Methods for Managing Variation in Clinical Drug Names

Lee Peters[1], M.S., Joan E. Kapusnik-Uner[2], Pharm.D., Olivier Bodenreider[1], M.D., PhD
[1] National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
[2] First DataBank, South San Francisco, California, USA

{lpeters|obodenreider}@mail.nih.gov
Joan_Kapusnik@firstdatabank.com

Abstract

Objectives*: To develop normalization methods for managing the variation in clinical drug names.* ***Methods****: Manual examination of drug names from RxNorm and local variants collected from formularies led to the identification of three types of drug-specific normalization rules: expansion of abbreviations (e.g., tab to tablet);reformatting of specific elements (e.g., space between number and unit); and removal of salt variants (e.g., succinate from metoprolol succinate).* ***Results****: After drug-specific normalization, recall of 3397 previously non-matching names from formularies reaches 45% overall (70% of some subsets), compared to 10-20% after generic normalization. Ambiguity has not increased significantly in the RxNorm dataset.* ***Conclusions****: A limited number of drug-specific normalization operations provide significant improvement over general language normalization.*

Introduction

What's In a Name? The name by which an individual patient, healthcare provider or an institution knows and recognizes a drug name, is highly variable, and for justifiable reasons. Drug naming conventions suitable for one process within the medication management loop may not however be suitable for another application or end-user. The Institute for Safe Medication Practices (ISMP), The Joint Commission and others patient safety organizations have identified confusion over medication names as a significant source of medication errors within institutions. Many applications have thus gotten away from using product-based drug naming e.g. Label Name from the NDC (National Drug Code) level and instead implement names that have editorial policy around conventions that minimize error (e.g., Tall Man lettering in Computerized Prescriber Order Entry (CPOE) systems [1]). Pharmacy dispensing systems that generate medication prescription vial labels and medication order and dispense history information use yet again different naming/coding conventions suitable for these environments. Institutional and payer formulary systems may also have localized names.

At points in care it is necessary to aggregate and communicate clearly the current medication history for a given patient. The Joint Commission sets National Patient Safety Goals, one of which describes Medication Reconciliation [2]. When a patient is referred or transferred from one organization to another, the complete and reconciled list of medications is communicated to the next provider of service and the communication is documented. Alternatively, when a patient leaves the organization's care directly to his or her home, the complete and reconciled list of medications is provided to the patient's known primary care provider, or the original referring provider, or a known next provider of service.

For safety, regulatory and other reasons, the exchange of medication information often requires that drug names be mapped across information systems. The mapping of drug names across drug vocabulary standards is greatly facilitated by the existence of specialized terminology integration systems such as RxNorm (described in more detail later). In contrast, the mapping of drug names across formulary systems and between formulary systems and drug vocabulary standards remains challenging, in part because of the presence of local variants in drug names. Such mapping is often handled manually.

Methods for managing variation in biomedical vocabularies have been developed for terms from general and clinical biomedical vocabularies [3], but they perform suboptimally on specialized terms, such as gene names, for which specific normalization tools have been developed (e.g., [4]). The same can be expected of drug names, as they exhibit specific variation, unlikely to be covered by generic normalization rule, as we demonstrate later.

The objective of this study is to identify a set of transformation rules to facilitate the mapping of clinical drug names (including local variants of these) to terms from the RxNorm vocabulary. An ideal transformation increases the chances of mapping arbitrary

clinical drug names to RxNorm without increasing the ambiguity of normalized strings.

Background

RxNorm is a standardized nomenclature for medications produced and maintained by the U.S. National Library of Medicine (NLM) in cooperation with proprietary vendors [5-6]. RxNorm concepts are linked by NLM to multiple drug identifiers for each of the commercially available drug databases within the Unified Medical Language System® (UMLS®) Metathesaurus® (including NDDF Plus). In addition to integrating names from existing drug vocabularies, RxNorm creates standard names for clinical drugs. RxNorm has established a rich set of editorial guidelines (naming conventions, conversion of units, etc), which inform both the creation of standard names and the mapping of proprietary names to standard names. However, the required transformations are only partially automated and the creation of RxNorm relies heavily on the work of human editors. For this study we used the February 2010 release of the RxNorm dataset.

