EVALUATION OF NOVEL BIOMARKERS FOR CORONARY ARTERY DISEASE AMONG SYMPTOMATIC PATIENTS: STATISTICAL METHODOLOGY AND APPLICATION

by

Daniel Lans

B.A., University of Minnesota - Twin Cities, 2011

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Daniel Lans

It was defended on

April 11th, 2013

and approved by

William LaFramboise, PhD, Associate Professor, Department of Pathology, School of Medicine, University of Pittsburgh

Andriy Bandos, PhD, Research Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Thesis Director: Gary Marsh, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

EVALUATION OF NOVEL BIOMARKERS FOR CORONARY ARTERY DISEASE AMONG SYMPTOMATIC PATIENTS: STATISTICAL METHODOLOGY AND APPLICATION

Daniel Lans, M.S.

University of Pittsburgh, 2013

ABSTRACT

Proteomics has led to the discovery of several biomarkers within an individual's bloodstream that can be used in the diagnostic process for disease. Identification of novel biomarkers have a significant impact in the area of public health, with the potential to replace existing diagnostic methods that are complicated, costly, and that pose considerable risk to the patient. Cardiac catheterization, the current diagnostic method for coronary artery disease, is such an invasive procedure. An over-abundance of negative test results leads to the inquiry whether exposing all symptomatic patients to the procedure is in a physician's best interest.

A statistical analysis involving multivariate logistic regression and evaluation of predictive models identified a panel of biomarkers that can be used to classify patient with coronary artery disease and those with "normal" coronary arteries. This panel was used in conjunction with common clinical risk factors for heart disease to examine the added predictive power of the multi-marker panel when combined with clinical characteristics.

A four-marker panel consisting of OPN, IL1β, Apo-B100, and Fibrinogen were found to be statistically significant predictors of coronary artery disease in a predictive logistic model adjusting for clinical risk factors, diabetes status and smoking status. The ability to identify

patients that did not have clinically relevant coronary disease based on currently used clinical risk factors increased greatly, from zero to approximately thirty percent of the patients, with the inclusion of the biomarker panel.

The use of a blood screening test for the diagnosis of coronary artery disease among symptomatic patients can limit the number of unnecessary cardiac catheterizations, reducing healthcare costs and patient risks associated with the invasive nature of the procedure. However, with such a test, there may be some discrimination error present, and the cost of misdiagnosing a patient with clinically relevant coronary artery disease needs to be weighed against the benefits of the test.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Meaning
ACRP-30	Adiponectin
APO (Apo-A1, Apo-B100)	Apolipoprotein
AUC	Area under the curve
B-H	Benjamini-Hochberg
CAD	Coronary artery disease
CRP	C-reactive protein
EKG	Electrocardiogram
ER	Emergency Room
E-Selectin	Endothelial leukocyte adhesion molecule
FDR	False discovery rate
IFNγ	Interferon-gamma
IL (IL6, IL10, IL1β)	Interleukin
L-Selectin	leukocyte selectin
MCP	Monocyte chemoattractant protein
MLE	Maximum likelihood estimator
MMP (MMP1, MMP7)	Matrix metalloproteinase protein
MPO	Myeloperoxidase
NT-pBNP	N-terminal fragment protein precursor brain natriuretic peptide
OPN	Osteopontin
PECAM	Platelet endothelial cell adhesion molecule
ROC	Receiver operating characteristic
SN	Sensitivity
SP	Specificity
TIMP	Tissue inhibitor of metalloproteinase
TM	Thrombomodulin
TNF	Tumor necrosis factor
VCAM	Vascular cell adhesion molecule

PREFACE

I would like to thank my family for all their support through my educational experiences and helping me reach such a focal part of my life – my mom and dad, my brother, Joshua, and my sisters, Alexa and Andrea.

Much gratitude goes out to my committee members – to Bill LaFramboise, whom I met through playing handball, who provided me with such a meaningful and thought-provoking public health problem to research that will undoubtedly lead to future success; to Dr. Gary Marsh for his outstanding advisement throughout my graduate program and essential role in helping me succeed as a biostatistician; and to Dr. Andriy Bandos, whose expertise, without question, was indispensable throughout the process of this thesis, and furthered my knowledge on the subject matter.

Last, but not least, I would like to thank my friends and classmates at the University of Pittsburgh Graduate School of Public Health, whose collaboration and camaraderie was essential to my success as a student. I could not have had as valuable an experience without the support of these individuals.

1.0 INTRODUCTION

Biological markers, more commonly referred to as “biomarkers,” refer to observable measurements derived from a patient that can be used to describe certain biological developments, including disease status, risk, or prognosis for that patient. According to an NIH working group, the definition of a biomarker is standardized to be “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [1]. These biomarkers can be classified into separate categories based on their clinical properties. For the purpose of this paper, the term biomarkers will be used to denote biological components that indicate disease status of an individual; they are disease biomarkers consisting of diagnostic properties. More specifically, this paper is interested in circulating biomarkers ascertained from advanced proteomics methods.

Previously, biomarkers were commonly found to be simple physiological measurements, such as one’s blood pressure or heart rate, but have now evolved into complex imaging techniques and multi-marker genomic/proteomic panels [2]. This revolution allows researchers to interrogate blood and serum samples for potential markers that may not correspond with a patient’s sense of well-being, but are evidently affecting the disease status of an individual. This method of diagnosis is especially attractive in areas where the incidence of disease is high and current diagnostic methods are both costly and invasive in nature.

Novel discoveries in biomarker research have a significant impact in the area of public health, providing alternative diagnostic methods for currently used invasive procedures, thus reducing the existing medical complications and economic burden of such procedures.

1.1 CORONARY ARTERY DISEASE

The application of biomarker research related to coronary artery disease (CAD) is the primary focus of this thesis. In the United States, CAD is the leading cause of mortality, accounting for about one of every six deaths. In 2009, 386,324 deaths due to CAD were recorded [3]. The disease occurs when the coronary arteries harden and narrow, due to atherosclerosis, preventing oxygen from reaching the heart. The most common symptom of coronary artery disease is angina, or chest pain, but a patient may also experience fatigue, light-headedness, or shortness of breath. However, sometimes an individual will experience myocardial infarction (a heart attack) or immediate death without having any of the previous symptoms. The use of screening procedures allows for early detection of the disease so successful interventions can be performed to reduce chances of infarct or death.

The current diagnostic method for CAD involves invasive coronary angiography, where medical imaging is used to detect a dye injected into the arteries by way of cardiac catheterization. This involves the insertion of a catheter, a thin and flexible tube, through a brachial or femoral artery and up to the aorta and chamber of the heart, where the dye is then released into the bloodstream [4]. Coronary angiography has been a highly efficient screening procedure for the detection of coronary stenosis and is regarded as the current gold standard for determining clinically significant CAD among symptomatic patients, but complications arising

from the procedure have been criticized [5]. Several common complications include arrhythmias (mostly attributed to anxiety about the procedure), bleeding and hematoma around the femoral artery region, allergic reactions to the injected dye, and anesthetic complications [6, 7]. Furthermore, a patient undergoing the catheterization procedure is exposed to localized x-ray radiation for an extended period of time, increasing the risk of cancer and other genetic effects [8].

The alarming rate of CAD has led to an increase in the number of cardiac catheterizations performed in hospitals, thus increasing the incidence of these complications. Almost half of the patients referred for catheterizations are found to have insignificant coronary lesions, and are unnecessarily exposed to procedural complications [9]. One alternative to the invasive procedure would include the identification of biomarkers existing in a patient's bloodstream. Biomarker discovery regarding CAD would reveal a safer, more pragmatic diagnostic procedure than coronary angiography with cardiac catheterization.

1.1.1 Symptomatic Patients

Patients referred for catheterization come in to the emergency room (ER) or heart clinic showing symptoms of CAD. This cohort of patients excludes those that have experienced a cardiac event, such as myocardial infarction, who skip the ER and are immediately sent for percutaneous intervention. For the patients received in the ER or heart clinic, an assessment of the individual is performed to determine the pretest probability of having CAD [10]. This may include looking at a patient's medical history and existing clinical characteristics (obesity, smoking, age, etc.). If the pretest probability for CAD is low to intermediate, a non-invasive stress test and/or

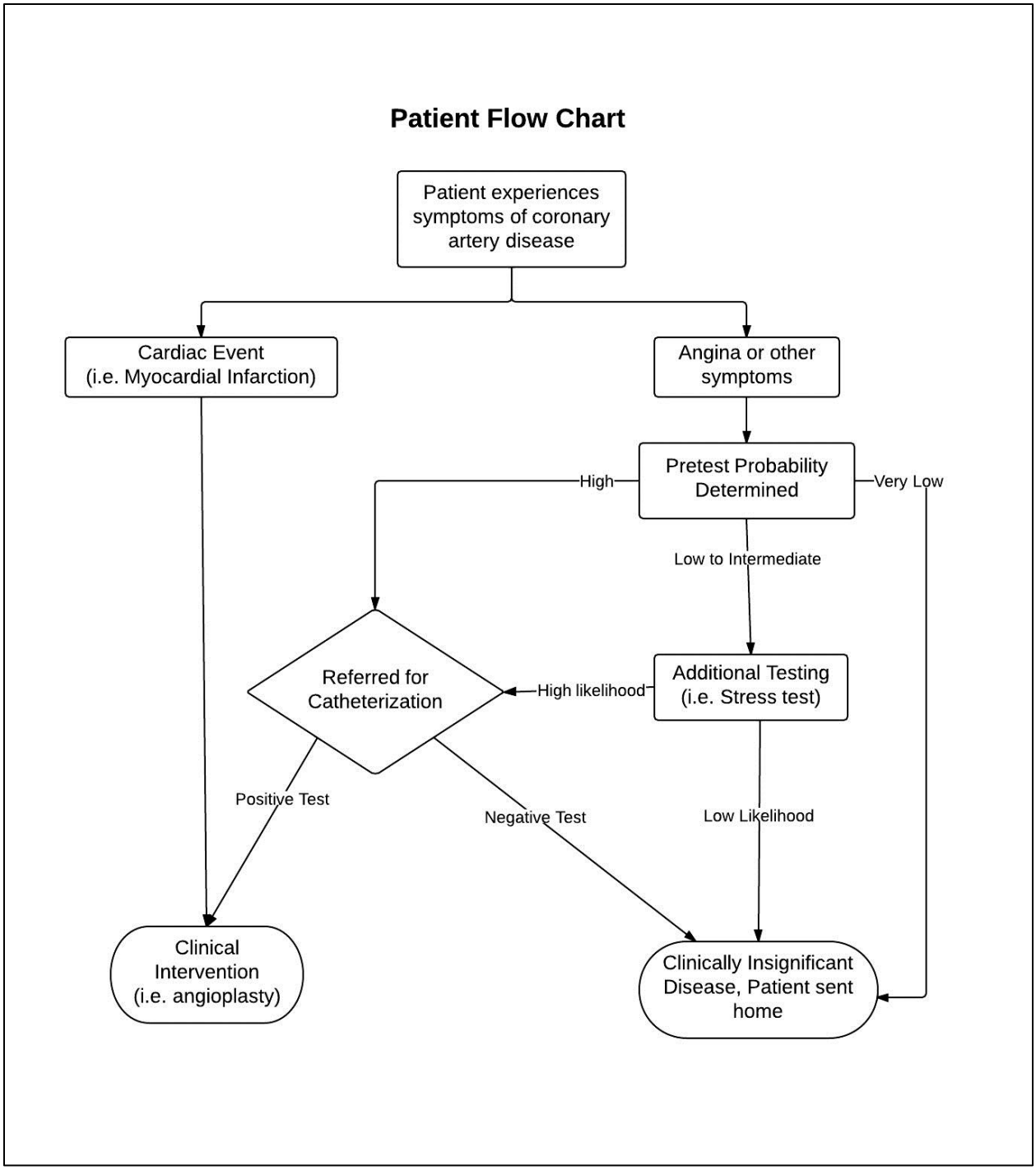

Figure 1. Patient experience during the diagnostic process of coronary artery disease

electrocardiogram (EKG) may also be taken into consideration to determine the likelihood of disease. Following the conclusion of an insignificant pretest probability, the patient may be

treated for his or her symptoms and sent home, being reassured there is insignificant evidence of disease. If the pretest probability and likelihood for CAD is high, a patient will be sent for catheterization to view any blockages in the arteries and determine the significance of CAD. Figure 1 demonstrates the patient flow from experiencing symptoms to diagnosis.

