Technical
White Paper

The data curation process

Watson Health Informatics – overview of mapping, standardization, and indexing

Watson Health™

IBM

Contents

About Watson Health™ Informatics and Analytics

The Watson Health Informatics and Analytics team is an essential component of the Watson Health solution delivery process that brings a diverse set of experiences in health care and life science with backgrounds in engineering, computer science, bioinformatics, and clinical informatics. The team is dedicated to unlocking the potential of the data from Watson Health's partners' systems and deriving meaningful knowledge. The team maps data from the source systems into the data model, performs data standardization, imputes and derives novel data points, and develops predictive and prescriptive analytics models. In addition, based on Watson Health's partners' needs, the team augments the IBM® Explorys EPM Application Suite's functionality through focused analytics and customized reports.

Data mapping and extraction

The Watson Health data integration process starts with data discovery and mapping tasks begun shortly after access to the partner source systems has been granted. The Informatics team works directly with healthcare partners to identify the data sources that are needed and how they will be ingested into the IBM® Explorys Platform. An extensive data mapping document is created showing the link between each data element in the source system and where it appears in the data model.

Table	Join	Element	Stored As
Encounter Dx		Patient ID	EHR_Patient_ID
Encounter Dx		Contact Date	DX_Date
Encounter Dx		Primary Dx	DX_Primary_Dx
Encounter Dx		ICD9_10_Code	DX_Code
Encounter Dx		Encounter ID	DX_Encounter_Record ID
Dx Lookup	Encounter Dx	Dx Name	DX_Diagnosis_Name

Table 1: An example of an encounter diagnosis connector

The Watson Health team works closely with healthcare partners to make sure that all relevant workflows and data sources are captured and prioritized by the partner's strategic imperatives. Any necessary changes to the Watson Health connectors are vetted with the partner to verify that accurate data is being ingested.

The data curation process

Watson Health maps each patient record in its system to a single set of ontologies, independent of the source data platform (Electronic Health Record, Billing or Claims System, ADT, etc.). This enables terms to be searched upon for inclusion into cohort designations or measurement filters. Watson Health utilizes a number of established licensed and open-source ontology maps in addition to internally developing libraries where standards do not currently exist.

Watson Health maps patient demographics into a series of standard categories that include age, language, religion, race, and insurance type, as well as geographic area. Watson Health maps are built upon International Organization for Standardization (ISO) and designed for compliance with HIPAA and HITECH standards.

Diagnoses, findings, and procedures are mapped into the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) hierarchy. The Observational Medical Outcomes Partnership (OMOP) provides a set of vocabularies to map ICD-9 and ICD-10 diagnosis codes and CPT procedure codes to SNOMED.[1]

