Risk Prediction in Cardiovascular Epidemiology: Considerations for Modeling Composite Endpoints

David Ezekiel-Herrera

A thesis
submitted in partial fullfillment of the
requirements for the degree of

Master of Science

University of Washington

2016

Reading Committee:

Robyn McClelland, Chair

Mary Lou Thompson

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Risk Prediction in Cardiovascular Epidemiology:
Considerations for Modeling Composite Endpoints

David Ezekiel-Herrera

Chair of the Supervisory Committee:
Robyn McClelland
Department of Biostatistics

This thesis reviews construction and evaluation of risk prediction scores in the context of cardiovascular disease. We give an overview of the clinical guidelines that make use of cardiovascular disease risk scores as well as current practices in recalibration of risk scores outside of the development population. Because cardiovascular disease is a composite of heart attack and stroke, we summarize approaches to modeling and recalibrating composite endpoints and give our major reservations about the clinical utility of such a composite endpoint and the ability to recalibrate such an endpoint. A method of predicting individual outcomes is proposed that is roughly statistically equivalent to the current practice in terms of calibration and discrimination but increases useful information for clinicians in that, for a given composite risk, risks for the component outcomes can vary greatly. Knowledge of these component risks should influence clinical recommendations. Limitations are discussed and next steps are suggested, largely in terms of simulating competing risks survival data and comparing the performance of recalibration methods for composite endpoint whose component risk models differ substantially.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author would like to acknowledge the statistical support and guidance provided by Robyn McClelland, Mary Lou Thompson, Scott Emerson, Andrew Spieker, Fedelis Mutiso, Lanae Schaal, and Leila Zelnick.

Chapter 1

BACKGROUND

Atherosclerotic cardiovascular disease (ASCVD), disease that affects the heart and blood vessels as a result of plaque accumulation in the arteries, is a matter of high public health importance which causes more morbidity and mortality than any other affliction in the developed world [Janic et al., 2013]. As of 2012, the yearly incidence of myocardial infarction (MI) in the United States was 600,000, and, among the people for whom this manifested as a sudden cardiac death, half of these events were not preceded by cardiovascular disease symptoms [Whelton et al., 2012]. To address the surreptitious nature of some cardiovascular events, scientists have devoted much effort to develop methods to detect and prevent cardiovascular disease prior to presentation of clinical symptoms. Accurate prediction could result in saved lives, improved primary prevention with treatment given to those whose need might previously have gone undetected, and saving medical costs by reducing ineffective spending in patients who are at low risk. Cardiovascular disease risk prediction is based on statistical modeling of outcomes as a function of subjects demographic, behavioral, and clinical characteristics. With these risk scores, medical organizations have developed guidelines that recommend treatment for patients at elevated risk of disease without needing to depend on the onset of clinical symptoms.

1.1 Guidelines in Risk Scoring

One of the preeminent sources of cardiovascular risk scoring and guidelines is the set of ACC/AHA guidelines to reduce cardiovascular risk. This effort is the product of collaboration between the American College of Cardiology (ACC), the American Heart Association (AHA), and the National Heart, Lung, and Blood Institute (NHLBI). This is an ongoing ef-

fort the NHLBI launched in 2008 with recommendations stratified into four classes based on difference of benefit and risk and three levels of strength of evidence. There are many levels of review to these guidelines including the evaluation of risk from treatment, potential for benefit from treatment, and 10-year risk of ASCVD. These will be used to evaluate certainty and estimated benefit for the ACC Cholesterol guidelines in Table 1.1.

The classes of treatment effect and harm risk are as follows:

- Class I
  - Benefit greatly exceeds risk
  - Procedure/treatment should be performed/administered
- Class IIa
  - Benefit somewhat exceeds risk
  - Additional studies with focused objectives needed
  - It is reasonable to perform procedure/administer treatment
- Class IIb
  - Benefit exceeds or equals risk
  - Additional studies with broad objectives needed; additional registry data would be helpful
  - Procedure/treatment may be considered
- Class III No Benefit or Class III Harm
  - No Benefit: Procedure/treatment is not helpful or has no proven benefit
  - Harm: Procedure/treatment has excess cost w/o benefit or is harmful.

The levels of strength of evidence:

- Level A
  - Multiple populations evaluated
  - Data derived from multiple randomized clinical trials or meta-analyses
- Level B
  - Limited populations evaluated
  - Data derived from single randomized trial or nonrandomized studies

- Level C
  - Very limited populations evaluated
  - Only consensus opinion of experts case studies, or standard of care

Table 1.1: ACC/AHA Blood Cholesterol Treatment Guidelines

Primary Prevention In Individuals With LDL-C 70-189 mg/dL	
With Diabetes	Without Diabetes
Moderate-intensity statin therapy should be initiated or continued for adults 40-75 years of age with diabetes Class: I Level: A	Pooled Cohort equations should be used to estimate 10 year risk for individuals with LDL-C 70-189 mg/dl without clinical ASCVD to guide initiation of statin therapy for primary prevention of ASCVD Class: I Level: B
High-intensity statin therapy is reasonable for adults 40-75 years of age with diabetes with a $\geq$7.5% estimated 10-year ASCVD risk unless contraindicated Class: IIa Level: B	Adults 40-75 years of age with LDL-C 70-189 mg/dL without clinical ASCVD or diabetes with an estimated 10-year ASCVD risk $\geq$7.5% should be treated with moderate- to high-intensity statin therapy. Class: I Level: A
In adults with diabetes, who are $<$ 40 years of age or $>$ 75 years of age it is reasonable to evaluate the potential for ASCVD benefits and for adverse effects and drug-drug interactions and to consider patient preferences when deciding to initiate, continue, or intensify statin therapy. Class: IIa Level: C	It is reasonable to offer treatment with a moderate-intensity statin to adults 40-75 years of age with LDL-C 70-189 mg/dL without clinical ASCVD or diabetes with an estimated risk of 5% to $<$ 7.5% Class: IIa Level: B

Ten year risk of ASCVD is defined as the risk of developing a first ASCVD event, an event composed of non-fatal myocardial infarction (MI), coronary heart disease (CHD) death (which includes fatal MI), and fatal or nonfatal stroke over the course of 10 years. Moving away from the use of the popular Framingham score, the ACC risk equations proportional hazards models built on a cohort pooled from several studies that included white and African-American subjects including the Atherosclerosis Risk in Communities study (ARIC), the Cardiovascular Health Study (CHS), and the Coronary Artery Risk Development in Young Adults study (CARDIA). The risk factors (covariates) included in these risk equations were: age, total cholesterol, high-density lipoprotein cholesterol, systolic blood pressure stratified by treatment status, diabetes mellitus, and current smoking status [Goff et al., 2014]. As outlined in Table 1.1, 7.5% or greater 10-year risk of ASCVD is the level of risk at which moderate to high-intensity statin-treatment is recommended to adults 40-75 years old, normal levels of LDL-cholesterol (70-189 mg/dL), and without diabetes while high-intensity treatment is recommended to similar adults with diabetes. Moderate-intensity statin treatment is recommended to all adults 40-75 with normal LDL-C levels and diabetes, but only recommended to their counterparts without diabetes if they have 5-7.5% 10-year risk of ASCVD. The ratings and levels in the table show that these thresholds are considered to be the points where the authors of these guidelines believe with confidence that the benefit of treatment outweighs potential harm. As such, the creation of valid and accurate risk scores is worth thorough investigation.

The equations used by the ACC/AHA determine the recommendations given to clinicians for treatment and come from a greatly expanded cohort compared to previous risk scores; however, all risk scores present challenges regardless of source population. Evaluation of the Framingham ASSIGN, and QRISK2 CVD risk scores has shown that population risk prediction is very good, as is identification of low-risk subjects, but high-risk subjects are not well identified when compared with a competing risk Cox proportional hazard model [Björnson et al., 2016]. Identification of high-risk subjects and calibration of risk scores will be discussed further in subsequent sections.

1.2 MESA: The Multi-Ethnic Study of Atherosclerosis

We motivate the work in this thesis using data from the Multi-Ethnic Study of Atherosclerosis (MESA) is a population-based cohort study. Examination of cohort members in MESA started in 2000 and follow-up examinations continue to this day. The objectives of MESA are described as follows [Bild et al., 2002]:

1. to determine characteristics related to progression of subclinical CVD to clinical CVD
2. to determine characteristics related to progression of subclinical CVD itself
3. to assess ethnic, age, and sex differences in subclinical disease prevalence, risk of progression, and rates of clinical CVD
4. to determine relations of newly identified factors with subclinical disease and to determine their incremental predictive value over established risk factors
5. to develop methods, suitable for application in future screening and intervention studies, for characterizing risk among asymptomatic persons

The study drew a population of 6,814 subjects aged 45-84 with no previous cardiovascular disease from 6 communities across the United States with representation from white, African-American, Hispanic, and Chinese American populations. In addition to an expanded population-base for the study, MESA also measures, among other biomarkers, Coronary Artery Calcification (CAC), a marker of atherosclerotic disease that is proving to be a valuable predictor of coronary heart disease. CAC and its use in our risk score will be discussed in Chapter 4. MESA's sampling design and measurement of predictors of disease is consistent with the ACC/AHA's preferences in moving away from the Framingham risk algorithm due to a need for a more ethnically and geographically diverse development population as well as the general goal of understanding, predicting, and preventing ASCVD before any clinical symptoms manifest.

1.3 Clinical Limitations of Composite Endpoints

Most major risk equations for cardiovascular disease rely on some composite measure of cardiovascular disease. Typically this measure is the first of a number of events centered

around fatal or nonfatal stroke or MI. Clinically, this is of limited value since these outcomes can have dramatically different relationships to risk factors. That is, one subject with high-risk of the composite endpoint may have that high-risk due to simultaneously high-risk of stroke or CHD, while another subject with identical composite risk could have that risk due primarily to one outcome or the other. To illustrate this point we analyzed the MESA cohort using a competing risks proportional hazards model with covariates including: the centered continuous variables age, systolic blood pressure, cholesterol, high-density lipoprotein (HDL) cholesterol, and the indicators for male gender, Black, Hispanic, and Chinese race, diabetic status, smoking status, lipid lowering medications, hypertension medication, and family history of heart attack. These covariate vectors were fit separately on the time to coronary heart disease events and stroke. Nonlinearities have been investigated for the continuous variables using generalized additive models and, while mild evidence of nonlinearities was found for age and systolic blood pressure, these higher order associations were not selected when using penalized regression techniques, this process has been described in further detail for CHD in McClelland et al. [2015]. The results are in Table 1.2, which shows the hazard ratio, the factor by which risk changes given a a one-unit increase in the covariate holding others constant, for each covariate as well as their respective p-values and 95% confidence-intervals, and the baseline survival at 10 years. Hazard ratios that differ substantially by outcome are bolded for emphasis.

The implications of the differences between these models are substantial. Depending on the gender, race, or use of lipid-lowering medications, an individual's cardiovascular disease risk could be very asymmetrically driven by two non-trivially different cardiovascular disease events. Figure 1.1 illustrates the range of risk combinations that could occur given one level of composite cardiovascular disease risk. In this figure, points are colored based on composite cardiovascular disease risk, cut off at 0.5, 0.1, and 0.2 which we have selected for purposes of illustrating the variation in relatively similar levels of composite risk.

As seen in Figure 1.1, an overall cardiovascular risk could be driven fairly evenly between the two component disease risks; however, there are some regions where one disease is largely

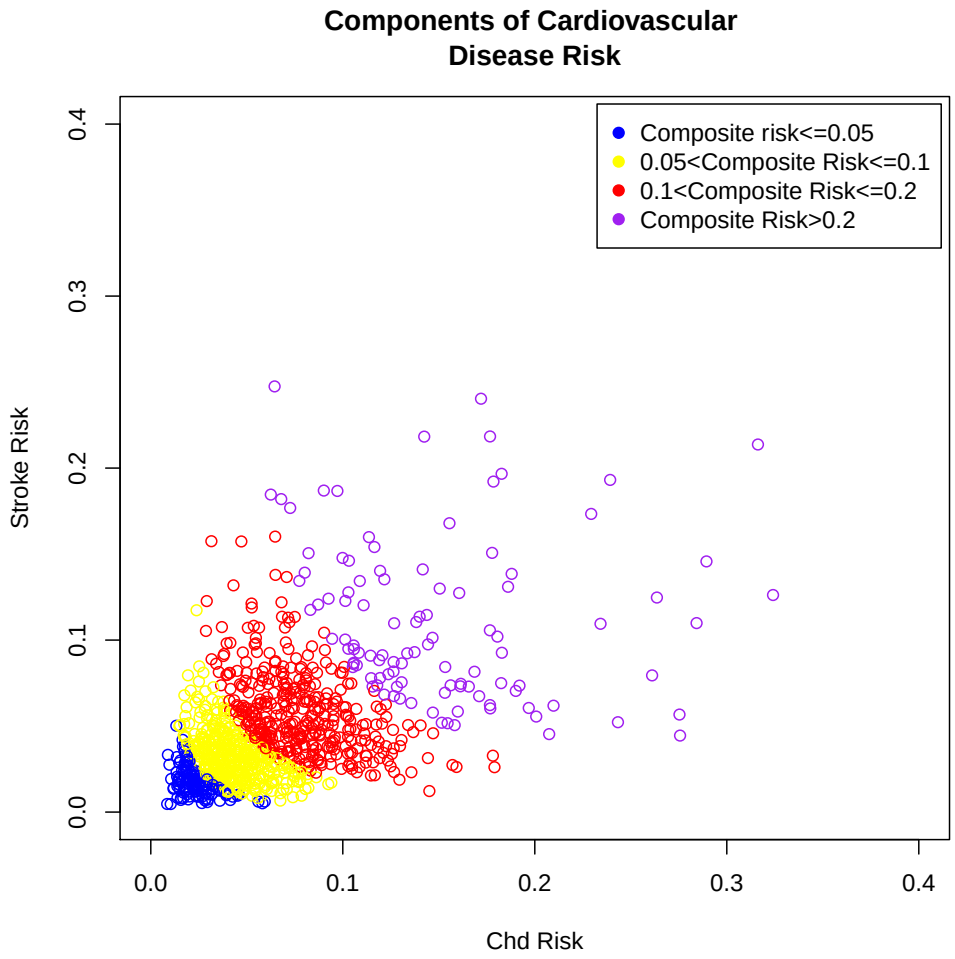
Table 1.2: MESA Coefficients

Regression Results								
	CHD				Stroke			
	HR	P-Value	95% CI		HR	P-Value	95% CI	
Age (years)	1.049	< 0.001	1.039	1.059	1.044	< 0.001	1.028	1.060
Male Gender	**2.096**	< 0.001	1.715	2.562	**1.034**	0.830	0.764	1.400
Chinese	0.721	0.040	0.527	0.985	0.746	0.290	0.436	1.280
Black	0.787	0.030	0.633	0.977	0.929	0.690	0.648	1.330
Hispanic	**0.765**	0.026	0.604	0.968	**1.309**	0.130	0.921	1.860
Systolic BP (mmHg)	1.009	< 0.001	1.005	1.013	1.017	< 0.001	1.010	1.020
Diabetes	1.762	< 0.001	1.421	2.185	1.504	0.024	1.056	2.140
Smoking Status	1.681	< 0.001	1.320	2.140	1.659	0.011	1.124	2.450
Cholesterol (mg/dL)	1.003	0.047	1.000	1.005	1.003	0.100	0.999	1.010
HDL Cholesterol (mg/dL)	0.988	0.002	0.981	0.996	0.992	0.120	0.981	1.000
Lipid Lowering Medications	**1.191**	0.110	0.961	1.477	**0.924**	0.670	0.641	1.330
Hypertension Medications	1.256	0.020	1.037	1.522	1.348	0.060	0.988	1.840
Family History of Heart Attack	1.546	< 0.001	1.296	1.843	1.077	0.600	0.814	1.430
$S_0(10)$[1]	0.973				0.985			

accountable for the overall risk. This is problematic for clinical decision making since stroke and MI have different health behaviors and clinical interventions that affect modify a patients risk. Without some idea of which cardiovascular disease event is driving a patients high

[1]Baseline survival at 0 values of all variables. Continuous variables are centered at the mean values in MESA, refer to Table 3.1 for these values.

Figure 1.1: Components of Composite ASCVD Risk



disease risk, interventions will not be precisely targeted which runs counter to the intent and presumed benefit of risk scoring.

1.4 Goals of the Thesis

The general goal of this thesis is to evaluate the relative merits of two different modeling strategies for cardiovascular disease risk-score development. For the first strategy, which we refer to in this thesis as the combined approach, model the two first event subtypes

separately using a competing risks framework, and then combine the event-specific risks to form a risk-score for the composite event. In the second strategy, model the overall composite endpoint. Both strategies treat non-CVD mortality as a competing risk. We hypothesize that the composite model, an average of the event-specific models weighted by the proportion of events of each subtype, will suffer from worse calibration when applied to a new population with a substantially different case mix of events. Additionally we hope that with the addition of a non-invasive biomarker (CAC) and a more carefully defined disease state, we will be able to contribute positively to primary prevention of cardiovascular disease in the United States and abroad.

Chapter 2

RISK SCORE DEVELOPMENT AND EVALUATION

The most basic setting to predict risk of an event is when the outcome is truly a single event. The use of composite events is an attempt to take multiple events and have them approximate the single event setting such that the single event procedure is appropriate. To take multiple, mutually exclusive, events into account; one must adopt a competing risks framework wherein one tries to isolate an outcome while acknowledging the effect of the other outcomes on the incidence of the outcome of interest. In this chapter we summarize the current approaches to modeling the risk of single events across time, and the approaches taken to modeling composite endpoints. We will review calibration, a popular measure to evaluate model validity and the range of methods to recalibrate a model to a new population at risk. Finally, we will conclude with our reservation about recalibration using composite endpoints that has motivated the competing risks aspect of our risk score.

2.1 General Approach for Single Endpoints: The Cox Model

One of the most frequently used tools in risk prediction is the Cox proportional hazards model. The Framingham risk score, for instance, uses the coefficients derived from the proportional hazards model to estimate 10-year CHD risk[Wilson et al., 1998]. The model is useful in that it models time to an event of interest making it possible to predict risk at a specified point in time as well as making it possible to account for the censoring due to loss-to-follow-up that is endemic in the sorts of cohort studies upon which risk scores are built.

