Statistical Analysis Plan

version 1

Hyejung Lee [hyejung.lee@utah.edu](mailto:hyejung.lee@utah.edu)

Thu Mar 20, 2025 02:08:07 PM

Table of Contents

# 1 Hypothesis

Advanced non-small cell lung cancer (aNSCLC) patients’ recommended course of treatments are identified in National Comprehensive Cancer Network (NCCN) guideline. The guideline suggests each patient to perform biopsy to identify presence of targetable mutations of certain biomarkers and PDL1 expression to make a informed decision on 1L therapy choice. This recommendation is made because targeted therapy is better for patient’s survival than a standard, genetic therapy such as chemotherapy or immunotherapy. Usually, at least 1 positive mutation or 2 negative mutation status among all targetable biomarkers should be obtained to make informed decision about targetable therapy. Thus having at least 1 positive mutation or 2 negative mutation status will be referred to as *useful* mutation test result from here onward. Similarly, obtaining PDL1 expression level (from 0-100%) level, as opposed to result such as missing expression level or indeterminate result, will be referred to as *useful* PDL1 test result.

The standard NCCN guideline, however, may not always be followed due to reasons such as clinician’s expert knowledge and ethics. For example, heavy smokers who develop lung cancer is most likely due to smoking rather than genetic mutation. Patients who fall into this category then may receive chemotherapy right away, instead of performing and waiting for genetic or PDL1 test results. One reason for not waiting until the useful test comes out is that a complete panel of test results (combined mutation and PDL1 expression tests) usually takes about 2-3 weeks to become available. It may take longer if the biopsy samples were not good and the test results are indeterminate, requiring another round of biopsy and waiting for the result. This period of waiting without doing any treatment may be detrimental to patient’s health and thus becomes an ethical issue. Therefore, the ultimate choice of whether waiting for the test result or not relies heavily on the clinician’s decision.

Just like the example of heavy smokers, there are some patient characteristics that are highly indicative of patient’s likelihood of targetable mutation status. However, there has not yet been a quantitative evaluation of effect of proceeding to 1L therapy prior to knowing targetable mutation status in patients. In this study, we evaluate causal effect of proceeding right 1L therapy in all aNSCLC patients, as well as in some subgroups that have previously been identified as associated with the mutation status.

# 2 Objectives

**Primary Objective**: Using marginal structural model, we will develop and evaluate the effect of proceeding to 1L therapy prior to obtaining useful targetable mutation status

* Limit the number of weeks.
* For entire observation.

**Secondary Objective**: Test impact of proceeding to 1L before useful tests become available in the following subgroups: male, female, have history of smoking, do not have history of smoking, baseline Eastern Cooperative Oncology Group (ECOG) score 0-2, baseline ECOG score 3-4, baseline albumin < 35g/L, baseline albumin >= 35 g/L, and Asian.

* These are the subgroups identified by Wally as having either higher or lower probability of having targetable mutation, and thus would more or less likely wait for the useful test results to become available prior to initiating 1L therapy.

# 3 Study Design and Population

## 3.1 Study Design

Retrospective Cohort Study

## 3.2 Study Population

**Data**: A nationwide Flatiron Health electronic health record (EHR)-derived de-identified database. The Flatiron Health database is a longitudinal database, comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction1,2. The de-identified data originated from 280 cancer clinics (~800 sites of care). Patients with a BirthYear of 1937 or earlier may have an adjusted BirthYear in Flatiron datasets due to patient de-identification requirements. For more information, please refer to their webpage3.

**Inclusion criteria**: People who got diagnosed with aNSLC (ICD-9 162.x or ICD-10 C34x or C39.9) from 01 January 2011, to 30 December 2022 from Flatiron Health network.

