Why normalizing your clinical and claims-based data into standard terminologies is critical in supporting forward-thinking initiatives such as big data analytics, population health management, and semantic interoperability among systems.

EXECUTIVE SUMMARY

Terminology is core to everything in healthcare—from procedures to results to diagnoses, healthcare IT systems (HIT) represent clinical concepts in various coded clinical terminologies or free text. Unfortunately, the explosive growth in HIT has resulted in patient data being scattered across an array of rapidly proliferating IT systems—each with their own way of representing clinical terms. The variety in terminologies and the variability in how they are used by the systems that compose the healthcare IT ecosystem has created an environment of data locked in silos. This terminology barrier must be overcome if we are to recognize the national effort around increased interoperability, transparency, and collaboration within our healthcare system.

Data normalization strategies that automatically map local content to terminology standards and translate data between standards are required to eliminate ambiguity of meaning in clinical data. From mapping of local codes to standards to creating crosswalks between terminologies to aggregating terminologies into clinical-friendly views of patient data, an enterprise-class data normalization solution can have a significant impact on an organization’s ability to achieve compliance to standards and to capitalize on new business models. By establishing a foundation for achieving semantic interoperability, data normalization enables the data sharing and aggregation aims of the federal Meaningful Use (MU) initiative and contributes to the financial success of emerging healthcare delivery approaches such as Accountable Care Organizations (ACOs).

Though data normalization in itself is not the endgame, it is the foundational process that enables healthcare organizations to answer critical questions, better report to registries, better report quality measures, and better analyze patient care.
**INTRODUCTION**

It’s a classic good news/bad news situation.

First the good news: The steady growth of healthcare information technology (HIT) in recent years has ushered in a new era of automation. Technology adoption is slated to continue. Research and Markets forecasts that the North American HIT market will reach $31.3 billion by 2017, compared with $21.9 billion in 2012. The use of enterprise health records (EHRs) among physicians has expanded dramatically since 2001. That year, only 18 percent of office-based physicians reported using some form of EHR system, according to the Centers for Disease Control and Prevention. But in 2013, 78 percent of physicians said they were using EHRs, the federal agency noted. A recent Oracle poll of 333 U.S. and Canadian C-level executives revealed that healthcare organizations were amassing 86 percent more data in 2012 than they did in the previous two years. The future of healthcare will be built on data—data to support population health management, and analytics to improve outcomes.

But there’s a downside to this unprecedented adoption of healthcare technology. Today, patient data is scattered across an array of rapidly proliferating IT systems—each with their own way of representing clinical terms. Healthcare has a rich history of using multiple descriptions—is it a heart attack, acute myocardial infarction, or cardiac arrest? Hypertension, arterial hypertension, or high blood pressure? Ibuprofen or Advil? The lack of a common clinical vocabulary across disparate systems creates communication barriers, which hinders the ability to coordinate care and aggregate data for analysis. These disparate terminology lexicons must be normalized into standard terminologies so that the meaning of the clinical data is unambiguous. When data is normalized across the continuum of care, a more accurate picture of performance can be achieved for better reporting and analytics.

**TRENDS DRIVING DATA NORMALIZATION**

Health information exchange, interoperability, big data, analytics, quality measurement, population health, and risk sharing are all leading buzzwords across the healthcare landscape that hold great promise toward moving the industry closer to improving quality, reducing costs, and increasing patient satisfaction. As the industry looks to increase momentum with these movements, key stakeholders are increasingly realizing that data normalization is a fundamental component of the equation.
A normalized information model lets organizations share a common vocabulary and promotes semantic interoperability.

A survey conducted by Premier of 115 C-suite healthcare executives revealed that nearly 75 percent report integrating clinical and claims data to better manage population health in order to support efforts with accountable care organizations (ACOs). "Making sense of data" was cited as one of the biggest barriers in making accountable care work.

