Data Mapping

Save to myBoK

by Susan McBride, PhD, RN; Richard Gilder, BSN, RN, CNOR, BCNI; Robert Davis, MS; and Susan Fenton, MBA, RHIA

Exchanging health data electronically requires some cartography. HIM professionals not involved with data mapping now will be soon.

One essential component of health IT interoperability, and the improved care and efficiency it offers, is data mapping. Data mapping involves "matching" between a source and a target, such as between two databases that contain the same data elements but call them by different names. This matching enables software and systems to meaningfully exchange patient information, reimbursement claims, outcomes reporting, and other data.

Data mapping is a broad technical function occurring in many different settings with many different databases, data sets, standards, and terminologies. If HIM professionals are not currently involved in data mapping in their organizations, there is every reason to expect they will be in the near future. Data mapping will continue to become integrated into US healthcare as the industry moves to electronic health records. Mapping will play a key role in moving data from setting to setting and use to use, from informing patient care to informing national policy decisions.

Many Maps to Draw

Maps are created for many purposes, including exchange of data for patient care purposes, access to longitudinal data, reimbursement, epidemiology, public health data reporting, and reporting to regulators and state data organizations. A data map's purpose is known as its use case.

Maps can be created between databases where the data element "FName" in one database equals the data element "First_Name" in the second database. Maps can also be created between data sets. For example, the old UB92 form for acute care hospital billing in the uniform billing standard maintains a source of payment code; the new HIPAA electronic transaction standard, 837 institutional (837i), requires a claim filing indicator code. This translation of similar information requires additional data to accurately map the transition from the older method to the new transaction format. Standards can be mapped to databases, which in turn translate the information from one standard to another. A database can act as a translation key from one source to the next, providing additional information needed to accurately map the information.

Mapping can also facilitate the exchange of data. A common example is the transmission of electronic data using the Health Level 7 (HL7) standard. The content of the different HL7 messages has to be defined or mapped from a database. Finally, terminologies, as well as classifications and coding systems, can be mapped between each other. An example that may soon become relevant to everyone in healthcare is mapping between ICD-9-CM and ICD-10-CM.

Data mapping in healthcare is ubiquitous and will become more obvious as data currently in paper records increasingly converts to electronic format. Its accuracy is vital because there is opportunity for error at each relay point in the system.

The Elements of Data Mapping

Designing a data map begins with identifying the source of the map (the database, data set, or terminology being mapped *from*) and the target (the database, data set, or terminology being mapped *to*). A basic data map is shown below. Versions of both the source and target should be tracked, and there must be a mechanism to reflect the version history of the map.

Data Map Basics

Data mapping matches from a source to a target so that the two may exchange data meaningfully. Common sources and targets include databases, data sets, standards, and terminologies.

Unidirectional mapping goes from the source to the target. Bidirectional maps translate in both directions. Not all maps can be bidirectional; for instance, when multiple and differing terms in the source map to a single term in the target.

In some maps, a database can act as a translation key from one source to the next, providing additional information needed to map the information.

Maps may be unidirectional (mapping only from the source to the target) or bidirectional (also mapping back from the target to the source). Not all maps can be bidirectional. For example, the map from SNOMED to ICD-9-CM cannot be reversed, since it is common for many detailed and different SNOMED concepts to map to a single ICD-9-CM code. Reversing the map is not possible because one ICD-9-CM code would point to many different SNOMED concepts.

Likewise, the map from the old UB92 billing standard to the new 837i standard cannot be bidirectional. The map from the payer classification structure in the old format cannot be mapped back on some of the fields once translated into the new transactions format. This is the case in   payer classification. In the old format, charity care was an option, whereas in the new format charity care is either mapped to unknown or self-pay. There is no option for charity care in the new transaction format.