1. Index date is date of aNSCLC diagnosis.

**Exclusion criteria**:

1. Initiate who receive first-line (1L) therapy on or after the index date.
2. Receive any useful test result (either PDL1 or targetable mutation) prior to the index date
3. Missing death or censoring date

Patient attrition diagram is shown below. It shows all exclusion criteria. We excluded all patients who don’t have valid survival end time and who start either the 1L therapy or valid test before advanced diagnosis date. We will explain what valid test means in the next section.

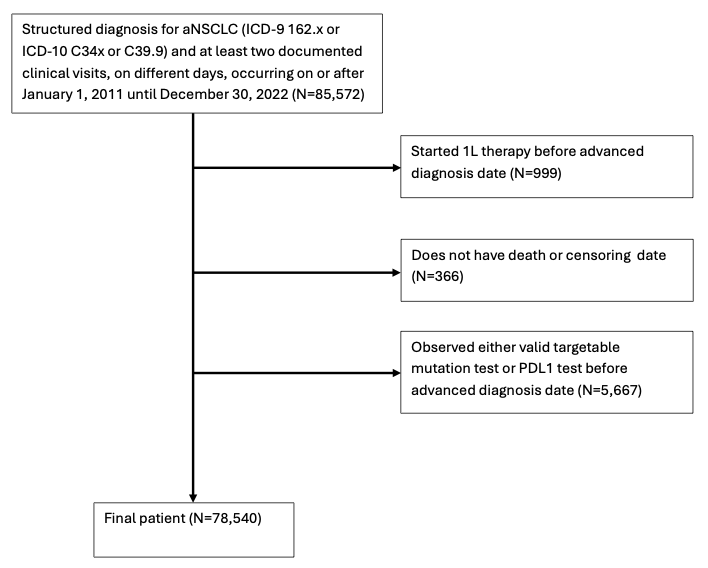


Figure XX: Patient attrition diagram

# 4 Data Strucutre

There are 9 data sets of wide and long formats. Data sets containing repeated measurements such as visits, PDL1 or targetable mutation test, or lab values will be long format. The unit of observation differs by the dataset.

1. Create a single date variable where useful PDL1 test results are observed in each patient.
2. Create a single date variable where useful targetable mutation test results are observed in each patient.
3. For each consecutive week after the index date, create a binary indicator variable indicating who initiated 1L therapy before receiving any useful test is out. (Cohort identification)
4. Create baseline covariates as listed in @ref(sec-appA)
5. report the percent missing of each variable overall by exposure group as in @ref(tab:baseline-missing-mock)
6. For each consecutive week after the index date, a long dataset is created for each patient’s weekly time-varying follow-up variables listed in Appendix B. There will be a variable named which shows number of weeks since index date. When , it refers to baseline. Suppose for an example, albumin measurement at represent measurements taken between days, and at will represent measurements taken between days.
7. Impute censoring date using visit and oral medication record

| PatientID | bmi | BirthYear | Gender | Race | Ethnicity | State | DiagnosisDate | AdvancedDiagnosisDate | Histology | GroupStage | SmokingStatus | DateOfDeath | DeathInd | RaceEthnicity | nLines | FirstLineTherapy | FirstLineStartDate | FirstLineEndDate | ALK | ALK\_MT\_SpecimenCollectedDate | ALK\_MT\_SpecimenReceivedDate | ALK\_MT\_ResultDate | EGFR | EGFR\_MT\_SpecimenCollectedDate | EGFR\_MT\_SpecimenReceivedDate | EGFR\_MT\_ResultDate | BRAF | BRAF\_MT\_SpecimenCollectedDate | BRAF\_MT\_SpecimenReceivedDate | BRAF\_MT\_ResultDate | KRAS | KRAS\_MT\_SpecimenCollectedDate | KRAS\_MT\_SpecimenReceivedDate | KRAS\_MT\_ResultDate | MET | MET\_MT\_SpecimenCollectedDate | MET\_MT\_SpecimenReceivedDate | MET\_MT\_ResultDate | RET | RET\_MT\_SpecimenCollectedDate | RET\_MT\_SpecimenReceivedDate | RET\_MT\_ResultDate | PDL1 | PDL1\_MT\_SpecimenCollectedDate | PDL1\_MT\_SpecimenReceivedDate | PDL1\_MT\_ResultDate | ROS1 | ROS1\_MT\_SpecimenCollectedDate | ROS1\_MT\_SpecimenReceivedDate | ROS1\_MT\_ResultDate | NTRK | NTRK\_MT\_SpecimenCollectedDate | NTRK\_MT\_SpecimenReceivedDate | NTRK\_MT\_ResultDate | PercentStaining | PercentStaining\_Date | ALK\_SpecimenCollectedDate | ALK\_SpecimenReceivedDate | ALK\_ResultDate | EGFR\_SpecimenCollectedDate | EGFR\_SpecimenReceivedDate | EGFR\_ResultDate | BRAF\_SpecimenCollectedDate | BRAF\_SpecimenReceivedDate | BRAF\_ResultDate | KRAS\_SpecimenCollectedDate | KRAS\_SpecimenReceivedDate | KRAS\_ResultDate | MET\_SpecimenCollectedDate | MET\_SpecimenReceivedDate | MET\_ResultDate | RET\_SpecimenCollectedDate | RET\_SpecimenReceivedDate | RET\_ResultDate | PDL1\_SpecimenCollectedDate | PDL1\_SpecimenReceivedDate | PDL1\_ResultDate | ROS1\_SpecimenCollectedDate | ROS1\_SpecimenReceivedDate | ROS1\_ResultDate | NTRK\_SpecimenCollectedDate | NTRK\_SpecimenReceivedDate | NTRK\_ResultDate | num\_tested | num\_valid\_result\_date | num\_result\_after\_firstline | result\_before\_or\_on\_firstline | missing\_test | earliest\_test\_names | earliest\_test\_date | min\_earliestTest\_1L\_baseline | earliest\_biop\_names | biopsy\_baseline | Albumin | Albumin\_Unit | Albumin\_nVal | Alkaline | Alkaline\_Unit | Alkaline\_nVal | ALT | ALT\_Unit | ALT\_nVal | Bilirubin | Bilirubin\_Unit | Bilirubin\_nVal | Calcium | Calcium\_Unit | Calcium\_nVal | Carbon dioxide | Carbon dioxide\_Unit | Carbon dioxide\_nVal | Chloride | Chloride\_Unit | Chloride\_nVal | Creatinine | Creatinine\_Unit | Creatinine\_nVal | eGFR | eGFR\_Unit | eGFR\_nVal | Potassium | Potassium\_Unit | Potassium\_nVal | Protein | Protein\_Unit | Protein\_nVal | Sodium | Sodium\_Unit | Sodium\_nVal | HCT | HCT\_Unit | HCT\_nVal | HGB | HGB\_Unit | HGB\_nVal | Lymphocyte\_N | Lymphocyte #\_Unit | Lymphocyte #\_nVal | Neutrophil\_N | Neutrophil #\_Unit | Neutrophil #\_nVal | Platelet | Platelet\_Unit | Platelet\_nVal | RBC | RBC\_Unit | RBC\_nVal | WBC | WBC\_Unit | WBC\_nVal | Albumin\_factor | Alkaline\_factor | ALT\_factor | Bilirubin\_factor | Calcium\_factor | Carbon dioxide\_factor | Chloride\_factor | Creatinine\_factor | eGFR\_factor | Potassium\_factor | Protein\_factor | Sodium\_factor | HCT\_factor | HGB\_factor | Lymphocyte\_N\_factor | Neutrophil\_N\_factor | Platelet\_factor | RBC\_factor | WBC\_factor | ECOG | ECOG\_nVal | age\_at\_advDiag | AdvancedDiagnosisDate\_numeric | time | time\_mth | cohort |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | 19.92212 | 1951 | M | White | Not Hispanic or Latino | KY | 2002-02-01 | 2019-08-23 | Squamous cell carcinoma | Stage IIA | History of smoking | 2020-02-27 | 0 | WhiteNonHisp | 1 | Carboplatin,Paclitaxel,Pembrolizumab | 2019-09-24 | 2020-01-01 | WT | NA | NA | NA | WT | NA | NA | NA | WT | NA | NA | NA | WT | NA | NA | NA | WT | NA | NA | NA | WT | NA | NA | NA | WT | NA | NA | NA | WT | NA | NA | NA | WT | NA | NA | NA | 0% | 2019-09-09 | 2019-08-29 | 2019-08-30 | 2019-09-11 | 2019-08-29 | 2019-08-30 | 2019-09-11 | 2019-08-29 | 2019-08-30 | 2019-09-11 | 2019-08-29 | 2019-08-30 | 2019-09-11 | 2019-08-29 | 2019-08-30 | 2019-09-11 | 2019-08-29 | 2019-08-30 | 2019-09-11 | 2019-08-23 | NA | 2019-09-09 | 2019-08-29 | 2019-08-30 | 2019-09-11 | 2019-08-29 | 2019-08-30 | 2019-09-11 | 9 | 9 | 0 | 9 | FALSE | PDL1 | 2019-09-09 | 2019-09-09 | PDL1 | 2019-08-23 | 36 | g/L | 1 | 142 | U/L | 1 | 13 | U/L | 1 | 0.4 | mg/dL | 1 | 8.9 | mg/dL | 1 | 30 | mmol/L | 1 | 105 | mmol/L | 1 | 0.81 | mg/dL | 1 | NA | NA | NA | 4.7 | mmol/L | 1 | 70 | g/L | 1 | 140 | mmol/L | 1 | 36.9 | % | 1 | 12.5 | g/dL | 1 | 1.40 | 10\*9/L | 1 | 3.77 | 10\*9/L | 1 | 211 | 10\*9/L | 1 | 4.03 | 10\*12/L | 1 | 5.9 | 10\*9/L | 1 | Average | Above | Average | Average | Average | Above | Average | Average | NA | Average | Average | Average | Average | Average | Average | Average | Average | Average | Average | 0 | 1 | 68 | 0.793438 | 188 | 6.176591 | tested |
| B | 26.69459 | 1950 | M | White |  | OK | 2018-01-08 | 2020-07-29 | Non-squamous cell carcinoma | Stage IIIA | History of smoking | 2021-05-15 | 1 | Other | 3 | Pembrolizumab | 2020-08-05 | 2020-09-01 | WT | NA | NA | NA | WT | NA | NA | NA | WT | NA | NA | NA | MT | 2020-07-29 | 2020-08-05 | 2020-08-18 | WT | NA | NA | NA | WT | NA | NA | NA | MT | 2020-07-29 | 2020-08-05 | 2020-08-11 | WT | NA | NA | NA | WT | NA | NA | NA | 90% - 99% | 2018-03-02 | 2020-07-29 | 2020-08-05 | 2020-08-18 | 2020-07-29 | 2020-08-05 | 2020-08-18 | 2020-07-29 | 2020-08-05 | 2020-08-18 | 2020-07-29 | 2020-08-05 | 2020-08-18 | 2020-07-29 | 2020-08-05 | 2020-08-18 | 2020-07-29 | 2020-08-05 | 2020-08-18 | 2020-07-29 | 2020-08-05 | 2020-08-11 | 2020-07-29 | 2020-08-05 | 2020-08-18 | 2020-07-29 | 2020-08-05 | 2020-08-18 | 9 | 9 | 9 | 0 | FALSE | PDL1 | 2020-08-11 | 2020-08-05 | PDL1 | 2020-07-29 | 41 | g/L | 1 | 116 | U/L | 1 | 21 | U/L | 1 | 1.0 | mg/dL | 1 | 9.1 | mg/dL | 1 | NA | NA | NA | NA | NA | NA | 1.20 | mg/dL | 1 | NA | NA | NA | NA | NA | NA | 72 | g/L | 1 | NA | NA | NA | 45.9 | % | 1 | 14.8 | g/dL | 1 | 1.27 | 10\*9/L | 1 | 3.70 | 10\*9/L & 10\*9/L | 2 | 267 | 10\*9/L | 1 | 4.48 | 10\*12/L | 1 | 6.4 | 10\*9/L | 1 | Average | Average | Average | Average | Average | NA | NA | Average | NA | NA | Average | NA | Average | Average | Average | Average | Average | Average | Average | 0 | 1 | 70 | 1.082807 | 290 | 9.527721 | tested |

| PatientID | k | Gender | SmokingStatus | Albumin | Platelet |
| --- | --- | --- | --- | --- | --- |
| A | 1 | M | History of smoking | 36 | 211 |
| A | 2 | M | History of smoking | 36 | 211 |
| A | 3 | M | History of smoking | 36 | 211 |
| B | 1 | M | History of smoking | 41 | 267 |
| B | 2 | M | History of smoking | 41 | 267 |
| B | 3 | M | History of smoking | 41 | 267 |
| B | 4 | M | History of smoking | 41 | 267 |
| B | 5 | M | History of smoking | 41 | 267 |
| B | 6 | M | History of smoking | 41 | 267 |

| PatientID | k | Gender | SmokingStatus | Albumin | Platelet |
| --- | --- | --- | --- | --- | --- |
| A | 0 | M | History of smoking | 36 | 211 |
| A | 1 | M | History of smoking | 36 | 211 |
| A | 2 | M | History of smoking | 36 | 211 |
| B | 0 | M | History of smoking | 41 | 267 |
| B | 1 | M | History of smoking | 41 | 267 |
| B | 2 | M | History of smoking | 41 | 267 |
| B | 3 | M | History of smoking | 41 | 267 |
| B | 4 | M | History of smoking | 41 | 267 |
| B | 5 | M | History of smoking | 41 | 267 |

| PatientID | k | Gender | SmokingStatus | Albumin | Platelet |
| --- | --- | --- | --- | --- | --- |
| A | 0 | M | History of smoking | 37.20139 | 211 |
| A | 1 | M | History of smoking | 17.87623 | 211 |
| A | 2 | M | History of smoking | 37.51583 | 211 |
| B | 0 | M | History of smoking | 29.80779 | 267 |
| B | 1 | M | History of smoking | 41.01908 | 267 |
| B | 2 | M | History of smoking | 52.88518 | 267 |
| B | 3 | M | History of smoking | 35.94656 | 267 |
| B | 4 | M | History of smoking | 40.00766 | 267 |
| B | 5 | M | History of smoking | 44.05353 | 267 |

| PatientID | k | Gender | SmokingStatus | Albumin | Platelet |
| --- | --- | --- | --- | --- | --- |
| A | 0 | M | History of smoking | 37.20139 | 214 |
| A | 1 | M | History of smoking | 17.87623 | 210 |
| A | 2 | M | History of smoking | 37.51583 | 198 |
| B | 0 | M | History of smoking | 29.80779 | 236 |
| B | 1 | M | History of smoking | 41.01908 | 262 |
| B | 2 | M | History of smoking | 52.88518 | 257 |
| B | 3 | M | History of smoking | 35.94656 | 262 |
| B | 4 | M | History of smoking | 40.00766 | 302 |
| B | 5 | M | History of smoking | 44.05353 | 283 |