There is a lot of noise in the system today, but two key trends consistently boil to the top: semantic interoperability to support new delivery models, and clinical data repositories to support big data analytics.

### Semantic Interoperability to Support New Delivery Models

Emerging healthcare delivery models such as ACOs, patient-centered medical homes (PCMHs), and pay-for-performance (P4P) all depend to some degree on interoperability—the ability to share data. ACO participants, primary care providers, specialists, hospitals, and labs, among others, must share information to coordinate care for the patient population under management. The PCMH model, which seeks to transform primary care, involves orchestrating care across physicians, physician assistants, nurses, social workers, and other professionals. And pay-for-performance programs depend on the ability to aggregate data from different clinical sources for quality reporting purposes.

Fueled by the Meaningful Use initiatives, the adoption of EHR and healthcare information exchange (HIE) technology has resulted in an exponential growth in health data and progress toward the first two levels of interoperability: Foundational interoperability, which enables data exchange without requiring IT systems to interpret data, and structural interoperability, which defines syntax of the data exchange, allowing systems to interpret data at the data field level. By making use of data transport and syntactic standards such as HL7 that define messaging structure, today's HIT maturity has gone a long way in establishing the foundational elements required for interoperability among disparate IT systems. As the industry moves into the more advanced stages of Meaningful Use and other national initiatives, healthcare organizations are now turning their attention to semantic interoperability—the ability for IT systems to understand the meaning of the data that is being shared. Two local drug codes, for instance, may describe the same drug in different terms. The lack of semantic interoperability limits communication in such situations.

Unfortunately, health IT innovation has developed in isolated pockets, with initial development taking place in billing and claims, followed by localized development of ancillary clinical systems such as laboratory,
Data normalization can assist in the transformation of stored data into an integrated, focused repository accessible to users across your enterprise.

radiology, pharmacy, and, more recently, EHR systems. Although MU has created new incentives for EHR vendors to adopt language standards such as SNOMED CT®,8 many HIT systems in today’s market still use local or proprietary codes to collect and store patient data. Due to the fragmented adoption of terminology standards, much of that data remains locked within isolated IT systems. Attempts to share data in this environment can lead to errors because systems represent data differently through different coding and terminology schemes. A recent report published by the Journal of the American Medical Informatics Association9 noted “615 observations of errors and data expression variation” across the 21 EHR technologies examined. The interoperability barriers underscore the need for data normalization—patient data must be normalized into standard code sets that can be easily accessed by all providers.

Clinical Data Repositories (CDRs) to Support Big Data Analytics

A clinical data repository is an aggregation of granular, patient-centric health data, usually collected from multiple-source IT systems and intended to support multiple uses such as monitoring caregiving processes in near real time, collecting data for quality measures, building predictive models, understanding cost of care, or identifying at-risk populations for intervention. Collectively, the industry refers to these initiatives as “big data.”

From administrative data sourced by payers to unstructured text and clinical data sourced by multiple provider EHRs, the fragmentation of data contained within CDRs has been identified as a significant obstacle to leveraging big data.10 Diagnosis data, for example, might be stored in a variety of forms, including text (such as “Diabetes Mellitus”); standardized diagnosis codes (such as the ICD-9-CM code “255.10”); or local, proprietary codes that have meaning only in internal applications. Many systems’ efforts are falling short due to the complexities in normalizing all this data into a format that renders it useful.

For example, Advocate Health Care required roughly 18 months of work to merge, clean, and organize patient data from multiple sources, including external insurance claims, internal financial and demographic records, and multiple electronic medical-record systems. The work was necessary, executives said, to deliver the community-wide health improvement—and savings—that come from treating fewer chronically...
ill patients who show up on a hospital’s doorstep with complex and hard-to-treat medical emergencies. The effort required data extraction across multiple databases with different codes that needed to be standardized. Comprehensive data “is the linchpin,” said Tina Esposito, vice president for the Center for Health Information at Advocate.