Mapping projects must carefully consider the data source and the intended use of the data for both primary and secondary purposes in order to ensure accuracy. The information collected (the system input) and the subsequent information produced from it (the system output) influence each other in what is known as a feedback control loop.[1] Data collection system design must first take into account the purpose of the data collection in order to produce a correct picture of the real world. Then the system must provide accurate and valid stored data, so people in an organization can create useful products or make rational decisions. This emphasis on purpose is also applicable to data sets mapped from one standard to the next within healthcare. For example, a system using hospital discharge data must carefully consider its limitations, since the primary purpose for this data set in the US is for billing, though it is also used for clinical care and quality management.[2]

Data Flow in Healthcare

The illustration below demonstrates the feedback loop of information as it flows from one source to the next via data maps. Following the flow illustrates how data within the electronic health record matures as it is distributed. Data must be accurately mapped from one system to the next, because there is opportunity for error at each relay point throughout the system.

In the diagram, data maps take data from a medical record, translate it into meaningful information in SNOMED, map to ICD-9-CM, which is transmitted into the electronic national uniform billing standard (837 institutional), paid by the

payers, and ultimately moved into an aggregated public use data file.

Common languages such as SNOMED and ICD-9-CM (in the future, ICD-10-CM) communicate data from one clinician to the next consistently and enable analysis of the information in aggregated data sets for quality, patient safety, and health policy decisions. Public use data files are submitted in many states to the Health Care Utilization Project National Inpatient Sample, which is used for health policy research and to investigate national trends in health. Data integrity in mapping is thus critical not only to patient care but critical also to health policy decisions for national health.
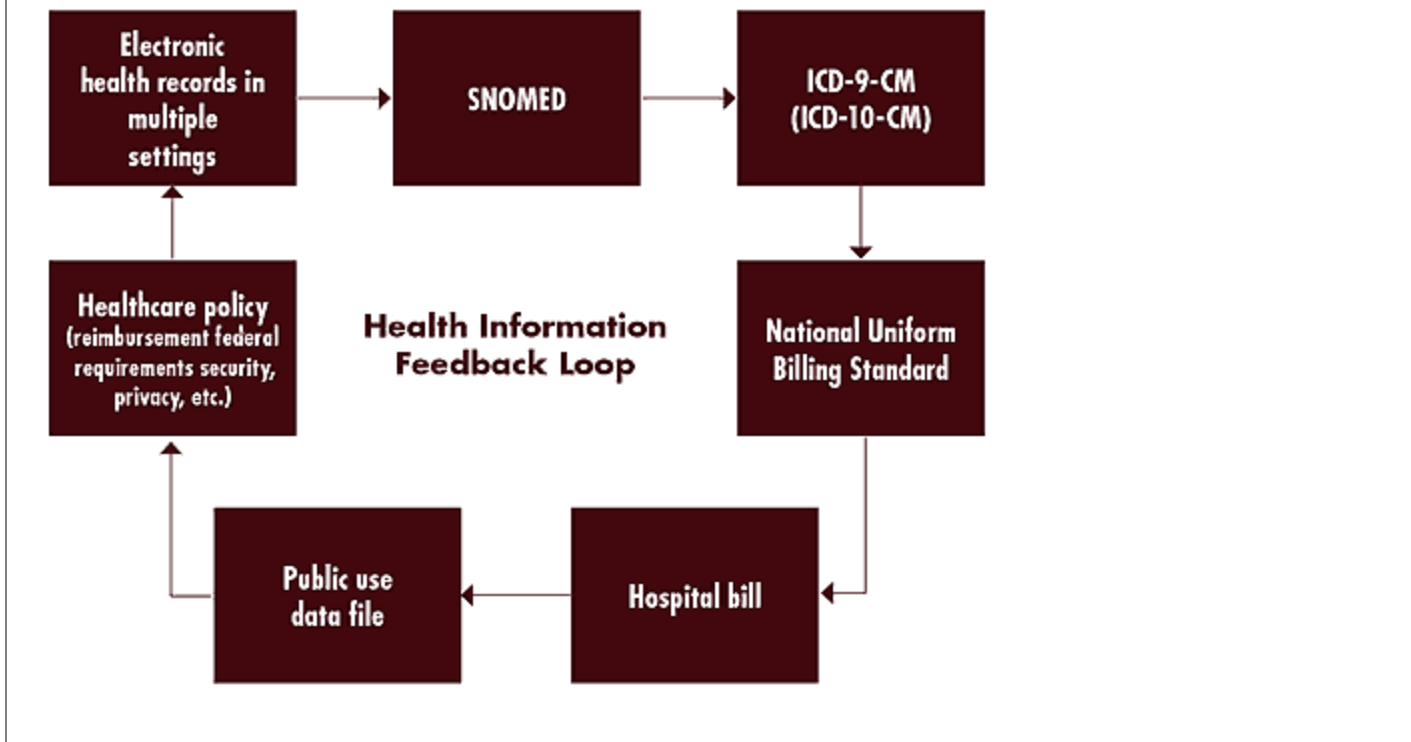
The data map's use case is the core upon which the guidelines and heuristics, or rules, for creating the map are specified. Guidelines and heuristics must be very detailed, because all data maps must be reproducibile. That is to say, once a map has been created, others not involved in the creation of the map should be able to verify the map's accuracy by reproducing the map following its guidelines and heuristics. The reliability of the factual content of data sets synthesized by mapping from one or more source data sets is often called into question, based upon the lack of demonstrated integrity in the map and mapping process itself as well as its stability over time. Critics focusing on real disparities of this nature have actually prevented some data set maps from ever being published, because of demonstrated flaws in the map.[2]

