**Project Development Specifications**

**Objective: create reusable codes of data preparation for predictive modeling by using TEVA Tardive Dyskinesia Predictive project**

**Data Source(s):**

* Patient mart data including:

1. [**V\_PMT\_PAT\_MTH**](http://gmr/NDW/NGPS_Detail_Object.asp?SystemID=NDW&ModelNm=NDW&ModelID=31&TypeCode=TABLE&TypeName=V%5FPMT%5FPAT%5FMTH&FilterTerm= &FilterObject= &FilterName= &FilterClass=N&ClassID=2)
2. [**V\_PMT\_PAT\_PROD\_MTH**](http://gmr/NDW/NGPS_Detail_Object.asp?SystemID=NDW&ModelNm=NDW&ModelID=31&TypeCode=TABLE&TypeName=V%5FPMT%5FPAT%5FPROD%5FMTH&FilterTerm= &FilterObject= &FilterName= &FilterClass=N&ClassID=2)
3. [**V\_PMT\_PAT\_DIAG\_MTH**](http://gmr/NDW/NGPS_Detail_Object.asp?SystemID=NDW&ModelNm=NDW&ModelID=31&TypeCode=TABLE&TypeName=V%5FPMT%5FPAT%5FDIAG%5FMTH&FilterTerm= &FilterObject= &FilterName= &FilterClass=N&ClassID=2)
4. [**V\_PMT\_PAT\_PRC\_MTH**](http://gmr/NDW/NGPS_Detail_Object.asp?SystemID=NDW&ModelNm=NDW&ModelID=31&TypeCode=TABLE&TypeName=V%5FPMT%5FPAT%5FPRC%5FMTH&FilterTerm= &FilterObject= &FilterName= &FilterClass=N&ClassID=2)
5. [**V\_PMT\_PAT\_PROVIDER\_MTH**](http://gmr/NDW/NGPS_Detail_Object.asp?SystemID=NDW&ModelNm=NDW&ModelID=31&TypeCode=TABLE&TypeName=V%5FPMT%5FPAT%5FPROVIDER%5FMTH&FilterTerm= &FilterObject= &FilterName= &FilterClass=N&ClassID=2)
6. [**V\_PMT\_PAT\_PROD\_PROV\_MTH**](http://gmr/NDW/NGPS_Detail_Object.asp?SystemID=NDW&ModelNm=NDW&ModelID=31&TypeCode=TABLE&TypeName=V%5FPMT%5FPAT%5FPROD%5FPROV%5FMTH&FilterTerm= &FilterObject= &FilterName= &FilterClass=N&ClassID=2)
7. [**V\_PMT\_PAT\_PROVIDER\_DIAG\_MTH**](http://gmr/NDW/NGPS_Detail_Object.asp?SystemID=NDW&ModelNm=NDW&ModelID=31&TypeCode=TABLE&TypeName=V%5FPMT%5FPAT%5FPROVIDER%5FDIAG%5FMTH&FilterTerm= &FilterObject= &FilterName= &FilterClass=N&ClassID=2)
8. [**V\_PMT\_PAT\_PROVIDER\_PRC\_MTH**](http://gmr/NDW/NGPS_Detail_Object.asp?SystemID=NDW&ModelNm=NDW&ModelID=31&TypeCode=TABLE&TypeName=V%5FPMT%5FPAT%5FPROVIDER%5FPRC%5FMTH&FilterTerm= &FilterObject= &FilterName= &FilterClass=N&ClassID=2)

Note: now institutional data is available in the patient mart DX data, we can set up the institutional data as an option to be included in the study.

When the study begin, need to confirm whether the study is required to include institutional data. If include, then all the data pull, and metrics calculations will include the institutional data

**Project Schema:**

The following project schema has been created on PAA:

Username: I1155657\_TEVA\_TD\_STD

Password: I1155657

**Project Folder:**

The following project folder has been created on the share drive:

S:\LRx\Best Practices\Programming\predictive modeling\Teva TD

**Market Definitions:**

* ***Clinical codes file:***

S:\LRx\Teva\1155657 Teva - TD Predictive Scoring Patient Analysis\DOI\ *Teva\_1155657\_TD Predictive Scoring\_Predictor Coding\_DOI\_HCPCS\_CPT\_DIAG CD\_NDW\_PRF008939\_20170321\_updated strength and UOM.xls*

* ***Predictors layout file***

S:\LRx\Teva\1155657 Teva - TD Predictive Scoring Patient Analysis\predictor layout\Teva *TD Flat File Layout\_v4 (3 20 2017).xlsx*