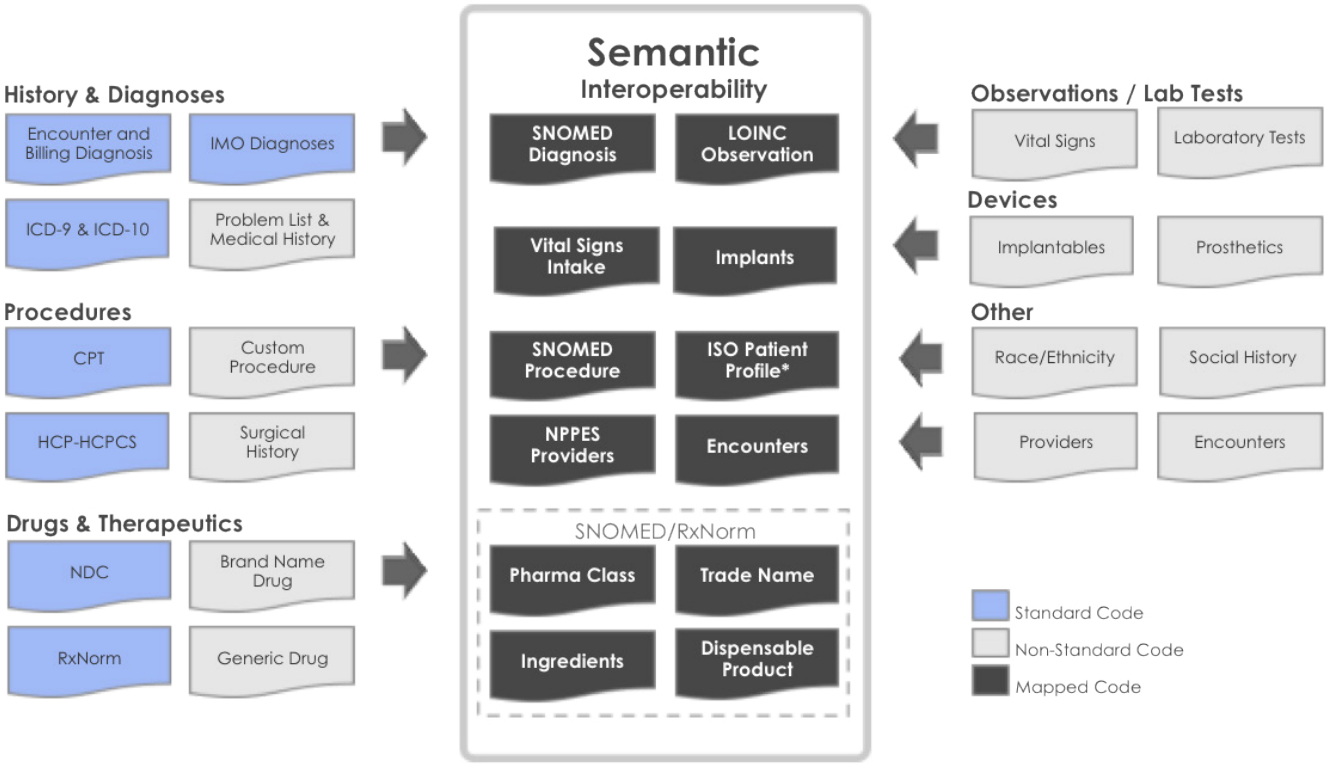

Figure 1: Representation of the data curation process; each source system gets mapped into one standard set of ontologies selected by Watson Health

Although healthcare providers and electronic health record systems vary in their adoption of standards for coding laboratory test observations, Watson Health has developed an extensive open map for translation into the Logical Observation Identifiers Names and Codes (LOINC) hierarchy established by the Regenstrief Institute.[2] The model continues to be expanded to leverage additional mapping and natural language processing techniques. Additionally, observational values are normalized into standard units of measurement as a basis for these curation algorithms. In cases where no logical target exists within the LOINC structure for a particular observation, Watson Health has developed an additional "Observations Auxiliary Map" (OAM) that is used for indexing and searching observations not yet included in the LOINC structure.

Watson Health performs mapping of pharmacy claims, medication orders, and drug administration records within the partner system into pharmaceutical classifications, ingredients, trade names, and dispensable product categorizations. The Watson Health map for these elements is based upon a combination of SNOMED and RxNorm that provides both standardization and drug class hierarchy. RxNorm provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction systems. By providing links between these vocabularies, RxNorm can mediate messages between systems not using the same software and vocabulary. RxNorm also includes the National Drug File – Reference Terminology (NDF-RT) from the Veterans Health Administration. NDF-RT is a terminology used to code clinical drug properties, including mechanism of action, physiologic effect, and therapeutic category. Through the use of RxNorm, Watson Health provides mapping of drugs into a SNOMED standard from a number of originating values including National Drug Code (NDC) and text processed by natural language processing techniques.

The curation workflow

The curation process begins with a structured, semi-structured or unstructured record (i.e., either discretely coded or within a text block) that is then processed by the Watson Health curation engine. Local codes are mapped to standardized codes and semi-structured and unstructured text is standardized when appropriate. Data elements that pass the verification process are automatically mapped and indexed into the IBM Explorys Platform to be made available to applications and platform services. Those that fail the verification process enter a queue that is manually inspected by members of the Watson Health curation team. Each partner's content change management queue is processed by a prioritization model that takes into consideration occurrence counts of unique data elements, as well as an ontology specific logic that provides additional weighting for certain types of less frequent, but clinically relevant, data elements. Figure 2 on the following page provides an overview of this process.

