9	81	14
Left Ventricular Ejection Fraction (%)	0	6	9	4	12	5	3	7	6	3	1	4	5	5	3	3	4	4	3	1	88	28
Blood Glucose (mg/dL)	0	3	0	3	1	4	1	6	3	3	8	4	7	2	4	4	14	12	7	1	87	13
Blood Urea Nitrogen (mg/dL)	2	1	1	0	2	1	2	3	3	3	5	6	8	3	6	11	3	8	10	10	88	12

Total Serum Protein Level (mg/dL)	4	2	1	6	3	3	3	7	7	6	3	6	5	7	7	7	0	4	7	5	93	18
Serum Albumin Level (mg/dL)	1	0	8	1	6	8	6	4	6	8	9	7	3	9	4	7	3	2	2	2	96	27
Serum Bilirubin Level (mg/dL)	2	4	0	5	9	7	10	6	11	14	5	4	4	6	4	1	4	1	1	1	99	30
Serum Alkaline Phosphatase Level (mg/dL)	0	0	2	2	3	3	2	1	5	1	7	3	4	3	3	8	6	7	6	6	72	13
Serum Creatine Level (mg/dL)	3	0	7	3	2	9	7	12	3	3	9	6	10	7	6	1	2	2	3	4	99	23

Table 4-12 Summary of Number of Times Selected Principal Component Variables are Selected into a Specific Model Position for the Clinical + Genotype Risk Index Model for Incident Hypertension

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
PC23	0	0	0	1	1	0	1	1	0	0	1	2	1	0	0	0	0	1	3	1	13	10
PC53	1	1	0	0	2	1	0	2	1	0	0	0	0	0	0	2	0	0	0	0	10	10
PC88	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	1	2	0	11	10
PC57	1	1	1	0	0	1	0	1	1	0	0	0	0	1	1	1	0	0	1	1	11	9
PC254	0	0	2	0	0	1	0	0	0	1	2	0	0	3	0	0	0	0	0	0	9	9
PC348	1	0	0	0	0	1	0	0	1	0	0	1	1	0	0	2	1	1	0	0	9	9
PC500	1	0	0	0	1	1	0	0	1	1	3	0	0	0	0	0	0	1	0	1	10	9
PC13	1	0	1	0	1	1	1	1	0	1	0	0	0	0	1	0	1	0	0	0	9	8
PC14	0	0	0	0	1	0	2	0	0	1	0	0	0	2	0	1	0	1	2	1	11	8
PC29	0	0	0	1	0	0	0	0	3	0	0	1	0	1	1	0	1	0	0	0	8	8
PC36	1	0	0	0	1	1	0	0	0	1	0	0	0	1	0	2	0	0	1	0	8	8
PC125	1	1	0	0	0	2	0	0	0	0	0	0	1	0	1	1	1	1	0	0	9	8
PC225	2	0	0	0	0	0	0	1	0	0	1	0	0	1	1	0	1	0	0	2	9	8
PC227	0	1	2	0	0	1	1	0	0	0	1	0	0	0	0	0	1	1	0	0	8	8
PC347	2	0	0	0	1	2	0	0	0	1	0	0	0	0	1	0	1	0	0	0	8	8
PC357	1	1	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	8	8
PC379	2	0	0	0	1	0	1	1	1	0	0	1	0	0	1	0	0	0	1	0	9	8

4.6.2 Models

Table 4-13 shows the trimmed Clinical risk index models for a selection of five random bootstrap samples. Figure 4-13 shows the distribution of risk index values in the optimization set for the Clinical risk index model from one randomly selected bootstrap sample (Bootstrap Sample #2), and Figure 4-14 shows the distribution of risk index values in the independent testing set for the Clinical risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #2). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as "high risk of developing hypertension" and those with a risk index value lower are predicted as "low risk of developing hypertension". Table 4-14 shows the risk index values for 25 randomly chosen individuals from the Independent Testing Set from these five Clinical risk index models along with that risk index model's prediction about each individual, where 0 indicates low risk of developing hypertension and 1 indicates high risk of developing hypertension.

Table 4-13 Five Randomly Selected Trimmed Clinical Risk Index Models for Incident Hypertension

Bootstrap Sample	Model
2	- 0.2706*Sex - 0.016*Height + 0.0649*Marital Status + 0.0133*lvef + 0.015*Blood Glucose + 0.0569*Blood Urea Nitrogen
16	0.0577*Age - 0.0407*Height + 0.0702*Systolic Blood Pressure - 0.0225*HDL + 0.0067*LDL + 0.0018*Triglycerides - 0.0038*Weekly Alcohol Consumption + 0.0602*Marital Status
39	0.1052*Diastolic Blood Pressure + 0.259*Marital Status
44	0.2289*Marital Status
99	0.0666*Age - 0.0215*Height + 0.0019*Triglycerides + 0.2416*Ever Smoked - 0.106*Current Smoking Status + 0.0262*Weekly Alcohol Consumption + 0.1361*Marital Status + 0.0037*Left Ventricular Mass + 0.0293*Left Ventricular Ejection Fractions + 0.068*Blood Urea Nitrogen - 0.045*Serum Albumin - 0.0038*Serum Bilirubin

Table 4-14 Risk Index Values for 25 Individuals from the Independent Testing Set from Five Randomly Selected Clinical Risk Index Models

Individual	Outcome	Bootstrap Sample #2 Cutoff Value = 0.313		Bootstrap Sample #16 Cutoff Value = 1.074		Bootstrap Sample #39 Cutoff Value = 4.257		Bootstrap Sample #44 Cutoff Value=0.458		Bootstrap Sample #99 Cutoff Value=0.370	
		Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	0.324	1	1.545	1	3.836	0	0.458	0	0.338	0
2	1	0.276	0	1.185	1	3.941	0	0.458	0	0.383	1
3	1	0.378	1	1.010	0	4.677	1	0.458	0	0.347	0
4	0	-0.095	0	0.912	0	4.257	0	0.458	0	0.589	1
5	1	0.345	1	1.186	1	4.677	1	0.458	0	0.364	0
6	0	0.322	1	0.911	0	4.046	0	0.458	0	0.332	0
7	0	0.304	0	1.028	0	3.941	0	0.458	0	0.296	0
8	0	0.287	0	0.736	0	3.310	0	0.458	0	0.168	0
9	0	0.216	0	1.042	0	3.941	0	0.458	0	0.236	0
10	1	0.321	1	1.111	1	3.941	0	0.458	0	0.418	1
11	1	0.270	0	1.050	0	4.677	1	0.458	0	0.283	0
12	0	0.324	1	1.311	1	4.467	1	0.458	0	0.343	0
13	0	0.235	0	1.028	0	3.941	0	0.458	0	0.270	0
14	1	-0.144	0	1.130	1	3.812	0	0.229	0	0.389	1
15	0	0.282	0	0.919	0	3.941	0	0.458	0	0.251	0
16	1	0.305	0	1.300	1	5.624	1	0.458	0	0.389	1
17	1	0.313	0	1.194	1	4.572	1	0.458	0	0.316	0
18	0	0.271	0	0.925	0	3.836	0	0.458	0	0.304	0
19	1	0.258	0	1.205	1	3.941	0	0.458	0	0.424	1
20	1	0.351	1	0.786	0	4.362	1	0.458	0	0.277	0
21	1	0.333	1	1.107	1	4.257	0	0.458	0	0.434	1
22	0	0.335	1	1.283	1	4.467	1	0.458	0	0.463	1
23	1	0.425	1	1.433	1	5.098	1	0.458	0	0.501	1
24	1	0.368	1	0.870	0	3.601	0	0.229	0	0.227	0
25	0	0.210	0	0.664	0	3.812	0	0.229	0	0.165	0

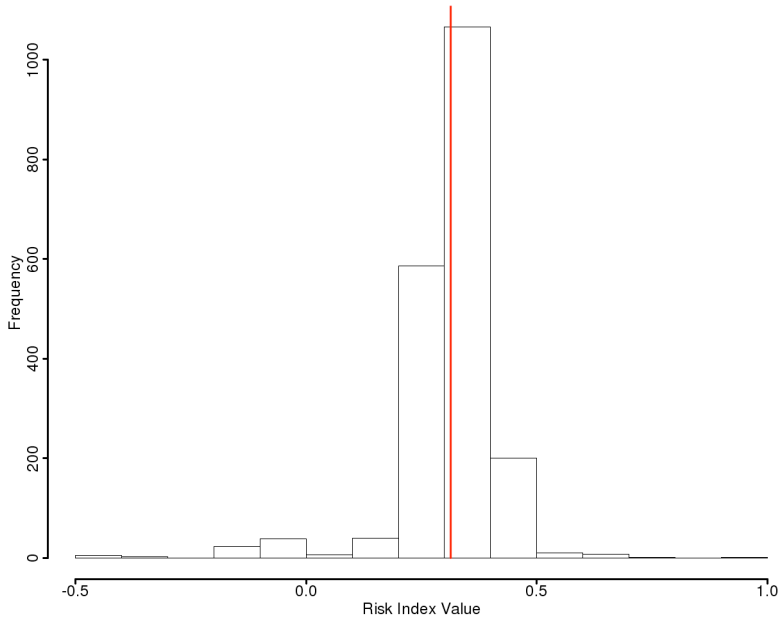


Figure 4-13 Distribution of Risk Index Values in the Optimization Set from the Clinical Risk Index Model for Bootstrap Sample #2

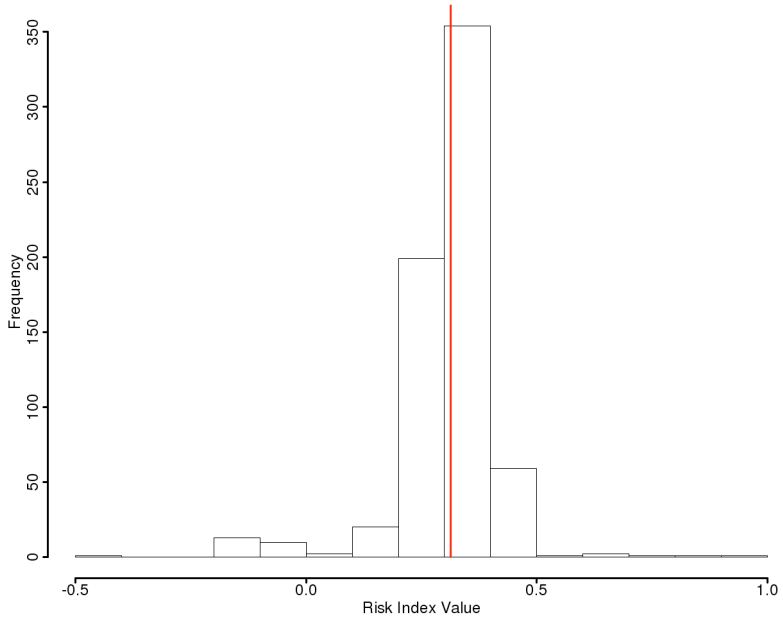


Figure 4-14 Distribution of Risk Index Values in the Independent Testing Set from the Clinical Risk Index Model for Bootstrap Sample #2

Table 4-15 shows the trimmed Clinical + Genotype risk index models for the same set of five bootstraps shown in Table 4-13, and Table 4-16 shows the risk index values and predictions of the same 25 randomly chosen individuals for the 5 bootstrap samples shown in Table 4-14. Figure 4-15 shows the distribution of risk index values in the optimization set for the Clinical + Genotype risk index model from one randomly selected bootstrap sample (Bootstrap Sample #2). Figure 4-16 shows the distribution of risk index values in the independent testing set for the Clinical + Genotype risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #2). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as “high risk of developing hypertension” and those with a risk index value lower are predicted as “low risk of developing hypertension”.

Table 4-15 Five Randomly Selected Clinical + Genotype Risk Index Models For Incident Hypertension

Bootstrap Sample	Model
2	- 0.2706*Sex - 0.016*Height + 0.0649*Marital Status + 0.0133*lvef + 0.015*Blood Glucose + 0.0569*Blood Urea Nitrogen - 0.8055*PC7 + 0.726*PC18 - 0.8886*PC28 + 0.5323*PC29 + 0.7198*PC43 + 0.334*PC52 - 1.8591*PC65 + 0.1256*PC73 - 4.362*PC95 - 1.1804*PC108 - 1.3473*PC115 - 0.6482*PC120 - 1.4195*PC130 + 4.2483*PC227 - 0.8451*PC263 - 0.3596*PC314 - 2.0389*PC348 - 0.4656*PC433 - 0.1687*PC464
16	0.0577*Age - 0.0407*Height + 0.0702*Systolic Blood Pressure - 0.0225*HDL + 0.0067*LDL + 0.0018*Triglycerides - 0.0038*Weekly Alcohol Consumption + 0.0602*Marital Status + 4.4639*PC18 + 0.8152*PC30 + 0.4643*PC107 + 0.1066*PC111 - 1.9152*PC130 + 4.2281*PC163 - 1.1749*PC167 - 1.345*PC196 - 0.9547*PC208 + 1.0625*PC211 - 2.6757*PC241 - 0.8091*PC245 + 1.9528*PC269 + 0.3945*PC306 - 0.8056*PC344 - 3.9691*PC393 + 5.869*PC435 + 7.5732*PC444 + 1.3192*PC448 + 0.8623*PC465
39	0.1052*Diastolic Blood Pressure + 0.259*Marital Status + 0.655*PC14 - 4.5884*PC28 - 5.0108*PC36 + 2.1823*PC146 - 4.5874*PC171 + 0.1737*PC205 + 0.5732*PC211 + 0.4999*PC230 + 0.6401*PC256 - 0.8464*PC282 + 0.2285*PC294 + 5.1635*PC334 + 1.4181*PC336 + 0.2759*PC342 + 0.6679*PC379 - 1.6134*PC397 + 2.7397*PC416 + 7.2186*PC427 + 0.5282*PC457
44	0.2289*Marital Status - 0.1024*PC23 + 0.1156*PC39 + 0.3855*PC60 + 0.4568*PC86 - 0.811*PC88 - 0.1054*PC94 + 0.7108*PC98 + 1.066*PC107 - 0.8398*PC156 + 0.5521*PC178 - 2.4885*PC179 + 0.4583*PC190 + 0.0081*PC195 - 0.166*PC197 + 0.4936*PC225 - 0.2339*PC243 - 0.0632*PC254 + 0.0569*PC358 - 0.6113*PC415 - 0.9741*PC455
99	0.0666*Age - 0.0215*Height + 0.0019*Triglycerides + 0.2416*Ever Smoked - 0.106*Current Smoking Status + 0.0262*Weekly Alcohol Consumption + 0.1361*Marital Status + 0.0037*Left Ventricular Mass + 0.0293*Left Ventricular Ejection Fractions + 0.068*Blood Urea Nitrogen - 0.045*Serum Albumin - 0.0038*Serum Bilirubin + 1.8893*PC24 - 1.1636*PC33 + 0.1657*PC93 - 2.3146*PC98 + 2.8609*PC186 + 1.1718*PC301 + 2.1745*PC332 - 1.6836*PC348 - 1.1166*PC356 + 1.5775*PC389 + 2.0837*PC394 - 0.6117*PC455 - 1.2163*PC493

Table 4-16 Risk Index Values for 25 Individuals in the Independent Testing Set for the Clinical + Genotype Risk Index Models from Five Randomly Selected Bootstrap Samples

Individual	Outcome	Bootstrap Sample #2 Cutoff Value = 0.312		Bootstrap Sample #16 Cutoff Value = 1.072		Bootstrap Sample #39 Cutoff Value = 4.447		Bootstrap Sample #44 Cutoff Value=0.459		Bootstrap Sample #99 Cutoff Value=0.371	
		Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	0.334	1	1.573	1	3.824	0	0.457	0	0.341	0
2	1	0.274	0	1.194	1	3.938	0	0.455	0	0.383	1
3	1	0.383	1	1.023	0	4.690	1	0.459	1	0.341	0
4	0	-0.086	0	0.914	0	4.258	0	0.460	1	0.596	1
5	1	0.353	1	1.214	1	4.705	1	0.456	0	0.349	0
6	0	0.333	1	0.908	0	4.044	0	0.460	1	0.319	0
7	0	0.308	0	1.039	0	3.946	0	0.457	0	0.293	0
8	0	0.281	0	0.723	0	3.320	0	0.457	0	0.169	0
9	0	0.216	0	1.043	0	3.941	0	0.453	0	0.227	0
10	1	0.320	1	1.113	1	3.933	0	0.459	0	0.405	1
11	1	0.272	0	1.050	0	4.689	1	0.455	0	0.286	0
12	0	0.308	0	1.323	1	4.468	1	0.455	0	0.355	0
13	0	0.238	0	1.037	0	3.947	0	0.458	0	0.277	0
14	1	-0.147	0	1.130	1	3.826	0	0.232	0	0.380	1
15	0	0.283	0	0.908	0	3.929	0	0.456	0	0.248	0
16	1	0.305	0	1.297	1	5.613	1	0.448	0	0.384	1
17	1	0.306	0	1.192	1	4.570	1	0.460	1	0.315	0
18	0	0.273	0	0.920	0	3.847	0	0.460	1	0.293	0
19	1	0.261	0	1.225	1	3.931	0	0.460	1	0.428	1
20	1	0.351	1	0.777	0	4.381	0	0.455	0	0.272	0
21	1	0.326	1	1.123	1	4.272	0	0.456	0	0.437	1
22	0	0.335	1	1.276	1	4.472	1	0.458	0	0.452	1
23	1	0.427	1	1.422	1	5.101	1	0.456	0	0.517	1
24	1	0.376	1	0.871	0	3.578	0	0.228	0	0.224	0
25	0	0.206	0	0.663	0	3.821	0	0.229	0	0.168	0

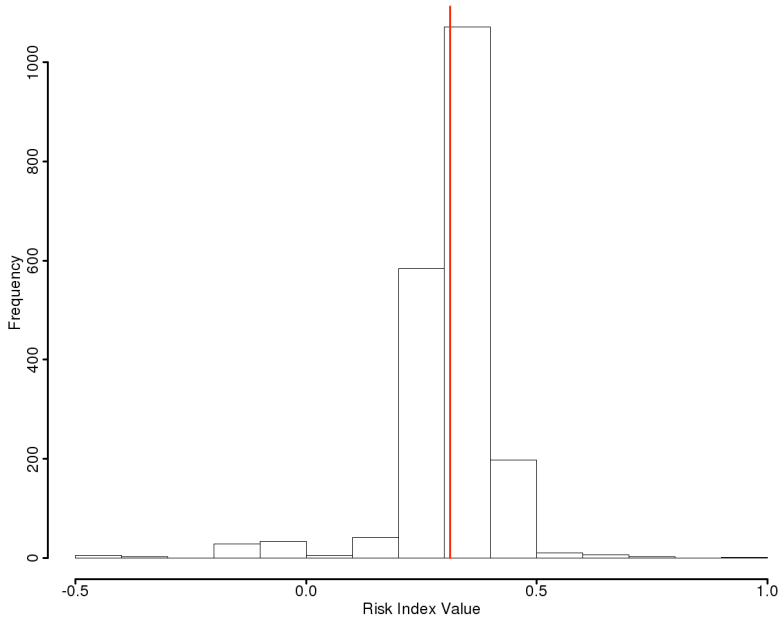


Figure 4-15 Distribution of Risk Index Values in the Optimization Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #2

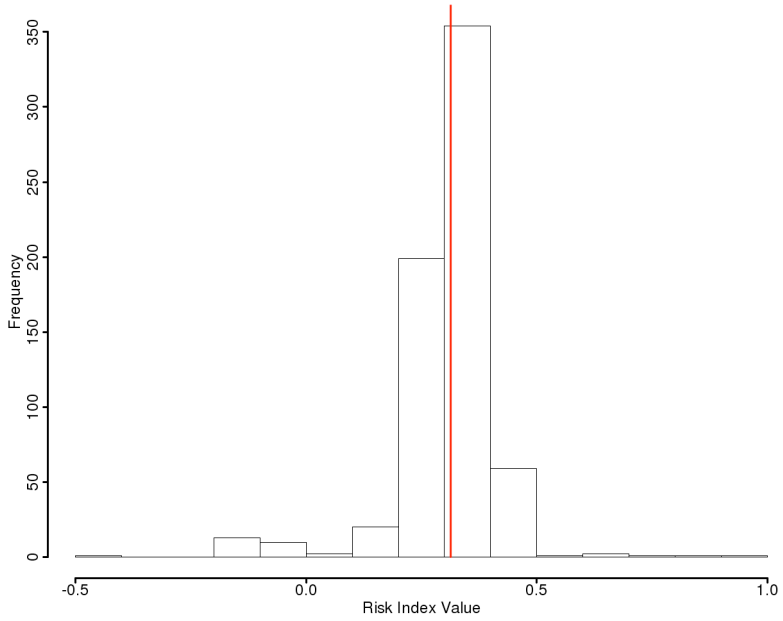


Figure 4-16 Distribution of Risk Index Values in the Independent Testing Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #2

4.6.3 Predictive Performance

Once predictions were made for each individual in the independent testing set the sensitivity, specificity, misclassification, and positive predictive value were calculated for the Clinical risk index model and the Clinical + Genotype risk index model as described in Section 4.5.3. The estimates and confidence intervals for sensitivity, specificity, misclassification, and positive predictive value are given in Table 4-17. Lastly, using the individual predictions from each of the 100 trimmed Clinical risk index models and 100 trimmed Clinical + Genotype risk index models for the individuals in the independent testing set, ROC curves were generated, and the AUC for the ROC curve was estimated for both the Clinical and Clinical + Genotype risk index models (Figure 4-17, Figure 4-18). For the Clinical risk index model the AUC for the ROC curve was 0.566, and for the Clinical + Genotype risk index model the AUC for the ROC curve was 0.563.

Table 4-17 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for Ten-year Incident Hypertension

Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
Clinical	0.608 (0.538-0.683)	0.505 (0.458-0.549)	0.468 (0.429-0.508)	0.299 (0.248-0.349)
Clinical + Genotype	0.591 (0.518-0.667)	0.544 (0.498-0.587)	0.444 (0.405-0.482)	0.31 (0.258-0.363)

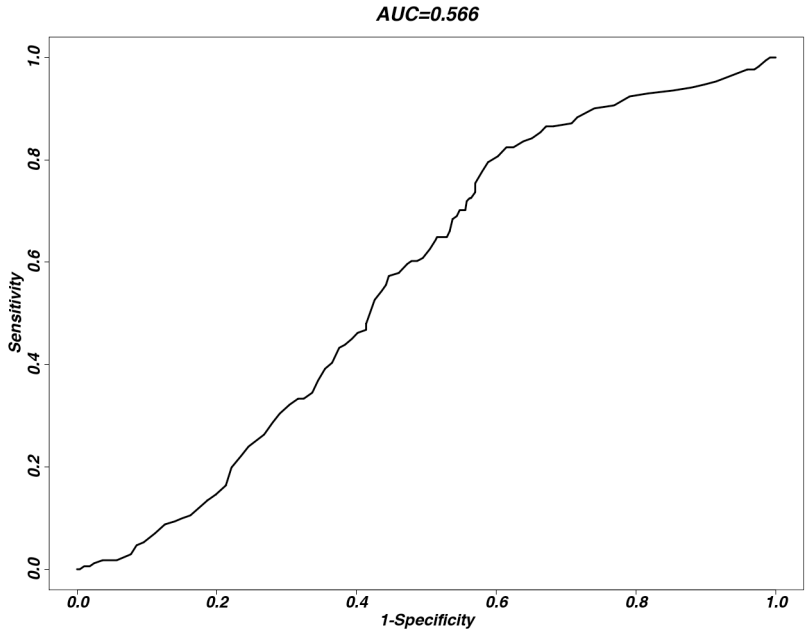


Figure 4-17 ROC Curve and AUC for the Incident Hypertension PCA Clinical Risk Index Model

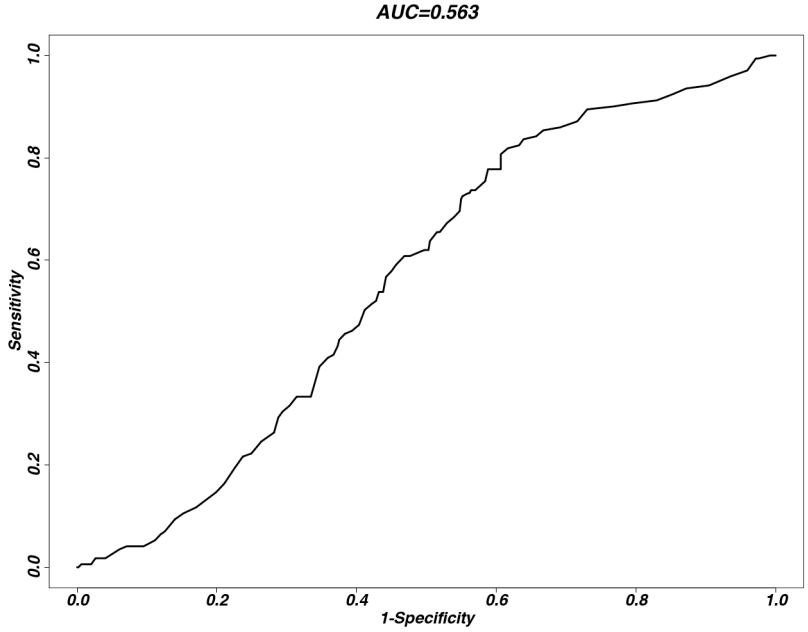


Figure 4-18 ROC Curve and AUC for the Incident Hypertension PCA Clinical + Genotype Risk Index Model

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. The proportion of models that predict that an individual is at high risk of developing hypertension, then, represents the predicted probability of an individual developing hypertension. Section 4.5.3 describes the calculation of a confidence interval for this predicted probability.

Figure 4-20 shows the distribution of the predicted probability of developing hypertension for the Clinical risk index model in the independent testing set, and Figure 4-11 shows the distribution of the predicted probability of developing hypertension for the Clinical + Genotype risk index model in the independent testing set. In both Figures, a density line is shown on the graph to indicate the density of a normal distribution with the mean and standard deviation matching that of the confidence score distribution.

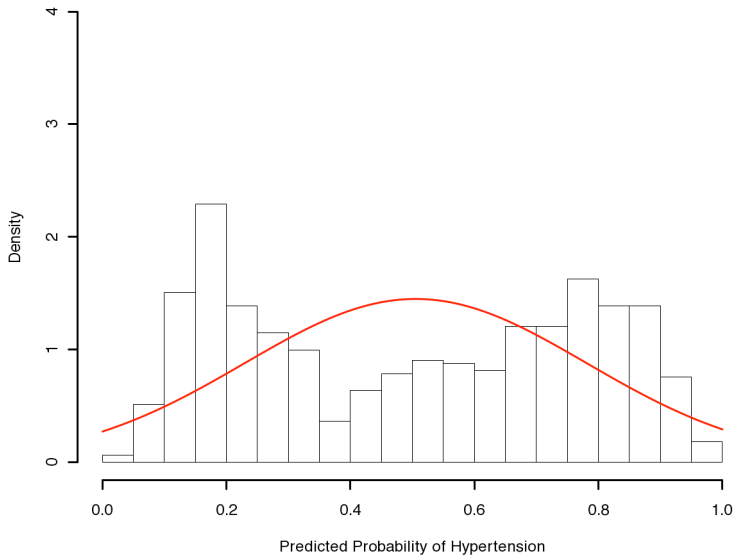


Figure 4-19 Histogram of the Predicted Probability of Developing Hypertension for the Clinical Risk Index Model

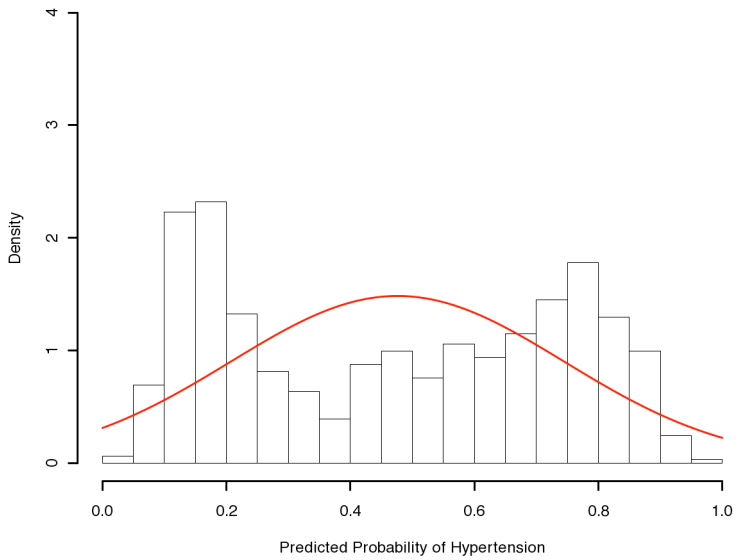


Figure 4-20 Histogram of the Predicted Probability of Developing Hypertension for the Clinical + Genotype Risk Index Model

4.6.4 Random Forests Comparison

A random forest was generated using the optimization set created by the risk index procedure. The forest had 500 individual trees, and the tuning procedure described in detail in Section 4.6.4 was used to find the number of variables *k* considered at each split that provided the lowest out-of-bag error estimate. The optimized *k* chosen was 45, which gave an out-of-bag error estimate of 31.4%.

