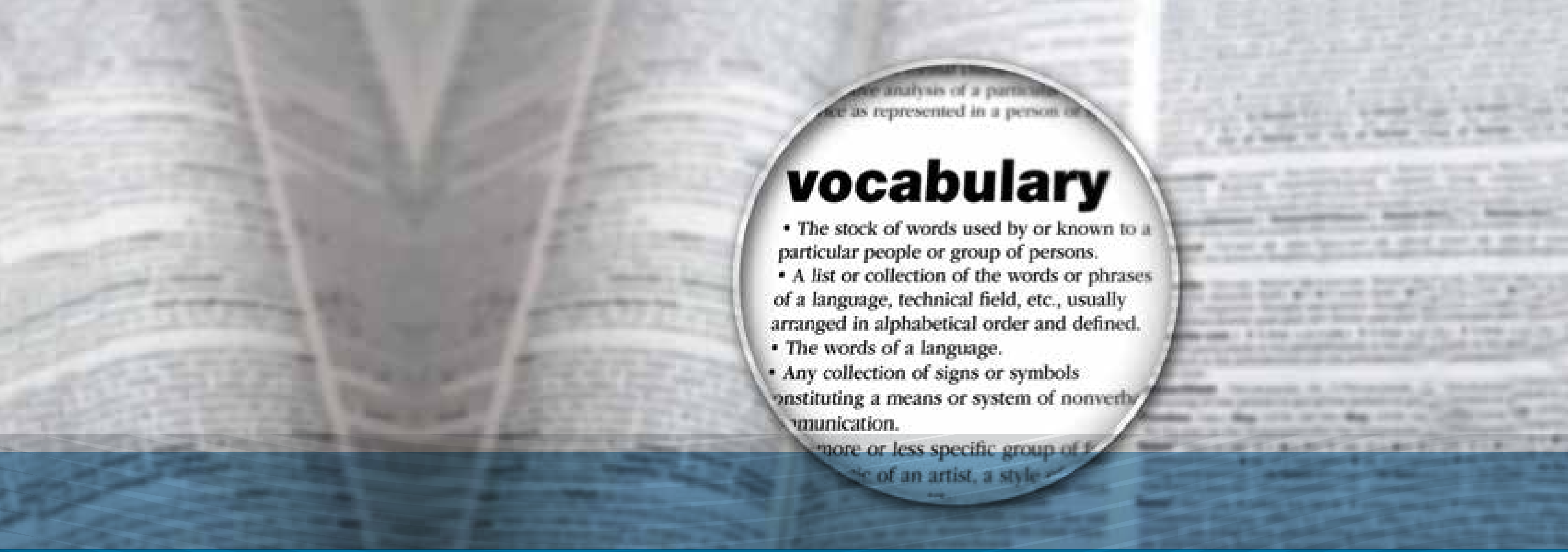
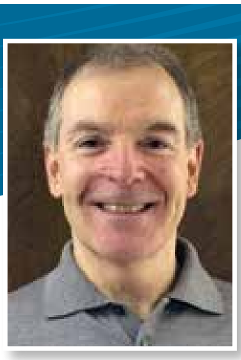

vocabulary
• The stock of words used by or known to a particular people or group of persons.
• A list or collection of the words or phrases of a language, technical field, etc., usually arranged in alphabetical order and defined.
• The words of a language.
• Any collection of signs or symbols onstituting a means or system of nonverb munication.

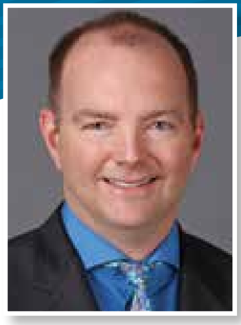
Clinical Vocabularies for Global RWE Analysis

Don O'Hara, MS
Senior Research Associate, Real-World Evidence, Evidera

Vernon F. Schabert, PhD
Senior Research Scientist, Real-World Evidence, Evidera



Don O'Hara



Vernon F. Schabert

Introduction

A significant volume of real-world evidence (RWE) analyses continue to be conducted with data repurposed from healthcare administrative databases. The range of sources represented by those databases has grown in response to demand for richer description of patient health status and outcomes. Data availability, including the range of available data sources, has grown unevenly across the globe in response to country-specific market and regulatory dynamics. Nonetheless, as demand globalizes for RWE insights from databases, those demands increase pressure on analysts to find ways to bridge differences between local data sources to achieve comparable insights across regions.

