The role of terminologies in health data analytics through common data models

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.
Outline

◆ The context of health data analytics
  ● Data models
  ● Terminology integration
◆ Observational Health Data Sciences and Informatics (OHDSI)
“Common” data models

- OMOP
- i2b2
- PCORnet
- Sentinel
- CDISC
OMOP

- OMOP – Observational Medical Outcomes Partnership

![Diagram showing data flow from multiple sources to analysis method and common data model](image-url)
- Standardized Clinical Data Tables
  - PERSON
  - OBSERVATION_PERIOD
  - SPECIMEN
  - DEATH
  - VISIT_OCCURRENCE
  - PROCEDURE_OCCURRENCE
  - DRUG_EXPOSURE
  - DEVICE_EXPOSURE
  - CONDITION_OCCURRENCE
  - MEASUREMENT
  - NOTE
  - NOTE_NLP (V5.5.2)
  - OBSERVATION
  - FACT_RELATIONSHIP
- Standardized Health System Data Tables
  - LOCATION
  - CARE_SITE
  - PROVIDER
- Standardized Health Economics Data Tables
  - PAYER_PLAN_PERIOD
  - COST (V5.0.1)
  - VISIT_COST - removed
  - PROCEDURE_COST - removed
  - DRUG_COST - removed
  - DEVICE_COST - removed
- Standardized Derived Elements
  - COHORT
  - COHORT_ATTRIBUTE
  - DRUG_ERA
  - DOSE_ERA
  - CONDITION_ERA
- Standardized Vocabularies
  - CONCEPT
  - VOCABULARY
  - DOMAIN
  - CONCEPT_CLASS
  - CONCEPT_RELATIONSHIP
  - RELATIONSHIP
  - CONCEPT_SYNONYM
  - CONCEPT_ANCESTOR
  - SOURCE_TO_CONCEPT_MAP
  - DRUG_STRENGTH
  - COHORT_DEFINITION
  - ATTRIBUTE_DEFINITION
- Standardized meta-data
  - CDM_SOURCE
i2b2

- i2b2 – Informatics for Integrating Biology & the Bedside
- Originally developed by the i2b2 National Center for Biomedical Computing (2004-2013)
  - Now i2b2 tranSMART Foundation
- Platform to support translational research
- Widely adopted worldwide

https://www.i2b2.org/
i2b2 data model – original “star schema”
i2b2-OMOP convergence

◆ i2b2 on OMOP
  ● Supports query formulation against an OMOP-compliant data source through i2b2 tools
PCORnet – National Patient-Centered Clinical Research Network

Initiative of the Patient-Centered Outcomes Research Institute (PCORI)

- Funded through the Patient Protection and Affordable Care Act of 2010

“designed to make it faster, easier, and less costly to conduct clinical research”

Made up of

- 13 Clinical Data Research Networks (CDRNs)
- 20 Patient-Powered Research Networks (PPRNs)
# PCORnet Common Data Model v3.0

<table>
<thead>
<tr>
<th>DEMOGRAPHIC</th>
<th>ENROLLMENT</th>
<th>DISPENSING</th>
<th>CONDITION</th>
<th>DEATH</th>
<th>DEATH_CONDITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>PATID</td>
<td>PATID</td>
<td>PATID</td>
<td>PATID</td>
<td>PATID</td>
<td>PATID</td>
</tr>
<tr>
<td>BIRTH_DATE</td>
<td>ENR_START_DATE</td>
<td>DISPENSINGID</td>
<td>ENCOUNTERID</td>
<td>DEATH_DATE</td>
<td>DEATH_CAUSE_CODE</td>
</tr>
<tr>
<td>BIRTH_TIME</td>
<td>ENR_END_DATE</td>
<td>(optional)</td>
<td>PATIENT_ID</td>
<td>DEATH_DATE</td>
<td>DEATH_CAUSE_TYPE</td>
</tr>
<tr>
<td>SEX</td>
<td>ENR_BASIS</td>
<td></td>
<td></td>
<td>DEATH_DATE</td>
<td>DEATH_CAUSE</td>
</tr>
<tr>
<td>HISPANIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DEATH_CAUSE_CONFIDENCE</td>
</tr>
<tr>
<td>RACE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DEATH_CAUSE_CONFIDENCE</td>
</tr>
<tr>
<td>BIOBANK_FLAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fundamental basis**

- Vital
  - VitalID
  - PATID
  - ENCOUNTERID (optional)
  - MEASURE_DATE
  - MEASURE_TIME
  - VITAL_SOURCE
  - HT
  - WT
  - DIABOLIC
  - SYSTOLIC
  - ORIGIN
  - TUMOR_TYPE

