Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Eric Naioti Date

Developing a Five-Biomarker Risk Score for Estimating Five-Year Risk of Myocardial Infarction or Death in a Sample of Subjects that Underwent Cardiac Catheterization

By

Eric Naioti
MSPH

Biostatistics

_____

Dr. Yi-An Ko
Committee Chair

_____

Dr. Michael Kutner
Committee Member

Developing a Five-Biomarker Risk Score Model for Estimating Five-Year Risk of Myocardial Infarction or Death in a Sample of Subjects that Underwent Cardiac Catheterization

By

Eric Naioti
B.S., Geneseo, The State University of New York, 2016

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2018

Abstract

Developing a Five-Biomarker Risk Score Model for Estimating Five-Year Risk of Myocardial Infarction or Death in a Sample of Subjects that Underwent Cardiac Catheterization

By Eric Naioti

Biomarkers of heat-shock protein 70 (HSP70), fibrin degradation products (FDP), soluble urokinase plasminogen activator receptor (suPAR), and C-reactive protein (CRP) have been shown to be associated with heart disease. A biomarker risk score (BRS) based on measures of circulating levels of these biomarkers has been shown to have utility in predicting adverse events in patients with coronary artery disease (CAD). High-sensitivity cardiac troponin I (HS-Trop) has been shown to indicate cardiomyocyte cell damage and could also be a predictor of heart disease. We examined how well a new BRS that included HS-Trop could be used in predicting an outcome of myocardial infarction (MI) or death in patients suspected of CAD. Two thousand eight hundred eighty-six participants were recruited from three Emory healthcare sites in Atlanta as part of the Emory Cardiovascular Biobank (EmCAB). Each participant had measures for each of these biomarkers. The five biomarkers were shown to be independent of each other and useful in predicting MI or all-cause death. It has been shown that using multiple cutoff points to stratify patients into multiple risk groups can be highly favorable. Therefore instead of only finding one cutoff point for each biomarker, optimal numbers and locations of cutoff points were found using the most significant splits of likelihood ratio tests. A BRS based on these stratified biomarkers was found, and a model was constructed using this BRS along with traditional risk factors to predict MI or all-cause death. In this cox proportional hazard model, BRS was found to be highly significant ($\beta = 0.60$, SE = 0.10, $Z=6.03$, $p<0.001$). When this model was compared to a model without BRS, we found using five-fold cross-validation that the model including our BRS improved prediction ability, with a mean increase in the concordance statistic (C-statistic) of 0.0411 [95% CI = (0.0054, 0.0769)] and a mean net reclassification index (NRI) of 0.259 [95% CI =(0.143, 0.376)].

Developing a Five-Biomarker Risk Score Model for Estimating Five-Year Risk of Myocardial Infarction or Death in a Sample of Subjects that Underwent Cardiac Catheterization

By

Eric Naioti
B.S., Geneseo, The State University of New York, 2016

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2018

Acknowledgments

I wish to thank Dr. Yi-An Ko for her guidance and assistance in this research project. Her willingness to giver her time is greatly appreciated.

I would also like to thank Sarah Reeves for her support, encouragement, and confidence in me throughout this process.

Table of Contents

Introduction

Evaluating the risk factors of cardiovascular disease is an area of research with much interest [1, 2] as heart disease remains the number one cause of death worldwide. [3] The ability to predict and quantify a subject's risk of heart disease is highly desirable. This would allow clinicians to identify subjects at highest risk in order to use more aggressive treatment options, risk factor control, and behavioral modification counseling, while identifying low risk subjects could prevent unnecessary testing.[2] Diagnostic risk scores that summarize various risk factors is a practical clinical tool for clinicians and subjects. The Framingham Heart Study led the way in establishing a cardiovascular risk profile for use in a primary care setting.[1] However, there is no established risk scores for clinical or research use for subjects with stable coronary artery disease (CAD).

Given that subjects with CAD often have multiple risk factors, searching for novel and informative biomarkers has been an active research area.[4] A previous study showed that when four biomarkers of heat-shock protein 70 (HSP70), fibrin degradation products (FDP), soluble urokinase plasminogen activator receptor (suPAR), and C-reactive protein (CRP) are added to a model with traditional risk factors, an improved prediction ability was observed. [2] These biomarkers were chosen specifically as they all represent a problem in the human body that can be linked to heart disease. HSP70 is representative of cell stress and FDP represents the coagulation pathway, which could lead to blood clots.[2] CRP is representative of inflammation and suPAR is a marker of immune activation.[2] It has also been seen that high-sensitivity cardiac troponin I (HS-Trop) could be a predictive biomarker for CAD patient outcomes as it indicates cardiomyocyte cell damage and has been associated with heart failure. [5, 6]

Previous biomarker studies such as Ghasemzadeh et al. have been limited due to their exclusion of HS-Trop from their biomarker calculations.[2] Another limitation of past biomarker studies is the use of biomarker cutoffs based on Youden's index, which does not take censored data and follow-up time into account. In this study, we investigated the associations between

these biomarkers and adverse cardiovascular outcomes and further redefined the cutoff points by using likelihood ratios in a survival analysis context. Specifically, we adopted the method proposed by Chang et al. to identify the optimal number of cutoff points and locations of cutoff points for biomarkers. [7] Furthermore, we developed a biomarker risk score (BRS) based on these five biomarkers as a summary of a subject's risk. Finally, we established a prognostic model that incorporates the BRS to predict the risk of adverse outcomes in subjects with CAD. We showed that this BRS-based model significantly improved prediction ability, compared to the one with only clinical risk factors.

Methods

Study Sample

Our study sample was 2,886 subjects with existing coronary artery disease enrolled in the Emory Cardiovascular Biobank (EmCAB). The EmCAB is a prospective cohort that was established in 2003 to identify novel factors associated with heart disease and allow for the discovery of unique predictive factors of subject outcomes such as protein biomarkers. [8] Participants were recruited from subjects that underwent heart catheterization at one of three Emory Healthcare sites in Atlanta.

