An Exploration and Evaluation of Automated Approaches to Cross-terminology Mapping

Michael Steigman
University of Texas School of Biomedical Informatics
Masters Practicum - December 2016
Advisor – Dean Sittig, Ph.D.

Introduction

Over the past several years, researchers at the Brigham and Women's Hospital in Boston and the University of Texas Health Science Center in Houston have successfully developed automated methods to infer patient problems from structured clinical data.[1] These methods hold great promise for improving the quality of healthcare in several ways. They can help clinicians maintain a more accurate problem list which can in turn drive more effective automated clinical decision support. They can also help in the identification of research cohorts.[2]

This research has been validated at external sites.[3] In the process, multiple knowledge bases containing medications and problems have been created based on different patient populations using different terminologies. The authors hypothesize that these knowledge bases would provide more complete coverage of conditions if they could be used in an ensemble fashion.

To combine the knowledge bases, the coded clinical data would need to be mapped to common terminologies. Some of the data sources included mappings to common terminologies while other did not. Initial efforts to map the terminologies involved the use of manual and partially automated methods. For this project, I chose to explore and evaluate the use of completely automated mapping methods to combine the knowledge bases.

Background and History

This paper describes several techniques that fall under the category of cross-terminology mapping - the process of mapping from raw biomedical text or one term in a standardized terminological or coding system to another term or set of terms. These mapping processes are usually the first part of a larger project to mine clinical data and involve transforming electronic health record (EHR) data such as medications, problems, observations, and lab results.

There are several reasons why coded clinical data may need to be transformed. First, there are some areas of the EHR where, despite much effort, no standard terminology has garnered enough acceptance to facilitate the easy study or exchange of data.[4] There are also many situations when researchers want to generalize from highly granular, "pre-coordinated" terms that can vary in use within or between systems, resulting in granularity mismatches.[5] It is also possible a researcher might want to normalize "post-coordinated" terms for similar reasons. Whether dealing with data

within a single EHR or across multiple EHRs, some manipulation is almost always required before meaningful analysis can proceed.

It might seem somewhat counterintuitive given the above but one of the enduring goals of the informatics discipline known as Knowledge Representation is to facilitate health data interoperability. Efforts in this area continue and standardized terminologies continue to evolve – consider the transition from ICD9 to ICD10 this past year in the United States - which also contributes to the complexity in mining clinical data.

This discipline has its roots in some of the earliest terminologies. The original version of SNOMED was developed by pathologists as a way to describe findings in the 1960s[6]. Although SNOMED evolved over subsequent decades and additional terminologies surfaced, research in this area received a big boost in the early 1990s when the National Library of Medicine (NLM) introduced the Unified Medical Language System (UMLS).[7] The UMLS is a combination of interlinked biomedical concepts from the standardized terminologies along with a semantic "type" and a lexicon to help identify variations of known concepts.

In 2001, National Library of Medicine released MetaMap with the goal of improving the information retrieval results of bibliographic information in tools such as MEDLINE.[8] MetaMap provided a way to map biomedical text or concepts to other concepts in the UMLS Metathesaurus, which had grown to include mappings to and from many of the commonly used terminologies in the field.

While MetaMap was not the first tool of its kind, it has eclipsed its competitors and outgrown its original purpose to become a widely-used tool for mapping biomedical text to UMLS concepts. MetaMap has applications in the areas of data mining and clinical decision support for example. I will describe how I utilized this tool below.

By the 2000s, standardized terminologies had been adopted within many parts of the EHR. However, because of the variation and overlap in content coverage[9] and usage as well as the more nuanced research needs mentioned above, many healthcare data interoperability challenges remain. Proposals have emerged[10,11] and informatics methods[12] and tools have evolved and continue to evolve to address these challenges but there are still many open questions and thus, this is likely to be an active area of research for some time to come.

Methods

It should be noted that this project was approached de novo. Several of the knowledge bases included mappings to one or more terminologies and initial efforts to combine the knowledge bases included some attempts at automated and manual mappings. However, as I was interested in examining the tools available for fully automated mapping, I chose to start with the raw medication and problem texts from the five

different sites – BCBSTX, BWH, CS UT, REP UT, UTH. This approach had its advantages and disadvantages, which I will discuss.

