Analysis of Cross-Institutional Medication Information Annotations in Clinical Notes

Sunghwan Sohn[1], Cheryl Clark[2], Scott Halgrim[3], Sean Murphy[1], Siddhartha Jonnalagadda[1], Kavishwar Wagholikar[1], Stephen Wu[1], Christopher Chute[1], Hongfang Liu[1]
[1]Mayo Clinic, [2]MITRE, [3]Group Health Research Institute
{sohn.sunghwan, murphy.sean, siddhartha, wagholikar.kavishwar, wu.stephen, chute, liu.hongfang}@mayo.edu, cclark@mitre.org, halgrim.s@ghc.org

Abstract

A large amount of medication information resides in unstructured text in electronic medical records, which requires advanced techniques to be properly mined. In clinical notes, medication information follows certain semantic patterns (e.g., medication, dosage, frequency, mode, etc.). Some medication descriptions contain additional word(s) between medication attributes. Therefore, it is essential to understand the semantic patterns as well as the patterns of the context interspersed among them (i.e., context patterns) to effectively extract comprehensive medication information. In this paper we examined both semantic and context patterns and compared those found in Mayo Clinic and i2b2 challenge data. We found that some variations exist between the institutions but the dominant patterns are common.

1 Introduction

Electronic medical records (EMRs) have grown rapidly, and a large amount of clinical data is stored in free-text format. Natural language processing (NLP) techniques, which can convert unstructured text to a structured format, have been successfully applied in various clinical applications including patient medical status extraction [1, 2], sentiment analysis [3], decision support [4, 5], genome-wide association studies [6, 7], and diagnosis code assignment [8, 9].

Over the past decade, multiple clinical NLP systems have been developed and applied in the clinical domain, such as cTAKES [10], YTEX [11], MTERMS [12], HiTEXT [13], MedLEE [14]. Although clinical NLP systems have proven to be successful in many information extraction tasks, their performance often varies across institutions and sources of data [15, 16]. As some researchers have demonstrated, this may be due to the fact that clinical language is not homogeneous but instead consists of heterogeneous sublanguage characteristics [17-21].

Medication information is one of the major data components in EMRs. Patient medication history is of great concern for future medical treatments and plays a vital role in enabling the secondary use of EMRs for clinical and translational research. Although some medication information can be extracted from the structured data, a substantial amount of it resides in an EMR's unstructured text and requires advanced techniques, such as NLP, to extract. Since the medication information is described in various ways in the clinical notes, understanding the semantic patterns—i.e., description patterns of a medication name and attributes—and context patterns—i.e., context variations between the semantic patterns—is essential to effectively extract comprehensive medication information and develop a general and customizable medication extraction tool.

In this paper, we investigated both the semantic and context patterns of medication descriptions in clinical notes from two different sources: Mayo Clinic and the 2009 i2b2 medication extraction challenge [22]. We provide various pattern statistics and compare them at both the section level and the institution level to better understand the usage of semantic and context variations.

2 Background

Early research on medication extraction focused on extracting the medication name itself [23, 24] and mapping it to a standardized nomenclature [25]. The 2009 i2b2 medication extraction challenge focused on extracting comprehensive medication information (i.e., medication name as well as the attributes such as dosage, mode,

frequency, duration, and reason) from clinical discharge summaries [22]. In this challenge, most top-performing teams used rule-based systems developed based on the examination of medication description patterns [22]. Because many of these systems performed well on most attributes, we may assume that medication information in clinical notes follows some patterns.

Some studies show that sublanguage-based clinical text processing systems, such as MedLEE [14], perform as accurately as medical experts on clinical concept extractions [13]. A sublanguage theory hypothesizes that the information content and structure from specific domains can be defined using a sublanguage grammar that specifies both domain-specific syntactic and semantic information [19, 26]. Aligned with sublanguage grammars, Xu et al. [27] performed a preliminary study on parsing medication sentences by assigning the probability to semantic-based grammar rules for medication detection. Their study showed promising results to disambiguate the complex sublanguage grammars of medication sentences.