- Enrollment
  - PATID
  - ENR_START_DATE
  - ENR_END_DATE
  - CHART
  - ENR_BASIS

- Dispensing
  - PATID
  - DISPENSINGID (optional)
  - DISPENSE_DATE
  - NDC
  - DISPENSE_SUP
  - DISPENSE_AMT

- Condition
  - CONDITIONID
  - PATID
  - ENCOUNTERID (optional)
  - REPORT_DATE
  - RESOLVE_DATE
  - ONSET_DATE
  - CONDITION_STATUS
  - CONDITION_TYPE
  - CONDITION_SOURCE

- Lab Result CM
  - PATID
  - ENCOUNTERID (optional)
  - LAB_NAME
  - SPECIMENT_SOURCE
  - LAB_LOINC
  - RESULT_NUM
  - RESULT_QUAL
  - RESULT_DATE
  - RESULT_TIME
  - RESULT_UNITS
  - RESULT_VENDOR

- Diagnosis
  - PATID
  - ENCOUNTERID
  - ENCTYPE (replacated)
  - PATENT (replacated)
  - PROVIDERID
  - PCX_DATE
  - PCX_TYPE

- Pro CM (replacated)
  - PATID
  - ENCOUNTERID (optional)
  - PRO_ITEM
  - PRO_LOINC
  - PRO_TIME
  - PRO_RESPOND
  - PRO_METHOD
  - PRO_MODE
  - PRO_CAT

- Procedures
  - PROCEDURE_ID
  - PATID
  - ENCOUNTERID
  - ENCTYPE (replacated)
  - PROVIDERID
  - PDX

- Prescribing
  - PRESCRIBINGID
  - PATID
  - ENCOUNTERID (optional)
  - RX_PROVIDERID
  - RX_ORDER_DATE
  - RX_ORDER_TIME
  - RX_END_DATE
  - RX_QUANTITY
  - RX_VITAL
  - RX_FREQUENCY

- Harvest
  - NETWORKID
  - NETWORK_NAME
  - DATAMART
  - DATAMART_NAME
  - DATAMART_PLATFORM
  - CDM_VERSION
  - DATAMART_CLAIMS
  - DATAMART_EHR
  - BIRTH_DATE
  - ENR_START_DATE
  - ENR_END_DATE
  - ADMIT_DATE
  - DISCHARGE_DATE
  - LAB_ORDER_DATE
  - SPECIMENT_DATE
  - SPECIMENT_TIME

**Data captured within multiple contexts:** healthcare delivery, registry activity, or directly from patients

**Data captured from healthcare delivery, direct encounter basis**

**Bold font indicates fields that cannot be null due to primary key definitions or record-level constraints.**
Sentinel

- Initiative of the Food and Drug Administration (FDA)
- Effort to create a national electronic system for monitoring the performance of FDA-regulated medical products (drugs, vaccines, and other biologics)
- Develop a system to obtain information from existing electronic health care data from multiple sources to assess the safety of approved medical products
- Distributed dataset reached 100 lives in 2011