1.1.2 Biomarkers and Coronary Artery Disease

Cardiovascular disease is often accompanied with sources of inflammation and plaque instability, followed by thrombosis within the arterial regions of the heart. Resulting ischemia may be followed by remodeling of the heart's ventricles. Investigation into the biological pathways for atherosclerosis involving inflammation, plaque instability, thrombosis, and remodeling of the extracellular matrix, has identified several biomarkers associated with acute coronary syndromes [11]. An up-regulation of proteins responding to these biological processes can provide useful diagnostic information for cardiac complications. For example, troponin is widely one of the most popular biomarkers for heart disease, where elevated levels of this protein points to the extent of injury to the heart during myocardial infarction [12]. However, there has been less success in regards to the clinical application of biomarkers to determine the degree of coronary artery disease among symptomatic patients. One suggestion is that a combination of protein changes in serum can address the severity of disease better than previous attempts that have focused on single markers [13-15]. Previous research has supported moderate improvement in risk models of coronary disease by implementing multiple biomarkers among other populations [16]. Adaptation of this multi-marker approach may point to a specific set of markers that would improve risk assessment among symptomatic patients, as defined by this paper.

Identifying biomarkers in serum of symptomatic patients could lead to the development of a clinical assay to use as a diagnostic method for those patients with a high pretest probability and likelihood for CAD. Instead of referring patients for catheterization based on a clinical assessment, stress test, and/or EKG, a less costly blood assay can be performed to filter out symptomatic patients that would otherwise be diagnosed negative for CAD. Data from a clinical study is analyzed later in this paper to demonstrate the effectiveness of using serum protein profiles and clinical characteristics as biomarkers for clinically relevant CAD.

2.0 LITERATURE REVIEW

Several methods for assessing biomarkers are currently used, and there is some debate in the best way to measure the predictive power of new biomarkers. Logistic regression is a common classification technique that is generally employed for problems involving biomarkers, while receiver operating characteristic (ROC) curves have been used to evaluate the predictive ability of these new biomarkers. In this literature review, these common statistical methods for classification of disease and the evaluation of biomarkers are covered.

2.1 STATISTICAL METHODS FOR PREDICTION

2.1.1 Odds ratios

Before delving into any of the more advanced statistical methods, it is important to grasp the concept of the odds ratio and how it used in clinical interpretation. In biomarker experiments, it is often desired to know the probability of an event, or the probability a patient is diagnosed with disease. Odds can then be defined as the ratio of the probability the event will occur versus the probability the event cannot occur [17]. In terms of patients who are symptomatic for coronary artery disease:

$$Odds = \frac{Probability\ the\ patient\ has\ coronary\ disease}{Probability\ the\ patient\ does\ not\ have\ coronary\ disease}$$

If p then equals the probability of disease for a patient, 1-p would equal the probability the patient does not have disease and the above equation can be reformulated as

$$Odds = \frac{p}{1-p}$$

For example, if a clinical test using biomarkers determined a patient to have a 60% risk of CAD, the odds this patient actually has the disease would be 60% / 40%, or 1.5. This means the patient is 50% more likely to be diagnosed with disease than disease-free by a gold standard assessment, i.e. coronary angiography. If the odds of disease equal 1, this means the patient has the same chance of being diagnosed positive or negative, and odds less than 1 means the patient has a lesser chance to be diagnosed with disease according to the angiographic test.

The odds ratio compares the odds of an event occurring between two patients:

$$Odds\ Ratio = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

If the odds of CAD for patient one was 1.5, and the odds of disease for patient two was 1.2, the odds ratio would then be 1.5 / 1.2, or 1.25. This means patient 1 has a 25% higher chance of having a positive result from coronary angiography than patient 2. This concept of odds ratios is carried over for logistic regression, and it will be shown how predicted probabilities of disease for patients can be derived from odds ratios.

2.1.2 Binary Logistic Regression

Binary logistic regression is a common statistical method for predicting the classification of subjects according to a dichotomous outcome. Many times, in health sciences, the goal is to differentiate those with and without a specific disease. Logistic regression has the ability to model the probability of disease, or any categorical outcome, and how the addition or subtraction of predictor variables affects that probability [18]. As opposed to linear regression methods, the logistic regression model estimates a linear function based on the log-odds of disease. This is because the relationship between the probability of the outcome and its predictors is usually nonlinear [19]. A unit increase in the predictor will have less of an impact when the probability of disease is close to 0 or 1, forming a logistic function, demonstrated by figure 2. Where $\hat{p}$ equals the probability of disease and β_0 and β represent numerical coefficients, the analytical form of this logit function is

$$\hat{p} = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}$$

. The regression model can then be written out as

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta^T x$$

Notice that the left side of the equation is the natural log of odds equation specified in section 2.1.1. Exponentiation both sides of the regression model then gives odds ratio estimates for the β's. With some simple algebra, the regression model can then be remodeled to match the logit function to calculate the estimated probability of disease.

Coefficients for main effects in the logistic regression model are generated through maximum likelihood estimators (MLE). Simply put, the MLE is an estimate for a parameter that

maximizes the probability of the outcome, or maximizing the agreement between observed data and expected values from the model [19]. Therefore, coefficients in a logistic regression model for coronary artery disease would be derived so that the most probable set of predictions can be made for the observed data. Statistical software can be used to easily compute these values when fitting a predictive model.

How well a model fits the data is determined by the deviance of the model. A simple way to describe deviance is the difference between the observed values and the expected values from the logistic model. The general idea would be to fit a model with a set of predictors that produces the lowest deviance, indicating a closer fit to the model [19].

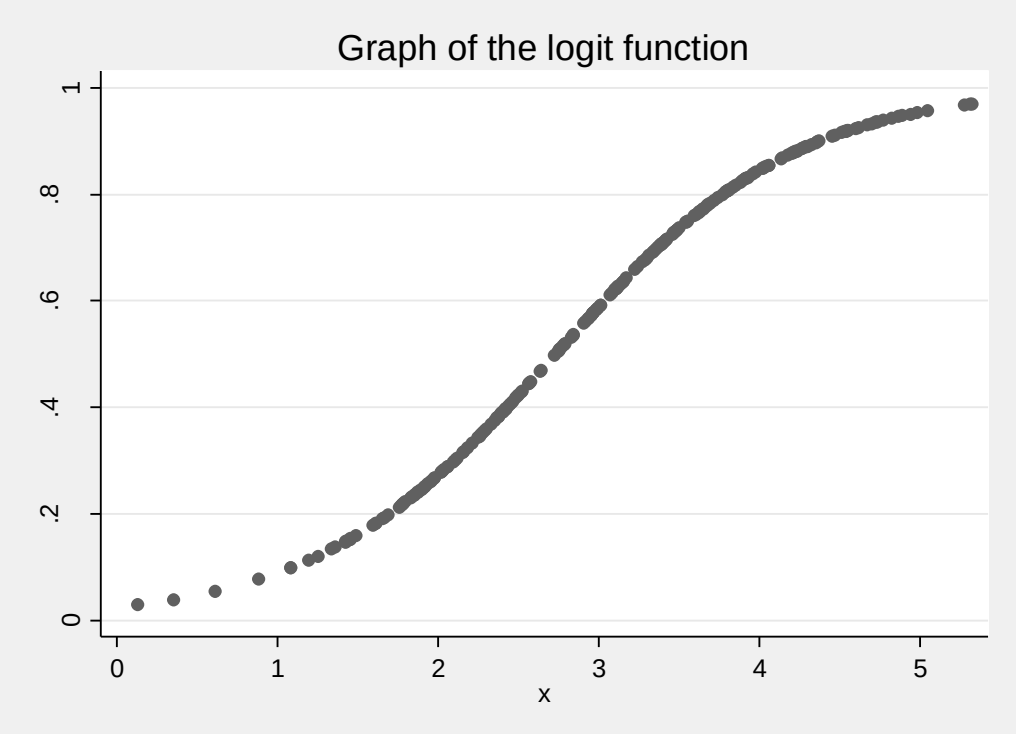

Figure 2. Graphical form of a logit function
A unit increase in x has little impact on the probability of disease when the probability is close to 0 or 1.

When combining several markers for prediction, which is becoming more and more popular with proteomic and genomic technologies, logistic regression serves as a useful tool for finding the best set of markers to use as a diagnostic tool [20]. However, using multiple signature of biomarkers for diagnostic tests leads to more difficulties in selecting the most predictive set of markers from a large list of candidates [21]. A simple approach to the variable selection process to obtain the most parsimonious model is forward stepwise selection, where variables are entered into the final model based on statistically significant relationships with the outcome. This differs from standard forward selection because variables that have entered the model in the stepwise method will also have potential to exit the model, based on the statistical significance of their relationship with the outcome once new predictors are added.

As previously discussed, the main assumption of logistic regression is that predictors within the model hold a linear relationship with the log-odds of the outcome. This is usually straightforward when dealing with categorical or ordinal predictors, but presents some difficulties when predictors are in a continuous form.

2.1.3 Continuous and Categorical Predictors

Protein biomarkers are usually reported on a continuous scale to reflect the concentration of the protein in a subject's serum. While the assumption of normality does not necessarily need to hold for variables used in logistic regression, if a predictor is normally distributed for both levels of the outcome, the logistic regression model will be better at describing a linear relationship between the predictor and the outcome [19]. However, most biomarkers have heavily skewed distributions and do not meet the assumptions of normality. For these variables, a log transformation to the data generally approximates a normal distribution [20]. More importantly,

proteins measured on the log-scale will produce more interpretable odds ratios than if they were left in their original scale. Figure 3 shows histograms of the frequency distributions of osteopontin (OPN) concentrations among 239 patients enrolled in a cohort study. It is seen that a natural logarithm transformation applied to the data (right), gives a much better approximation to the normal curve than the regular, skewed data (left).

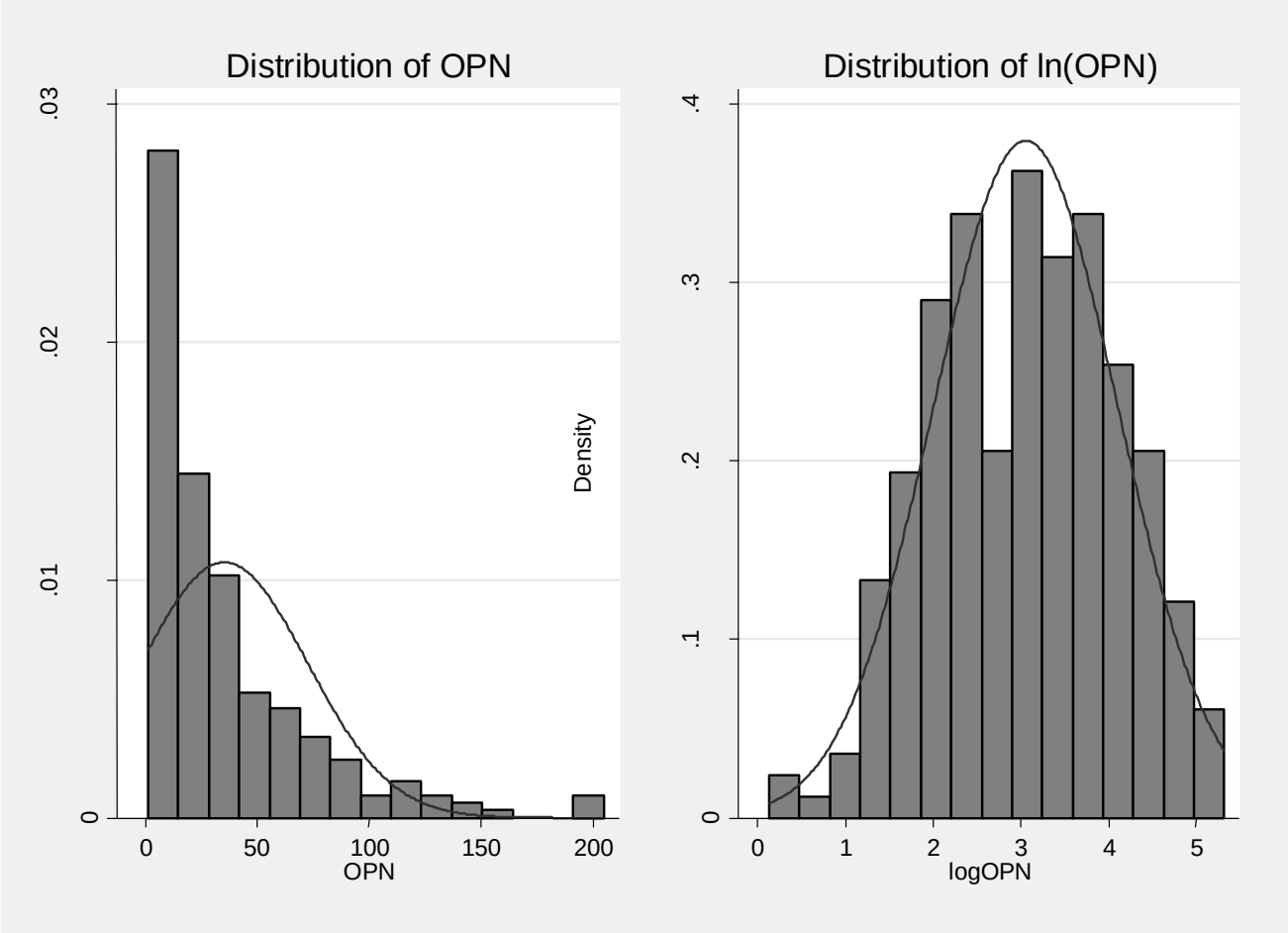

Figure 3. Histograms of the distribution of OPN and ln(OPN)
OPN on the original scale (left) has a heavily skewed right distribution. A natural logarithmic transformation to the variable more closely approximates a normal curve (right).