One of the challenges in bridging differences across databases is the codes used to represent key clinical facts. Historically, RWE database studies have leveraged local code sets for cost-bearing healthcare services such as drugs, procedures, and laboratory tests. While diagnosis codes have long been globalized (the International Classification of Diseases, or ICD, is maintained by the World Health Organization), adoption of specific diagnosis code revisions has occurred inconsistently by country and region.

Two dynamics are increasing pressure to use more globalized codes for the full range of clinical facts in RWE database analyses. One is the increased set of incentives for providers' administrative systems to exchange information for improved quality and coordination of care, often using standardized messaging systems such as Health Level 7 (HL7). These messages are only as good as the standardization of codes between message senders and receivers, which motivates the encoding of facts using common code sets. The second is the increased availability of common data models to standardize the extraction and analysis of these data for RWE and drug safety purposes. While common data models make compromises on the structure of tables and fields extracted from healthcare systems such as electronic medical records (EMR) and billing systems, they can improve consistency and replicability of analyses by mapping data values to globally standardized clinical codes.

Analysts faced with using more clinically rich or globally standardized data will need to master new coding systems. This paper provides a brief primer on several of these global clinical terminologies: LOINC, SNOMED CT, and RxNorm. We'll highlight the origins, structure, content, and overlap of each, and will also highlight novel ways to leverage these global code sets even when they have not been included within a particular database.

Table 1. LOINC Codes Related to "Hemoglobin A1c."

LOINC	LongName	Component	Property	Timing	System	Scale	Method	Units
4548-4	Hemoglobin A1c/ Hemoglobin.total in Blood	Hemoglobin A1c/ Hemoglobin.total	MFr	Pt	Bld	Qn		%
55454-3	Hemoglobin A1c in Blood	Hemoglobin A1c	—	Pt	Bld	—		
41995-2	Hemoglobin A1c [Mass/ volume] in Blood	Hemoglobin A1c	MCnc	Pt	Bld	Qn		g/dL
17855-8	Hemoglobin A1c/ Hemoglobin.total in Blood by calculation	Hemoglobin A1c/ Hemoglobin.total	MFr	Pt	Bld	Qn	Calculated	%
4549-2	Hemoglobin A1c/ Hemoglobin.total in Blood by Electrophoresis	Hemoglobin A1c/ Hemoglobin.total	MFr	Pt	Bld	Qn	Electrophoresis	%
17856-6	Hemoglobin A1c/ Hemoglobin.total in Blood by HPLC	Hemoglobin A1c/ Hemoglobin.total	MFr	Pt	Bld	Qn	HPLC	%
62388-4	Hemoglobin A1c/ Hemoglobin.total in Blood by JDS/JSCC protocol	Hemoglobin A1c/ Hemoglobin.total	MFr	Pt	Bld	Qn	JDS/JSCC	%
71875-9	Hemoglobin A1c/ Hemoglobin.total [Pure mass fraction] in Blood	Hemoglobin A1c/ Hemoglobin.total	MFr.DF	Pt	Bld	Qn		
59261-8	Hemoglobin A1c/ Hemoglobin.total in Blood by IFCC protocol	Hemoglobin A1c/ Hemoglobin.total	SFr	Pt	Bld	Qn	IFCC	mmol/ mol

MFr = Mass Fraction, **MCnc** = Mass Concentration, **MFR.DF** = Mass Decimal Fraction, **SFr** = Substance Fraction, **Pt** = Point in Time, **Bld** = Blood, **Qn** = Quantitative

LOINC

Logical Observation Identifiers Names and Codes (LOINC) is a coding system focused on structured "observations." Most of those observations are laboratory tests, although the LOINC system extends to systematic observations such as radiology reports, clinician rating scales, and tumor registries. It was developed at the Regenstrief Institute of the Indiana University School of Medicine, which developed one of the first U.S.-based electronic medical records in the 1970s. Development began in 1994, and the first list of codes was released in 1996.[1]

LOINC's original developer, Clem McDonald, had previously been a founding developer of the HL7 2.x messaging standard used in virtually all EMRs today. The HL7 2.x standard provided a structure to exchange clinical content, but the widespread use of proprietary codes limited the value of exchanging laboratory orders and results. LOINC set out to solve the problem of reconciling proprietary lists of lab codes from each HL7 message sender and recipient.