The Cox model is a semi-parametric regression model with a parametric covariate structure but a nonparametric dependence on time. The formulation of the Cox proportional

hazards model is as follows:

$$h(t, X, \beta) = h_0(t)e^{X\beta} \tag{2.1}$$

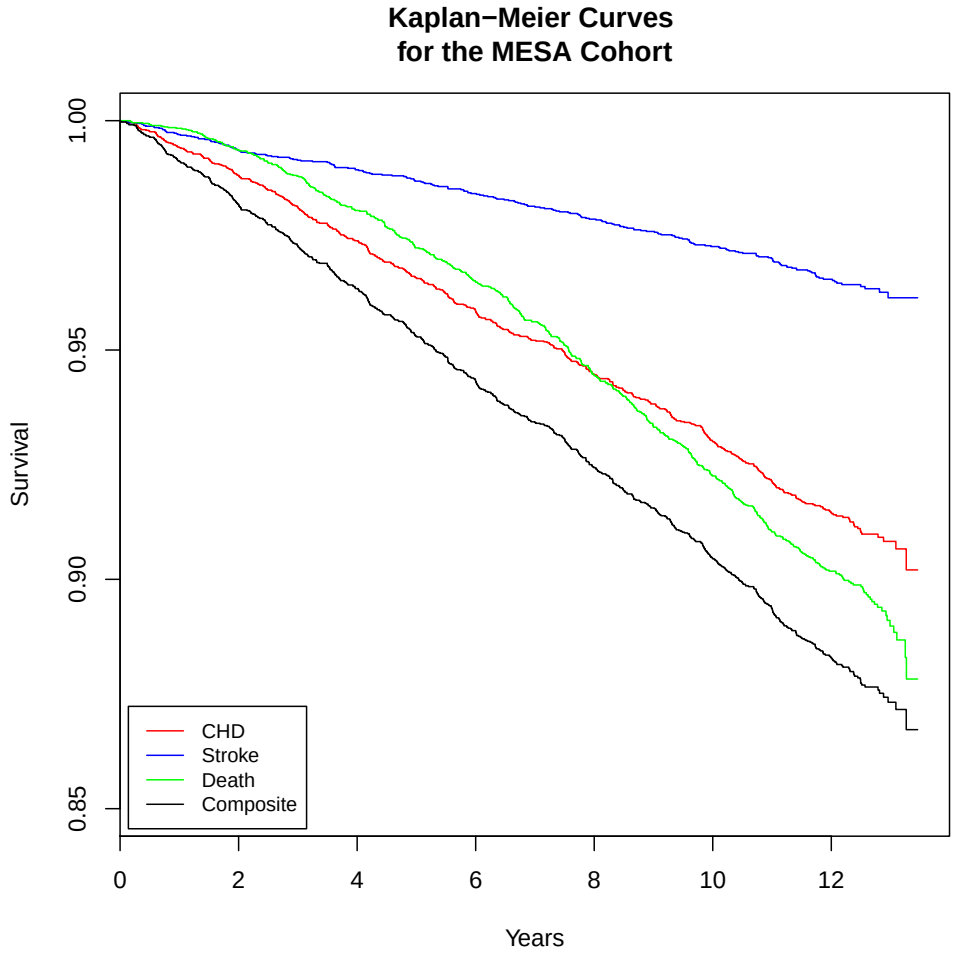
Here, h(t,X,β) is the risk of an event at time t, and letting $i \in 1, \ldots, I$ be an index of observations in the study and $j \in 1, \ldots, J$ be an index of the covariates of interest, we have a design matrix $X_{I\times J}$ of covariates obtained from the population that is analyzed, and the corresponding vector of coefficients that parameterize the relationship between the covariates and the change in risk of an event β. The exponentiated coefficient, e^{β_j} is equal to the hazard ratio associated with a one unit increase in the corresponding covariate X_j holding the other covariates constant, ie subjects who have a value of X_j one unit higher than otherwise similar subjects have an average e^{β_j} times higher risk of the event at time t. Parameterization based on the exponentiated regression coefficient β is the basis for the proportional hazards assumption, meaning that an additive increase in a risk factor corresponds to a proportional change in the hazard. The nonparametric proportion of the model is $h_0(t)$ which is the baseline hazard of the population for whom the vector of covariates X equals zero at time t, this is the same as one minus the height of the Kaplan-Meier curve at time t for observations with X=0. See Figure 2.1 for the Kaplan-Meier curve for the whole MESA cohort. The resulting $h(t, X, \beta)$ is the estimated hazard at time t, which is converted into the cumulative hazard that practitioners will ultimately use to stratify people into risk groups to guide behavioral interventions and clinical therapy recommendations. The formula for the cumulative hazard, $H(t, X, \beta)$ up to time t has the following form:

$$H(t, X, \beta) = \int_0^t h(u, X, \beta)du \tag{2.2}$$

Here, equation 2.1 is integrated across time to aggregate the time-specific hazards into a total hazard. In a clinical or public health setting, it is highly unlikely that a population is at risk for only one type of disease event. As such, a non-trivial amount of what is regarded as censoring in the Cox model may be due to a different event that precludes our event of

interest. This is referred to as a competing risk. The higher the proportion of events that the competing risks take, the more important it is explicitly account for them in our modeling approach.

Figure 2.1: Kaplan Meier Plots in MESA



2.2 Approaches for Multiple Endpoints

The Cox model is the long-standing gold standard for modeling a single endpoint and is widely used in current prognostic models for single outcomes of interest and outcomes whose

status as a single event is debatable. For our purposes, we are explicitly interested in predicting more than one type of disease outcome. Our proposed risk score would predict the risk of cardiovascular disease, defined as either a fatal or nonfatal coronary heart disease (CHD) event or a fatal or nonfatal stroke. To succinctly categorize risk of cardiovascular disease, many researchers and clinicians rely on a composite endpoint. The composite approach to modeling cardiovascular disease combines the component endpoints of CHD and stroke and gives an indicator of whether either event has occurred. The time to the composite event is the time to the first of the events to occur whether or not a second event occurs.

When modeling the individual components of a composite endpoint, adopting a method that takes competing risks into account is essential. Methodologically, competing events affect the censoring distribution of the event of interest. Clinically, it is important to make the distinction between an event whose risk was genuinely reduced and one whose risk was reduced because another negative outcome precluded it. In terms of reducing overall morbidity and mortality, a competing risks framework allows us to differentiate between a success and a failure.

A widely used regression model that estimates cumulative incidence is the Fine and Gray model[Fine and Gray, 1999] which estimates a baseline subdistribution hazard and proportional hazards related to the covariates in the model. If we let $k \in (1, \ldots, K)$ be an index of possible events, what the Fine and Gray model ultimately estimates is the cumulative incidence for event k, $I_k(t|X)$, which has the following form:

$$I_k(t|X) = 1 - e^{-e^{X\beta} \int_0^t \bar{\lambda}_{k,0}(s)ds} \tag{2.3}$$

Again, X is a matrix of covariates just as one would use in a Cox proportional-hazards regression. Also similar to the formulation of the Cox model, $\bar{\lambda}_{k,0}$ is the subdistribution baseline hazard of event k at time t when all covariates values in X are zero. We integrate this across time from the beginning of time at risk (or initial observation) to time t to obtain the cumulative subdistribution baseline hazard. The regression coefficients that form the vector β are the log hazard ratios associated with the unit change in a covariate and have

the same proportional hazards assumptions and interpretations as the coefficients in a Cox proportional hazards model. Of the competing risks approaches, this one is favored for its direct estimation of cumulative incidence and ease of interpretation of model components.

2.3 Calibration

Risk equations can vary in quality as predictors of disease risk depending on the selection of covariates, the quality of the data on which the equation is modeled, and the similarity of the population on which the score is applied to the original dataset. Calibration, or the observed risk relative to the predicted risk, can be evaluated if there is a cohort on which a substantial amount of follow-up time and events have been accrued. Hosmer and Lemeshow's goodness-of-fit test is a standard measure of calibration that takes a risk set and divides them into levels of risk, often into ten equal-sized risk groups. Then squared difference between expected and observed numbers of events scaled by the expected number of events is compared to a chi-squared distribution with g-2 degrees of freedom where g is the number of risk groups into which the population is divided. Thresholds for significantly poor goodness of fit can vary, the conventional threshold of $p < 0.05$ is not unreasonable, though many papers that evaluate the calibration of the Framingham risk score use a threshold of $p < 0.01$.

2.4 Discrimination

In addition to calibration as a measure of the performance of risk prediction scores, we also consider discrimination. In brief, discrimination is the ability to differentiate between observations with and without the outcome. The method we use is the area under the receiver operating characteristic (ROC) curve, referred to from here on as the AUC. The ROC curve plots sensitivity against 1-specificity such that, varying the proportion we allow for non-cases to be identified as cases, the height of the curve represents the proportion of cases predicted to have the event. The resulting plot shows how well the predictive model classifies cases and non-cases at each classification cut-off. A test that perfectly predicted

events would have an AUC of 1 and a worthless test would have an area of 0.5. The AUC is, in the binary outcome case, equivalent to the concordance or c-statistic. The c-statistic is calculated in the following way:

1. Split all observations into cases and non-cases
2. Compare each possible pair of cases and non-cases
3. Assign a value of 1 to each concordant pair, ie where the case has a higher risk estimate than the non-case
4. Assign a value of 0.5 to each tie
5. Take the sum of the values assigned to the total number of pairs

This statistic, though insensitive to systematic errors in calibration, is a valuable metric to use in conjunction with the calibration statistics we will be using as our main evaluation measure.

2.5 Recalibration

2.5.1 Techniques for Single Endpoints

Recalibration is a technique of improving the fit of a predictive model to a new population or context outside of the one in which is was developed. Recalibration techniques have been used to recalibrate to groups differing in age, ethnicity, geographic location and other factors that would impact the components of their risk estimates. Additionally, recalibration can be used to validate a model within the same population or even quantify the impact a new predictor has on the model as a whole. In their 2010 paper, Royston and collaborators in England gave an in-depth review of external validation of prognostic survival models including a review of recalibration of survival models that form the risk scores like the ones we have reviewed[Royston and Parmar, 2010]. The authors propose a flexible parametric method of estimating baseline hazard that is beyond the scope of this investigation. Limiting to the parameterization offered by the semi-parametric Cox model, the authors describe four potential recalibration techniques. In Table 2.1, β_2 refers to a linear scalar on the $X\beta$ fit by

the original predictive model.

Table 2.1: Recalibration Methods

Recalibration Cases		
	$\beta_2 = 1$	β_2 Estimated
h_0 Not Re-Estimated	Case 1	Case 3
h_0 Re-Estimated	Case 2	Case 4

Table 2.1 outlines the four cases that are possible using the non-parametric estimation of h_0 that the Cox proportional hazards model uses. The first is the ideal case where the original dataset has provided a model that is suitable to use in its original unrecalibrated form. There is no need to re-estimate neither the baseline hazard nor a scalar on the $x\beta$ to provide reasonable calibration. Case 2 is a situation where the covariate coefficient relationship is preserved from the original model but the baseline hazard of an event is believed to be substantially different in the validation dataset and must be re-estimated. Case 3 assumes no need to re-estimate the baseline hazard but does assume there is a mis-estimation of the covariate risk relationship and places a scale adjustment on the $x\beta$. Finally, Case 4 is a recalibration technique that allows for both variation in the baseline hazard and the relationship between the covariate set and the proportional hazard of the event. Here, h_0 is re-estimated and a scalar factor is placed on the $x\beta$. Case 2 is frequently used for recalibration to populations with different baseline risks like older populations, otherwise frail populations like those with related comorbidities, or, as we will discuss below, populations in different countries. There are not many practical uses for Case 3 recalibration; it is rarely used if at all. Case 4 is often used for introduction of new biomarkers into a risk prediction model. Yeboah et al[Yeboah et al., 2012] and Polonsky et al[Polonsky et al., 2010] use the technique in Case 4 to recalibrate the Framingham risk score to the MESA cohort as a point of comparison to assess the added value of novel biomarkers such as CAC, CIMT and ankle-brachial index (ABI) to the traditional Framingham risk score. We will explore

Case 2 and Case 4 recalibration. We will also make a brief statement on a subset of Case 2 recalibration where, rather than estimating a baseline hazard using data, researchers will use the cumulative incidence of the event, a common yet statistically and conceptually poor strategy.

2.5.2 Techniques for Composite Endpoints

There are two options for creating and recalibrating risk scores for composite endpoints. The first of these is the standard approach that is taken by the majority of researchers, that being that the composite endpoint is constructed as in section 2.2 and creating an initial risk score in the development population using either a Cox proportional hazards model or a competing risks model with a known competing risk. This composite risk score can be recalibrated using Case 2 or Case 4 recalibration methods depending on the availability of cohort data in the population one is recalibrating to. In practice, Case 2 recalibration often involves a fairly crude estimate of baseline hazard based on Kaplan-Meier estimates, cumulative incidence, or raw proportions, we will discuss this more in chapter 3.

The alternative that we propose uses a competing risks modeling approach to estimate cause specific hazards for each component of the composite endpoint for the original risk score. This gives a risk estimate for each endpoint of interest within the composite endpoint. After this, we combine the event-specific risks into a risk score for a composite endpoint using the following formula:

$$p(event_1 \cup event_2 \cup \cdots \cup event_K = 1 | T = t) = 1 - S(t) \tag{2.4}$$

where overall event-free survival is

$$S(t) = e^{-H(t)} \tag{2.5}$$

and overall cumulative hazard H(t) is equal to the sum of cause-specific cumulative hazards $H^{(k)}(t)$:

$$H(t) = \sum_{k=1}^{K} H^{(k)}(t). \tag{2.6}$$

This formulation provides results similar to the risk estimates given by a single risk model based on the cumulative hazard for a composite endpoint but is composed of individual hazard estimates. This affords us the ability to re-estimate cause-specific cumulative baseline subdistribution hazards in both Case 2 and Case 4 recalibrations and cause specific predictive indices $X_k\beta_k$ which are combined into a recalibrated risk estimate using formula 2.4.

2.5.3 Examples

Epidemiological and clinical cardiovascular research is full of instances in which recalibration is a useful tool. As described above, one may be interested in extending the risk score to ethnic groups not included in the development population[D'Agostino et al., 2001], applying the risk score to a different country with different baseline risks[Brindle et al., 2003, Barzi et al., 2007], or to frail populations with concomitant risk factors that elevate their risk for the outcome of interest[Koller et al., 2007]. In studies assessing the predictive accuracy of the Framingham cohort in predicting coronary heart disease (CHD), a cohort of largely European descended older people in western Massachusetts from a period of particularly high cardiovascular disease risk, the unrecalibrated score typically over-predicts cardiovascular disease risk in other populations. The 2001 Framingham paper by D'Agostino and the CHD Risk Prediction Group of the Framingham study compares outcomes to predicted Framingham risk in the ARIC Study, the Physicians Health Study (PHS), the Honolulu Heart Program (HHP), the Puerto Rico Heart Health Program (PR), the Strong Heart Study (SHS) and the Cardiovascular Health Study (CHS) all of which include substantial membership of ethnic groups not included in the Framingham study. In Japanese American men in HHP, Hispanic men in PR, and Native American men and women in SHS, unadjusted predicted risk greatly exceeded actual event proportions. Similarly in the Brindle paper of Framingham in British men, and the Bari paper of the Asia Pacific Cohort Studies meta analysis of over 44 cohort studies in the Asia Pacific region showed systematic over-prediction of risk when compared to populations that did not match the Framingham population.

Despite the poor calibration, there is benefit to using an established score over inventing

a new one for each subpopulation of interest. There is considerable work and resource expenditure in developing a risk score and difficulty in communicating so many risk scores to non-statistically inclined practitioners. As such, each of these papers used recalibration techniques to adjust the Framingham risk score. With the exception of the Brindle paper, which divided risk by the over-prediction factor, the researchers in each paper adjusted the risk score by replacing the baseline survival from the Framingham cohort with that of the cohort being analyzed. This can be done using Cox regression or Kaplan Meier techniques. In each case, the recalibration dramatically improved goodness of fit. The Asia Pacific Cohort Studies paper, for example, reduced the chi-squared test statistic for the Hosmer-Lemeshow goodness-of-fit test in cohort members in China from 557.5 ($p < 0.001$) for men and 608 ($p < 0.001$) for women to 16.7 (p=0.032) and 20.5 (p=0.009), respectively.

2.5.4 Issues with Recalibrating using Composite Endpoints

The statistical motivation for our approach to estimating and recalibrating composite events is the concern that a baseline hazard and predictive index for a composite endpoint, which are weighted averages of baseline hazards and predictive indices for the component events in a composite endpoint, are dependent on the case mix in the development population. As such, re-estimating an overall cumulative baseline hazard while restricting the adjustment to a scalar adjustment on the original predictive index would not be able to adequately account for the different proportions of events that would lead to different weightings of the predictive indices and cumulative baseline subdistribution hazards. Our hypothesis was that a recalibrated composite risk score on a population with a case mix of component events that is very different from that in the development population will be poorly calibrated due to this inability to properly weight component events contributions to the overall risk equation.

Chapter 3

SIMULATION STUDIES

3.1 Simulating Competing Risks Survival Data

To assess the impact of our estimation and recalibration method on populations with differing case mixes at baseline we simulated competing risks survival for populations based largely on the MESA populations, their covariate profiles and the relationship of those covariates with the risk of hard coronary heart disease events, stroke, and non-CVD death. Initially we used the method described in Bender et al. [2005] wherein we simulated event times separately for both the component events CHD and stroke as well as the competing event non-CVD death based on a manipulation of the cumulative distribution of the Cox model for each event. The event times are compared for each simulated individual and, as these are mutually exclusive events, the minimum event time for an individual is considered to be their event time and the event to which that time corresponds is their event. In their 2005 paper, Jan Beyersmann and his colleagues found the method in the Bender article (known as the latent failure time method) to be the method of simulation for 60 percent of the articles published in *Statistics in Medicine* since the year 2000 that used simulation to study competing risks data [Beyersmann et al., 2009]. The authors take issue with the prevalence of this method as they have strong reservations about its plausibility in applied settings, and the non-identifiability of the dependence between the latent failure times in the observed data. The solution that Beyersmann and colleagues propose has dependence between events that is built into the simulation method. Their approach begins with taking cause-specific hazards α_{0k}, where $k \in 1, 2, \ldots, K$ represents one of the competing events, for each of the competing events and summing to obtain an all-cause hazard. This all-cause hazard is used to simulate survival times T. At each simulated survival time T, simulate a

single multinomial trial where the probability of an event k is $\alpha_{0k}(T)/\sum_{k=1}^{K}\alpha_{0k}(T)$. Rather than having an unidentifiable dependence between competing events, this makes it such that each event is a value of a random variable in a poisson process with exponential waiting time and multinomial probability for event type. Interestingly, the way that this article suggests to simulate censoring times is to use the latent failure times approach and take the first of the event and censoring times. Since our models assume independence between event times and censoring times, we feel comfortable proceeding with this method.

3.2 The 'survsim' R Package

Researchers David Moriña of the Center for Research in Environmental Epidemiology (CREAL) and Albert Navarro of Universitat Autónoma de Barelona have written **survsim** [Moriña and Navarro, 2014], an R package that simulates many sorts of survival data. Of interest to our project is the function crisk.sim, a function that simulates survival data for competing events using the methods from the Beyersmann article. This function takes as inputs the desired number of observations, maximum follow-up time, and censoring and event distributions from an accelerated failure time model to generate a matrix of simulated competing-risks survival data with observation id, event type, and event time. Technical details can be found in the R-Cran repository. This package was instrumental in simulating data with survival event values and covariate distributions that mimicked our development dataset; however, there was a limitation in that the simulations did not allow for administrative censoring that reflected the conditions in the MESA study. In order to recreate those conditions, we found it necessary to simulate events beyond the actual maximum follow-up time and cut time off at the point we determined by using the maximum follow-up time from MESA.

3.3 Simulation Environment

As alluded to in the previous section, the design of our simulations was based upon the observed distributions in the MESA cohort, in order to mimic a realistic population. We simulated cohorts of comparable size that would have a similar censoring distribution, similar

covariate profiles and corresponding relationships to cardiovascular disease risk.

