Distribution Agreement

____

Signature of Student Date

Towards a Common Data Model: An Evaluation of Standard Healthcare Data Models

By

Shailesh Nair

Degree to be awarded: MPH

Executive MPH

Jeff Weaver, MBA Committee Chair	Date	3/29/2018
Andre Bosman, BS Committee Member	Date	3/29/2018
Dr. Laurie Gaydos, PhD Associate Chair for Academic Affairs, Executive MPH	Date	

Towards a Common Data Model:

An Evaluation of Standardized Healthcare

Reporting Data Models

Shailesh Nair

Master's Degree in Public Health

APHI

Rollins School of Public Health

Introduction

Over the past several years, we are witnessing the widespread adoption of electronic health records systems by hospitals and clinics to streamline, improve and enhance patient-centered care. The heightened emphasis on coupling patient care with the reuse and synthesis of clinical and administrative data, demands a robust informatics infrastructure. Today health data varies significantly among healthcare organizations. These variations in data are common despite the existence of the standard of care models for most diseases. This difference raises the specter of the difficulty in standardizing centralized data repositories that will support data analytics, decision support systems and evaluation of interventions versus outcomes.

All software applications and data processing frameworks whether at information ingestion stage or during reporting and analytics are built on an information storage substrate. The growing need for data standardization, agile application development, and cost-saving potential of crowd-sourced solutions adoption is disrupting the healthcare analytics landscape and spurring institutions to explore consensus data models.

The Emory University's Library and Information Technology Services (LITS) department's mission to support health care providers, researchers and academic health sciences through the design, development, and adoption of information technology solutions involve extracting, compiling and transforming data from disparate sources. LITS believes that the most significant benefits of data standardization will only be achieved with the adoption and use of a single standardized data model across the entire enterprise analytics ecosystem. Thus streamlining the solutions and platforms within Emory University. To this end, the Library and Information Technology Services (LITS) department is assessing a flexible, scalable and standardized data model that will support clinical research, health registry reporting, data analytics and clinical decision support systems.

This thesis will evaluate two popular common data models optimized for information retrieval that constitute the core of today's analytics, dashboards, machine learning and artificial intelligence algorithms. These models were created to support a standards-based ecosystem that enables development of analytics platforms with efficiency, reusability and reproducibility.

What is a Data Model?

Data modeling is the process of determining how data are to be stored in a database. A data model defines characteristics of the information domain and the relationships between them. A data model specifies the data types contained within the entity (e.g., number, date, character), constraints imposed on the attributes of the entity (e.g., uniqueness, null or missing values allowed or not, list of valid values). It also defines the links between the data contained in entities, hierarchical relationships, parent-child data relationships. The structure and the metadata of a data model design affect the way the data is stored and accessed. A data model design is crucial in determining how quickly the information is queryable and extractable from the data repository. The architectural components of a data model are typically conveyed schematically in diagrams that use symbols and notations to denote the features and relationships among data items. A widely used format for representing a data model is the Entity-Relationship diagram (ERD). These diagrams (Fig.1) show each entity with its component attribute items and their data types (numeric, date character) and size. Each of the entity is expected to have a Primary Key (PK) or a Foreign Key (FK). ERD typically do not contain the metadata (descriptions or data dictionary) which are usually stored as a separate document and linked to the ER Diagram.

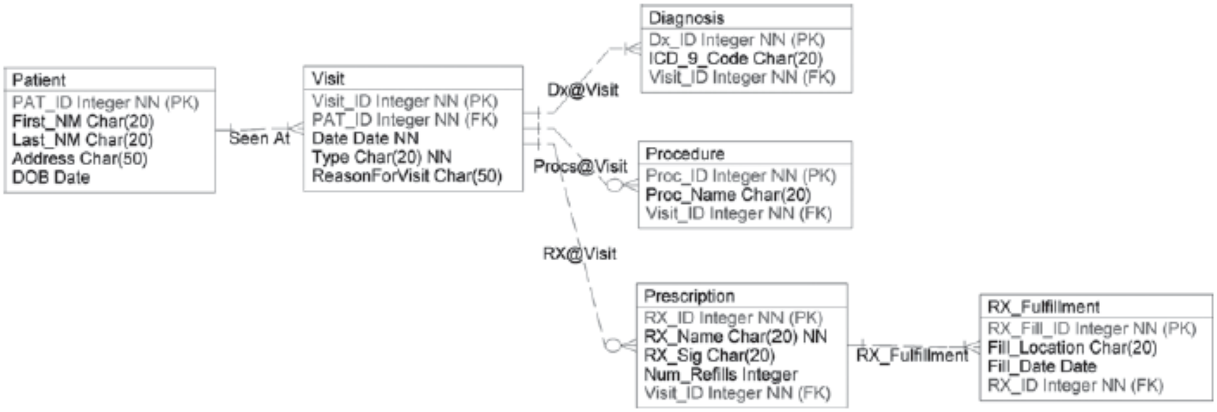


Fig 1. Example of Entity Relationship Diagram (ERD)

An ERD is used by multidisciplinary teams typically consisting of informatics professionals, clinical analysts, and biostatisticians to assess a data model’s ability to meet their respective information needs.

Data Model standardization is fraught with significant challenges to meet the broad and flexible requirements to accommodate diverse data sources and data types. In a healthcare institution, these sources include electronic health records, patient billing systems, laboratory information management system, claims processing systems, anatomic pathology applications. Within an organization, some of these information systems may belong to the same vendor, but it is seldom the case where all the source system share the same data model. In order to sustain enterprise-wide analytics, research and decision support, combining related data from disparate sources with different data structures, variable formats, definitions, and quality are crucial. This process is called data integration.

Data integration requires rigorous attention to how the data from different systems will be rendered in the data model. How differences in data definitions, procedures, and sources will be resolved to ensure compatibility and comparability of the resulting data values. Database architects design the data integration architecture, and data modelers develop of data models that house the aforementioned health information based on the requirements of the user community within the organization. These models customarily drive large enterprise data warehouses (EDW) and usually follow specific design principles that are optimized for fast retrieval. EDWs also include prebuilt aggregates known to be of interest to the user community for speeding up analysis, taking into account the business rules and data governance. At its core, a data model embodies the different assumptions made around the questions that are likely to be asked by the consumers of the information.

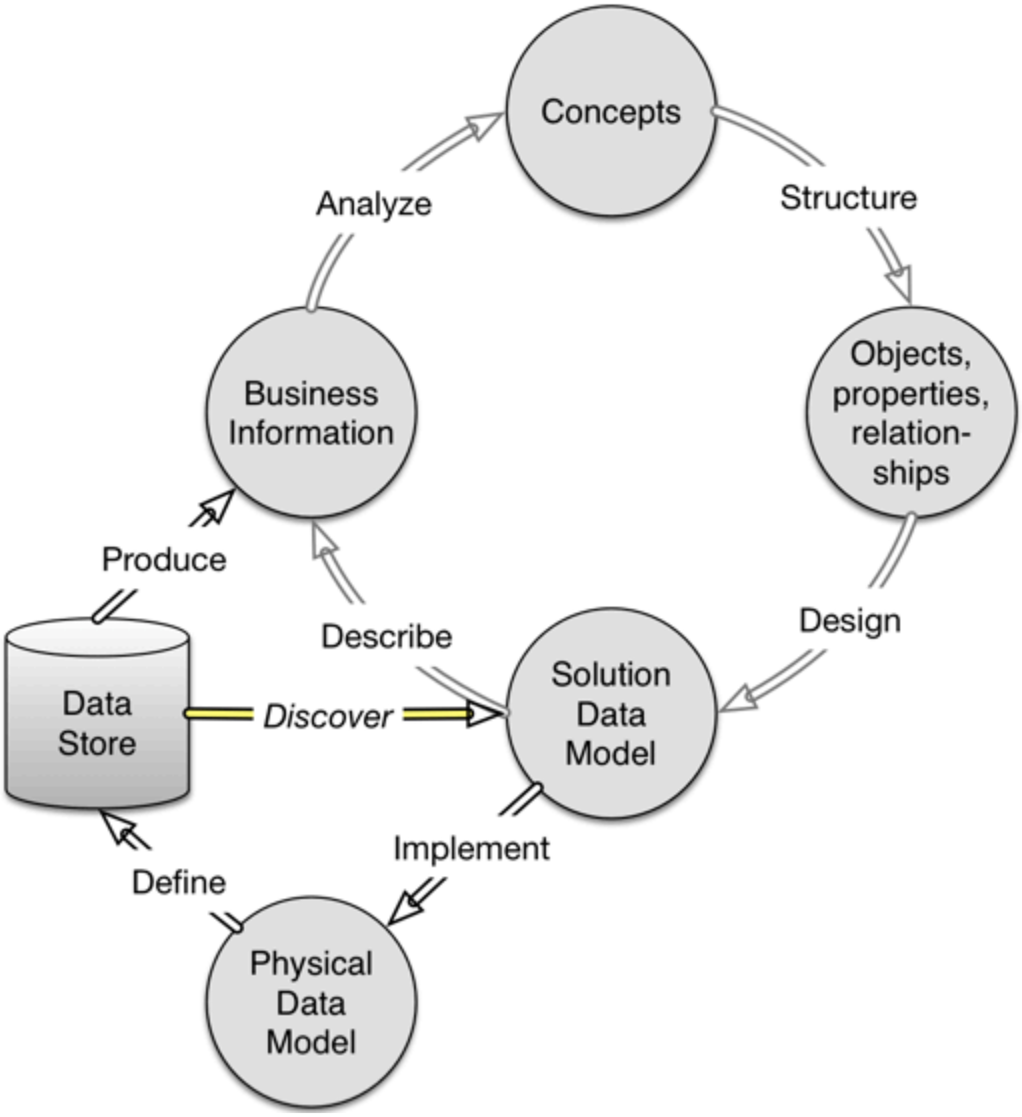


Fig. 2. Shows the data model build process from business information and requirements

http://www.dataversity.net/artificial-intelligence-vs-human-intelligence-hey-robo-data-modeler/

The above process of building a data model from the ground up within the organization on a project by project basis can be expensive and can result in inconsistencies and uneven quality. These ground realities factor into the decision whether to build-from-scratch or adopt a standardized model (Fig.2) to underpin the current and future information needs of Emory University.

Rationale

In the healthcare field, there is often a broad consensus on how to diagnose and treat an individual's disease. Typically patient care is delivered through an amalgam of disciplines be it in the clinic setting, emergency medicine or in the hospital. Each of these settings has their healthcare workflows and system that assist providers in the delivery of care. Hence the data models that drive each of these systems are as different as the disparate systems that support clinical care. Healthcare data analytics primarily rely on Electronic Health Records, Laboratory Information Management Systems (LIMS), administrative systems like health care finance and administration systems as well as databases that house information collected during research studies.

Given the heterogeneous information sources and the need for supporting multiple use-cases, healthcare organizations have responded by developing Clinical Data Warehouses (CDW) or Data Lakes. These Clinical Data warehouses are typically based on custom data models and a mix of standard and organization-specific terminologies. Data modeling constitutes a small proportion of the total systems development effort, yet its impact on the quality of the final software product is perhaps more significant than any other (Witt, et al., 2000).

The data model is one of the primary determinant of systems development cost, integration with other systems and flexibility of the system to meet as yet to be determined organizational needs. Empirical studies based on packaged software vendors have shown that "uncorrected errors become exponentially more costly with each phase in which they are unresolved" (Westland, 2002). This research suggests putting more efforts upstream in the systems development process to reduce defects and in many instances prevent them from occurring in the first place.

A majority of the reporting and analytics applications are built on custom-designed data models which is almost entirely dependent on the skill of the data modeling personnel. Thus the quality of the data model is prone to inconsistencies. Data modeling task is often the realm of data architects and modelers who are often in dedicated teams serving a multitude of projects. This often leads to different project teams defining the same data in different ways resulting in data redundancies and overlap between various application data models. Application development staff and Database Administrators may find the data model to be impractical, inflexible or complex at implementation time due to the complexity of the design. A data model may look impressive on paper but may need to re-working for the sake of application query performance especially in the case of data model designed for supporting analytics.

In the past decade, there is a national emphasis on data sharing especially information gleaned from research funded by agencies like the National Institutes of Health (NIH). Federal agencies like NIH has put greater emphasis on data sharing aiming to "expedite the expedite the translation of research results into knowledge, products, and procedures to improve human health" ("NIH Data Sharing Policy and Implementation Guidance," 2003).

NIH cites many reasons for this policy towards grants of five hundred thousand dollars or higher. Among them is to encourage open scientific inquiry, diversity of data analysis, facilitating the education of new researchers, enabling collaborative research which engenders creating new datasets by combining multi-institutional data as well as enabling exploring newer hypotheses not envisioned by the initial investigators. Sharing information via a standard data model decreases the burden of specifying what standards and best practices should be proposed and created before concrete information sharing (Curtis et al., 2012). Additionally, CDMs also ensure that the data is shared with fewer transformations needed thus reducing the administrative costs of research projects for partner institutions and time between data collection and release. NIH recognizes the timeliness of the information is valuable and expects data to be released in a timely fashion. Grantees are also expected to archive the research data for three years after grant agreement closeout. Data retention in standard data model format with appropriate version information can easily be brought online incase its needed.