Because of a potential impact upon outcomes, the specter of clinical and policy errors resulting from decisions based upon poorly mapped data cannot be taken lightly. When mapping any healthcare data, this is a prime data integrity issue. All maps must be subject to quality controls and validation during the creation and updating processes. All organizations creating and updating maps must have a process in place to ensure the quality of the data map. In some instances, especially where patient care is the final product, it will make sense to perform 100-percent quality review and validation prior to use of the map. For other less critical applications, the organization may decide to select a random sample to review and validate.

Once a data map has been created and validated, it will need to be tested to determine if it is "fit for purpose." This testing is crucial for ensuring the map meets the needs for which it was created. End users of the data must be involved in this process. For example, if the purpose of a data map is to exchange lab results between systems, both the sender and receiver must be involved in determining whether or not data are being exchanged accurately. Once a map is used within other software applications, processes external to the map can affect the data, particularly if the software uses the information in an interpretive way, such as the ICD-9-CM codes translated into risk of mortality or severity of illness groupings.

Frequently in this type of mapping project, in the event the random sample did not identify mapping errors, the use of the information will trigger further investigation. It is important for healthcare professionals to closely examine electronic

information. In the event they see something that appears to be inaccurate, they need to question the information. Should a data map introduce questionable information, the investigator may need to track backward, up the stream of information, to determine where the discrepancy began.

Finally, the data map must be maintained and updated. This is an often-overlooked part of data mapping. Many believe that once a map is created it is valid forever. This is not so, since data reporting requirements, standards, databases, and terminology and classification systems change all the time. In fact, some organizations using data maps have staff titled "map managers" employed to manage and update maps. An inadequately maintained map is sometimes worse than no map because it has the potential to transmit the wrong data, subsequently introducing error.[2]

Case Studies

Two case studies appearing in the online appendix to this article highlight the experiences of very different data mapping projects. The first case study is the experience of the Texas Health Care Information Collection Agency, responsible for collecting data for public domain in the state of Texas. It recently underwent a major transition from requiring the UB92 electronic billing standard moving to the new HIPAA electronic standard, referred to in Texas as the Texas 837. The second case study examines AHIMA's experience validating the SNOMED to ICD-9-CM map being created for the Unified Medical Language System.

Case Studies

Mapping the UB92 to the Texas 837

In 1995 the state of Texas passed legislation requiring hospitals to submit inpatient electronic billing data to the state so that quality reports could be produced for consumers. The public domain program was funded in 1997, and the first data collection took place in 1998. The statute called for the data to be submitted according to the "National Uniform Billing Committee (Uniform Hospital Billing Format UB 92) and HCFA-1500 or their successors..." This required that the state move its submission data set from the old format to the new format.

