**HW01 BIOST 518/514**

***ID: 1538***

Question 1.

Within the dataset, the time to death variable characterizes the follow-up time for all observations including censored observations. The minimum value (1480 days) is just over four years. Thus, at the four year timepoint we have data on the vital statuses for every participant.

Question 2.

**Methods:** In order to provide baseline descriptives for these data, a binary indicator variable was created for four-year vital status. There was no censoring of deaths within the four-year period. Descriptive statistics are shown in table 1 for all subjects and by vital status subgroup. Within this table the mean, standard deviation (SD), and range are shown for continuous variables (age, body mass index (BMI), cholesterol, and c-reactive protein (crp), while percentages are shown for binary variables (male, smoking history, and prior cardiovascular disease (CVD)). All computing was done with Stata version 13.1.

**Results:**  Data was available for 5000 participants; however, 67 participants had missing crp measures. These individuals were not included in the subsequent analyses.

Among 4933 participants, 4449 remained alive at four years post-enrollment while 484 had died. Table 1 provides descriptive statistics within these groups and for all participants. On average, individuals who died within four years tended to be older and were more likely to be male, smokers, and have a history of CVD. These individuals also showed trends for a lower cholesterol and a higher crp than individuals surviving for at least four years. The mean crp was 5.38 (mg/L) among those who died within four years compared with 3.42 mg/L among those who survived at least four years.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Four Year Vital Status** | |
| **Baseline Characteristics\*** | **All Subjects** | **Alive Within Four Years** | **Death Within Four Years** |
| Number | 4933 | 4449 | 484 |
| Age | 72.8 (5.56, 65-100) | 72.4 (5.29, 65-98) | 76.24 (6.70, 65-100) |
| Percent male (%) | 42.0% | 40.0% | 60.1% |
| BMI | 26.7 (4.72, 14.7-58.8) | 26.7 (4.69, 14.7-58.8) | 26.32 (5.0, 14.8-48.1 |
| Cholesterol | 211.7 (39.23, 73-430) | 212.5 (38.90, 78-430) | 204.13 (41.43, 73-396) |
| Smoking History (%) | 12.2% | 12.0% | 14.3% |
| Prior CVD (%) | 22.9% | 20.9% | 41.9% |
| crp (mg/L) | 3.61 (6.15, 0-108) | 3.42 (5.87, 0-108) | 5.38 (8.10, 0-55) |
| \*All baseline characteristics are mean (SD, range) unless otherwise indicated. | | |  |

Question 3.

**Methods:** We compared the mean crp levels between participants who survived at least four years and those who died within four years. A two sample t test was performed, which provided an estimate of the difference as well as the 95% confidence intervals (CI). An assumption of unequal variances for groups facilitated the use of Satterthwaite degrees of freedom. We used a significance level of 0.05 for this statistical test.

**Results:** Among the 4449 participants who survived at least four years the mean crp was 3.42 mg/L at baseline. The mean crp was 5.38 mg/L among those who died within four years. With 95% confidence, the observed data are consistent with participants who died within four years having a mean crp between 1.21 and 2.70 mg/L higher than those who survived to four years (difference in means: 1.95). This finding is statistically significant with a two-sided p-value of 0.0000. We thus reject the null hypothesis of no difference in the mean crp between these two groups. There is evidence to suggest that the distributions of crp level is different between those who survive and do not survive past four years.

Question 4.

**Methods:** We compared the geometric mean crp levels between participants who survived at least four years and those who died within four years. Crp was collected as integer data, with a crp level of 1 mg/L representing the lower limit of detection. In order to account for participants with crp 0 mg/L (N=21) within the calculation of geometric mean, we created a new crp value such that all zeros were assigned half of the lower limit of detection, that is 0.5 mg/L, and all other values remained the same. A two sample t test was performed with the log of crp, which provided the difference an estimate of the difference well as the 95% confidence intervals (CI). An assumption of unequal variances for groups facilitated the use of Satterthwaite degrees of freedom. In order to make inference on geometric means, estimates and CIs were exponentiated. We used a significance level of 0.05 for this statistical test.