Table 3.1: Summary Statistics for the MESA Population

MESA Properties (N=6,647)	
	N(%) or Mean (Standard Deviation)
Male Gender	3145(47%)
Chinese	798(12%)
Black	1816(27%)
Hispanic	1470(22%)
Diabetes	837(13%)
Current Smoking	862(13%)
Stenosis	2755(41%)
Medications:	
Lipid Lowering	1076(16%)
Anti-Hypertensive	2462(37%)
Family History:	
Heart Attack	2656(40%)
Stroke	2157(32%)
Age (years)	62.16(10.25)
Systolic BP (mmHg)	126.60(21.51)
Cholesterol (mg/dL)	194.2(35.7)
HDL Cholesterol (mg/dL)	50.98(14.85)
log(CAC) (ln(HU))	2.19(2.52)

The feature of this population we wanted to modify was baseline risk of the competing disease events, or in terms of the Fine and Gray model, the baseline cumulative subdistribution

hazard. In our simulations we based our model parameters on the fitted values of accelerated failure time survival models for each of the outcomes in our composite ASCVD event in MESA. The accelerated failure time for an observation 'i' and event 'k' is parameterized as $log(T_{i,k}) = \beta_0 + \beta' x + \sigma\epsilon_{i,k}$ where β is a vector of regression coefficients, x is a matrix of covariates and $\sigma\epsilon_{i,k}$ is the scaled residual error for observation i and event k. The accelerated failure time coefficients for each event are summarized in Table 3.2. We note that we use accelerated failure time coefficients to fit the specifications of the crisk.sim function, but that it is different from the Fine and Gray model we ultimately use to predict risk.

Table 3.2: Accelerated Failure Time Covariate-Risk Relationships in Simulations

Regression Coefficients			
	CHD	Stroke	Non-CVD Death
Age (years)	-0.025	-0.042	-0.092
Male Gender	-0.451	0.138	-0.269
Chinese	0.272	0.525	0.307
Black	-0.024	0.009	-0.173
Hispanic	0.111	-0.257	0.067
Systolic BP (mmHg)	-0.008	-0.018	-0.001
Log(CAC) (ln(HU))	-0.264	-0.106	-0.057
Diabetes	-0.46	-0.421	-0.292
Smoker	-0.516	-0.618	-0.841
Cholesterol (mg/dL)	-0.002	-0.004	0.003
HDL-C (mg/dL)	0.009	0.011	0.0002
Lipid RX	-0.083	0.265	0.246
Hypertension RX	-0.193	-0.315	-0.131
FH Heart Attack	-0.313	-0.086	0.092

Our intercepts were all 5.79 to get a 10-year baseline cumulative subdistribution hazard

(BCSH) of approximately 3% for each outcome at 0 values of all binary variables and mean values of all continuous variables (see Table 3.1). These intercepts were then modified in such a way to investigate the effect that a differing case mix would have on recalibration when the relationship with risk factors remained unchanged. We scaled the baseline sub-hazards for CHD and stroke by a factor ranging from 0.1 to 3 and held the other outcome, and death, constant at its original baseline cumulative subdistribution hazard. The notation we have adopted for the relative scaling factors has been summarized in Table 3.3. Each element in the notation represents a different event. The first element, (**X**,1,1) represents CHD, the second element (1,**Y**,1) represents stroke and the third element (1,1,**Z**) represents non-CVD death. In the first example, (**X**,1,1), we have inflated the BCSH by a factor of **X** and left the baseline hazards for stroke and non-CVD death unchanged. Similarly in the (1,**Y**,1) case mix we have multiplied the baseline hazard for stroke by **Y**. Though stroke and CHD vary between 0.1 and 3, the inflation factor for the baseline hazard for death is always 1.

Table 3.3: Case Mix Convention

Notation for Case Mixes							
	Inflation Factor						
Event	0.1	0.5	1	1.5	2	2.5	3
CHD	(0.1,1,1)	(0.5,1,1)	(1,1,1)	(1.5,1,1)	(2,1,1)	(2.5,1,1)	(3,1,1)
Stroke	(1,0.1,1)	(1,0.5,1)	(1,1,1)	(1,1.5,1)	(1,2,1)	(1,2.5,1)	(1,3,1)

For each of the 13 unique scenarios listed in Table 3.3, we ran 500 simulations with a sample size of 5,000. Each observation is followed across time until they experience one of the three competing events, are lost to follow up (6% 10 year probability of loss to follow up), or until 12.2 years have elapsed. The population has been simulated to have the covariate distributions listed in table 3.1 and accelerated failure times for the three events related to those covariates as listed in table 3.2. Within a simulation, we estimated risk of CHD, stroke, and the composite ASCVD outcome without recalibrating, recalibrating only on

the baseline subdistribution hazard (Case 2 recalibration), recalibrating substituting the cumulative incidence at time t for the baseline hazard (typical Case 2 calibration in practice), and recalibrating on the subdistribution hazard and fitting a linear scalar on our $X\beta$ from the original risk equation. Ten-year CHD and stroke risk equations were combined as described in section 2.4.2 to create a composite risk score based on cause-specific hazards. Each risk score was evaluated for goodness-of-fit using the Hosmer-Lemeshow calibration test which follows a χ^2_{g-2} distribution where g is the number of risk groups used to evaluate the probability of events. Also collected was the area under the receiver operating curve, a measure of discrimination, and calibration in-the-large which is the absolute difference between the mean predicted risk and the proportion of observations that experience an event. These summary measures, as well some implications of recalibration based on the ACC/AHA risk management guidelines for an example simulation are presented in the next section.

3.4 Simulation Results

3.4.1 Unrecalibrated Risk Scores

Figures 3.1-3.3 give a general impression of the effect of varying case mixes on an unrecalibrated risk score, presenting the 25th percentile, median, and 75th percentile of the Hosmer-Lemeshow test statistic, area under the ROC curve and calibration in-the-large respectively. As one might expect, an unrecalibrated risk score performs very poorly in settings where the case mix is not identical to that of the development dataset. As the inflation factor varies on CHD (Figure 3.1, top panel), the only case mix where the minimum Hosmer-Lemeshow test statistic does not show a significantly poor fit to the data is the (1,1,1) case mix, this is true for both the composite risk score and the risk score that combines cause specific hazards. For reference, our Hosmer-Lemeshow tests follow a χ^2_8 and a significant test statistic at the $\alpha = 0.05$ level is approximately 15.5. Our tests fare a little better as we vary the inflation factor on the baseline subdistribution hazard for stroke, though only under a case mix identical to that in the development dataset does the either model predict risk

with acceptable accuracy (Figure 3.1, bottom panel). AUCs give moderate results overall (Figure 3.2) and calibration in the large echoes the results given by the Hosmer-Lemeshow test statistics (Figure 3.3) as we have mentioned previously, discrimination is insensitive to systematic issues in calibration so the negative calibration occurring simultaneously with positive discrimination is not terribly surprising. Additionally, we see an improvement in discrimination as the inflation factor increases. This is likely due to the increase in events for the simulations with inflated baseline cumulative sub-hazards. The tendency when varying the baseline risk for either of the main component outcomes is for the combined approach to outperform the composite approach in cases where the case mix has a diminished baseline risk whereas the composite approach performs better in cases where the case mix gives an elevated baseline risk. This is trivial; however, given how poor the fit is in general and we would not recommend using an unrecalibrated risk score unless there is strong evidence that the baseline hazard is consistent for all outcomes of interest.

Figure 3.1: Hosmer-Lemeshow Statistics for Unrecalibrated Risk Scores

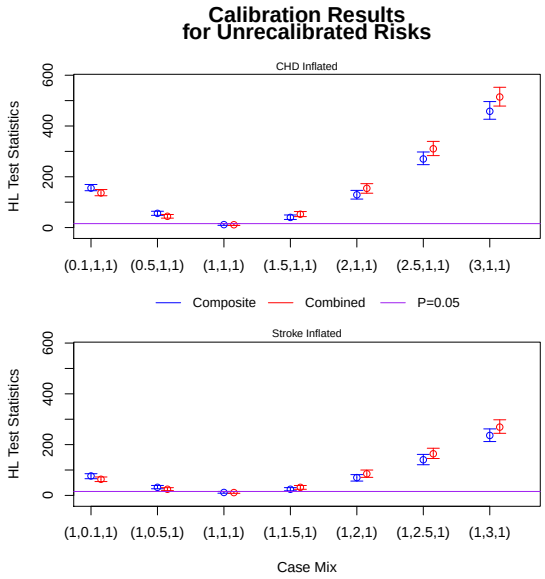


Figure 3.2: Area under ROC Curve for Unrecalibrated Risk Scores

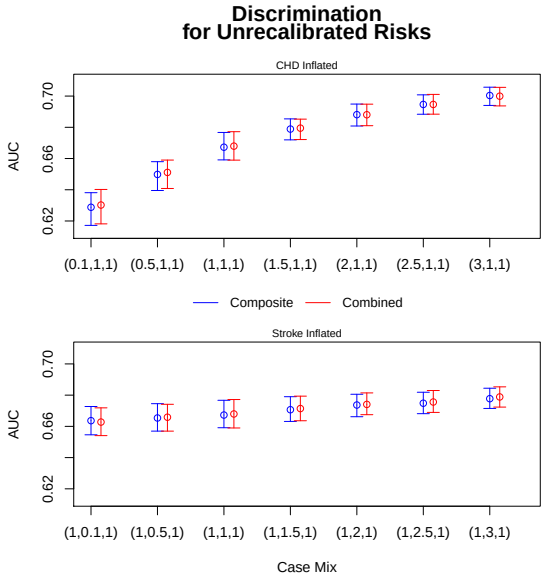
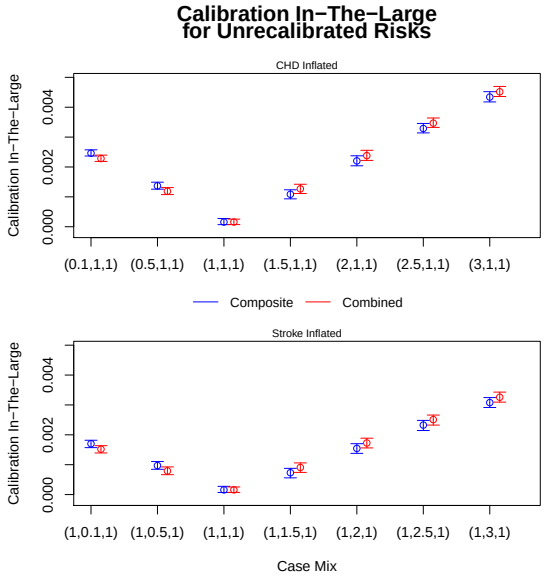


Figure 3.3: Calibration In-the-Large for Unrecalibrated Risk Scores



3.4.2 Case 2 Recalibrated Risk Scores

Case 2, where only the baseline subdistribution hazard is re-estimated begins to fare better. Varying the baseline hazards for both CHD (Figure 3.4, top panel) and stroke (bottom panel), we see reasonably good performance of both composite and combined recalibration methods in nearly all case mixes. The combined risk score shows mild improvement in calibration over the composite risk score as the inflation factor for the baseline subdistribution hazard for stroke increases; however, this improvement is not large and the result is not consistent enough for us to declare a preference in risk estimation approaches. Our results for AUC, shown in Table 3.5 are similar to those in the unrecalibrated risk-scores. This simply shows the lack of impact on ordering of predicted risk that recalibration has, as well as the limits of discrimination measures in evaluating risk-scores when the scientifically meaningful measure is the absolute level of risk rather than the relative level of risk compared to other observations. Figure 3.6 shows the calibration in-the-large, which largely reflects the tendencies in the calibration statistics.

Figure 3.4: Hosmer-Lemeshow Statistics for Case 2 Recalibrated Risk Scores

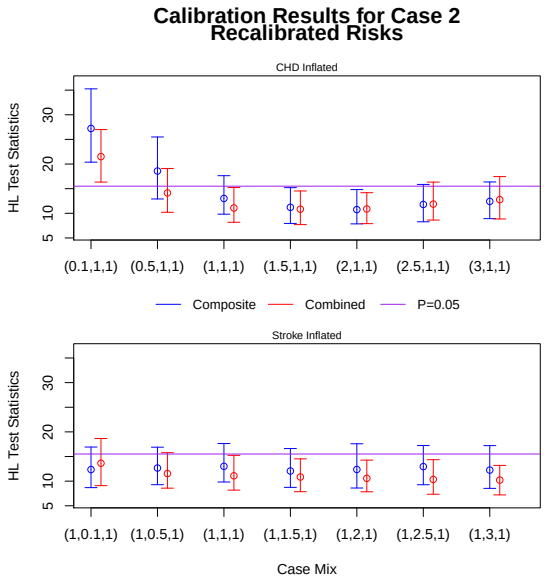


Figure 3.5: Area under ROC Curve for Case 2 Recalibrated Risk Scores

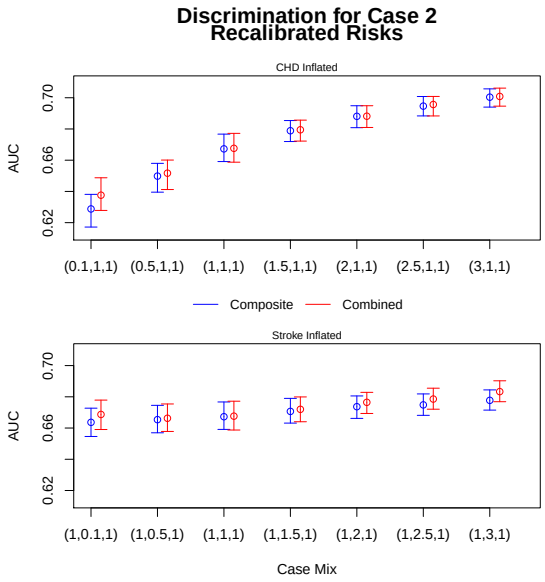
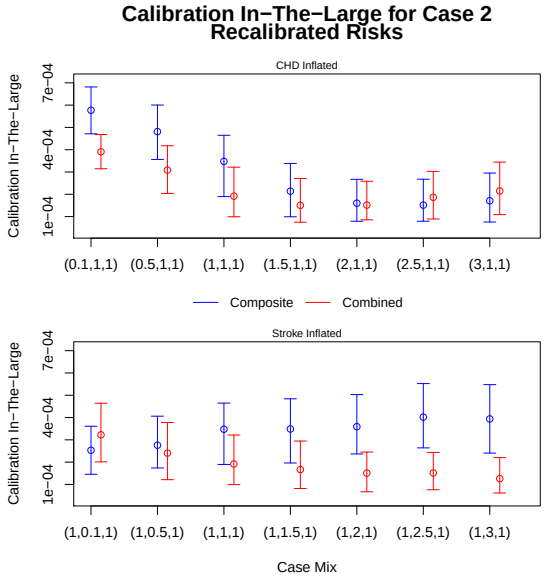


Figure 3.6: Calibration In-the-Large for Case 2 Risk Scores



3.4.3 Case 4 Risk Scores

Case 4 recalibration of our risk score displays some interesting behavior. Overall, calibration is improved substantially over both the unrecalibrated risk scores and the risk scores that only re-estimate the baseline subdistribution hazards. This recalibration technique, wherein the baseline subdistribution hazard is re-estimated and a scalar adjustment is placed on the $X\beta$ matrix has overall good and comparable calibration as the inflation factor for CHD baseline hazard varies. In the series where we vary the baseline hazard for CHD only, at least three quarters of the simulations show good calibration regardless of technique, neither technique nears the threshold for significantly poor fit as displayed in the top panel of Figure 3.7. However, as we vary the inflation factor for stroke risk (Figure 3.7, bottom panel), the combined recalibration approach begins to lose precision. Though over half of the simulations have good model fit to the data, at about an inflation factor of two, we see the 3rd quartile of recalibrated risk scores reaching significantly poor fit to the simulated populations whose risks they are estimating. While results for the combined recalibration technique suffer in these settings with highly inflated stroke hazard, calibration for the composite approach is consistent across case mixes.

Figure 3.7: Hosmer-Lemeshow Statistics for Case 4 Recalibrated Risk Scores

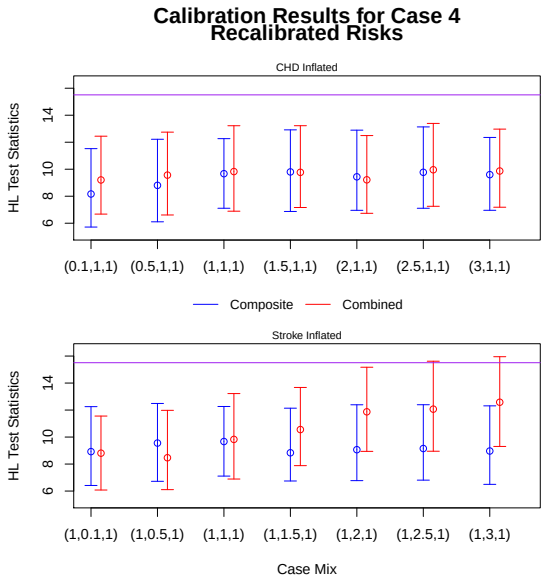


Figure 3.8: Area under ROC Curve for Case 4 Recalibrated Risk Scores

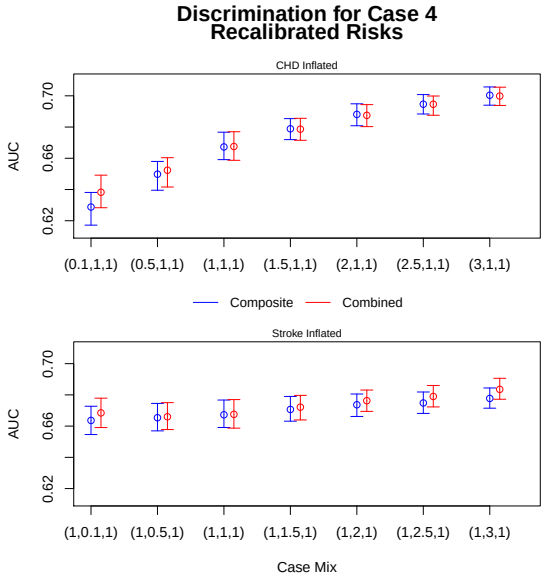
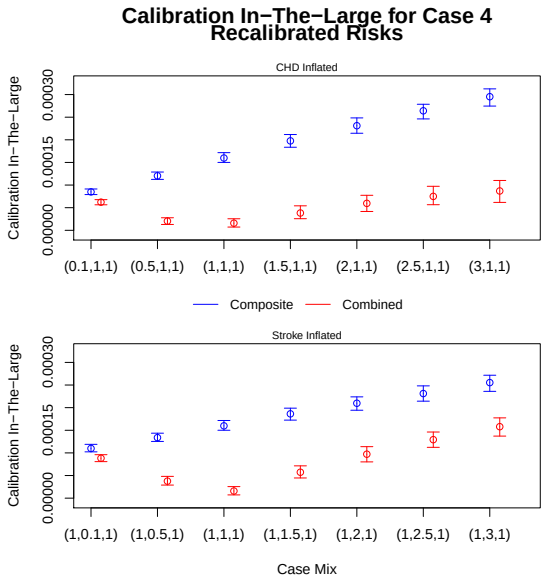


Figure 3.9: Calibration In-the-Large for Case 4 Risk Scores



3.4.4 Example Simulation

For illustrative purposes, we present two individual simulations and compare our estimated risk categories to the recommendations for treatment of blood cholesterol given by the American College of Cardiology/American Heart Association. The ACC/AHA recommendations are based on a risk categorization where an estimated 10-year ASCVD risk greater-than or equal-to 7.5% constitutes a high-risk and below that threshold is considered a low risk. We examine how the recalibration of risk affects the proportion of subjects in the high and low risk groups who go on to have an ASCVD event over the course of 10 years. We are estimating how many true positives are captured as well as how many false negatives we have let occur by our estimation methods.