When working with a dataset that has two possible classes, the standard procedure for a random forest is to assign a prediction to an individual based on a simple majority of votes, when the prevalence of the outcome is less than 50% changing the proportion of

votes needed to classify an individual can significantly impact the estimates of performance. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set on a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of 5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the individual to be "high risk". Table 4-18 shows the results of this investigation.

Predictions were then made about each individual in the independent testing set created by the risk index procedure, and the sensitivity, specificity, misclassification, and positive predictive value of the predictions was assessed. One thousand bootstrap samples of the independent testing set were generated, and predictions were made about each individual in each of the bootstrap samples. This data was used to create 95% confidence intervals for the sensitivity, specificity, misclassification, and positive predictive value estimates. Table 4-19 shows the sensitivity, specificity, misclassification, and positive predictive value estimates for the random forest as well as the 95% confidence interval for each estimate. Lastly, using the class votes for the individuals in the independent testing set, an ROC curve was created, and the AUC for the ROC curve was estimated (Figure 4-21).

Table 4-18 Performance Estimates of the Random Forest

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	1.000	0.000	0.686	0.314
0.1	1.000	0.003	0.684	0.315
0.15	1.000	0.031	0.665	0.321
0.2	0.979	0.116	0.613	0.336
0.25	0.952	0.310	0.488	0.387
0.3	0.836	0.480	0.409	0.424
0.35	0.712	0.646	0.333	0.479
0.4	0.534	0.762	0.310	0.506
0.45	0.329	0.843	0.318	0.490
0.5	0.123	0.950	0.310	0.529
0.55	0.027	0.994	0.310	0.667
0.6	0.000	0.997	0.316	0.000
0.65	0.000	1.000	0.314	-
0.7	0.000	1.000	0.314	-
0.75	0.000	1.000	0.314	-
0.8	0.000	1.000	0.314	-
0.85	0.000	1.000	0.314	-
0.9	0.000	1.000	0.314	-
0.95	0.000	1.000	0.314	-

Table 4-19 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for the Random Forest Model

Proportion of Votes for "High Risk" Class	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
0.2	0.979 (0.954-1.000)	0.116 (0.082-0.151)	0.613 (0.570-0.658)	0.336 (0.292-0.380)
0.25	0.952 (0.916-0.985)	0.310 (0.262-0.358)	0.488 (0.441-0.531)	0.387 (0.340-0.436)
0.3	0.836 (0.764-0.890)	0.480 (0.428-0.540)	0.409 (0.359-0.454)	0.424 (0.366-0.486)
0.35	0.712 (0.642-0.788)	0.646 (0.583-0.688)	0.333 (0.297-0.381)	0.479 (0.410-0.537)
0.4	0.534 (0.448-0.613)	0.762 (0.712-0.804)	0.310 (0.269-0.353)	0.506 (0.425-0.581)
0.45	0.329 (0.236-0.387)	0.843 (0.807-0.886)	0.318 (0.277-0.361)	0.490 (0.384-0.584)
0.5	0.123 (0.074-0.177)	0.950 (0.925-0.974)	0.310 (0.271-0.348)	0.529 (0.366-0.696)
0.55	0.027 (0.006-0.054)	0.994 (0.984-1.000)	0.310 (0.267-0.355)	0.667 (0.200-1.000)

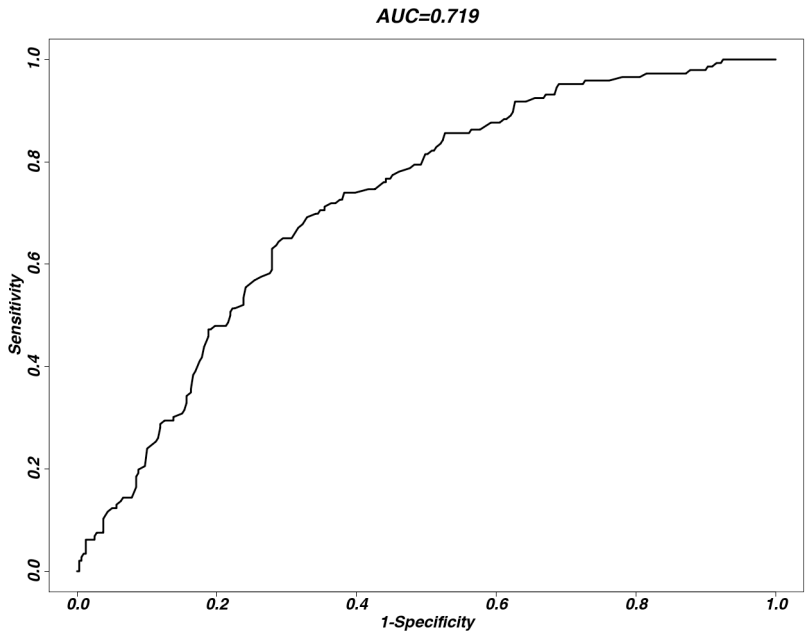


Figure 4-21 ROC Curve and AUC for the Incident Hypertension PCA Random Forest Model

4.6.5 Conclusion

As expected, the performance of the Clinical risk index model for this dataset is comparable to the performance of the Clinical risk index model built for the Incident Hypertension dataset with the 500 most highly associated SNPs. The inclusion of the top 500 principal components, however, increases the AUC of the Clinical + Genotype risk index model from 0.475 to 0.563. The random forest, however, did not perform as well when built using the top 500 principal components. The AUC for the random forest built using the principal components was 0.719, as opposed to 0.811 for the random forest built using the 500 most highly associated SNPs. Tables 4-18 and 4-19 show that there is not a proportion cutoff to assign a prediction of "high risk" for the random forest that achieves the same level of fairly high, balanced predictive performance as was available

when the random forest was built using the top 500 most highly associated SNPs. Even still, for the cutpoint that yields the most balanced predictive performance, the random forest performed better than either risk index model in overall classification accuracy, prediction specificity, and positive predictive value. However, the both the Clinical and Clinical + Genotype risk index model produced a sensitivity that was equal to that of the random forest.

4.7 Ten-Year Incident Diabetes Results Using 500 Most Highly Associated SNPs

4.7.1 Variable Selection

Using the procedure described in Section 4.5.1, the risk index procedure was performed to predict risk of developing diabetes within a ten-year time frame. The 500 SNPs most highly associated with this outcome (i.e., which had the lowest p-values from a logistic regression analysis of this outcome) were identified and used to build the risk index.

Table 4-20 gives a summary of the order in which variables were selected for the 100 bootstrap samples of the optimization set used to build the Clinical risk index model. Forty-seven out of the 100 trimmed Clinical risk index models contained marital status as a variable, and 46 contained weight. Blood glucose was included in 42 out of 100 of the 100 trimmed Clinical risk index models. Having ever smoked, current smoking status, and age were also frequently included in the set of 100 trimmed Clinical risk index models, appearing in 34, 31, and 29 trimmed Clinical risk index models, respectively.

Table 4-21 gives a summary of the order in which variables were selected for the 100 bootstrap samples of the Optimization Set used to build the Clinical + Genotype risk index model. Several SNPs were selected into the first five positions 15 or more times. Table 4-21 shows a summary of variable selection process for the 11 SNPs that were selected into 50 or more untrimmed Clinical + Genotype risk index models. All of these variables also appear in at least 30 trimmed Clinical + Genotype risk index model.

Table 4-20 Summary of Number of Times Each Variable is Selected into a Specific Model Position for the Clinical Risk Index Model for Incident Diabetes

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
Age (yrs)	2	10	11	4	9	1	7	8	4	4	2	5	3	6	4	3	3	2	3	4	95	29
Sex	5	6	12	5	4	6	4	5	4	5	5	2	5	1	1	7	4	2	6	7	96	24
Weight (lbs)	21	15	3	8	2	4	3	3	3	3	2	3	1	3	2	1	5	5	3	5	95	46
Height (in)	1	1	7	2	10	12	9	10	8	2	8	2	6	5	4	2	4	2	2	1	98	21
Systolic Blood Pressure (mm Hg)	4	2	5	2	4	5	6	4	2	6	9	8	3	5	3	3	4	1	6	5	87	15
Diastolic Blood Pressure (mm Hg)	1	0	5	0	0	5	6	2	7	8	9	6	6	6	3	3	7	3	2	6	85	17
Total Cholesterol (mg/dL)	3	0	1	2	7	2	10	3	8	6	9	6	5	9	7	4	4	6	3	3	98	16
High-density Lipoprotein Level (mg/dL)	3	10	2	8	5	4	2	3	0	5	1	11	7	4	6	5	2	5	6	7	96	28
Low-density Lipoprotein Level (mg/dL)	0	0	2	0	2	4	1	6	7	5	8	4	14	9	9	6	4	6	6	5	98	8
Triglycerides (mg/dL)	6	4	1	2	1	3	1	1	1	3	5	3	2	9	8	10	3	9	11	6	89	20
Ever Smoked	1	5	11	10	7	6	4	6	6	11	9	5	7	4	1	2	2	1	0	2	100	34
Currently Smokes	2	12	8	9	12	9	6	8	7	3	3	8	4	1	1	3	0	2	1	0	99	31
Weekly Alcohol Consumption	3	4	5	4	10	5	9	5	7	3	4	7	5	5	6	5	4	2	4	1	98	24
Marital Status	7	17	10	22	7	8	4	7	7	1	2	2	0	4	0	0	1	0	1	0	100	47
Left Ventricular Mass (g)	1	1	0	1	0	1	2	1	0	0	1	1	0	2	12	3	1	1	2	8	38	2
Left Ventricular Ejection Fraction (%)	1	1	1	1	0	0	2	1	0	1	2	1	2	0	3	13	2	4	2	3	40	5
Blood Glucose (mg/dL)	32	3	2	1	2	0	0	2	3	3	0	3	1	2	0	1	17	4	3	3	82	42
Blood Urea Nitrogen (mg/dL)	0	2	3	1	1	6	3	6	7	7	4	5	8	2	4	5	6	22	2	2	96	18

Total Serum Protein Level (mg/dL)	0	1	2	2	2	1	5	1	4	6	3	3	5	5	7	5	4	7	24	3	90	13
Serum Albumin Level (mg/dL)	3	4	1	4	6	6	5	4	7	2	5	3	6	5	6	4	2	4	1	18	96	23
Serum Bilirubin Level (mg/dL)	3	1	2	7	5	3	4	8	4	8	3	7	5	3	2	4	3	5	0	4	81	23
Serum Alkaline Phosphatase Level (mg/dL)	0	1	4	1	1	1	4	3	1	3	0	2	0	6	3	2	11	6	10	7	66	9
Serum Creatine Level (mg/dL)	1	0	2	4	3	8	3	3	3	5	6	3	5	4	8	9	7	1	2	0	77	16

Table 4-21 Summary of Number of Times Selected Genotype Variables are Selected into a Specific Model Position for the Clinical + Genotype Risk Index Model for Incident Diabetes

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
rs6891442	0	4	5	4	7	5	5	5	8	4	3	4	5	3	4	0	1	1	1	1	70	54
rs2021319	4	7	4	7	3	9	7	4	6	4	1	2	3	2	3	3	2	1	2	1	75	54
rs4986893	2	3	4	6	4	3	6	3	3	1	2	5	6	1	3	4	4	4	1	0	65	50
rs11975965	1	4	4	7	4	9	4	0	2	3	5	4	1	3	2	2	0	3	4	2	64	46
rs2720533	2	3	4	6	1	4	2	4	5	6	1	4	3	0	4	0	3	1	0	3	56	39
rs7114437	0	0	5	1	3	2	2	1	8	1	3	3	8	3	1	1	1	0	2	3	48	38
rs12610412	1	0	3	2	3	4	2	0	5	2	3	4	8	3	3	2	5	5	3	2	60	37
rs2069168	0	4	3	1	6	7	6	5	1	2	2	2	2	2	1	3	0	2	0	1	50	36
rs3821406	3	0	4	2	2	1	1	6	3	1	3	4	3	4	2	4	2	3	2	3	53	36
rs1804254	0	4	0	2	3	3	0	3	2	4	1	3	1	4	6	2	3	1	2	0	44	32
rs12459238	1	1	4	2	4	2	3	3	3	1	4	6	2	1	2	2	4	4	5	3	57	31

4.7.2 Models

Table 4-22 shows the trimmed Clinical risk index models for a selection of five random bootstrap samples. Figure 4-22 shows the distribution of risk index values in the optimization set for the Clinical risk index model from one randomly selected bootstrap sample (Bootstrap Sample #2), and Figure 4-23 shows the distribution of risk index values in the independent testing set for the Clinical risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #2). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as "high risk of developing diabetes" and those with a risk index value lower are predicted as "low risk of developing diabetes". Although Figures 4-22 and 4-23 appear discontinuous in their distribution, this is simply due to the presence of only one categorical variable, marital status, which takes only a small number of values, in the Clinical risk index model for bootstrap sample #2. Table 4-23 shows the risk index values for 25 randomly chosen individuals from the Independent Testing Set from these five Clinical risk index models along with that risk index model's prediction about each individual, where 0 indicates low risk of developing diabetes and 1 indicates high risk of developing diabetes.

Table 4-22 Five Randomly Selected Trimmed Clinical Risk Index Models for Incident Diabetes

Bootstrap Sample	Model
2	-0.2017*Marital Status
45	0.0396*Age + 0.0605*Height + 0.0096*Total Cholesterol - 0.0655*HDL + 0.0092*LDL + 0.2239*Ever Smoked + 0.1754*Current Smoking Status + 0.0213*Weekly Alcohol Consumption + 0.0531*Marital Status + 0.1124*Blood Glucose + 0.0491*bun + 0.03*Serum Albumin + 0.0021*Serum Bilirubin + 0.0451*Serum Alkaline Phophatase + 0.0375*Serum Creatine
54	0.0208*Weight + 1e-04*Height + 0.0583*diastolicBP + 0.0018*Triglycerides + 0.2566*Ever Smoked + 0.0709*Current Smoking Status + 0.121*Marital Status + 0.1016*Blood Glucose + 0.0369*Total Serum Protein - 0.0446*Serum Albumin - 0.0031*Serum Bilirubin + 0.0349*Serum Creatine
59	0.0255*Weight + 0.0735*Height + 0.0062*Total Cholesterol - 0.0738*HDL + 0.0294*Weekly Alcohol Consumption
81	0.0317*Age - 0.5642*Sex + 0.0229*Weight + 0.0258*Height - 0.0571*HDL + 0.1177*Marital Status

Table 4-23 Risk Index Values for 25 Individuals from the Independent Testing Set for Clinical Risk Index Models for Incident Diabetes from Five Randomly Selected Bootstrap Samples

		Bootstrap Sample #2 Cutoff Value = -0.403		Bootstrap Sample #45 Cutoff Value = 1.536		Bootstrap Sample #54 Cutoff Value = 1.765		Bootstrap Sample #59 Cutoff Value=1.890		Bootstrap Sample #76 Cutoff Value=0.958	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	-0.403	0	1.146	0	1.602	0	0.681	0	0.029	0
2	0	-0.403	0	1.423	0	1.756	0	1.886	0	0.871	0
3	0	-0.403	0	1.430	0	1.701	0	1.434	0	0.549	0
4	0	-0.403	0	1.220	0	1.699	0	1.607	0	0.712	0
5	1	-0.403	0	1.533	0	1.910	1	1.803	0	0.964	1
6	0	-0.403	0	1.416	0	1.568	0	0.711	0	0.080	0
7	0	-0.202	1	1.350	0	1.214	0	3.037	1	0.985	1
8	0	-0.807	0	1.316	0	1.571	0	1.156	0	0.431	0
9	0	-0.403	0	1.507	0	1.704	0	1.263	0	0.594	0
10	0	-0.403	0	1.509	0	1.879	1	1.380	0	0.599	0
11	1	-0.202	1	1.399	0	1.587	0	1.517	0	0.660	0
12	0	-1.009	0	1.296	0	1.734	0	1.502	0	0.699	0
13	0	-0.403	0	1.417	0	1.829	1	1.580	0	0.646	0
14	0	-0.403	0	1.495	0	1.525	0	1.076	0	0.352	0
15	0	-0.403	0	1.451	0	1.945	1	1.396	0	0.601	0
16	1	-0.403	0	1.724	1	1.603	0	1.174	0	0.382	0
17	0	-0.403	0	1.166	0	1.713	0	1.120	0	0.357	0
18	0	-0.403	0	1.166	0	1.830	1	3.653	1	1.538	1
19	0	-0.605	0	1.369	0	1.345	0	0.454	0	0.010	0
20	0	-0.403	0	1.386	0	1.712	0	1.925	1	0.968	1
21	0	-0.403	0	1.610	1	1.592	0	1.633	0	0.766	0
22	0	-0.403	0	1.351	0	1.474	0	1.071	0	0.323	0
23	0	-0.403	0	0.794	0	1.654	0	1.546	0	0.703	0
24	0	-0.403	0	1.346	0	1.527	0	1.031	0	0.393	0
25	1	-0.807	0	1.560	1	1.704	0	1.602	0	0.785	0

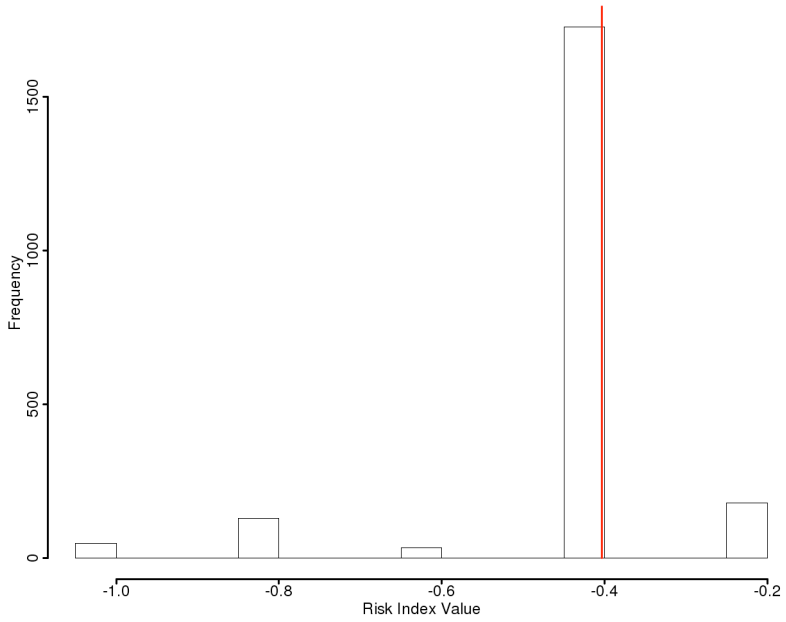


Figure 4-22 Distribution of Risk Index Values in the Optimization Set from the Clinical Risk Index Model for Bootstrap Sample #2

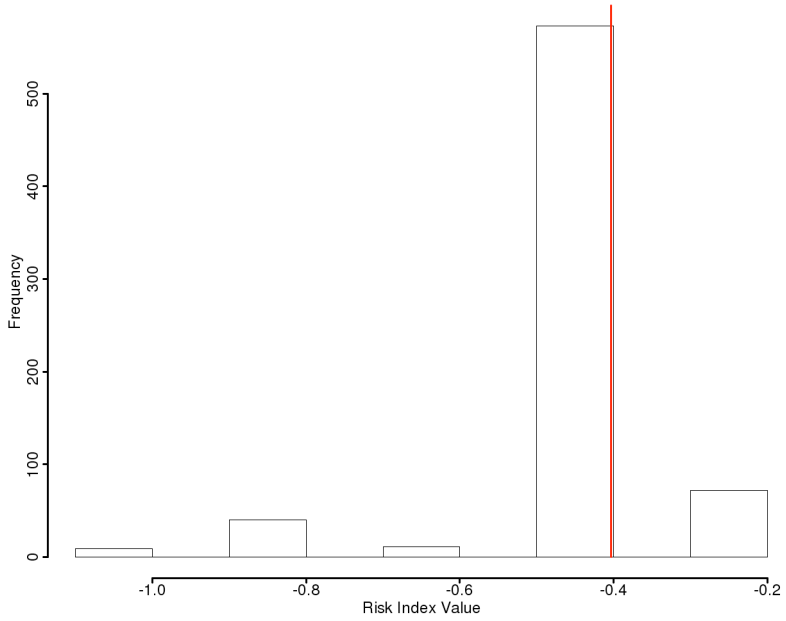


Figure 4-23 Distribution of Risk Index Values in the Independent Testing Set from the Clinical Risk Index Model for Bootstrap Sample #2

Table 4-24 shows the trimmed Clinical + Genotype risk index models for the same set of five bootstraps shown in Table 4-22, and Table 4-25 shows the risk index values and predictions of the same 25 randomly chosen individuals for the 5 bootstrap samples shown in Table 4-23. Figure 4-24 shows the distribution of risk index values in the optimization set for the Clinical + Genotype risk index model from one randomly selected bootstrap sample (Bootstrap Sample #2). Figure 4-25 shows the distribution of risk index values in the independent testing set for the Clinical + Genotype risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #2). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as “high risk of developing diabetes” and those with a risk index value lower are predicted as “low risk of developing diabetes”.

Table 4-24 Five Randomly Selected Clinical + Genotype Risk Index Models for Incident Diabetes

Bootstrap Sample	Model
2	-0.2017*Marital Status + rs12610412(C_G=0.4812) + rs2239811(A_C=0.3766, C_C=-0.9967) + rs2021319(G_G=-9.8516) + rs2720533(A_G=15.0516) + rs3821406(A_C=-0.8067) + rs2908780(G_G=-0.4694) + rs3918021(G_G=0.4714)
45	0.0396*Age + 0.0605*Height + 0.0096*Total Cholesterol - 0.0655*HDL + 0.0092*LDL + 0.2239*Ever Smoked + 0.1754*Current Smoking Status + 0.0213*Weekly Alcohol Consumption + 0.0531*Marital Status + 0.1124*Blood Glucose + 0.0491*bun + 0.03*Serum Albumin + 0.0021*Serum Bilirubin + 0.0451*Serum Alkaline Phophatase + 0.0375*Serum Creatine + rs2069168(G_G=96.6672) + rs3736352(A_C=-0.1512, C_C=-712.1793) + rs3745489(C_T=-0.752) + rs12072734(A_G=1.5317, G_G=1.645) + rs3745581(C_G=27.0546, G_G=27.7574) + rs16939879(A_G=-0.546, G_G=-0.5386) + rs1017842(C_G=98.7711, G_G=39.6089) + rs6891442(T_T=3.2488) + rs2021319(G_G=3.0098) + rs2720533(A_G=-2.3309) + rs1804254(C_C=1.2153) + rs10514767(C_G=0.6335, G_G=0.71) + rs4711000(C_T=-0.2313, T_T=-0.5737) + rs17010210(G_G=0.4383) + rs3918021(G_G=-0.092)
54	0.0208*Weight + 1e-04*Height + 0.0583*diastolicBP + 0.0018*Triglycerides + 0.2566*Ever Smoked + 0.0709*Current Smoking Status + 0.121*Marital Status + 0.1016*Blood Glucose + 0.0369*Total Serum Protein - 0.0446*Serum Albumin - 0.0031*Serum Bilirubin + 0.0349*Serum Creatine + rs2289622(G_G=0.9044) + rs1800361(G_G=0.6614) + rs3111222(A_G=-0.0578, G_G=-0.0525) + rs9635334(A_T=0.2905, T_T=-0.8468) + rs3764633(A_G=-0.7265, G_G=-0.676) + rs17787561(C_T=-0.4008) + rs4986893(G_G=1.8691) + rs9724933(A_G=0.5904, G_G=0.6278) + rs6891442(T_T=3.9009) + rs2021319(G_G=2.2705) + rs2720533(A_G=-2.2751) + rs1804254(C_C=1.2692) + rs12818539(A_C=0.2794, C_C=0.3856) + rs12136578(C_T=0.2783, T_T=0.276) + rs17010210(G_G=0.8918) + rs3918021(G_G=0.1882) + rs11975965(C_T=-1.6993)
59	0.0255*Weight + 0.0735*Height + 0.0062*Total Cholesterol - 0.0738*HDL + 0.0294*Weekly Alcohol Consumption + rs4084639(G_G=0.6644) + rs2290427(A_C=1.5041, C_C=2.2818) + rs2289622(G_G=0.3645) + rs3745489(C_T=-0.5684) + rs17179966(A_G=-0.324, G_G=0.324) + rs9257940(A_G=-0.0026, G_G=-1.5549) + rs12464093(G_T=0.1202, T_T=0.3303) + rs2288663(A_G=1.2017, G_G=0.9963) + rs7373862(A_G=0.0795, G_G=-0.2653) + rs751191(G_T=0.1059, T_T=-0.3885) + rs12528104(C_T=0.5498, T_T=0.549) + rs6891442(T_T=1.8469) + rs2021319(G_G=0.6368) + rs3821406(A_C=-0.1651) + rs3918021(G_G=0.4557) + rs11975965(C_T=-2.142)
81	0.0317*Age - 0.5642*Sex + 0.0229*Weight + 0.0258*Height - 0.0571*HDL + 0.1177*Marital Status + rs2289622(G_G=1.1543) + rs7114437(A_T=-0.4356) + rs12610412(C_G=-0.018) + rs6086342(A_C=0.415, C_C=0.4348) + rs555990(C_G=1.0379, G_G=1.1079) + rs4986893(G_G=1.818) + rs11220285(A_G=-0.0185, G_G=0.4039) + rs6924468(A_G=0.8622, G_G=0.9034) + rs3745581(C_G=-0.1301, G_G=0.1301) + rs11889528(A_G=-15.5226, G_G=-15.4986) + rs6891442(T_T=2.2284) + rs4910163(A_G=0.1847, G_G=-0.0056) + rs3821406(A_C=-1.8263) + rs1804254(C_C=2.2329) + rs2286975(A_G=0.1277, G_G=0.2051) + rs12085435(A_G=0.9808, G_G=1.1449)

Table 4-25 Risk Index Values for 25 Individuals in the Independent Testing Set for the Clinical + Genotype Risk Index Models for Incident Diabetes from Five Randomly Selected Bootstrap Samples

		Bootstrap Sample #2 Cutoff Value = -1.810		Bootstrap Sample #45 Cutoff Value = 13.125		Bootstrap Sample #54 Cutoff Value = 2.494		Bootstrap Sample #59 Cutoff Value=2.370		Bootstrap Sample #76 Cutoff Value=0.691	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	-1.810	0	12.723	0	2.298	0	1.208	0	-0.231	0
2	0	-1.878	0	13.022	0	2.482	0	2.330	0	1.578	1
3	0	-1.810	0	13.006	0	2.431	0	1.881	0	0.294	0
4	0	-1.810	0	12.748	0	2.436	0	2.092	0	0.452	0
5	1	-1.757	1	13.126	1	2.644	1	2.282	0	0.701	1
6	0	-1.810	0	13.007	0	2.392	0	1.219	0	-0.197	0
7	0	-1.555	1	12.949	0	1.950	0	3.551	1	0.723	1
8	0	-2.214	0	13.669	1	2.322	0	1.683	0	0.158	0
9	0	-1.810	0	13.083	0	2.404	0	1.740	0	0.326	0
10	0	-1.757	1	13.123	0	2.615	1	1.907	0	0.265	0
11	1	-1.609	1	12.948	0	2.320	0	2.000	0	0.399	0
12	0	-2.416	0	12.867	0	2.470	0	1.990	0	0.371	0
13	0	-1.810	0	12.979	0	2.557	1	1.960	0	0.371	0
14	0	-1.757	1	13.093	0	2.269	0	1.603	0	0.062	0
15	0	-1.810	0	12.936	0	2.682	1	1.908	0	0.308	0
16	1	-1.810	0	13.328	1	2.336	0	1.678	0	0.123	0
17	0	-1.810	0	12.742	0	2.447	0	1.605	0	0.098	0
18	0	-1.810	0	12.734	0	2.558	1	4.179	1	1.265	1
19	0	-2.012	0	12.944	0	1.845	0	0.871	0	-0.393	0
20	0	-1.810	0	12.984	0	2.445	0	2.358	0	0.633	0
21	0	-1.757	1	13.167	1	2.369	0	2.133	0	0.505	0
22	0	-1.810	0	12.949	0	2.216	0	1.585	0	0.062	0
23	0	-1.810	0	12.408	0	2.390	0	2.019	0	0.373	0
24	0	-1.810	0	12.960	0	2.257	0	1.549	0	0.124	0
25	1	-2.160	0	13.122	0	2.438	0	2.087	0	0.511	0

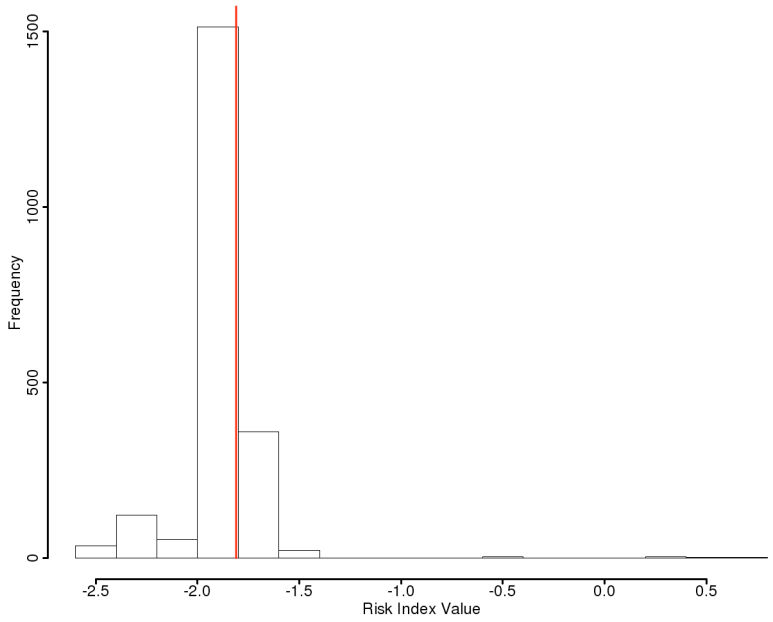


Figure 4-24 Distribution of Risk Index Values in the Optimization Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #2

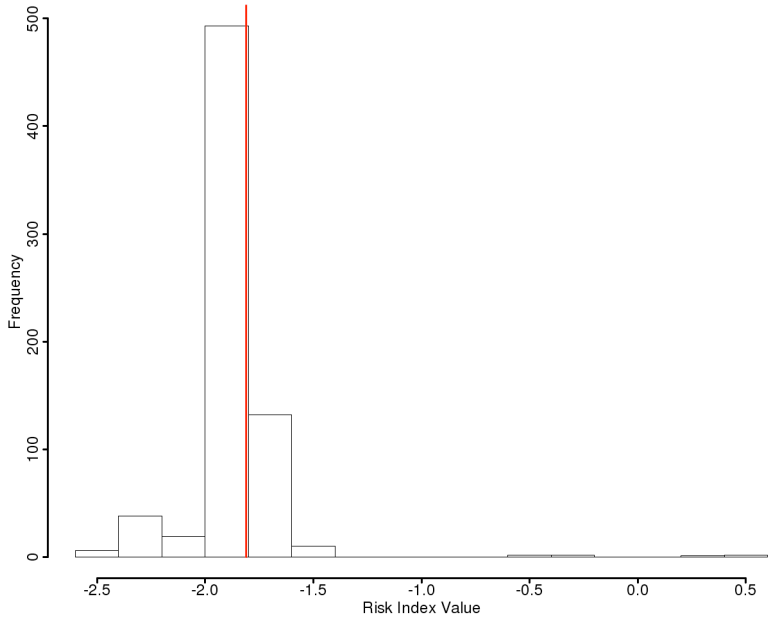


Figure 4-25 Distribution of Risk Index Values in the Independent Testing Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #2

4.7.3 Predictive Performance

Once predictions were made for each individual in the independent testing set the sensitivity, specificity, misclassification, and positive predictive value were calculated for the Clinical risk index model and the Clinical + Genotype risk index model as described in Section 4.5.3. The estimates and confidence intervals for sensitivity, specificity, misclassification, and positive predictive value are given in Table 4-26. Lastly, using the individual predictions from each of the 100 trimmed Clinical risk index models and 100 trimmed Clinical + Genotype risk index models for the individuals in the independent testing set, ROC curves were generated, and the AUC for the ROC curve was estimated for both the Clinical and Clinical + Genotype risk index models (Figure 4-26, Figure 4-27). For the Clinical risk index model the AUC for the ROC curve was 0.722, and for the Clinical + Genotype risk index model the AUC for the ROC curve was 0.683.

