

# NHANES Blood Pressure-Based Mortality Risk - Appendix

Rscripts by Hamish Patten, DW Bester and David Steinsaltz

20/09/2023

## Contents

<b>Appendix A – Exploring the data</b>	<b>1</b>
Summarizing the data . . . . .	1
Addressing errors in blood pressure measurement . . . . .	2
Last digit preference . . . . .	2
Pseudo-replicates . . . . .	4
Missing or implausible measurements . . . . .	4
<b>Appendix B – Model details</b>	<b>5</b>
Survival Analysis (Time-to-Event) . . . . .	5
Longitudinal Modelling . . . . .	6
Combined Hierarchical Model . . . . .	7
The modelling variants . . . . .	7
Methodology . . . . .	8
Empirical Bayes Parameters . . . . .	11
Hamiltonian Monte Carlo (HMC) . . . . .	16
Centering the Linear Predictor . . . . .	17
Code Description . . . . .	18
Data Cleaning . . . . .	18
Main . . . . .	18
Stan . . . . .	18
Centering . . . . .	18
Empirical Bayes Estimation . . . . .	18
Post-processing . . . . .	19
Results . . . . .	19
Convergence of Simulations . . . . .	19
Results - Model Parameterization . . . . .	21
Results - Model Performance . . . . .	22
Results - Exploring $\Delta$ Directionality . . . . .	27

## Appendix A – Exploring the data

### Summarizing the data

The empirical means of the home and clinic measures in population B are tabulated in Table 1. We note that the home measures are systematically higher than the clinic measures, within every demographic group, with greater differences for subjects who are white or Mexican, and female. The average difference is about 2.3 for diastolic and 2.7 for systolic, which is small compared with the general range of the differences, which have SD of 10.1 (diastolic) and 14.0 (systolic).

[1] TRUE [1] TRUE There were 19592 subjects in the initial data set. Of these 4573 were excluded because they had missing data or were not followed up, or belonged to the “Other” ethnic group. This left 15019 subjects for further consideration. A small number of subjects were excluded because their blood pressure measurements were outside the normal range, as described below in section @ref{sec:BPrange}.

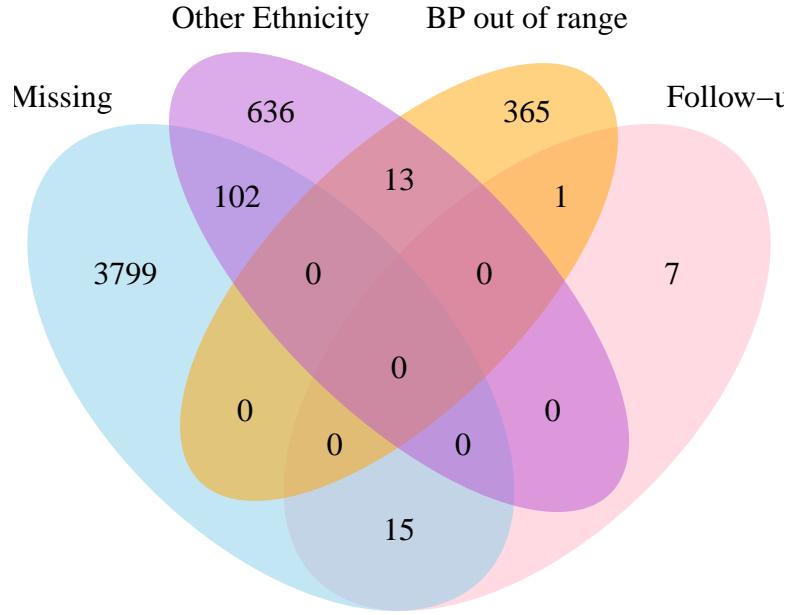


Figure 1: Venn diagram of subjects excluded from the analysis.

We note that the home measures are systematically higher than the clinic measures, within every demographic group, with greater differences for subjects who are white or Mexican, and female. The average difference is about 2.3 for diastolic and 2.7 for systolic, which is small compared with the general range of the differences, which have SD of 10.1 (diastolic) and 14.0 (systolic).

We see that there is relatively little correlation between empirical SD and empirical mean SD for the different BP types and places. This is reassuring, as it avoids the possibility of a collinearity effect confounding the sampling of mean and SD, which are being treated as independent covariates in the model.

## Addressing errors in blood pressure measurement

The blood pressure errors were found in the at home measurements. While these did not destroy the usefulness of the home measurements, they did require some attention and decisions for how to work with these defects. We also consider them inherently interesting, and worth registering for future researchers working on these or similar data. In particular, the problem we have called “pseudo-replication” was entirely unexpected, although not unprecedented, and is of particular concern to researchers trying to estimate individual variation in clinically relevant measures.

### Last digit preference

Mild tendency for observers to prefer certain last digits in reporting BP measurements has been reported in other studies, though an analysis of the 1999 wave of NHANES reported no last-digit preference. The last-digit preference in NHANES III, on the other hand, is substantial, with about 26% of all the diastolic BP measurements ending in 0, but only about 30% ending in 4 or 6. Because the shifts due to last-digit preference are presumably small, we expect them to have little effect on the main effects that we are examining in this paper, but they do increase the probability of two measurements being rounded to the same value, something that needs to be taken into account in examining the problem of pseudo-replication.

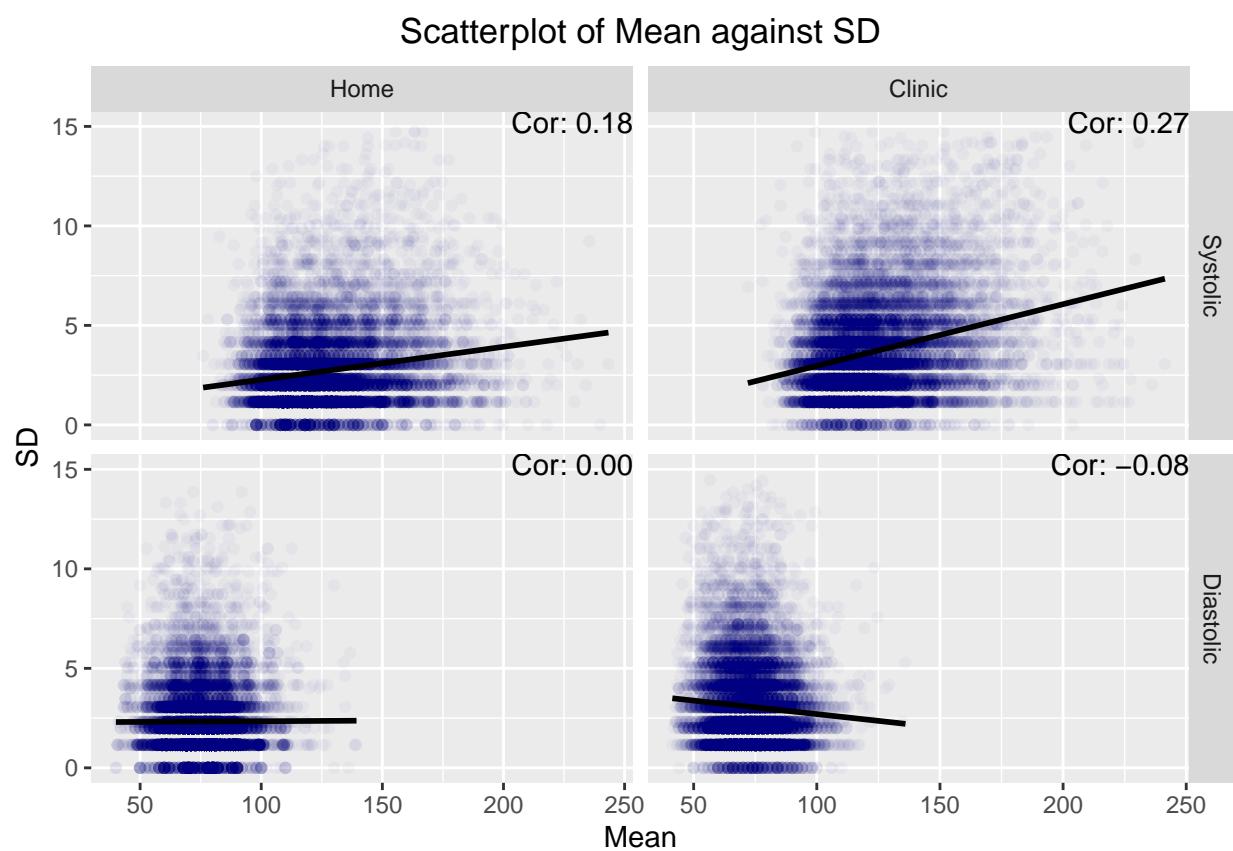


Figure 2: Scatterplot of individual mean BP against individual SD of BP

Table 1: Summary data for blood pressure

Sys/Dias	Place	Sex	Ethnicity	Mean	SD
Systolic	Home	Male	Black	128.8	2.7
Systolic	Home	Male	White	131.2	2.8
Systolic	Home	Male	Mexican	127.1	2.5
Systolic	Home	Female	Black	123.2	2.7
Systolic	Home	Female	White	127.2	2.9
Systolic	Home	Female	Mexican	121.2	2.7
Systolic	Clinic	Male	Black	127.7	3.5
Systolic	Clinic	Male	White	128.0	4.2
Systolic	Clinic	Male	Mexican	123.5	3.5
Systolic	Clinic	Female	Black	122.4	3.6
Systolic	Clinic	Female	White	123.5	4.1
Systolic	Clinic	Female	Mexican	117.8	3.4
Diastolic	Home	Male	Black	78.7	2.4
Diastolic	Home	Male	White	77.1	2.3
Diastolic	Home	Male	Mexican	76.8	2.4
Diastolic	Home	Female	Black	75.0	2.4
Diastolic	Home	Female	White	73.7	2.3
Diastolic	Home	Female	Mexican	72.0	2.4
Diastolic	Clinic	Male	Black	77.8	3.1
Diastolic	Clinic	Male	White	75.4	3.1
Diastolic	Clinic	Male	Mexican	74.9	3.2
Diastolic	Clinic	Female	Black	72.4	3.0
Diastolic	Clinic	Female	White	70.5	3.0
Diastolic	Clinic	Female	Mexican	69.2	3.0

### Pseudo-replicates

While the protocol calls for each subject to have three independent BP measures taken, it appears from looking at the distribution of the results that this is not what happened in all cases, at least in the home measurements. Instead, it may be that one or two measurements were taken correctly, and the remaining “observations” were either copied in, or matched the earlier results because of the observer’s unconscious inclination to expect the same number to come up. We analyze this problem at length below, two out of twelve examiners who collected the data may have committed some amount of copying (whether intentional or inadvertent), potentially affecting approximately 16% of the subjects. This is a significant problem, because copying affects the variance of the reported measurements, shifting it consistently downward, to an extent that may be very large, and is impossible to reconstruct. The effect is to create a class of spurious low-variance individuals that dilute the effect of variance on survival.

### Missing or implausible measurements

Some of the reported measures were extremely implausible, particularly for diastolic BP. NAsubjects had at least one diastolic BP measure recorded as 0, in addition to the 3916 subjects who were missing at least one measurement. We excluded all of these subjects, and indeed any subject who had at least one measurement recorded outside the ranges (40,140) for diastolic and (60,250) for systolic BP, as recommended by the CDC. There was just one subject with systolic BP measures that were too low, but NA subjects with low diastolic BP (in addition to those with measures recorded as 0). One subject was excluded for diastolic BP 156, and three were excluded for systolic BP that was too high, with the maximum being 264.

## Appendix B – Model details

This appendix aims to add more detail about the numerical modelling than was provided in the article. This is to ensure that the research methods are transparent and entirely reproducible. The numerical modelling presented in this paper was performed using R combined with Rstan. More detail will be provided here about the model, about the specific methodology used to parameterize the model, and more results are provided that were not included in the main text.

The model used in this research is built from the theory of joint modelling of longitudinal and time-to-event data. This will be described in detail later on in this section, however, in brief, this allows the simultaneous modelling of both longitudinal observation data (in this article, this is blood pressure measurements) and also the time-to-event outcome. In this research the event of interest is either death from any cause, or death from specifically heart attack, cardiovascular disease (CVD) or cerebrovascular disease (CeVD). We henceforth will refer to this latter mortality as HA-CVD-CeVD. In the latter case, death from a different cause is treated as a noninformative censoring event.

### Survival Analysis (Time-to-Event)

The basic survival model is a Gompertz hazard rate with proportional hazards influences of the blood pressure covariates. The Gompertz equation

$$h_0(t) = B \exp(\theta(x + T)), \quad (1)$$

describes the baseline hazard of the population to a particular risk, which, for this article, investigates HA-CVD-CeVD mortality specifically, as well as studying mortality risk in general.  $x \in \mathbb{N}^N$  is the age of the individual at the initial interview time, for  $N$  the number of individuals, and  $T \in \mathbb{R}^{+,N}$  the time since the individual entered the survey. Note that both  $B$  and  $\theta$  have 6 different values, depending on the sex reported at the initial interview — female or male — or the race — black, white or ‘other’. Note that ‘other’ in the race category is a combination of all non-black or non-white racial identities, such as Hispanic populations. The log-linear proportional hazards model links the covariates of the model (mean systolic blood pressure, variance in the diastolic blood pressure, etc) to the survival outcome of the individual via the equation

$$h(t) = h_0(t) \exp(\boldsymbol{\beta} \cdot (\mathbf{X} - \hat{\mathbf{X}})), \quad (2)$$

where  $\mathbf{X} \in \mathbb{R}^{+,N \times d}$  is a vector of summary statistics of the blood pressure measurements of individual covariates in our model,  $\hat{\mathbf{X}} \in \mathbb{R}^{+,d}$  is the centering of the covariates such that the equation  $\sum_i^N \exp(\boldsymbol{\beta} \cdot (\mathbf{X} - \hat{\mathbf{X}})) = 0$  is approximately satisfied (more on this later), and  $\boldsymbol{\beta} \in \mathbb{R}^d$  implies the strength of the influence of the covariate on the mortality risk. The majority of mortality events are censored — not yet known at the time of data collection — the censoring indicator being notated as  $\delta \in \{0, 1\}$ . When HA-CVD-CeVD mortality is the event being analysed, deaths due to other causes are treated as noninformative censoring events. In this study, we explored the following covariates:

Variable Name	Support	Description
$FRS - 1998$	$R^N$	1998 version of the FRS score
$FRS - ATP$	$R^N$	ATP version of the FRS score
$M_S$	$R^{+,N}$	Mean systolic blood pressure
$M_D$	$R^{+,N}$	Mean diastolic blood pressure
$\Delta_S$	$R^{+,N}$	Difference between Home and Clinic mean systolic blood pressure
$\Delta_D$	$R^{+,N}$	Difference between Home and Clinic mean diastolic blood pressure
$\sigma_{\{S,H\}}$	$R^{+,N}$	Standard deviation of the systolic blood pressure taken at home
$\sigma_{\{D,H\}}$	$R^{+,N}$	Standard deviation of the diastolic blood pressure taken at home
$\sigma_{\{S,C\}}$	$R^{+,N}$	Standard deviation of the systolic blood pressure taken at the clinic
$\sigma_{\{D,C\}}$	$R^{+,N}$	Standard deviation of the diastolic blood pressure taken at the clinic
$\tau_{\{S,H\}}$	$R^{+,N}$	Precision of the systolic blood pressure taken at home
$\tau_{\{D,H\}}$	$R^{+,N}$	Precision of the diastolic blood pressure taken at home
$\tau_{\{S,C\}}$	$R^{+,N}$	Precision of the systolic blood pressure taken at the clinic
$\tau_{\{D,C\}}$	$R^{+,N}$	Precision of the diastolic blood pressure taken at the clinic

Please note that the last four elements of this list, the precision values, were only carried out to ensure model consistency with the use of standard deviation instead. Note as well that the  $\Delta$  covariates, representing the medium-term variability, enter into the log relative risk sum as an **absolute value**.

For the parametrization of this model, we assume that the Gompertz parameters and the parameters in the linear predictor term are distributed as follows:

$$\begin{aligned} \boldsymbol{B} &\sim \mathbb{C}(\mu_B, \sigma_B), \\ \boldsymbol{\theta} &\sim \mathfrak{N}(\mu_\theta, \sigma_\theta), \\ \boldsymbol{\beta} &\sim \mathfrak{N}(\mu_\beta, \sigma_\beta), \end{aligned} \tag{3}$$

noting that  $\mathbb{C}(\mu, \sigma)$  is the Cauchy distribution.

The likelihood for this Gompertz proportional hazards model, over all individuals in the census, is as follows:

$$L_S(\boldsymbol{v}, \boldsymbol{\delta}) = \prod_i^N f(v_i, \delta_i | B_i, \theta_i, \beta_i, \mathbf{X}, \hat{\mathbf{X}}) = \prod_i^N h(v_i | B_i, \theta_i, \beta_i, \mathbf{X}, \hat{\mathbf{X}})^{\delta_i} \exp \left( - \sum_i^N H(v_i | B_i, \theta_i, \beta_i, \mathbf{X}, \hat{\mathbf{X}}) \right), \tag{4}$$

with  $H(v) = \int_0^v h(w)dw$  the cumulative hazard.

### Longitudinal Modelling

The mortality hazard rates are assumed to be influenced by individual-level blood pressure means and variability characteristics. These characteristics are not directly observed, but are inferred from their influence on the individual blood pressure measurements, which have been observed. Let  $Y_i(t_j)$  be the observed blood pressure for patient  $i$  at time  $t_j$ , for the individual  $i \in 1, 2, \dots, N$  and the number of blood pressure measurements per individual  $j \in 1, 2, \dots, k$ . Due to the fact that the blood pressure measurement data was taken at both the home and clinic (written using subscripts H and C, respectively), with approximately 6 months between these two measurements, we model the blood pressure using the following model, assuming the diastolic  $Y_i^D$  and systolic  $Y_i^S$  blood pressure to be Gaussian-distributed:

$$\begin{aligned} (Y_i^D)_H &\sim \mathfrak{N}(M_i^D + \Delta_i^D, (\sigma_i^D)_H), \\ (Y_i^D)_C &\sim \mathfrak{N}(M_i^D - \Delta_i^D, (\sigma_i^D)_C), \\ (Y_i^S)_H &\sim \mathfrak{N}(M_i^S + \Delta_i^S, (\sigma_i^S)_H), \\ (Y_i^S)_C &\sim \mathfrak{N}(M_i^S - \Delta_i^S, (\sigma_i^S)_C), \end{aligned} \tag{5}$$

where superscripts  $D$  and  $S$  refer to diastolic and systolic blood pressure, respectively.

The blood pressure characteristics — the individual-level parameters — are themselves distributed according to a hierarchical model, determined by population-level parameters (also called “hyperparameters’ ’):

$$\begin{aligned} M_i^{\{D,S\}} &\sim \mathcal{N}(\mu_M^{\{D,S\}}, \sigma_M^{\{D,S\}}), \\ \Delta_i^{\{D,S\}} &\sim \mathcal{N}(\mu_D^{\{D,S\}}, \sigma_D^{\{D,S\}}), \\ \sigma_{i,C}^{\{D,S\}} &\sim \Gamma(r_C^{\{D,S\}}, \lambda_C^{\{D,S\}}), \\ \sigma_{i,H}^{\{D,S\}} &\sim \Gamma(r_H^{\{D,S\}}, \lambda_H^{\{D,S\}}). \end{aligned} \quad (6)$$

The longitudinal outcome modelling therefore aims to infer these hyperparameters

$$\Theta = \left\{ \mu_M^{\{D,S\}}, \mu_D^{\{D,S\}}, \sigma_M^{\{D,S\}}, \sigma_D^{\{D,S\}}, r_C^{\{D,S\}}, \lambda_C^{\{D,S\}}, r_H^{\{D,S\}}, \lambda_H^{\{D,S\}} \right\}, \quad (7)$$

and to use the implied uncertainty about the individual-level parameters to inform the inference about the survival parameters. The likelihood for the longitudinal measurements is therefore (combining the systolic and diastolic into a single parameter for simplicity):

$$L_L(\Theta|Y) = \prod_{i=1}^N \left( \prod_{j=1}^k f(y_{ij}|M_i, \Delta_i, \sigma_i) \right) f(M_i|\mu_M, \sigma_M) f(\Delta_i|\mu_D, \sigma_D) f(\tau_{i,C}|r_C, \lambda_C) f(\tau_{i,H}|r_H, \lambda_H) \quad (8)$$

### Combined Hierarchical Model

Combining the longitudinal outcome and time-to-event partial likelihoods, and for a given parameter space value of  $\Omega = \{\beta, B, \theta\} \cup \Theta$ , the joint likelihood is

$$\begin{aligned} L(\Omega|Y) = \prod_{i=1}^N \left( \prod_{j=1}^k f(y_{ij}|M_i, \Delta_i, \sigma_i) \right) &f(v_i, \delta_i|B_i, \theta_i, \beta_i, \mathbf{X}, \hat{\mathbf{X}}) f(M_i|\mu_M, \sigma_M) \\ &f(\Delta_i|\mu_D, \sigma_D) f(\tau_{i,C}|r_C, \lambda_C) f(\tau_{i,H}|r_H, \lambda_H). \end{aligned} \quad (9)$$

One approach to estimating the complete set of hyperparameters

$$\Omega_H = \{\mu_B, \sigma_B, \mu_\theta, \sigma_\theta, \mu_\beta, \sigma_\beta, \mu_M^{\{D,S\}}, \sigma_M^{\{D,S\}}, \mu_D^{\{D,S\}}, \sigma_D^{\{D,S\}}, r_C^{\{D,S\}}, \lambda_C^{\{D,S\}}, r_H^{\{D,S\}}, \lambda_H^{\{D,S\}}\} \quad (10)$$

is to impose a higher-level prior distribution, and use the machinery of Bayesian inference to produce posteriors for everything. This approach runs into computational difficulties, which have led us to a two-stage “empirical Bayes’ ’ approach, where the hyperparameters for the longitudinal model are first fixed by a maximum-likelihood calculation, after which the remaining hyperparameters and individual-level parameters can be estimated with Bayesian machinery. For the time-to-event parameters we choose flat hyperpriors, selecting the hyperparameters  $\mu_B = \mu_\theta = \mu_\beta = 0$ ,  $\sigma_B = \sigma_\theta = 2$ , and  $\sigma_\beta = 100$ .

