

Observational Medical Outcomes Partnership:

Points to Consider in Developing a Common Semantic Data Model and Terminology Dictionary for Observational Analyses

Patrick Ryan¹, Don Griffin², Luann Whittenburg², Dan Foltz², Marc Overhage³

¹GlaxoSmithKline, ²Computer Sciences Corporation, ³Regenstrief Institute

Last revised: 3 March 2009

Table of Contents

1. Purpose of Document.....	2
2. Overview of Problem.....	2
2.1. Introduction to the Observational Medical Outcomes Partnership.....	2
2.2. Introduction to the Needs of a Common Data Model (CDM).....	3
2.3. Types of Observational Data	4
2.4. Types of Observational Analyses	5
2.5. Portability of research methods	6
2.6. Process for Developing the Common Data Model	7
3. Defining the Common Data Model.....	9
3.1. CDM Requirements	10
3.2. CDM Design Approach	11
3.3. Proposed CDM Design	12
3.4. The Role of the Terminology Dictionary in Querying the CDM	17
3.5. Mapping Raw Source Data to Standard Concept Codes	21
3.6. Deciding Upon a Standard Concept Hierarchy for the Terminology Dictionary	21
4. Overview of Observational Healthcare Data Elements	23
4.1. Guidelines for Data Element Definitions.....	23
4.2. Data Transformation Example.....	26
4.3. Guidelines for Extending the CDM with Additional Data Elements	30
5. Conclusion	31

1. Purpose of Document

This document outlines key points to consider when developing a uniform query and interface environment applicable to medical observational databases (administrative claims and electronic health records) for active drug surveillance. The document introduces the specific problem of active drug surveillance in observational data, describes the observational data elements expected to be within scope, and provides initial Common Data Model and Terminology Dictionary designs to stimulate dialogue. The immediate intentions of the Common Data Model and Terminology Dictionary designs are to facilitate analyses of methods and outcomes research across the Partnership's research data sets, and the expectation that the OMOP design, implementation, and objective evidence about the Common Data Model and Terminology Dictionary may inform future efforts as well.

2. Overview of Problem

2.1. Introduction to the Observational Medical Outcomes Partnership

The Observational Medical Outcomes Partnership (OMOP) is a public-private partnership designed to protect human health by improving the monitoring of drugs for safety and effectiveness. The partnership will conduct a two-year research initiative to determine whether it is feasible and useful to identify and evaluate safety issues of drugs on the market.

The Partnership's methodological research will be conducted across multiple disparate observational databases (administrative claims and electronic health records) and plans to engage in collaborations with qualified organizations in a number of different ways: The Partnership is funding data provider organizations to participate in the initiative, either as a Research Core contributor by providing de-identified patient-level data into OMOP's centralized IT research environment, or as a distributed partner conducting the analysis within its organization and reporting back aggregate summary results to the Partnership. In addition, the Partnership will be encouraging other organizations with access to observational data to participate in conducting supplementary analyses as part of an Extended Consortium.

In order to facilitate this research, there is a need to develop a common structure and framework for organizing and standardizing observational data that will enable the consistent application of analyses across disparate data sources. A Common Data Model and a method for standardizing its content (via a Terminology Dictionary) will ensure methods can be systematically applied to produce meaningfully comparable results across sources. More generally, it is recognized that no single observational data source is likely to be sufficient to meet all expected drug safety analysis needs, so it can be anticipated that solutions will be required that can analyze combined (integrated) disparate sources. The Partnership will study the use of a Common Data Model and a Terminology Dictionary as potential enabling technologies for its own methodological research and with intentions for future applications to support broader pharmacovigilance activities.

2.2. Introduction to the Needs of a Common Data Model (CDM)

A data model defines the categories and the relationships between disparate data entities within a particular information system environment, thus establishing the context within which those entities have meaning. Several data models have been proposed within the healthcare domain. It is important to recognize and learn from these past efforts, while clearly focusing the intended purposes.

The CDM acknowledges the Health Level Seven (HL7) Reference Information Model (RIM). HL7 provides a framework and related standards for the exchange, integration, sharing and retrieval of electronic health information. The standards can support clinical practice and the management, delivery, and evaluation of health services. The RIM expresses the data content needed in a specific clinical or administrative context, and provides an explicit representation of the semantic and lexical connections that exist between the information carried in the fields of HL7 messages.

The Healthcare Information and Management Systems Society (HIMSS) Electronic Health Record Definitional Model is another effort to develop guidelines for a common data model. This work focuses on the needs for standards in developing electronic health records databases to support clinical practice.

The OMOP Common Data Model is intended to facilitate observational analyses from disparate healthcare databases (administrative claims and electronic health records). It is expected that the Common Data Model should include all data elements that may be used in such analyses, but the goal is not necessarily to provide a mechanism to archive all healthcare data elements. For example, cost information, which is an important aspect of healthcare administrative processes, may not play a prominent role in identifying associations between drug exposure and outcome occurrence.

OMOP intends to design a Common Data Model and agree to standard practice in transforming observational data into the common model content as a preliminary step to enabling analysis. The design and implementation will require the development of a Terminology Dictionary that contains standard terminologies and that drives agreed algorithms for data transformation. All analysis methods and code used to execute the research protocols will be developed for the Common Data Model, with the express purpose of enabling a common set of procedures to be applied to each participating data source. OMOP intends to test the feasibility of both distributed and centralized network architectures to enable observational analyses across disparate observational data sources. It is expected that all participating data sources will be transformed into the Common Data Model structure and terminology standards, regardless of where the data reside.

2.3. Types of Observational Data

Two classes of observational data that are the focus of this effort are administrative claims databases and clinical records (electronic health records: EHRs). Each type has its own advantages and limitations, and specific data sources may have unique features that need to be well-understood and carefully considered when conducting observational analyses and interpreting results.

Administrative claims databases contain health information for large numbers of persons with period of coverage based on insurance. Patient medication can be extracted from pharmacy claims of filled prescriptions. Conditions can be captured, with some limitations, from diagnosis codes on inpatient and outpatient medical billing claims. Insurance claims data have the advantage of very large sample size, and of representing a generally comprehensive summary of health-related activities during enrollment. However, claims databases may not adequately capture symptoms or other important aspects of the patients' medical histories, or may reflect a reimbursement translation of a clinical problem (e.g. coding myocardial infarction to justify a procedure when the procedure result in fact rules out the diagnosis of MI). Some claims databases represent a biased, privately insured population of employees, which may not be generalizable to other populations of interest.

Electronic health records reflect information collected during clinical care of the patient. Medication data can be extracted from a variety of sources, including prescriptions written by the provider, medication history lists, and prescriptions filled. Conditions can be identified from problem lists of diagnoses, symptoms, and other components of medical history. Because the data are collected with the intent of enabling the physician to provide quality care, electronic health records offer the promise of having more granular and precise information about personal health status from that provider. However health services received at other locations (such as inpatient care) may not be adequately represented; data may not be well or consistently structured and incomplete.

Because EHR and claims databases record and maintain different data elements for different purposes, the structures of these data types are often quite different. Claims data may be organized by the type of claim (inpatient, outpatient, pharmacy) with information contained within each claim from a given person that can be linked by some personal identifier to other claims for that same person. In contrast, EHR data may be organized, for example, by personal health encounter, with drugs and conditions observed being logically associated with a particular episode of care. Several data models have been proposed for these specific purposes. For example, the HMO Research Network developed the Virtual Data Warehouse as a tool to enable analyses across multiple organizations with claims data¹. In contrast, the Informatics for Integrating Biology and the Bedside (i2b2) initiative developed an entity-attribute-value (EAV) model to store clinical observations like those that could be collected in an EHR in a scalable form². While both models appear to be quite successful in satisfying their original purpose, it is

¹ http://www.hmoresearchnetwork.org/resources/toolkit/HMORN_VDWDetailedDataStructures.pdf

² https://www.i2b2.org/software/projects/datarepo/CRC_Design_Doc_13.pdf

quite possible that OMOP's common data model for active drug surveillance across both claims and EHR sources can be informed by these efforts but will result in a new model accommodating both perspectives.