Medication Mapping and Normalization

The first step in normalizing the medication texts in each knowledge base was to map them to a common terminology. Mapping each text in the knowledge bases to an identifier within a terminology was not enough, however, as we would certainly encounter many different texts that represent the same drug but with a different dose form or strength.

Since one of the desired outcomes was the evening out of granularity differences in the different knowledge bases, another layer of abstraction was going to be required. There was also the consideration that it might be necessary in some cases to perform a "roll-up", i.e., to look at types of a drug grouped together to increase the overlap between the knowledge bases. If I mapped the normalized medications to a drug class hierarchy, I could potentially achieve both ends. In my literature searches, I discovered another group of researchers who had developed a novel approach[13] to this challenge and so I based my approach on their work.

For terminologies, I considered SNOMED-CT and RxNorm. SNOMED-CT is characterized by a poly-hierarchy (e.g. aspirin would be classified as both an analgesic and a blood thinner). This type of representation is certainly necessary in some cases and was chosen by one research group for a similar integration effort.[5] However, RxNorm has become the de facto standard for medications over the past decade.[14] As such, it was a logical choice for a common medication terminology.

Next, I had to choose a classification system (i.e., drug class hierarchy), which led to an examination of the drug classification hierarchies linked to RxNorm, including SNOMED-CT, ATC and the NDF-RT classes. Although two of the sources for the knowledge bases utilized First Databank's (FDB) Extended Therapeutic Classes (ETC) drug hierarchy, FDB is proprietary and was not available across all sources.
Thus, for the initial mapping of medication names, I chose a strategy based on Pathak, et al,[15] which uses the RxNorm and the VA legacy drug classes in NDF-RT.

NDF-RT contains multiple multi-axial classification hierarchies; a general drug hierarchy that is known as the VA Legacy class as well as a handful of hierarchies based on drugs characteristics such as mechanism of action, target disease, etc. The integration of NDF-RT as a source vocabulary in RxNorm in 2010 provided clinicians and researchers with some new tools[16] and has engendered a new approach to data integration.[17,13,18,15]

As with previous efforts on this project, I began the mapping process by utilizing the natural language processing (NLP) tool called MedEx[19] to analyze the medication strings from each location. The initial MedEx pass resulted in the numbers displayed in Results and Discussion. The last two columns show the number of meds extracted

by MedEx and the subset that could be mapped to a VA class, using the algorithm described next.

Once the database was populated with all the unique meds and their respective RxCUIs, I developed a Python script to fetch the drug classes. Although Pathak, et al utilized complex query and mapping algorithms and the UMLS Metathesaurus data tables, I have attempted to simplify the mapping process with the help of RxNorm's API. Here is the algorithm:

1. Retrieve an XML drug classification for RxCUI using the RxNorm API. The specific classification does not matter as all classifications return either ingredient level and/or "VA Product" level drug nodes based on the RxCUI.

    a. If the RxNorm classification contains a Node with a path/value of Node/nodeAttr/attrName=level and an attrValue of "VA Product", this node represents a classified VA drug. I can obtain the legacy VA class from the RxNorm API by using the NUI (NDF-RT Unique Identifier) in the nodeId.
    b. If the RxNorm classification does not contain a "VA Product" level node - ingredient level RxCUIs (i.e., generics) do not have one - I must find an ingredient level identifier NUI in the classification and query Bioportal's SPARQL endpoint to find a "VA Product" level subclass of that generic NUI. An example would be Cetirizine, which appears in our medication list several times. The generic preparation falls under the "Pharmaceutical Preparations/Drug Products by Generic Ingredient Combinations/C Preparations" hierarchy in the NDF-RT hierarchy. NDF-RT populates these generic preparations with subclasses of type "VA Product". In the case of Cetirizine, one of the "VA Product" subclasses has the label "CETIRIZINE HCL 5MG TAB, CHEW". The ID for this product, a drug "instance" of the generic Cetirizine preparation, can be used to retrieve a VA class for Cetirizine.
    c. If the RxNorm classification returns neither "VA Product" level node nor an ingredient level node, I contact the BioPortal's Annotator service with the medication name and search the NDF-RT ontology for a match, taking the first. Two examples of medication names that were assigned RxCUIs from MedEx but which yielded no NDF-RT mappings were "INSULIN SYRG MIS 1ML/27G" (RxCUI 763000) and "Alendronate Sodium" (RxCUI 203152).