**Study time**

* **Study begin time**: 20100101
* **Study end time:** the most recent available month in patient mart Dx and RX data

**Specification**

1. Data extraction:

Based on the ***Clinical codes file***, pull all the data from patient mart tables where service date is from study begin to study end (201001 to most recent available month)

* V\_PMT\_PAT\_PROD\_MTH
  + –need to select records where Data\_typ\_cd=”RX”
* V\_PMT\_PAT\_DIAG\_MTH
  + –need to select records where Data\_typ\_cd=”DX” and Claim\_typ\_cd=(’P’, ‘I’)
* V\_PMT\_PAT\_PRC\_MTH
  + need to select records where Data\_typ\_cd=”DX” and Claim\_typ\_cd=(’P’, ‘I’)

1. Positive cohort identify

From the above extracted data, selection all patients who satisfy the following selection criteria:

1. patients who ever had at least one claim in DIAG CD of Interest in the market definition during study begin to study end
2. patients who ever had at least one claim in Procedure CD of Interest in the market definition during study begin to study end
3. patients who ever had at least one claim in NDC CD of Interest in the market definition during study begin to study end

---in TD example, patients meet the first criteria, i.e. had at least one TD diagnosis claim in DIAG CD of Interest in ***Clinical codes file*** during the study time will be the positive patients

Save the eligible patients as Positive\_cohort\_initial , else patients without selection criteria of interested would be Scoring \_cohort \_initial

1. For each patient in Positive\_cohort\_initial, create the below standard variables:

|  |  |
| --- | --- |
| Column Name | Column Definition |
| Patient\_ID |  |
| LRx\_FLAG | 1 if patient has at least 1 claim in LRx in the Study Time Period for ANY clinical code in any market; 0 otherwise |
| Dx\_FLAG | 1 if patient has at least 1 claim in Dx in the Study Time Period for ANY clinical code in any market; 0 otherwise |
| FRST\_RX\_CLM\_DT | Date of first LRx claim in the Study Time Period for ANY clinical code in any market |
| FRST\_DX\_CLM\_DT | Date of first Dx claim in the Study Time Period for ANY clinical code in any market |
| LOOKBACK\_DT | Maximum of FRST\_RX\_CLM\_DT and FRST\_DX\_CLM\_DT |
| LST\_RX\_CLM\_DT | Date of latest LRx claim in the Study Time Period for ANY clinical code in any market |
| LST\_DX\_CLM\_DT | Date of latest Dx claim in the Study Time Period for ANY clinical code in any market |
| LST\_INDEX\_DT | Maximum of LST\_RX\_CLM\_DT and LST\_DX\_CLM\_DT |
| [Selection Criteria #1]\_FLAG | 1 if patient has at least 1 claim for [Selection Criteria #1] in the Study Time Period; 0 otherwise |
| [Selection Criteria #1]\_Claim Count | Number of claims for [Selection Criteria #1] in the Study Time Period |
| [Selection Criteria #1]\_FRST\_EXP\_DT | Date of the first claim for [Selection Criteria #1] in the Study Time Period |
| [Selection Criteria #1]\_HCP\_SPEC | Physician Specialty of the HCP on the first claim for [Selection Criteria #1] in the Study Time Period |
| [Selection Criteria #n]\_FLAG | [Same as for Selection Criteria #1; Repeat for all Selection Criteria] |
| [Selection Criteria #n]\_Claim Count | [Same as for Selection Criteria #1; Repeat for all Selection Criteria] |
| [Selection Criteria #n]\_FRST\_EXP\_DT | [Same as for Selection Criteria #1; Repeat for all Selection Criteria] |
| [Selection Criteria #n]\_HCP\_SPEC | [Same as for Selection Criteria #1; Repeat for all Selection Criteria] |
| [Disease]\_FRST\_EXP\_DT | The minimum date of the (subset of) [Selection Criteria]\_FRST\_EXP\_DTs. |
| LOOKBACK\_DYS | Number of days between [Disease]\_FRST\_EXP\_DT and LOOKBACK\_DT |
| LOOKBACK\_MNTHS | Number of months between [Disease]\_FRST\_EXP\_DT and LOOKBACK\_DT |
| PAT\_AGE | Patient age on [Disease]\_FRST\_EXP\_DT |
| PAT\_GENDER | Patient Gender |
| [OTHER] | [Every study is in some way unique and will require its own metrics for the initial Positive Cohort.] |

QC:

* If [Selection Criteria #n]\_FLAG=1, [Selection Criteria #n]\_Claim Count will >0, [Selection Criteria #n]\_FRST\_EXP\_DT will not null. But [Selection Criteria #n]\_HCP\_SPEC could be blank if the prescriber’s specialty is missing from the data base.
* If [Selection Criteria #n]\_FLAG=0, [Selection Criteria #n]\_Claim Count will =0, [Selection Criteria #n]\_FRST\_EXP\_DT will be null.

