Observational Outcomes Data Science Workshop

Creating Reliable Evidence with Standardized Databases

Health Data Science Day

February 10, 2020

Brian Paciotti, PhD, MS

IT Health Informatics

UC Davis

Workshop Agenda

1. Health Sciences and Observational Data
2. Standardized Clinical Data (OMOP/OHDSI)
3. EHR Systems at UC
4. UC Standardization Efforts
5. DataPATH -- De-Identified Data
6. Data Quality and Validation
7. Accessing DataPATH data
8. **Interactive Session (second hour)**

Conflicts of Interest

- The UC Davis DataPATH team does not have any conflicts of interests to report

Acknowledgements! Impressive Collaboration

- **OHDSI Columbia University**
  - Data model and code
  - Documentation
  - I copied slide material!
- **OHSDI Worldwide**
  - S. Korea's entire population in OHDSI common data model

- **UC Health Team**
  - Lisa Dahm
  - Atul Butte
  - Ayan Patel
- **UC Teams**
  - UCLA
  - UCSD
  - UCI
  - UCSF

- **UCD IT Health Informatics**
  - Kent Anderson
  - Doug Berman
  - Steve Covington
  - Calvin Chang
  - Hemanth Tatiparthi
  - Duke Letran
- **UCD Public Health Informatics**
  - Nick Anderson
  - Bill Riedl
  - Chris Lambertus

Health Sciences and Observational Data

Opportunities to Create Knowledge with Observational Data

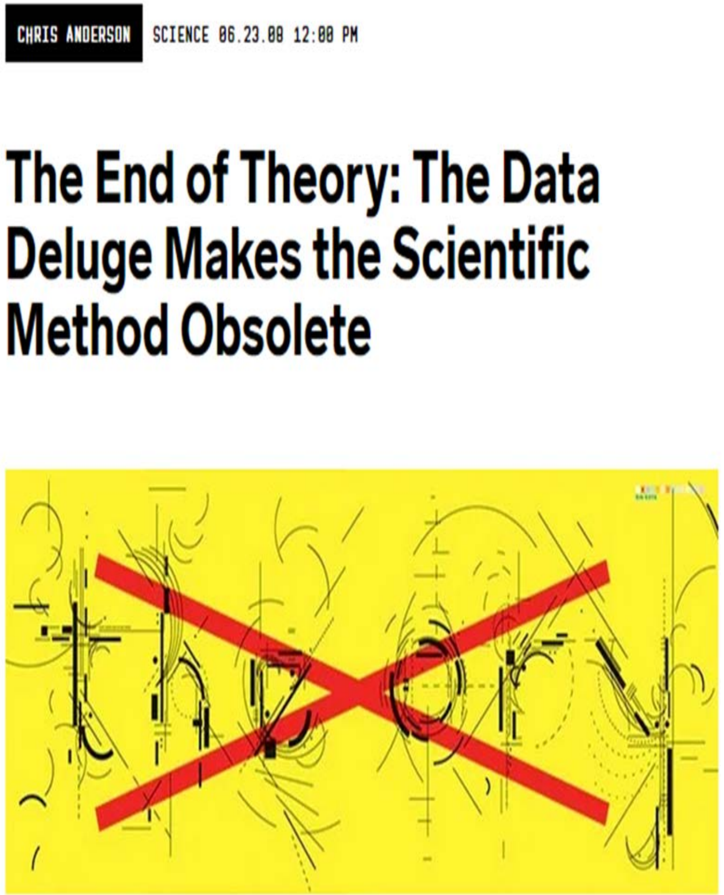
Challenges Associated with Secondary Use of Data for Observational Research

Healthcare Science – Creating Evidence

- Science – Create knowledge (evidence)
  - Symbolic communication (thinking, writing)
  - Models and theory

    “All models are wrong, but some are useful”
- Approaches to quantitative science vary
  - Symbolic vs connectionist
  - Traditional statistical methods vs. data mining
- Evidence implemented by systems (people + technology)
  - **Data** + **Knowledge** = Information
- My assumptions:
  - We need theory/models/metaphors
  - There is no “raw” data.

    Meaning of data is grounded in context


CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Why are Healthcare Data Created?

- Data: external human knowledge
  - Symbolic representations that capture meaning in complex ways
- Primary use of healthcare data
  - EHR systems to manage patient care
  - Billing data systems to request payment
  - Insurance systems to process claims and encounters
- Organizations and processes are complex and variable
  - Clinical workflows
  - Different payers (e.g., Medicaid vs commercial)
- Observational data is complex!
  - “**Data Archeology**”

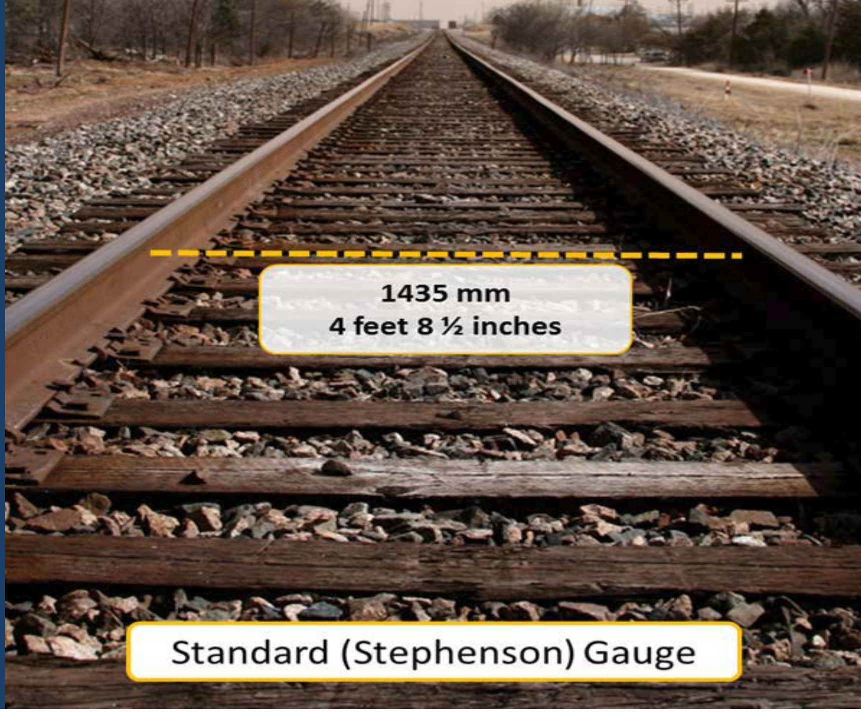
Why is Observational Research Challenging?

- Creating scientific knowledge in the medical domain is challenging
- Teams must have skills and knowledge of the following:
  - Understanding what types of evidence is useful
  - Databases and “data archaeology”
  - Health informatics
  - Modeling and Algorithms
  - Clinical Knowledge
  - Collaboration/teamwork

Challenges of Working with Non-Standard Database

- Time-consuming to map theoretical concepts to database fields
  - Researchers request specific types of labs, meds, or procedures—but they do not know the Clarity lab/procedure codes … “LABSC00026” HBV Core Ab, total
- Clinical databases often have thousands of fields
  - Which ones are important?
- Databases evolve through time as culture evolves (e.g., technology, terminologies)
  - ICD-9 to ICD-10 Transition (Oct 1st, 2016)
  - New EHR Modules lead to changes—often for the good
- Analytic cohorts require specific events, often with particular sequences
  - Events can have many dates and associated concepts
  - Which ones do we choose?

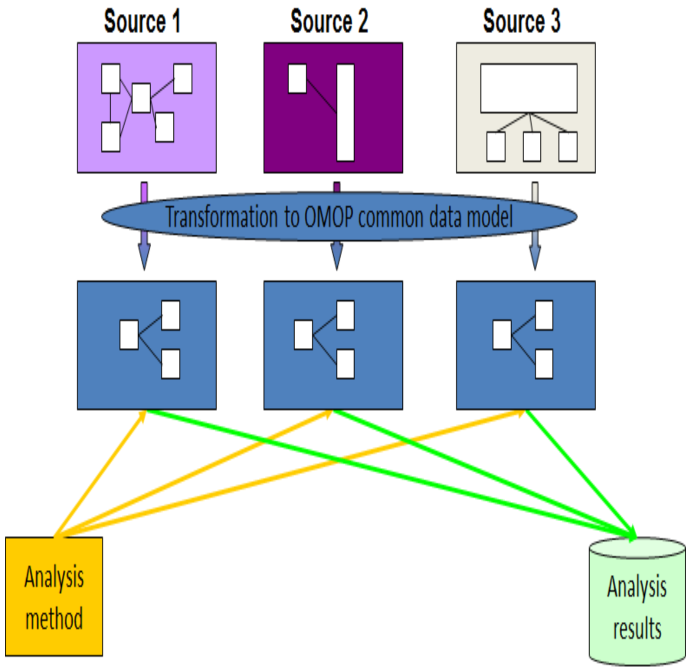
Standardized Clinical Data: OHDSI



Overview of OMOP and OHDSI Efforts to Standardize Clinical Data

Common Data Models (CDMs)

- Common Data Model (CDM) -- a way of organizing data into a **standard** structure
- Observational databases have different purposes and designs
  - Electronic health record systems (EHRs) support clinical practice at the point of care
  - Administrative claims data are built for the insurance processes
- Each collected for a different purpose, resulting in different logical organizations and physical formats
  - Terminologies used to describe the medicinal products and clinical conditions vary from source to source.



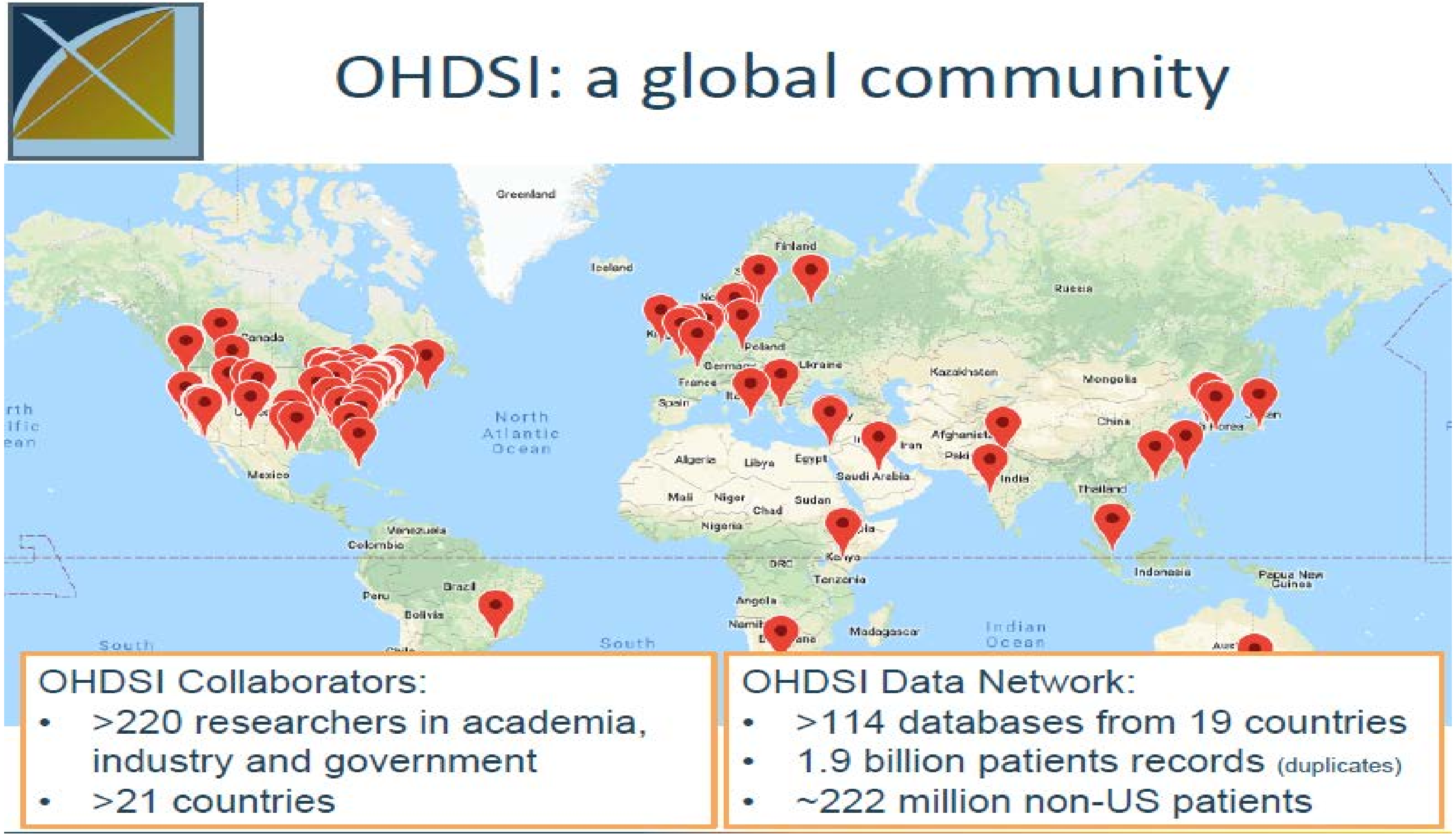
Common Data Models Create Value with Standards

- Standards: Industries have increased efficiency/productivity (e.g., shipping containers, standard railroads)
- Standardizing Healthcare Science
  - Efficiency - opportunity for more efficient knowledge creation
  - Reproduceable knowledge – a way to standardize science!
  - Transparency – use of standardized medical concepts
  - Data is more manageable for data owners and more useful for data users
- CDMs can integrate both administrative claims, EHR data and other sources
  - ETL and mapping processes to standardize data
  - Users to generate evidence from a wide variety of sources
  - Support collaborative research across data sources both within and outside the United States

What is OMOP and OHDSI?

