Building the observational medical outcomes partnership's T-MSIS Analytic File common data model

Nick Williams Ph.D
Lister Hill National Center for Biomedical Communications
National Library of Medicine (USA)
National Institutes of Health
2023

About the paper

- We document the complexity of and build an OMOP CDM out of T-MSIS Analytic extract Files (current vintage Medicaid data).
- Towards complexity, 591,633 variables, 50T, and 320,240,249,348 records in MAX-TAF-RIF/PDE. This is the largest clinical observational database in the world?
- The ETL for the TAF subset found 119 million distinct individuals, while the total set-to-date had 297,494,554 individuals.
- We review the ETL mapping, utilization of DX-PX-RX codes and Databricks.

Project:

We investigate the

-interoperability

-impact

-value add

of complex clinical observational data via

non-trivial real world use cases

Product:

We are building OMOP CDM ETL code for ‘Internal Use’

We are making ETLs for TAF (ASPE), RIF+PDE(OAR) and MAX(OAR)

Progress:

We have an ETL and production CDM for TAF.
RIF+PDE CDM was done the first week of June.
MAX should be done by the end of July-early August.

What does internal use mean?

We are not supporting or developing code for external partners at this time.

We are sharing the code to support your development; you are welcome to use it.

Do note that this code is ‘prototype’, by which we mean that CMS is highly likely to develop and maintain an official CMS OMOP CDM, which the NLM would\will use when available.

This is a stop gap measure, as CMS OMOP CDM will not be available to us for several(?) years.

The real value of OMOP (for us) is normalization (58,800 TVP vs 1 TVP).

Thanks for the background… can we see it?

Not really!

The data lives in a Federal Data Center. FDCs have strict access and environment controls.

We should try a standalone replication of a network study to assess how much of the OMOP codebase is FDC ready.

The end or the beginning:

Thoughts?

Thanks for your time!