1932

Bistats 518 HW 1

12 Jan 15

1. The first shortest amount of follow up any participant was in the study without and event (death) before the end of the study was 4 years (1480 days). The first censored observation is censored at 4 years so we have full information about death in all participants up to 4 years.

2. Methods: One indicator variable was created for death within 4 years of study enrollment such that no observations are censored in the present dataset. Variables are presented by blood C Reactive Protein (CRP) levels indicating risk of heart disease. CRP was defined as less than 1 mg/L (low risk), 1-3 mg/L inclusive (average risk), and greater than 3 mg/L (high risk). Statistics presented for continuous variables (age, body mass index [BMI], and cholesterol) include mean, standard deviation, minimum and maximum. Percentages are presented for categorical variables (sex, smoking status, prior cardiovascular disease [CVD], and death within 4 years).

Results: The original data set included 5000 study participants. Data on variables included in this analysis (CRP, BMI, cholesterol, or smoking status) are missing for 89 participants. These participants were excluded in the present analysis, resulting in a total of 4911 subjects included in this analysis.

Descriptive statistics within levels of CRP are presented in Table 1. Of the 4910 subjects, 426 (9%) had CRP levels indicating low risk for heart disease ( < 1mg/L), 3313 (67%) had levels indicating average risk (1-3 mg/L), and 1172 (24%) had levels indicating high risk (> 3 mg/L). Subjects with low risk group were more likely to be male and have lower BMIs. Additionally, subjects in the high risk group were more likely have been smokers, had a history of CVD and died within 4 years. Smokers represented 16% of the high risk group and only 10% and 11% in the low and average risk groups respectively. Consistently the high risk group had the highest prevalence of prior CVD (29%) compared to the lower risk groups (18 % in low group, 11% in average risk). The increase of these risk factors for all cause mortality (BMI, smoking, and prior CVD) across groups of CRP level was consistent with increased death within 4 years of study enrollment. A high proportion of the high risk CRP group died (16%) when compared to the average (8%) and low (5%) risk groups. Lastly there appears to be no trend in age and cholesterol by CRP levels.



3. Method: The analysis compares the means of CRP level between two groups of subjects classified by vital status at 4 years after study enrollment (ie, those that died before 4 year point compared to those who survived). The differences of means were tested using a t test that does not assume equal variances in the two groups (Satterthwaite approximation). Similarly the 95% confidence interval was calculated using methods that do not assume equal variances.

Results: The mean blood CRP level among subjects that survived the 4 year period was lower (n=4429; 3.42 mg/l) compared to subjects that did not survived the 4 year period since enrollment (n=482; 5.39 mg/l). We estimate a difference in mean CRP level to be 1. 97 mg/L lower among those that survived the 4 year period compared to those that did not. Our data would not be unusual, based on a 95% confidence interval, if the difference in mean CRP levels was between 1.23 mg/l and 2.72 mg/l lower among those who survived the 4 year enrollment period compared to those that did not. These results of a t test allowing for unequal variances (two sided p <0.001) are statistically significant at alpha level of 0.05. We can therefor reject the null hypothesis that the mean blood CRP levels are no different between the groups defined by 4 year survival (those that survived 4 years past enrollment compared to those who don’t).

4. Method: The analysis compares the geometric means of CRP level between two groups of subjects classified by vital status at 4 years after study enrollment (ie, those that died before 4 year point compared to those who survived). To retain the zero values of CRP differentiated from a missing value, the value 0.5 was used to represent a true CRP blood level of zero. The differences of mean log transformed blood CRP level were tested using a t test that does not assume equal variances in the two groups (Satterthwaite approximation). Similarly the 95% confidence interval for the difference in overall population log mean was calculated using methods that do not assume equal variances. Log mean estimates and confidence intervals were exponentiated back to the geometric mean scale presented below.

Results: The mean blood CRP level among subjects that survived the 4 year period was lower (n=4429; 2.03 mg/l) compared to subjects that did not survived the 4 year period since enrollment (n=482; 2.98 mg/l). We estimate a difference in geometric mean CRP level to be 32% lower among those that survived the 4 year period compared to those that did not. Our data would not be unusual, based on a 95% confidence interval, if the true ratio of population geometric mean CRP levels was between 25% and 38% lower among those who survived the 4 year enrollment period compared to those that did not. These results of a t test allowing for unequal variances (two sided p <0.001) on log transformed CRP level are statistically significant at alpha level of 0.05. We can therefor reject the null hypothesis that the geometric mean blood CRP levels are no different between the groups defined by 4 year survival (those that survived 4 years past enrollment compared to those who don’t).

5. Method: The analysis compares the proportion of subjects who die within 4 years of study enrollment who had CRP levels categorized as high risk for heart disease, greater than 3 mg/L and those that have levels less than or equal to 3 mg/l. The differences of proportions of those who died within 4 years (probability of death) were tested using a Pearson’s chi test and the 95% confidence interval was calculated using Wald statistics.