LOINC was formally adopted as a code set for HL7 messaging in 1999. LOINC has registered users in 177 countries around the world, with documentation available in 20 languages or linguistic variants. Within the U.S., LOINC has also been adopted as a coding standard for EMR meaningful use regulations and was proposed as a code set for electronic transactions in the HIPAA administrative simplification rules. LOINC has helped individual providers accelerate mapping of their local codes to its standard through the release of RELMA (Regenstrief LOINC Mapping Assistant), an application that facilitates side-by-side comparison of uploaded codes to the LOINC standard.

The numeric part of LOINC codes are structured as one to five digits, a hyphen, and a single check digit. For example, the most frequent code used to describe Hemoglobin A1c tests (as a percentage of total blood) is "4548-4" *(Table 1)*. There is no order or structure to the numeric value before the hyphen, and the allowed digit length may expand once LOINC contains more than 100,000 records. The check digit is a feature allowing message receivers to confirm that the first part of the code is completely and accurately specified.

For each LOINC code, up to six text fields (parts) may be included in the description. These parts include the component (analyte), measurement property, measurement time (duration), body system providing the measurement sample, measurement scale, and reference method. Separating these parts is an important detail when describing labs, because our lay descriptions of specific labs often combine the analyte with the measurement property ("% hematocrit") or with the timing or sample ("fasting blood glucose") in ways that complicate the grouping and ordering of lab results across a population.

Summarizing laboratory results poses several challenges for the RWE analyst; the structure and taxonomy of LOINC codes helps with some, but not all, of these challenges. The LOINC database stores multiple synonyms for lab tests in addition to the fully specified name, which can help accelerate the mapping of imprecise text descriptions for lab tests. In addition, because many labs are ordered as panels of analytes measured from the same sample, LOINC links codes for the panel (57021-8 for "CBC W Auto Differential panel - Blood") to the (in this case, 30) results typically returned from the panel. Finally, for tests whose results are delivered as categorical values (e.g., tumor stages), LOINC provides standardized codes for answer sets (indicated with a character prefix of "LA") that reduce the risk of alternate spellings disrupting the grouping process ("Stage 4" vs. "Stage IV").

On the other hand, LOINC has developed a fairly open policy for accepting proposals of new lab tests for coding, which has greatly accelerated the scope of tests covered at the expense of enforcing canonical values for tests. That Hemoglobin A1c code above is actually one of nine different values that could be used, with some specifying variants in the reference standard or the analysis method *(Table 1)*. Unlike many of the diagnosis and procedure coding systems with which analysts are familiar, LOINC code values are not logically grouped together (HbA1c values are in a non-contiguous range from "4548-4" to "71875-9"), and while notes in the LOINC database indicate preferences for some codes over others, none are officially deprecated or retired. Therefore, the selection of appropriate codes by an analyst requires careful attention, and often requires consultation of LOINC's published list of the 2,000 most frequent codes observed by ordering volume to determine the preferential values among a range of alternates.

Access to LOINC reference materials is free, with some material requiring the creation of a free user account at https://loinc.org. The online search tool for LOINC codes is at https://search.loinc.org, although downloading the RELMA desktop application offers a few additional features not found in the online search tool. LOINC provides a quick start guide and helpful FAQs, as well as a more detailed user guide both for the LOINC code set and for the RELMA application.

SNOMED CT

SNOMED Clinical Terms (SNOMED CT) is an ambitious attempt to encode the full range of concepts that might be entered in an EMR. It is truly international in nature, resulting from the 1999 merger of one terminology project from the College of American Pathologists (formerly called the **S**ystematized **NO**menclature of **MED**icine), and the READ code project from the UK's National Health Service (NHS).

