Project 3 Report

Analyst: Max McGrath

Investigator: Nichole Carlson

Report generated: 12/06/2021

Introduction

We received data of 11,627 observations from 4,434 individuals with up to three observations per individual tracking subjects' cardiovascular outcomes for up to 24 years. Most relevant to this study, each observation records whether the subject has had a stroke since the last observation and, if so, the timing of that stroke. Each observation also measures four continuous cardiovascular risk factors: (1) age, (2) systolic blood pressure (SBP), (3) cholesterol levels, and (4) BMI, and four dichotomous variables: (5) smoking status, (6) diabetic status, (7) history of cardiovascular disease (CVD), and (8) use of anti-hypertension medication.

The aim of this study will be to use this data to identify statistically significant risk factors that are associated with greater risk of stroke 10 years after the start of observation by testing the hypotheses that each of the eight previously listed risk factors are associated with 10-year risk of stroke. The second aim is to determine 10-year probabilities of strokes for different risk profiles using the significant risk factors. The last aim of this analysis is to identify whether their is meaningful change in risk factors over the first 10 years of the study to determine whether a longitudinal analysis approach is warranted. This study is stratified by sex, meaning males and females are treated as two separate populations with potentially differing risk factors.

Methods

First, subjects who have had a stroke prior to the start of observation were removed from the data set. Next, for each subject, a record of whether they had a stroke in the first 10 years of observation was created along with the timing of that stroke (which we will refer to as events and event times, respectively). If they did not have an event in those years, they were classified as right censored. Based on the findings of Wolf et al. (1990) as well as data availability, age, systolic blood pressure (mmHg) (SBP), use of anti-hypertensive medication, diabetes mellitus, cigarette smoking, past cardiovascular disease (CVD), BMI, and serum

total cholesterol (mg/dL) were included in the analysis as potential risk factors (all measured at baseline). Age, SBP, cholesterol, and BMI will be treated as continuous variables with the rest treated as dichotomous. Only observations with no missing values for any of the potential risk factors were used.

The first part of the analysis provides descriptive statistics for potential risk factors, stratified by sex and whether the subject had an event or was censored. Next, estimated Kaplan-Meier survival curves were fit to crudely estimate event probability over time. For continuous risk factors, the K-M curves were fit according to the risk factor quantiles. Then, to fully quantify each risk factor's relationship with stroke occurrence, a Cox proportional hazards (CPH) regression was fit for each sex with the number of days after beginning observation that a stroke occurred or the last observation time in days as the outcome and the potential risk factors described above as variables of interest. The risk factors were then narrowed using backwards selection with statistical significance (alpha < 0.05) as the criteria for whether each factor was kept in the model. After predictor selection, a final CPH model was fit for each sex. Coefficient estimates from these models were then used to quantify the relationship between risk factors and stroke occurrence, with Wald tests quantifying these relationships' statistical significance. 10-year stroke probabilities were then calculated for a predefined set of risk profiles.

Lastly, we provide summary statistics of the changes in potential risk factors from the first observation to the second and discuss the potential impact of using a more complex, time-varying covariate analysis rather than simply evaluating relationships between baseline values and stroke occurrence. As a part of this discussion, the proportional hazards assumption for Cox proportional hazards models was evaluated in the context of this study to identify whether the relationship between covariates and stroke risk varies with time thereby altering the interpretation of this study's results.

Results

After preparing the data, there were 2378 subjects with sex listed as female including 57 who had an event and 2321 who are censored, and there were 1897 subjects with sex listed as male with 48 who had an event and 1849 who are censored.