In the TD example, create the below metrics which include the standard metrics and also the metrics specific to TD.

Red highlighted is the standard, it can be one standard code. The specific metrics could be another code.

|  |  |
| --- | --- |
| Column Name | Column Definition |
| Patient\_ID |  |
| LRx\_FLAG | 1 if patient has at least 1 claim in LRx in the Study Time Period for ANY clinical code in any market; 0 otherwise |
| Dx\_FLAG | 1 if patient has at least 1 claim in Dx in the Study Time Period for ANY clinical code in any market; 0 otherwise |
| FRST\_RX\_CLM\_DT | Date of first LRx claim in the Study Time Period for ANY clinical code in any market |
| FRST\_DX\_CLM\_DT | Date of first Dx claim in the Study Time Period for ANY clinical code in any market |
| LOOKBACK\_DT | Maximum of FRST\_RX\_CLM\_DT and FRST\_DX\_CLM\_DT |
| LST\_RX\_CLM\_DT | Date of latest LRx claim in the Study Time Period for ANY clinical code in any market |
| LST\_DX\_CLM\_DT | Date of latest Dx claim in the Study Time Period for ANY clinical code in any market |
| LST\_INDEX\_DT | Maximum of LST\_RX\_CLM\_DT and LST\_DX\_CLM\_DT |

|  |  |
| --- | --- |
| TD\_flag | 1 since it is all positive patients |
| TD\_claim\_count | total TD claims counts during study time (201010 to 201612) |
| TD\_First\_Exposure\_Date | First exposure date for Disease TD |
| TD\_First\_Exposure\_Mo\_Yr | First exposure month year for Disease TD |
| TD \_HCP\_specilty | Specialty of the HCP issuing the TD diagnosis on the First TD Exposure date |
| LOOKBACK\_DYS | Days between First TD Exposure date and Lookback date |
| LOOKBACK\_MNTHS | Months between First TD Exposure date and Lookback date |
| Patient Gender | Patient gender |
| Patient Age | Patient age at the first TD exposure date |
| APD\_flag | 1 if patient has at least one APD claim from 20100101 to first disease TD exposure date, else 0 |
| APD\_claim\_count | Total APD claim count from 201001o1 to first disease TD exposure date |
| METOCLOPRAMIDE\_flag | 1 if patient has at least one METOCLOPRAMIDE claim from 20100101 to first disease TD exposure date, else 0 |
| METOCLOPRAMIDE\_claim\_count | Total METOCLOPRAMIDE claim count from 201001o1 to first disease TD exposure date |
| Diseases Treated with AntiPsych | 1 if patient has at least one Diseases Treated with AntiPsych claim from 20100101 to first disease TD exposure date, else 0 |
| Diseases Treated with AntiPsych\_claim\_count | Total Diseases Treated with AntiPsych claim count from 201001o1 to first TD exposure date |
| First\_APD\_Exposure\_Date | First\_APD\_Exposure\_Date prior first TD exposure date |
| First\_APD\_Exposure\_Mo\_Yr | First\_APD\_Exposure\_Date prior first TD exposure month year |
| First\_Metoclopramide\_Exposure\_Date | First\_ Metoclopramide \_Exposure\_Date prior first TD exposure date |
| First\_Metoclopramide\_Exposure\_Mo\_Yr | Month year of First\_ Metoclopramide \_Exposure\_Date |
| First\_APD/Metoclopramide\_Exposure\_Date | The minimum date of First\_APD\_Exposure\_Date and First\_APD\_Exposure\_Date |
| First\_APD/Metoclopramide\_Exposure\_Mo\_Yr | Month year of First\_APD/Metoclopramide\_Exposure\_Date |
| Time Between First TD and APD Exposure (in days) | Days between first APD exposure day to first TD date |
| Time Between First TD and APD Exposure (in months) | Months between first APD exposure day to first TD date |
| Time Between First TD and Metoclopramide Exposure (in days) | Days between first Metoclopramide exposure day to first TD date |
| Time Between First TD and Metoclopramide Exposure (in months) | Months between first Metoclopramide exposure day to first TD date |
| Time Between First TD and APD/Metoclopramide Exposure (in days) | Days between first APD/Metoclopramide exposure day to first TD date |
| Time Between First TD and APD/Metoclopramide Exposure (in months) | Months between first APD/Metoclopramide exposure day to first TD date |