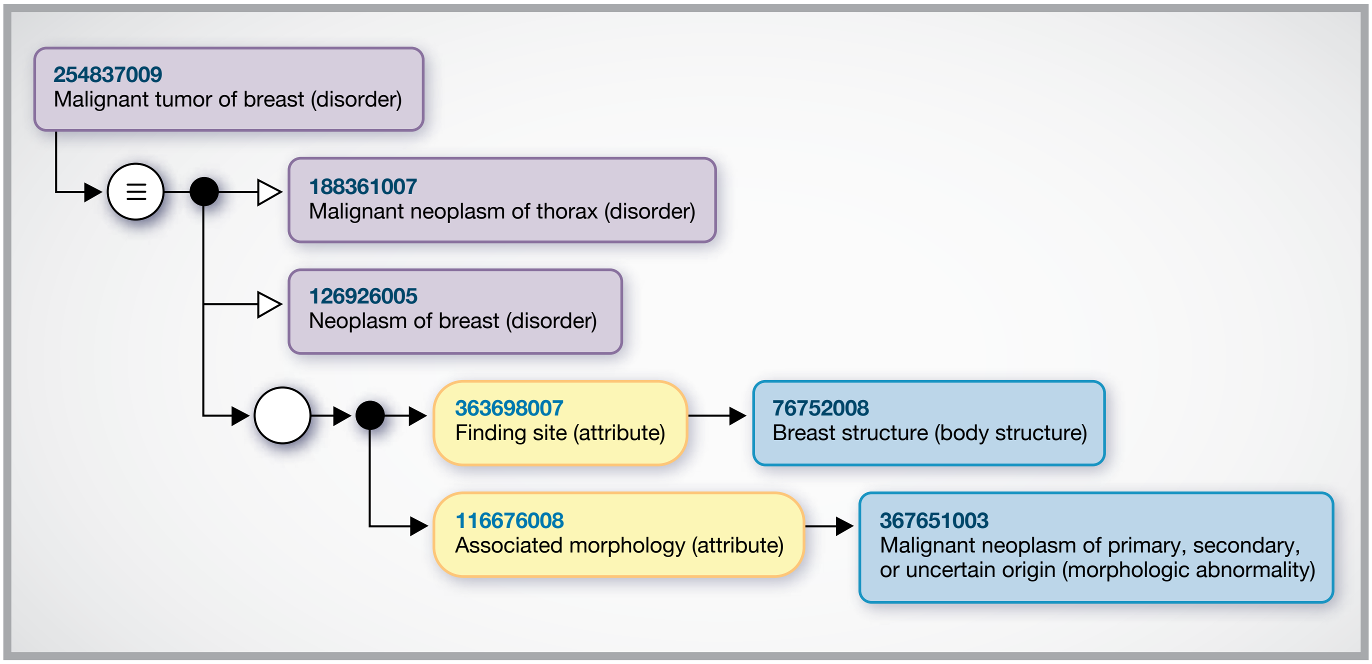
Nine countries with leading roles in health IT created the International Health Terminology Standards Development Organisation (IHTSDO) to acquire the rights to SNOMED CT in 2007. Membership in IHTSDO has since expanded to 24 member countries. Currently, IHTSDO maintains English and Spanish translations of SNOMED descriptions, and member countries have released 8 additional language or dialect translations.

SNOMED codes are between 6 and 18 numeric digits long, and all codes begin with a non-zero digit. Like LOINC, they contain a single check-digit at the end, and

Table 2. SNOMED CT Top Level Domains

Body structure (body structure)
Clinical finding (finding)
Environment or geographical location (environment / location)
Event (event)
Observable entity (observable entity)
Organism (organism)
Pharmaceutical / biologic product (product)
Physical force (physical force)
Physical object (physical object)
Procedure (procedure)
Qualifier value (qualifier value)
Record artifact (record artifact)
Situation with explicit context (situation)
SNOMED CT Model Component (metadata)
Social context (social concept)
Special concept (special concept)
Specimen (specimen)
Staging and scales (staging scale)
Substance (substance)

Figure 1. SNOMED CT Diagram Illustrating Multiple Relationships and Hierarchies for Concept "Breast Cancer"


254837009
Malignant tumor of breast (disorder)
≡
188361007
Malignant neoplasm of thorax (disorder)
126926005
Neoplasm of breast (disorder)
363698007
Finding site (attribute)
76752008
Breast structure (body structure)
116676008
Associated morphology (attribute)
367651003
Malignant neoplasm of primary, secondary, or uncertain origin (morphologic abnormality)

there is no order or structure to the numeric value of the code. For reasons beyond the scope of this article, SNOMED concept codes also contain a "00" in the second- and third-last digits (e.g., "73211009") that can help the RWE analyst recognize SNOMED concept codes.