2. Using the NUI obtained above retrieve the VA class from the RxNorm API.

Problem Mapping and Normalization

The first choice for problem mapping was, again, terminology. As ICD9/10CM is a shallow hierarchy when compared with SNOMED-CT - max depth of 6 vs 28 – I decided to go with the shallow hierarchy initially. The simpler hierarchy resulted in an easier choice for a target when it came to picking a roll-up level to reduce granularity.

The roll-up technique used was, at the time (2014), a novel aspect of the project. It involved leveraging BioPortal's[20] SPARQL interface to roll identifiers up to a common ancestor and thus reduce the granularity of the mapped problem and increase overlap between data sets. No other examples of this approach were found in the literature, perhaps because the tooling was so new. The details of this approach will be discussed below.

Regarding granularity, I experimented with MetaMap input options looking for ways to improve the overlap in the initial mapping from text to concept as well. I didn't find much in the literature to guide my efforts, especially when compared to normalizing the medication texts. However, after re-reading the final ensemble report for ideas, the approach I settled on was:

1. Pre-process problems, generating unique list, removing items that contain characters that upset MetaMap and tweaking problems that should map but didn't (i.e., "Parkinsons" did not generate a mapping candidate but "Parkinson's Disease" did)
2. Using MetaMap batch, restrict to five vocabulary sources – CHV, ICD9CM, ICD10CM, MTH, MTHICD9 - and 16 semantic types to reduce the number of mappings and candidates for each problem.
3. Generate problem IDs based on mappings per following:

    a. Concatenate CUIs for problems that map to multiple CUIs.
    b. Use the raw problem text for unmapped problems.
    c. Use the single CUI for non-ICD9/10CM mapping candidates.
    d. For all problems with a single ICD9/10CM mapping candidate (see numbers in Figure 2 below) deeper than level 3 in the ICD9/10CM hierarchy, use BioPortal's ICD9/10CM ontologies and SPARQL to roll up to the ancestor CUI at level 3. (Level 3 was just an initial guess based on browsing the ICD hierarchies; many of the codes at the first and second level seem to general to be useful. This will likely be adjusted after further analysis.)

I ran into some issues with MetaMap's tokenizer during this part of the work. After some debugging and email communication with the MetaMap team, I decided to lower-case all problems to prevent MetaMap from tokenizing the input (e.g., "H. Pylori") into multiple phrases. Multiple phrases had necessitated the handling more special cases, making my MetaMap output parsing code more complex.

Results and Discussion

Medication Mappings

Source	Meds	RxCUIs	VA Class Mappings
BCBSTX	281577	7329	5477
BWH	9819	1411	1143
CS UT	2945	2594	2256
REP UT	572	39	32
UTH	25058	2411	1657

Figure 1. Number of unique meds, meds mapped to RxCUIs and RxCUIs mapped to VA Classes

The VA Class Mappings in Figure 1 represent higher-level groupings retrieved for the RxCUIs and thus is a subset. Since the higher-level groupings can facilitate comparisons across systems, the objective was to achieve or approach a 1/1 ratio of VA Class Mappings to RxCUIs.

Problem Mappings

Source	Problems	ICD10CM	Single CUI
UTH	563	536	322
CS UT	1576	1557	901
REP UT	369	364	233
BWH	205	198	136
BCBSTX	1633	1626	1128

Figure 2. Number of unique problems, problems mapped to one or more ICD10CM CUIs and subset of those mapped which included a single CUI only.

The single-CUI ICD9/10CM mapping candidates in Figure 2 represent opportunities to further reduce granularity; with a single term, it is possible to retrieve a less-granular parent from the hierarchy. Much the same as with medications, the objective was to maximize these types of mappings. This was done by experimenting with MetaMap's configuration settings - the vocabulary and semantic type options – as described above.