Hospitals with exports pulling data in the UB 92 format (the former format was HCFA-1500 6.0 Version) had to be compliant with this new format. Data beginning with 2004 submissions for the state of Texas require this format for submission to the state agency. Many hospital exports in place retrieved the UB 92 format, with a number of different ways to convert to a new format, including several different mapping options offered by vendors and submission agents. The state submission process requires the standard billed information and some additional information, including race, ethnicity, and additional external cause of injury codes. The added elements create additional complexity when mapping. Many hospitals in Texas used the following options for mapping to the new standard:

- Flat file format mapped to a Texas 837
- UB 92 plus a supplement file mapped to the Texas 837
- 837 institutional (standard bill format) plus a small supplement (including race, ethnicity, and additional E codes) to the Texas 837

Once hospitals had mapped their data from one of the existing formats into the new format, they were required to submit a successful test to the state in the new format. They were required to pass both format edits and some content edits. This individual, hospital-by-hospital process was followed with quality controls by the state agency to ensure that errors had not been introduced into the public domain data file during the transition.

One key area of concern introduced with this new format was the lack of charity care coding in the 837 format. Many hospitals in Texas recommended that charity care accounts be mapped to self-pay, whereas the state preferred that charity care be mapped to unknown. This issue created considerable

debate, as the data for this public domain program is often used to approximate the indigent care burden within the state. There were no clear instructions from the National Uniform Billing Committee to advise how mapping should address the issue.

This example illustrates a common challenge in mapping projects when a piece of information does not match precisely from one format to the next, with some information being lost in the transfer. It is important that the group controlling the specifications understands the end use of the data to determine important decisions such as the payer example above. In Texas's case, the primary use of the format was for billing; however, many states use the billing format for public domain reporting. The primary use of the data does not necessarily have a need to track charity care, given a bill is not distributed for known charity cases. However, in this example, the public use data file clearly has a need to accurately track the number of charity cases in the state.

Once data are in use, it is important that they be monitored after transition. The use of additional ICD-9-CM fields in the 837 created issues with data quality, requiring post-transition monitoring. The old format had 9 diagnoses fields, 6 procedure fields, and 1 E-code field. The new Texas 837 has up to 25 diagnoses, 25 procedures, and up to 10 E codes. This conversion to additional diagnoses, procedures, and E codes has been tracked after implementation to determine if all hospitals are using the additional codes. In the event some hospitals are and some are not, the variability may introduce bias into the quality reports when using this information for risk-adjusted mortality rates.[1]

A quality report on the additional fields of the new format post-implementation indicated that few hospitals in the state were taking advantage of the additional codes in the early quarterly submissions following implementation. Data improved over time with respect to increased fields used. Based on these post-implementation quality checks the state decided to restrict the public use data file to the former format requirements until use of the additional fields reaches a quality threshold. This is projected to begin with the release of the 2005 public use data file.

Although the data mapping project in Texas has been fairly significant for hospitals, the added number of codes to account for the clinical comorbidities is equally important to hospitals. Coding data are often criticized by providers as having limitations for use in quality and patient safety.[2] However, it is also recognized that additional codes may increase the validity and reliability of using public domain data for patient safety and quality reports. Therefore, the conclusions on this data mapping project are as follows:

1. Data mapping of the older reporting formats for public domain data create challenges to providers, vendors, and data submission agents.
2. The new HIPAA transactions format involves new complexities not encountered with the older formats.
3. There is value in using this new format to improve the clinical information available with the increased codes in the new format. This improvement takes time to implement going beyond the introduction of the new format.

SNOMED to ICD-9-CM

In September 2004 AHIMA signed a contract with the National Library of Medicine to validate the SNOMED to ICD-9-CM rules-based, reimbursement use-case map being created by SNOMED for inclusion in the Unified Medical Language System. AHIMA has been involved for several years in the SNOMED Mapping Working Group.