Like LOINC, the text description of each code contains a fully standardized name and accepted synonyms. SNOMED CT code 73211009 corresponds to "Diabetes mellitus (disorder)" (fully specified name), "Diabetes mellitus" (preferred synonym), and "DM - Diabetes mellitus" (acceptable synonym). SNOMED CT organizes all of its codes in hierarchies, which include 19 top-level domains *(Table 2)*. The level of organization within these hierarchies varies widely, and is defined by one or more relationships (also with their own SNOMED codes) between concepts. Diagnoses for conditions, for example, are found within the "Clinical finding" domain, but often belong to multiple hierarchies based on relationships to concepts in the "Body structure" domain. To help keep all of this complexity organized, SNOMED has developed a diagramming system to show definitions of key concepts and their relationships *(Figure 1)*. Data analysts will occasionally need to dig into these concepts and relationships when determining which level of a hierarchy to use for selecting codes (and child codes) for a particular research question.

The ambitious scope of SNOMED CT means that its content will overlap with many of the coding systems used for diagnoses, drugs, labs, and procedures. Because of this, IHTSDO has supported multiple projects to map SNOMED CT codes to ICD-9, ICD-10, and LOINC. Other organizations have developed mappings of their own coding systems (e.g., RxNorm) to relevant SNOMED terms. In the near term, this means that one of SNOMED's great values for data analysts will be to offer alternative ways to group concepts when other coding systems fall short.

Access to SNOMED reference materials is free for research use. A variety of reference materials are available at http://www.snomed.org/, ranging from quick start guides all the way to technical implementation guides. The online search tool for SNOMED CT codes is at http://snomed.info/, which includes all of the currently published language translations.

RxNorm

RxNorm is a collection of drug names that have been normalized by the United States National Library of Medicine (NLM). The drug terms have been formalized to represent the primary components of a drug (ingredient[s], strength[s], and dose form) in a standard format, while linking the standardized name to the names found in commonly used drug vocabularies.

The desire to share the variety of existing drug terminologies used by healthcare systems and pharmaceutical manufacturers, and to develop a system to overcome known defects in the existing coding systems (such as National Drug Codes [NDC]) motivated the HL7 Vocabulary Technical Committee in 1998 to develop a better model for representing drug terms. In response, the RxNorm project began in 2002.

Table 3. RxNorm Drug Records Related to "Fluoxetine"

Term type (TTY) Name	Description	Example	RxNorm Concept Unique ID (RXCUI)
Ingredient	A compound or moiety that gives the drug its distinctive clinical properties	Fluoxetine	4493
Precise Ingredient	A specified form of the ingredient that may or may not be clinically active	Fluoxetine Hydrochloride	227224
Multiple Ingredients	Two or more ingredients appearing together in a single drug preparation	Fluoxetine / Olanzapine	406024
Semantic Clinical Drug Component	Ingredient + Strength	Fluoxetine 4 MG/ML	315953
Semantic Clinical Drug Form	Ingredient + Dose Form	Fluoxetine Oral Solution	372232
Semantic Clinical Drug	Ingredient + Strength + Dose Form	Fluoxetine 4 MG/ML Oral Solution	310386
Brand Name	A proprietary name for a family of products containing a specific active ingredient	Prozac	58827
Semantic Branded Drug Component	Ingredient + Strength + Brand Name	Fluoxetine 4 MG/ML [Prozac]	563784

EMR records that utilize RxNorm vocabularies achieve compliance with the 'Meaningful Use' requirements for electronic health records, which has greatly increased adoption in the U.S. RxNorm assimilates drug taxonomies from several global sources to expand the system's reach beyond the U.S.

