Name:	Sharidan Parr
Affiliation:	Tennessee Valley Healthcare System
Email:	Sharidan.Parr@va.gov
Presentation type(s):	Poster

Machine Learning to Classify Laboratory Tests in a National Electronic Health Record System Database

Sharidan K. Parr, MD, MSCI[1,2], Thomas A. Lasko, MD, PhD[2], Matthew S. Shotwell, PhD[2], Scott L. DuVall, PhD[3,4], Michael E. Matheny MD, MS, MPH[1,2]

[1]Tennessee Valley Healthcare System VA, Nashville, TN; [2]Vanderbilt University, Nashville, TN; [3]VA Salt Lake City Healthcare System, Salt Lake City, UT; [4]University of Utah, Salt Lake City, UT

Abstract:

Electronic health record (EHR) systems are a rich source of data accumulated through routine clinical care. This data can be leveraged for analytics, research, and quality improvement. However, idiosyncratic, ambiguous local data representations can hinder data aggregation across healthcare institutions. Standards, such as Logical Observation Identifiers Names and Codes (LOINC®), are critical for transforming data into common data models (CDM), such as the Observational Medical Outcomes Partnership (OMOP) CDM. Using national Department of Veterans Affairs (VA) laboratory data, we developed a completely automated algorithm for classifying laboratory test LOINC codes. We present an automated laboratory classification algorithm that could reduce the manual effort required to transform data into the OMOP CDM.

Background

Secondary use of EHR data for analytics, research, and quality improvement is prevalent. Data aggregation and centralization are also increasingly common. When multiple data sources are integrated, the need for data cleaning and transformation increases. This process is often manual and may not be scalable. Standards, such as LOINC are critical for integrating data into CDMs, but are inconsistently used. Additionally, even when LOINC codes are used, in many cases they are incorrectly mapped. We hypothesized that machine learning techniques could be leveraged to automate classification of laboratory test LOINC codes in data source with existing partial LOINC mappings.

Objective

To develop a scalable, automated algorithm for classifying laboratory test LOINC codes in a in a large, heterogeneous, aggregate data source.

Methods

Data sources included 1) the Department of Veterans Affairs (VA) Corporate Data Warehouse (CDW), from January 2002 through December 2016, across 130 sites, limiting to 150 most common laboratory tests per site, 2) publically-available LOINC controlled vocabulary. Data were aggregated by test name identifier, specimen type identifier, units, and LOINC code, with numeric test results summarized as mean, median and percentiles. Unlabeled data (i.e. missing LOINC code), were included in the aggregation roll-up. Using an automated text feature normalization and mapping algorithm, local test names and specimen types were tokenized, with punctuation and stop-words removed. Then, using an automated string-distance comparison algorithm, test name tokens were mapped to LOINC Long Name tokens, and specimen name tokens were mapped to LOINC System tokens. The model features are shown in Table 1.

The labeled dataset was partitioned such that LOINC codes with only 1 occurrence by test volume or comprising <5 rows in the aggregate dataset were placed in separate dataset for reclassification. We used 5-fold cross-validation, with 80/20 split by sites. We implemented a random forest multi-class classifier and a one-versus-rest ensemble of binary random forest classifiers. To evaluate classification accuracy, we conducted a manual single reviewer validation comparing source data elements (test Name, topography, units, and LOINC (when present)) with the fields from the model-assigned LOINC code (Component, Property, Time Aspect, System, Scale, and Units). For unlabeled data, we reviewed 200 randomly-selected aggregate data rows. In the labeled data, we reviewed 100 randomly-selected aggregate data rows where predicted LOINC matched original source data LOINC code (concordant), and 100 randomly-selected aggregate data rows where predicted LOINC did not match original source

data LOINC code (discordant).

Table 1: Model Features	
Text	**Numeric**
Test Name mapped to LOINC Long Name (JW)	Test result 5th percentile
Test Name mapped to LOINC Long Name (LV)	Test result 25th percentile
Topography mapped to LOINC System (JW)	Test result median
Topography mapped to LOINC System (LV)	Test result mean
Predicted LOINC Component (JW)	Test result 75th percentile
Component Match Distance (JW)	Test result 95th percentile
Predicted LOINC Component (LV)	Test result minimum
Component Match Distance (LV)	Test result maximum
Predicted LOINC System (JW)	Normalized test frequency*
System Match Distance (JW)	
Predicted LOINC System (LV)	
System Match Distance (LV)	
Units	

JW, Jaro-Winkler; LV, Levenshtein; *Normalized test frequency calculation = Test frequency/Total tests per site

Results

Test results numbered 6.6 billion, ranging 2.5 – 183 million per site. LOINC codes were missing for 7% of tests by volume, and 31% by row count in the aggregate dataset. Among the top 150 laboratory tests, there were 2,223 distinct LOINC codes in the source data. Tables 2 and 3 detail the performance of the model on unlabeled and labeled data.

Table 2: Evaluating LOINC Codes Assigned to Unlabeled Data		
Predicted LOINC Code	**Aggregate Data Rows (N=200)**	**Associated Tests (N=3,364,245)**
Correct	182 (91%)	3,097,999 (92.1%)
Incorrect	10 (5%)	263,744 (7.8%)
Insufficient Information to Determine	8 (4%)	2,502 (0.1%)

Table 3: Model Performance within Labeled Data		
Concordant Original and Predicted LOINC	**Aggregate Data Rows (N=100)**	**Associated Tests (N=14,910,333)**
Correct	92	14,895,799 (99.9%)
Incorrect	2	316 (<0.1%)
Insufficient Information to Determine	6	14,218 (0.1%)
Discordant Original and Predicted LOINC	Aggregate Data Rows (N=100)	Associated Tests (N=1,515,472)
Predicted LOINC Correct	45	251,498 (16.6%)
Original LOINC Correct	19	85,113 (5.6%)
Both Correct (Equivalent LOINC Codes)	23	1,176,470 (77.6%)
Both Incorrect	9	1,805 (0.1%)
Insufficient Information to Determine	4	585 (<0.1%)

Conclusions

Using a completely automated process, we are able to assign LOINC codes to unlabeled data with >90% accuracy. In VA CDW data, the prevalence of incorrect existing LOINC codes in the source data is non-trivial. This model demonstrates utility in LOINC code reclassification, suggesting that the algorithm could be improved by including an iterative training phase. Our current model incorporates features created from raw source data aggregation, and as such, could be implemented as an initial step in the OMOP transformation pipeline to derive the LOINC CONCEPT_ID. In addition, this model could be used to compare laboratory MEASUREMENT in multiple OMOP instances to evaluate possible areas of mapping discordance, as the proposed model uses both value and string-based assessments. In summary, this scalable, automated algorithm may improve data quality and interoperability, while substantially reducing the manual effort currently needed to accurately map laboratory data.