Adoption of a Common Data Model opens the door for open source application development, and independent software vendors to build apps that can be used across institutions. Standardization creates marketplace opportunities like building middle-tier application services like software development kits (SDK) for instance, which in turn enables building low code applications. SDKs accelerates building robust, rich client apps leveraging the standard entities and their relationships that are at the core of a CDM [Diagram]. Once the enterprise health data is persisted in a CDM format, LITS will be empowered to tap into a growing ecosystem of tools and applications developed to act upon this data.

Need for a Common Data Model (CDM)

The absence of a consensus standard data model poses a significant barrier to the development and adoption of marketplace solutions that help healthcare enterprises towards a standard architecture.

A Common Data Model benefits research teams by enabling healthcare data management staff to provide standard data sets with robust clinical data that's easily de-identified. Today's research ideas translate to improving both future kinds of research and advanced clinical practice. Research scientists and investigators across the enterprises like Academic Medical Centers (AMC) benefit by building tools and services based on standard database schemas. Cohort identification and data extraction for research purposes is a typical, recurring activity that is significantly enabled when there are standardized tools available to empower researchers to interact with data iteratively. Scientific knowledge is often incremental. Each subsequent research paper builds on previous research findings published in scientific literature.

A substantial benefit of implementing a CDM is the necessary move towards standard vocabularies and coding terminologies like the Systematized Nomenclature of Medicine (SNOMED), Logical Observation Identifiers Names and Codes (LOINC), and National Drug Code (NDC). Mapping custom or proprietary source codes into standard vocabularies dramatically enables dataset for analytics and collaborative research work. Another critical advantage is multicenter research initiatives benefit considerably from sharing data in a standard format, which catalyzes data integration with limited transformations.

Improving population and public health is a data-driven endeavor. The Meaningful Use (MU) Stage 2 regulations added a new public health-related option for Eligible Professionals (EPs) to identify and report specific cases to specialized registries. MU Stage 2 regulations broadly define a specialized registry type as those which include congenital disabilities registries, chronic disease registries, and traumatic injury registries among others. These repositories are operated by patient safety, and quality improvement organizations with an intention to enable knowledge engendering or process enhancement concerning diagnosis, therapy, and prevention of health conditions at a population level ("Specialized Registry | Meaningful Use | CDC," 2018). Although there are no certifications and standards criteria specified by Office of the National Coordinator for Health Information Technology (ONC), the data submission formats represent an opportunity for reporting organizations to propose standardized data formats. The intent for MU program is to support interoperability, usability, and innovation through health IT through flexibility, development, and certification of health systems ("2015 Edition Health Information Technology (Health IT) Certification Criteria", 2015). Adoption of CDMs like OMOP signals a readiness and intent for meeting MU reporting milestones.

In a connected world where collaboration between institutions at various levels are expected, mandated and incentivized, a CDM provides the underlying standard substrate which instantiates the information models (IM) ("Information model," 2017). Information modeling is the essential first step towards capturing the concepts, constraints, business rules and processes that specify the knowledge domain.

Decision Support System (DSS)

The integration of decision support software which leverages data residing in healthcare information systems is a typical scenario. DSS or Clinical Decision Support (CDS) provides timely information, customarily at the point of care, to help inform clinicians' decisions about a patient's care. DSS is typically used, among other things, to detect drug interaction analyses, prescription-mix safety analyses such as adverse drug events (ADE) monitoring.

- Drug Dosage and timing recommendations based individual patient characteristics. Real-time or near real-time patient surveillance for early warning of deteriorating patient condition in hospital settings.
- Aid providers in identifying candidates for alternative treatment plans for improved treatment efficacy as well as improving provider efficiency. These measures directly lead to optimized care plans to minimize hospital length of stay and order sets tailored to the standard of care, thus improve patient health outcomes.
- Control or lower cost of care from alerting providers at the point of care of potentially duplicative test.
- Preventive care recommendations.

DSS is apparently gaining traction (Agency for Healthcare Research & Quality, n.d.) and is here to stay due to growing concerns about the quality of care, cost of care, incentives at the Federal level for meaningful use (MU) and increasing realization for better cognitive support for clinicians.

Method

Purpose, representation, and process drive data modeling. A successful data model satisfies the purpose of the information by accounting for the context of the data collected. Contextual knowledge is essential for all use cases whether it is research, decision support, analytics or specialized registry reporting. The context of information-use guides how it is modeled. Understanding the structure and meaning of this context is paramount for developing a quality data model.

In the last decade, short project delivery times and mainstreaming of agile methodologies has disrupted the art of data modeling by pushing organizations to look for alternative approaches like adopt-and-modify instead of custom development. Secondly, the rise of open-source software development has permeated organizations to lean towards community wisdom rather than organization-specific builds to keep pace with the exploding data generation and even greater emphasis towards treating data as a valuable resource to be leveraged. CDMs represent this evolution.

Evaluation Process

The evaluation process consisted of identifying candidate data models, reviewing past trove of data requests received by LITS from the user community and building a representative list for evaluation. Additionally, identify electronic Health record (EHR) based registry reporting use case to validate the CDMs (Fig.3). Reporting to health registries often involve identifying and prospectively following specific disease-related cohorts which then form the basis of the information being collected, cleansed or in some cases, de-identified and transmitted.

Queryability of the data models were also assessed, to satisfy the chosen sample use cases. Emory University is one of the leading research universities in the nation with health sciences research conducted in over forty specialties. Hence, secondary uses of clinical data is a significant driver of information demand. Additionally, this aspect also translates to a need for integrating disparate data sources like survey data and specimen information that gives a complete picture of the patient information that may be of interest to retrospective and prospective research studies.

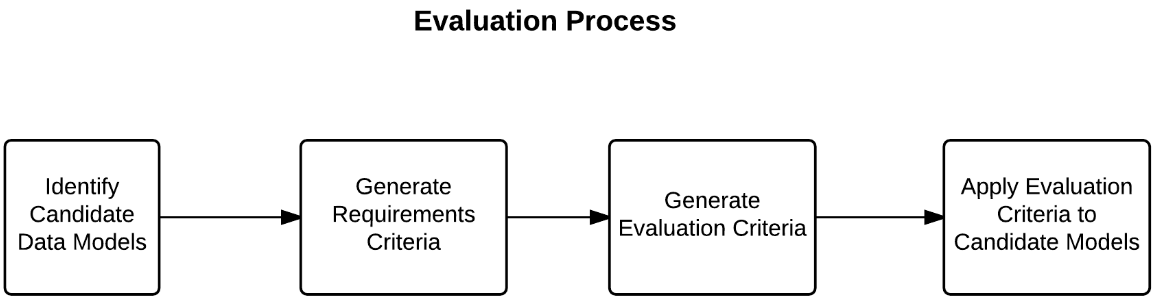

Evaluation Process
Identify Candidate Data Models
Generate Requirements Criteria
Generate Evaluation Criteria
Apply Evaluation Criteria to Candidate Models

Fig 3. Showing the steps in the Evaluation Process

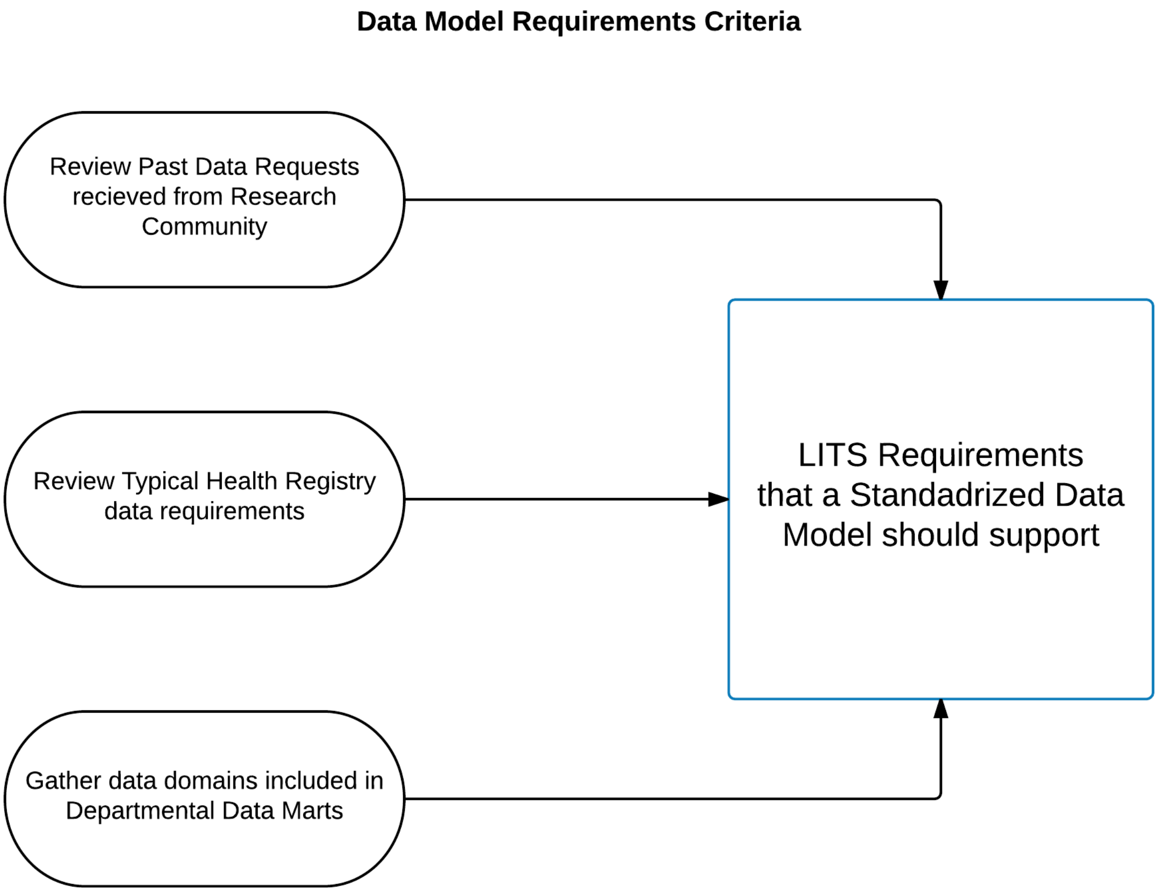

Data Model Requirements Criteria
Review Past Data Requests recieved from Research Community
Review Typical Health Registry data requirements
Gather data domains included in Departmental Data Marts
LITS Requirements that a Standadrized Data Model should support

Fig. 4. Diverse data requirements contributing to the generation of LITS-specific Requirements that a standardized Data Model should ideally support.

As before-mentioned, the requirements for sustaining the diverse data requirements of the Emory departmental users (Fig.4), the CDM should support the following aspects:

1. Incorporate identifiable health information flowing from EHR and other ancillary systems.
2. Represent common health information domains like demographics, diagnoses, encounters, visits, medications, labs, vital signs.

3. Support multiple sources of similar and related data like diagnosis flowing from different settings - billing, problem lists, pathology.

4. Enable linkages across primary care visits, emergency care, and hospital encounters.

5. Support standard controlled terminologies and the ability to include custom Emory-specific codes.

6. Design derived-data cohort building for research projects or specialized health registry reporting.

7. Enable extraction of de-identified health data.

8. Incorporate unstructured or narrative text data.

9. Storing of patient-level secondarily calculated data such as mortality prognosticating indexes like Charlson comorbidity indexes.

10. Specimen information.

11. Support temporal aspects of longitudinal data.

These requirements informed the generation of the evaluation criteria based on established assessment frameworks like Moody and Shanks, which was ultimately evolved further by Kahn et al. The choice of candidate data models for this thesis was gleaned by the assessment of CDMs by Garza and colleagues who operationalized the previously mentioned evaluation criteria. These refinements ultimately helped narrow down the standardized data model chosen for this evaluation.

Candidate Common Data Models

Several research collaborations and working groups have developed and implemented CDMs to promote efficiency in evidence generation practices and to provide better interoperability among diverse study partners. These networks briefly described below, provide a public setting for advancing the ability of analysis standardization.

There are multiple CDMs developed to support the secondary use of data collected in the course of care, including claims and health financial data like costs and billing. These two leading healthcare CDMs include the Observational Medical Outcomes Partnership (OMOP) developed and maintained by the Observational Health Data Sciences and Informatics (OHDSI) program which is a multi-stakeholder, interdisciplinary collaborative with a mission to enable insights from health data through large-scale analytics.

The National Patient-Centered Clinical Research Network (PCORnet) is a distributed network-of-networks, where partner organizations in the PCORnet Distributed Research Network transform data from electronic health record (EHR) and medical claims data sources into a Common Data Model (Microsoft, 2017). PCORnet CDM is designed to make it faster, convenient, and less costly to conduct clinical research than what is now possible by harnessing the power of large amounts of health data and patient partnerships.

OMOP CDM

Background

CDM Name	Observational Medical Outcomes Partnership (OMOP)	
Version	5.2	
Release Date	2017-11-21	
Source	https://github.com/OHDSI/CommonDataModel/blob/master/OMOP_CDM_v5_3.pdf	

Introduction

A fundamental goal for OMOP was to integrate data from multiple data sources by overcoming significant barriers related to the diverse information sources. The OMOP Common Data Model ("OMOP Common Data Model – OHDSI," 2017) is an open-source, community contributed standard for observational healthcare data. The data model specifications and associated work products are available in the public domain. The OMOP CDM is designed to enable collaboration across various sites by unifying data structures and mapping data to standardized vocabularies when possible. This aim is accomplished in the CDM design by six categories of database tables - clinical data, health system data, health economics, derived elements, metadata, and standardized vocabularies (Fig. 5).

