Top of Form

|  |  |  |
| --- | --- | --- |
| **Client:** | Patrick Kiel | **File:** 25-027 |
| **Dept:** | Pharmacy | **Faculty:**  **Student:** |
| **Date:** | 3/4/25 | **Initial Meeting:**  **Follow-up:** |
| **Consultant and Attendees:** Patrick Kiel, Sumeeth Guda, Dr. Michael Preston, Dr. Chong Gu | | |
| **Statement of Problem:** To see if drug targeted therapy improves time-to-treatment discontinuation (TTD) and overall survival (OS) in patients with advanced NSCLC (Non-Small Cell Lung Cancer) compared to carboplatin-based chemotherapy. | | |
| **Goal of this Project:** PhD Dissertation / Journal Publication | | |
| **Background:**  The client is a PhD Student in the pharmacy department who needs help with the analysis of their data for a journal publication with respect to chemotherapy. They have 2 goals which they are investigating. The first would be to see if drug targeted therapy improves time-to-treatment discontinuation (TTD) and overall survival (OS) in patients with advanced NSCLC (Non-Small Cell Lung Cancer) compared to carboplatin-based chemotherapy. Additionally, their second goal is to see if Patients with higher socioeconomic status (e.g., college education, higher income, or employment) have more access to biomarker testing and targeted therapies compared to those receiving conventional carboplatin-based chemotherapy. All of the data has been collected and are present within an NIH database which the client isn’t able to share with the SCS. But overall, they needed help with their analysis methods, since they have missing data, and they need help analyzing the data. Additionally, they want to do model selection and want to know the best approach of either backward, forward, or stepwise regression are variable selection methods used for multiple linear regression. | | |
| **Progress of project at this time:** Analysis (All data collected) | | |
| **Relevant information presented at meeting:**  In the meeting, Patrick presented his research on real-world evidence generation using the NIH dataset, focusing on drug-induced chemotherapy treatments for Stage 4 lung cancer patients. His first research question examines the time to treatment discontinuation and overall survival associated with these treatments, while the second question investigates how demographic factors influence treatment plans, particularly regarding whether patients have access to genome-based cancer treatments versus standard chemotherapy.  Dr. Gu inquired about the decision-making process in treatment plans, asking whether the patient or physician has more influence on the decision. Patrick clarified that while decisions are ideally mutual, they are primarily physician-driven, with physicians following standardized treatment guidelines, though training differences could influence the approach. Patrick is working with a dataset of 340 patients, analyzing how various social determinants (six variables based on domain knowledge) affect the choice of treatment. For this survey question, most of the responses for the categories were binary and hence Patrick used a logistic regression model for his analysis. Dr. Gu suggested using an unsupervised learning approach such as clustering rather than logistic regression for the second question, as it might provide better insights.  With respect to analyzing the continuous variables to compare the treatment groups, Dr. Gu recommended starting with simple visualizations, such as boxplots for continuous variables and two-way tables for categorical variables, to identify any clear differences between the treatment groups. Dr. Gu emphasized that exploratory data analysis (EDA) and visualization can reveal important patterns, even though these patterns may be difficult to detect in two-dimensional plots. If clear separation between treatment groups is found, classification techniques might be more appropriate than logistic regression, as logistic regression can struggle to detect such distinctions in the presence of social determinant factors.  Regarding the modeling of the data, Patrick had not yet completed model selection or checked for multicollinearity. Dr. Gu recommended conducting stepwise regression to identify any significant interaction terms, emphasizing the importance of incorporating domain knowledge to guide the model selection process rather than relying solely on machine-driven methods. He also suggested using AIC to assess the models and check for collinearity among categorical predictors.  There was some concern regarding whether the Cox model would be appropriate for the project. For the first question Dr. Gu highlighting that the Cox model, which tracks failure rates and survival times, might not be the best fit due to significant data issues such as heavy censoring, missing death dates, and the lack of baseline performance data. However, Dr. Gu noted that the Cox model could still be useful for assessing instantaneous failure rates, as many patients are still alive and under surveillance. Given that 20-25% of the data is missing, and with the lack of clear death dates, Dr. Gu recommended considering a multinomial approach to analyze the data, focusing on different treatment categories and their dependence on covariates. | | |
| **Recommendations for Design and/or Analysis:**   * As shown in the meeting, the overall survival model isn't feasible due to the censoring of the current data. If he has more complete data, use the time to failure and ordinal regression for the overall response without censoring. It was recommended to look at the lifetime data and regress on the independent variables. The overall feel is like a count response compared to a continuous response. When he converts to count, the time periods won't be uniform in terms of the length. * With respect to missing data, a lot of the treatment data is missing, but what is more concerning is some of the data from the baseline is missing / censored. Missing data from the baseline needs to be redacted because any inference from the estimated baseline is essentially useless. If imputation is used, it could impose bias and make insignificant variables appear significant. | | |
| **Who will carry out these actions?**  Client:   * Change the survival model to become more of a count model. * Clean the dataset further. * Let the consultant know if there are any issues regarding the interpretation of the results or if help is needed with the data processing.   Consultant:   * Help the client with the clarification and check in with him regularly when he needs help. | | |
| **Status:** Follow up not needed. | | |

Bottom of Form