https://www.fda.gov/Safety/FDAsSentinelInitiative/
# Sentinel Common Data Model

## List of Tables

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Enrollment</td>
<td>Created by Data Partners using Data Partner data.</td>
<td>The SCDM Enrollment Table has a start/stop structure that contains one record per continuous enrollment period. Members with medical coverage, drug coverage, or both should be included. A unique combination of PatiID, Enr_Start, Enr_End, MedCov, DrugCov, and Chart identifies a unique record. A break in enrollment (of at least one day) or a change in either the medical or drug coverage variables should generate a new record.</td>
</tr>
<tr>
<td>2. Demographic</td>
<td>Created by Data Partners using Data Partner data.</td>
<td>The SCDM Demographic Table contains one record per PatiID with the most recent information on Birth_Date, Sex, Race/Ethnicity, and Zip Code.</td>
</tr>
<tr>
<td>3. Dispensing</td>
<td>Created by Data Partners using Data Partner data.</td>
<td>The SCDM Outpatient Pharmacy Dispensing Table contains one record per unique combination of PatiID, NDC, and RxDate. Each record represents an outpatient pharmacy dispensing. Rollback transactions and other adjustments should be processed before populating this table.</td>
</tr>
<tr>
<td>4.1 Encounter</td>
<td>Created by Data Partners using Data Partner data.</td>
<td>The SCDM Encounter Table contains one record per PatiID and EncounterID. Each encounter should have a single record in the SCDM Encounter Table. Each diagnosis and procedure recorded during the encounter should have a separate record in the Diagnosis or Procedure Tables. Multiple visits to the same provider on the same day should be considered one encounter and should include all diagnoses and procedures that were recorded during those visits. Visits to different providers on the same day, such as a physician appointment that leads to a hospitalization, should be considered multiple encounters. Rollback transactions and other adjustments should be processed before populating this table.</td>
</tr>
<tr>
<td>4.2 Diagnosis</td>
<td>Created by Data Partners using Data Partner data.</td>
<td>The SCDM Diagnosis Table contains one record per unique combination of PatiID, EncounterID, DX, and DX_CodeType. This table should capture all uniquely recorded diagnoses for all encounters.</td>
</tr>
<tr>
<td>4.3 Procedure</td>
<td>Created by Data Partners using Data Partner data.</td>
<td>The SCDM Procedure Table contains one record per unique combination of PatiID, EncounterID, PX, PX_CodeType. This table should capture all uniquely recorded procedures for all encounters.</td>
</tr>
<tr>
<td>5.1 Death</td>
<td>Created by Data Partners using Data Partner data.</td>
<td>The SCDM Death Table contains one record per PatiID. When legacy data have conflicting reports, make a local determination as to which to use. There is typically a 1-2 year lag in death registry.</td>
</tr>
<tr>
<td>5.2 Cause of Death</td>
<td>Created by Data Partners using Data Partner data.</td>
<td>The SCDM Cause of Death Table contains one record per unique combination of PatiID and COD. Legacy data have conflicting reports, please make a local determination as to which to use. There is typically a 1-2 year lag in death registry data.</td>
</tr>
<tr>
<td>6.1 Laboratory Result</td>
<td>Created by Data Partners using Data Partner data.</td>
<td>The SCDM Laboratory Result Table contains one record per result/entry. Only include results if Data Partners are strongly encouraged to review the comprehensive Sentinel Common Data Model Laboratory Result Table Documentation for details on how to populate each variable.</td>
</tr>
</tbody>
</table>
Role of terminologies

◆ Normalization
  ● Different datasets may be annotated in reference to different terminologies
  ● Identify equivalent (or close) concepts across terminologies

◆ Aggregation
  ● Queries are generally formulated at a high-level
  ● Terminologies support aggregation
    ▪ Transitive closure
    ▪ Value sets
Terminologies used for health data analytics

- **Main clinical terminologies for the Meaningful Use incentive program (clinical documentation; clinical quality measures)**
  - SNOMED CT
  - LOINC
  - RxNorm

- **Legacy terminologies (billing)**
  - [ICD9-CM]; ICD10-CM
  - CPT

- **Other terminologies (CDISC)**
  - NCI Thesaurus
Binding between terminology and information model

◆ Often involves mapping
◆ Can happen at various stages along of the integration process
  ● ETL – Extract / Transform / Load
  ● Query (query translation)
◆ Issues
  ● Integrating heterogeneous datasets
    ■ Different terminologies
    ■ Different levels of granularity
Terminology integration

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Anatomy
- Model organisms
- Other subdomains
Terminology integration

Addison's disease (363732003)

Clinical repositories

NCBI Taxonomy

SNOMED CT

OMIM

Model organisms

FMA

Anatomy

Genome annotations

Other subdomains

Genetic knowledge bases

Biomedical literature

Addison Disease (D000224)

MeSH

Clinical repositories

SNOMED CT

OMIM

FMA

Anatomy

Genome annotations

Other subdomains

Genetic knowledge bases

Biomedical literature

Addison Disease (D000224)
Observational Health Data Sciences and Informatics (OHDSI)
OHDSI Outline

- From OMOP to OHDSI
- Foundational principles
- OHDSI software, test data and methods
- Use cases and research
  - PNAS paper
From OMOP to OHDSI

◆ OMOP – Observational Medical Outcomes Partnership
  ● Public-private partnership established to inform the appropriate use of observational healthcare databases for studying the effects of medical products (2008-2013)
  ● Community of researchers from industry, government, and academia
  ● Achievements
    ■ Conduct methodological research to empirically evaluate the performance of various analytical methods on their ability to identify true associations and avoid false findings
    ■ Develop tools and capabilities for transforming, characterizing, and analyzing disparate data sources across the health care delivery spectrum
    ■ Establish a shared resource so that the broader research community can collaboratively advance the science

http://omop.org
From OMOP to OHDSI

OHDSI – Observational Health Data Sciences and Informatics

- Multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics
- International network of researchers and observational health databases with a central coordinating center housed at Columbia University
- Continues to actively use the OMOP Common Data Model and Standardized Vocabularies
- Develops open-source solutions [with Greek names]
- Annual symposium

https://www.ohdsi.org/
Foundational principles

◆ Data standardization through
  ● Common data model (OMOP CDM)
  ● Standard vocabularies
◆ Conversion (ETL) of the local clinical data warehouse to the OMOP CDM and standard vocabularies
  ● Supported by the WhiteRabbit tool
◆ Applicable to various types of observational data (EHR, claims)
◆ Data remain local to a clinical institution
◆ The same query can be executed at each site and the results aggregated across sites
◆ Research projects are based on rigorous protocols
◆ Open-source software
OHDSI software