Demographic information, medical history, medication use, and health behaviors such as cigarette use were collected at enrollment into the cohort. Medical records were reviewed to confirm medical history and medication use. One and five years of follow up were planned and follow up events of death and cardiovascular events were recorded via follow up phone interviews, combined with chart review and national death index verification.

Inflammatory Biomarkers

Circulating levels of our five protein biomarkers were measured in a subset of the EmCAB participants. Levels of high-sensitivity troponin-I were measured by Abbott Laboratories. A sandwich immunoassay by FirstMark was used to determine levels of serum CRP and FDP. HSP-70 measurements were determined by a sandwich ELISA (R&D Systems). Finally, plasma suPAR levels were determined using a commercially available suPARnostic kit (Virogates). [8] All of the 2,886 participants had measurements for the five biomarkers at enrollment.

Analysis Plan

The sample of 2,886 subjects was split randomly into two equal datasets of 1,443 patients: a training dataset, and a testing dataset. A map of the analysis plan can be seen in Figure 1.

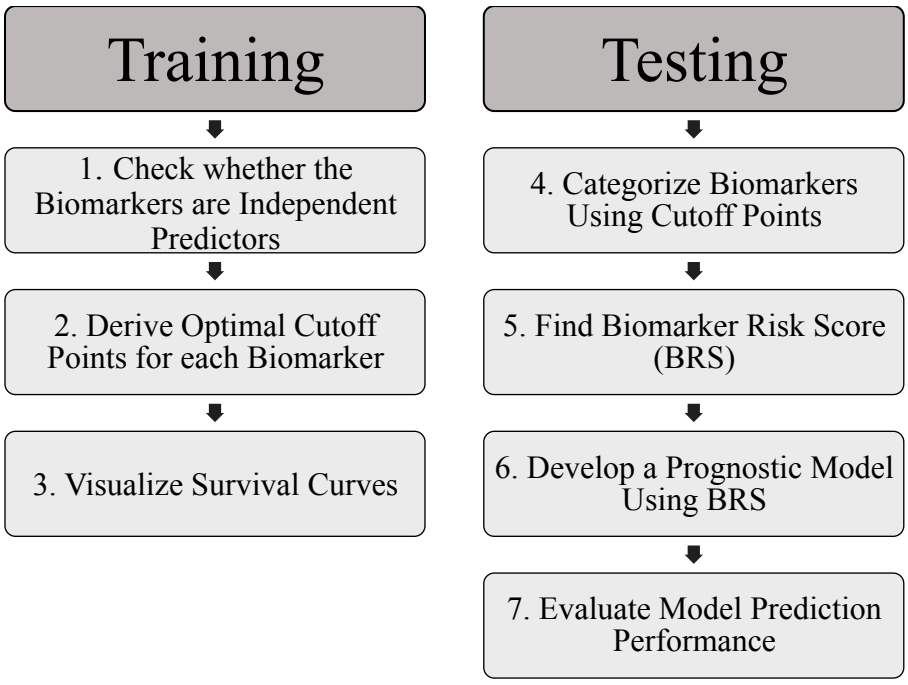

Figure 1. Analysis plan using equal, random samples of our full data set split into a training and a testing set.

Check whether the Biomarkers are Independent

First, we used the training dataset to investigate the correlations among the five biomarkers (HS Troponin, Hsp-70, FDP, suPAR, CRP). The distributions of all five biomarkers appear to stray significantly from a normal distribution, and we verified this using a Shapiro-Wilk normality test. The normality tests were performed in R using the Shapiro test function.[9] Because of the non-normal nature of our distributions, Spearman correlations were used to evaluate associations between biomarkers. Spearman correlations (θ) were determined using the following formula via the cor function in R.[9] In the formula, R_i represents the rank of the i^{th} observation for one biomarker, and S_i is the rank of the same individual's observation of their second biomarker.

$$\theta = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum_i (S_i - \bar{S})^2}}$$

We were further interested in seeing how each biomarker is associated with time to death or myocardial infarction. To do this we constructed a Cox proportional hazards model. The lowest measurement of each biomarker was considered as the limit of detection (LOD); values of zero were replaced with the LOD divided by the square root of two. [10] To better characterize the associations between biomarkers and outcomes, our model was fit using the log-transformed biomarkers, along with the following covariates: age, BMI, gender, race, estimated glomerular filtration rate (eGFR), Gensini score, smoking history, hypertension, diabetes, hypercholesteria, history of heart failure, and previous myocardial infarction. We also included statin use and angiotensin-receptor blocker/angiotensin-converting enzyme inhibitor use (ARB/Ace inhibitor). Age was treated as a continuous variable with knots at ages 60 and 70 using linear splines to assign different slopes for ages ≤60, 60-70, and >70 years. Continuous BMI was used in the model. About 10% of BMI data that were missing have been previously imputed using the entire EmCAB database. Gender and race were dichotomized into male/female and black/non-black, respectively. The data for eGFR were dichotomized into values above or below 60 mL/min/1.73

m^2 and Gensini scores were divided into those above or below 20. Smoking status was categorized into past smoker, current smoking, and non-smoker.

Deriving Optimal Cutoff Points

To facilitate clinical use of the proposed biomarker risk score, we divided each biomarker into categories and the risk score was calculated based on these categories. Rather than simply using the median or previously identified cut points, we attempted to identify the ideal number and location of cutoff points for each protein biomarker. For this we used a method described by Chang et al. [7] For each of our five biomarkers, the optimal number of cutoff points needed to be determined when modeling a survival analysis outcome of either death or a myocardial infarction. The optimal cutoff numbers were found by minimizing the Akaike information criterion (AIC), which takes into account both the model's flexibility and complexity. [7]

To find the location of the cutoff points after determining the number of cutoff points (K) we split the biomarker risk factor (Z) into K+1 ordinal risk groups (Z*) represented by K dummy variables $Z_1,\ldots, Z_k$. A Cox proportional hazards model was then fit where $\beta_1,\ldots, \beta_k$ represent model coefficients. The hazard function at time t is

$$h(t|Z^*) = h_0(t)e^{\sum_{k=1}^{K}\beta_k Z_k}$$

Furthermore, the most advantageous cutoff locations were defined to be those with the most significant likelihood ratio tests. This was all done using the findcutnum and findcut functions. [7] Cutoff numbers and locations for our five biomarkers were determined using the *training* dataset.