1. After the initial analysis, the Positive\_cohort\_initial was further restricted requiring patients to satisfy all of the following criteria

* LRX\_FLAG =1 and DX\_FLAG\_=1 (i.e. in both Dx and RX data base)
* Valid age and gender. i.e. no negative age, no unknown age or unknown gender
* [Disease]\_FRST\_EXP\_DT >= a specific date
* LOOKBACK\_MNTHS >=N months
* [Selection Criteria #1]\_FLAG =1
* [Selection Criteria #n]\_FLAG=1
* [Selection Criteria #1]\_Claim Count >=n (n=1,2,3..)
* [Selection Criteria #n]\_Claim Count>=n (n=1,2,3..)
* Other specific criteria

In TD example, the criteria is

* LRX\_FLAG =1 and DX\_FLAG\_=1 (i.e. in both Dx and RX data base)
* Valid age and gender. i.e. no negative age, no unknown age or unknown gender
* Disease\_FRST\_EXP\_DT >= 01Jan2012
* LOOKBACK\_MNTHS >=24
* TD\_FLAG =1
* TD\_Claim Count>=1
* APD\_Flag=1 or METOCLOPRAMIDE\_flag=1

Save the eligible patients and all the initial metrics as Positive\_cohort

1. Similar for the Scoring \_cohort\_initial, create the same variables as positive\_cohort\_initial:

Note: treat index date (LST\_INDEX\_DT) of scoring cohort as [Disease]\_FRST\_EXP\_DT in positive cohort

For each patient in Scoring\_cohort\_initial, create the below standard variables:

|  |  |
| --- | --- |
| Column Name | Column Definition |
| Patient\_ID |  |
| LRx\_FLAG | 1 if patient has at least 1 claim in LRx in the Study Time Period for ANY clinical code in any market; 0 otherwise |
| Dx\_FLAG | 1 if patient has at least 1 claim in Dx in the Study Time Period for ANY clinical code in any market; 0 otherwise |
| FRST\_RX\_CLM\_DT | Date of first LRx claim in the Study Time Period for ANY clinical code in any market |
| FRST\_DX\_CLM\_DT | Date of first Dx claim in the Study Time Period for ANY clinical code in any market |
| LOOKBACK\_DT | Maximum of FRST\_RX\_CLM\_DT and FRST\_DX\_CLM\_DT |
| LST\_RX\_CLM\_DT | Date of latest LRx claim in the Study Time Period for ANY clinical code in any market |
| LST\_DX\_CLM\_DT | Date of latest Dx claim in the Study Time Period for ANY clinical code in any market |
| LST\_INDEX\_DT | Maximum of LST\_RX\_CLM\_DT and LST\_DX\_CLM\_DT |
| LOOKBACK\_DYS | Number of days between LST\_INDEX\_DT and LOOKBACK\_DT |
| LOOKBACK\_MNTHS | Number of months between LST\_INDEX\_DT and LOOKBACK\_DT |
| PAT\_AGE | Patient age on LST\_INDEX\_DT |
| PAT\_GENDER | Patient Gender |
| [OTHER] | [Every study is in some way unique and will require its own metrics for the initial Positive Cohort.] |

In the TD example, create the below metrics which include the standard metrics and also the metrics specific to TD.

Red highlighted is the standard, it can be one standard code. The specific metrics could be another code.

|  |  |
| --- | --- |
| Column Name | Column Definition |
| Patient\_ID |  |
| LRx\_FLAG | 1 if patient has at least 1 claim in LRx in the Study Time Period for ANY clinical code in any market; 0 otherwise |
| Dx\_FLAG | 1 if patient has at least 1 claim in Dx in the Study Time Period for ANY clinical code in any market; 0 otherwise |
| FRST\_RX\_CLM\_DT | Date of first LRx claim in the Study Time Period for ANY clinical code in any market |
| FRST\_DX\_CLM\_DT | Date of first Dx claim in the Study Time Period for ANY clinical code in any market |
| LOOKBACK\_DT | Maximum of FRST\_RX\_CLM\_DT and FRST\_DX\_CLM\_DT |
| LST\_RX\_CLM\_DT | Date of latest LRx claim in the Study Time Period for ANY clinical code in any market |
| LST\_DX\_CLM\_DT | Date of latest Dx claim in the Study Time Period for ANY clinical code in any market |
| LST\_INDEX\_DT | Maximum of LST\_RX\_CLM\_DT and LST\_DX\_CLM\_DT |

