Weekly STATUS REPORT

Report Date	TRACK NAME	Prepared By
Sep 26th, 2021	IntegrityM (Data cleaning and Data wrangling)	Yihang Zhao, Yuwen Luo, SeungHeon Han

Status Summary

We received the source of the new datasets, synthetic CMS databases (DE-SynPUF), from our clients.

As a result of the exploration of the datasets, we found a few limits of the datasets in using them for combining, labeling, and modeling processes. On the basis of the outcomes, we and our clients decided to keep using the original CMS datasets.

Project Overview

Main task	Sub tASK	% Done	Due date	Assigned to	NOTES
Loaded data	Concat the 4 dataset "Beneficiary, Inpatient Claims, Outpatient Claims, Carrier from 7 samples	100%	Sep 26th, 2021	Yihang Zhao, Yuwen Luo, SeungHeon Han	
Explorate the new dataset 'DE-SynPUF' clients provided	Further cleaning like grouping the 'Beneficiary' dataset by ID.	100%	Sep 26th, 2021	SeungHeon Han	
	Checking every variable to see the availability of Primary key and forgin key for each dataset for merging them	100%	Sep 26th, 2021	SeungHeon Han	

	Observe any invalid NPI values.	100%	Sep 26th, 2021	SeungHeon Han	
Decision on selecting the new 'DE-SynPUF' data set or the original CMS data set	Meet with clients for decision making.	100%	Sep 26th, 2021	Yihang Zhao, Yuwen Luo, SeungHeon Han	Decide to use the original 'CMS' data set.

Risk and issue history

ISSUE	ASSIGNED TO	DATE
the synthesized Chronic diseases-related features in the Beneficiary dataset are not valid to use for modeling	Yihang Zhao, Yuwen Luo, SeungHeon Han	Sep 26th, 2021
the synthesized NPI variables are not valid to be labeled by LEIE dataset	Yihang Zhao, Yuwen Luo, SeungHeon Han	Sep 26th, 2021
there are not synthesized Claim_ID variables matched across the datasets - Carrier, Inpatient & Outpatient datasets cannot be combined	Yihang Zhao, Yuwen Luo, SeungHeon Han	Sep 26th, 2021

CONCLUSIONS & OUTLOOK

We decided to use the original 'CMS' data sets instead of the new 'DS-Synpuf' datasets.