Since it has been shown in the literature that sex influences how suPAR predicts heart disease, we decided to find separate suPAR cutoff points for men and women. [11] In order to make sure that no biomarker range would be too small or encompass too few subjects to be relevant in further uses, we forced each ordinal group of Z* to contain at least 100 subjects. For

suPAR cutoffs, since the data were split further into males and females, we forced only one cutoff value to be found in order to assure that we were not forcing too few patients into too many categories.

Visualizing Survival Curves

To examine the effects of the identified cutoff points for each biomarker, Kaplan-Meier survival curves were calculated and plotted (Figure 2.) for each risk group determined by the optimal cutoff points. [12] The Kaplan-Meier survival estimator at time *t* is calculated by the following equation, where d_i represents the number of events and n_i represents the number of those in the risk set at time *t*.

$$\hat{S}(t) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right)$$

Categorizing Biomarkers Using Cutoff Points

The cutoff points derived using the training data set were used to categorize biomarkers in the testing data set. The optimal number of cutoff points derived in the training data set determined the number of subgroups in the testing data set.

Finding a Biomarker Risk Score (BRS)

In order to develop a biomarker risk score (BRS), a Cox proportional hazards model was fit using covariates of only the dummy variables associated with the risk groups of each biomarker. At time *t,* the hazard is

$$h(t|\bar{X}) = h_0(t)e^{\sum_{k=1}^{K} M_k X_k}.$$

In this model, the number of X variables (K) was determined by the optimal number of cutoff points for each biomarker, where M_k is the coefficient corresponding to group k. The BRS was

calculated for each individual by simply taking the sum of all S coefficients multiplied by the X values: $\sum_{k=1}^{K} \widehat{M}_k X_k$.

Developing a BRS-based Risk Prediction Model

Once BRS's are assigned to each individual, we developed another Cox proportional hazards model using the BRS as a continuous variable. The single BRS replaced all of the continuous biomarkers previously used in the model, while the other covariates remained the same: age, BMI, gender, race, eGFR, Gensini score, smoking history, and histories of hypertension, diabetes, hypercholesteria, history of heart failure, previous myocardial infarction, and histories of statin use and ARB/Ace inhibitor use. In this model, however, we wanted to simplify future risk calculations for clinicians; therefore, age and BMI variables were stratified. Age cutoffs were set at 60 and 70 years of age, while BMI cutoffs were set at the standards for low(less than 18.5 kg/m^2), normal (18.5 to 25 kg/m^2) , mildly obese (25 to 30 kg/m^2), and obese (over 30 kg/m^2). This model will provide an estimate of a patient's chance of surviving 5 years without dying or suffering from a myocardial infarction given medical background and biomarker data.

Calibrating the Model

We then checked whether the prediction model was well calibrated by dividing the test data set into deciles based on predicted risk using the BRS-based model, and comparing the average predicted risk of each group to (1 - Kaplan-Meier estimate) for each group. Predicted risks were calculated using the following formula,[1] where the baseline survival S_0 is the estimated survival at t = 5 years.

$$\hat{p} = 1 - S_0(t)^{\exp\left(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \overline{X_i}\right)}$$

Evaluating Discrimination of the Risk Prediction Models

To evaluate the prediction performance of the new model, we evaluated the difference in the concordance statistic (C-statistic) of the model, compared to a model using only the non-BRS

covariates. The C-statistic is commonly used to give a global assessment of the fitted survival function. [13] It measures the area under the receiver operating characteristic curve in order to measure the predictive accuracy of a model. [14] We used the method proposed by Uno et al. [REF], which calculates C-statistics and the inference of risk prediction models with censored survival data. To perform an internal validation for the change in the C-statisticis, the testing data set was split into five equal data sets in order to perform 5-fold cross validation. In this process we used four-fifths of the testing data set to find β coefficient estimates as in Step 5, to be used for defining the weight of each biomarker. These β estimates were applied to the final fifth of the testing data set to calculate the BRS of those individuals. A Cox proportional hazards model was then constructed as in Step 6, and this model was compared to a model on the same data set using only the non-BRS covariates. This process was repeated four more times, choosing a separate fifth of the testing data set each time to be left out for finding the β estimates and to be used in the model to calculate the C-statistic. The C-statistic for each model, and 95% confidence intervals were calculated using the Inf.Cval.Delta function in the SurvC1 R package. [15]

Commonly used alongside the C-statistic in the evaluation of a new risk score is the continuous net reclassification index (NRI) that offers an objective measurement in the improvement of risk classification.[2, 16] Thus, a NRI was also calculated for our model including the BRS against a model without our BRS.

Results

Baseline Characteristics of Study Participants

The baseline characteristics of the study participants are summarized in Table 1. Means and standard deviations are reported for continuous variables and frequencies and percentages are given for binary variables. Our study population consisted of 2,886 participants aged 23 to 91 years old. Sixty-three percent of participants were male, 17% of participants were black. The

mean BMI of all participants was 30 kg/m^2. Thirty-one percent of participants had diabetes, 74% had hypertension, 8% were currently smokers, 58% had smoked in the past, and 70% had high cholesterol. Of our participants, 70% had a history of CAD, 22% had a history of an MI, and 27% had a history of heart failure. Fifty-eight percent of our participants had a history of ACE-inhibitor/ARB use, and 71% had a history of statin use. At five years of follow-up, 3% of all study participants experienced a myocardial infarction, and 13% of participants died. In total, 15% of participants experienced either death or MI at five years.

