Supplemental Material

Prognostic factors for premature cardiovascular disease mortality in Malaysia: A modelling approach using semi-parametric and parametric survival analysis with National Health and Morbidity Survey linked mortality data

List of Supplementary:

- Description of the survival analysis used in this study
- Calculation of age-standardized incidence rate; overall (**Table S1**) and by sex (**Table S1)**.
- **Table S3**. Baseline characteristics of samples by each NHM survey
- **Figure S1**. Schoenfeld residual test and plot
- **Figure S2**. Schoenfeld residual test and plot after stratifying the Cox proportional hazard model by age of entry
- Sensitivity analysis: treated imbalance data
- **Figure S3**. Parametric distribution fittings for premature CVD mortality data after treatment using under-sampling method on imbalanced data
- **Table S4**. Sensitivity analysis compared between imbalanced data and original data: univariable analysis
- **Table S5**. Sensitivity analysis compared between imbalanced data and original data: adjusted models

Description of survival analysis method used in this study

Semi-parametric survival analysis

The semi-parametric Cox PH model was first employed for univariable and multivariable regression analyses of premature CVD mortality modelling using ***survival*** package[1] in R software. The Cox PH model is usually written in terms of the hazard model where this model gives an expression for the hazard at time t for an individual with a given specification of a set of explanatory variables[2]. The important property of Cox model is that the baseline hazard, $h_0(t)$ is unspecific (we do not require to assume any probability distribution, shape or form of the survival times)[2] . This property makes the Cox model a semi-parametric. A simple specification of the model is as follows[2]:

$$h_i(t) = h_0(t)e^{\sum_{i=1}^{p}\beta_i X_i} \quad (1)$$

In Equation 1, $h_i(t)$ denotes a hazard or risk at time t pertaining to the individual *i.* Basically, this formula is derived by the product of two quantities. The first of these, $h_0(t)$ signifies a baseline hazard function. The second quantity is the exponential expression, *e* to the linear sum of $\beta_i X_i$ where *X* represent a collection or a vector of explanatory variables and β is a vector of coefficients to be estimated. The hazard ratio (HR) is obtained by dividing both sides of equation 1 by $h_0(t)$ and taking log, as follows:

$$\ln\left[\frac{h_i(t)}{h_0(t)}\right] = b_1X_1 + b_2X_2 + , \ldots\ldots + b_kX_k \quad (2)$$

Using algebra involving exponentials, the HR formula simplifies to the exponential expression (Equation 3). Thus, the HR is computed by exponentiating the sum of each β_i "hat" time the difference between X_i* (predictor for one individual) and X_i (predictors for the other individuals).

$$\widehat{HR} = \exp\left[\sum_{i=1}^{p} \beta_i (X_i^* - X_i)\right] \quad (3)$$

In this study, HR and their respective 95% confidence intervals (95% CI) were used to assess the risk factors. The model building involved several steps: 1) univariable Cox regression (unadjusted analysis), 2) variable selection using the backward method with Akaike's information criteria (AIC) to determine the best-fitting model, 3) assessing the assumption of hazard proportionality for the chosen model, and 4) checking for interactions between prognostic factors. The PH assumption necessitates a constant relationship between the outcome and the covariate vectors over time[2]. To assess the PH assumption, we used the *cox.zph* function from the ***survival*** package[1]. This function calculates scaled Schoenfeld residuals for each covariate and includes a global test to assess their correlation with time. A significant p-value from the global test indicates a violation of the PH assumption in our model. The Schoenfeld residual plots were also used to visualized the pattern of hazard ratios over time for each covariate. If the PH assumption was violated, we proceeded to conduct the stratified Cox regression model. Stratified Cox models extend standard Cox models to accommodate covariates with non-proportional hazards.

Parametric survival analysis

Additionally, we explored an alternative survival analysis approach by conducting a fully parametric survival analysis based on selected covariates from a multivariable Cox PH model. In the parametric analysis, the outcome (survival time) is assumed to follow a known distribution whose probability density function, *f (t)* can be expressed in terms of unknown parameters[2]. Once a probability density function is specified for survival time, the corresponding survival function, S(t) and hazard functions, h(t) can be determined. The relationship between S(t) and h(t) can be expressed equivalently in either of two calculus formulae[2];

$$S(t) = \exp\left[-\int_0^t h(u)du\right] \quad (4)$$

$$h(t) = \frac{-d[S(t)]/dt}{S(t)} \quad (5)$$

Where S(t) equals the exponential of the negative integral of the hazard function between integration limits of 0 and t (Equation 4). The second formula describes how the hazard function h(t) can be written in terms of a derivative involving the survivor function. This formula says that h(t) equals minus the derivative of S(t) with respect to t divided by S(t) (Equation 5). Finally, the probability density function, f(t) can be expressed as the product of the hazard and the survival function (Equation 6),

$$f(t) = h(t)S(t) \quad (6)$$

The effect sizes of parametric model can be expressed either as proportional hazards or accelerated failure time (AFT) models[2,3]. In the AFT model, the survival function is assumed to have the same shape for all individuals, and explanatory variables are assumed to affect survival by changing the speed at which individuals move on the curve. In simpler terms, some individuals move along the survival timeline faster or slower than others, based on certain factors[3]. In contrast to a proportional hazards model, where hazard is multiplied by a factor, in this model, the survival time is multiplied by a factor. This can be expressed as $T = \psi T_0$ or $T = T_0 e^{-\beta}$. This formula gives the survival function[3];

$$S(t;x) = S_0\left(e^{\beta x}t\right) \quad (7)$$

Where S_0 represents the baseline survival for an individual with an explanatory variable value of zero. The factor e^{β} is referred to as the acceleration factor, indicating the speed at which an individual would move along the survival curve for a one-unit increase in the explanatory variable X[4].

In order to find the best functional form of the survival time, we constructed five types of parametric survival models including the Exponential, Weibull, Gompertz, Log Normal and Log logistic distributions. We used the *flexsurvreg* function from ***flexsurv*** package[4] in R to perform parametric regression model. The parameter of primary interest (in ***flexsurv***) is known as the location parameter and typically governs the mean or location for each distribution[4]. These

parameters impact the hazard function, which can take a variety of shapes depending on the distribution, where; 1) the Exponential distribution only supports a constant hazard; 2) the Weibull and Gompertz distributions support monotonically increasing and decreasing hazards; 3) the lognormal distribution supports arc-shaped and monotonically decreasing hazards[4].