### The modelling variants

In this article, we researched into 16 variants of the model-fitting problem, but focussed mainly on 8 of them. The 8 main models use the standard deviation,  $\sigma$ , as the measure of the influence of blood-pressure variability on mortality. We also produced the same 8 models but using precision,  $\tau = 1/\sigma^2$ , as the measure of the influence of blood-pressure variability on mortality. However, this was only to ensure that there were no differences between the use of one over the other. Throughout the remainder of this appendix, we refer to the 8 main models using the following run numbers:

1. All participants (15,295), using mean systolic and diastolic blood pressure (not FRS) in the linear predictor term, with the outcome data as death specifically from HA-CVD-CeVD.
2. All participants (15,295), using mean systolic and diastolic blood pressure (not FRS) in the linear predictor term, with the outcome data as all-causes of death.

Table 2: Parameters for distribution of blood pressure, for the full population

BP	Variable	Mean	SD
Systolic	Overall Mean	125.36	19.48
Systolic	Delta	5.24	4.86
Systolic	Home SD	2.74	2.05
Systolic	Clinic SD	3.78	2.61
Diastolic	Overall Mean	74.34	10.29
Diastolic	Delta	3.90	3.14
Diastolic	Home SD	2.34	1.75
Diastolic	Clinic SD	3.08	2.08

3. Only participants that had data from which FRS values could be computed ( $N=9,418$ ) — the “FRS population” but using mean systolic and diastolic blood pressure (not FRS) in the linear predictor term, with the outcome data as death specifically from HA-CVD-CeVD.
4. FRS population, but using mean systolic and diastolic blood pressure (not FRS) in the linear predictor term, with the outcome data as all-causes of death.
5. FRS population, and using the FRS ATP-III value in the linear predictor term, with the outcome data as death specifically from HA-CVD-CeVD.
6. FRS population, and using the FRS ATP-III value in the linear predictor term, with the outcome data as all-causes of death.
7. FRS population, and using the FRS 1998-version value in the linear predictor term, with the outcome data as death specifically from HA-CVD-CeVD.
8. FRS population, and using the FRS 1998-version value in the linear predictor term, with the outcome data as all-causes of death.

We also include Directed Acyclical Graph (DAG) sketches to help visualize the different models, as shown in figures 3 and 4. In order to read the DAGs, note that each square background layer that appears as a stack of layers represents different measured outcomes that were made in the first wave of the survey. The outcome variables measured are represented by a square-shaped text box, and a parameter of the model is represented by a circular-shaped text box. If either a square or circular text box is placed on top of a stacked rectangular layer, it means that multiple values of that variable (as many as there are layers to the stack) are either measured (for outcome variables) or simulated (for parameters of the model). Please note that the number of layers in the stack is written in the text box that does not contain a frame which is intentionally displayed on top of the stacked layer that it represents. For example,  $i = 1, \dots, N$ . Finally, the direction of the arrows implies causality assumed in the model.

For convenience, we provide an overview of the different blood pressure values in the full and FRS populations in tables 2 and 3.

## Methodology

The methodology for this research can be split into three main sections: 1) calculating the empirical Bayes’ parameters, 2) parameterizing the model using Hamiltonian Monte Carlo (HMC) and 3) re-centering the variables in the linear predictor equation. By applying empirical Bayes’, Maximum Likelihood Estimates (MLEs) of some of the parameter distributions are provided. Note that the parameters estimated here are only the prior distribution of the global (not individual) blood pressure means and the variances, for both systolic and diastolic and home and clinic measurements. These estimates are then provided as prior distributions for the Stan MCMC simulations using HMC, where estimates can be made for all the parameter distributions of the model, given the specific centering applied. Finally, section (3) recalculates the centering values based on

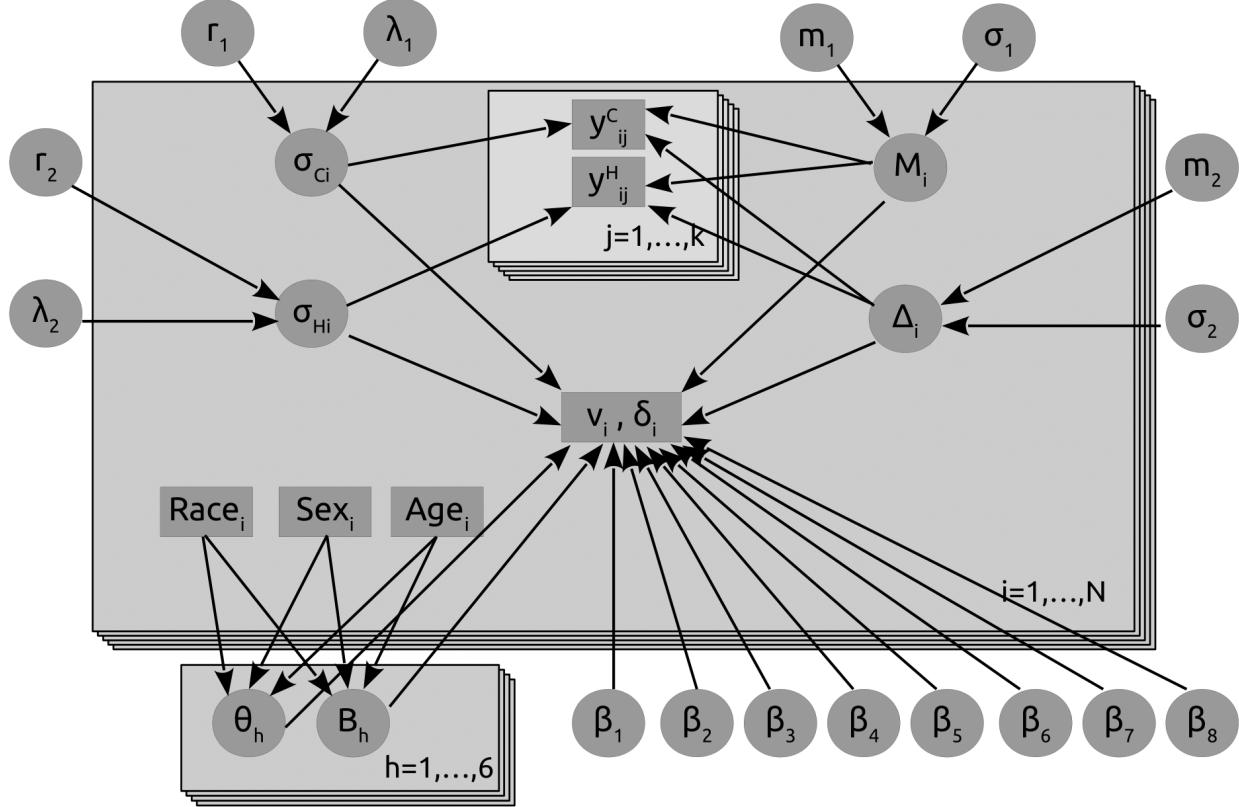


Figure 3: An illustration of the DAG of the mean blood pressure-based model presented in this article.

Table 3: Parameters for distribution of blood pressure, for the FRS population

BP	Variable	Mean	SD
Systolic	Overall Mean	125.89	18.33
Systolic	Delta	5.23	4.71
Systolic	Home SD	2.78	2.06
Systolic	Clinic SD	3.78	2.52
Diastolic	Overall Mean	76.41	10.00
Diastolic	Delta	3.84	3.10
Diastolic	Home SD	2.28	1.71
Diastolic	Clinic SD	2.92	1.97

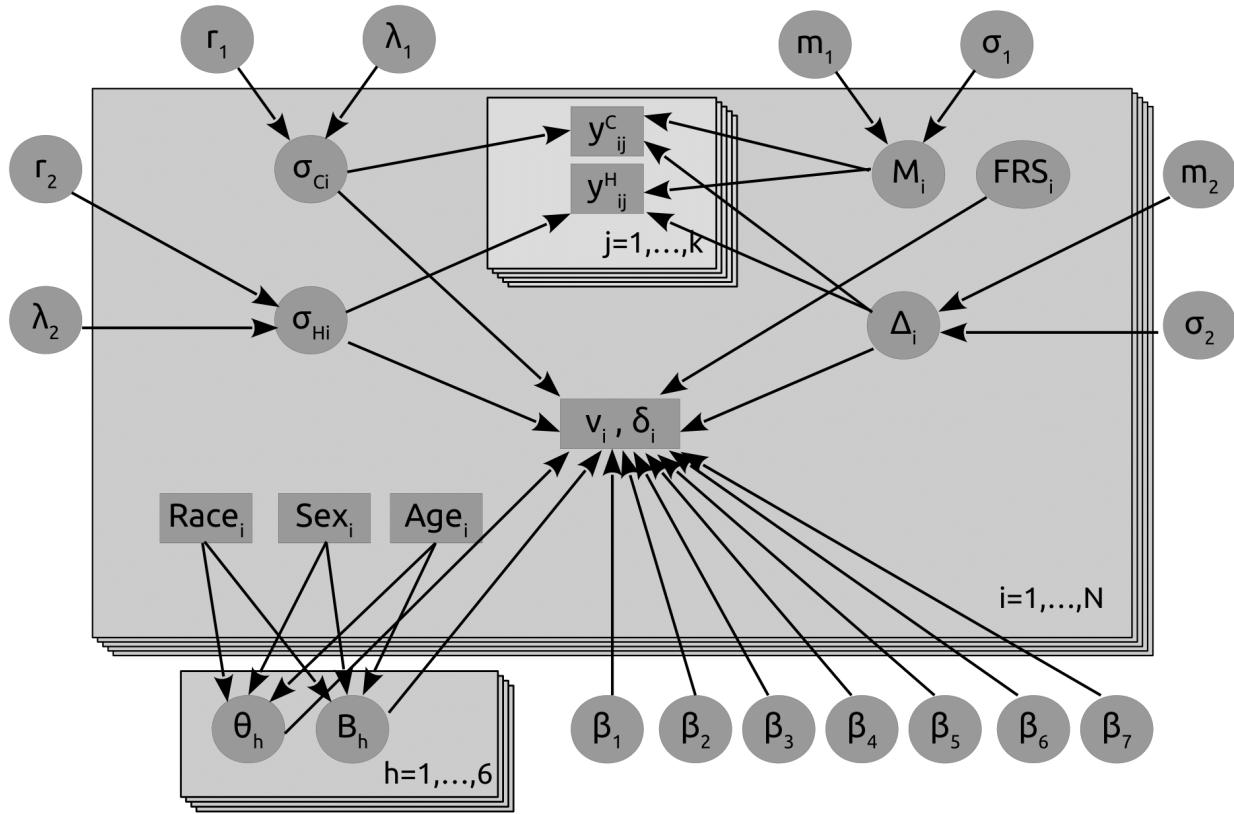


Figure 4: An illustration of the DAG of the FRS-based model presented in this article.

the previous MCMC iteration, and sets of the next iteration, while simultaneously checking for convergence in both the MCMC simulations and the centering values.

### Empirical Bayes Parameters

First, we extract the intervals for the digits in the blood pressure measurement recordings. Suppose the fractions of digits 0,2,4,6,8 are  $b_0, b_2, b_4, b_6, b_8$ . Letting  $B_0 = 0$  and  $B_k = 10 \sum_{j=0}^{k-1} b_{2j}$  for  $k = 1, \dots, 5$ , we want to choose a positive  $a$  and place breaks at  $-a + B_k$ , so that measurements between  $-a + B_k$  and  $-a + B_{k+1}$  modulo 10 are assigned the final digit  $2k$ , for  $k = 0, \dots, 4$ . We choose  $a$  to minimise the total distance of the intervals from the rounded value:

$$\sum_{k=0}^4 \int_{-a+B_k}^{-a+B_{k+1}} |x - 2k| dx = \frac{1}{2} \sum_{k=0}^4 (-a + B_k - 2k)^2 + (-a + B_{k+1} - 2k)^2,$$

as long as  $2k$  is in the appropriate interval. This is minimized at

$$a = \frac{1}{5} (B_1 + B_2 + B_3 + B_4 - 15) = \sum_{j=0}^3 (8 - 2j)b_{2j} - 3.$$

Next step, we fit the BP distribution parameters. We suppose that each individual has BP measures  $\tilde{y}_{ij}^l$  for  $i = 1, \dots, n$ ,  $j = 1, \dots, k$  (default  $k = 3$ ), and  $l = 1, 2$ , which are rounded versions of

$$y_{ij}^l \sim \mathcal{N}(\mu_i^l, (\tau_i^l)^{-1}),$$

where

$$\begin{aligned} \mu_i^1 &= (M_i + \Delta_i)/2, \\ \mu_i^2 &= (M_i - \Delta_i)/2, \\ M_i &\sim \mathcal{N}(m_M, \sigma_M^2) \text{ and } \Delta_i \sim \mathcal{N}(m_\Delta, \sigma_\Delta^2) \text{ independent,} \\ \tau_i^l &\sim \text{Gamma}(\alpha^l, \alpha^l/\theta^l). \end{aligned}$$

(Note that  $\alpha^l$  is the usual shape parameter, while  $\theta^l$  is the expectation.)

We wish to estimate the eight parameters

$$(m_M, m_\Delta, \sigma_M^2, \sigma_\Delta^2, \alpha^1, \alpha^2, \theta^1, \theta^2)$$

We begin by assuming  $y_{ij}^l$  observed directly. We estimate by maximising the partial likelihood on the observations

$$\begin{aligned} \bar{y}_{i+} &:= \frac{1}{2k} \sum_{j=1}^k y_{ij}^1 + y_{ij}^2, \\ \bar{y}_{i-} &:= \frac{1}{2k} \sum_{j=1}^k y_{ij}^1 - y_{ij}^2, \\ s_i^l &:= \frac{1}{k-1} \sum_{j=1}^k \left( y_{ij}^l - \frac{1}{k} \sum_{j=1}^k y_{ij}^l \right)^2. \end{aligned}$$

Note that

$$(k-1)s_i^l \tau_i^l = \sum_{j=1}^k \left( z_{ij}^l - \frac{1}{k} \sum_{j=1}^k z_{ij}^l \right)^2.$$

where  $z_{ij}^l$  are i.i.d. standard normal is independent of  $\tau_i^l$ , thus has a chi-squared distribution with  $k-1$  degrees of freedom — hence  $\frac{k-1}{2} \cdot s_i^l \tau_i^l$  is gamma distributed with parameters  $(\frac{k-1}{2}, 1)$ . Since  $\frac{\alpha}{\theta} \tau_i^l$  is independent of

$s_i^l \tau_i^l$ , with  $\text{Gamma}(\alpha, 1)$  distribution, we see that  $\frac{\theta(k-1)}{2\alpha} s_i^l$  is the ratio of two independent gamma random variables, hence has beta-prime distribution with parameters  $(\frac{k-1}{2}, \alpha)$ , so log partial likelihood

$$\ell_{\text{Beta}}(\alpha, \theta; s^l) = n\alpha \log \frac{\alpha}{\theta} + n \log \Gamma\left(\alpha + \frac{k-1}{2}\right) - n \log \Gamma(\alpha) + \frac{k-1}{2} \sum_{i=1}^n \log s_i^l - \left(\alpha + \frac{k-1}{2}\right) \sum_{i=1}^n \log \left(s_i^l + \frac{\alpha}{\theta}\right).$$

Note as well that these quantities  $(k-1)s_i^l$  should correspond to empirically observed individual variances; hence we will compare these empirical variances (with imputed fractional parts) divided by the normalization factor  $2\alpha/(k-1)\theta$  to the beta-prime distribution below as a goodness-of-fit test.

The partial Fisher Information has entries

$$\begin{aligned} -\frac{\partial^2 \ell}{\partial \alpha^2} &= n\psi_1(\alpha) - n\psi_1\left(\alpha + \frac{k-1}{2}\right) - \frac{n}{\alpha} + \sum_{i=1}^n \frac{2\theta s_i^l + \alpha - (k-1)/2}{(\theta s_i^l + \alpha)^2} \\ -\frac{\partial^2 \ell}{\partial \theta^2} &= -\frac{n\alpha}{\theta^2} + \frac{\alpha}{\theta^2} \left(\alpha + \frac{k-1}{2}\right) \sum_{i=1}^n \frac{2\theta s_i^l + \alpha}{(\theta s_i^l + \alpha)^2} \\ -\frac{\partial^2 \ell}{\partial \theta \partial \alpha} &= \frac{n}{\theta} - \frac{1}{\theta} \sum_{i=1}^n \frac{\alpha^2 + 2\alpha\theta s_i^l + \frac{k-1}{2}\theta s_i^l}{(\theta s_i^l + \alpha)^2}. \end{aligned}$$

where  $\psi_1$  is the trigamma function.

Let  $(\hat{\alpha}^l, \hat{\beta}^l)$  be the maximum partial likelihood estimators. Conditioned on  $(\tau_i^l)$  we have

$$\begin{aligned} \bar{y}_{i+} &\sim \mathcal{N}\left(m_M, \sigma_M^2 + \frac{1}{4k} \left(\frac{1}{\tau_i^1} + \frac{1}{\tau_i^2}\right)\right), \\ \bar{y}_{i-} &\sim \mathcal{N}\left(m_\Delta, \sigma_\Delta^2 + \frac{1}{4k} \left(\frac{1}{\tau_i^1} + \frac{1}{\tau_i^2}\right)\right). \end{aligned}$$

We would then have MLEs

$$\begin{aligned} \hat{m}_M &= \frac{1}{n} \sum_{i=1}^n \bar{y}_{i+}, \\ \hat{m}_\Delta &= \frac{1}{n} \sum_{i=1}^n \bar{y}_{i-}, \end{aligned}$$

which are approximately normally distributed, with means  $m_M$  and  $m_\Delta$  respectively, and conditional on  $\tau_i^l$  standard errors

$$\frac{\sigma_M^2}{n} + \frac{1}{4kn^2} \sum_{i=1}^n (\tau_i^1)^{-1} + (\tau_i^2)^{-1} \quad \text{and} \quad \frac{\sigma_\Delta^2}{n} + \frac{1}{4kn^2} \sum_{i=1}^n (\tau_i^1)^{-1} + (\tau_i^2)^{-1},$$

which we may approximate — with error on the order of  $n^{-3/2}$  — replacing the mean of  $(\tau_i^l)^{-1}$  by its expected value  $\beta^l/(\alpha^l - 1)$  to obtain

$$\begin{aligned} \text{Var}(\hat{m}_M) &\approx \frac{\sigma_M^2}{n} + \frac{1}{4kn} \left( \frac{\beta^1}{\alpha^1 - 1} + \frac{\beta^2}{\alpha^2 - 1} \right) \\ \text{Var}(\hat{m}_\Delta) &\approx \frac{\sigma_\Delta^2}{n} + \frac{1}{4kn} \left( \frac{\beta^1}{\alpha^1 - 1} + \frac{\beta^2}{\alpha^2 - 1} \right) \end{aligned}$$

Finally, conditioned on the  $\tau_i^l$  we have that the random variables  $\bar{y}_{i+}$  are normal with variance

$$\sigma_M^2 + \frac{1}{4k} ((\tau_i^1)^{-1} + (\tau_i^2)^{-1}),$$

so the unconditional variance is the expected value, or

$$\sigma_M^2 + \frac{1}{4k} \left( \frac{\beta^1}{\alpha^1 - 1} + \frac{\beta^2}{\alpha^2 - 1} \right).$$

This yields the estimators

$$\begin{aligned}\hat{\sigma}_M^2 &= \frac{1}{n-1} \sum_{i=1}^n \left( \bar{y}_{i+} - n^{-1} \sum_{i=1}^n y_{i+} \right)^2 - \frac{1}{4k} \left( \frac{\hat{\beta}^1}{\hat{\alpha}^1 - 1} + \frac{\hat{\beta}^2}{\hat{\alpha}^2 - 1} \right), \\ \hat{\sigma}_\Delta^2 &= \frac{1}{n-1} \sum_{i=1}^n \left( \bar{y}_{i-} - n^{-1} \sum_{i=1}^n y_{i-} \right)^2 - \frac{1}{4k} \left( \frac{\hat{\beta}^1}{\hat{\alpha}^1 - 1} + \frac{\hat{\beta}^2}{\hat{\alpha}^2 - 1} \right).\end{aligned}$$

Using the delta method, and the fact that we see that the variance of  $\hat{\beta}/(\hat{\alpha} - 1)$  is approximately

$$\frac{\sigma_\beta^2}{(\hat{\alpha} - 1)^2} + \frac{\hat{\beta}^2 \sigma_\alpha^2}{(\hat{\alpha} - 1)^4},$$

where  $\sigma_\alpha$  and  $\sigma_\beta$  are the standard errors for  $\hat{\alpha}$  and  $\hat{\beta}$  respectively, so the standard errors for  $\hat{\sigma}_M^2$  and  $\hat{\sigma}_\Delta^2$  are approximately

$$\begin{aligned}\text{SE}(\hat{\sigma}_M^2) &\approx \frac{1}{2k} \left( \frac{8k^2 \hat{\sigma}_M^2}{n} + \frac{\sigma_\beta^2}{(\hat{\alpha}^1 - 1)^2} + \frac{(\hat{\beta}^1)^2 \sigma_\alpha^2}{(\hat{\alpha}^1 - 1)^4} + \frac{\sigma_\beta^2}{(\hat{\alpha}^2 - 1)^2} + \frac{(\hat{\beta}^2)^2 \sigma_\alpha^2}{(\hat{\alpha}^2 - 1)^4} \right)^{1/2}, \\ \text{SE}(\hat{\sigma}_\Delta^2) &\approx \frac{1}{2k} \left( \frac{8k^2 \hat{\sigma}_\Delta^2}{n} + \frac{\sigma_\beta^2}{(\hat{\alpha}^1 - 1)^2} + \frac{(\hat{\beta}^1)^2 \sigma_\alpha^2}{(\hat{\alpha}^1 - 1)^4} + \frac{\sigma_\beta^2}{(\hat{\alpha}^2 - 1)^2} + \frac{(\hat{\beta}^2)^2 \sigma_\alpha^2}{(\hat{\alpha}^2 - 1)^4} \right)^{1/2}\end{aligned}$$

Now we compute the combined variance. For a parameter like  $\alpha$  we estimate the variance of  $\hat{\alpha}$  by

$$\text{Var}(\hat{\alpha}) = \mathbb{E}[\text{Var}(\hat{\alpha} | I)] + \text{Var}(\mathbb{E}[\hat{\alpha} | I]).$$

Here  $I$  represents the randomly imputed fractional part. We can estimate the first term by averaging the estimated variance (from Fisher Information) over all random imputations. We estimate the second term by the variance of the  $\alpha$  estimates over imputations. Note that this is not quite right, since what we really want the variance of is  $\alpha_0(I)$  — effectively, the “true” parameter consistent with the imputation. This is a plug-in estimate, as is the Fisher Information estimate of the variance.