Table 4-26 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for Ten-year Incident Diabetes

Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
Clinical	0.224 (0.13-0.323)	0.901 (0.877-0.922)	0.163 (0.136-0.193)	0.192 (0.113-0.29)
Clinical + Genotype	0.104 (0.04-0.183)	0.923 (0.901-0.943)	0.155 (0.129-0.182)	0.125 (0.05-0.22)

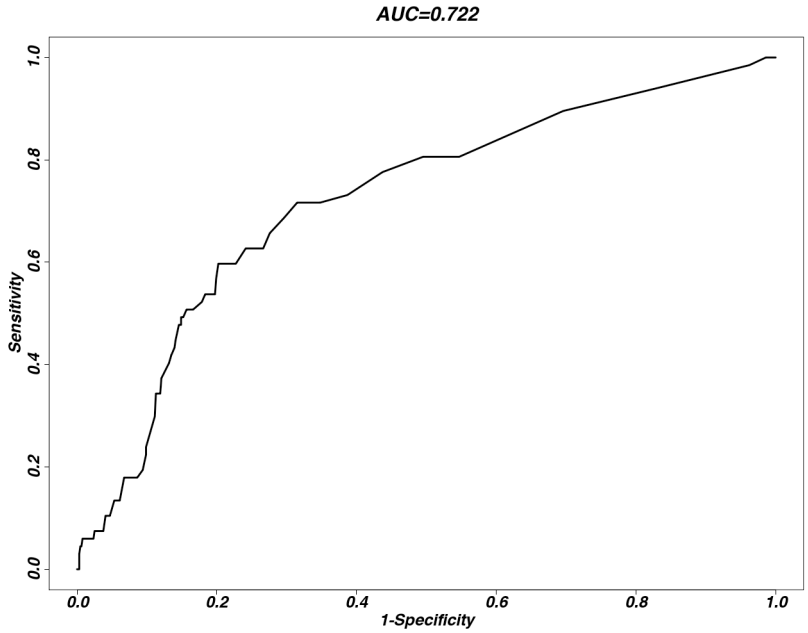


Figure 4-26 ROC Curve and AUC for the Incident Diabetes Clinical Risk Index Model

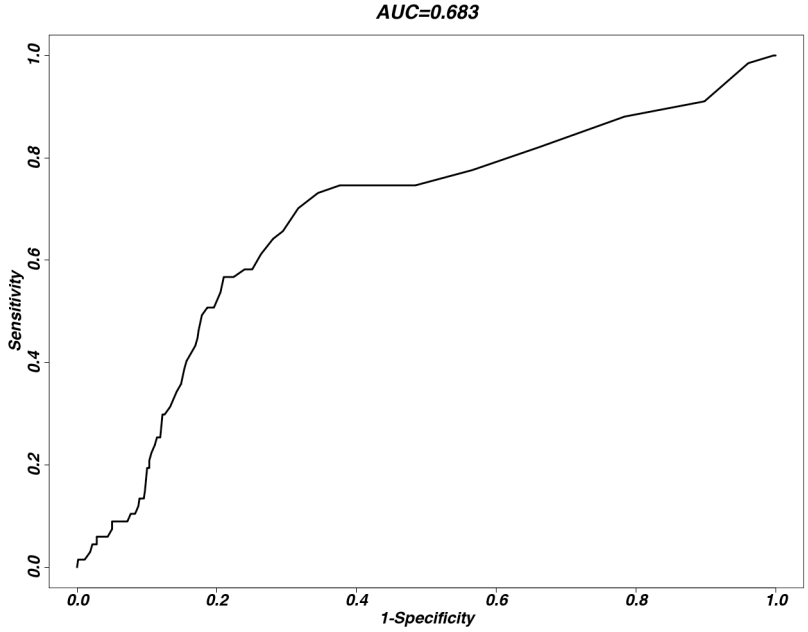


Figure 4-27 ROC Curve and AUC for the Incident Diabetes Clinical + Genotype Risk Index Model

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. The proportion of models that predict that an individual is at high risk of developing diabetes, then, represents the predicted probability of an individual developing diabetes. Section 4.5.3 describes the calculation of a confidence interval for this predicted probability.

Figure 4-28 shows the distribution of the predicted probability of developing diabetes for the Clinical risk index model in the independent testing set, and Figure 4-29 shows the distribution of the predicted probability of developing diabetes for the Clinical + Genotype risk index model in the independent testing set. In both Figures, a density line is shown on the graph to indicate the density of a normal distribution with the mean and standard deviation matching that of the predicted probability distribution.

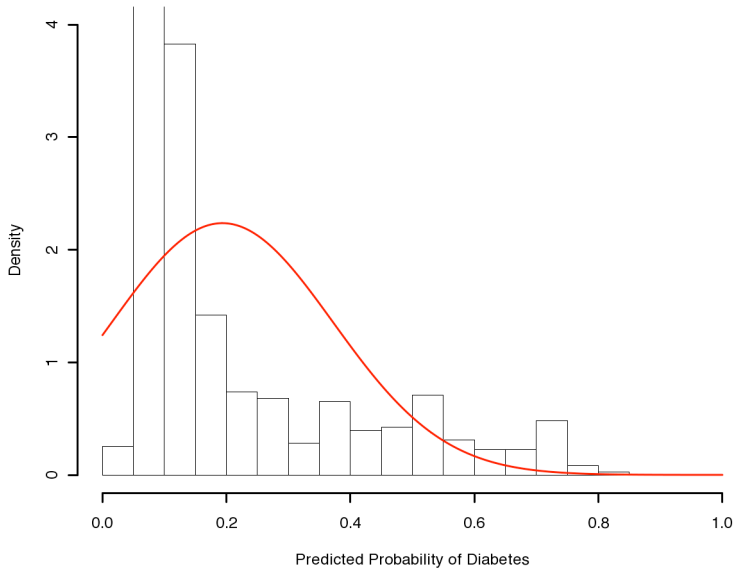


Figure 4-28 Histogram of the Predicted Probability of Developing Diabetes for the Clinical Risk Index Model

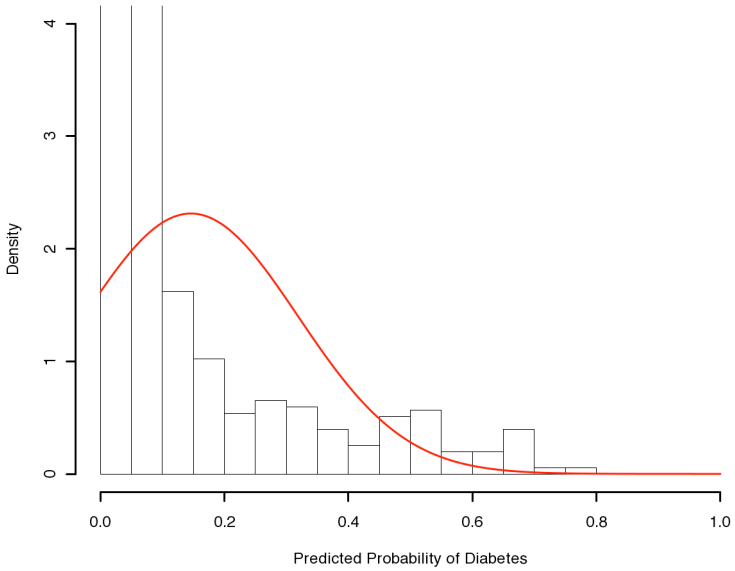


Figure 4-29 Histogram of the Predicted Probability of Developing Diabetes for the Clinical + Genotype Risk Index Model

4.7.4 Random Forests Comparison

A random forest was generated using the optimization set created by the risk index procedure. The forest had 500 individual trees, and the tuning procedure described in detail in Section 4.6.4 was used to find the number of variables *k* considered at each split that provided the lowest out-of-bag error estimate. The optimized *k* chosen was 10, which gave an out-of-bag error estimate of 8.84%.

When working with a dataset that has two possible classes, the standard procedure for a random forest is to assign a prediction to an individual based on a simple majority of votes, when the prevalence of the outcome is less than 50% changing the proportion of

votes needed to classify an individual can significantly impact the estimates of performance. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set on a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of 5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the individual to be "high risk". Table 4-27 shows the results of this investigation.

Predictions were then made about each individual in the independent testing set created by the risk index procedure, and the sensitivity, specificity, misclassification, and positive predictive value of the predictions was assessed. One thousand bootstrap samples of the independent testing set were generated, and predictions were made about each individual in each of the bootstrap samples. This data was used to create 95% confidence intervals for the sensitivity, specificity, misclassification, and positive predictive value estimates. Table 4-28 shows the sensitivity, specificity, misclassification, and positive predictive value estimates for the random forest as well as the 95% confidence interval for each estimate. Lastly, using the class votes for the individuals in the independent testing set, an ROC curve was created, and the AUC for the ROC curve was estimated (Figure 4-30).

Table 4-27 Performance Estimates of the Random Forest

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	1.000	0.065	0.850	0.097
0.1	0.956	0.635	0.335	0.209
0.15	0.556	0.946	0.089	0.510
0.2	0.089	0.996	0.087	0.667
0.25	0.000	1.000	0.091	-
0.3	0.000	1.000	0.091	-
0.35	0.000	1.000	0.091	-
0.4	0.000	1.000	0.091	-
0.45	0.000	1.000	0.091	-
0.5	0.000	1.000	0.091	-
0.55	0.000	1.000	0.091	-
0.6	0.000	1.000	0.091	-
0.65	0.000	1.000	0.091	-
0.7	0.000	1.000	0.091	-
0.75	0.000	1.000	0.091	-
0.8	0.000	1.000	0.091	-
0.85	0.000	1.000	0.091	-
0.9	0.000	1.000	0.091	-
0.95	0.000	1.000	0.091	-

Table 4-28 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for the Random Forest Model

Proportion of Votes for "High Risk" Class	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
0.1	0.956 (0.881-1.000)	0.635 (0.589-0.678)	0.335 (0.295-0.378)	0.209 (0.156-0.263)
0.15	0.556 (0.412-0.700)	0.946 (0.926-0.966)	0.089 (0.065-0.114)	0.510 (0.368-0.646)
0.2	0.089 (0.019-0.182)	0.996 (0.989-1.000)	0.087 (0.063-0.114)	0.667 (0.200-1.000)

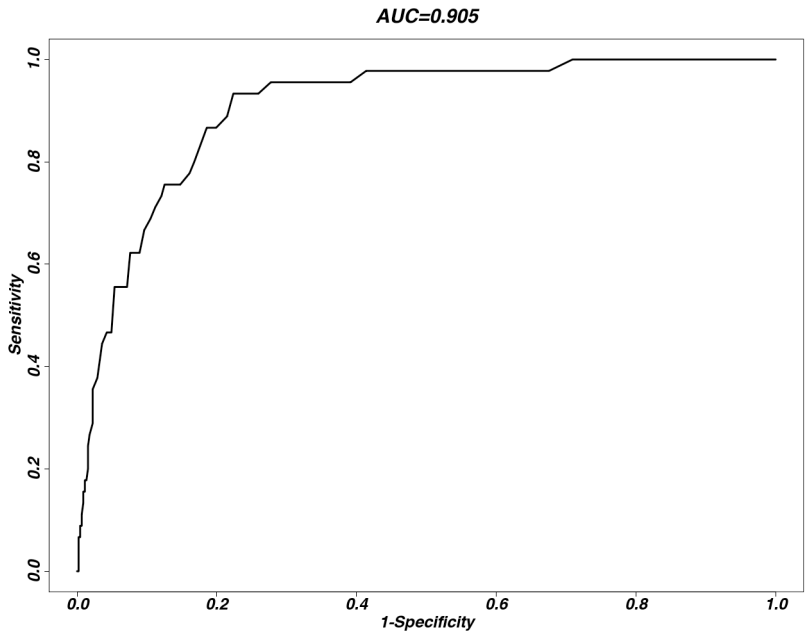


Figure 4-30 ROC Curve and AUC for the Incident Hypertension PCA Random Forest Model

4.7.5 Conclusion

Both the Clinical and Clinical + Genotype risk index models had very high specificities but very low sensitivities. However, the AUC for both is higher than the AUCs achieved by the Clinical and Clinical + Genotype risk index models in the Incident Hypertension analysis. Choosing the proportion cutoff that gives the random forest model the most balanced performance (0.15), the sensitivity, misclassification, and positive predictive value of the random forest model is greater than either the Clinical or the Clinical + Genotype risk index model. However, the specificity of the random forest is in the same range as the Clinical and Clinical + Genotype risk index models. The ROC curves for each of the risk index models also exhibit a greater curve than that seen in the Incident Hypertension analysis. This suggests that modifying the proportion of votes required to

give a prediction of "high risk" might be able to increase the predictive performance of the models.

4.8 Ten-Year Incident Diabetes Results Using Top 500 Principal Components

4.8.1 Variable Selection

Using the procedure described in Section 4.5.1, the risk index procedure was performed to predict risk of developing diabetes within a ten-year time frame. In place of the 500 SNPs most highly associated with ten-year incident diabetes, the top 500 principal components from a principal components analysis of all available SNPs were used.

Table 4-29 gives a summary of the order in which variables were selected for the 100 bootstrap samples of the optimization set used to build the Clinical risk index model. Forty-seven out of the 100 trimmed Clinical risk index models contained marital status as a variable, and 46 contained weight. Marital status was included in 56 out of 100 of the 100 trimmed Clinical risk index models. Weight, weekly alcohol consumption, and current smoking status, were also frequently included in the set of 100 trimmed Clinical risk index models, appearing in 51, 36, and 31 trimmed Clinical risk index models, respectively. Table 4-30 gives a summary of the order in which the most commonly chosen principal components were selected into the 100 Clinical + Genotype risk index models. Each principal component in Table 4-30 appears in at least 10 trimmed Clinical + Genotype risk index models.

Table 4-29 Summary of Number of Times Each Variable is Selected into a Specific Model Position for the Clinical Risk Index Model for Incident Diabetes

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
Age (yrs)	1	8	8	7	6	6	5	11	3	2	7	6	4	6	3	3	4	2	2	2	96	26
Sex	3	6	4	11	7	14	7	4	5	2	1	6	2	2	6	4	2	1	3	6	96	26
Weight (lbs)	32	6	5	3	6	3	4	6	3	3	0	4	5	2	0	2	2	3	3	1	93	51
Height (in)	2	4	7	6	14	13	7	9	8	5	4	2	6	5	1	2	2	0	0	3	100	27
Systolic Blood Pressure (mm Hg)	2	5	6	3	2	3	5	3	13	13	3	3	3	7	6	4	4	4	2	3	94	19
Diastolic Blood Pressure (mm Hg)	6	5	1	1	3	2	4	7	8	10	8	4	4	2	5	1	3	10	3	3	90	22
Total Cholesterol (mg/dL)	2	2	1	2	0	4	4	3	2	5	8	14	7	7	4	9	8	4	5	7	98	10
High-density Lipoprotein Level (mg/dL)	5	2	5	5	2	3	1	5	5	1	12	10	10	6	2	5	6	1	6	3	95	21
Low-density Lipoprotein Level (mg/dL)	1	1	3	2	3	2	1	2	6	8	3	7	14	7	5	5	8	8	4	4	94	14
Triglycerides (mg/dL)	6	4	0	2	2	0	1	5	1	6	8	2	7	13	5	4	2	7	11	6	92	16
Ever Smoked	0	8	10	14	8	10	11	9	6	1	2	3	2	4	4	4	2	0	1	1	100	28
Currently Smokes	1	12	15	6	6	5	11	7	4	8	5	3	4	3	2	3	0	0	2	2	99	31
Weekly Alcohol Consumption	5	6	9	11	6	10	10	6	5	4	4	5	4	3	3	3	3	0	2	0	99	36
Marital Status	7	17	18	13	12	8	8	2	6	3	1	2	0	2	0	0	1	0	0	0	100	56
Left Ventricular Mass (g)	0	0	0	0	1	0	0	0	0	2	1	0	0	0	18	2	2	6	1	9	42	1
Left Ventricular Ejection Fraction (%)	1	2	0	2	0	0	1	0	4	1	2	2	0	0	0	15	2	4	4	3	43	6
Blood Glucose (mg/dL)	18	5	1	2	0	2	3	1	1	4	0	3	2	2	1	2	16	6	5	5	79	28
Blood Urea Nitrogen (mg/dL)	3	0	0	4	4	6	1	7	5	2	3	4	6	7	7	5	5	20	7	2	98	17

Total Serum Protein Level (mg/dL)	1	0	1	0	2	1	2	1	2	2	8	2	2	4	8	5	5	9	21	10	86	9
Serum Albumin Level (mg/dL)	1	2	1	3	7	3	4	5	2	8	8	5	8	6	4	4	6	3	0	17	97	20
Serum Bilirubin Level (mg/dL)	1	3	3	3	8	1	6	5	5	6	4	3	3	5	7	6	4	4	2	0	79	16
Serum Alkaline Phosphatase Level (mg/dL)	2	1	1	0	0	2	2	0	1	2	2	3	3	4	5	4	7	5	9	7	60	9
Serum Creatine Level (mg/dL)	0	1	1	0	1	2	2	2	5	2	6	7	4	3	4	8	6	3	7	6	70	11

Table 4-30 Summary of Number of Times Selected Principal Components Variables are Selected into a Specific Model Position for the Clinical + Genotype Risk Index Model for Incident Diabetes

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
PC3	1	1	3	0	1	6	6	2	2	2	1	4	0	0	0	2	2	1	1	1	36	29
PC10	0	0	1	1	1	1	0	2	2	1	3	0	0	1	0	1	0	0	1	0	15	13
PC11	0	1	1	0	1	3	0	1	1	1	2	1	0	0	1	1	0	0	2	0	16	12
PC148	2	0	1	1	1	1	0	0	1	1	2	0	0	0	0	1	0	1	0	0	12	11
PC198	0	0	2	0	1	1	0	2	0	0	1	2	0	0	0	0	0	1	1	1	12	11
PC4	0	0	0	1	0	1	0	0	0	2	2	0	0	2	0	1	1	0	0	1	11	10
PC57	0	0	0	1	0	0	0	1	2	1	0	1	1	2	0	0	0	1	0	0	10	10
PC104	0	2	0	1	0	1	1	2	0	0	0	0	1	0	0	0	0	0	1	1	10	10
PC152	0	1	0	0	1	0	1	1	0	1	2	1	0	1	1	1	0	1	2	1	15	10
PC490	1	1	1	0	0	1	1	1	1	0	2	0	1	1	0	0	1	0	0	1	13	10

4.8.2 Models

Table 4-31 shows the trimmed Clinical risk index models for a selection of five random bootstrap samples. Figure 4-31 shows the distribution of risk index values in the optimization set for the Clinical risk index model from one randomly selected bootstrap sample (Bootstrap Sample #32), and Figure 4-32 shows the distribution of risk index values in the independent testing set for the Clinical risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #32). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as "high risk of developing diabetes" and those with a risk index value lower are predicted as "low risk of developing diabetes". Table 4-32 shows the risk index values for 25 randomly chosen individuals from the Independent Testing Set from these five Clinical risk index models along with that risk index model's prediction about each individual, where 0 indicates low risk of developing diabetes and 1 indicates high risk of developing diabetes.

Table 4-31 Five Randomly Selected Trimmed Clinical Risk Index Models for Incident Diabetes

Bootstrap Sample	Model
32	0.0304*Height + 0.0348*Systolic Blood Pressure + 0.0515*Diastolic Blood Pressure + 0.0093*Total Cholesterol - 0.0546*HDL + 0.0123*LDL + 0.0019*Triglycerides - 0.0663*Ever Smoked - 0.1709*Current Smoking Status - 0.0277*Weekly Alcohol Consumption + 0.0507*Marital Status
36	0.0763*Diastolic Blood Pressure - 0.0908*Marital Status
41	0.0326*Systolic Blood Pressure - 0.0574*HDL + 0.0017*Triglycerides + 0.2083*Ever Smoked + 0.2976*Current Smoking Status + 0.0343*Weekly Alcohol Consumption + 0.0125*Marital Status + 0.0941*Blood Glucose - 0.0238*Blood Urea Nitrogen + 0.0243*Serum Albumin - 4e-04*Serum Bilirubin + 0.0409*Serum Alkaline Phosphotase
90	0.2025*Marital Status
96	0.0256*Weight

Table 4-32 Risk Index Values for 25 Individuals from the Independent Testing Set from Five Randomly Selected Clinical Risk Index Models for Incident Diabetes

		Bootstrap Sample #32 Cutoff Value = 1.205		Bootstrap Sample #36 Cutoff Value = 3.724		Bootstrap Sample #41 Cutoff Value = 1.206		Bootstrap Sample #90 Cutoff Value=0.405		Bootstrap Sample #96 Cutoff Value=5.555	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	1.152	0	3.038	0	1.182	0	0.405	0	4.557	0
2	0	1.063	0	3.038	0	1.118	0	0.405	0	4.352	0
3	1	1.398	1	3.953	1	1.344	1	0.405	0	4.813	0
4	0	1.013	0	3.495	0	0.977	0	0.405	0	3.098	0
5	0	1.288	1	3.572	0	1.274	1	0.405	0	4.454	0
6	0	1.350	1	3.190	0	1.234	1	0.405	0	5.018	0
7	1	1.312	1	3.343	0	1.304	1	0.405	0	5.478	0
8	0	0.741	0	2.122	0	0.864	0	0.405	0	3.277	0
9	0	0.799	0	2.732	0	0.857	0	0.405	0	3.558	0
10	0	1.162	0	2.809	0	1.179	0	0.405	0	4.582	0
11	0	1.066	0	3.266	0	1.263	1	0.405	0	4.275	0
12	1	1.271	1	3.266	0	1.217	1	0.405	0	3.789	0
13	0	1.252	1	3.114	0	1.713	1	0.405	0	4.275	0
14	0	1.116	0	3.572	0	0.971	0	0.405	0	4.019	0
15	0	0.987	0	2.198	0	0.921	0	0.405	0	3.251	0
16	0	1.003	0	2.809	0	0.952	0	0.405	0	3.507	0
17	0	0.632	0	2.031	0	0.722	0	0.810	1	2.918	0
18	0	0.612	0	2.549	0	0.765	0	0.203	0	3.533	0
19	0	0.860	0	2.732	0	0.981	0	0.405	0	3.558	0
20	0	1.121	0	3.266	0	1.125	0	0.405	0	4.710	0
21	0	1.404	1	3.495	0	1.269	1	0.405	0	5.990	1
22	0	1.267	1	3.038	0	1.107	0	0.405	0	6.323	1
23	0	0.856	0	2.885	0	0.782	0	0.405	0	3.226	0
24	1	1.010	0	3.023	0	1.224	1	0.810	1	4.659	0
25	0	1.050	0	3.343	0	0.938	0	0.405	0	2.970	0

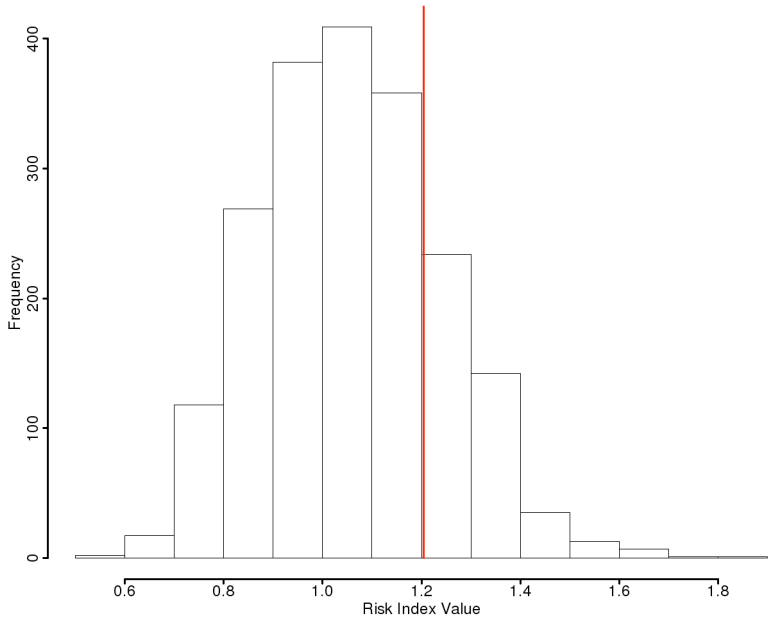


Figure 4-31 Distribution of Risk Index Values in the Optimization Set from the Clinical Risk Index Model for Bootstrap Sample #32

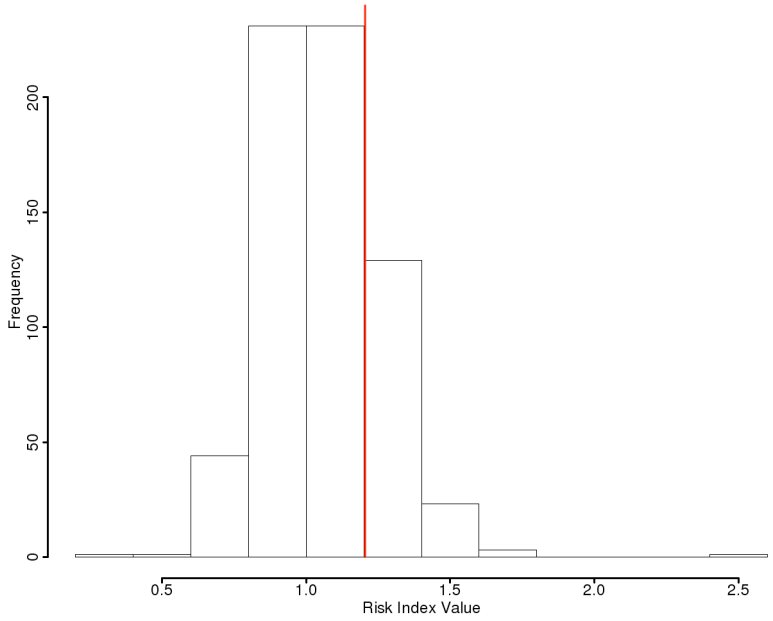


Figure 4-32 Distribution of Risk Index Values in the Independent Testing Set from the Clinical Risk Index Model for Bootstrap Sample #32

Table 4-33 shows the trimmed Clinical + Genotype risk index models for the same set of five bootstraps shown in Table 4-31, and Table 4-34 shows the risk index values and predictions of the same 25 randomly chosen individuals for the 5 bootstrap samples shown in Table 4-32. Figure 4-33 shows the distribution of risk index values in the optimization set for the Clinical + Genotype risk index model from one randomly selected bootstrap sample (Bootstrap Sample #32). Figure 4-34 shows the distribution of risk index values in the independent testing set for the Clinical + Genotype risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #32). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as "high risk of developing diabetes" and those with a risk index value lower are predicted as "low risk of developing diabetes".