Table 3.4 shows the proportions of subjects in the high-risk group who have an ASCVD within 10 years. For both the (3,1,1) case mix and the (1,3,1) case mix, the unrecalibrated risk score has the highest positive predictive value. In the (3,1,1) scenario, 22% of subjects

classified as high-risk using either the composite or combined risk scoring method were true-positives. Likewise, in the (1,3,1) case mix scenario, 27% of high-risk subjects using either risk scoring method had events. Case 2 and Case 4 recalibrations have similar levels of ASCVD events within 10 years, all near 20%. In other words, if you put the people in this population on statins under the guidelines placed by ACC/AHA, you would be treating a group of people among whom 20% would have gone on to have an ASCVD event by the end of a decade. To address any alarm that these figures would give, the consequences of not treating high-risk individuals are much worse than treating low-risk individuals and the treatment guideline thresholds reflect that. The mean risk for case mix (1,3,1) is much closer to the proportion of predicted events than the guidelines would indicate. Mean case 2 composite risk is 0.18, combined is 0.19, in fact only the mean unrecalibrated risk is very far from the proportion of events in this category with a mean predicted risk of 0.11.

Table 3.4: High-Risk Events

$Events for Risk \geq 0.075$			
Case Mix: (3,1,1)			
	Unrecalibrated	Case 2	Case 4
Composite	0.267	0.217	0.218
Combined	0.267	0.217	0.215
Case Mix: (1,3,1)			
	Unrecalibrated	Case 2	Case 4
Composite	0.223	0.189	0.189
Combined	0.223	0.189	0.188

Table 3.5 shows the proportion of observations on whom no treatment is recommended who would go on to have an ASCVD event within 10 years. Here, we are concerned with the ability of our risk scores to detect high-risk individuals. Capture for these cases is similar all around. Between 4 and 7% of the subjects classified as low risk will go on to have an

ASCVD event within the next ten years. Composite and combined risk scores have similar errors within 1% of each other and between recalibration cases there is not a substantial difference either, with the greatest gains being made between not recalibrating at all and using either of our recommended recalibration techniques.

Table 3.5: Low Risk Events

$Events for Risk < 0.075$			
Case Mix: (3,1,1)			
	Unrecalibrated	Case 2	Case 4
Composite	0.091	0.046	0.044
Combined	0.094	0.036	0.039
Case Mix: (1,3,1)			
	Unrecalibrated	Case 2	Case 4
Composite	0.089	0.072	0.069
Combined	0.093	0.070	0.065

3.5 Additional Comments

During the course of this research we came across two matters of interest that had not formed a central part of our initial hypothesis, but that merit some commentary. One is a matter that has great potential as a future research direction, and the other is a word of warning to practitioners about an approximate method of recalibration and risk scoring.

3.5.1 Comment 1: Differing Predictors

While developing the clinical risk score, penalized regression models were selecting quite different sets of covariates to include in our predictive models. This would present a problem for a composite endpoint; however, it would be simple to build two event-specific risk scores

and combine the event-specific risks for an overall risk. To test the predictive ability of such a set of risk scores, we ran simulations similar to those we ran for the main investigation of this thesis, except we set family history of heart attack to have no relation to stroke. We ran simulations comparing the composite risk score to the combined risk score with 500 simulations of a population of 5,000 each. We recalibrated in the three ways discussed in the previous section: no recalibration, Case 2 recalibration, and Case 4 recalibration. Calibration tests for Case 2 and Case 4 recalibration will be discussed and other summary plots can be found in the appendix in Figures A.1-A.2. Figure 3.10 compares calibration between the composite risk score and a combined risk score that excludes family history of heart attack in the stroke risk equation. When baseline hazard for CHD is attenuated, calibration is not very good, though the combined model tends to perform better, perhaps due to the increased prevalence in stroke events. As CHD baseline hazard increases, this advantage of the combined model appears to disappear. Similarly, when baseline hazard for stroke is reduced, the combined method does not perform as well as the composite method; however, as stroke events increase in proportion, the combined method improves in calibration while the composite method worsens. Case 4 recalibration, as shown in Figure 3.11, does not show the same trends, across inflation factors for CHD, performance is good overall and differences in calibration appear close to null. When we increase stroke events, Case 4 recalibrated risk scores perform substantially worse for the combined risk scores, it does not appear to make a difference for the composite model.

This appears to show some potential for modeling a composite of two substantially different outcomes; however, the advantage appears to disappear when one has sufficient data to re-predict the population's baseline hazard and linearly adjust the $X\beta$ predictive index as is done in Case 4 recalibration.

Figure 3.10: Hosmer-Lemeshow Statistics for Case 2 Recalibrated Risk Scores Where Models Differ

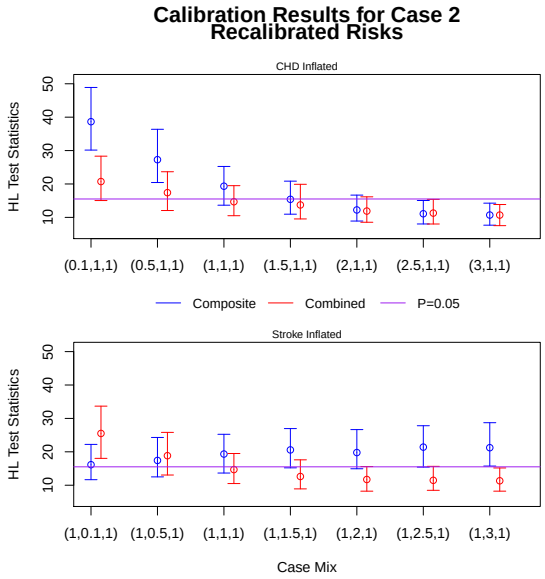
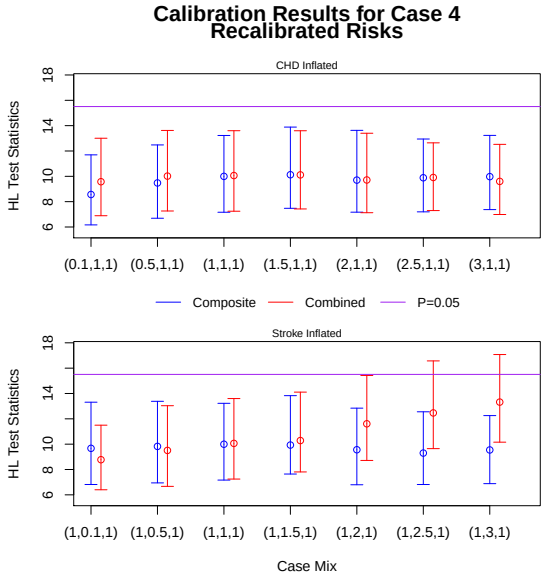


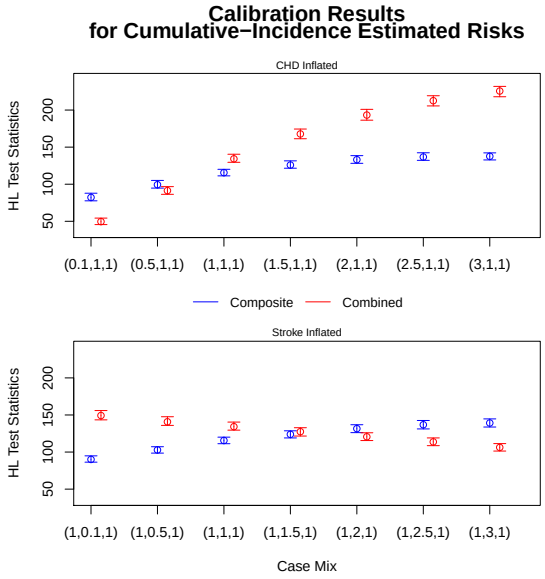
Figure 3.11: Hosmer-Lemeshow Statistics for Case 4 Recalibrated Risk Scores Where Models Differ



3.5.2 Comment 2: Substituting Cumulative Incidence for Baseline Hazard

As mentioned in Chapter 2, there is a tendency in clinical practice when recalibrating to substitute the cumulative incidence of an event in a population for the baseline hazard. This is conceptually wrong as this is the average risk of an event in a total population, not the risk in a population for observations with the average covariate values. This means that the cumulative incidence is likely inflated due to the outsized risk from observations with high-risk profiles. This is demonstrated in Figure 3.12 where no method performs remotely well under any case mix. In the (1,1,1) case mix where all baseline hazards and covariate-risk relationships are perfectly preserved from the development population, both the composite and combined model perform so badly that the Hosmer-Lemeshow test statistics are in the hundreds range when a significantly poor model threshold is around 15. This is substantially worse than the unrecalibrated risk score results and the results are comparably bad for more highly altered case mixes. These results are included to raise awareness about using this approximation in practice.

Figure 3.12: Hosmer-Lemeshow Statistics for Models with Cumulative Incidence Substituted for Baseline Hazard



3.6 Conclusions

Our simulation studies did not present us with any conclusive evidence about the superiority of one recalibration technique over the other. There was some cause for concern with the combined risk scoring approach in case 4 when we alter the baseline subdistribution hazard for stroke; however, recalibrated risk scores performed reasonably well across the board. Lending strength to this argument, our inspection of the effect of recalibration on treatment guidelines in two individual simulations with extreme case mix alteration showed that the greatest benefit to capturing high-risk individuals was in recalibrating generally as opposed to not recalibrating, though that increase in sensitivity does result in a loss in specificity as seen by the proportions of events in the high-risk groups.

Chapter 4

APPLICATION: COMPARISON OF COMPOSITE ENDPOINT MODELING APPROACHES IN THE MESA COHORT

Guided by the conclusions from our simulations, we are in the process of creating a new risk score using data from the MESA study that will incorporate the event-specific risk scores as well as the currently used composite modeling approach. In addition to the information we gain through the use of competing-risks modeling, we are incorporating a biomarker, Coronary Artery Calcium (CAC), a marker of the amount of atherosclerotic disease in a person, that has been found to improve the predictive power of cardiovascular risk scores. We elaborate on the measurement and predictive contribution of CAC in the following section.

4.1 Coronary Artery Calcium

Coronary Artery Calcium (CAC) is a marker of atherosclerotic disease level measured using the Agatston Scoring Method[McClelland et al., 2009, Bild et al., 2002] which uses tomography imaging of the coronary artery to get a calcium volume. The Agatston scoring method multiplies a density factor for the calcified section of the coronary by the area of the section, then the score of each section is summed across the coronary artery to give a total score. The density factor is split into four levels of highest attenuation value in Hounsfield Units (HU):

1. 130-199 HU
2. 200-299 HU
3. 300-399 HU
4. 400+ HU

Two scans are taken for each subject and independently measured. Both total score estimates are averaged to give a final CAC score. Agatston-based CAC scores can be roughly translated to severity of coronary artery disease as follows:

- No evidence of coronary artery disease: 0 CAC score
- Minimal CAD: 1-10
- Mild CAD: 11-100
- Moderate CAD: 101-400
- Severe CAD: > 400

In MESA, CAC ranges from 0 Agatston units to above 4600 in the highest scoring group and, typical of a population free of CVD at the time, a high proportion (50%) of subjects have CAC scores of 0 Agatston units. Analysis of the distribution of CAC shows that men tend to have more CAC, whites tend to have more CAC than otherwise similar nonwhite subjects, and CAC tends to increase with age.

In recent years, this marker and variations on it have gained traction in cardiovascular risk prediction. Recent studies show CAC having a higher impact than other biological measures. The ARIC study shows higher predictive value than Carotid intima-media thickness (CIMT) and the Womens Health Study and Physicians Health Study show improvement over using high sensitivity C-reactive Protein (hsCRP) in predicting ASCVD. CAC improves risk scores in these studies by reclassifying intermediate risk patients more precisely into high and low risk categories where treatment recommendations are more definite [Janic et al., 2013, Kerut, 2011, deGoma et al., 2013, Pletcher et al., 2004]. In the following sections I will summarize our risk score and the effects of adding CAC to the risk models in the MESA cohort.

4.2 Main Effects Models for CHD, Stroke, and Composite CVD

The following risk equations are presented for the members of the MESA cohort for whom data is available on all covariates of interest and outcomes. Ultimately, 6,647 of a total of 6,814 cohort members were included in the risk models. The tables in the following subsection are the regression coefficients that result from a Fine and Gray competing risks

model for survival data. The outcome in the model is the one listed at the top of the table and the remaining outcome and non-CVD death are included as competing risks. For example, the model for CHD includes Stroke and non-CVD death as competing risks and only observations that left the risk set due to loss to follow-up or administrative censoring were considered censored observations. This preliminary model considers only main-effects. In a true risk score, interactions will be accounted for, and the method that MESA uses to select interactions to include in the risk model is the Lasso penalized regression technique. The main effects that are included for CHD and stroke differ because we have don't believe all the predictors for stroke have a strong predictive contribution to estimating CHD risk. CHD has age, systolic blood pressure, cholesterol, and high-density lipoprotein cholesterol as well as indicators for the presence of: male gender, Chinese, Black, and Hispanic races, diabetes, current smoking, lipid lowering medication use, hypertension medication use, and family history of heart attack. The stroke and composite CVD outcomes include all the above covariates and family history of stroke and presence of carotid artery stenosis. All continuous covariates were centered on their mean. To get a composite risk score in the case that CAC is unavailable; practitioners would plug the values of the covariates of their patient into equation 2.3 as follows for CHD:

1. Multiply covariates and coefficients to get the predictive index $X\beta$=-0.047*(age - 62.16) - 0.740*male gender - 0.327*Chinese Race - 0.240*Black Race - 0.268*Hispanic Race - 0.009*(Systolic Blood Pressure - 126.6)) + 0.566*Diabetes + 0.003*(Cholesterol - 194.2) - 0.012*(HDL-Cholesterol - 50.98)) + 0.175*Lipid Lowering Medications + 0.228*Hypertension + 0.435*Family History of Heart Attack
2. Take the cumulative baseline sub-hazard for CHD $H_0(10) = 0.027$ estimated at the mean values of the continuous variables given in Table 4.1, or in the case of the risk score with CAC, Table 4.2.
3. Plug the two resulting values $X\beta$ and $H_0(10)$ into the formula for ten year risk: Risk(10)=$1 - e^{-e^{X\beta} * H_0(10)}$

The same formulation can be applied to the stroke and composite outcomes, and to

generate a combined risk-score one must simply add the inverse of the two event-specific $H_0(10)e^{x\beta}$ for stroke and CHD and plug that resulting value into the formula. In this chapter we are applying the strategy we believe is indicated by our research from Chapter 3, that we are modeling both composite and component outcomes to obtain risk scores using the composite and combined approaches. This allows us to include models that include different covariates for each endpoint and will eventually provide information about the driver of the composite risk. Table 4.1 shows the hazard ratios and baseline hazards for the preliminary risk-score without including CAC, and Table 4.2 shows the coefficients and baseline hazards to use if a CAC measurement is available.

Table 4.1: Main-Effects Risk Equation Without CAC

	HR(e^{β})			Mean(%)
	CHD	Stroke	Composite	MESA
Age (years)	1.049	1.041	1.045	62.16
Male Gender	2.096	1.026	1.716	47%
Chinese	0.721	0.801	0.762	12%
Black	0.787	0.939	0.822	27%
Hispanic	0.765	1.366	0.937	22%
Systolic BP (mmHg)	1.009	1.016	1.012	126.6
Diabetes	1.762	1.498	1.725	13%
Smoker	1.681	1.580	1.635	13%
Cholesterol (mg/dL)	1.003	1.003	1.003	194.2
HDL-C (mg/dL)	0.988	0.992	0.989	50.98
Lipid RX	1.191	0.893	1.085	16%
Hypertension RX	1.256	1.332	1.282	37%
FH Heart Attack	1.546	1.042	1.370	40%
FH Stroke	–	1.252	1.161	32%
Stenosis	–	1.326	1.319	41%
$H_0(10)$[2]	0.027	0.012	0.036	–

[2]Baseline hazard is estimated at 0 values of all categorical variables. Continuous variables are centered at the mean values in MESA shown in the right-hand column

Table 4.2: Main-Effects Risk Equation With CAC

	HR(e^{β})			Mean(%)
	CHD	Stroke	Composite	MESA
Age (years)	1.021	1.035	1.026	62.16
Male Gender	1.527	0.949	1.349	47%
Chinese	0.850	0.815	0.825	12%
Black	1.009	0.993	0.995	27%
Hispanic	0.899	1.418	1.055	22%
Systolic BP (mmHg)	1.007	1.016	1.011	126.6
Diabetes	1.569	1.467	1.602	13%
Smoker	1.475	1.537	1.510	13%
Cholesterol (mg/dL)	1.002	1.002	1.002	194.2
HDL-C (mg/dL)	0.990	0.992	0.990	50.98
Lipid RX	1.058	0.873	1.009	16%
Hypertension RX	1.137	1.296	1.187	37%
FH Heart Attack	1.357	1.008	1.238	40%
FH Stroke	–	1.260	1.182	32%
Stenosis	–	1.252	1.093	41%
Log(CAC) (ln(HU))	1.279	1.065	1.217	2.19
$H_0(10)$[3]	0.028	0.013	0.041	–

[3]Baseline hazard is estimated at 0 values of all categorical variables. Continuous variables are centered at the mean values in MESA shown in the right-hand column

4.3 Assessment of Model Fit and Predictive Power

To illustrate the predictive impact of adding CAC to a predictive model of cardiovascular disease, I have summarized calibration and the area under the receiver operating characteristic curve, which measures discrimination, in Table 4.3. For all outcomes save for stroke, adding CAC to the model improves model fit and discrimination. The combined and composite approaches appear to be about the same and both have fairly poor model fit. Since the Hosmer-Lemeshow test is sensitive to non-linearities and interactions, the poor fit in the main-effects model is not immediately troubling. We have some additional reservations about the Hosmer-Lemeshow test but will go over that more fully in the limitations section. Figures 4.1 - 4.6 show grouped calibration and ROC plots for component events as well as the composite and combined cardiovascular disease outcome. The Hosmer-Lemeshow plots show smoothed series of proportions of events plotted against risk estimates. The risk binnings that occur in the estimation of the Hosmer-Lemeshow test-statistic are seen between the tick-marks along the x-axis. The ROC plots show sensitivity plotted against 1-specificity as described in section 2.4. These plots illustrate the marginal benefit that adding CAC to the predictive model has in terms of calibration and discrimination. Obviously, the benefit is greatest when predicting CHD, as previously described literature would suggest, and the attenuation of that benefit makes sense since we are not seeing as strong a link between CAC and stroke. In summation, we have a statistical method that allows us to see cause-specific risks, choose covariates separately for each component outcome, and have a biomarker that improves CHD prediction. Choosing interactions and whether to include CAC in the stroke model requires further reflection; regardless, we have a strong foundation for our risk-score.

