# Assessing Risk Score Calculation in the Presence of Uncollected Risk Factors

By

Alice Toll

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

January 31, 2019

Nashville, Tennessee

Approved:

Dandan Liu, Ph.D.

Qingxia Chen, Ph.D.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

Figure Page

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AFib | Atrial Fibrillation |
| AntiHypT | Anti-Hypertension Rx |
| CVD | cardiovascular disease |
| ECG | echocardiogram |
| FHS | Framingham Heart Study |
| FSRP | Framingham Stroke Risk Profile |
| LVH | Left Ventricular Hypertrophy |
| NHLBI | National Heart, Lung, and Blood Institute |
| SBP | Systolic Blood Pressure |

# CHAPTER 1

## INTRODUCTION

Risk prediction models developed from clinical studies have been widely used for individualized risk prediction to aid clinical decision making in the clinical setting and risk adjustments in clinical studies. It uses selected predictors to estimate the probability that an individual will develop a disease or experience an event such as a stroke within a specific period of time. A well-developed risk prediction model could provide accurate individualized risk prediction when all the requested predictors are available.

In general, a risk prediction model is developed from a regression model relating a linear combination of predictors to the outcome of interest. For the ease of utility, linear combinations of predictors are usually presented as a points system, where points are assigned for each predictor based on patient profile and the summation of the points are calculated as a risk score. The risk of event could then be calculated as a monotonic function of the risk score. A detailed example of such a point system is provided in Chapter 2 for the Framingham stroke risk profile (FSRP) (Wolf et al., 1991). Other than individualized risk prediction, risk scores are also commonly used for risk stratification to facilitate optimal resource allocation. For example, the Model for End-Stage Liver Disease (MELD) measures disease severity (Kamath et al., 2001). This risk score is incorporated into deceased donor liver allocation. Liver waitlist candidates with higher MELD scores are given prioritized access to broader sharing of organs, thereby increasing their access to transplant. In clinical research, risk scores are often used for risk adjustment purpose, in which risk scores are used as adjusting variables. For example, Charlson Comorbidity Index is often included as an adjusting variable to avoid potential confounding effects when intervention effect on some clinical outcomes are being evaluated. Adjusting for a single risk score rather than all relevant risk factors allows for more degrees of freedom while ensuring that important risk factors that have already been reviewed by the scientific community are included.

Despite of popular utility of risk prediction modeling in clinical research, its application can present a set of challenges when the required risk factors are missing or uncollected. A predictor might be missing for some but not all participants, as is often the case when the risk factor is self-reported or requires expensive tested and was only ordered when deemed necessary. Such missingness can introduce bias into the study because the test may only have been ordered if the physician thinks the patient might have related adverse outcome. Three approaches are generally adopted to handle such missing risk factors, i.e. treating missingness as absence, complete case analysis, and imputation.

In the following, three examples on the application of Framingham Risk Score were provided to illustrate the three approaches. Sara et al. (2016) looked at the utility of the Framingham risk score in predicting secondary cardiovascular outcomes for a cohort of patients diagnosed with coronary heart disease who had received percutaneous coronary intervention, where missing risk factors were assumed to be absent in the calculation of the risk score. Obviously, this approach will underestimate the risk for high risk patients and lead to biased evaluation in the utility of the Framingham Risk Score. Towfighi et al. investigated the Framingham risk score's ability to predict MI in patients who had recently experienced a stroke but were not known to have coronary heart disease (2012). Authors noted that patients with missing Framingham risk factors had higher values/prevalence of the other risk factors than patients with complete data. For example, they had higher systolic blood pressure (143.9 vs 141.2 mm Hg, $p = 0.001$) and higher prevalence of diabetes (32% vs 25%, $p = 0.07$). However, only patients with complete data were included in the analysis, which potentially lead to biased results since high risk patients may have been excluded due to missing data. Ankle Brachial Index Collaboration (2008) used the Framingham risk score to assess the accuracy of the ankle brachial index in predicting cardiovascular disease. Missing risk factors from the Framingham risk score were imputed. This approach is the most appropriate if the imputation is conducted appropriately. Another scenario of missing risk factors is uncollected (i.e. systematically missing) risk factors, where relevant risk factors are not collected for any subjects in the study. This often occurs when a risk score is used in studies that are different from where it was originally developed and thus not all risk factors of the risk score are included in the study protocol In the presence of uncollected risk factors, the above mentioned complete case analysis and imputation approach cannot be used. Most often, the first approach treating missingness as absence is adopted, which often lead to biased assessment on the utility of a risk prediction model.

In this thesis, we will assess the impact of uncollected risk factor on the utility of risk prediction models. Extensive simulation studies will be conducted to evaluated characteristics of uncollected risk factors in relation to risk prediction model performance. We will then illustrate our findings using the 10-year stroke risk prediction model developed from Framingham Heart Study (Wolf et al., 1991).

# CHAPTER 2

# MOTIVATING EXAMPLE

## 2.1 Framingham Stroke Risk Profile

The Framingham Heart Study (FHS) is an epidemiologic study funded by National Heart, Lung, and Blood Institute (NHLBI), which has committed to identifying the common factors or characteristics that contribute to cardiovascular disease (CVD) (Dawber and Moore, 1952). The study began in 1948 and recruited an Original Cohort of 5,209 respondents of a random sample of 2/3 of the adult population from the town of Framingham, Massachusetts, who had not yet developed overt symptoms of cardiovascular disease or suffered a heart attack or stroke. Subjects have continued to return to the study every two years for a detailed medical history, physical examination, and laboratory tests. Although the study cohort is primarily Caucasian, and lacks geographic, socio-economic, and environmental variability, major CVD risk factors identified in FHS have been shown in other studies to apply almost universally among racial and ethnic groups with varying patterns of distribution. Since then three generations of participants have enrolled in the study.

The Framingham Stroke Risk Profile (FSRP) was developed using FHS to predict the 10-year risk of developing stroke Wolf et al. (1991). The study included two cohorts of patients aged 55-84 and free of stroke at the time of two examination cycles. The patients had to be free of stroke during the initial examination, either examination 9 (1964-1968) or 14 (1975-1978), and followed for 10 years. Patients who were enrolled in the examination 9 cohort and followed for more than 10 years could be enrolled in examination 14 cohort again if they were free of stroke at examination 14. However, for the purpose of analysis, such patients were treated as two unique patients. Cox proportional hazards model was used with time from the initial examination to stroke event as the primary outcome. Separate models were built for males and females and variables were included based on stepwise selection methods. Only significant variables (p-value $\leq 0.05$) were included in the final model. The risk factors identified were age, systolic blood pressure, antihypertensive therapy, diabetes mellitus, cigarette smoking, cardiovascular disease, atrial fibrillation, and left ventricular hypertrophy (LVH). For females an interaction between systolic blood pressure and antihypertensive therapy was also included.

FSRP risk score is developed as a linear combination of those risk factors and is calculated using a point system for convenience use. Points are a scaled version of regression coefficients. For binary risk factors, a fixed point will be assigned if the risk factor is present. For continuous risk factors, different points are assigned depending on the risk factor's value. FSRP risk score is the summation of these points and the corresponding 10-year risk of stroke could be obtained from the probability look-up table.

### 2.2 Using Stroke Risk Scores in the Presence of Uncollected Risk Factors

Application of risk profiles requires information from all risk factors and thus might be challenging when data collection for some risk factors are not included in the protocol of a research study. For example, the diagnosis of left ventricular hypertrophy (LVH) is a risk factor included in the FSRP requires an echocardiogram (ECG) which is not routinely administered in clinical practice and is usually not conducted in research studies not directly related to cardiovascular diseases. Calculating FSRP in such studies usually do not account for LVH which is equivalent to assuming all patients do not have LVH.

A retrospective analysis that compiled data from the English Longitudinal Study of Ageing and Health Survey for England did not have access to LVH since neither data source included an ECG. This study used a modified form of FSRP, omitting LVH, as an adjusting covariate for modeling cognitive function (Llewellyn et al., 2008). Such naive approach for treating unknown risk factors as absent might result in bias when a risk prediction model is externally validated for a different population or a risk score is used for risk adjustment purposes. The goal of this thesis is to assess the impact of excluding a risk factor from the risk score calculation.

# CHAPTER 3

# METHODS

## 3.1 Cox Regression Models for Risk Prediction

Risk prediction models are usually developed for binary outcomes or time-to-event outcomes. For binary outcomes representing prevalence of the event of interest, logistic regression models are commonly used, which models, the probability of event occurrence, $p$, through logit link function using a linear combination of risk factors, $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}$. The risk score is developed from the linear predictor, $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}$. Binary outcomes were mostly used in risk prediction model development for acute illness where patients are rarely subject to loss of follow-up.

For chronic diseases that take longer to develop, the occurrence of the event of interest is often subject to censoring due to loss of follow-up or competing risks of death. Therefore modelings for time-to-event outcome using survival analysis is more appropriate. Survival analysis provides an advantage over logistic regression in that it models the instantaneous risk (i.e., hazard) of the event of interest while accounting for censoring. The observed outcome for a survival model is a composite of two random variables, the time to event occurrence denoted as $T$, and the time to a censoring event, denoted as $C$. For the development of FSRP, $T$, would be the time to stroke and $C$ would be the time until any of the following: self-withdraw from the study, move away from the study area, or experience an unrelated death. We assume that censoring events are independent of the event of interest, conditioning on risk factors. For each patient, let $\delta = I\left(T < C\right)$ denote an event indicator, $Y = \min\left(T, C\right)$ denote the observed time, and $Z$ denote a $p$-vector covariate.