From Table 1, the majority of individuals observed did not have strokes within 10 years of observation (2.5% of males and 2.4% of females had strokes). Among those who did have strokes, generally they are older, have higher SBP, are more likely to use anti-hypertensive medication, are more likely to have diabetes, and are more likely to have a history of CVD. From the Kaplan-Meier survival curves shown in Figures 1 and 2, age, SBP, and all dichotomous measures appear to be associated with greater incidence of strokes in males, and all continuous measures and all dichotomous measures except current smoking status appear to be associated with greater incidence of strokes in females. Both the descriptive statistics and Kaplan-Meier curves lend credence to the selection of the initial set of potential risk factors to consider.

Following backwards model selection, age, SBP, diabetes, and smoking status remained as significant predictors of stroke incidence in males, while only age and SBP remained as significant predictors of stroke incidence in females. For males, having diabetes at baseline was associated with an estimated hazard of stroke within 10 years of observation of 4.79 (95% CI: (2.1, 10.99)) times that of those who did not have diabetes at baseline holding age, BPD, and smoking status constant (p-value < 0.001), and being a smoker at baseline was associated with an estimated hazard of having a stroke of 1.92 (95% CI: (1.03, 3.56)) times that of those who were not smokers at baseline, holding age, SBP, and diabetic status constant (p-value $= 0.039$). For both males and females, the estimated hazard of stroke was 1.07 times greater for each additional year of age (95% CI's: (1.03, 1.11) and (1.03, 1.12), respectively), and the estimated hazard of stroke was 1.03 times greater for each additional mmHg of blood pressure (95% CI's: (1.02, 1.05) and (1.02, 1.03), respectively) with p-values $\leq .001$ for both age and SBP in each sex's respective models.

Table 3 provides the probabilities of stroke within 10 years for a variety of risk profiles as estimated by the Cox proportional models. The average probability of stroke within ten years is 2% at 55 and 15% at 85 for males and 2% at 55 and 16% at 85 for females. For females, the probability of stroke at each age remains constant for each risk profile, as none of the reported risk profiles are significantly associated with stroke probability. For males, being a smoker and having diabetes both increase the probability of having a stroke within 10 years for all ages, with the conjunction of the two carrying the highest probability of stroke within 10 years of any risk profile at all ages (13% at 55, 61% at 85).

For the longitudinal comparison between the first and second period, Table 4 provides descriptive statistics of the differences between the two periods. Only 1594 males and 2001 subjects that were present for period one (after removing those who had previously had a stroke) were also present for period two. For the continuous risk factors, BMI slightly increased for those who were censored and decreased for those who had strokes, while cholesterol and SBP increased for both censored and those who had a stroke. For dichotomous factors, more individuals stopped smoking than started smoking, and similarly more individuals started taking anti-hypertension medication than stopped taking it. There was also a small percentage of individuals who were recognized as diabetic between the first and second period or who had a history of CVD in the second period but not in the first. The proportional hazards assumption of the CPH models were also evaluated using Schoenfeld tests, and for both the male and female models there was no evidence of violation of the proportional hazards assumption (p-values of 0.35 and 0.95, respectively).

Conclusions

From the CPH models, only age and systolic blood pressure were significantly associated with increased hazard of stroke for females, and age, systolic blood pressure, smoking status, and diabetic status were significantly associated with increased hazard of stroke in males. From the 10-year stroke probabilities calculated using these models, the probability of stroke

increases with age for all risk profiles. For females, the probability of stroke is constant across all risk profiles for a fixed age, and for males those who are current smokers, those who are diabetic, or those who are both are associated with increased 10-year probabilities of stroke, the latter the most of any risk profile.

From the comparison of the first and second period, the across-the-board increase in SBP coupled with SBP's significant association with stroke hazard is indicative that a longitudinal approach may be preferable for fully assessing stroke hazard, although the proportional hazards assumption for this analysis was met indicating a non-longitudinal approach may have been sufficient. A longitudinal analysis would also account for observed changes in smoking habits, anti-hypertension medication use, and diabetic status.