- **ATLAS** – unified interface to multiple OHDSI tools
- **ATHENA** – access to standardized vocabularies
- **ACHILLES** – database characterization and data quality assessment
- **CALYPSO** – analytical component for clinical study feasibility assessment
- **CIRCE** – cohort creation
- **HERACLES** – cohort-level analysis and visualization
- **LAERTES** – system for investigating the association of drugs and health (adverse events)
- **DRUG EXPOSURE EXPLORER** – visualize drug exposures (an experimental deployment using the SynPUF 1% simulated patient data set)
OHDSI methods

◆ Population-Level Estimation
  ● Safety surveillance
  ● Comparative effectiveness

◆ Patient-Level Prediction

◆ Implemented with open-source tools for large-scale analytics
  ● R packages
Examples of network research studies

- Comparison of combination treatment in hypertension
- Comparative effectiveness of alendronate and raloxifene in reducing the risk of hip fracture
- Levetiracetam and risk of angioedema in patients with seizure disorder
- Drug utilization in children
- Characterizing treatment pathways at scale using the OHDSI network
Characterizing treatment pathways at scale using the OHDSI network

George Hripcsak\textsuperscript{a,b,c,1}, Patrick B. Ryan\textsuperscript{c,d}, Jon D. Duke\textsuperscript{c,e}, Nigam H. Shah\textsuperscript{c,f}, Rae Woong Park\textsuperscript{c,g}, Vojtech Huser\textsuperscript{c,h}, Marc A. Suchard\textsuperscript{c,i,j,k}, Martijn J. Schuemie\textsuperscript{c,d}, Frank J. DeFalco\textsuperscript{c,d}, Adler Perotte\textsuperscript{a,c}, Juan M. Banda\textsuperscript{c,f}, Christian G. Reich\textsuperscript{c,l}, Lisa M. Schilling\textsuperscript{c,m}, Michael E. Matheny\textsuperscript{c,n,o}, Daniella Meeker\textsuperscript{c,p,q}, Nicole Pratt\textsuperscript{c,r}, and David Madigan\textsuperscript{c,s}

www.pnas.org/cgi/doi/10.1073/pnas.1510502113

PNAS | July 5, 2016 | vol. 113 | no. 27 | 7329–7336
Characterizing treatment pathways at scale using the OHDSI network

- Objectives: analyze the variability of pharmacological treatment interventions over three years across three diseases (type-2 diabetes mellitus, hypertension, or depression)

- Inclusion criteria: exposure to an antidiabetic, antihypertensive, or antidepressant medication for 3 years, as well as presence of at least one diagnostic code for the corresponding disease

- Exclusion criteria: based on diagnostic data (e.g., exclusion of schizophrenia patients from the depression cohort)
Characterizing treatment pathways at scale using the OHDSI network

- **Materials:** 11 datasets representing a total of 255 million patients
  - EHR data (South Korea, U.K., U.S.) 67M
  - Claims data (U.S., Japan) 188M

- **Methods:** Analyze the sequences of medications that patients were placed on during those 3 years, to reveal patterns and variation in treatment among data sources and diseases
Characterizing treatment pathways at scale using the OHDSI network

◆ Results

● Patients with 3 years of uninterrupted therapy
  - 327,110 diabetes patients
  - 1,182,792 hypertension patients
  - 264,841 depression patients

● Treatment pathways

A Diabetes
A Diabetes

- Metformin
- pioglitazone
- sitagliptin
- Glipizide
- glimepiride
- Gliclazide
- Glyburide
- rosiglitazone
- Insulin, Glargine, Human
- exenatide
- Insulin, Aspart, Human
- liraglutide
- saxagliptin
- Insulin, Lispro, Human
- Glucose
- Insulin, Isophane, Human
Differences across diseases

- **Diabetes**
  - Metformin is the first line of treatment and often the only treatment

- **Hypertension**
  - Slight predominance of HCTZ, frequently paired with other medications

- **Depression**
  - Even spread of medications

- **Unique treatment pathways (within a cohort)**
  - 10% TDM
  - 25% HTN
Differences across countries

- Metformin less often used in Japan
- Wide variety of starting medications
- The most common medication varies by source
All of Us – Precision Medicine Initiative

The future of health begins with All of Us

The All of Us Research Program is a historic effort to gather data from one million or more people living in the United States to accelerate research and improve health. By taking into account individual differences in lifestyle, environment, and biology, researchers will uncover paths toward delivering precision medicine.
Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: https://mor.nlm.nih.gov

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA