1683

Biost 515/518 hw01

1. *The observations of time to death in this data are subject to (right) censoring. Nevertheless, problems 2 – 6 ask you to dichotomize the time to death according to death within 4 years of study enrollment or death after 4 years. Why is this valid? Provide descriptive statistics that support your answer.*

**Answer:** When we look at the the total time(in days) that the censored participants (death=0) was observed, the minimum time is 1480 days, or just over 4 years. Thus, the status that a participant was censored or not is known at 4 year. It is reasonable to dichotomize the time to death according to death within 4 years of study enrollment or death after 4 years.

1. *Provide a suitable descriptive statistical analysis for selected variables in this dataset as might be presented in Table 1 of a manuscript exploring the association between serum CRP and 4 year all-cause mortality in the medical literature. In addition to the two variables of primary interest, you may restrict attention to age, sex, BMI, smoking history, cholesterol, and prior history of cardiovascular disease.*

**Methods:** The total time (in days) that the participant was observed on study between the date of study enrollment and death or data analysis were dichotomized to within 4 years of study enrollment or death after 4 years. Descriptive statistics are presented within groups defined by serum CRP level( less than 1mg/L, between 1 and 3 mg/L, and greater than 3 mg/L). For continuous variables(age, cholesterol, systolic blood pressure and BMI) I adopted the mean, standard deviation, minimum and maximum. For binary variables(male, smoker, prevdis and death within 4 years), I used percentages.

**Results:** Data is available on 5000 participants, **67 of those are missing data on serum C reactive protein (CRP).** **In addition to the two variables of primary interest, I restricted my attention to another 7 variables, which were age, sex, BMI, smoker, cholesterol, systolic blood pressure, prevdis**. **After omitting observations with missing values on any of those interested variables, 4901 participants met the criteria (another 32 subjects were excluded). A total of 99 subjects were excluded from the descriptive analyses in this part.** We don’t know whether omitting these missing data have impact on the generalizability of our results or not.

**Interpretation:** Of 4901 participants with available data on CRP level, 425 had serum CRP measurements less than 1mg/L, 3304 had measurements between 1mg/L and 3mg/L, 1172 had measurements greater than 3mg/L. Table 1 presents descriptive statistics within each group. No consistent trend was seen across groups in age, serum cholesterol level. Subjects having serum CRP in the highest interval (>3mg/L)were more likely to have higher mean systolic blood pressure (137.48 mmHg)compared to those in the other two intervals( 132.83 and 136.54 mmHg). Participants in the highest serum CRP interval (>3mg/L)also appeared to have higher average BMI than those in the other two intervals.

Subjects having serum CRP in the lowest interval were more likely to be male than the other two intervals. Participants in the highest serum CRP interval (>3mg/L) also had the highest proportion of smokers(16%), highest percentage of prevalent atherosclerotic disease at study enrollment(29%) and highest proportion of death within 4 years of follow-up(16%).

*Table 1. Basic characteristics of participants by serum C Reactive Protein Level.*

|  |  |  |
| --- | --- | --- |
|  | Serum C Reactive Protein (CRP) Level |  |
|  | <1 mg/L (n=425) | 1 - 3 mg/L (n=3304) | >3 mg/L (n=1174) | Total |
| Age (yrs)1 | 73.44 (5.80; 65 -94) | 72.73 (5.53; 65 -100) | 72.73 (5.58; 65 -93) | 72.83 (5.60; 65 -100) |
| Cholest2 (mg/dL)1 | 206.16(40.50; 109.00 -407.00) | 212.79(38.56; 73.00 -363.00) | 210.52(40.40; 97.00 -430.00) | 211.70（39.29; 73.00 -430.00) |
| SysBP3 (mmHg)1 | 132.83(21.24; 89.00 -199.00) | 136.54(21.83; 82.00 -235.00) | 137.48(21.85; 77.00 -230.00) | 136.60(21.86; 77.00 -235.00) |
| BMI (kg/cm2)1 | 23.82(3.65; 15.60 -38.60) | 26.38(4.30; 14.70 -53.20) | 26.45(5.47; 15.30 -58.80) | 26.67(4.735; 14.70 -58.80) |
| Male (%) | 46 | 43 | 37 | 41.92 |
| Smoker (%) | 10 | 11 | 16 | 12.09 |
| Prevdis4(%) | 18 | 21 | 29 | 22.98 |
| Death within 4 years (%) | 5 | 8 | 16 | 9.90 |

1Descriptive statistics presented are the mean (standard deviation; minimum -maximum)

2Serum cholesterol

3Systolic blood pressure

4Indicator of prevalent atherosclerotic disease at study enrollment

1. *Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing mean CRP values across groups defined by vital status at 4 years.*

**Methods:** I chose mean serum CRP as a measurement of serum CRP. Mean serum CRP levels were compared between participants who survived at least 4 years and those who died within 4 years since being enrolled. Differences in the mean were tested using a t test allowing for unequal variances. 95% confidence intervals for the difference in population means were calculated based on the same handling of variances.

**Results:** Mean serum CRP was 5.392mg/L among the 482 participants who died within 4 years after being enrolled and 3.422mg/L among the 4419 participants who survived at least 4 years.

**Interpretation:** Based on a 95% confidence interval computed allowing for unequal variances, the tendency that participants with a 1.970mg/L higher mean CRP level dying earlier would not be unusual if the true difference population means of CRP level were between 1.224mg/L and 2.716mg/L higher on average among those who die within 4 years.The t-test allowing for unequal variances showed that this observation is statistically significant at a 0.05 level of significance(two-sided P= 0.0000003054), we can reject the null hypothesis that the mean serum CRP levels are not different by survival status at 4 years.