Table 4-33 Five Randomly Selected Clinical + Genotype Risk Index Models for Incident Diabetes

Bootstrap Sample	Model
32	0.0304*Height + 0.0348*Systolic Blood Pressure + 0.0515*Diastolic Blood Pressure + 0.0093*Total Cholesterol - 0.0546*HDL + 0.0123*LDL + 0.0019*Triglycerides - 0.0663*Ever Smoked - 0.1709*Current Smoking Status - 0.0277*Weekly Alcohol Consumption + 0.0507*Marital Status + 3.9359*PC3 + 9.9229*PC29 + 0.5186*PC71 + 2.3951*PC87 - 0.8883*PC95 + 1.7747*PC100 + 0.2637*PC160 - 1.0093*PC225 - 2.0324*PC250 + 10.1536*PC423 + 0.4997*PC490 - 6.6034*PC491
36	0.0763*Diastolic Blood Pressure - 0.0908*Marital Status + 8.3048*PC58 + 4.5072*PC75 - 1.3113*PC91 - 2.5058*PC276 - 0.2329*PC336 + 3.5587*PC382 - 1.4082*PC406 + 12.8929*PC487 - 0.1383*PC490
41	0.0326*Systolic Blood Pressure - 0.0574*HDL + 0.0017*Triglycerides + 0.2083*Ever Smoked + 0.2976*Current Smoking Status + 0.0343*Weekly Alcohol Consumption + 0.0125*Marital Status + 0.0941*Blood Glucose - 0.0238*Blood Urea Nitrogen + 0.0243*Serum Albumin - 4e-04*Serum Bilirubin + 0.0409*Serum Alkaline Phosphatase + 1.4758*PC11 - 6.9019*PC12 - 5.9903*PC199 + 2.8352*PC215 + 2.5872*PC237 + 13.172*PC248 - 0.2214*PC376 - 0.8552*PC388 - 2.2939*PC476 + 15.5728*PC480 + 3.9395*PC484
90	0.2025*Marital Status + 1.2357*PC10 + 2.4205*PC20 - 0.3959*PC24 - 0.0429*PC32 + 2.2258*PC51 - 0.4129*PC54 + 0.2607*PC95 - 7.282*PC130 + 1.5002*PC136 - 0.0582*PC146 - 2.134*PC148 - 0.533*PC152 - 2.2258*PC156 - 0.2734*PC160 + 2.0539*PC162 - 0.1559*PC184 + 1.5615*PC233 + 0.6637*PC316 + 1.6871*PC318 + 2.0245*PC382
96	0.0256*Weight + 3.3456*PC50 + 0.14*PC57 + 0.8653*PC59 + 2.0555*PC98 + 1.3727*PC127 + 0.1076*PC128 + 13.5397*PC140 - 0.6265*PC148 + 1.7555*PC178 + 5.8116*PC245 - 0.319*PC338 + 14.536*PC343 + 1.674*PC417 - 1.0469*PC442 + 0.8239*PC453 + 8.2504*PC489 + 0.8037*PC497

Table 4-34 Risk Index Values for 25 Individuals in the Independent Testing Set for the Clinical + Genotype Risk Index Models from Five Randomly Selected Bootstrap Samples for Incident Diabetes

		Bootstrap Sample #32 Cutoff Value = 1.205		Bootstrap Sample #36 Cutoff Value = 3.761		Bootstrap Sample #41 Cutoff Value = 1.209		Bootstrap Sample #90 Cutoff Value=0.416		Bootstrap Sample #96 Cutoff Value=5.571	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	1.150	0	3.070	0	1.160	0	0.405	0	4.584	0
2	0	1.061	0	3.043	0	1.205	0	0.404	0	4.346	0
3	1	1.336	1	3.927	1	1.379	1	0.394	0	4.810	0
4	0	1.004	0	3.470	0	1.009	0	0.413	0	3.128	0
5	0	1.260	1	3.569	0	1.237	1	0.384	0	4.483	0
6	0	1.309	1	3.207	0	1.289	1	0.404	0	4.980	0
7	1	1.348	1	3.337	0	1.269	1	0.404	0	5.474	0
8	0	0.755	0	2.108	0	0.936	0	0.414	0	3.249	0
9	0	0.819	0	2.730	0	0.903	0	0.400	0	3.578	0
10	0	1.162	0	2.789	0	1.192	0	0.400	0	4.575	0
11	0	1.060	0	3.268	0	1.224	1	0.406	0	4.275	0
12	1	1.296	1	3.293	0	1.226	1	0.397	0	3.803	0
13	0	1.282	1	3.066	0	1.713	1	0.416	0	4.268	0
14	0	1.116	0	3.528	0	0.953	0	0.415	0	4.017	0
15	0	1.013	0	2.220	0	0.920	0	0.406	0	3.295	0
16	0	1.012	0	2.807	0	0.940	0	0.400	0	3.505	0
17	0	0.721	0	2.016	0	0.730	0	0.818	1	2.856	0
18	0	0.599	0	2.519	0	0.693	0	0.188	0	3.522	0
19	0	0.854	0	2.676	0	0.947	0	0.396	0	3.615	0
20	0	1.148	0	3.259	0	1.126	0	0.407	0	4.687	0
21	0	1.410	1	3.413	0	1.277	1	0.414	0	5.945	1
22	0	1.261	1	3.025	0	1.063	0	0.399	0	6.350	1
23	0	0.851	0	2.925	0	0.774	0	0.399	0	3.193	0
24	1	0.988	0	3.070	0	1.242	1	0.814	1	4.676	0
25	0	1.075	0	3.329	0	0.906	0	0.420	1	2.978	0

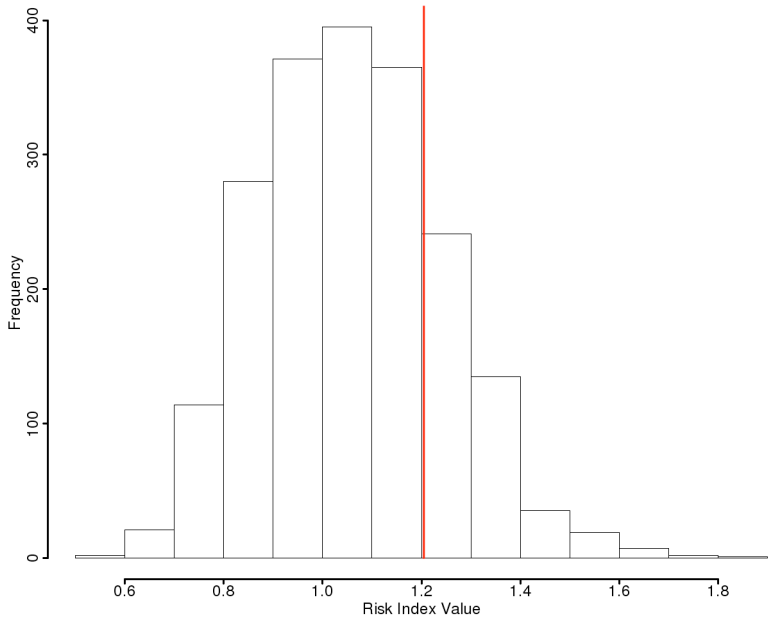


Figure 4-33 Distribution of Risk Index Values in the Optimization Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #32

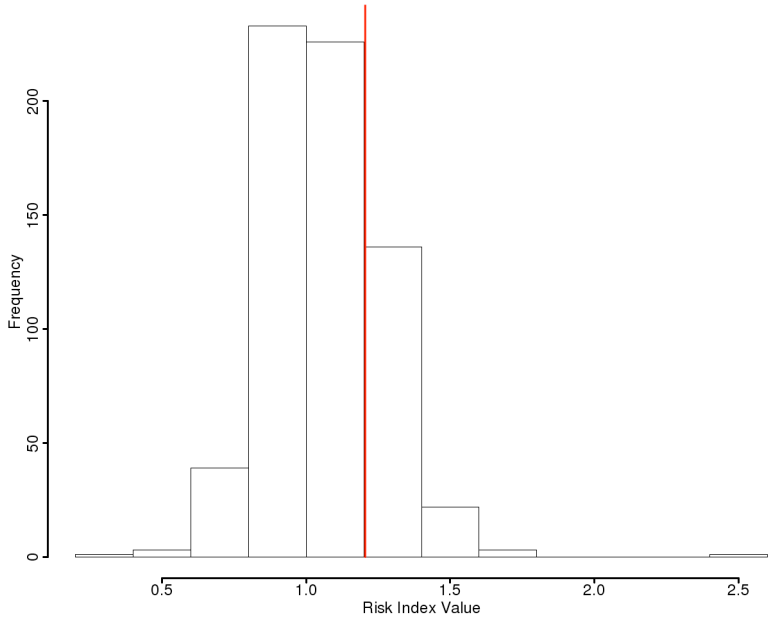


Figure 4-34 Distribution of Risk Index Values in the Independent Testing Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #32

4.8.3 Predictive Performance

Once predictions were made for each individual in the independent testing set the sensitivity, specificity, misclassification, and positive predictive value were calculated for the Clinical risk index model and the Clinical + Genotype risk index model as described in Section 4.5.3. The estimates and confidence intervals for sensitivity, specificity, misclassification, and positive predictive value are given in Table 4-35. Lastly, using the individual predictions from each of the 100 trimmed Clinical risk index models and 100 trimmed Clinical + Genotype risk index models for the individuals in the independent testing set, ROC curves were generated, and the AUC for the ROC curve was estimated for both the Clinical and Clinical + Genotype risk index models (Figure 4-35, Figure 4-36). For the Clinical risk index model the AUC for the ROC curve was 0.768, and for the Clinical + Genotype risk index model the AUC for the ROC curve was 0.782.

Table 4-35 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for Ten-year Incident Diabetes

Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
Clinical	0.204 (0.1-0.322)	0.959 (0.943-0.975)	0.102 (0.078-0.125)	0.306 (0.15-0.469)
Clinical + Genotype	0.13 (0.043-0.224)	0.974 (0.96-0.987)	0.095 (0.074-0.119)	0.304 (0.118-0.5)

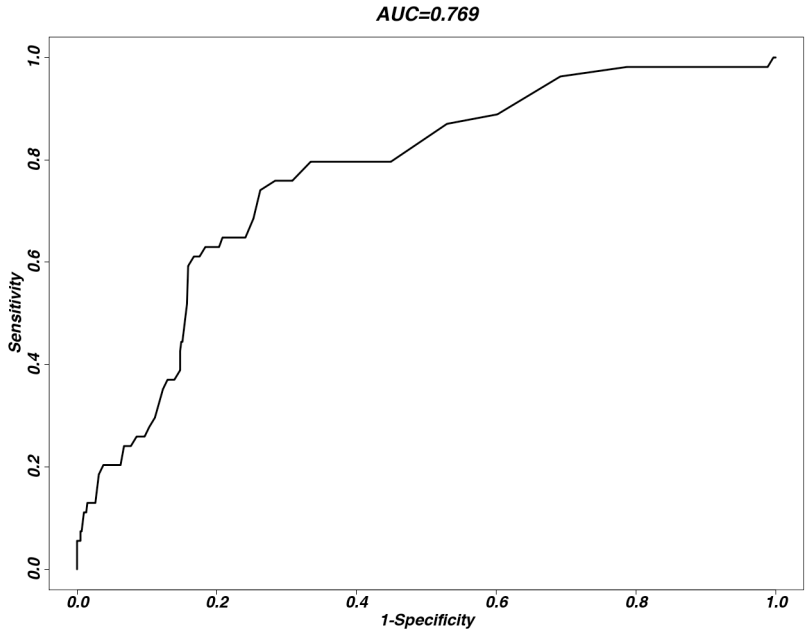


Figure 4-35 ROC Curve and AUC for the Incident Diabetes Clinical Risk Index Model

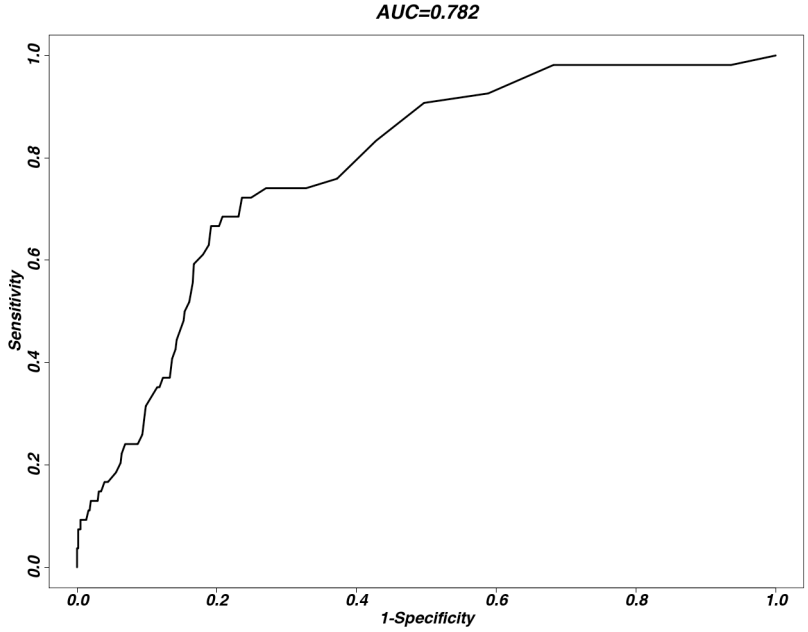


Figure 4-36 ROC Curve and AUC for the Incident Diabetes Clinical + Genotype Risk Index Model

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. The proportion of models which predict that an individual is at high risk of developing diabetes, then, represents the predicted probability of an individual developing diabetes. Section 4.5.3 describes the calculation of a confidence interval for this predicted probability.

Figure 4-37 shows the distribution of the predicted probability of developing diabetes for the Clinical risk index model in the independent testing set, and Figure 4-38 shows the distribution of the predicted probability of developing diabetes for the Clinical + Genotype risk index model in the independent testing set. In both Figures, a density line is shown on the graph to indicate the density of a normal distribution with the mean and standard deviation matching that of the predicted probability distribution.

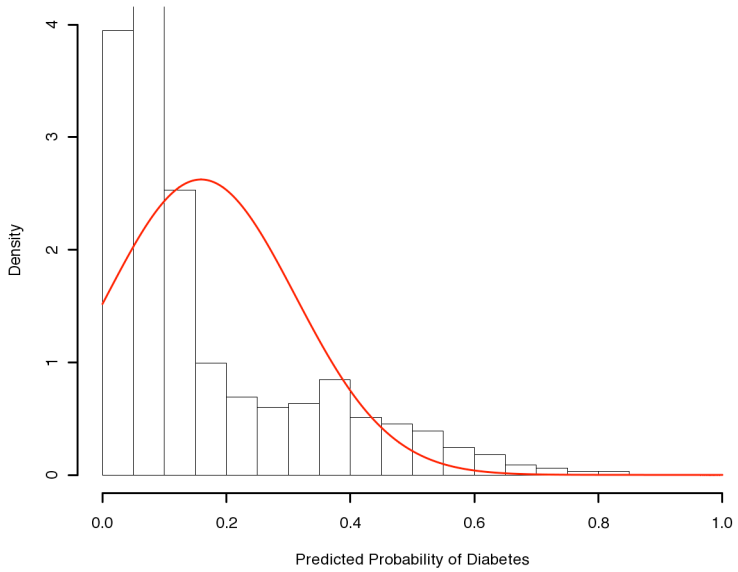


Figure 4-37 Histogram of the Predicted Probability of Developing Diabetes for the Clinical Risk Index Model

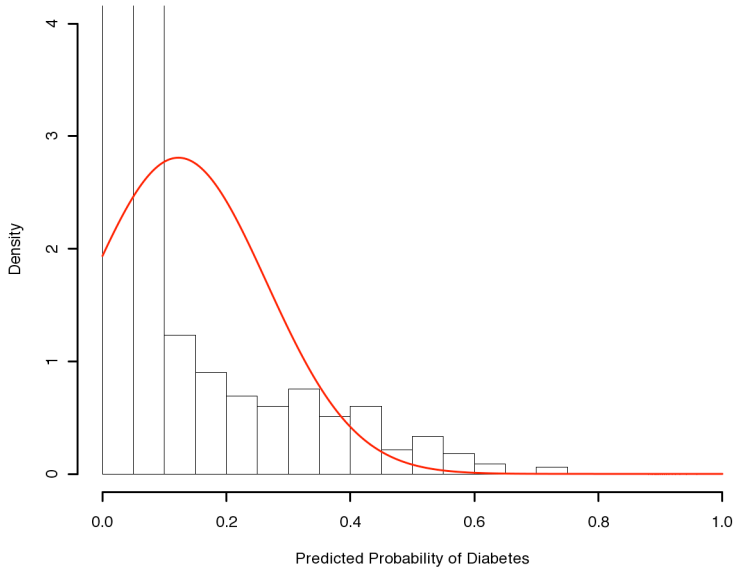


Figure 4-38 Histogram of the Predicted Probability of Developing Diabetes for the Clinical + Genotype Risk Index Model

4.8.4 Random Forests Comparison

A random forest was generated using the optimization set created by the risk index procedure. The forest had 500 individual trees, and the tuning procedure described in detail in Section 4.6.4 was used to find the number of variables *k* considered at each split that provided the lowest out-of-bag error estimate. The optimized *k* chosen was 45, which gave an out-of-bag error estimate of 8.36%.

When working with a dataset that has two possible classes, the standard procedure for a random forest is to assign a prediction to an individual based on a simple majority of votes, when the prevalence of the outcome is less than 50% changing the proportion of

votes needed to classify an individual can significantly impact the estimates of performance. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set on a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of 5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the individual to be "high risk". Table 4-36 shows the results of this investigation.

Predictions were then made about each individual in the independent testing set created by the risk index procedure, and the sensitivity, specificity, misclassification, and positive predictive value of the predictions was assessed. One thousand bootstrap samples of the independent testing set were generated, and predictions were made about each individual in each of the bootstrap samples. This data was used to create 95% confidence intervals for the sensitivity, specificity, misclassification, and positive predictive value estimates. Table 4-37 shows the sensitivity, specificity, misclassification, and positive predictive value estimates for the random forest as well as the 95% confidence interval for each estimate. Lastly, using the class votes for the individuals in the independent testing set, an ROC curve was created, and the AUC for the ROC curve was estimated (Figure 4-39).

Table 4-36 Performance Estimates of the Random Forest

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	0.941	0.337	0.611	0.119
0.1	0.765	0.624	0.364	0.162
0.15	0.510	0.810	0.216	0.203
0.2	0.294	0.918	0.136	0.254
0.25	0.176	0.963	0.105	0.310
0.3	0.098	0.983	0.094	0.357
0.35	0.039	0.991	0.092	0.286
0.4	0.000	0.993	0.094	0.000
0.45	0.000	0.998	0.088	0.000
0.5	0.000	0.998	0.088	0.000
0.55	0.000	1.000	0.087	-
0.6	0.000	1.000	0.087	-
0.65	0.000	1.000	0.087	-
0.7	0.000	1.000	0.087	-
0.75	0.000	1.000	0.087	-
0.8	0.000	1.000	0.087	-
0.85	0.000	1.000	0.087	-
0.9	0.000	1.000	0.087	-
0.95	0.000	1.000	0.087	-

Table 4-37 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for the Random Forest Model

Proportion of Votes for "High Risk" Class	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
0.1	0.765 (0.646-0.877)	0.624 (0.582-0.664)	0.364 (0.325-0.401)	0.162 (0.117-0.210)
0.15	0.510 (0.370-0.655)	0.810 (0.777-0.843)	0.216 (0.182-0.247)	0.203 (0.133-0.277)
0.2	0.294 (0.170-0.421)	0.918 (0.894-0.940)	0.136 (0.109-0.165)	0.254 (0.143-0.369)
0.25	0.176 (0.081-0.288)	0.963 (0.945-0.978)	0.105 (0.082-0.129)	0.310 (0.154-0.484)
0.3	0.098 (0.021-0.188)	0.983 (0.972-0.993)	0.094 (0.071-0.117)	0.357 (0.111-0.667)
0.35	0.039 (0.000-0.102)	0.991 (0.981-0.998)	0.092 (0.070-0.116)	0.286 (0.000-0.714)

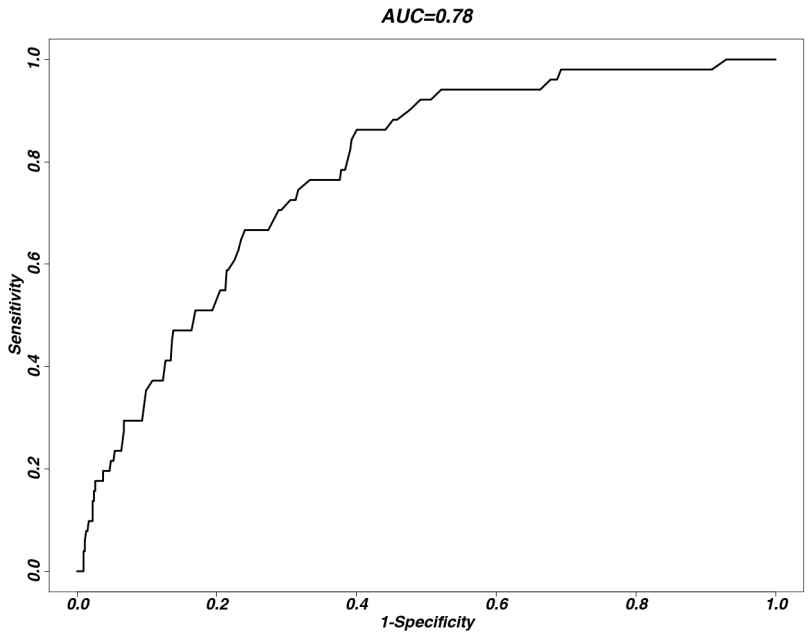


Figure 4-39 ROC Curve and AUC for the Incident Hypertension PCA Random Forest Model

4.8.5 Conclusion

As with the risk index models for ten-year incident diabetes built using the 500 most highly associated SNPs, both the Clinical and Clinical + Genotype risk index models had very high specificities but very low sensitivities. However, as was observed for the 10-year incident hypertension analyses, the AUC for both is higher than the AUCs achieved by the Clinical and Clinical + Genotype risk index models in the ten-year incident diabetes analysis using the 500 most highly associated SNPs. Additionally, the AUC for the Clinical + Genotype model is basically equivalent to the AUC for the random forests model. The ROC curves for each of the risk index models also exhibit a greater curve than that seen in the incident hypertension analysis. This suggests that modifying the

proportion of votes required to give a prediction of "high risk" might be able to increase the predictive performance of the models.

4.9 Prevalent Hypertension Using 500 Most Highly Associated SNPs

4.9.1 Variable Selection

Using the procedure described in Section 4.5.1, the risk index procedure was performed to predict risk of developing diabetes within a ten-year time frame. The 500 SNPs most highly associated with this outcome (i.e., which had the lowest p-values from a logistic regression analysis of this outcome) were identified and used to build the risk index.

Table 4-38 gives a summary of the order in which variables were selected for the 100 bootstrap samples of the optimization set used to build the Clinical risk index model. Sixt-two out of the 100 trimmed Clinical risk index models contained age as a variable, and 46 contained weight. Marital status was included in 46 out of the 100 trimmed Clinical risk index models. Serum albumin levels, height, and blood glucose levels, were also frequently included in the set of 100 trimmed Clinical risk index models, appearing in 38, 35, and 33 trimmed Clinical risk index models, respectively. Table 4-39 gives a summary of the order in which the most commonly chosen principal components were selected into the 100 Clinical + Genotype risk index models. Each SNP in Table 4-39 appears in at least 25 trimmed Clinical + Genotype risk index models.

Table 4-38 Summary of Number of Times Each Variable is Selected into a Specific Model Position for the Clinical Risk Index Model for Prevalent Hypertension

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
Age (yrs)	46	8	5	4	7	3	3	4	3	0	8	2	1	0	3	0	1	1	1	0	100	62
Sex	1	7	3	7	9	10	5	6	7	6	1	3	5	2	4	4	3	4	6	3	96	18
Weight (lbs)	3	14	5	1	2	7	14	9	5	6	2	3	1	3	4	4	3	6	6	1	99	24
Height (in)	2	12	11	11	10	5	14	9	8	6	7	2	0	1	0	0	1	0	1	0	100	35
Systolic Blood Pressure (mm Hg)	0	0	1	2	1	2	3	4	9	7	5	2	3	4	3	10	10	9	9	8	92	6
Diastolic Blood Pressure (mm Hg)	0	0	0	1	0	2	2	5	6	14	10	6	8	6	8	5	7	13	4	1	98	4
Total Cholesterol (mg/dL)	0	0	0	0	0	0	2	1	1	8	7	8	6	7	6	5	9	12	11	6	89	0
High-density Lipoprotein Level (mg/dL)	5	1	1	2	1	0	0	2	3	6	7	13	9	4	7	9	9	9	5	4	97	10
Low-density Lipoprotein Level (mg/dL)	0	15	12	12	7	9	7	9	4	3	4	5	5	1	0	6	1	0	0	0	100	31
Triglycerides (mg/dL)	4	5	12	12	6	10	5	3	2	3	4	4	6	11	4	0	0	4	1	2	98	24
Ever Smoked	0	3	5	9	9	9	3	6	0	7	6	4	9	15	5	5	1	0	1	2	99	23
Currently Smokes	5	13	15	11	16	9	9	9	1	1	2	1	0	4	2	0	1	1	0	0	100	46
Weekly Alcohol Consumption	2	3	3	4	5	6	2	5	2	2	2	0	2	2	5	4	5	8	6	18	86	8
Marital Status	1	3	3	1	1	2	4	2	6	2	3	4	4	2	11	6	3	1	10	18	87	9
Left Ventricular Mass (g)	8	9	7	3	5	2	2	4	4	0	3	4	3	5	7	13	8	2	5	2	96	33
Left Ventricular Ejection Fraction (%)	0	0	1	4	2	0	3	2	4	4	8	8	6	7	7	11	13	8	4	4	96	8
Blood Glucose (mg/dL)	0	2	8	4	2	2	8	8	6	3	2	5	8	3	4	3	7	12	8	3	98	22
Blood Urea Nitrogen (mg/dL)	18	2	3	1	6	5	7	4	11	4	8	9	4	2	2	2	3	4	5	0	100	38

Total Serum Protein Level (mg/dL)	4	3	2	6	5	6	2	4	11	11	4	5	7	3	6	2	1	0	10	6	98	16
Serum Albumin Level (mg/dL)	1	0	2	2	3	5	4	1	4	1	4	4	5	8	8	5	7	4	3	18	89	8
Serum Bilirubin Level (mg/dL)	0	0	1	3	3	6	1	3	3	6	3	8	8	10	4	6	7	2	4	4	82	13
Serum Alkaline Phosphatase Level (mg/dL)	46	8	5	4	7	3	3	4	3	0	8	2	1	0	3	0	1	1	1	0	100	62
Serum Creatine Level (mg/dL)	1	7	3	7	9	10	5	6	7	6	1	3	5	2	4	4	3	4	6	3	96	18

Table 4-39 Summary of Number of Times Selected Genotype Variables are Selected into a Specific Model Position for the Clinical + Genotype Risk Index Model for Prevalent Hypertension

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
rs3087481	0	7	10	10	8	6	12	2	1	0	8	4	1	2	1	1	0	1	1	0	75	58
rs12024717	3	4	4	3	3	3	3	2	2	3	6	3	1	3	2	1	2	3	3	2	56	36
rs11880330	1	2	4	4	2	6	6	2	3	1	2	5	5	2	0	1	4	0	1	4	55	32
rs4302331	0	2	1	2	3	5	5	0	5	6	5	2	4	0	3	4	2	2	2	1	54	31
rs16858033	0	2	4	4	2	4	2	1	1	3	3	2	2	3	5	5	4	3	0	2	52	31
rs6844109	0	2	2	2	5	1	1	2	2	6	1	6	1	1	2	1	3	1	2	1	42	29
rs6475322	1	1	4	7	7	5	4	4	1	4	2	1	3	3	0	5	0	1	1	2	56	29
rs4768264	0	2	2	1	4	4	5	6	5	1	2	4	2	2	2	5	2	2	1	1	53	28
rs12039283	1	1	0	3	5	3	3	2	3	3	5	4	2	2	2	2	1	2	2	2	48	27
rs3006870	0	1	2	2	2	2	4	3	3	2	7	1	1	2	3	1	1	2	3	1	43	25
rs4986893	0	1	0	2	3	5	1	9	1	4	1	6	4	0	1	3	4	3	1	2	51	25

4.9.2 Models

Table 4-40 shows the trimmed Clinical risk index models for a selection of five random bootstrap samples. Figure 4-40 shows the distribution of risk index values in the optimization set for the Clinical risk index model from one randomly selected bootstrap sample (Bootstrap Sample #27), and Figure 4-41 shows the distribution of risk index values in the independent testing set for the Clinical risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #27). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as "high risk of having hypertension" and those with a risk index value lower are predicted as "low risk of having hypertension". Table 4-41 shows the risk index values for 25 randomly chosen individuals from the Independent Testing Set from these five Clinical risk index models along with that risk index model's prediction about each individual, where 0 indicates low risk of having hypertension and 1 indicates high risk of having hypertension.