To illustrate the distinct shapes of parametric distributions, we initially apply the parametric model using a baseline estimate without covariate inclusion. We then fit the parametric regression model with covariates. The exponential and Gompertz distributions can be parametrized as a PH model. The Log Normal and Log Logistic distribution can be parametrized as an AFT model. While the Weibull distribution can be parameterized as both an AFT model and as a PH model[4]. In the case of each model, we derived the relevant values by extracting from the exponential term (exp(β coefficient)). Consequently, the outcome from the PH model can be interpreted as HRs, while the interpretation for the AFT model is the Event Time Ratio (ETR). This approach providing a quantitative measure of the impact of covariates on the risk of premature CVD mortality. Within the PH model, the estimation of HRs for each covariate reveals that an HR greater than 1 signifies an elevated risk of premature CVD mortality, whereas an HR less than 1 indicates a reduced risk. In contrast, the AFT model estimates with an ETR value below 1 suggesting a shorter survival time (increased risk), and an ETR value exceeding 1 suggesting a longer survival time (decreased risk). We used AIC value to assess the best fit model of survival estimates against time together. The model exhibiting the lowest AIC score was chosen as the best.

Calculation of age standardized incidence rate of premature CVD mortality

We calculated the incidence of premature mortality from CVD in our study population, expressed as rates per 1000 person-years. Person-years were calculated based on the total sum of the number of years that each sample of the study population has been under observation. The age-specific incidence rates were obtained based on observed deaths and person-years for the overall population and sex strata. Subsequently, we used the direct standardization method by multiplying the age-specific death rate with the appropriate weight in the WHO standard population[5] to estimate the age-standardized rate for each age group/sex strata. Then, the overall ASIR was obtained from the summation of these age groups of age-standardized rate. This step of calculation follows the WHO Pan American Health Organization (PAHO) approach[6].

Table S1: Calculation of age standardized incidence rate: Premature CVD mortality (ICD10: I00-I99) aged 30-70 years

Age group (years) $_i$	No. of cases (pemature CVD mortality aged 30-69 years) d_i	Person-years at risk y_i	Age-specific incidence (per 10^5 years) $10^3(d_i/y_i)$	Standard world population (WHO)	Weight aged 30-69 w_i	Age-standardized rate (per age group) $10^3(di/yi)*wi/100$
30–34	12	79226.58	0.151	76.07	17.30	0.026
35–39	34	78771.29	0.432	71.47	16.25	0.070
40–44	53	84394.51	0.628	65.88	14.98	0.094
45–49	60	79365.22	0.756	60.38	13.73	0.104
50–54	107	70830.74	1.511	53.68	12.21	0.184
55–59	180	58311.78	3.087	45.48	10.34	0.319
60–64	207	39034.21	5.303	37.19	8.46	0.448
65–69	233	28374.61	8.212	29.59	6.73	0.553
Total	886	518308.93	20.0792	439.75	100.00	1.799

*Crude incidence rate of premature CVD mortality per 100,000-person year
$=\sum (d_i/y_1) = 886/518{,}308.93 *1000 = 1.71$
*Age adjusted incidence rate of premature CVD mortality per 1000-person year
$=\sum(10^3(d_i/y_1)* w_i/100) = 1.80$

Table S2: Calculation of age standardized incidence rate: Premature CVD mortality (ICD10: I00-I99) aged 30-70 years by sex

Age group (years) $_i$	No. of cases (pemature CVD mortality aged 30-69 years) d_i	Person-years at risk y_i	Age-specific incidence (per 10^5 years) $10^3(d_i/y_i)$	Standard world population (WHO)	Weight aged 30-69 w_i	Age-standardized rate (per age group) 10^3(di/yi)*wi/100
30–34	11	35363.21	0.311	76.07	17.30	0.054
35–39	22	33524.54	0.656	71.47	16.25	0.107
40–44	37	36789.16	1.006	65.88	14.98	0.151
45–49	45	34567.76	1.302	60.38	13.73	0.179
50–54	71	31378.29	2.263	53.68	12.21	0.276
55–59	120	27026.61	4.440	45.48	10.34	0.459
60–64	127	18315.74	6.934	37.19	8.46	0.586
65–69	145	12572.37	11.533	29.59	6.73	0.776
Total	578	229537.68	28.4448	439.75	100.00	2.588

Age group (years) $_i$	No. of cases (pemature CVD mortality aged 30-69 years) d_i	Person-years at risk y_i	Age-specific incidence (per 10^5 years) $10^3(d_i/y_i)$	Standard world population (WHO)	Weight aged 30-69 w_i	Age-standardized rate (per age group) 10^3(di/yi)*wi/100
30–34	1	43863.37	0.023	76.07	17.30	0.004
35–39	12	45246.75	0.265	71.47	16.25	0.043
40–44	16	47605.35	0.336	65.88	14.98	0.050
45–49	15	44797.46	0.335	60.38	13.73	0.046
50–54	36	39452.45	0.912	53.68	12.21	0.111
55–59	60	31285.17	1.918	45.48	10.34	0.198
60–64	80	20718.46	3.861	37.19	8.46	0.327
65–69	88	15802.25	5.569	29.59	6.73	0.375
Total	308	288771.25	13.2194	439.75	100.00	1.154

*Male

- Crude incidence rate of premature CVD mortality per 100,000-person year =$\sum (d_i/y_1)$ = 578/229,537.68 *1000 = 2.52
- Age adjusted incidence rate of premature CVD mortality per 1000-person year =$\sum(10^3(d_i/y_1)* w_i/100)$ = 2.59

*Female

- Crude incidence rate of premature CVD mortality per 100,000-person year =$\sum (d_i/y_1)$ = 308/288,771.25 *1000 = 1.07
- Age adjusted incidence rate of premature CVD mortality per 1000-person year =$\sum(10^3(d_i/y_1)* w_i/100)$ = 1.15

Distribution of sample characteristics

Table S3: Baseline characteristics and premature cardiovascular mortality in the study population for each National Health and Morbidity Survey

