EPI 231 - Lecture 1

Intro to Databases and UCSF Data

Anobel Odisho, MD MPH

Course Objectives

- Explain the differences between a relational database model and data warehouse model as it pertains to EHR data
- Demonstrate knowledge of medical vocabularies and ontologies and their use in analyzing data in the EHR.
- Construct a patient cohort based on structured data, such as diagnosis codes, encounters, and procedures.
- Summarize demographic and clinical characteristics for a cohort from an analytic file.
- Extract relevant clinical data for a specified patient cohort, including labs, orders, medications, and history.
- Formulate a clinical research question and create a standard analytic file ready for analysis.

Labs

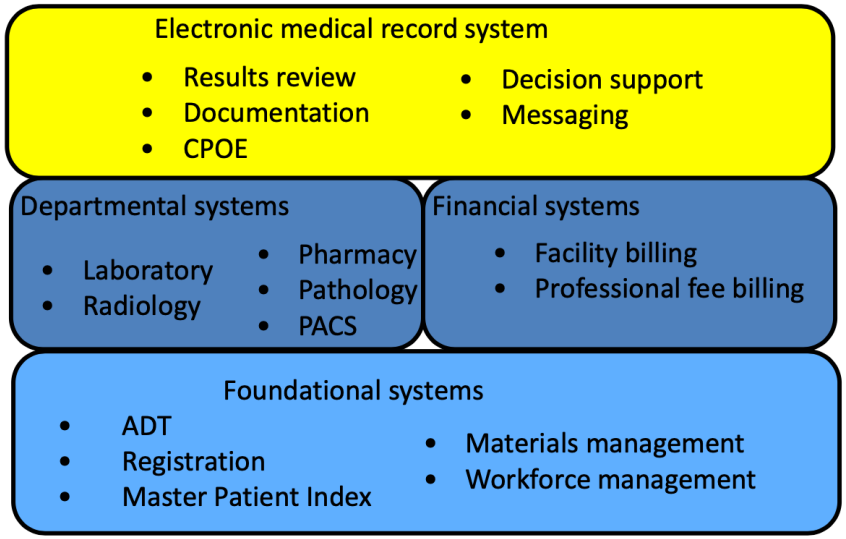
- The first two labs are designed to help you get access to all of the resources necessary for the class, access data, and execute simple queries
- Each lab after that will focus on a specific type of clinical data (labs, medications, etc) _ Each lab will have two general sections.
  1. A set of pre-defined exercises you will have to complete
  2. Work on your own project using what you have learned in this lab
- By the end of the course, you will be able to create your own research cohort, extract basic demographic and clinical data, prepare your data for analysis, and perform bivariate analyses.

Lecture Outline

- Overview of the EHR
- Databases: Important Terms and Concepts
- Clinical Data Warehouses
- The UCSF Data Landscape

What Does the EHR Do?

- Initially, EHR served to support billing, insurance claims, and clinical care (probably in that order of importance)
- Scope has increased



Departmental Systems

- Laboratory: Clinical pathology, Anatomic pathology, Blood bank
- Radiology: RIS, PACS
- Pharmacy: Unit dose, Retail
- Cardiology: ECG, Echo, Cath, PACS
- Dietary
- Neurology: EMG, NCV
- Pulmonary: PFT
- GI endoscopy

EHR Functionality

- Message box (proprietary names vary but functionality similar)
- Results review (lab, path, imaging, notes)
- Documentation (direct entry, structured/unstructured, dictation, mixed)
- Order management
- Patient summary displays
- Medication administration record / Bar code admin
- Patient lists, schedule, rounding/handoff tools
- Patient monitoring review
- Quality metrics, dashboards
- Billing, Professional fee, Facility fee
- Administrative
- Electronic communication (with team, with patients)

Database Terms

- **Table**: made of rows and columns
- **Relational Database**: multiple tables, linked together using common columns
- **Data Warehouse**: a database that is designed for analysis and incorporates data from many disparate databases.
- **CDW**: Clinical Data Warehouse
- **ETL**: Extract, Transform, Load: the process of taking data from one source, transforming it in some way, and then storing it in another database
- **SQL**: Structured Query Language

ETL: Extract, Transform, Load

Process of taking data from one database, making changes to the data, and storing it in another data source

