**Q1:** The sample (n = 209) consists of four age groups: <65 (n=77), 66-75 (n=44), 76-85 (n=40), and 85+ (n=48).

**Q2:** The patients’ average and median durations were 83 months and 85 months, respectively, with roughly 80% of durations at 50 to 120 months.

**Q3:** For the first treatment, 102 patients reported receiving Anti-Androgen, 158 reported receiving Docetaxel, 146 reported receiving Abiraterone, and 11 reported receiving Other.For the second treatment, 103 patients reported receiving Anti-Androgen, 162 reported receiving Docetaxel, 166 reported receiving Abiraterone, and 7 reported receiving Other. Chi-square test results indicated that treatment type frequencies were not signicantly different across treatment phase, χ²(3, N = 209) = 1.71, p = 0.63.

**Q4:** A one-way between subjects ANOVA was conducted to compare the effect of treatment on the duration of first-line treatment in Anti-androgen (mean = 30.59, se = 20.65), Docetaxel (mean = 28.45, se = 21.31), Abiraterone (mean = 31, se = 20.7), and Other (mean 31, se = 20.7) treatment conditions. There was not a significant effect of treatment on the duration of first-line treatment for the four conditions [F(3, 499) = 0.476, p = 0.699]. Tukey's Post-Hoc Test further supported that the treatment groups' length of time were not statistically significant.

**Q5:** A two-samples t-test was conducted to compare docetaxel use in first-line treatment duration in under 75 and over/equal to 75 age groups. There was a not a significant difference in the scores for under 75 (M=27.82, SD=21.33) and equal to and over 75 (M=29.29, SD=21.44) age groups; t(129.04)=-0.41, p = 0.68.

**Q6:** Linear Regression and Random Forest Regression techniques were applied to predict the number of hospitalizations based on age, symptomatic status, years diagnosed, and first-line treatment. Overall, both models reported great overall fit, with Linear Regression at 84% variance explained and was found to be statistically significant [F(57,151)=20.14, p<.001], and Random Forest at 82% variance explained (Note that the results are for the entire dataset, not from the training dataset). The data were split evenly into training and testing data for validation purposes for both statistical techniques. Results for both models were very similar, indicating consistent predictability capabilities and with low errors (shown in the boxplots). Further, the coefficients table for linear regression and variable importance from random Forest indicate the importance of age and symptomatic status on predicting the number of hospitalizations.

**Q7:** Linear Regression and Random Forest Regression techniques were applied to predict the number of hospitalizations based on age, symptomatic status, years diagnosed, and first-line treatment. Overall, both models reported great overall fit, with Linear Regression at 89% variance explained and was found to be statistically significant [F(57,151)=30.16, p<.001], and Random Forest at 81% variance explained (Note that the results are for the entire dataset, not from the training dataset). The data were split evenly into training and testing data for validation purposes for both statistical techniques. Results for both models were very similar, indicating consistent predictability capabilities and with low errors (shown in the boxplots). Further, the coefficients table for linear regression and variable importance from random Forest indicate the importance of age and symptomatic status on predicting the number of hospitalizations.

**Q8:** Logistic regression and support vector machine classification techniques were applied to predict the impact of age, symptomatic status, years diagnosed, and first-line treatment on the presence or absence of a bone metastatis. Generally, both models accurately classified the presence or absence of a bone metastatis (results from the testing/training split classifications were compared to the observed results).

**Q9:** Survival Analysis was conducted to predict patients’ overall survival-ability from the time of diagnosis to the time of death (if applicable). Censored date was set a May, 2016, for patients’ who were not reported as deceased at the time of the data collection.

Survival Analysis was conducted three times.

For the first model, Age, symptomatic status, ECOG score, number of metastases, and the average number of yearly hospitalizations were used as predictors, since they were found to be significant and/or were mentioned in previous questions. Results indicate that neither the full model nor expected predictors were found to be significant. Thus, hospitalizations and ECOG score were removed because they have lowest impact on the full model. The subset model then found a stronger overall model fit with Age and symptomatic status having a significant impact as predictors on overall survival duration. However, a caveat between the full model and subset model is that the differences weren’t too different from each other – also indicated in the survival graphs.

As such, random forest and lasso were used for variable selection procedures in the variable selection process in selecting candidate predictors – random Forest for variable importance selection and lasso for regularization purposes, both of which are different techniques to tackle variable selection.

Further, the survival object response variable was modified for parsimony purposes and now included durations between diagnosis date and first-line treatment, between first-line treatment and second-line treatment, and if possible second-line treatment and censor/deceased date (whichever was later or was available), instead of the time difference between initial diagnosis and censor/deceased date. Further, the dates for diagnosis, first-line and second-line treatment dates were considered as candidate predictors.