If the log-linear assumption of logistic regression is violated, the predictive model will produce inaccurate estimates for the odds-ratios. Dichotomization or categorization of continuous predictors is commonly used in exploratory stages to fit logistic regression models when the linear relationship is questionable [22].

Categorization of variables into two or more categories is often done in medical research as a way to simplify the interpretation of odds ratios, creating regression models with step functions. Factoring by tertiles, quartiles, or quintiles is commonly seen in proteomic analysis when clinically relevant thresholds are not available [20]. This crude approach to categorization can be used to identify a log-linear relationship between the outcome and its predictor.

Moving forward with the factored continuous variables may present complications for clinical interpretation. First, the cutpoints used to factor continuous variables need to be explicitly defined when translating results into other research. Second, categorization of these variables discards information that may be relevant to the analysis. It is improbable that a subject's risk for disease will suddenly increase when one of the thresholds is crossed. If a linear assumption is validated, continuous variables will provide more powerful statistical results than their factored counterparts. Therefore, categorization of continuous variables is valid in an exploratory process, but final analysis should be conducted on the continuous form of the data. If the data is truly expected to be non-linear with respect to the log-odds of the outcome, some more advanced modeling techniques can be used to address the situations.

2.1.3.1 Fractional Polynomials

The idea of fractional polynomials in regression is discussed in detail by Royston and Altman [23]. Fractional polynomials are used to transform continuous data to investigate improvements in model fit, compared to the straight line model $\beta_0 + \beta_1 x$, or in other words, the model without

a transformation applied to the covariate x [24]. Transformations are usually applied to continuous covariates in the event there is a nonlinear relationship with the dependent variable, in order to obtain better estimates for model coefficients. In regular polynomial regression, the independent variable, x, is taken to an nth power (i.e. x^2, x^3, x^4, etc.) to describe a nonlinear relationship it may have with the dependent variable (in the case of logistic regression, this would be the log-odds). The fractional polynomial method extends the current theory of polynomial regression by including negative and fractional powers for the covariates, usually from the set -2, -1, -0.5, 0, 0.5, 1, 2, 3 [22]. In this sequence of powers, 0 refers to a natural log transformation. Deviance of the model with the nonlinear transformation is compared to the deviance of the model without the transformation, and this difference is formally tested to determine whether or not the transformation should be used [23]. Using one of these fractional polynomial transformations can be referred to as first-degree polynomials.

Fitting a main effect with first-degree polynomials may not provide enough flexibility to fit a model, and in such cases, second-degree polynomials can be explored [24]. If p is a first-degree fractional polynomial, a regression model with the nonlinear transformation can be written: $\beta_0 + \beta_1 x^p$. Where q is a second-degree fractional polynomial, the equation featuring the transformations would be written: $\beta_0 + \beta_1 x^p + \beta_1 x^q$. Both p and q are chosen from the same set of values, -2, -1, -0.5, 0, 0.5, 1, 2, 3 and all possible combinations of first and second degree polynomials are explored to provide the model fit.

Fractional polynomials provide a flexible and more practical approach to modeling continuous covariates in an appropriate functional form, as opposed to categorization of these covariates which may present several disadvantages and statistically significant loss of information. Provided a nonlinear relationship exists between the dependent and independent

variables, fitting a logistic regression model with fractional polynomials will produce more accurate odds-ratios for covariates within the model.

2.2 EVALUATION OF BIOMARKERS

One of the main uses of biomarkers is to make a diagnosis more reliable, more rapidly, and inexpensive compared to existing methods [25]. However, proper evaluation of a biomarker needs to be assessed before it can be determined useful. Clinicians looking to implement biomarkers into their clinical practice are most concerned with diagnostic accuracy. Diagnostic accuracy refers to the ability of a biomarker to classify subjects into clinically relevant groups and is the general purpose of biomarker analyses [26]. That is, can a biomarker accurately distinguish between those patients that truly do have a disease and those that in fact do not have the disease? Several statistical tools can be used to measure diagnostic accuracy to determine if it should be used in clinical practice.

2.2.1 Multiple Comparisons

Valid biomarkers should have a greater presence in the affected individuals than the unaffected individuals. [27]. Statistical comparative tests such as the Student's t-test or Wilcoxon rank-sum test are commonly used to detect statistically significant differences in biomarker concentrations among disease categories. In most research studies, multiple biomarkers are assessed from a sample, creating inflation in type I errors. The Bonferroni adjustment for p-values is a common method to use for multiple comparisons, but when the number of comparisons is large, this

method can be too conservative. Letting k equal the number of comparisons and α equal the selected type I error rate, the Bonferroni method adjusts the error rate by α/k. For large values of k, the adjustment becomes radically small, reducing the chance that any hypothesis be rejected.

Controlling for the false discovery rate (FDR) using a method proposed by Benjamini and Hochberg [28] is more practical for proteomic or genomic experiments comparing several potential biomarkers among patients [29, 30]. In this method, unadjusted p-values are first ordered from smallest to largest, and the rank is recorded. The adjustment to the error rate is calculated as $\alpha * m/k$ for each p-value, where α is the error rate, m is the rank, and k is the total number of comparisons made. The adjusted p-value is referred to as the Q-value in the Benjamini-Hochberg approach. This correction is more suitable for experiments with large k as it is less likely to overlook statistically significant results that may be masked by more conservative approaches.

2.2.2 Sensitivity and Specificity

Sensitivity and specificity are common statistical measures used to assess the diagnostic accuracy of a biomarker [25]. Sensitivity refers to the probability of identifying a disease when it is actually present in the individual, whereas specificity refers to the probability of correctly dismissing individuals without the disease. Results for sensitivity and specificity can be classified as true positives (sensitivity), false positives (1-specificity), true negatives (specificity), and false negatives (1-sensitivity). Figure 4 illustrates these measures.

Sensitivity and Specificity

		Disease Status		
		+	-	
Test Result	+	**True Positive (A)**	**False Positive (B)**	Total, positive
	-	**False Negative (C)**	**True Negative (D)**	Total, negative
		Total, disease	Total, no disease	

$$Sensitivity = \frac{A}{A + C} \qquad False\ Negative\ Rate = \frac{C}{A + C}$$

$$Specificity = \frac{D}{B + D} \qquad False\ Positive\ Rate = \frac{B}{B + D}$$

Figure 4. Sensitivity and specificity calculations for a diagnostic test

In most cases, it is desired to find a certain threshold that will maximize both sensitivity and sensitivity. For diagnostic tests, this will provide the most accurate results for discrimination between patients with and without disease. Sometimes, it is more convenient to control for higher levels of sensitivity if the benefit of identifying true positives highly outweighs the cost of false positives. This is such the case in biomarker analysis for CAD. The cost of misdiagnosing patients with clinically relevant CAD is too great, while misdiagnosing a symptomatic patient without CAD will only expose them to a cardiac catheterization procedure. Valid diagnostic tests should maintain very high levels of sensitivity. In order to characterize measures of sensitivity and specificity, receiver operating characteristic curves are usually generated.

2.2.3 Receiver Operating Characteristic Curves

The Receiver Operating Characteristic (ROC) curve is a way to visualize and gauge the performance of a set of classifiers [31]. ROC analysis is the principal method for evaluating sensitivity and specificity of a classifier and proves to be a useful tool in the evaluation of biomarkers [26]. In general, a measurement of the area under the ROC curve (AUC) is reported to compare the intrinsic accuracy of different tests [32]. The ROC curve is generated by plotting a set of thresholds according to their corresponding true-positive and false positive rates, or sensitivity and 1-specificity.

When generating ROC curves for logistic models with several predictors, retrospective calculations of the ROC curve tend to give inflated assessments of score performance [33]. A predictive model will almost always fit better to the data it was constructed around than when applied to future data. This is because the ROC curve is generated by a process of resubstitution, where the model is constructed using the available data, and then validated on the same data [34]. Some more advanced techniques for handling the upward bias of the ROC curve have been discussed [35], but among the simplest and most common methods are k-fold cross validation and external validation.

K-fold cross-validation is an internal validation method for estimating the prediction error. In this process, the data is split into *k* number of blocks. A predictive model is generated based on the K-1 partitions and used to score the Kth block (figure 5). This process is repeated until predicted probabilities have been generated for all the observations. 5-fold and 10-fold cross-validation are the most common forms of K-fold cross validation, using 80% of the data to score the other 20%, or 90% of the data to score the other 10%, respectively [36].

External validation is one of the best ways to determine how well a logistic model can perform in clinical applications [37]. The data made available to the researchers or the statistician is used as a training set, where the predictive model is generated. The test set comes from additional experiments where the data has not been seen prior to developing the model. This can sometimes be emulated in an internal validation process where the data set is split into a training set and validation set prior to analysis. The validation data is then scored using the predictive model from the training set. In order to use this technique to estimate the prediction error of a model, the data set needs to be sizable enough to split into two separate data sets (ex. 2/3 data for the training set, 1/3 data for the validation set) [36].

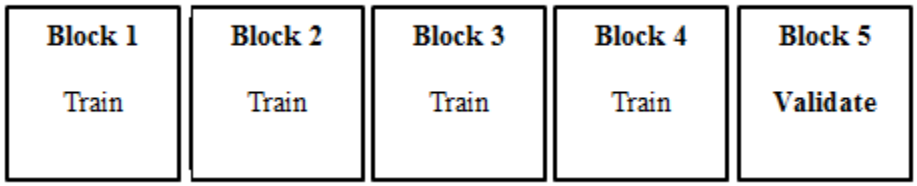

Figure 5. Example of 5-fold cross-validation
Four blocks are used as the training set and the model is validated on the fifth block of data. This process is repeated until cross-validated probabilities are calculated for all the blocks of data.

The ROC curve has been widely used to illustrate the sensitivity-specificity trade of in medical diagnostic testing, but newer methods are currently being applied to analyses to increase the clinical usefulness of statistics reporting on the added predictive ability of new biomarkers.

3.0 CLINICAL APPLICATION

The data set examined in this thesis originates from a study by LaFramboise et al., focusing on the identification of circulating proteins for the diagnosis of coronary artery disease [14]. This single-center study interrogated 359 serum samples for proteins from symptomatic patients referred for cardiac catheterizations from an emergency room or heart clinic. The proteins chosen for research have been previously identified as potential biomarkers for cardiac conditions or pathways involved with these conditions. The hypothesis of the study suggests that a combination of proteins known to be involved with multiple pathways of atherosclerosis can be used to develop a non-invasive alternative to coronary angiography for detection of CAD. This study is unique, in that it applies to the specific population of symptomatic patients that are referred for cardiac catheterization.

The proteomics analysis was conducted in two stages. In stage one, 239 samples (138 with CAD and 101 with normal coronary arteries) were assayed for 24 proteins. A scoring algorithm was generated off these 239 samples to measure the predictive ability of the proteins. This scoring algorithm was developed with a Monte Carlo optimization technique using a Metropolis algorithm [38] to derive the numerical coefficients, rather than the maximum likelihood approach used in logistic regression. 5-fold cross validation was used to estimate the bias in the ROC curve and to generate relevant sensitivity and specificity measurements. In the

following analysis, this process is duplicated using logistic regression rather than methods used to generate the previous scoring algorithm.

In stage 2, assays were run on 120 additional samples (71 with CAD and 49 with normal coronary arteries) for validation of the algorithm, but for economic reasons, the researchers excluded assaying for proteins the scoring algorithm found to be poor predictors for disease. Therefore, patient samples in this stage were only assayed for 11 of the proteins in the study. The addition of the 120 samples in stage 2 of the study was intended to externally validate the predictive ability of the scoring algorithm, comparable to the external validation process mentioned in section 2. Through statistical processes, it was determined the 11 proteins in the stage 2 data set were sufficient for this analysis.

Clinical characteristics for these subjects were obtained retrospectively, so there are some missing data encountered where clinical information could not be determined for the patient. No clinical characteristics were made available for the validation data set. Figure 6 shows a summary of the data that was available for this project [14].

