Using RxNorm to Extract Medication Data from Electronic Health Records in the Rochester Epidemiology Project

Jyotishman Pathak[1], Sean P. Murphy[1], Brian N. Willaert[1], Hilal M. Kremers[1], Christopher G. Chute[1], Barbara P. Yawn[2], Walter A. Rocca[1]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[2]Department of Research, Olmsted Medical Center, Rochester, MN, USA

Abstract. RxNorm is a standardized terminology for clinical drugs developed by the U.S. National Library of Medicine (NLM) in order to facilitate exchange and public availability of medication information. In this study, we evaluate the applicability of RxNorm for representation of medication data from two institutions that are part of the Rochester Epidemiology Project (REP). We detail the researchers' analysis objectives and subsequent requirements for a drug terminology that is comprehensive and easily accessible. We also explore the completeness of mappings between RxNorm and a commercial drug database, Multum, for this sample of REP medications.

1 Introduction

In this study, we develop approaches for structured and unstructured querying of medication data from the out-patient prescription records of two institutions that are part of the Rochester Epidemiology Project (REP [1]). We evaluate RxNorm [2] for standardized data representation from the following aspects: (1) Coverage: What coverage of terms and concepts does RxNorm provide for the REP medication data? (2) Mapping consistency: Are the mappings between RxNorm and external drug databases, such as Multum, accurate and consistent?

2 Background: The Rochester Epidemiology Project

The Rochester Epidemiology Project (REP [3]) is a collaborative effort between several healthcare providers in Olmsted County, MN. It is a medical records-linkage system encompassing the care delivered to all residents of Olmsted County including the Mayo Clinic and Olmsted Medical Center. For this study, we investigate the utility of RxNorm for standardization of REP medication data to facilitate data exchange and interoperability. In particular, based on Mayo's cTAKES platform [4] and RxNorm, we develop natural language processing (NLP) techniques to standardize medication data extracted from the EHR systems of the two REP providers.

3 Materials

The following materials were used in this study:

- RxNorm January 3, 2011 and November 17, 2008 Full Update release data were used.
- Medication history, as part of out-patient clinical notes (out-patient prescriptions referred to as Orders97), for 212,974 unique individuals was retrieved from the Mayo Clinic's EHR system between January, 2004 and October, 2010. The dataset contained more than 180,000 unique mentions of medications and was retrieved by processing approximately 5 million rows of data from out-patient prescriptions by the cTAKES NLP techniques.
- Out-patient prescription data for 105,151 unique individuals was retrieved from the Olmsted Medical Center EHR system (based on Microsoft® SQL Server), between August, 2002 and November, 2010. The dataset contained 1375 unique mention of medications and corresponding Multum codes extracted via SQL queries.

4 Methods

4.1 Extracting Medication Data via Structured Querying

The Olmsted Medical Center prescription data was extracted from the EHR system (InteGreat IC-Chart) through a scheduled job that queries the data directly from the EHR source database tables. The prescription data, ranging between August, 2002 and November, 2010, was retrieved for 105,151 unique individuals from the primary prescription table, which also included Multum drug codes.

For representing this data using RxNorm, we mapped the Multum codes to RxNorm codes (RxCUIs). Specifically, we loaded the RxNorm January 3, 2011 Full Update release data in a MySQL database, and queried the RXNCONSO table to retrieve the mappings. For example, Hydrogen Peroxide 300 MG/ML Topical Solution with a Multum code=16282 was mapped to RxCUI=91348.

4.2 Extracting Medication Data via Natural Language Processing

To extract drug mentions from Mayo's Orders97 clinical notes data, we first created a dictionary comprising more than 265,000 terms, identified uniquely via a RxCUI code, to assist in the look-up process using RxNorm. This dictionary also comprised of RxNorm codes that were deemed as "obsolete" by the NLM since the patient corpora included more than a decade old drug information. Furthermore, we supplemented the dictionary with 1717 additional terms primarily comprising drug misspellings and abbreviations that were not available in RxNorm. We used open-source Apache Lucene for implementing the dictionary.

For the extraction process, we used Mayo's open-source cTAKES NLP toolkit [4]. In particular, only those terms from the Orders97 data that were composed of seven tokens or less considered for a match in the dictionary. (A token is considered any text surrounded by blanks with the exception of punctuation characters, which are considered as tokens as well.) Therefore, if a match was discovered for the first term in a drug mention, then all permutations of up to the six remaining tokens were considered for a match. Note that limiting the number of permutations is necessary to minimize the computation time necessary to handle larger drug names. For instance, a ten token drug name would take 362,880 permutation lookup tasks, whereas a seven token name would only take 720 lookups. Each term in the dictionary was run through a tokenizer process that separates the terms into distinct tokens to automatically discover drug mentions and relevant attributes, such as dosage, route, form, frequency, duration, and drug change status.