Table 1. Baseline Characteristics

Characteristics by Dataset	Training Dataset	Testing Dataset
Baseline Characteristics	Total (N=1,443)	Total (N=1,443)
Demographics		
Age, y	63±11	63±12
Male sex, N (%)	916 (63%)	910 (63%)
White, N (%)	1,184 (82%)	1,182(82%)
Black, N (%)	243 (17%)	245 (17%)
BMI, kg/m^2	30±6	30±6
Five-Year Follow-up Events		
MI, N (%)	52 (4%)	49 (3%)
All-cause death, N (%)	196 (14%)	187 (13%)
Death or MI, N (%)	222 (15%)	211 (15%)
Biomarkers		
HS Troponin, pg/mL, median (IQR)	5.0 (2.8-11.4)	4.7 (2.7-10.9)
HSP-70, ng/mL, median (IQR)	0.0 (0.0-0.0)	0.0 (0.0-0.0)
FDP, µg/mL, median (IQR)	0.5 (0.4-0.8)	0.5 (0.4-0.8)
suPAR, ng/mL, median (IQR)	3.0 (2.3-4.0)	3.0 (2.4-4.0)
CRP, mg/L, median (IQR)	2.7 (1.1-6.4)	2.9 (1.3-6.7)
Disease Histories		
Diabetes, N (%)	448 (31%)	448 (31%)
Hypertension, N (%)	1,035 (72%)	1,086 (75%)
Current Smoking, N (%)	107 (7%)	138 (10%)
Past Smoking, N (%)	822 (57%)	839 (58%)
High Cholesterol, N (%)	997 (69%)	1,037 (71%)
History of CAD, N (%)	1,009 (70%)	1018 (71%)
History of MI, N (%)	316 (22%)	329 (23%)
History of Heart Failure, N (%)	391 (27%)	388 (27%)
eGFR ≥ 60 mL/min/1.73 m^2, N (%)	1,084 (75%)	1,081 (75%)
Medications		
ACE-inh/ ARB use, N (%)	842 (58%)	857 (59%)
Statin use, N (%)	1,029 (71%)	1,029 (71%)
Angiographic Findings		
Gensini angiographic score, median (IQR)	60 (0-162)	80 (0-192)
Gensini score ≥ 20, N (%)	830 (61%)	853 (63%)

Correlations of Biomarkers

Spearman correlations among the five biomarkers are reported in Table 2. All the pairwise correlations are less than 0.33, suggesting the potential utility of five biomarkers.

Table 2. Spearman correlations for each pair of biomarkers.

	HS-Trop	**HSP70**	**FDP**	**suPAR**	**CRP**
HS-Trop	1.00	0.08	0.25	0.32	0.15
HSP70	0.08	1.00	0.19	0.16	0.12
FDP	0.25	0.19	1.00	0.30	0.23
suPAR	0.32	0.16	0.30	1.00	0.28
CRP	0.15	0.12	0.23	0.28	1.00

The results of the Cox proportional hazards model showed that all biomarkers were significantly associated with the outcome ($p<0.01$) except for FDP ($p=0.21$). Since our primary aim was to have a model with improved prediction ability, it was not vital that each biomarker was significant in this model, so long as the model was a good fit. To evaluate the model good fit we visualized a Cox-Snell residual plot. [17]

$$r_j = \widehat{H}_0(T_j)\exp(\hat{\beta}' Z_j), j = 1, \dots, n$$

In this equation r_j's are censored sample from a unit exponential distribution, given the assumed Cox model holds and $\hat{\beta}, \widehat{H}_0$ close to the true values β, H_0. When we plot $\widehat{H}_r(r_j)$ versus r_j and see a straight line through the origin, with a slope of 1, this indicates a perfect model fit. Our residual plot can be seen in Figure 2. The plot diverges very little from the $\widehat{H}_r(r_j) = r_j$ line, showing a good overall fit for our model.

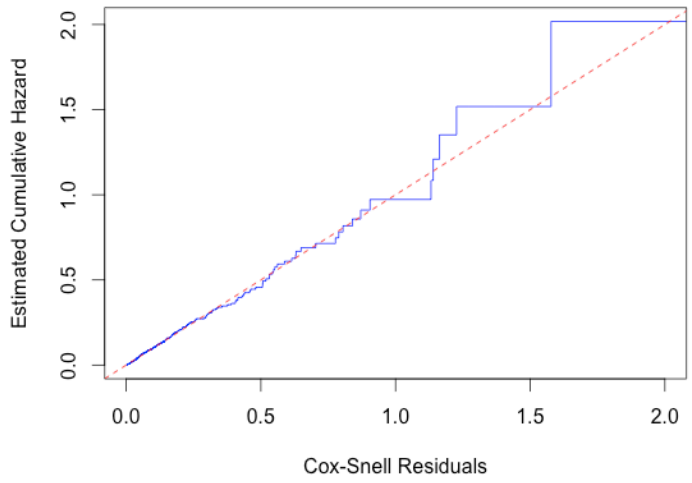

Figure 2. Cox-Snell residual plot for full model using training data with all five biomarkers.

We further examined the functional form of the FDP covariate using martingale residuals. [18] Figure 3 shows the martingale residual plot. The linear nature of the fitted line suggests that this is the correct functional form for the covariate.

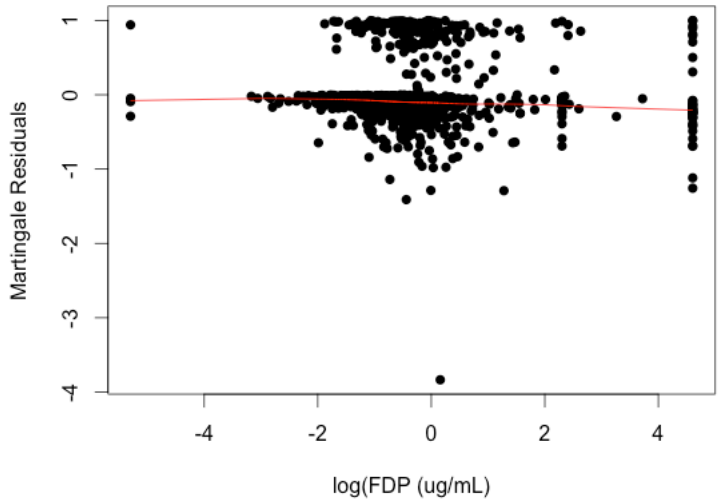

Figure 3. Martingale residual plot for checking the functional form of FDP in the Cox proportional hazards model.