In order to conduct analyses to identify and evaluate associations between drugs and outcomes, observational data must be maintained in a structure that facilitates rapid access while maintaining all necessary relations across data elements. The Common Data Model is critical to this end, because it "neutralizes" the source-specific variations in data structure, while preserving the relevant relationships and meanings inherent in the data sources. The Terminology Dictionary plays a critical role, too, by allowing—indeed, requiring—the Partnership's researchers to query the Common Data Model in a way that further neutralizes source-specific idiosyncrasies in the data, by standardizing the data based on meaning rather than representation. However, the Common Data Model and Terminology Dictionary cannot be of arbitrary design. They must be based on the specific types of observational analyses that the Partnership's researchers intend to conduct.

2.4. Types of Observational Analyses

There are three distinct types of analysis within scope of the Partnership's research. Each analysis type presents different methodological challenges, may require different algorithms and may utilize different data elements within the common data model. The three analysis types are:

- **Identification of non-specified associations:** This exploratory analysis aims to generate hypotheses from observational data by identifying associations between drugs and conditions for which the relationships were previously unknown. This type of analysis is likely to be considered an initial step of a triaged review process, where many drug-outcome pairs are simultaneously explored to prioritize the drugs and outcomes that warrant further attention. It could be expected that a primary consideration for identification analyses is developing an efficient model to allow high-throughput computing across large sets of potential hypotheses about drug-outcome relationships. Method thresholds can be evaluated on the basis of the tradeoff between observed true positives, false positives, true negatives, and false negatives across the results for a given analysis.
- **Monitoring of Health Outcomes of Interest:** The goal of this surveillance analysis is to monitor the relationship between a series of drugs and specific outcomes of interest. These analyses require an effective definition of the events of interest in the context of the available data (e.g. 'acute liver injury' may best be defined by a combination of medical diagnoses, pharmacy records, procedure codes, and/or laboratory results). This is in contrast to the first analysis type, which may concurrently explore many outcomes for a given drug. Where possible, outcomes definitions may be validated within the observational sources to provide broader context for interpreting analysis results.

- **Evaluation of a drug-condition association:** This hypothesis-strengthening analysis is consistent with traditional pharmacoepidemiology practice for comprehensive observational studies. Evaluation studies may require particular data elements specific to the study in question, and will commonly apply multivariate statistics like linear, logistic, or Poisson regression. Evaluations may require specific customization if standard transformations are deemed inappropriate for the particular hypothesis being tested.

For observational analyses, it is important to recognize that the goal is to provide information about associations between drugs and outcomes across a population of interest. The intended objective is not necessarily to conclusively ascertain whether a specific person had a particular outcome due to a particular drug, but instead to infer whether a population of individuals exposed to a product experiences more of the outcome than otherwise expected had they been unexposed. This population-based approach differs from the spontaneous adverse event reporting systems, which considers each data record a specific self-report of a suspected causal association between a drug and an event.

For each type of analysis, the particular method in use may require an analysis dataset in a specific format, such as a single table in which rows represent persons and columns represent indicator variables for the existence of drug exposure or condition occurrence. The OMOP Common Data Model must facilitate the efficient production of analysis datasets. Consideration should be made to determine what data management activities can be pre-processed within the data model as an expedient, and what data manipulations should be reserved for run-time as part of the analysis process. Given that different analysis types may have different specific requirements, it could be anticipated that the Common Data Model may include multiple parallel data tables to facilitate specific tasks. It could also be anticipated that the Terminology Dictionary may include multiple terminologies, to facilitate encoding query result sets with multiple standard codes or terms.

2.5. Portability of research methods

One of the Partnership's goals is to create and promulgate data-driven research methods that are portable across the Research Core data providers. The Partnership intends to ensure this portability of research methods through a variety of means, including establishing policies, procedures, and controls on research data, applications, methods, and the governance thereof. One of the vehicles for enforcing these policies, procedures, and controls is the OMOP Common Data Model (CDM).

The Common Data Model, and the source-specific data extraction, transformation, and loading (ETL) logic that targets it, allows each disparate data source to be standardized from its native form into a structure that is common across all data sources. If all data sources can be consistently transformed into one common data structure (i.e., the CDM), then all data analysis methods can be conducted on the CDM, obviating the need to develop separate (i.e., custom) analysis methods for each data source. Ideally, we would develop one custom transformation of

each data source into the CDM, and then develop for each analysis method one custom application that accesses or extracts the necessary data from the CDM.

Note: The Partnership does not intend to use the CDM as target for combining data from multiple sources into one consolidated database since source-specific characteristics need to be preserved. The goal of the CDM is to allow data standardization. That is, the OMOP Common Data Model accommodates divergent data sources, by transforming them into a consistent form for analysis.

Organizations wishing to participate in OMOP-sponsored research, either as Distributed Partners or within the Extended Research Consortium, must adhere to the OMOP data standardization strategy. This requires the following three actions.

1. **Instantiate the OMOP Common Data Model** in a credible relational database management system (RDBMS)
2. **Transform the content of existing source database(s)**, by linking selected raw source data to the **standard concept hierarchy** in the **OMOP Terminology Dictionary**
3. **Load the newly-instantiated CDM database** with the appropriate raw (i.e., untransformed) and transformed data

If a participating organization transforms its data in this way, then queries designed to run against its CDM database will successfully run, unedited, against another organization's CDM database. This, in turn, will allow organizations to immediately compare and contrast the results of their CDM queries. In that CDM database queries are the foundation of OMOP's methodological research, this comparability of query results is a key enabler of the research collaboration.

2.6. Process for Developing the Common Data Model

We expect that developing a best-practice Common Data Model and Terminology Dictionary for observational analysis will be an iterative process with the model potentially evolving over time. The process is:

1. Draft a document outlining 'points to consider' for CDM and Terminology Development, and propose an initial conceptual design (this document)
2. Elicit public review and expert comment before finalizing the initial design
3. Apply the Common Data Model and Terminology Dictionary to a sample of observational databases within OMOP's centralized IT research environment
4. Report lessons learned, implementation code, and suggestions for model refinement (as necessary)
5. Support Research Core distributed partners in applying the Common Data Model to their observational database
6. Report lessons learned, implementation code, and suggestions for model refinement (as necessary)

7. Encourage Extended Consortium participants to apply the Common Data Model and Terminology Dictionary to their own data sets, and share lessons learned
8. Execute analyses for methodological research
9. Publish final report about Common Data Model and Terminology Dictionary design, implementation, technical requirements and utility in facilitating analyses

3. Defining the Common Data Model

The OMOP pilot infrastructure contains several separately query-able databases, one for each raw source data set, and one for each transformed (as defined above) data set. The raw databases are each of unique design, because each reflects the underlying structure of its associated raw source data set. However, the transformed databases are all of the same design; each transformed database instantiates the Common Data Model. We use the term “CDM” to mean the Common Data Model design, and the terms “CDM database” and “transformed database” to mean a database that instantiates the CDM design.

Each organization wishing to participate in OMOP-sponsored research will create at least one instance of its data transformed according to the CDM design. The organization may choose to have a single CDM database that collectively contains all the transformed versions of all of its data, regardless of data source. Alternatively, the organization may choose to have multiple CDM databases, one for each transformed data source. In the latter case, the databases will be identical in all respects (e.g., number and names of tables, names and order of columns, keys, indexes, constraints, etc.) except for the content within them.

Each CDM database will contain data that have been transformed from the original raw dataset as well as augmented with concept codes from the OMOP standard concept hierarchy. The OMOP standard concept hierarchy is the “backbone” of the OMOP Terminology Dictionary which is described in detail later in this document. For the purposes of this section, the Terminology Dictionary may be considered a set of reference metadata (i.e., data about the data) about the CDM, because it defines the meanings of the data contained in a CDM database.

Transforming the raw data to augment them with standard concept codes is what allows us to query a CDM database based on the meanings of the data rather than on the data themselves. This presents at least two advantages over querying a CDM database that contains non-transformed (i.e., only raw) data.

1. Researchers will, for example, be able to query a CDM database for all male subjects, without needing to know the potentially myriad and disparate ways in which the raw data sources may represent the concept (i.e., meaning) of “male” (e.g., “M,” “1,” “Male,” etc.) This will allow researchers to focus more on their research, and worry less about whether they are using the right data.
2. Researchers will also be able to query a CDM database for higher-level classes of concepts, without needing to know the individual concepts that those classes subsume. Continuing the previous example, researchers will be able to query a CDM database for all subjects regardless of gender, without needing to know that the possible concepts within the gender class are “male,” “female,” “other,” “unknown,” and “not specified.” More to the point, researchers will be able to query a CDM database for all drugs within a specific therapeutic class without needing to know which drugs those are, and will be able to query a CDM database for a particular medical condition without necessarily knowing which individual diagnoses comprise that medical condition.