Figure 4.1: Calibration Plots for Composite CVD in MESA

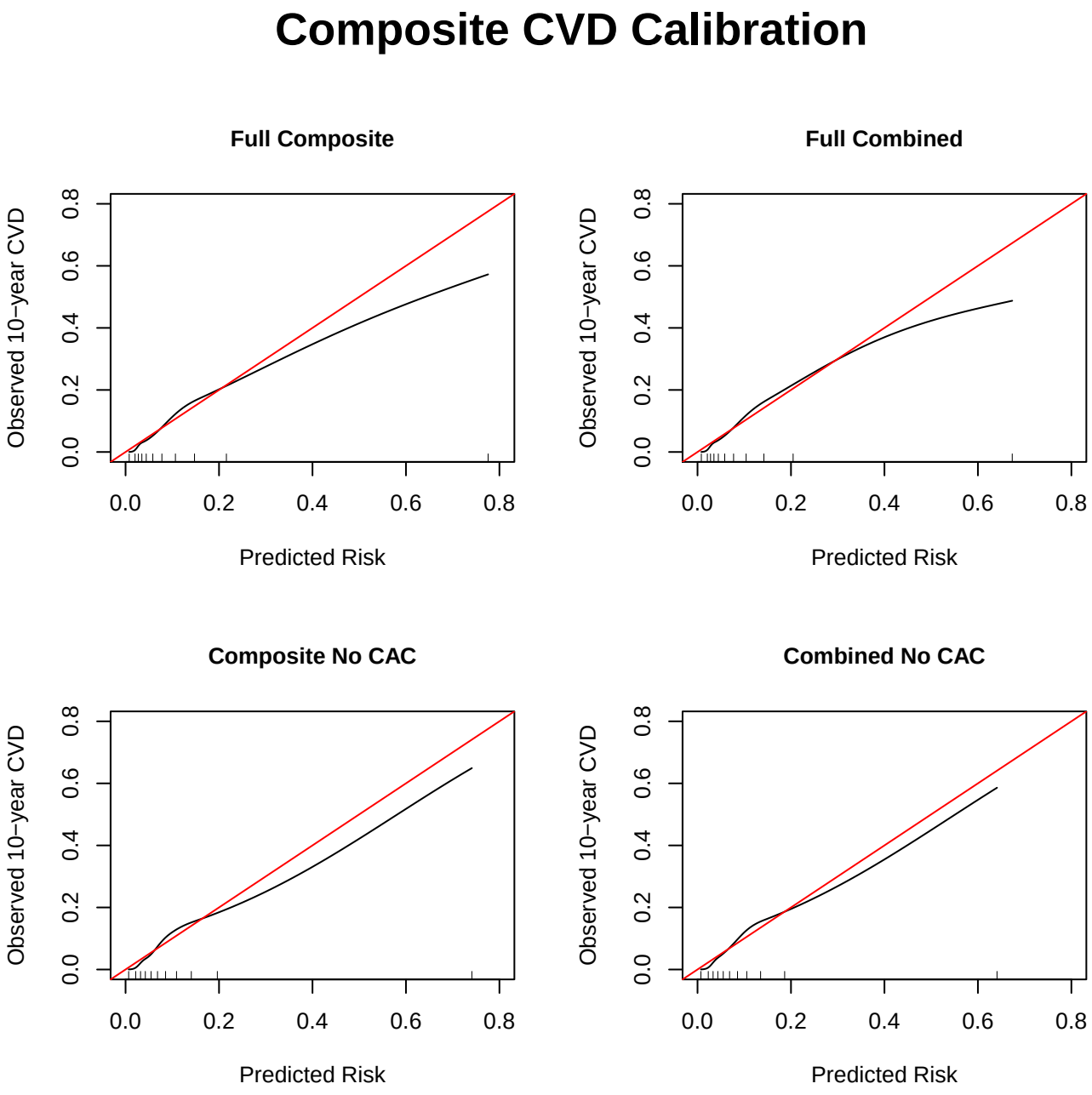


Figure 4.2: Calibration Plots for Stroke and CHD in MESA

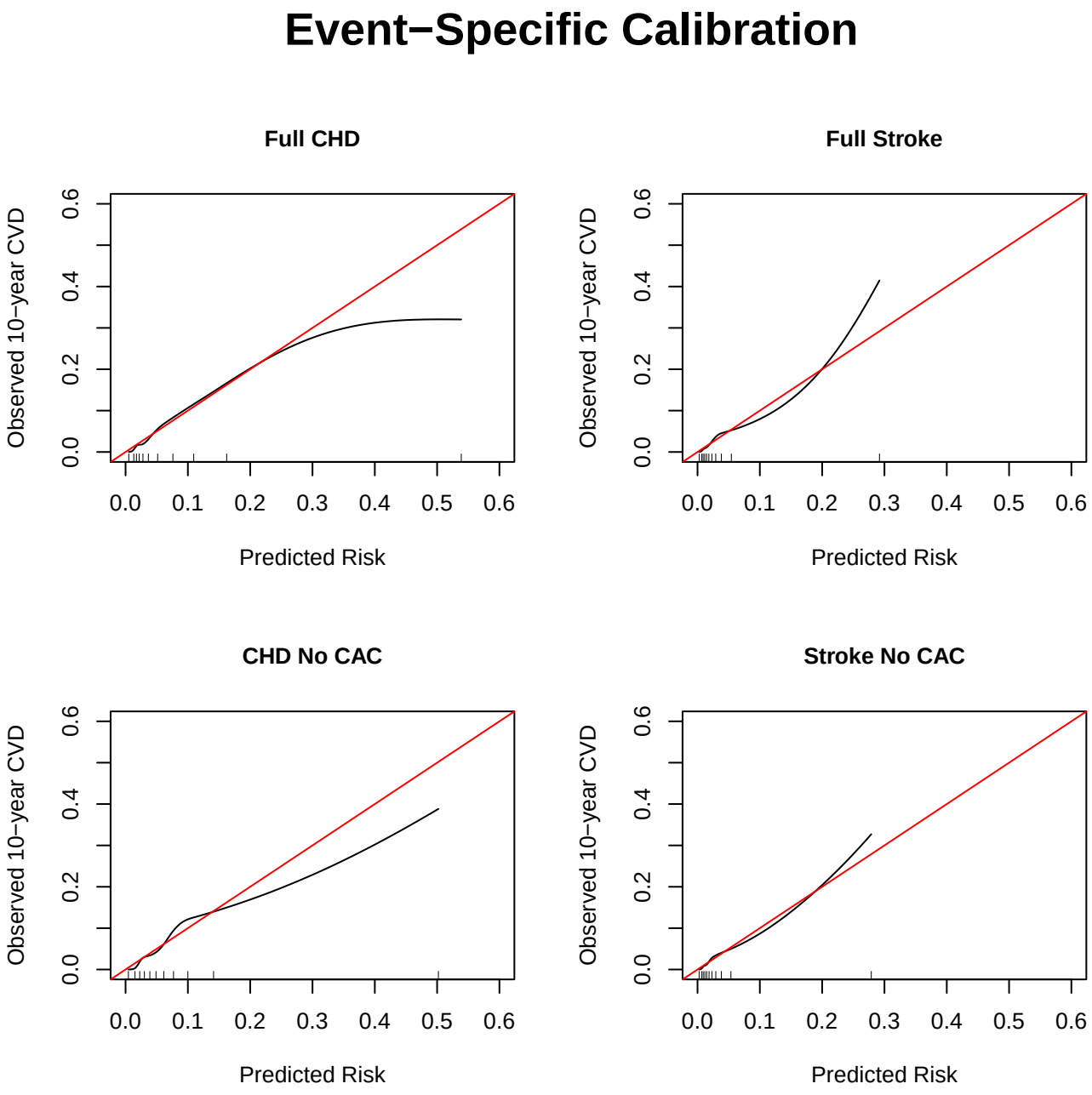


Figure 4.3: ROC Curves for CHD in MESA

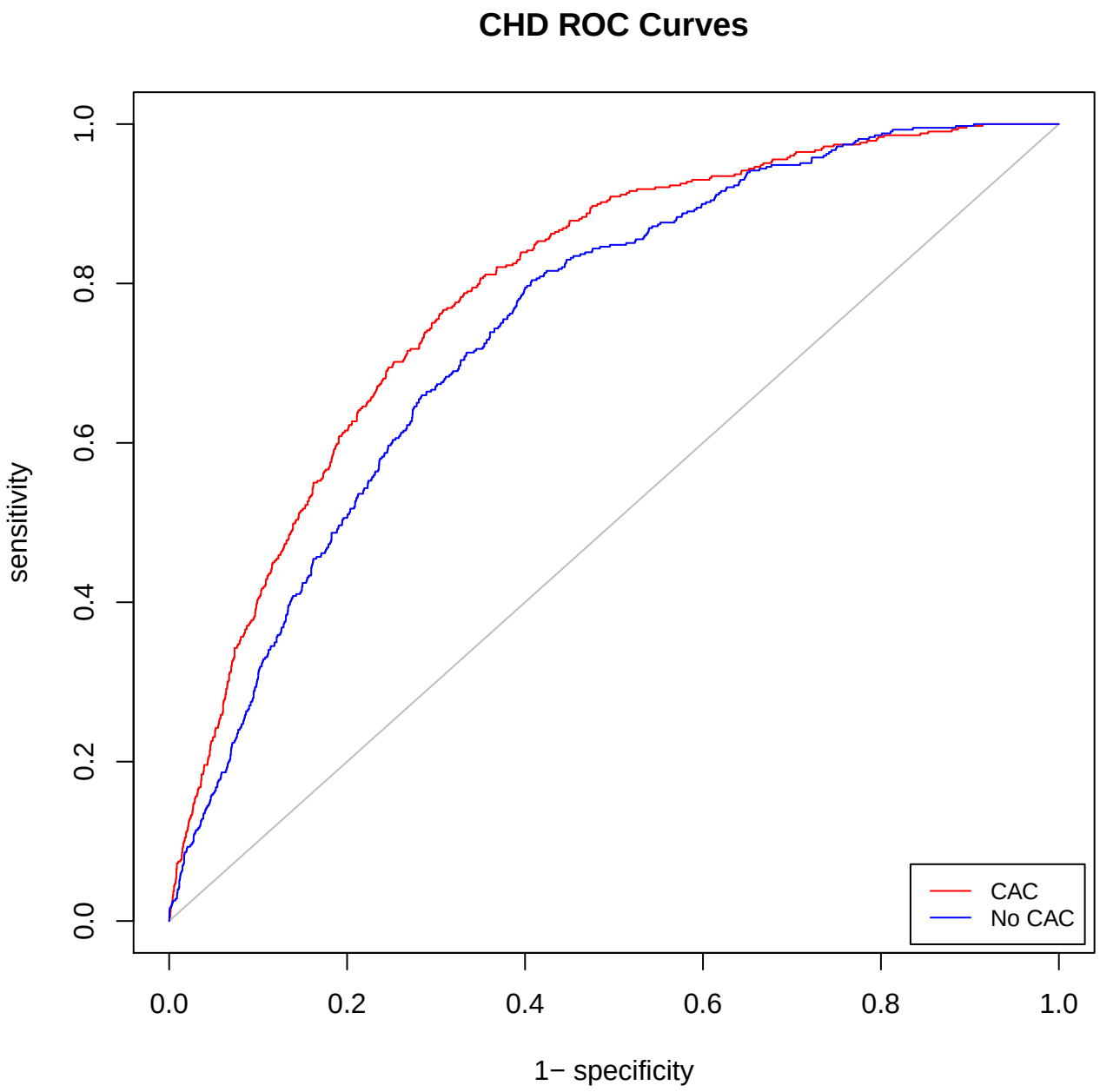


Figure 4.4: ROC Curves for Stroke in MESA

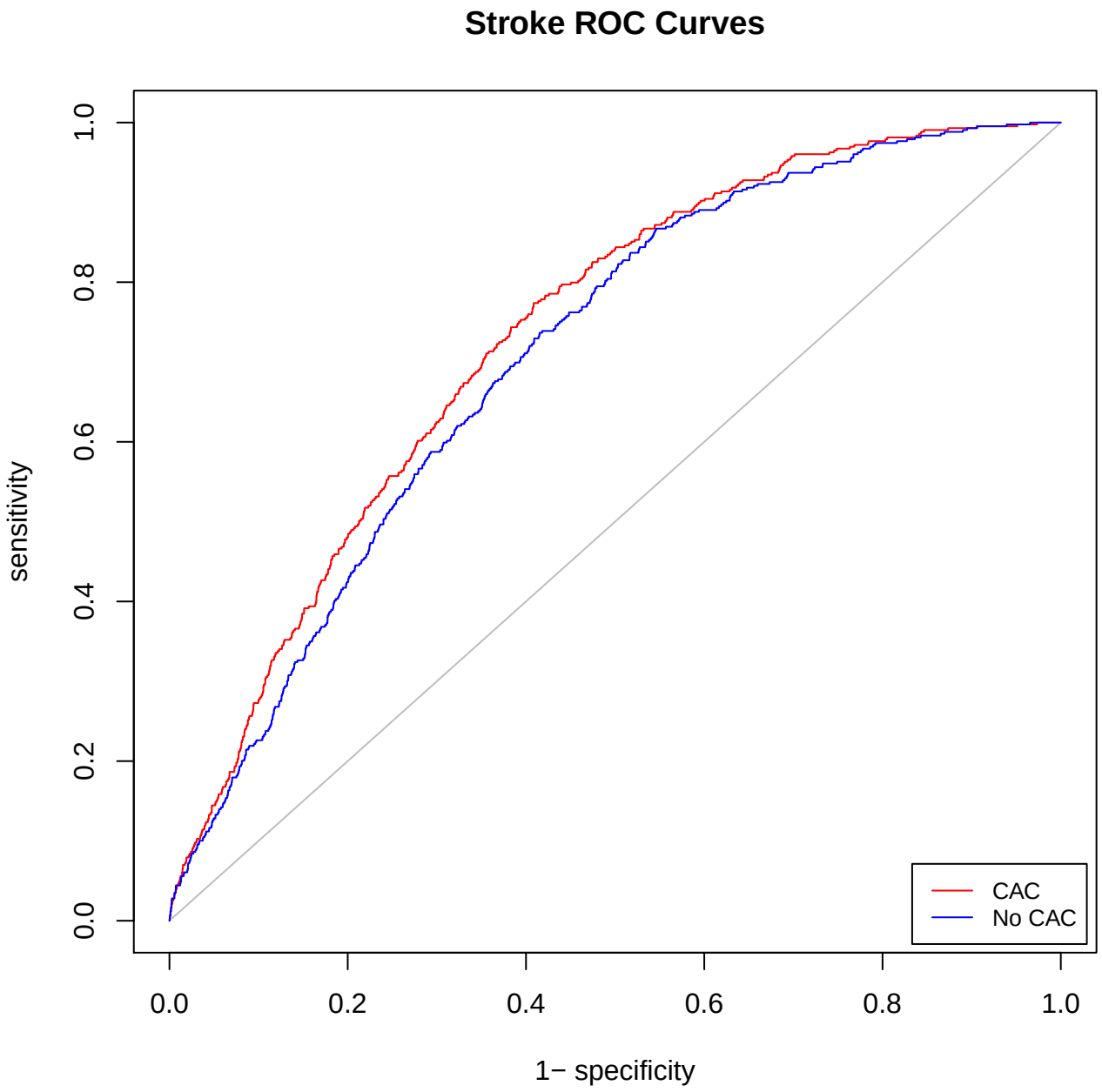


Figure 4.5: ROC Curves for for Composite CVD in MESA, Composite Approach

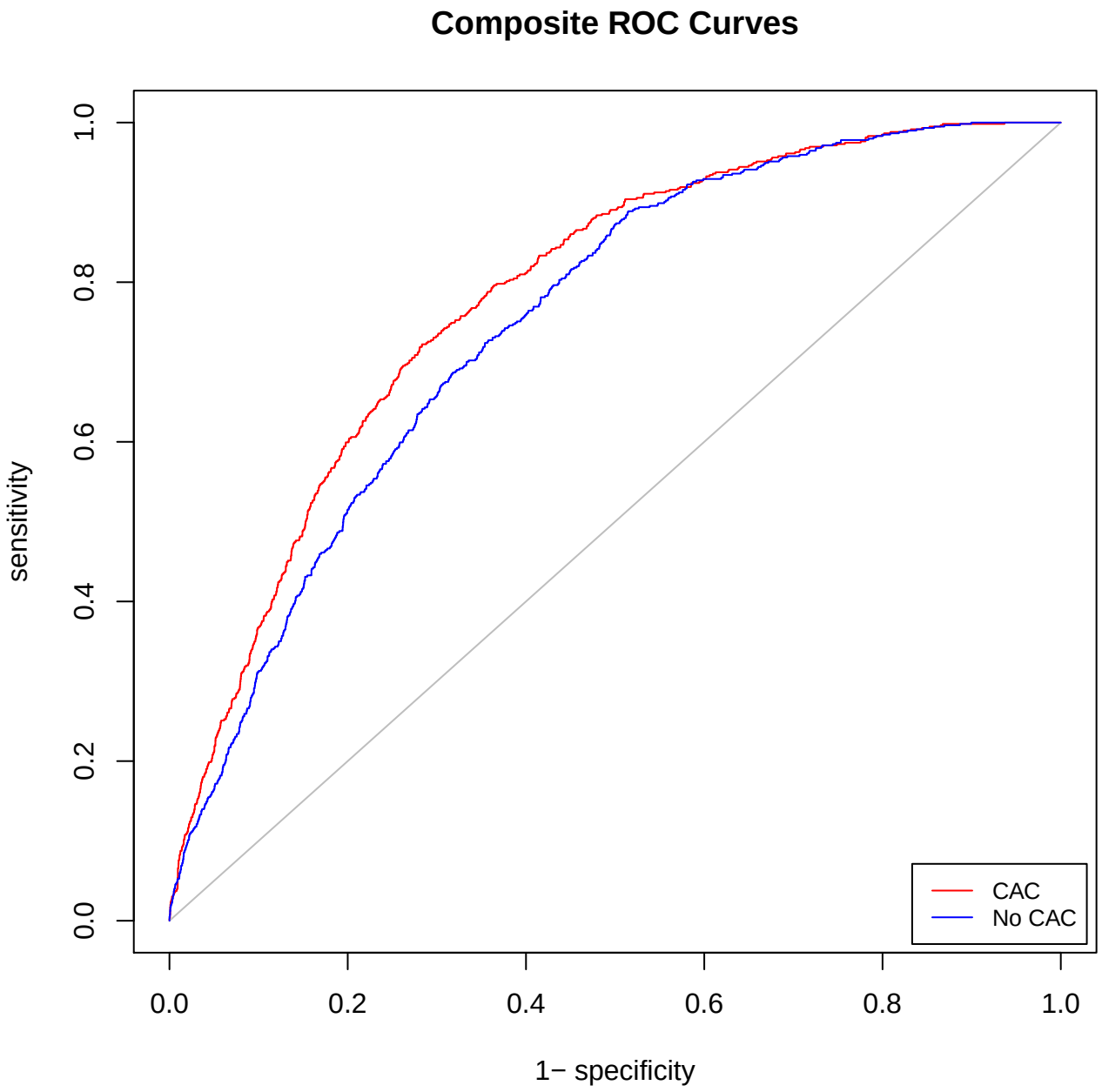


Figure 4.6: ROC Curves for for Composite CVD in MESA, Combined Approach

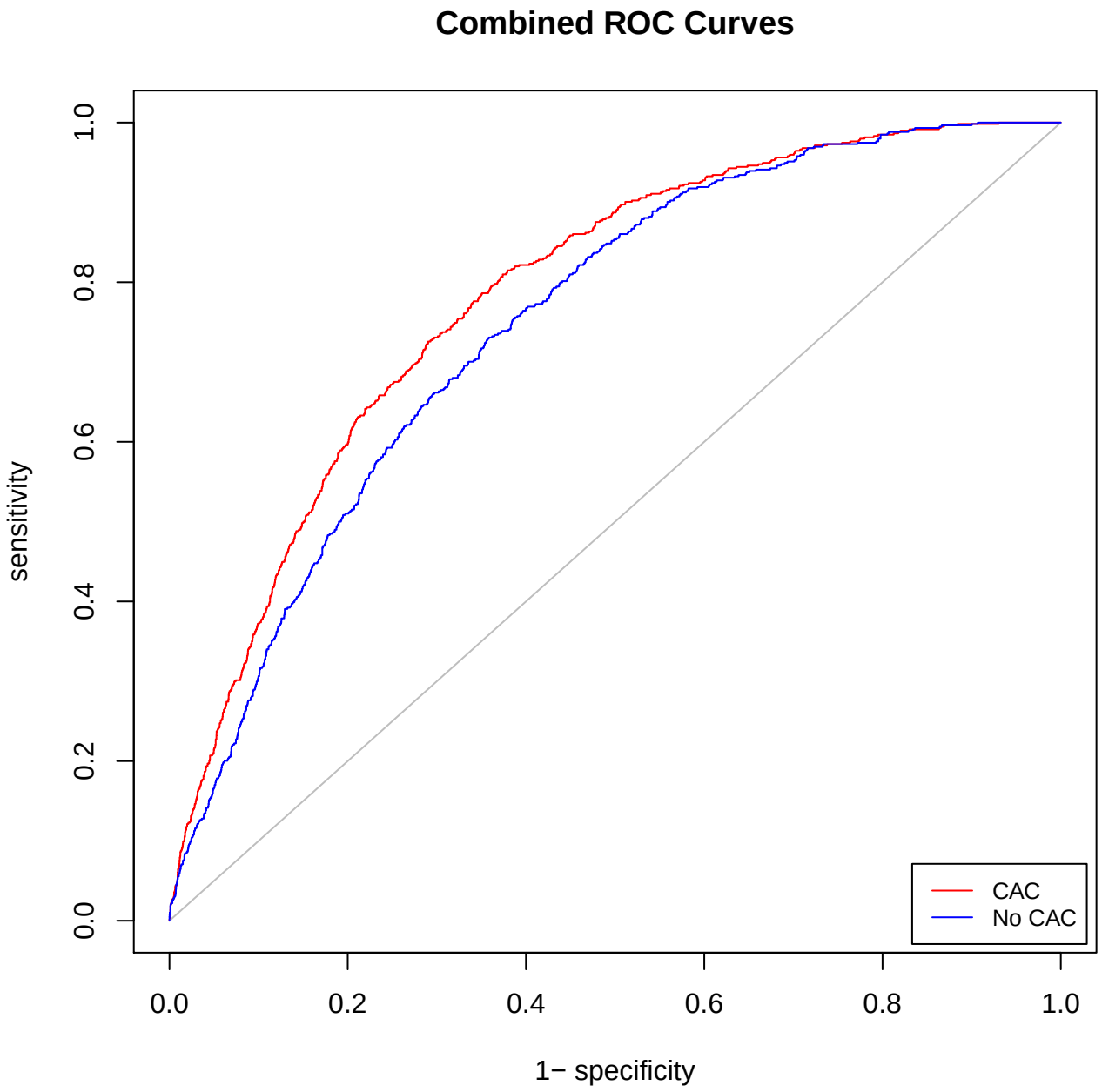


Table 4.3: Measures of Model Fit

	HL Statistic: χ^2_8	P-value	AUC	At Risk Threshold 0.075	
				False Positive Rate (1-Specificity)	True Positive Rate (Sensitivity)
Composite*	23.197	0.003	0.782	0.376	0.798
Composite	27.603	0.001	0.750	0.427	0.791
Combined*	22.602	0.004	0.784	0.371	0.803
Combined	25.876	0.001	0.749	0.426	0.783
CHD*	8.820	0.358	0.793	0.275	0.718
CHD	23.687	0.003	0.749	0.289	0.660
Stroke*	12.134	0.145	0.733	0.040	0.121
Stroke	8.649	0.373	0.709	0.038	0.103
*Model Including CAC					

Chapter 5
CONCLUSIONS

While we were not able to demonstrate uniform superiority for either method for modeling composite endpoints across recalibration techniques and case mixes, we have been able to thoroughly present the different options and the advantages and limitations of both. A strategy that we encourage is to model both composite risk and combined risk as the composite risk is the conventional method and tends to work well, especially when there is data to perform case 4 recalibration. However, for clinical reasons, the component risks should be estimated to gain an understanding of whether it is one event that is driving the composite risk or if it is spread evenly across events. In Case 2 recalibration the combined method of risk scoring seems a reasonable substitute for the composite approach; however, there does seem to be a benefit in using the composite method to get the overall risk when using Case 4 recalibration.

These recommendations are based on preliminary results and are subject to some limitations regarding testing and implementation. These limitations are acknowledged in the following section and further research is recommended in the section after that.

5.1 Limitations

5.1.1 Hosmer-Lemeshow Goodness-of-Fit

The measure of model fit on which we assessed our risk scoring methods was the Hosmer-Lemeshow goodness-of-fit test. This test bins the data by quantiles, usually deciles, and as such is highly dependent on the number of groups used in setting up the test. The pitfall of this is illustrated in Figures 4.1-4.2 where the main deviance between predicted risk and observed risk occurs in the highest decile. This top decile contains a large range of predicted

risks, and to take the mean of those risks to get an expected number of events is unrealistic, especially in the presence of a skewed distribution of risks. The Hosmer-Lemeshow test has been critiqued, and in their textbook *Applied Logistic Regression*, Hosmer, Lemeshow, and Sturdivant suggest several atlernative tests including the Osius Rojek test and the Stukel test. While these tests do not depend on grouping, they do require approximate variances that are based on a single covariate matrix linearly regressed on a single outcome. This would not be an appropriate technique, as it would not consider the stratified, cause-specific variances we will encounter using the combined approach, nor will it allow for specification of more than one covariate matrix. For the time being, we have used the Hosmer-Lemeshow test because of its wide use and our ability to use it without specifying a single model, an impossible task in the combined risk score; however, we currently use it for lack of a test that is better suited to our risk modeling methods.

5.1.2 Simulation Methods

While we are extremely grateful for the development of the **survsim** R package, there were several ways in which our simulations did not match the conditions we had hoped to simulate. First, our data is administratively censored, as the MESA was 12 years old when we created the simulations and is still on going. This means that our hazards were based on data that was right censored due to loss to follow-up and administrative censoring. This is not possible to simulate using the package and we had to simulate events out to a very large time-span and cut them off at 12 years in order to have reasonable event proportions and times that were not concentrated at the beginning of the twelve year period. In addition, the risk factors we included in our model are unlikely to be independent but we were unable to model the correlations between risk factors, meaning that, for instance, cholesterol and HDL cholesterol could be completely unrelated in the simulation, as could smoking status and blood pressure. Further research should incorporate these considerations into the simulation method.

5.2 Further Research

This project brought up many opportunities for further research in the fields of risk scoring and cardiovascular epidemiology. We detail a few below.

5.2.1 Methods Research

Our research was limited, as previously mentioned by the measures by which we could assess goodness-of-fit. The test we used was limited by the grouping aspect and our alternatives were unfeasible for our combined risk approach, making a direct comparison between the two approaches impossible with these tests. Research into a model-free ungrouped test for goodness-of-fit would be valuable for the general body of statistical knowledge as well as for use in clinical risk scoring.

Use of different models for risk factors that are typically included in a composite endpoint could be a very valuable tool in cardiovascular epidemiology as well as other areas of research where estimation of composite risks occur. In particular, how does a composite modeling approach on a risk score built with modern variable selection techniques compare to a risk score built on a combined modeling approach with the same variable selection procedures? The combined approach seemed to have some advantages when we did not include a variable whose coefficient was estimated to be zero in a way that did not occur when we included the variable in both models but the coefficients had opposite signs (ie, the two hazard ratios were on opposite sides of one). Research into the advantages and disadvantages of independently modeling these decomposed composite endpoints for the purposes of estimating their composite risk would be quite valuable.

5.2.2 Applied Research