We decided to keep FDP in the model in order to have improved prediction ability due to the model being such a strong fit, and FDP not being highly associated with other biomarkers.

Additionally we checked the proportional hazard assumption of each of the biomarkers by visualizing the Schoenfeld residuals (Figure 4). [19] These plots show that the residuals of four of the five biomarkers (HS-Troponin, FDP, SuPAR, CRP) appear to be centered about zero, which means that the proportional hazards assumption holds. HSP-70 generate a banded residual pattern, which can be explained by the large amount of values for HSP-70 that were below the limit of detection (1,140/1,443 individuals). We concluded that the proportional hazards assumption still held for the biomarker.

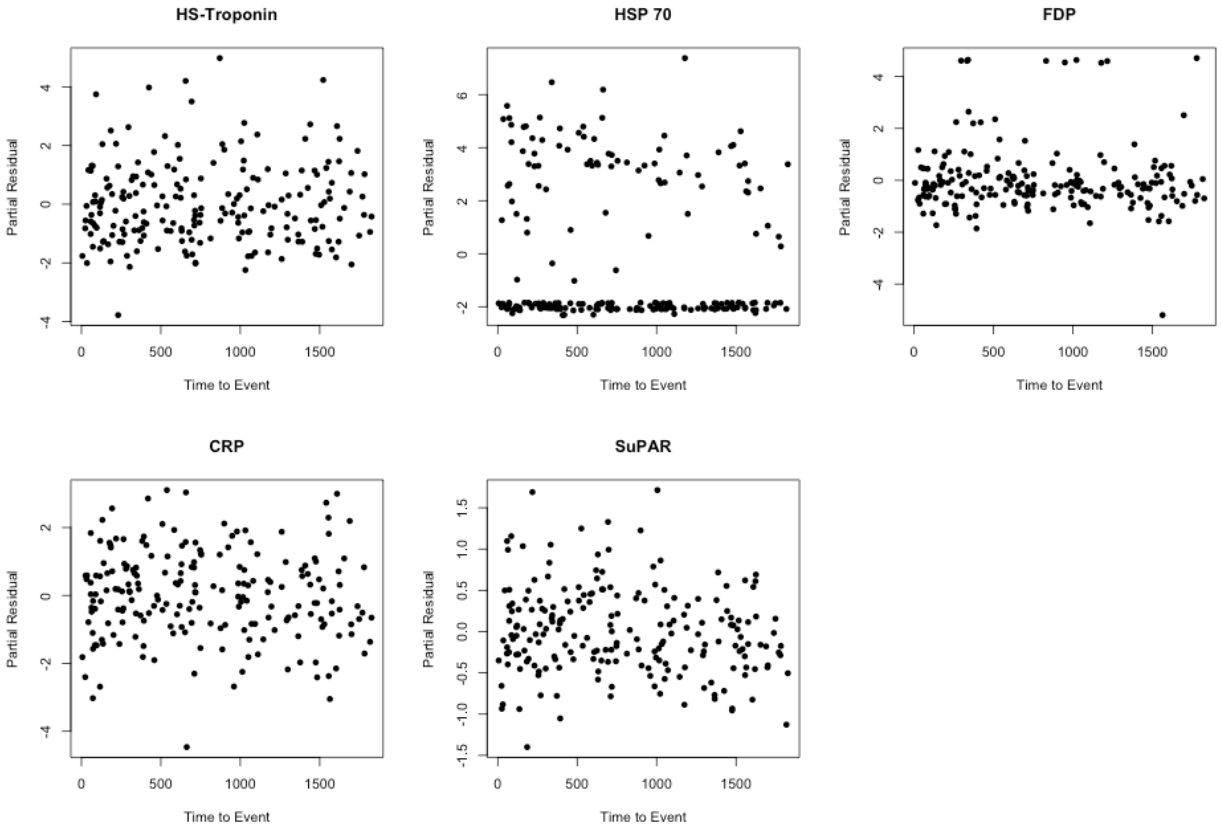

Figure 4. Schoenfeld residuals of the five biomarkers from the Cox proportional hazards model. Residuals centered around zero show that the proportional hazards assumption holds.

Deriving Optimal Cutoff Points

The optimal number and locations of cutoff points for each of the five biomarkers are given in Table 3. The number of individuals from the training and the testing data set that fall into each subgroup after cuts are made can be seen in Table 4.

Table 3. Locations of cutoff points for each of the five biomarkers to be used in calculating a BRS.

Outcome: Death or MI						
Biomarker Name	Scale	Cutoff Points	Cut 1	Cut 2	Cut 3	Cut 4
HS Troponin	(pg/mL)	4	1.6	2.1	2.6	6.9
Hsp-70	(ng/mL)	2	1.0	151.0	--	--
FDP	(µg/mL)	3	0.27	0.48	0.88	--
SuPAR (Men)	(ng/mL)	1*	3.2	--	--	--
SuPAR (Women)	(ng/mL)	1*	3.6	--	--	--
CRP	(µg/mL)	3	1.6	5.8	16.9	--

*We restricted SuPAR to one cutoff point for men and one for women, since it was shown that sex influences how SuPAR affects cardiovascular disease, and we did not want to have too few people represented in each of subgroup after cuts were made.

Table 4. The number of individuals in each data set that fall into the biomarker subgroups defined by the optimal cutoff points derived using the training data set. A minimum of 100 individuals from the training data set were forced into each subgroup to ensure no subgroup would be made too small.