1. *Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing geometric mean CRP values across groups defined by vital status at 4 years. (Note that there are some measurements of CRP that are reported as zeroes. Make clear how you handle these measurements.)*

**Methods:** There were 425 out of 4901 participants with a serum CRP value of 0 mg/L. Thus I replaced all these “0 mg/L” with half of the minimum non-zero value of serum CRP level in these sample, which was 0.5 mg/L. Then geometric mean serum CRP levels were compared between participants who died within 4 years since study enrollment and who survived at least 4 years. Differences in mean of log transformed serum CRP levels were tested using a t test allowing for unequal variances. 95% confidence intervals for the difference in population means for log transformed serum CRP were also calculated allowing for unequal variances. Estimates and confidence interval were exponentiated to get the geometric means.

**Results:** Geometric mean serum CRP was 2.03 mg/L among 4419 participants who survived at least 4 years after enrollment and 2.98 mg/L among 482 participants who died within 4 years.

**Interpretation:** Based on a 95% confidence interval allowing for unequal variances, this observed tendency of 31.9% (1-68.2%) higher geometric mean serum CRP among participants died within 4 years would not be deemed as unusual if the true ratio of population geometric CRP means were between 61.9% and 74.9% (those who die within 4 years Vs those who survive at least 4 years). Using a t test on log transformed CRP level that allows for unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P=0.000000000000011143), thus we can reject the null hypothesis that the geometric mean serum CRP levels are not different by survival status at 4 year.

1. *Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing the probability of death within 4 years across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).*

**Methods:** The proportion of participants dying within 4 years since study enrollment were compared between participants who had serum CRP greater than 3mg/L and those whose serum CRP was 3mg/L or less. Differences in the probability death within 4 years were tested using Person’s chi squared test for independence. 95% confidence intervals for the difference in population 4-year mortality probabilities were computed using Wald statistics.

**Results:** Of the 3729 participants whose CRP was lower than or equal to 3mg/L, 8.02% were observed to die within 4 years, while 15.61% of the participants with serum CRP higher than 3 mg/L died within 4 years since study enrollment.

**Interpretation:** Based on a 95% confidence interval, this 7.59% higher death probability in participants with higher serum CRP would not be unusual if the true difference in death probability is between 5.34% to 9.85% higher in the populations with serum CRP higher than 3mg/L. Using a chi squared test, this observation is statistically significant at a 0.05 level of significance (two-sided P=0.00000000000003994), and we can reject the null hypothesis that the 4-year all-causes mortality is not associated with serum CRP levels with confidence.

1. *Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing the odds of death within 4 years across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).*

***Methods:*** The odds of participants dying within 4 years of study enrollment were compared between participants who had serum CRP greater than 3 mg/L and those whose CRP was lower than or equal to 3 mg/L. An odds ratio different from 1 was tested using Chi squared test. 95% confidence interval for the odds ratio was also calculated using Woolf’s method.

***Result:*** Of the 3729 participants whose CRP was lower than or equal to 3mg/L, the odds of dying within 4 years since enrollment was 0.0802, while for participants with serum CRP higher than 3 mg/L the odds of dying within 4 years since study enrollment was 0.1561.

***Interpretation:*** Based on a 95% confidence interval, this observed odds ratio of 2.1226 for the comparison of the over 3mg/L serum CRP group to those with a serum CRP ≤3mg/L would not be unusual if the true odds ratio were between 1.7429 to 2.5851. A chi squared test two sided P value of 0.000000000000025868 suggests that we can, with confidence, reject the null hypothesis that the odds of 4 year mortality is not associated with serum CRP level.

1. *Perform a statistical analysis evaluating an association between serum CRP and all-cause mortality over the entire period of observation of these subjects by comparing the instantaneous risk of death across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).*

***Methods:*** I used Kaplan-Meier estimated with stratas defined by serum CRP ≤3mg/L and serum CRP>3mg/L to estimate the survival distribution. Difference in survival distributions between those two groups was tested using logrank statistic. Then I used Cox proportional hazards regression to get the hazard ratio and 95% CI .

***Results:*** The following graph shows Kaplan-Meier estimates of survival probability for 3729 participants whose CRP was lower than or equal to 3mg/L and 1172 participants with serum CRP higher than 3 mg/L. Obviously, we can see from the graph the tendency for lower survival probabilities for the higher CRP group at every point in time. The instantaneous risk of death is estimated to be 68.7% higher for the participants with a serum CRP level over 3mg/L compared to those equal to or below 3mg/L.

***Interpretation:*** Based on a 95% confidence interval, this observed hazard ratio of 1.687 would not be seen as unusual if the true hazard ratio were between 1.485 to 1.917. A logrank test two sided P value of 0.0000000000000003331 suggest hat we can, with confidence, reject the null hypothesis that the probability of survival is not associated with serum CRP level.

1. *Supposing I had not been so redundant (in a scientifically inappropriate manner) and so prescriptive about methods of detecting an association, what analysis would you have preferred a priori in order to answer the question about an association between mortality and serum CRP? Why?*

I would prefer a multi-variate survival analysis, which is the cox proportional hazards model. By doing this, I could summarize the survival distribution conditioned on a multicative level of CRP. And I could also include other data like male, smoker, etc as co variables in the model , and I could get better understanding of interactions between these variables. However, a simple test of geometric means across survival groups defined by CRP level would also do the job. I prefer geometric means because of a better precision.