Limitations

The range of ages for which risk probabilities (55 - 85 years) were calculated was based upon the methods of Wolf (1990), but it may be problematic to extend the inferences of this study to the higher ages in that range. Specifically, at baseline the largest ages present in the study were 69 for male and 70 for female subjects, with averages of about 50 years for both sexes. Coupled with the low incidence of stroke among the studied populations, the studied data may not provide helpful insight into higher age ranges.

Further, several of the risk factors that were identified as significant by Wolf et al. were not identified as significant in this study, and no factors which Wolf et al. did not consider were identified as significant. As such, this study is likely under powered and would be improved by increasing the sample size, particularly to capture a greater number of individuals who have strokes in the period of observation.

Reproducibility

The code used to generate this analysis is available on GitHub at https://github.com/BIOS6624-UCD/bios6624-MaxMcGrath/tree/main/Project3 . The `Background` folder contains information pertinent to understanding the analysis but unnecessary for reproducing it. The `Code` folder contains six files: `1_ProcessData.R`, `2_EDA.R`, `3_KM.R`, `4_ModelSelection.R`, `5_ProbabilitiesM.R`, `6_ProbabilitiesF.R`, and `7_TimeAnalysis.R`. These R scripts are dependent upon a data file `Data/frmgham2.csv` which is not available on GitHub, but may be requested by emailing `max.mcgrath@ucdenver.edu`. To run the complete analysis, each script should be run in the order of the number prefixing its filename. The last directory, `Report`, contains the RMarkdown file `report.Rmd` which may be used to generate this report (note that it also depends on the aforementioned data and scripts).

The complete details of the `R` version, package versions, and machine details for the instance which generated this report are provided below.

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

```
## other attached packages:
## [1] survminer_0.4.9  ggpubr_0.4.0     ggplot2_3.3.5    survival_3.2-13
## [5] table1_1.4.2     dplyr_1.0.7      gridExtra_2.3    kableExtra_1.3.4
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.2        tidyr_1.1.4       viridisLite_0.4.0 splines_4.1.1
##  [5] carData_3.0-4     Formula_1.2-4     cellranger_1.1.0  yaml_2.2.1
##  [9] pillar_1.6.4      backports_1.2.1   lattice_0.20-45   glue_1.4.2
## [13] digest_0.6.28     ggsignif_0.6.3    snakecase_0.11.0  rvest_1.0.2
## [17] colorspace_2.0-2  htmltools_0.5.2   Matrix_1.3-4      pkgconfig_2.0.3
## [21] broom_0.7.9       haven_2.4.3       purrr_0.3.4       xtable_1.8-4
## [25] scales_1.1.1      webshot_0.5.2     svglite_2.0.0     km.ci_0.5-2
## [29] openxlsx_4.2.4    rio_0.5.27        KMsurv_0.1-5      tibble_3.1.5
## [33] generics_0.1.0    car_3.0-11        ellipsis_0.3.2    withr_2.4.2
## [37] janitor_2.1.0     magrittr_2.0.1    crayon_1.4.1      readxl_1.3.1
## [41] evaluate_0.14     fansi_0.5.0       rstatix_0.7.0     forcats_0.5.1
## [45] xml2_1.3.2        foreign_0.8-81    tools_4.1.1       data.table_1.14.2
## [49] hms_1.1.1         lifecycle_1.0.1   stringr_1.4.0     munsell_0.5.0
## [53] zip_2.2.0         compiler_4.1.1    systemfonts_1.0.2 rlang_0.4.12
## [57] grid_4.1.1        rstudioapi_0.13   labeling_0.4.2    rmarkdown_2.11
## [61] gtable_0.3.0      abind_1.4-5       curl_4.3.2        R6_2.5.1
## [65] lubridate_1.8.0   zoo_1.8-9         knitr_1.36        fastmap_1.1.0
## [69] survMisc_0.5.5    utf8_1.2.2        stringi_1.7.5     Rcpp_1.0.7
## [73] vctrs_0.3.8       tidyselect_1.1.1  xfun_0.27
```