Table 4-40 Five Randomly Selected Trimmed Clinical Risk Index Models for Prevalent Hypertension

Bootstrap Sample	Model
27	0.0995*Age - 0.0519*Ever Smoked + 0.0365*Marital Status
38	0.0043*Height - 0.0226*Ever Smoked + 0.0517*Serum Albumin
44	0.1061*Age + 0.1173*Marital Status
63	0.1131*Age - 0.3159*Sex + 0.0455*Marital Status + 0.0458*Blood Glucose + 0.0494*Serum Alkaline Phosphatase + 0.0197*Serum Creatine
82	0.0019*Height + 0.012*Marital Status

Table 4-41 Risk Index Values for 25 Individuals from the Independent Testing Set from Five Randomly Selected Clinical Risk Index Models for Prevalent Hypertension

		Bootstrap Sample #27 Cutoff Value = 1.765		Bootstrap Sample #38 Cutoff Value = 0.901		Bootstrap Sample #44 Cutoff Value = 2.823		Bootstrap Sample #63 Cutoff Value=2.023		Bootstrap Sample #82 Cutoff Value=0.082	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	1.756	0	0.822	0	2.993	1	1.788	0	0.083	1
2	1	1.931	1	0.881	0	3.194	1	1.941	0	0.076	0
3	0	1.665	0	0.910	1	2.770	0	1.767	0	0.072	0
4	0	1.557	0	0.828	0	2.675	0	1.738	0	0.087	1
5	0	1.650	0	0.850	0	2.717	0	1.971	0	0.073	0
6	0	1.500	0	0.848	0	2.505	0	1.828	0	0.077	0
7	1	1.815	1	0.874	0	2.982	1	1.935	0	0.066	0
8	0	1.201	0	0.841	0	2.027	0	1.448	0	0.072	0
9	0	1.433	0	0.804	0	2.398	0	1.594	0	0.070	0
10	0	1.325	0	0.756	0	2.304	0	1.602	0	0.085	1
11	0	0.990	0	0.792	0	1.650	0	1.293	0	0.068	0
12	1	1.732	0	0.791	0	2.876	1	1.802	0	0.073	0
13	0	1.997	1	0.840	0	3.300	1	2.026	1	0.071	0
14	1	1.201	0	0.914	1	2.027	0	1.675	0	0.075	0
15	0	1.384	0	0.793	0	2.292	0	1.637	0	0.069	0
16	0	1.400	0	0.851	0	2.345	0	1.601	0	0.079	0
17	0	1.367	0	0.809	0	2.292	0	1.597	0	0.074	0
18	1	1.138	0	0.878	0	2.044	0	1.372	0	0.092	1
19	0	1.152	0	0.910	1	1.921	0	1.679	0	0.079	0
20	0	1.318	0	0.810	0	2.186	0	1.424	0	0.069	0
21	0	1.716	0	0.817	0	2.823	0	2.344	1	0.074	0
22	0	0.941	0	0.896	0	1.544	0	1.505	0	0.074	0
23	0	1.002	0	0.811	0	1.709	0	1.374	0	0.075	0
24	0	1.500	0	0.916	1	2.505	0	1.980	0	0.076	0
25	0	1.599	0	0.965	1	2.664	0	1.853	0	0.074	0

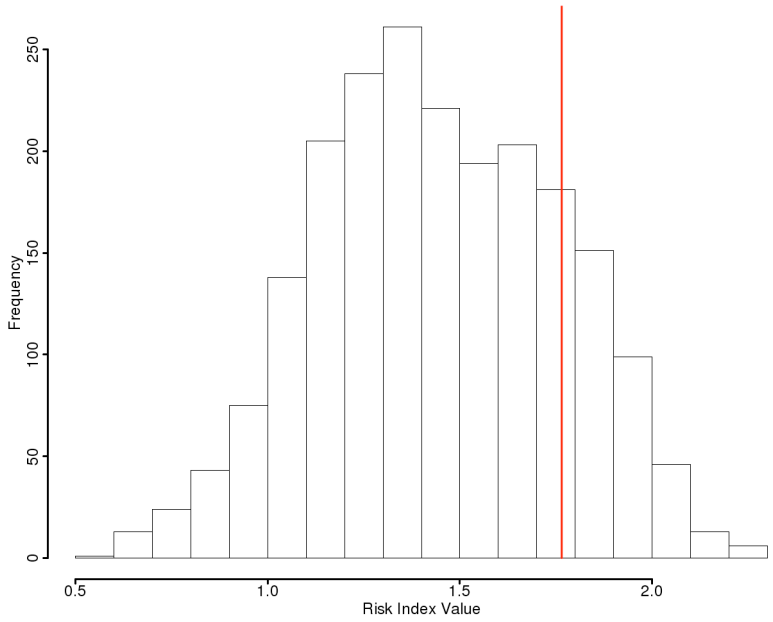


Figure 4-40 Distribution of Risk Index Values in the Optimization Set from the Clinical Risk Index Model for Bootstrap Sample #27

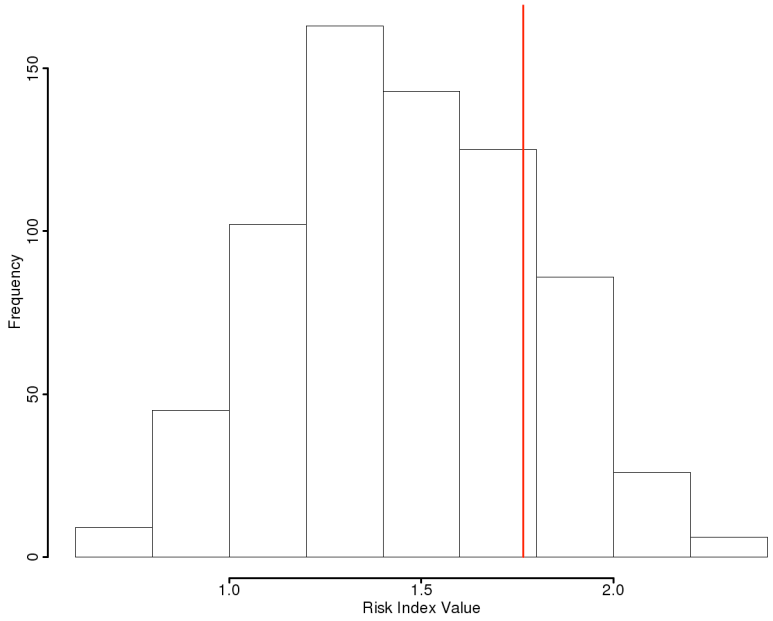


Figure 4-41 Distribution of Risk Index Values in the Independent Testing Set from the Clinical Risk Index Model for Bootstrap Sample #27

Table 4-42 shows the trimmed Clinical + Genotype risk index models for the same set of five bootstraps shown in Table 4-40, and Table 4-43 shows the risk index values and predictions of the same 25 randomly chosen individuals for the 5 bootstrap samples shown in Table 4-41. Figure 4-42 shows the distribution of risk index values in the optimization set for the Clinical + Genotype risk index model from one randomly selected bootstrap sample (Bootstrap Sample #27). Figure 4-43 shows the distribution of risk index values in the independent testing set for the Clinical + Genotype risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #27). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as “high risk of having hypertension” and those with a risk index value lower are predicted as “low risk of having hypertension”.

Table 4-42 Five Randomly Selected Clinical + Genotype Risk Index Models for Prevalent Hypertension

Bootstrap Sample	Model
27	0.0995*Age - 0.0519*Ever Smoked + 0.0365*Marital Status + rs9994289(C_T=-0.2157, T_T=-0.6068) + rs17031297(C_C=1.4318)
38	0.0043*Height - 0.0226*Ever Smoked + 0.0517*Serum Albumin + rs13256239(A_C=0.6367, C_C=1.0273) + rs13242(C_T=1.049, T_T=18.9896) + rs940224(A_G=-0.9546, G_G=-1.2444) + rs4302331(G_G=2.181) + rs5491(A_T=-2.6136) + rs16858033(A_G=-3.4422) + rs4507748(A_G=-0.8966, NA) + rs3087481(G_G=3.8066, NA) + rs289059(C_T=-0.2222, T_T=-0.3453) + rs12024717(C_T=-1.9, T_T=-32.1701) + rs17439459(C_T=0.8757, T_T=1.0827) + rs7945609(C_T=-0.1973, T_T=-1.4705) + rs2291256(C_T=0.3162, T_T=-1.1085) + rs9976886(A_C=0.4639, C_C=0.6816) + rs1510955(A_G=-0.4285, G_G=-0.6372) + rs17867624(G_T=-0.3673, T_T=0.3673) + rs3802384(G_G=0.0639) + rs2269714(C_T=-0.044, T_T=8942.447) + rs11703393(A_G=-0.0554, G_G=0.0624) + rs2915400(C_T=0.2492, T_T=-0.0072)
44	0.1061*Age + 0.1173*Marital Status + rs2292664(G_G=0.3336) + rs7729495(T_T=-0.5198) + rs3087481(G_G=2.5349) + rs45497698(A_G=-642.1136, G_G=-641.8706) + rs11705259(A_G=10.7366, G_G=11.028) + rs17867624(G_T=-0.0562, T_T=0.0562) + rs2269714(C_T=-0.0916, T_T=-0.0537)
63	0.1131*Age - 0.3159*Sex + 0.0455*Marital Status + 0.0458*Blood Glucose + 0.0494*Serum Alkaline Phosphatase + 0.0197*Serum Creatine + rs2961944(A_G=0.2217, G_G=0.3388) + rs16858033(A_G=-0.7358) + rs12510552(C_G=0.0464, G_G=0.0584) + rs6475322(A_C=0.3663)
82	0.0019*Height + 0.012*Marital Status + rs2292664(G_G=2.1354) + rs9497762(C_T=-0.5166, T_T=-19.6005) + rs11003001(C_G=0.653, G_G=0.7901) + rs13395300(A_G=-0.3285, G_G=-0.4343) + rs2171497(C_G=-0.2303, G_G=-0.5966) + rs6512087(C_T=-1.7552, T_T=-1.8575) + rs17354559(C_G=1.5243, G_G=26.314) + rs12039283(G_G=1.664) + rs3006870(A_G=-2.009) + rs4768264(G_G=1.5142) + rs1867435(C_C=1.5142) + rs4768268(C_C=1.5142) + rs2493151(A_G=0.4256, G_G=0.6677) + rs17031297(C_C=0.4971) + rs9976886(A_C=0.3079, C_C=0.4395) + rs6869755(C_T=0.4382, T_T=0.2873) + rs11880330(A_G=-1.6592) + rs10521004(C_T=0.1105, T_T=0.4325) + rs10111520(G/T=1.1709, T/T=0.7244) + rs3827760(A_G=-0.4295)

Table 4-43 Risk Index Values for 25 Individuals in the Independent Testing Set for the Clinical + Genotype Risk Index Models from Five Randomly Selected Bootstrap Samples for Prevalent Hypertension

		Bootstrap Sample #27 Cutoff Value = 2.414		Bootstrap Sample #38 Cutoff Value = 1.248		Bootstrap Sample #44 Cutoff Value = -86.954		Bootstrap Sample #63 Cutoff Value=2.095		Bootstrap Sample #82 Cutoff Value=0.584	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	2.472	1	1.026	0	-86.784	1	1.843	0	0.507	0
2	1	2.539	1	1.237	0	-86.583	1	2.008	0	0.536	0
3	0	2.273	0	1.252	1	-87.007	0	1.852	0	0.476	0
4	0	2.273	0	1.096	0	-87.102	0	1.822	0	0.563	0
5	0	2.365	0	1.127	0	-87.060	0	2.026	0	0.544	0
6	0	2.215	0	1.181	0	-87.272	0	1.828	0	0.565	0
7	1	2.531	1	1.179	0	-86.808	1	1.991	0	0.497	0
8	0	1.809	0	1.226	0	-87.750	0	1.515	0	0.654	1
9	0	2.041	0	1.141	0	-87.391	0	1.649	0	0.563	0
10	0	1.933	0	0.953	0	-87.473	0	1.658	0	0.497	0
11	0	1.706	0	1.160	0	-88.127	0	1.360	0	0.504	0
12	1	2.340	0	1.097	0	-86.901	1	1.857	0	0.514	0
13	0	2.410	0	1.258	1	-86.477	1	2.093	0	0.534	0
14	1	1.917	0	1.314	1	-87.750	0	1.675	0	0.630	1
15	0	2.100	0	1.111	0	-87.484	0	1.722	0	0.535	0
16	0	2.008	0	1.151	0	-87.431	0	1.686	0	0.569	0
17	0	1.975	0	1.155	0	-87.484	0	1.694	0	0.516	0
18	1	1.854	0	1.255	1	-87.733	0	1.428	0	0.568	0
19	0	1.868	0	1.254	1	-87.869	0	1.776	0	0.592	1
20	0	1.926	0	1.164	0	-87.591	0	1.523	0	0.539	0
21	0	2.324	0	1.190	0	-86.954	0	2.400	1	0.591	1
22	0	1.657	0	1.289	1	-88.233	0	1.505	0	0.552	0
23	0	1.718	0	1.196	0	-88.103	0	1.470	0	0.552	0
24	0	1.912	0	1.264	1	-87.314	0	2.064	0	0.534	0
25	0	2.315	0	1.223	0	-87.113	0	1.909	0	0.642	1

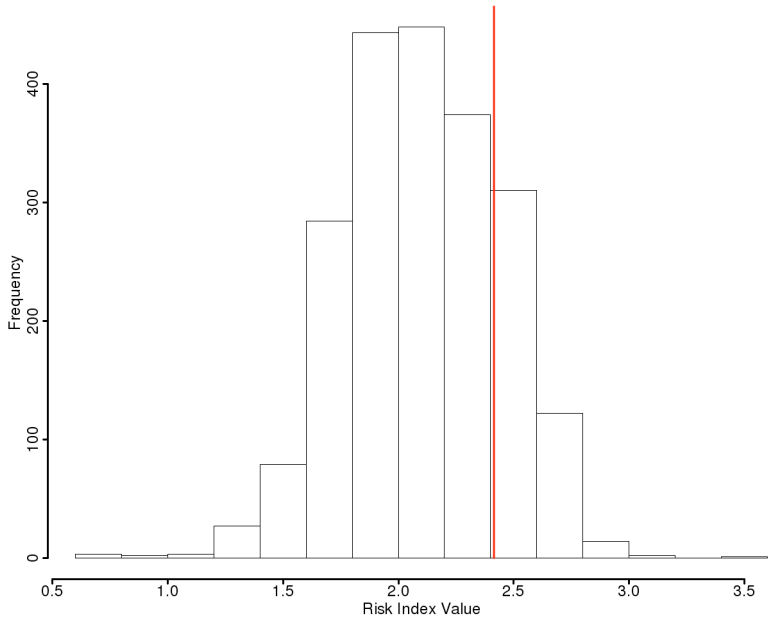


Figure 4-42 Distribution of Risk Index Values in the Optimization Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #27

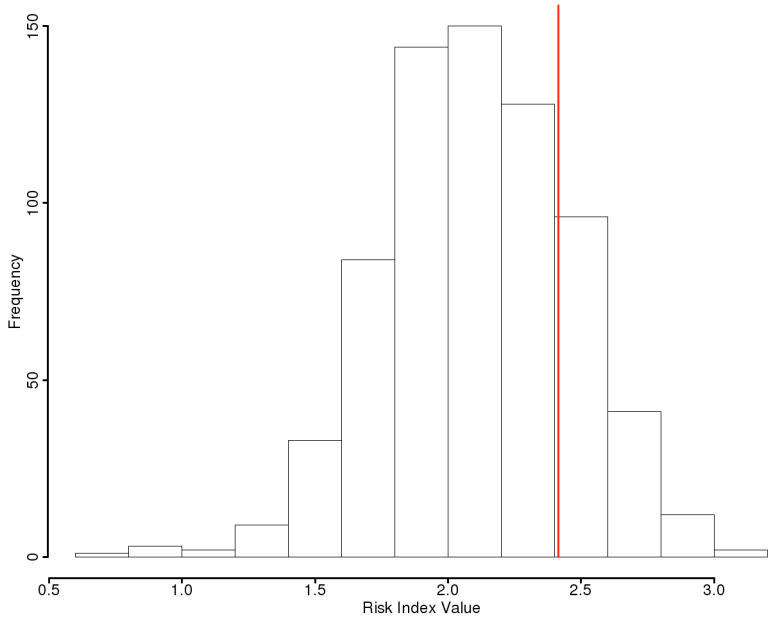


Figure 4-43 Distribution of Risk Index Values in the Independent Testing Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #

4.9.3 Predictive Performance

Once predictions were made for each individual in the independent testing set the sensitivity, specificity, misclassification, and positive predictive value were calculated for the Clinical risk index model and the Clinical + Genotype risk index model as described in Section 4.5.3. The estimates and confidence intervals for sensitivity, specificity, misclassification, and positive predictive value are given in Table 4-44. Lastly, using the individual predictions from each of the 100 trimmed Clinical risk index models and 100 trimmed Clinical + Genotype risk index models for the individuals in the independent testing set, ROC curves were generated, and the AUC for the ROC curve was estimated for both the Clinical and Clinical + Genotype risk index models (Figure 4-44, Figure 4-45). For the Clinical risk index model the AUC for the ROC curve was 0.733, and for the Clinical + Genotype risk index model the AUC for the ROC curve was 0.692.

Table 4-44 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for Prevalent Hypertension

Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
Clinical	0.459 (0.375-0.547)	0.844 (0.816-0.875)	0.228 (0.197-0.26)	0.407 (0.326-0.489)
Clinical + Genotype	0.323 (0.245-0.405)	0.846 (0.817-0.875)	0.252 (0.22-0.285)	0.328 (0.25-0.413)

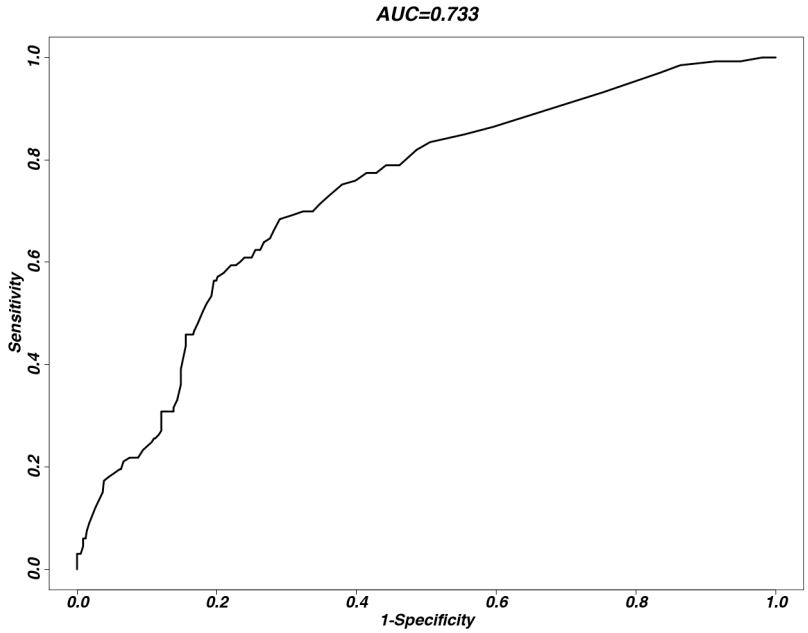


Figure 4-44 ROC Curve and AUC for the Incident Diabetes Clinical Risk Index Model

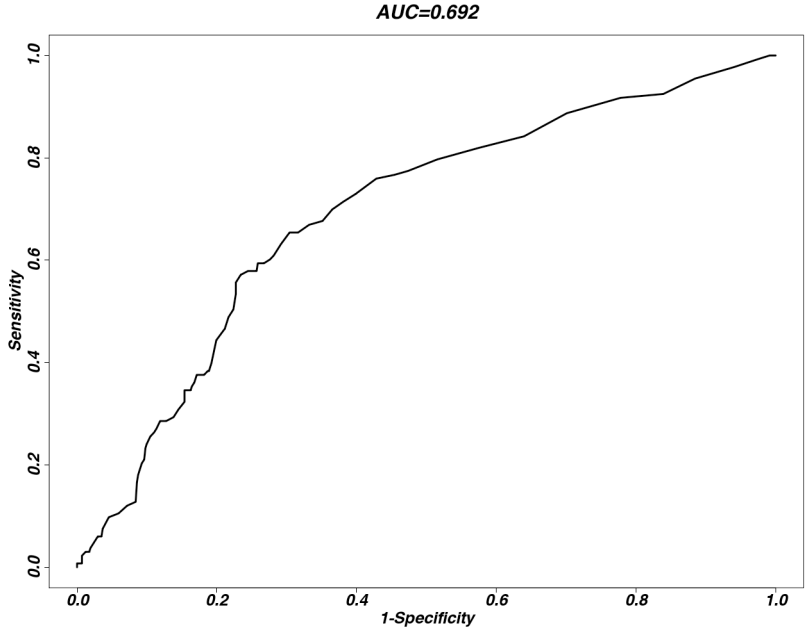


Figure 4-45 ROC Curve and AUC for the Incident Diabetes Clinical + Genotype Risk Index Model

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. The proportion of models that predict that an individual is at high risk of having hypertension, then, represents the predicted probability of an individual having hypertension. Section 4.5.3 describes the calculation of a confidence interval for this predicted probability.

Figure 4-46 shows the distribution of the predicted probability of having hypertension for the Clinical risk index model in the independent testing set, and Figure 4-47 shows the distribution of the predicted probability of having hypertension for the Clinical + Genotype risk index model in the independent testing set. In both Figures, a density line is shown on the graph to indicate the density of a normal distribution with the mean and standard deviation matching that of the predicted probability distribution.

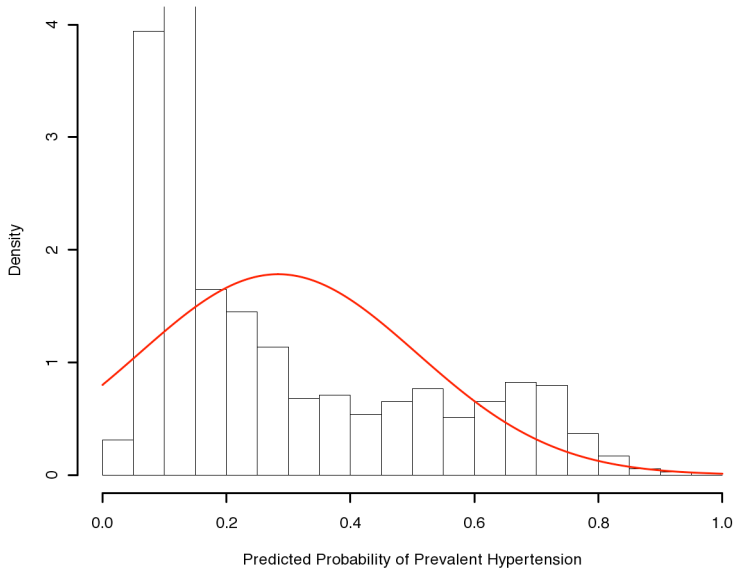


Figure 4-46 Histogram of the Predicted Probability of Developing Diabetes for the Clinical Risk Index Model

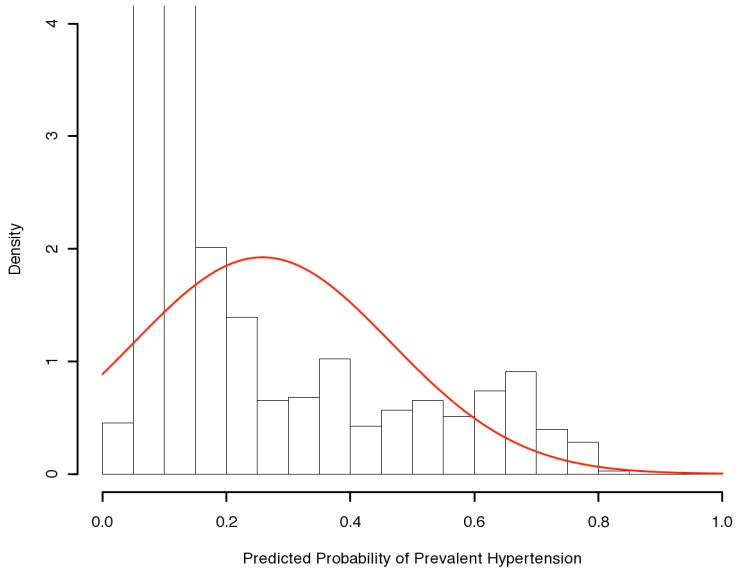


Figure 4-47 Histogram of the Predicted Probability of Developing Diabetes for the Clinical + Genotype Risk Index Model

4.9.4 Random Forests Comparison

A random forest was generated using the optimization set created by the risk index procedure. The forest had 500 individual trees, and the tuning procedure described in detail in Section 4.6.4 was used to find the number of variables *k* considered at each split that provided the lowest out-of-bag error estimate. The optimized *k* chosen was 44, which gave an out-of-bag error estimate of 18.32%.

When working with a dataset that has two possible classes, the standard procedure for a random forest is to assign a prediction to an individual based on a simple majority of votes, when the prevalence of the outcome is less than 50% changing the proportion of

votes needed to classify an individual can significantly impact the estimates of performance. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set on a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of 5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the individual to be "high risk". Table 4-45 shows the results of this investigation.

Predictions were then made about each individual in the independent testing set created by the risk index procedure, and the sensitivity, specificity, misclassification, and positive predictive value of the predictions was assessed. One thousand bootstrap samples of the independent testing set were generated, and predictions were made about each individual in each of the bootstrap samples. This data was used to create 95% confidence intervals for the sensitivity, specificity, misclassification, and positive predictive value estimates. Table 4-46 shows the sensitivity, specificity, misclassification, and positive predictive value estimates for the random forest as well as the 95% confidence interval for each estimate. Lastly, using the class votes for the individuals in the independent testing set, an ROC curve was created, and the AUC for the ROC curve was estimated (Figure 4-48).