|  |  |
| --- | --- |
| LOOKBACK\_DYS | Days between LST\_INDEX\_DT and Lookback date |
| LOOKBACK\_MNTHS | Months between LST\_INDEX\_DT and Lookback date |
| Patient Gender | Patient gender |
| Patient Age | Patient age at LST\_INDEX\_DT |
| APD\_flag | 1 if patient has at least one APD claim from 20100101 to LST\_INDEX\_DT, else 0 |
| APD\_claim\_count | Total APD claim count from 201001o1 to LST\_INDEX\_DT |
| METOCLOPRAMIDE\_flag | 1 if patient has at least one METOCLOPRAMIDE claim from 20100101 to LST\_INDEX\_DT, else 0 |
| METOCLOPRAMIDE\_claim\_count | Total METOCLOPRAMIDE claim count from 201001o1 to LST\_INDEX\_DT |
| Diseases Treated with AntiPsych | 1 if patient has at least one Diseases Treated with AntiPsych claim from 20100101 to LST\_INDEX\_DT, else 0 |
| Diseases Treated with AntiPsych\_claim\_count | Total Diseases Treated with AntiPsych claim count from 201001o1 to LST\_INDEX\_DT |
| First\_APD\_Exposure\_Date | First\_APD\_Exposure\_Date prior LST\_INDEX\_DT |
| First\_APD\_Exposure\_Mo\_Yr | First\_APD\_Exposure\_Date prior LST\_INDEX\_DT month year |
| First\_Metoclopramide\_Exposure\_Date | First\_ Metoclopramide \_Exposure\_Date prior LST\_INDEX\_DT |
| First\_Metoclopramide\_Exposure\_Mo\_Yr | Month year of First\_ Metoclopramide \_Exposure\_Date |
| First\_APD/Metoclopramide\_Exposure\_Date | The minimum date of First\_APD\_Exposure\_Date and First\_ Metoclopramide \_Exposure\_Date |
| First\_APD/Metoclopramide\_Exposure\_Mo\_Yr | Month year of First\_APD/Metoclopramide\_Exposure\_Date |
| Time Between First TD and APD Exposure (in days) | Days between first APD exposure day to LST\_INDEX\_DT |
| Time Between First TD and APD Exposure (in months) | Months between first APD exposure day to LST\_INDEX\_DT |
| Time Between First TD and Metoclopramide Exposure (in days) | Days between first Metoclopramide exposure day to LST\_INDEX\_DT |
| Time Between First TD and Metoclopramide Exposure (in months) | Months between first Metoclopramide exposure day to LST\_INDEX\_DT |
| Time Between First TD and APD/Metoclopramide Exposure (in days) | Days between first APD/Metoclopramide exposure day to LST\_INDEX\_DT |
| Time Between First TD and APD/Metoclopramide Exposure (in months) | Months between first APD/Metoclopramide exposure day to LST\_INDEX\_DT |

1. After the initial analysis, the Scoring cohort\_initial was further restricted requiring patients to satisfy all of the following criteria, similar criteria as positive cohort.

* LRX\_FLAG =1 and DX\_FLAG\_=1 (i.e. in both Dx and RX data base)
* Valid age and gender. i.e. no negative age, no unknown age or unknown gender
* LST\_INDEX\_DT >= a specific date
* LOOKBACK\_MNTHS >=N months
* Other specific criteria

In TD example, the criteria is

* LRX\_FLAG =1 and DX\_FLAG\_=1 (i.e. in both Dx and RX data base)
* Valid age and gender. i.e. no negative age, no unknown age or unknown gender
* LST\_INDEX\_DT >=01jan2012
  + LOOKBACK\_MNTHS >=24
  + APD\_Flag=1 or METOCLOPRAMIDE\_flag=1