Works Cited

Wolf, P.A., D'Agostino, R.B., Belanger, A.J., and Kannel, W.B. (1991) Probability of Stroke: A Risk Profile From the Framingham Study. *Stroke, 22* (3), 312-318

Appendix

Table 1: Data summary

	Male		Female	
	Censored	Stroke	Censored	Stroke
	(N=1849)	(N=48)	(N=2321)	(N=57)
Time to censoring or stroke (days)				
Mean (SD)	3460 (623)	2160 (1100)	3550 (477)	2040 (1150)
Median [Min, Max]	3650 [26.0, 3650]	2290 [294, 3640]	3650 [58.0, 3650]	1980 [22.0, 3620]
Age (years)				
Mean (SD)	49.5 (8.69)	55.7 (7.57)	49.7 (8.53)	57.2 (7.51)
Median [Min, Max]	49.0 [33.0, 69.0]	57.0 [36.0, 68.0]	49.0 [32.0, 70.0]	59.0 [38.0, 68.0]
BMI				
Mean (SD)	26.2 (3.37)	26.6 (4.13)	25.5 (4.49)	27.2 (5.51)
Median [Min, Max]	26.1 [15.5, 40.4]	26.6 [18.6, 38.4]	24.8 [16.0, 56.8]	27.2 [18.1, 42.5]
SBP (mmHg)				
Mean (SD)	131 (18.7)	151 (27.9)	133 (23.6)	162 (34.3)
Median [Min, Max]	128 [83.5, 217]	145 [104, 232]	128 [83.5, 244]	156 [102, 295]
Use of anti-hypertensive medication				
No	1813 (98.1%)	44 (91.7%)	2238 (96.4%)	48 (84.2%)
Yes	36 (1.95%)	4 (8.33%)	83 (3.58%)	9 (15.8%)
Total cholesterol (mg/dL)				
Mean (SD)	234 (42.4)	238 (46.2)	239 (46.1)	252 (45.6)
Median [Min, Max]	231 [113, 696]	228 [164, 405]	237 [129, 600]	253 [166, 358]
Diabetes				
No	1800 (97.4%)	41 (85.4%)	2268 (97.7%)	53 (93.0%)
Yes	49 (2.65%)	7 (14.6%)	53 (2.28%)	4 (7.02%)
Smoking status				
No	733 (39.6%)	17 (35.4%)	1383 (59.6%)	37 (64.9%)
Yes	1116 (60.4%)	31 (64.6%)	938 (40.4%)	20 (35.1%)
History of CVD				
No	1740 (94.1%)	44 (91.7%)	2263 (97.5%)	51 (89.5%)
Yes	109 (5.90%)	4 (8.33%)	58 (2.50%)	6 (10.5%)