Characteristics	NHMS 2006		NHMS 2011		NHMS 2015	
	Total	Premature CVD mortality	Total	Premature CVD mortality	Total	Premature CVD mortality
	n = 29,853[a]	n = 504[b]	n = 15,620[a]	n = 211[b]	n = 18,249[a]	n = 171[b]
Follow up time[c]	15.75 (15.75, 15.75)	8.87 (4.87, 12.92)	10.71 (10.71, 10.71)	6.40 (3.79, 8.13)	6.84 (6.84, 6.84)	3.90 (2.2772, 5.4531)
Age at baseline[c]	41 (29, 52)	51 (44, 57)	40 (28, 52)	54 (46, 58)	41 (29, 53)	56 (50, 61)
Sex						
Female	16,646 (55.8%)	188 (1.1%)	8,339 (53.4%)	66 (0.8%)	9,563 (52.4%)	54 (0.6%)
Male	13,207 (44.2%)	316 (2.4%)	7,281 (46.6%)	145 (2.0%)	8,686 (47.6%)	117 (1.3%)
Location						
Urban	17,760 (59.5%)	266 (1.5%)	9,209 (59.0%)	103 (1.1%)	10,650 (58.4%)	86 (0.8%)
Rural	12,093 (40.5%)	238 (2.0%)	6,411 (41.0%)	108 (1.7%)	7,599 (41.6%)	85 (1.1%)
Ethnicity						
Chinese	6,365 (21.3%)	69 (1.1%)	3,079 (19.7%)	30 (1.0%)	2,815 (15.4%)	20 (0.7%)
Indian	2,584 (8.7%)	61 (2.4%)	1,301 (8.3%)	24 (1.8%)	1,308 (7.2%)	14 (1.1%)
Malay	16,911 (56.6%)	316 (1.9%)	9,414 (60.3%)	140 (1.5%)	11,390 (62.4%)	125 (1.1%)
Other Bumiputeras	3,553 (11.9%)	50 (1.4%)	1,529 (9.8%)	14 (0.9%)	1,619 (8.9%)	11 (0.7%)
Others	440 (1.5%)	8 (1.8%)	297 (1.9%)	3 (1.0%)	1,117 (6.1%)	1 (0.1%)
Education level						
Tertiary	3,225 (10.9%)	20 (0.6%)	3,641 (23.6%)	13 (0.4%)	4,307 (23.9%)	18 (0.4%)
Secondary	15,941 (53.7%)	205 (1.3%)	7,631 (49.4%)	97 (1.3%)	8,884 (49.3%)	80 (0.9%)
Primary	8,196 (27.6%)	231 (2.8%)	3,393 (21.9%)	84 (2.5%)	3,982 (22.1%)	65 (1.6%)
No formal education	2,302 (7.8%)	45 (2.0%)	795 (5.1%)	14 (1.8%)	863 (4.8%)	8 (0.9%)
Occupation						
Working	17,953 (62.9%)	310 (1.7%)	10,196 (65.5%)	130 (1.3%)	11,774 (64.5%)	84 (0.7%)
Not Working	10,579 (37.1%)	154 (1.5%)	5,381 (34.5%)	80 (1.5%)	6,474 (35.5%)	87 (1.3%)
Income group						
top20%	2,892 (10.1%)	31 (1.1%)	4,106 (27.3%)	43 (1.0%)	5,854 (33.6%)	44 (0.8%)
middle40%	9,768 (34.1%)	147 (1.5%)	6,887 (45.7%)	100 (1.5%)	7,707 (44.2%)	69 (0.9%)
bottom40%	15,967 (55.8%)	315 (2.0%)	4,064 (27.0%)	63 (1.6%)	3,859 (22.2%)	45 (1.2%)

Table S3: Baseline characteristics and premature cardiovascular mortality in the study population for each National Health and Morbidity Survey, continue.

Characteristics	NHMS 2006		NHMS 2011		NHMS 2015	
	Total	Premature CVD mortality	Total	Premature CVD mortality	Total	Premature CVD mortality
	n = 29,853[a]	n = 504[b]	n = 15,620[a]	n = 211[b]	n = 18,249[a]	n = 171[b]
Marital status						
Never married	6,565 (22.1%)	38 (0.6%)	3,857 (24.7%)	18 (0.5%)	4,093 (22.4%)	11 (0.3%)
Married	21,394 (72.0%)	425 (2.0%)	10,880 (69.7%)	174 (1.6%)	12,848 (70.4%)	145 (1.1%)
Widow/Widower/Divor	1,769 (6.0%)	40 (2.3%)	878 (5.6%)	19 (2.2%)	1,306 (7.2%)	15 (1.1%)
Diabetes status						
Normal	27,108 (90.9%)	334 (1.2%)	13,424 (87.9%)	125 (0.9%)	15,587 (85.5%)	98 (0.6%)
Diagnosed	2,062 (6.9%)	145 (7.0%)	1,323 (8.7%)	69 (5.2%)	1,746 (9.6%)	65 (3.7%)
Undiagnosed	666 (2.2%)	25 (3.8%)	518 (3.4%)	14 (2.7%)	905 (5.0%)	8 (0.9%)
Hypertension status						
Normal	18,772 (63.0%)	166 (0.9%)	9,985 (64.3%)	58 (0.6%)	12,109 (66.4%)	62 (0.5%)
Diagnosed	3,954 (13.3%)	167 (4.2%)	2,243 (14.4%)	88 (3.9%)	2,589 (14.2%)	69 (2.7%)
Undiagnosed	7,061 (23.7%)	170 (2.4%)	3,296 (21.2%)	63 (1.9%)	3,541 (19.4%)	40 (1.1%)
Hypercholestrolemia status						
Normal	22,733 (76.4%)	306 (1.3%)	9,440 (61.2%)	83 (0.9%)	8,718 (47.8%)	67 (0.8%)
Diagnosed	1,394 (4.7%)	59 (4.2%)	1,437 (9.3%)	48 (3.3%)	1,882 (10.3%)	42 (2.2%)
Undiagnosed	5,641 (18.9%)	138 (2.4%)	4,559 (29.5%)	76 (1.7%)	7,638 (41.9%)	62 (0.8%)
General obesity	12,917 (45.0%)	294 (2.3%)	7,145 (48.9%)	122 (1.7%)	9,050 (53.0%)	88 (1.0%)
Abdominal obesity	12,137 (42.4%)	280 (2.3%)	6,975 (49.2%)	127 (1.8%)	9,159 (53.8%)	94 (1.0%)
Current smoker	6,781 (22.9%)	178 (2.6%)	3,566 (23.0%)	68 (1.9%)	4,125 (22.6%)	60 (1.5%)
Physical Inactive	12,446 (42.3%)	216 (1.7%)	5,417 (34.9%)	82 (1.5%)	5,459 (30.3%)	66 (1.2%)

*NHMS = National Health and Morbidity Survey; Premature CVD mortality defined as those who died due to cardiovascular disease (ICD-10: I00-I99) between the age of 30-70 years.
[a]n (%), percent by column; [b]n (%), percent by row; [c]Median (IQR).