We consider Cox proportional hazards model Cox (1972), the most commonly used model for time-to-event outcomes. Cox regression is a semi-parametric model with a non-parametric baseline hazard and a parametric model for covariates that is proportional to the baseline hazard. This model is defined as

$$\lambda(t; \boldsymbol{Z}) = \lambda_0(t) \exp\left(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}\right) \tag{3.1}$$

where $\boldsymbol{\beta}$ is the $p$-vector parameter for $\boldsymbol{Z}$ and $\lambda_0(t)$ is the baseline hazard. This semi-parametric model, with $\lambda_0$ un-restricted, puts the focus of the model interpretation on the parameter estimates, $\boldsymbol{\beta}$.

### 3.2 Calculating Risk Scores

Risk scores are usually expressed as a linear combination of risk factors, as $R = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}$. To assess the impact of uncollected risk factors on the utility of the risk score we considered two approaches to calculate risk scores. Without loss of generality, we assumed the last risk factor, $Z_p$ is not collected. The first approach calculates a risk score by omitting the uncollected risk factor from the calculation, which is equivalent to assuming the risk factor is absent. This naive approach is commonly adopted in practice. Let $\widetilde{R} = \boldsymbol{\beta}_{-p}\boldsymbol{Z}_{-p}$ denote this risk score with the subscript $_{-p}$ indicating omitting the $p$-th element from the corresponding vector. The second approach calculates a risk score by refitting the model excluding the uncollected risk factor. Let $\widehat{R} = \widehat{\boldsymbol{\beta}}_{-p}\boldsymbol{Z}_{-p}$ denote this alternative risk score where $\widehat{\boldsymbol{\beta}}_{-p}$ denotes the parameter estimates from model refitting and reflects reassigned weights supplementing information loss due to the uncollected risk factor. This approach is less commonly used, but might provide better caliber when the original data used to develop the risk score is available.

### 3.3 Performance Measures

In the following we consider several measures to evaluate the performance of these risk scores calculated in the presence of uncollected risk factors relative to the gold standard full risk score where all risk factors are available.

#### 3.3.1 Correlation

Measurements of correlation between the naive risk score ($\widetilde{R}$ ) or refit risk score ($\widehat{R}$) and the full risk score($R$) are considered. Both Pearson and Spearman correlation coefficients were evaluated. Pearson correlation evaluates if the correlation with the full risk score is linear and potentially assess the impact of uncollected risk factors when risk scores are to be used for risk adjustment. Spearman correlation evaluates if the correlation is rank based and potentially evaluates risk scores being used for discrimination purposes.

#### 3.3.2 C-Index

Discrimination is the model's ability to separate patients into outcome groups (Harrell et al., 1996). As the most commonly used discrimination measure for survival outcome, the C-index (Harrell et al., 1982) is the proportion of all pairs of patients for which we could determine the ordering of survival times such that the predictions are concordant. A C-index value ranges from 0.5, no discrimination, to 1.0, perfect discrimination. C-index is based on ranks of predicted risks and thus cannot be used to assess accuracy of individual risk prediction. In the case where two models yields the same concordant pairs, but one model consistently predicts a larger risk the the other model, the rank-based C-index would be the same for the two models. This presents a challenge in comparing the true and refit models, but is less of a concern for the the true and naive models as we know differences in concordant pairs are attributable solely to the omitted risk factor.

### 3.3.3 IDI

As we have discussed, discrimination is a useful metric for evaluating risk prediction. A simple method for evaluating the addition of the risk factor is to build two models, one with and one without the addition, and evaluate the difference in the area under the receiver-operating-characteristic (ROC) curve (AUC). However, this difference can be very small, especially when the new risk factor is a biomarker, making it difficult to determine if this change in AUC is clinically significant.

Integrated Discrimination Increment (IDI) quantifies the effect on overall discrimination when a new risk factor is added to a risk prediction model (Pencina et al., 2008). Generally this is utilized when a risk score has already been developed and a new risk factor is separately been identified as having predictive value to the outcome of interest. IDI quantifies the improvement of model performance with the addition of a new risk factor.

IDI calculates the difference between the old and new models' sensitivity and the complement of the specificity which are integrated over the entire spectrum of risk threshold. The sensitivity curve allows us to estimate the mean predicted probability for patients who develop the event and the complement of the specificity curve allows us to estimate the mean predicted probability for patients who do not develop the event. This metric is asymptotically the same as the difference in discrimination slopes.

In our setting, we have a well-developed risk score, and we want to assess the overall discrimination decrement when a risk factor is not collected at all. Therefore, this measure could be used reversely such that the naive method and the refit method are used as the "old" models and the full model is used as the "new" model.

### 3.3.4 Calibration

Calibration describes the extent of bias in a model (Harrell et al., 1996) by its ability to accurately predict patient risk. A broad metric is calibration in the large, the difference between mean observed risk and mean predicted risk. This measure is most valuable when the model is applied to a new dataset, because high agreement is expected when using the training dataset. The Hosmer-Lemeshow test was a global test to assess calibration in the large (Hosmer Jr et al., 2013). When the outcome is continuous, a plot of expected versus observed outcomes can be used. This can be extended to survival outcomes by plotting mean predicted risks against the observed event rate within pre-defined risk stratified groups, where the predicted risks is calculated from the model at a pre-specified time-point and the observed event rate is calculated using Kaplan-Meier estimates at the time-point. This method is a visual extension of the Hosmer-Lemeshow test and provides a visual description of the model fit.

Calibration slope is a regression of the observed outcome on predicted risk and could be visually examined using calibration plots. The plot helps to describe which populations the model has high bias and may not be very predictive for. For example, when a model is fit using a cohort of mostly low risk patients, the model may be biased for high risk patients. While calibration plots are very helpful for describing a single model they are not practical for a simulation; we needed a way to summarizes a calibration plot numerically. Crowson et al. (2016) proposed regression-based methods to quantify calibration in the large and calibration slope Crowson et al. (2016). For binary or continuous outcomes these models regress the observed outcomes on the linear predictors. Extensions to survival data are also provided leveraging Poisson regression with appropriately defined off-set. These regression-based calibration measures are used in simulation studies.

To further understand the goodness of fit, we can assess the calibration by risk groups. Generally these groups are defined using quantiles of predicted risk, but other categorizations may be used as appropriate. This model, as described by Crowson et al. (2016), allows for a separate coefficient, $\gamma_{1-k}$ for each risk group, and then tests for the group effect, $\gamma_1 = \gamma_2 = \cdots = 0$. This approach clearly shows for which groups the model doesn't not accurately predict.

### 3.3.5 Differences in Predicted Risk

After building a Cox model we can estimate risks for individual patients, $\rho_i$, $i = 1, \ldots, n$. For the naive and refit models we denote the risk as $\tilde{\rho}_i$ and $\widehat{\rho}_i$ respectively. We looked at three metrics relative to the true model: mean difference, maximum absolute difference, and mean absolute relative difference. The mean difference, $\frac{\sum_i^n (\tilde{\rho}_i - \rho_i)}{n}$, describes, on average how far of a given prediction is. The maximum absolute difference, $\max\left(|\tilde{\rho}_i - \rho_i|\right)$, will show the largest possible prediction error you can make when using one of the alternative models. The mean absolute relative difference, $\frac{\sum_i^n \left(\frac{|\tilde{\rho}_i - \rho_i|}{\rho_i}\right)}{n}$ also let's us see the average error, but relative to the estimate from the true model.

# CHAPTER 4

## SIMULATION ANALYSIS

The simulation analysis was conducted using the Cox model (3.1) with five covariates $\boldsymbol{Z} = \{Z_1, Z_2, Z_3, Z_4, Z_5\}$ where $Z_1$ is a continuous variable following standard normal distribution. $Z_2$, $Z_3$, $Z_4$, and $Z_5$ are binary variables. The frequencies of $Z_2$ and $Z_5$ varied in the simulation at 5% or 20%. The distributions of $Z_3$ and $Z_4$ were consistent throughout the simulation with 15% frequency. To allow correlation between covariates, we first simulated multivariate normal variables $\boldsymbol{Z}^* \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ with the correlation matrix $\boldsymbol{\Sigma}$ specified in Table 4.1. Then we let $Z_1 = Z_1^*$, $Z_2 = I(Z_2^* > z_{1-\alpha})$, $Z_3 = I(Z_3^* > z_{1-0.15})$, $Z_4 = I(Z_3^* > z_{1-0.15})$, and $Z_5 = I(Z_5^* > z_{1-\alpha})$ with $\alpha$=0.2 or 0.05 and $z_{1-\alpha}$ denoting quantiles from standard normal distribution. Therefore $Z_2$ is a binary variable with varying degree of correlation with the three covariates, whereas $Z_5$ is a binary variable completely independent of all other covariates.

A total of 16 scenarios were considered with varying frequencies and effect sizes for $Z_2$ and $Z_5$ (Table 4.2). In addition, for each scenario, we separately considered $Z_2$ and $Z_5$ being omitted, which results in 32 scenarios overall. We only considered different scenarios for binary variables because our preliminary literature search did not suggest omitting continuous risk factors was a common practice. This is especially true when considering risk factors with wide ranges such as age and systolic blood pressure because the omission would have such a large impact on the risk score. Patient survival time, $T_i$ was simulated as the exponentiated linear predictor (with error) divided by $\lambda$. Patient censoring time, $C_i$ was simulated using a uniform distribution. The observed follow-up time for each patient was recorded as $\min(T_i, C_i)$ with and event indicator $D_i = I(T_i \leq C_i)$. $\lambda$ was adjusted for each simulation in order to achieve a 90% censoring rate to approximate the Framingham cohort.