Each RxNorm concept is identified by an 8-digit Concept Unique ID (RXCUI). Those familiar with existing drug coding systems understand that the existence of combination ingredients, multiple dosing and packaging variants, and different routes of administration, create substantial complexity for how drug concepts are represented and organized. RxNorm assigns RXCUI values at various levels of specificity, called term types or TTYs, in addition to a drug's complete clinical drug name (ingredient, strength, and dose form). Table 3 shows the many different levels at which the antidepressant fluoxetine may be represented, including its appearance in fixed dose combinations.

To manage the links between all of these RXCUIs for a single drug, RxNorm maintains a rich set of relationships among concepts. Each relationship between concept A and B has an exact reverse relationship mapped between concept B and A, as is the case in SNOMED CT. Examples of common relationship pairs in RxNorm include "Has brand name/Brand name of," "Has form/Form of," "Has ingredient/Ingredient of," "Has tradename/Tradename of," "Is a/Inverse is a," and "Has precise ingredient/Precise ingredient of." These relationship links allow analysts to navigate the variety of challenges associated with brand versus generic names; dose, form, and route variations; and fixed dose combinations to select the set of concepts most useful for analysis. However, they also demand greater precision from the analyst to understand which level(s) of specificity is required for selecting the drugs and forms of interest. Selecting RXCUIs usually also requires simultaneously selecting the relevant TTYs, or being prepared to navigate RxNorm's relationship links to filter and capture all the concepts of interest.

The RxNorm datasets and documentation are available for download at no cost from http://www.nlm.nih.gov/research/umls/rxnorm. The National Library of Medicine also provides free access to RxNav, a web-based tool for searching and traversing the RxNorm vocabulary. https://rxnav.nlm.nih.gov/. A desktop version of RxNav is also available for download.

Applications

Analysts working with data that include these newer coding systems will not need to be convinced of the need to understand and use them. An increasing number of data sources are leveraging these code sets to document clinical data, even if the codes were not used in the original data system. This is most clear in the case of datasets formatted for the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), now maintained by the Observational Health Data Sciences and Informatics (OHDSI) program. The

Table 4. ICD-9 and ICD-10 Concepts Mapped to SNOMED Concept for Breast Cancer in OHDSI ATLAS Browser

Code	Name	Standard?	Domain	Vocabulary
254837009	Malignant tumor of breast	Standard	Condition	SNOMED
174	Malignant neoplasm of nipple and areola of female breast	Non-Standard	Condition	ICD9CM
174	Malignant neoplasm of female breast	Non-Standard	Condition	ICD9CM
174.1	Malignant neoplasm of central portion of female breast	Non-Standard	Condition	ICD9CM
174.2	Malignant neoplasm of upper-inner quadrant of female breast	Non-Standard	Condition	ICD9CM
174.3	Malignant neoplasm of lower-inner quadrant of female breast	Non-Standard	Condition	ICD9CM
174.4	Malignant neoplasm of upper-outer quadrant of female breast	Non-Standard	Condition	ICD9CM
174.5	Malignant neoplasm of lower-outer quadrant of female breast	Non-Standard	Condition	ICD9CM
174.6	Malignant neoplasm of axillary tail of female breast	Non-Standard	Condition	ICD9CM
174.8	Malignant neoplasm of other specified sites of female breast	Non-Standard	Condition	ICD9CM
174.9	Malignant neoplasm of breast (female), unspecified	Non-Standard	Condition	ICD9CM
175	Malignant neoplasm of nipple and areola of male breast	Non-Standard	Condition	ICD9CM
175	Malignant neoplasm of male breast	Non-Standard	Condition	ICD9CM
175.9	Malignant neoplasm of other and unspecified sites of male breast	Non-Standard	Condition	ICD9CM
198.81	Secondary malignant neoplasm of breast	Non-Standard	Condition	ICD9CM
C50	Malignant neoplasm of breast	Non-Standard	Condition	ICD10
C50.0	Malignant neoplasm: Nipple and areola	Non-Standard	Condition	ICD10
C50.1	Malignant neoplasm: Central portion of breast	Non-Standard	Condition	ICD10
C50.2	Malignant neoplasm: Upper-inner quadrant of breast	Non-Standard	Condition	ICD10
C50.3	Malignant neoplasm: Lower-inner quadrant of breast	Non-Standard	Condition	ICD10
C50.4	Malignant neoplasm: Upper-outer quadrant of breast	Non-Standard	Condition	ICD10
C50.5	Malignant neoplasm: Lower-outer quadrant of breast	Non-Standard	Condition	ICD10
C50.6	Malignant neoplasm: Axillary tail of breast	Non-Standard	Condition	ICD10
C50.8	Malignant neoplasm: Overlapping lesion of breast	Non-Standard	Condition	ICD10
C50.9	Malignant neoplasm: Breast, unspecified	Non-Standard	Condition	ICD10

