Smoking features for PLCO Calculation

This is based on Baldwin code on CPRD. Most of the code is processing of the “raw” data of CPRD into a “signal” filtered before index date(prediction time) and handling conflict, multiple sources of data from more than 1 table. For example, intensity is given in more than 1 table.

What we left is:

* **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 in CPRD data

Additionally, there is commented-out code suggesting a more complex approach to cessation age. For example, assuming a patient stopped smoking at age 55, or adjusting to ages 45 or 35 if the index date is before age 55. However, this logic does not appear to be implemented.

Anyway, I think they have a bug, and they took starting age as 0 and not 18 for starting date– this might impact calibration and not discrimination performance.