We then compute the residuals. We define the deviance for an individual  $i$  with observations  $(Y_i)$  given the hyperparameters  $h = (m_M, m_\Delta, \sigma_M^2, \sigma_\Delta^2, \alpha^H, \theta^H, \alpha^C, \theta^C)$

$$D = \sum_{i=1}^n \log \mathbb{P}\{\mathbf{Y}_i | \text{hyperparameters} = h\}.$$

Since the  $\mathbf{Y}_i$  are independent conditioned on  $h$ ,

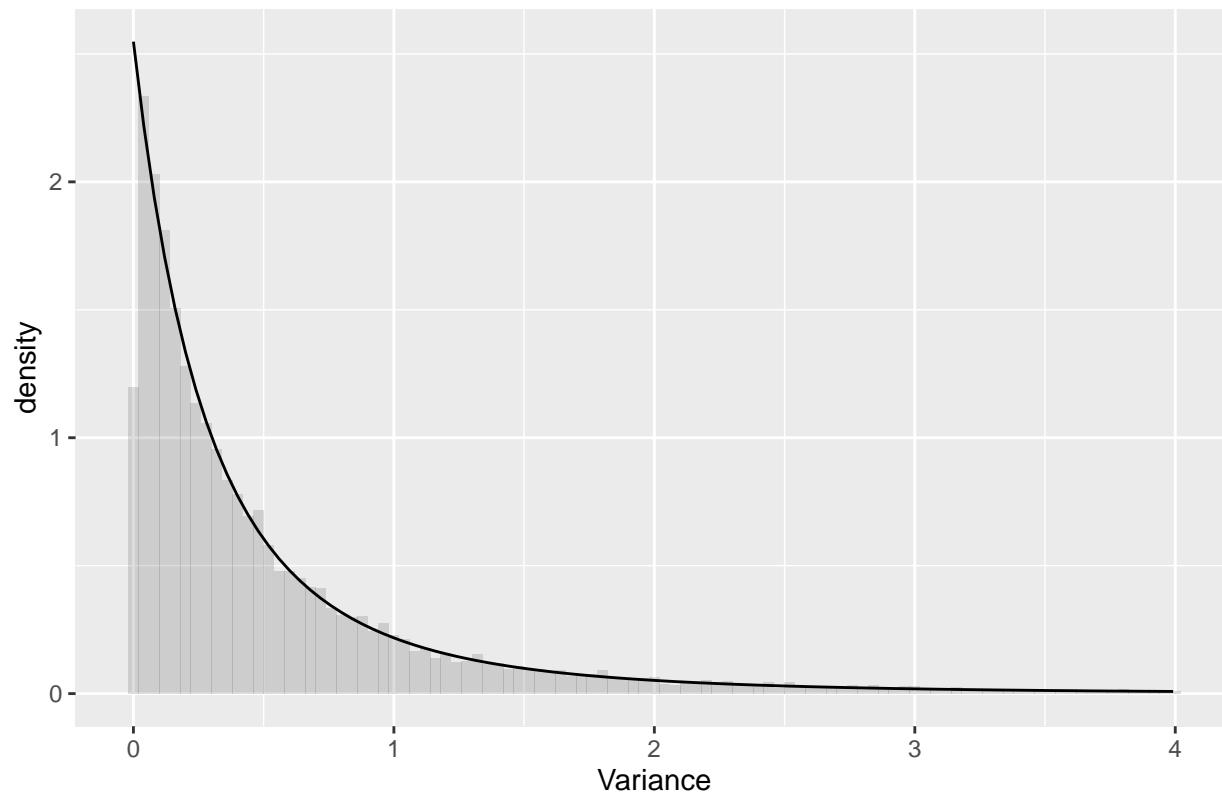
$$\begin{aligned}D &= \sum_{i=1}^n \log \mathbb{E}_h [\mathbb{P}\{\mathbf{Y}_i | M_i, \Delta_i, \tau_i^C, \tau_i^H\}] \\ &\approx \sum_{i=1}^n \log \frac{1}{R} \sum_{r=1}^R [\mathbb{P}\{\mathbf{Y}_i | M_{i,r}, \Delta_{i,r}, \tau_{i,r}^C, \tau_{i,r}^H\}] \frac{\pi_h(M_{i,r}, \Delta_{i,r}, \tau_{i,r}^C, \tau_{i,r}^H)}{q(M_{i,r}, \Delta_{i,r}, \tau_{i,r}^C, \tau_{i,r}^H | h, \mathbf{Y}_i)},\end{aligned}$$

where  $(M_{i,r}, \Delta_{i,r}, \tau_{i,r}^C, \tau_{i,r}^H)$  are independent samples from a distribution  $q$  that may depend on  $\mathbf{Y}_i$  and  $h$ , and  $\pi_h$  is the true density of those individual parameters given hyperparameters  $h$ .

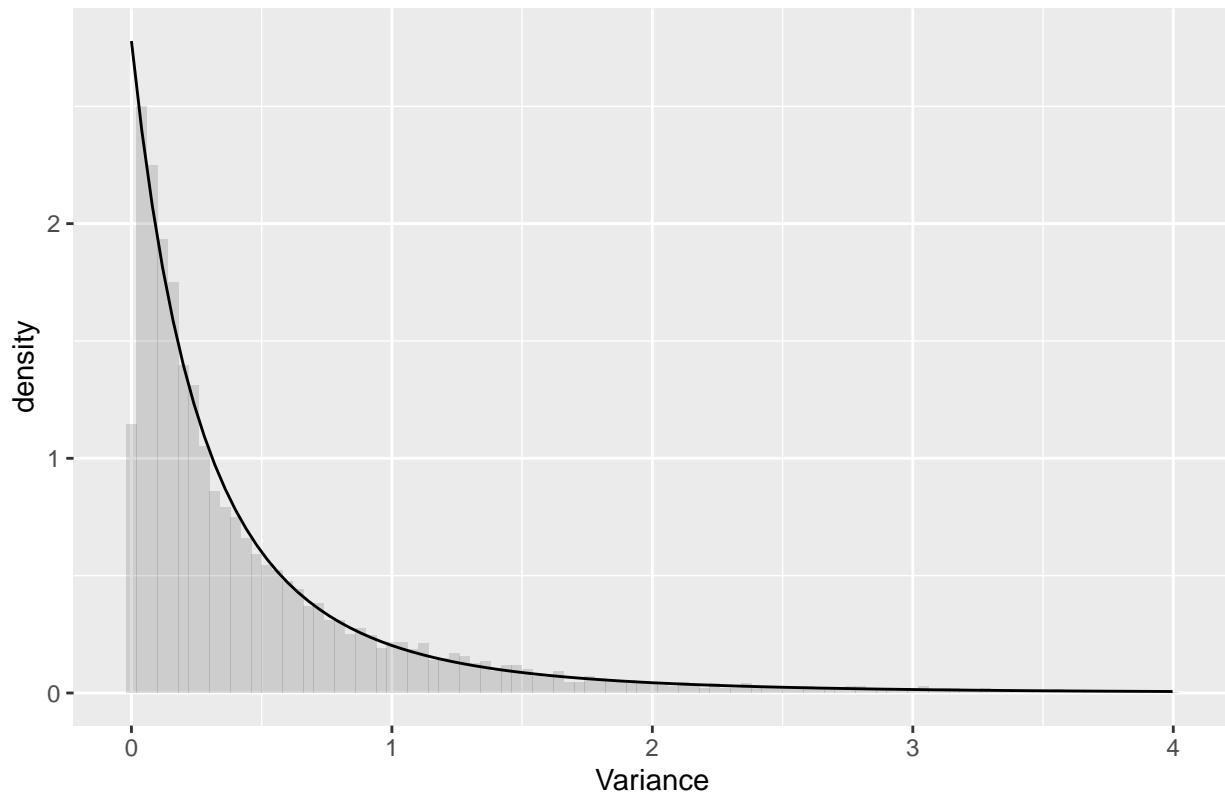
Finally, we check the variance distribution empirically, to check whether the continuous distribution we have fit for individual variances describes the true distribution of variances in the population reasonably well. The first

thing we do is to compare the empirical variances (with imputed fractional parts) to the theoretical beta-prime distribution. To match the standard distribution, the variances are normalized by being divided by the factor  $\alpha/\theta$ . Note that the distribution has a very long tail, and we have truncated the plot at a point where about

### Systolic Clinic : Histogram of Variances/beta



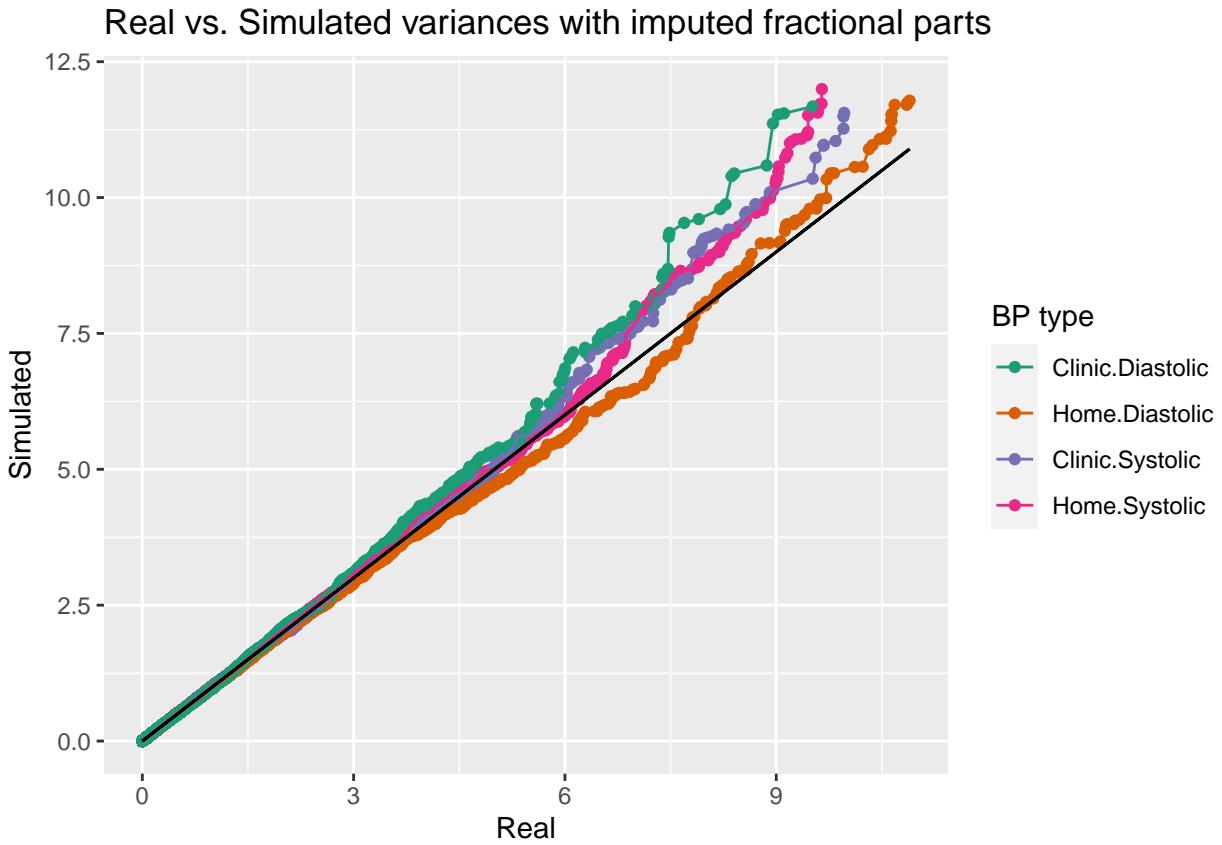
## Diastolic Clinic : Histogram of Variances/beta



Now we generate data from the inferred model that mimic the true data, with three systolic and three diastolic BP measures per person. These cannot be directly compared with the observed data, which are rounded to the nearest 2 (in a somewhat biased way), so we randomly impute the fractional part, which we add to the true observations. This gives us a set of true variances and a set of simulated variances, which we hope will have approximately the same distribution. We compare these — for each of the four combinations of systolic/diastolic and home/clinic — by Q-Q plots.

Table 4: Empirical Bayes prior hyperparameter estimates for the systolic blood pressure for the NHANES III, full population.

Parameter	Estimate
Clinic SD alpha	2.5486+/-0.0058688
Clinic SD theta	0.075434+/-4.9887e-05
Delta mu	1.3468+/-0.057983
Delta sigma	6.8127+/-1.1526
Home SD alpha	2.1597+/-0.0032284
Home SD theta	0.1496+/-7.1299e-05
Mean mu	123.34+/-0.16096
Mean sigma	19.412+/-1.3567



To finish this section, we include a table of the empirical Bayes priors, shown in tables 4 and 5.

### Hamiltonian Monte Carlo (HMC)

The model, as described in the article, is a Bayesian hierarchical model. In order to parameterize such an intricate model, traditional Maximum Likelihood Estimation methods can no longer be applied. Therefore, we apply the Hamiltonian Monte Carlo (HMC) method. HMC is a form of Markov Chain Monte Carlo methods, which samples potential parameter space values of the model, then calculates directly the likelihood function based on that choice of parameters. The derivative of the likelihood function,  $\phi$ , guides parameter space exploration in  $\theta$  towards the modal value of the joint posterior distribution. This method is ideal for complicated, non-Gaussian distribution forms. The three steps of HMC are:

1. Draw a sample of the derivative  $\phi$  using the posterior distribution of  $\phi$ , which is the same as its prior.

Table 5: Empirical Bayes prior hyperparameter estimates for the diastolic blood pressure for the NHANES III, full population.

Parameter	Estimate
Clinic SD alpha	2.7786+/-0.0078203
Clinic SD theta	0.10888+/-8.2224e-05
Delta mu	1.096+/-0.040361
Delta sigma	4.6823+/-0.94718
Home SD alpha	2.3167+/-0.0040113
Home SD theta	0.19507+/-0.00010161
Mean mu	72.312+/-0.085055
Mean sigma	10.201+/-1.0485

2. Update the values of  $\theta^*$  and  $\phi^*$  using

$$\theta^* \leftarrow \theta + \epsilon M^{-1} \phi, \quad (11)$$

and

$$\phi \leftarrow \phi + \epsilon \frac{1}{2} \frac{d \log\{p(\theta|y)\}}{d\theta}, \quad (12)$$

where  $M$  is the jacobian of the parameters. This can be set to a diagonal matrix for no correlation between parameters, and is pointwise updated throughout the calculation. This is the leapfrog method, whereby  $\epsilon$  dictates the scale size of the step to ensure convergence on the correct point is made, and  $L$  is the number of steps to be ‘leaped’.

3. Compute the rejection parameter:

$$r = \frac{p(\theta^*|y)p(\phi^*)}{p(\theta^{t-1}|y)p(\phi^{t-1})} \quad (13)$$

4. Set  $\theta^t$  to  $\theta^*$  with probability  $\min\{1, r\}$ , or otherwise keep  $\theta^{t-1}$ .

The tuning parameters  $\epsilon$  and  $L$  should be chosen according to a desired acceptance rate. The No-U-Turn Sampler of Stan automates the calculation of these tuning parameters. A more detailed overview of HMC and the NUTS algorithm integrated into the Stan package, see ‘*The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*’ by M. Hoffman and A. Gelman, Journal of Machine Learning Research, 15, 1351-1381 (2014).

### Centering the Linear Predictor

During the MCMC simulations, the centering values play a non-negligible role in shaping the model parameterization. If the centering parameters are held constant throughout all of the MCMC simulations, then the equation  $\sum_i^N \exp(\beta \cdot (\mathbf{X} - \hat{\mathbf{X}})) = 0$  is no longer guaranteed. However, automatically defining the centering values based on the model parameters sampled at the current MCMC iteration is not advisable as it can lead to poor parameter convergence. This is because it modifies the likelihood function at every MCMC iteration. Therefore, we iterate the MCMC algorithm multiple times. At every iteration, we recalculate the centering parameters to satisfy the requirement that the average of the linear predictor term going to zero, based on the posterior distributions of the previous MCMC simulation. This iteration is carried out until the centering parameters converge. Convergence is defined by optimising on two factors. The first is that the sum of the linear predictor term across all MCMC samples needs to tend to negligible values (we define this as the average difference being less than  $10^{-7}$ ), see figure 5. The second convergence criteria is that the average Root Mean-Squared Error (RMSE) of the model predictions on the survival outcomes in the MCMC simulations needs to also decrease towards zero, see figure 5 (top). For the second criteria, we stopped the simulations when either the difference in the RMSE stopped decreasing (below a threshold of 1%), or the RMSE value was less than 20, see figure 5 (bottom). Illustration of the convergence is shown in figure 5.

## Code Description

The code can be found at <https://github.com/hamishwp/Nhanes2021>. The numerical code has been built in multiple stages. Below, we explain the principal files required to replicate the entire analysis presented in the article. There are 5 main groups for the code:

1. Data cleaning scripts
2. Main file
3. Stan files for HMC
4. Centering recalculation scripts
5. Post-processing analysis

We provide a brief description of each of these below.

### Data Cleaning

This is found in the file `Dataclean2021.R`. Provided the raw NHANES dataset (in CSV format), it extracts all the data required for the simulations, and stores it in a structure that can be directly read in to the main file (`MCMC_DiasSyst_v3.R`) of this research.

### Main

The main file is `MCMC_DiasSyst_v3.R`. It reads in the cleaned NHANES data, the specific choice of simulation parameters (for example, whether to use the FRS number or mean systolic & diastolic blood pressure), and runs the correct RStan scripts for that specific selection of simulation parameters. This script is intended for use on computing clusters.

### Stan

There are eight Stan files:

1. `mystanmodel_DS_sigma_v2_autopred.stan`
2. `mystanmodel_DS_tau_v2_autopred.stan`
3. `mystanmodelFRS_DS_sigma_v2_autopred.stan`
4. `mystanmodelFRS_DS_tau_v2_autopred.stan`
5. `mystanmodel_DS_sigma_v2.stan`
6. `mystanmodel_DS_tau_v2.stan`
7. `mystanmodelFRS_DS_sigma_v2.stan`
8. `mystanmodelFRS_DS_tau_v2.stan`

These correspond to the following alternative simulation parameters:

- For the blood-pressure variability, choosing to use the standard-deviation  $\sigma$  or the precision  $\tau = 1/\sigma^2$
- Using the FRS score or the mean diastolic and systolic blood pressure as a covariate in the analysis
- Whether the centering parameters,  $\hat{X}$ , in the linear predictor term are automatically calculated to satisfy  $\sum_i^N \exp(\beta \cdot (\mathbf{X} - \hat{X})) = 0$  for every MCMC iteration, or whether the centering is held constant across all iterations

### Centering

The centering of the linear predictors, which is required as input to every MCMC simulation iteration, is recalculated in the files `AutoPred_Recalc.R` and `ManPred_Recalc.R`. This is then provided to the Main script, `MCMC_DiasSyst_v3.R`, which provides these centering values to the Stan code for the MCMC simulations.