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

The majority of tables are person-centric, with connections to Health system databases and vocabulary tables to provide further information and to maintain data provenance. OMOP strives to mitigate information loss when mapping source system data to target CDM by preserving the original data representation as source values.

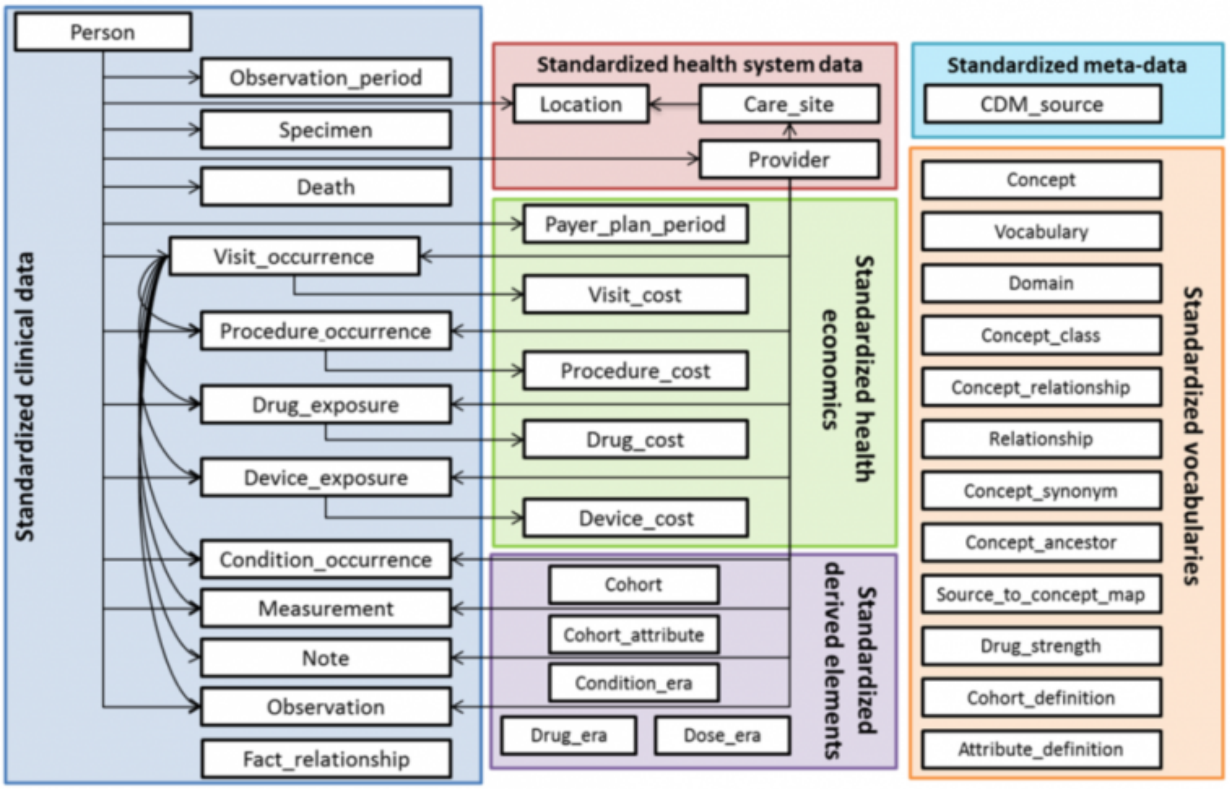


Fig.5 OMOP Data Model with categorization

Standardized Vocabularies are at the core of the OMOP CDM design principles. The CDM strives to include all essential approved healthcare concepts leveraged from national organizations or initiatives, like National Library of Medicine (NLM), Department of Veterans' Affairs (VA), the Center of Disease Control and Prevention (CDC). Standard Concepts encompasses well-established Vocabularies that have comprehensive coverage of the health data domain, and well-defined concepts. For example, SNOMED Vocabulary utilized to codify the Condition Domain. Classification concepts represent categorization of standard vocabularies, thus have a hierarchical relationship to vocabularies. It is possible to source Classification Concepts from different Vocabularies than the Standard Concepts. Note that Classification Concepts are not unique. All concepts that do not belong to the above are assigned as source concepts. They represent the codes specific to the source data. Source concepts reflect the OMOP CDM's flexibility incorporating organization-specific vocabularies different from standard vocabularies (Fig.6).

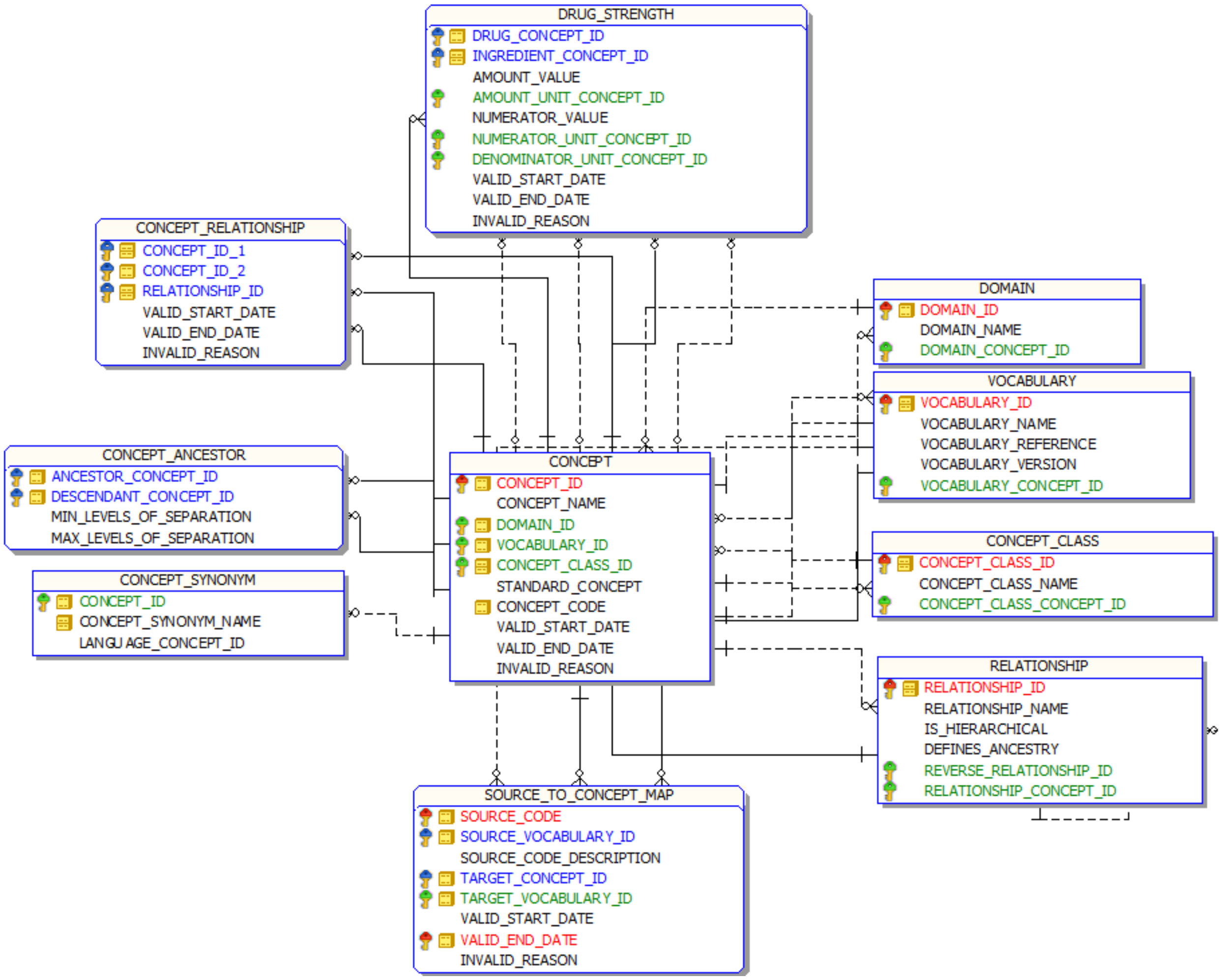


Fig. 6. Physical Model of the Standardized Vocabulary

http://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:standard_classification_and_source_concepts

OMOP Table Names and Description

	Standardized Clinical Data Tables	Description

	PERSON	The Person Table contains records that uniquely identify each patient in the source data for whom there are clinical observation(s) recorded within the source systems.
	OBSERVATION_PERIOD	The OBSERVATION_PERIOD table contains records which uniquely define the spans of time for which a Person has clinical events recorded within the source systems
	SPECIMEN	The SPECIMEN table contains the records identifying biological samples from a person.
	DEATH	The DEATH table contains the clinical event for how and when a Person died.
	VISIT_OCCURRENCE	The VISIT_OCCURRENCE table contains the spans of time a Person continuously receives medical services from one or more providers at a Care Site (location) in a given setting within the health care system. Visits are classified as: ● outpatient care, ● inpatient confinement, ● emergency room, and ● long-term care.

	PROCEDURE_OCCURRENCE	The PROCEDURE_OCCURRENCE table contains records of activities or processes ordered by/performed by, a healthcare provider on the patient to have a diagnostic or therapeutic purpose.
	DRUG_EXPOSURE	The DRUG_EXPOSURE table contains records about the utilization of a Drug. Drugs include prescription, over-the-counter medicines, vaccines, and large-molecule biologic therapies.
	DEVICE_EXPOSURE	The DEVICE_EXPOSURE table contains person's exposure to a foreign physical object or instrument, used for diagnostic or therapeutic purposes. Devices include implantable objects (e.g. pacemakers, stents, artificial joints), medical equipment and supplies (e.g. bandages, crutches, syringes), other instruments used in medical procedures (e.g. sutures, defibrillators) and material used in clinical care (e.g. adhesives, body material, dental material, surgical material).
	CONDITION_OCCURRENCE	The CONDITION_OCCURRENCE table contains records of a Person suggesting the presence of a

		disease or medical condition stated as a diagnosis, a sign or a symptom, which is either observed by a Provider or reported by the patient. Conditions are recorded in different sources as ICD9/10, for instance.
	MEASUREMENT	The MEASUREMENT table contains records of orders and numerical or categorical results of measurement, obtained through systematic and standardized examination or testing of a Person or sample. Measurements include laboratory tests, vital signs, quantitative findings from pathology reports
	NOTE	The NOTE table captures unstructured information that was recorded by a provider about a patient in free text notes.
	NOTE_NLP	The NOTE_NLP table will stores all output of Natural Language Processing (NLP) on clinical notes. Each row represents a single extracted term from a note.
	OBSERVATION	The OBSERVATION table captures clinical facts about a Person obtained in the context of

		examination, questioning or a procedure. Any data that cannot be represented by any other domains, such as social and lifestyle facts, medical history, family history, etc. are stored here.
	FACT_RELATIONSHIP	The FACT_RELATIONSHIP table contains records about the relationships between facts stored as records in any table of the OMOP CDM. Relationships can be defined between facts from the same domain (table), or different domains. Example:Person relationships (parent-child),

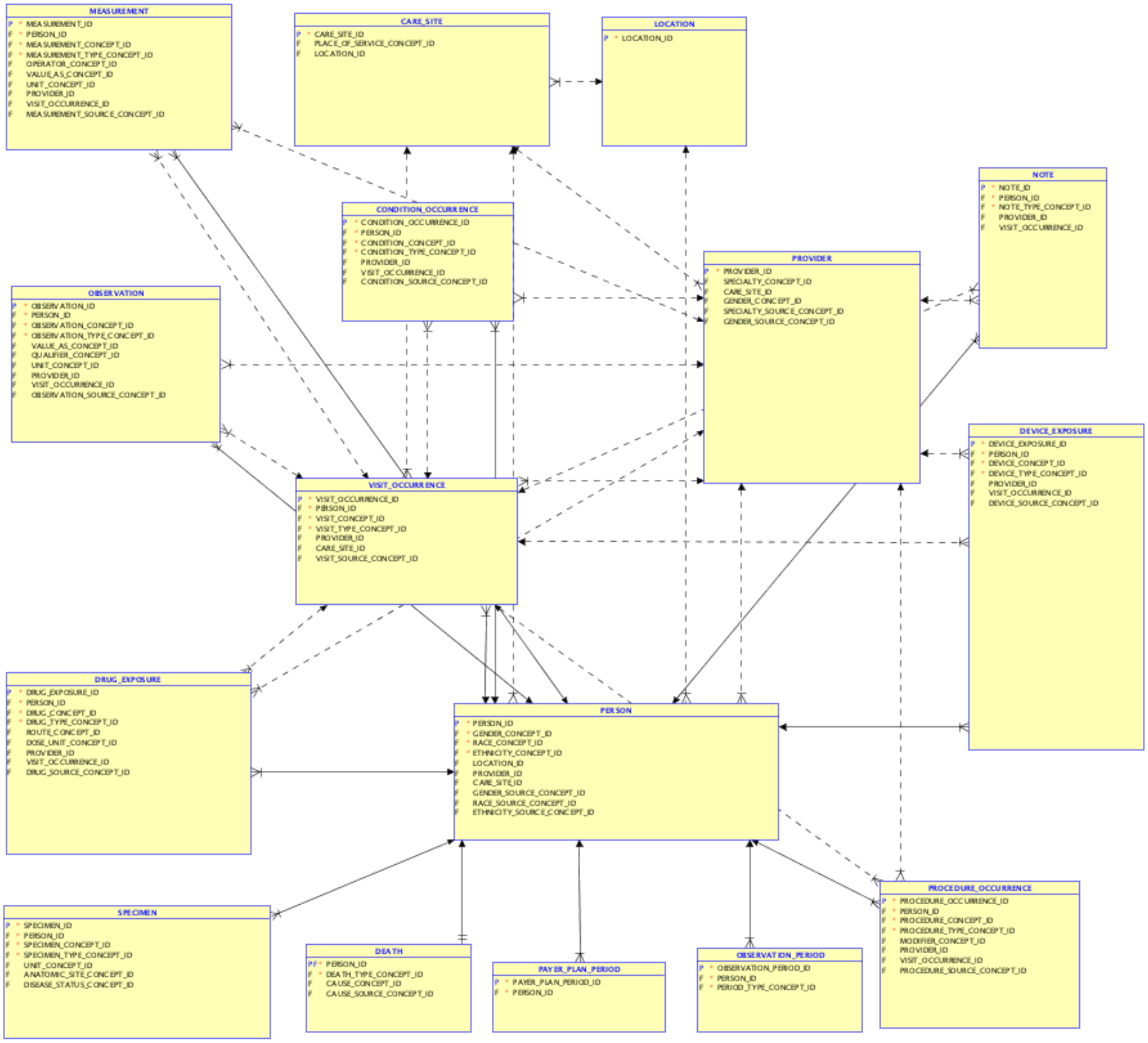


Fig:7: Entity Relationship Diagram Depicting the Standardized Clinical Data Tables. Primary Key and Foreign Keys and relationships shown for brevity.