- Observational Medical Outcomes Partnership (OMOP)
  - Public-private partnership, chaired by the US Food and Drug Administration, administered by the Foundation for the National Institutes of Health
  - Consortium of pharmaceutical companies, academic researchers, and health data partners to advance the science of active medical product safety surveillance using observational healthcare data
  - OMOP produced an effective CDM now used around the world
- Observational Health Data Sciences and Informatics program
  - **OHDSI**, pronounced "Odyssey"
  - Multi-stakeholder (coordination at Columbia University)
  - Interdisciplinary collaborative: value of health data through large-scale analytics
  - Non-pharma funded
  - All solutions are open-source
- OMOP partnership has now evolved into the OHDSI program
  - “**OMOP**” is still a common term for this evolving and popular CDM

The OHDSI Community – Special Shout-Out to S. Korea!



OHDSI: a global community

OHDSI Collaborators:

- >220 researchers in academia, industry and government
- >21 countries

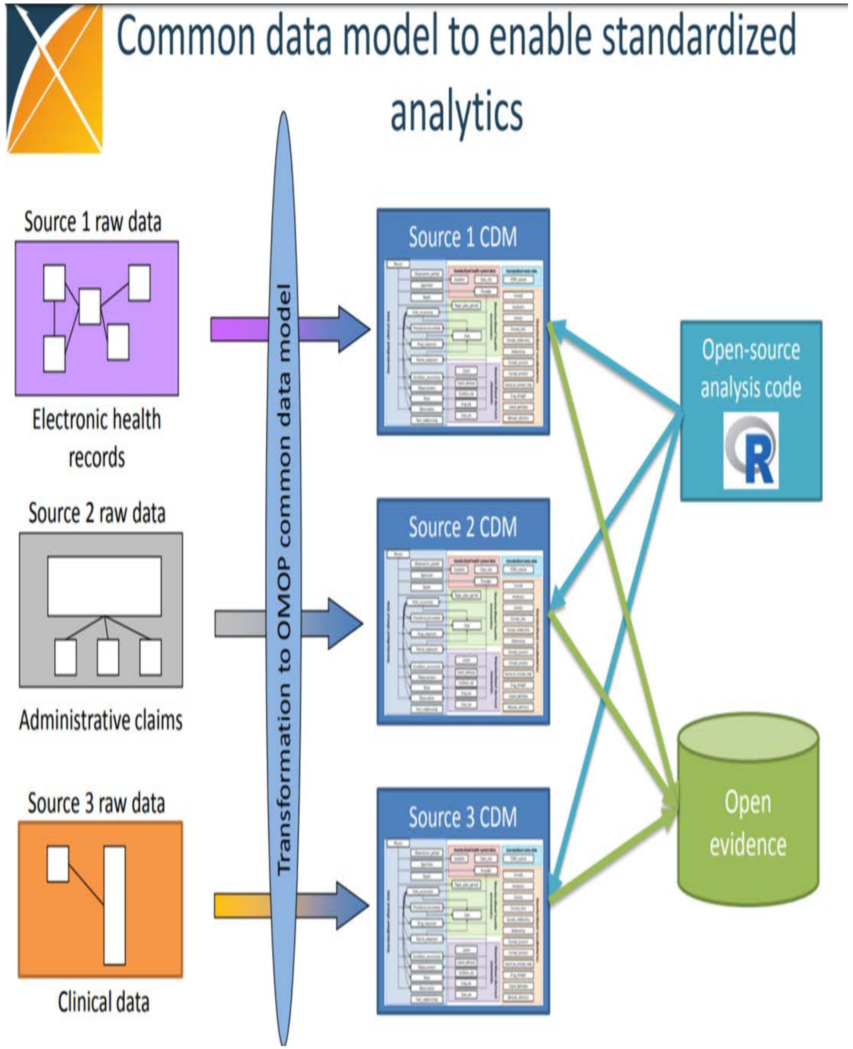
OHDSI Data Network:

- >114 databases from 19 countries
- 1.9 billion patients records (duplicates)
- ~222 million non-US patients

UCDAVIS HEALTH

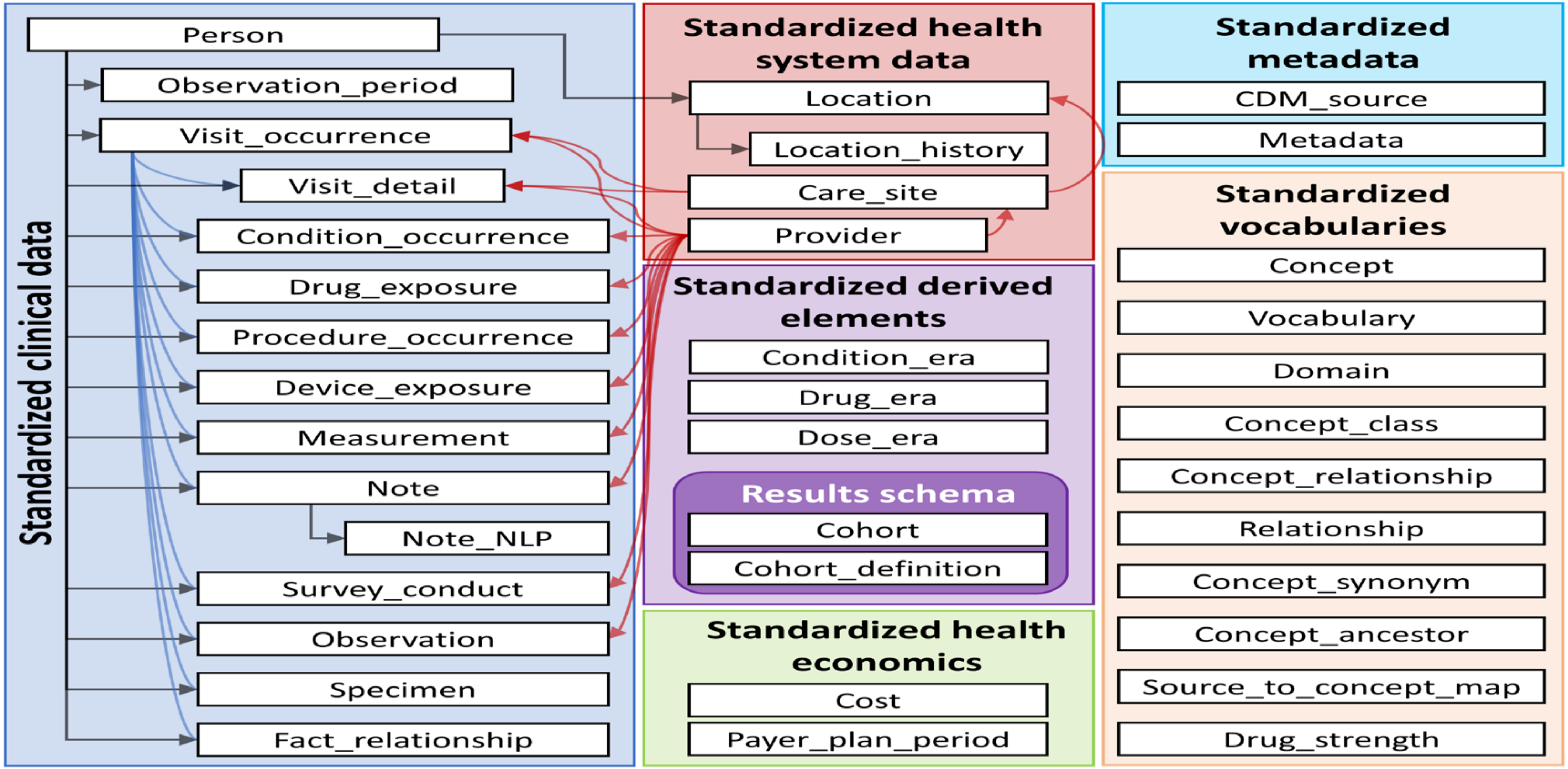
OMOP Common Data Model + Standard Analytics

- Harmonize Disparate Source Systems:
  - Data from information models
  - Varying institutional workflows and underlying conceptual representations
  - Transform data into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes)
- Collaboration using Standard Tools/Algorithms:
  - Disparate teams work together, sharing workload, processes and code
  - Systematic analyses using a library of standard analytic routines written based on the common format



Common data model to enable standardized analytics
Source 1 raw data
Electronic health records
Source 2 raw data
Administrative claims
Source 3 raw data
Clinical data
Transformation to OMOP common data model
Source 1 CDM
Source 2 CDM
Source 3 CDM
Open-source analysis code
R
Open evidence

Common Data Model – OMOP Version 6



Standardized clinical data

Person

Observation_period

Visit_occurrence

Visit_detail

Condition_occurrence

Drug_exposure

Procedure_occurrence

Device_exposure

Measurement

Note

Note_NLP

Survey_conduct

Observation

Specimen

Fact_relationship

Standardized health system data

Location

Location_history

Care_site

Provider

Standardized derived elements

Condition_era

Drug_era

Dose_era

Results schema

Cohort

Cohort_definition

Standardized health economics

Cost

Payer_plan_period

Standardized metadata

CDM_source

Metadata

Standardized vocabularies

Concept

Vocabulary

Domain

Concept_class

Concept_relationship

Relationship

Concept_synonym

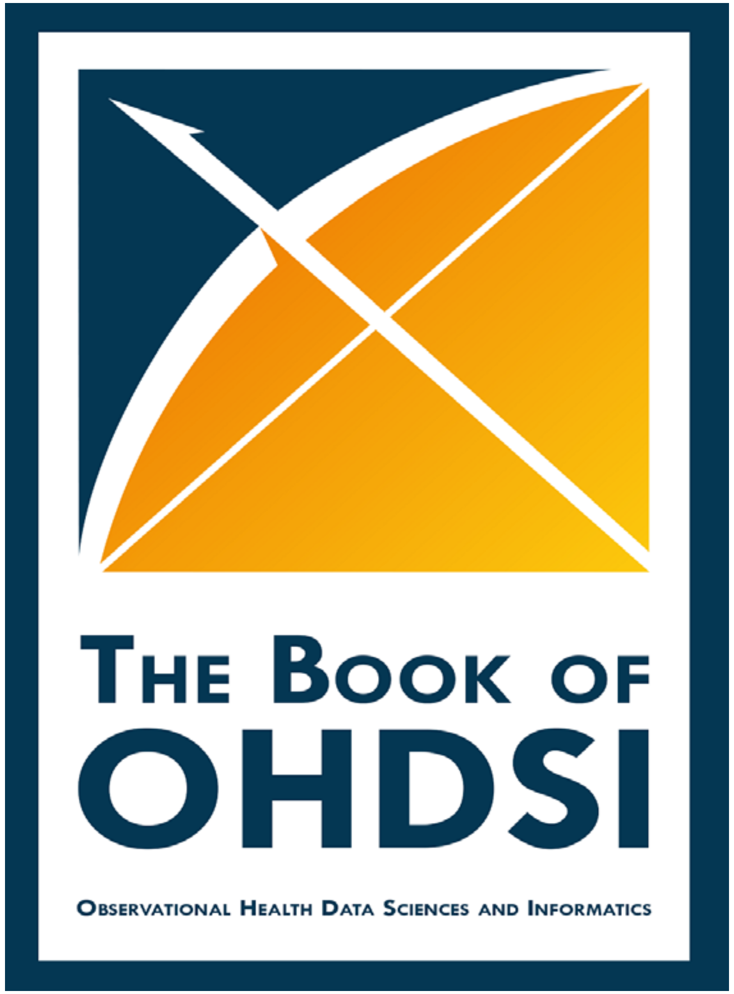
Concept_ancestor

Source_to_concept_map

Drug_strength

The Gift from the OHDSI Community !!!