A CDR that has not normalized its terminologies could become problematic quickly. Take, for example, a CDR that stores diagnosis data as text strings. A simple query asking for all the instances in which “diabetes mellitus” is the admitting diagnosis will retrieve only records in which the diagnosis was recorded as the text string “diabetes mellitus.” Such a query, however, will not recognize “diabetes,” “IDDM,” “Type II DM,” “DM,” and “adult onset diabetes”—all legitimate terms used by clinicians. This form of unrecognized equality is one of the primary roadblocks to retrieving meaningful information from these types of data warehouses. Conversely, a query using the strings “diabetes” and “DM” would be equally inappropriate, incorrectly identifying patients with “diabetes insipidus” and “family history of diabetes” as having diabetes mellitus. A better approach is to employ a normalization solution that recognizes the many term variants used to describe diabetes mellitus and their precise mappings to terminology standards.

Mapping terms to standards is a good start but it’s only a partial solution. Organizations often need to define code groups (also known as “value sets”) to ensure that analyses are consistent across the enterprise. Continuing with the diabetes example, suppose terms have been mapped to ICD-9-CM, and an analyst needs to identify all patients being treated for diabetes. She may be tempted to define diabetes using “all diagnosis codes that are descendants of the ICD-9-CM code for diabetes mellitus (250).” But what about patients with secondary diabetes mellitus (249)? Or gestational diabetes mellitus (648.8)? What about patients who clearly have diabetes mellitus based on complications such as diabetic retinopathy (362.0)? Because there isn’t a single right answer to these questions, different analysts are likely to employ different approaches, resulting in apples-to-oranges comparisons. Analysts are able to achieve more consistent results when they use a normalization solution that maintains common organizational definitions of code groups. Ideally, a normalization solution should also help keep code groups up to date when terminologies add or invalidate codes – a common occurrence in SNOMED CT, RxNorm, and many other standards.

These processes demonstrate how data normalization solutions can help transform a hodgepodge of structured and unstructured data into
Data normalization is required to standardize local content to terminology standards, and to semantically translate data between standards to eliminate ambiguity of meaning.

**LAYING THE NEEDED FOUNDATION TO SUPPORT THE FUTURE OF HEALTHCARE: STANDARDIZING LOCAL CONTENT TO TERMINOLOGY STANDARDS AND SEMANTICALLY TRANSLATING AMONG STANDARDS**

As data is aggregated from disparate systems across the continuum, one of the first steps in using this data rests with the ability to normalize clinical and claims data into standardized terminologies to ensure an accurate picture of health across identified patient populations. Data normalization is required to standardize local content to terminology standards, and to semantically translate data between standards to eliminate ambiguity of meaning.

Lab and Drug Mapping Examples
In general, data normalization establishes a foundation for achieving semantic interoperability and creates an infrastructure that enables data sharing and aggregation.

**Standardizing Local Content to Terminology Standards**

Vendor applications are starting to use standardized languages to get everyone on the same page. For example, medical terminology is moving toward the use of SNOMED CT; differences in drug descriptions can be resolved through the use of RxNorm, which provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management. Still, many healthcare systems have their own local content, and differences in terminology usage still exist among legacy vendors. An example (and there are thousands of scenarios such as this) is the concept of a Hemoglobin A1C test. Different systems will call this test different things: HbA1C, A1C, HA1C, A1C Hemoglobin, Hb A1c (%), HEMOGLOBIN A1c, and so on. Attempting to trend Hemoglobin A1C results to ascertain whether you are properly controlling your diabetic population in this scenario is next to impossible. However, if all your lab data was standardized to the lab standard LOINC, HIT systems could automate lab trend analytics.

The key to making analytic initiatives work is ensuring your ecosystem consists of structured data—data that can be mined and shared with other systems. Structured data is available in a controlled format or vocabulary, rather than in free text. The consistency of structuring data allows for statistical research, business intelligence reporting, and data interoperability that cannot be obtained with unstructured or narrative data.