3.1. CDM Requirements

Based on the foregoing, the CDM design must fulfill four broad requirements:

1. The CDM design must accommodate all of the logical entities, attributes, and relationships relevant to OMOP-sponsored research.

From a practical point-of-view, this means that the CDM design must accommodate, from each raw data set from each of the OMOP's data partners, that subset of data that the OMOP researchers wish to analyze. Ideally, though, the CDM design should accommodate any data that OMOP researchers might ever wish to analyze, whether or not those data actually appear in the raw data sets of OMOP's data partners. We expect that this will include, but will not necessarily be limited to, data on the following.

- Persons (i.e., patients, with associated demographics)
- Medications
- Conditions (i.e., diagnoses)
- Procedures
- Laboratory tests
- Image (i.e., radiography) interpretations
- Assessments
- Inpatient hospital stays and outpatient visits

All other things being equal, we believe that it is preferable to have a CDM design that is capable of accommodating any data, rather than a CDM design that is limited in some artificial or arbitrary way.

2. The CDM design must allow the augmentation of the data with the Terminology Dictionary's standard concept codes that represent the meanings of the data.

Theoretically, any entity, attribute, or relationship modeled in the CDM design may have a corresponding standard concept code in the Terminology Dictionary. We believe that it is preferable, if not required, to have a CDM design that permits the augmentation of any individual datum with its corresponding standard concept code.

3. The CDM design, and the databases that instantiate it, must be usable.

That is, the CDM design must ultimately be intuitive, not overly complex, and otherwise "researcher-friendly." Researchers who find it difficult to understand the CDM design will find it difficult to formulate an accurate and efficient query against a CDM database.

4. The CDM design must enable queries to perform at "acceptable" rates on OMOP's central hardware or its distributed partner's hardware.

3.2. CDM Design Approach

In arriving at a CDM design, we must choose from among three general approaches to data modeling:

- Entity-Attribute-Value (EAV) modeling,
- Entity-Relational (ER) modeling, and
- Multi-Dimensional Modeling (MDM).

CDM Requirement 1 above (accommodate all OMOP-relevant data) makes an EAV data model the most attractive. An EAV data model places no artificial or arbitrary limits on the numbers or kinds of entities, attributes, or relationships that may be accommodated by the data model. EAV data models also make extensive use of metadata, rather than rely on static table structures and “hard-wired” foreign keys to describe the characteristics of and relationships among the data modeled. This makes an EAV data model supremely flexible, while isolating the EAV design from changes in source data structures, forms, and formats. This, in turn, makes EAV data modeling an attractive choice in research environments, where rapid and profound change is expected. In addition, EAV more closely captures the organizations of clinical events as they are recorded in transaction-oriented clinical systems, which may be important to some researchers.

CDM Requirement 3 above (usability, simplicity, and intuitiveness) makes an ER data model or an MDM especially attractive. ER data models comprise tables with familiar names that connote real-world entities of interest (e.g., patient, diagnosis, procedure, medication, etc.), and table columns with familiar names that connote real-world attributes of interest (e.g., medication name, NDC, diagnosis name, ICD-9-CM diagnosis code, etc.) Indeed, the central theme of ER data modeling is intuitiveness, and the goal is to obviate the need for the data user to know anything about how the data are physically stored in the database. MDM databases are essentially ER databases that have been de-normalized (i.e., that tolerate some data redundancy) to improve the performance of certain kinds of queries.

This tension—EAV on the technical side and ER/MDM on the user side—is, in most business settings, usually short-lived. Virtually always, the user’s need for a database design that matches his logical “view of the world” wins out over the flexibility and data-storage efficiency of the EAV design. Most business databases are of ER and MDM designs, because most of the time user acceptance is a primary determinant of success.

Arguably, in the OMOP’s case, there is no clear winner between EAV and ER/MDM. User acceptance, based on usability and intuitiveness, is important, but so is the flexibility of the CDM to accommodate new and different source data sets in the future. The OMOP is considering adding additional source data sets to the pilot infrastructure, and is hoping to promulgate OMOP research methods to additional partners with different source data sets. With this in mind, designing the CDM as an ER data model or an MDM would likely place an unacceptable maintenance burden on the data modeling, database administration, and Extract, Transform, and Load (ETL) development resources of the OMOP and its partners. However, designing the CDM as anything other than an ER model or an MDM will likely place additional user documentation and training requirements on the organization.

3.3. Proposed CDM Design

Considering the aforementioned tension—EAV on the technical side and ER/MDM on the user side—we propose a compromise. Building on the strengths of each of these approaches, our proposed CDM design combines an ER approach for selected, well-understood, and less likely to evolve portions of the model, with an EAV component to support unanticipated and more dynamic data.

Researchers who are comfortable with the EAV portion of the CDM will find modeled in it all of the data entities, attributes, and relationships necessary for their research (provided that those data are actually inherent in the source data sets used to load the CDM database). Researchers who are more comfortable with the ER “view of the world” will still be required to learn the EAV portion of the CDM, but may be able to concentrate the majority of their analyses on the ER portion of the CDM.

The EAV Portion of the Proposed CDM

To understand the EAV portion of the proposed CDM, first consider the following excerpt from a hypothetical “Prescriptions” source data set that we desire to load into a CDM database.

Patient_ID	Prescribed Medication	Prescribed Form	Prescribed Dosage
12345	Aspirin	Tablet	81 mg
12345	Acetaminophen	Tablet	500 mg
56789	Tylenol	Caplet	500 mg

One may think of the EAV portion of the proposed CDM as a single database table that holds all values, of all attributes, of all entities represented in the source data. For the example source data shown above, the EAV table in the CDM database might resemble the following.

Entity	Attribute	Value
Patient ID 12345	Has Prescription	Prescription ID 23456 ³
Prescription ID 23456	Medication	Aspirin
Prescription ID 23456	Form	Tablet
Prescription ID 23456	Dosage	81 mg
Patient ID 12345	Has Prescription	Prescription ID 34567 ³
Prescription ID 34567	Medication	Acetaminophen
Prescription ID 34567	Form	Tablet
Prescription ID 34567	Dosage	500 mg
Patient ID 56789	Has Prescription	Prescription ID 67890 ³
Prescription ID 67890	Medication	Tylenol
Prescription ID 67890	Form	Caplet
Prescription ID 67890	Dosage	500 mg

³ In this example, Prescription IDs are not provided in the source data. In such a case, Prescription ID numbers would be system-generated (i.e., by ETL logic) in a manner that makes them unique for each Patient ID.

Linking the Terminology Dictionary to the EAV Portion of the CDM

Theoretically, any entity, attribute, or relationship captured in the EAV table data exemplified above may have a corresponding standard concept code in the Terminology Dictionary. Consider the following excerpt from a hypothetical Terminology Dictionary.

Value	Relationship	Standard Concept Code
Patient	Has standard concept code	C001
Prescription	Has standard concept code	C002
Aspirin	Has standard concept code	C003
Acetaminophen	Has standard concept code	C004
Tylenol	Has standard concept code	C005
Tylenol	Is a brand name of	C004 (i.e., Acetaminophen)
Tablet	Has standard concept code	C006
Caplet	Has standard concept code	C007
Caplet	Is a synonym of	C006 (i.e., Tablet)
...

Augmenting the CDM database with standard concept codes from the Terminology Dictionary affects the EAV table in two ways.

1. For each entity (in this case, patient and prescription) in the EAV table, we add to the EAV table a new row that links the entity to its corresponding standard concept code.
2. For each value in the EAV table that has a corresponding standard concept code in the Terminology Dictionary, we replace the raw source value with the standard concept code.

Entity	Attribute	Value
12345	Is a	C001
23456	Is a	C002
34567	Is a	C002
56789	Is a	C001
67890	Is a	C002
12345	Has Prescription	23456
23456	Medication	C003
23456	Form	C006
23456	Dosage	81 mg
12345	Has Prescription	34567
34567	Medication	C004
34567	Form	C006
34567	Dosage	500 mg
56789	Has Prescription	67890
67890	Medication	C005
67890	Form	C007
67890	Dosage	500 mg

Refining the EAV Portion of the CDM

In the process of loading the CDM database from the source data set(s), we should separate units of measure (UOMs) from the values that they modify. This will create an additional EAV table row for each such value already in the EAV table.

Entity	Attribute	Value
12345	Is a	C001
23456	Is a	C002
34567	Is a	C002
56789	Is a	C001
67890	Is a	C002
12345	Has Prescription	23456
23456	Medication	C003
23456	Form	C006
23456	Dosage_Value	81
23456	Dosage_UOM	mg
12345	Has Prescription	34567
34567	Medication	C004
34567	Form	C006
34567	Dosage_Value	500
34567	Dosage_UOM	mg
56789	Has Prescription	67890
67890	Medication	C005
67890	Form	C007
67890	Dosage_Value	500
67890	Dosage_UOM	mg

Ideally, we should make the Terminology Dictionary as complete as possible. In this case, completeness would be achieved when there is a standard concept code in the Terminology Dictionary for every attribute and easily standardized value loaded into the EAV table.

Value	Relationship	Standard Concept Code
Patient	Has standard concept code	C001
Prescription	Has standard concept code	C002
Aspirin	Has standard concept code	C003
Acetaminophen	Has standard concept code	C004
Tylenol	Has standard concept code	C005
Tylenol	Is a brand name of	C004 (i.e., Acetaminophen)
Tablet	Has standard concept code	C006
Caplet	Has standard concept code	C007
Caplet	Is a synonym of	C006 (i.e., Tablet)
Unit of Measure	Has standard concept code	C008
Milligram	Has standard concept code	C009
Milligram	Is a sub-class of	C008 (i.e., Unit of Measure)

Medication	Has standard concept code	C010
Aspirin	Is a sub-class of	C010 (i.e., Medication)
Acetaminophen	Is a sub-class of	C010 (i.e., Medication)
Medication Form	Has standard concept code	C011
Medication Form	Describes	C010 (i.e., Medication)
Medication Dosage	Has standard concept code	C012
Medication Dosage	Describes	C010 (i.e., Medication)
Medication Dosage Units	Has standard concept code	C013
Medication Dosage Units	Describes	C010 (i.e., Medication)
Semantic Relation	Has standard concept code	C014
Is a	Has standard concept code	C015
Is a	Is a sub-class of	C014 (i.e., Semantic Relation)
Has Prescription	Has standard concept code	C016
Has Prescription	Is a sub-class of	C014 (i.e., Semantic Relation)
...

With a complete Terminology Dictionary such as that depicted in the above example, we can fully augment the EAV portion of the proposed CDM database with standard concept codes. This fully augmented EAV table is exemplified on the following page.

The value stored may be one of several types, including numeric or string. For situations in which the Terminology Dictionary is incomplete, the original attribute name from the source data set column header will persist in the CDM database.

Entity	Attribute	Value
12345	C015	C001
23456	C015	C002
34567	C015	C002
56789	C015	C001
67890	C015	C002
12345	C016	23456
23456	C010	C003
23456	C011	C006
23456	C012	81
23456	C013	C009
12345	C016	34567
34567	C010	C004
34567	C011	C006
34567	C012	500
34567	C013	C009
56789	C016	67890
67890	C010	C005
67890	C011	C007
67890	C012	500
67890	C013	C009

Some EAV data modeling practitioners prefer to break the single EAV table into many tables, one for each non-standardized data type that may be contained in the Value column. This approach results in separate EAV tables for short text strings (e.g., 255 characters or less), long text strings (e.g., character large objects, or CLOBs), Boolean values, integers, floating-point numbers, dates/times, binary large objects (BLOBs), etc. The advantage of this approach is that it allows the database management system to handle some tasks at which it excels, such as data type validation. It also preserves the ability to perform calculations on numeric values. However, the disadvantage of this approach is that it complicates CDM database queries (which one might argue are difficult enough against a single EAV table).

The ER Portion of the Proposed CDM

As noted above, we propose a CDM design that is mostly EAV, but that uses supplemental database tables and views to present, to those users who require it, some of the most frequently accessed data in a manner that reflects the more logical ER view of those entities, attributes, and relationships. We expect to have “ER-style” tables or views for at least three domains of data: persons (i.e., patients), medications, and health outcomes of interest (HOIs). Although these tables have yet to be designed, we anticipate that they might contain the following kinds of data (subject to final data privacy policy decisions), augmented with standard concept codes.

- Persons (Patients)
 - Patient ID
 - Date of Birth
 - Gender
 - Race
 - Ethnicity
 - Date of death

- Medications
 - Patient ID
 - Drug ID
 - Dose
 - Compliance/MPR or other
 - Method including version
 - Exposure Start Date
 - Exposure End Date (must be able to represent explicitly an unknown date)

- Outcome
 - Patient ID
 - Outcome ID
 - Certainty
 - Method including version
 - Outcome Start Date
 - Outcome End Date (must be able to represent explicitly an unknown date)

In addition to these three tables or views, we could envision the need to produce other tables or views for researchers who need them. It is recognized that these views can be created from data elements contained within the EAV model, but the convenience of standard transformation may outweigh the cost of redundancy across the tables.

Questions for the reader:

- What other views should be made available in the ER model to supplement the EAV data representation?

3.4. The Role of the Terminology Dictionary in Querying the CDM

The Terminology Dictionary does not just enable the mechanism for transforming raw data into standardized data. It also plays a role in searching and querying the transformed data in a CDM database, browsing and navigating the hierarchies of classes and abstractions inherent in the transformed data, and interpreting the results returned by those operations. Indeed, the ability to efficiently and effectively generate actionable information from a CDM database depends, to a large degree, on the Terminology Dictionary and its relationship to the CDM.

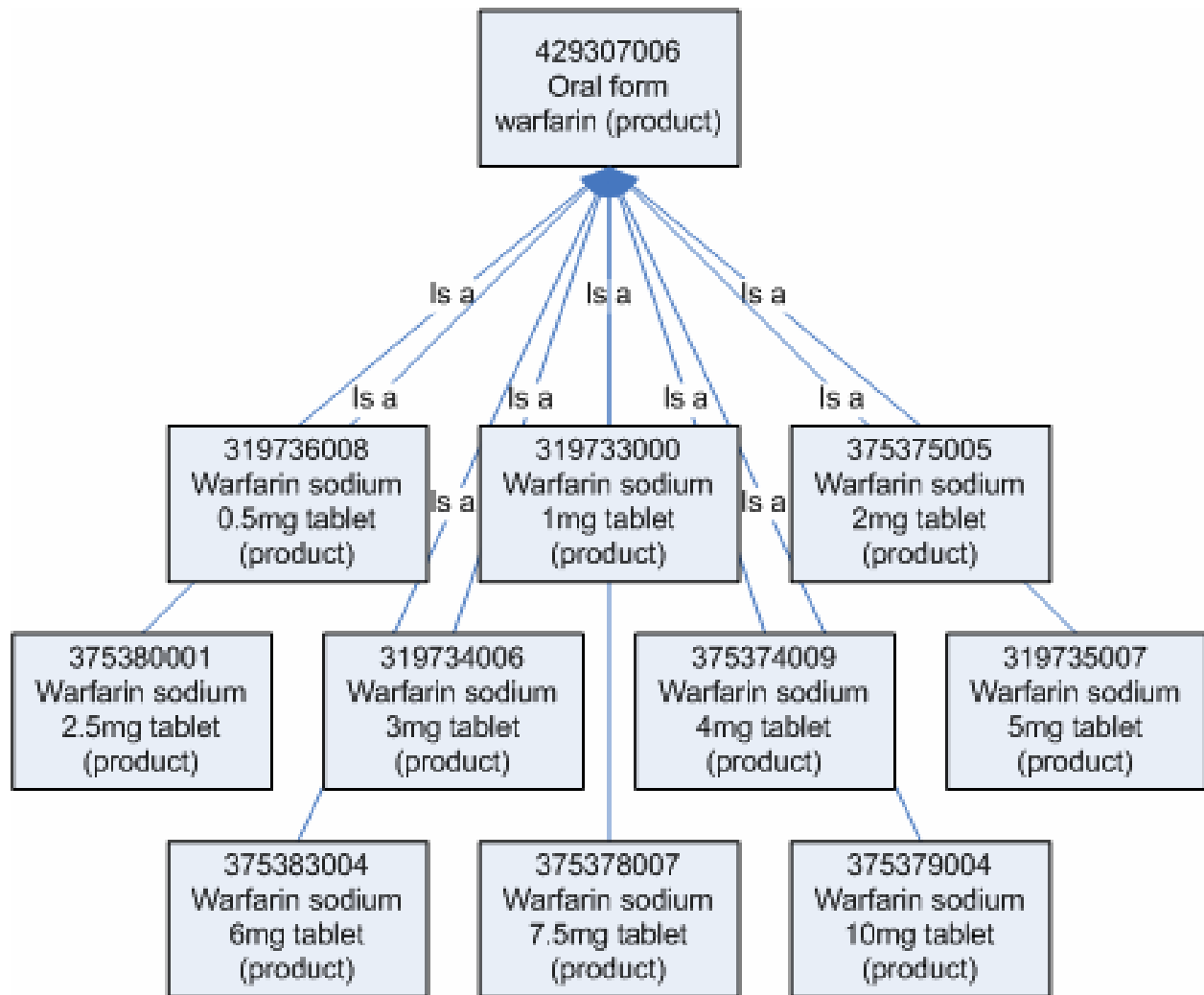
The Terminology Dictionary contains all of the code sets, terminologies, vocabularies, nomenclatures, lexicons, thesauri, ontologies, taxonomies, classifications, abstractions, and other such data that are required for:

1. Creating the transformed (*i.e.*, standardized) data from the raw data sets,
2. Searching and querying the transformed data, and browsing and navigating the hierarchies of classes and abstractions inherent in the transformed data, and
3. Interpreting the meanings of the data.

Researchers will be able to query a CDM database for classes of concepts (*i.e.*, meanings) without needing to know the individual concepts that those classes subsume. For example, a researcher will be able to query the common data repository for all patients regardless of gender, without needing to know that the possible concepts within the gender class are “male,” “female,” “other,” “unknown,” and “not specified.” More to the point, researchers will be able to query a CDM database for all drugs within a specific therapeutic class without needing to know which drugs those are, and will be able to query a CDM database for a particular medical condition without necessarily needing to know which individual diagnoses comprise that medical condition. The researcher will search or browse the Terminology Dictionary to find the class of concepts on which to query or analyze, and will transfer the class’ appropriate concept code from the Terminology Dictionary to the query and analysis tool.

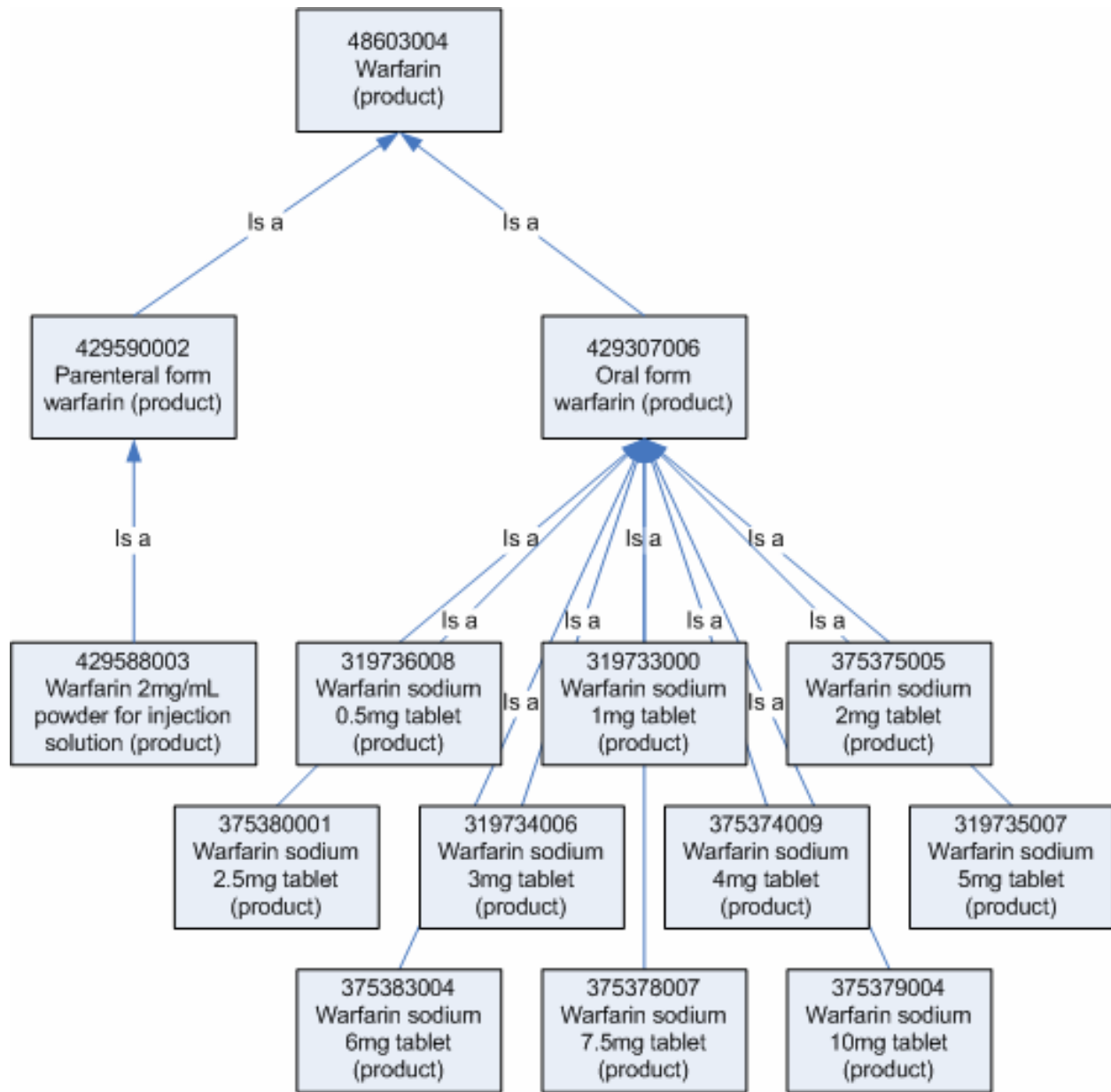
Consider the researcher who wants to create a portable (as defined above) query for all patients who took warfarin and who later had a diagnosis of gastrointestinal bleeding. For the query to be

portable, the query must use OMOP standard concept codes for warfarin and gastrointestinal bleeding. As an example of a potential starting point, the researcher searches the Terminology Dictionary for oral forms of warfarin and could find the following.



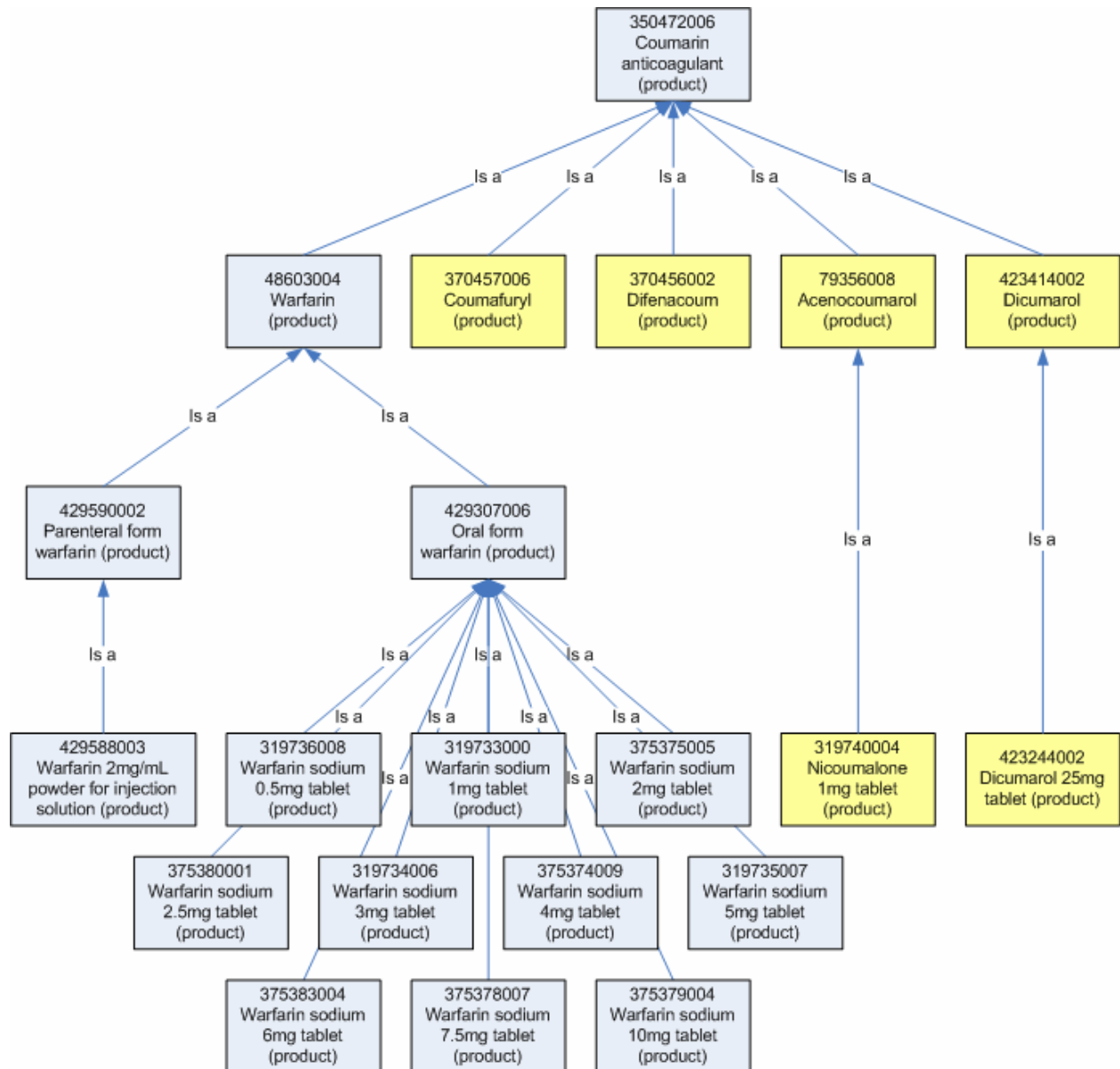
The class subsuming all oral forms of warfarin

The researcher understands that querying for patients only on these oral forms of warfarin would leave out patients on other forms of warfarin. Accordingly, the researcher navigates up the standard concept hierarchy, to all forms of warfarin, as depicted below.



Navigate up, to the class subsuming all forms of warfarin

Wanting to make sure that no study patients that should be included are left out, the researcher navigates up the standard concept hierarchy one more time, to the concept that represents all coumarin anticoagulants, as depicted below.



Navigate up again, to the class subsuming all Coumarin anticoagulants

The researcher decides that the class he wants to include in his query is the previous class (all forms of warfarin). Accordingly, he notes that the standard concept code that he must include in the medication portion of his query is “48603004.” He follows a similar process to determine the standard concept code for the condition (or higher-level class of conditions) that accurately represents the type of gastrointestinal bleeding that he wishes to correlate (or fail to correlate) to warfarin use.

3.5. Mapping Raw Source Data to Standard Concept Codes

Section 3.2 of this document illustrates how the data in a CDM database will link to the standard concept codes in the Terminology Dictionary. Implicit in that example is the existence of a Terminology Dictionary that is pre-populated with all of the standard concepts to which the selected raw data values correspond. We propose to achieve this by building the Terminology Dictionary around an existing, comprehensive healthcare concept hierarchy, such as, but not necessarily limited to, 3M's Healthcare Data Dictionary (HDD) or the Unified Medical Language System (UMLS) Metathesaurus. For each class of data (such as conditions, medications, procedures and labs), a necessary step in the data normalization process will be to apply the standard concept codes to the raw data. For example, if SNOMED were chosen to represent conditions but raw data had diagnoses coded using ICD9, then semantic maps between ICD9 and SNOMED would need to be established so that SNOMED concepts could be used in analysis in lieu of ICD9 codes.

That being said, we recognize that even the most complete of these alternatives may not be pre-populated with all of the mappings from the OMOP's raw data sources to its standard concept hierarchy. We may expect these products to be pre-populated with concept mappings for standard terminologies, such as ICD-9-CM, RxNorm, LOINC, CPT4, HCPCS, etc. However, it may be less likely for mappings between proprietary code sets to be available, in which case, procedures to establish mappings will need to be created when necessary.

3.6. Deciding Upon a Standard Concept Hierarchy for the Terminology Dictionary

The standard concept hierarchy used in the preceding examples originated in SNOMED CT, but it could also have come from the UMLS Metathesaurus, the 3M HDD, or some other comprehensive, standard concept hierarchy that covers medications and conditions. The point here is that, in choosing a standard concept hierarchy around which to build the OMOP Terminology Dictionary, we have options.

One option we have is to use a single, already-existing standard concept hierarchy that provides sufficient and appropriate coverage of all of the OMOP data domains (i.e., demographics, medications, conditions, procedures, laboratory tests) Examples include, but are not necessarily limited to, the 3M HDD, the UMLS Metathesaurus, and SNOMED CT. Another option we have is to "piece together" a single standard concept hierarchy from portions of existing concept hierarchies or commercial healthcare terminologies. For example, we might build "from scratch" the portion of our standard concept hierarchy that covers demographics, but we might use SNOMED CT for medications, ICD-9-CM for conditions, CPT4 or HCPCS for procedures, and LOINC for laboratory tests. A final option we have is to build "from scratch" the entire concept hierarchy ourselves. This seems least attractive, but it is an option nonetheless.

As an aid to deciding which option to choose, and depending on that option, which existing standard concept hierarchies or commercial healthcare terminologies to use, we propose the following process:

1. Bottom up inventory: Make a list of the commercial healthcare terminologies that are used in the source data for each domain (i.e., demographics, medications, conditions, etc.) to answer the question, “which codes from which terminologies show up in the raw data?”
2. Top down inventory: Make a list of the standard concept hierarchies (e.g., 3M HDD, UMLS Metathesaurus, etc.) that may be able to provide the necessary concepts for each domain. This requires understanding what standard and class-based queries the OMOP researchers will want to perform, and what semantic and ontological information is required of the Terminology Dictionary to make those things possible.
3. Conduct an evaluation of the enumerated standard concept hierarchies to determine which is best for each domain (note: there may not be a single choice that is best for all domains). The cross-product of the number of data domains and the number of candidate concept hierarchies will make this a potentially time-consuming process.
4. Select the best concept hierarchy for each domain, seeking public comment and approval from the OMOP Principal Investigators (PIs) and Advisory Boards.
5. Create the Terminology Dictionary to accommodate the selections made in step 3, and ratified in step 4.
6. Load the CDM database(s) with transformed data, as described earlier in this document.

Given the time and resource constraints on the OMOP project, we anticipate that it will be preferable to use an existing concept hierarchy (e.g., that of the 3M HDD, UMLS Metathesaurus, etc.) than to build one. Therefore, regarding the evaluation process of step 3 above, we recommend that the evaluation criteria include, but not necessarily be limited to, cost of use (e.g., licensing cost, software cost, maintenance cost, renewal cost, etc.), ease of use to apply the concepts in the Common Data Model, relevance and specificity of the concept hierarchy in supporting observational analyses, and the quality of mapping between source data terminologies and reference concepts. We also recommend that the evaluation process include the following.

- An **objective** assessment of the **quantity** of coverage of the data domain(s) by the candidate standard concept hierarchies. This will answer the question, “How much of the data in the domain(s) of interest can be correlated to a single standard concept?”
- A **subjective** assessment of the **quality** of the coverage of the data domain(s) by the candidate standard concept hierarchies. This will answer the question, “does the concept hierarchy allow researchers to accurately isolate the classes of medications and classes of conditions necessary to define the drug-HOI pairs of interest?”

Possible criteria for selection of ontologies include

- Subjective feasibility of mapping from codes found in observational data to the chosen ontology

- Availability and cost of the ontology for use by distributed partners

The following questions remain open.

- What criteria should be applied to inform the selection of the ontologies?
- Are there any suggestions for objective measures to be applied to the ontology, mapping, or source data to evaluate the criteria?

4. Overview of Observational Healthcare Data Elements

This section outlines the relevant data elements considered for inclusion with the Common Data Model. The Common Data Model will be designed to support de-identified data with anonymous linkage, and essential data from the OMOP centralized data sources will be constrained based on the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor principle (i.e., removing all direct and certain indirect individual identifiers from datasets). De-identification provides a balance by removing identifiers but still retaining individual characteristics about the dataset, such as demographics, medications, or conditions. The OMOP de-identified data set will be coded with a unique identifier that cannot be traced back to the individual, and so cannot be used to re-identify the data at a later date. The OMOP will use datasets for the stated purpose only, and will not attempt to re-identify any individual's data.