Curation transparency

Watson Health provides a transparent process of how diagnoses, findings, procedures, demographics, drugs, and observations are mapped from their originating values into target ontological codes. Not only is transparency supported within the user applications, but also meta-data mapping is stored along with preserved original values in the underlying data structures.

Record mapping may be observed within Limited Data Set (LDS) downloads in IBM® Explorys EPM: Explore, Patient Summary in IBM® Explorys EPM: Measure, and the data within IBM® Explorys SuperMart. Within the LDS files, Patient Summary, and SuperMart, the originating code/value is displayed along with the curated code/value for each record. Users also get full visibility into records that have not been mapped into standard ontologies and classifications.

Watson Health also provides the ability to inspect how observational values are normalized into standard units of measurement.

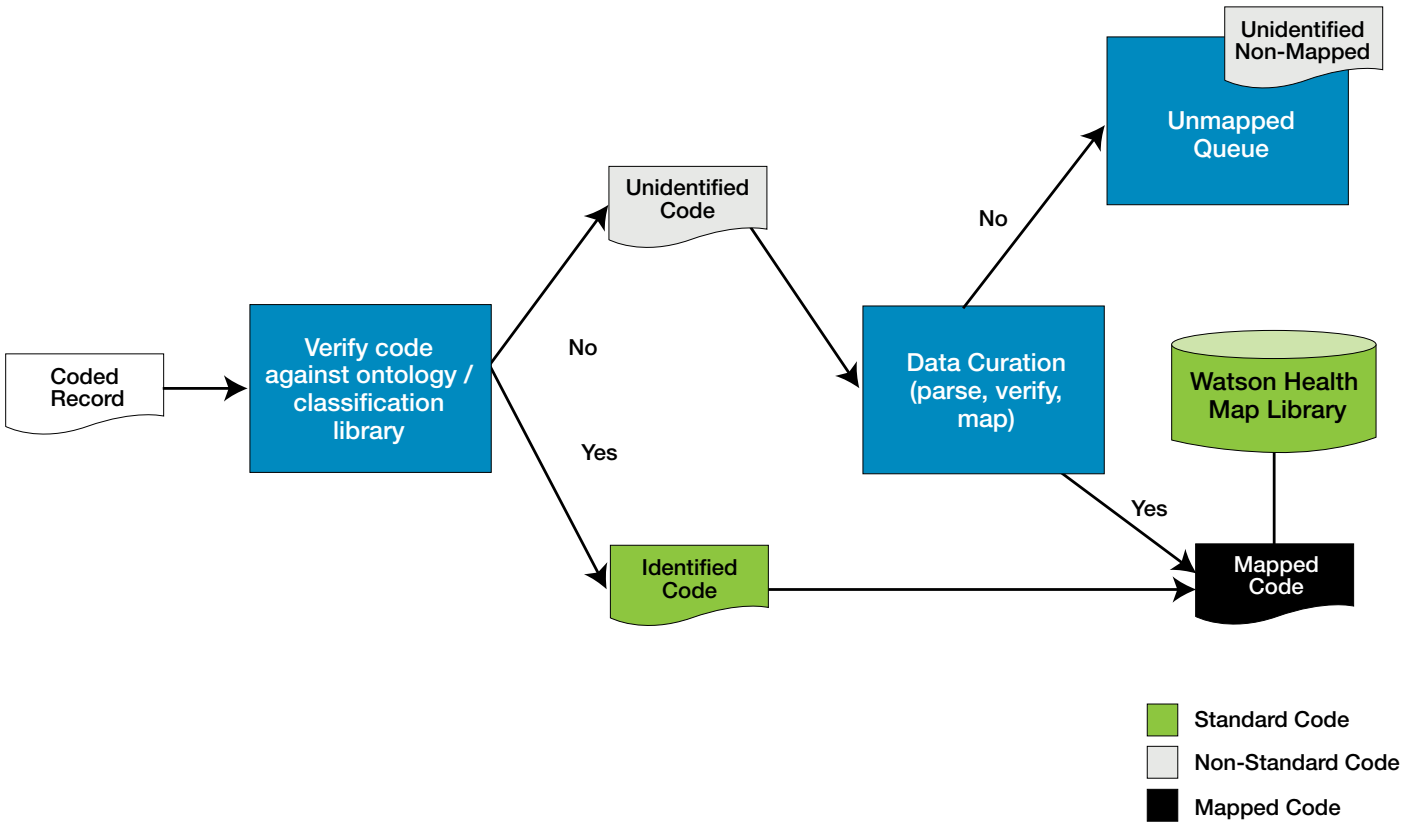

Figure 2: Overview of the data curation workflow process at Watson Health

Data scrubbing

Data is stored in the platform in its native form. The indexing process standardizes the data and maps it to common ontologies. Additional rules are put in place to remove poor quality data. Some of these rules include:

– Data cleansing:

- Records with invalid dates are not indexed (unless a valid date can be imputed)
- Removal of known test patients
- Technically implausible data points as well as physiologically impossible are flagged and not indexed using clinically meaningful rules (e.g., BMI > 100, A1c > 40 percent)

– Standardization:

- Height is standardized to centimeters
- Weight is standardized to kilograms
- Medication prescription meta-data such as “SIG” is standardized into a common format, e.g., “Take 2 tablets twice a day”
- Qualitative lab tests with a semi-structured value (“Strong Positive”) are standardized (“Abnormal”)

Imputation:

- For drug records, if the Prescription Date is null, the listed drug start date is used
- Patient vital status is imputed as alive or deceased based upon date of death or recorded status
- Laboratory observation data lacking a Unit of Measure have an imputed unit (e.g., mg/dL) based on probabilistic algorithms

- BMI results are imputed based on height and weight coming from the nearest encounters (if both do not appear on the same encounter)
- BSA results are imputed based on height and weight coming from the nearest encounters (if both do not appear on the same encounter)

– Advanced imputation

- Pediatric BMI percentiles are imputed based on age, gender, weight and height of patient record
