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!

- **OHDSI 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
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

STANDARDS
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
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 – OMOP Version 6
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
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 MEDICATION_IDs can vary
  - Flowsheets are created by local clinical workflows—different IDs are created for similar concepts
- Data harmonization and analytics requires standard codes
EPI C Query Tools -- Slicer/Dicer

- Standardized databases developed by UCs are not the only available databases for analytics
- EPI C 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.
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!
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

Extract Transform Load
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

- **Research**
  - All of Us “Precision medicine”
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.
All Patients -- (Last Two Years of Data)
DataPATH: Standardized De-Identified Clinical Data at UC Davis

What is DataPATH?
Multiple Databases and Processes

OMOP Identified
(Extension tables, Identifiers)

OMOP EZ
(Analyst Friendly, Identified)

DataPATH
(OMOP De-Identified)

UC Health Data Warehouse

CLARITY

UC Health

Epic
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 offset 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 offset by the same value

- “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

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

North America  Southeast Asia  China
Europe  UK  Japan  India
So Africa  Switzerland  Italy  Israel
How Do I Request Access?  ServiceNow

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

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

OHDSI Tools

OMOP CDM

North America • Southeast Asia • China

Europe • UK • Japan • India

So Africa • Switzerland • Italy • Israel

Standardized data
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?
    - Sub-Category: VALUE
      In agreement with a pre-specified architecture?
    - Sub-Category: RELATIONAL
      In agreement with additional structural constraints?

- Category: PLAUSIBILITY
  Are Data Values Believable?
    - 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?

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?
    - Sub-Category: VALUE
      In agreement with a pre-specified architecture?
    - Sub-Category: RELATIONAL
      In agreement with additional structural constraints?

- Category: PLAUSIBILITY
  Are Data Values Believable?
    - 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?
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