| PatientID | k | Gender | SmokingStatus | Albumin | Platelet | cohort |
| --- | --- | --- | --- | --- | --- | --- |
| A | 0 | M | History of smoking | 37.20139 | 214 | 0 |
| A | 1 | M | History of smoking | 17.87623 | 210 | 0 |
| A | 2 | M | History of smoking | 37.51583 | 198 | 0 |
| B | 0 | M | History of smoking | 29.80779 | 236 | 0 |
| B | 1 | M | History of smoking | 41.01908 | 262 | 0 |
| B | 2 | M | History of smoking | 52.88518 | 257 | 0 |
| B | 3 | M | History of smoking | 35.94656 | 262 | 0 |
| B | 4 | M | History of smoking | 40.00766 | 302 | 0 |
| B | 5 | M | History of smoking | 44.05353 | 283 | 0 |

| PatientID | k | Gender | SmokingStatus | Albumin | Platelet | cohort |
| --- | --- | --- | --- | --- | --- | --- |
| A | 0 | M | History of smoking | 37.20139 | 214 | 0 |
| A | 1 | M | History of smoking | 17.87623 | 210 | 0 |
| A | 2 | M | History of smoking | 37.51583 | 198 | 0 |
| B | 0 | M | History of smoking | 29.80779 | 236 | 0 |
| B | 1 | M | History of smoking | 41.01908 | 262 | 0 |
| B | 2 | M | History of smoking | 52.88518 | 257 | 0 |
| B | 3 | M | History of smoking | 35.94656 | 262 | 0 |
| B | 4 | M | History of smoking | 40.00766 | 302 | 0 |
| B | 5 | M | History of smoking | 44.05353 | 283 | 0 |

| PatientID | k | Gender | SmokingStatus | Albumin | Platelet | cohort |
| --- | --- | --- | --- | --- | --- | --- |
| A | 0 | M | History of smoking | 37.20139 | 214 | 0 |
| A | 1 | M | History of smoking | 17.87623 | 210 | 0 |
| A | 2 | M | History of smoking | 37.51583 | 198 | 0 |
| B | 0 | M | History of smoking | 29.80779 | 236 | 0 |
| B | 1 | M | History of smoking | 41.01908 | 262 | 0 |
| B | 2 | M | History of smoking | 52.88518 | 257 | 0 |
| B | 3 | M | History of smoking | 35.94656 | 262 | 1 |
| B | 4 | M | History of smoking | 40.00766 | 302 | 1 |
| B | 5 | M | History of smoking | 44.05353 | 283 | 1 |

Example dataset: An expected sample of the dataset for the baseline and follow-up variables for two patients. Gender and smoking status are baseline covariates that do not change over time. Albumin and Platelet are the tiem-dependent covariates that are measured each week (k). Cohort is the exposure variables that can change over time.

| Patient.ID | k | Gender | SmokingStatus | Albumin | Platelet | cohort |
| --- | --- | --- | --- | --- | --- | --- |
| A | 0 | M | History of smoking | 37.20139 | 214 | 0 |
| A | 1 | M | History of smoking | 17.87623 | 210 | 0 |
| A | 2 | M | History of smoking | 37.51583 | 198 | 0 |
| B | 0 | M | History of smoking | 29.80779 | 236 | 0 |
| B | 1 | M | History of smoking | 41.01908 | 262 | 0 |
| B | 2 | M | History of smoking | 52.88518 | 257 | 0 |
| B | 3 | M | History of smoking | 35.94656 | 262 | 1 |
| B | 4 | M | History of smoking | 40.00766 | 302 | 1 |
| B | 5 | M | History of smoking | 44.05353 | 283 | 1 |

1. Ma, X., Long, L., Moon, S., Adamson, B. J. & Baxi, S. S. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron health, SEER, and NPCR. *Medrxiv* 2020–03 (2020).

2. Birnbaum, B. *et al.* Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. *arXiv preprint arXiv:2001.09765* (2020).

3. Zhang, S. BirthYear overview and data considerations. <https://flatironlifesciences.zendesk.com/hc/en-us/articles/11186051513229-BirthYear-Overview-and-Data-Considerations>.