**Results:** Among the 4449 participants who survived at least four years the geometric mean crp was 2.03 mg/L at baseline. The geometric mean crp was 2.97 mg/L among those who died within four years. With 95% confidence, the observed data are consistent with participants who died within four years having a geometric mean crp between 62.1% and 75.1% higher than those who survived to four years (difference in means: 0.683). This finding is statistically significant with a two-sided p-value of 0.0000. We thus reject the null hypothesis of no difference in the geometric mean crp between these two groups. There is evidence to suggest that the distributions of crp level is different between those who survive and do not survive past four years.

Question 5.

**Methods:** We compared the probability of death within four years based on crp status, with “high crp” defined as crp > 3 mg/L per the Mayo Clinic Guidelines for cardiovascular risk. Crp was collected as a continuous variable, so a binary variable was created to represent “high” crp status and “low to average” crp status. A Pearson’s chi squared test was performed with the 95% CI for probability of death within four years calculated via the Wald statistic. We used a significance level of 0.05 for this statistical test.

**Results:** Among 1175 participants with high crp, 183 (15.6%) died within four years. Among those 3758 participants with normal to average crp, 301 (8.01%) died within four years. With 95% confidence, the observed data are consistent with participants with high crp having probability of death between 5.32% and 9.81% higher than those with low to average crp (risk difference: 7.56% in absolute terms). This finding is statistically significant with a two-sided p-value of 0.0000. We thus reject the null hypothesis that probability of death and high crp are independent.

Question 6.

**Methods:** We compared the odds of death within four years based on crp status, with “high crp” defined as crp > 3 mg/L per the Mayo Clinic Guidelines for cardiovascular risk. Crp was collected as a continuous variable, so a binary variable was created to represent “high” crp status and “low to average” crp status. A Pearson’s chi squared test was performed to test to independence of the odds with the 95% CI for the odds ratio (OR) calculated via Woolf’s method. We used a significance level of 0.05 for this statistical test.

**Results:** Among 1175 participants with high crp, the odds of death within four years was 0.184. Among those 3758 participants with low to average crp, the odds was 0.0871. With 95% confidence, the observed data are consistent with participants with high crp having an odds of death between 1.74 and and 2.58-fold higher than those with low to average crp (OR: 2.12). This finding is statistically significant with a two-sided p-value of 0.0000. We thus reject the null hypothesis that probability of death and high crp are independent.

Question 7.

**Methods:** Kaplan-Meier methods were used to estimate survival distributions within the groups with “high” crp (defined as crp > 3 mg/L) and “low to average” crp (crp </= 3 mg/L). The hazard ratio (HR) and corresponding 95% CI for the instantaneous risk of death was estimating using a Cox regression with proportional hazards. A logrank statistic was computed in order to test the difference in these survival distributions. We used a significance level of 0.05 for this test.

**Results:** A Kaplan-Meier curve is shown below, which illustrates the survival distributions for the group with high crp (N=1175) and the group with low to average crp (N=3758). A marked trend for reduced survival is shown for the high crp group. Estimates of survival probabilities at one, two, three, and four years are also shown in the below table.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | **Probability of Survival As Estimated from Kaplan-Meier Curve** | | | |
| **Group** | **One Year** | **Two Years** | **Three Years** | **Four Years** |
| Low to Average crp (<=3 mg/L) | 0.988 | 0.971 | 0.948 | 0.92 |
| High crp (>3 mg/L) | 0.966 | 0.926 | 0.881 | 0.844 |
|  |  |  |  |  |

The group with high crp demonstrated a nearly 70% increase in the instantaneous risk of death. With 95% confidence, a HR of 1.69 is consistent with a true HR between 1.49 and 1.92. Based on two-sided p-value of 0.0000 in a logrank test, we can reject the null hypothesis of equality of survival functions. Findings suggest the probability of survival is associated with crp level > 3 mg/L.

Question 8.

I prefer to use a method where the least number of continuous variables are dichotomized or made categorical because it is more statistically precise. A method of analysis that allowed for both crp and survival time and crp to be continuous variables would be most ideal on this point; however, we have not technically learned these methods yet.

Simplicity of the measure for testing usually optimizes communicating the results. Thus, testing based on basic means or frequencies would probably be more easily understood than hazards, odds, or geometric means. Thus, I would probably choose the t-test of crp means by four-year vital status. While geometric means might more precise (that is, in normalizing the range of values for crp being averaged), we also have to account for the lower limit of detection in its calculation. We change values of the data to be slightly non-zero. We do not have to make any data changes to calculate the means, thus *a priori* we will perform a more “pure” analysis.