Patterson and Hurdle [28] experimentally demonstrated that different clinical domains use their own sublanguages in clinical narratives. They conducted clinical document clustering in 17 different clinical domains and showed that note types in a broad clinical scope form the same cluster, but note types in a narrow clinical extent form different clusters. This study suggests that the performance of a clinical NLP system may depend on the source of the clinical notes and that the semantic and context information of the target notes should be seriously considered in the tool development process.

Previous studies support the idea that medication descriptions also exhibit sublanguage characteristics, and therefore an understanding of those characteristics in clinical notes within and across institutions is fundamental to properly extract medication information.

3 Materials and Methods

This study used annotated medication corpora from two sources: Mayo Clinic and the 2009 i2b2 medication extraction challenge data. The Mayo medication data consist of 159 clinical notes randomly selected from clinical notes. There are 659 manually annotated medication mentions with attributes. For i2b2 medication data, 253 discharge summaries from Partners Healthcare were manually annotated by challenge teams and released to the community for research [29]. In the combined corpora there are 9,003 medication mentions along with attributes. The semantic types of medication mentions and their abbreviations for both data sets are described in Table 1 (i2b2 annotations are excerpted from [29]).

We investigated both semantic and context patterns of medication descriptions in various aspects. The Mayo clinical notes consist of multiple sections explicitly separated by section tags. The i2b2 medication annotations include format information where the medication information is described—i.e., list vs. narrative. We included pattern variations within an individual section as well as a list vs. narrative format.

A. Semantic pattern parsing

The medication description in clinical notes consists of a medication name and its attributes such as dosage, frequency, mode, etc. We refer to a sequence of such semantics tags (ignoring other text between them) as the *medication semantic pattern*. For example,

Aspirin was reduced to *1 tablet po twice daily*.
Mayo: m do fm mo fn fu
semantic pattern: *m|do|fm|mo|fn|fu*

Aspirin was reduced to *1 tablet po twice daily*.
i2b2: m do mo f
semantic pattern: *m|do|mo|f*

We investigated these patterns in each data set and also compared them across the data sets (Mayo vs. i2b2). Since the two corpora have different annotation guidelines, we mapped semantically equivalent Mayo annotations to i2b2 annotations in the following way (Mayo → i2b2): mo → mo; sn, su, do → do; fn, fu →f; fm → N/A. In order to simplify the expression of semantic patterns we assigned a short symbol to dominant patterns. Table 2 contains those pairs and actual text examples.

TABLE 1. SEMANTIC TYPE ANNOTATION OF MEDICATION INFORMATION IN MAYO AND I2B2. PARENTHESIZED STRINGS IN THE SEMANTIC TYPE COLUMN ARE ABBREVIATIONS.

i2b2		Mayo	
Semantic type	**Definition**	**Semantic type**	**Definition**
medication (m)	medication name	medication (m)	medication name
dosage (do)	the amount of a single medication used in each administration (e.g., *"one tab" "4 units" "30 mg"*)	dosage (do)	how many of each medication the patient is taking (e.g., "*2*" in "2 daily", "*1*" in "1 tablet")
		strength number (sn)	e.g., "*30*" in "30 mg"
		strength unit (su)	e.g., "*mg*" in "30 mg"
frequency (f)	how often each dose of the medication should be taken (e.g., "*daily" "once a month" "3 times a day"*)	freq number (fn)	e.g., "*twice*" in "twice a day"
		freq unit (fu)	e.g., "*day*" in "twice a day"
mode (mo)	the route for administering the medication (e.g., "*oral" "intravenous" " topical"*)	route (mo)	same as the i2b2
duration (du)	how long the medication is to be administered (e.g., "*for a month" "during spring break"*)	duration (du)	same as the i2b2 but not include preposition (e.g., "*a month*" in "for a month)
		form† (fm)	the physical appearance of the medication (e.g., *aerosol, capsule, cream, tablet*)

†In i2b2, there is no form annotation but it may be part of the medication or dosage (e.g., m: "Aspirin tablet", do: "1 tab")

