**Smoking features for PLCO/LLP Calculation**

This is derived from the code D.Baldwin provided.

The code, written in Stata, is tied to specific raw data files and cannot be used directly. Its primary function is to process the raw data and consolidate signals from multiple sources into a single location.

Features calculation:

* **Smoking Intensity**: Use the most recent data available before the index date.
* **Smoking Status**: Use the most recent status, but if "smoking intensity" information is available, use it to determine smoking status. If no smoking information is present, assume the patient is a non-smoker.
* **Smoking Duration**: Simplified into a single range representing the smoking period. If the patient has multiple quit dates, they are disregarded.
  1. **Start Date/Age**: Use the earliest smoking start date (from column data5). If no data is available, assume the patient began smoking at age 18.
  2. **Cessation Date/Age**: Use the latest cessation date (from column data6). If no data is available, assume the cessation occurred 17 years prior to the index date. According to their analysis, the median gap between the index date and the latest cessation date (when available) is 19 years. I don’t know we they used 19
* **BMI** – BMI will be imputed by taking median value based on CPRD data

PLCO2012 – adjusted model without family history information (received from MT), very similar coeff to regular model with family history:

* Age: age center: 62, coef = 0.0741942
* Race: all white coef= 0
* Education: center 4 (Some college), coeff = -0.0949644
* BMI: center BMI: 27, coef = -0.0279899
* COPD in the past: coef = 0.3748469
* Other cancer int the past: coef = 0.4493221
* no family history in this model
* Current\_Smoker: coef = 0.2874487, EX smoker = 0
* Smoking Duration in years: center: 27 years. coef=0.0337509
* Time since quit in years: center\_smoking\_quit\_time = 10 years, coef = -0.0276701
* Smoking Intensity: x cigarettes per day. (exp(-x/10) - 0.4016839021) \* (-1.826223)
* b0: -4.453382

LLP V3:

<https://pubmed.ncbi.nlm.nih.gov/33082166/>

Table to calculate “alpha\_xy” by age and sex. We will refer this table as “alpha” as a function that accepts age and sex

|  |  |  |
| --- | --- | --- |
| **Age** | **Sex** | **alpha** |
| 40 | Male | -9.84 |
| 45 | Male | -8.94 |
| 50 | Male | -8.09 |
| 55 | Male | -7.41 |
| 60 | Male | -6.75 |
| 65 | Male | -6.34 |
| 70 | Male | -6.09 |
| 75 | Male | -5.61 |
| 80 | Male | -5.46 |
| 40 | Female | -10.37 |
| 44 | Female | -8.53 |
| 50 | Female | -7.93 |
| 55 | Female | -6.97 |
| 60 | Female | -6.69 |
| 65 | Female | -6.46 |
| 70 | Female | -5.96 |
| 75 | Female | -5.7 |
| 80 | Female | -5.89 |

alpha\_xy := (alpha(age//5,sex) \* (5 - (age % 5)-0.5) + alpha(age//5+5,sex) \*((age % 5)+0.5)) / 5

* Cancer history coef = 0.6754
* Pneumonia coef = 0.6025
* Family history, before age 60, after age 60 – we don’t have. So all are considered without family history
* Exposer to asbestos more than 1 year working life - doesn’t exists in data – all patients are treated as not exposed
* Smoking duration in years coef:

if smoking\_years < 1:

return 0

if smoking\_years >= 1 and smoking\_years < 20:

return 0.7692

if smoking\_years >= 20 and smoking\_years < 40:

return 1.4516

if smoking\_years >= 40 and smoking\_years < 60:

return 2.5072

if smoking\_years >= 60:

return 2.72434

LLP score: = 1 / (1+exp( -( alpha\_xy + sum(betas)) ))