Paper draft: Predicting Erectile Function in Patients Undergoing Stereotactic Body Radiation Therapy

Dylan Sun and Zijiang Yang

2017-01-26

Abstract

Stereotactic Body Radiation Therapy (SBRT) is a new method for treating prostate cancer. This study looks at the potential for SBRT patients to experience erectile dysfunction, an important side-effect that is associated with the treatment. Using a logistic regression model, this study finds that baseline variables including age, score on a quality of life survey, erectile function at baseline, and prostate-specific antigen levels can be used to predict the probability of having erectile function two years after recieving SBRT. The model performs well, with XX area under curve and XX% predictive probability. Following the results from the model, we conclude that around XX% of patients who undergo SBRT are likely to lose erectile function.

Introduction

Stereotactic Body Radiation Therapy (SBRT) is a new treatment option available for prostate cancer, the most common form of male cancer. The treatment is characterized by relatively few instances of high dose radiation over a short period of time, as opposed to previous treatment methods with small doses of radiation over a longer period of time. Given that SBRT has around the same tumor kill effectiveness as the previous treatment, and that prostate cancer is quite common and very treatable, the largest concern associated with treatment selection is the lifestyle impact of side-effects. In particular, the most relevant side-effect associated with SBRT is the potential for erectile dysfunction.

In this study, data collected from SBRT patients is used to fit a logistic regression model exploring the likelihood of losing erectile function following treatment with SBRT. The model shows that variables such as age, score on a health survey, and prostate-specific antigen levels can be used to predict the probability of erectile function two years post-treatment with reasonably high accuracy.

Methods

Our analysis consists of two parts. In the first part, we carry out a series of exploratory data analysis in order to understand the data in general so that we can choose an appropriate statistical model to build, and in the second part we build a logistic regression model to predict the probability that one will have erectile function at two years after SBRT.

Exploratory Data analysis

To get a big picture of how the data look like, we first generated a table of descriptive statistics for each variable, including the mean, standard deviation, median, minimum, maximum, etc. For categorical variables(Gleason Score, T-Stage Group, and ADT), we also calculated the frequency of each category. With respect to the "Erectile Function at Baseline", and "Erectile Function at 2 Years", we calculated the frequency of changing from 0 to 0, 0 to 1, 1 to 0 and 1 to 1 respectively. We could use this frequency table to estimate the rate of erectile function following SBRT.

We then plotted a histogram for each of the continuous variable, in order to see if there are any outliers or something strange with the distribution. If we found outliers in a histogram, we would first look into those outliers to determine if those values are theoretically possible. If they were clearly typos we could safely remove them. If they were theoretically possible, then we needed to determine the goodness of fit of the models with and with out including those outliers to determine whether we should keep those subjects.

Collinearity was another potential issue, which would not reduce the predictive power of the model overall, but could reduce the preciseness of the calculations of individual predictors. Hence we calculated the correlation between each of the potential predictors and visualized the correlation matrix using the corrplot function in R, so that we could directly find any pair of predictors that have high correlations.

Logistic Regression

Since our outcome variable "Erectile Function at 2 Years" was a binary variable, it was reasonable for us to use a logistic regression model, because it met one of our objectives, which was to estimate the probability of having function at 2 years, and the interpretation of the effects of predictors in the logistic regression model was straightforward. We also considered using other classification methods such as random forest classifier and supported vector classifier, but since it would be hard to interpert the effects of the predictors in those models, we decided to focus on logistic regression.

We evaluated the performance of our model based on its predicting accuracy rate, goodness of fit, and the area under receiver operating characteristic curve (the ROC curve). We calculated the predicting accuracy rate as the proportion of subjects that we predicted correctly among all 325 subjects. We tested the goodness of fit by using the Hosmer-Lemeshow test, which was a commonly used statistical test for goodness of fit for logistic regression models. We created the ROC curve by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, which could compare the performance of different binary classifiers. We used a threshold of 0.5, so that if the predicted probability was greater or equal to 0.5, then we predicted that the subject would have erectile function at 2 years. (Do we need to try different threshold and compare their AUC?)

Results

Discussion

Final Conclusions

Appendix