Save the eligible patients and all the initial metrics as Scoring\_cohort

## Appending Patient Medical History

For each patient in positive cohort, scoring cohorts,

1. Extract all relevant predictors (diagnoses, procedures, symptoms, products.) during the studying time (from Jan 2010 through study end month). Subset to the metrics available between [Disease]\_FRST\_EXP\_DT for the positive cohort (or LST\_INDEX\_DT date for the scoring cohort) and the LOOKBACK\_DT
2. Create 0/1 flags for each predictor
3. Keep the date of first exposure for each predictor (the earliest activity for each predictor since LOOKBACK\_DT)
4. Count the number of occurrences of each predictor ( i.e. count the number of unique visits and claims)
5. Calculate average claim (visit) count as claim (visit) count/available days\*365. Available days is calculated between the LOOKBACK\_DT and the [Disease]\_FRST\_EXP\_DT for positive patients (LST\_INDEX\_DT for scoring patients) which is LOOKBACK\_DYS
6. For each predictor category, calculate the unique predictor category count, for example:
   1. Unique Misdiagnosis Count
   2. Unique Symptom Count
   3. Unique Comorbidities Count
   4. Unique Treatment Count
   5. Unique Procedure Count
   6. Unique Specialty Visit Count

(Note: about the specialty visit. Although Specialty visit is a predictor, we didn’t use the specialty as a selection criteria to select patient cohort. Patient cohort is defined by the clinical codes. After we select the patients with at least one clinical codes, then we will look at the specialty visit of the patients. )

Here is the relevant predictor’s layout in TD:



Note:

* When select the relevant predictor claims,
  + For positive cohort, select the claims where claim service date>= LOOKBACK\_DT and the claims service date<= [Disease]\_FRST\_EXP\_DT
  + For Scoring cohort, select the claims where service date>= LOOKBACK\_DT and the claims service date<= LST\_INDEX\_DT
* When calculate the specialty visit count,
  + Using patient mart table [**V\_PMT\_PAT\_PROVIDER\_MTH**](http://gmr/NDW/NGPS_Detail_Object.asp?SystemID=NDW&ModelNm=NDW&ModelID=31&TypeCode=TABLE&TypeName=V%5FPMT%5FPAT%5FPROVIDER%5FMTH&FilterTerm= &FilterObject= &FilterName= &FilterClass=N&ClassID=2)
  + For positive cohort, select the claims where claim service month < month of the [Disease]\_FRST\_EXP\_DT and the claims service month >= month of the LOOKBACK\_DT
  + For Scoring cohort, select the claims where service month < month of the LST\_INDEX\_DT and the claims service month >= month of the LOOKBACK\_DT
  + The first specialty visit exposure date could be set as the last day of the minimum claim service month

After append the medical history predictor variables, the positive\_cohort \_output and scoring\_cohort\_output would have the below layouts:

* Patient\_Id
* Predictor\_1\_flag
* Predictor\_2\_flag
* ….
* Predictor\_N\_flag
* Specialty\_1\_visit\_flag
* Specialty\_n\_visit\_flag
* Predictor\_1\_count
* Predictor\_2\_count
* ….
* Predictor\_N\_count
* Specialty\_1\_visit\_count
* Specialty\_n\_visit\_count
* Predictor\_1\_ave\_count
* Predictor\_2\_ ave\_count
* ….
* Predictor\_N\_ ave\_count
* Specialty\_1\_visit\_ave\_count
* Specialty\_n\_visit\_ ave\_count
* Predictor\_1\_first\_exp\_dt
* Predictor\_2\_ first\_exp\_dt
* ….
* Predictor\_N\_ first\_exp\_dt
* Specialty\_1\_visit\_ first\_exp\_dt
* Specialty\_2\_visit\_ first\_exp\_dt
* …
* Specialty\_N\_visit\_ first\_exp\_dt
* Unique Misdiagnosis Count
* Unique Symptom Count
* Unique Comorbitities Count
* Unique Treatment Count
* Unique Procedure Count
* Unique Specialty Visit Count

QC steps:

* If Predictor\_n\_flag=1, then Predictor\_n\_count >0, Predictor\_N\_ ave\_count>0 Predictor\_N\_ first\_exp\_dt is not missing
* If Predictor\_n\_flag=0, then Predictor\_n\_count =00, Predictor\_N\_ ave\_count=0 Predictor\_N\_ first\_exp\_dt is blank
* All the predictor exposure date is >= LOOKBACK\_DT and <=[Disease]\_FRST\_EXP\_DT for the positive patients (<=LST\_INDEX\_DT for the scoring patients)
* The unique predictor category count should be equal to all the Predictor\_n\_flag add up. And The unique predictor category count should be less than or equal to all the Predictor\_n\_count add up (for example, the Unique Misdiagnosis Count=2, prodctor1 and predictor 2 belongs to Misdiagnosis, Predictor\_1\_flag+ Predictor\_2\_flag should be 2. Predictor\_1\_count+ Predictor\_2\_count>=2)

1. Negative cohort

Each patient in the Positive Cohort will be matched to a number of patients from the Scoring Cohort. The value of the match number will differ from project to project and will be dependent on the expected prevalence of the disease.