HS Troponin	Training Dataset	Testing Dataset
< 1.6	106	99
1.6-2.1	113	110
2.1-2.6	114	146
2.6-6.9	560	568
≥ 6.9	550	520
Hsp-70		
< 1.0	1145	1147
1.0-151.0	150	154
≥ 151.0	148	152
FDP		
< 0.27	177	159
0.27-0.48	432	440
0.48-0.88	504	538
≥ 0.88	330	306
suPAR (Men)		
< 3.2	584	515
≥ 3.2	332	395
suPAR (Women)		
< 3.6	307	374
≥ 3.6	219	156
CRP		
< 1.6	526	472
1.6-5.8	517	556
5.8-16.9	282	302
≥ 16.9	118	113

Visualizing Survival Curves

Kaplan-Meier survival curves will approximate differences in survival rates for the *training* group based on their biomarker subgroups. We expected to see steeper survival curves for those in the subgroups with the highest biomarker measures, and a more gradual curve for the low biomarker subgroups. This is exactly what we see when we look at the Kaplan-Meier curves in Figure 5. We can see that the four lowest groups of HS-Troponin have similar survival curves, and the two lowest groups of FDP have similar survival curves. Here, there is a possible concern, however, the groups with the highest measures clearly still have steeper survival curves.

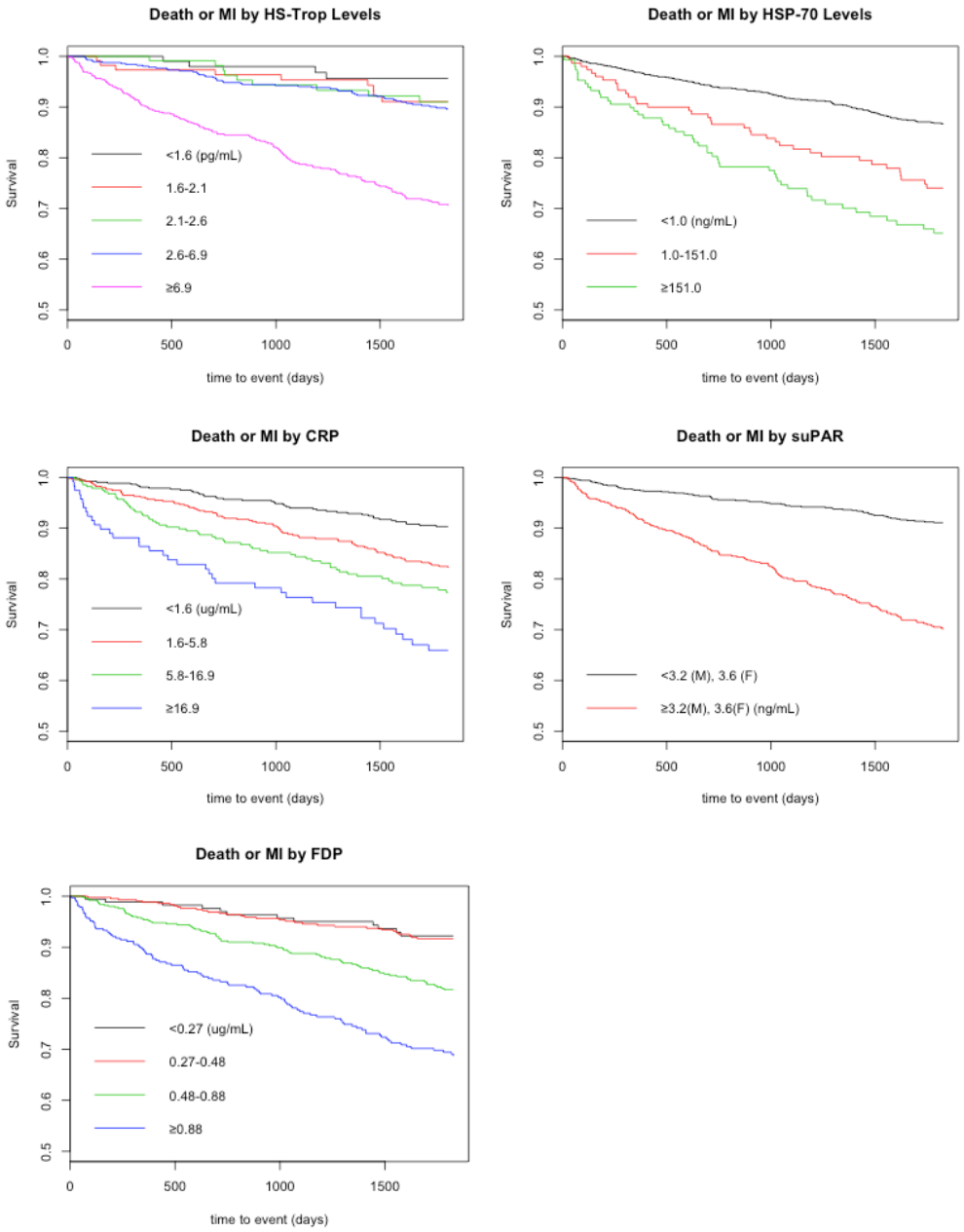

Figure 5. Kaplan-Meier survival curves for subgroups of the five biomarkers for individuals in the *training* data set.

Categorizing Biomarkers Using Cutoff Points

The number of individuals in the testing data set that fall into each of the biomarker subgroup are outlined in Table 5. These subgroups will be used to define dummy variables in a Cox proportional hazards model for finding the BRS.

Finding the BRS

A Cox proportional hazards model was constructed using only dummy variables defined by the biomarker subgroups. BRS was then calculated as BRS = $\sum_{k=1}^{K} \widehat{M}_k X_k$.

$$h(t|\bar{X}) = h_0(t)e^{\sum_{k=1}^{K} M_k X_k}$$

Table 5. X values for the Cox proportional hazards model for finding a BRS.