- Pediatric BP percentiles are imputed based on age, gender, blood pressure and height of patient record
- Risk Models such as Framingham Cardiovascular Risk Score, Charlson-Deyo Comorbidity Index, etc. are generated for relevant patients
- Calculation of established scores, including 3M APR-DRG, 3M EAPG, CMS MS-DRG, and CMS HCC
- Calculation of advanced risk scores, including the CMS Yale readmission risk models

Data monitoring and validation process

Watson Health continually monitors the data coming from partners to verify that data is flowing into the system and it is being standardized correctly. Watson Health uses the MapCheck tool to identify how many records are being standardized and which record values are not being standardized along with their counts. Patient data is further reviewed through the validation of measures and registries within the IBM Explorys EPM Application Suite. The Informatics team reviews a sample of patients in each measure and registry to validate that the patients' data is being represented accurately. Similarly, Watson Health healthcare partners can review the quality of the data in the IBM Explorys Platform through Mapcheck, the patient's virtual chart, dataset downloads, and the IBM Explorys SuperMart.

ICD-10 Support

Watson Health supports ICD-10 coding just as it does with ICD-9. If a diagnosis is coded with an ICD-10 code, that code will be available in the IBM Explorys Platform along with the SNOMED translation of that diagnosis. Clients should work with their Account Manager and Project Manager to make sure that the Watson Health team understands the locations of any new ICD-10 coded workflows.

About IBM Watson Health

In April 2015, IBM launched IBM Watson Health and the Watson Health Cloud platform. The new unit will work with doctors, researchers and insurers to help them innovate by surfacing insights from the massive amount of personal health data being created and shared daily. The Watson Health Cloud can mask patient identities and allow for information to be shared and combined with a dynamic and constantly growing aggregated view of clinical, research and social health data.

For more information on IBM Watson Health, visit: ibm.com/watsonhealth.

Footnotes

1. 1 Observational Medical Outcomes Partnership (OMOP); Latest Release Version 4.4 2014-04-11; http://omop.org/Vocabularies

2 Logical Observation Identifiers Names and Codes (LOINC) hierarchy established by the Regenstrief Institute, 6/29/2015, http://www.loinc.org

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
June 2016

HPW03025-USEN-02

Watson Health™