Table 4.1: Correlation matrix of $\boldsymbol{Z}^*$.

| | ${Z_1}^*$ | ${Z_2}^*$ | ${Z_3}^*$ | ${Z_4}^*$ | ${Z_5}^*$ |
|---|---|---|---|---|---|
| ${Z_1}^*$ | 1.00 | 0.30 | 0.20 | 0.10 | 0.00 |
| ${Z_2}^*$ | 0.30 | 1.00 | 0.05 | 0.05 | 0.00 |
| ${Z_3}^*$ | 0.20 | 0.05 | 1.00 | 0.00 | 0.00 |
| ${Z_4}^*$ | 0.10 | 0.05 | 0.00 | 1.00 | 0.00 |
| ${Z_5}^*$ | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

Table 4.2: Summary of simulation scenarios with varying weights and frequencies of $Z_2$ and $Z_5$. The event rates were set to approximate the Framingham Heart Study.

| **Scenario** | $Z_2$ | | $Z_5$ | | **Event Rates** | |
|---|---|---|---|---|---|---|
| | $\beta$ | Frequency | $\beta$ | Frequency | Mean | Median |
| 1 | 0.8 | 0.20 | 0.8 | 0.20 | 0.106 | 0.106 |
| 2 | 0.2 | 0.20 | 0.8 | 0.20 | 0.103 | 0.103 |
| 3 | 0.8 | 0.05 | 0.8 | 0.20 | 0.104 | 0.104 |
| 4 | 0.2 | 0.05 | 0.8 | 0.20 | 0.100 | 0.100 |
| 5 | 0.8 | 0.20 | 0.2 | 0.20 | 0.103 | 0.103 |
| 6 | 0.2 | 0.20 | 0.2 | 0.20 | 0.103 | 0.102 |
| 7 | 0.8 | 0.05 | 0.2 | 0.20 | 0.100 | 0.100 |
| 8 | 0.2 | 0.05 | 0.2 | 0.20 | 0.102 | 0.102 |
| 9 | 0.8 | 0.20 | 0.8 | 0.05 | 0.104 | 0.104 |
| 10 | 0.2 | 0.20 | 0.8 | 0.05 | 0.104 | 0.104 |
| 11 | 0.8 | 0.05 | 0.8 | 0.05 | 0.101 | 0.101 |
| 12 | 0.2 | 0.05 | 0.8 | 0.05 | 0.101 | 0.100 |
| 13 | 0.8 | 0.20 | 0.2 | 0.05 | 0.100 | 0.100 |
| 14 | 0.2 | 0.20 | 0.2 | 0.05 | 0.106 | 0.106 |
| 15 | 0.8 | 0.05 | 0.2 | 0.05 | 0.101 | 0.100 |
| 16 | 0.2 | 0.05 | 0.2 | 0.05 | 0.102 | 0.102 |

## 4.1 Simulation Results for Performance Measures

### 4.1.1 Correlation

We observed similar results in Pearson and Spearman correlations (Figure 4.1) of predicted risks using an alternative compared to the true model. With the naive modeling approach, the omitted risk factor's frequency has the largest impact. There is generally minor improvement when other risk factors have high prevalence. The weight of the omitted risk factor also affects correlation. Correlation is higher when the omitted risk factor has a small weight. The omitted risk factor's relationship with other predictors in the model does not has much impact on the correlation in the naive model. For the refit model, when the omitted risk factor is independent of other predictors the model performs slightly better when measured using Spearman correlation, but this effect is not seen when using Pearson correlation. In the Spearman correlation we see larger separation based on risk factor frequency.

### 4.1.2 C-Index

Similar to correlation, we saw model performance was highly impacted by the omitted risk factor's frequency and weight. To assess the overall model performance we looked at the C-index relative difference from the true model (Figure 4.2). When the omitted risk factor has a small weight, the frequency of other risk factors in the dataset is not relevant, but when the risk factor's weight is large some information can be gained if the present risk factors have high prevalence. The omitted risk factor's correlation with other predictors in the model did not impact the C-index.

### 4.1.3 IDI

Model performance can also be summarized with IDI (Figure 4.3). The weight of the omitted risk factor had the largest impact on IDI. When the simulated coefficient for the omitted risk factor was small there was minimal effect on IDI, but when the risk factor was large we saw a substantial decrease in IDI. This effect was modified by the frequency of the omitted risk factor, again showing that omitting a high frequency risk factor has a detrimental effect on model performance. The IDI also showed the model performed slightly worse if one of the remaining risk factors in the model had a large weight. The performance was not effected by the omited risk factor's correlation with other variables The refit model often only showed minor improvements, if any, compared to the naive model.

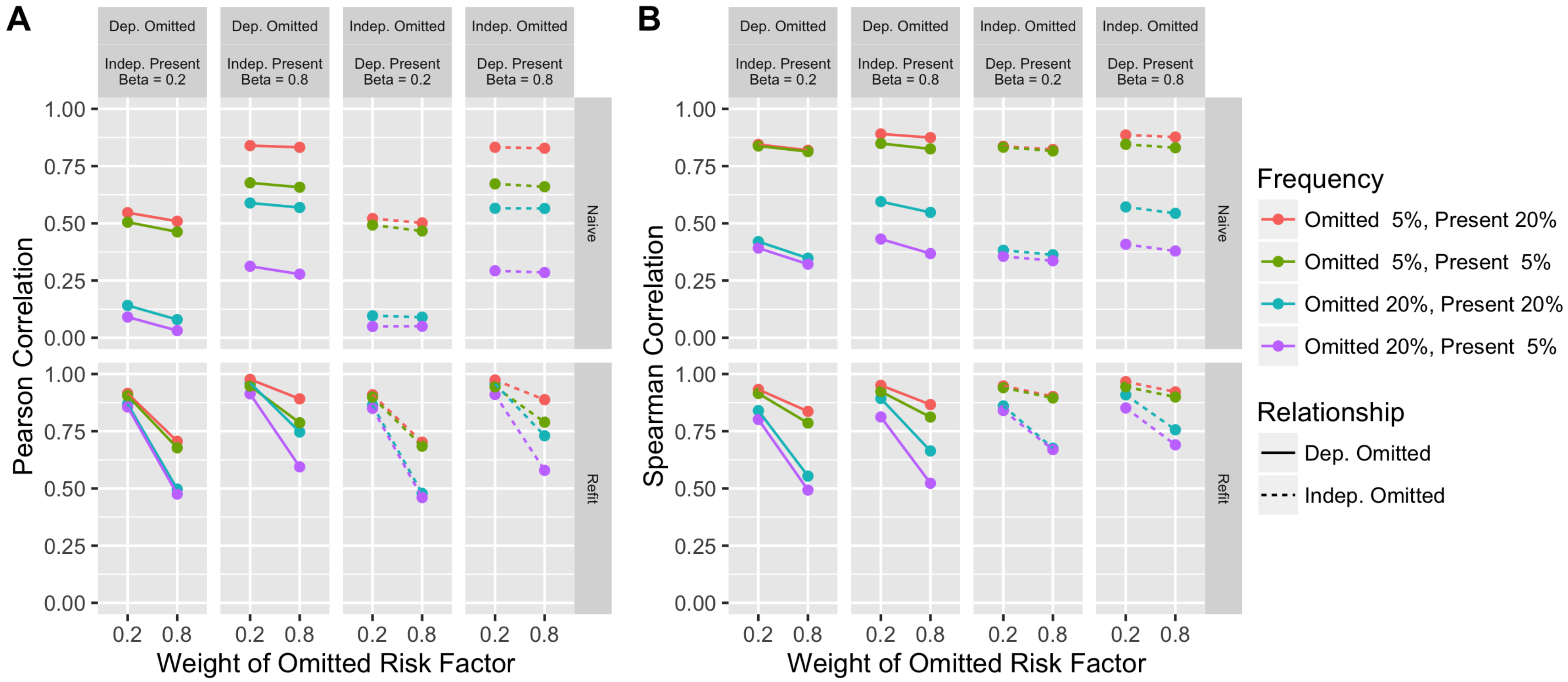

Figure 4.1: Mean Pearson (A) and Spearman (B) correlations for each model, where "Dep" denote $Z_2$ and "Indep" denote $Z_5$.

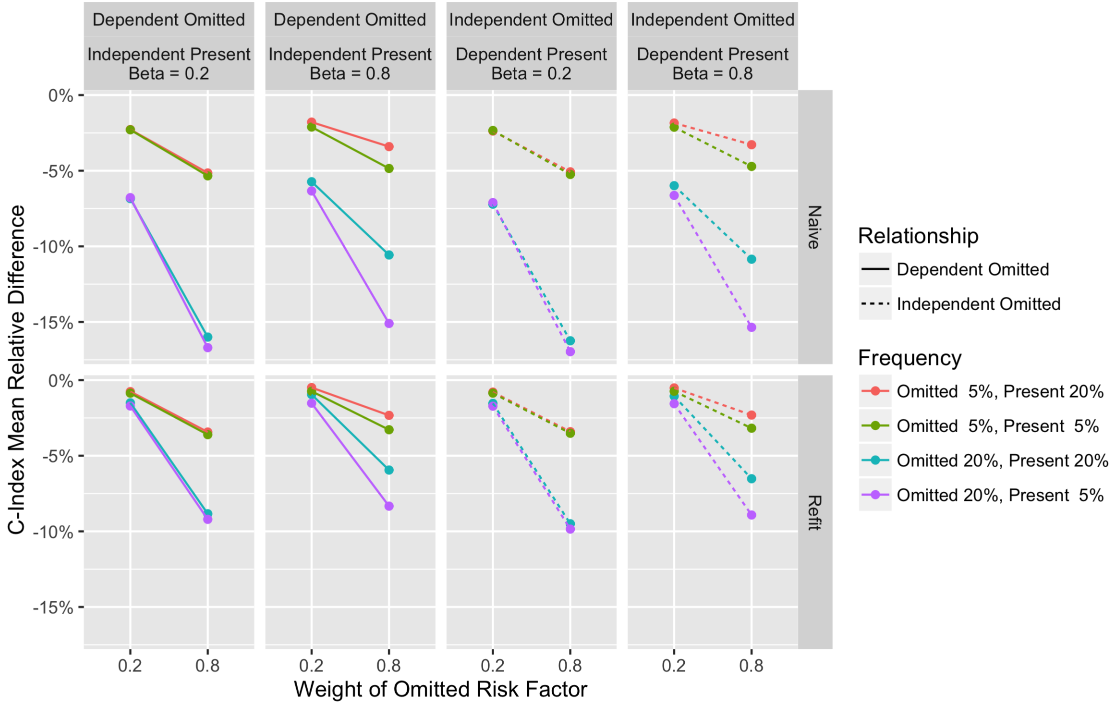

Figure 4.2: Mean C-index relative difference for each model, where “Dep” denote $Z_2$ and “Indep” denote $Z_5$.