	Standardized Derived Elements	Description
	COHORT	The COHORT table contains records of subjects that satisfy a given set of criteria for a duration of

		time. Cohorts can be constructed of patients (Persons), Providers or Visits. Often used for research or disease registry reporting.
	COHORT_ATTRIBUTE	The COHORT_ATTRIBUTE table contains attributes associated with each subject within a cohort.
	DRUG_ERA	A Drug Era is defined as a span of time when the Person is assumed to be exposed to a particular active ingredient. A Drug Era is different from Drug Exposure: Exposures are records when a drug was delivered to the Person. Successive periods of Drug Exposures are combined under certain rules to produce continuous Drug Eras.
	DOSE_ERA	A Dose Era is defined as a span of time when the Person is assumed to be exposed to a constant dose of a specific drug.
	CONDITION_ERA	Condition Eras are chronological periods of Condition Occurrence. Allows aggregation of chronic conditions that require frequent ongoing care. Also allows aggregation of multiple, closely timed doctor visits for the same Condition.

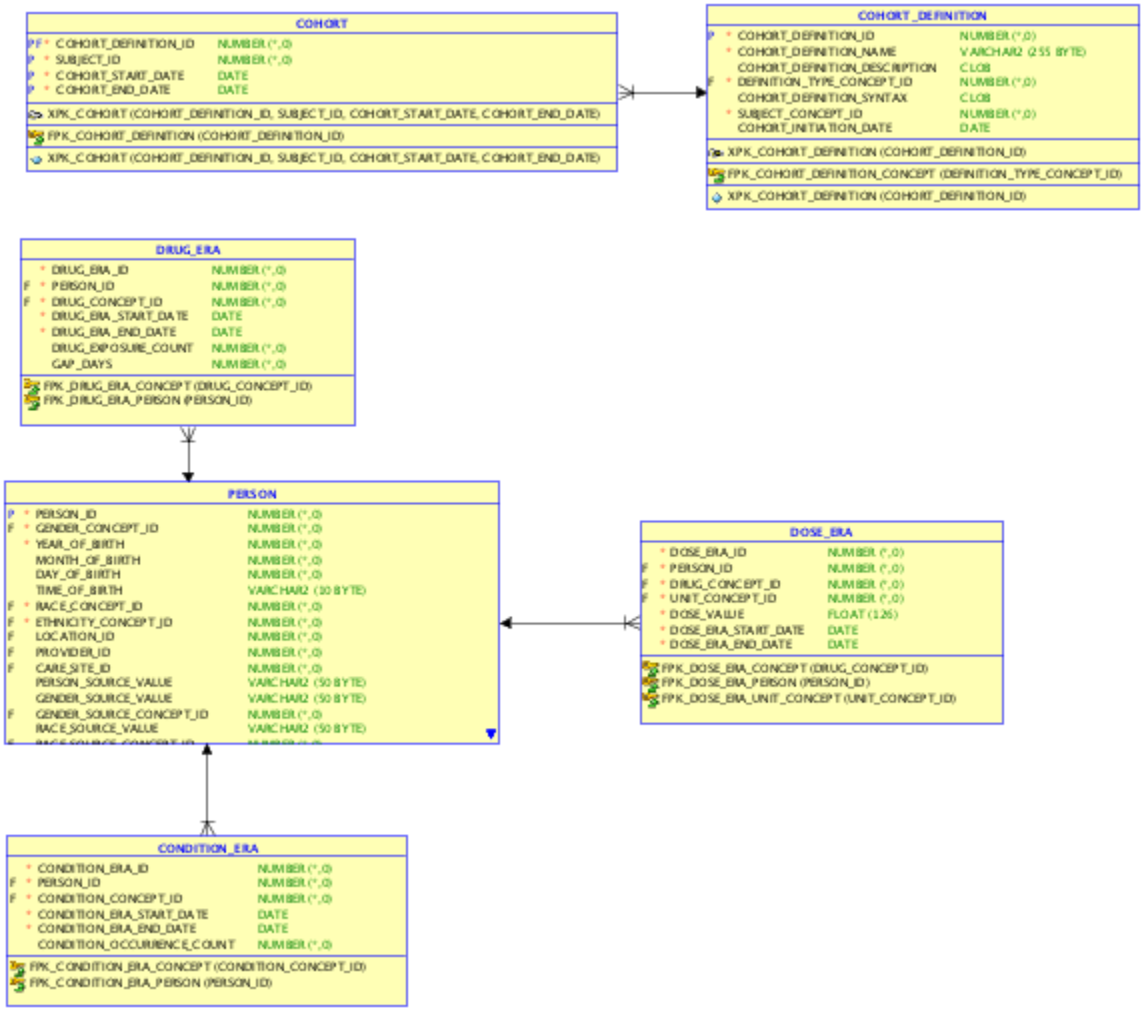

COHORT
PF * COHORT_DEFINITION_ID NUMBER (*,0)
P * SUBJECT_ID NUMBER (*,0)
P * COHORT_START_DATE DATE
P * COHORT_END_DATE DATE
XPK_COHORT (COHORT_DEFINITION_ID, SUBJECT_ID, COHORT_START_DATE, COHORT_END_DATE)
FPK_COHORT_DEFINITION (COHORT_DEFINITION_ID)
XPK_COHORT (COHORT_DEFINITION_ID, SUBJECT_ID, COHORT_START_DATE, COHORT_END_DATE)
COHORT_DEFINITION
P * COHORT_DEFINITION_ID NUMBER (*,0)
* COHORT_DEFINITION_NAME VARCHAR2 (255 BYTE)
COHORT_DEFINITION_DESCRIPTION CLOB
F * DEFINITION_TYPE_CONCEPT_ID NUMBER (*,0)
COHORT_DEFINITION_SYNTAX CLOB
* SUBJECT_CONCEPT_ID NUMBER (*,0)
COHORT_INITIATION_DATE DATE
XPK_COHORT_DEFINITION (COHORT_DEFINITION_ID)
FPK_COHORT_DEFINITION_CONCEPT (DEFINITION_TYPE_CONCEPT_ID)
XPK_COHORT_DEFINITION (COHORT_DEFINITION_ID)
DRUG_ERA
* DRUG_ERA_ID NUMBER (*,0)
F * PERSON_ID NUMBER (*,0)
F * DRUG_CONCEPT_ID NUMBER (*,0)
* DRUG_ERA_START_DATE DATE
* DRUG_ERA_END_DATE DATE
DRUG_EXPOSURE_COUNT NUMBER (*,0)
GAP_DAYS NUMBER (*,0)
FPK_DRUG_ERA_CONCEPT (DRUG_CONCEPT_ID)
FPK_DRUG_ERA_PERSON (PERSON_ID)
PERSON
P * PERSON_ID NUMBER (*,0)
F * GENDER_CONCEPT_ID NUMBER (*,0)
* YEAR_OF_BIRTH NUMBER (*,0)
MONTH_OF_BIRTH NUMBER (*,0)
DAY_OF_BIRTH NUMBER (*,0)
TIME_OF_BIRTH VARCHAR2 (10 BYTE)
F * RACE_CONCEPT_ID NUMBER (*,0)
F * ETHNICITY_CONCEPT_ID NUMBER (*,0)
F LOCATION_ID NUMBER (*,0)
F PROVIDER_ID NUMBER (*,0)
F CARE_SITE_ID NUMBER (*,0)
PERSON_SOURCE_VALUE VARCHAR2 (50 BYTE)
GENDER_SOURCE_VALUE VARCHAR2 (50 BYTE)
F GENDER_SOURCE_CONCEPT_ID NUMBER (*,0)
RACE_SOURCE_VALUE VARCHAR2 (50 BYTE)
DOSE_ERA
* DOSE_ERA_ID NUMBER (*,0)
F * PERSON_ID NUMBER (*,0)
F * DRUG_CONCEPT_ID NUMBER (*,0)
F * UNIT_CONCEPT_ID NUMBER (*,0)
* DOSE_VALUE FLOAT (126)
* DOSE_ERA_START_DATE DATE
* DOSE_ERA_END_DATE DATE
FPK_DOSE_ERA_CONCEPT (DRUG_CONCEPT_ID)
FPK_DOSE_ERA_PERSON (PERSON_ID)
FPK_DOSE_ERA_UNIT_CONCEPT (UNIT_CONCEPT_ID)
CONDITION_ERA
* CONDITION_ERA_ID NUMBER (*,0)
F * PERSON_ID NUMBER (*,0)
F * CONDITION_CONCEPT_ID NUMBER (*,0)
* CONDITION_ERA_START_DATE DATE
* CONDITION_ERA_END_DATE DATE
CONDITION_OCCURRENCE_COUNT NUMBER (*,0)
FPK_CONDITION_ERA_CONCEPT (CONDITION_CONCEPT_ID)
FPK_CONDITION_ERA_PERSON (PERSON_ID)

Fig.8: ERD showing the OMOP Standardized Derived Elements

PCORNet CDM

Background

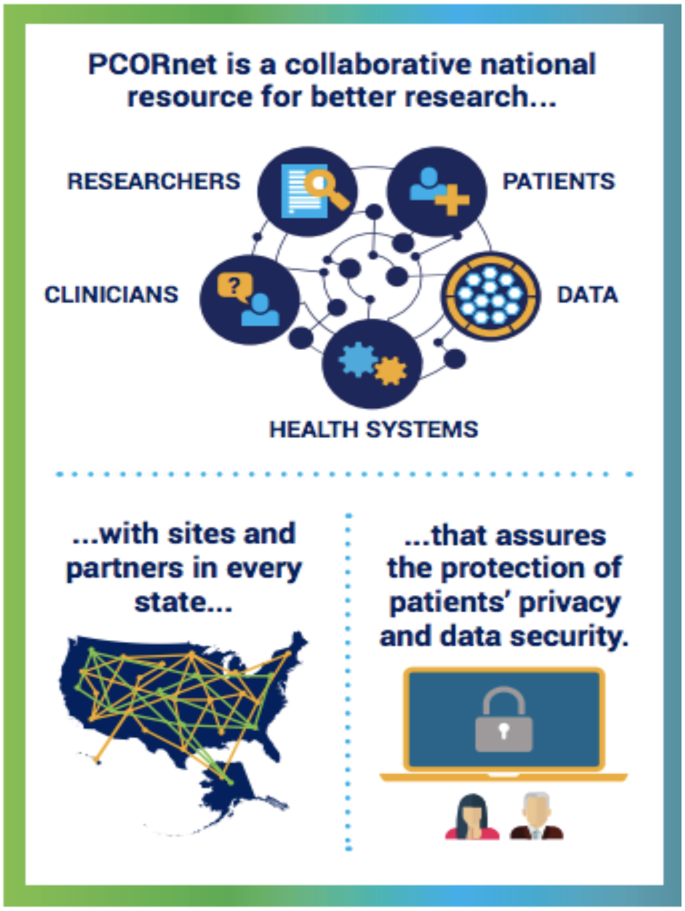
CDM Name	Patient Centered Outcomes Research Network (PCORnet)	
Version	4.0	
Release Date	2018-01-31	
Source	http://www.pcornet.org/wp-content/uploads/2018/01/PCORnet-Common-Data-Model-v4.0_Specification.pdf	

Introduction

The Patient-Centered Outcomes Network (PCORnet) is a nationally distributed research network (DRN) funded by the Patient-Centered Outcomes Research Institute (PCORI). This “network of networks” brings together participating institutions across the United States to form an infrastructure to elucidate critical scientific questions based on the infrastructure of secondary data generated through healthcare delivery - clinical and administrative - in Electronic Health Records (EHR), health plans, and claims data sources.

A crucial component of PCORnet was to develop the PCORnet Common Data Model (CDM), a standardized representation of data elements and domains, selected and structured to optimize rapid implementation of distributed analytical functionality, to support PCORnet DRN objectives ("PCORnet Common Data Model (CDM) - PCORnet," 2018) (Fig. 9,10,11).

pcornet®



PCORnet is a collaborative national resource for better research...