The input to MetaMap was 3,414 unique problems, lower-cased and sorted with dupes removed using the `uniq`, `sort` and `tr` Linux commands. 1,244 of the 2,720 single-CUI mappings were "granularized", i.e. the Python script retrieved a less granular concept from the ICD10CM hierarchy. 1,320 were left "as is"; there was either a single mapping to a ICD10CM concept that was already at level 3 in the hierarchy - the level I had targeted - or was mapped to a concept in another vocabulary that didn't support the same type of hierarchical treatment. The rest mapped to multiple UMLS concepts.

Using materialized database views which layered the new mappings on top of the original texts, I generated overlap numbers. The unique overlapping pairs between the five sources after processing the mappings are shown in Figure 3. Joining all the materialized views together resulted in 47 common pairs.

Source Pair	**Overlap Count**
BWH <-> UTH	984
BWH <-> BCBSTX	461
BWH <-> REP UT	120
BWH <-> CS UT	296
CS UT <-> BCBSTX	650
CS UT <-> UTH	722
CS UT <-> REP UT	473
UTH <-> REP UT	222
UTH <-> BCBSTX	1171
REP UT <-> BCBSTX	156

Figure 3. Overlap between each of the source pairs.

Although I settled on ICD10CM for the "granularizing" vocabulary, I generalized the Python module to allow for the use any of the ontologies on BioPortal. I also refined the database model to allow for multiple versions of the mappings. Those two factors will allow me to try additional combinations of vocabularies and semantic types on the MetaMap side and iteratively improve the mappings.

Conclusions, Limitations and Future Work

The automated methods used in this project yielded promising results. The number of overlapping pairs between all sources was approximately the same as in previous efforts and there is room for further improvement. For example:

- Whereas the pairs in the knowledge bases had been culled from a much larger data set based on the application of the statistical procedure chi-square, the current effort attempted to normalize **all** the pairs from the raw data. The subset of significant pairs found in the knowledge bases would likely result in better coverage, consistency and overlap as some of the less common, custom or incorrect terms had been removed.
- NLP-based tools have evolved further in the time since the medications were initially processed. RxNAV's web API now supports search and approximate search. Initial explorations into using this new tool were promising. This approach could eliminate the MedEx step and yield better mappings.
- Although more research has been carried out using the VA classes from NDF-RT as described above, it is possible that a mapping of medications to SNOMED-CT's poly-hierarchical classification may yield better results or provide better data for the statistical methods being used.

- On the problem side, one further step might be to attempt to reduce granularity using UMLS-CORE, a problem list subset of SNOMED-CT designed for this purpose.[21]

Also, because I was exploring completely automated methods, I did not address the medications MedEx was unable to map to RxCUIs. In most cases, a complete mapping or nearly complete mapping is desired. Since no NLP tool can deliver an F-measure of 100, the automated NLP step is usually the first of several. What usually follows is a semi-automated or manual process of mapping the medications the NLP tool could not map.

While it is possible that the combination of improvements discussed above – specifically, RxNAV's new API with only the subset of significant pairs - would result in a much more complete mapping, a thorough approach would include an analysis of any unmapped medications.

Finally, some interesting new approaches have surfaced over the past year. In Europe, one research group implemented a BioPortal-based terminology service to provide the types of mappings I attempted.[22] This service likely includes a more sophisticated version of the granularity-reducing approach I developed based on BioPortal's APIs.

In closing, I would like to thank Dr. Sittig and Dr. Wright for their guidance, support and patience during this project and Dr. Cui Tao for her help with the BioPortal API.