- Extract: data collected from data sources, held in temporary storage. Validation rules test whether data has expected values
- Transform: Data processed to make values and structure consistent across all data. Date formatting, resorting rows or columns, joining data from two values into one or splitting data from one value into two. Goal of transformation is to make all the data conform to a uniform schema
- Load: Copying data to the new location (data warehouse)

Structured Data

- Codes (CPT, Procedure, ICD-9/10)/Billing
- Labs
- Orders
- Events
- Metadata
- Digital Dust
- Flowsheets
- Smart Date Elements
- Manual Abstraction

Unstructured Data

- Clinical Notes
- Radiology Images, Reports
- Pathology Images, Reports
- PROs
- Photos, Videos
- Email, Phone Calls, MyChart Messages
- Scanned Documents

Database vs Data Warehouse

	Database	Data Warehouse
Definition	Collection of data organized for storage, accessibility, and retrieval	Integrates copies of transaction data from disparate source systems for analytical use
Data Sources	Usually a single application	From multiple applications and sources
Purpose	Optimized for transactions	Optimized for analysis
Uptime	Requires very high uptime	Can tolerate downtime for data loads/maintenance
Data Organization	Fully normalized, to minimize how often data is repeated, so often requires many	Ok to have denormalized data, allowing for fewer joins and more

CDW Characteristics

- **Subject-oriented data**: Data are stored by subjects, not by applications. Subjects: patient, provider, modality (lab)
- **Integrated data**: Data related to a subject come from several operational systems (multiple lab systems).
- **Nonvolatile data**: Data cannot update in real time; they are read-only, and the users cannot manipulate them.
- **Time-variant data**: Data are stored historically, meaning there are histories of events of over time, for example membership in a Registry.
- **Data granularity**: High level of detail (hour, minutes, seconds) allowing different aggregations.

Karimi, Health Care Manager 2017

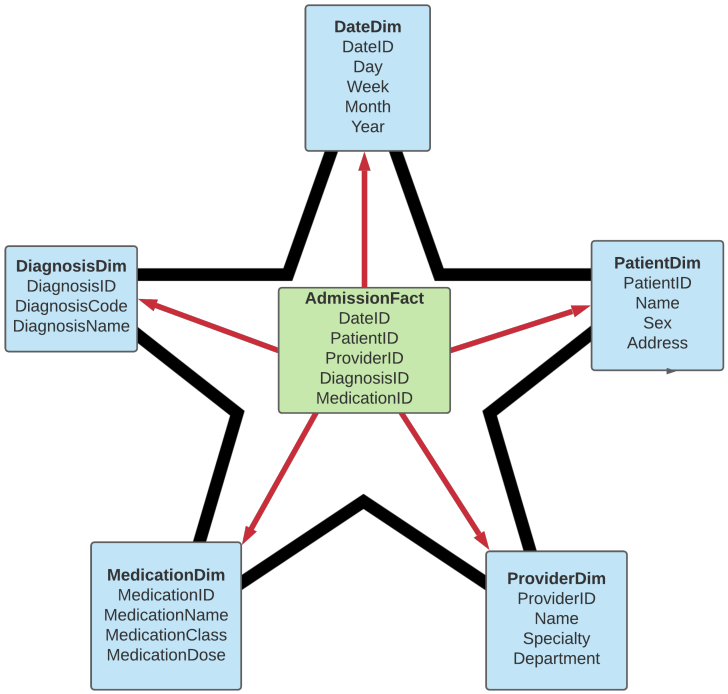
CDW Functions

- **Data extraction**: Data is extracted from various systems (such as PACS for radiology, LIS for labs, EHR, etc)
- **Data cleansing**: Correction of conflicts, duplication, and sometimes providing defaults values for missing data
- **Data standardization**: Standardization of data types and field length for the same data elements from multiple sources, and semantic standardization
- **Data summarization**: Purging source data that are not useful and sorting and merging of data in the data staging area. When transformation ends, a collection of integrated, cleaned, standardized, and summarized data is ready to load into CDW.
- **Data storage**: This is the end product that you see and want to use!