Lexical variant generation (lvg). The NLM lexical variant generation tool Norm (referred to hereafter as *lvg-norm*) is widely used in terminology applications [3]. For example, use of *lvg-norm* in the Unified Medical Language System (UMLS) forms the basis for automatic identification of synonym candidates. The normalization process is linguistically motivated and involves stripping genitive marks, transforming plural forms into singular, replacing punctuation (including dashes) with spaces, removing stop words, lower-casing each word, breaking a string into its constituent words, and sorting the words in alphabetic order. For example, the two terms Cancer of the Lung and Lung cancer share the same normalized form cancer lung. In this study, we use *lvg-norm* as a baseline normalization technique for evaluation purposes.

Ambiguity and variability in terminologies. Tsuruoka et al [7] define ambiguity and variability to quantify the utility of normalization rules. Ambiguity measures how many concepts share a given term on average:

$$\text{(ambiguity)} = \frac{1}{N}\sum_{i=1}^{N} C(t_i),$$

where N is the number of terms in the dataset, and $C(t_i)$ is the number of concept IDs that include a term whose spelling is identical to t_i.

Variability quantifies how many different names concepts have on average:

$$\text{(variability)} = \frac{1}{M}\sum_{j=1}^{M} T(c_i),$$

where M is the number of concept IDs in the dataset, and $T(c_i)$ is the number of unique terms that the concept c_i includes.

Complexity defines the overall "goodness" of the normalization rule. It is calculated as follows:

$$\text{(complexity)} = \text{(ambiguity)} \times \text{(variability)}^{\alpha}$$

where α is the constant that determines the trade-off between ambiguity and variability. For this study we use $\alpha = 1$. We calculate ambiguity, variability and complexity to assess the effect of various normalization techniques on the RxNorm terms.

Materials

Two different datasets are used in this study, namely a set of RxNorm terms (collected from existing vocabularies) and a set of drug names from formularies (i.e., local variants of drug names).

RxNorm terms. We use the set of over 127,000 names available in RxNorm for Semantic Clinical Drug (SCD) concepts, in order to identify frequent variability patterns, as well as to test the degree to which various normalization methods increase ambiguity. These terms encompass all ten of the source vocabularies in the RxNorm dataset, as well as the normalized names created by RxNorm.

Drug names from formularies. Drug formularies represented as data dictionaries are ubiquitous and are legendary for not being coded or derived from standard drug terminology. Several state Medicaid formularies were randomly selected for their diverse formats and structure. A set of 3,397 drug names was collected by First DataBank. The names in these three sample sets are local variants of drug names with no direct correspondence to RxNorm. We use this collection for identifying specific variability patterns and for testing the degree to which various normalization methods improve the mapping to RxNorm terms. Some sample names include:

KETOPROFEN ORAL 50MG CAPSULE
CARBIDOPA/LEVODOPA ORAL 50MG-200MG TABLET SA
FELODIPINE TAB.SR 24H 5 mg
Hydroxyzine HCl 10 5 mg/ml Syrup Oral

Methods

Identifying frequent variation patterns among clinical drug names in RxNorm. In addition to variation patterns already covered by *lvg-norm*, we identified several patterns specific to clinical drug names (by manual inspection of a subset of terms from both datasets, including specific outlier examples). Based on these specific variation patterns, we developed transformation rules to enhance the normalization process. These include:

Expanding abbreviated words. A number of source vocabularies use abbreviations in their clinical drug names. Shortened dose forms such as tab (tablet), susp (suspension) and cap (capsule) are common. Drug name abbreviations also appear, especially for multiple-ingredient drugs. For example, P-EPD TAN/CHLOR-TAN are shortened forms for Pseudoephedrine tannate and Chlorpheniramine tannate. We expand shortened names to the full names.