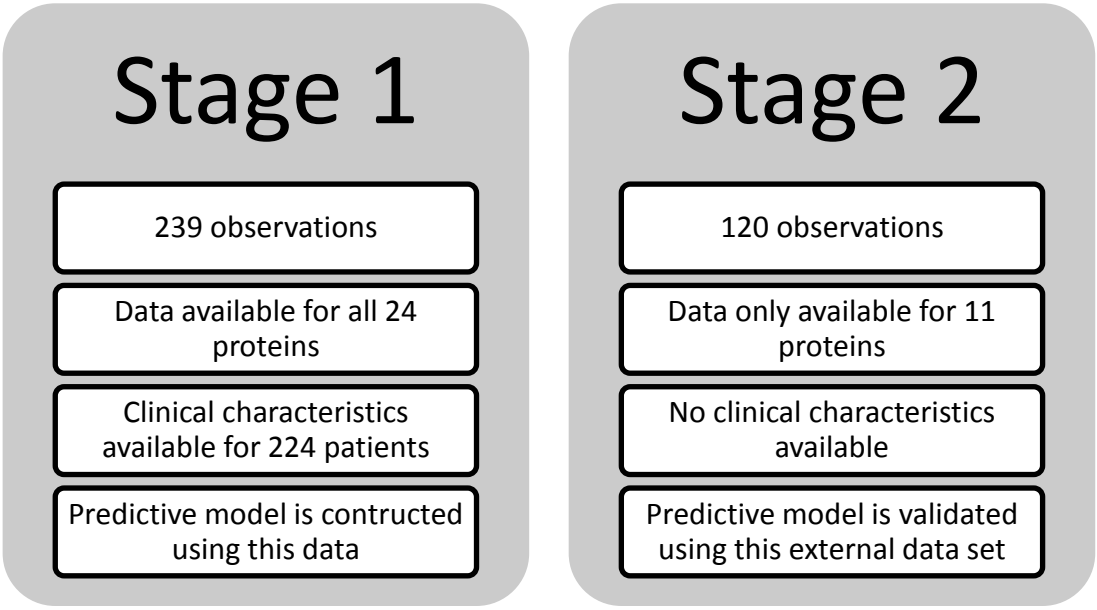

Figure 6. Data set components in stage 1 and stage 2 of the original study

The goal of the following analysis was to suggest a multi-marker panel, derived from the statistical methods covered in section 2, to be considered, in addition to previous results, for future studies. The analysis expands upon the original study by incorporating clinical characteristics and measuring the added effect of a multi-marker panel to these characteristics in the prediction of CAD.

A secondary analysis was included, following discovery of biomarkers with high predictive power, to measure whether or not a serum protein profile could be used on its own as a diagnostic tool, or if adding clinical characteristics to the algorithm could enhance predictive measures.

To sum up, the analysis will answer three questions:

1. Can a combination of circulating proteins in a symptomatic patient's serum predict coronary artery disease?
2. Can the predictive power of currently used clinical risk factors for heart disease be enhanced by protein biomarkers?
3. If a predictive model is generated based only on proteomic factors, can this model be enhanced by clinical risk factors for heart disease?

3.1 STATISTICAL METHODS

A preliminary analysis was conducted to emulate the objective of the original study by LaFramboise et al – to derive a statistically significant multi-marker panel for that can accurately discriminate between patients with and without coronary artery disease. The rest of the analysis

addresses the added predictive ability of multiple biomarkers to clinical characteristics that are commonly used to assess risk of coronary disease. Logistic regression and ROC curves are the primary statistical methods used for the analysis.

3.1.1 Preliminary analysis: evaluation of proteomic biomarkers

Before analysis was conducted, one-third of the data (120 samples in stage 2) were set aside to use as a validation set. The remaining 239 samples were used as a training data set.

Descriptive analysis of proteins

Descriptive analysis was performed on the proteins of the 239 samples (stage 1) in the training data set, and values of the proteins were compared across the disease groups to identify which proteins projected higher concentrations in patients with CAD. Imputation of 3 missing values was performed for MPO, Fibrinogen, and Leptin by replacing the missing value with the average level of the protein, conditional on the disease group. The Wilcoxon rank sum test was used under the null hypothesis that there were no statistically significant differences in protein concentrations among the groups. P-values were adjusted by the Benjamini-Hochberg (B-H) method [28] to control for multiple comparisons. The false discovery rate (FDR) was set at .05 and resulting Q-values less than this value were determined to be statistically significant. Results from the statistical tests identified proteins that would serve as useful biomarkers due to more or less of a presence in patients with disease.

Histograms of proteomic factors were generated to identify the distributions of the protein concentrations among the patients in the study. Heavily skewed concentrations were normalized by a natural logarithmic transformation to the data.

Descriptive statistics were also generated for the external data set of 120 patients (stage 2) for comparison to the cohort tested in the training data. Differences between the validation set and the training set may be a cause of variability in the technical procedures during the assay process.

Multivariate logistic regression of protein biomarkers

Protein concentrations, recorded on the original-scale or log-scale, were first factored by quartiles to categorize variables into low, medium, high, and very high intensities. The categorization of the variables was used to identify the pattern of association during univariate analyses. Univariate logistic regressions performed on the quartiled predictors produced odds ratios for quartiles 2-4, using the first quartile as the reference category, and a trend test was used to formally test whether or not there is an uniform increase (or decrease) in odds ratios across the quartiles. A multivariate logistic regression model was constructed based on a stepwise variable selection process with these factored covariates to assess the probability that patients with higher levels of a given set of proteins have CAD. This was followed up by analysis on the continuous form of the variables to produce more clinically relevant results.

For variables that appeared to follow a trend with the log-odds of disease, the functional form of the continuous covariate was investigated using first and second degree fractional polynomials. Fractional polynomial transformations that produced statistically significant improvements to the fit of the model were used for each covariate. Once the functional form of the variable was determined, a subsequent univariate analysis was carried out on the continuous form of the covariates. OPN unquestionably held the most statistically significant relationship with CAD from the univariate tests. Since the purpose of the analysis was to identify an effective multi-marker panel for discrimination among patients with and without CAD, logistic

regressions were run on the main effects once again, adjusting for OPN. Associations were tested at the .05 significance level.

After appropriate functional forms were found, a forward stepwise variable selection process was conducted on remaining variables using a probability of entry into the model of 0.05 and probability of removal from the model to be 0.1. All variables were entered into the model building process using their functional form as determined by the univariate analyses.

ROC analysis of the protein biomarker model

A naïve estimator of the area under the ROC curve (AUC) using the resubstitution method was recorded for comparison to the cross-validation methods. *K*-fold cross-validation was performed on the data set to estimate the error from the overly optimistic curve generated by resubstitution. 5-fold cross-validation was performed first to mimic the methods of the original study and compare the predictive model presented in this thesis with panels of markers identified by the original scoring algorithm. This was followed up by 10-fold cross-validation to detect any bias-variance tradeoff in the selection of *k* for this sample.

A validation procedure, using an additional 120 samples from a similar cohort, was conducted to measure the predictive model's ability to discriminate among "unseen" data. These samples were scored according to the final model derived from the first 239 samples, and statistics from an ROC analysis were compared to the cross-validated estimates.

3.1.2 Addition of proteomic biomarkers to clinical characteristics

The main focus of this work was to reveal how the proteomic biomarkers discovered in the preliminary analysis contributed to the predictive power of common clinical risk factors for heart

disease. This extended the hypothesis of the original study to include clinical variables in the discriminatory process. Age, smoking status (never, former, current), diabetes status (yes/no), and BMI calculations were studied to identify whether or not the factors had associations with coronary artery disease. BMI calculations were split into categories according to the Center for Disease Control (CDC), listed in table 1. This analysis was restricted to 224 observations, omitting patients whose clinical characteristics were unavailable.

Table 1. BMI Categories outlined by the Center for Disease Control

BMI (kg/m^2)	Weight Status
Below 18.5	Underweight
18.5 – 24.9	Normal
25.0 – 29.9	Overweight
30.0 and above	Obese

Descriptive analysis and logistic regression for protein biomarkers and clinical characteristics

Descriptive analyses were carried out on each of the clinical variables to determine any differences between the CAD and Normal groups. A Student's t-test was used to test the continuous variable age, and Chi-Square tests were used to test the other categorical variables in order to determine statistically significant associations with the disease group.

Univariate logistic regressions were carried out for each clinical variable to identify statistically significant linear relationships with CAD. Factors showing strong associations with disease were entered into a multivariate predictive model for CAD. The predictive model was then enhanced by adding proteomic biomarkers to the modeling process, both as quartiles and in

the continuous form in separate analyses. A final model that represented both a multi-marker panel from the preliminary proteomics analysis and important clinical risk factors was evaluated for predictive accuracy.

ROC analysis of the multivariate model

An ROC curve was generated for the predictive model consisting of clinical characteristics to use as a baseline measurement for comparison when new biomarkers are added to the model. For this part of the analysis, an external validation set was not available; the data set only provided clinical characteristics for 224 samples. Therefore, evaluation methods for the predictive models were restricted to k-fold cross validation – specifically, 10-fold cross validation, using 90% of the data as the training set and 10% as the validation set.

Biomarkers from the multi-marker panel identified in the proteomics analysis were added sequentially to the model consisting of clinical characteristics to measure the added predictive ability of the biomarkers. Improvement was determined by differences in the specificity at high sensitivities.

Unless otherwise stated, all statistical tests were conducted at a 0.05 significance level. The B-H method for multiple comparisons was restricted to comparative tests across the disease groups, as the conservative nature of the method did not affect the variable selection process in logistic regression.

Software

The analysis was performed mainly using SAS 9.3 statistical software. Fractional polynomials were analyzed using the fracpoly command in Stata 12. Graphics were generated by Stata 12.

3.2 RESULTS

3.2.1 Preliminary Results: Evaluation of proteomic biomarkers

3.2.1.1 Descriptive Analysis of Proteins

Tables 2-4 include descriptive statistics – the 24 proteins across 239 patient samples. Proteins were grouped by the scale they were measured on (μg/mL, ng/mL, pg/mL). The "CAD" column represents patients with clinically relevant coronary artery disease, and the "Normal" column represents patients with normal coronary arteries, or clinically irrelevant coronary disease. The Q-values are the Benjamini-Hochberg equivalent of the p-value. Fourteen proteins were found to have statistically significant differences in patients with significant CAD versus those with normal coronary arteries, controlling for a false discovery rate (FDR) of 0.05. All significant findings were found to be elevated in the CAD group, except for Apo-A1, which appeared to be of lesser quantities in the CAD group. These 14 proteins were carried over for further analysis.

Distributions of the 14 proteins up-regulated/down-regulated during times of coronary artery disease were examined. Apo-B100 and MPO appeared to follow a normal distribution in the CAD and Normal groups. All other proteins had heavily skewed-right distributions in both disease groups. A natural logarithmic transformation was applied to this data to fix the skew. Descriptive statistics for the stage 2 data can be found in Appendix A.

Table 2. Stage 1 Protein Descriptive Statistics (nanograms/mL)

	Combined	CAD						Normal							
Proteins	N	N	AVE	STD	25 PCT	50 PCT	75 PCT	N	AVE	STD	25 PCT	50 PCT	75 PCT	P-Value	Q-Value
OPN (ng)	239	138	49.6	41.6	13.8	28.6	55.7	101	15.8	15.9	6.7	10.9	21.6	<.0001	<.0001
VCAM (ng)	239	138	1175.1	500.9	940.0	1152.1	1538.8	101	856.7	365.2	683.8	903.6	1193.5	<.0001	<.0001
IL6 (ng)	239	138	0.9	1.2	0.04	0.3	1.3	101	0.6	1.1	0.02	0.04	0.6	<.0001	.0001
MPO (ng)	239	138	619.8	370.1	380.7	586.4	915.2	101	451.2	267.3	297.2	441.0	700.3	.0003	.0006
MMP7 (ng)	239	138	5.4	2.4	3.7	4.8	6.1	101	4.8	3	3.0	4.2	5.3	.0006	.0013
Resistin (ng)	239	138	104.6	70.6	52.8	90.1	139.7	101	81.9	62.8	47.0	71.8	110.8	.0020	.0038
MMP1 (ng)	239	138	5.3	2.4	3.7	5.0	6.6	101	4.8	2.2	3.1	4.4	6.0	.0452	.0723
Leptin (ng)	239	138	10.6	15.5	3.1	5.0	13.8	101	13.9	17.6	3.1	5.9	16.8	.2625	.3210
TIMP1 (ng)	239	138	329.3	87.9	269.3	323.4	370.9	101	320.1	112.2	238.9	310.5	372.1	.2675	.3210
TM (ng)	239	138	1.4	0.4	1.1	1.3	1.5	101	1.4	0.9	1.0	1.3	1.6	.3341	.3739
Pecam-1 (ng)	239	138	35	25.3	26.6	26.6	41.8	101	32.1	29	18.1	26.6	40.4	.3437	.3739
MCP1 (ng)	239	138	3.3	3.4	1.2	2.4	4.5	101	3.1	3.5	1.3	2.1	3.6	.3583	.3739
E-Selectin (ng)	239	138	34.3	16.5	21.2	34.2	44.3	101	33.9	19.1	22.5	31.7	44.7	.5686	.5686

Number of observations, average, standard deviation and interquartile range for CAD and Normal groups. The Q-value is the B-H analog of the P-value.