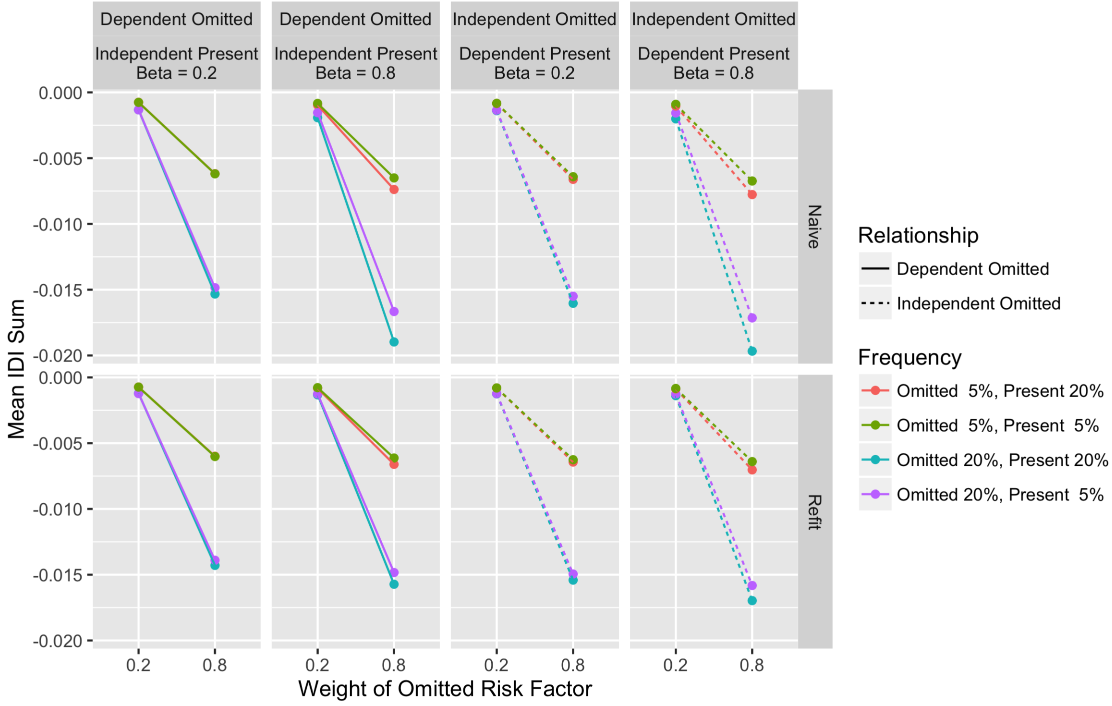

Figure 4.3: Mean IDI for each model, where “Dep” denote $Z_2$ and “Indep” denote $Z_5$.

### 4.1.4 Calibration

Calibration in the large should be 0 for a perfect model. All of the models performed relatively well, but performed best when the omitted risk factor had a large weight or low frequency (Figure 4.4A). Calibration slope should be 1 for a perfect model. We see a stark difference between the naive model and the refit model as measured by the calibration slope (Figure 4.4B). Similar to calibration in the large, the naive model performs best when the omitted risk factor has a large effect size or low frequency. The coefficients correlation do not appear to have very much of an effect. The calibration slope for the refit model indicates a perfect fit, but this is an artifact of using the same data to build the model and calibrate.

For calibration by risk group, the cohort was divided into 5 equally sized groups based on their linear predictor, with risk increasing from Group #1 to Group #5 (Figure 4.5). Groups #2 - #5 have negative coefficients, whereas Group #1 has substantially large positive coefficients. A well fit model would have equal coefficients for each risk group. This shows us that the naive model severely overestimates risk for inherently low-risk patients and underestimates for high-risk patients. This effect is exaggerated when the omitted risk factor has a high frequency. For most scenarios the model also performs worse when the omitted risk factor has a small weight.

### 4.1.5 Differences in Predicted Risk

The mean difference between the alternative and true model shows a similar pattern as we have seen in many of our results and informs us about the population dynamics (Figure 4.6.A). Simple arithmetic explains why the magnitude of the mean difference is related to the omitted risk factor's frequency and weight in the naive model. In the refit model we see the mean difference is 0 because the sum of the risk for the refit and true models is the same. For the average patient the refit model incorrectly estimates risk to a larger extent than the naive model. However, the maximum absolute difference shows that the naive model has the greatest opportunity to incorrectly estimate risk (Figure 4.6B). This metric shows the worst-case scenario for incorrectly estimating risk.

The absolute relative risk provides more insight as it does not allow the over- and under-estimation of individual risk in the refit model to cancel out (Figure 4.6C).

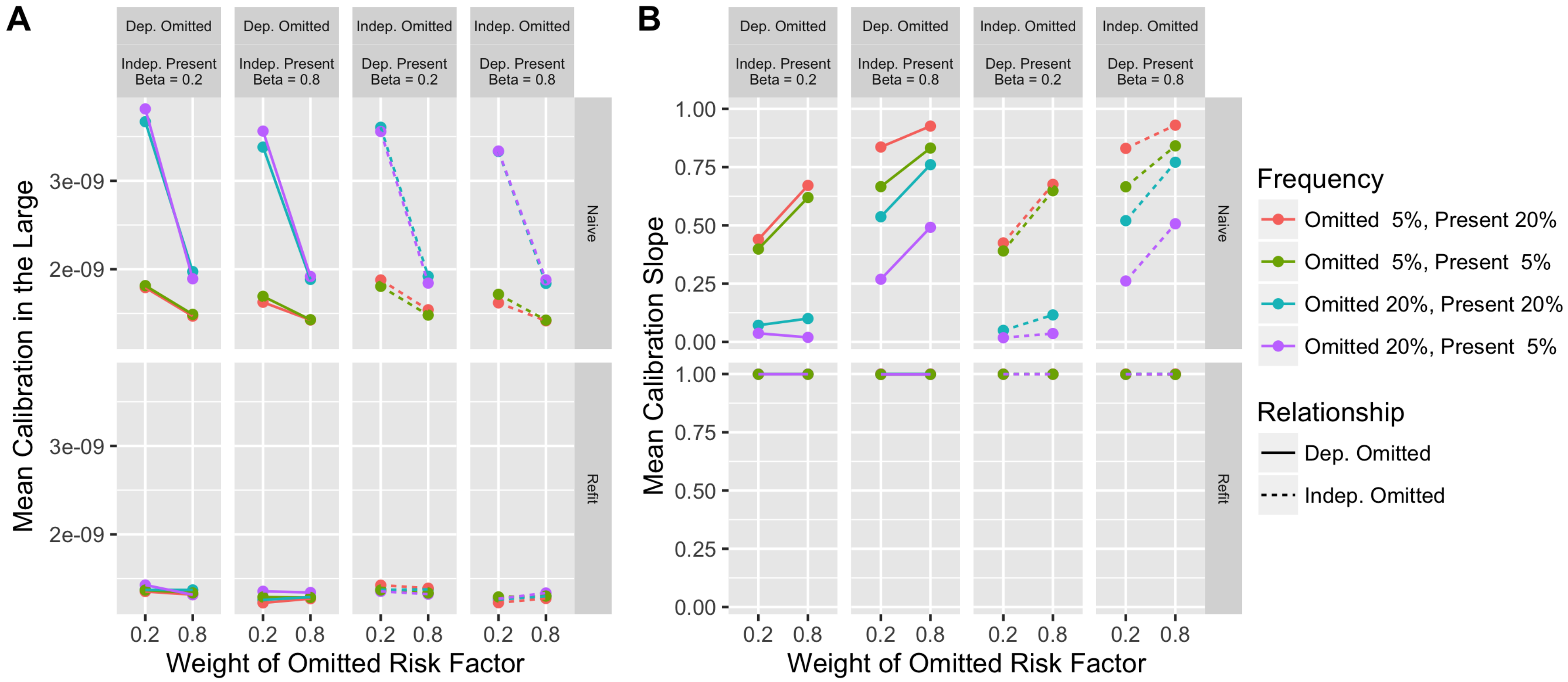

Figure 4.4: Calibration in the Large (A) and Calibration Slope (B)