References

1 Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. *J Biomed Inform* 2010;**43**:891–901. doi:10.1016/j.jbi.2010.09.009

2 Wright A, Pang J, Feblowitz JC, *et al.* A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. J. Am. Med. Informatics Assoc. 2011;**18**:859–67. doi:10.1136/amiajnl-2011-000121

3 Wright A, McCoy A, Henkin S, *et al.* Validation of an association rule mining-based method to infer associations between medications and problems. *Appl Clin Inform* 2013;**4**:100–9. doi:10.4338/ACI-2012-12-RA-0051

4 Hammond WE, Cimino JJ. Standards in Biomedical Informatics. *Biomed Informatics Comput Appl Heal Care Biomed* 2006;:47. doi:10.1007/0-387-36278-9_7

5 Saitwal H, Qing D, Jones S, *et al.* Cross-terminology mapping challenges: A demonstration using medication terminological systems. *J Biomed Inform* 2012;**45**:613–25. doi:10.1016/j.jbi.2012.06.005

6 Cornet R, de Keizer N. Forty years of SNOMED: a literature review. *BMC Med Inform Decis Mak* 2008;**8**:S2. doi:10.1186/1472-6947-8-S1-S2

7 Lindberg C. The Unified Medical Language System (UMLS) of the National

Library of Medicine. *J Am Med Rec Assoc* 1990;**61**:40–2.http://www.ncbi.nlm.nih.gov/pubmed/10104531 (accessed 8 Dec2016).
8 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001;:17–21.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243666&tool=pmcentrez&rendertype=abstract (accessed 18 Jan2016).
9 Chute CG, Cohn SP, Campbell KE, *et al.* The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. *J Am Med Informatics Assoc* 1996;**3**:224–33. doi:10.1136/jamia.1996.96310636
10 Hamm RA, Knoop SE, Schwarz P, *et al.* Harmonizing clinical terminologies: driving interoperability in healthcare. *Stud Health Technol Inform* 2007;**129**:660–3.http://www.ncbi.nlm.nih.gov/pubmed/17911799 (accessed 27 Aug2014).
11 Halper M, Gu H, Perl Y, *et al.* Abstraction networks for terminologies: Supporting management of 'big knowledge'. *Artif Intell Med* 2015;**64**:1–16. doi:10.1016/j.artmed.2015.03.005
12 Richesson RL. An informatics framework for the standardized collection and analysis of medication data in networked research. J. Biomed. Inform. 2014;**52**:4–10. doi:10.1016/j.jbi.2014.01.002
13 Pathak J, Murphy SP, Willaert BN, *et al.* Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: experiences from the Rochester Epidemiology Project. *AMIA Annu Symp Proc* 2011;**2011**:1089–98.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243205&tool=pmcentrez&rendertype=abstract (accessed 12 Aug2014).
14 Zhou L, Plasek JM, Mahoney LM, *et al.* Mapping Partners Master Drug Dictionary to RxNorm using an NLP-based approach. *J Biomed Inform* 2012;**45**:626–33. doi:10.1016/j.jbi.2011.11.006
15 Palchuk MB, Klumpenaar M, Jatkar T, *et al.* Enabling Hierarchical View of RxNorm with NDF-RT Drug Classes. *AMIA Annu Symp Proc* 2010;**2010**:577–81.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041416&tool=pmcentrez&rendertype=abstract (accessed 30 Sep2014).
16 Nelson SJ, Zeng K, Kilbourne J, *et al.* Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc*;**18**:441–8. doi:10.1136/amiajnl-2011-000116
17 Pathak J, Chute CG. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. *J Am Med Inform Assoc*;**17**:432–9. doi:10.1136/jamia.2009.001289
18 Pathak J, Richesson RL. Use of standard drug vocabularies in clinical research: a case study in pediatrics. *AMIA Annu Symp Proc* 2010;**2010**:607–11.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041282&tool=pmcentrez&rendertype=abstract (accessed 30 Sep2014).
19 Xu H, Stenner SP, Doan S, *et al.* MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc*;**17**:19–24. doi:10.1197/jamia.M3378
20 Noy NF, Dorf M, Montegut MJ, *et al.* BioPortal: A Web repository for

biomedical ontologies and data resources. In: *CEUR Workshop Proceedings*. 2008.

21 Fung KW, McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. *J Am Med Informatics Assoc* 2010;**17**:675–80. doi:10.1136/jamia.2010.007047

22 Zhao L, Lim Choi Keung SN, Arvanitis TN. A BioPortal-Based Terminology Service for Health Data Interoperability. *Stud Health Technol Inform* 2016;**226**:143–6.http://www.ncbi.nlm.nih.gov/pubmed/27350488 (accessed 8 Dec2016).