Table 4-45 Performance Estimates of the Random Forest

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	1.000	0.034	0.806	0.170
0.1	1.000	0.172	0.691	0.193
0.15	0.913	0.345	0.561	0.216
0.2	0.739	0.586	0.388	0.262
0.25	0.565	0.793	0.245	0.351
0.3	0.435	0.905	0.173	0.476
0.35	0.348	0.948	0.151	0.571
0.4	0.087	0.974	0.173	0.400
0.45	0.087	0.991	0.158	0.667
0.5	0.087	1.000	0.151	1.000
0.55	0.043	1.000	0.158	1.000
0.6	0.000	1.000	0.165	-
0.65	0.000	1.000	0.165	-
0.7	0.000	1.000	0.165	-
0.75	0.000	1.000	0.165	-
0.8	0.000	1.000	0.165	-
0.85	0.000	1.000	0.165	-
0.9	0.000	1.000	0.165	-
0.95	0.000	1.000	0.165	-

Table 4-46 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for the Random Forest Model

Proportion of Votes for "High Risk" Class	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
0.1	1.000 (1.000-1.000)	0.172 (0.107-0.248)	0.691 (0.612-0.770)	0.193 (0.124-0.265)
0.15	0.913 (0.786-1.000)	0.345 (0.259-0.435)	0.561 (0.482-0.640)	0.216 (0.140-0.300)
0.2	0.739 (0.545-0.913)	0.586 (0.491-0.678)	0.388 (0.302-0.475)	0.262 (0.152-0.377)
0.25	0.565 (0.357-0.773)	0.793 (0.712-0.857)	0.245 (0.180-0.317)	0.351 (0.206-0.512)
0.3	0.435 (0.233-0.647)	0.905 (0.843-0.956)	0.173 (0.115-0.237)	0.476 (0.261-0.700)
0.35	0.348 (0.150-0.550)	0.948 (0.905-0.983)	0.151 (0.094-0.216)	0.571 (0.300-0.833)
0.4	0.087 (0.000-0.222)	0.974 (0.939-1.000)	0.173 (0.108-0.237)	0.400 (0.000-1.000)

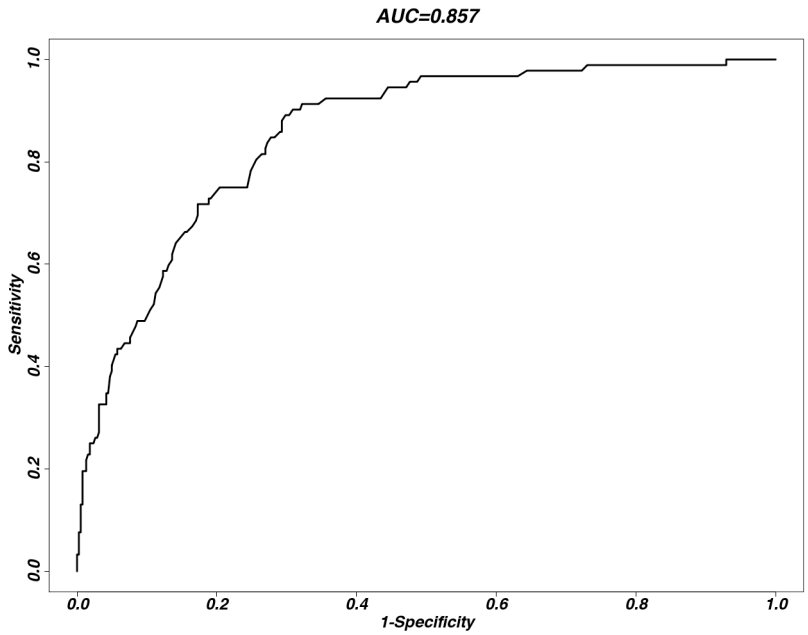


Figure 4-48 ROC Curve and AUC for the Incident Hypertension PCA Random Forest Model

4.9.5 Conclusion

Prevalent hypertension occurs in 19% of subjects, making it intermediate in commonness to incident hypertension (30%) and incident diabetes (10%). As might be expected, then, the predictive performance for prevalent hypertension falls between these outcomes as well, with a sensitivity less than that for incident hypertension but greater than that for incident diabetes, with a similar pattern observed for specificity, misclassification, and PPV. As for the risk index for incident hypertension and incident diabetes built with the 500 most highly associated SNPs the AUC for both the Clinical and Clinical + Genotype risk index model is less than that form the random forest model, and modifying the voting procedure for the random forest yields predictive performance exceeding that of either risk index model. However, no single set of predictions from the random forest provides

predictive performance where the 95% confidence intervals do not overlap with the performance estimates from the risk index models.

4.10 Prevalent Hypertension Results Using Top 500 Principal Components

4.10.1 Variable Selection

Using the procedure described in Section 4.5.1, the risk index procedure was performed to predict risk of developing diabetes within a ten-year time frame. In place of the 500 SNPs most highly associated with prevalent hypertension, the top 500 principal components from a principal components analysis of all available SNPs were used.

Table 4-47 gives a summary of the order in which variables were selected for the 100 bootstrap samples of the optimization set used to build the Clinical risk index model. Sixty-seven out of the 100 trimmed Clinical risk index models contained age as a variable, and 46 contained weight. Marital status was included in 58 out of the 100 trimmed Clinical risk index models. Weight, height, and having ever smoked, were also frequently included in the set of 100 trimmed Clinical risk index models, appearing in 39, 39, and 28 trimmed Clinical risk index models, respectively. Table 4-48 gives a summary of the order in which the most commonly chosen principal components were selected into the 100 Clinical + Genotype risk index models. Each principal component in Table 4-48 appears in at least 9 trimmed Clinical + Genotype risk index models.

Table 4-47 Summary of Number of Times Each Variable is Selected into a Specific Model Position for the Clinical Risk Index Model for Prevalent Hypertension

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
Age (yrs)	54	8	4	2	1	1	3	2	1	3	1	0	3	3	1	1	2	4	2	0	96	67
Sex	1	7	4	6	8	9	8	11	7	5	2	0	4	2	1	4	4	6	3	6	98	16
Weight (lbs)	1	26	6	5	2	1	5	1	6	4	7	2	1	7	7	1	4	5	3	1	95	39
Height (in)	5	2	14	14	14	8	6	10	5	8	3	1	5	2	0	0	1	1	0	0	99	39
Systolic Blood Pressure (mm Hg)	0	0	0	0	1	1	2	0	4	7	7	7	2	5	6	8	11	11	7	11	90	2
Diastolic Blood Pressure (mm Hg)	0	0	2	1	5	7	4	5	11	10	9	13	8	3	9	1	5	3	2	2	100	13
Total Cholesterol (mg/dL)	1	0	0	0	1	2	3	4	0	3	9	7	6	11	2	7	11	5	12	9	93	5
High-density Lipoprotein Level (mg/dL)	3	2	1	2	1	2	2	1	3	3	11	5	8	6	6	11	7	3	11	7	95	9
Low-density Lipoprotein Level (mg/dL)	0	4	17	16	15	15	8	5	5	5	0	2	5	2	1	0	0	0	0	0	100	28
Triglycerides (mg/dL)	2	4	11	5	7	9	13	12	3	4	3	1	8	3	3	2	4	1	2	1	98	25
Ever Smoked	3	2	9	7	7	12	6	3	7	6	3	3	3	4	4	5	4	4	2	2	96	27
Currently Smokes	10	16	12	18	14	10	2	4	4	1	1	3	1	1	1	1	0	1	0	0	100	58
Weekly Alcohol Consumption	2	8	1	4	1	5	1	6	11	3	1	5	2	2	3	4	6	3	8	13	89	10
Marital Status	1	0	2	1	4	4	9	6	4	10	3	4	4	0	5	3	5	4	9	12	90	8
Left Ventricular Mass (g)	3	4	2	6	2	3	2	1	1	5	1	5	7	5	7	7	5	13	6	8	93	17
Left Ventricular Ejection Fraction (%)	0	2	1	0	0	0	6	3	3	3	12	6	6	5	11	11	9	11	4	3	96	9
Blood Glucose (mg/dL)	1	2	6	4	1	0	2	3	3	3	8	7	3	7	11	7	4	7	6	11	96	16
Blood Urea Nitrogen (mg/dL)	2	7	2	3	3	4	9	6	3	8	7	5	4	8	3	6	5	3	8	3	99	15

Total Serum Protein Level (mg/dL)	11	3	1	4	2	4	7	5	6	2	5	6	7	8	9	7	4	3	2	3	99	24
Serum Albumin Level (mg/dL)	0	0	0	0	0	2	1	4	3	6	3	10	8	7	2	9	5	6	9	6	81	5
Serum Bilirubin Level (mg/dL)	0	3	5	2	11	1	1	8	10	1	4	8	5	9	8	5	4	6	4	2	97	16
Serum Alkaline Phosphatase Level (mg/dL)	54	8	4	2	1	1	3	2	1	3	1	0	3	3	1	1	2	4	2	0	96	67
Serum Creatine Level (mg/dL)	1	7	4	6	8	9	8	11	7	5	2	0	4	2	1	4	4	6	3	6	98	16

Table 4-48 Summary of Number of Times Selected Principal Component Variables are Selected into a Specific Model Position for the Clinical + Genotype Risk Index Model for Prevalent Hypertension

Variable	Variable Position																				Total # of Times in Untrimmed Model	Total # of Times in Trimmed Model
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
PC55	1	0	0	1	0	0	1	1	1	0	0	0	1	1	1	1	0	0	1	1	11	10
PC160	1	0	1	1	0	1	0	1	2	0	1	0	0	0	0	0	2	0	1	0	11	10
PC3	0	0	0	2	1	0	1	2	0	0	1	1	1	0	0	0	1	1	0	0	11	9
PC36	0	0	0	0	1	1	0	1	0	1	0	0	1	0	2	1	0	1	1	1	11	9
PC104	1	1	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	1	1	0	9	9
PC116	0	1	0	0	0	1	2	0	1	0	1	0	1	0	0	1	0	0	1	1	10	9
PC150	2	0	1	0	1	0	0	2	0	0	2	0	1	0	1	0	3	0	0	1	14	9

4.10.2 Models

Table 4-49 shows the trimmed Clinical risk index models for a selection of five random bootstrap samples. Figure 4-49 shows the distribution of risk index values in the optimization set for the Clinical risk index model from one randomly selected bootstrap sample (Bootstrap Sample #14), and Figure 4-50 shows the distribution of risk index values in the independent testing set for the Clinical risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #14). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as "high risk of having hypertension" and those with a risk index value lower are predicted as "low risk of having hypertension". Table 4-50 shows the risk index values for 25 randomly chosen individuals from the Independent Testing Set from these five Clinical risk index models along with that risk index model's prediction about each individual, where 0 indicates low risk of having hypertension and 1 indicates high risk of having hypertension.

Table 4-49 Five Randomly Selected Trimmed Clinical Risk Index Models for Prevalent Hypertension

Bootstrap Sample	Model
14	0.0207*Height - 0.0167*HDL + 0.0114*LDL + 0.028*Blood Urea Nitrogen + 0.0569*Serum Albumin + 0.0029*Serum Bilirubin + 0.0404*Serum Creatine
49	0.1028*Age - 0.2738*Sex + 0.0181*Weight + 0.0062*Height + 0.0603*Weekly Alcohol Consumption - 0.0429*Marital Status
58	0.0963*Age - 0.0231*HDL + 0.002*Triglycerides - 0.6903*Current Smoking Status + 0.1074*Marital Status + 0.0405*Blood Glucose + 0.0798*Serum Creatine
61	0.1074*Marital Status
97	0.1019*Age + 0.0172*Weight - 0.7112*Current Smoking Status + 0.068*Marital Status

Table 4-50 Risk Index Values for 25 Individuals from the Independent Testing Set from Five Randomly Selected Clinical Risk Index Models for Prevalent Hypertension

		Bootstrap Sample #14 Cutoff Value = 0.943		Bootstrap Sample #49 Cutoff Value = 1.549		Bootstrap Sample #58 Cutoff Value = 1.333		Bootstrap Sample #61 Cutoff Value=0.322		Bootstrap Sample #97 Cutoff Value=1.996	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	0.875	0	1.470	0	1.615	1	0.215	0	2.065	1
2	1	0.754	0	0.921	0	0.977	0	0.215	0	1.446	0
3	0	0.785	0	1.494	0	1.345	1	0.215	0	2.209	1
4	0	0.651	0	1.377	0	1.342	1	0.322	0	2.134	1
5	0	0.858	0	1.273	0	1.280	0	0.215	0	1.941	0
6	0	0.747	0	1.024	0	1.000	0	0.215	0	1.498	0
7	1	0.795	0	1.429	0	1.610	1	0.215	0	2.081	1
8	0	0.781	0	0.908	0	0.792	0	0.107	0	1.176	0
9	1	0.754	0	1.353	0	1.388	1	0.107	0	2.034	1
10	1	0.948	1	1.736	1	1.970	1	0.215	0	2.523	1
11	0	0.783	0	1.651	1	1.584	1	0.215	0	2.488	1
12	0	0.708	0	1.156	0	1.258	0	0.430	1	1.799	0
13	0	0.878	0	1.406	0	1.163	0	0.215	0	1.993	0
14	0	0.723	0	0.983	0	1.138	0	0.215	0	1.492	0
15	0	0.906	0	1.194	0	1.420	1	0.215	0	1.846	0
16	0	0.792	0	0.959	0	1.180	0	0.215	0	1.502	0
17	0	0.709	0	1.210	0	1.115	0	0.215	0	1.520	0
18	1	0.779	0	1.411	0	1.357	1	0.215	0	2.066	1
19	0	0.811	0	1.437	0	1.487	1	0.430	1	2.107	1
20	0	0.802	0	1.095	0	1.107	0	0.215	0	1.462	0
21	0	0.903	0	1.332	0	1.278	0	0.430	1	1.996	0
22	0	0.771	0	1.288	0	1.159	0	0.215	0	1.681	0
23	0	0.876	0	1.338	0	1.330	0	0.215	0	2.050	1
24	0	0.713	0	1.282	0	1.277	0	0.322	0	1.982	0
25	0	0.821	0	1.166	0	1.324	0	0.537	1	1.860	0

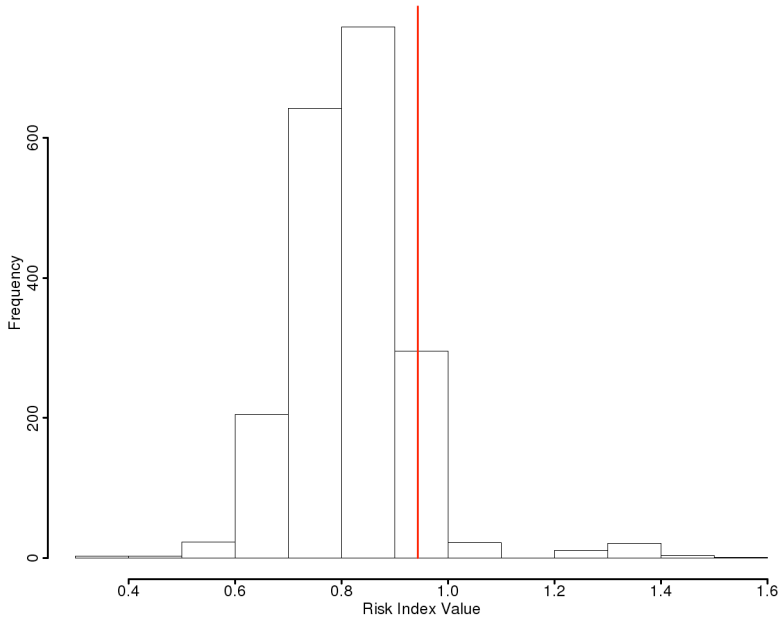


Figure 4-49 Distribution of Risk Index Values in the Optimization Set from the Clinical Risk Index Model for Bootstrap Sample #14

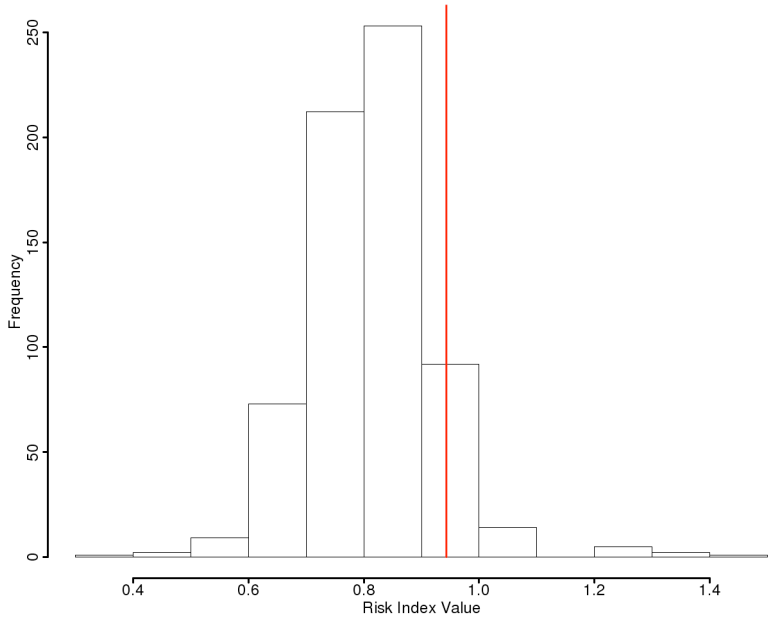


Figure 4-50 Distribution of Risk Index Values in the Independent Testing Set from the Clinical Risk Index Model for Bootstrap Sample #14

Table 4-51 shows the trimmed Clinical + Genotype risk index models for the same set of five bootstraps shown in Table 4-49, and Table 4-52 shows the risk index values and predictions of the same 25 randomly chosen individuals for the 5 bootstrap samples shown in Table 4-50. Figure 4-51 shows the distribution of risk index values in the optimization set for the Clinical + Genotype risk index model from one randomly selected bootstrap sample (Bootstrap Sample #14). Figure 4-52 shows the distribution of risk index values in the independent testing set for the Clinical + Genotype risk index model from the same randomly selected bootstrap sample (Bootstrap Sample #14). The red line on each graph marks the cutoff point for the model. All individuals with a risk index value greater than this are predicted as "high risk of having hypertension" and those with a risk index value lower are predicted as "low risk of having hypertension".

Table 4-51 Five Randomly Selected Clinical + Genotype Risk Index Models for Prevalent Hypertension

Bootstrap Sample	Model
14	0.0207*Height - 0.0167*HDL + 0.0114*LDL + 0.028*Blood Urea Nitrogen + 0.0569*Serum Albumin + 0.0029*Serum Bilirubin + 0.0404*Serum Creatine + 1.1469*PC30 - 2.7755*PC41 - 0.2129*PC81 + 0.6202*PC104 - 0.446*PC106 - 1.164*PC109 + 0.3898*PC135 - 0.2404*PC153 - 0.393*PC172 - 1.6513*PC223 + 0.1298*PC260 - 0.2362*PC316 - 0.0653*PC337 - 1.0113*PC343 - 0.747*PC363 - 0.1273*PC371 + 0.9937*PC436 + 4.0547*PC454
49	0.1028*Age - 0.2738*Sex + 0.0181*Weight + 0.0062*Height + 0.0603*Weekly Alcohol Consumption - 0.0429*Marital Status + 0.4226*PC36 - 0.5782*PC68 + 0.2086*PC116 - 6.1775*PC131 - 1.2208*PC248 + 1.3316*PC282 + 0.5426*PC394 + 0.6511*PC470
58	0.0963*Age - 0.0231*HDL + 0.002*Triglycerides - 0.6903*Current Smoking Status + 0.1074*Marital Status + 0.0405*Blood Glucose + 0.0798*Serum Creatine + 3.6541*PC6 - 1.9015*PC50 + 0.3906*PC70 + 0.3068*PC128 - 0.9822*PC171 + 13.5045*PC213 + 1.6875*PC241 + 3.6425*PC329 + 1.8774*PC363 - 3.5805*PC395 - 0.7359*PC403 + 4.4064*PC427 - 1.2497*PC478 + 1.2648*PC481
61	0.1074*Marital Status - 0.5996*PC7 - 3.7335*PC16 + 0.0024*PC110 + 0.9075*PC134 + 1.1272*PC141 - 0.3245*PC171 - 0.643*PC173 - 1.6777*PC212 - 0.1376*PC233 + 5.5252*PC273 + 1.2982*PC336 - 1.7907*PC364 - 0.3527*PC371 + 1.6016*PC388 + 2.5273*PC397 + 0.0266*PC405 - 1.4463*PC418 + 0.9719*PC430 + 1.0295*PC465
97	0.1019*Age + 0.0172*Weight - 0.7112*Current Smoking Status + 0.068*Marital Status - 0.9318*PC22 - 0.7419*PC54 + 2.3027*PC55 - 0.538*PC58 - 0.2581*PC86 - 0.675*PC88 - 0.3091*PC100 - 2.0593*PC101 - 1.1611*PC107 - 3.3506*PC170 + 0.3638*PC175 - 1.7034*PC182 - 2.6485*PC215 - 1.455*PC225 + 0.1955*PC276 + 2.887*PC355 + 0.8472*PC363 + 0.808*PC394 - 2.5515*PC413 - 0.0194*PC450

Table 4-52 Risk Index Values for 25 Individuals in the Independent Testing Set for the Clinical + Genotype Risk Index Models from Five Randomly Selected Bootstrap Samples for Prevalent Hypertension

		Bootstrap Sample #14 Cutoff Value = 0.943		Bootstrap Sample #49 Cutoff Value = 1.544		Bootstrap Sample #58 Cutoff Value = 1.334		Bootstrap Sample #61 Cutoff Value=0.311		Bootstrap Sample #97 Cutoff Value=1.998	
Individual	Outcome	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction	Risk Index Value	Prediction
1	0	0.874	0	1.475	0	1.653	1	0.217	0	2.051	1
2	1	0.765	0	0.920	0	0.977	0	0.215	0	1.436	0
3	0	0.781	0	1.511	0	1.351	1	0.202	0	2.202	1
4	0	0.657	0	1.359	0	1.304	0	0.317	1	2.131	1
5	0	0.851	0	1.284	0	1.293	0	0.213	0	1.935	0
6	0	0.747	0	1.031	0	1.001	0	0.215	0	1.511	0
7	1	0.792	0	1.422	0	1.585	1	0.210	0	2.066	1
8	0	0.783	0	0.901	0	0.774	0	0.112	0	1.168	0
9	1	0.770	0	1.373	0	1.408	1	0.100	0	2.032	1
10	1	0.949	1	1.760	1	1.969	1	0.205	0	2.516	1
11	0	0.783	0	1.677	1	1.608	1	0.234	0	2.484	1
12	0	0.709	0	1.162	0	1.271	0	0.434	1	1.796	0
13	0	0.874	0	1.413	0	1.160	0	0.214	0	1.987	0
14	0	0.719	0	0.991	0	1.147	0	0.220	0	1.491	0
15	0	0.907	0	1.172	0	1.414	1	0.205	0	1.842	0
16	0	0.792	0	0.955	0	1.192	0	0.210	0	1.512	0
17	0	0.703	0	1.214	0	1.091	0	0.224	0	1.514	0
18	1	0.773	0	1.407	0	1.357	1	0.211	0	2.066	1
19	0	0.806	0	1.443	0	1.506	1	0.397	1	2.109	1
20	0	0.797	0	1.100	0	1.074	0	0.246	0	1.460	0
21	0	0.890	0	1.328	0	1.280	0	0.427	1	1.987	0
22	0	0.769	0	1.284	0	1.182	0	0.211	0	1.683	0
23	0	0.871	0	1.359	0	1.327	0	0.223	0	2.054	1
24	0	0.703	0	1.290	0	1.257	0	0.328	1	1.976	0
25	0	0.819	0	1.176	0	1.341	1	0.532	1	1.857	0

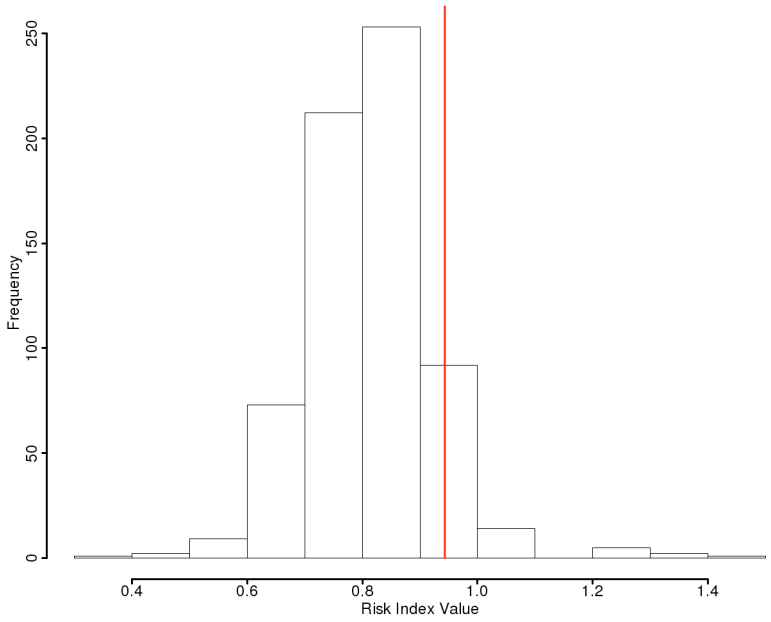


Figure 4-51 Distribution of Risk Index Values in the Optimization Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #14

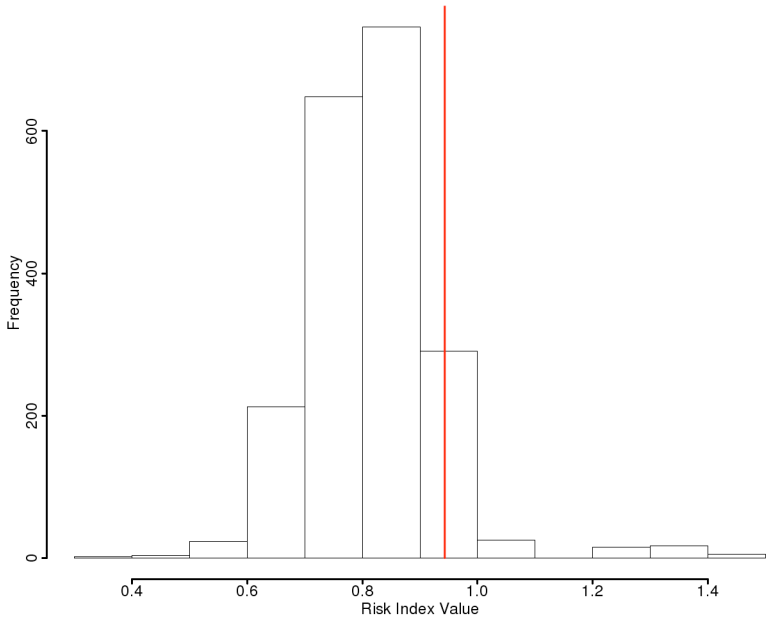


Figure 4-52 Distribution of Risk Index Values in the Independent Testing Set from the Clinical + Genotype Risk Index Model for Bootstrap Sample #14

4.10.3 Predictive Performance

Once predictions were made for each individual in the independent testing set the sensitivity, specificity, misclassification, and positive predictive value were calculated for the Clinical risk index model and the Clinical + Genotype risk index model as described in Section 4.5.3. The estimates and confidence intervals for sensitivity, specificity, misclassification, and positive predictive value are given in Table 4-53. Lastly, using the individual predictions from each of the 100 trimmed Clinical risk index models and 100 trimmed Clinical + Genotype risk index models for the individuals in the independent testing set, ROC curves were generated, and the AUC for the ROC curve was estimated for both the Clinical and Clinical + Genotype risk index models (Figure 4-53, Figure 4-54). For the Clinical risk index model the AUC for the ROC curve was 0.722, and for the Clinical + Genotype risk index model the AUC for the ROC curve was 0.712.

Table 4-53 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for Prevalent Hypertension

Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
Clinical	0.263 (0.192-0.343)	0.895 (0.868-0.921)	0.232 (0.2-0.265)	0.385 (0.286-0.489)
Clinical + Genotype	0.203 (0.138-0.274)	0.932 (0.91-0.953)	0.214 (0.182-0.245)	0.429 (0.313-0.552)

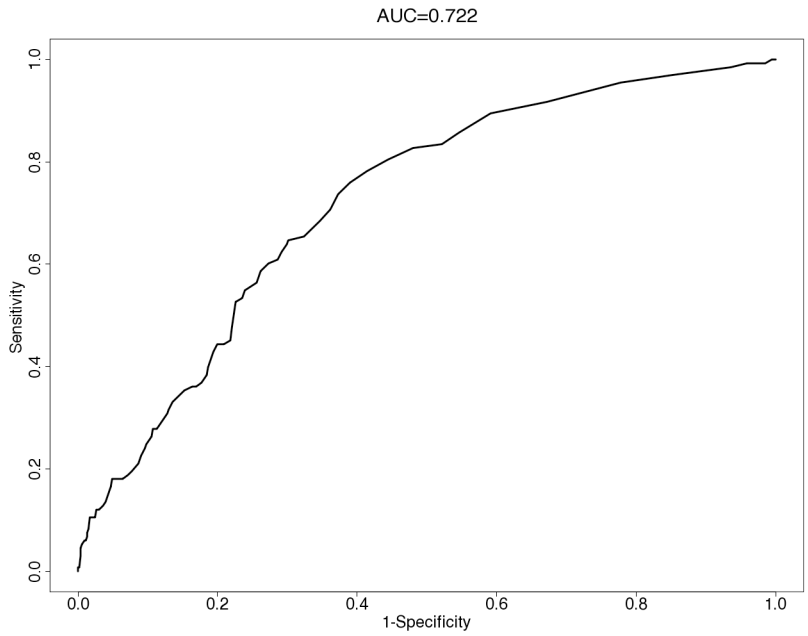


Figure 4-53 ROC Curve and AUC for the Incident Diabetes Clinical Risk Index Model

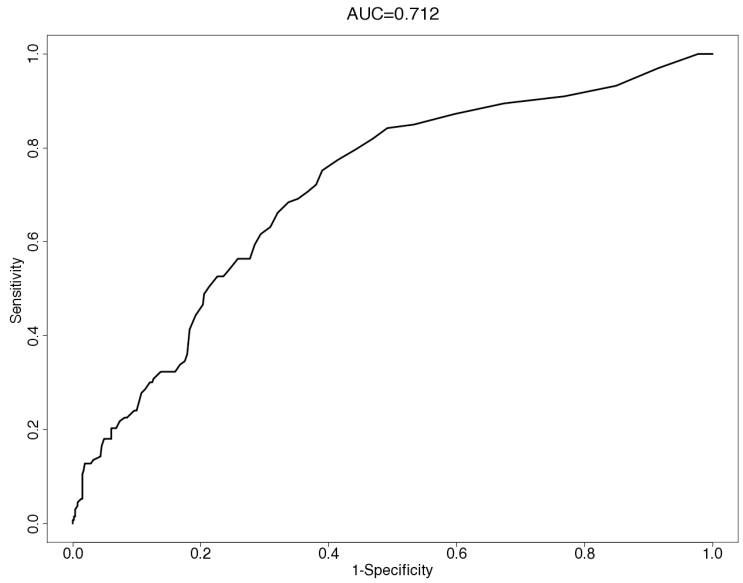


Figure 4-54 ROC Curve and AUC for the Incident Diabetes Clinical + Genotype Risk Index Model

The ensemble nature of the final risk index prediction means that there is a consensus prediction based on votes from the individual bootstrap samples. The proportion of models that predict that an individual is at high risk of having hypertension, then, represents the predicted probability of an individual having hypertension. Section 4.5.3 describes the calculation of a confidence interval for this predicted probability.