Dummy Variables ($\bar{X}$)		M coefficients
$X_1 = 1$ if HS-Troponin < 1.6,	o/w $X_1 = 0$	Ref (0)
$X_2 = 1$ if 1.6 ≤ HS-Troponin < 2.1,	o/w $X_2 = 0$	0.27
$X_3 = 1$ if 2.1 ≤ HS-Troponin < 2.6,	o/w $X_3 = 0$	0.34
$X_4 = 1$ if 2.6 ≤ HS-Troponin < 6.9,	o/w $X_4 = 0$	1.00
$X_5 = 1$ if 6.9 ≤ HS-Troponin,	o/w $X_5 = 0$	1.67
$X_6 = 1$ if HSP-70 < 1.0,	o/w $X_6 = 0$	Ref (0)
$X_7 = 1$ if 1.0 ≤ HSP-70 < 151.0,	o/w $X_7 = 0$	0.39
$X_8 = 1$ if 151.0 ≤ HSP-70	o/w $X_8 = 0$	0.58
$X_9 = 1$ if FDP < 0.27,	o/w $X_9 = 0$	Ref (0)
$X_{10} = 1$ if 0.27 ≤ FDP < 0.48,	o/w $X_{10} = 0$	0.69
$X_{11} = 1$ if 0.48 ≤ FDP < 0.88,	o/w $X_{11} = 0$	0.76
$X_{12} = 1$ if 0.88 ≤ FDP,	o/w $X_{12} = 0$	1.11
$X_{13} = 1$ if suPAR < 3.2 (males) or < 3.6 (females)	o/w $X_{13} = 0$	Ref (0)
$X_{14} = 1$ if suPAR ≥ 3.2 (males) or ≥ 3.6 (females)	o/w $X_{14} = 0$	0.73
$X_{15} = 1$ if CRP < 1.6,	o/w $X_{15} = 0$	Ref (0)
$X_{16} = 1$ if 1.6 ≤ CRP < 5.8,	o/w $X_{16} = 0$	0.17
$X_{17} = 1$ if 5.8 ≤ CRP < 16.9,	o/w $X_{17} = 0$	0.18
$X_{18} = 1$ if 16.9 ≤ CRP,	o/w $X_{18} = 0$	0.49

Developing a New Model Using BRS

A Cox proportional hazards model for the outcome of death or MI with commonly used clinical variables along with the aforementioned BRS was fit. The results are shown in Table 6 and follow the equation below. Reference categories for the categorical variables were as follows: age < 60 years, 18.5 kg/m^2 ≤ BMI < 25 kg/m^2, eGFR < 60 mL/min/1.73 m^2, Gensini < 20, female, non-black, non-smoker, with negative histories of hypertension, diabetes, hypercholesteria, heart failure, previous myocardial infarction, statin use, and ARB/Ace inhibitor. Subsequently, baseline hazard or baseline survival function was estimated using these reference categories to calculate the estimated risk for each patient according to his/her clinical risk profile and biomarker levels

(represented by BRS). In this model BRS was found to be highly significant ($p<0.01$) with an increase in BRS of 1 corresponding to a 1.81-fold increase in five-year risk of MI or death.

$$h(t|Z) = h_0(t)e^{\sum_{k=1}^{K}\beta_k Z_k}$$

Table 6. Parameter estimates and corresponding P-values for coefficients of cox proportional hazard model including BRS.

Covariate	$\hat{\beta}$	SE($\hat{\beta}$)	P-value
Age 60-70 vs ≤60	0.27	0.21	0.19
Age >70 vs ≤60	0.46	0.21	0.03
Overweight	-0.02	0.20	0.93
Obese	-0.09	0.21	0.66
Underweight	1.38	0.40	<0.01
High eGFR	-0.63	0.16	<0.01
High Gensini	0.01	0.17	0.95
Male	0.09	0.16	0.56
Black	-0.42	0.22	0.06
Past Smoker	-0.31	0.16	0.06
Current Smoker	0.55	0.25	0.02
Hypertension	0.12	0.19	0.53
Diabetes	0.26	0.16	0.10
Hypercholesteria	0.15	0.18	0.41
MI history	0.05	0.17	0.75
HF history	0.32	0.16	0.04
Statin Use	-0.45	0.18	0.01
ARB/Ace Use	-0.05	0.16	0.78
BRS	0.60	0.10	<0.01

The average 5-year predicted risks of event of each decile of the testing group based on the model are shown in Figure 6. The predicted risk is plotted against Kaplan-Meier estimates for the actual proportion of individuals with events to show that this model is well calibrated.

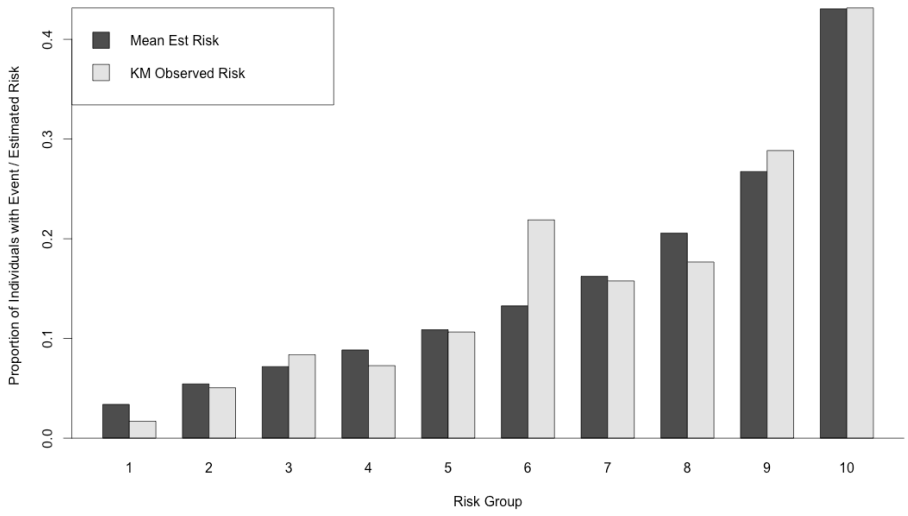

Figure 6. Estimated 5-year risk based on BRS model versus actual proportion of individuals with events for each decile group sorted by increasing estimated risk.

Finding the C-statistic and NRI for a New Model

Five-fold cross validation as used in computing the C-statistic and the NRI of this model when compared to a model with no BRS. The changes in C-statistic (Δ) were calculated for each iteration. The positive Δ values for each iteration show that there is a better discrimination for each iteration when the model with BRS is used. The averaged improvement in C-statistic is 0.0411, and the corresponding 95% confidence interval is (0.0054, 0.0769).