RESEARCHERS

PATIENTS

CLINICIANS

DATA

HEALTH SYSTEMS

...with sites and partners in every state...

...that assures the protection of patients' privacy and data security.

Fig. 9: https://www.pcori.org/sites/default/files/PCORI-PCORnet-Fact-Sheet.pdf

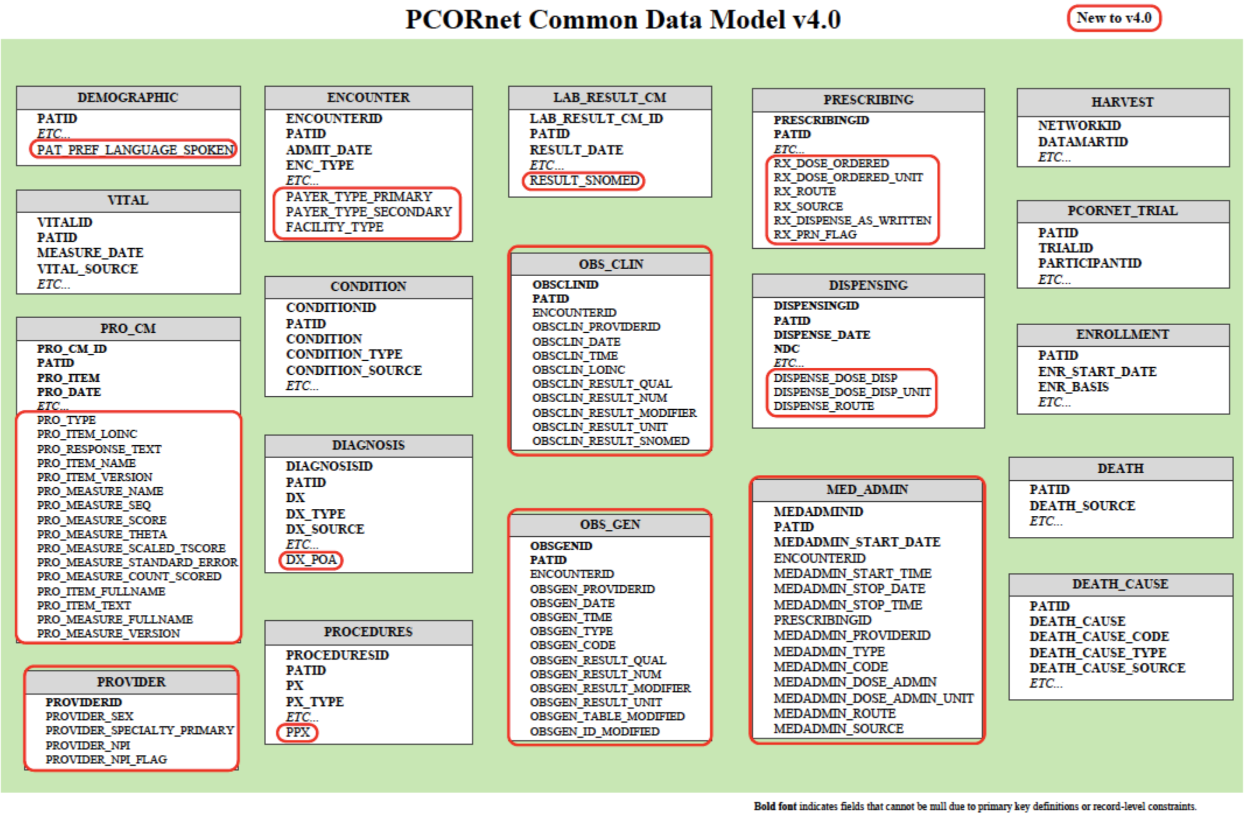


Fig. 10: http://www.pcornet.org/wp-content/uploads/2018/01/PCORnet-Common-Data-Model-v4.0_Specification.pdf

PCORnet Table Names and Description

	Table Name	Description
	DEMOGRAPHIC	Demographics record the direct attributes of individual patients.
	ENROLLMENT	Enrollment is a concept that defines a period of time during which a person is expected to have complete data capture. This concept is often insurance-based, but other methods of defining enrollment are

		possible.
	ENCOUNTER	Encounters are interactions between patients and providers within the context of healthcare delivery.
	DIAGNOSIS	Diagnosis codes indicate the results of diagnostic processes and medical coding within healthcare delivery. Data in this table are expected to be from healthcare-mediated processes and reimbursement drivers.
	PROCEDURES	Procedure codes indicate the discreet medical interventions and diagnostic testing, such as surgical procedures and lab orders, delivered within a healthcare context.
	VITAL	Vital signs (such as height, weight, and blood pressure) directly measure an individual’s current state of attributes.
	DISPENSING	Prescriptions filled through a community, mail-order or hospital pharmacy. Outpatient dispensing may not be directly captured within healthcare systems.
	LAB_RESULT_CM	This table is used to store quantitative and

		qualitative measurements from blood and other body specimens.
	CONDITION	A condition represents a patient’s diagnosed and self-reported health conditions and diseases. The patient’s medical history and current state may both be represented.
	PRO_CM	This table is used to store responses to patient-reported outcome measures (PROs) or questionnaires. This table can be used to store item-level responses as well as the overall score for each measure.
	PRESCRIBING	Provider orders for medication dispensing and/or administration. These orders may take place in any setting, including the inpatient or outpatient basis.
	PCORNET_TRIAL	Patients who are enrolled in PCORnet clinical trials and PCORnet studies.
	DEATH	Reported mortality information for patients.
	DEATH_CAUSE	The individual causes associated with a reported death.
	MED_ADMIN	Records of medications administered to patients by

		healthcare providers. These administrations may take place in any setting, including inpatient, outpatient or home health encounters.
	PROVIDER	Data about the providers who are involved in the care processes documented in the CDM.
	OBS_CLIN	Standardized qualitative and quantitative clinical observations about a patient.
	OBS_GEN	Table to store everything else.
	HARVEST	Attributes associated with the specific PCORnet datamart implementation, including data refreshes.

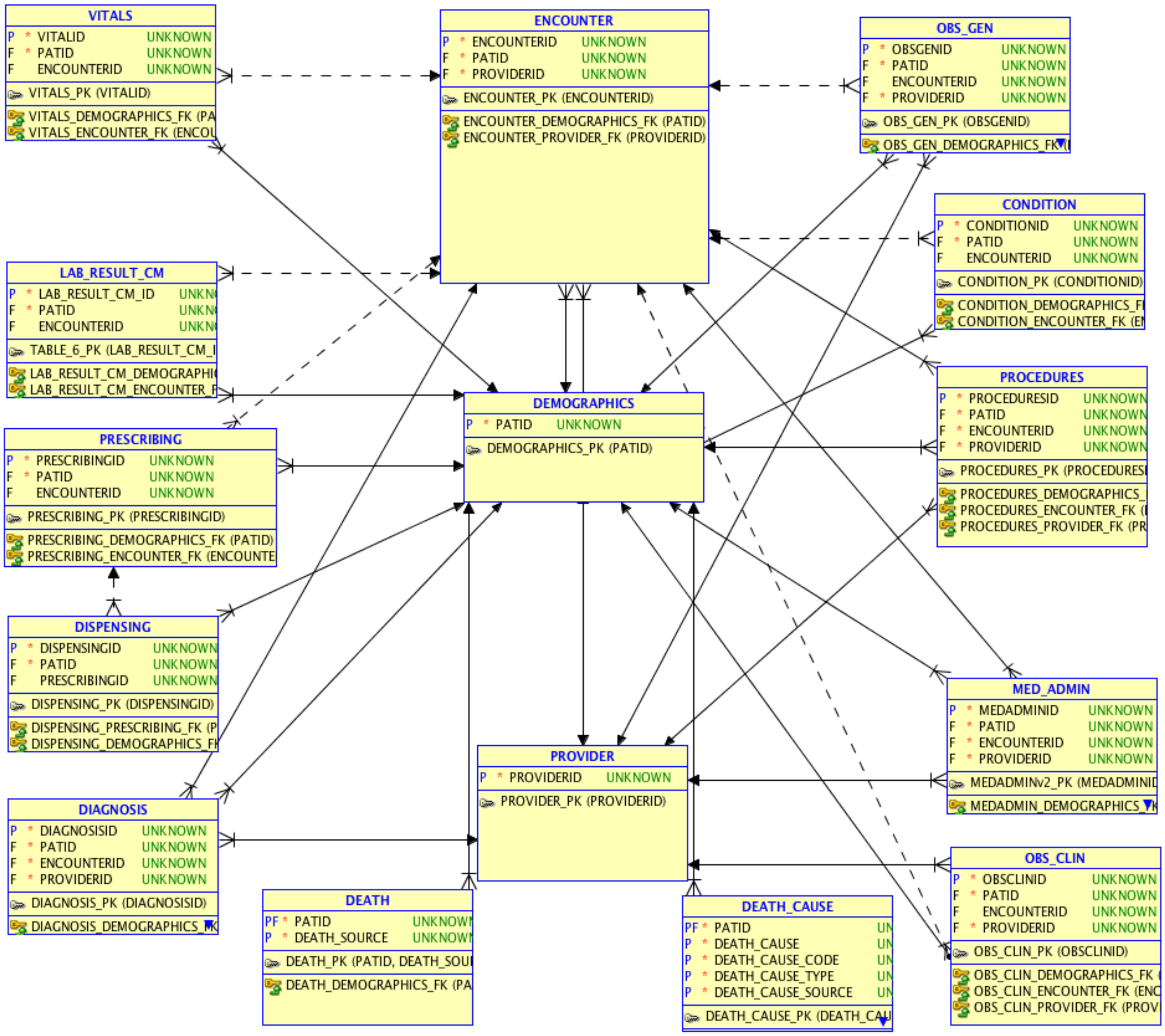


Fig: 11: Entity Relationship Diagram Depicting the PCORNet tables Primary Key and Foreign Key and relationships shown for brevity.

Evaluation Criteria

Due to the diverse aspects of LITS department's needs in mind as well as balancing the realities of custom data models versus adopting a standard data models, we synthesized the Moody and Shanks (2003) evaluation framework as well as the Kahn et al (2012). criteria. This data model quality evaluation methodology provides a comprehensive framework that represents an evolutionary approach that considers all potential data model features of interest for LITS department. Moody and Shanks's approach evaluates both the product quality and process quality that produces the final product - a standard data model - for this discussion.

There are eight "quality factors" that are considered for data model evaluation. These factors take into account the various stakeholders typically involved in the systems development from business analysis to data persistence design and implementation. In addition to the broad assessment criteria, there is a recognition of the need for flexibility, practicality, and rigor in the assessment dimension. The evaluation criteria make a deliberate effort to overcome the barriers that exist between information systems academics (theory) and practitioners by relying on action research process of plan, act, observe and reflect. These elements bolster the empirical and iterative nature of the evaluation framework, thus enabling its applicability in a wide variety of settings.

Several measures from the Moody and Shanks evaluation criteria are excluded from this assessment including correctness and understandability. Moody and Shanks (Moody & Shanks, 2003) define correctness as "whether the model conforms to the rules of the data modeling technique." As both OMOP and PCORnet models are collaboratively developed, publicly available and successfully implemented by institutions, it is expected that these models meet the data modeling technical conventions. Moody and Shanks criteria of understandability are debatable for CDMs as they are not abstract and undecipherable. Moreover, LITS department has the informatics workforce who are trained and eminently qualified to interpret these data models and have the domain knowledge to supplement this awareness. At a general level, understandability is essential and can bolster any assessment but is less relevant given this specific situation.

From Kahn et al. the scalability measure is less applicable to this evaluation. Scalability is defined as "can the model be sized to a smaller or larger data set." With cloud computing and on-premise powerful servers and databases, a data model should store the source data in a scalable manner. Relational databases offer many tools and techniques to improve scalability out of the box - such as indexing, clustering, data partitioning, aggregate-aware query optimization, columnar data storage.

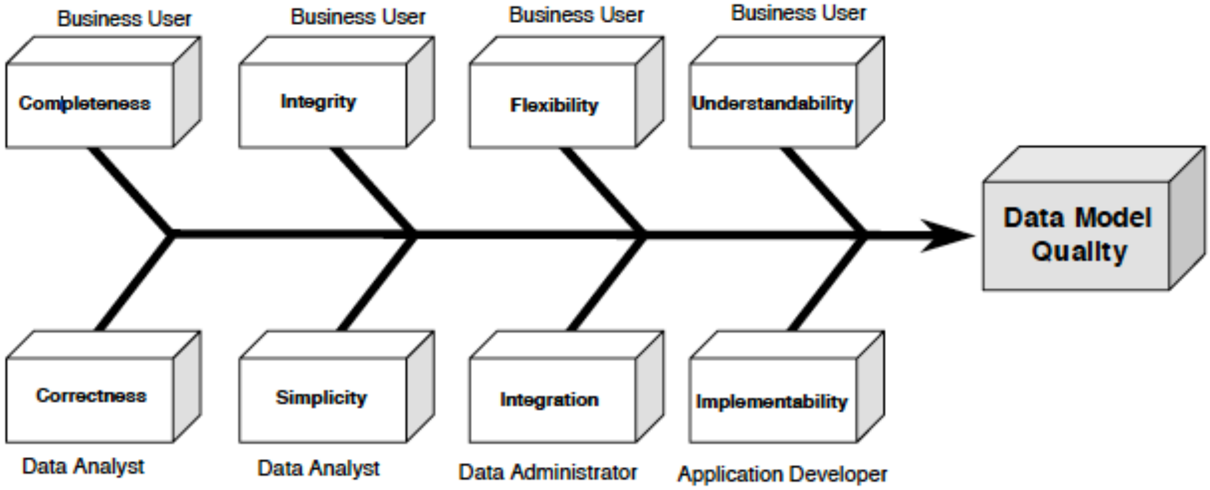


Fig.11: Moody and Shanks Quality Factors

Moody-Shanks Evaluation Model

Quality Factor	**Moody and Shanks Definition (2003)**
Completeness	refers to whether the data model contains all user requirements
Simplicity	means that the data model contains the minimum possible entities and relationships.
Flexibility	defined as the ease with which the data model can cope with business and/or regulatory change.
Integrity	defined as the consistency of the data model with the rest of the organisation’s data.
Understandability	defined as the ease with which the concepts and structures in the data model can be understood.
Implementability	defined as the ease with which the data model can be implemented within the time, budget and technology constraints of the project.
Correctness	Defined as the data model’s adherence to sound principles.

