Data Manipulation for VAERS data

(R code for Data_PAT, data_VAX, data_ae, data_AE, full_dds.csv: manipulation.R)

data_PAT
521,831 obs.

- Patient data based on VAERS_ID
- From "VAERSDATA" (2000~2018)
- Variables: VAERS_ID (unique key), CAGE_YR (Vaccine date - Birthday), NUMDAYS, STATE
- 126,381 obs. are missing in CAGE_YR

+

data_ae
1,651,640 obs.

- AE data based on VAERS_ID
- From "VAERSSYMPTOMS"
- Variables: VAERS_ID (**NOT** unique key), AE_NAME, year

+

meddra_ptlist.csv
8230 PT terms
(Code: CleaningMeddraTerms.R)

Web scraping

RESTful API

BioPortal HTML source

+

data_AE
1,651,640 obs.
= data_ae + meddra_ptlist.csv

- In 'data_ae', a lot of AE_NAME are the informal (or old) terms
- Also, 'data_ae' doesn't have their Meddra ID
- Thus, we do a data cleaning by web-scraping

Merge

full_dds.csv
2,551,766 obs.

- 3806 obs. are missing in MEDDRA_AE_NAME (Unique: 52 AE names are unmatched)
- vaxnum = 1 vs vaxnum > 1: 1,159,223obs. vs 1,392,541 obs.
- 338,747 obs. are missing in CAGE_YR

Future Manipulation

- Remove and/or combine some PT terms including unmatched AEs
- Handling missing values
- Handling data based on 'vaxnum' and/or 'wt_report'

data_VAX
798,561 obs.
521,830 unique VARES_ID
(Since a VARES_ID has 1~5 Symptoms)

- Vaccine data based on VAERS_ID
- From "VAERSVAX" (2000~2018)
- Variables: VARES_ID (**NOT** unique key), VAX_NAME, VAX_TYPE
- There is no missing (NA) in this dataset.