Throughout our project, stroke has provided us with difficulty in modeling and in calibration. As we varied stroke, more problems seemed to appear and our models for stroke typically had less accuracy. Our predictor of risk, CAC, which was so useful for predicting CHD, events

did not help in predicting stroke, in fact in our main effects model it appeared to worsen estimates. We speculate that this is due to missing interaction terms, but this remains to be further investigated. To gain a better understanding of atherosclerotic disease, we should investigate etiologic aspects of stroke, how to predict and prevent it, in ways that are independent of its relation to coronary heart disease.

5.3 Concluding Remarks

Much of the utility of the research in this thesis has been to demonstrate current limitations in cardiovascular risk scoring and recalibration. We have shed light on some of the limitations in the cumulative incidence recalibration technique, stroke prediction, and modeling composite events. Future research should capitalize on the idea of tailoring models to component endpoints prior to composite risk estimation. Our risk scores and simulations demonstrate application of the competing-risks framework to our risk scoring method that should always be considered for modeling cardiovascular disease and any other disease that has death as a competing risk. We hope that our contribution will help lead to better methods that will have an appreciable impact on scientific inquiry and public health.

BIBLIOGRAPHY

F. Barzi, A. Patel, D. Gu, P. Sritara, T.H. Lam, Rodgers A., and M. Woodward. Cardiovascular risk prediction tools for populations in asia. *Journal of Epidemiology & Community Health*, 61:115–121, 2007.

R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics In Medicine*, 24:1713–1723, 2005.

J. Beyersmann, A. Latouche, A. Buchholz, and M. Schumacher. Simulating competing risks data in survival analysis. *Statistics In Medicine*, 28:956–971, 2009.

D.E. Bild, D.A. Bluemke, G.L. Burke, R. Detrano, A.V. Diez Roux, A.R. Folsom, P. Greenland, D.R. Jacobs Jr., R. Kronmal, K. Liu, J.C. Nelson, D. O'Leary, M.F. Saad, S. Shea, M. Szklo, and R.P Tracy. Multi-ethnic study of atherosclerosis: Objectives and design. *American Journal of Epidemiology*, 156(9), 2002.

E. Björnson, J. Borén, and A. Mardinoglu. Personalized cardiovascular disease prediction and treatment—a review of existing strategies and novel systems medicine tools. *Frontiers in Physiology*, 7, 2016.

P. Brindle, J. Emberson, F. Lampe, M. Walker, P. Whincup, T. Fahey, and S. Ebrahim. Predictive accuracy of the framingham coronary risk score in british men: Prospective cohort study. *The BMJ*, 327, 2003.

R.B. D'Agostino, S. Grundy, L.M. Sullivan, and P.W.F. Wilson. Validation of the framingham coronary heart disease prediction scores. *The Journal of the American Medical Association*, 286(2), 2001.

E.M. deGoma, R.L. Dunbar, D. Jacoby, and B. French. Differences in absolute risk of cardiovascular events using risk-refinement tests: A systematic analysis of four cardiovascular risk equations. *Atherosclerosis*, (172-177), 2013.

J.P. Fine and R.J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

D. Goff, D.M. Lloyd-Jones, G. Bennett, S. Coady, R.B. D'Agostino, R. Gibbons, C.J. O'Donnel, J.G. Robinson, J.S. Schwartz, S.T. Shero, S.C. Smith Jr, P. Sorlie, N.J. Stone, and P.W.F. Wilson. 2013 ACC/AHA guideline on the assessment of cardiovascular risk a report of the American College of Cardiology/American Heart Association task force on practice guidelines. *Circulation*, 129[suppl 2]:S49–S73, 2014.

M. Janic, M. Lunder, and M. Sabovic. A new anti-ageing strategy focused on prevention of arterial ageing in the middle-aged population. *Medical Hypotheses*, 80:837–840, 2013.

E.K. Kerut. Coronary risk assessment and arterial age calculation using coronary artery calcium scoring and the framingham risk score. *Echocardiography*, 28:686–693, 2011.

M.T. Koller, E.W. Steyerberg, M. Wolbers, T. Stijnen, H.C. Bucher, M.G.M. Hunink, and J.C.M. Witteman. Validity of the framingham point scores in the elderly: Results from the rotterdam study. *American Heart Journal*, 154(1):87–93, 2007.

R.L. McClelland, K. Nasir, M. Budoff, R. Blumentha, and R. Kronmal. Arterial age as a function of coronary atery calcium (from the Multi-Ethnic Study of Atherosclerosis [MESA]). *American Journal of Cardiology*, 2009.

R.L. McClelland, N.W. Jorgensen, M. Budoff, M.J. Blaha, S.W. Post, R. Kronmal, D.E. Bild, S. Shea, K. Liu, K.E. Watson, A.R. Folsom, A. Khera, C. Ayers, A.A. Mahabadi, N. Lehmann, K.H. Jöckel, S. Moebus, J.J. Carr, R. Erbel, and G.L. Burke. 10-year coronary heart disease risk prediction using coronary artery calcium and traditional risk factors. *Journal of the American College of Cardiology*, 66(15):1643–53, 2015.

D. Moriña and A. Navarro. The r package survsim for the simulation of simple and complex survival data. *Journal of Statistical Software*, 59(2), 2014.

M.J. Pletcher, J.A. Tice, M. Pignone, and W.S. Browner. Using the coronary artery calcium score to predict coronary heart disease events. *Archives of Internal Medicine*, 164:1285–1292, 2004.

T.S. Polonsky, R.L. McClelland, N.W. Jorgensen, D.E. Bild, G.L. Burke, A.D. Guerci, and P. Greenland. Coronary artery calcium score and risk classification for coronary heart disease prediction: The multi-ethnic study of atherosclerosis. *Journal of the American Medical Association*, 303(16), 2010.

P. Royston and M.K.B. Parmar. External validation and updating of a prognostic survival model. Research Report 307, Department of Statistical Science, University College London, 2010.

S.P. Whelton, K. Nasir, M.J. Blaha, and ...R.S. Blumenthal. Coronary artery calcium and primary prevention risk assessment: What is the evidence? an updated meta-analysis on patient and physician behavior. *Circulation: Cardiovascular Quality and Outcomes*, 5: 601–607, 2012.

P.W.F Wilson, R.B. D'Agostino, D. Levy, A. Belanger, H. Silbershatz, and W.B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97:1837–1847, 1998.

J. Yeboah, R.L. McClelland, T.S. Polonsky, G.L. Burke, C.T. Sibley, D. O'Leary, J.J. Carr, D. Goff, P. Greenland, and D.M. Herrington. Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate risk individuals. the multi-ethnic study of atherosclerosis. *Journal of the American Medical Association*, 308(8):788–795, 2012.

Appendix A

SUPPLEMENTARY FIGURES

Figure A.1: Area under ROC Curve for Case 2 Recalibrated Risk Scores Where Models Differ

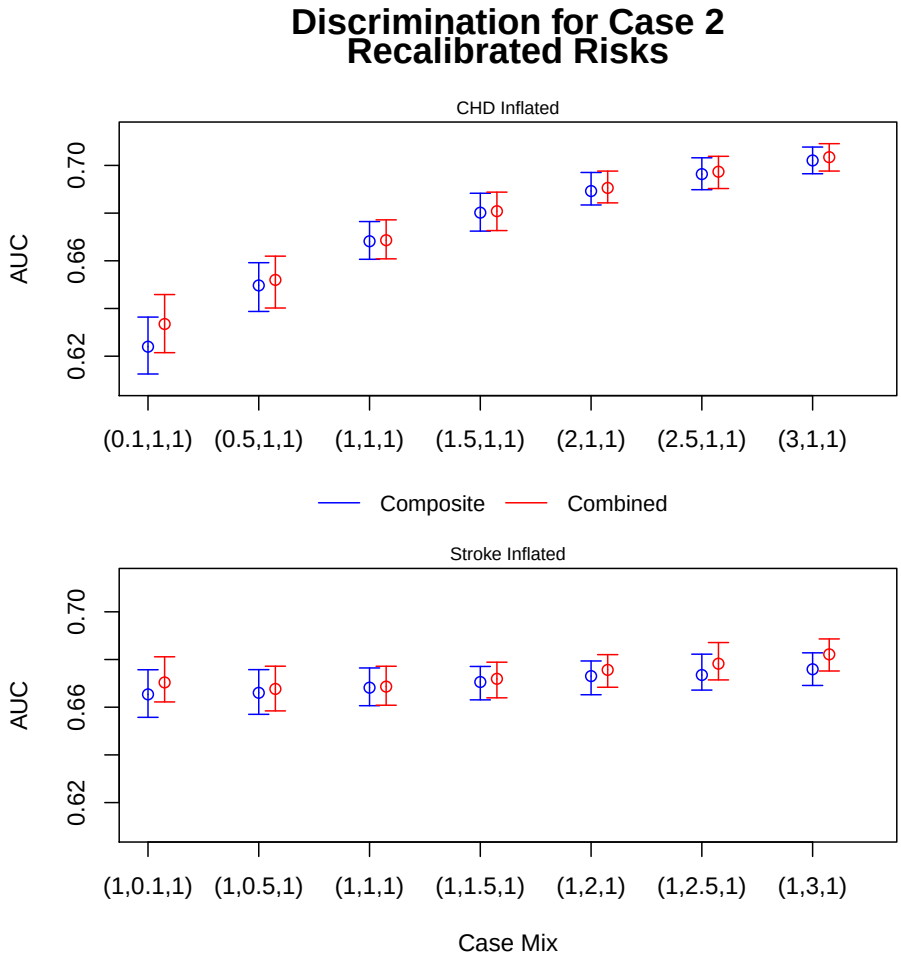


Figure A.2: Area under ROC Curve for Case 2 Recalibrated Risk Scores Where Models Differ