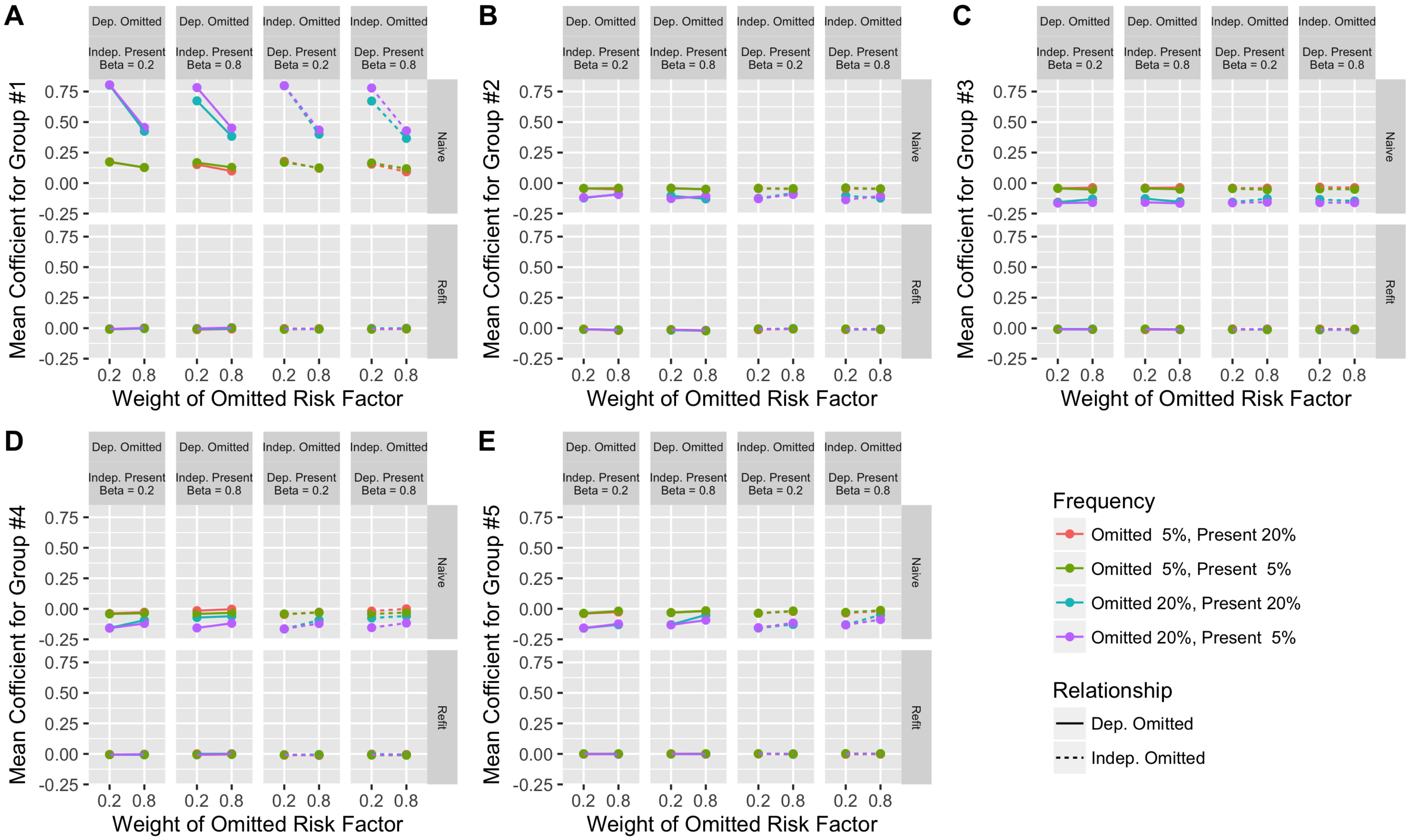

Figure 4.5: Calibration by Risk Group: Mean coefficient for Group #1 (A), Group #2 (B), Group #3 (C), Group #4 (D), and Group #5 (E)

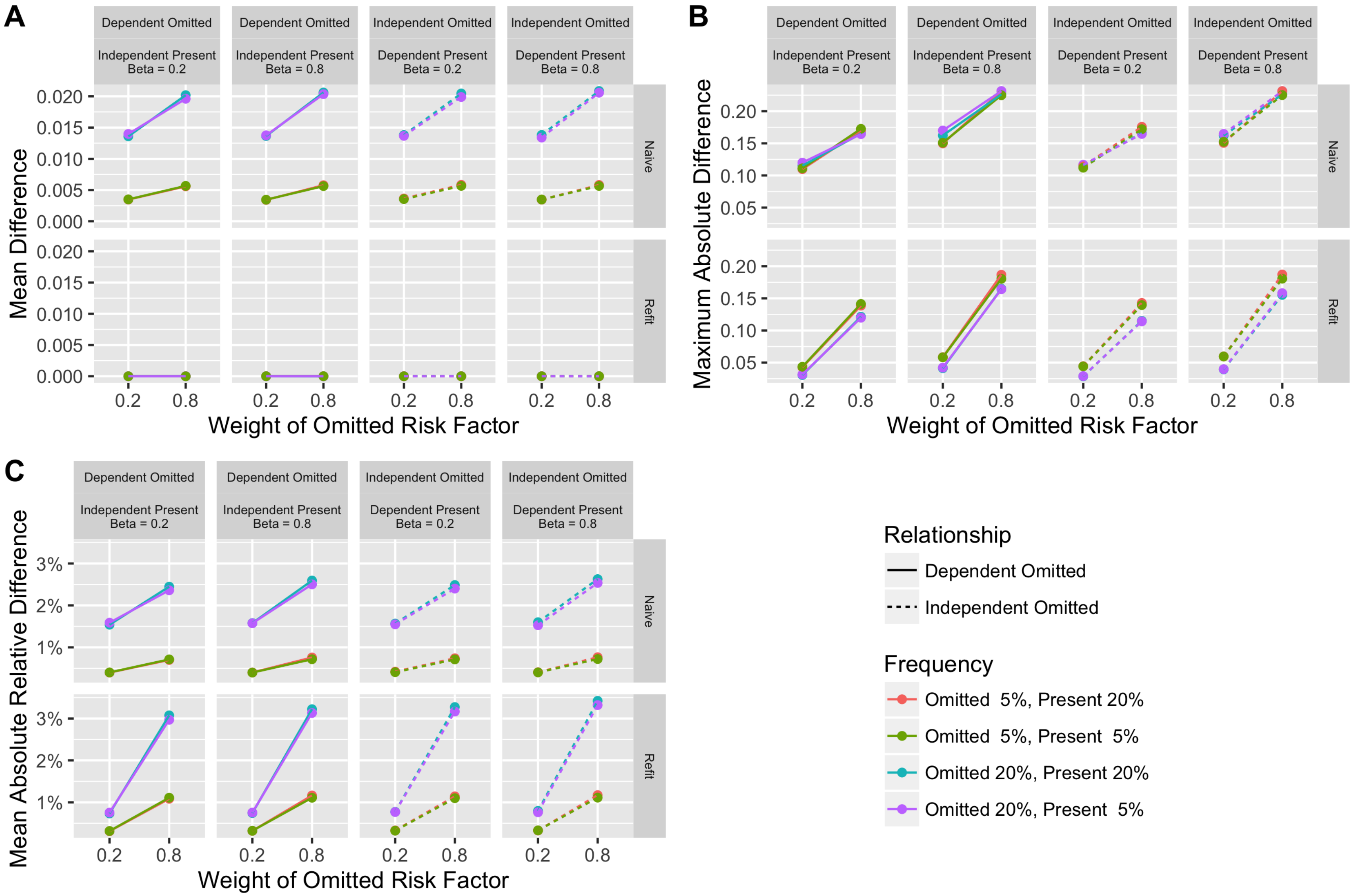

Figure 4.6: Differences in predicted risk between alternative and true models: Mean difference (A), Maximum absolute difference (B), and Mean Absolute Relative Difference (C)

# CHAPTER 5

# APPLICATION

## 5.1 Study Data for Framingham Stroke Risk Profile

Using the original data from the Framingham Study we built Cox proportional hazard regression models using all of the risk factors in attempt to replicate the work by Wolf et al. (1991). We built both naive and refit models to understand the influence of a single missing covariate.

Patients eligible for the study were members of the Framingham Study during examinations 9 or 14 cycles. Additionally patients must have been between 55 and 84 years old and free of stroke at baseline. Patients were followed up for 10 years. If a patient met the age requirements and was free of stroke at the time of both exam 9 and exam 14 the patient could appear in the dataset twice.

The original dataset published by Wolf et al. (1991) included 5734 patients (2372 men, 3362 women) and 472 stroke events over 10 years of followup. Our dataset contains 4918 patients (2064 men, 2854 women) and 342 stroke events. We suspect differences in the datasets are due to patients withdrawing consent after the paper had been published. For illustration purposes we limited our application to focus only on the model built for male patients. Model details are provided in Tables 5.1 and 5.3 below.