Integration	Defined as the Extent of the DM supporting controlled vocabularies.

Completeness

The typical use cases for evaluating the completeness of the data model was generated by evaluating the data requests received by the LITS Data Solutions Team (DST). These data requests typically are received from the research community representing diverse departments within the School of Medicine (SOM). These data requests back research studies, analytical data sets, departmental quality initiatives or take the form of data projects for specialized registry projects and departmental data marts. The standard data model should satisfy these diverse requirements to the most considerable extent possible to satisfy LITS' objectives for a CDM. The primary domains of health data in the current data warehouse like patient, encounters, labs, vitals, medications and additionally data from ancillary sources like LIMS for specimens and survey data like REDCap should be adequately represented in the CDM. A few of the illustrative use cases for the type of data requested are:

Use Cases
All patients with End Stage Renal Disease based on Diagnosis code and dialysis status for the last ten years. For this cohort calculate the encounter Length of Stay (LOS), add the comorbidities as well as the administered medication during the hospital stay.

Identify the following population of melanoma cancer population based on diagnosis. For this cohort, extract positive mutations based on the cancer mutation panel recorded in the anatomic pathology report along with Diabetes Mellitus type 2 comorbid conditions.

Simplicity

Moody and Shanks' describes simplicity as a limited number of entities and links between them. The simplicity of model defined this way is inversely correlated to completeness. The framework shows that simplicity has a positive correlation to the overall quality of the model, completeness often leads to a more complex model regarding implementation friendliness and understandability. Ideally, a model should strive to distill the possible universe of entities and relationships. Simplicity can potentially challenge the data transformation process from source systems as the meaning of the data will need to be maintained while feeding a model with a reduced number of entities. Denormalization of data would probably be required.

For this evaluation, from a LITS departmental perspective query-ability of the data, i.e., the number of table joins needed to query the needed information and the transformation required to output the final resultset is of particular interest.

Flexibility

The types of healthcare data captured in the course of clinical care will change over time with newer modalities of care. Data models should cope with evolving number of data elements over time which can also be due to regulatory changes. US Healthcare is subject government regulations especially the secondary use of the data for comparative effectiveness research, quality of care indicators, population health reporting. HIPAA regulations around protected health information (PHI) and the resultant need for de-identification is an example of changing regulations that models should be amenable to be extended without significant changes. For LITS, this is a pertinent quality dimension as the future needs of the research community often backed by federal and state supported grants are difficult to predict. Complicating this picture is the changing technological landscape like the advent of cloud-based applications and data processing.

CDM entities should be readily extensible by way of adding attributes that reflect site-specific elements that allow data provenance which is vital to health data (Buneman, et al.,2000). Flexibility can also be judged by the granularity of data the model can consume. Data that is in its most granular non-aggregated state support a wider variety of uses in the community. Aggregation at the point of use allows for flexible definitions and has a higher chance of keeping up with changing clinical definitions over the years.

Understandability

Emory University has a diverse data user community, that which is made of a multitude of disciplines and medical specialties. These user groups vary from principal investigators, data analysts, biostatisticians. A data model should be intuitive by design, one where the entities reflect the health data domains and the linkages between them. This simplifies the metadata which describes the data contained in the entities making up the model to bring out the assumptions, intended use, and restrictions of the attributes. Eventually, understandability is subjective, depending on the stakeholder familiarity with healthcare domain, institution-specific workflow, and processes. As Moody and Shanks (2003) have pointed out understandability has a significant influence on perceptions of data model quality as it reflects the data models function of communicating with users.

Integration

No data model exists in isolation in an organization. Even if there are no explicit links between models that form the universe of data representation, it should conform to the "big picture" or the commonly agreed upon data definitions. For this evaluation, integrity is weighed in favor of CDM that adopts controlled vocabulary standards of the healthcare data like International Classification of Diseases (ICD-10-CM), RxNorm ("RxNorm," 2004), Logical Observation Identifiers Names and Codes (LOINC), National Provider Identifier Standard ("National Provider Identifier Standard (NPI) - Centers for Medicare & Medicaid Services," 2015) have much higher chance of being integrated with source models without loss of data fidelity. High levels of data reuse behoove that the data model should integrate well with existing information ecosystem. This capability will ensure that the CDM concepts map readily with the domain definitions and entity attributes of the source system where the data originates.

Implementability

This dimension looks at the essence of the process of implementing the data model. Implementability of a CDM is affected by the number of revisions that it has undergone. It is desirable for the model to be mostly stable so that future changes can be implemented incrementally rather than a significant overhaul and the resultant changes rippling through the application stack.

The application ecosystem is another excellent dimension to evaluate the maturity of the data model under consideration. The more third-party tools that are available indicate the rate of adoption and plausible reduction in the barriers to transitioning to the CDM. Besides the ecosystem surrounding the model, the diversity of intended uses for the CDM by the designers can be illuminating to inform the breadth of its applicability. Although the standard terminology coverage and harmonization of various vocabularies speaks directly to the CDM's strength and applicability, the time and costs associated with vocabulary mapping are not inconsequential. However, for this evaluation, it should be noted that this cost is considered on a qualitative scale of high, medium and low since it is a reality no matter the CDM under assessment. So the emphasis is placed on the availability of utilities to possibly reduce the burden of mapping and development at implementation phase. Lastly, implementability criterion is also influenced by the rate of adoption of the model among institutions.

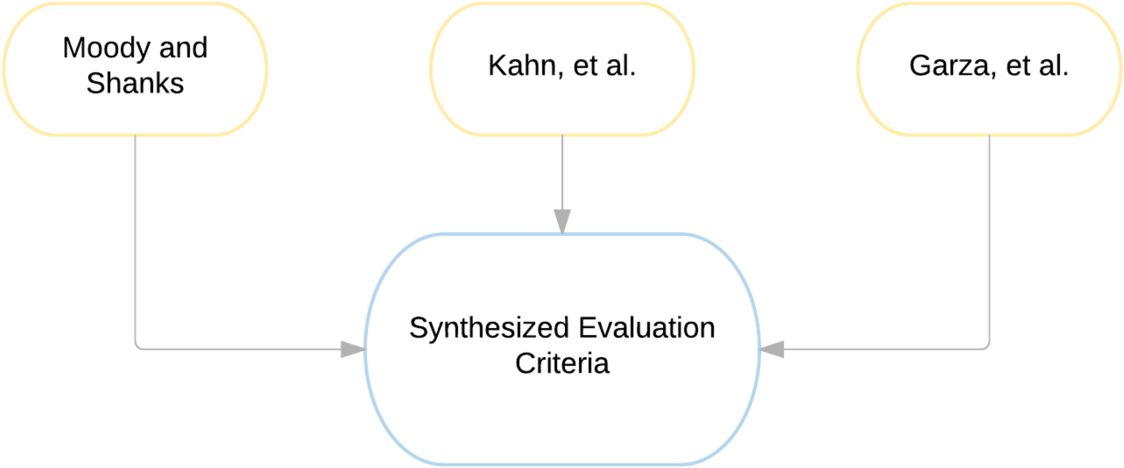


Fig 12: Synthesized Evaluation Criteria

Evaluation Criteria	Moody and Shanks	Kahn et al.	Garza et al.	Synthesized Criteria
Completeness	X			X
Domains		X	X	X

Domain Attributes			X	X
Integration	X			
Standard Terminology Support	X	X	X	X
Flexibility	X	X	X	
Extensibility		X	X	X
Adaptability		X		
Scalability		X		
Understandability	X	X		X
Correctness	X	X		
Simplicity	X		X	X
Ease of Querying			X	X
Ease of Anonymizing & De-Identification			X	X
Implementability	X	X	X	X
Field Experience		X		
Stability		X	X	X
Adoption		X	X	X
Grid Friendliness		X	X	
Cost		X	X	X

Fig13 Shows the Synthesized Evaluation Criteria for assessing the candidate CDMs

Results

Completeness - Domain and attribute coverage.

OMOP CDM has the most coverage of the domains of interest for LITS evaluation. While the latest version of the PCORnet CDM added more entities and enhanced the attributes of the existing objects, it still lacks domains like Specimen, Unstructured notes, and Survey results. For example, PCORnet version 4 added the OBS_GEN table to store general observations which may be appropriated for storing survey results. But these operational decisions at implementation time limits the standardized applications from being utilized to their fullest extent. In particular, the PCORNet CDM design is optimized for patient-centered comparative effectiveness research. Hence, the CDM includes some data domains that are relatively more highly specialized to target specific use cases when compared to other common CDMs. The broad aim of the CDM is to impose uniformity across data partners to enable the platform-driven app ecosystems and interoperability. Persisting unstructured data like clinical notes and impressions allows Emory to be ready for future, yet-to-be-determined uses of the data or the rapidly evolving big data analytics. To be sure, PCORnet provides columns that flag the source of the data derived from data mining, machine learning, and natural language processing (NLP). Specimen information is a critical need for biomedical research and clinical trials reporting. The results show the domains and data points that each model include.

Domain/Attributes	**OMOP v5.2**	**PCORnet v4.0**
Demographics	Yes	Yes
Personal demographics	Yes	Yes
Contact information	Yes	Yes
Social History	Yes	Yes

Medical History	Yes	Yes
Immunization History	Yes	No
Insurance	Yes	Yes
Death	Yes	Yes
Cause of Death	Yes	Yes
Encounter	Yes	Yes
Type	Yes	Yes
Admit/Discharge Date	Yes	Yes
Discharge Disposition	Yes	Yes
Location	Yes	Yes
Provider	Yes	Yes
Vitals	Yes	Yes
Lab Results	Yes	Yes
Reference Range	No	Yes
Diagnosis	Yes	Yes
Diagnosis Source (admit/discharge)	Yes	Yes
Diagnosis Data Source		
Diagnosis Code	Yes	Yes
Survey Results	Yes	No
Procedures	Yes	Yes
Date	Yes	Yes
Duration	No	No
Procedure code	Yes	Yes
Provider	Yes	Yes

Medication	Yes	Yes
Context	Yes	Yes
Item	Yes	Yes
Dose	Yes	Yes
Quantity	Yes	Yes
Frequency		Yes
Route of Administration	Yes	Yes
Start and End date	Yes	Yes
Provider	Yes	Yes
Provider	Yes	Yes
Demographics	Yes	Yes
Specialty	Yes	Yes
Practice location	Yes	No
National Provider ID (NPI)	Yes	Yes
Specimen		
Type	Yes	No
Location	No	No
Quantity	Yes	No
Anatomic Site	Yes	No
Collection Date	Yes	No
Notes		
Unstructured	Yes	No
Semi Structured	Yes*	Yes*
Cohort Definition	Yes	No

Integration

By creating a standardized mapping between vocabularies like RxNorm and other medication data references, such as First Databank, Anatomical Therapeutic Chemical (ATC), it's possible to link drugs to chemicals, protein targets, genes, and disease associations. The OMOP vocabulary mapping metadata tables serve as a useful guide for determining the destination tables where the source data should be stored. This data-driven approach lowers the barriers for institutions like Emory LITS department looking to adopt CDMs. Once adopted, harmonization process of source concepts to standardized concepts are significantly enhanced. PCORnet has some standardized vocabularies, but there is no metadata support out of the box as in the case of OMOP Standardized vocabularies which includes extensive mapping and higher order categorization of concepts. PCORnet uses a significant number of predefined value sets unique to PCORnet, while the rest are sourced from Sentinel
Common Data Model (SCDM) ("Sentinel Common Data Model | Sentinel Initiative," 2017), Mini-Sentinel Common Data Model (SCDM), Center for Medicaid and Medicare Services (CMS), RxNorm, LOINC to name a few. In OMOP CDM tables, the meaning of the content of each record is represented using Concepts. The OMOP's extensive standardized vocabularies coverage, concept classification, and mapping represent a stronger emphasis on uniformity (Fig.14). When institutions consider moving to a standard data model, this uniformity is invaluable at the outset.

