**Carlos J Gallego**

**BIOST 536 HW#3**

1. Provide suitable descriptive statistics for this dataset as might be presented in Table 1 of a manuscript appearing in the medical literature. (Because the primary question is comparing 24 month relapse free survival across groups defined by nadir PSA, you might consider presenting descriptive statistics in groups according to some dichotomization of nadir PSA levels. Alternatively, you could provide descriptive statistics within groups defined by whether the subjects relapse within 24 months or not.)

|  |  |  |
| --- | --- | --- |
|   | **PSA <=0.95** | **PSA > 0.95** |
|   | **n** | **%** | **n** | **%** |
| **Age**  |   |   |   |   |
|  **<60** | 1 | 4 | 1 | 4 |
|  **60-69** | 20 | 80 | 14 | 56 |
|  **70-79** | 4 | 16 | 8 | 32 |
|  **80-89** | 0 | 0 | 2 | 8 |
| **Grade**  |   |   |   |   |
|  **1** | 5 | 22.73 | 5 | 26.32 |
|  **2** | 8 | 36.36 | 7 | 36.84 |
|  **3** | 9 | 40.91 | 7 | 36.84 |
| **Performance status**  |   |   |   |   |
|  **<=70** | 3 | 12.5 | 7 | 29.17 |
|  **80** | 10 | 41.67 | 11 | 45.83 |
|  **>=90** | 11 | 45.83 | 6 | 25 |
| **BSS** |   |   |   |   |
|  **1** | 5 | 20.83 | 0 | 0 |
|  **2** | 6 | 25 | 7 | 29.17 |
|  **3** | 13 | 54.17 | 17 | 62.5 |

1. Perform logistic regression analyses to determine whether the distribution of relapse within 24 months differs across groups defined by nadir PSA level after adjustment for bone scan score and performance status. For each of the following models, provide full statistical inference for your measure of association.
	1. Perform an adjusted logistic regression comparing the odds of relapse within 24 months across groups defined by the nadir PSA level when modeled as a continuous, untransformed variable.

After adjusting for performance status and bone scan score, for every one ng/ml increase in nadir PSA, the odds ratio of relapse within 24 months was 1.034 (0.943-1.133), but this odds ratio was non-significant (p=0.476).

* 1. Perform an adjusted logistic regression comparing the odds of relapse within 24 months across groups defined by the nadir PSA level when modeled as a continuous, log transformed variable.

After adjusting for performance status and bone scan score, for every 2.72-fold (e) increase in nadir PSA, the odds ratio of relapse within 24 months was 2.36 (1.268-4.402), and this odds ratio was significant (p=0.007).

* 1. Perform an adjusted logistic regression comparing the odds of relapse within 24 months across groups defined by the nadir PSA level when modeled as linear splines with knots at 1, 4, and 16 ng/ml.

After adjusting for performance status and bone scan score, for every one ng/ml increase in nadir PSA when nadir PSA is <1, the odds ratio of relapse within 24 months was significant (p=0.031) at 29.67 (96% CI: 1.359-645.633), for every one ng/ml increase in nadir PSA when nadir PSA is between 1 and <4, the odds ratio of relapse within 24 months was non-significant (p=0.845) at 0.90 (95% CI: 0.327-2.496), for every one ng/ml increase in nadir PSA when nadir PSA is between 4 and <16, the odds ratio of relapse within 24 months was non-significant (p=0.096) at 1.37 (95% CI: 0.944-2.017), for every one ng/ml increase in nadir PSA when nadir PSA is 16 or higher, the odds ratio of relapse within 24 months was significant (p=0.043) at 0.98 (95% CI: 0.964-0.999).

* 1. For each of the above regression models, provide an interpretation of the intercept.

For 2a: the intercept is the odds of relapse within 24 months for subjects with a PSA of 0 when bone scan score and performance status are 0.

For 2b: the intercept is the odds of relapse within 24 months for subjects with a PSA of 0 when bone scan score and performance status are 0.

For 2c: the intercept is the odds of relapse within 24 months for subjects with a PSA of 0 when bone scan score and performance status are 0.

1. In this longitudinal study, we could instead have considered the “reverse” analyses in which nadir PSA is used as the response and the predictor is the indicator of relapse within 24 months.
	1. Perform linear regression analyses to determine whether there is an association between mean nadir PSA level and relapse within 24 months after adjustment for bone scan score and performance status. Make clear the statistical analysis you perform. Provide full statistical inference for your measure of association.

After adjusting for performance status and bone scan score, when comparing people that had a relapse within 24 months with those who didn’t, PSA was 23.5 ng/ml higher (95% CI: 0.476-46.558). This association was significant (p=0.046).

* 1. Perform linear regression analyses to determine whether there is an association between geometric mean nadir PSA level and relapse within 24 months after adjustment for bone scan score and performance status. Make clear the statistical analysis you perform. Provide full statistical inference for your measure of association. (Recall that inference on the geometric mean is obtained by performing linear regression on log transformed response variables.)

After adjusting for performance status and bone scan score, when comparing subjects who relapsed within 24 months with those who didn’t, the PSA was 13.66 times higher (95% CI: 4.130-45.163). This association was significant (p<0.001).

1. Consider the analyses performed in problems 2 and 3 above.
	1. What are the relative merits of the five analyses. Which might you prefer *a priori*? Why?

The logistic models have an intrinsic advantage when using a binary outcome because it deals with probabilities. We will not use a priori the linear models in question 3 because they don’t address a clinically relevant question: you want to predict relapse based on your nadir PSA and not all the way around. The linear spline model is neither the best one because its effect is mainly limited to the first spline, and there are better alternatives. Between the logistic models using nadir PSA and log nadir PSA as predictors of interest, I would prefer the first one, because from a clinical standpoint additive changes in your PSA are more relevant.

* 1. All of these analyses suffer from a serious definitional problem inherent in this study. Can you deduce this problem? (Hint: There is no analysis that you can do to address this problem. It is a problem with the study design.)

The problem is inherent in the observational nature of this study (a retrospective cohort). Although you can control for confounding, like we did for bone scan score and performance status, there is always some residual confounding or non-observed confounders that can affect this association. Another problem is that the prediction of clinical outcomes based on biomarkers is not susceptible to randomization, which would be the ideal assessment of a clinical question. A final problem would be the temporary relation of the predictors of interest and outcome variables, although it is unlikely that a relapse within those two years preceded the nadir PSA.