Targeted reformatting. Spacing between numbers and units, and large number formatting are used inconsistently. For example, some vocabularies may designate a dosage as 2mg (no space between the number and unit) and others as 2 mg (space between number and unit). We add spaces to the first case. Large numbers are formatted with and without commas (e.g. 1,000 or 1000). We remove commas in numbers. For example: 1,000 mg becomes 1000 mg.

Removing salt modifiers in ingredient names. Clinical drug names sometimes contain the salt modifiers and sometimes do not. We remove most salt modifiers. For example: Pseudoephedrine tannate becomes Pseudoephedrine. However, certain ingredients such as zinc have clinical names containing more than one salt form. For example Zinc acetate 50 mg oral capsule and Zinc gluconate 50 mg oral capsule. We do not remove salt forms of these ingredients having more than one salt form.

This set of specific transformations is meant to complement the generic transformations already implemented in *lvg-norm*. Therefore, the application of these specific rules can be thought of as additional pre-processing of the terms. We implemented the transformation rules into a program (*RxNormNorm*). This program expands abbreviated words, reformats specific parts of the clinical drug names, removes salt modifiers from ingredients, and finally applies the general transformation supported by *lvg-norm*.

Example of original and transformed strings:

Original name:
METOPROLOL SUCCINATE 200MG SA TAB
After lvg-norm (alone):
200mg metoprolol sa succinate tab
After RxNormNorm:
200 action metoprolol mg sustain tablet

In the example, *RxNormNorm* expands tab into tablet and sa into sustained action, separates 200 from mg, and removes the salt modifier succinate. Then the string is lower cased and the words are sort alphabetically.

Evaluating the benefit on recall. In order to assess the specific contribution of various normalization methods to recall, we measure the proportion of names from formularies with mapping to RxNorm without transformation, after applying *lvg-norm*, and after applying *RxNormNorm*.

Evaluating the impact on ambiguity. In order to assess the specific contribution of various normalization methods to ambiguity, we measure ambiguity (as well as variability and complexity) in the original RxNorm set of clinical drug names, after applying *lvg-norm*, and after applying *RxNormNorm*.

Results

Normalization rules. *RxNormNorm* currently includes two formatting rules (for dosage and large number formatting). It also contains a table of 46 salt modifiers to remove, and 216 different abbreviated words to expand.

RxNorm dataset. Of the 127,000 clinical drug names in RxNorm processed by *RxNormNorm*, 37,285 (29%) were modified for dosage formatting and 2,182 (2%) for large number formatting. Salt modifiers were removed from 30,763 terms (24%), and at least one word was expanded in 40,997 of the names (32%).

Formulary dataset. Of the 3,397 drug names collected from formularies, 937 (28%) were reformatted for dosage and 18 (1%) for large number. There were 1,263 terms (37%), which had salt modifiers removed, and 2,442 terms (72%), which had one or more word expansions for abbreviated words.

Benefit on recall. Table 1 (at the end of the manuscript) summarizes the mapping results for the set of names from formularies using exact name matching (i.e., without normalization), *lvg-norm* and *RxNormNorm*. Both *lvg-norm* and *RxNormNorm* significantly increased the number of mappings of the names from formularies. However, *RxNormNorm* shows significant improvement over *lvg-norm*. While the overall recall for *RxNormNorm* is 45%, it must be noted that the recall in the first two sets averages 70%. We discuss this difference in the next section.

Some sample mappings from *RxNormNorm*:

Formulary variant term:
ACETAMINOPHEN/PHENYLTOLX CIT ORAL 500MG-30MG TABLET
RxNormNorm mapped concept:
Acetaminophen 500 MG / phenyltoloxamine 30 MG Oral Tablet

Formulary variant term:
PROCHLORPERAZINE MALEATE SUPP.RECT 25 mg
RxNormNorm mapped concept:
Prochlorperazine 25 MG Rectal Suppository

Note that in these examples the word expansion, dosage separation and salt modifier removal are needed to achieve the matching.