Table 5.1: Cox proportional hazards model fit

| | | Model Tests | | Discrimination Indexes | |
|---|---|---|---|---|---|
| Obs | 1992 | LR $\chi^2$ | 69.82 | $R^2$ | 0.052 |
| Events | 145 | d.f. | 8 | $D_{xy}$ | 0.382 |
| Center | 6.0601 | Pr$(>\chi^2)$ | 0.0000 | $g$ | 0.720 |
| | | Score $\chi^2$ | 79.27 | $g_r$ | 2.053 |
| | | Pr$(>\chi^2)$ | 0.0000 | | |

## 5.2 Performance Measure Comparisons

We assumed risk factors are uncollected one at a time and compared the naive approach and the refit approach using aforementioned performance measures. Regression coefficients obtained in the refit approach under each case of uncollected risk factor were provided in Table 5.2

Table 5.2: Coefficient comparison between true and refit models

| **Risk Factor** | **Wolf et. al** | **True Model** | **Refit Model for Removed Covariate** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Age | SBP | AntiHypT | Diabetes | Cigs | CVD | AFib | LVH |
| Age | 0.050 | 0.044 | | 0.046 | 0.045 | 0.044 | 0.040 | 0.047 | 0.044 | 0.044 |
| SBP | 0.014 | 0.021 | 0.021 | | 0.022 | 0.021 | 0.021 | 0.021 | 0.021 | 0.022 |
| Anti-Hypertensive Rx | 0.326 | 0.321 | 0.348 | 0.522 | | 0.328 | 0.284 | 0.330 | 0.325 | 0.332 |
| Diabetes | 0.338 | 0.139 | 0.141 | 0.210 | 0.162 | | 0.103 | 0.176 | 0.138 | 0.145 |
| Cigarettes | 0.515 | 0.408 | 0.320 | 0.419 | 0.384 | 0.401 | | 0.420 | 0.404 | 0.408 |
| CVD | 0.520 | 0.328 | 0.427 | 0.322 | 0.335 | 0.337 | 0.342 | | 0.340 | 0.350 |
| AFib | 0.606 | 0.212 | 0.340 | 0.211 | 0.242 | 0.209 | 0.160 | 0.316 | | 0.191 |
| LVH | 0.842 | 0.473 | 0.506 | 0.844 | 0.491 | 0.476 | 0.473 | 0.514 | 0.469 | |

Table 5.3: Cox proportional hazards model coefficients

| | Coefficient | S.E. | Wald $Z$ | Pr($>|Z|$) |
|---|---|---|---|---|
| Age | 0.0439 | 0.0120 | 3.66 | 0.0003 |
| Systolic Blood Pressure (SBP) | 0.0208 | 0.0040 | 5.22 | $< 0.0001$ |
| Anti-Hypertensive Rx | 0.3212 | 0.1913 | 1.68 | 0.0931 |
| Diabetes | 0.1390 | 0.2349 | 0.59 | 0.5540 |
| Cigarettes | 0.4085 | 0.1772 | 2.31 | 0.0211 |
| CVD | 0.3280 | 0.1835 | 1.79 | 0.0739 |
| Atrial Fibrillation (AFib) | 0.2121 | 0.4252 | 0.50 | 0.6179 |
| Left Ventricular Hypertrophy (LVH) | 0.4733 | 0.2849 | 1.66 | 0.0967 |

### 5.2.1 Correlation

Risk factors with higher frequencies (anti-hypertensive medication, CVD, cigarette smoking) show lower Pearson and Spearman correlations (Figure 5.1). There is not much improvement with the refit model.

### 5.2.2 C-Index

In general the C-indexes for the naive and refit models were very similar. Compared to the true model, discrimination was most harmed when an continuous predictor was omitted (Figure 5.2). Cigarette smoking was the categorical risk factor which caused the largest impact to C-index when removed from the model.

Table 5.4: C-Index for models by omitted predictor. C-Index for fake and refit models when each predictor is omitted. Full model C-Index = 0.6912

| | Naive | Refit |
|---|---|---|
| Age | 0.6787 | 0.6793 |
| Systolic Blood Pressure (SBP) | 0.6588 | 0.6636 |
| Anti-Hypertensive Rx | 0.6874 | 0.6871 |
| Diabetes | 0.6903 | 0.6903 |
| Cigarettes | 0.6808 | 0.6808 |
| CVD | 0.6841 | 0.6845 |
| Atrial Fibrillation (AFib) | 0.6902 | 0.6901 |
| Left Ventricular Hypertrophy (LVH) | 0.6880 | 0.6878 |

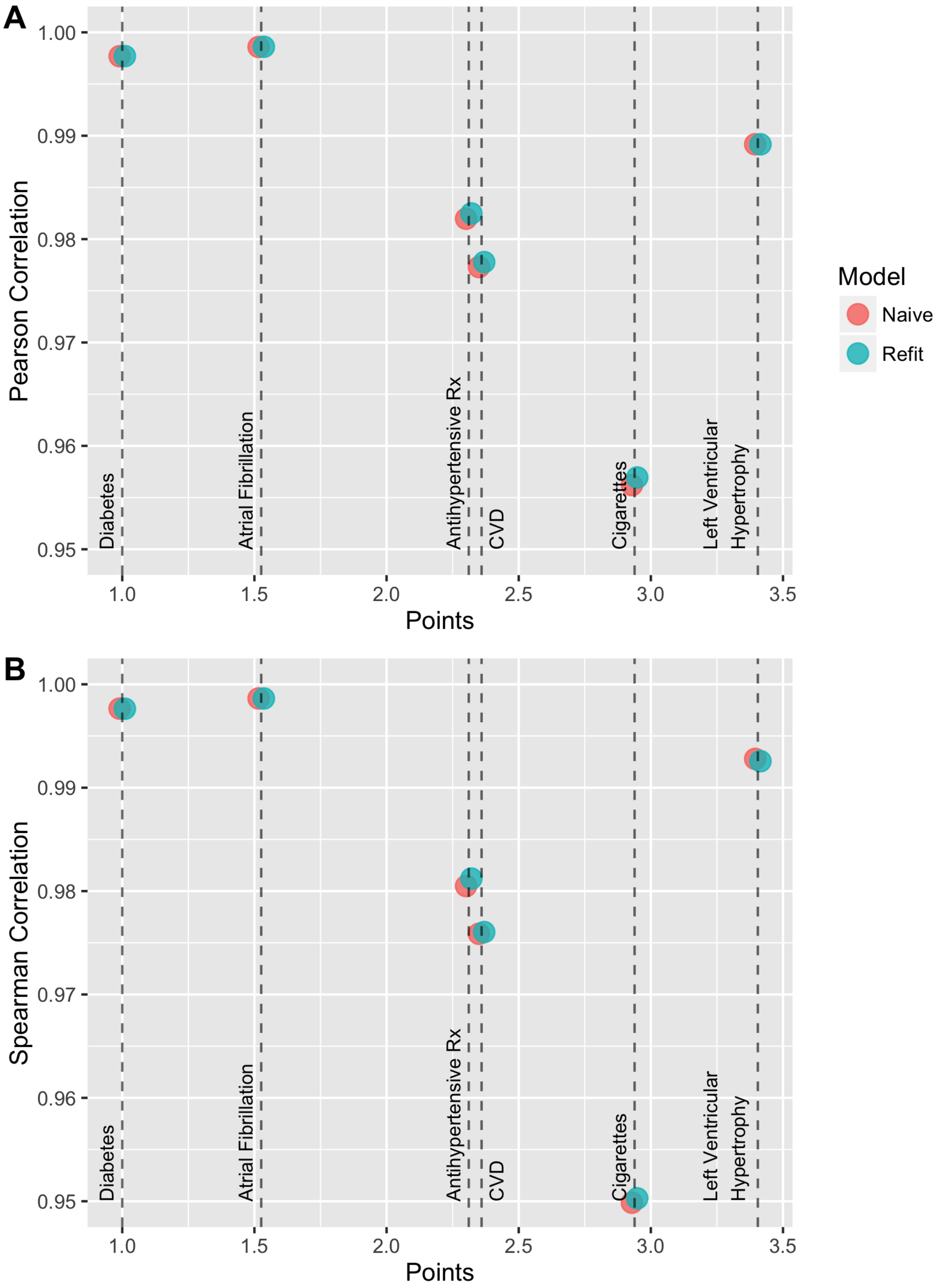

Figure 5.1: Pearson and Spearman correlation with true model

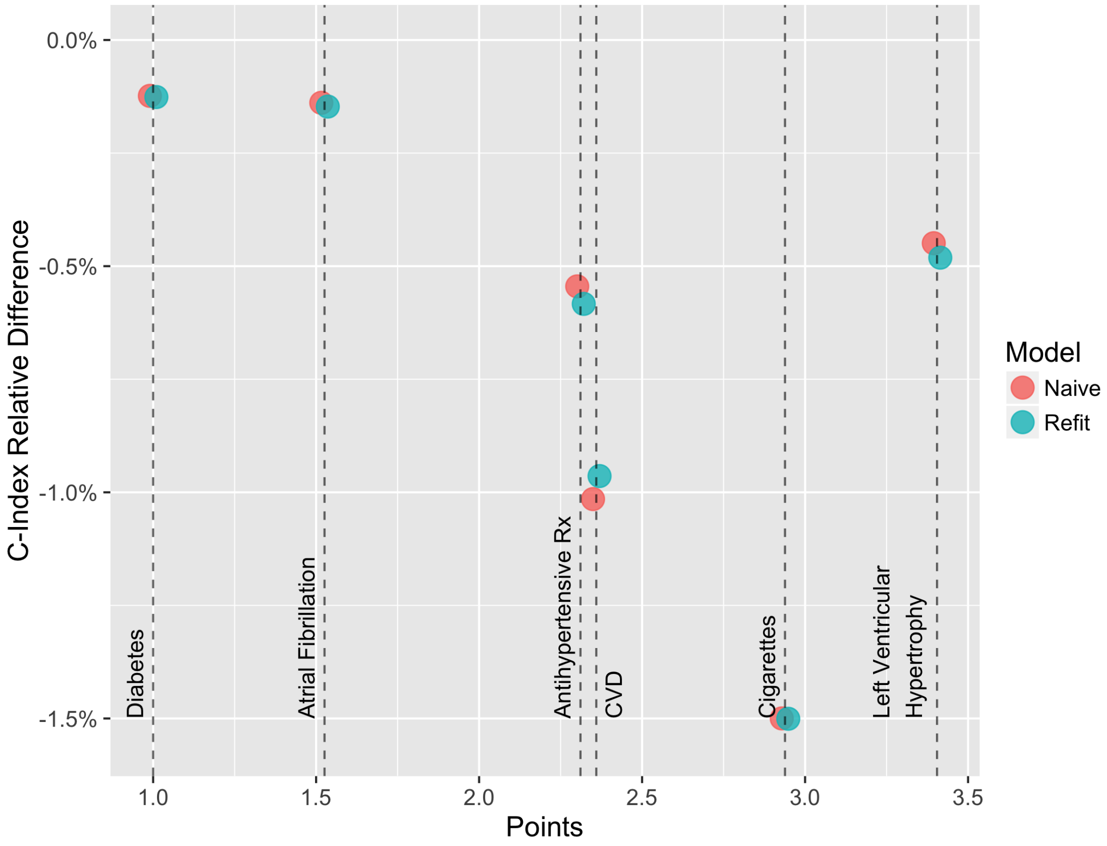

Figure 5.2: C-index relative difference from true model when each risk factor is removed

### 5.2.3 IDI

Changes in IDI were very small and do not provide meaningful information in this example.