4.1. Guidelines for Data Element Definitions

The selection of the OMOP uniform data elements is based on the mission and organizational experience of collecting administrative and observational data. One objective is to provide uniform data elements that harmonize with prevailing standards for electronic data exchange. Therefore, a structured data element format is being used to document each data element providing the variable name, definition, code description, and reference to any applicable data standards used to define the element and the code description.

In addition, the OMOP data element definitions are also resourced from government initiatives that support the use and implementation of electronic data exchange such as the Health Information Technology Standards Panel (HITSP) and the U.S. Health Information Knowledgebase (USHIK) so that public and private organizations can harmonize information formats from the CDM with existing and emerging healthcare standards. The Agency for Healthcare Research and Quality (AHRQ) provides and maintains the United States Health Information Knowledgebase (USHIK) definitions, values and information models that enables browsing, comparison, synchronization and harmonization of health data elements within a uniform query and interface environment.

Essential OMOP data elements are:

Person and Demographics

- PersonID: Unique identifier within the database itself used to link tables
 Health Information Technology Standards Panel (HITSP) modified element description:
 (PersonID) An identifier that uniquely identifies the individual to which the data refers and connects that data to the individual record.
- Audit ID: Identifier to trace back to the source record
 Health Level Seven (ID): A unique identifier for an entity. Ideally each entity will have only one identifier assigned to it, however, since different systems will maintain different data bases, there may be different instance identifiers assigned by different systems. Note that an instance identifier is a pure identifier and not a classifier. The identifier is not used to store information about the kind or type of entity.
- Year of Birth: Field to calculate Patient Age. Data transformation may be necessary from data variables of age and/or date of birth.
 ANSI ACS X12: Four position designation of the year expressed as CCYY
- Patient Gender: Code indicating the administrative sex of the individual
 National Committee on Vital and Health Statistics (NCVHS) Core Health Data Elements (Gender): The gender of the patient may be classified as recommended by the Centers for Medicare and Medicaid Services (CMS) Uniform Hospital Discharge Data Set (UHDDS) and Uniform Ambulatory Care Data Set (UACDS) 1 = Male; 2 = Female; 3 = Unknown/not stated
- Race and Ethnicity: Refers to a person's race (Consolidated Health Informatics Initiative) Federal standards do not conceptually define ethnicity and recognize ethnicity is a social-political construct in which an individual's own identification with a particular ethnicity is preferred to observer identification. Federal standards for classifying data on race and ethnicity exert a strong influence on categorization by state and local agencies and private sector organizations. The collections of race and ethnicity have required definitions for Federal data collection in Office of Management and Budget (OMB) Directive 15. The Centers for Medicare and Medicaid Services (CMS) definition of race was extended from the OMB requirement: A. Race 1. American Indian/Eskimo/Aleut, 2.Asian or Pacific Islander (specify) 3. Black, 4. White, 5. Other (specify), 6. Unknown/not stated. B. Ethnicity 1. Hispanic Origin (specify), 2. Other (specify), 3. Unknown/not stated.

A1	Race - American Indian/Eskimo/Aleut
A2	Race – Asian
A3	Race – Black
A4	Race – White
A5	Race - Other (Specify)
A6	Race - Unknown/not stated
B1	Ethnicity - Hispanic Origin (specify)
B2	Ethnicity - Other (specify)
B3	Ethnicity - Unknown/not stated

Medication use

- **PersonID:** Unique identifier within the database itself used to link tables
Health Information Technology Standards Panel (HITSP) modified element description: (PersonID) An identifier that uniquely identifies the individual to which the data refers and connects that data to the individual record.
- **Drug Identifier:** Unique identifier of pharmaceutical and medicinal products. Data transformations may be necessary to identify drug characteristics that together unambiguously identify a pharmaceutical product, substance, excipient, ingredient or medicinal product (Over-the Counter). The variables may include: product name, generic name (ingredient), drug class, as well as dose form, strength and route of administration to sufficiently differentiate the drug dispensed
National Council on Prescription Drug Programs (NCPDP): The code to indicate the type of drug dispensed. Note: a person may have several records for the same drug since use may be interrupted.
- **Start Date:** Field contains the date and time of service when medication was initiated.
- **End Date:** Field contains when medication should have stopped.

End Date – Start Date can be used to define ‘Duration of Therapy’. Duration of treatment can be recorded in multiple different ways within observation databases. Some sources capture drug exposure as a span of time (e.g. start date and end date), prescription dispensing claims may contain a fill date and ‘days supply’, and records for prescriptions written may only have the prescription date and quantity or number of refills.

There are three options of varying levels of required transformation:

1. Establish a common data structure that has multiple available fields for all the different types of ways duration of therapy could be inferred (e.g. end date, days supply, refills). Each data source will only use a subset of the available fields. No data transformation would be applied.
2. Create a common concept for ‘duration of therapy’ and define one field to represent that concept (ex. end date). Each source database would require documented, standard data transformation process for how to infer end date from the available data elements (e.g. end date = start date + days supply)
3. Aggregate prescription records to create periods of drug exposure. Ex. if a person has three concurrent 30-day prescriptions for the same drug; collapse the records into one 90-day drug era. A standard data aggregation process would need to be implemented to apply to common records in #2, outlining assumptions around persistence windows and comparable drugs that can be combined.

Questions for the reader:

- What data transformation processes have you used to define drug exposure in your studies?
- What lessons learned do you have from your research?
- What approach(es) should be recommended within the OMOP common data model to support the observational analyses?
- If multiple approaches should be considered, what criteria could be used to select a best practice?

Outcome

- **PersonID:** Unique identifier within the database itself used to link tables
Health Information Technology Standards Panel (HITSP) modified element description: (PersonID) An identifier that uniquely identifies the individual to which the data refers and connects that data to the individual record.
- **Outcome ID:** Code(s) used to identify condition(s) relating to the patient.
- **Start Date:** Field contains the initial date the outcome could be documented
- **End date:** Field contains the last date the outcome could be documented

Analogous to the data transformations offered to define ‘duration of therapy’ for medications, it is possible to consider aggregating diagnoses that are observed within a specified period of time as representing a period of condition persistence within an episode of care. It remains an outstanding question of what degree of transformation would be desired for the common data model.

4.2. Data Transformation Example

One example of potential data transformations is based on research at GlaxoSmithKline, as highlighted at the FDA Sentinel Network (7 Mar 2007) and the AMIA pharmacovigilance summit (12 Jun 2007), with publications forthcoming. ***This example is offered as one potential approach to consider, but it is expected that the development of a solution for OMOP will require further investigation.***

To facilitate a common structure to use across disparate data sources for representing drug utilization and condition incidence, GSK developed the concept of ‘eras’. ‘Eras’ are periods of time that a person has persistent health status for one characteristic. ‘Drug eras’ and ‘condition eras’ are two types of eras used to represent observational data that allow for the exploration of temporal associations between drugs and conditions.

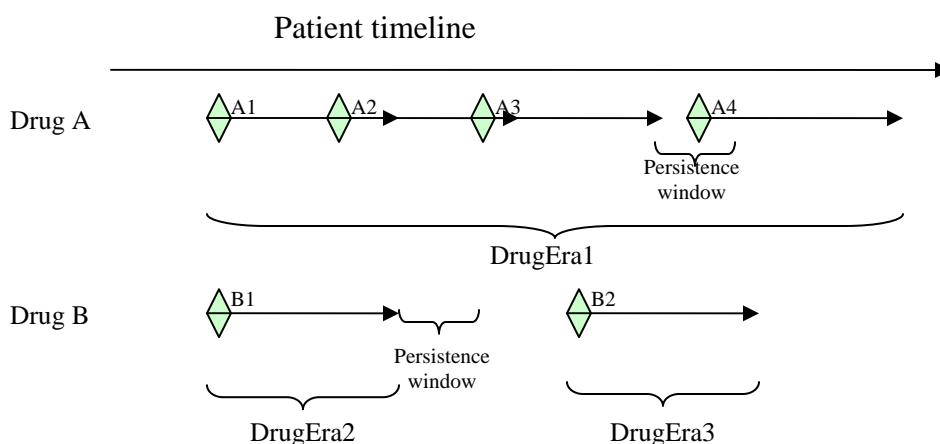
Drug Eras

‘Drug eras’ represent the span of time a given person is assumed to be persistently using a given drug concept. This approach can allow for considerations of length of exposure and can be modified to accommodate cumulative exposure.