Table 3. Stage 1 Protein Descriptive Statistics (micrograms/mL)

	Combined	CAD						Normal							
Proteins	N	N	AVE	STD	25 PCT	50 PCT	75 PCT	N	AVE	STD	25 PCT	50 PCT	75 PCT	P-Value	Q-Value
Fibrinogen (µg)	239	138	19	59.2	3.4	6.0	12.6	101	4.1	6.3	1.8	3.2	5.6	<.0001	<.0001
Apo-A1 (µg)	239	138	154.2	134.8	57.8	117.0	195.6	101	300.6	258.8	114.0	223.2	385.7	<.0001	<.0001
Apo-B100 (µg)	239	138	339.1	80.6	233.5	299.9	370.9	101	295.6	80.6	206.8	265.5	328.9	.0001	.0003
CRP (µg)	239	138	0.8	1.5	0.1	0.5	1.4	101	0.3	0.6	0.1	0.3	0.8	.0038	.0065
L-Selectin (µg)	239	138	1.1	0.3	0.9	1.1	1.3	101	1.1	0.3	1.0	1.1	1.3	.1482	.2760
Acrp30 (µg)	239	138	4.8	3.5	2.7	4.1	6.2	101	5.3	3.8	3.0	4.8	7.7	.2650	.3210

Number of observations, average, standard deviation and interquartile range for CAD and Normal groups. The Q-value is the B-H analog of the P-value.

Table 4. Stage 1 Protein Descriptive Statistics (picograms/mL)

	Combined	CAD						Normal							
Proteins	N	N	AVE	STD	25 PCT	50 PCT	75 PCT	N	AVE	STD	25 PCT	50 PCT	75 PCT	P-Value	Q-Value
IL10 (pg)	239	138	7.5	18.2	2.6	3.65	5.9	101	3.2	3.6	1.2	2.1	3.5	<.0001	<.0001
IL1β (pg)	239	138	113.7	168.2	4	23.2	152.3	101	48.9	120.1	1.9	7.0	38.0	<.0001	<.0001
NT-pBNP (pg)	239	138	101.7	202.5	12.3	31.8	93.9	101	41.1	111.6	7.8	15.8	32.6	<.0001	.0001
IFNγ (pg)	239	138	4.2	7.7	0.5	2.1	4.6	101	4	12.9	0.2	1.4	2.9	.0010	.0020
TNFα (pg)	239	138	15	19.4	3.1	8.7	17.8	101	22.4	73.7	0.0	5.4	16.9	.0998	.1497

Number of observations, average, standard deviation and interquartile range for CAD and Normal groups. The Q-value is the B-H analog of the P-value.

3.2.1.2 Logistic Regression Modeling of Proteins

Univariate regressions for the 14 up-regulated/down-regulated proteins (identified by the previous comparative tests) uncovered statistically significant relationships with disease status. For proteins entered as categorical variables into a univariate logistic regression model, higher quartiles produced higher odds ratios with disease compared to concentrations in the first quartile. Trend tests on the quartiles of the continuous covariates identified increasing odds ratios among all variables (table 5). Graphs of the odds ratios for the quartiles of variables can be found in Appendix A.

Univariate regressions on the continuous form of the 14 proteins also produced statistically significant relationships with CAD. While categorization had resulted in increasing odds ratios for most all of the proteins, nonlinearity in the proteins' continuous form was still assessed. Fitting each univariate regression (using the continuous form of the variables) with fractional polynomials identified a nonlinear function for IL1β that produced a better fit for the model. A cubic transformation was applied to capture the functional form of this variable. Analysis of second degree fractional polynomial transformations produced statistically non-significant results for all main effects.

Table 6 shows the results of both the unadjusted logistic regressions and regressions adjusted for OPN. Apo-A1 and NT-pBNP produced statistically non-significant relationships with disease when adjusted for OPN. Odds ratios in this table are calculated according to unit increases in the standard deviation of each protein.

Table 5. Odds ratios and trend tests for univariate regressions of factored protein concentrations

Variable (quartiles)	Odds-ratio	P-Value	Trend Test	Variable (quartiles)	Odds-ratio	P-Value	Trend Test
OPN				NT-pBNP			
2	2.32	.0322	<.0001	2	1.16	.6863	.0009
3	5.72	<.0001		3	2.07	.0538	
4	55.73	<.0001		4	5.28	<.0001	
Fibrinogen				IL-6			
2	2.250	.0075	<.0001	2	0.96	.9185	.0006
3	4.211	.0002		3	2.23	.0012	
4	8.421	<.0001		4	2.53	.0256	
VCAM				Apo-B100			
2	2.21	.0355	<.0001	2	1.347	.5466	.0016
3	4.00	.0003		3	3.214	.0157	
4	7.23	<.0001		4	3.897	.0045	
IL-10				CRP			
2	1.45	.3350	<.0001	2	1.71	.1445	.0224
3	8.11	<.0001		3	1.22	.5840	
4	4.93	<.0001		4	3.07	.0041	
Apo-A1				MMP7			
2	.551	.1619	<.0001	2	2.042	.0554	.0151
3	0.273	.0017		3	2.585	.0115	
4	0.118	<.0001		4	3.070	.0035	
IL1β				Resistin			
2	2.259	.0292	.0004	2	1.31	.4636	.0116
3	2.519	.0142		3	1.66	.1711	
4	6.245	<.0001		4	3.48	.0017	
MPO				IFNγ			
2	2.66	.0061	.0004	2	1.52	.2634	.0275
3	1.97	.0574		3	1.73	.1217	
4	5.74	.0001		4	2.96	.0031	

Proteins are factored by quartile, and the first quartile is used as the reference category.

Table 6. Univariate regressions for the continuous form of the protein concentrations

		Unadjusted		*Adjusted for OPN*	
Variable	STD	Odds-ratio	P-Value	Odds-ratio	P-Value
OPN*	1.05	4.06	<.0001	-	-
VCAM*	0.44	2.34	<.0001	1.39	.0835
Apo-A1*	1.00	0.43	<.0001	0.77	.1530
IL-10*	3.16	3.24	<.0001	2.18	.0104
Fibrinogen*	1.12	2.53	<.0001	1.96	.0012
IL1β*	2.92	1.92	<.0001	2.73	<.0001
NT-pBNP*	1.83	1.92	.0002	1.11	.5979
Apo-B100	83.26	2.29	.0003	1.79	.0014
MPO	340.32	1.97	.0009	2.77	<.0001
IL-6*	2.17	1.76	.0011	2.04	<.0001
CRP*	1.51	1.57	.0019	1.43	.0315
MMP7*	0.46	1.54	.0034	1.34	.0825
Resistin*	0.74	1.57	.0034	2.93	<.0001
IFNγ*	5.54	1.70	.0054	1.78	.0004

*log-transformed variable

A forward stepwise selection process of the predictors to determine a final model resulted in a panel of 4 distinct markers – OPN, IL1β, Apo_B100, and Fibrinogen. The resulting model was fit using both categorical (quartiles) and continuous forms of the variables.

The multivariate model containing the quartiles of the covariates (table 7) indicate that the odds of disease are elevated for higher quartiles compared to the lowest quartile. The increasing trend of the factored covariates provided justification for using proteins in their continuous form.

Table 7. Multivariate logistic regression model for proteins factored by quartile

Variable	Coefficient (β)	Standard Error	P-Value	Odds Ratio	95% CI
Intercept	-4.76	0.92	<.0001	--	--
OPN					
2	1.15	0.49	0.0187	3.15	1.21 to 8.18
3	2.19	0.51	<.0001	8.95	3.32 to 24.09
4	4.35	0.76	<.0001	77.54	17.66 to 340.60
IL1β					
2	0.95	0.51	0.0634	2.58	0.95 to 7.01
3	1.75	0.53	0.001	5.75	2.03 to 16.27
4	2.71	0.57	<.0001	15.01	4.94 to 45.61
Fibrinogen					
2	0.84	0.49	0.0868	2.31	0.89 to 6.04
3	1.26	0.52	0.0152	3.54	1.28 to 9.82
4	1.45	0.55	0.0081	4.26	1.46 to 12.46
Apo-B100					
2	0.41	0.76	0.5922	1.51	0.34 to 6.73
3	1.74	0.77	0.0239	5.68	1.26 to 25.66
4	1.85	0.75	0.0135	6.34	1.46 to 27.44

For the model featuring continuous covariates (table 8), Apo_B100 was recorded on the original scale while the other 3 markers were measured on the natural logarithmic scale. IL1β entered the model in its functional form as determined by the fractional polynomial transformation. All first-order interactions among the covariates were determined to be not statistically significant during the modeling process.

Table 8. Multivariate logistic regression model for continuous covariates

Variable	Coefficient (β)	Standard Error	P-Value	Odds Ratio	95% CI	Min	Max
Intercept	-7.31	1.16	<.0001	--	--	--	--
OPN*	1.43	0.22	<.0001	4.18	2.70 to 6.47	0.13	5.32
IL1β**	0.52	0.11	<.0001	1.68	1.35 to 2.09	-4.38	4.07
Apo_B100	0.008	0.002	0.0015	1.01	1.00 to 1.01	0.36	529.21
Fibrinogen*	0.53	0.19	0.0065	1.70	1.16 to 2.48	-0.63	6.10

*natural logarithmic scale **functional form: $\ln(IL1\beta)^3$

All coefficients for main effects in this final model were statistically significant (all $p \leq 0.006$), suggesting that removal of one of the variables would significantly reduce the fit of the model. Interpretations of coefficients can be made in respect to the log odds of CAD, taking into account the functional form of the main effects within the model, while adjusting for other covariates. For example, the log odds of CAD would increase at a rate of 1.43 for every unit increase in the natural log of OPN. It may make more sense to report these numbers as odds ratios though. The odds of a patient having CAD are approximately 4 times greater for every unit increase in the natural log of OPN. Odds of disease increase at a rate of 68% for every unit increase in the cubic function of the natural log of IL1β, 0.8% for every unit increase in Apo-B100, and 70% for every unit increase in the natural log of Fibrinogen.

An ROC analysis was conducted for this final model including continuous covariates in the next steps.

3.2.1.3 ROC Analysis of Protein Model

The optimistic estimator for the AUC was .8816. 5-fold and 10-fold cross-validation was performed 10 times each, and results were averaged to obtain a correction for the predicted probabilities and the ROC curve. The AUC estimates from 5-fold and 10-fold cross-validation were 0.8611 and 0.8632, respectively, which are lower than the original estimate, but only by a small amount. 5-fold cross-validation did not appear to introduce much bias to the prediction error, as the results obtained by both 5-fold and 10-fold cross-validation were similar. Table 9 shows the AUC estimate for each method, the exact sensitivity (SN) of ~95%, and specificities, the ability to identify true negatives, at 90%, 95%, and 98% sensitivities (SP90, SP95, and S98). Where an exact sensitivity of 90, 95, or 98 was attainable, the closest level of sensitivity to each

value was used. Controlling for 95% sensitivity, 38% specificity was attained using the predictive model to classify patients within the sample of 239.

Validation on the external data set, the 120 patients of "stage 2," saw lower discrimination among the patients, but with an AUC statistic at .7077, an impressive level for a predictive model. Performance for validation usually declines because the predictive model was built without "seeing" this data. Maintaining 95% sensitivity, the predictive model generated on the 239 patients had a specificity of 10% when applied to external data. This may have implications for instability in the predictive model in clinical practice, or that patients in the additional cohort displayed significantly different characteristics than the first cohort. Further mention of this variability can be found in the discussion.

Table 9. AUC, Sensitivity, and Specificity from Cross-Validated Results for the protein-only model

Method	AUC	SN	SP90 [95% CI]	SP95 [95% CI]	SP98 [95% CI]
Resubstitution	.8816	94.9	.505 [.404, .605]	.436 [.338, .538]	.307 [.221, .408]
5-Fold Cross Validation	.8611	94.9	.515 [.414, .615]	.376 [.283, .479]	.356 [.265, .459]
10-Fold Cross Validation	.8632	94.9	.515 [.414, .615]	.386 [.292, .489]	.347 [.256, .448]
External Validation	.7077	95.7	.225 [.122, .370]	.102 [.038, .230]	.061 [.016, .179]

Column SN is the closest rounded estimate for 95% sensitivity. The last 3 columns represent corresponding specificities at 90, 95, and 98 percent fixed sensitivities with 95 percent confidence intervals.