Figure 4-55 shows the distribution of the predicted probability of having hypertension for the Clinical risk index model in the independent testing set, and Figure 4-56 shows the distribution of the predicted probability of having hypertension for the Clinical + Genotype risk index model in the independent testing set. In both Figures, a density line is shown on the graph to indicate the density of a normal distribution with the mean and standard deviation matching that of the predicted probability distribution.

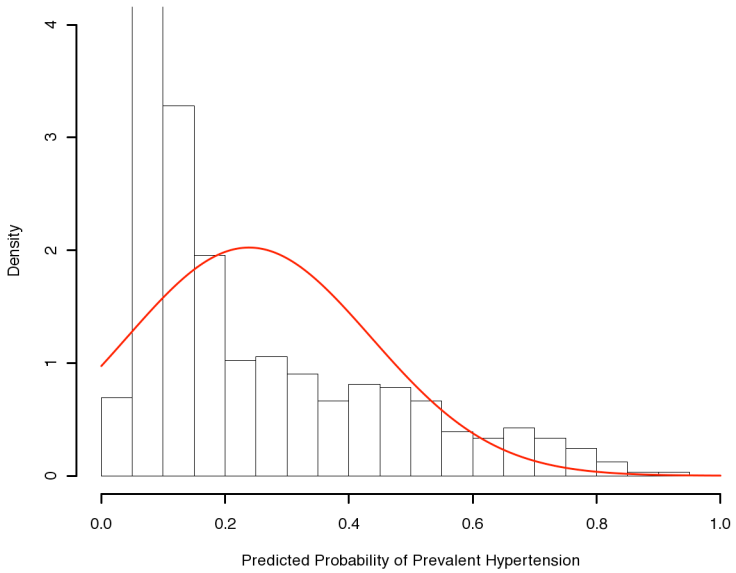


Figure 4-55 Histogram of the Predicted Probability of Developing Diabetes for the Clinical Risk Index Model

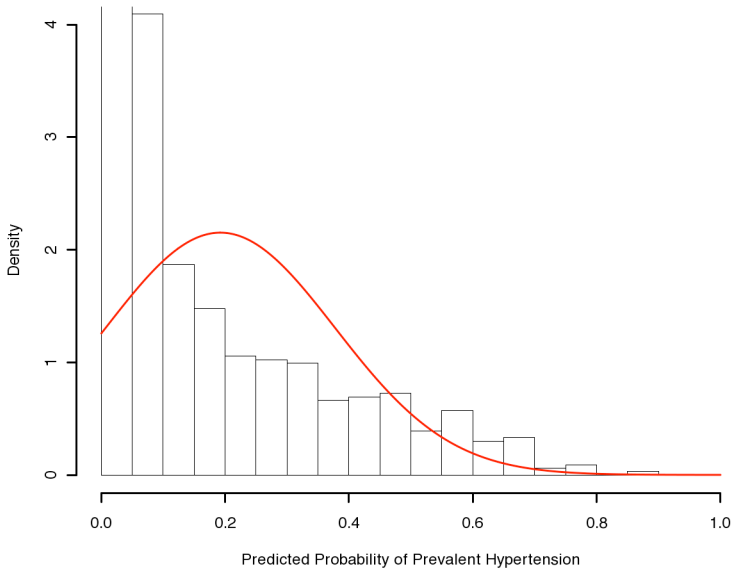


Figure 4-56 Histogram of the Predicted Probability of Developing Diabetes for the Clinical + Genotype Risk Index Model

4.10.4 Random Forests Comparison

A random forest was generated using the optimization set created by the risk index procedure. The forest had 500 individual trees, and the tuning procedure described in detail in Section 4.6.4 was used to find the number of variables *k* considered at each split that provided the lowest out-of-bag error estimate. The optimized *k* chosen was 45, which gave an out-of-bag error estimate of 18.86%.

When working with a dataset that has two possible classes, the standard procedure for a random forest is to assign a prediction to an individual based on a simple majority of votes, when the prevalence of the outcome is less than 50% changing the proportion of

votes needed to classify an individual can significantly impact the estimates of performance. To fully examine the performance of the random forest, predictions were made about each individual in the independent testing set on a range of proportions. First, an individual was assigned a prediction of "high risk" if 5% or more of the trees in the forest predicted the individual to be "high risk". This was then repeated in increments of 5% until individuals were assigned a prediction of "high risk" only if 95% or more of the trees in the forest predicted the individual to be "high risk". Table 4-54 shows the results of this investigation.

Predictions were then made about each individual in the independent testing set created by the risk index procedure, and the sensitivity, specificity, misclassification, and positive predictive value of the predictions was assessed. One thousand bootstrap samples of the independent testing set were generated, and predictions were made about each individual in each of the bootstrap samples. This data was used to create 95% confidence intervals for the sensitivity, specificity, misclassification, and positive predictive value estimates. Table 4-55 shows the sensitivity, specificity, misclassification, and positive predictive value estimates for the random forest as well as the 95% confidence interval for each estimate. Lastly, using the class votes for the individuals in the independent testing set, an ROC curve was created, and the AUC for the ROC curve was estimated (Figure 4-57).

Table 4-54 Performance Estimates of the Random Forest

Proportion of Votes for "High Risk" Class	Sensitivity	Specificity	Misclassification	PPV
0.05	0.990	0.043	0.803	0.167
0.1	0.929	0.170	0.707	0.178
0.15	0.909	0.391	0.525	0.224
0.2	0.818	0.577	0.384	0.273
0.25	0.576	0.740	0.287	0.300
0.3	0.424	0.849	0.220	0.353
0.35	0.303	0.939	0.164	0.492
0.4	0.162	0.969	0.162	0.500
0.45	0.051	0.992	0.161	0.556
0.5	0.000	0.998	0.164	0.000
0.55	0.000	1.000	0.162	-
0.6	0.000	1.000	0.162	-
0.65	0.000	1.000	0.162	-
0.7	0.000	1.000	0.162	-
0.75	0.000	1.000	0.162	-
0.8	0.000	1.000	0.162	-
0.85	0.000	1.000	0.162	-
0.9	0.000	1.000	0.162	-
0.95	0.000	1.000	0.162	-

Table 4-55 Estimates and 95% Confidence Intervals of Sensitivity, Specificity, Misclassification, and Positive Predictive Value for the Ranomd Forest Model

Proportion of Votes for "High Risk" Class	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)
0.1	0.929 (0.871-0.978)	0.170 (0.139-0.202)	0.707 (0.672-0.741)	0.178 (0.148-0.211)
0.15	0.909 (0.846-0.963)	0.391 (0.349-0.431)	0.525 (0.487-0.559)	0.224 (0.186-0.265)
0.2	0.818 (0.740-0.892)	0.577 (0.534-0.617)	0.384 (0.348-0.420)	0.273 (0.223-0.325)
0.25	0.576 (0.478-0.670)	0.740 (0.702-0.776)	0.287 (0.251-0.323)	0.300 (0.232-0.366)
0.3	0.424 (0.320-0.520)	0.849 (0.816-0.877)	0.220 (0.189-0.254)	0.353 (0.265-0.444)
0.35	0.303 (0.211-0.398)	0.939 (0.916-0.958)	0.164 (0.136-0.193)	0.492 (0.362-0.615)
0.4	0.162 (0.087-0.236)	0.969 (0.954-0.984)	0.162 (0.131-0.193)	0.500 (0.333-0.696)
0.45	0.051 (0.011-0.093)	0.992 (0.984-0.998)	0.161 (0.133-0.190)	0.556 (0.200-0.875)

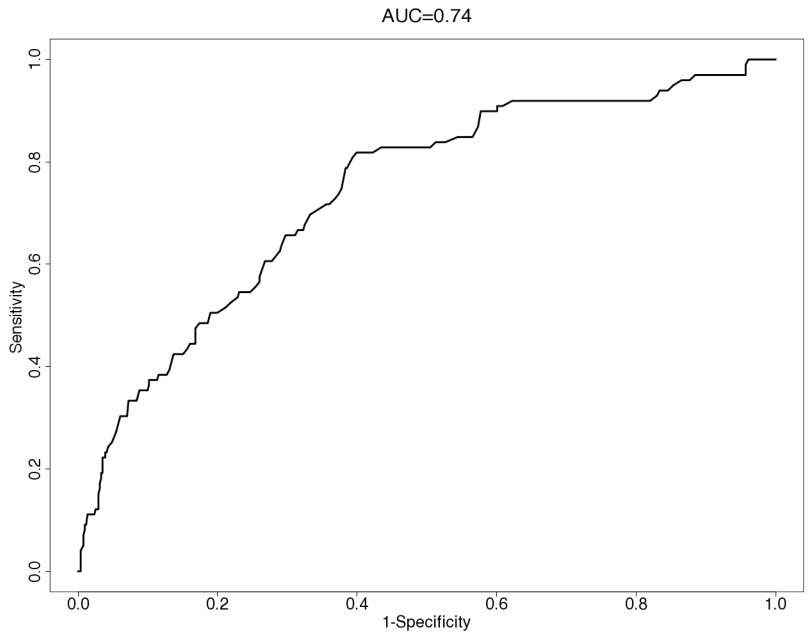


Figure 4-57 ROC Curve and AUC for the Incident Hypertension PCA Random Forest Model

4.10.5 Conclusion

Although the predictive performance estimates for the Clinical and Clinical + Genotype risk index models for prevalent hypertension built with the top 500 principal components is somewhat lower than those for the Clinical and Clinical + Genotype risk index models for prevalent hypertension built with the 500 most highly associated SNPs, the AUC of the Clinical + Genotype model is higher for the top 500 principal components than for the 500 most highly associated SNPs. This trend was also observed for the incident hypertension and incident diabetes outcomes. This suggests that the risk index is able to take advantage of this additional information to improve prediction. The random forests model built with the top 500 principal components, however, has a lower AUC than the random forest model built using the 500 most highly associated SNPs, and this was also

observed for the incident hypertension and incident diabetes outcomes. Because the random forest methodology classifies individuals based on finding context-dependent relationships in variables, the uncorrelated nature of the principal components, even though it captures a larger amount of genetic information, is not as well-suited to prediction with random forests as a smaller number of polymorphisms that are somewhat correlated and may be involved in the context-dependent effects that the tree-structure of the random forest method is designed to exploit.

Chapter 5

Conclusion

5.1 Development of the Risk Index

The risk index procedure created and tested in this dissertation is intended as an expansion of genetic risk score methods used in a number of studies as a means to harness genetic information to make predictions about disease risk. Sequence variations, especially single nucleotide polymorphisms (SNPs), which can now be easily and inexpensively genotyped in large numbers, provide a solid starting place for disease risk prediction. An individual's genetic polymorphisms are, in large part, static, and so they can be queried long before a disease process has even begun (Plomin, et al, 2007a). In contrast, transcriptomic and proteomic markers that indicate disease are unlikely to be detected until an individual has already begun to develop that disease, even if they are outwardly asymptomatic (Plomin, et al, 2007b). When SNPs that can be used to identify a pool of individuals at increased risk of a disease are identified, doctors can then monitor these individuals more closely and track the development and progression of the disease (Ziogas, et al, 2009).

By creating a robust framework in which risk models can be constructed and tested, the risk index procedure is intended to develop the genetic risk score methods proposed and used in other contexts into a machine learning algorithm that can combine clinical data with genotype data to classify an individual at high or low risk for a particular disease.

The risk index procedure developed in this dissertation can thus move genetic risk scores from an ad-hoc strategy that is feasible only on a small scale with a limited number of polymorphisms to a large-scale process that incorporates statistical techniques in order to provide high quality predictions. Using a forward-selection procedure, prediction models that contain clinical variables were created for bootstrap samples of a dataset. With the addition of each variable the model is assessed using the Brier score, a metric developed to assess a model's predictive accuracy. Once the forward selection is complete, the model is pared back (ie. "trimmed") so that the best performing model remains. Using this model made up of clinical variables as a base the procedure is then repeated for the genotype variables. Once Clinical + Genotype models are created for each of the bootstrap samples of the dataset, they are used to make predictions about a fully independent testing set, with the prediction from each of the models acting as a vote. Each individual in the independent testing set is assigned a prediction of either high risk or low risk based on the majority vote of these models. Figure 5-1 shows a graphical overview of the risk index procedure. The models created by the risk index procedure can be easily applied to patients in a clinic, giving doctors a prediction about the individual's risk of developing the disease, a predicted probability of the individual developing the disease, and a 95% confidence interval around that predicted probability.

Chapter 2 describes in detail the risk index procedure, as well as the use of random forests (Breiman, 1996) as a standard metric against which to compare the performance of the risk index procedure. Chapter 3 discusses two simulation studies undertaken to characterize the performance of the risk index procedure. The first uses a small dataset of

1000 individuals, each with a binary outcome variable (with a 30% prevalence for the disease),

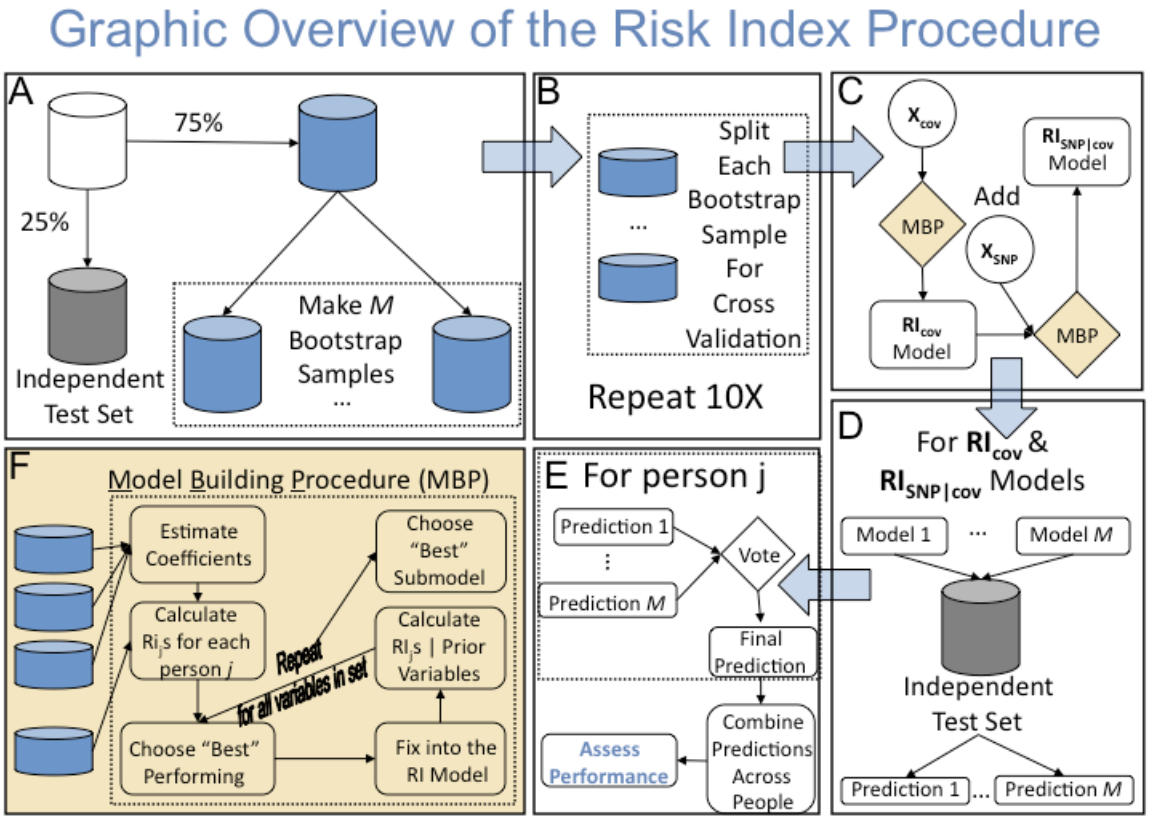


Figure 5-1 Graphic Overview of the Risk Index Procedure

eight clinical covariates (four associated with the outcome and four noise variables), and 500 SNPs (four associated with the outcome and the remainder noise variables). The second simulation is of a larger dataset of 10,000 individuals, again with a binary outcome that has 30% prevalence. Each individual has 29 clinical covariates (one very strongly associated with the outcome, three highly associated with the outcome, 11 moderately associated with the outcome, and the remainder weakly associated with the outcome) and 38,835 SNPs, simulating the SNPs on chromosome one from the

Affymetrix Genome-wide Human SNP Array 5.0 (Affymetrix, 2007), six of which were simulated to be associated with the outcome. Chapter 4 describes the application of the risk index procedure to Framingham Heart Study (FHS), and examines its ability to predict ten-year incident hypertension, ten-year incident diabetes, and prevalent hypertension.

5.2 Small-scale Simulation

5.2.1 Complete SNP Set

Table 5-1 shows a summary of the results from the small-scale suimulation. The first analysis in the small-scale simulation study was to determine the performance of the risk index procedure on a small set of clinical variables and 500 SNPs. On average across the 100 small-scale simulation datasets the Clinical risk index model produced a sensitivity of 60.6% and a specificity of 88.6%, and the Clinical + Genotype risk index model produced a sensitivity of 58.9% and a specificity of 89.6%. The average area under the receiver operating characteristic (ROC) curve (AUC) of the Clinical risk index model was 0.832 (SD = 0.027), and the average AUC of the Clinical + Genotype risk index model was 0.846 (SD = 0.024). The Clinical + Genotype risk index model does have a significantly higher mean AUC than the Clinical risk index model, however the difference is fairly small, and its effect on prediction in a real-world situation is unclear. The performance of the risk index here is fairly good, however the random forest excels at predicting these datasets, with a mean AUC of 0.987 (SD=0.006), suggesting that, particularly for these fairly simplistic datasets, random forests are better positioned to provide accurate classification.

Table 5-1 A Summary of the Results from the Small-scale Simulation Study

Simulation Study	Model	Sensitivity (SD)	Specificity (SD)	Misclassification (SD)	PPV (SD)	AUC (SD)
Small-scale Simulation (500 SNPs)	Clinical	0.606 (0.067)	0.886 (0.031)	0.199 (0.021)	0.704 (0.064)	0.832 (0.027)
	Clinical + Genotype	0.589 (0.065)	0.896 (0.030)	0.197 (0.022)	0.717 (0.062)	0.846 (0.024)
Small-scale Simulation (Top PCs)	Clinical	0.607 (0.066)	0.883 (0.032)	0.201 (0.027)	0.607 (0.067)	0.826 (0.033)
	Clinical + Genotype	0.590 (0.063)	0.891 (0.031)	0.200 (0.026)	0.708 (0.067)	0.839 (0.032)

5.2.2 Principal Components of the Complete SNP Set

The second analysis in the small-scale simulation study was to determine the performance of the risk index procedure on a small set of clinical variables and the principal components that explain 90% of the variance in the set of 500 SNPs simulated for the dataset. On average across the 100 small-scale simulation datasets the Clinical risk index model produced a sensitivity of 60.7% and a specificity of 88.3%, and the Clinical + Genotype risk index model produced a sensitivity of 59.0% and a specificity of 89.1%. The average AUC of the ROC curve for the Clinical risk index model was 0.826 (SD = 0.033), and the average AUC of the Clinical + Genotype risk index models was 0.839 (SD = 0.032). The random forest models that were created for each of the 100 small-scale simulation datasets had, on average, an AUC of 0.821 (SD = 0.031). Tuning the class voting procedure for these random forests models did not produce the same level of predictive performance that was achieved for the random forests models created using the full set of 500 SNPs; while some cut-off values could lead to higher sensitivities or specificities than what was observed for the Clinical + Genotype risk index model, no

cut-off demonstrated predictive performance that exceeded the Clinical + Genotype risk index model for all four predictive performance metrics.

The Clinical + Genotype risk index model built using the principal components of the SNPs has a significantly higher mean AUC than the Clinical risk index model (p=0.008) and the Clinical + Genotype model built using the SNPs. Again, however, these increases are relatively small. The Clinical + Genotype risk index model built using the principal components of the SNPs also has a significantly higher mean AUC than the random forest built using the clinical variables and the principal components of the SNPs (p=8.4E-5), however both of these means are much lower than the mean AUC for the random forest built with clinical covariates and SNP genotypes. This suggests that while the inclusion of the principal components improves the predictive ability of the risk index, it hampers the predictive ability of the random forest, likely because the orthoganalization of the genotype data breaks up the context-dependent effects the random forest leverages to make predictions.

5.3 Large-scale Simulation

5.3.1 500 Most Highly Associated SNPs

Table 5-2 shows a summary of the results from the large-scale simulation study. The first analysis in the large-scale simulation study was to determine the performance of the risk index procedure on a set of 29 clinical variables and 500 SNPs that were identified as highly associated with the outcome after a logistic regression analysis of the full set of 38,835 SNPs. On average across the 25 large-scale simulation datasets the Clinical risk

index model produced a sensitivity of 73.5% and a specificity of 93.3%, and the Clinical + Genotype risk index model produced a sensitivity of 73.4% and a specificity of 94.0%. The average AUC of the ROC curves for the Clinical risk index model was 0.926 (SD = 0.015), and the average AUC for the Clinical + Genotype risk index models was 0.932 (SD = 0.012). The random forest models that were created for each of the 25 large-scale simulation datasets on average had an AUC of 0.915 (SD = 0.013). Tuning the class voting procedure for these random forests models did not produce predictive performance that equaled either the Clinical or the Clinical + Genotype risk index models.

As with the small-scale simulation results, the Clinical + Genotype risk index model has a mean AUC significantly higher than the mean AUC for the Clinical risk index model ($p=0.001$), but again the difference in mean AUCs is small and it's impact on predictive performance is unclear. The random forest built using the clinical covariates and SNP genotypes, however, had a mean AUC that was significantly lower than the Clinical + Genotype model built using the most highly associated SNPs ($p=1.5E-8$), and the difference between the AUCs here is considerably more sizeable. One explanation for this is that the larger number of clinical covariates and fairly low correlation among the clinical covariates means that the data is too complex for fairly simple trees to be effective predictors, but there are too few context-dependent effects for the random forest to use.

Table 5-2 A Summary of the Results from the Large-Scale Simulation Study

Simulation Study	Model	Sensitivity (SD)	Specificity (SD)	Misclassification (SD)	PPV (SD)	AUC (SD)
Large-scale Simulation (Top 500 SNPs)	Clinical	0.725 (0.038)	0.933 (0.016)	0.132 (0.015)	0.831 (0.035)	0.926 (0.015)
	Clinical + Genotype	0.734 (0.034)	0.940 (0.015)	0.124 (0.013)	0.849 (0.033)	0.939 (0.012)
Large-scale Simulation (Top 500 PCs)	Clinical	0.758 (0.052)	0.927 (0.024)	0.125 (0.021)	0.827 (0.043)	0.931 (0.011)
	Clinical + Genotype	0.749 (0.052)	0.930 (0.024)	0.126 (0.021)	0.832 (0.044)	0.931 (0.011)

5.3.2 Principal Components of Complete SNP Set

The second analysis in the large-scale simulation study was to determine the performance of the risk index procedure on a set of 29 clinical variables and 500 most highly ranked principal components after a principal components analysis of the full set of 38,835 SNPs. On average across the 25 large-scale simulation datasets the Clinical risk index model produced a sensitivity of 73.8% and a specificity of 92.7%, and the Clinical + Genotype risk index model produced a sensitivity of 74.9% and a specificity of 93.0%. The average AUC of the ROC curves for the Clinical risk index model was 0.931 (SD = 0.021), and the average AUC for the Clinical + Genotype risk index models was 0.931 (SD = 0.022). The random forest models that were created for each of the 25 large-scale simulation datasets on average had an AUC of 0.856 (SD = 0.022). Tuning the class voting procedure for these random forest models did not produce predictive performance that equaled either the Clinical or the Clinical + Genotype risk index model.

The mean AUCs for the Clinical and Clinical + Genotype risk index model here are effectively identical, with neither model providing significantly higher values (p=0.98).

The Clinical + Genotype risk index model built with the principal components of the SNPs does have a significantly higher mean AUC than the random forests model built with principal components (p=4.4E-16). This provides further evidence that the uncorrelated nature of the principal components makes them a better choice for the risk index than for the random forest. As observed for the small-scale simulation, the principal components do not perform as well in the random forest as do the most highly associated SNPs (p=6.5E-14).

5.4 Simulation Study Conclusions

The results from the simulation study offer some important insights into the functioning and performance of the risk index. First of all, the results from the Clinical + Genotype risk models for the small-scale simulation study using a set of 500 SNPs and the large-scale simulation using the set of 500 SNPs most highly associated with the outcome show that while the predictive performance of the these models is quite good, the genotype variables selected into the models do not reflect the known, true positive genotype variables. This suggests that while the risk index may be useful for predictive applications, the resulting models are not interpretable and are probably not useful as starting points for investigation into the role of implicated polymorphisms in the disease process being investigated.

Second, the Clinical + Genotype risk index models outperformed the Clinical risk index models in all but one case. T-tests show that the mean AUC for the Clinical + Genotype risk index model was statistically significantly higher than the mean AUC for the Clinical

risk index model for the small-scale simulation using the full set of 500 SNPs (p=0.006), the small-scale simulation using the principal components that accounted for 90% of the variance in the genotype variables (p=0.008), and the large-scale simulation using the top 500 most highly associated SNPs (p=0.001). For the large-scale simulation using the top 500 principal components of the 38,835 SNPs, however, the Clinical and Clinical + Genotype risk index models do not have statistically significantly different performance (p=0.98).

Lastly, the difference in performance between the risk index models constructed using a set of SNPs and the risk index models constructed using sets of principal components suggests that the risk index procedure performs best with the uncorrelated data provided by principal components. Given the statistical procedures used to build the risk index, this is an understandable result, but it is important to note that the fact that the principal components data is uncorrelated is not the sole reason that the risk index performance is improved. The fact that the top 500 principal components also encode more information than the 500 most highly associated SNPs also likely plays a role in the performance improvement (Raychaudhuri, et al, 2000).

5.5 Framingham Heart Study

The analysis of the FHS data focused on three outcomes: ten-year incident hypertension, ten-year incident diabetes, and prevalent hypertension. Table 5-3 shows an overview of the results from the analysis of the FHS data.

Table 5-3 A Summary of the Risk Index's Predictive Performance on the Framingham Heart Study Data

Outcome	Model	Sensitivity (95% CI)	Specificity (95% CI)	Misclassification (95% CI)	PPV (95% CI)	AUC
Incident Hypertension (Top 500 SNPs)	Clinical	0.667 (0.594 - 0.736)	0.486 (0.444 - 0.53)	0.468 (0.433 - 0.504)	0.308 (0.26 - 0.352)	0.567
	Clinical + Genotype	0.539 (0.464 - 0.609)	0.457 (0.413 - 0.497)	0.522 (0.487 - 0.559)	0.254 (0.21 - 0.298)	0.475
Incident Hypertension (Top 500 PCs)	Clinical	0.608 (0.538-0.683)	0.505 (0.458-0.549)	0.468 (0.429-0.508)	0.299 (0.248-0.349)	0.566
	Clinical + Genotype	0.591 (0.518-0.667)	0.544 (0.498-0.587)	0.444 (0.405-0.482)	0.31 (0.258-0.363)	0.563
Incident Diabetes (Top 500 SNPs)	Clinical	0.224 (0.13-0.323)	0.901 (0.877-0.922)	0.163 (0.136-0.193)	0.192 (0.113-0.29)	0.722
	Clinical + Genotype	0.104 (0.04-0.183)	0.923 (0.901-0.943)	0.155 (0.129-0.182)	0.125 (0.05-0.22)	0.683
Incident Diabetes (Top 500 PCs)	Clinical	0.204 (0.1-0.322)	0.959 (0.943-0.975)	0.102 (0.078-0.125)	0.306 (0.15-0.469)	0.769
	Clinical + Genotype	0.13 (0.043-0.224)	0.974 (0.96-0.987)	0.095 (0.074-0.119)	0.304 (0.118-0.5)	0.782
Prevalent Hypertension (Top 500 SNPs)	Clinical	0.459 (0.375-0.547)	0.844 (0.816-0.875)	0.228 (0.197-0.26)	0.407 (0.326-0.489)	0.733
	Clinical + Genotype	0.323 (0.245-0.405)	0.846 (0.817-0.875)	0.252 (0.22-0.285)	0.328 (0.25-0.413)	0.692
Prevalent Hypertension (Top 500 PCs)	Clinical	0.263 (0.192-0.343)	0.895 (0.868-0.921)	0.232 (0.2-0.265)	0.385 (0.286-0.489)	0.722
	Clinical + Genotype	0.203 (0.138-0.274)	0.932 (0.91-0.953)	0.214 (0.182-0.245)	0.429 (0.313-0.552)	0.712

5.5.1 Ten-Year Incident Hypertension Using 500 Most Highly Associated SNPs

Using the 500 SNPs most highly associated with ten-year incident hypertension, the risk index methodology was able to build a Clinical risk index model with a sensitivity of 66.7% and a specificity of 46.8% and a Clinical + Genotype risk index model with a sensitivity of 53.9% and a specificity of 45.7%. The Clinical risk index model had an AUC of 0.567, while the Clinical + Genotype risk index model had an AUC of 0.475. The random forest model constructed with this data had an AUC of 0.811, and tuning the

class voting procedure produced predictive performance greater than either of the risk index models.