Karimi, Health Care Manager 2017

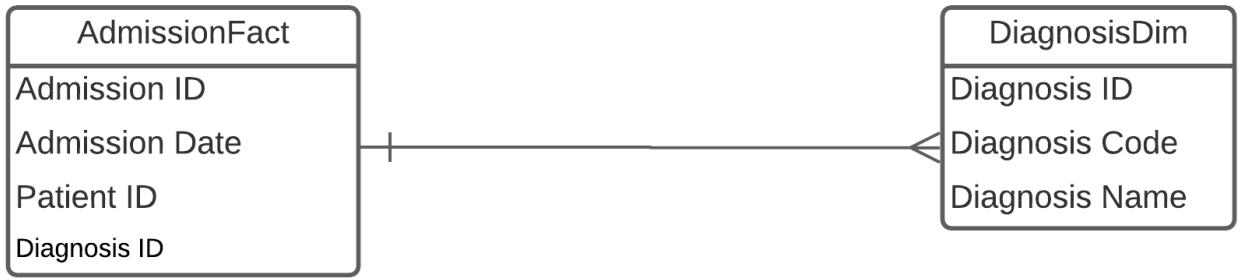
CDW Star Schema: Dimensions and Facts

- Fact: a value, or measurement, which represents an observation/event
- Dimension: provides structured labeling information and context to facts



CDW: Bridge Tables

Bridge Tables “bridge” the gap between tables of different levels of granularity - Some tables have 1 row per event (Admission) - Some tables have 1 row per value (diagnosis) - This works well for 1:1 relationship

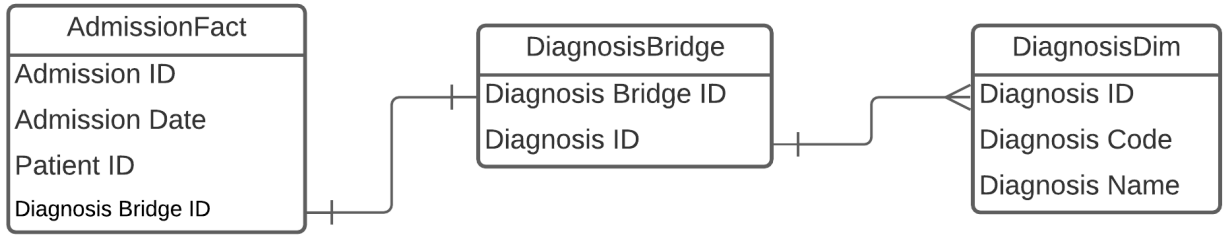


AdmissionFact			
Admission ID	Admission Date	Patient ID	Diagnosis ID
1	1/1/2021	4	1

DiagnosisDim		
Diagnosis ID	Diagnosis Code	Diagnosis Name
1	C61	Prostate
2	J44.9	COPD

CDW: Bridge Tables

What happens when you have multiple diagnoses for a single admission?



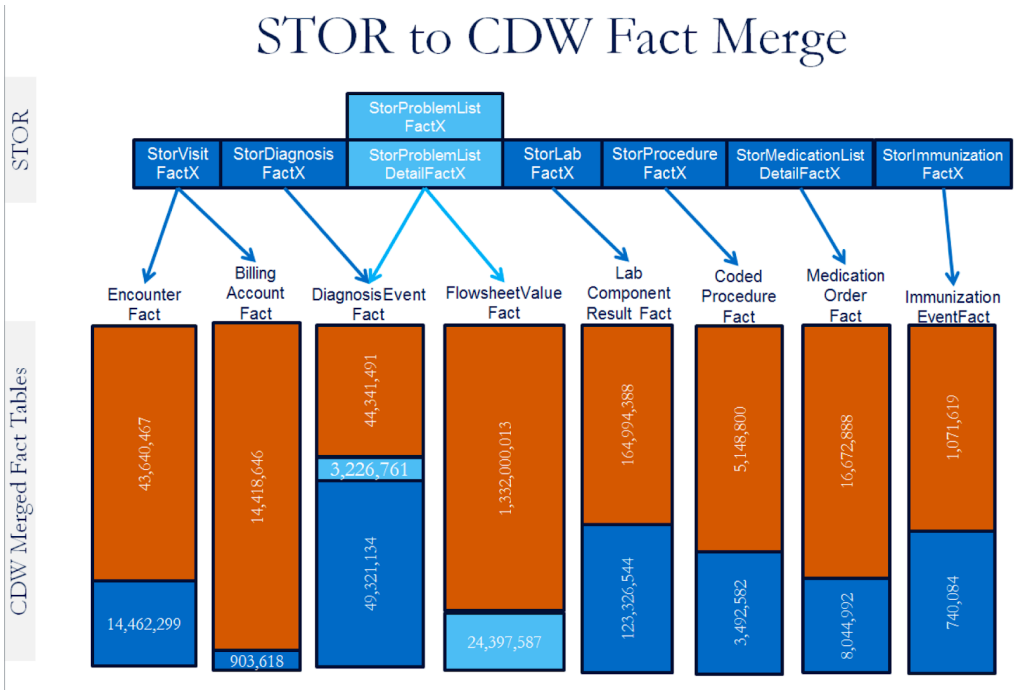
AdmissionFact			
Admission ID	Admission Date	Patient ID	Diagnosis Bridge ID
1	1/1/2021	4	1