5 Results

5.1 Using RxNorm for Olmsted Medical Center Data

Since the medication information for Olmsted Medical Center was retrieved using straightforward SQL queries, our focus was primarily on evaluating the mapping between this data, represented using Multum, to RxNorm codes. All of the 1375 unique mention of medications and corresponding Multum codes in the dataset were mapped to RxNorm by querying the RXNCONSO table from the RxNorm release. In order to check accuracy, we randomly selected and manually reviewed (by an experienced pharmacist) the 500 drug mentions of top 50 frequently administered drugs, specifically focusing on medication names and their descriptions. We found that the entire set of mappings for the 500 drug mentions were accurate, thereby validating that the curation of mappings, at least between RxNorm and Multum, done by the RxNorm curators is of high quality and consistency.

RxNorm Release	Unique terms identified by cTAKES NLP medication pipeline	Unique RxCUIs identified	Unique terms with RxCUIs	Unique terms without RxCUIs
Nov. 2008	181,722	7,908	114,653	67,069
Jan. 2011	181,727	8,058	135,988	45,739

Table 1. RxNorm results for Orders97 dataset processing

5.2 Using RxNorm for Mayo Clinic Data

As mentioned above, for Mayo's Orders97 data, we extracted medication information for 212,974 unique individuals using the cTAKES NLP process, and represented it using both the November 2008 and January 2011 releases of RxNorm (both versions were used to facilitate a comparative analysis). In particular, we analyzed 4,964,022 rows representing medication mentions for 212,974 different individuals in the Orders97 data (there were one or more mentions on each row). 181,722 and 181,727 unique terms for 2008 and 2011 RxNorm releases, respectively, were identified as valid medication "related" mentions by the cTAKES pipeline. From this, 7,908 and 8,058 unique RxNorm concept codes were identified for the 2008 and 2011 RxNorm releases, respectively.

DTaP/IPV/Hib Vaccine	Inactivated polio vaccine
Haemophilus Influenzae Type b (Hib) Vaccine	MMRV (measles, mumps, rubella varicella) vaccine
H1N1 swine flu vaccine	Typhoid vaccine
Meningococcal Vaccine	Hepatitis A vaccine

Table 2. Sample list of vaccines

6 Discussion

RxNorm with its vastness provides a comprehensive coverage of drugs and medications. However, in this study, we found that coverage for vaccines require further improvement. Table 2 shows a sample list of vaccines for which no corresponding RxNorm codes were discovered while processing the Orders97 data (this list was manually identified from Orders97). Specifically, we identified 103 distinct vaccine related terms in our Orders97 dataset, of which 35 had recognizable terms in RxNorm, although variations of those terms were missing from RxNorm. For example, the text span "H1NI" produced no hits, but the text span "Influenza A (H1N1) Vaccine 2009" had a corresponding RxCUI. Furthermore, terms such as "influenza H1N1 vaccine" which were not present in the 2008 release of RxNorm, but present in the updated 2011 release. This leads us to hypothesize that coverage for vaccines will potentially improve in future RxNorm releases. In addition to above, our investigation also highlighted that drug and medication terms commonly occurring in the Orders97 dataset which were represented using abbreviations, hyphenations, or aliases (e.g., "vit. b3" to represent "Vitamin B3") did not have a corresponding RxNorm code.

RxNorm also includes terms such as air and water, and forms, such as cream, powder and oil, and ingredient mentions that are associated with minerals or common elements, such as calcium, glucose and magnesium. Such terms tend to create noise when extracting drug content from unstructured text. Therefore, we compiled a list of such terms to prevent the cTAKES NLP pipeline from propagating the annotated named entities in this group, which lead to significant performance improvement in named entity detection.

Limitations and Future Work. In this study, we limited our investigation to analysis of outpatient medication data from only 2 institutions. In future, we plan to incorporate in-patient medication data and other institutions as well. Furthermore, an important requirement for REP is classification and categorization of RxNorm coded medication data using standardized drug classifications, such as NDF-RT. To this end, we are currently incorporating recently released NLM's NDF-RT API [5] within the cTAKES NLP pipeline for mapping and classifying RxNorm coded data from REP to NDF-RT drug classes.

Acknowledgments

Research funded in part by the Rochester Epidemiology Project (AG034676-45) and Mayo Clinic Early Career Award to the first author.

References

1. St. Sauver, J., et al., *Use of a Medical Records Linkage System to Enumerate a Dynamic Population Over Time.* Am. J. Epidemiol 2011.
2. Liu, S., et al., *RxNorm: Prescription for Electronic Drug Information Exchange.* IT Professional, 2005. **7**(5): p. 17-23.
3. Melton, L., *History of the Rochester Epidemiology Project.* Mayo Clinic Proceedings, 1996. **71**(3): p. 266-274.
4. Savova, G., et al., *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.* JAMIA, 2010. **17**(5): p. 507-513.
5. *National Library of Medicine NDF-RT Web Services API. Last updated on: February 16th, 2011.* http://rxnav.nlm.nih.gov/NdfrtAPI.html.