- Researchers can benefit from cumulative cultural evolution
  - Join, Copy, Collaborate, Share
  - Enjoy structures, tools, and code created by others
- What tools can researchers borrow?:
  - Excellent books, slides, demos, wiki posts
  - ATLAS interface
  - R modules for statistical modeling
  - Research designs / cohort definition methods
  - Data model summaries
- I only have time to share a tiny fraction of the available materials!



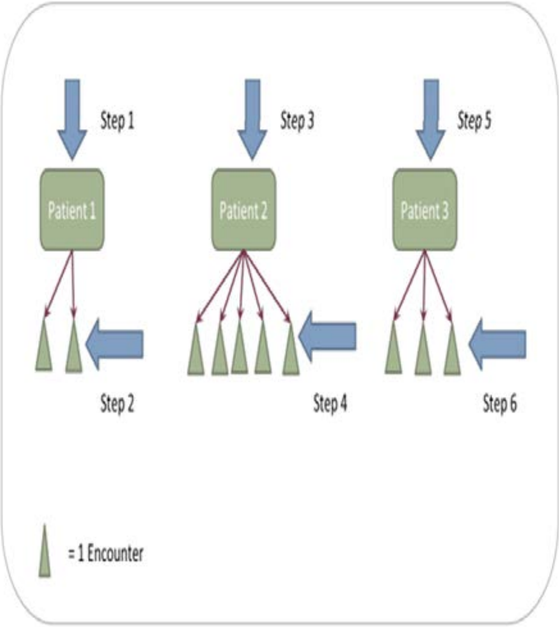
EHR Data at UC Institutions

Brief Summary of EPIC Electronic Health Record System

EPIC Systems and Associated Databases

Clinicians manage patient care using EPIC's *Hyperspace* graphical user interface

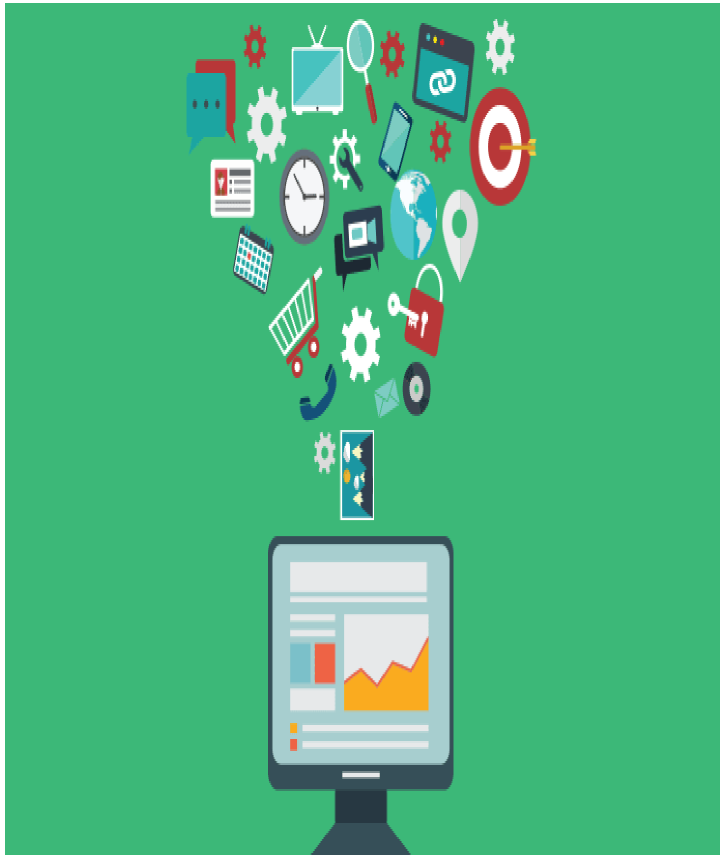
Data in Hyperspace are stored in a hierarchical database called *Chronicles*



Data extracted nightly into relational database called *Clarity* for population reporting

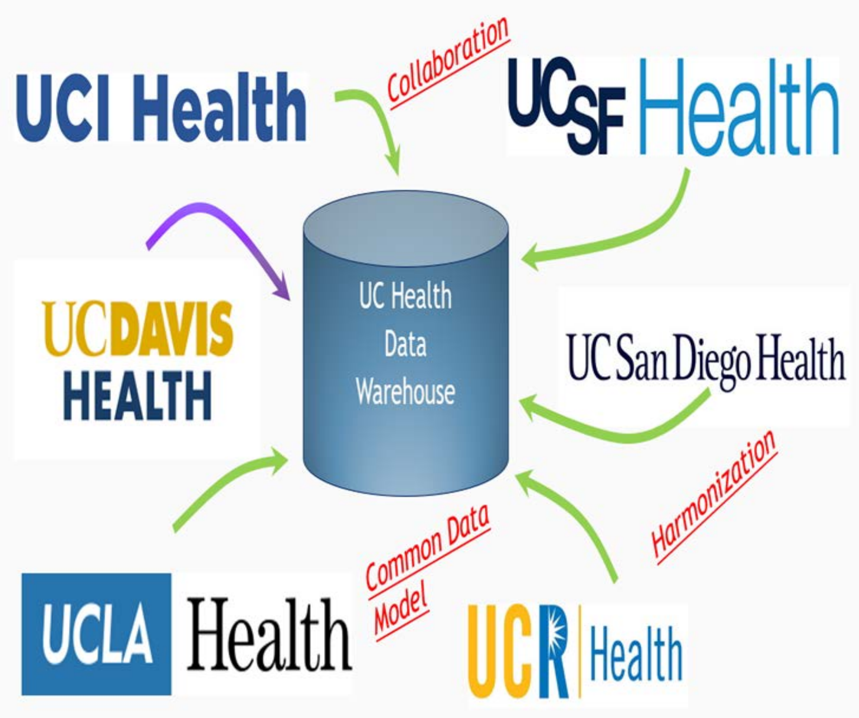
Epic Systems among UCs Are NOT Standardized

- Clinical workflows, billing systems, and other processes vary among healthcare organizations
- With different workflows and preferences, Epic has historically allowed organizations to use different codes and modules
  - Medications referenced by MEDICATATION_IDs can vary
  - Flowsheets are created by local clinical workflows—different IDs are created for similar concepts
- Data harmonization and analytics requires standard codes



EPIC Query Tools -- Slicer/Dicer

- Standardized databases developed by UCs are not the only available databases for analytics
- EPIC systems has developed sophisticated tools within Hyperspace to query data and extract patient data
  - Tools such as “SlicerDicer” have a lot of promise for research and analytics
- OMOP CDM and EPIC tools likely can complement one another



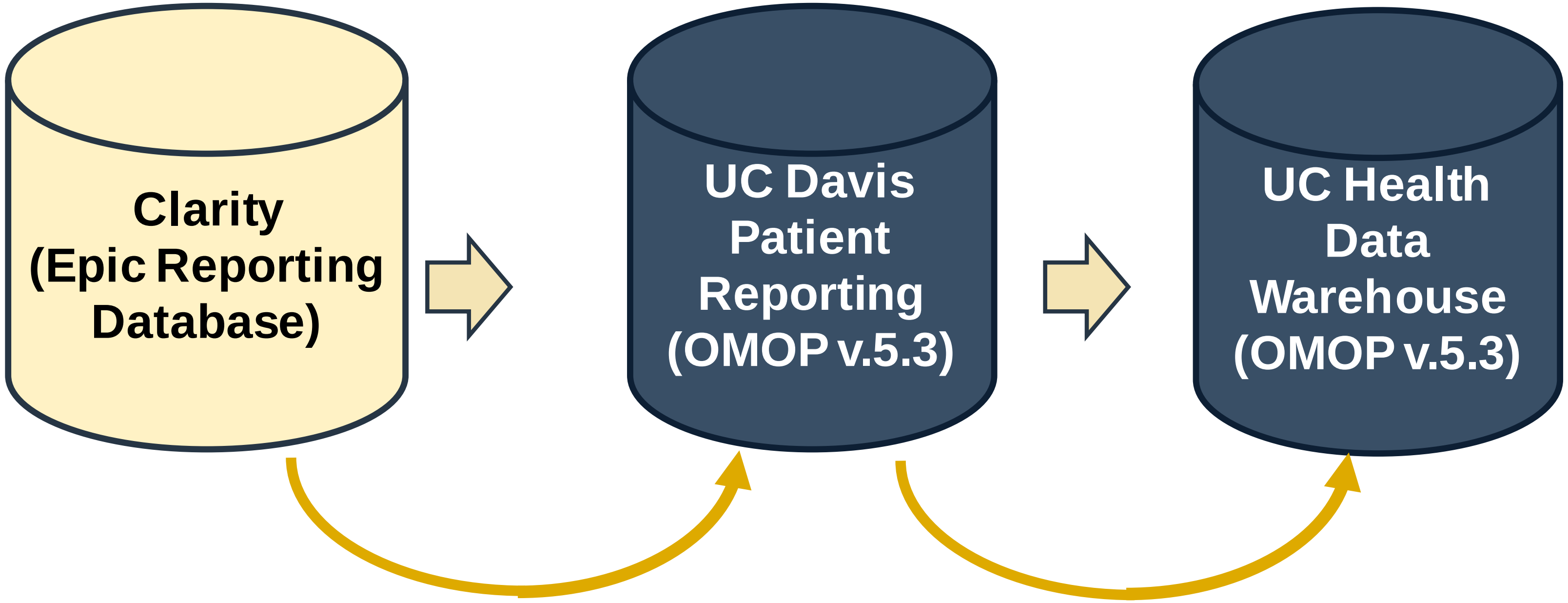
Standardized Data at UC Health

Implementing the OHDSI (OMOP) Model at UC Health

UC Health Data Warehouse

- A centralized, secure, healthcare data warehouse and analytics platform covering all UC Health sites that supports strategic data driven initiatives

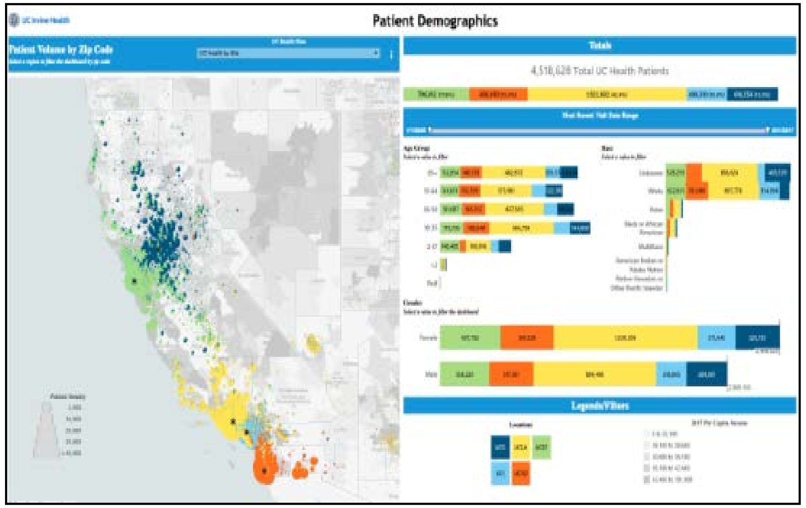
UC HEALTH



Clarity (Epic Reporting Database)

UC Davis Patient Reporting (OMOP v.5.3)

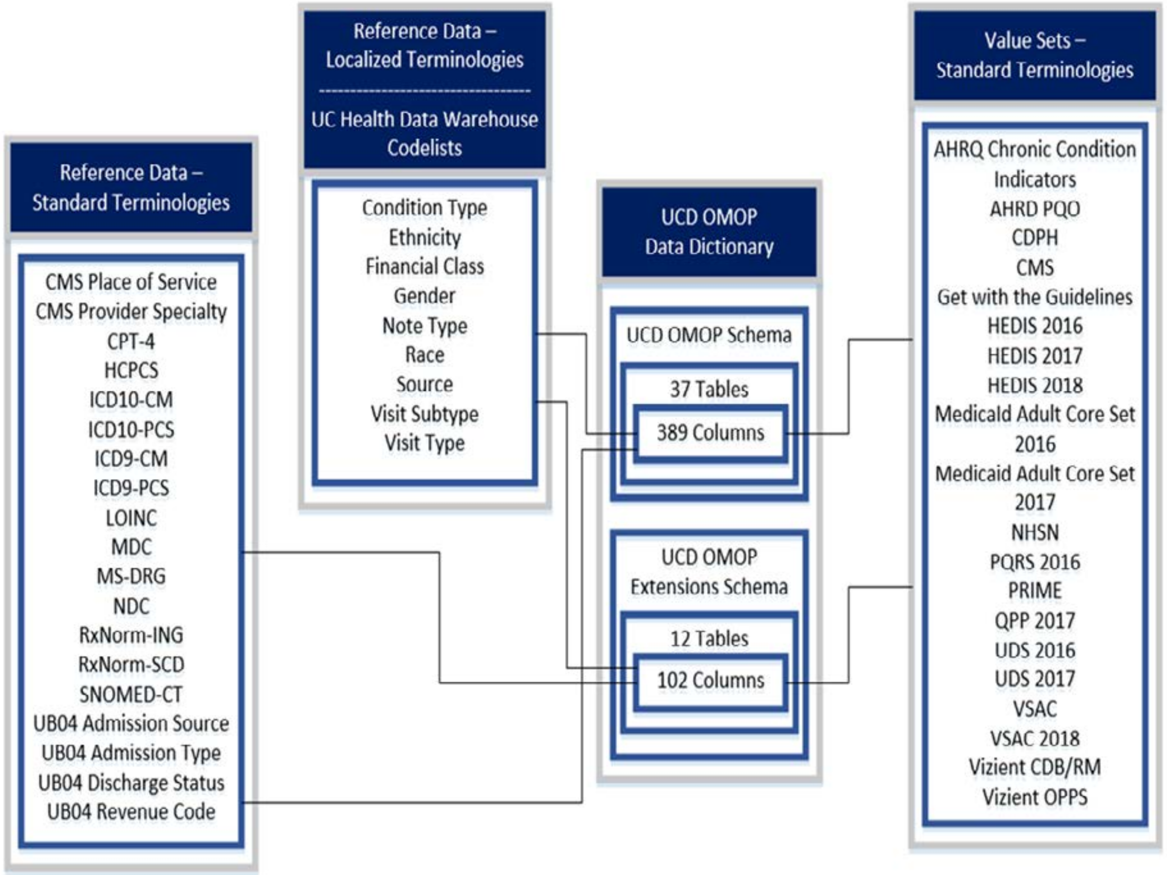
UC Health Data Warehouse (OMOP v.5.3)