DiagnosisBridge	
Diagnosis Bridge ID	Diagnosis ID
1	1
1	2

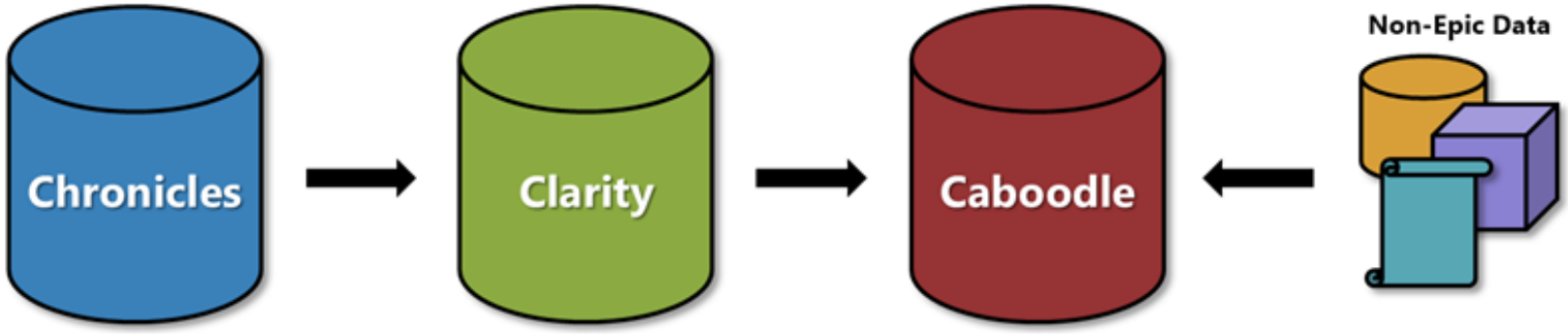
DiagnosisDim		
Diagnosis ID	Diagnosis Code	Diagnosis Name
1	C61	Prostate
2	J44.9	COPD

UCSF Core Clinical Databases

- STOR (Summary Time Oriented Record): primary EHR at UCSF 1988-2005
- UCare: Primary EHR at UCSF 2005-2012