### Empirical Bayes Estimation

The file `gamma_fits.Rmd` contains all the necessary routines in order to replicate the calculation of the empirical Bayes' priors for the hyperparameters of the model.

## Post-processing

The post-processing script is called `PostProcessing.R`, which heavily relies on the `Functions.R` script which contains all the necessary functions to analyse the data. The post-processing script generates many useful plots of the MCMC posterior distribution for the user, including Bayes' factors, violin plots of the normalised beta and gompertz posteriors, and more.

## Results

In this section, we add some additional detail to the results section covered in the article. Extra information is given to explain how convergence of the simulations was ensured, and to also include more visualisations of the converged model parameterizations. The authors feel that this is particularly useful to provide confidence in the model parameterization and the predictions.

### Convergence of Simulations

Convergence of the simulations required to parameterize the model presented in this work is required for the MCMC simulations performed by Stan, as well as convergence in the centering values that requires repeating the Stan calculations several times. Convergence of the latter is shown in figure 5. The upper plot in figure 5 illustrates convergence in the average Root Mean-Squared Error (RMSE) of the model predictions on the survival outcomes in the MCMC simulations. The lower plot in figure 5 illustrates convergence in the average sum of the linear predictor terms over all MCMC chain iterations.

With respect to convergence of the MCMC simulations, defining convergence first involves discarding the burn-in period of the simulations. When the time-evolution marker chain has a large number of samples, sequence thinning is used to reduce the amount of data storage - after convergence, take only the  $k$ th value of the simulations (after having discarded the burn-in phase values) and discard the rest. One measure of convergence is to bin similar markers and check that for each bin, the variation of the individual marker movement over a few time steps is larger than the variation of the ensemble markers in-between one-another. Other methods of convergence are stationarity and mixing. The former occurs by ensuring that the gradients of movements in the chains in time are in the same direction, the latter ensures that the amplitude of the movements in the chains are similar. To calculate the mixing and stationarity, one can do the following:

- Take the supposedly converged marker population, where there are  $N$  markers in total each of index length  $\tau$  (thus of total physical time quantity  $t\tau$ ). Split it  $k$  times, where  $k$  is a common denominator of  $\tau$ .
- Now you have  $kN$  MCMC chains each of length  $|\tau/k|$
- For the marker  $\psi_{ij}$  with  $i$  and  $j$  the chain length (time) and marker number indices respectively, then the mean marker value over the chain length (time) is

$$\bar{\psi}_{|j} = \frac{k}{\tau} \sum_{i=1}^{\tau/k} \psi_{ij} \quad (14)$$

and the total average quantity of  $\psi$  over all markers, over all chain lengths is therefore

$$\bar{\psi}_{||} = \frac{1}{kN} \sum_{j=1}^{kN} \bar{\psi}_{|j} \quad (15)$$

- Stationarity: compare the inter-marker variance (between sequence B):

$$B = \frac{\tau}{k(kN-1)} \sum_{j=1}^{kN} (\bar{\psi}_{|j} - \bar{\psi}_{||})^2 \quad (16)$$

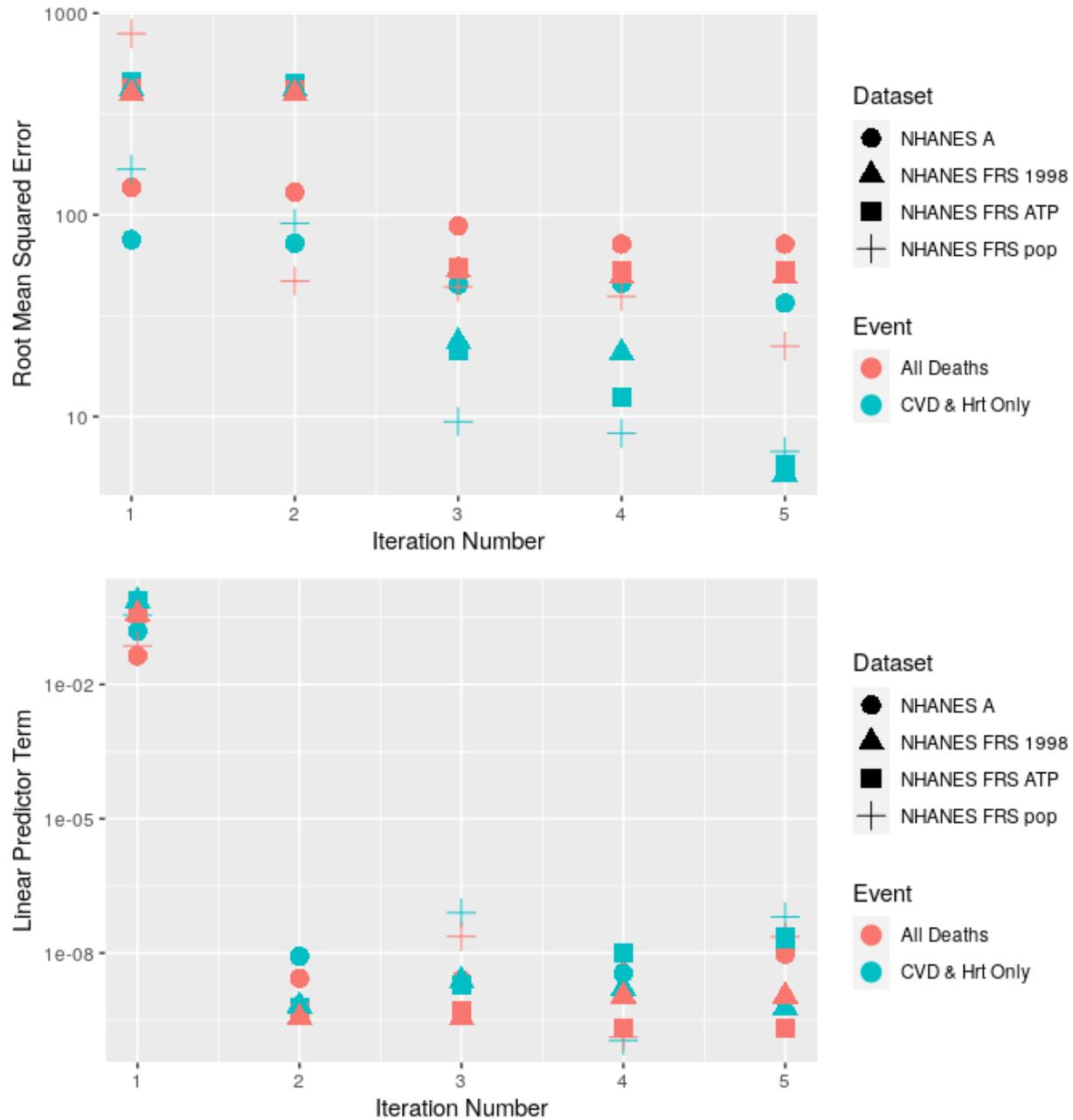


Figure 5: Illustration of the convergence of the centering parameters of the model.

- Mixing: compare the variance along each markers chain length (within-sequence W):

$$W = \frac{1}{n(\tau - k)} \sum_{j=1}^{kN} \sum_{i=1}^{\tau/k} (\psi_{i,j} - \bar{\psi}_{|j})^2 \quad (17)$$

- Therefore, to estimate the marginal posterior variance of  $p(\psi|y)$ , then we use a weighted average

$$\hat{\text{Var}}^+(\psi|y) = \frac{\tau - k}{N} W + \frac{1}{Nk} B \quad (18)$$

Note that this quantity overestimates the marginal posterior variance, but it is unbiased under stationarity: this can be used to infer convergence. When the variation in

$$\hat{R} = \sqrt{\frac{\hat{\text{Var}}^+(\psi|y)}{W}} \quad (19)$$

should approach close to 1 for converged simulations.

Another convergence parameter is the number of effective independent marker draws. Upon convergence, the time evolution of each marker should be uncorrelated and independent to previous time steps. To find the average time-correlation over all particles, we use the variogram  $V_t$ :

$$V_t = \frac{1}{Nk(\tau/k - \tilde{t})} \sum_{j=1}^{kN} \sum_{i=1}^{\tau/k} (\psi_{i,j} - \psi_{i-\tilde{t},j})^2, \quad (20)$$

where  $\tilde{t} \in 1, 2, \dots, \tau/k$  is a time index. Then we get the time-correlations:

$$\hat{\rho}_t = 1 - \frac{V_t}{2\hat{\text{Var}}^+} \quad (21)$$

This comes from the expectation of the variance  $E[(\psi_i - \psi_{i-t})^2] = 2(1 - \rho_t)\text{Var}(\psi)$ . This can be used to infer the effective number of independent marker draws:

$$\hat{n}_{eff} = \frac{mn}{1 + 2 \sum_{\tilde{t}=1}^T \hat{\rho}_t} \quad (22)$$

Where T is the index at which the sum of the autocorrelation estimates  $\hat{\rho}_{t'} + \hat{\rho}_{t'+1}$  is negative. As a general guide, we should have  $\hat{n}_{eff} \sim 10N/k$  effective independent marker draws and that  $\hat{R} \rightarrow 1 \sim 1.1$ . In this research, we continued running the MCMC simulations until these two criteria were met (and went beyond:  $\hat{R} < 1.05$  for all parameters in all models and that  $\hat{n}_{eff} > 750$  for all parameters in all models).

## Results - Model Parameterization

We remind the reader of the list of numbers of the different models explored in this research, provided in the list found in section ‘Proposed Models’. The authors will use the numbers in the list, referred to as the run number, in the following plots. One of the most important set of parameters of the model is the vector  $\beta$  of covariates in the Cox’ proportional hazards model. When the  $\beta$  vector is normalised, the larger (in absolute terms) the value of  $\beta$ , the larger the correlation between that specific covariate and the risk of mortality. Positive values of  $\beta$  imply a higher risk of mortality, and the inverse for negative values of  $\beta$ . As we can see from the violin plots of the MCMC posterior samples of the  $\beta$  parameters in figure 6, the parameter that correlated the highest with both the mortality risk of HA-CVD-CeVD and for all mortalities, in absolute terms, was the 1998 version of the FRS score, shown in the top-right plot under run numbers 7 and 8. The FRS-1998 score correlated, on average over all the MCMC iterations, approximately 25% more with mortality risk of HA-CVD-CeVD than the (more recently developed) FRS ATP III score. A similar, but slightly weaker, correlation was found between the two FRS scores for all mortality-based risk. The middle-left plot in figure

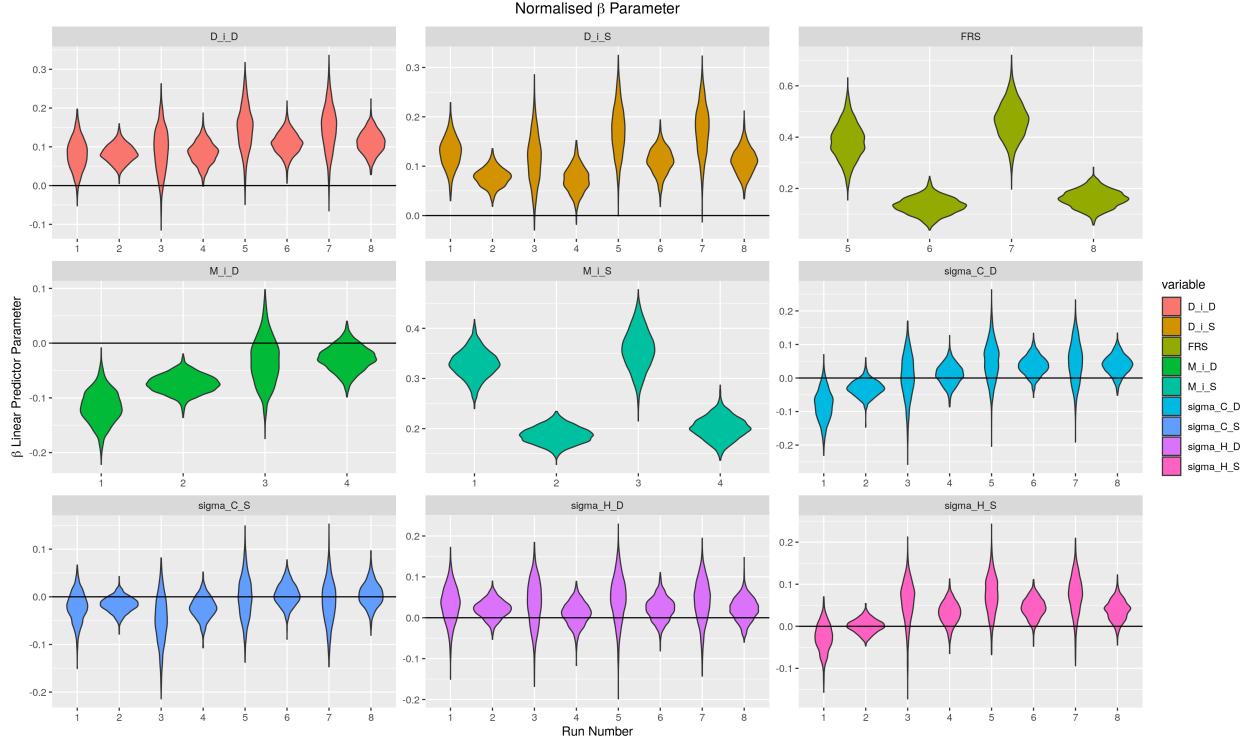


Figure 6: Violin plots of the normalised  $\beta$  parameters of the different models.

6 shows that the mean diastolic blood pressure acts to decrease mortality risk. Finally, the influence of the longer-term difference in the mean blood pressure, displayed in the top-left and top-middle plots of figure 6, is also shown to increase mortality risk across all run numbers. The influence of the blood-pressure variability on mortality is illustrated to not be consistent across simulations, whereby the statistical significance of the effect is lower than for the other parameters in the linear predictor term.

With respect to the time-independent Gompertz parameter, described using  $B$  in this article, the results between all models that simulate HA-CVD-CeVD mortality risk, and all the models that simulation all-cause mortality risk are consistent with one-another. This is illustrated by the similarity between plots on the left hand side and the right hand side of figure 7. The consistency appears across sex assigned at birth and race.

Figure 8 reflects the same level of consistency for the Gompertz parameter that influences the temporal evolution of the mortality risk. It is worth noting that both figures 7 and 8 have inverse trends between the values of  $B$  and theta for each demographic group. This makes it difficult to imagine, based on these two plots, what the mortality risk is at different ages across demographics, yet it is evident that the form of the change in the mortality risk curve in time is different for each demographic group. Women are observed to have lower initial values of risk, but mortality risk later in life begins to increase much faster than for men. Additionally, hispanic populations are shown to have a larger initial mortality risk than black populations who are shown to have a larger initial mortality risk than white populations in the USA. However, mortality risk increases at a faster rate for white populations than for black populations, for which it increases faster than hispanic populations in the USA. For ease of comparison, we also present here tables of the mean and standard deviation values of the time dependent and independent Gompertz parameters in tables 6 to 8.

## Results - Model Performance

To measure the performance of the model to predict the survival outcome of individuals in the population, figure 9 shows, ordered by individual age, the cumulative hazard  $H(t)$  predicted against the cumulative number of deaths in the populations, for each model explored in this research. Each model is shown to predict

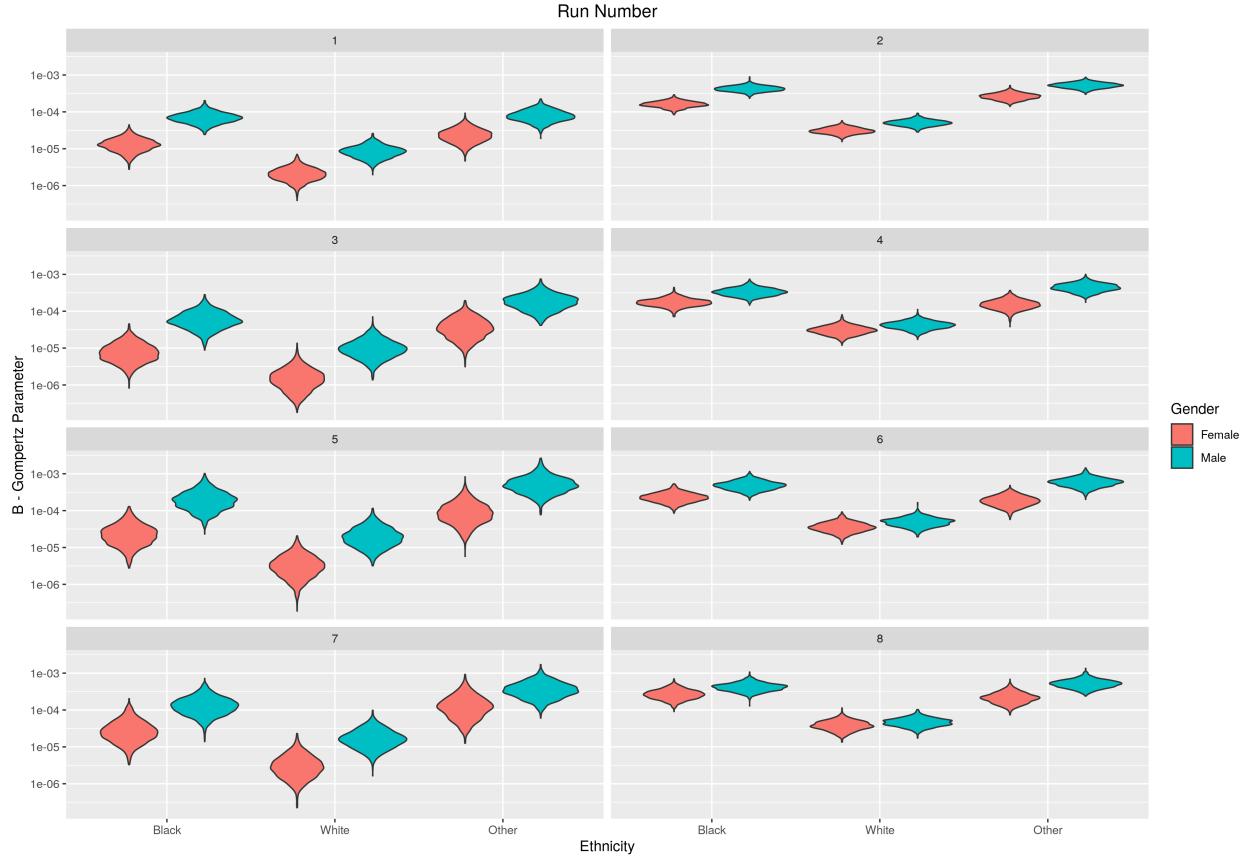


Figure 7: Violin plots of the normalised B parameter (from the Gompertz equation) of the different models.

Table 6: Parameters for survival model for HA-CVD-CeVD and all-cause mortality, NHANES III, Full population, using the systolic and diastolic mean model.

Sex	Race/Ethnicity	B-Mean	B-SD	$\theta$ -Mean	$\theta$ -SD
Female	Black	1.36e-05	5.70e-06	0.0883	0.00534
Female	White	2.20e-06	9.00e-07	0.1080	0.00488
Female	Other	2.44e-05	1.07e-05	0.0775	0.00565
Male	Black	7.20e-05	2.49e-05	0.0712	0.00459
Male	White	8.70e-06	3.40e-06	0.0973	0.00466
Male	Other	8.17e-05	3.00e-05	0.0678	0.00498

Table 7: Parameters for survival model for HA-CVD-CeVD and all-cause mortality, NHANES III, FRS-population only, using the systolic and diastolic mean model.