5.5.2 Ten-year Incident Hypertension Using Principal Components of Complete SNP Set

Using the top 500 principal components of the full set of SNPs from the Affymetrix 50K SNP genotyping platform, the risk index methodology was able to build a Clinical risk index model with a sensitivity of 60.8% and a specificity of 50.5% and a Clinical + Genotype risk index model with a sensitivity of 59.1% and a specificity of 54.4%. The Clinical risk index model had an AUC of 0.566, while the Clinical + Genotype risk index model had an AUC of 0.563. The random forest model constructed with this data had an AUC of 0.719, and tuning the class voting procedure produced predictive performance greater than either of the risk index models.

5.5.3 Ten-Year Incident Diabetes Using 500 Most Highly Associated SNPs

Using the 500 SNPs most highly associated with ten-year incident diabetes, the risk index methodology was able to build a Clinical risk index model with a sensitivity of 22.4% and a specificity of 90.1% and a Clinical + Genotype risk index model with a sensitivity of 10.4% and a specificity of 92.3%. The Clinical risk index model had an AUC of 0.722, while the Clinical + Genotype risk index model had an AUC of 0.683. The random forest model constructed with this data had an AUC of 0.905, and tuning the class voting procedure produced predictive performance greater than either of the risk index models.

5.5.4 Ten-Year Incident Diabetes Using Principal Components of Complete SNP Set

Using the top 500 principal components of the full set of SNPs from the Affymetrix 50K SNP genotyping platform, the risk index methodology was able to build a Clinical risk index model with a sensitivity of 20.5% and a specificity of 95.9% and a Clinical + Genotype risk index model with a sensitivity of 13.0% and a specificity of 97.4%. The Clinical risk index model had an AUC of 0.769, while the Clinical + Genotype risk index model had an AUC of 0.782. The random forest model constructed with this data had an AUC of 0.78. While tuning the class voting procedure could produce sensitivities and specificities better than either risk index model, however, in order to match the risk index models' misclassification rates (10.2% and 9.5% for the Clinical and the Clinical + Genotype risk index models, respectively) and PPV (30.6% and 30.4% respectively) the random forests model had a sensitivity and specificity comparable to that of the two risk index models.

5.5.5 Prevalent Hypertension Using 500 Most Highly Associated SNPs

Using the 500 SNPs most highly associated with prevalent hypertension, the risk index methodology was able to build a Clinical risk index model with a sensitivity of 45.9% and a specificity of 84.4% and a Clinical + Genotype risk index model with a sensitivity of 32.3% and a specificity of 84.6%. The Clinical risk index model had an AUC of 0.733, while the Clinical + Genotype risk index model had an AUC of 0.692. The random forest model constructed with this data had an AUC of 0.857, and tuning the class voting procedure produced predictive performance greater than either of the risk index models.

5.5.6 Prevalent Hypertension Using Principal Components of Complete SNP Set

Using the top 500 principal components of the full set of SNPs from the Affymetrix 50K SNP genotyping platform, the risk index methodology was able to build a Clinical risk index model with a sensitivity of 26.3% and a specificity of 89.5% and a Clinical + Genotype risk index model with a sensitivity of 20.3% and a specificity of 93.2%. The Clinical risk index model had an AUC of 0.722, while the Clinical + Genotype risk index model had an AUC of 0.712. The random forest model constructed with this data had an AUC of 0.74, and tuning the class voting procedure produced predictive performance greater than either of the risk index models.

5.5.7 Framingham Heart Study Conclusions

The real-world application of the risk index to the FHS provided performance that is noticeably worse than the results of the simulation study. In all but one case (diabetes using the top 500 principal components) the Clinical risk index model had an AUC greater than that of the Clinical + Genotype risk index model. Likewise, in all but that same case the random forest model produced an AUC greater than either of the risk index models. However, for ten-year incident diabetes and prevalent hypertension, the risk index methodology's performance, while less than that of random forests, is still fairly comparable, and does not have nearly the performance gap observed between the risk index models created for ten-year incident hypertension and the random forests created to predict ten-year incident hypertension.

One unexplained occurrence is the extremely poor performance that the risk index demonstrated in predicting ten-year incident hypertension. The poor performance seems to be isolated to this outcome, as the predictive performance for ten-year incident diabetes and prevalent hypertension were noticeably better. It seems unlikely that developing hypertension is an inherently more complex trait than developing diabetes, but this result could be explained by a combination of several factors. First, it may be that the variables included in this analysis are simply not the optimal predictors. The set of variables, however, did include the most commonly used predictors of hypertension risk: age, weight, and current blood pressure. Secondly, it may be that in this particular sample the individuals who develop hypertension do so through a number of heterogeneous pathways, while the individuals who develop diabetes do so in a relatively homogenous way. Lastly, the apparent increase in predictive performance may simply stem from the lower relative frequencies of incident diabetes and prevalent hypertension as compared to incident hypertension. The less frequent outcomes mean that a lower misclassification can be achieved by simply marking every individual as "low risk" and misclassifying every high risk individual. The much lower sensitivities and high specificities for incident diabetes and prevalent hypertension support the idea that the lower relative frequencies are playing a role in the apparent increase in performance.

One notable consistency between the FHS study and the simulation study is the relative improvement in predictive performance that occurs when the risk index methodology is presented with principal components data compared to when it is presented with SNP genotype data. Conversely, random forest performance decreases when using principal

components data compared to using SNP genotypes. This is discussed in further detail Section 5.7 below.

5.6 Methodological Limitations

Although the risk index offers reasonably good predictive performance in certain cases, there are some limitations to its use that mist be addresses. First, because of the way the data is divided first between the optimization set and independent testing set and then into cross-validation sets, the risk index is poorly suited to predicting risk for very uncommon outcomes.

Also, although the individual models created by the risk index from each bootstrap sample of the optimization set are much simpler to interpret than the decision trees making up a random forest, the method still relies on ensemble prediction, and so the end result is not one single model but rather a set of models, making interpretation much more difficult. Additionally, the fact that this method provides set of models that are difficult to interpret and the fact that the modeling was done with the intention of prediction and not biological interpretability means that it is not a strong starting point for further biological investigation. Although the most commonly selected variables may play some biological role in the disease process this is not necessarily the case.

5.7 Methodological Expansions and Future Directions

In several cases the overall predictive performance of the risk index procedure is noticeably poorer than that of random forests. However, examining the best aspects of the

risk index procedure and the difference between the approach of the risk index procedure and random forests offers some insights into how to improve the risk index procedure in the future. First and foremost, the primary difference between random forests and the risk index procedure is the presence of interactions. The decision tree structures that compose a random forest account for interactions at each level. Each new split finds the variable that best divides a subset of the data given the context of the previous splits. The risk index procedure, however, does not account for interactions, but rather composes a set of individual linear models that are applied equally to each individual. Although this is a shortcoming of the risk index procedure, some information about interactions would still be expected to be captured by this approach. Cheverud suggests that SNPs involved in interactions, when modeled univariately, typically show some marginal univariate effect (Cheverud, et al, 1995).

It would be possible, however, to include interaction terms directly in the risk index. In addition to the univariate logistic regression models currently used to estimate coefficients, it would be straightforward to include interaction terms. This must be done with caution, though, because with even a moderate number of variables the available number of interaction terms increases tremendously. Taking a cue from Cheverud, the variables considered could be limited to those that demonstrate a marginally significant effect, for example, those variables with a p-value from logistic regression modeling of 0.2 or lower. By reducing the search space in this way the effect of interactions on the risk index's predictive performance could be investigated manageably while still focusing on those variables most likely to demonstrate interaction.

As currently implemented the risk index methodology allows the user to specify the maximum number of variables to grow a risk index model to. For variables types (e.g., clinical covariates) that include only a small number of variables (e.g., 10 or 15) it would make sense to have these models grown to include all possible variables. For variables types (e.g., genotypes) that include a much larger number of variables (e.g., 500), growing the risk index model to include all possible variables is prohibitive. Instead, some reasonable maximum size should be used. In this dissertation, that maximum was 20 variables, a number chosen because it is sufficiently large but not prohibitively so. A better approach might be to develop a stopping rule of some type that followed the performance of the risk index model as it was built and stop the model building process when the performance gains drop below a certain threshold. Implementing this, however, would require a thorough investigation of the Brier score in real-world prediction, because if the score rate of change varies widely as more variables are added then a programmatically defined stopping rule may not be feasible.

Another possible enhancement to the risk index methodology would be the addition of a sequential prediction procedure, in which individuals with very high or very low predicted probabilities of developing disease are assigned a prediction while those with more intermediate values are not. The remaining, unclassified individuals could then be used to perform the risk index procedure a second time, hopefully producing more accurate predictions for these individuals. The underlying rationale for this approach is that while only one coefficient is estimated for each variable, if the sample of individuals

used to create the risk index is composed of individuals at varying stages of disease, the relationship between their outcome and a particular variable may not be constant across the entire sample. Taking a sequential approach might allow for these differing relationships to be explicitly considered leading to better predictive performance.

A further possible enhancement to the risk index methodology concerns the assignment of final predictions to new individuals. As it is currently implemented, a prediction is made for a new individual for each of the *n* Clinical or Clinical + Genotype risk index models created using the bootstrap samples of the optimization set. The final prediction (i.e., "high risk" or "low risk") is chosen based on a majority vote of these *n* predictions. This may not provide the best possible predictive performance, however, and could be addressed by a simple optimization step. Once all *n* Clinical or Clinical + Genotype risk index models have been built, they could be applied to the full optimization set. The proportion p^* of votes required to assign an individual a prediction of "high" risk could be examined over a range of values, for example from 0.05 to 0.95 in steps of 0.05. The Brier score, described in detail in Chapter 2, Section 2, could be calculated for each examined value of p^*, and the value which minimizes the Brier score could then be chosen and used to make predictions about any new individuals the risk index is applied to. For example, if $p^*=0.35$ gives the lowest Brier score, then for any new individual to whom the risk index the risk index is applied if 35% or more of the *n* Clinical or Clinical + Genotype risk index models predict the individual is at high risk then they would be assigned a prediction of "high risk". However, if 34% or fewer of the *n* Clinical or

Clinical + Genotype risk index models predict the individual is at high risk then they would be assigned a prediction of "low risk".

Alternatively, a weighted voting procedure could be implemented in which the predictions from each of the *n* Clinical or Clinical + Genotype risk index models could be weighted by the inverse of their Brier score (because the optimal Brier score value is the lowest, this would give greatest weight to those models with the lowest Brier score). The predictions could then be summed, and the prediction (i.e., "high risk" or "low risk") with the highest value would be the prediction assigned to a new individual.

5.8 Conclusion

Considering the results from the risk index procedure, a few trends become evident. First, and most importantly, the risk index methodology performs better when provided with a set of principal components from a large set of SNPs compared to when it is provided with a set of SNPs that have been selected because of high association with the outcome being examined. Considering the way in which the risk index methodology builds predictive models makes the reason behind this clear. The risk index methodology creates a linear combination, and so implicitly makes the assumption that each variable that is added to the model has the same effect for all individuals. However, it is well known that the effect of polymorphisms on a particular phenotype is influenced both by environmental factors as well as by the individual's other polymorphisms. Random forests, with its tree-based structure, accounts for this differential impact by selecting the variable with the best predictive power in the context of all of the variables that have

been previously selected. By using principal components, however, the risk index methodology is supplied with a set of variables that are completely uncorrelated and therefore are not subject to the interactive effects observed between SNPs. Conversely, random forests tend to perform more poorly when presented with principal components as compared to a set of highly associated SNPs. This is likely because the uncorrelated nature of these variables makes it difficult for random forests to identify the context-dependent effects it uses for classification.

Second, while the simulation studies showed that the Clinical + Genotype risk index models had significantly better average AUC than the Clinical risk index models in all but one case, the application of the risk index methodology to the FHS data showed an improvement in the AUC of the Clinical + Genotype risk index model over the AUC of the Clinical risk index model in only one case. It is not immediately apparent why this would be the case, but one possible explanation is the use of the 50K Affymetrix genotypes. Perhaps the 500K Affymetrix will offer better performance because of the improved genome coverage of this technology.

Overall, the goal of this dissertation was to develop the risk index described by Beer, at al. (Beer, et al, 2002) into a flexible risk prediction system capable of predicting an individual's risk of developing a particular chronic disease. The simulation study performed in Chapter 3 suggests that for very large datasets in certain circumstances the risk index methodology may perform quite well and may even outperform random forests. The application of the risk index methodology to the FHS data in Chapter 4 did

not provide strong support for the potential of the risk index methodology to have greater predictive performance than random forests in real-world datasets, but for both ten-year incident diabetes and prevalent hypertension the performance of the risk index methodology was comparable to that of random forests. Additionally, these results add support to the observation in the large-scale simulation data that using the top available principal components for the risk index procedure led to improved performance over the use of a set of highly associated SNPs.

Taken as a whole, this dissertation demonstrates the potential of genetic risk score methodologies for large-scale prediction. By including important statistical enhancements, such as forward selection to improve the model and the use of cross-validation and an independent testing set to reduce misclassification, the concept of a genetic risk score has been modified from a typically ad-hoc, small-scale procedure focusing on a small number of polymorphisms to a more robust, statistically focused method that is more in line with the requirements of a clinical risk prediction method. Although this method was not designed for and would likely be poor at identifying the biological underpinnings of chronic disease risk, as a pure risk prediction method it has the potential to improve public health through the identification of individuals most in need of interventions. Additionally, by incorporating both clinical covariates and genotypes, it represents an attempt to integrate information that is typically treated separately and to leverage the information in our genome for risk prediction.

References

Aaronson, K. D., Schwartz, J. S., Chen, T. M., Wong, K. L., Goin, J. E., and Mancini, D. M. 1997. Development and prospective validation of a clinical index to predict survival in ambulatory patients referred for cardiac transplant evaluation. *Circulation* **95:** 2660-2667.

Affymetrix. 2007. Affymetrix Genome-Wide Human SNP Nsp/Sty Assay 5.0.

Agarwal, A., Williams, G. H., and Fisher, N. D. 2005. Genetics of human hypertension. *Trends Endocrinol. Metab.* **16:** 127-133.

Antoniou, A., Pharoah, P. D. P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Loman, N., Olsson, H., Johannsson, O., Borg, Å., et al. 2003. Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies. *The American Journal of Human Genetics* **72:** 1117-1130.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., et al. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* **8:** 816-824.

Breiman, L. 1996. Random Forests. *Mach. Learn.* **45:** 5-32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. 1984. *Classification and Regression Trees.* Chapman & Hall, New York.

Burke, W. and Psaty, B. M. 2007. Personalized medicine in the era of genomics. *JAMA* **298:** 1682-1684.

Cheverud, J. M. and Routman, E. J. 1995. Epistasis and its contribution to genetic variance components. *Genetics* **139:** 1455-1461.

Cho, M. K. 2009. Translating genomics into the clinic: moving to the post-Mendelian world. *Genome Med.* **1:** 7.

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L.,Jr, Jones, D. W., Materson, B. J., Oparil, S., Wright, J. T.,Jr, et al. 2003. Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension* **42:** 1206-1252.

Collins, J. G. 1986. Prevalence of selected chronic conditions, United States, 1979-1981. *Vital Health Stats* **10:** .

Dawber, T. R., Meadors, G. F., and Moore, F. E.,Jr. 1951. Epidemiological approaches to heart disease: the Framingham Study. *Am. J. Public Health Nations Health* **41:** 279-281.

Dawber, T. R., Moore, F. E., and Mann, G. V. 1957. Coronary heart disease in the Framingham study. *Am. J. Public Health Nations Health* **47:** 4-24.

DeVol, R., Bedroussian, A., Charuworn, A., Chatterjee, A., Kim, I. K., Kim, S., and Klowden, K. 2007. An Unhealthy America: The Economic Burden of Chronic Disease. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316:** 1331-1336.

Edwards, T. L., Bush, W. S., Turner, S. D., Dudek, S. M., Torstenton, E. S., Schmidt, M., Martin, E., and Ritchie, M. D. 2008. Generating Linkage Disequilibrium Patterns in Data Simulations Using genomeSIMLA. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (eds. E. Marchiori and J. Moore), pp. 24-35. Springer Berlin.

Emery, J. and Hayflick, S. 2001. The challenge of integrating genetic medicine into primary care. *BMJ* **322:** 1027-1030.

Feinleib, M., Kannel, W. B., Garrison, R. J., McNamara, P. M., and Castelli, W. P. 1975. The Framingham Offspring Study. Design and preliminary data. *Prev. Med.* **4:** 518-525.

Fischer, M., Broeckel, U., Holmer, S., Baessler, A., Hengstenberg, C., Mayer, B., Erdmann, J., Klein, G., Riegger, G., Jacob, H. J., et al. 2005. Distinct heritable patterns of angiographic coronary artery disease in families with myocardial infarction. *Circulation* **111:** 855-862.

Golden Helix, I. HelixTree Software.

Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., et al. 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38:** 320-323.

Grosse, S. D. and Khoury, M. J. 2006. What is the clinical utility of genetic testing? *Genet. Med.* **8:** 448-450.

Haddow, J. E. and Palomaki, G. E. 2004. ACCE: A Model Process for Evaluating Data on Emerging Genetic Tests. In *Human genome epidemiology: a scientific foundation for using genetic information to improve health and prevent disease* (eds. M. J. Khoury, J. Little, and W. Burke), pp. 217-233. Oxford University Press, Oxford.

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., Jonasdottir, A., Sigurdsson, A., Baker, A., Palsson, A., et al. 2007. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316:** 1491-1493.

Hoffmann, K., Firth, M. J., Beesley, A. H., de Klerk, N. H., and Kees, U. R. 2006. Translating microarray data for diagnostic testing in childhood leukaemia. *BMC Cancer* **6:** 229.

Holtzman, N. A. and Watson, M. S. Promoting Safe and Effective Genetic Testing in the United States. Final Report of the Task Force on Genetic Testing, 1997.

Horne, B. D., Anderson, J. L., Carlquist, J. F., Muhlestein, J. B., Renlund, D. G., Bair, T. L., Pearson, R. R., and Camp, N. J. 2005. Generating genetic risk scores from intermediate phenotypes for use in association studies of clinically significant endpoints. *Ann. Hum. Genet.* **69:** 176-186.

Hosmer, D. and Lemeshow, S. 2000. *Applied Logistic Regression.* John Wiley & Sons, New York.

Jackson, J. F. 2003. *A user's guide to principal components.* Hoboken, N.J. : Wiley-Interscience, c2003., .

Klerk, M., Verhoef, P., Clarke, R., Blom, H. J., Kok, F. J., Schouten, E. G., and MTHFR Studies Collaboration Group. 2002. MTHFR 677C-->T polymorphism and risk of coronary heart disease: a meta-analysis. *JAMA* **288:** 2023-2031.

Knudsen, L. E., Loft, S. H., and Autrup, H. 2001. Risk assessment: the importance of genetic polymorphisms in man. *Mutat. Res.* **482:** 83-88.

Koelling, T. M., Joseph, S., and Aaronson, K. D. 2004. Heart failure survival score continues to predict clinical outcomes in patients with heart failure receiving β-blockers. *The Journal of Heart and Lung Transplantation* **23:** 1414-1422.

Lee, J. M. 2008. Why Young Adults Hold the Key to Assessing the Obesity Epidemic in Children. *Arch. Pediatr. Adolesc. Med.* **162:** 682-687.

Li, L. 2006. Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics* **22:** 466-471.

Lu, C., Van Gestel, T., Suykens, J. A., Van Huffel, S., Vergote, I., and Timmerman, D. 2003. Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines. *Artif. Intell. Med.* **28:** 281-306.

McEwan, P., Williams, J. E., Griffiths, J. D., Bagust, A., Peters, J. R., Hopkinson, P., and Currie, C. J. 2004. Evaluating the performance of the Framingham risk equations in a population with diabetes. *Diabet. Med.* **21:** 318-323.

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., Hinds, D. A., Pennacchio, L. A., Tybjaerg-Hansen, A., Folsom, A. R., et al. 2007. A common allele on chromosome 9 associated with coronary heart disease. *Science* **316:** 1488-1491.

Milne, R., Gamble, G., Whitlock, G., and Jackson, R. 2003. Framingham Heart Study risk equation predicts first cardiovascular event rates in New Zealanders at the population level. *N. Z. Med. J.* **116:** U662.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. 2005. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21:** 3301-3307.

Morrison, A. C., Bare, L. A., Chambless, L. E., Ellis, S. G., Malloy, M., Kane, J. P., Pankow, J. S., Devlin, J. J., Willerson, J. T., and Boerwinkle, E. 2007. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. *Am. J. Epidemiol.* **166:** 28-35.

National Cancer Institute. Breast Cancer Risk Assessment Tool. **2009:** .

National Center for Health Statistics. Crude and Age-Adjusted Incidence of Diagnosed Diabetes per 1,000 Population Aged 18–79 Years, United States, 1980–2007. *Centers for Disease Control and Prevention (CDC), Division of Health Interview Statistics* .

National Diabetes Data Group. 1979. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. National Diabetes Data Group. *Diabetes* **28:** 1039-1057.

Opitz, D.; Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* **11**: 169–198.

Org, E., Eyheramendy, S., Juhanson, P., Gieger, C., Lichtner, P., Klopp, N., Veldre, G., Doring, A., Viigimaa, M., Sober, S., et al. 2009. Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. *Hum. Mol. Genet.*

Patterson, N., Price, A. L., and Reich, D. 2006. Population structure and eigenanalysis. *PLoS Genet.* **2:** e190.

Pawitan, Y., Bjohle, J., Amler, L., Borg, A. L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., et al. 2005. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.* **7:** R953-64.

Permutt, M. A., Wasson, J., and Cox, N. 2005. Genetic epidemiology of diabetes. *J. Clin. Invest.* **115:** 1431-1439.

Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., Chen, C. M., Bild, A., Iversen, E. S., Huang, A. T., et al. 2004. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl. Acad. Sci. U. S. A.* **101:** 8431-8436.

Pleis, J. R. and Lethbridge-Cejku, M. 2007. Summary Health Statistics for U.S. Adults: National Health Interview Study, 2006. *Vital Health Stats* **10:** .

Plomin, R. and Schalkwyk, L. C. 2007a. Microarrays. *Dev. Sci.* **10:** 19-23.

Price, A. L., Butler, J., Patterson, N., Capelli, C., Pascali, V. L., Scarnicci, F., Ruiz-Linares, A., Groop, L., Saetta, A. A., Korkolopoulou, P., et al. 2008. Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* **4:** e236.

Ramachandran, S., French, J. M., Vanderpump, M. P., Croft, R., and Neary, R. H. 2000. Using the Framingham model to predict heart disease in the United Kingdom: retrospective study. *BMJ* **320:** 676.

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.* 455-466.

Rebbeck, T. R., Spitz, M., and Wu, X. 2004. Assessing the function of genetic variants in candidate gene association studies. *Nat. Rev. Genet.* **5:** 589-597.

Rice, T., Rankinen, T., Province, M. A., Chagnon, Y. C., Perusse, L., Borecki, I. B., Bouchard, C., and Rao, D. C. 2000. Genome-wide linkage analysis of systolic and diastolic blood pressure: the Quebec Family Study. *Circulation* **102:** 1956-1963.

Rigat, B., Hubert, C., Alhenc-Gelas, F., Cambien, F., Corvol, P., and Soubrier, F. 1990. An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. *J. Clin. Invest.* **86:** 1343-1346.

Rosamond, W., Flegal, K., Furie, K., Go, A., Greenlund, K., Haase, N., Hailpern, S. M., Ho, M., Howard, V., Kissela, B., et al. 2008. Heart disease and stroke statistics--2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation* **117:** e25-146.

Russell, L. B. 2009. Preventing chronic disease: an important investment, but don't count on cost savings. *Health. Aff. (Millwood)* **28:** 42-45.

Ryall, R. G., Staples, A. J., Robertson, E. F., and Pollard, A. C. 1992. Improved performance in a prenatal screening programme for Down's syndrome incorporating serum-free hCG subunit analyses. *Prenat. Diagn.* **12:** 251-261.

Schunkert, H., Gotz, A., Braund, P., McGinnis, R., Tregouet, D. A., Mangino, M., Linsel-Nitschke, P., Cambien, F., Hengstenberg, C., Stark, K., et al. 2008. Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. *Circulation* **117:** 1675-1684.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316:** 1341-1345.

Seldin, M. F., Shigeta, R., Villoslada, P., Selmi, C., Tuomilehto, J., Silva, G., Belmont, J. W., Klareskog, L., and Gregersen, P. K. 2006. European population substructure: clustering of northern and southern populations. *PLoS Genet.* **2:** e143.

Sing, C. F., Stengard, J. H., and Kardia, S. L. 2003. Genes, environment, and cardiovascular disease. *Arterioscler. Thromb. Vasc. Biol.* **23:** 1190-1196.

Splansky, G. L., Corey, D., Yang, Q., Atwood, L. D., Cupples, L. A., Benjamin, E. J., D'Agostino RB, S., Fox, C. S., Larson, M. G., Murabito, J. M., et al. 2007. The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am. J. Epidemiol.* **165:** 1328-1335.

Spurgeon, S. E., Hsieh, Y. C., Rivadinera, A., Beer, T. M., Mori, M., and Garzotto, M. 2006. Classification and regression tree analysis for the prediction of aggressive prostate cancer on biopsy. *J. Urol.* **175:** 918-922.

Stephenson, A. J., Smith, A., Kattan, M. W., Satagopan, J., Reuter, V. E., Scardino, P. T., and Gerald, W. L. 2005. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer* **104:** 290-298.

Tatsioni, A., Zarin, D. A., Aronson, N., Samson, D. J., Flamm, C. R., Schmid, C., and Lau, J. 2005. Challenges in systematic reviews of diagnostic technologies. *Ann. Intern. Med.* **142:** 1048-1055.

Thuerigen, O., Schneeweiss, A., Toedt, G., Warnat, P., Hahn, M., Kramer, H., Brors, B., Rudlowski, C., Benner, A., Schuetz, F., et al. 2006. Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer. *J. Clin. Oncol.* **24:** 1839-1845.

Thuluvath, P. J., Yoo, H. Y., and Thompson, R. E. 2003. A model to predict survival at one month, one year, and five years after liver transplantation based on pretransplant clinical characteristics. *Liver Transpl.* **9:** 527-532.

Tu, K., Chen, Z., Lipscombe, L. L., and Canadian Hypertension Education Program Outcomes Research Taskforce. 2008. Prevalence and incidence of hypertension from 1995 to 2005: a population-based study. *CMAJ* **178:** 1429-1435.

Turner, S. T., Boerwinkle, E., and Sing, C. F. 1999. Context-dependent associations of the ACE I/D polymorphism with blood pressure. *Hypertension* **34:** 773-778.

Wang, Y., O'Connell, J. R., McArdle, P. F., Wade, J. B., Dorff, S. E., Shah, S. J., Shi, X., Pan, L., Rampersaud, E., Shen, H., et al. 2009. From the Cover: Whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proc. Natl. Acad. Sci. U. S. A.* **106:** 226-231.

Ward, M. M., Pajevic, S., Dreyfuss, J., and Malley, J. D. 2006. Short-term prediction of mortality in patients with systemic lupus erythematosus: classification of outcomes using random forests. *Arthritis Rheum.* **55:** 74-80.

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22:** 209-212.

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22:** 209-212.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. 1998. Prediction of coronary heart disease using risk factor categories. *Circulation* **97:** 1837-1847.

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R., Rayner, N. W., Freathy, R. M., et al. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316:** 1336-1341.

Ziogas, D. and Roukos, D. H. 2009. Genetics and personal genomics for personalized breast cancer surgery: progress and challenges in research and clinical practice. *Ann. Surg. Oncol.* **16:** 1771-1782.