* + 1. For all the patients in the scoring cohort, check their activities in each month during study time from patient mart table V\_PMT\_PAT\_MTH. As long as a patient has either an Rx or Dx claim in the month regardless of the market definition/across any clinical codes, this patient has activities in this month. Save the table as Scoring\_patient\_activity\_month with:
  + Patient\_ID
  + Month\_ID
  + Claim\_count
  + First\_svc\_dt
  + Last\_svc\_dt
    1. For positive patient, randomly split them into 2 cohorts. One positive cohort has x% of the total positive patients and another cohort has Y% positive patients. X%+Y%=100%. X\_positive\_cohort will match N negative patients from scoring cohort. Y\_positive\_cohort will match M negative patients from scoring cohort. Matched negative patients cannot be repeated.

1. For each patient in the Positive Cohort output, take the month of [Disease]\_FRST\_EXP\_DT and the month of LOOKBACK\_DT
2. From the Scoring Cohort output patients, select N or M patients who satisfy the following criteria:
   1. Have ANY activity in any market within +/- 1 month of months of [Disease]\_FRST\_EXP\_DT (call the last service date of the match exposure month as INDEX\_DT\_2) AND
   2. Have the month of LOOKBACK\_DT within +/- 1 month of the month of LOOKBACK\_DT for the Positive Cohort output patient(call the first service date of the match lookback month as LOOKBACK\_DT\_2)
3. Recalculate LOOKBACK\_DYS\_2 and LOOKBACK\_MNTHS\_2 for each selected patient from the Scoring Cohort Output as LOOKBACK\_DYS\_2 = INDEX\_DT\_2– LOOKBACK\_DT\_2 and LOOKBACK\_MNTHS\_2 = month of INDEX\_DT\_2– month of LOOKBACK\_DT\_2

Note: we will prior the Disease\_exp\_month as the matching month if a scoring patient has activities in all three eligible matching month. for example, if positive patient’s disease\_exp\_month is 201605, a scoring patient has activities at 201604, 201605, 201606, we will take 201605 as the first prior matching month.

Save the table as Match\_positive\_negative\_patients\_X and Match\_positive\_negative\_patients\_Y with below variables:

|  |
| --- |
| PATIENT\_ID\_negative |
| PATIENT\_ID\_positive  Disease\_exp\_month |
| Disease\_lookback\_dt\_month |
| Match\_index\_month  Match\_lookback\_dt\_month  INDEX\_DT\_2  LOOKBACK\_DT\_2  LOOKBACK\_DYS\_2  LOOKBACK\_MNTHS\_2 |

Patient\_id\_negative is from scoring cohort output. Patient\_ID\_positive is from positive cohort.

QC:

* Patient\_id\_negative is unique patient ID, Patient\_id\_positive should repeat N / M times.
* Match\_index\_month -1 <=Disease\_exp\_month <= Match\_index\_month +1
* Match\_lookback\_dt\_month -1 <=Disease\_lookback\_dt\_month <= Match\_lookback\_dt\_month +1

Match\_lookback\_dt\_month=Month of LOOKBACK\_DT\_2

Match\_lindex\_month=Month of INDEX\_DT\_2

(For example,

* If Disease\_exp\_month is 201610, and then the Match\_index\_month should be wihtin 201609 to 201611.
* if Disease\_lookback\_dt\_month is 201210, and then the Match\_lookback\_dt\_month should be within 201209 to 201211)
* Match\_index\_month- Match\_lookback\_dt\_month should be >=24 (the LOOKBACK\_MNTHS)

1. For each matched negative patients (Patient\_id\_negative), recreate all the predictor variables that already created in the scoring cohort output but just change to the new time frame(from LOOKBACK\_Date\_2 to INDEX\_date\_2) instead of the original LOOKBACK\_Date to INDEX\_date. i.e. treat LOOKBACK\_Date\_2 as LOOKBACK\_Date, INDEX\_date\_2=INDEX\_date.

Save it as negative\_output with:

Patient\_Id (this is unique matched scroing patient ID)

Patient\_ID\_test (each ID will repeat N/M times)

LRx\_FLAG

Dx\_FLAG

FRST\_RX\_CLM\_DT

FRST\_DX\_CLM\_DT

LOOKBACK\_DT (=LOOKBACK\_Date\_2)

LST\_RX\_CLM\_DT

LST\_DX\_CLM\_DT

LST\_INDEX\_DT (=INDEX\_date\_2)

LOOKBACK\_DYS

LOOKBACK\_MNTHS

Patient Gender

Patient Age (recalculate from INDEX\_DT\_2 –birth year)

….