Sex	Race/Ethnicity	B-Mean	B-SD	$\theta$ -Mean	$\theta$ -SD
Female	Black	8.60e-06	5.40e-06	0.0937	0.00795
Female	White	1.70e-06	1.30e-06	0.1100	0.00844
Female	Other	4.06e-05	2.65e-05	0.0708	0.00879
Male	Black	6.36e-05	3.53e-05	0.0697	0.00765
Male	White	1.08e-05	6.50e-06	0.0896	0.00737
Male	Other	1.94e-04	9.69e-05	0.0539	0.00709

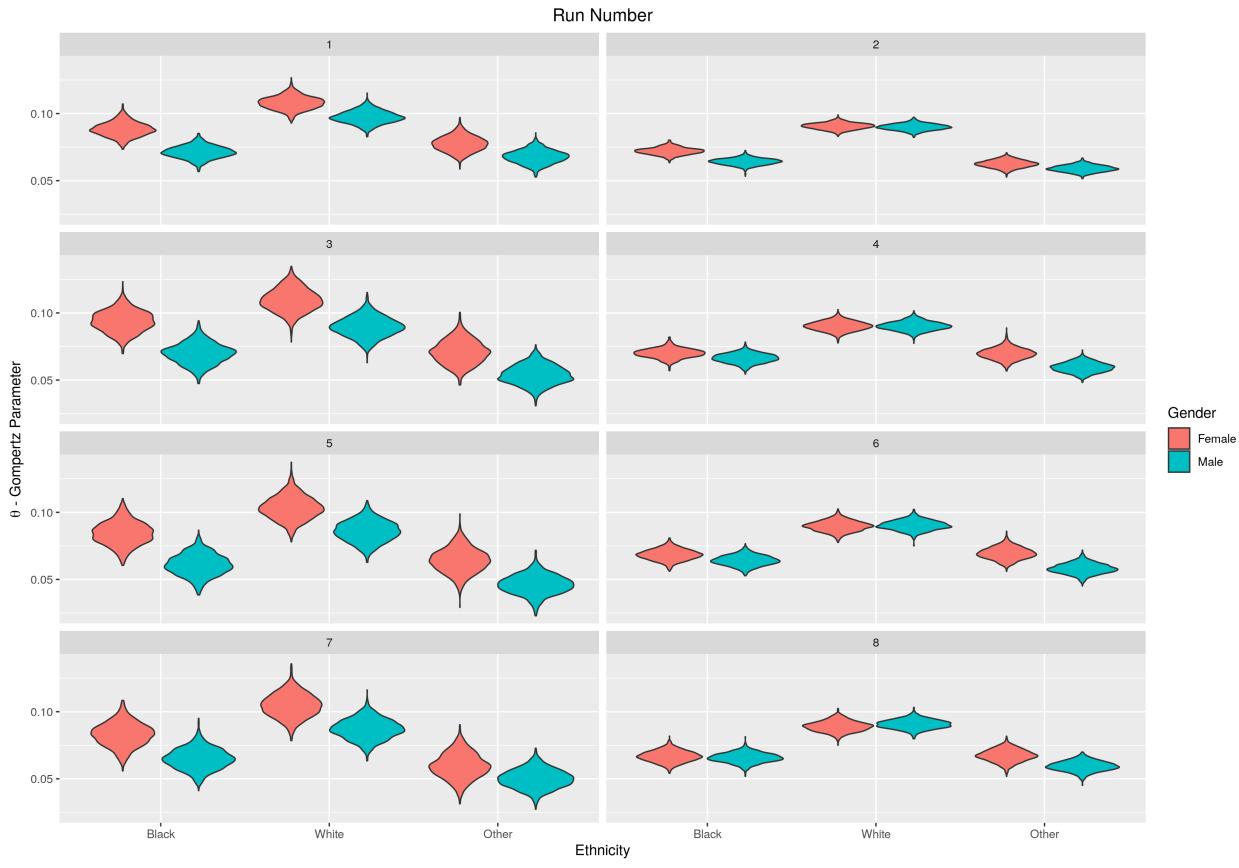


Figure 8: Violin plots of the normalised  $\theta$  parameter (from the Gompertz equation) of the different models.

Table 8: Parameters for survival model for HA-CVD-CeVD and all-cause mortality, NHANES III, FRS-population only, using the 1998 FRS-based model.

Sex	Race/Ethnicity	B-Mean	B-SD	$\theta$ -Mean	$\theta$ -SD
Female	Black	2.62e-05	1.82e-05	0.0843	0.00859
Female	White	3.70e-06	2.70e-06	0.1030	0.00861
Female	Other	1.02e-04	7.46e-05	0.0641	0.00896
Male	Black	2.14e-04	1.34e-04	0.0608	0.00793
Male	White	2.22e-05	1.46e-05	0.0859	0.00778
Male	Other	5.58e-04	3.25e-04	0.0461	0.00720

survival outcomes reliably, across the entire age range of the population.

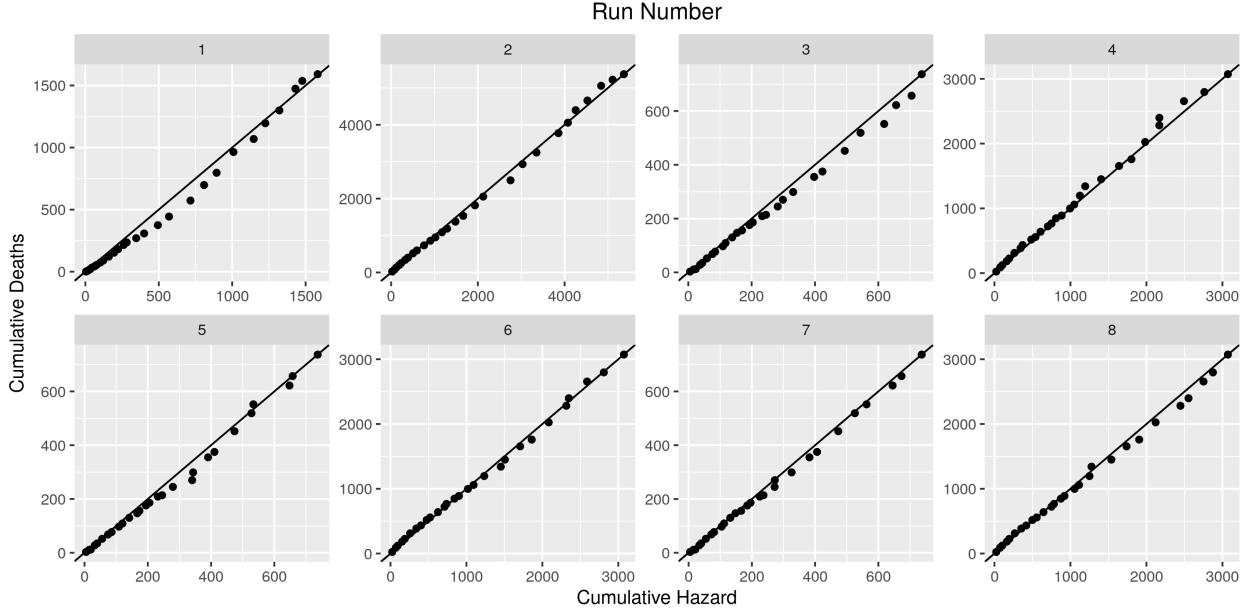


Figure 9: Predicted cumulative hazard against cumulative number of deaths in the population, ordered by the age of the individual.

A common metric that is used to evaluate the performance of models such as presented in this article is called the Receiver Operating Characteristic (ROC) curve. With continuous predictor values such as cumulative hazard  $H(T_i)$ , a threshold can be defined whereby any individual who has a cumulative risk larger than the threshold  $H(T_i) > \epsilon$  is predicted to die. The ratio of the number individuals that were predicted to die compared to the total number who die corresponds is referred to as the True Positive Ratio (TPR)

$$TPR(\epsilon) = \frac{\sum_i (\mathbb{I}(H(T_i) > \epsilon \text{ } \& \text{ } \delta_i = 1))}{\sum_i (\mathbb{I}(\delta_i = 1))}. \quad (23)$$

Note that TPR is also referred to as the recall or sensitivity. Conversely, the ratio of the number of individuals predicted to die but survive compared to the total number of individuals that survived is referred to as the False Positive Ratio (FPR)

$$FPR(\epsilon) = \frac{\sum_i (\mathbb{I}(H(T_i) > \epsilon \text{ } \& \text{ } \delta_i = 0))}{\sum_i (\mathbb{I}(\delta_i = 0))}. \quad (24)$$

Note that the FPR is also referred to as 1 – specificity. The ROC curve is produced by varying the threshold value that is then used to calculate both the TPR and FPR, and plotting them against one another. The area under this curve is a metric that indicates performance of the model to predict survival outcomes. AUROC=1 implies perfect predictions and AUROC=0.5 implies the contrary. However, our model is formulated such that the variables age and time since starting the survey both form part of Cox's proportional hazard. Furthermore, the Gompertz model is stratified by demographic group. Therefore, in this work, we present a modified ROC curve, which calculates the individuals cumulative hazard at a given time since the start of the survey,  $T_{surv} \in 5, 10, 15$  years, and calculates whether the model correctly predicted an event to occur before or after this time. Note that to do this, we split the ROC population by ages: 45-64 and 65-84. The modified TPR is then calculated via:

$$TPR(\epsilon) = \frac{\sum_i (\mathbb{I}(\delta_i = 1 \text{ } \& \text{ } H(T_i) \geq \epsilon \text{ } \& \text{ } T_i < T_{surv}))}{\sum_i (\mathbb{I}(\delta_i = 1 \text{ } \& \text{ } T_i < T_{surv}))}, \quad (25)$$

and the modified FPR:

$$FPR(\epsilon) = \frac{\sum_i (\mathbb{I}(H(T_i) \geq \epsilon \text{ } \& \text{ } T_i \geq T_{surv}))}{\sum_i (\mathbb{I}(T_i \geq T_{surv}))}. \quad (26)$$

Before presenting any ROC or AUC values, we first present the density distributions of the median posterior systolic  $\Delta$  values for all individuals, split by demographic.

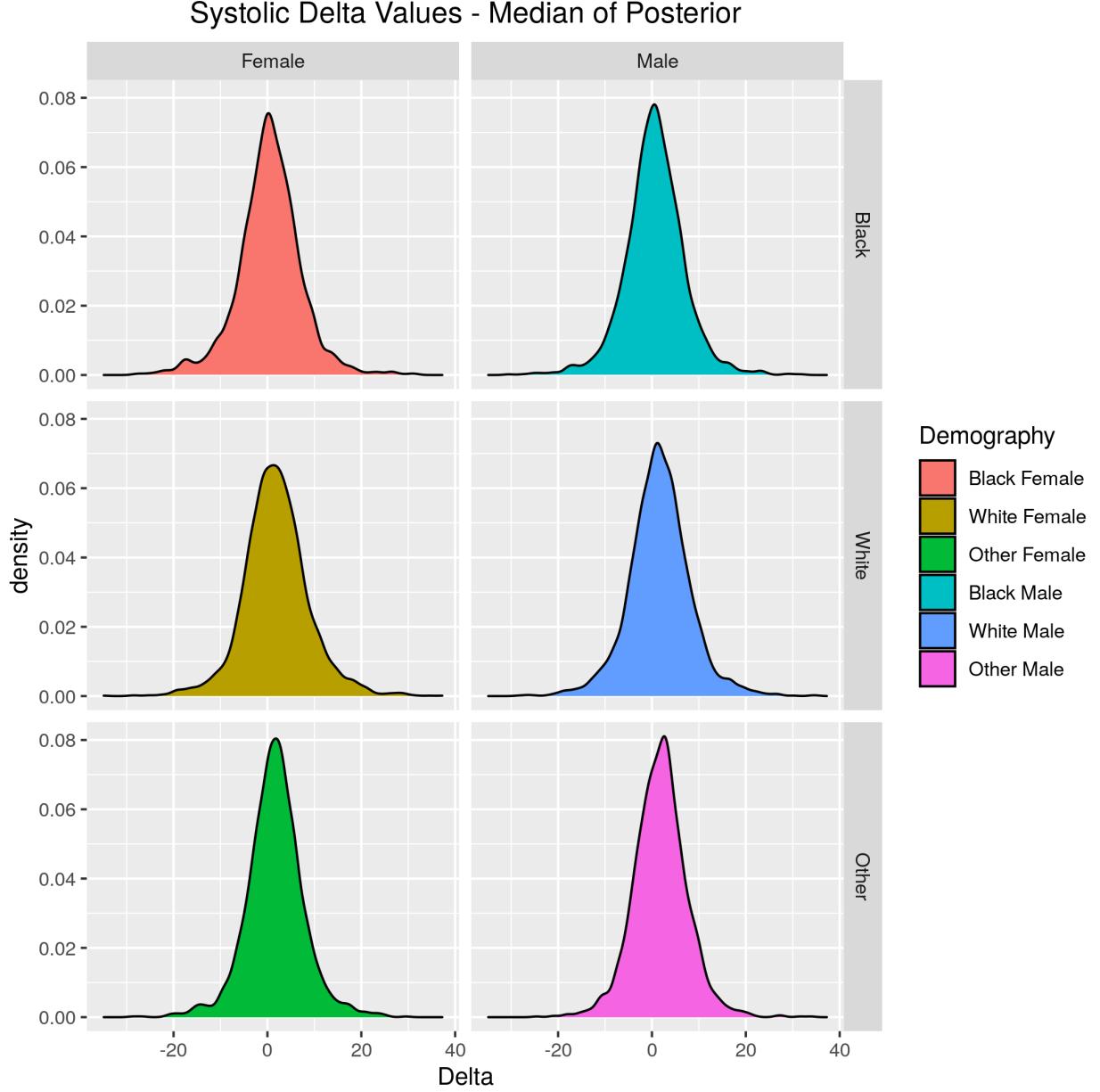


Figure 10: Density of the (median posterior) systolic  $\Delta$  values, per demographic.

Using Welch's ANOVA test, we calculated that  $p < 1 \times 10^{-9}$  for all demographics, including when split between the male and female populations. Figures 11 and 12 show the ROC curves (including the AUC values) of the model, the former using the mean systolic and diastolic blood pressure as covariates in the linear predictor term (in the Cox's proportional hazards component) and the latter using the FRS value instead. By making predictions of the 5, 10 and 15 year survival between the middle aged and old aged sub-groups, for the three different mortality causes, we start to build a picture of the performance of the model. For figure 11, we notice that the AUC value of the predictions for the middle aged compared to the older aged population is higher, independent of the survival year prediction or the different mortality causes. The highest AUC is for the 45-64 year old population with a focus on HA-CVD-CeVD and heart attack-related mortality, for all three

survival year periods. We also note that the TPR seems to start increasing at a faster rate for the population aged 45-64 than the 65-84 group, implying that it is possible to choose a threshold level,  $\epsilon$ , for the survival predictions that could correctly identify people at risk without incorrectly predicting as many people to be at risk of mortality as for the group aged 65-84. The results also reflect that the influence of choosing a 5, 10 or 15 year prediction period does not seem to significantly influence the results. Figure 12 displays similar results to the mean systolic and diastolic model when using the FRS value instead, with the main difference that the predictions of the middle aged group for HA-CVD-CeVD and heart attack-related mortality for 5 year survival seems to be lower than the equivalent in the older group or as compared to the mean blood pressure model. This is caused by a reduced mortality before 5 years for the middle aged population that had their FRS value calculated, where 36, 83 and 213 HA-CVD-CeVD and heart attack-related deaths occurred before 5, 10 and 15 years in this sub-group, respectively. This can be compared to 141, 356 and 952 all-cause deaths in this same sub-group (red-curve in figure 12 and 11). Alternatively, when compared to the full-population (not just those who had the FRS value), the HA-CVD-CeVD and heart attack-related deaths for the middle aged population are 55, 115 and 282 over the 5, 10 and 15 year range, respectively.

Comparison of the ROC curves and AUC values is also presented for the different demographic groups, see figure 13. This figure shows the differences in the prediction performance (w.r.t. the ROC and AUC values) of the full-population model for the 45-64 age population for their 10 year survival outcome, using the mean systolic and diastolic blood pressure model. This plot illustrates that potentially only the all-cause mortality has enough outcomes in each demographic group to separate the ROC curves. The model seems to most accurately predict the 10 year survival outcome of the black and other ethnic groups, as well as the white female demographic as compared to the black female, other female and white male population. To provide insight into this, we also provide the frequency table of deaths for each demographic group, mortality cause and survival year, see table 9 and 10.

To finalise the section on the performance of the models using ROC and AUC values, we present a series of figures that provide ROC curves and AUC values for different linear predictor (Cox's proportional hazards model) covariate configurations. By setting the covariate-specific  $\beta$  parameter values to zero, we can measure the additional prediction performance that adding different covariates provides to the model. In figures 14 to 19, we present three main formulations: using only the systolic and diastolic  $\Delta$  terms, using the mean systolic and diastolic (figures 14-16) or the FRS value (figures 17-19) terms as well as the systolic and diastolic  $\Delta$  terms (figures 14-19) and, finally, using the systolic mean (figures 14-16) or FRS value only (figures 17-19). Figures 14, 15 and 16 apply to the full-population with the models trained on HA-CVD-CeVD and heart-attack related mortality, all-cause, and other mortality, respectively. Figures 17, 18 and 19 apply to the population with an FRS value, with the models trained on HA-CVD-CeVD and heart-attack related mortality, all-cause, and other mortality, respectively. The first thing to note as a commonality between all these different figures is that the use of the long-term variability,  $\Delta$ , in the model has comparable performance with that of using the systolic mean or FRS values only. Additionally, where the number of deaths permits for prediction, the use of both the  $\Delta$  and mean/FRS values results in higher AUC values than using models that use one or the other. Finally, the use of the FRS value consistently under-performs the mean diastolic and systolic blood pressure-based model.

## Results - Exploring $\Delta$ Directionality

This section of the appendix is to explore whether the directionality of the difference in clinic-home blood pressure (represented through the non-absolute value of the  $\Delta$  covariate) may have an influence on the survival outcome in the population. In the work presented in this article,  $\Delta$  is the absolute value of the differences in the means of the blood pressure measurements at the clinic and at home, respectively, for both diastolic and systolic blood pressure. By 'directionality', we refer to whether the difference between the clinic and home mean measurements are positive or negative. Figure 20 shows the clinic-home directionalities, split by demographic group, indicating no significant difference between the different demographic groups. There is a general trend that the directionality for systolic and diastolic blood pressure is more likely to be the same than opposite.

In order to explore whether the directionality of the clinic-home measurements influences survival outcome, we will use a combination of Kaplan-Meier curves and Cox's proportional hazards regression. The latter will

### Mean Blood Pressure Model

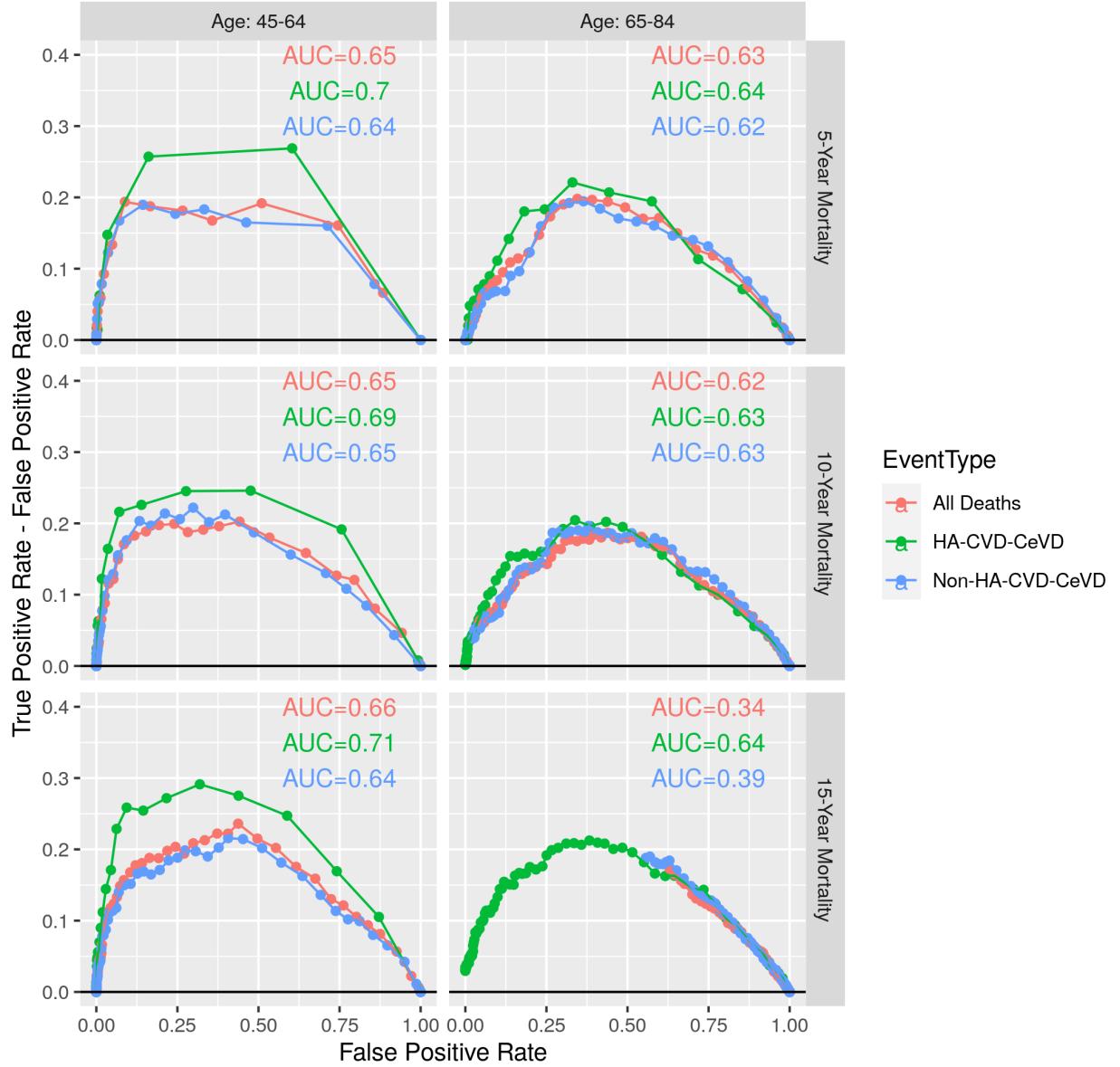


Figure 11: ROC curves for the model that used mean systolic and diastolic blood pressure as covariates in the linear predictor, stratified by the event type (cause of mortality). The columns split two groups in the population: those who start the survey aged between 45 to 64 and 65-84 years old. The rows split the model predictions between 5, 10 and 2 year survival.