Results: A higher proportion of those with high levels of CRP died within 4 years of study enrollment compared to those with low or average levels of CRP. Of the 3041 subjects with low or average levels of CRP (≤ 3 mg/l), 7.27% died compared to 14.0% of the 1870 subjects with high levels of CRP (> 3 mg/l). We estimate a difference in the probability of death within 4 years of enrollment to be 6.69% lower among those with low or average CRP levels compared to those with high CRP levels. Our data would not be unusual, based on a 95% confidence interval, if the difference in absolute probability of 4 year death was between 4.87% and 8.51% lower among those with low or average CRP levels compared to those with high CRP levels. These results of a chi squared test (two sided p <0.001) are statistically significant at alpha level of 0.05. We can therefor reject the null hypothesis that the distributions of CRP levels by 4 year death probabilities.

6. Method: The analysis compares the odds of death within 4 years of study enrollment for subjects who had CRP levels categorized as high risk for heart disease, greater than 3 mg/L and those that have levels less than or equal to 3 mg/l. The odds ratio (OR) of 1 was tested using a Pearson’s chi test and the 95% confidence interval was calculated using Wald statistics.

Results: The odds of death within 4 years was 0.0784 among those with average/low levels of blood CRP (≤ 3 mg/L) and 0.162 among those with high levels of blood CRP (> 3 mg/L). We estimate an OR of 2.07 when comparing those with high CRP levels relative to those with average/low levels. Our data would not be unusual, based on a 95% confidence interval, if the true OR was between 1.71 and 2.50. These results of a chi squared test (two sided p <0.001) are statistically significant at alpha level of 0.05. We can therefor reject the null hypothesis that the distribution of blood CRP levels do not differ by blood CRP levels.

7. Methods: The analysis compares the survival distribution estimated with Kaplan-Meier estimates between subjects who had CRP levels categorized as high risk for heart disease, greater than 3 mg/L and those that have levels less than or equal to 3 mg/l. Cox proportional hazards regression with Huber-White sandwich estimator for the standard of errors was used to calculate the hazard ratio estimate and the 95% confidence interval. The difference in distributions between the two groups was assed using the log rank test.

Results:

The graph and table below summarize the Kaplan Meier estimates of survival between two groups defined by high CRP blood levels (> 3mg/l) and average/low CRP blood levels (≤ 3 mg/l). The results indicate a higher probability of survival at all time points among the group with average/low CRP levels compared to the group with high CRP levels. We estimate the instantaneous risk of death is estimated to be 32.4% higher for those who had a high blood CRP levels (HR 1.32). Based on a 95% confidence interval our data is consistent if the true hazard estimate was between 13% and 55% higher among those who had higher blood CRP levels (95% CI: 1.12, 1.55). The results of a log rank test are statistically significant (two sided p=0.0006) using at alpha equal 0.05. We can therefor reject the null hypothesis that the Kaplan-Meier survival distribution do not differ between those with CRP blood levels greater than 3 mg/l (high risk group) and those with CRP levels less than or equal to 3 mg/l (average/low risk group).



|  |  |  |
| --- | --- | --- |
|  | **Survivor probabilities (Kaplan Meier estimate)** | |
|  | CRP ≤ 3 mg/l | CRP >3 mg/l |
| **1 year** | 0.989 | 0.972 |
| **2 years** | 0.975 | 0.935 |
| **3 years** | 0.954 | 0.898 |
| **4 years** | 0.927 | 0.860 |
| **5 years** | 0.891 | 0.819 |
| **6 years** | 0.860 | 0.780 |
| **7 years** | 0.820 | 0.732 |
| **8 years** | 0.776 | 0.679 |

8. Analysis plans should be prepared during the study planning and design stage of an investigation rather than after data collection and analysis has began.

You probably first want to interpreted the scientific question into a statistical question. This starts with identifying the outcome and exposure/ dependent and independent variables. CRP is the logical exposure and independent variable on which to condition the analysis. CRP it always is measured before the event death and can be divided into levels. Next, you must decide the best way to compare the survival distribution by CRP levels. This is the definitive step of the analysis. I would carefully look at the distribution of CRP variable and gather information about the biological importance and clinical relevance of different levels. While you give up some statistical precision and potential depth when categorizing a continuous variable, you gain the ability to compare the distribution of survival between a few groups. I would decide to categorize in this case because it 1) I don’t know how to complete hazard regression with continuous variables and 2) we can gain interpretability if we carefully categorize the groups.

The next step is to carefully categorize the groups. I would consider biological or clinical importance of CRP levels and groups. It seems here that a level above 3 represents high risk of heart disease and less than 3 represents average to low risk. We would then have to choose if we are interested in those with low or increase risk for heart disease. We are most often interested in increased risk and so dichotomizing CRP into two levels at 3 mg/l seems appropriate. You may also want to explore more categories and the potential for multiplicative, ratio rather than additive scale for the group intervals. Two variables simplifies the analysis to a level that I am currently able to handle and interpret.

After exploring the descriptive statistics I would assess the difference between the distributions of survival in the two groups. First using a measure of central tendency such as means and geometric means (or both). Second, I would compare the distributions with Kaplan meier estimates because of all the censored data. It is important to consider is the interpretation of the analysis results. An analysis comparing geometric means in two groups might give us the most precision, but is also difficult to interpret. If the goal of the analysis was to present results interpretable for the general population I would report difference in means and survival probabilities.