NOTE: !.Codes are cleaned when processing perday data. So “Comb\_PerDay\_data.csv” data codes are cleaned

“filtered\_inValidMonth\_comb\_perday\_df.csv” holds final analysis ID = 27607 less than Final\_analysis\_ID.csv before valid month

**Steps:**

1. Split large claim data
2. Separate process Medicare and Medicaid Claim data, Generate patient unique code per day data and clean codes
3. Code:
   1. 2A\_PerDayData\_CleanCodes\_Medicare.R
   2. 2B\_PerDayData\_CleanCodes\_Medicaid.R
   3. 2C\_Combine\_PerDay\_BothData.R
4. Output: Patient per day table
   1. All\_PerDay\_Data\_Medicare.csv
   2. All\_PerDay\_Data\_Medicaid.csv
   3. Comb\_PerDay\_data.csv
5. Example output:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| study\_id | claims\_date | Diag\_Codes | Proc\_Codes | Drug\_Codes |
| 1 | 1/14/00 | NA | NA | 281604$$$$50111036801 |
| 1 | 1/18/00 | 72690 | 99213$$$$73020 | 28080408$$$$283228$$$$00025152531$$$$00173045900 |

1. Get unique clean codes from per day data
   1. Code: 3A\_UniqueCodes\_FromPerDayData.R
   2. #'@NOTE1: because some of the drug codes are prepend 0s, but they are actually should not be prepended #'When processing the drug data in perday data, remember to do this also,

#TODO: It is better to regenerate this in perday data from the beginning, do not use the same process to procede drug code as diag, or directly map drug code to names.

* 1. Output: unique codes
     + 1. All\_unique\_Diag\_codes\_Cleaned.csv
       2. All\_unique\_Proc\_codes\_Cleaned.csv
       3. All\_unique\_Drug\_codes\_Cleaned.csv
  2. Example output:

|  |
| --- |
| unique\_Diag\_codes |
| 4589 |
| 585 |
| 25001 |
| V451 |
| 40391 |
| 25000 |

1. Group unique clean codes
   1. Code: 3B\_Grouping\_UniqueCodesInClaims.R
   2. Output: codes with group category
      * 1. Grouped\_Diag\_codes.csv
        2. Grouped\_Proc\_codes.csv
        3. Grouped\_Drug\_codes.csv
   3. Example output:

Diagnose:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Code | Chubak\_type | Chubak\_catogory | Ritzwoller\_catogory | CCS\_catogory |
| 4589 | Low blood pressure | Patient symptoms | NA | 117 |
| 585 | NA | NA | NA | 158 |

Procedure:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Code | Chubak\_type | Chubak\_catogory | Ritzwoller\_catogory | CCS\_catogory |
| 96410 | Chemotherapy | Systemic Therapy | Chemotherapy | NA |
| J1100 | NA | NA | Chemotherapy | NA |

Drug:

|  |  |  |  |
| --- | --- | --- | --- |
| unique\_Drug\_codes | drug\_name | specific\_group | general\_group |
| 401819 | NA | NA | NA |
| 52268020001 | PHOSLO | NA | NA |

1. Get outcome and outcome date (1st,2nd 3rd event, and death)
2. Code: 4A\_Get\_RecurrencePatientsDates
3. Output: 1st,2nd 3rd event date and site
4. updated\_All\_event\_df.csv
5. Example output:



1. Get valid claims month and Truncate valid claim month for the duration of predefined starting and ending point
   1. Start point: the diagnosis date of the first primary breast cancer + 6 months
   2. End point:
      1. without an SBCE: end of recorded claims
      2. with an SBCE, but no third event: end of recorded claims
      3. with an SBCE, the subsequent cancer is a non-breast primary cancer: 3 months before the registry-based diagnosis date or, subsequent is a breast cancer event (Recurrence or diagnose of breast cancer): 1 month before the first subsequent breast cancer event.
   3. Code: 4B\_Get\_ValidClaimsMonths.R
   4. Output: All\_Final\_Valid\_month.csv
2. Get analysis ID: the intersection of event data, comb per day data, and valid month
   1. Code: 5C\_GetFinal\_AnalysisID.R
3. Filtered per day data: the claims per day data must has a date >= min enrollment month, and max enrollment month
   1. Code: 5D\_Filtered\_PerDayData
   2. Output : filtered\_inValidMonth\_comb\_perday\_df.csv
4. Get patient per day data
   1. Code: 5C\_Get\_PerPatientData.R
   2. NOTE2: remember to convert drug codes to numeric when processing perday data due to NOTE1