### Framingham Risk Score Model

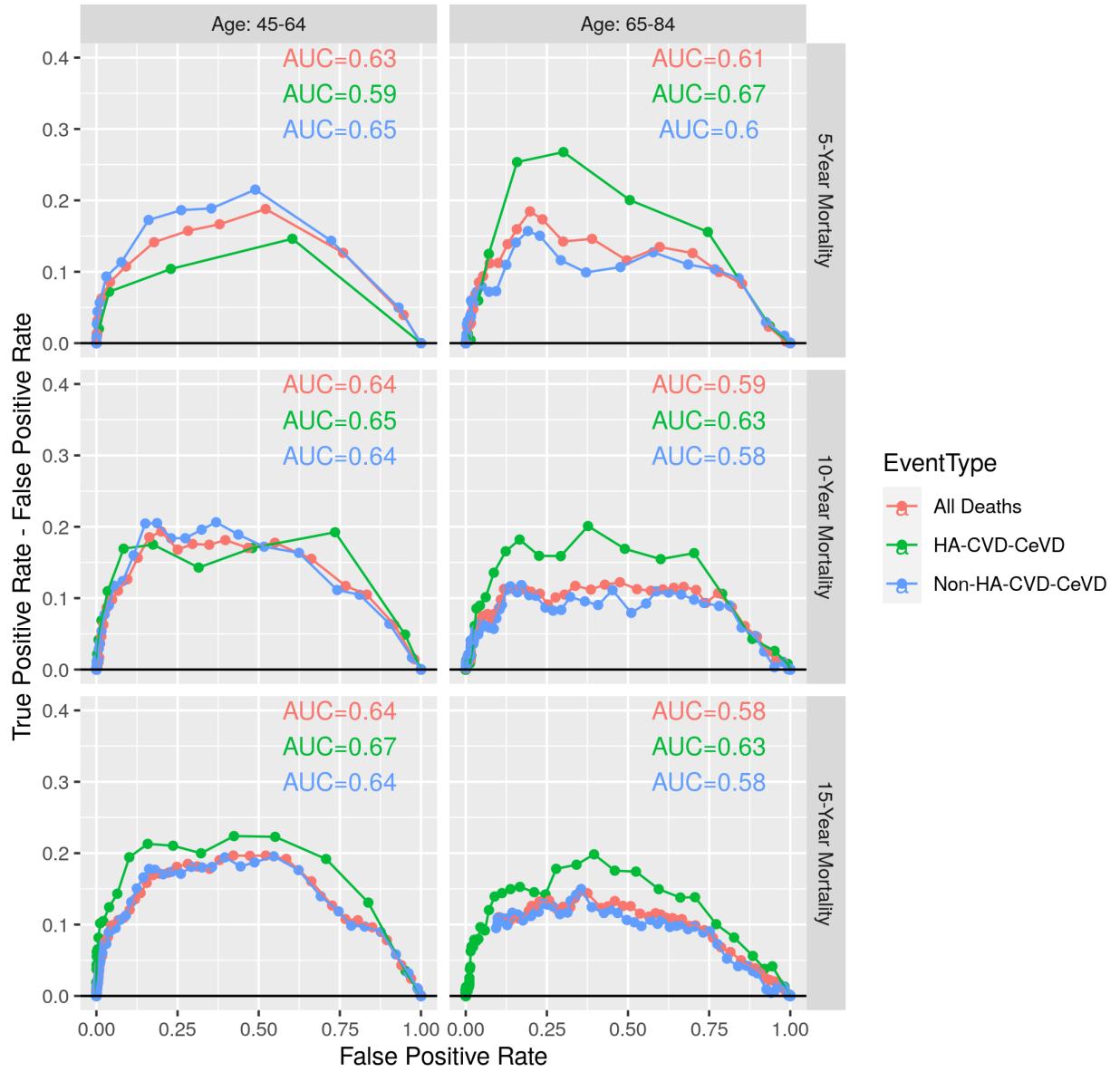


Figure 12: ROC curves for the model that used the FRS value as covariates in the linear predictor, stratified by age group and the number of years the survival outcome was predicted since participant starting the survey.

## Performance by Demographic

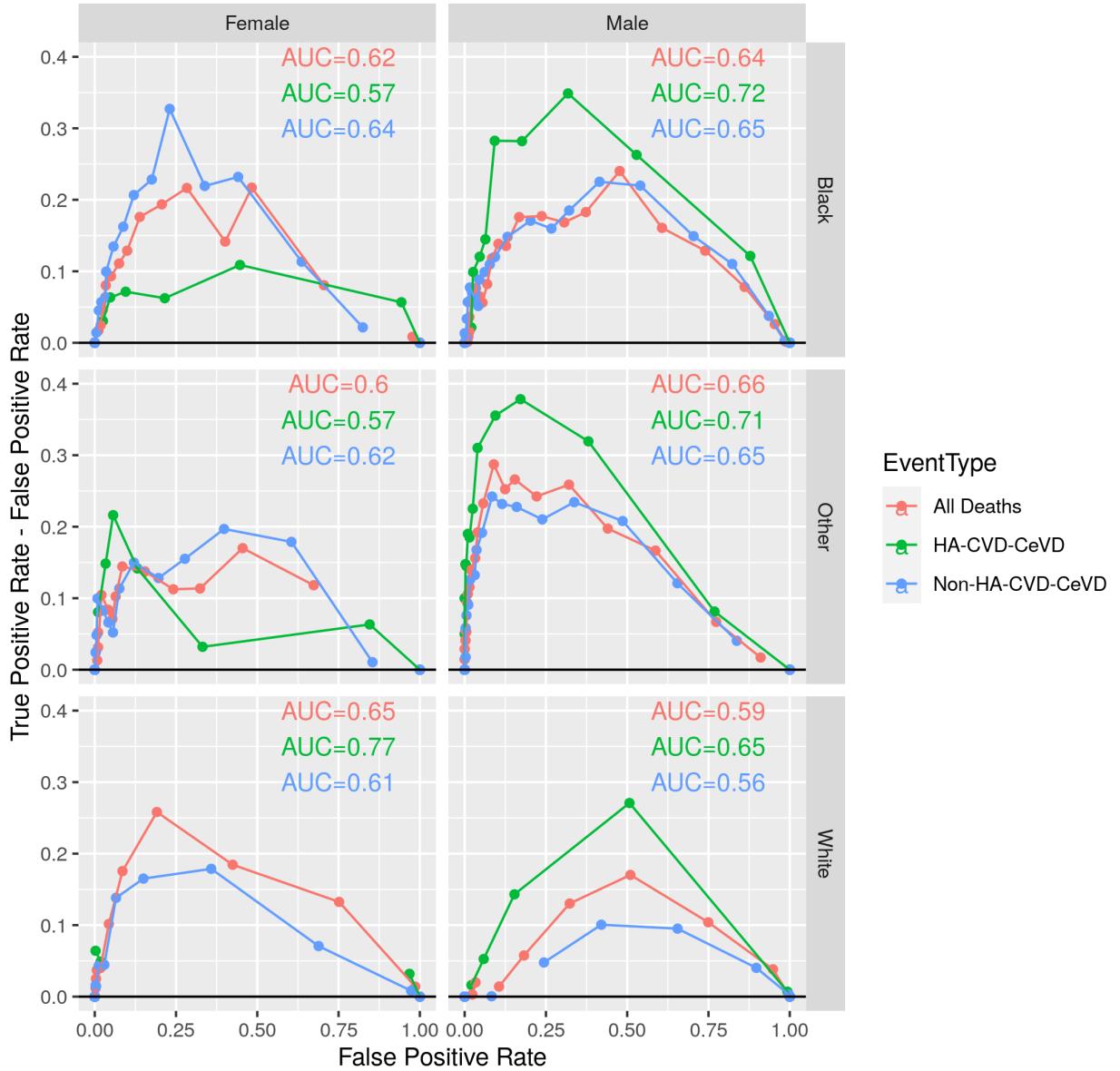


Figure 13: ROC curves stratified by the different demographic groups used in this research. The point and line colours represent the different event types that were used to predict on.

Table 9: Frequency table of the population aged 45-64 for the N-year survival outcomes as separated by demographic group and mortality cause.

Year	EventType	Ethnicity	Gender	Deaths
5-Year Mortality	HA-CVD-CeVD	Black	Female	15
5-Year Mortality	All Deaths	Black	Female	44
5-Year Mortality	HA-CVD-CeVD	White	Female	9
5-Year Mortality	All Deaths	White	Female	33
5-Year Mortality	HA-CVD-CeVD	Other	Female	9
5-Year Mortality	All Deaths	Other	Female	24
5-Year Mortality	HA-CVD-CeVD	Black	Male	6
5-Year Mortality	All Deaths	Black	Male	25
5-Year Mortality	HA-CVD-CeVD	White	Male	7
5-Year Mortality	All Deaths	White	Male	24
5-Year Mortality	HA-CVD-CeVD	Other	Male	6
5-Year Mortality	All Deaths	Other	Male	21
10-Year Mortality	HA-CVD-CeVD	Black	Female	23
10-Year Mortality	All Deaths	Black	Female	87
10-Year Mortality	HA-CVD-CeVD	White	Female	21
10-Year Mortality	All Deaths	White	Female	67
10-Year Mortality	HA-CVD-CeVD	Other	Female	19
10-Year Mortality	All Deaths	Other	Female	63
10-Year Mortality	HA-CVD-CeVD	Black	Male	17
10-Year Mortality	All Deaths	Black	Male	67
10-Year Mortality	HA-CVD-CeVD	White	Male	13
10-Year Mortality	All Deaths	White	Male	62
10-Year Mortality	HA-CVD-CeVD	Other	Male	10
10-Year Mortality	All Deaths	Other	Male	44
15-Year Mortality	HA-CVD-CeVD	Black	Female	37
15-Year Mortality	All Deaths	Black	Female	132
15-Year Mortality	HA-CVD-CeVD	White	Female	44
15-Year Mortality	All Deaths	White	Female	139
15-Year Mortality	HA-CVD-CeVD	Other	Female	35
15-Year Mortality	All Deaths	Other	Female	105
15-Year Mortality	HA-CVD-CeVD	Black	Male	26
15-Year Mortality	All Deaths	Black	Male	113
15-Year Mortality	HA-CVD-CeVD	White	Male	29
15-Year Mortality	All Deaths	White	Male	117
15-Year Mortality	HA-CVD-CeVD	Other	Male	18
15-Year Mortality	All Deaths	Other	Male	67

Table 10: Frequency table of the population aged 65-84 for the N-year survival outcomes as separated by demographic group and mortality cause.

Year	EventType	Ethnicity	Gender	Deaths
5-Year Mortality	HA-CVD-CeVD	Black	Female	35
5-Year Mortality	All Deaths	Black	Female	87
5-Year Mortality	HA-CVD-CeVD	White	Female	92
5-Year Mortality	All Deaths	White	Female	253
5-Year Mortality	HA-CVD-CeVD	Other	Female	15
5-Year Mortality	All Deaths	Other	Female	45
5-Year Mortality	HA-CVD-CeVD	Black	Male	19
5-Year Mortality	All Deaths	Black	Male	49
5-Year Mortality	HA-CVD-CeVD	White	Male	55
5-Year Mortality	All Deaths	White	Male	138
5-Year Mortality	HA-CVD-CeVD	Other	Male	10
5-Year Mortality	All Deaths	Other	Male	29
10-Year Mortality	HA-CVD-CeVD	Black	Female	56
10-Year Mortality	All Deaths	Black	Female	153
10-Year Mortality	HA-CVD-CeVD	White	Female	187
10-Year Mortality	All Deaths	White	Female	501
10-Year Mortality	HA-CVD-CeVD	Other	Female	37
10-Year Mortality	All Deaths	Other	Female	105
10-Year Mortality	HA-CVD-CeVD	Black	Male	50
10-Year Mortality	All Deaths	Black	Male	121
10-Year Mortality	HA-CVD-CeVD	White	Male	144
10-Year Mortality	All Deaths	White	Male	357
10-Year Mortality	HA-CVD-CeVD	Other	Male	44
10-Year Mortality	All Deaths	Other	Male	91
15-Year Mortality	HA-CVD-CeVD	Black	Female	79
15-Year Mortality	All Deaths	Black	Female	231
15-Year Mortality	HA-CVD-CeVD	White	Female	251
15-Year Mortality	All Deaths	White	Female	694
15-Year Mortality	HA-CVD-CeVD	Other	Female	56
15-Year Mortality	All Deaths	Other	Female	170
15-Year Mortality	HA-CVD-CeVD	Black	Male	76
15-Year Mortality	All Deaths	Black	Male	185
15-Year Mortality	HA-CVD-CeVD	White	Male	225
15-Year Mortality	All Deaths	White	Male	589
15-Year Mortality	HA-CVD-CeVD	Other	Male	57
15-Year Mortality	All Deaths	Other	Male	139

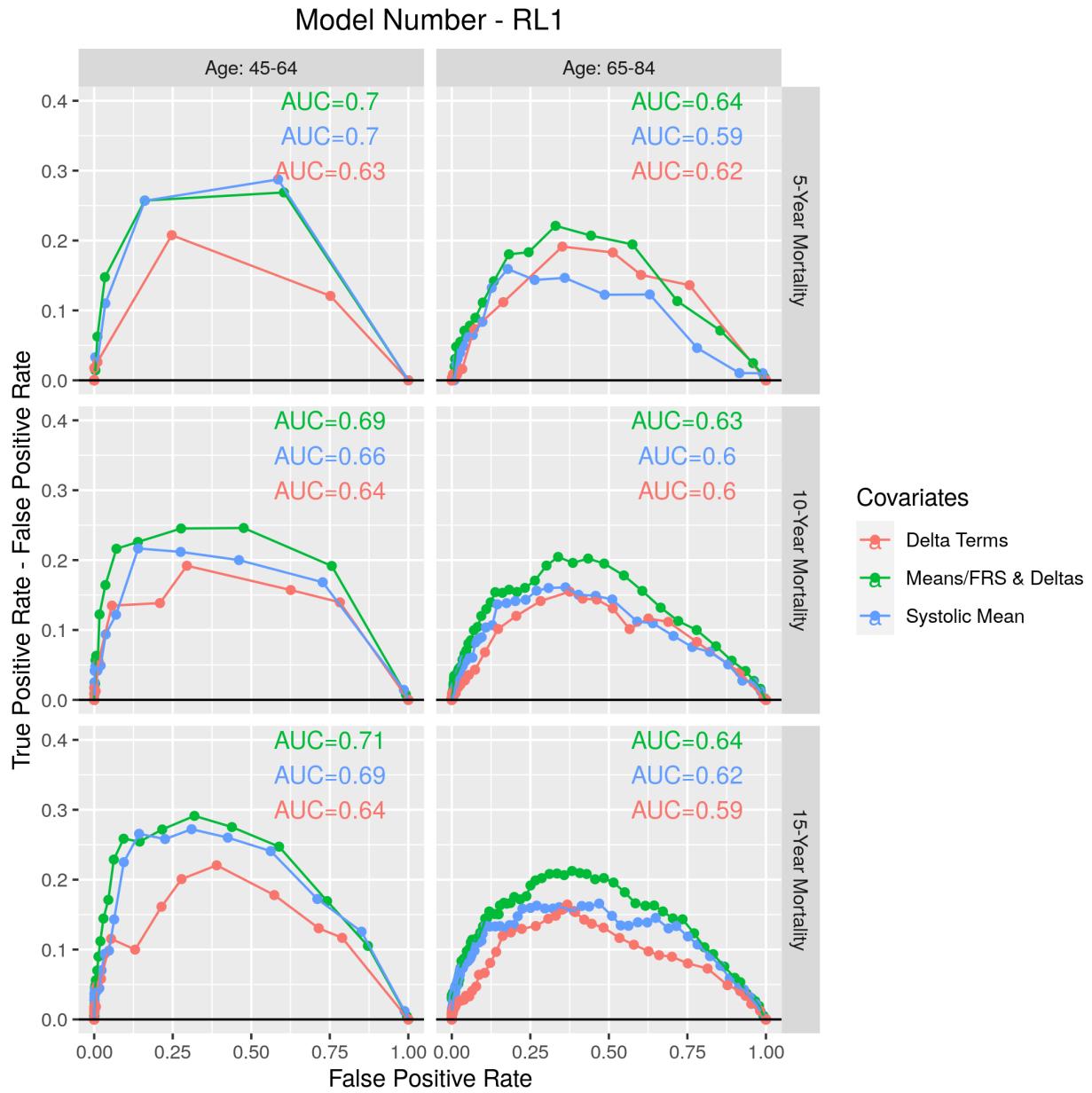


Figure 14: ROC curves for the mean systolic and diastolic model, looking specifically at HA-CVD-CeVD and heart attack-related deaths, stratified by age group and the number of years the survival outcome was predicted since participant starting the survey. The colour of the points and lines represents the different linear predictor covariate models possible.

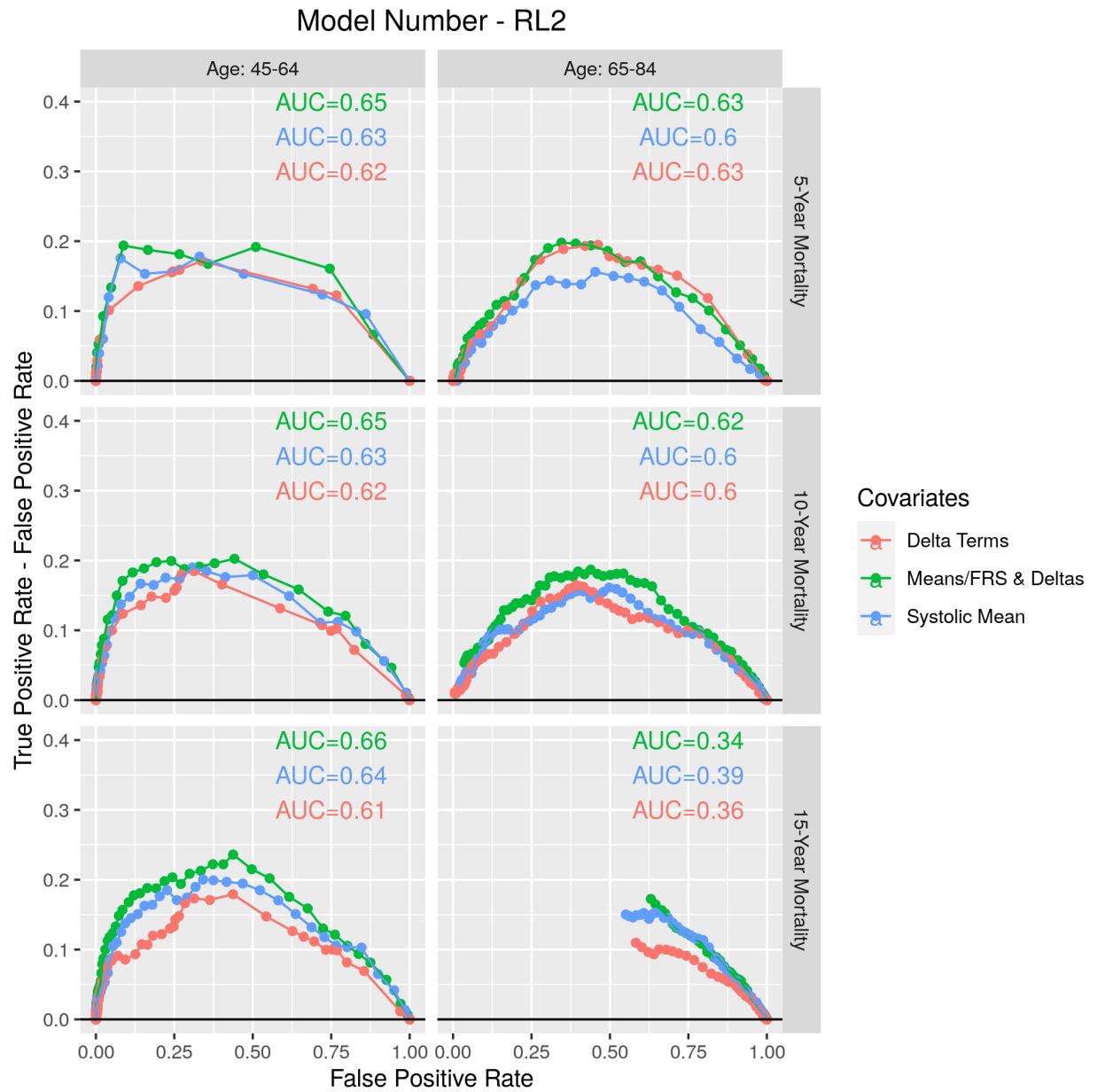


Figure 15: ROC curves for the mean systolic and diastolic model, looking at all-cause deaths, stratified by age group and the number of years the survival outcome was predicted since participant starting the survey. The colour of the points and lines represents the different linear predictor covariate models possible.