Impact on ambiguity. Table 2 summarizes the number of ambiguous terms from the SCD names for each normalization type. The numbers in the second column represent the number of terms which mapped to more than one RxNorm concept. The ambiguity, variability and complexity values are displayed in columns 4-6 respectively. Note that the complexity value decreases compared to no normalization by using *lvg-norm*, and decreases again by using *RxNormNorm*, indicating that our normalization rules are effective in removing minor variants, but do not add unnecessary ambiguity.

Type of normalization	Ambiguous terms		Ambiguity	Variability	Complexity
	#	%			
None	459	0.36%	1.009	6.677	6.740
Lvg-norm	727	0.57%	1.013	6.476	6.558
RxNormNorm	1483	1.17%	1.019	5.890	6.003

Table 2 – Ambiguity in clinical drug names

Examples of ambiguity introduced by *lvg-norm*:

Original term:
10 ML Ephedrine 5 MG/ML Prefilled Syringe
Original and other mapped concepts:
10 ML Ephedrine 5 MG/ML Prefilled Syringe
5 ML Ephedrine 10 MG/ML Prefilled Syringe

In the above example, both concepts contain the numbers "5" and "10". Normalization blurs the ordering which is important here. (This phenomenon is similar to what was observed – in a limited number of cases – in general medical language, with classical examples, such as "nursing school" and "school nursing").

Original term:
Aspirin, 81 mg oral tablet
Original and other mapped concepts:
Aspirin 81 MG Enteric Coated Tablet
Aspirin 81 MG Oral Tablet

The original term was actually a term of the enteric coated tablet concept, and normalization produced a good match in this case.

Examples of ambiguity introduced by *RxNormNorm*:

Original term:
ACETAMINOPHEN 100MG/ML DROPS,ORAL
Original and other mapped concepts:
Acetaminophen 100 MG/ML Oral Solution
Acetaminophen 100 MG/ML Oral Suspension

The oral solution concept contained the original term. The oral suspension concept contained the term ACETAMINOPHEN 100 mg/ml ORAL DROPS.

Finally, in the following example, removing the salt modifier nitrate by *RxNormNorm* caused a mapping ambiguity.

Original term:
butoconazole 20 mg/ml vaginal cream
Original and other mapped concepts:
butoconazole 20 MG/ML Vaginal Cream
Butoconazole nitrate 20 MG/ML Vaginal Cream

Discussion

Findings. The major finding of this study is that the drug-specific normalization rules we developed significantly improve recall without negatively impacting ambiguity.

Benefit on recall. The results show adding additional normalization rules greatly increased the mappings in the set of names from formularies. The 385 additional mappings gained by using *lvg-norm* instead of exact match indicate the usefulness of applying generic normalization rules on clinical drug data. The increase in mappings from 389 to 1,525 using *RxNormNorm* instead of *lvg-norm* suggests that the additional normalization rules specific to the dataset were effective.

Impact on ambiguity. Using RxNormNorm exposed equivalent normalized terms in multiple concepts resulting in increased ambiguity in the RxNorm dataset. Many of the terms are not fully specified relative to one of the mapped concepts. For example, the original term ropinirole 2mg extended release tablet is mapped to both ropinirole 2 mg extended release tablet and 24 HR ropinirole 2 mg extended release tablet. The term is underspecified relative to the latter concept, as "24 HR" is implicit in the original term.

There were only 13 term ambiguous term mappings in the 1,525 total mappings from *RxNormNorm*, indicating that for the names from formularies, ambiguity was not a significant issue. Mappings using *lvg-norm* contained no ambiguity for the names from formularies.

Practical applications. The limited set of drug-specific normalization rules we developed can be easily implemented and extended to specific local variants.