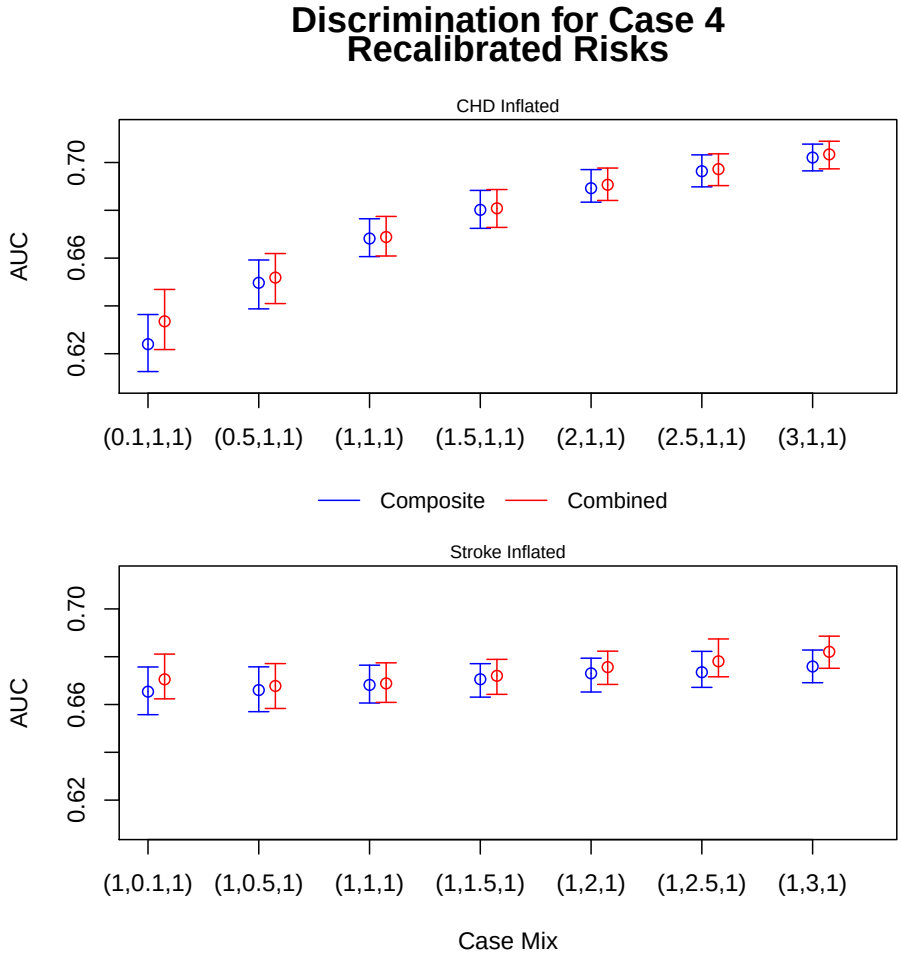


Figure A.3: Baseline Cumulative Sub-hazards Without CAC (See Table 3.1 for Centered Covariate Values)

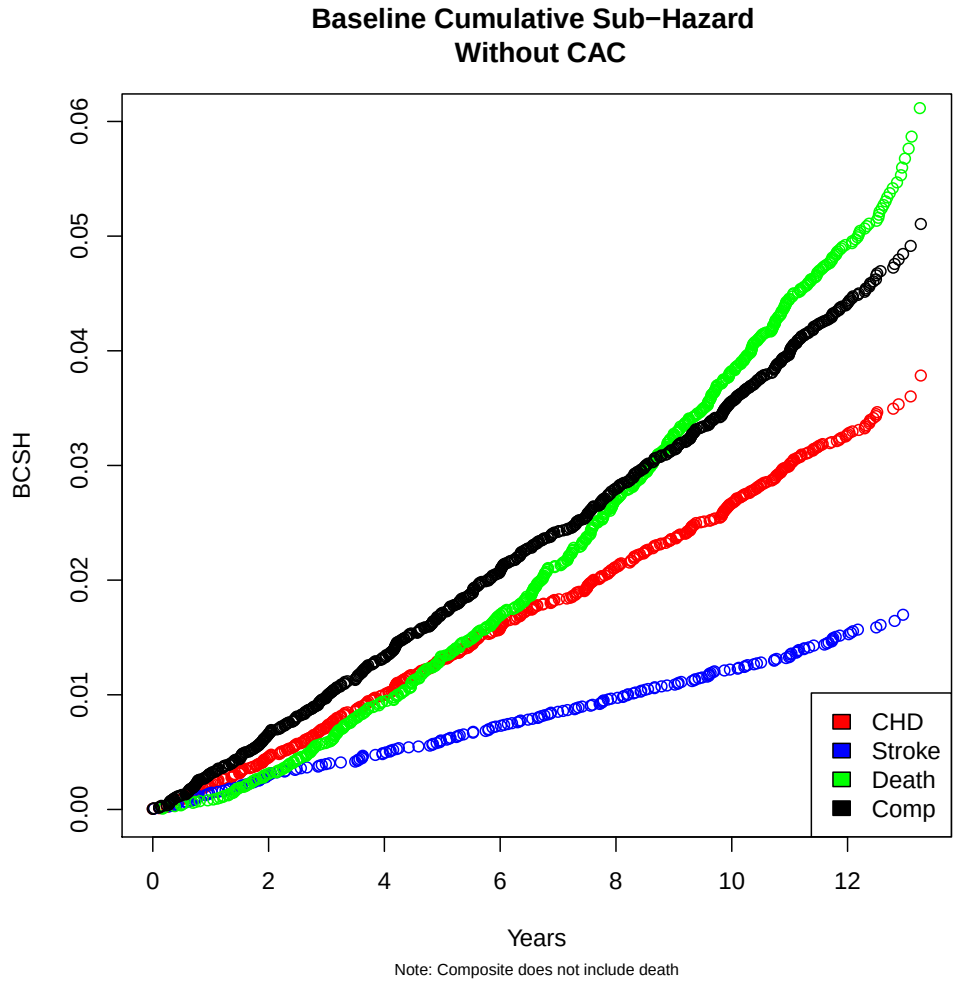
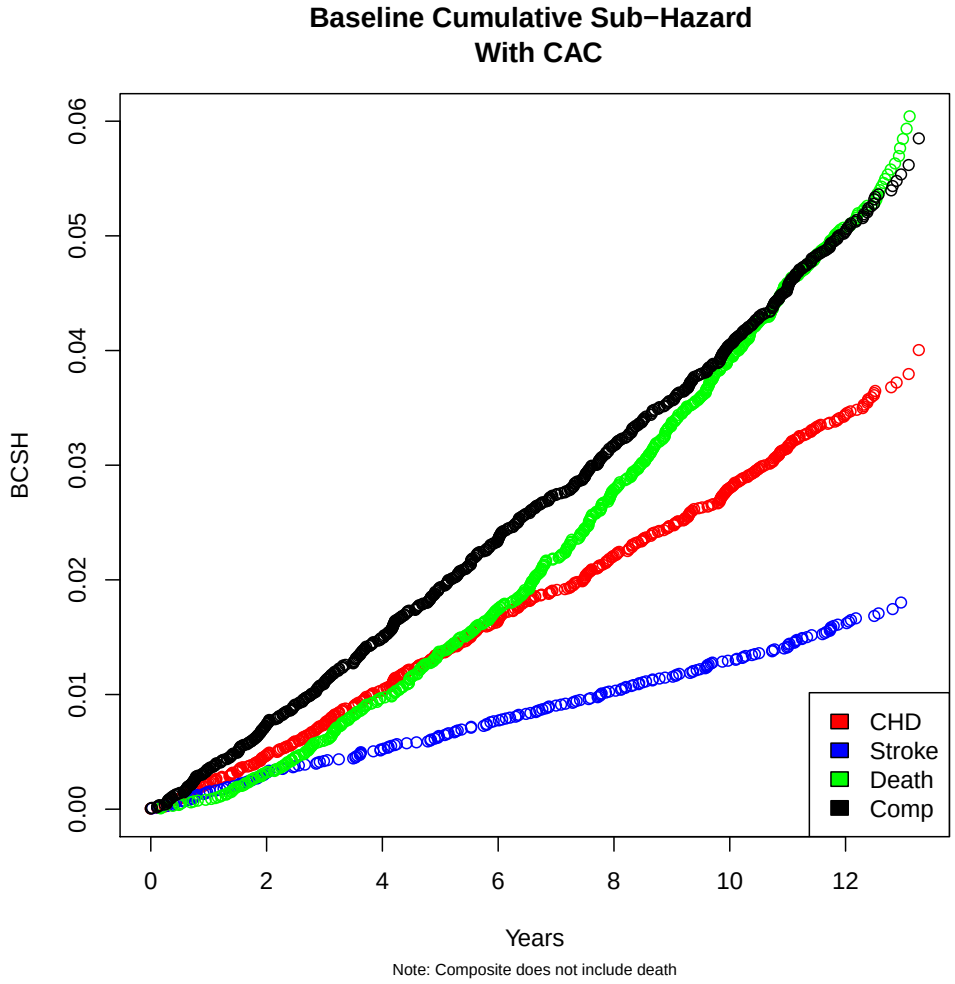


Figure A.4: Baseline Cumulative Sub-hazards With CAC (See Table 3.1 for Centered Covariate Values)



Appendix B

ANALYTICAL CODE

The programming for this project was done in R and made heavy use of the following R packages:

- survsim
- survival
- foreign
- cmprsk
- gtools
- ResourceSelection
- plotrix
- plyr
- parallels
- AUC
- gbm
- usd

B.1 Analysis Functions

```
#function that takes a data from a model and extracts coefficients and covariate
values to create an XB vector
xbeta <- function(data, model){
 x <- as.matrix(data[, names(model$coef)])
 beta <- (as.numeric(model$coef))
 return((x %*% beta))
}

#function that takes a model and time and calculates the position of the failure
```

```
maximum failure time less than or equal to the follow - up time of interest
max.uftime <- function(model, time){ max(which(model$uftime <= time))}

#function that takes a model and Breslow jumps in the estimate of the base Cumulative
Sub-hazard and returns an estimate of the base CSH for the time of interest
base.csh <- function(model, time){ sum(model$bfitj[1:max.uftime(model, time)])}

#Function that extracts the discrimination slope based on a data frame position
of the column containing the outcome and a vector of predicted risks
disc.slope <- function(data, cOutcome, predrisk){
     risk <- predrisk
     p <- list(Discrim.Slope = round(mean(risk[data[, cOutcome] == 1]) -
          mean(risk[data[, cOutcome] == 0]), 3))
     return(p)
}

#function that returns a data frame that has no missing values for the columns
of interest
completeFun <- function(data, desiredCols){
     completeVec <- complete.cases(data[, desiredCols])
     return(data[completeVec,])
}

#Function that takes XBs and BCSHs and returns a corresponding risk estimate
fgrisk <- function(xbeta, bcsh){1 - exp(- (exp(xbeta) %*%t(bcsh)))}
```

B.2 Cleaning and Forming the Development MESA Dataset

```
#Eliminate observations with missing values of the covariates and outcomes of
interest
cols <- colnames(mesa)
mesa <- completeFun(mesa, cols)

#Create Centered Variables
mesa$cage1c <- mesa$age1c - mean(mesa$age1c)
mesa$csbp1c <- mesa$sbp1c - mean(mesa$sbp1c)
mesa$clogcac1 <- mesa$logcac1 - mean(mesa$logcac1)
mesa$cchol1 <- mesa$chol1 - mean(mesa$chol1)
mesa$chdl1 <- mesa$hdl1 - mean(mesa$hdl1)
```

```
#Define covariate matrices for survival models
chd.cov <- with(mesa, cbind(cage1c, gender1, race2, race3, race4, csbp1c, clogcac1,
     diab, smoker, cchol1, chdl1, lipid1c, htnmed1c, fhhx))
stk.cov <- with(mesa, cbind(cage1c, gender1, race2, race3, race4, csbp1c, clogcac1,
     diab, smoker, cchol1, chdl1, lipid1c, htnmed1c, fhhx, fhxstroke,  maxcarotid2))
comp.cov <- with(mesa, cbind(cage1c, gender1, race2, race3, race4, csbp1c, clogcac1,
     diab, smoker, cchol1, chdl1, lipid1c, htnmed1c, fhhx, fhxstroke,  maxcarotid2))
chd.cov.nocac <- with(mesa, cbind(cage1c, gender1, race2, race3, race4, csbp1c,
     diab, smoker, cchol1, chdl1, lipid1c, htnmed1c, fhhx))
stk.cov.nocac <- with(mesa, cbind(cage1c, gender1, race2, race3, race4, csbp1c,
     diab, smoker, cchol1, chdl1, lipid1c, htnmed1c, fhhx, fhxstroke,  maxcarotid2))
comp.cov.nocac <- with(mesa, cbind(cage1c, gender1, race2, race3, race4, csbp1c,
     diab, smoker, cchol1, chdl1, lipid1c, htnmed1c, fhhx, fhxstroke,  maxcarotid2))

#Re-scale time-to-event
mesa$strktt <- mesa$strktt/365.25
mesa$chdhtt <- mesa$chdhtt/365.25
mesa$dthtt <- mesa$dthtt/365.25
mesa$revctt <- mesa$revctt/365.25
mesa$angtt <- mesa$angtt/365.25

#Create single event variable for competing cardiovascular disease events:
#Hard CHD Event
#Fatal or NonFatal Stroke
#Non-CVD Death(Competing Risk)
#Hard CHD events included MI, resuscitated cardiac arrest, fatal CHD, and
#revascularization if the participant also had prior or concurrent angina.
mesa$chdhtt[mesa$revc == 1 & mesa$ang == 1 & mesa$angtt <= mesa$revctt &
     mesa$chdh == 0] <- mesa$revctt[mesa$revc == 1 & mesa$ang == 1 &
     mesa$angtt <= mesa$revctt & mesa$chdh == 0]
mesa$chdh[(mesa$revc == 1 & mesa$ang == 1 & mesa$angtt <= mesa$revctt) |
     (mesa$dth == 1 & mesa$dthtype <= 4 & mesa$dthtype! = 2)] <- 1
mesa$chdhtt[mesa$dth == 1 & mesa$dthtype <= 4 & mesa$dthtype! = 2] <-
     pmin(mesa$dthtt[mesa$dth == 1 & mesa$dthtype <= 4 & mesa$dthtype! = 2],
     mesa$chdhtt[mesa$dth == 1 & mesa$dthtype <= 4 & mesa$dthtype! = 2])
event.1 <- vector(length = nrow(mesa))
event.1[(mesa$chdh == 1 & mesa$chdhtt <= mesa$strktt &
     is.na(mesa$strktt)  == F) | (mesa$chdh == 1 & mesa$strk == 0)] <- 1
event.1[(mesa$strk == 1 & mesa$strktt <= mesa$chdhtt &
     is.na(mesa$chdhtt)  == F) | (mesa$strk == 1 & mesa$chdh == 0)] <- 2
event.1[mesa$chdh == 0 & mesa$strk == 0 & mesa$dth == 1] <- 3
```

```
fu.time.1 <- vector(length = nrow(mesa))
fu.time.1[event.1 == 0] <- pmin(mesa$chdhtt[event.1 == 0],
     mesa$strktt[event.1 == 0], mesa$dthtt[event.1 == 0])
fu.time.1[event.1 == 1] <- mesa$chdhtt[event.1 == 1]
fu.time.1[event.1 == 2] <- mesa$strktt[event.1 == 2]
fu.time.1[event.1 == 3] <- mesa$dthtt[event.1 == 3]
comp.cvd.1 <- vector(length = nrow(mesa))
comp.cvd.1[event.1 == 1 | event.1 == 2] <- 1
comp.cvd.1[event.1 == 3] <- 2
cvd.10yr.1 <- vector(length = nrow(mesa))
cvd.10yr.1[event.1> = 1 & event.1 <3 & fu.time.1 <= 10.01] <- 1
chd.10yr.1 <- as.numeric(event.1 == 1 & fu.time.1 <= 10.01)
stk.10yr.1 <- as.numeric(event.1 == 2 & fu.time.1 <= 10.01)
cvd.10yr.1f <- as.factor(cvd.10yr.1)
chd.10yr.1f <- as.factor(event.1 == 1 & fu.time.1 <= 10.01)
stk.10yr.1f <- as.factor(event.1 == 2 & fu.time.1 <= 10.01)

#Examples of Fine and Gray Competing Risks Models

#With CAC
chd.1 <- crr(fu.time.1, event.1, cov1 = chd.cov, failcode = 1, cencode = 0)

#Without CAC
chd.2 <- crr(fu.time.1, event.1, cov1 = chd.cov.nocac, failcode = 1,
     cencode = 0)

#Calculate Baseline Cumulative Sub-Hazards at ~10 Years and XBs
chd.bcsh.1 <- base.csh(chd.1, 10.01)
chd.bcsh.2 <- base.csh(chd.2, 10.01)
chd.xbeta.1 <- xbeta(mesa, chd.1)
chd.xbeta.2 <- xbeta(mesa, chd.2)

#Compute Risk Scores
chd.risk.1 <- fgrisk(chd.xbeta.1, chd.bcsh.1)
chd.risk.2 <- fgrisk(chd.xbeta.2, chd.bcsh.2)

#Calcuate Baseline Cumulative Sub-Hazards Across Time
chd.bht.1 <- predict.crr(chd.1, cov1 = 0)
chd.bht.2 <- predict.crr(chd.2, cov1 = 0)
```

```
#Estimate Kaplan-Meier Curves for Each Event
chd.km <- survfit(Surv(fu.time.1, event.1 == 1) ~1, data = mesa)

#Plot Kaplan-Meier Curves
pdf("filepath/mesakaplanmeier.pdf")
plot(chd.km, ylim = c(0.85, 1), mark.time = F, conf.int = F, col = 'red',
     main = "Kaplan-Meier Curves \nfor the MESA Cohort", xlab = "Years",
     ylab = "Survival")
lines(stk.km, mark.time = F, conf.int = F, col = 'blue')
lines(dth.km, mark.time = F, conf.int = F, col = 'green')
lines(comp.km, mark.time = F, conf.int = F, col = 'black')
legend('bottomleft', inset = 0.01, cex = 0.85,
     legend = c("CHD", "Stroke", "Death", "Composite"),
     col = c('red', 'blue', 'green', 'black'), lty = c(1, 1, 1, 1))
dev.off()

#make tables of event coefficients and bcshs
#no CAC
c(chd.2$coef, chd.bcsh.2)

#With CAC
c(chd.1$coef, chd.bcsh.1)

#Plot CHD vs. Stroke risks
#No CAC
colcode.2 <- vector(length = length(comp.risk.2))
colcode.2[comp.risk.2 <= 0.05] <- "blue"
colcode.2[comp.risk.2>0.05 & comp.risk.2 <= 0.1] <- "yellow"
colcode.2[comp.risk.2>0.1 & comp.risk.2 <= 0.2] <- "red"
colcode.2[comp.risk.2>0.2] <- "purple"

pdf("filepath/stratifiedchdstk.nocac.pdf")
plot(chd.risk.2, stk.risk.2, xlim = c(0, 0.6), ylim = c(0, 0.6), col = colcode.2,
     xlab = "Chd Risk", ylab = "Stroke Risk", main = "Components of Cardiovascular
     \nDisease Risk")
legend("topright", legend = c("Composite risk<=0.05", "0.05<Composite Risk<=0.1",
     "0.1<Composite Risk<=0.2", "Composite Risk>0.2"), cex = 0.85, inset = 0.01,
     col = c("blue", "yellow", "red", "purple"), pch = 19)
dev.off()
```

```
#Example of Plotting Goodness of Fit
pdf("filepath/comp.hl.1.pdf")
calibrate.plot(cvd.10yr.1, comp.risk.1, main = "Composite Endpoint",
     xlab = "Full Model")
dev.off()

#Compute Receiver Operating Characteristic Curves and Areas Under the Curve
chd.roc.1 <- roc(chd.risk.1, chd.10yr.1f)
chd.auc.1 <- auc(chd.roc.1, 0, 1)
chd.roc.2 <- roc(chd.risk.2, chd.10yr.1f)
chd.auc.2 <- auc(chd.roc.2, 0, 1)

#Plot AUCS
pdf("filepath/chd.auc.1.pdf")
plot(chd.roc.1, main = "CHD")
legend('bottomright', inset = 0.01, legend = "AUC = 0.793", bty = "n")
dev.off()

#Plot BCSHs Across Time
pdf("filepath/bcshscac.pdf")
plot(chd.bht.1[, 1], - log(1 - chd.bht.1[, 2]), col = 'red',  ylim = c(0, 0.06),
     ylab = "BCSH", xlab = "Years", main = "Baseline Cumulative Sub-Hazard
     \n With CAC")
points(stk.bht.1[, 1], - log(1 - stk.bht.1[, 2]), col = 'blue')
points(dth.bht.1[, 1], - log(1 - dth.bht.1[, 2]), col = 'green')
points(comp.bht.1[, 1], - log(1 - comp.bht.1[, 2]), col = 'black')
legend('bottomright', legend = c("CHD", "Stroke", "Death", "Comp"),
     col = c('red', 'blue', 'green', 'black'),
     fill = c('red', 'blue', 'green', 'black'))
dev.off()
```