### Model Number - RL20th

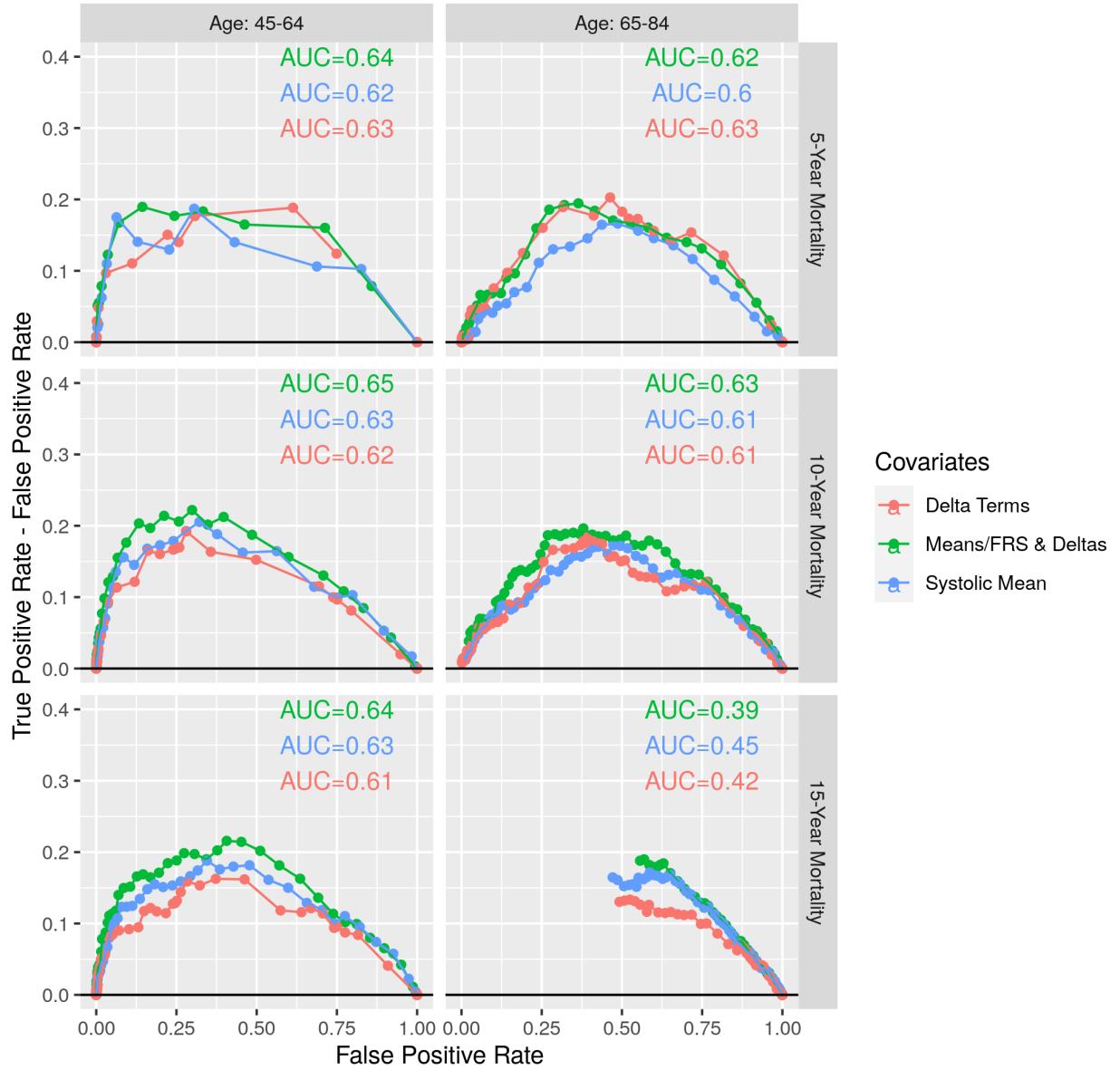


Figure 16: ROC curves for the mean systolic and diastolic model, looking non-HA-CVD-CeVD and heart attack-related deaths, stratified by age group and the number of years the survival outcome was predicted since participant starting the survey. The colour of the points and lines represents the different linear predictor covariate models possible.

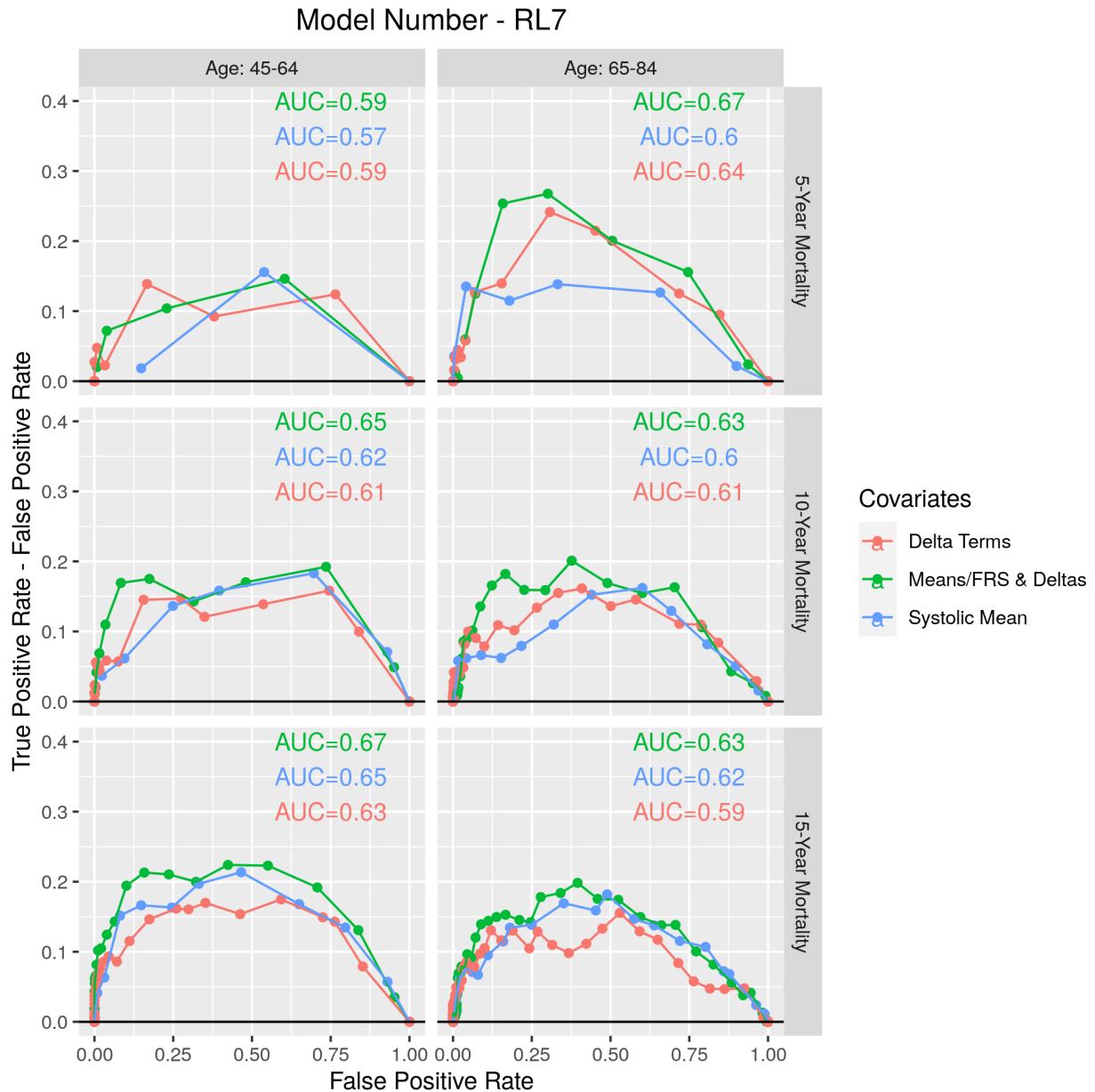


Figure 17: ROC curves for the FRS-based model, looking specifically at HA-CVD-CeVD and heart attack-related deaths, stratified by age group and the number of years the survival outcome was predicted since participant starting the survey. The colour of the points and lines represents the different linear predictor covariate models possible.

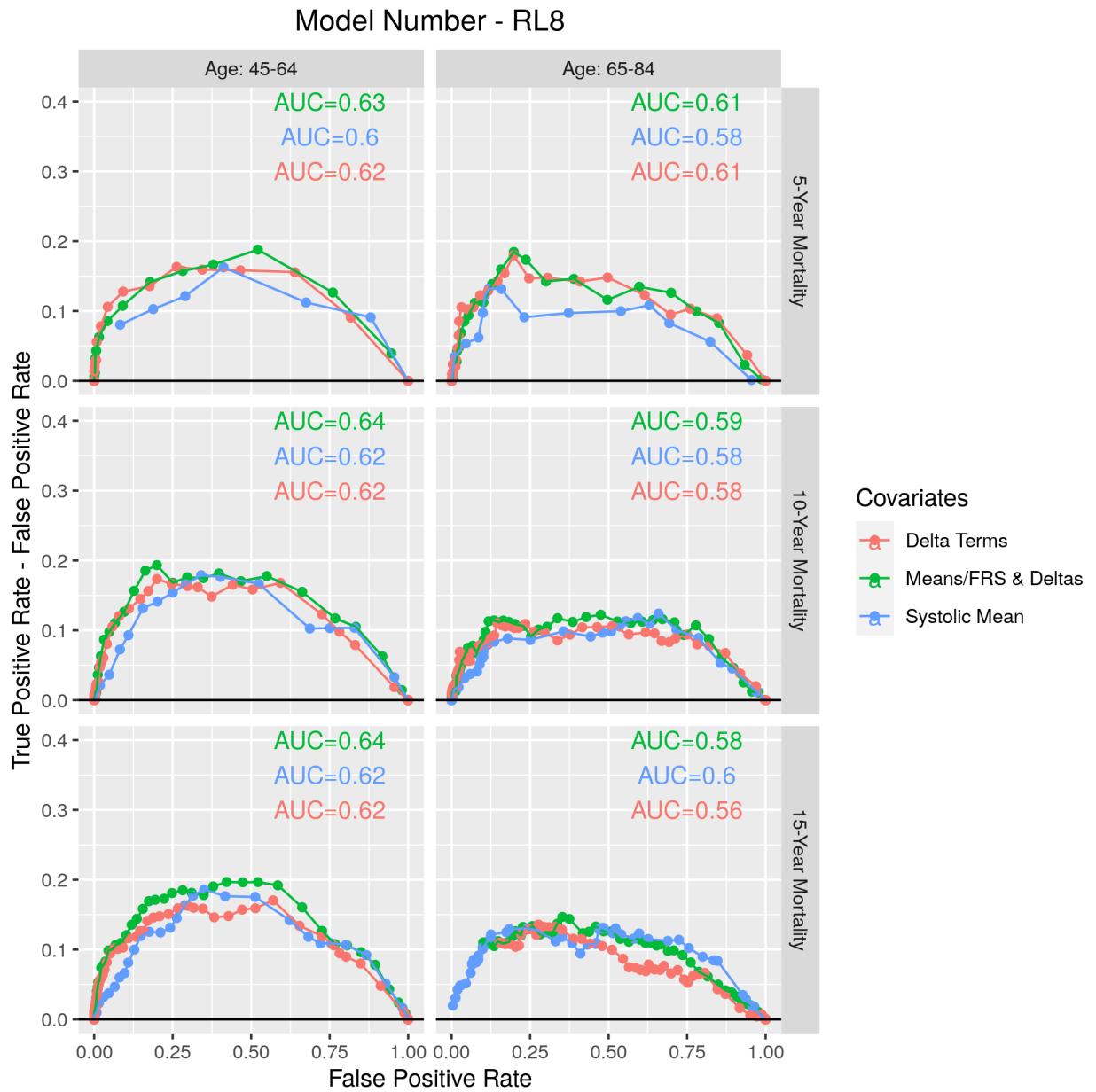


Figure 18: ROC curves for the FRS-based model, looking at all-cause deaths, stratified by age group and the number of years the survival outcome was predicted since participant starting the survey. The colour of the points and lines represents the different linear predictor covariate models possible.

### Model Number - RL80th

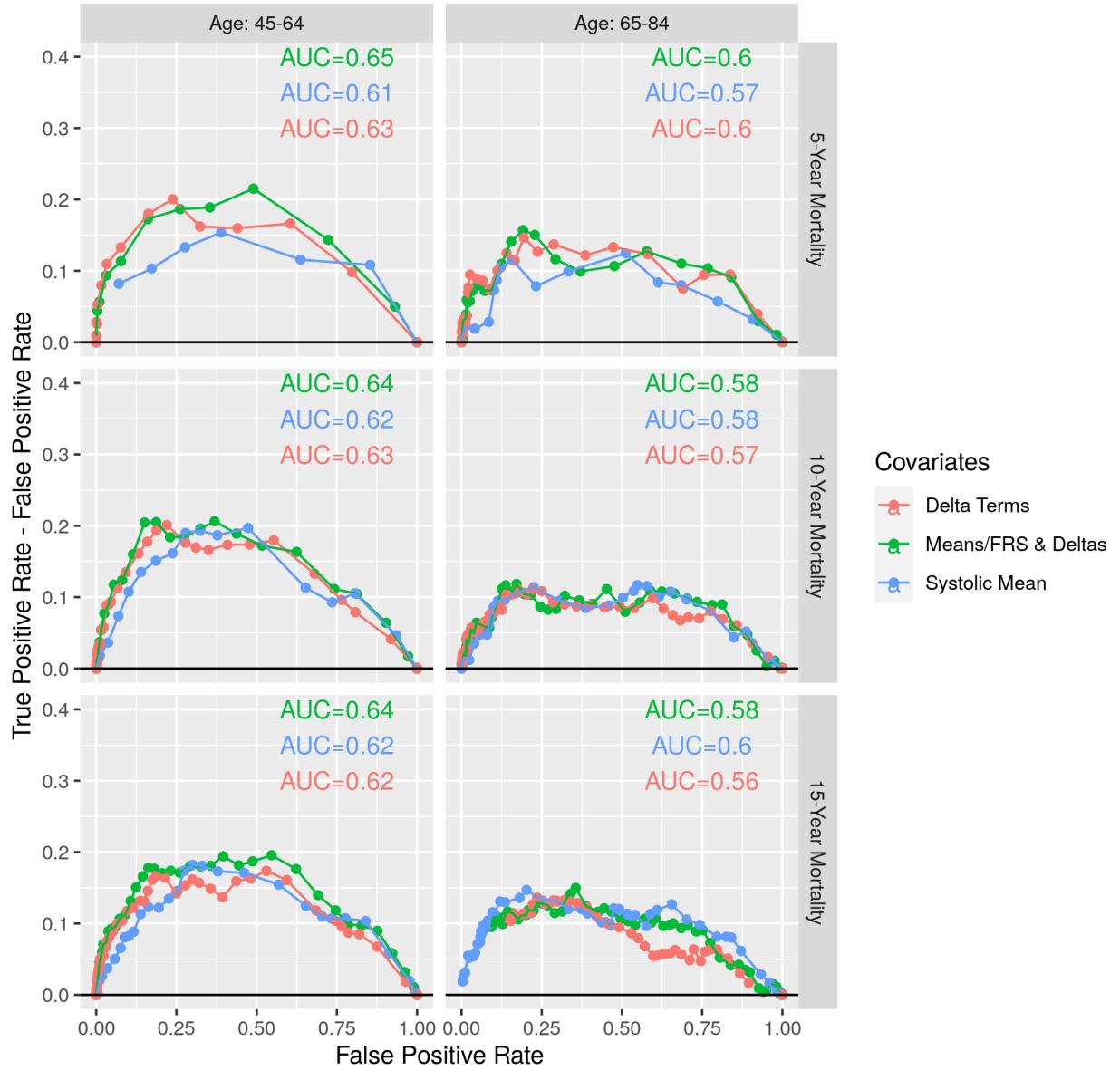


Figure 19: ROC curves for the FRS-based model, looking at non-HA-CVD-CeVD and heart attack-related deaths, stratified by age group and the number of years the survival outcome was predicted since participant starting the survey. The colour of the points and lines represents the different linear predictor covariate models possible.

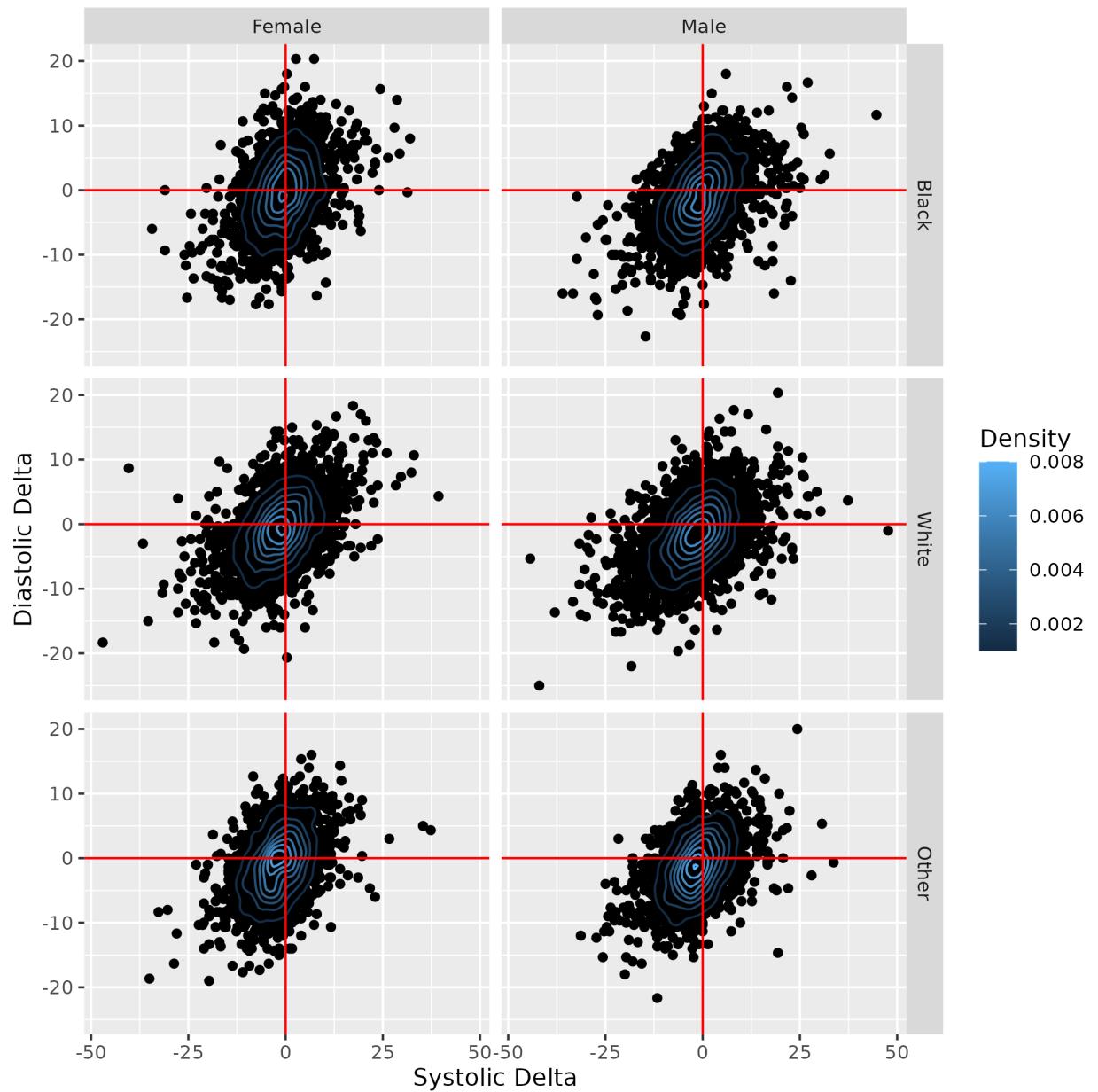


Figure 20: The range of the non-absolute  $\Delta$  values in the systolic and diastolic blood pressure measurements, split by demographic group. This reflects the differences between the average measurements at the clinic and at home.

Table 11: Parameters for distribution of blood pressure, for the full population

covariate	coef	exp(coef)	se(coef)	z	Pr(> z )
DeltaRegionSys. -ve, Dys. +ve	0.0468181	1.0479314	0.0738547	0.6339213	0.5261322
DeltaRegionSys. +ve, Dys. -ve	-0.0021526	0.9978497	0.0688467	-0.0312669	0.9750567
DeltaRegionSys. +ve, Dys. +ve	-0.0558931	0.9456403	0.0634611	-0.8807454	0.3784556
GenderMale	-0.4125926	0.6619319	0.0498157	-8.2823782	0.0000000
EthnicityWhite	-0.2478299	0.7804927	0.0613030	-4.0427047	0.0000528
EthnicityOther	-0.1695250	0.8440656	0.0745991	-2.2724800	0.0230575
age	0.0995937	1.1047220	0.0019693	50.5727891	0.0000000

implement a simple Maximum Likelihood Estimation (MLE) method based on summary statistics of the Bayesian posterior blood pressure values, not the Bayesian-HMC method applied elsewhere in this article. The Kaplan-Meier curve is a plot of the change in survival probability of a population in time since the start of a survey/census. The survival distribution is calculated using

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right), \quad (27)$$

for  $d_j$  the number of individuals who die within the time interval  $t_j$  and  $r_j$  the population that are alive (at risk of death) and not censored. Greenwood's formula is used to calculate the variance of the Kaplan-Meier estimation

$$\hat{\sigma}(t)^2 = \hat{S}(t)^2 \sum_{t_j \leq t} \left( \frac{d_j}{r_j(r_j - d_j)} \right). \quad (28)$$

The  $100(1-\alpha)\%$  confidence intervals of the Kaplan-Meier estimate are assumed to be normally distributed

$$\hat{S}(t) \pm z_{1-\alpha/2} \hat{\sigma}(t). \quad (29)$$

Figures 21 and 22 show the Kaplan-Meier estimates for the full NHANES population for HA-CVD-CeVD mortality, split by demographic group, for the age range 45-64 and 65-84, respectively. The survival probability of the older population decreases faster in time than the middle-aged (45-64) population.

By splitting the populations into the respective regions of the  $\Delta$  directionality, we can use the different Kaplan-Meier plots to try to identify differences in the survival outcomes. Figures 23 and 24 show the Kaplan-Meier estimates for the full NHANES population for HA-CVD-CeVD mortality, split by  $\Delta$  directionality and demographic group, for the age range 45-64 and 65-84, respectively. With the broad confidence intervals, all of the different  $\Delta$  directionality regions overlap, for all demographic groups and both age range groups (where  $\hat{S}(t) \neq 1$ ).

We further quantify this insignificant relationship between  $\Delta$  directionality and survival outcome via the use of a Cox's Proportional Hazards (CPH) model. Via the use of the 'coxph' function in the 'survival' R package, we fit (using MLE) a CPH model. The covariates used in the model are  $\Delta$  directionality region, gender, ethnicity and age. Table 11 shows the summary of the model fit, which reflects that being in different  $\Delta$  directionality regions has a non-statistically significant influence on survival outcomes. As shown in the remainder of this article, ethnicity, gender and age are shown to have statistically significant effects on survival outcome.

Finally, we wish to confirm that the performance of the model does not depend on the directionality of  $\Delta$ . Figure 25 plots the AUC values of 10-year HA-CVD-CeVD mortality for the all-covariate mean blood pressure-based model trained on the full NHANES population, split by the two age-ranges (45-64 and 65-84) and demographic groups. There is no clear trend between mode performance for the different regions of  $\Delta$  directionality.

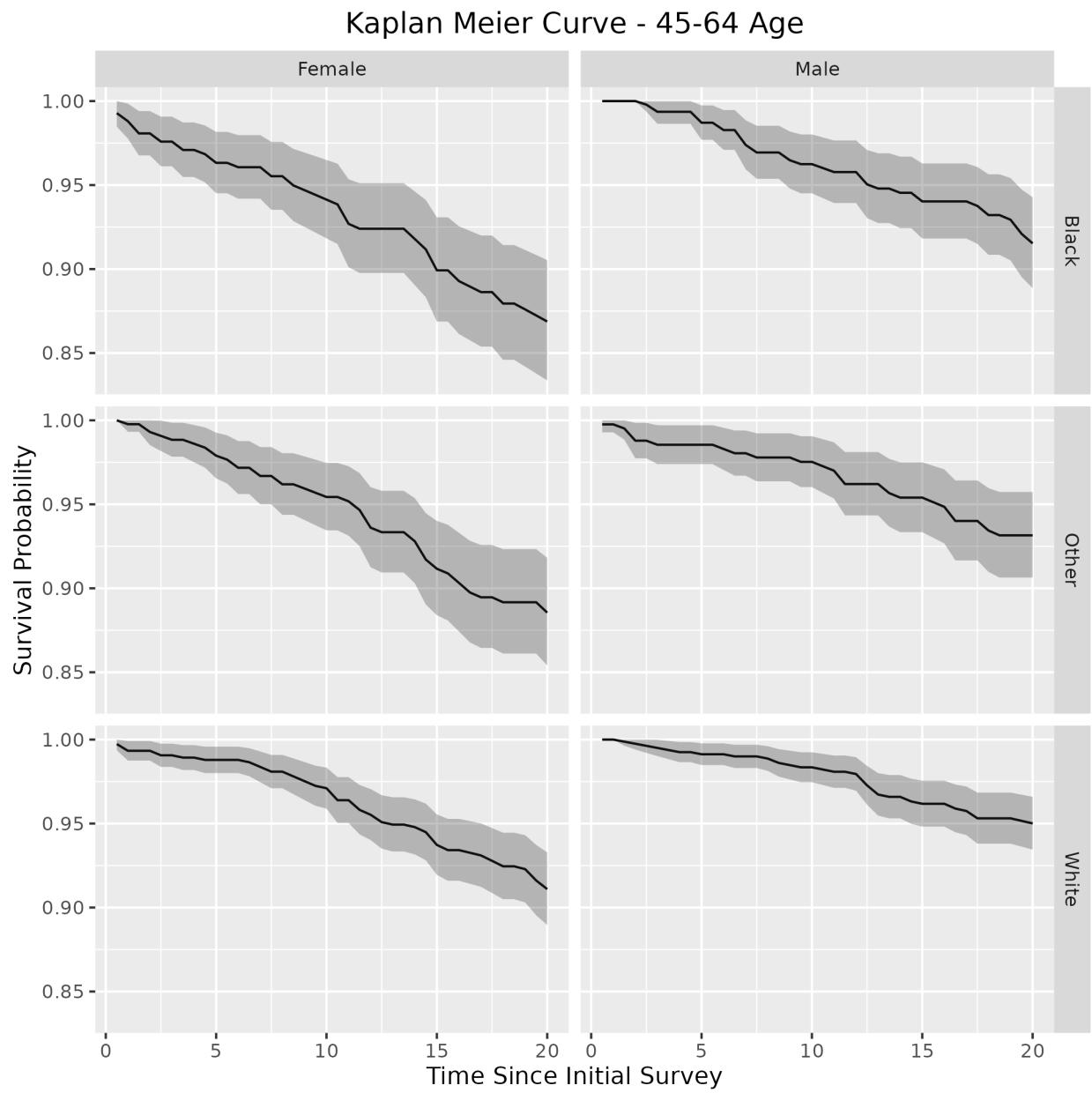


Figure 21: Kaplan-Meier plots of the full-population for HA-CVD-CeVD mortality, for ages between 45-64, split by demographic group.

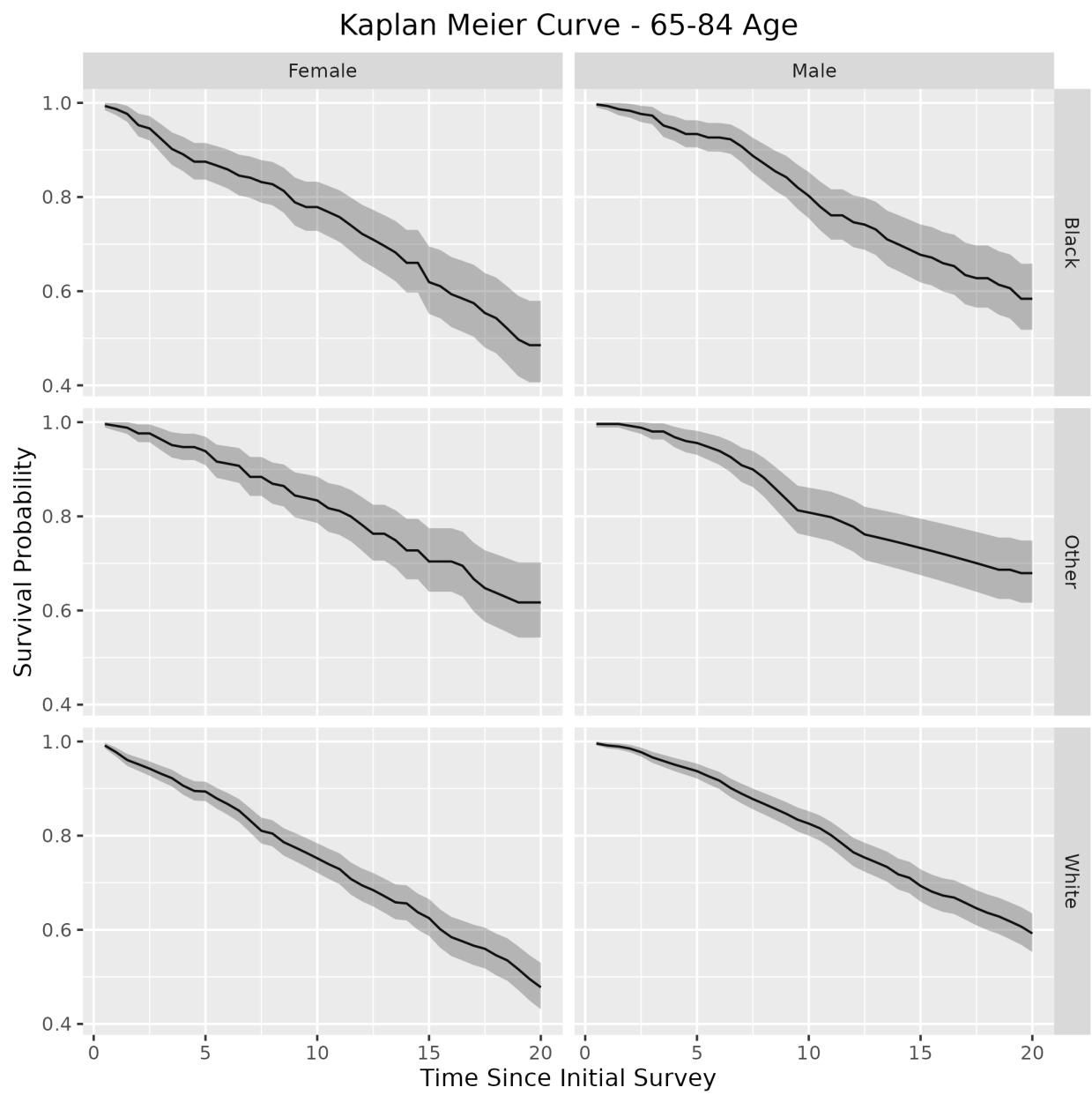


Figure 22: Kaplan-Meier plots of the full-population for HA-CVD-CeVD mortality, for ages between 65-84, split by demographic group.

### Age 45-64 Kaplan Meier Curve

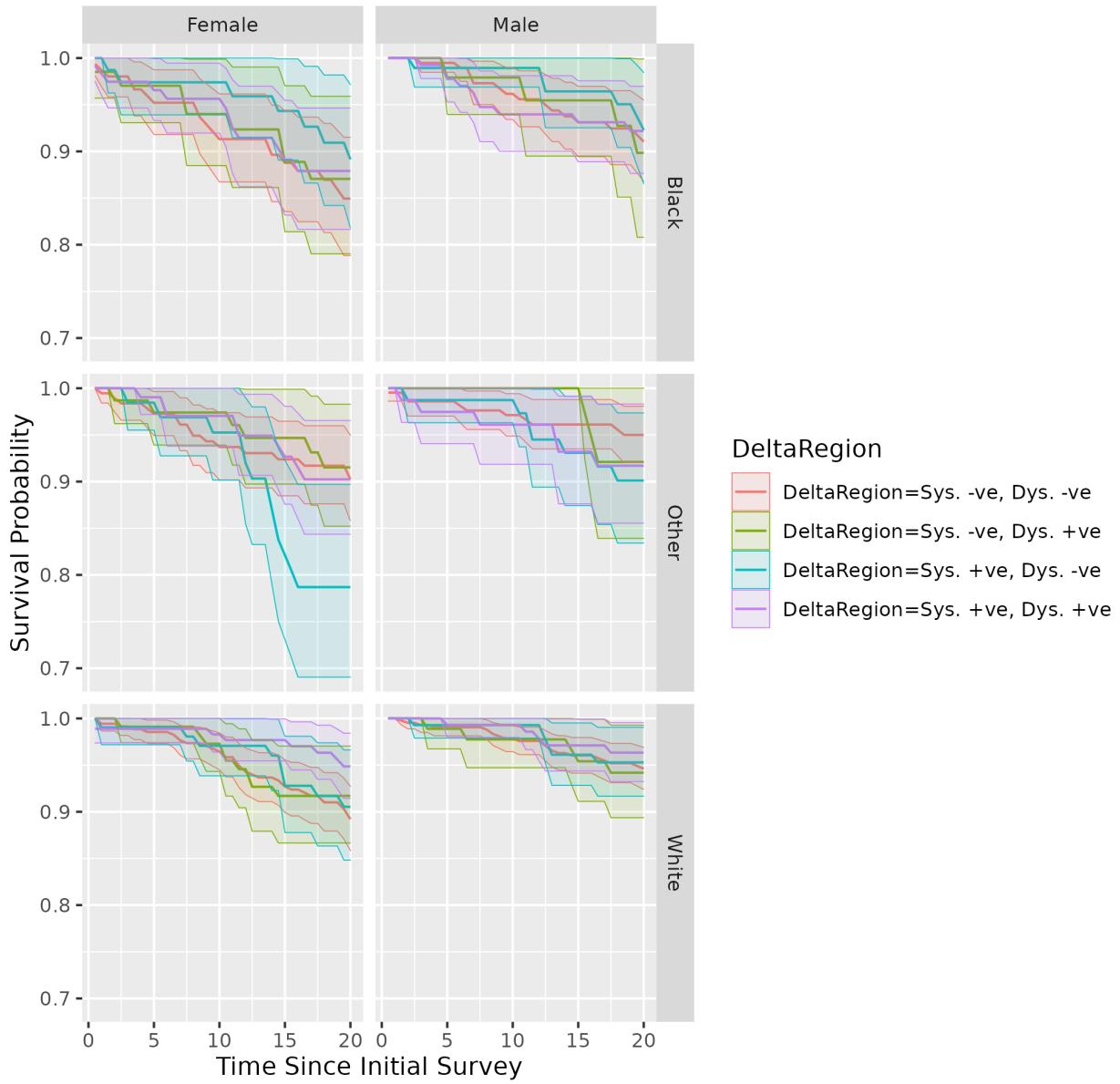


Figure 23: Kaplan-Meier plots of the full-population for HA-CVD-CeVD mortality, for ages between 45-64, split by demographic group and region in systolic-diastolic  $\Delta$  space.

### Age 65-84 Kaplan Meier Curve

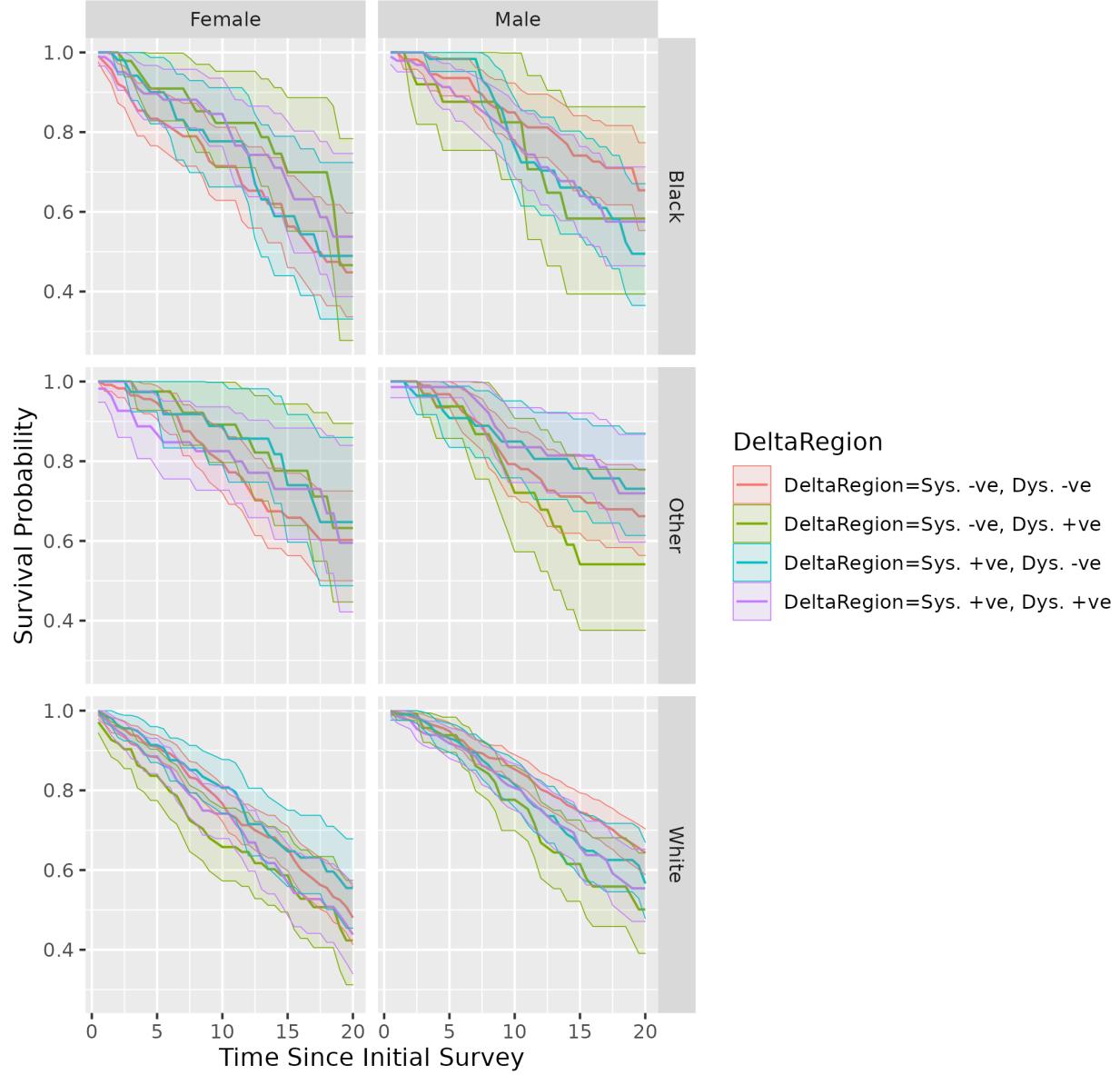


Figure 24: Kaplan-Meier plots of the full-population for HA-CVD-CeVD mortality, for ages between 65-84, split by demographic group and region in systolic-diastolic  $\Delta$  space.

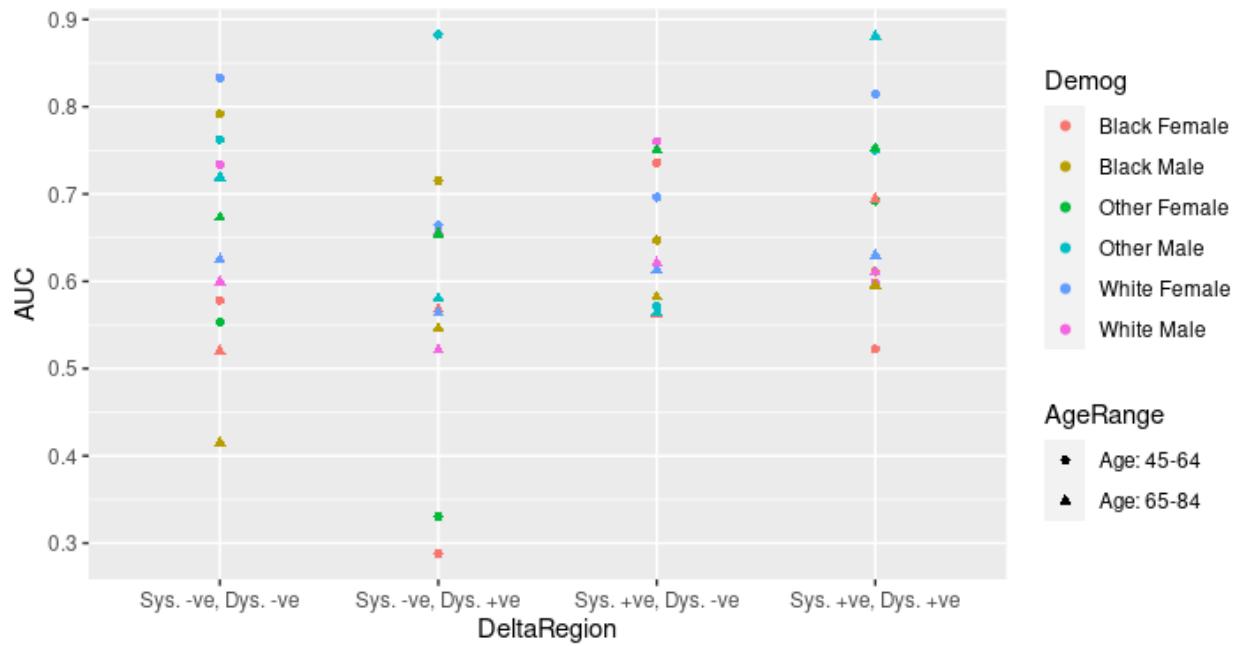


Figure 25: AUC values of the full-population, all covariate model with the mean blood pressure model, based on HA-CVD-CeVD 10-year mortality.