Predictor\_1\_flag

Predictor\_2\_flag

….

Predictor\_N\_flag

Predictor\_1\_count

Predictor\_2\_count

….

Predictor\_N\_count

Predictor\_1\_ave\_count

Predictor\_2\_ ave\_count

….

Predictor\_N\_ ave\_count

Predictor\_1\_first\_exp\_dt

Predictor\_2\_ first\_exp\_dt

….

Predictor\_N\_ first\_exp\_dt

Specialty\_1\_visit\_count

Specialty\_2\_visit\_count

…

Specialty\_N\_visit\_count

Unique Misdiagnosis Count

Unique Symptom Count

Unique Comorbitities Count

Unique Treatment Count

Unique Procedure Count

Unique Specialty Visit Count

QC steps:

* If Predictor\_n\_flag=1, then Predictor\_n\_count >0, Predictor\_N\_ ave\_count>0 Predictor\_N\_ first\_exp\_dt is not missing
* If Predictor\_n\_flag=0, then Predictor\_n\_count =00, Predictor\_N\_ ave\_count=0 Predictor\_N\_ first\_exp\_dt is blank
* All the predictor exposure date is >= LOOKBACK\_DT and <=LST\_INDEX\_DT
* The unique predictor category count should be equal to all the Predictor\_n\_flag add up. And The unique predictor category count should be less than or equal to all the Predictor\_n\_count add up (for example, if the Unique Misdiagnosis Count=2, prodctor1 and predictor 2 belongs to Misdiagnosis, so Predictor\_1\_flag+ Predictor\_2\_flag should be 2. Predictor\_1\_count+ Predictor\_2\_count should >=2. )
* Patient age is not negative or unknown
* Patient gender is F or M

10, combine all the variables into one table for each cohort.

**Positive cohort layout:**

Patient\_ID

LRx\_FLAG

Dx\_FLAG

FRST\_RX\_CLM\_DT

FRST\_DX\_CLM\_DT

LOOKBACK\_DT

LST\_RX\_CLM\_DT

LST\_DX\_CLM\_DT

LST\_INDEX\_DT

[Selection Criteria #1]\_FLAG

[Selection Criteria #1]\_Claim Count

[Selection Criteria #1]\_FRST\_EXP\_DT

[Selection Criteria #1]\_HCP\_SPEC

[Selection Criteria #n]\_FLAG

[Selection Criteria #n]\_Claim Count

[Selection Criteria #n]\_FRST\_EXP\_DT

[Selection Criteria #n]\_HCP\_SPEC

[Disease]\_FRST\_EXP\_DT

LOOKBACK\_DYS

LOOKBACK\_MNTHS

PAT\_AGE

PAT\_GENDERPredictor\_1\_flag

Predictor\_2\_flag

….

Predictor\_N\_flag

Predictor\_1\_count

Predictor\_2\_count

….

Predictor\_N\_count

Predictor\_1\_ave\_count

Predictor\_2\_ ave\_count

….

Predictor\_N\_ ave\_count

Predictor\_1\_first\_exp\_dt

Predictor\_2\_ first\_exp\_dt

….

Predictor\_N\_ first\_exp\_dt

Specialty\_1\_visit\_count

Specialty\_2\_visit\_count

…

Specialty\_N\_visit\_count

**Scoring cohort layout**:

Patient\_ID

LRx\_FLAG

Dx\_FLAG

FRST\_RX\_CLM\_DT

FRST\_DX\_CLM\_DT

LOOKBACK\_DT

LST\_RX\_CLM\_DT

LST\_DX\_CLM\_DT

LST\_INDEX\_DT

LOOKBACK\_DYS

LOOKBACK\_MNTHS

Patient Gender

Patient Age

Predictor\_1\_flag

Predictor\_2\_flag

….

Predictor\_N\_flag

Predictor\_1\_count

Predictor\_2\_count

….

Predictor\_N\_count

Predictor\_1\_ave\_count

Predictor\_2\_ ave\_count

….

Predictor\_N\_ ave\_count

Predictor\_1\_first\_exp\_dt

Predictor\_2\_ first\_exp\_dt

….

Predictor\_N\_ first\_exp\_dt

Specialty\_1\_visit\_count

Specialty\_2\_visit\_count

…

Specialty\_N\_visit\_count