UC Irvine Health
Patient Demographics
Patient Volume by Zip Code
Totals
4,518,028 Total UC Health Patients

Dashboards & Reporting

The Mapping Magic: Local Code to Standard Vocabularies

- Encounter types, departments
  - Map to Visit Types
- MEDICATION_IDs
  - Map to RxNorm
- Lab COMPONENT_IDs
  - Map to LOINC
- Condition ICD9/10
  - Map to SNOMED
- And many more mappings!



Reference Data – Standard Terminologies

CMS Place of Service
CMS Provider Specialty
CPT-4
HCPCS
ICD10-CM
ICD10-PCS
ICD9-CM
ICD9-PCS
LOINC
MDC
MS-DRG
NDC
RxNorm-ING
RxNorm-SCD
SNOMED-CT
UB04 Admission Source
UB04 Admission Type
UB04 Discharge Status
UB04 Revenue Code

Reference Data – Localized Terminologies

UC Health Data Warehouse Codelists

Condition Type
Ethnicity
Financial Class
Gender
Note Type
Race
Source
Visit Subtype
Visit Type

UCD OMOP Data Dictionary

UCD OMOP Schema
37 Tables
389 Columns

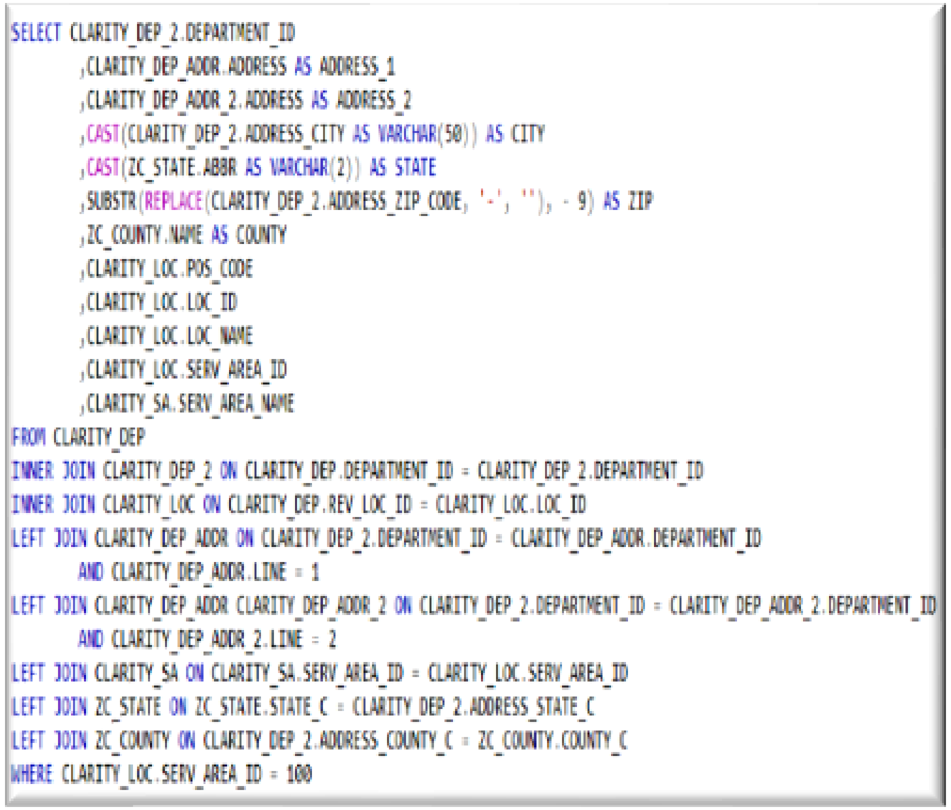
UCD OMOP Extensions Schema
12 Tables
102 Columns

Value Sets – Standard Terminologies

AHRQ Chronic Condition Indicators
AHRD PQO
CDPH
CMS
Get with the Guidelines
HEDIS 2016
HEDIS 2017
HEDIS 2018
Medicaid Adult Core Set 2016
Medicaid Adult Core Set 2017
NHSN
PQRS 2016
PRIME
QPP 2017
UDS 2016
UDS 2017
VSAC
VSAC 2018
Vizient CDB/RM
Vizient OPPS

Extract, Transform, Load (ETL)

- Using SQL programming language (and additional tools)
  - Extract specific data from Clarity (EPIC) relational databases into staging tables
  - Apply transformations to data and apply the mapping (local to standard) from the terminology database
  - Load data into CDM standard tables



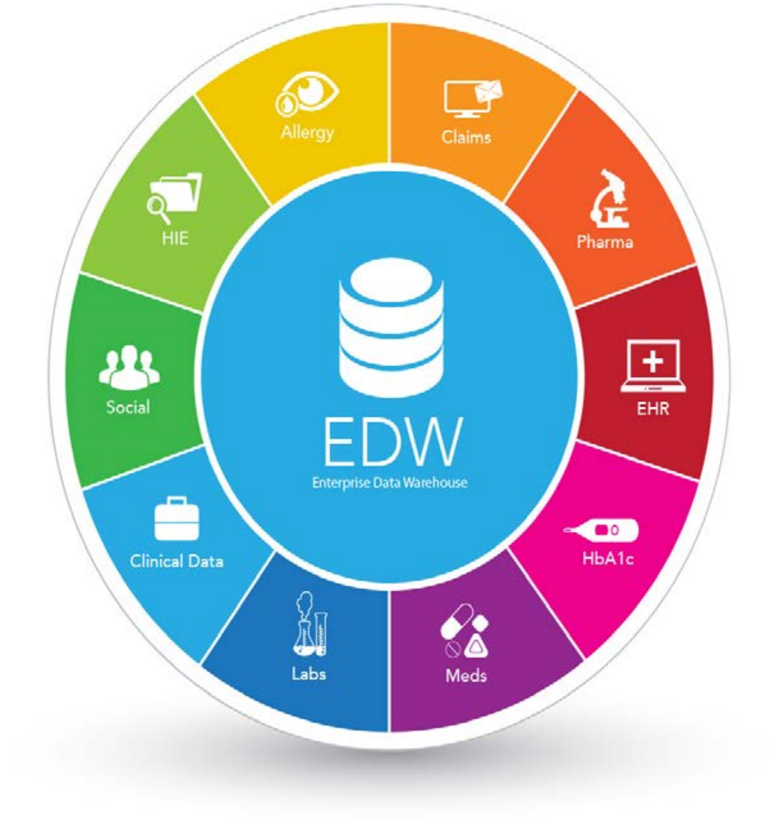
```sql
SELECT CLARITY_DEP_2.DEPARTMENT_ID
        ,CLARITY_DEP_ADDR.ADDRESS AS ADDRESS_1
        ,CLARITY_DEP_ADDR_2.ADDRESS AS ADDRESS_2
        ,CAST(CLARITY_DEP_2.ADDRESS_CITY AS VARCHAR(50)) AS CITY
        ,CAST(ZC_STATE.ABBR AS VARCHAR(2)) AS STATE
        ,SUBSTR(REPLACE(CLARITY_DEP_2.ADDRESS_ZIP_CODE, '-', ''), - 9) AS ZIP
        ,ZC_COUNTY.NAME AS COUNTY
        ,CLARITY_LOC.POS_CODE
        ,CLARITY_LOC.LOC_ID
        ,CLARITY_LOC.LOC_NAME
        ,CLARITY_LOC.SERV_AREA_ID
        ,CLARITY_SA.SERV_AREA_NAME
FROM CLARITY_DEP
INNER JOIN CLARITY_DEP_2 ON CLARITY_DEP.DEPARTMENT_ID = CLARITY_DEP_2.DEPARTMENT_ID
INNER JOIN CLARITY_LOC ON CLARITY_DEP.REV_LOC_ID = CLARITY_LOC.LOC_ID
LEFT JOIN CLARITY_DEP_ADDR ON CLARITY_DEP_2.DEPARTMENT_ID = CLARITY_DEP_ADDR.DEPARTMENT_ID
        AND CLARITY_DEP_ADDR.LINE = 1
LEFT JOIN CLARITY_DEP_ADDR CLARITY_DEP_ADDR_2 ON CLARITY_DEP_2.DEPARTMENT_ID = CLARITY_DEP_ADDR_2.DEPARTMENT_ID
        AND CLARITY_DEP_ADDR_2.LINE = 2
LEFT JOIN CLARITY_SA ON CLARITY_SA.SERV_AREA_ID = CLARITY_LOC.SERV_AREA_ID
LEFT JOIN ZC_STATE ON ZC_STATE.STATE_C = CLARITY_DEP_2.ADDRESS_STATE_C
LEFT JOIN ZC_COUNTY ON CLARITY_DEP_2.ADDRESS_COUNTY_C = ZC_COUNTY.COUNTY_C
WHERE CLARITY_LOC.SERV_AREA_ID = 100
```

What Types of Data are Included?

- Patients with an encounter after 1/1/2012
- Observations: Discrete data derived from patient interaction
  - Mapped to SNOMED, otherwise LOINC
- Conditions: ICD9/10-CM
- Procedures: ICD9/10-PCS; CPTII/CPT4/HCPCS
  - Health Maintenance, Orders, and Referrals mapped to SNOMED
- Drug exposure: Medications mapped to RxNorm SCD
- Measurements: Discrete data derived from equipment (e.g., labs, scales, thermometer)
- What patients are excluded?:
  - Patients from Marshal Medical Center (MMC) – a health system in Placerville, CA that has partnered with UCD

How Much Data?

- Over 5 million patients seen since 2012
  - 600,000+ of these patients are primary care patients
- Treated by nearly 100,000 healthcare providers
- Over 100 million encounters
- Over 300 million procedures
- More than 250 million medication orders
- Over 1 billion vital signs measurements and test results
- Claims data from self-funded plans now included
- Continually harmonizing elements



How Do UCs Collaborate?

- Workload is Shared
  - Development
    - Quality metrics – development distributed to teams (e.g., UCLA develops 5 of 20 QIP metrics)
  - Validation
    - Each team validates a sub-set of metric algorithms that were developed by other teams
- Code Repository – GitHub
  - All project code is shared on Github
- Effective technical discussions using “Slack” channels
- Meetings
  - Weekly business/clinical and CDM meetings
- Knowledge sharing
  - Open sharing of knowledge and problems

What Projects Use Standardized Databases?

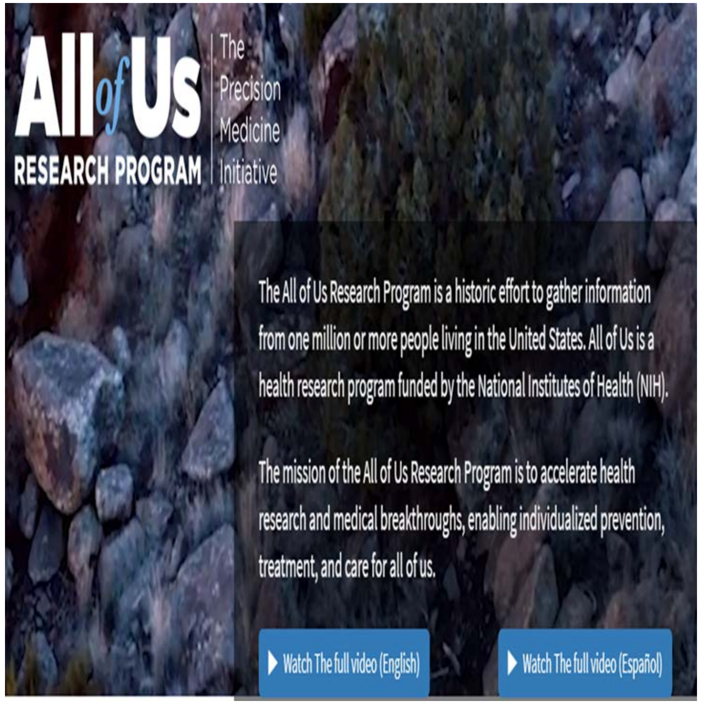
- **Quality improvement and P4P**
  - Quality Incentive Program (QIP)
  - Medicare Shared Saving Program (MSSP)
  - O35 cancer quality measure

Department of Health Care Services

DHCS oversees comprehensive health care, including mental health and drug treatment programs, for about 13 million Californians—nearly one-third of the state's population. Our largest program is Medi-Cal, the backbone of California's health-care safety net.

Medicare Shared Savings Program
Better Care for **Individuals.**
Better Health for **Populations.**
Lowering Growth in **Expenditures.**

- **Research**
  - All of Us "Precision medicine"



The All of Us Research Program is a historic effort to gather information from one million or more people living in the United States. All of Us is a health research program funded by the National Institutes of Health (NIH).

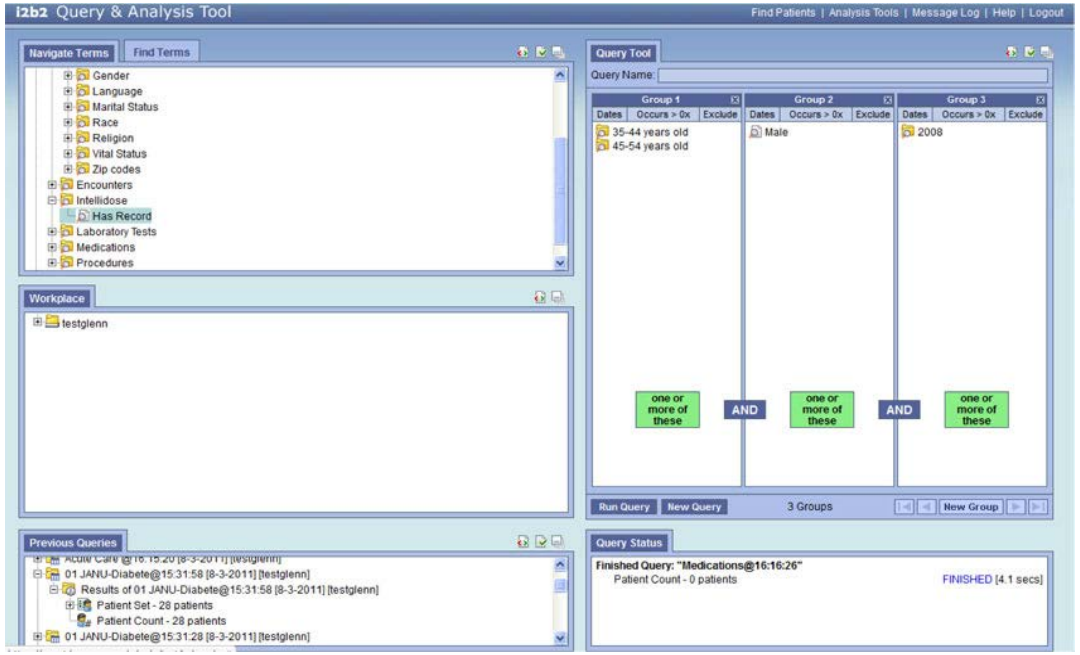
The mission of the All of Us Research Program is to accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us.

Watch The full video (English)
Watch The full video (Español)

Cohort Discovery for Research: i2b2 and “Data Explorer

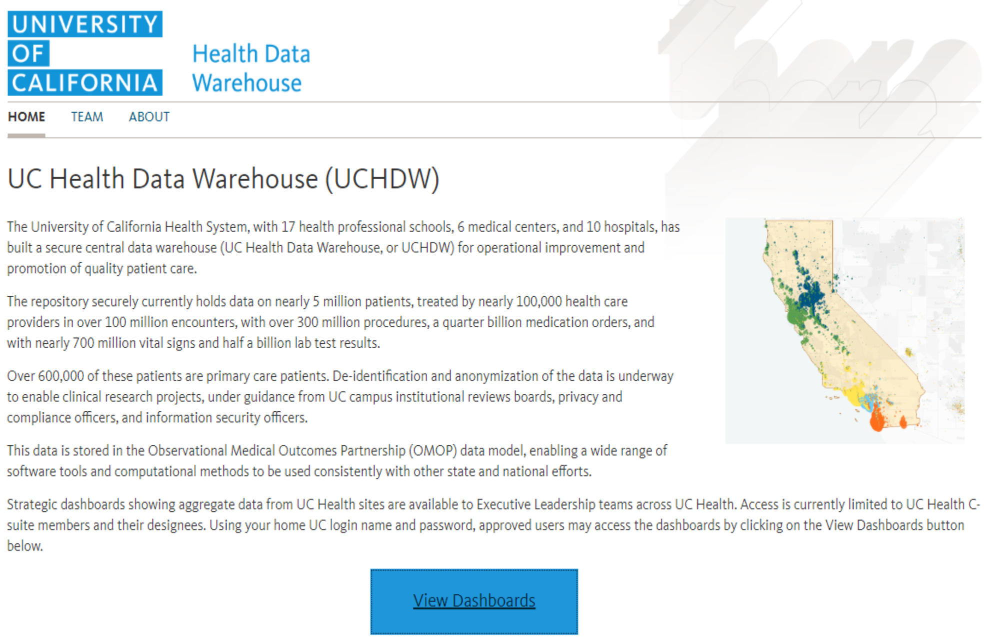
- Clinical researchers at UCD have used a cohort discovery tool known as i2b2
  - Identify counts of patients based on clinical traits. Enough patients for analysis?

- UC Health has developed similar tool called “Data Explorer”
  - UCD is developing this for local use



Acute Care@16:15:20 [8-3-2011] [testglenn]
01 JANU-Diabete@15:31:58 [8-3-2011] [testglenn]
Results of 01 JANU-Diabete@15:31:58 [8-3-2011] [testglenn]
Patient Set - 28 patients
Patient Count - 28 patients
01 JANU-Diabete@15:31:28 [8-3-2011] [testglenn]

Finished Query: "Medications@16:16:26"
Patient Count - 0 patients
FINISHED [4.1 secs]



UC Health Data Warehouse (UCHDW)

The University of California Health System, with 17 health professional schools, 6 medical centers, and 10 hospitals, has built a secure central data warehouse (UC Health Data Warehouse, or UCHDW) for operational improvement and promotion of quality patient care.

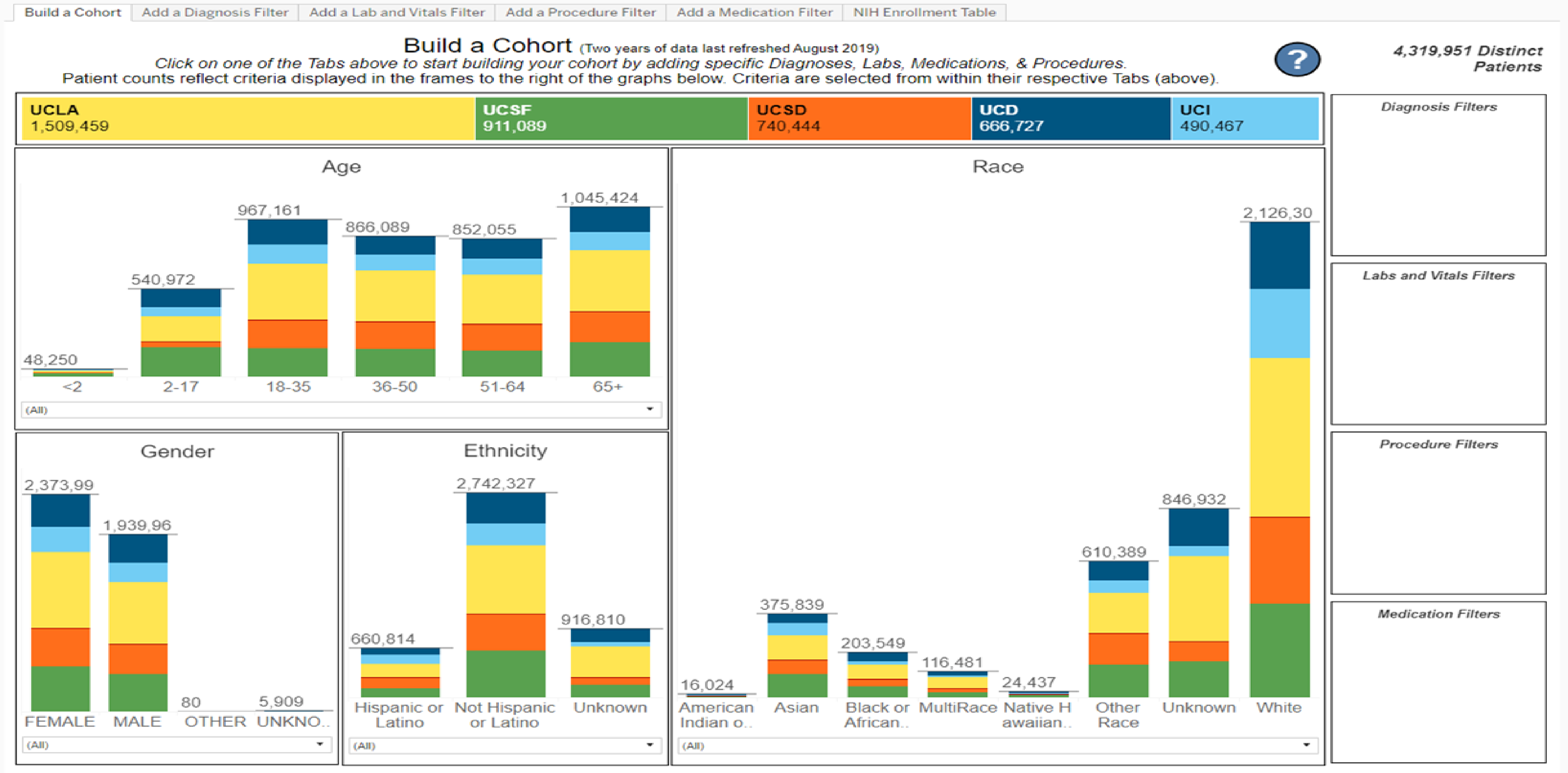
The repository securely currently holds data on nearly 5 million patients, treated by nearly 100,000 health care providers in over 100 million encounters, with over 300 million procedures, a quarter billion medication orders, and with nearly 700 million vital signs and half a billion lab test results.

Over 600,000 of these patients are primary care patients. De-identification and anonymization of the data is underway to enable clinical research projects, under guidance from UC campus institutional reviews boards, privacy and compliance officers, and information security officers.

This data is stored in the Observational Medical Outcomes Partnership (OMOP) data model, enabling a wide range of software tools and computational methods to be used consistently with other state and national efforts.

Strategic dashboards showing aggregate data from UC Health sites are available to Executive Leadership teams across UC Health. Access is currently limited to UC Health C-suite members and their designees. Using your home UC login name and password, approved users may access the dashboards by clicking on the View Dashboards button below.

All Patients -- (Last Two Years of Data)



Build a Cohort (Two years of data last refreshed August 2019)

Click on one of the Tabs above to start building your cohort by adding specific Diagnoses, Labs, Medications, & Procedures.