### 5.2.4 Calibration

Calibration in the large is very close to 0 for the refit models (Figure 5.3A). This is due to using the same data the model was built on for testing. We may see different results if we separated the cohort into training and testing datasets. For calibration slope we see the refit model coefficients very close to one due to using the same data for training and testing (Figure 5.3B). In the naive model we see a largest impact from one of our continuous risk factors, systolic blood pressure. For categorical risk factors, we see the largest impact in the naive model when left ventricular hypertrophy is omitted. When separating the cohort into risk groups, we see differences in the coefficients for each omitted risk factor across risk groups (Table 5.7). This tells us

Table 5.5: Comparison of IDI of True Model to Naive and Refit Models when each predictor is omitted

| | **IDI** | | **IDI D1 (Cases)** | | **IDI D0 (Controls)** | |
|---|---|---|---|---|---|---|
| | Naive | Refit | Naive | Refit | Naive | Refit |
| Age | −0.040 | −0.009 | −0.106 | −0.008 | −0.066 | 0.000 |
| SBP | −0.041 | −0.018 | −0.107 | −0.017 | −0.066 | 0.001 |
| Anti-Hypertension Rx | −0.006 | −0.002 | −0.011 | −0.002 | −0.005 | 0.000 |
| Diabetes | −0.001 | 0.000 | −0.003 | 0.000 | −0.001 | 0.000 |
| Cigarettes | −0.005 | −0.002 | −0.013 | −0.002 | −0.008 | 0.000 |
| CVD | −0.006 | −0.002 | −0.012 | −0.001 | −0.006 | 0.000 |
| Atrial Fibrillation | 0.000 | 0.000 | −0.001 | 0.000 | −0.001 | 0.000 |
| LVH | −0.006 | −0.002 | −0.008 | −0.002 | −0.002 | 0.000 |

that the model does not accurately estimate risk across risk groups by underestimating high risk patients.

### 5.2.5 Differences in Predicted Risk

The average difference in a patients estimated individual risk when comparing the naive or refit models to the true models can be seen in Figure 5.4A. The difference is not only influenced by the risk factor's weight, but also by the frequency of the risk factor in the original dataset. Cigarette smoking, CVD, and antihypertensive medication were fairly common in the Framingham dataset, so we see ignoring those risk factors would be more detrimental to the population risk estimates than LVH even though LVH has the largest coefficient. The refit model shows almost no loss of information when estimating patient risk for the population.

In contrast to the mean difference, the maximum absolute difference shows how far off an individual estimate could be (Figure 5.4B). For an individual's estimate, both the naive and refit models are inaccurate. Here we are concerned with an individual patient's estimate, so the frequency of the risk factor in the sample is irrelevant. The risk factor with the largest weight will have the greatest impact. In the Framingham dataset, assuming LVH is negative when it is unknown will result in underestimating a patient's risk by 15% (Figure 5.4C).

Table 5.6: Calibration in the Large (true model: -0.00002) and Slope (true model: -0.00002, 0.99998)

| | **Fit 1 (Calibration in the Large)** | | **Fit 2 (Calibration Slope)** | | | |
|---|---|---|---|---|---|---|
| | Naive | Refit | Naive $\gamma_1$ | Refit $\gamma_1$ | Naive $\gamma_2$ | Refit $\gamma_2$ |
| Age | $-2.200e-05$ | $-2.157e-05$ | 0.09906966 | $-1.565e-05$ | 1.03902598 | 0.99998314 |
| SBP | $-1.785e-05$ | $-1.829e-05$ | 0.48371965 | $-1.383e-05$ | 1.17991291 | 0.99998382 |
| AntiHypT | $-2.292e-05$ | $-2.315e-05$ | $-0.01322052$ | $-1.619e-05$ | 1.03910522 | 0.99998343 |
| Diabetes | $-2.498e-05$ | $-2.503e-05$ | $-0.00288077$ | $-1.575e-05$ | 1.00688833 | 0.99997867 |
| Cigs | $-2.397e-05$ | $-2.374e-05$ | 0.00911311 | $-1.473e-05$ | 0.96942879 | 0.99997774 |
| CVD | $-2.534e-05$ | $-2.518e-05$ | $-0.01010989$ | $-1.608e-05$ | 1.03114294 | 0.99997825 |
| AFib | $-2.468e-05$ | $-2.472e-05$ | $-0.00202497$ | $-1.549e-05$ | 1.00467937 | 0.99997886 |
| LVH | $-2.548e-05$ | $-2.601e-05$ | $-0.01717795$ | $-1.243e-05$ | 1.04534289 | 0.99996823 |

Table 5.7: Crowson calibration by risk group: For the true model, the coefficients are -0.1998, 0.1351, -0.1391, 0.1947, -0.0720.

| | **Group #1** | | **Group #2** | | **Group #3** | | **Group #4** | | **Group #5** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Naive | Refit | Naive | Refit | Naive | Refit | Naive | Refit | Naive | Refit |
| Age | 0.0429 | 0.1461 | −0.6348 | −0.4109 | 0.0885 | −0.0617 | 0.1203 | 0.1543 | 0.0344 | 0.0028 |
| SBP | −0.2004 | 0.0122 | −0.1611 | −0.3211 | −0.0845 | −0.1052 | −0.0120 | 0.1691 | 0.1591 | 0.0299 |
| AntiHypT | −0.2652 | −0.3426 | −0.1533 | 0.0031 | 0.2238 | 0.1311 | −0.0563 | 0.0432 | 0.0198 | −0.0234 |
| Diabetes | −0.2102 | −0.2030 | 0.0675 | 0.0180 | −0.1423 | −0.0423 | 0.1908 | 0.1917 | −0.0462 | −0.0689 |
| Cigs | −0.6306 | −0.6790 | 0.3887 | 0.3667 | 0.0064 | −0.0527 | 0.0489 | 0.1096 | −0.0704 | −0.0692 |
| CVD | −0.8492 | −0.6282 | 0.0886 | 0.1143 | 0.2163 | 0.1580 | 0.0971 | 0.1004 | −0.0673 | −0.0710 |
| AFib | −0.3234 | −0.3172 | 0.1310 | 0.1881 | −0.0458 | −0.1398 | 0.1620 | 0.1655 | −0.0657 | −0.0535 |
| LVH | −0.3542 | −0.3099 | 0.0984 | 0.2400 | −0.1183 | −0.1997 | 0.1766 | 0.2082 | −0.0349 | −0.0824 |