Checking propotional hazards (PH) assumption of Cox model PH model

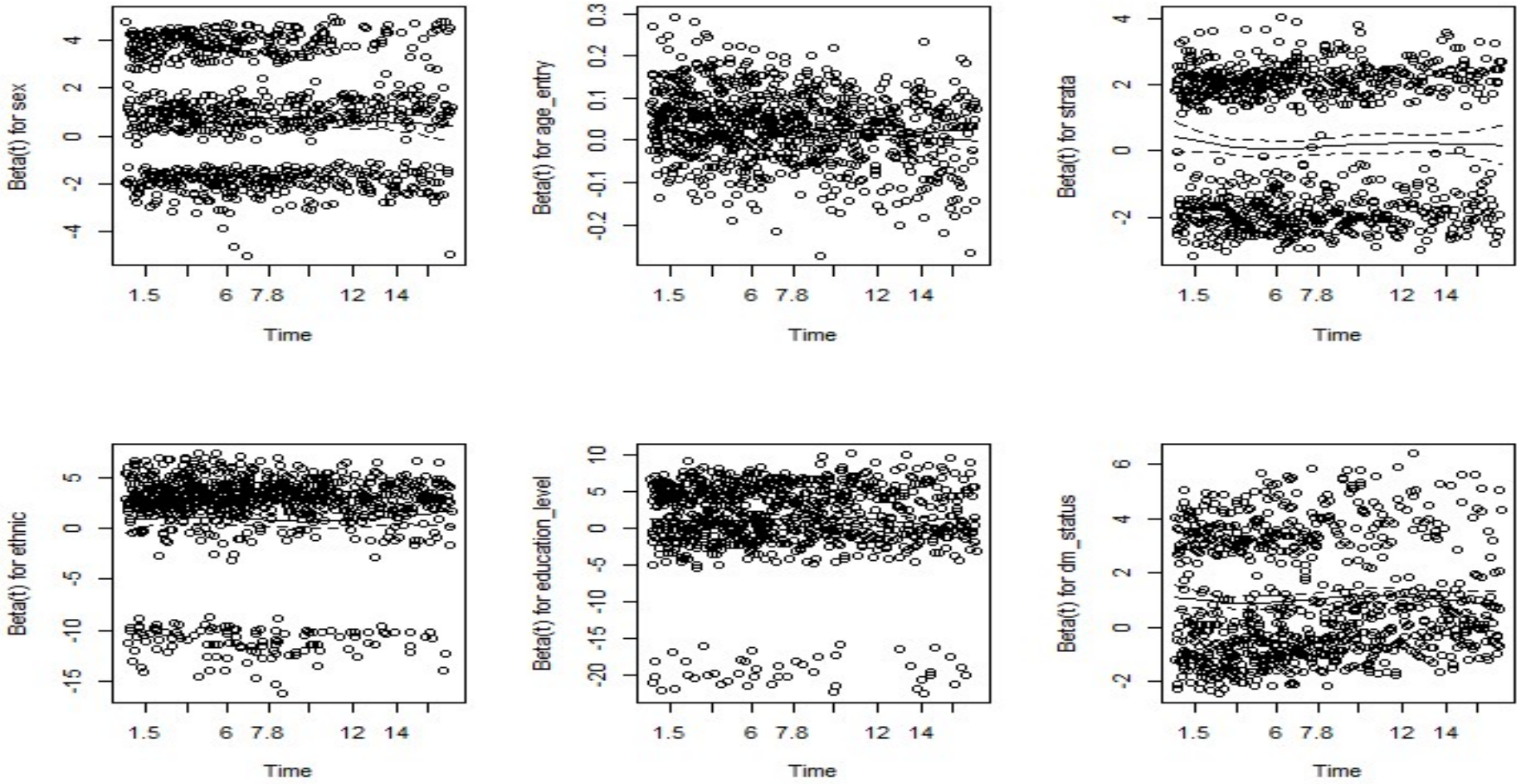
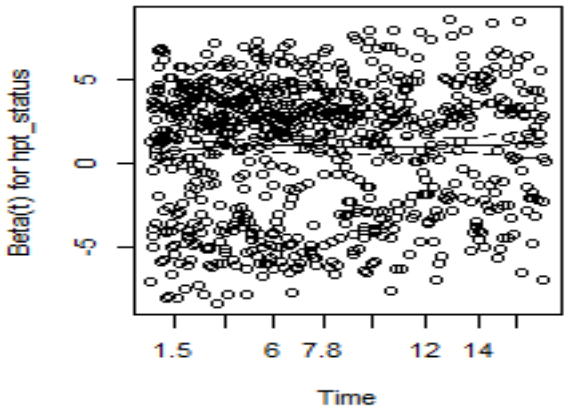
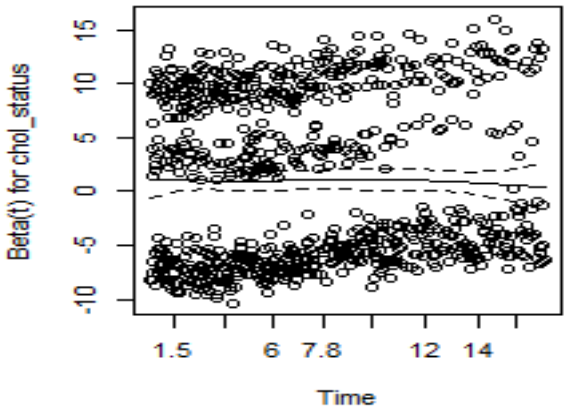
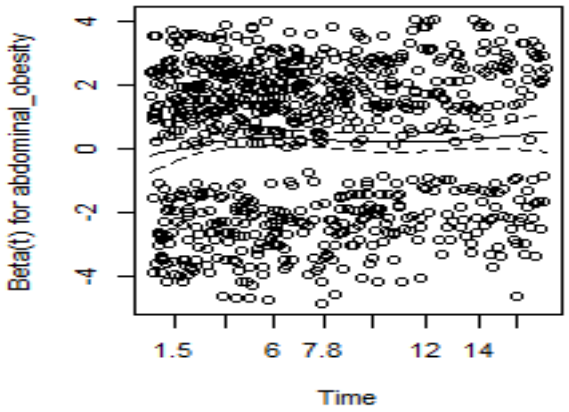
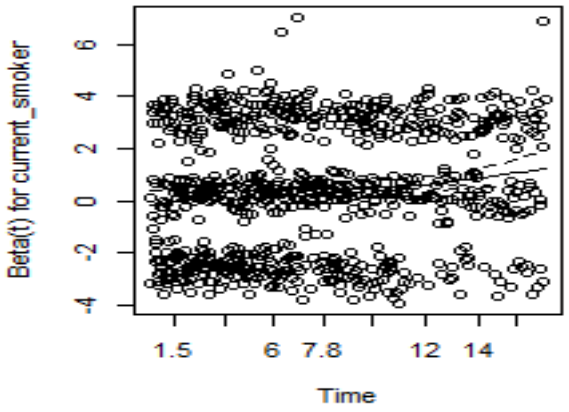