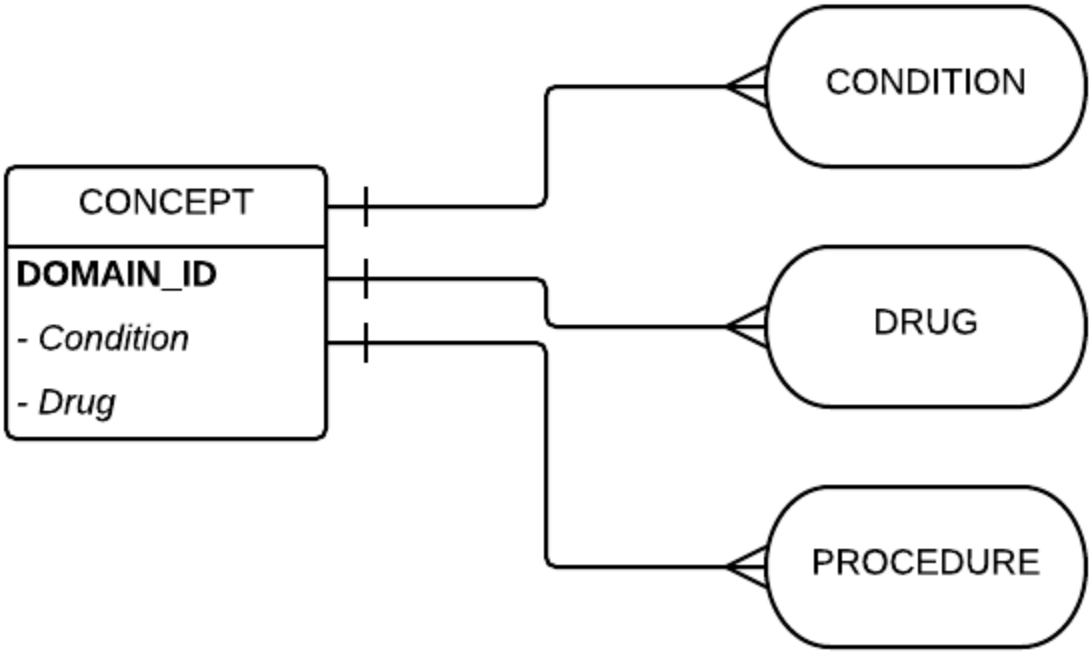


Fig14: Metadata driven design where the concept's domain informs where the data belongs

Flexibility

Overall both OMOP and PCORnet CDMS were found to be extensible to adding attributes to domains without changing the critical constraints of the domains. Adding new entities to the model were also possible without changing the key associations. However, the OMOP model is found to be more flexible regarding accommodating heterogeneous data due to its higher emphasis on entity-attribute-value (EAV) model. Furthermore, the metadata-driven vocabulary mapping and categorization are also adaptable to incorporate complex concepts. The vocabularies tables include the description of the relationship between concepts and their classification.

With the OMOP Version 5 release, one saw the addition of relationship tables that described relationships between observations or across domains. This enhancement meant that survey data like those originating from popular research data capture tools like REDCap (Harris, et al., 2009) could be incorporated into the existing model without the need to create site-specific tables and attributes which would be beyond the reach of standardized tool sets developed for the OMOP CDM.

Simplicity - Ease/Complexity of Querying

Both OMOP and PCORnet CDM are similar regarding the number of joins between tables, union statements and nested queries for retrieving information from the data model. The queries, written in Structured Query Language (SQL), were compared by analyzing the scripts and associated files published by the Informatics Common Metric ("ncats/CTSA-Metrics", 2017) initiative. The Clinical and Translational Science Awards Program (CTSA) ("Common Metrics," n.d.) under the auspices of the National Center for Advancing Translational Sciences (NCATS) publishes the queries/scripts to enable standardized automated query against the data repository that uses the standardized data models like OMOP and PCORnet. Additionally, for OMOP, sample queries were investigated by analyzing the queries hosted by the Reagan-Udall Foundation ("OMOP CDM Queries," n.d.).

Implementation - Field Experience, Stability, Adoption.

The OMOP CDM version was released in 2009 while the PCORnet CDM has been existence since 2014. OMOP has been in existence longer than and is reflected in the evolution of the content coverage as well as the diverse adoption comprising nationally and internationally. PCORI's particular intent on patient-centered multi-site research has lead to member institutions adopting it. Even within these members, a few have implemented OMOP CDM. There was also a collaborative initiative that mapped PCORnet elements into OMOP standardized vocabularies making it amenable to populate PCORnet tables from OMOP Data model. As can be expected from a newer model, PCORnet CDM has undergone major releases every year since inception. The latest version (4.0) was released end of January 2018 with additional attributes and tables. Meanwhile, the OMOP CDM has seen minor releases from the multi-disciplinary stakeholder Observational Health Data Sciences and Informatics (OHDSI) collaborative, the latest in November 2017.

The OMOP CDM community actively engages and supports users through the online forums and yearly symposia. The OHDSI collaborative seeks volunteers to participate in working groups looking to enhance the CDM. The meeting agendas and discussion notes are publicly available. This transparency and inclusive culture to bring together individuals of varying backgrounds and experience bode well for the robustness of the data model. For Emory or other large institutions, this represents an opportunity to help shape its future direction. At the very least a supportive community reduces the burden of transitioning to a standardized model by influencing implementation estimations, clarifying guiding principles and ultimately lets Emory tap into community wisdom. PCORnet also has a forum providing documentation and reference material for best practice recommendations. This public facing forum (https://github.com/CDMFORUM) accepts enhancement requests from data partners and maintains a list of errors reported by member organizations and individuals. It should be noted that the PCORnet community is less active when compared to OMOP CDM forums. For organizations considering moving to a CDM, active user population contributing ideas, reporting issues, discussing enhancements and future roadmaps for improvements are vital.

Measure	**OMOP CDM v5.3**	**PCORnet CDM v4**
Completeness - Domain coverage by Standard Vocabularies	All Domains covered	Partial Coverage
Integration - Source Concepts to Standard Concepts Mapping	Completely mapped. OMOP CDM offers source to standard concepts mapping as well as categorization of standard concepts. These mapping are through metadata driven Standardized vocabulary tables Enables a terminology driven ETL process	None.
Integrity - Standard vocabularies supported	Comprehensive	Partial Coverage
Flexibility (adding Data Elements)	Yes	Yes
Understandability	Yes	Yes
Implementability - Adoption		
Number of Adopters (organizations)	34	34
International	Yes	No
Implementability - Stability		
Model updates in the last three years	1	3
Minor Updates	0	1
Major Updates	1	2
Implementability - App EcoSystem		
Application ecosystem	Yes	No
Implementability - ETL Tools		
ETL Tools	Yes	No
Community ETL Support	Yes (forum)	No
Implementability - ETL Tools		

Community Support	Yes	Yes

Discussion

There is a critical need to improve the infrastructure supporting the reuse of Health data both within an enterprise and in diverse cross-institutional collaborative efforts. Data management standardization is an essential conduit for knowledge discovery, decision support, and downstream uses of data like research and quality improvement. The various stakeholders representing academia, funding agencies, industry, and the public behoove the information to be in a standard format that is easily discoverable, reusable and concise. Thus a standard data substrate such as a Common Data Model is essential for innovation through data science that facilitates machine learning, integrating and analyzing new and existing data to advance discovery. The emergence of machine learning has brought a computational dimension to data discovery, transformation and data integration that emphasizes pattern recognition over temporal relations among health events. These further bolster the need for standards-based data, storage, and cataloging (Choi, et al., 2017).

Clinical data collected in the process of patient care, research data accumulated during studies and data aggregated from standard external sources comprise a rich tapestry of information that advances the goal of precision medicine. The knowledge gained from the synthesis of these health data pushes Public Health policies and agendas at a national level which eventually percolate to state and local jurisdictions. As described in this paper, there is a growth regarding the standardized, curated data sharing among institutions which form the basis of population-level research. This landscape requires health information to be shareable, standardized and transportable. A Common Data Model is the foundation on which such a vision can translate into reality.

Since 2013 the NIH launched the "Big Data to Knowledge" (BD2K) initiative with the express aim to "support the research and development of innovative and transformative approaches and tools to maximize and accelerate the integration of big data and data science into biomedical research." ("About BD2K | Data Science at NIH," 2012). One of the core aspects of BD2K initiative is an effort to make data Findable, Accessible, Interoperable, and Reusable (FAIR). The NIH Data Commons Pilot seeks to test the feasibility of storing, accessing and sharing of NIH funded common data sets as well as making computational tools and resources (cloud platform) available through public, collaborative platforms. Through this effort, the NIH hopes to develop best practices, architectures, and standards for biomedical big data sets. To this end, cloud service providers like Amazon Web Services (AWS) have published guides for architecting solutions that leverage OMOP CDM. These types of knowledge dissemination help enterprises to reuse architectural strategies, methodologies and realize cost savings in implementing CDMs.

Multiple CDMs exist for merging clinical research and EHR data; each is developed for specifics uses like Patient-centered research to medical device outcomes tracking. Thus the relevance of the CDM depends on the planned uses of the data model. Although this evaluation tried to address the most comprehensive use cases possible, no standard model can cover all the use cases for a diverse community such as the case in Emory University. The focus of this evaluation has been to take a generalized approach by synthesizing the evaluation criteria published in the literature. These synthesized criteria resulted in an amalgam of dimensions first proposed by Moody and Shanks, which was operationalized by Kahn et al., and Garza et al. by applying the assessment to Common Data Models for specific project use cases. For Emory LITS departmental requirement this translated to the criteria of content coverage, integration to exists sources, vocabulary coverage, and mapping, flexibility and ease of querying as well as the readiness of implementing the model.

Overall the OMOP CDM was rated favorably across most evaluation criteria. The ontology coverage, mapping of concepts to source concepts that come preloaded in the standardized vocabulary tables, and the classification concepts make it attractive to LITS. This representation expressivity (Wand and Weber, 1993) means the initial move would require a non-trivial effort to map homegrown custom codes from source systems into the standard terminologies stored in the Observational Medical Outcomes Partnership (OMOP) vocabulary. The Observational Health Data Sciences and Informatics (OHDSI) team created the Usagi software tool ("Usagi," 2017) for translating source specific codes into standard concepts like SNOMED, RxNorm. Usagi uses a term similarity approach for mapping to standardized vocabularies, thus automating the vocabulary mapping and classification which significantly reduces manual curation and speeds implementation. WhiteRabbit ("WhiteRabbit," 2016) is a data profiling tool that scans the source tables, columns, and values to provide a reference report for ETL design (Fig. 15). The growing suite of open source tools provided by OHDSI has made OMOP attractive option for projects like US Precision Medicine Initiative All of US Research Program ("About the All of Us Research Program," n.d.). As the opportunities for accessing open data sources increase, the impact of these tools that work with the OMOP CDM has the potential to be much more significant than those developed with less standardized approaches.

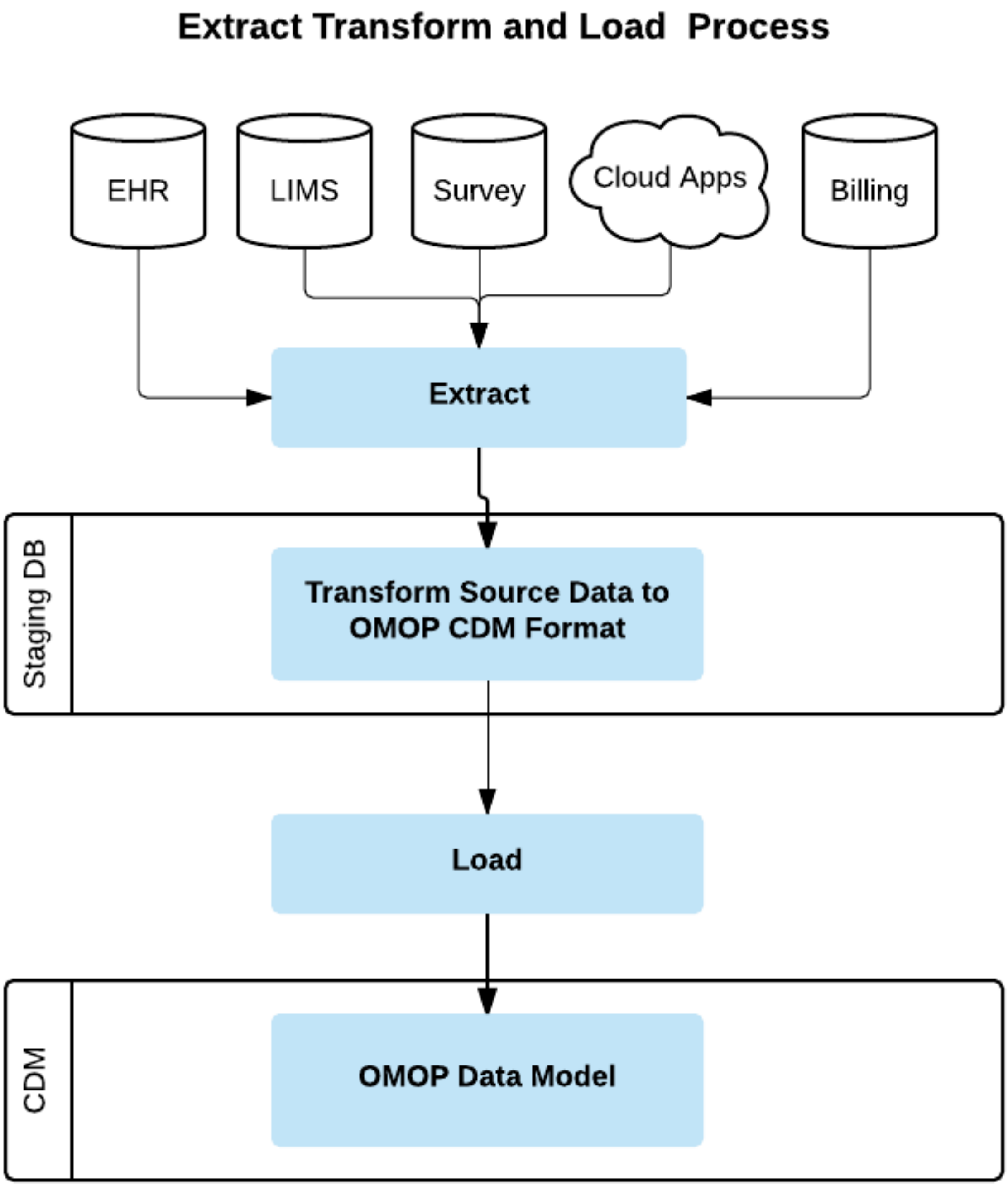

Extract Transform and Load Process
EHR
LIMS
Survey
Cloud Apps
Billing
Extract
Staging DB
Transform Source Data to OMOP CDM Format
Load
CDM
OMOP Data Model