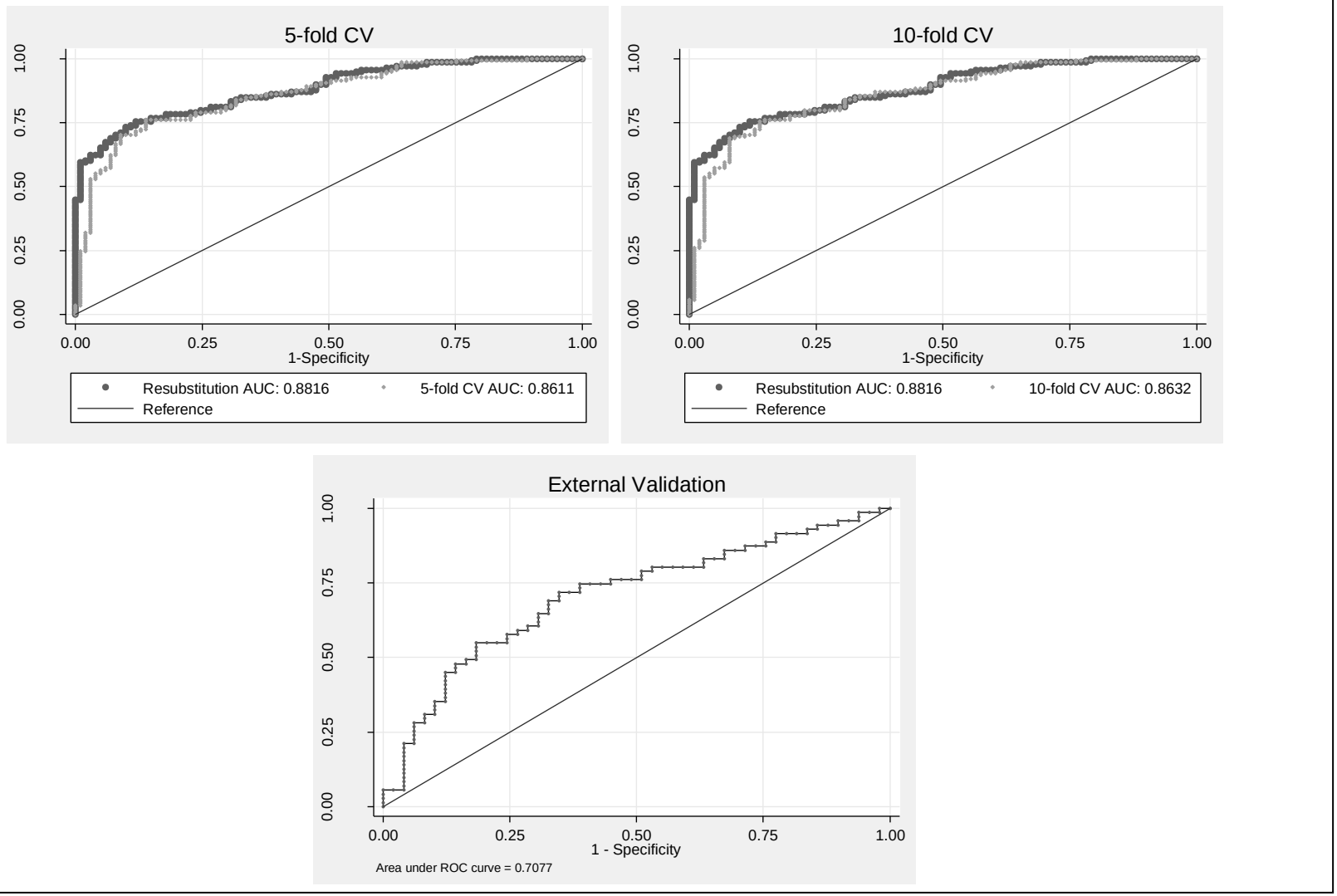

Figure 7. Validation methods for the ROC curve
In 5-fold and 10-fold cross-validation (top) the lighter line represents the cross-validated probabilities and the darker line is the curve generated by the resubstitution method.

3.2.2 Added effect of biomarkers to clinical characteristics

The results in this section demonstrate the added predictive ability of a multi-marker panel for CAD among symptomatic patients. As clinical characteristics were scarce, only gender, diabetes, BMI, hypertension, smoking, and age factors were analyzed.

3.2.2.1 Descriptive Analysis of Clinical Characteristics

For the full data set, 138 symptomatic patients were diagnosed with CAD and 101 patients were considered to have insignificant disease according to catheterization results. Fifteen of these patients were removed from the analysis due to the absence of clinical information for those subjects, leaving 123 samples in the CAD group and 101 samples in the Normal group. Of the clinical variables examined, only age, diabetes status, and smoking status were significantly different among the CAD and Normal groups. Patients in the CAD group were, on average, older than the Normal group ($p = .0096$), and positive associations with CAD were confirmed for patients with diabetes ($p = .0087$) and smokers ($p = .0071$), according to the Student's t-test for continuous variables and the chi-square test for categorical variables. BMI categories, hypertension status, and gender were determined to have statistically non-significant associations with CAD (all $p \geq .40$).

Univariate regressions on the clinical variables produced significant linear relationships with the log-odds of significant CAD for age, diabetes status, and a current smoking status. All other clinical variables that showed insignificant associations with disease status from the descriptive analysis did not produce significant results from the univariate analyses (all $p \geq .40$).

Table 10. Descriptive statistics of clinical characteristics

Clinical Characteristics	Combined N	CAD N	CAD AVE	CAD STD	NORMAL N	NORMAL AVE	NORMAL STD	P-Value
AGE (years)	224	123	62.3	12.1	101	58.1	11.6	.0096

Categorical Variables	Combined N	CAD Frequency	NORMAL Frequency	P-Value
Gender				
-Male	114	62	52	.8723
-Female	110	61	49	
BMI Category				
-Normal	34	19	15	.9881
-Overweight	88	48	40	
-Obese	101	56	45	
Diabetes Status				
-Yes	59	41	18	.0087
-No	165	82	83	
Hypertension Status (HTN)				
-Yes	135	77	58	.4309
-No	89	46	43	
Smoking Status (SMO)				
-Never	75	35	40	.0071
-Former	110	58	52	
-Current	39	30	9	

Table 11. Univariate regressions on clinical characteristics

Variable	Odds-Ratio	P-Value
SMO		
-Former	1.274	.5190
-Current	3.810	.0034
Diabetes	2.306	.0097
Age	1.030	.0107
HTN	1.241	.4311
Gender	.9578	.8723
BMI		
-overweight	.9474	.8821
-obese	.9825	.9372

3.2.2.2 Multivariate model building involving proteins and clinical characteristics

A full multivariable regression process was carried out on proteins and clinical characteristics in the data set (categorical form of covariates can be found in Appendix A). The final multivariate model (table 12) produced the same 4-marker panel derived from the preliminary analysis, along with indicator variables for diabetes and current smokers. According to the model, these clinical factors accounted for a 3-fold increase in odds of disease for smoking patients or diabetics. Age dropped out of the final model due to lack of a statistically significant relationship with disease status once adjusted for additional predictors. The estimates for the four protein biomarkers in the final model remained statistically significant even with the addition of further clinical risk factors (age, BMI, and HTN), proving the validity of these predictors.

The final model including both clinical characteristics and proteomic biomarkers had a high level of discrimination among patients (AUC = .8805), and there was improved ability to detect true negative results while controlling for high levels of sensitivity (see table 13).

Table 12. Multivariate model containing proteins and clinical characteristics

Variable	Coefficient (β)	Standard Error	P Value	Odds Ratio	95% CI	Min	Max
Intercept	-8.412	1.339	<.0001	--	--	--	--
Smoking Status	1.210	.5270	.0217	3.353	1.194 to 9.420	--	--
Diabetes	1.083	.4373	.0133	2.952	1.253 to 6.569	--	--
OPN*	1.462	.2367	<.0001	4.313	2.712 to 6.860	.130	5.32
IL1β**	.5015	.1229	<.0001	1.651	1.298 to 2.101	-4.38	4.07
Apo_B100	.0091	.0028	.0010	1.009	1.004 to 1.015	.363	529.21
Fibrinogen*	.5942	.2168	.0061	1.812	1.812 to 2.771	-.628	6.10

Method	AUC	SP90 [95% CI]	SP95 [95% CI]	SP98 [95% CI]
Resubstitution	.9006	.6536 [.552, .744]	.4753 [.376, .577]	.3861 [.292, .489]
10-Fold Cross Validation	.8805	.5941 [.482, .680]	.3861 [.292, .489]	.3366 [.248, .438]

*natural logarithmic scale **functional form: $\ln(IL1\beta)^3$

Table 13. AUC, Sensitivity, and Specificity from Cross-Validated Results for the protein-only model

Model	AUC	SP90	SP95	SP98
Current Smoker, Diabetes	.5787	.0396	0	0
Additional marker (Adjusted for Current Smoker and Diabetes)	**AUC (difference)**	**SP90 [95% CI] (difference)**	**SP95 [95% CI]**	**SP98 [95% CI]**
OPN	.8340 (+.2553)	.4951 [.395, .596] (+.4555)	.3564 [.265, .459]	.1683 [.104, .259]
IL1b	.6883 (+.1096)	.2574 [.169, .345] (+.2178)	.0693 [.031, .142]	.0198 [.003, .077]
Fibrinogen	.7315 (+.1528)	.3861 [.292, .489] (+.3465)	.2574 [.178, .356]	.0891 [.044, .167]
Apo-B100	.6725 (+.0938)	.2970 [.212, .397] (+.2574)	.2277 [.152, .324]	.0396 [.012, .104]
OPN, IL1b, Fibrinogen, Apo-B100	.8805 (+.3018)	.5941 [.482, .680] (+.2574)	.3861 [.292, .489]	.3366 [.248, .438]

4.0 DISCUSSION

Preliminary analysis: proteomic biomarkers

The final multi-marker panel, consisting of OPN, IL1β, Apo-B100, and Fibrinogen, differed from any of the multi-marker panels in the original study. Each individual marker appeared in at least one of the proposed panels in the original study, however there were no 4-marker panels identical to the one derived from the logistic regression process. Differences in results between this analysis and the original study may be attributed to the methods used in generating a predictive model or scoring function. The maximum likelihood approach used in this analysis for estimating the model's coefficients provided a statistical advantage to developing a multi-marker panel with the best diagnostic ability. The 4-marker panel performed less well when applied to external data from an additional cohort, but this was done without first calibrating the model. Smoothing parameters were not used in this analysis, but may be used in a future analysis or study to improve prediction on external data. There may have also some variability in the protein measurements during the proteomics phase of the experiment due to the validation samples being run on a separate lot of plates and reagents than the samples used for the training data.

Compared with the multi-marker panels suggested by LaFramboise et al, the 4-marker panel derived from the logistic regression process was similar. The best 4-marker panels derived by the study's scoring algorithm had AUC measurements ranging from .82-.84 and specificities

of 43% - 58% at 95% sensitivity, while the predictive model generated here produced an AUC of .86 and 38% specificity achieved at 95% sensitivity.

Concentrations of OPN produced the strongest relationship of any protein, which is a significant result of this analysis in itself, as OPN has been linked to heart disease through recent studies [39]. In regards to its relationship with coronary disease, OPN is a glycoprotein/cytokine of the extracellular matrix that has shown implications of roles in cardiac remodeling and fibrosis [40]. Other studies suggest that OPN is associated with calcification in coronary arteries [41, 42]. However, up-regulation of this protein has been linked to many other pathologies as well, including myeloma, multiple sclerosis, bone destruction, and cancer, preventing a direct association with heart disease to be made [43]. This makes it difficult to draw conclusions on the clinical usefulness of OPN as a biomarker for CAD without further understanding of the protein's precise function.

IL1β is a cytokine of the interleukin family known to be involved in inflammatory response. The inflammatory process has been discussed as a significant mediator in the development of atherosclerosis, and the gene encoding the IL1β protein has been linked to coronary artery disease in Brazilian populations [44]. Apo-B100 is a lipid binding protein that is responsible for carrying low density lipoproteins (LDL, aka "bad cholesterol") to tissues. Apo-B100 has been declared as a more reliable indicator of risk of heart disease than LDL and a standardized assay for the protein can be used clinically [45]. Fibrinogen is known to play a role in blood clot formation and concentration levels have been notably increased in patients with cardiovascular disease. Thrombosis, the formation of blood clots, has been recognized as the basis for many cardiac cases involving myocardial infarction, ischemic death, and unstable angina pectoris [46].

Proteomic biomarkers add predictive value to clinical characteristics

The addition of a proteomic multi-marker panel to common clinical risk factors of heart disease greatly improved discrimination among symptomatic patients. For the sample of symptomatic patients studied, the data resembled the current status quo, where the use of clinical characteristics provides zero ability to detect a patient who does not need to undergo catheterization. None of the clinical characteristics had adequate discriminatory power before the inclusion of proteins. The added discriminatory power of OPN, alone, accounted for most of the increase in predictive power, but the ability to classify patients with normal coronary arteries while maintaining a high sensitivity was the best when Fibrinogen, IL1β, and Apo-B100 were all present. The analysis demonstrated that the use of protein biomarkers for coronary artery disease can identify approximately thirty percent of patients for whom cardiac catheterization would be unnecessary.

Lastly, the analysis involving clinical characteristics was very limited. It would have been beneficial to incorporate patient cholesterol profiles and family history among other risk factors into the analysis. Ideally, it would be of interest to measure the impact biomarkers have on currently used scoring systems for pretest probability or likelihood of CAD. Additional studies should measure all risk factors of a patient that are currently used in practice to diagnose the risk of heart disease and the effect new biomarkers will have in determining a patient's pretest probability or likelihood of disease. This alludes to a chief argument against biomarker studies – that efforts should rather be focused on improving current diagnostic methods. While one obvious approach to this argument would be to improve medical imaging techniques for

coronary angiography, one could also argue to improve current diagnostic methods by adding new biomarkers to the current set of clinical risk factors used by professionals.

