**Client 25-027 IM Agenda**

**3/4/25 9:30 AM**

**Goal: PhD Dissertation / Journal Publication**

**Phase: Analysis (All Data Collected)**

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.
* They tested both of these questions through conducting a study.

Design & Variable Notes:

* Their retrospective cohort study collects data from the NIH, encompassing biomarker testing, clinical outcomes of TTD and OS, and SDoH factors from Jan 2017 to June 2022.
* The study focuses on patients over 18 years old who were diagnosed with advanced non-small cell lung cancer (NSCLC) and received first-line treatment. Patients were either treated with a standard chemotherapy regimen (with or without immunotherapy) or targeted therapy based on NCCN guidelines.
* Researchers identified NSCLC cases using specific medical codes and checked for accuracy. Patients were excluded if they had more than one type of cancer, received treatment before their official diagnosis, or had a diagnosis/treatment unrelated to the main NSCLC types (squamous or adenocarcinoma).

The 2 variables pertaining to their first research question are TTD and OS:

* TTD was measured as the number of days from initiation of first line therapy to the last documented date of administration. Censoring occurred for patients who remained on treatment within 120 days of July 1, 2022.
* OS was measured as the number of days from initiation of first line therapy to the documented date of death. Patients without a death event were censored at the end of the database or last observation.

The variables related to their second research question were:

1. Race: (White, non-white),
2. Ethnicity: Hispanic, non-hispanic
3. Smoking status: (Smoker / Non-smoker)
4. Disability: At least one of the following: Deaf, blind, difficulty with errands alone, difficulty concentrating, difficulty dressing or bathing, difficulty walking or climbing stairs.
5. Education level:(College educated, not college educated)
6. Employment status: Employed, unemployed, or retired.
7. Sex: (What was documented at birth).
8. Marital: (Married / Not married)
9. Neighborhood safety: (Safe / not safe)
10. Insurance type,
11. Income: (> 75K, < 75K).

All of the above variables were self-reported by the patients.

Analysis Methods:

* The base analysis techniques they used so far with the data include:
  + Student t-test and chi-square test, to compare baseline characteristics, excluding missing data. They defined statistical significance with a 2-sided alpha level of 0.05 or a 95% confidence interval.
  + Treatment outcomes were analyzed using Kaplan-Meier analysis, and differences between groups were tested with log-rank tests.
    - A Cox regression model estimated hazard ratios while accounting for factors like age, race, sex, smoking status, and biomarker testing.
* Social factors were evaluated using logistic regression to assess their impact on biomarker testing and targeted therapy use.

Statistical Issues

1. They have missing data within the dataset, and they need help analyzing the data. (Possibly through deletion or imputation).
2. 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.

Questions to ask client:

1. What does the dataset look like?
2. Considering they use a survival analysis approach for their second question, why do they need to use model selection in this case?
   1. What is the distribution they use for the cox proportional analysis (Weibull or Exponential).
3. How many patients are in their study?