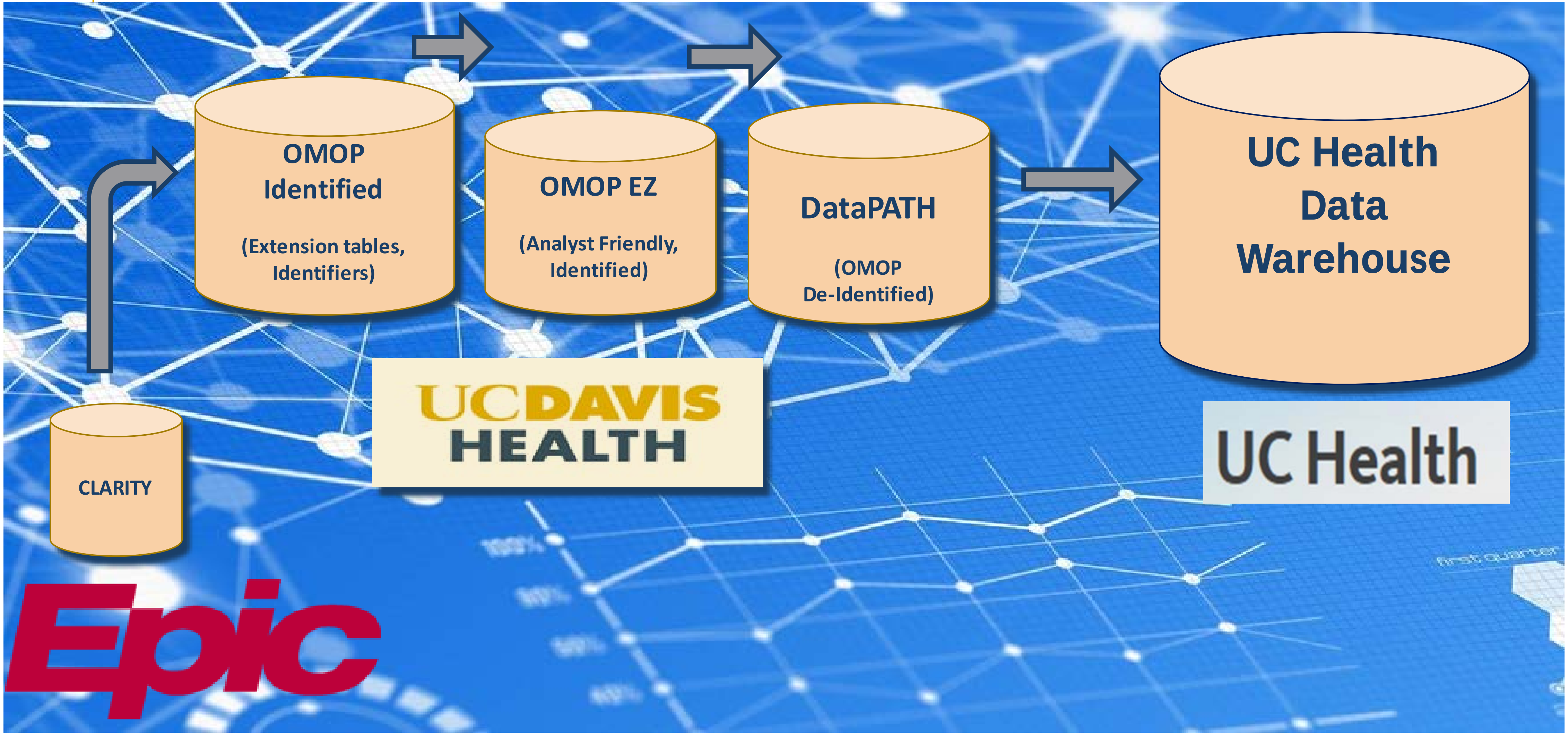
Patient counts reflect criteria displayed in the frames to the right of the graphs below. Criteria are selected from within their respective Tabs (above).

4,319,951 Distinct Patients

DataPATH: Standardized De-Identified Clinical Data at UC Davis

What is DataPATH?

Multiple Databases and Processes



What Concept Vocabularies are in the Identified Tables?

- **CONDITION_OCCURRENCE**

– ICD10CM:	61,711,261
– ICD9CM:	38,720,288
– SNOMED:	469,855
– None:	290

- **COST**

– Currency:	64,476,314

- **DEVICE EXPOSURE**

– None:	21,690,403

- **DRUG_EXPOSURE**

– RxNorm:	78,331,807
– None:	9,672,108

- **MEASUREMENT**

– LOINC:	322,921,653

- **NOTE**

– None:	2,573,941

- **OBSERVATION**

– SNOMED:	72,134,862
– LOINC:	6,109

- **PROCEDURE_OCCURRENCE**

– CPT4:	38,310,233
– SNOMED:	22,306,550
– HCPCS:	18,649,248
– None:	4,474,582
– CVX:	1,303,861
– ICD10PCS:	336,359
– ICD9Proc:	142,560

How is DataPATH De-Identified?

- **DataPATH**: legally de-identified database
  - Application of algorithms and/or data (e.g., voter data) could result in re-identification
  - Users must adhere to rules for keeping data secure (e.g., data cannot be exported)
- Mapping tables used to obfuscate the primary keys as found in the source database
  - Original primary key and a new primary key randomly assigned

    The last step of the ETL process is to delete the data in the mapping tables
- Dates are <u>offset</u> with additional column in Person mapping table.
  - Column stores an offset for all dates associated with the patient
  - Column value is used to offset all dates with a value randomly assigned between -365 to 365 (exclusion of zero)
  - All dates associated with the patient are <u>offset by the same value</u>
- “Source value” columns that hold data as found in the source database are either set to NULL or a value that can not be used to re-identify a patient

Data Quality and Validation

What Data Validation Efforts are Underway?

Are the Data “Good Enough”

Current Validation Efforts

- **ETL Validation**: Are data in target database correctly represented in the source databases?
  - (e.g., compare counts between Clarity and OMOP)
- **Content Validation**: What data issues are present in the source data (Clarity) as a result of how data were input/collected
  - Changing workflows
  - Data entry errors by clinical staff
- **CDM Conformance**: Do our tables, fields, and values conform to constraints imposed by data model and data coding standards?

Mapping Validation: Local to Standard Mappings are Correct?

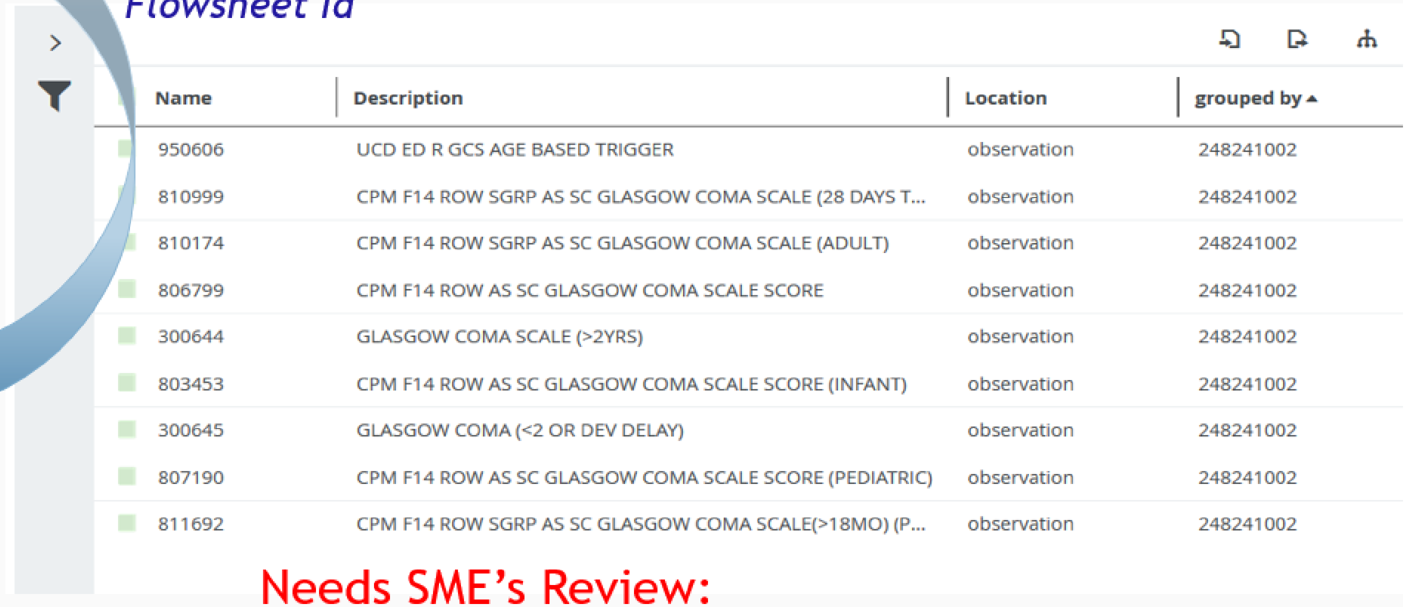
Glasgow Comma Score

Local Component IDs

Standard Concepts

UCD EPIC Flowsheet Id

SNOMED-CT Code



Name	Description	Location	grouped by ▲
950606	UCD ED R GCS AGE BASED TRIGGER	observation	248241002
810999	CPM F14 ROW SGRP AS SC GLASGOW COMA SCALE (28 DAYS T...	observation	248241002
810174	CPM F14 ROW SGRP AS SC GLASGOW COMA SCALE (ADULT)	observation	248241002
806799	CPM F14 ROW AS SC GLASGOW COMA SCALE SCORE	observation	248241002
300644	GLASGOW COMA SCALE (>2YRS)	observation	248241002
803453	CPM F14 ROW AS SC GLASGOW COMA SCALE SCORE (INFANT)	observation	248241002
300645	GLASGOW COMA (<2 OR DEV DELAY)	observation	248241002
807190	CPM F14 ROW AS SC GLASGOW COMA SCALE SCORE (PEDIATRIC)	observation	248241002
811692	CPM F14 ROW SGRP AS SC GLASGOW COMA SCALE(>18MO) (P...	observation	248241002

Needs SME's Review:

- Is mapping correct?

Are the Data Good Enough?

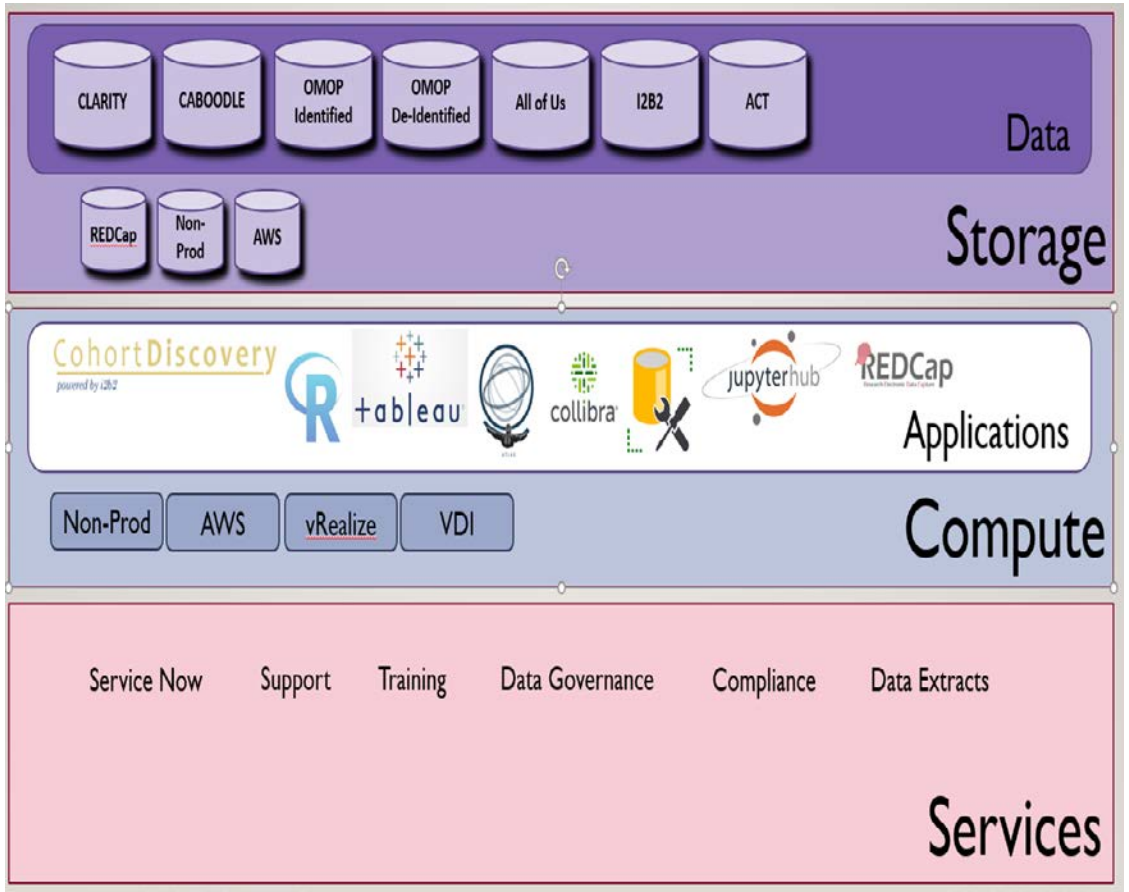
- Reliable research requires high-quality data
- Quality improvement projects at UCs show that data are generally well represented in the CDM
  - Errors found during QA processes lead to continuous quality improvement of data
  - What gets used will be improved !
- Researchers can start with cohort identification and preliminary analyses
  - If research proves promising, or data quality appears to be an issue, we can validate data against source data (Hyperspace and Clarity)

Accessing the DataPATH Data

How Do Researchers Get Access to the De-Identified Data?