OMOP CDM standardizes data for more interchangeable, globally consistent analyses by relying heavily on these three systems as the standard vocabularies for most clinical facts. Data that are translated into OMOP CDM format have their NDC drug codes converted to RxNorm, their labs converted to LOINC, and their diagnoses converted to SNOMED CT.

OHDSI has created its own browser of codes that can be used within an OMOP CDM, called ATLAS (http://www.ohdsi.org/web/atlas/#/home). This tool allows users to search for specific code values or text descriptions from any of the preferred clinical vocabularies or the non-preferred vocabularies that OHDSI has mapped to them. A search for "diabetes mellitus" returns over 1,000 different records, to which several filters can be applied, including coding system, "domain" (type of clinical fact), and whether the concept is preferred ("standard") in OMOP CDM.

Given the mapping between code sets in ATLAS, the browser has the helpful capability of searching related concepts within and across code sets. This can be useful even if an analyst is working with a dataset that does not contain these newer coding systems. For example, many U.S. data sources in the next several years will include a mixture of ICD-9-CM and ICD-10 diagnosis codes for similar conditions. ICD-9 to ICD-10 mapping schemes exist, but the process of using them can be cumbersome, and there is a reasonable risk of using them improperly.

However, the cross-mappings available in ATLAS can permit users to start with concepts that are closer to their concept of interest, and then find the mapped values in their code sets of interest.

For example, selecting the SNOMED CT code for "Malignant tumor of breast (disorder)" (254837009), and then selecting its related concepts within the ATLAS browser, identifies the 14 distinct ICD-9-CM codes and the 10 ICD-10 codes that have been directly mapped *(Table 4)*. Indeed, if the analyst also needed to find codes to replicate the analysis in a British data source, the same search could be used to select the 31 READ codes linked to the same SNOMED concept.

Despite its power, the ATLAS browser has its limitations when exploring the utility of these newer code sets. The browsers specific to each code do a better job of preserving some of the more detailed documentation and the concept relationships within each code set. The SNOMED browser represents its synonyms and concept diagrams better than ATLAS; the LOINC browser excels at linking analytes to their panels and answer sets; and, the RxNav application includes RXCUI values at more TTY levels than does ATLAS. Analysts will be well served by toggling between each code set's own browser and the ATLAS browser to narrow down the clinical concepts most useful to their research question.

Conclusions

An increasing number of provider-based data sources use or reference global code sets such as LOINC, SNOMED CT, and RxNorm. Local systems are turning to global code sets because of pressure to exchange clinical information with other providers' data systems, and are often incentivized to use global codes by payers or regulatory authorities. As RWE analyses increase in complexity, command of these code sets will become a foundational skill for the RWE analyst. Conversion of databases to common data models will also accelerate the importance of understanding global code sets in greater detail.

As we have shown, however, understanding these global codes can help manage confusion inherent in traditional local code systems, even before they appear in a desired data source. The mapping initiatives required to make these code sets global can assist the RWE analyst with code translation and replication. The hierarchies and other relationships embedded in global code sets can also help the RWE analyst define concepts more precisely without reliance on local billing or coding experts. Free tools and documentation exist for learning most of these code sets, as well as understanding their overlap and relationships to older coding systems. Few barriers exist to developing the coding skills required of the next generation RWE analyst!

For more information, please contact Don.O'Hara@evidera.com or Vernon.Schabert@evidera.com.

REFERENCE

[1] Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, Fiers T, Charles L, Griffin B, Stalling F, Tullis A, Hutchins K, Baenziger J. Logical Observation Identifier Names and Codes (LOINC) Database: A Public Use Set of Codes and Names for Electronic Reporting of Clinical Laboratory Test Results. *Clin Chem.* 1996 Jan; 42(1):81-90.