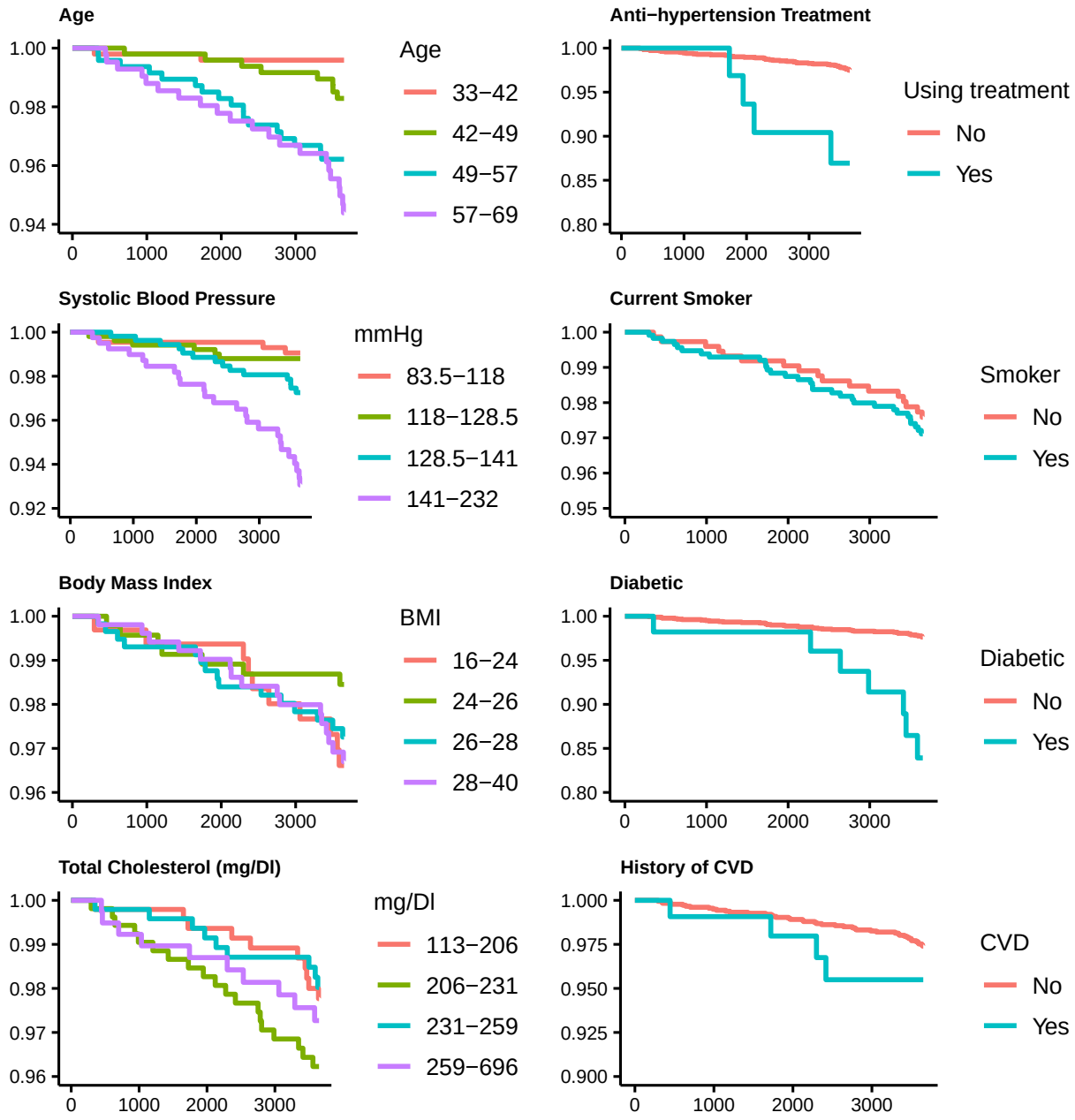

Figure 1: Kaplan-Meier Survival Curves for Male Sex

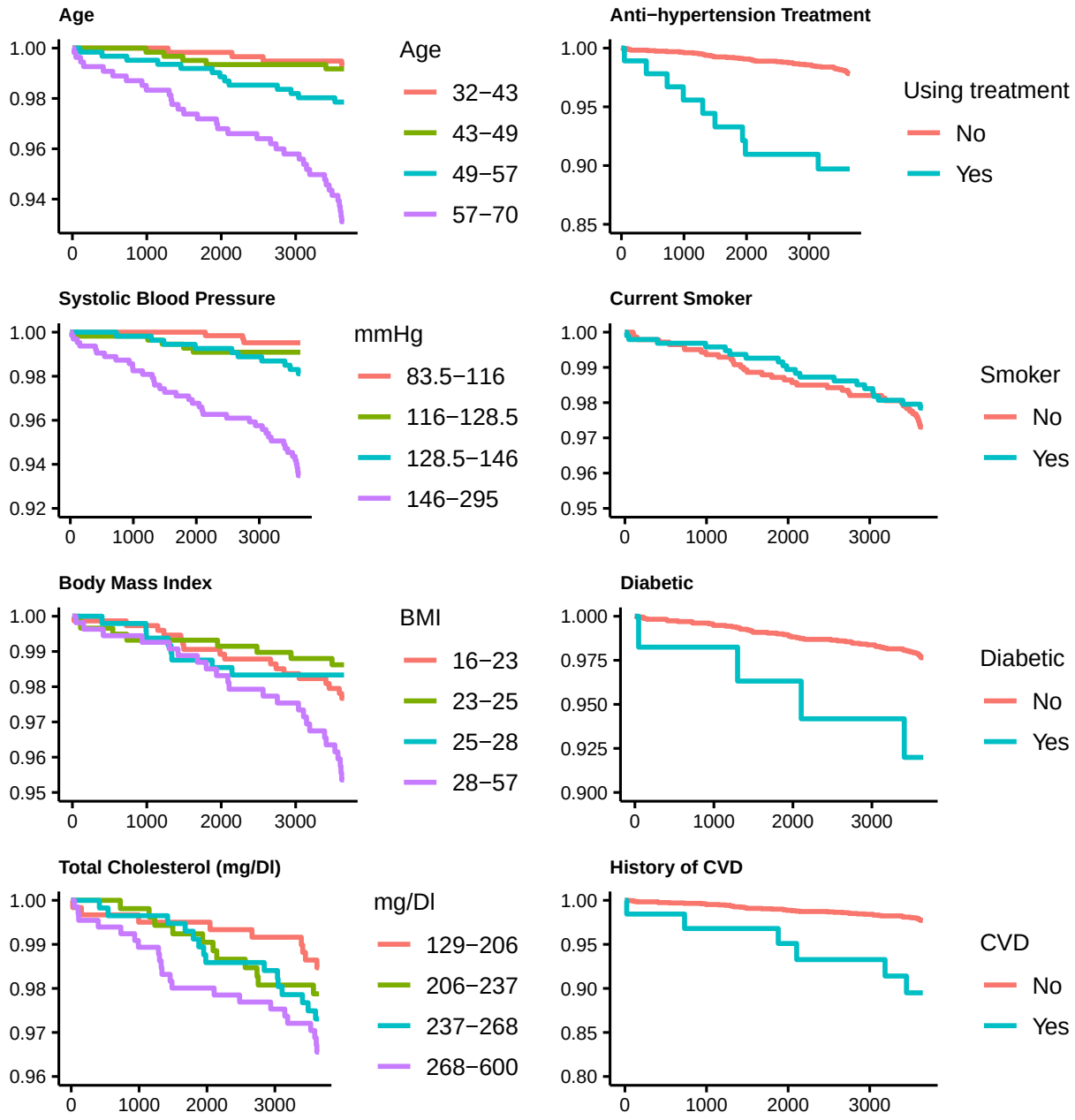

Figure 2: Kaplan-Meier Survival Curves for Female Sex

Table 2: Model coefficients (stratified by sex)

	Male			Female		
Variable	Hazard Ratio	95% CI	p-value	Hazard Ratio	95% CI	p-value
Age	1.07	(1.03, 1.11)	0.001	1.07	(1.03, 1.12)	<0.001
SBP	1.03	(1.02, 1.05)	<0.001	1.03	(1.02, 1.03)	<0.001
Diabetic	4.80	(2.1, 10.99)	<0.001			
Current Smoker	1.92	(1.03, 3.56)	0.039			

Table 3: Probability of Stroke at 10 Years stratified by baseline age in years and sex