5.0 CONCLUSION

Throughout this thesis, evaluation of diagnostic biomarkers has been discussed in regards to coronary artery disease, where the current diagnostic method is starting to become costly. Common statistical methodologies regarding biomarker experiments, including logistic regression and ROC curves, were demonstrated on a clinical study intending to identify multiple proteins that can accurately diagnose cardiac patients that are normally referred for coronary angiography. The statistical analysis confirmed the predictive ability of protein biomarkers that was discovered in the previous study, but also added further insight into the public health matter by incorporating clinical characteristics. The primary focus of the analysis resulted in the following conclusion: A multi-marker panel featuring OPN, IL1β, Apo-B100, and Fibrinogen, based on the serum protein profiles of symptomatic patients, can be used in conjunction with clinical risk factors of heart disease (specifically diabetes and smoking status) to improve discrimination between those with clinically significant coronary artery disease and those with "normal" coronary arteries that do not require cardiac catheterization.

Screening symptomatic patients prior to cardiac catheterization with a blood assay can be advantageous if the goal is to reduce the number of patients exposed to unnecessary operations. In order to prevent a high number of patients that actually do have coronary artery disease from being misclassified, only a small number of patients that do not require angiographic testing could be identified. This may seem like a poor outcome, but the current status quo identifies zero

percent of these patients. Therefore, even identifying one of these symptomatic patients with clinically insignificant coronary artery disease can be labeled a success.

The main issue to consider before implementing a blood test as a new diagnostic procedure would be whether or not it is worth the risk of dismissing a single patient that has clinically significant coronary artery disease. While a multi-marker panel has proven ability to classify patients into disease and non-disease groups, there is still a chance of error with this type of test, and it is still uncertain whether or not the benefits outweigh the costs when used in clinical practice. However it is also undeniable that the future landscape of healthcare in the United States, particularly with the inclusion of the Affordable Care Act, will find these types of screening procedures more and more attractive, especially for procedures such as cardiac catheterization, where the costs are increasing into the tens of thousands of dollars.

APPENDIX

ADDITIONAL TABLES AND FIGURES

1. DESCRIPTIVE STATISTICS OF THE STAGE 2 (VALIDATION) DATA SET

Table 14. Stage 2 Protein Descriptive Statistics (nanograms/mL)

	Combined	CAD						Normal					
Proteins	N	N	AVE	STD	25 PCT	50 PCT	75 PCT	N	AVE	STD	25 PCT	50 PCT	75 PCT
VCAM (ng)	120	71	1425.6	435.1	1073.3	1415.7	1658.0	49	1250.6	602.9	906.7	1106.8	1388.0
MPO (ng)	120	71	950.2	710.7	397.4	816.8	1155.3	49	676.7	420.5	390.1	629.3	838.5
OPN (ng)	120	71	23.0	20.3	9.4	17.3	29.1	49	17.6	14.1	7.3	13.6	23.5
Resistin (ng)	120	71	123.3	87.7	55.3	97.7	167.7	49	102.8	62.6	64.3	82.3	127.8
MMP7 (ng)	120	71	4.9	2.3	3.1	4.6	6.0	49	4.8	3.7	3.3	4.3	4.8

Table 15. Stage 2 Protein Descriptive Statistics (micrograms/mL)

	Combined	CAD						Normal					
Proteins	N	N	AVE	STD	25 PCT	50 PCT	75 PCT	N	AVE	STD	25 PCT	50 PCT	75 PCT
Fibrinogen (µg)	120	71	37.7	111.6	6.3	9.4	20.0	101	4.1	6.3	1.8	3.2	5.6
Acrp30 (µg)	120	71	5.9	4.8	3.3	4.5	6.6	101	5.3	3.8	3.0	4.8	7.7
CRP (µg)	120	71	3.6	7.2	0.6	1.1	3.0	101	0.3	0.6	0.1	0.3	0.8
Apo-B100 (µg)	120	71	229.0	68.2	172.9	217.4	273.4	101	295.6	80.6	206.8	265.5	328.9

Table 16. Stage 1 Protein Descriptive Statistics (picograms/mL)

	Combined	CAD						Normal					
Proteins	N	N	AVE	STD	25 PCT	50 PCT	75 PCT	N	AVE	STD	25 PCT	50 PCT	75 PCT
IFNγ (pg)	120	71	5.6	13.2	0.95	3.0	4.7	49	2.8	3.2	0.8	2.0	3.9
IL10 (pg)	120	71	103.6	157.0	2.42	22.13	148.8	49	71.4	130.9	1.8	14.6	54.0

2. GRAPHS OF ODDS RATIOS BY QUARTILE FROM UNIVARIATE LOGISTIC REGRESSION

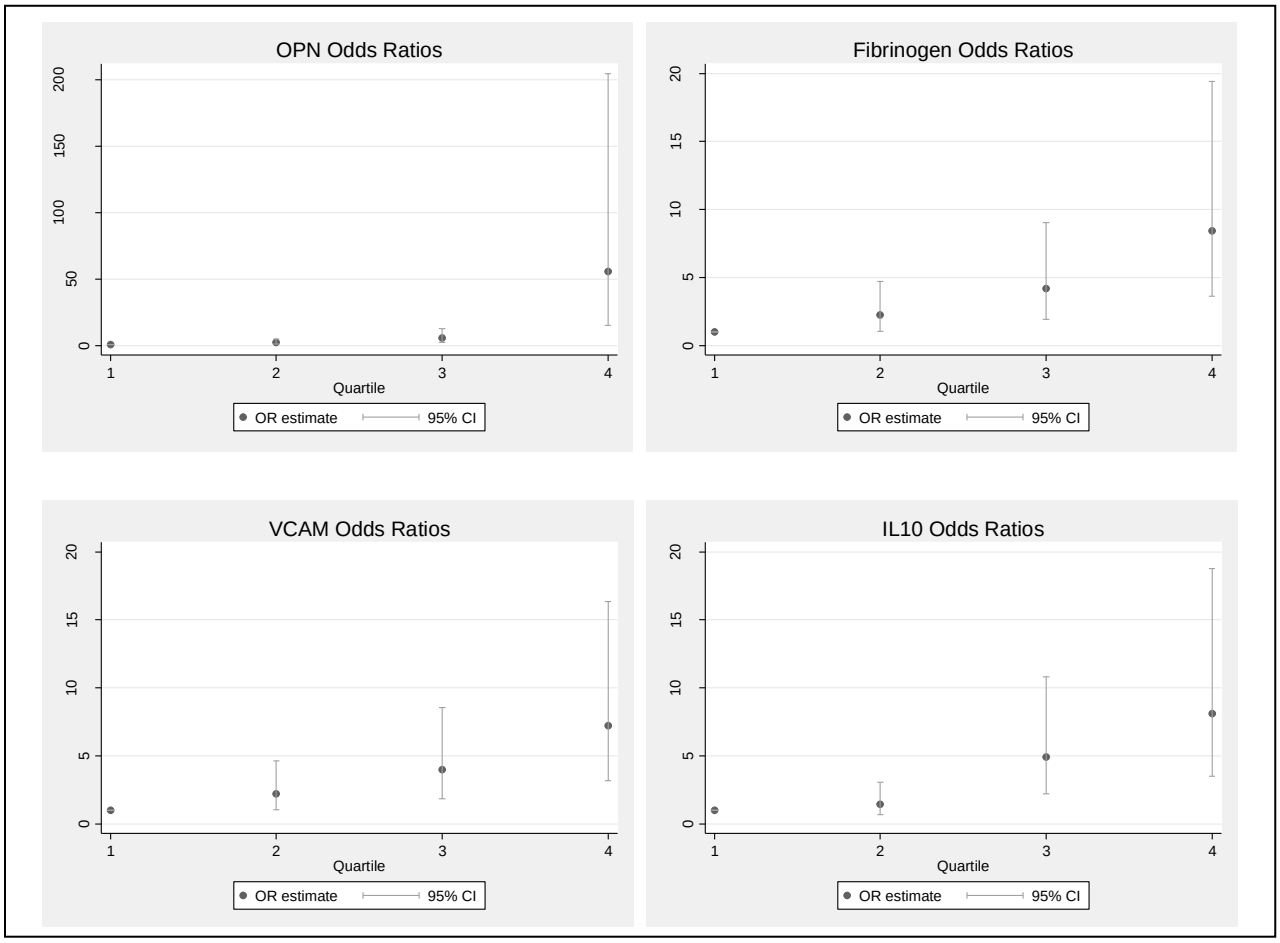

Figure 8. Graphs of odds ratios from univiariate regression for OPN, Fibrinogen, VCAM, and IL10

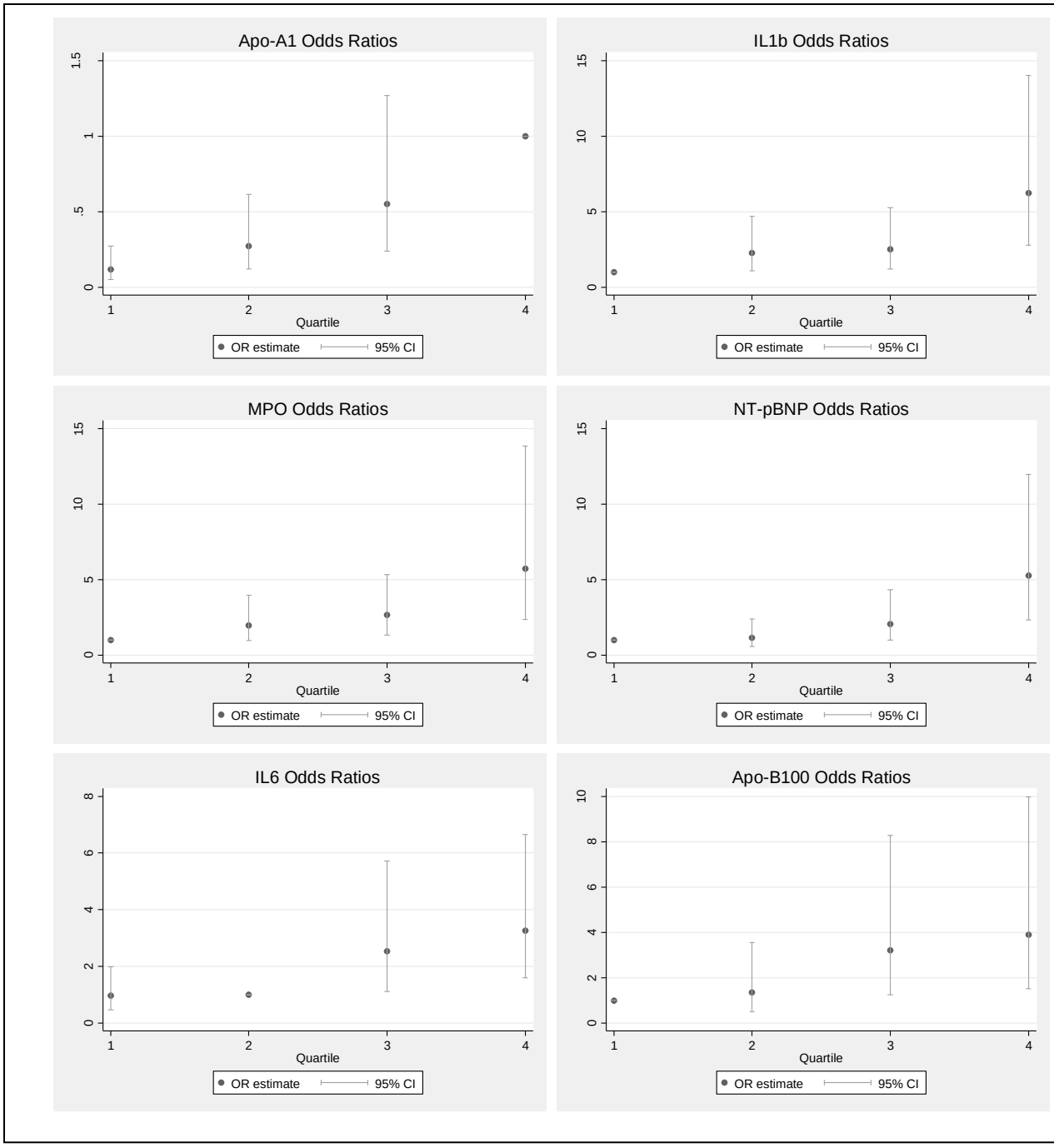

Figure 9. Graphs of odds ratios from univiariate regression for Apo-A1, IL1b, MPO, NT-pBNP, IL6, and Apo-B100