From being able to correct the 50 different misspellings of the drug Lisinopril and correctly classifying it as an ACE inhibitor, to understanding that CPR and cardiopulmonary resuscitation refer to the same medical procedure, standardizing the concepts that those terms represent ensures that analytic results include all the data contained within the healthcare system.

**Semantic Translations Among Different Terminology Standards**

A single reference terminology or classification system cannot, by itself, serve all of the purposes for which health information is currently used or will be used in the future. Terminologies and classifications are designed for distinctly different purposes and satisfy diverse user data requirements. Multiple terminologies as well as classification systems are necessary to capture and effectively use the breadth and depth of clinical data in an EHR.

Classification systems such as ICD-9-CM, ICD-10-CM, and ICD-10-PCS, which group similar diseases and procedures and organize related
entities for easy retrieval, were never intended for the primary documentation of clinical care. They are typically used in scenarios for which aggregation is helpful, such as measuring quality, monitoring resource utilization, or processing claims for reimbursement. On the other hand, due to its ability to finely define individual clinical concepts and their relationships, the SNOMED CT reference terminology is ideal for codifying clinical information captured in an EHR during an encounter. Although classification systems and reference terminologies serve distinct purposes, organizations that can semantically translate (or map) between these systems are better positioned to use this information for secondary purposes such as determining compliance to evidence-based protocols irrespective of the underlying terminology contained within the healthcare system’s CDR.

Let’s say, for example, you are an ACO that employs an incentive-based payment model from Centers for Medicare & Medicaid Services. Asthma hospital admissions are, for the most part, preventable, and your incentive payment model is set up such that if your admission rates are above 5 percent of your population of patients with asthma, you are penalized. The measure is a simple formula:

\[
\frac{\text{Patients discharged from the hospital with principal diagnoses code (ICD-9) for asthma}}{\text{Total population of patients with principal diagnoses code (ICD-9) for asthma}}
\]

Now, let’s say your quality measure analysis engine responsible for computing this measure is solely based on claims (ICD-9 codes) data that is sourced from a payer that is part of the ACO. However, you are also getting EMR data feeds that contain patients who have SNOMED CT codes for asthma. If you’re not able to map (semantically translate) the SNOMED CT codes to ICD-9, you may be missing patients who belong in your denominator. You face the same issue or risk if you’re getting free-text problem and diagnosis data feeds. This situation will skew your stats in a bad way, making it look like a high percentage of your asthma patients are ending up in the hospital—which appears that you’re doing a lousy job of taking care of them. It also means your ACO reimbursement will be lowered because you aren’t meeting this measure.

As another example, consider the case in which your health system’s CDR has been populated with inpatient data coded in ICD-10-PCS and outpatient data coded in CPT-4. In order to interrogate your data, you must be able to semantically translate between these code systems. If you need to pull all patients who had an upper GI endoscopy, an
Localized terminologies have been proliferating among individual providers, labs, and other healthcare entities over the years. The key task is to be able to map between local terminologies and standards.

informatics analyst would need to find CPT 91110 gastrointestinal tract imaging, intraluminal (e.g., capsule endoscopy), esophagus through ileum, with interpretation and report and the ICD-10-PCS 0WJP8ZZ inspection of gastrointestinal tract, via natural or artificial opening endoscopic approach codes. A normalization platform could enable analysts to more easily choose the known CPT® code and quickly find the relevant ICD-10-PCS codes to search for.

A mediation layer that can map between clinical terminologies enables health systems to build business applications and workflows regardless of the underlying coding systems.

WHAT TO LOOK FOR IN A DATA NORMALIZATION PLATFORM

For many resource-strapped IT departments, the business case for leveraging the expertise of third-party consultation and an enterprise terminology management platform is an easy one. By leveraging a platform that provides the software, content, and consulting solutions to effectively map, translate, update, and manage standard and enhanced clinical terminologies on an enterprise scale, healthcare organizations across the board can take the guesswork out of this complex process. Vendors lacking expertise in this area are also turning to third-party groups to build infrastructures in order to support the future of some of healthcare’s greatest challenges—ACOs, data warehousing, population health management, ICD-10, and MU.