Making RxNormNorm available. The work done in this study is a precursor to the development of a normalized matching function for the RxNorm API [8]. In July 2010, a normalized matching function based on this work was incorporated in the RxNorm API (function *findRxcuiByString*).

Improving Recall. The normalization rules implemented in *RxNormNorm* are relatively conservative and not specific to any particular drug vocabulary or set of local variants. As a consequence, *RxNormNorm* will likely perform suboptimally on specific local

variants. As *RxNormNorm* adds drug-specific normalization features on top of *lvg-norm*, a number of "localized" transformations can be performed prior to using *RxNormNorm*, in order to improve recall.

Recall results for the formulary data were low due to a number of observed characteristics of the terms. The most frequent characteristic of unmapped terms was the fact that many were missing dosage units. For example the term LEVONORGESTREL-ETHIN ESTRADIOL ORAL 0.1-0.02 TABLET is missing mg (twice). This particular issue accounts for most of the failed mappings in sample #3 and artificially lowers the overall recall.

Another frequent characteristic of unmapped formulary terms were the drug forms were partially specified. For example, the term IPRATROPIUM BROMIDE SPRAY 21 mcg needs the word nasal added to it to match an existing term. Other forms such as cream, ointment, solution and lotion need generally need modifier words such as topical, oral or ophthalmic to obtain matching.

Future work. In future work, we would like to apply drug normalization rules to quality assurance in RxNorm and develop alternative approaches to mapping drug names.

Application to quality assurance. The ambiguous results also showed some possible errors in the association between terms and concepts and provide targets for additional curation. For example:

Original term:
Dalteparin Sodium Inj 10000 Unit/ML

RxNormNorm mapped concepts:
Dalteparin 10000 UNT/ML Injectable Solution
Dalteparin 25000 UNT/ML Injectable Solution

The strength in the original term (10000 Unit/ML) seems inconsistent with that in the second concept (25000 UNT/ML).

Alternative approach. An alternate, more complex approach for mapping clinical drug names is to identify the specific components of a clinical drug name (ingredient, strength and dose form) and use the relationships in the RxNorm dataset to map the components to clinical drug names. We plan to investigate and evaluate this approach in the future.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. FDA and ISMP Lists of Look-Alike Drug Name Sets With Recommended Tall Man Letters: http://www.ismp.org/tools/tallmanletters.pdf
2. National Patient Safety Goals - Goal 8: Accurately and completely reconcile medications across the continuum of care: http://www.jointcommission.org/nr/rdonlyres/98572685-815e-4af3-b1c4-c31b6ed22e8e/0/07_hap_npsgs.pdf
3. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994:235-9
4. Lau WW, Johnson CA, Becker KG. Rule-based human gene normalization in biomedical text with confidence estimation. Comput Syst Bioinformatics Conf 2007;6:371-9
5. Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Professional 2005;7(5):17-23
6. RxNorm: http://www.nlm.nih.gov/research/umls/rxnorm/
7. Tsuruoka Y, McNaught J, Ananiadou S. Normalizing biomedical terms by minimizing ambiguity and variability. BMC Bioinformatics 2008;9 Suppl 3:S2
8. RxNorm API: http://rxnav.nlm.nih.gov/RxNormAPI.html

Formulary		**Exact Match**		***Lvg-norm***		***RxNormNorm***		**% match increase**	
Set	# terms	# matches	Recall	# matches	Recall	# matches	Recall	*Lvg-norm* /Exact	*RxNormNorm*/ *lvg-norm*
1	1103	0	0.00 %	299	27.1 %	672	61.0 %		125%
2	979	4	0.41 %	79	8.1 %	761	76.7 %	1875%	863%
3	1307	0	0.00 %	11	0.8 %	92	7.0 %		736%
Total	**3389**	**4**	**0.11 %**	**389**	**11.5 %**	**1525**	**45.0 %**	**9625%**	**292%**

Table 1 – Matching results for the names from formularies