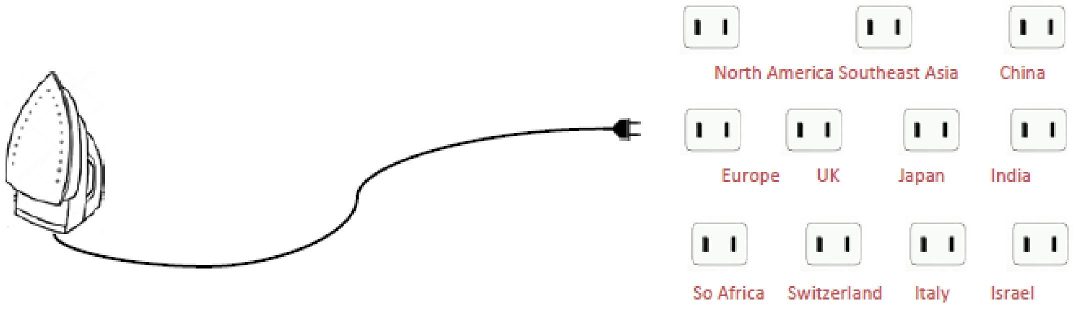
Where Are the Data? What Analytical Tools Available?

- Researchers can analyze data within a secure compute environment behind firewall
  - Data cannot be extracted or copied!
- Storage
  - DataPATH and other databases
- Applications -- Jupyter, Tableau, etc.
  - Phase I – SQL database access
  - Phase II - What do researchers want?
- Services – access, training, support

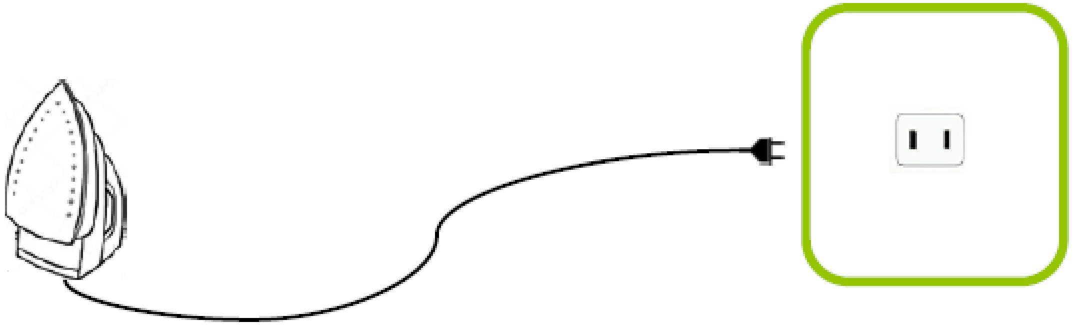


Analytics Can Be...

Remote

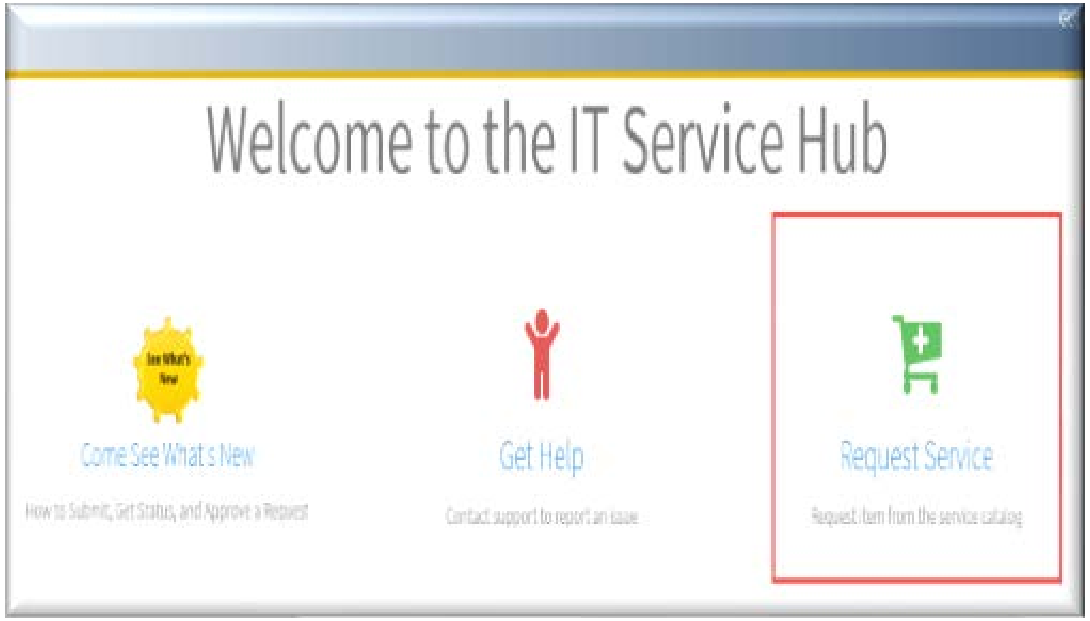


Safety done behind firewall



How Do I Request Access? ServiceNow

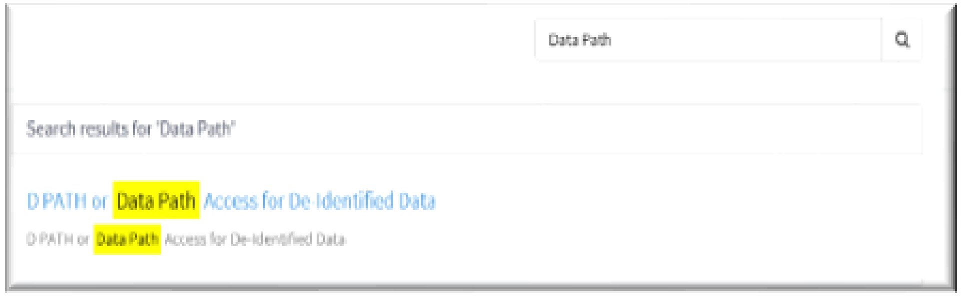
- Go to ServiceNow
  - https://ucdh.service-now.com/itss
- Click, “Request Service”



Welcome to the IT Service Hub

See What's New
Come See What's New
How to Submit, Get Status, and Approve a Request
Get Help
Contact support to report an issue
Request Service
Request item from the service catalog

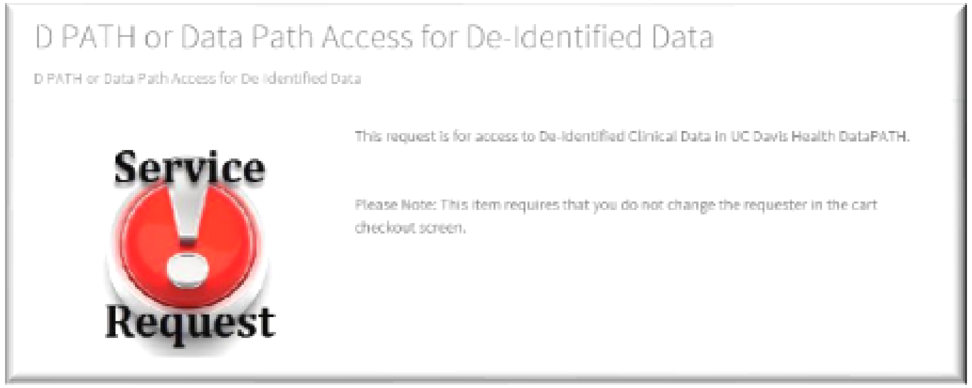
- Search
  - **“D PATH”** or ”**Data Path**”



Search results for 'Data Path'

D PATH or Data Path Access for De Identified Data

D PATH or Data Path Access for De-Identified Data



D PATH or Data Path Access for De-Identified Data

D PATH or Data Path Access for De Identified Data

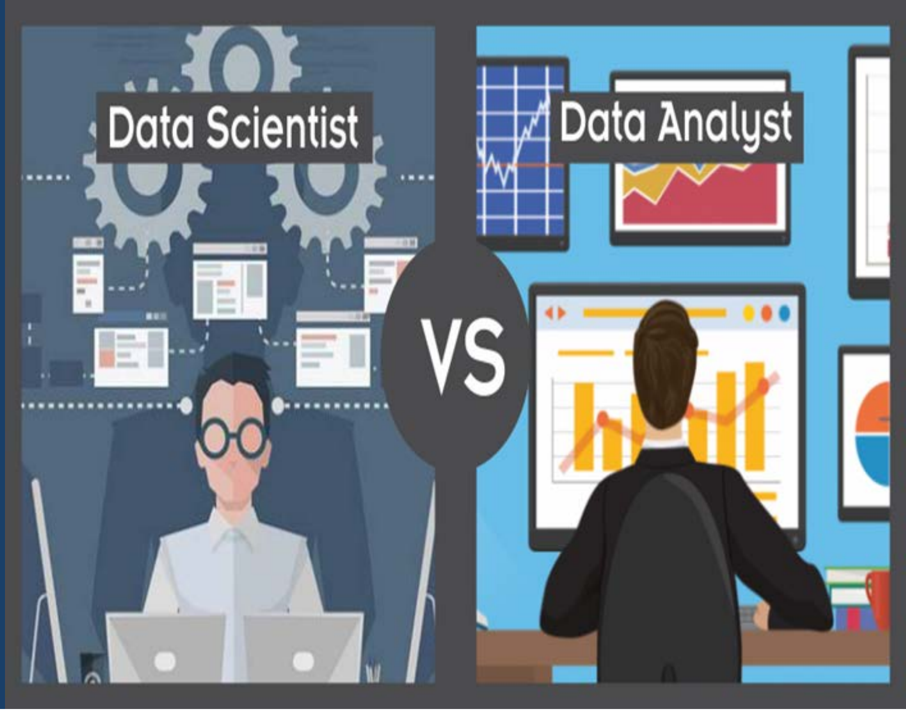
This request is for access to De-Identified Clinical Data in UC Davis Health DataPATH.

Please Note: This item requires that you do not change the requester in the cart checkout screen.

The Data Journey

- Standardized clinical databases can help researchers more efficiently create knowledge
- The Data Journey
  - Metaphor of “path” important because researchers will be on an increasingly complex journey to learn and access more detailed data
- Researchers are encouraged to implement analyses in the following order:
  1. Training database (CMS Synthetic data)
  2. De-Identified database (**DataPATH)**
  3. UC Davis Identified Database
  4. UC Health Data Warehouse

Interactive Session



Data Use Agreement

Working Together to Execute SQL and Python Code

Data Use Agreement

The data have been de-identified, and we can explore the data together *within* the secure Jupyter environment

- Your login will expire at the end of the session

All Attendees Must Agree:

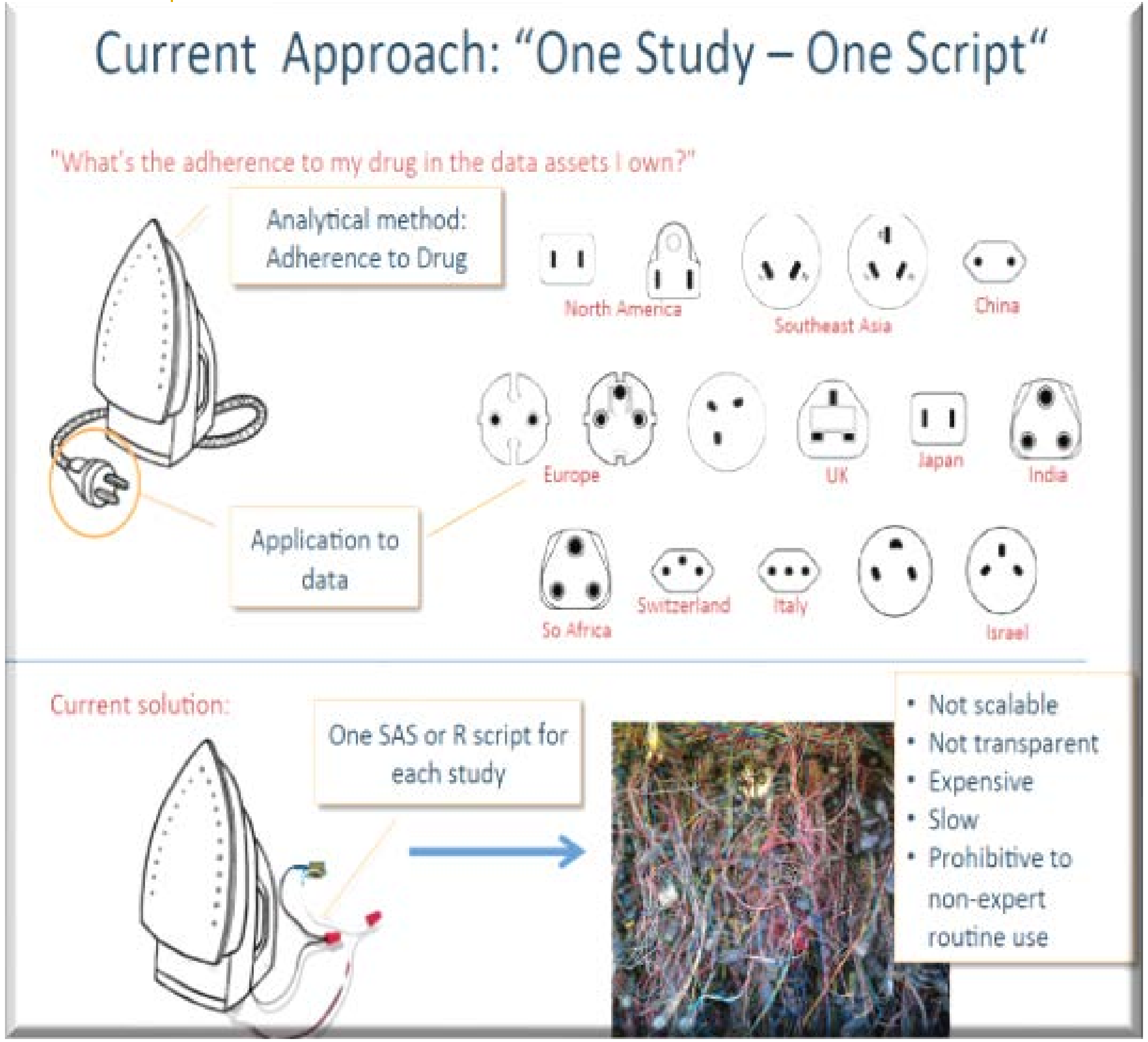
- To run the pre-created queries, or follow instructions to modify queries
- **To NOT extract or copy data onto personal machines**