A solution that takes on all of the facets of data normalization will include several characteristics. The main focus of any solution will be on terminology mapping. Today’s HIT mapping needs are immense, and although the solution must have web-based support for manual mapping workflows, the solution should be able to automate the majority of enterprise mapping needs. Automated mapping algorithms that can be seamlessly integrated into enterprise systems via real-time web service calls can normalize full catalogs or distinct elements from cryptic and poorly maintained source data into standard terminologies. At the core of any solution are the underlying content databases that the mapping algorithms and end user-facing tools tap into to accomplish their goals.

Alternatively, a healthcare entity might prefer to use a professional services group associated with the data normalization solution. That group would use the same mapping tools, and apply its own workflow expertise, to assist the healthcare customer with its mapping efforts.
Localized terminologies have been proliferating among individual providers, labs, and other healthcare entities over the years. The key task is to be able to map between local terminologies and standards. Moreover, this mapping should occur automatically because the vast amount of data healthcare organizations will need to exchange every day dictates that a large percentage of it must be handled automatically as well. Manual interpretation won’t work in a high-volume environment. Even though maps are always subject to human review, the goal is to automate as much of the mapping process as possible using automated algorithms. Furthermore, these algorithms must be callable in real time via web services so that they can be seamlessly integrated into enterprise-class HIT systems. As automated mapping algorithms, including the ability to learn over time, are developed for multiple use cases and become increasingly sophisticated, the level of human review at the individual code level diminishes, and workplace roles turn to the development and maintenance (including quality control) of maps for a variety of use cases and the development of algorithmic translation rules.

Roles-Based Workflow for Modeling Local Content and Collaboration

While the goal is to automate as much mapping as possible, a normalization solution should support roles-based collaboration workflows for the modeling of local content and the subsequent mapping of that content to standards. Here’s why: The need to manage custom content becomes inevitable, regardless of whether an organization is a provider or a health plan. There’s more than one way to go about mapping data; it all depends on a healthcare organization’s objectives for a given data exchange. These workflows should be tightly integrated into the aforementioned automated matching algorithms. For instance, in cases when an automated match is not possible, a user-facing interface should present a task list of “fallouts” that can be manually mapped. Furthermore, these manual maps should be fed into the algorithms so that they can learn over time.

A Comprehensive Terminology Database and Content Management System

At the core of any solution are the underlying content databases that the mapping algorithms and end user-facing tools tap into to accomplish their goals. Today’s solutions must include a database that contains:
By ensuring the clinical data which the ACO is built is complete and semantically understood, data normalization is one tool an ACO can use to make that delivery model work and qualify for shared savings.

- All relevant standard terminologies (e.g. ICD-9, ICD-10, SNOMED CT, RxNorm, LOINC)

- Multilateral maps between various terminology code sets (e.g. ICD-9/ICD-10 to SNOMED CT)

- The ability to manage code groups that can be used to represent a single clinical concept (e.g., ICD-9-CM, ICD-10-CM, and SNOMED CT codes defining patients who have a history of myocardial infarction)

- A synonym library of provider (e.g., “ank fx”) and consumer-friendly (e.g., “nosebleed”) terminologies and their mappings to standards

- Management interfaces so that users can manage terminology updates and model their own local content as needed

A data normalization platform helps healthcare organizations on a number of fronts.

In general, data normalization establishes a foundation for achieving semantic interoperability and creates an infrastructure that enables data sharing and aggregation. But there are also specific payoffs from the profit and compliance perspectives.