Fig. 15. Extract Transform and Load Process integrating disparate source data to OMOP CDM

Limitations

No discussion for data model standardization would be complete without debating the trade-offs of OMOP or CDMs in general. Standardization by its very nature is moving towards a one size fits all approach, hence opportunities for iterative enhancements exist. One of the limitations of OMOP from Emory LITS perspective is the less than ideal approach to storing survey data. Emory's research community uses REDCap extensively to collect study related information that is outside of the EHR. Currently, the OBSERVATION table will persist the survey data, and the FACT_RELATIONSHIP table will hold the relationships between the fields. This approach makes the SQL queries more complicated and conflicts with the OMOP design principles than if survey data had its entity. Surveys are a domain by itself that is extensively used in the research as well as in public health case reporting forms. Recently the proposal ("SURVEY data in OMOP CDM Issue #137 · OHDSI/CommonDataModel," 2017) has been accepted for adding a SURVEY table in the CDM which preserves the level of normalization as other domains in the model.

Another limitation in OMOP is representing oncology treatment (Fig.16). Cancer treatment involves specific regimens in the complete course of systemic therapy that includes multiple modalities of treatment, as well as the intent of treatment - diagnostic, curative, palliative. Analytics use cases delving into oncology treatment timelines require a higher level abstraction to condition_era and procedure_era data in the OMOP CDM. Presently, the working group has proposed options ("Oncology Treatment Proposal · Issue #163 · OHDSI/CommonDataModel," 2018) for modifying the CDM and is inviting comments.

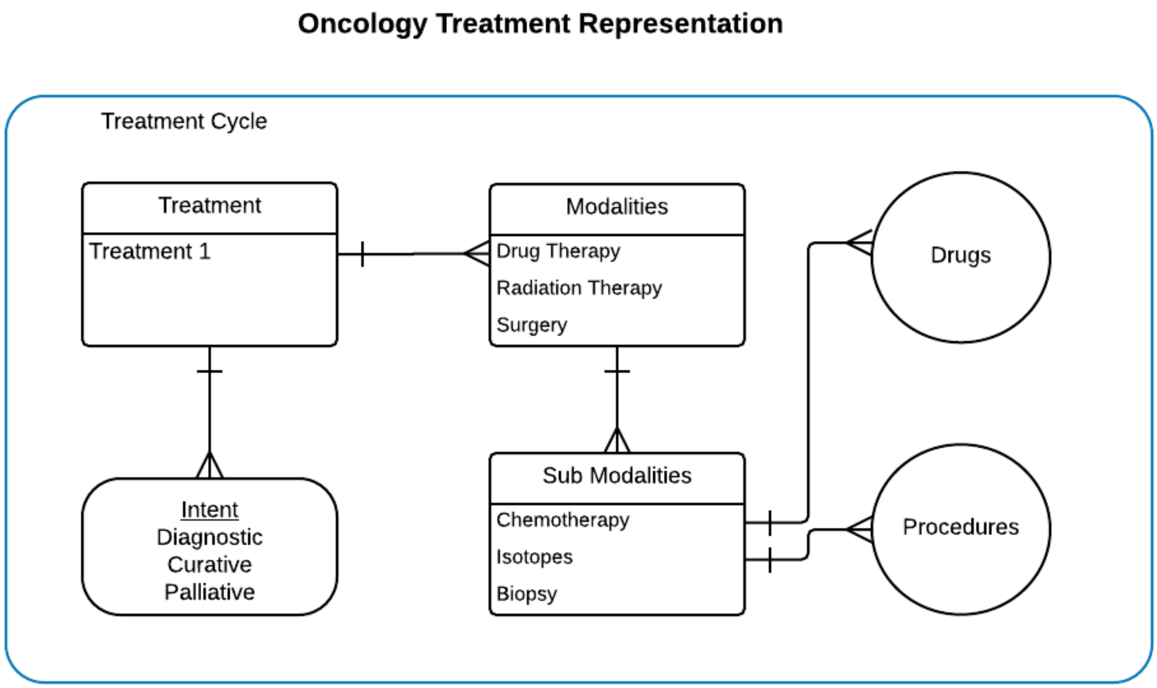


Fig. 16. Shows the Proposed Oncology Treatment representation for OMOP.

Then there is the question of how best to represent discrete events that occur within an inpatient encounter, such as multiple surgeries during a hospital stay. Recording outcomes, complications, medications pertinent to each surgery is a challenge even in healthcare data warehouses. This scenario results in having to use time-based queries at a finer granularity than what typical data warehouses are designed to handle, which inevitably results in multi-pass queries leading to degradation of data retrieval performance. One could handle the surgeries and related information similar to oncology treatments wherein they are categorized as derived time-based events related to PROCEDURE table with an ending date and time to capture procedure duration. Until such time when there are enhancements to the CDM, Emory will need to approach limitations described above with a synergistic data modeling strategy. This strategy involves creating derived custom entities in conjunction with OMOP standardized concept tables. This approach enables Emory University to put forth enhancement proposals to the OMOP CDM designer community based practical implementation experience and serves to further the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles.

Conclusion

The OMOP and PCORnet CDMs were developed for specific purposes, and their fitness for other use cases depend on the secondary uses which are often particular to the project or site that plans to implement the CDM. The evaluation methodology used to assess their fitness to Emory LITS requirements were broad and found OMOP CDM satisfied most of the generalized criteria. The community based approach to data model enhancement is a welcome paradigm where organizations that adopt the CDM can propose, discuss and accept enhancement requests. A participatory approach to model changes leads to robust design-decisions that incorporate best of breed solutions while still being within the design principles of the CDM. Moreover, OHDSI's yearly symposia and tutorials that offer hands-on training on extract transformation and load strategies, open-source toolsets, and recommendations on putting together project teams for a successful implementation are valuable for estimating the costs, resources, and timelines for adopting the CDM. Tools like ATLAS provide an interactive platform to visualize the data in the CDM which can then be used for data quality assurance to ensure that data profile matches the source systems. It even provides data quality reporting and visualizations that notifies gaps in mapping to target entities and helps drill down on potential ETL issues.

References:

1. Corley, D. A., Feigelson, H. S., Lieu, T. A., & McGlynn, E. A. (2015). Building Data Infrastructure to Evaluate and Improve Quality: PCORnet. Journal of Oncology Practice, 11(3), 204–206. http://doi.org.proxy.library.emory.edu/10.1200/JOP.2014.003194
2. Chen, PP-S. The Entity-Relationship Model—Toward a Unified View of Data. ACM Trans Database Syst 1976;1:9–36.
3. G.C. Witt, G.C. Simsion, Data Modeling Essentials: Analysis, Design, and Innovation, The Coriolis Group, 2000.
4. Westland, J.C. (2002). The cost of errors in software development: evidence from industry. Journal of Systems and Software, 62, 1-9.
5. NIH Data Sharing Policy and Implementation Guidance. (2003, March). Retrieved from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
6. Specialized Registry | Meaningful Use | CDC. (2018, March 5). Retrieved March 14, 2018, from http://www.cdc.gov/ehrmeaningfuluse/specialized_registry.html
7. Information model. (2017, December 13). Retrieved March 14, 2018, from https://en.wikipedia.org/wiki/Information_model
8. Agency for Healthcare Research & Quality. (n.d.). Clinical Decision Support. Retrieved March 14, 2018, from https://www.ahrq.gov/professionals/prevention-chronic-care/decision/clinical/index.html
9. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, Raebel MA, Beaulieu NU, Rosofsky R, Woodworth TS, Brown JS. (2012) Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiology and Drug Safety 21(S1):23-31

10. Sentinel Common Data Model | Sentinel Initiative. (2017, October 4). Retrieved from https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/sentinel-common-data-model

11. ncats/CTSA-Metrics. (18, January 17). Retrieved March 6, 2018, from https://github.com/ncats/CTSA-Metrics

12. Common Metrics. (n.d.). Retrieved March 14, 2018, from https://clic-ctsa.org/common_metrics

13. 2015 Edition Health Information Technology (Health IT) Certification Criteria, 2015 Edition Base Electronic Health Record (EHR) Definition, and ONC Health IT Certification Program Modifications. (2015, October 16). Retrieved March 14, 2018, from https://www.federalregister.gov/documents/2015/10/16/2015-25597/2015-edition-health-information-technology-health-it-certification-criteria-2015-edition-base

14. Microsoft. (2017, April 24). Common Data Model home page. Retrieved from https://docs.microsoft.com/en-us/common-data-service/entity-reference/common-data-model

15. Moody D.L., Shanks G.G. (1994) What makes a good data model? Evaluating the quality of entity relationship models. In: Loucopoulos P. (eds) Entity-Relationship Approach — ER '94 Business Modelling and Re-Engineering. ER 1994. Lecture Notes in Computer Science, vol 881. Springer, Berlin, Heidelberg

16. Moody DL, Shanks GG. Improving the quality of data models: Empirical validation of a quality management framework. Inf Syst.2003;28:619–650.

17. Kahn, M. G., Batson, D., & Schilling, L. M. (2012). Data Model Considerations for Clinical Effectiveness Researchers. Medical Care, 50(0), 10.1097/MLR.0b013e318259bff4. http://doi.org/10.1097/MLR.0b013e318259bff4

18. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016;64:333–41.

19. Buneman P., Khanna S., Tan WC. (2000) Data Provenance: Some Basic Issues. In: Kapoor S., Prasad S. (eds) FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science. FSTTCS 2000. Lecture Notes in Computer Science, vol 1974. Springer, Berlin, Heidelberg

20. OMOP Common Data Model – OHDSI. (2017). Retrieved from https://www.ohdsi.org/data-standardization/the-common-data-model/

21. PCORnet Common Data Model (CDM) - PCORnet. (2018, February 19). Retrieved from http://www.pcornet.org/pcornet-common-data-model/

22. ICD - ICD-10-CM - International Classification of Diseases, Tenth Revision, Clinical Modification. (2017). Retrieved from https://www.cdc.gov/nchs/icd/icd10cm.htm

23. LOINC Basics — LOINC. (2018). Retrieved from https://loinc.org/faq/basics/

24. National Provider Identifier Standard (NPI) - Centers for Medicare & Medicaid Services. (2015, March 6). Retrieved from https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/NationalProvIdentStand/

25. RxNorm. (2004, March 22). Retrieved from https://www.nlm.nih.gov/research/umls/rxnorm/

26. OMOP CDM Queries. (n.d.). Retrieved March 14, 2018, from http://cdmqueries.omop.org/home

27. Paul A. Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, Jose G. Conde, Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support, J Biomed Inform. 2009 Apr;42(2):377-81

28. About the All of Us Research Program. (n.d.). Retrieved March 18, 2018, from https://allofus.nih.gov/about/about-all-us-research-program

29. About BD2K | Data Science at NIH. (2012). Retrieved March 18, 2018, from https://datascience.nih.gov/bd2k/about

30. Edward Choi, Andy Schuetz, Walter F Stewart, Jimeng Sun; Using recurrent neural network models for early detection of heart failure onset, Journal of the American Medical Informatics Association, Volume 24, Issue 2, 1 March 2017, Pages 361–370, https://doi.org/10.1093/jamia/ocw112

31. Wand, Y. and Weber, R. (1993), On the ontological expressiveness of information systems analysis and design grammars. Information Systems Journal, 3: 217-237. doi:10.1111/j.1365-2575.1993.tb00127.x

32. Usagi. (2017, February 28). Retrieved March 18, 2018, from http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:usagi

33. WhiteRabbit. (2016, December 22). Retrieved March 18, 2018, from http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:whiterabbit

34. SURVEY data in OMOP CDM - updated · Issue #137 · OHDSI/CommonDataModel. (2017, November 23). Retrieved March 18, 2018, from https://github.com/OHDSI/CommonDataModel/issues/137

35. Oncology Treatment Proposal · Issue #163 · OHDSI/CommonDataModel. (2018, February 9). Retrieved March 18, 2018, from https://github.com/OHDSI/CommonDataModel/issues/163

36. Rosenbloom, S. T., Carroll, R. J., Warner, J. L., Matheny, M. E., & Denny, J. C. (2017). Representing Knowledge Consistently Across Health Systems. *Yearbook of Medical Informatics,26*(01), 139-147. doi:10.15265/iy-2017-018