	Male							Female						
	55	60	65	70	75	80	85	55	60	65	70	75	80	85
Average	0.02	0.03	0.04	0.06	0.08	0.11	0.15	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With smoking	0.03	0.04	0.05	0.07	0.10	0.14	0.18	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With treated hyptertension	0.02	0.03	0.04	0.06	0.08	0.11	0.15	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With untreated hypertension	0.02	0.03	0.04	0.06	0.08	0.11	0.15	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With diabetes	0.10	0.13	0.18	0.24	0.31	0.41	0.51	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With CVD	0.02	0.03	0.04	0.06	0.08	0.11	0.15	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With smoking and diabetes	0.13	0.17	0.23	0.30	0.39	0.49	0.61	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With smoking and CVD	0.03	0.04	0.05	0.07	0.10	0.14	0.18	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With smoking and treated hypertension	0.03	0.04	0.05	0.07	0.10	0.14	0.18	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With smoking and untreated hypertension	0.03	0.04	0.05	0.07	0.10	0.14	0.18	0.02	0.03	0.04	0.06	0.08	0.11	0.16
With all conditions	0.13	0.17	0.23	0.30	0.39	0.49	0.61	0.02	0.03	0.04	0.06	0.08	0.11	0.16

Table 4: Change in risk factors from first observation period to second

	Male		Female	
	Censored	Stroke	Censored	Stroke
	(N=1564)	(N=30)	(N=1956)	(N=45)
Started Smoking				
No	1511 (96.6%)	30 (100%)	1894 (96.8%)	44 (97.8%)
Yes	53 (3.39%)	0 (0%)	62 (3.17%)	1 (2.22%)
Stopped Smoking				
No	1391 (88.9%)	26 (86.7%)	1855 (94.8%)	42 (93.3%)
Yes	173 (11.1%)	4 (13.3%)	101 (5.16%)	3 (6.67%)
Change in Smoking Status				
No	1338 (85.5%)	26 (86.7%)	1793 (91.7%)	41 (91.1%)
Yes	226 (14.5%)	4 (13.3%)	163 (8.33%)	4 (8.89%)
Recognized Diabetic				
No	1531 (97.9%)	30 (100%)	1928 (98.6%)	44 (97.8%)
Yes	33 (2.11%)	0 (0%)	28 (1.43%)	1 (2.22%)
Started BP Medication				
No	1490 (95.3%)	22 (73.3%)	1773 (90.6%)	33 (73.3%)
Yes	74 (4.73%)	8 (26.7%)	183 (9.36%)	12 (26.7%)
Stopped BP Medication				
No	1554 (99.4%)	30 (100%)	1941 (99.2%)	44 (97.8%)
Yes	10 (0.639%)	0 (0%)	15 (0.767%)	1 (2.22%)
Change in BP Medication Use				
No	1480 (94.6%)	22 (73.3%)	1758 (89.9%)	32 (71.1%)
Yes	84 (5.37%)	8 (26.7%)	198 (10.1%)	13 (28.9%)
Change in history of CVD				
No	1482 (94.8%)	28 (93.3%)	1912 (97.8%)	42 (93.3%)
Yes	82 (5.24%)	2 (6.67%)	44 (2.25%)	3 (6.67%)
Change in SBP				
Mean (SD)	5.10 (16.0)	4.87 (27.1)	5.14 (16.6)	5.44 (20.8)
Median [Min, Max]	5.00 [-79.0, 75.0]	7.50 [-77.0, 45.0]	4.50 [-79.0, 106]	3.50 [-47.5, 41.0]
Change in BMI				
Mean (SD)	0.00525 (1.61)	-0.183 (1.85)	0.130 (1.92)	-0.635 (2.57)
Median [Min, Max]	0.00500 [-9.03, 5.37]	-0.395 [-4.51, 3.67]	0.180 [-10.5, 10.4]	-0.500 [-9.67, 5.42]
Change in Cholesterol				
Mean (SD)	8.44 (30.2)	10.9 (39.5)	17.3 (34.6)	4.96 (39.2)
Median [Min, Max]	9.00 [-159, 96.0]	18.5 [-101, 76.0]	17.0 [-137, 321]	9.00 [-136, 78.0]