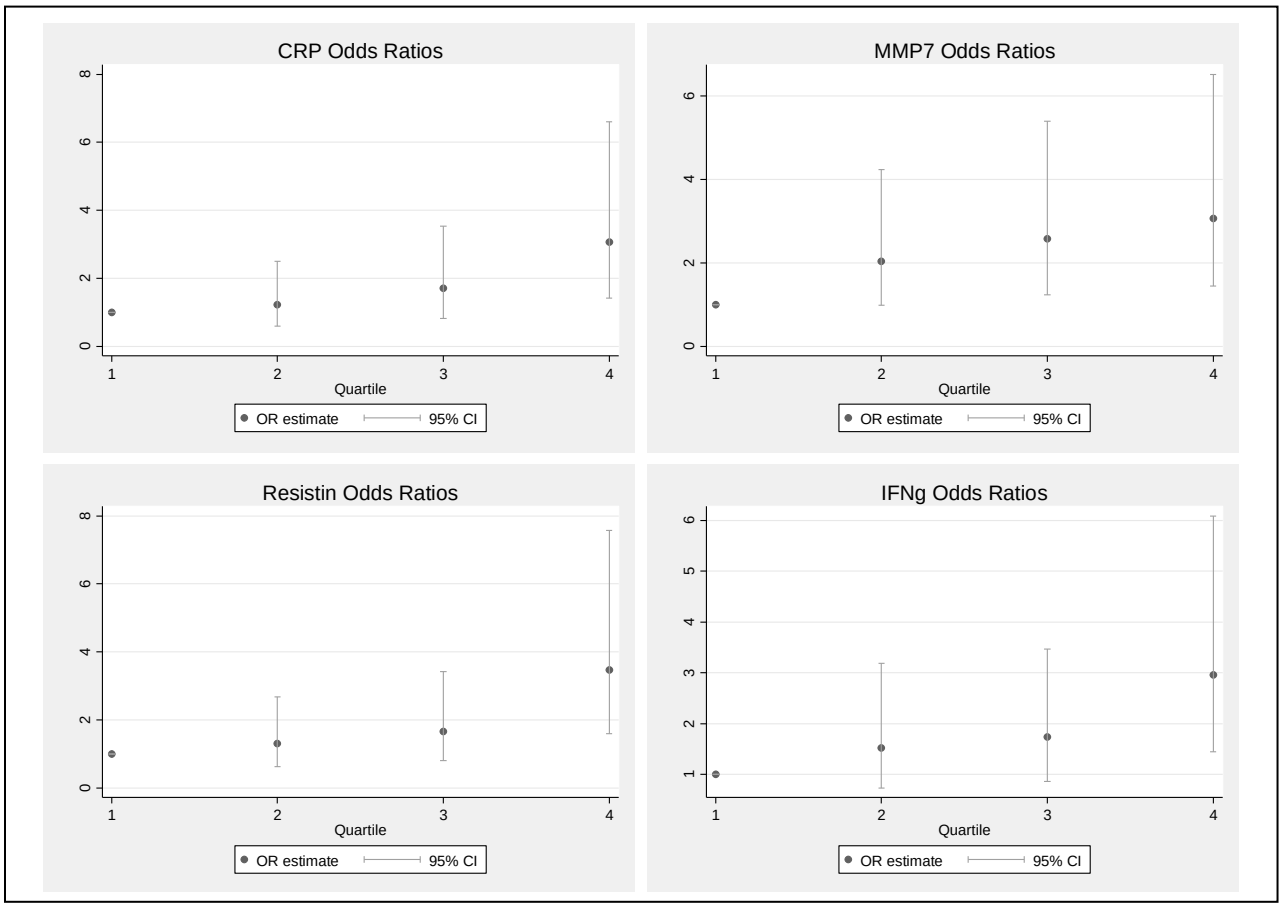

Figure 10. Graphs of odds ratios from univiariate regression for CRP, MMP7, Resistin, and IFNg

3. MULTIVARIATE LOGISTIC REGRESSION MODEL INVOLVING CLINICAL CHARACTERISTICS AND FACTORIED PROTEINS (BY QUARTILE)

Table 17: Multivariate model with proteins categorized by quartile

Variable	Coefficient (β)	Standard Error	P-Value	Odds Ratio	95% CI
Intercept	-5.39	1.01	<.0001	--	--
OPN					
2	1.29	0.54	0.0167	3.62	1.26 to 10.41
3	1.98	0.55	0.0003	7.24	2.45 to 21.41
4	4.83	0.83	<.0001	125.39	24.71 to 636.16
IL1β					
2	0.80	0.56	0.156	2.22	0.74 to 6.74
3	1.60	0.57	0.0053	4.95	1.61 to 15.21
4	2.58	0.63	<.0001	13.24	3.82 to 45.22
Fibrinogen					
2	0.86	0.53	0.1058	2.37	0.83 to 6.77
3	1.52	0.58	0.0082	4.59	1.48 to 14.23
4	1.40	0.60	0.0199	4.07	1.25 to 13.26
Apo-B100					
2	0.11	0.80	0.8931	1.11	0.23 to 5.37
3	1.66	0.82	0.0413	5.28	1.07 to 26.09
4	1.91	0.79	0.0151	6.75	1.45 to 31.46
Diabetes	1.24	0.46	0.0072	3.46	1.40 to 8.57
Smoking Status	1.49	0.54	0.0055	4.46	1.55 to 12.81

BIBLIOGRAPHY

1. Atkinson, A.J., et al., *Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework**. Clin Pharmacol Ther, 2001. **69**(3): p. 89-95.
2. Baker, M., *In biomarkers we trust?* Nat Biotechnol, 2005. **23**(3): p. 297-304.
3. Go, A.S., et al., *Heart Disease and Stroke Statistics--2013 Update: A Report From the American Heart Association.* Circulation, 2013. **127**(1): p. e6-e245.
4. Sedlacek, M.A., and Newsome, Joseph, *Identification of Vascular Bleeding Complications After Cardiac Catheterization Through Development and Implementation of a Cardiac Catheterization Risk Predictor Tool.* Dimensions of Critical Care Nursing, 2010. **29**(3): p. 143-152.
5. Gorenoi, V., M.P. Schonermark, and A. Hagen, *CT coronary angiography vs. invasive coronary angiography in CHD.* GMS Health Technol Assess, 2012. **8**: p. Doc02.
6. Mori, Y., K. Takahashi, and T. Nakanishi, *Complications of cardiac catheterization in adults and children with congenital heart disease in the current era.* Heart Vessels, 2012.
7. Batyraliev, T., et al., *Complications of cardiac catheterization: a single-center study.* Angiology, 2005. **56**(1): p. 75-80.
8. Sulieman, A.A., et al., *Evaluation of Effective Dose to Patients Undergoing Cardiac Catheterization.* 5th European Conference of the International Federation for Medical and Biological Engineering, 2012. **37**: p. 512-515.
9. Patel, P., et al., *Effect of abacavir on acute changes in biomarkers associated with cardiovascular dysfunction.* Antivir Ther, 2012. **17**(4): p. 755-61.
10. Breen, D.P., *Stress tests: how to make a calculated choice.* J Fam Pract, 2007. **56**(4): p. 287-93.
11. Vasan, R.S., *Biomarkers of cardiovascular disease: molecular basis and practical considerations.* Circulation, 2006. **113**(19): p. 2335-62.
12. Babuin, L. and A.S. Jaffe, *Troponin: the biomarker of choice for the detection of cardiac injury.* CMAJ, 2005. **173**(10): p. 1191-202.
13. Sidransky, D., *Emerging molecular markers of cancer.* Nat Rev Cancer, 2002. **2**(3): p. 210-9.
14. LaFramboise, W.A., et al., *Serum protein profiles predict coronary artery disease in symptomatic patients referred for coronary angiography.* BMC Med, 2012. **10**(1): p. 157.
15. Packard, C.J., et al., *Lipoprotein-Associated Phospholipase A2 as an Independent Predictor of Coronary Heart Disease.* New England Journal of Medicine, 2000. **343**(16): p. 1148-1155.
16. Kim, H.C., et al., *Multimarker prediction of coronary heart disease risk: the Women's Health Initiative.* J Am Coll Cardiol, 2010. **55**(19): p. 2080-91.
17. Gordis, L., *Epidemiology*. 4th ed. 2009, Philadelphia: Elsevier/Saunders. xv, 375 p.

18. Katz, M.H., *Multivariable analysis: a primer for readers of medical research.* Annals of Internal Medicine, 2003. **138**(8): p. 644-50.
19. Agresti, A., *An introduction to categorical data analysis*. Wiley series in probability and statistics Applied probability and statistics. 1996, New York: Wiley. xi, 290 p.
20. Grund, B. and C. Sabin, *Analysis of biomarker data: logs, odds ratios, and receiver operating characteristic curves.* Curr Opin HIV AIDS, 2010. **5**(6): p. 473-9.
21. Yuan, Z. and D. Ghosh, *Combining multiple biomarker models in logistic regression.* Biometrics, 2008. **64**(2): p. 431-9.
22. Styerberg, E.W., *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health. 2009: Springer US.
23. Royston, P., Altman, DF, *Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling.* Appl Stat, 1994. **43**: p. 429-467.
24. Royston P, Ambler G, and S. W, *The use of fractional polynomials to model continuous risk variables in epidemiology.* Int J Epidemiol, 1999. **28**(5): p. 654-657.
25. Ray, P., et al., *Statistical evaluation of a biomarker.* Anesthesiology, 2010. **112**(4): p. 1023-40.
26. Soreide, K., *Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research.* J Clin Pathol, 2009. **62**(1): p. 1-5.
27. Kalogeropoulos, A.P., V.V. Georgiopoulou, and J. Butler, *Clinical adoption of prognostic biomarkers: the case for heart failure.* Prog Cardiovasc Dis, 2012. **55**(1): p. 3-13.
28. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society, 1995. **57**(1): p. 289-300.
29. Bailey, D.J., et al., *Proteomic and phosphoproteomic comparison of human ES and iPS cells.* Nature Methods, 2011. **8**: p. 821+.
30. Becker, C.H. and M. Bern, *Recent developments in quantitative proteomics.* Mutat Res, 2011. **722**(2): p. 171-82.
31. Fawcett, T., *An Introduction to ROC Analysis.* Pattern Recognition Letters, 2006. **27**(8): p. 861-874.
32. Zhou, X.-h., D.K. McClish, and N.A. Obuchowski, *Statistical methods in diagnostic medicine*, in *Wiley series in probability and statistics*2011, Wiley: Hoboken, N.J.
33. Copas, J.B.a.C., P., *Overestimation of the receiver operating characteristic curve for logistic regression.* Biometrika, 2002(89): p. 315-331.
34. Özsu, M.T., L. Liu, and SpringerLink (Online service), *Encyclopedia of database systems*, 2009, Springer: New York ; London.
35. Huang, X., G. Qin, and Y. Fang, *Optimal Combinations of Diagnostic Tests Based on AUC.* Biometrics, 2011. **67**(2): p. 568-76.
36. Hastie, T., R. Tibshirani, and J.H. Friedman, *The elements of statistical learning data mining, inference, and prediction*, in *Springer series in statistics*,2009, Springer: New York, NY.
37. Steyerberg, E.W., *Clinical prediction models : a practical approach to development, validation, and updating*. Statistics for biology and health. 2009, New York: Springer. xxviii, 497 p.
38. Chib, S. and E. Greenberg, *Understanding the Metropolis-Hastings Algorithm.* The American Statistician, 1995. **49**(4): p. 327-335.

39. Waller, A.H., et al., *Osteopontin in cardiovascular disease: a potential therapeutic target.* Cardiol Rev, 2010. **18**(3): p. 125-31.
40. Rosenberg, M., et al., *Osteopontin, a new prognostic biomarker in patients with chronic heart failure.* Circ Heart Fail, 2008. **1**(1): p. 43-9.
41. Abdel-Azeez, H.A.-Z., M., *Plasma osteopontin as a predictor of coronary artery disease: association with echocardiographic characteristics of atherosclerosis.* J Clin Lab Anal, 2010. **24**(3): p. 201-206.
42. Mazzone, A., et al., *Osteopontin plasma levels and accelerated atherosclerosis in patients with CAD undergoing PCI: a prospective clinical study.* Coron Artery Dis, 2011. **22**(3): p. 179-87.
43. Vassiliadis, E., et al., *Novel cardiac-specific biomarkers and the cardiovascular continuum.* Biomark Insights, 2012. **7**: p. 45-57.
44. Rios, D.L., et al., *Interleukin-1 beta and interleukin-6 gene polymorphism associations with angiographically assessed coronary artery disease in Brazilians.* Cytokine, 2010. **50**(3): p. 292-6.
45. Contois, J.H., et al., *Apolipoprotein B and cardiovascular disease risk: position statement from the AACC Lipoproteins and Vascular Diseases Division Working Group on Best Practices.* Clin Chem, 2009. **55**(3): p. 407-19.
46. Eriksson, M., et al., *Relationship Between Plasma Fibrinogen and Coronary Heart Disease in Women.* Arteriosclerosis, Thrombosis, and Vascular Biology, 1999. **19**: p. 67-72.