Here, the ‘drug concept’ can be variable based on the data source. Drugs are extracted from each source based on unique ‘product name + strength’ identifiers, as constructed by the source drug references. For example, assume a person has taken ‘DrugX 25mg’ and ‘DrugX 50mg’: these prescriptions are treated as distinct drug concepts that would be represented as separate eras. Data can then be normalized using reference drug ontology to aggregate drug concepts to higher levels of generality. For example, assume a person has taken ‘DrugX 25mg’ and ‘DrugX 50mg’: since these drug concepts both annotate to a common product name, ‘DrugX’, any overlapping eras will be aggregated into a common drug era. Drug eras can be further aggregated to the generic name level or a higher-level drug classes as necessary.

The ‘span of time’ that is covered by the drug era must be derived based on the data elements available for each drug concept contained within the source data. Each period of time is represented by a start date and an end date, so the drug exposure for a given drug era can be calculated as $ERA_END - ERA_START$. However, each data source may represent the period of drug use in different ways. For example, pharmacy claims contain date of prescription and days supply, so end date could be calculated as date of prescription + days supply. However, many patients receive recurring prescriptions for the same product such that one may consider multiple prescriptions to represent one continuous period of drug use. We combine these independent prescriptions into a common drug era through the use of a ‘persistence window’, which is the allowable span of time after the prescription should have completed before another prescription of the same drug needs to be filled to maintain drug persistence. This persistence window allows for tolerance of non-compliance or logistics of getting a new prescription filled and assumes that the drug can still be considered to be in the patient’s system during this window. The ‘persistence window’ could be a parameter in the data normalization process and can be modified as necessary.

An example is illustrated graphically below. Imagine a patient who has taken two drugs during his or her insurance coverage, Drug A and Drug B. Drug A has had 4 prescriptions filled (A1, A2, A3, A4), each with 60 days supply. Drug B has had 2 prescriptions (B1, B2).



To define drug persistence for Drug A, we look at the timing of successive prescriptions in the context of when the previous prescription should have been completed. Here, A2 has been filled before the expected completion of A1. Similarly, A3 has been filled before the expected completion of A2. A4 was filled after A3 is completed, but within the ‘persistence window’ specified. Thus, the 4 prescriptions for Drug A can be collapsed into DrugEra1, with $EraStartDate = A1StartDate$ and $EraEndDate = A4EndDate$.

Drug B, however, had significant time between filling the two prescriptions. Because this time exceeded our ‘persistence window’, we define two distinct ‘drug eras’ for Drug B; DrugEra2 has $EraStartDate = B1StartDate$, $EraEndDate = B1EndDate$, DrugEra3 has $EraStartDate = B2StartDate$, $EraEndDate = B2EndDate$.

A typical ‘persistence window’ may be 7 or 30 days, but a decision would need to be made for this initiative.

The assumption about persistent use requires further clarification. ‘Drug eras’ are an approximation of the period of drug utilization: EHR records prescriptions written and medications self-reported, but does not contain data about prescriptions filled at pharmacies or actual patient consumption; claims databases record prescriptions filled as captured on pharmacy claims, but do not capture prescriptions written or filled out-of-system or actual patient consumption.

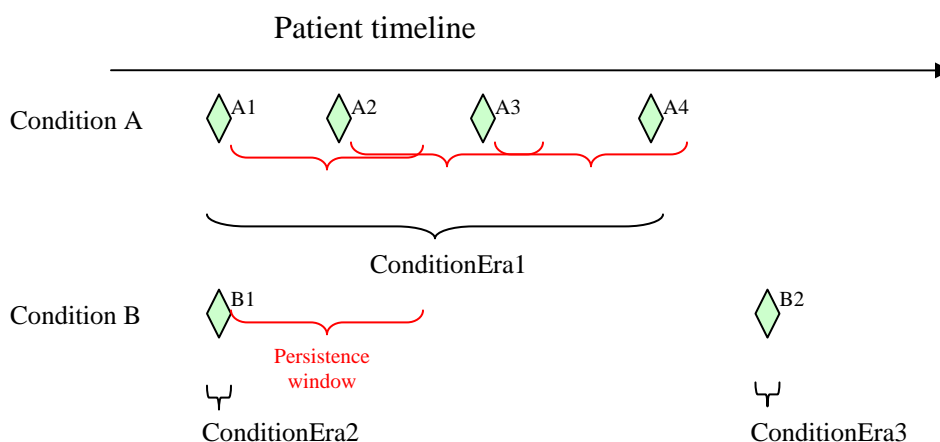
Condition Eras

‘Condition eras’ represent the span of time for which a patient can be considered to have a given condition concept.

Health conditions can be inferred from observational data source through the diagnoses that are recorded on claims or electronic medical records. Claims data may supply diagnoses that were submitted as part of claims for health service visits and procedures. These diagnoses, when taken in aggregate across a patient’s participation, provide context for a patient’s health

conditions, either new or preexisting. However, any given claim may not provide a complete summary of patient health. Similarly, EHRs have diagnoses and symptoms recorded within the electronic health record, but do not provide any context for out-of-system health experiences. Many condition concepts are originally represented as ICD-9 diagnosis codes. For initial extraction of data at the highest degree of granularity, ‘condition concepts’ could be defined as diagnoses with the same 5-digit ICD-9 code, with an initial ‘persistence window’ of 7 days. Similarly with drugs, outcomes references could be further normalized to a common reference condition ontology.

An example is illustrated graphically below. Imagine a patient who has been diagnosed with two conditions during his or her insurance coverage: Condition A and Condition B. Condition A has had 4 diagnoses (A1, A2, A3, A4). Condition B has had 2 diagnoses (B1, B2).



To define condition persistence for Condition A, we look at the timing of successive diagnoses in the context of when the previous diagnoses. Here, A2 is within the ‘persistence window’ of A1. Similarly, A3 is within the ‘persistence window’ of A2 and A4 is within the ‘persistence window’ of A3. Thus, the 4 diagnoses for Condition A can be collapsed into ConditionEra1, with EraStartDate = A1StartDate and EraEndDate=A4StartDate.

Condition B, however, had significant time between diagnoses. Therefore, we cannot assume dependence between the diagnoses. Because this time exceeded our ‘persistence window’, we define two distinct ‘condition eras’ for Condition B; ConditionEra2 has EraStartDate=B1StartDate and EraEndDate=B1StartDate, ConditionEra3 has EraStartDate=B2StartDate and EraEndDate=B2StartDate.

Consolidation of diagnoses into ‘condition eras’ has several purposes. First, we will be able to aggregate chronic conditions that require frequent ongoing care, instead of treating each diagnosis as independent. Second, multiple doctor visits for the same condition in near proximity will be aggregated so as not to double-count the same event occurrence. ‘Condition eras’ follow a conservative structure that best fits an acute condition model; multiple health care

visits within a short time frame for the same condition are combined into one episode of condition occurrence. For example, assume a patient goes to a PCP, who diagnoses a particular condition and recommends the patient to seek supplemental care for that condition at a specialist office. The patient visits the specialist one week later, where the specialist confirms the diagnosis and provides the appropriate treatment to reach condition resolution without further care required. These two independent health visits would be aggregated into one ‘condition era’. It is recognized this model generally fits well for acute conditions, but may be less robust for chronic conditions. Chronic conditions that do not require regular follow-up may be recorded as multiple ‘condition eras’ because the data source does not have any information in the intermittent periods of time to justify the aggregation of disparate eras. Because our ‘persistence window’ is small, we are likely to capture multiple visits in rapid succession for the same condition, but unlikely to capture infrequent visits for chronic conditions (e.g. RA patient visits rheumatologist every 3 months). However the small window also reduces the likelihood that we will falsely classify independent events into the same ‘condition era’. At present time, a consistent approach is applied to all conditions. Further research is needed to develop a systematic mechanism for identifying acute and chronic conditions such that different heuristics can be placed within the aggregation process based on the condition.

4.3. Guidelines for Extending the CDM with Additional Data Elements

It is expected that there will be research interest in data elements beyond the Common Data Model as outlined above. Therefore, the OMOP recommendation for extending the Common Data Model is through the addition of additional CDM elements within the EAV table to contain the additional data.

5. Conclusion

This document outlines points-to-consider in establishing a common data model to facilitate observational analyses to support active drug safety surveillance. The model posited is intended to be a starting point to facilitate an active conversation. Please share your experiences, opinions and recommendations such that OMOP can move forward with a common data model that will be agreeable to all stakeholders and may inform future research. Feedback will be collected on the OMOP website, and releases of the common data model design will be made available as soon as possible.