Figure S1: Schoenfeld residual test and plot for each covariate before stratified Cox model









	chisq	df	p
sex	1.277	1	0.2584
age_entry	41.744	1	1.0e-10
strata	0.116	1	0.7339
ethnic	9.317	4	0.0537
education_level	6.003	3	0.1115
dm_status	4.300	2	0.1165
hpt_status	5.841	2	0.0539
chol_status	1.528	2	0.4659
abdominal_obesity	0.127	1	0.7216
current_smoker	7.962	1	0.0048
GLOBAL	70.170	18	4.2e-08

Figure S1: Schoenfeld residual test and plot for each covariate before stratified Cox model (continue)

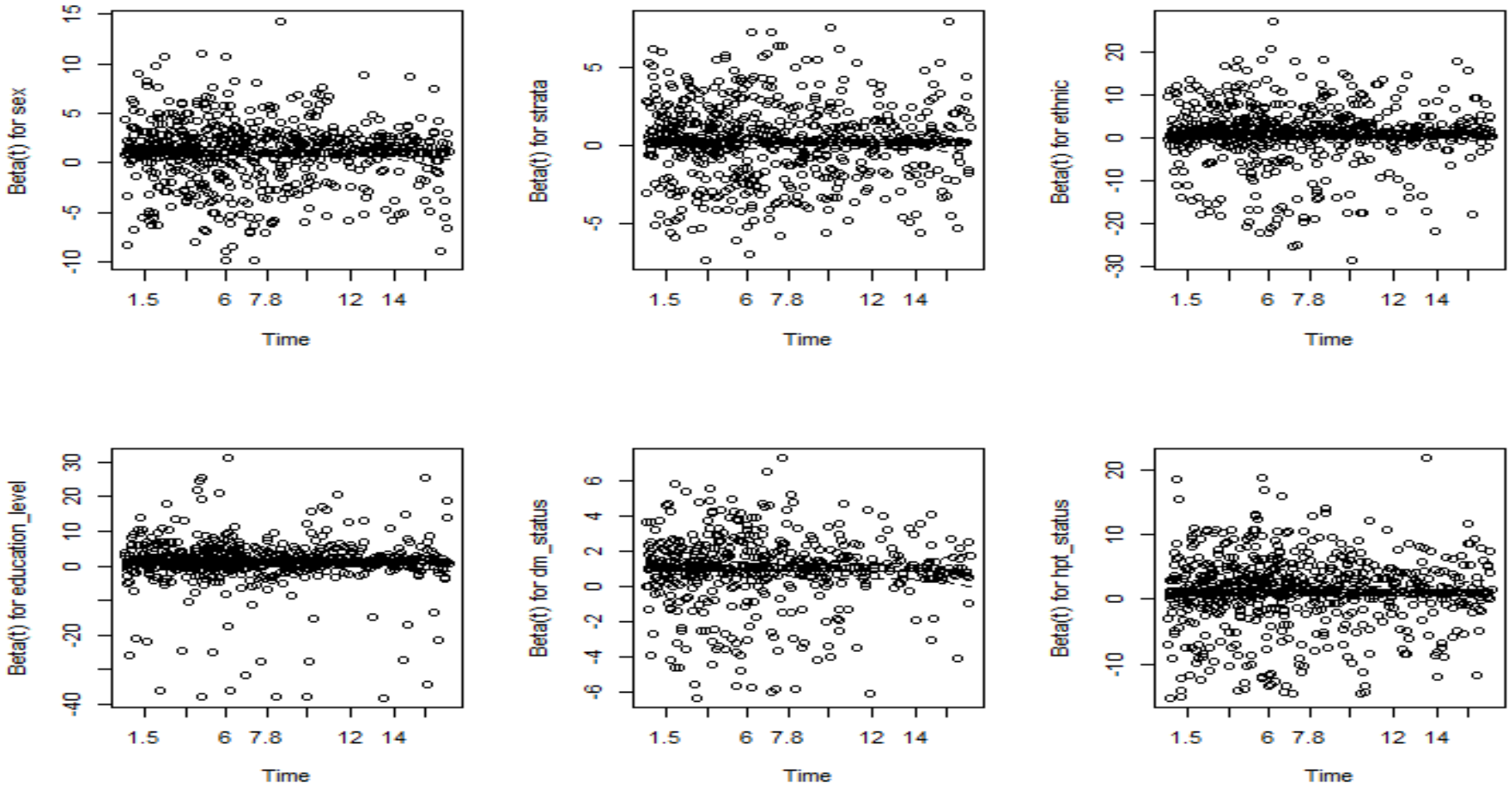
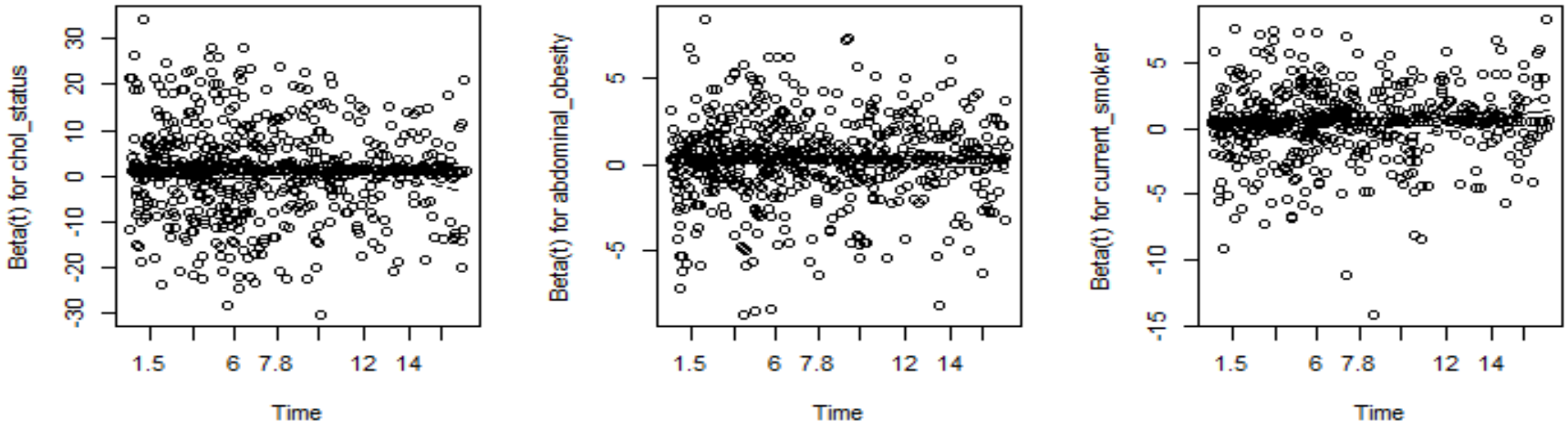


Figure S2: Schoenfeld residual test and plot for each covariate after stratified Cox model by age of entry



	chisq	df	p
sex	0.0815	1	0.775
strata	1.1745	1	0.278
ethnic	2.2257	4	0.694
education_level	2.8296	3	0.419
dm_status	6.3698	2	0.041
hpt_status	2.2656	2	0.322
chol_status	2.6738	2	0.263
abdominal_obesity	0.6192	1	0.431
current_smoker	4.7397	1	0.029
GLOBAL	26.2700	17	0.070

Figure S2: Schoenfeld residual test and plot for each covariate after stratified Cox model by age of entry (continue)

Sensitivity analysis

Since our study data had an imbalance between the outcome (event premature CVD mortality) and censored data, we conducted sensitivity analyses to address potential issues. This disparity was characterized by small event cases (886 of premature CVD mortality, 1.4%) compared to large number of 62,836 censored cases (98.6%). To address this imbalance, an under-sampling method was employed. Specifically, censored data were randomly deleted at a rate of 90%, effectively attaining a balanced events-to-censored ratio of 1:7. Consequently, the final sample size encompassed 886 events and 6,284 censored cases, yielding a total sample of 7,170. The subsequent analysis involved a semi-parametric Cox model and selected the best fit model from parametric regression model. We presented the result in comparison of the original imbalanced data with the treated balanced data.

Remarkably, the pattern observed for each distribution remained almost identical for both balanced (**Figure S3**) and imbalanced data (**Figure 2** – from main manuscript), suggesting that the data treatment did not significantly alter the distribution characteristics. This consistency was further affirmed by employing semi-parametric Cox PH models and the parametric Log Normal distribution (AFT) model to compare the outcomes of the analysis on the balanced and imbalanced datasets. **Table S4** presented comparison of unadjusted models for both semi-parametric and parametric Log-normal models. The comparison of adjusted model is presented in **Table S5**. Across a range of covariates, which included diagnosed and undiagnosed DM, HPT status, HPL status, obesity, AO, current smoking, and physical inactivity, we consistently observed effect sizes represented by HRs or ETRs, all demonstrating consistent statistical significance. These results underscore that the initial data imbalance does not significantly impact the conclusions drawn from the analysis across various covariates and models.

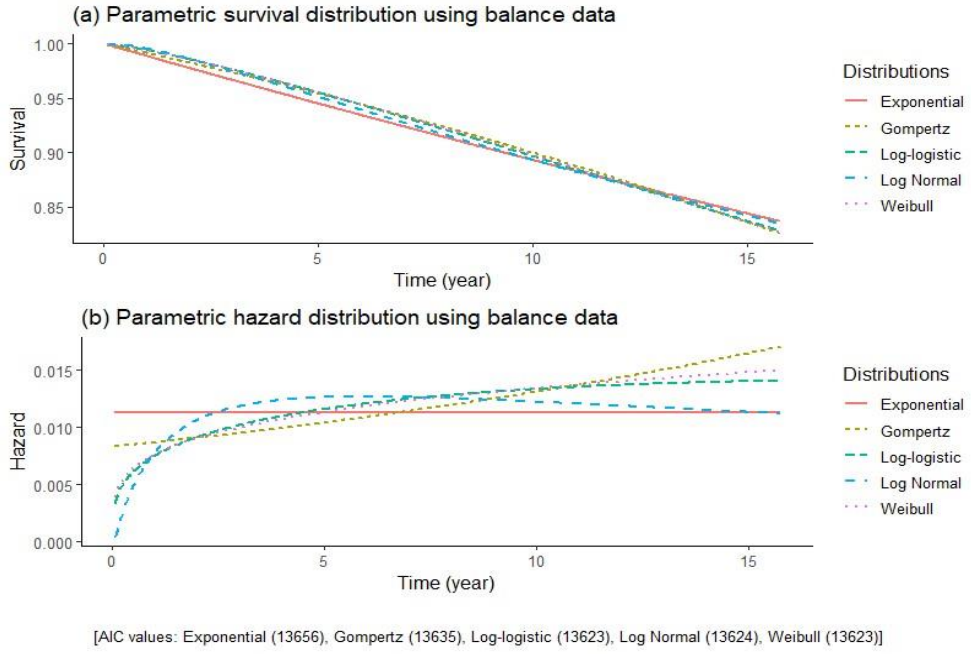


Figure S3: Parametric distribution fittings for premature CVD mortality data after treatment using under-sampling method on imbalanced data

Table S4: Sensitivity analysis treated of imbalance data for unadjusted survival analysis model of premature cardiovascular disease mortality in Malaysia

Covariates	Semi-parametric model						Parametric model			
	Unadjusted Cox PH model						Unadjusted Log Normal (AFT model)			
	Imbalance data			Balance data			Imbalance data		Balance data	
	HR	95% CI	p-value	HR	95% CI	p-value	ETR	95% CI	ETR	95% CI
Sex										
Female	—	—		—	—		—	—	—	—
Male	2.33	2.03, 2.67	<0.001	2.19	1.90, 2.51	<0.001	0.47	0.41, 0.53	0.49	0.43, 0.56
Age at baseline	1.06	1.06, 1.07	<0.001	1.06	1.05, 1.07	<0.001	0.95	0.94, 0.95	0.95	0.94, 0.95
Location										
Urban	—	—		—	—		—	—	—	—
Rural	1.40	1.23, 1.60	<0.001	1.40	1.23, 1.60	<0.001	0.73	0.65, 0.83	0.73	0.64, 0.82
Ethnicity										
Chinese	—	—		—	—		—	—	—	—
Indian	2.00	1.53, 2.62	<0.001	1.93	1.48, 2.53	<0.001	0.54	0.42, 0.68	0.5	0.39, 0.64
Malay	1.70	1.40, 2.07	<0.001	1.71	1.41, 2.09	<0.001	0.63	0.53, 0.75	0.62	0.52, 0.74
Other Bumiputeras	1.15	0.86, 1.54	0.300	1.18	0.89, 1.58	0.300	0.91	0.71, 1.17	0.85	0.66, 1.10
Others	0.89	0.49, 1.61	0.700	0.89	0.49, 1.61	0.700	1.13	0.69, 1.86	1.20	0.74, 1.94
Education level										
Tertiary	—	—		—	—		—	—	—	—
Secondary	2.31	1.72, 3.11	<0.001	2.33	1.74, 3.14	<0.001	0.50	0.39, 0.64	0.51	0.40, 0.64
Primary	4.98	3.71, 6.69	<0.001	4.71	3.51, 6.32	<0.001	0.25	0.19, 0.32	0.27	0.21, 0.34
No formal education	3.48	2.41, 5.02	<0.001	3.17	2.20, 4.57	<0.001	0.35	0.26, 0.48	0.38	0.28, 0.52
Occupation										
Working	—	—		—	—		—	—	—	—
Not Working	1.11	0.97, 1.27	0.140	1.17	1.02, 1.35	0.026	0.88	0.78, 1	0.83	0.73, 0.94
Income group										
top20%	—	—		—	—		—	—	—	—
middle40%	1.23	1.00, 1.52	0.053	1.23	0.99, 1.52	0.057	0.82	0.68, 0.98	0.79	0.66, 0.96
bottom40%	1.46	1.19, 1.79	<0.001	1.43	1.16, 1.76	<0.001	0.71	0.59, 0.85	0.7	0.58, 0.84
Marital status										
Never married	—	—		—	—		—	—	—	—
Married	3.67	2.86, 4.72	<0.001	3.56	2.77, 4.57	<0.001	0.33	0.27, 0.41	0.34	0.28, 0.42
Widow/Widower/Divorcee	4.57	3.29, 6.36	<0.001	4.31	3.10, 6.00	<0.001	0.26	0.2, 0.35	0.28	0.21, 0.38

Table S4: Sensitivity analysis treated of imbalance data for unadjusted survival analysis model of premature cardiovascular disease mortality in Malaysia (continue)

Covariates	Semi-parametric model						Parametric model			
	Unadjusted Cox PH model						Unadjusted Log Normal (AFT model)			
	Imbalance data			Balance data			Imbalance data		Balance data	
	HR	95% CI	p-value	HR	95% CI	p-value	ETR	95% CI	ETR	95% CI
Diabetes status										
Normal	—	—		—	—		—	—	—	—
Diagnosed	6.95	6.02, 8.03	<0.001	5.97	5.17, 6.90	<0.001	0.18	0.15, 0.21	0.21	0.18, 0.25
Undiagnosed	2.81	2.08, 3.78	<0.001	2.55	1.89, 3.44	<0.001	0.43	0.33, 0.56	0.49	0.37, 0.65
Hypertension status										
Normal	—	—		—	—		—	—	—	—
Diagnosed	5.93	5.05, 6.95	<0.001	5.2	4.44, 6.10	<0.001	0.21	0.18, 0.24	0.22	0.19, 0.26
Undiagnosed	2.83	2.40, 3.34	<0.001	2.52	2.14, 2.98	<0.001	0.42	0.36, 0.49	0.45	0.39, 0.52
Hypercholestrolemia status										
Normal	—	—		—	—		—	—	—	—
Diagnosed	3.72	3.08, 4.48	<0.001	3.64	3.02, 4.38	<0.001	0.31	0.26, 0.37	0.34	0.28, 0.41
Undiagnosed	1.74	1.49, 2.02	<0.001	1.65	1.42, 1.92	<0.001	0.63	0.55, 0.72	0.63	0.55, 0.72
Obesity status										
Non-Obese	—	—		—	—		—	—	—	—
Obese	1.78	1.55, 2.05	<0.001	1.81	1.57, 2.08	<0.001	0.62	0.55, 0.7	0.65	0.57, 0.74
Abdominal obesity										
No	—	—		—	—		—	—	—	—
Yes	1.91	1.66, 2.19	<0.001	1.92	1.67, 2.21	<0.001	0.58	0.51, 0.66	0.62	0.54, 0.70
Current smoker										
No	—	—		—	—		—	—	—	—
Yes	1.83	1.59, 2.10	<0.001	1.72	1.50, 1.98	<0.001	0.59	0.52, 0.68	0.65	0.56, 0.74
Physical activity										
Active	—	—		—	—		—	—	—	—
Inactive	1.14	1.0, 1.30	0.059	1.15	1.00, 1.31	0.043	0.87	0.77, 0.98	0.84	0.74, 0.95

Abbreviations: PH = Proportional Hazard, AFT = Accelerated Failure Time, HR = Hazard Ratio, ETR = Event Time Ratio, CI = Confidence Interval.

Imbalanced data were obtained from the original premature CVD mortality dataset (1.4% event cases out of 63,722 total data). Balanced data (12.4% event cases out of 7,107 total data) were obtained using an under-sampling method for censored data. All parametric and semi-parametric models were adjusted for the following covariates: sex, age at entry, ethnicity, education level, diabetes status, hypertension status, hypercholesterolemia status, abdominal obesity, and current smoking status.

* Indicates a statistically significant p-value of <0.05

Table S5: Sensitivity analysis comparing treated imbalance data to original data for the premature cardiovascular disease mortality survival analysis model

Covariates	Semi-parametric Cox PH model								Parametric model			
	Adjusted model without stratification				Adjusted model stratified by age				Log Normal (AFT model)			
	Imbalance data		Balance data		Imbalance data		Balance data		Imbalance data		Balance data	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI	ETR	95% CI	ETR	95% CI
Sex (ref: Male)	2.34	1.96, 2.79	2.32	1.94, 2.77	2.68	2.08, 3.44	3.05	1.66, 5.58	0.50	0.43, 0.58	0.51	0.44, 0.59
Age at baseline	1.04	1.03, 1.04	1.04	1.03, 1.05	—	—	—	—	0.97	0.96, 0.98	0.97	0.96, 0.97
Rural (ref: Urban)	1.19	1.03, 1.38	1.19	1.03, 1.39	1.29	1.03, 1.60	1.31	0.77, 2.23	0.87	0.77, 0.98	0.86	0.77, 0.97
Ethnicity (ref: Chinese)												
Indian	1.80	1.35, 2.39	1.65	1.24, 2.19	1.93	1.27, 2.94	1.31	0.48, 3.60	0.62	0.49, 0.78	0.70	0.56, 0.87
Malay	1.69	1.36, 2.10	1.61	1.30, 2.00	1.60	1.19, 2.15	1.2	0.59, 2.44	0.69	0.58, 0.81	0.73	0.62, 0.86
Other Bumiputera	1.48	1.08, 2.02	1.55	1.13, 2.12	1.57	1.03, 2.40	1.13	0.41, 3.12	0.75	0.59, 0.95	0.75	0.60, 0.96
Others	0.87	0.44, 1.73	0.92	0.46, 1.82	1.17	0.51, 2.69	1.51	0.24, 9.69	1.11	0.68, 1.82	1.11	0.69, 1.79
Education level (ref: Tertiary)												
Secondary	1.78	1.30, 2.43	1.90	1.39, 2.60	1.89	1.23, 2.91	2.85	0.88, 9.27	0.65	0.51, 0.81	0.63	0.51, 0.79
Primary	2.16	1.56, 3.00	2.22	1.60, 3.08	1.95	1.24, 3.06	3.1	0.91, 10.5	0.55	0.43, 0.70	0.54	0.43, 0.69
No formal education	1.72	1.14, 2.60	1.77	1.18, 2.67	2.21	1.24, 3.93	3.1	0.64, 15.0	0.67	0.49, 0.91	0.68	0.50, 0.92
Diagnosed DM	3.28	2.75, 3.91	3.10	2.61, 3.69	3.26	2.48, 4.30	2.37	1.28, 4.36	0.37	0.32, 0.44	0.41	0.35, 0.48
Undiagnosed DM	1.86	1.37, 2.52	1.70	1.25, 2.31	1.63	1.04, 2.58	2.98	0.87, 10.2	0.63	0.49, 0.81	0.67	0.52, 0.87
Diagnosed HPT	2.23	1.82, 2.73	2.05	1.68, 2.50	1.84	1.39, 2.44	1.32	0.69, 2.53	0.53	0.45, 0.62	0.57	0.49, 0.67
Undiagnosed HPT	1.69	1.41, 2.03	1.52	1.27, 1.82	1.46	1.13, 1.88	1.06	0.58, 1.95	0.68	0.59, 0.78	0.75	0.65, 0.86
Diagnosed Hyperchol	1.21	0.97, 1.50	1.32	1.06, 1.65	1.08	0.77, 1.50	1.92	0.86, 4.24	0.88	0.74, 1.06	0.86	0.71, 1.04
Undiagnosed Hyperchol	1.35	1.15, 1.59	1.42	1.21, 1.66	1.31	1.04, 1.65	1.21	0.75, 1.97	0.80	0.70, 0.91	0.80	0.70, 0.91
Abdominal obesity	1.29	1.10, 1.50	1.38	1.18, 1.61	1.38	1.11, 1.72	2.00	1.18, 3.41	0.81	0.72, 0.92	0.78	0.69, 0.88
Current smoker	1.56	1.31, 1.86	1.46	1.22, 1.73	1.60	1.23, 2.08	1.52	0.81, 2.87	0.71	0.61, 0.82	0.76	0.66, 0.88

Abbreviations: PH = Proportional Hazard, AFT = Accelerated Failure Time, HR = Hazard Ratio, ETR = Event Time Ratio, CI = Confidence Interval. DM = diabetes mellitus, HPT = hypertension and HPL = hypercholesterolemia. Imbalanced data were obtained from the original premature CVD mortality dataset (1.4% event cases out of 63,722 total data). Balanced data (12.4% event cases out of 7,107 total data) were obtained using an under-sampling method for censored data. All parametric and semi-parametric models were adjusted for the following covariates: sex, age at entry, ethnicity, education level, diabetes status, hypertension status, hypercholesterolemia status, abdominal obesity, and current smoking status.

References

1. Therneau TM, Lumley T. Package ‘survival.’ R Top Doc. 2015;128(10):28–33.

2. Kleinbaum DG, Klein M. Survival analysis a self-learning text. Springer; 1996.

3. Kartsonaki C. Survival analysis. Diagn Histopathol. 2016;22(7):263–70.

4. Jackson C, Metcalfe P, Amdahl J, Jackson MC. Package ‘flexsurv.’ Flexible parametric survival and multi-state models, https://cran r-project org/web/packages/flexsurv/flexsurv pdf (2018, accessed 26 September 2019). 2023;

5. Ahmad OB, Boschi-Pinto C, Lopez AD, Murray CJL, Lozano R, Inoue M, et al. Age standardization of rates: a new WHO standard. Geneva: World Health Organization. 2001;9(10):1–14.

6. Pan American Health Organization (PAHO). PAHO. Age standardized mortality rates per 100,000 population deaths less 70 years, 2009 [Internet]. 2009 [cited 2023 Sep 7]. Available from: https://www.paho.org/en/documents/paho-age-standardized-mortality-rates-100000-population-deaths-less-70-years-2009