CDMV5 Test Data

Lee Evans

LTS Computing LLC

My Background

- levans@ltscomputingllc.com
- OHDSI community member
- My OHDSI activities:
  - Manage OHDSI cloud infrastructure
    - web site/wiki/forums/databases/WebAPI
  - LAERTES knowledgebase data loads
- Owner of LTS Computing LLC
  - http://ltscomputingllc.com
- My company
  - Develops custom ETL processes
  - Provides commercial support for big data analytics, databases & applications (installation/upgrades/administration)

Why Do We Need Test Data?

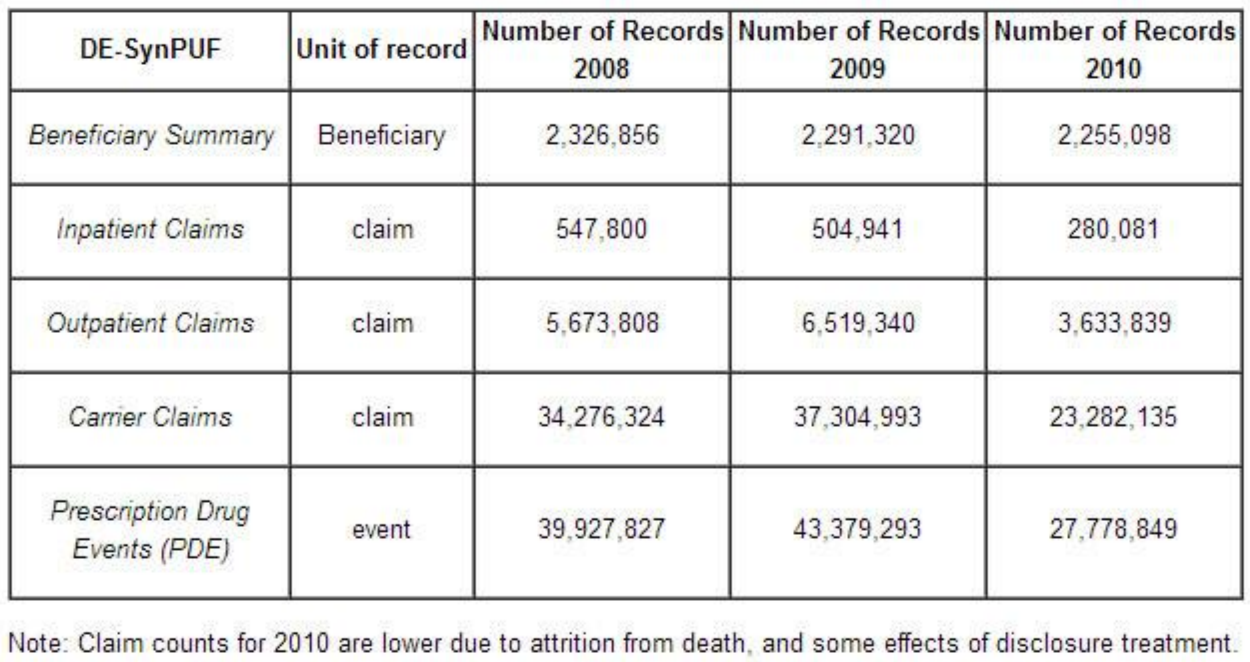
- Public data to demo the OHDSI tools
- Benchmark performance
  - platforms and methods
- Developing & testing OHDSI tools
- OHDSI tools training

Test Data Now Available

- 1000 patient sample of CMS 2008-2010 Data Entrepreneurs’ Synthetic Public Use File (DE-SynPUF)
- Synthetic patients & medicare claims/prescription data
- Converted to OMOP CDM V5 format
- Zip file download from LTS Computing LLC website:

http://www.ltscomputingllc.com/downloads/

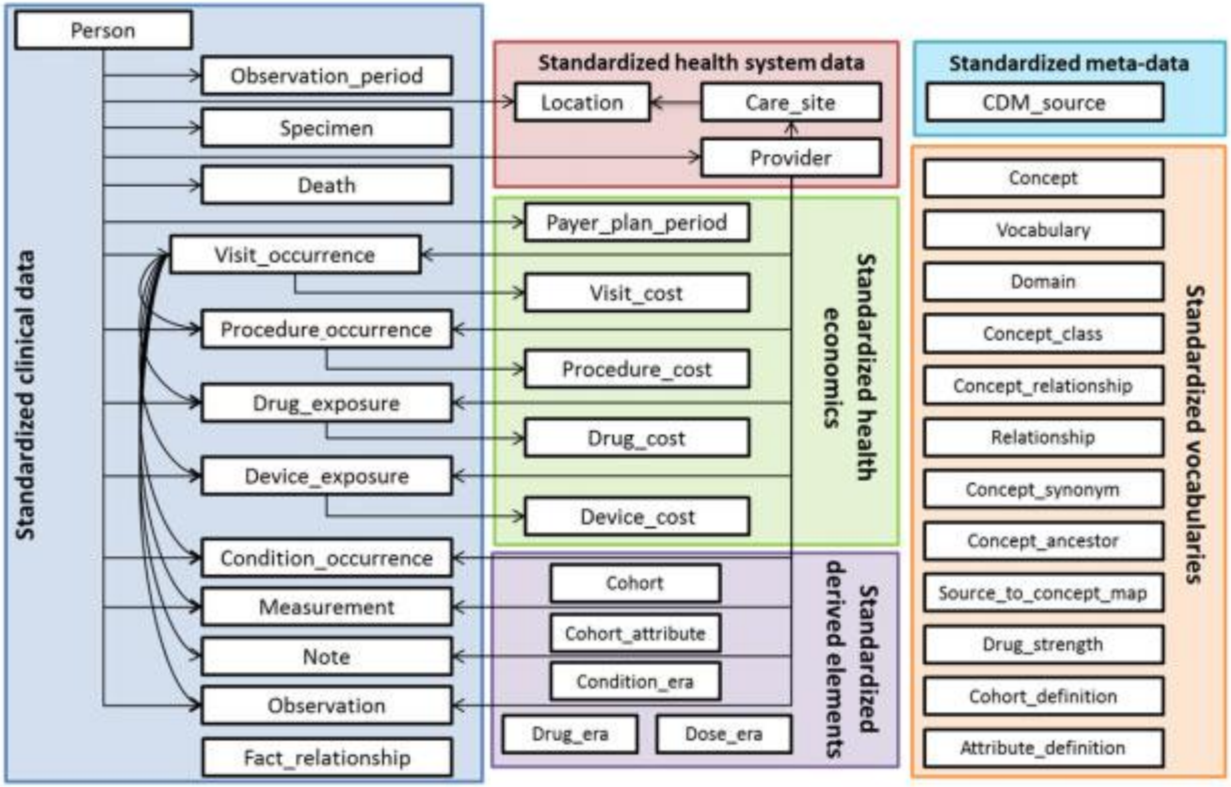
SynPUF source data files



DE-SynPUF	Unit of record	Number of Records 2008	Number of Records 2009	Number of Records 2010
Beneficiary Summary	Beneficiary	2,326,856	2,291,320	2,255,098
Inpatient Claims	claim	547,800	504,941	280,081
Outpatient Claims	claim	5,673,808	6,519,340	3,633,839
Carrier Claims	claim	34,276,324	37,304,993	23,282,135
Prescription Drug Events (PDE)	event	39,927,827	43,379,293	27,778,849

Note: Claim counts for 2010 are lower due to attrition from death, and some effects of disclosure treatment.

CDMV5 data model



Standardized clinical data

Person

Observation_period

Specimen

Death

Visit_occurrence

Procedure_occurrence

Drug_exposure

Device_exposure

Condition_occurrence

Measurement

Note

Observation

Fact_relationship

Standardized health system data

Location

Care_site

Provider

Standardized health economics

Payer_plan_period

Visit_cost

Procedure_cost

Drug_cost

Device_cost

Standardized derived elements

Cohort

Cohort_attribute

Condition_era

Drug_era

Dose_era

Standardized meta-data

CDM_source

Standardized vocabularies

Concept

Vocabulary

Domain

Concept_class

Concept_relationship

Relationship

Concept_synonym

Concept_ancestor

Source_to_concept_map

Drug_strength

Cohort_definition

Attribute_definition

Test data SynPUF ETL challenges

- Multiple source files to combine
  - 1 file per year (3 years)
  - Large claims files split into 20 individual files
- The files have a number of repeating groups
  - Multiple claim line items
  - 6 tax numbers & 13 physician numbers
  - 12 ICD9 codes
  - 15 HCPCS codes
  - 13 payment amounts
- Multiple claim types
  - Carrier, in-patient and out-patient
- The great OHDSI CMS workgroup ETL spec really helped!

Test Data - Some Limitations

- Very small dataset - only 1k patients
- Simulated Medicare population
- No Laboratory data
- No Health Economics cost data converted
- Complex visit logic not implemented
- Converted data not verified in detail

OHDSI Cloud Demo

- ACHILLES - visualize dataset
  - – http://www.ohdsi.org/web/achilles/index.html#/Demo data - 1K synthetic patients/dashboard
- CIRCE - create cohort
  - – http://www.ohdsi.org/web/circe/#/155
- HERACLES - visualize cohort
  - – http://www.ohdsi.org/web/heracles/viewer.html?cohortId=155

Questions & Discussion