Data normalization, as noted, helps promote the interoperability aims of the federal Meaningful Use initiative. Healthcare providers and hospitals that comply with the government’s MU criteria—which will increasingly involve an interoperability component—become eligible for incentive payments. Provider incentive payments will be available through 2016. On the other hand, financial penalties are slated to go into effect in 2015 for providers that fail to transition to EHR technology and adopt the Meaningful Use requirements. So, data normalization can help providers obtain the carrot, avoid the stick, and improve profitability.

Data normalization also contributes to the financial success of emerging healthcare delivery approaches such as ACOs. An ACO will need to integrate claims and clinical data in order to provide an accurate picture of health across the patient population it manages.

ACOs affiliated with the Medicare Shared Services Program have an opportunity to receive a cut of the savings that result from improved patient care and greater efficiency. But, as with Meaningful Use, there is the risk of a financial penalty for missing the mark. By ensuring the clinical data which the ACO is built is complete and semantically understood, data normalization is one tool an ACO can use to make that delivery model work and qualify for shared savings.
1 North American Healthcare IT Market by Application Delivery Mode & Component – Forecasts to 2017
   http://www.researchandmarkets.com/research/Hzq965/north_american

   http://www.cdc.gov/nchs/data/databriefs/db143.htm

3 From Overload to Impact: An Industry Scorecard on Big Data Business Challenges

4 Spooner, William, 2013. Making Information Flow: Even with EHRs, There’s More Than Meets The Eye
   http://www.cio-chime.org/chime/press/CaseStudy/Sharp_Case_Study.pdf

5 Accountable Care Organization and Population Health Management Trends

6 What is Interoperability?
   http://www.himss.org/library/interoperability-standards/what-is?navItemNumber=17333

7 Health Level Seven
   http://www.hl7.org

8 Systematized Nomenclature of Medicine--Clinical Terms
   http://www.ihtsdo.org/snomed-ct/

9 Are Meaningful Use Stage 2 certified EHRs ready for interoperability? Findings from the SMART C-CDA Collaborative
   http://jamia.bmj.com/content/early/2014/06/26/amiajnl-2014-002883.short?g=w_jamia_ahead_tab

10 Big Data in Healthcare

11 Data Collection Could Stump Next Phase in Predictive Analytics
    http://www.modernhealthcare.com/article/20140712/MAGAZINE/307129969

12 http://www.nlm.nih.gov/research/umls/rxnorm/

13 Spooner, William, 2013. Making Information Flow: Even with EHRs, There’s More Than Meets The Eye
   http://www.cio-chime.org/chime/press/CaseStudy/Sharp_Case_Study.pdf

14 Data Management for Population Health

15 Logical Observation Identifiers Names and Codes
    www.loinc.org

16 Futrell, Kim, 2013. Structured Data: Essential for Healthcare Analytics and Interoperability

17 Bowman, Sue 2013. Coordinating SNOMED-CT and ICD-10: Getting the most out of Electronic Health Record Systems
    http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_027621.hcsp?dDocName=bok1_027621

18 International Classification of Diseases
    http://www.who.int/classifications/icd/en/

19 Current Procedural Terminology
OUR ENTERPRISE TERMINOLOGY MANAGEMENT SOLUTION

We provide you with software, content, and consulting solutions that map, translate, update, and manage standard and enhanced clinical terminologies.

Our global team of developers, clinical professionals, terminology domain experts, and other healthcare information technology specialists have built an enterprise clinical terminology management platform that enables the information liquidity required to support some of healthcare’s toughest challenges, such as Meaningful Use compliance, ICD-10 conversion, population health management, analytics, ACOs, and semantic interoperability among systems.

ABOUT THE AUTHOR

Jason Wolfson is Vice President of Marketing and Product Management at Wolters Kluwer Health.

Jason can be reached at jason.wolfson@wolterskluwer.com

Additional information
To learn more about Data Normalization, visit and subscribe to our blog at http://blog.healthlanguage.com/

To learn more about our Enterprise Terminology Platform, visit http://info.healthlanguage.com/demo-request