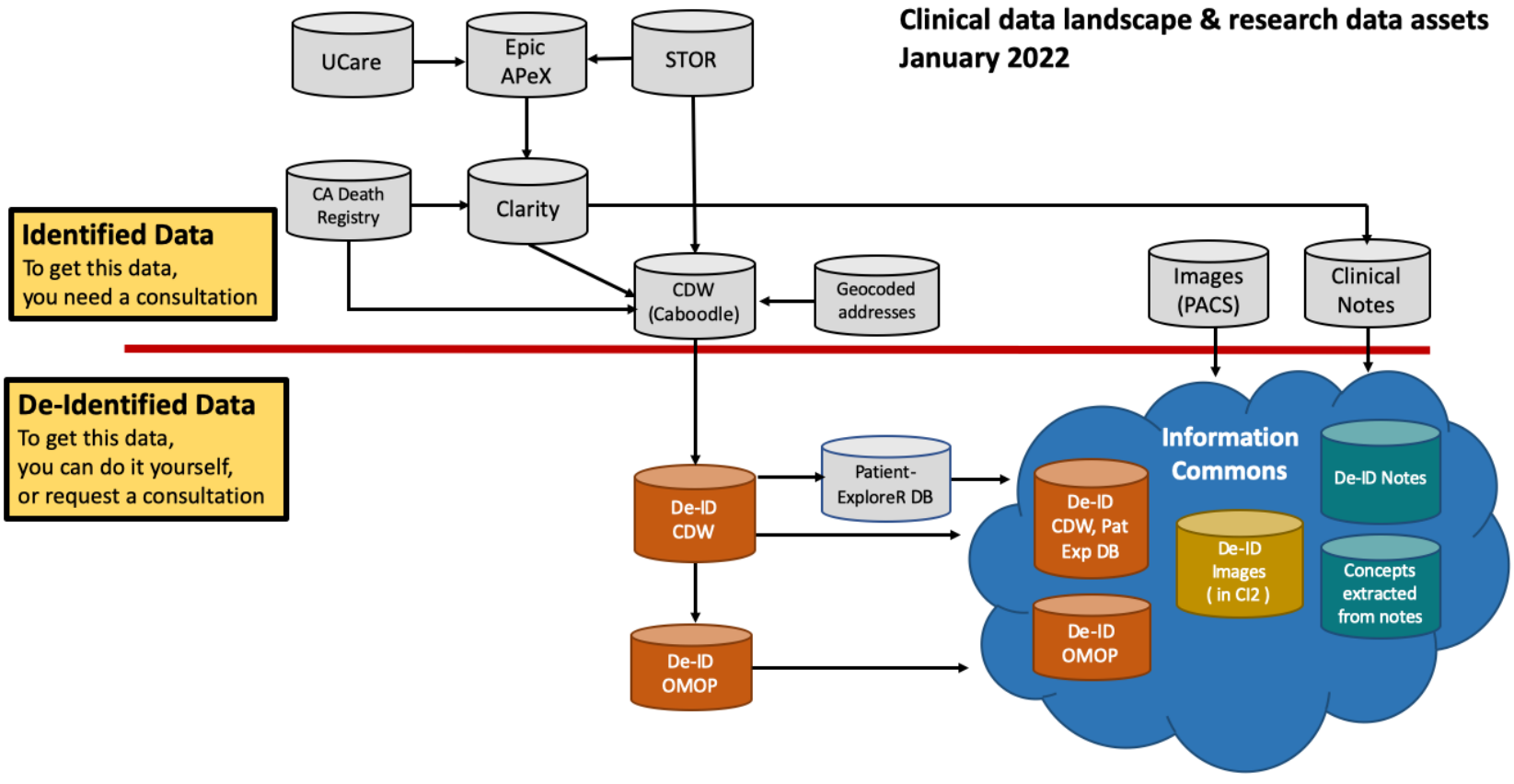


Current UCSF EHR Data Sources



- These are just the brand names given to these databases by the vendor
- Chronicles is a hierarchical database, underpins the day-to-day EHR.
- Every night most of Chronicles is copied over to a relational SQL database designed for reporting (Clarity) through an ETL process (~24k tables)
- Some data from Clarity is then copied into a data warehouse, called Caboodle

UCSF Data Landscape



De-ID DW

- implementation of EPIC Caboodle
- 18 PHI Identifiers removed
- No IRB required
- No images, clinical notes
- As of Jan, 2021
  - 5,147,226 distinct patients
  - 3,556,008 distinct patients with encounters
  - 102,330,846 encounters
  - 79,085,969 from Epic
  - 12,264,327 from Stor
  - 8,613,726 derived from billing

De-identification Process

Deid Type	Description
show	The data element is not PHI, and can be displayed
hide	The data is PHI and will be hidden (patient name, SS#)
key	The data is a value that is unique to a patient, and is to be replaced with a randomly generated surrogate key. The map between the two is maintain for future reidentification with IRB approval
date	Dates will be offset by a random number of days from 0 to 364
agekey	Age will be adjusted so that patients over 90 years old will appear to be 90 years old on the date of the build
birthdate	Birth dates will be adjusted so that patients over 90 years old will appear to be 90 years old on the date of the build.
zipcode	Zip codes of patients will be limited to 3 digits in the deidentified views unless the total population of the zip codes with those 3 digits is < 20,000, in which case, the zip code 000 will be provided.
varvalue	see next page

Conclusion

- Understanding the how the EHR is used will help you understand where the data come from and how reliable they are
- The CDW has a different structure from a traditional database to support analytic use
- Understanding data provenance and types will help you design your study, as not all studies can be done with EHR data
- UCSF has a complex data infrastructure