Using the five-fold cross validation, comparing the model with BRS against the model without BRS, the average NRI was calculated to be 0.259 [95% CI =(0.143, 0.376)]. This shows that the addition of BRS to the model improves the risk classification significantly. Results from each iteration of the cross validation study are shown in Table 7.

Table 7. Results from each iteration of our five-fold cross validation study. Estimates for concordance C-statistics of each model, as well as their differences are shown, as well as estimates for NRI.

Iteration	C-stat, No BRS	C-stat, with BRS	Δ	NRI
1	0.753	0.784	0.031	0.173
2	0.711	0.756	0.044	0.334
3	0.742	0.746	0.004	0.206
4	0.733	0.752	0.019	0.201
5	0.701	0.810	0.109	0.385

Discussion

Our results show that the BRS is statistically significant in our Cox proportional hazards model, even when adjusting for several other covariates that are known to affect patient outcomes. This model was also shown to give greater prediction ability than a model that did not include the BRS using a continuous NRI and C-statistic as metrics.

Our results not only give a unique and improved prediction of death or myocardial infarction in our sample but also outlines new cutoff points for five biomarkers that could be used in future studies with similar outcomes. The use of optimal cutoff points in calculating the biomarker risk score is something that should be done in all future studies, and is why we believe our model is an improvement over similar BRS models that have been developed. [2]

Although we see significant results, and are confident in our BRS model, our study has many limitations. Our sample size was relatively small, having only 2,886 patients compared to the 8,491 patients used in the Framingham Heart Study. This limited our study in that we were unable to include a separate replication cohort for validation, instead only using internal validation methods. Participants in the study were all recruited from Emory Healthcare sites in Atlanta, and therefore our results may not be generalizable to the broader population. Cutoff points that were determined for our biomarkers may also not be conclusive for other patient populations.

Based on Figure 6 we can see that the observed risk of risk group 6 was not estimated very well, and we were skeptical of our model because of this. We verified our findings by using a model that combined our lowest two BMI categories (low and normal BMI) and by using a model that included an interaction term between BRS and gender. The results of these models were consistent with our model and we decided to keep the model at it is described, but further investigation may be useful.

Future studies could utilize a much larger patient sample to have an additional validation data set. Moreover, collaborations with different institutions that have the same patient-level data

in the CAD patient population will provide a tremendous opportunity to conduct external replication studies. More importantly, it will allow for further improvement upon this risk prediction model to generate a simple model that provides good prediction accuracy for clinical use.

References

1. D'Agostino, R.B., et al., *General Cardiovascular Risk Profile for Use in Primary Care.* Circulation, 2008. **117**(6): p. 743.
2. Ghasemzadeh, N., et al., *Pathway-Specific Aggregate Biomarker Risk Score Is Associated With Burden of Coronary Artery Disease and Predicts Near-Term Risk of Myocardial Infarction and Death.* Circulation: Cardiovascular Quality and Outcomes, 2017. **10**(3).
3. Lozano, R., et al., *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010.* The Lancet, 2012. **380**(9859): p. 2095-2128.
4. Shlipak, M.G., et al., *Biomarkers to Predict Recurrent Cardiovascular Disease: The Heart and Soul Study.* The American Journal of Medicine. **121**(1): p. 50-57.
5. McEvoy, J.W., et al., *6-Year Change in High Sensitivity Cardiac Troponin-T and Risk for Subsequent Coronary Heart Disease, Heart Failure and Death.* JAMA cardiology, 2016. **1**(5): p. 519-528.
6. Zhu, K., et al., *High-sensitivity cardiac troponin I and risk of cardiovascular disease in an Australian population-based cohort.* Heart, 2017.
7. Chang, C., et al., *Determining the optimal number and location of cutoff points with application to data of cervical cancer.* PLOS ONE, 2017. **12**(4): p. e0176231.
8. Ko, Y.-A., et al., *Cohort profile: the Emory Cardiovascular Biobank (EmCAB).* BMJ Open, 2017. **7**(12).
9. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2017, R Foundation for Statistical Computing.
10. C Croghan, P.P.E., *METHODS OF DEALING WITH VALUES BELOW THE LIMIT OF DETECTION USING SAS*. Presented at Southeastern SAS User Group, St. Petersburg, FL, September 22-24, 2003.
11. Eapen, D.J., et al., *Soluble Urokinase Plasminogen Activator Receptor Level Is an Independent Predictor of the Presence and Severity of Coronary Artery Disease and of Future Adverse Events.* Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease, 2014. **3**(5): p. e001118.
12. Therneau TM, L., Thomas, *Survival Analysis*. 2016: github.
13. Uno, H., et al., *On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data.* Statistics in medicine, 2011. **30**(10): p. 1105-1117.
14. Austin, P.C. and E.W. Steyerberg, *Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable.* BMC Medical Research Methodology, 2012. **12**(1): p. 82.
15. Uno, H., *survC1*. 2013, : https://cran.r-project.org/web/packages/survC1/index.html.
16. Pencina, M.J., E.W. Steyerberg, and R.B. D'Agostino, *Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers.* Statistics in Medicine, 2011. **30**(1): p. 11-21.
17. Cox, D.R. and E.J. Snell, *A General Definition of Residuals.* Journal of the Royal Statistical Society. Series B (Methodological), 1968. **30**(2): p. 248-275.
18. Therneau, T.M., P.M. Grambsch, and T.R. Fleming, *Martingale-based residuals for survival models.* Biometrika, 1990. **77**(1): p. 147-160.
19. Schoenfeld, D., *Partial residuals for the proportional hazards regression model.* Biometrika, 1982. **69**(1): p. 239-241.
20. Lindholm, D., et al., *Biomarker-Based Risk Model to Predict Cardiovascular Mortality in Patients With Stable Coronary Disease.* J Am Coll Cardiol, 2017. **70**(7): p. 813-826.