Types of Analyses We Will Review

- Data Dictionary -- available online in HTML format
  - Access data dictionary and meta-data (Folder within Jupyter)
  - Review simple summary statistics about the data
- Concepts and Vocabularies
  - Explore how medical ontologies are represented in OMOP
  - OMOP concept model great for computers, more work for humans to understand
- Characterization
  - A Tableau dashboard shown
  - Use SQL and Python to characterize the data
- Complex Cohorts
  - Illustrates steps to create cohorts using SQL
  - *We may not have time to review—examples can be shared later with researchers*
- Can I Modify Queries?
  - No, at least not until the end of the session
  - A large query on larger tables could crash our server!

Appendixes – Extra Backup Slides

Standardizing Science



Current Approach: “One Study – One Script“

"What's the adherence to my drug in the data assets I own?"

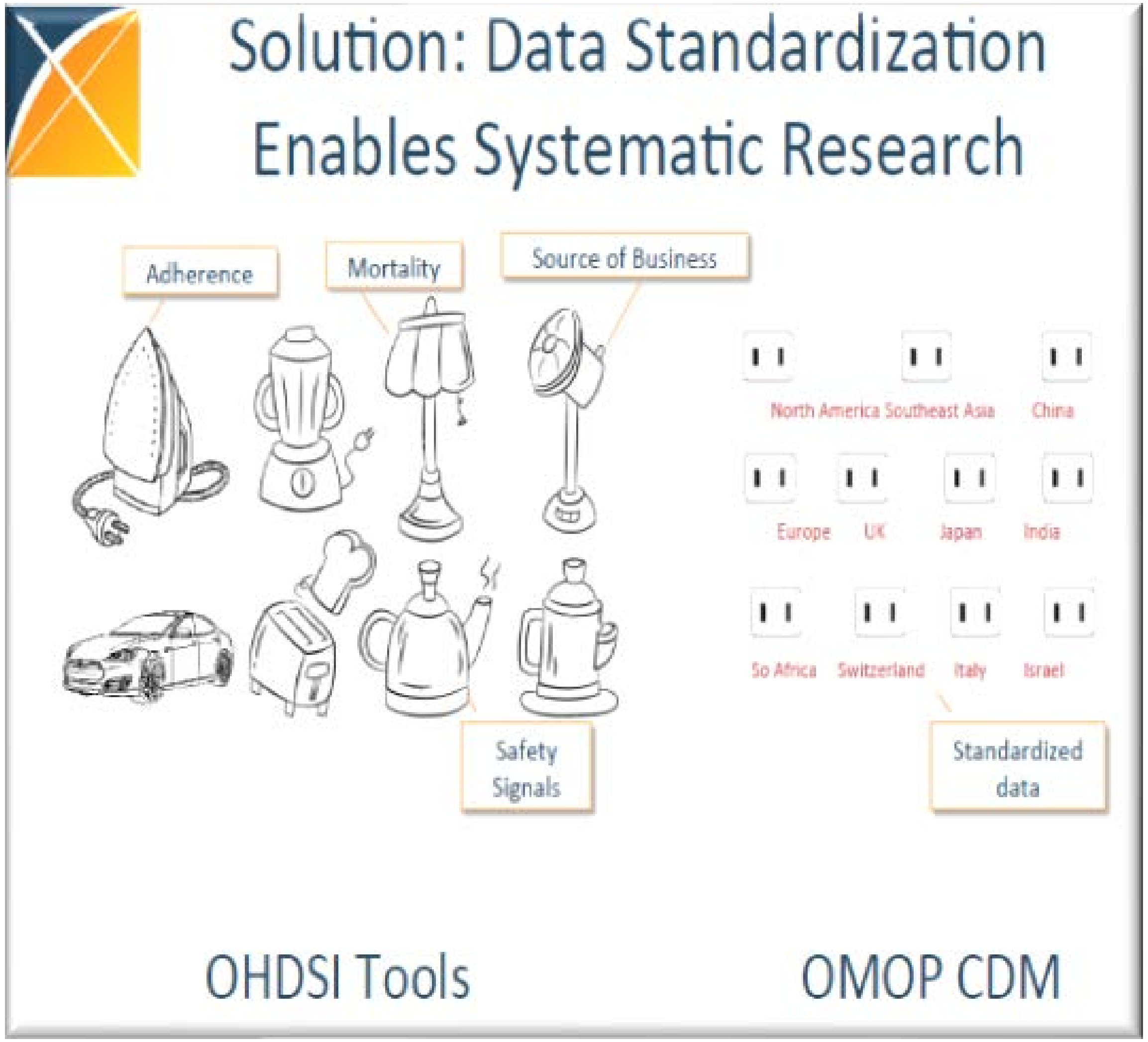
Analytical method: Adherence to Drug

Application to data

Current solution:

One SAS or R script for each study

- Not scalable
- Not transparent
- Expensive
- Slow
- Prohibitive to non-expert routine use



Solution: Data Standardization Enables Systematic Research

Adherence

Mortality

Source of Business

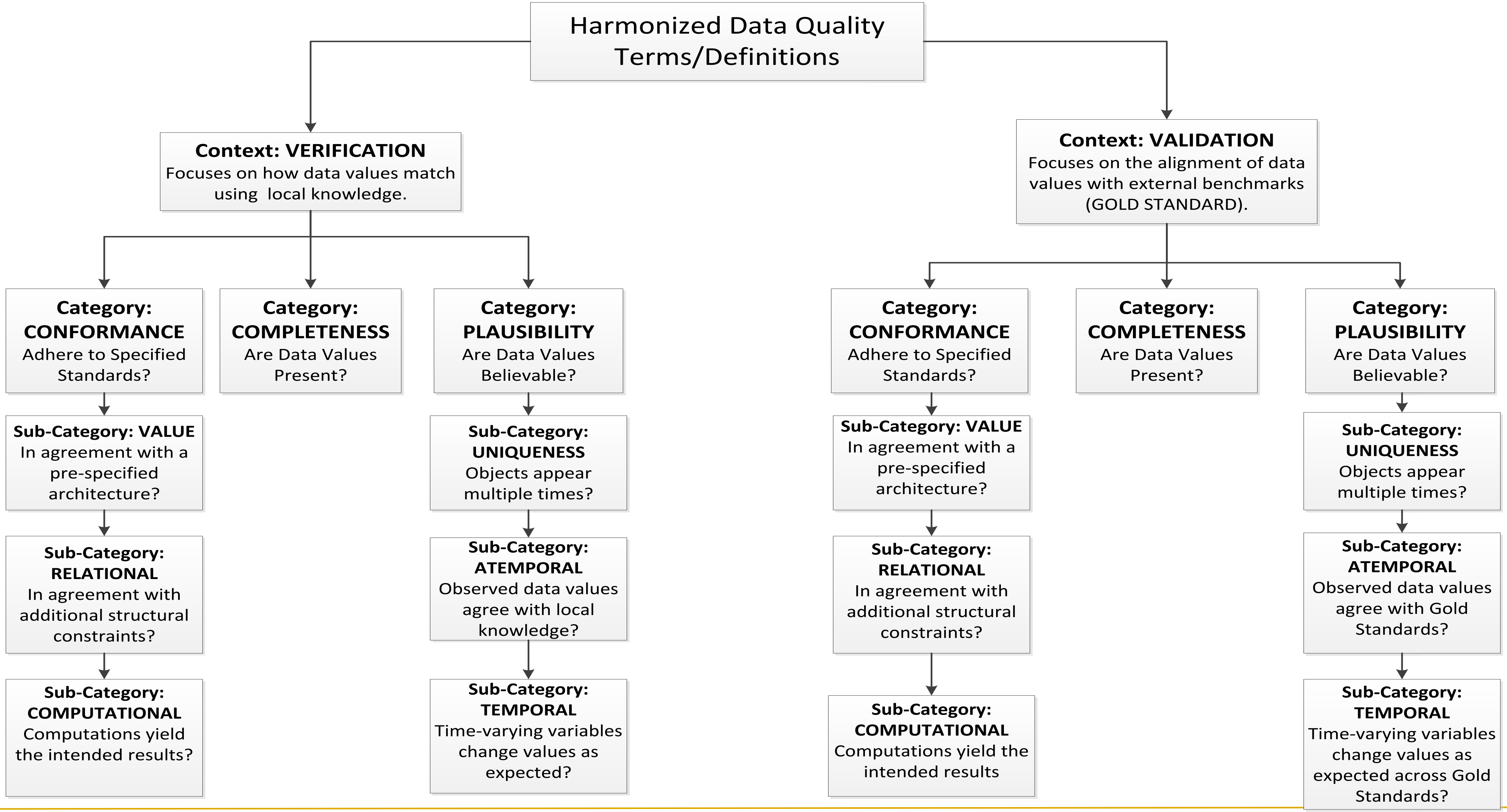
Safety Signals

Standardized data

OHDSI Tools

OMOP CDM

Data Quality Terminology



Harmonized Data Quality Terms/Definitions

Context: VERIFICATION
Focuses on how data values match using local knowledge.

- **Category: CONFORMANCE** Adhere to Specified Standards?
  - **Sub-Category: VALUE** In agreement with a pre-specified architecture?
  - **Sub-Category: RELATIONAL** In agreement with additional structural constraints?
  - **Sub-Category: COMPUTATIONAL** Computations yield the intended results?
- **Category: COMPLETENESS** Are Data Values Present?
- **Category: PLAUSIBILITY** Are Data Values Believable?
  - **Sub-Category: UNIQUENESS** Objects appear multiple times?
  - **Sub-Category: ATEMPORAL** Observed data values agree with local knowledge?
  - **Sub-Category: TEMPORAL** Time-varying variables change values as expected?

Context: VALIDATION
Focuses on the alignment of data values with external benchmarks (GOLD STANDARD).

- **Category: CONFORMANCE** Adhere to Specified Standards?
  - **Sub-Category: VALUE** In agreement with a pre-specified architecture?
  - **Sub-Category: RELATIONAL** In agreement with additional structural constraints?
  - **Sub-Category: COMPUTATIONAL** Computations yield the intended results
- **Category: COMPLETENESS** Are Data Values Present?
- **Category: PLAUSIBILITY** Are Data Values Believable?
  - **Sub-Category: UNIQUENESS** Objects appear multiple times?
  - **Sub-Category: ATEMPORAL** Observed data values agree with Gold Standards?
  - **Sub-Category: TEMPORAL** Time-varying variables change values as expected across Gold Standards?

What Concept Domains are in the Identified Tables?

CONDITION_OCCURRENCE

Condition:	82,495,062
Observation:	9,459,075
Procedure:	7,954,723
Measurement:	992,544
Metadata:	290

COST

Currency:	64,476,314

DEVICE EXPOSURE

Metadata:	21,690,403

DRUG_EXPOSURE

Drug:	78,331,807
Metadata:	9,672,108

MEASUREMENT

Measurement:	322,917,495
Observation:	4,158

NOTE

Metadata:	2,573,941

OBSERVATION

Procedure:	36,525,491
Measurement:	27,519,140
Observation:	8,085,711
Condition:	10,629

PROCEDURE_OCCURRENCE

Procedure:	39,769,996
Drug:	18,508,223
Measurement:	15,861,487
Observation:	5,547,303
Metadata:	4,474,582
Device:	1,361,802