B.3 Simulating Competing-Risks Survival Data and Comparing Models

```
#Make use of coefficients from single development simulation
load("~/filepath/expregs.RData")
exp.chd.1 <- expregs[[1]]
exp.stk.1 <- expregs[[2]]
exp.dth.1 <- expregs[[3]]
exp.comp.1 <- expregs[[4]]
```

```
#Extract BCSHs from development models
exp.chd.bcsh.1 <- base.csh(exp.chd.1, 10.01)

#Simulation Function using crisk.sim from the survsim package
sim.one <- function(obs, chd.factor, stk.factor, dth.factor)
{
sim.data <- crisk.sim(n = obs, foltime = 200,
     dist.ev = c("weibull", "weibull", "weibull"), anc.ev = c(1, 1, 1),
     beta0.ev = c(5.793952 - log(chd.factor), 5.793952 - log(stk.factor),
     5.793952 - log(dth.factor)), dist.cens = "weibull", anc.cens = 1,
     beta0.cens = 5.1, z = NULL,
     beta = list(c(- 0.024503283, - 0.041850203, - 0.0915710699),
          c(- 0.450638035, 0.138054608, - 0.2687836345),
          c(0.271551575, 0.525474540, 0.3067037024),
          c(- 0.024066640, 0.009199193, - 0.1728923271),
          c(0.110728392, - 0.257048380, 0.0673535691),
          c(- 0.007844862, - 0.017702946, - 0.0006914177),
          c(- 0.263545179, - 0.105682297, - 0.0566456539),
          c(- 0.460086721, - 0.421180953, - 0.2919739704),
          c(- 0.516185877, - 0.618397964, - 0.8413763997),
          c(- 0.002189288, - 0.003636525, 0.0028858239),
          c(0.009034325, 0.010695192, 0.0002666374),
          c(- 0.082614084, 0.265239552, 0.2458083013),
          c(- 0.192557999, - 0.314757029, - 0.1305906408),
          c(- 0.312645352, - 0.085815272, 0.0919088031)),
     x = list(c("normal", 0, 10.2470563),
          c("bern", 0.4719988),
          c("bern", 0.1188561),
          c("bern", 0.2743521),
          c("bern", 0.2207328),
          c("normal", 0, 21.51129),
          c("normal", 0, 2.52078),
          c("bern", 0.1255585),
          c("bern", 0.1301758),
          c("normal", 0, 35.70227),
          c("normal", 0, 14.84637),
          c("bern", 0.1633899),
          c("bern", 0.3725052),
          c("bern", 0.4008043)), nsit = 3)
```

```
#rename data so covariate vectors conform in xbeta() function
sim.data <- rename(sim.data, replace = c(x = "cage1c", "x.1" = "gender1",
     "x.2" = "race2", "x.3" = "race3", "x.4" = "race4", "x.5" = "csbp1c",
     "x.6" = "clogcac1", "x.7" = "diab", "x.8" = "smoker", "x.9" = "cchol1",
     "x.10" = "chdl1", "x.11" = "lipid1c", "x.12" = "htnmed1c", "x.13" = "fhhx"))

#format event and time data for use in crr() function
sim.data$cause[sim.data$time>12.2] <- 0
sim.data$time[sim.data$time>12.2] <- 12.2
sim.data$cause[is.na(sim.data$cause) == TRUE] <- 0
fu.time.1 <- sim.data$time
event.1 <- sim.data$cause
comp.cvd.1 <- vector(length = nrow(sim.data))
comp.cvd.1[event.1 == 1 | event.1 == 2] <- 1
comp.cvd.1[event.1 == 3] <- 2
chd.10yr.1 <- as.factor(event.1 == 1 & fu.time.1 <= 10.01)
stk.10yr.1 <- as.factor(event.1 == 2 & fu.time.1 <= 10.01)
cvd.10yr.1 <- vector(length = nrow(sim.data))
cvd.10yr.1[event.1> = 1 & event.1 <3 & fu.time.1 <= 10.01] <- 1
cvd.10yr.1f <- as.factor(cvd.10yr.1)

#define covariate matrix
cov = with(sim.data, cbind(cage1c, gender1, race2, race3, race4, csbp1c, clogcac1,
     diab, smoker, cchol1, chdl1, lipid1c, htnmed1c, fhhx))

#fit coefficients from development dataset to simulated data
exp.chd.xbeta = xbeta(sim.data, exp.chd.1)
exp.stk.xbeta = xbeta(sim.data, exp.stk.1)
exp.dth.xbeta = xbeta(sim.data, exp.dth.1)
exp.comp.xbeta = xbeta(sim.data, exp.comp.1)

#Case 2 and Case 4
#Fit Fine and Gray Models
chd.1 <- crr(fu.time.1, event.1, cov1 = exp.chd.xbeta, failcode = 1,
     cencode = 0)

#Fit Baseline Cumulative Sub - Hazards for all events
chd.bcsh.1 <- base.csh(chd.1, 10.01)

#Fit Linear Adjustment for Case 4 Recalibration
chd.beta.1 <- as.numeric(chd.1$coef)
```

```
#Approximate version of Case 2 using the cumulative incidence to approximate the
BCSH
tp <- timepoints(cuminc(fu.time.1, event.1), 10.01)
comp.tp <- timepoints(cuminc(fu.time.1, comp.cvd.1), 10.01)
chd.cuminc <- tp$est[1]

#Testing Accuracy of Simulation Method by Refitting Original Models
chd.2 <- crr(fu.time.1, event.1, cov1 = cov, failcode = 1, cencode = 0)

#Re - estimating BCSHs
chd.bcsh.2 <- base.csh(chd.2, 10.01)

#Case 2 Risks
chd.risk.1 <- fgrisk(xbeta = exp.chd.xbeta, chd.bcsh.1)
combined.risk.1 <- fgrisk(xbeta = cbind(exp.chd.xbeta, exp.stk.xbeta),
     bcsh = cbind(chd.bcsh.1, stk.bcsh.1))

#Case 4 Risks
chd.risk.2 <- fgrisk(xbeta = chd.beta.1*exp.chd.xbeta, bcsh = chd.bcsh.1)
combined.risk.2 <- fgrisk(xbeta = cbind((stk.beta.1*exp.stk.xbeta),
     (chd.beta.1*exp.chd.xbeta)), bcsh = cbind(stk.bcsh.1, chd.bcsh.1))

#Unrecalibrated Risks
ur.chd.risk <- fgrisk(xbeta = exp.chd.xbeta, bcsh = exp.chd.bcsh.1)

#Using Cumulative Incidence Estimate of 10 year Risk
emp.chd.risk <- fgrisk(xbeta = exp.chd.xbeta, bcsh = chd.cuminc)
emp.combined.risk <- fgrisk(xbeta = cbind((exp.stk.xbeta),(exp.chd.xbeta)),
     bcsh = cbind(stk.cuminc, chd.cuminc))

#Hosmer - Lemeshow Test Statistics
chd.hl.1 <- as.numeric(hoslem.test((event.1 == 1 & fu.time.1 <= 10.01),
     chd.risk.1)$statistic)

#HL P - values
chd.pval.1 <- as.numeric(hoslem.test((event.1 == 1 & fu.time.1 <= 10.01),
     chd.risk.1)$p.value)

#Area Under the Receiver Operating Characteristic Curve
chd.auc.1 <- auc(roc(chd.risk.1, chd.10yr.1), 0, 1)
```

```
#extensive output of calibration and discrimination tests as well as checks for
simulation validity
sums <- c("chd.hl.1"=chd.hl.1, "stk.hl.1"=stk.hl.1, "comp.hl.1"=comp.hl.1,
     "combined.hl.1"=combined.hl.1, "chd.hl.2"=chd.hl.2, "stk.hl.2"=stk.hl.2,
     "comp.hl.2"=comp.hl.2, "combined.hl.2"=combined.hl.2,  "chd.hl.3"=chd.hl.3,
     "stk.hl.3"=stk.hl.3, "comp.hl.3"=comp.hl.3, "combined.hl.3"=combined.hl.3,
     "comp.hl.4"=comp.hl.4, "combined.hl.4"=combined.hl.4, "chd.cbcsh"=chd.bcsh.2,
     "chd"=chd.2$coef, "stk.bcsh"=stk.bcsh.2, "stk"=stk.2$coef, "dth.bcsh"=dth.bcsh.2,
     "comp"=comp.2$coef, "comb.bcsh"=comp.bcsh.2, "chd.pval.1"=chd.pval.1,
     "stk.pval.1"=stk.pval.1, "comp.pval.1"=comp.pval.1,
     "combined.pval.1"=combined.pval.1, "chd.pval.2"=chd.pval.2,
     "stk.pval.2"=stk.pval.2, "comp.pval.2"=comp.pval.2,
     "combined.pval.2"=combined.pval.2, "chd.pval.3"=chd.pval.3, "
     stk.pval.3"=stk.pval.3, "comp.pval.3"=comp.pval.3,
     "combined.pval.3"=combined.pval.3, "comp.pval.4"=comp.pval.4,
     "combined.pval.4"=combined.pval.4, "chd.auc.1"=chd.auc.1, "stk.auc.1"=stk.auc.1,
     "comp.auc.1"=comp.auc.1, "combined.auc.1"=combined.auc.1, "chd.auc.2"=chd.auc.2,
     "stk.auc.2"=stk.auc.2, "comp.auc.2"=comp.auc.2, "combined.auc.2"=combined.auc.2,
     "chd.auc.3"=chd.auc.3, "stk.auc.3"=stk.auc.3, "comp.auc.3"=comp.auc.3,
     "combined.auc.3"=combined.auc.3, "comp.auc.4"=comp.auc.4,
     "combined.auc.4"=combined.auc.4,
     "comp.calib.itl.1"=((sum(comp.risk.1)-sum(cvd.10yr.1))/length(sim.data)),
     "combined.calib.itl.1"=((sum(combined.risk.1)-sum(cvd.10yr.1))/length(sim.data)),
     "comp.calib.itl.2"=((sum(comp.risk.2)-sum(cvd.10yr.1))/length(sim.data)),
     "combined.calib.itl.2"=((sum(combined.risk.2)-sum(cvd.10yr.1))/length(sim.data)),
     "comp.calib.itl.3"=((sum(ur.comp.risk)-sum(cvd.10yr.1))/length(sim.data)),
     "combined.calib.itl.3"=((sum(ur.combined.risk)-sum(cvd.10yr.1))/length(sim.data)),
     "comp.calib.itl.4"=((sum(emp.comp.risk)-sum(cvd.10yr.1))/length(sim.data)),
     "combined.calib.itl.4"=((sum(emp.combined.risk)-sum(cvd.10yr.1))/length(sim.data)))
return(sums)
}

#function to replicate simulation in a cluster computing environment
sim.setup <- function(seed, obs, chd.factor, stk.factor, dth.factor, reps){
set.seed(seed)
dataname <- lapply(1:reps, function(z)
     sim.one(obs, chd.factor, stk.factor, dth.factor))
dataname <- t(simplify2array(dataname))
dataname <- cbind(dataname, "chd.prop.sig.1" = mean(dataname[, "chd.pval.1"] <0.05),
     "stk.prop.sig.1" = mean(dataname[, "stk.pval.1"] <0.05),
```

```
     "comp.prop.sig.1" = mean(dataname[, "comp.pval.1"] <0.05),
     "comb.prop.sig.1" = mean(dataname[, "combined.pval.1"] <0.05),
     "chd.prop.sig.2" = mean(dataname[, "chd.pval.2"] <0.05),
     "stk.prop.sig.2" = mean(dataname[, "stk.pval.2"] <0.05),
     "comp.prop.sig.2" = mean(dataname[, "comp.pval.2"] <0.05),
     "comb.prop.sig.2" = mean(dataname[, "combined.pval.2"] <0.05),
     "chd.prop.sig.3" = mean(dataname[, "chd.pval.3"] <0.05),
     "stk.prop.sig.3" = mean(dataname[, "stk.pval.3"] <0.05),
     "comp.prop.sig.3" = mean(dataname[, "comp.pval.3"] <0.05),
     "comb.prop.sig.3" = mean(dataname[, "combined.pval.3"] <0.05))
save(dataname, file = paste0("~/crsims/output/sim", chd.factor, stk.factor, dth.factor,
     seed, ".RData"))
return(dataname)
}

sim.setup(seed, obs, chdfactor, stkfactor, dthfactor, reps)
```

B.4 Plotting Simulation Results

In this section the generic variable 'x' is substituted for a seed set in the simulations produced in the previous section.

```
#Load in data produced by simulations
load("Data/sim0.111x.RData")
sim.0.111.x <- dataname

#Create arrays of simulations with inflation factors on the same endpoint
sims1.x <- array(data = list(sim.0.111.x, sim.0.511.x, sim.111.x,
     sim.1.511.x, sim.211.x, sim.2.511.x, sim.311.x))
sims2.x <- array(data = list(sim.10.11.x, sim.10.51.x, sim.111.x, sim.11.51.x,
     sim.121.x, sim.12.51.x, sim.131.x))
axis1 <- c("(0.1,1,1)", "(0.5,1,1)", "(1,1,1)", "(1.5,1,1)", "(2,1,1)", "(2.5,1,1)",
     "(3,1,1)")
axis2 <- c("(1,0.1,1)", "(1,0.5,1)", "(1,1,1)", "(1,1.5,1)", "(1,2,1)", "(1,2.5,1)",
     "(1,3,1)")

#Function to plot the 25th, 50th, and 75th percentile of the values of the simulation
results
pctile.graph <- function(col, n, sims, axislab, ...){
x <- 1:n
x2 <- 1:n+.15
```

```
compp05 <- rbind(quantile(sims[[1]][, col[1]], .25),
     quantile(sims[[2]][, col[1]], .25),
     quantile(sims[[3]][, col[1]], .25),
     quantile(sims[[4]][, col[1]], .25),
     quantile(sims[[5]][, col[1]], .25),
     quantile(sims[[6]][, col[1]], .25),
     quantile(sims[[7]][, col[1]], .25))
compmed <- rbind(quantile(sims[[1]][, col[1]], .5),
     quantile(sims[[2]][, col[1]], .5),
     quantile(sims[[3]][, col[1]], .5),
     quantile(sims[[4]][, col[1]], .5),
     quantile(sims[[5]][, col[1]], .5),
     quantile(sims[[6]][, col[1]], .5),
     quantile(sims[[7]][, col[1]], .5))
compp95 <- rbind(quantile(sims[[1]][, col[1]], .75),
     quantile(sims[[2]][, col[1]], .75),
     quantile(sims[[3]][, col[1]], .75),
     quantile(sims[[4]][, col[1]], .75),
     quantile(sims[[5]][, col[1]], .75),
     quantile(sims[[6]][, col[1]], .75),
     quantile(sims[[7]][, col[1]], .75))

plotCI(x, compmed, ui = compp95, li = compp05, xlim = c(1, n+.5),
     ylim = c(min(compp05, combp05), max(compp95, combp95)), col = "blue",
      scol = "blue", xaxt = "n", ...)
plotCI(x2, combmed, ui = combp95, li = combp05, axes = F, add = T, col = "red",
     scol = "red", ...)
axis(labels = axislab, side = 1, at = c(1, 2, 3, 4, 5, 6, 7))
}

#Formatted plot of results for Hosmer - Lemeshow Test Statistics
pdf("filepath/unrecalibratedhlsx.pdf")
par(mfcol = c(2, 1), oma = c(0, 0, 4, 0), mar = c(5, 4, 1, 4), cex.main = 1.5,
     xpd = NA)
pctile.graph(col = c("comp.hl.3", "combined.hl.3"), n = 7, sims = sims1.x,
     axislab = axis1, ylab = "HL Test Statistics", xlab = "")
pctile.graph(col = c("comp.hl.3", "combined.hl.3"), n = 7, sims = sims2.x,
     axislab = axis2, ylab = "HL Test Statistics", xlab = "Case Mix")
title(c("Calibration Results", "for Unrecalibrated Risks"), outer = TRUE)
legend(x = 2.75, y = 430, horiz = TRUE, legend = c("Composite", "Combined"),
     col = c("blue", "red"), pch = 19, cex = .95, bty = "n")
```

```
dev.off()

#Same approach works for AUC plots and Calibration In - The - Large
pdf("filepath/unrecalibratedaucsx.pdf")
par(mfcol = c(2, 1), oma = c(0, 0, 4, 0), mar = c(5, 4, 1, 4), cex.main = 1.5,
     xpd = NA)
pctile.graph(col = c("comp.auc.3", "combined.auc.3"), n = 7, sims = sims1.x,
     axislab = axis1, ylab = "AUC", xlab = "")
pctile.graph(col = c("comp.auc.3", "combined.auc.3"), n = 7, sims = sims2.x,
     axislab = axis2, ylab = "AUC", xlab = "Case Mix")
title(c("Discrimination", "for Unrecalibrated Risks"), outer = TRUE)
legend(x = 2.75, y = .695, horiz = TRUE, legend = c("Composite", "Combined"),
     col = c("blue", "red"), pch = 19, cex = .95, bty = "n")
dev.off()

#As well as AUC
pdf("filepath/case2aucsx.pdf")
par(mfcol = c(2, 1), oma = c(0, 0, 4, 0), mar = c(5, 4, 1, 4), cex.main = 1.5,
     xpd = NA)
pctile.graph(col = c("comp.auc.1", "combined.auc.1"), n = 7, sims = sims1.x,
     axislab = axis1, ylab = "AUC", xlab = "")
pctile.graph(col = c("comp.auc.1", "combined.auc.1"), n = 7, sims = sims2.x,
     axislab = axis2, ylab = "AUC", xlab = "Case Mix")
title(c("Discrimination for Case 2", "Recalibrated Risks"), outer = TRUE)
legend(x = 3, y = .7, horiz = TRUE, legend = c("Composite", "Combined"),
     col = c("blue", "red"), pch = 19, cex = .95, bty = "n")
dev.off()
```