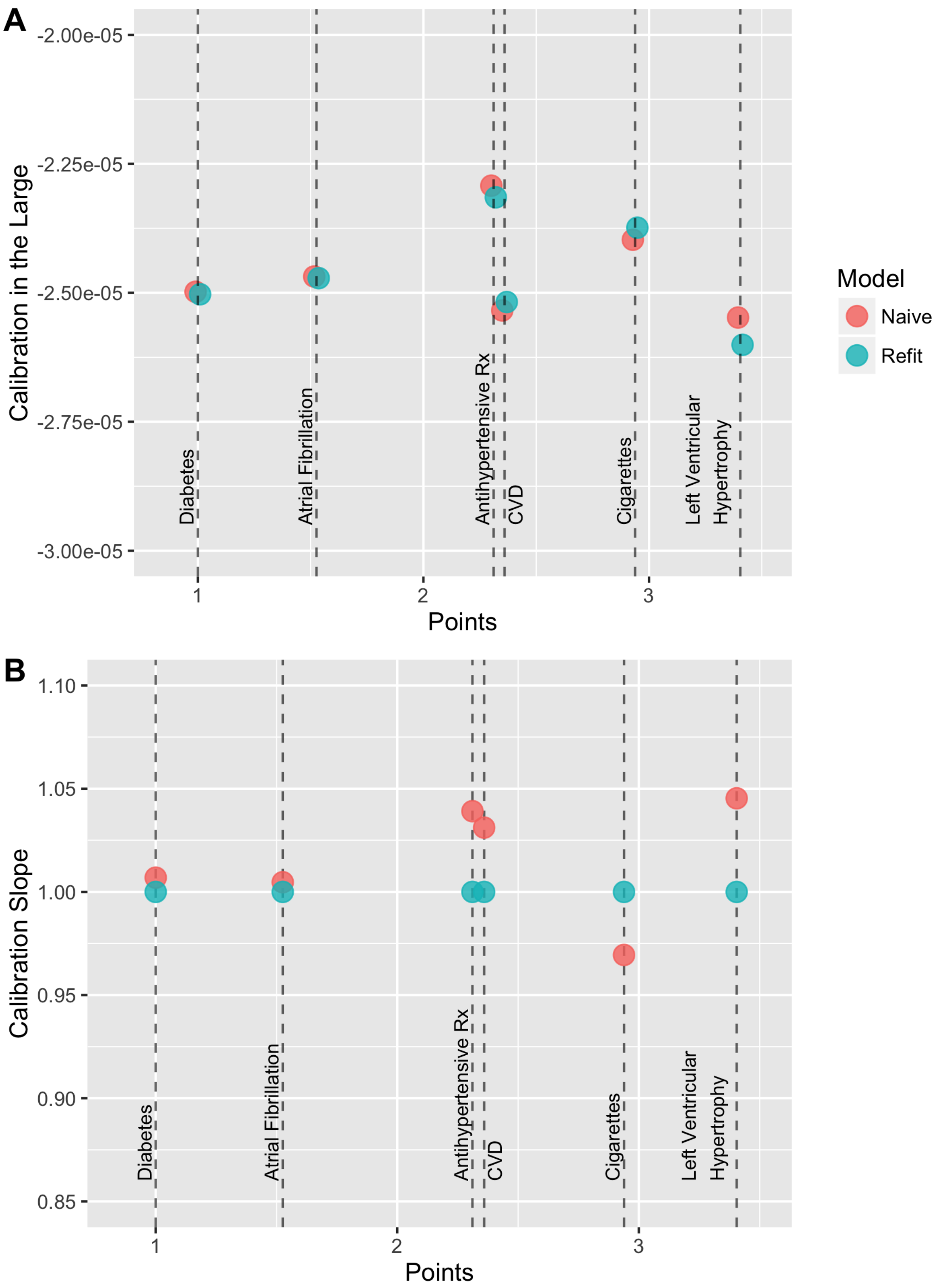

Figure 5.3: Calibration in the Large and Calibration Slope

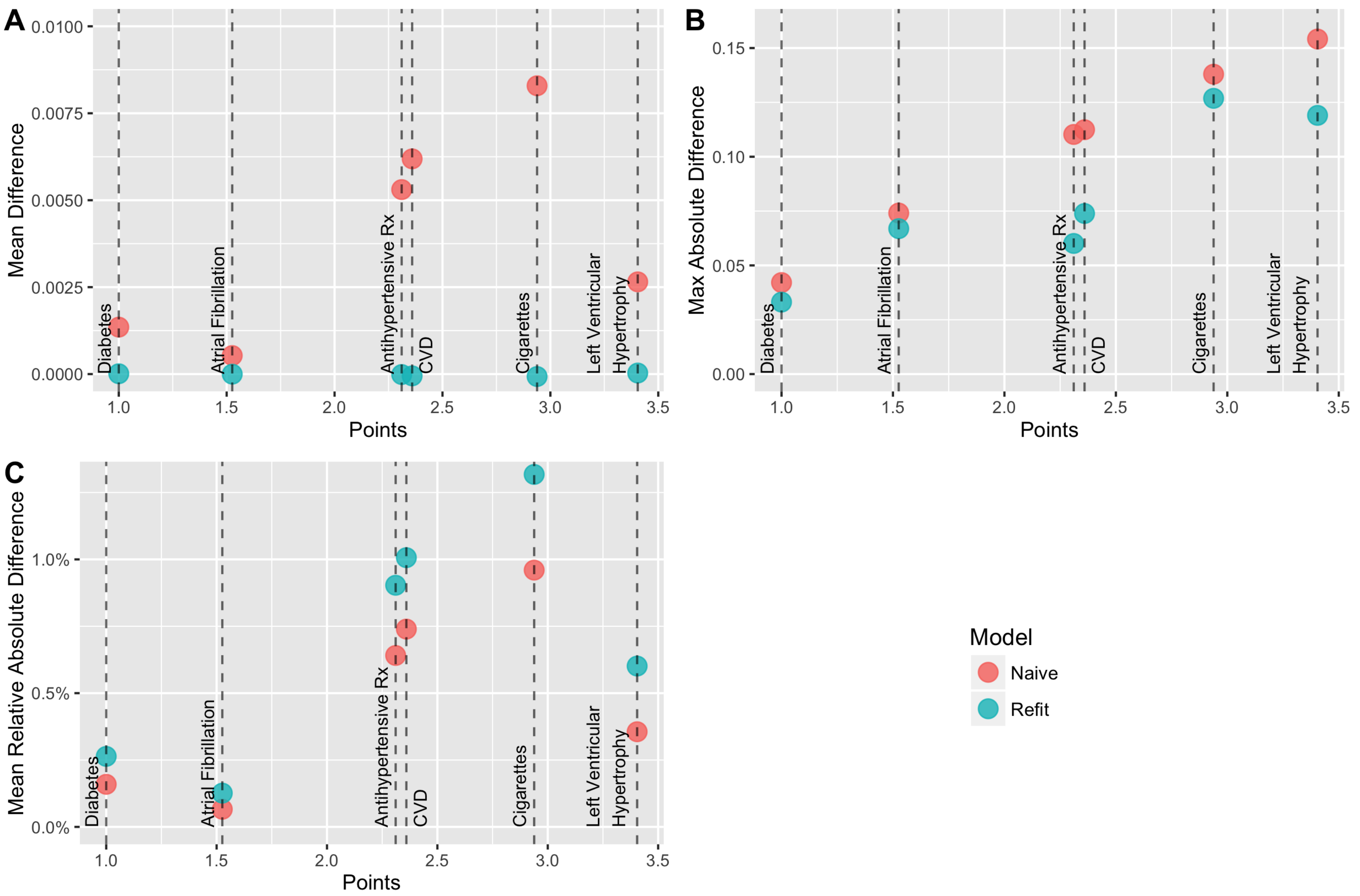

Figure 5.4: Differences in predicted risk between alternative and true models: Mean difference (A), Maximum absolute difference (B), and Mean Absolute Relative Difference (C)

# CHAPTER 6

# DISCUSSION

In this thesis we investigated alternative models for risk prediction for use when a risk factor is uncollected in the dataset. We considered two models: (1) use the same coefficients as the published model but omit the uncollected risk factor's component from the linear predictor; (2) refit the model without the risk factor effectively reassigning coefficients in the absence of the uncollected risk factor. To understand how these alternative model effect predictive ability and model performance we looked at them in terms of correlation, discrimination (C-index and IDI), calibration, and predicted risk. A simulation analysis examined how the uncollected risk factor's frequency, weight, and dependence with other risk factors affected the metrics. An application analysis using the Framingham Heart Study confirmed many of the simulation conclusions.

In general, our simulation analysis showed that the frequency of the uncollected risk factor had the largest impact on the overall performance of a revised risk score, followed by the weight of the uncollected risk factor in the original risk score. The uncollected risk factor's correlation with other risk factors in the model held little significance. These findings hold true regardless of whether the risk scores are being used as an adjusting variable, overall performance, or individual risk prediction.

When considering the global model performance, refitting the model in the absence of the uncollected risk factor allows us to retain the most information. However, if the model is to be used for individual risk prediction, refit models can have larger differences in predicted risk for patients. While on the average patient prediction is improved, the impact to a single patient may be very large.

Refitting the model provides an additional hurdle to the researchers. They must either have access to the original data, or have a secondary dataset separate from their analysis that includes all of the risk factors and outcomes needed for the model. Rather than refitting the model without the risk factor, although not studied as part of this thesis, an improved approach would be to identify an alternative risk factor that can be used as a "surrogate" measure. Such a measure would need to be correlated with the omitted risk factor and provide a similar level of information. There should be clinical knowledge to support such a substitution. For example, heavy alcohol consumption could be a surrogate for cigarette smoking because both are types of addiction. Although they target different organ systems, both are known to have harmful health effects.

# REFERENCES

Ankle Brachial Index Collaboration (2008), Ankle brachial index combined with framingham risk score to predict cardiovascular events and mortality: a meta-analysis, *JAMA: the journal of the American Medical Association* **300**(2), 197.

Cox, D. R. (1972), Regression models and life tables, *Journal of the Royal Statistical Society* **34**(2), 187–220.

Crowson, C. S., Atkinson, E. J. and Therneau, T. M. (2016), Assessing calibration of prognostic risk scores, *Statistical methods in medical research* **25**(4), 1692–1706.

Dawber, T. R. and Moore, F. E. (1952), Longitudinal study of heart disease in framingham, massachusetts: an interim report, *in* 'Research in Public Health, Papers presented at the 1951 Annual Conference of the Milbank Memorial Fund', 241–247.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., Rosati, R. A. et al. (1982), Evaluating the yield of medical tests, *Jama* **247**(18), 2543–2546.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996), Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in medicine* **15**(4), 361–387.

Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013), *Applied logistic regression*, Vol. 398, John Wiley & Sons.

Kamath, P. S., Wiesner, R. H., Malinchoc, M., Kremers, W., Therneau, T. M., Kosberg, C. L., D'Amico, G., Dickson, E. R. and Kim, W. (2001), A model to predict survival in patients with end-stage liver disease, *Hepatology* **33**(2), 464–470.

Llewellyn, D. J., Lang, I. A., Xie, J., Huppert, F. A., Melzer, D. and Langa, K. M. (2008), Framingham stroke risk profile and poor cognitive function: a population-based study, *BMC Neurology* **8**(1), 12.

Pencina, M. J., D'Agostino, R. B. and Vasan, R. S. (2008), Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond, *Statistics in medicine* **27**(2), 157–172.

Sara, J. D., Lennon, R. J., Gulati, R., Singh, M., Holmes, D. R., Lerman, L. O. and Lerman, A. (2016), Utility of the framingham risk score in predicting secondary events in patients following percutaneous coronary intervention: A time-trend analysis, *American heart journal* **172**, 115–128.

Towfighi, A., Markovic, D. and Ovbiagele, B. (2012), Utility of framingham coronary heart disease risk score for predicting cardiac risk after stroke, *Stroke* **43**(11), 2942–2947.

Wolf, P. A., D'agostino, R. B., Belanger, A. J. and Kannel, W. B. (1991), Probability of stroke: a risk profile from the framingham study., *Stroke* **22**(3), 312–318.