The alpha data map consisting of 500 mapped concepts was received by AHIMA in late March 2005. AHIMA decided to validate the map by reproducing it using the guidelines and heuristics provided by the working group.

Six AHIMA validators were assigned to validate approximately 150 concepts each. The work of each validator overlapped with the work of two other validators. This was done so AHIMA could examine internal inter-rater reliability (IRR) statistics for training and learning purposes since this was a new project.

AHIMA developed a mapping workflow document for the validators from the SNOMED International Alpha Testing documentation. This was done to specify the decision steps when mapping a concept. The workflow document and MS Access tool were tested prior to training, using a small number of concepts.

Training for the map validators consisted of a two-hour Web seminar. The validators were provided with the AHIMA workflow document, the SNOMED Alpha Test documentation, and the references and resources as outlined in the SNOMED Alpha Test documentation. Several of the testing concepts were selected, and maps were determined collaboratively and documented in the tool using the workflow and documentation.

Following the training, the database and the related documents were placed on an internal AHIMA server supporting version control. AHIMA validators checked out the database to perform validation on their assigned concepts, checking it back in when they had completed their work. Database use was reserved using a shared calendar.

Following completion of the validation work, AHIMA proceeded with data analysis. SAS statistical software was chosen for this analysis since it is anticipated that future validation databases may consist of close to 10,000 concepts, each potentially having multiple maps.

The validation data were loaded into SAS, and inter-rater reliability statistics between AHIMA validator pairs were run for ICD-9-CM code assignment, map category, and map rule. Overall, the results were good, usually with an inter-rater reliability greater than 70 percent. Variances in the mapping were most often due to misunderstandings of the instructions provided and the difficulty in determining when some terms are specific or nonspecific for reimbursement or subjectivity in classifying a concept that is not indexed in ICD nor commonly encountered for statistical aggregation or reimbursement reporting. These data will be used to refine AHIMA's training and education for the future validation efforts.

In addition to results detailed in a report delivered to both SNOMED International and NLM, AHIMA drew the main conclusions from its work:

- The specification of the guidelines and heuristics for data mapping must be very detailed to enable understanding and reproducibility.
- HIM professionals are well-suited to data mapping because of their training in health information and related processes.

Notes

1. Iezzoni, L. "Using Administrative Diagnostic Data to Assess the Quality of Hospital Care: The Pitfalls and Potential of ICD-9-CM." *International Journal of Technology Assessment in Health Care* 6, no. 2 (1990): 272-81.
2. Iezzoni, L. *Risk Adjustment for Measuring Healthcare Outcomes* , 2nd ed. Chicago, IL: Health Administration Press, 1997.

Notes

1. Orr, K. "Data Quality and Systems Theory." *Communications of the ACM* 41, no. 2 (1998): 66-71.

2. Iezzoni, L. *Risk Adjustment for Measuring Healthcare Outcomes* , 2d ed. Chicago, IL: Health Administration Press, 1997.
3. Gilder, Richard. Clinical information systems administrator, Texas Health Resources, Inc. Personal communication, November 7, 2005.
4. Dolin, R., et al. *Kaiser Permanente's Convergent Terminology* . San Francisco, CA: AMIA, 2004.

Susan McBride *(smcbride@dfwhc.org) is vice president of the Dallas/Fort Worth Hospital Council Data Initiative, Irving, TX.* **Richard E. Gilder** *is a senior clinical data analyst and a programmer in at Texas Health Resources, Arlington, TX.* **Robert Davis** *is health data standards consultant at the National Association of Health Data Organizations, Salt Lake City, UT.* **Susan Fenton** *is a professional practice manager at AHIMA.*

Article citation:
McBride, Susan; Gilder, Richard; Davis, Robert; Fenton, Susan H.. "Data Mapping" *Journal of AHIMA* 77, no.2 (February 2006): expanded online edition.

Driving the Power of Knowledge