TABLE 2 EXAMPLES OF MEDICATION SEMANTIC PATTERNS AND SYMBOLS. IN EXAMPLE, SEMANTIC TYPES ARE ANNOTATED AS XML-STYLE TAGS.

symbol	**semantic pattern**	**example**
A	m	<m>Aspirin</m>
B	m|do|mo|f	<m>Zocor</m> <do>5 mg</do> <mo>p.o.</mo> <f>q.h.s.</f>
C	m|do|f	<m>Glipizide</m> <do>5 mg</do> <f>b.i.d</f>.
D	m|mo	<m>Lasix</m> <mo>drip</mo>
E	m|do	<m>Coumadin</m> <do>alternating doses of 4 mg</do>
F	mo|m	<mo>IV</mo> <m>ACE inhibitors</m>
G	do|m	<do>one unit</do> of <m>packed red cells</m>
H	m|du	<m>Dilantin</m> <du>for less than a year</du>
I	m|f	<m>Pepcid AC</m> <f>QHS</f>
J	m|do|mo|f|du	<m>KEFLEX (CEPHALEXIN)</m> <do>250 MG</do> <mo>PO</mo> <f>QID</f> <du>X 12 doses</du>
K	m|do|mo	<m>acetylsalicylic acid</m> <do>325 mg</do> <mo>p.o.</m>
L	m|do|f|du	<m>Lasix</m> <do>40 mg</do> <f>QD</f> <du>x3 doses</du>
M	du|m	<du>two days</du> of <m>Indocin</m>
N	f|m	stable on <f>b.i.d.</f> <m>torsemide</m>
O	m|mo|f	<m>nitroglycerin</m> <mo>sublingual</m> <f>p.r.n.</f>
P	m|mo|do|f	<m>Lovenox</m> <mo>subcutaneously</mo> <do>90 mg</do> <f>daily</f>
Q	do|mo|m	<do>10 mg</do> of <mo>IV</mo> <m>Lopressor</m>
R	mo|m|du	<mo>IV</mo> <m>cefotaxime</m> <du>for the 7-day period</du>
S	do|m|f	<do>low dose</do> <m>dilaudid</m> <m>as needed</m>
T	m|do|du	<m>heparin</m> <do>500 units</do> <du>for 48 hours</du>

B. Context pattern discovery

We use the phrase *medication context pattern* to refer to the semantic pattern plus the text that occurs between the semantic tags. Any text that occurs before the first semantic tag or after the last semantic tag is excluded. For example,

"The patient said that *Aspirin* was reduced to *1 tablet po twice daily* then stopped it"
Context pattern:
<m> was reduced to <do><mo><f> (i2b2)
<m> was reduced to <do><fm><mo><fn><fu> (Mayo)

C. Large corpus analysis (Mayo)

Mayo 159 clinical notes have a relatively lower number of medication descriptions (659 in Mayo vs. 9,003 in i2b2) and therefore it might cause sampling bias. To alleviate this issue we also examined 10,000 Mayo Clinic notes randomly selected from the Mayo patient cohort in the eMERGE study [30]. They were processed by the drug NER annotator in cTAKES [10] and their semantic patterns were parsed for comparison with i2b2. Drug NER uses rule-based pattern match implemented by finite state machine to parse medication and its attributes. These semantic patterns were not manually reviewed.

4 Results

A.Medication Semantic Patterns

1) Mayo data

In the Mayo corpus, there are a total of 89 unique semantic patterns, but 85% of the medication mentions occur in the 20 most frequent patterns; 70% of the medication mentions occur in the 5 most frequent patterns: *m* (322 mentions), *m|sn|su|fu* (85), *m|fu* (19), *m|sn|su|fn* (17), *m|sn|su|fn|fu* (16). Medication alone is the most dominant pattern, comprising almost half of all the medication mentions.

The Mayo clinical notes consist of multiple sections, and each section contains specific content. To further examine the content-based medication description, we analyzed each section separately and obtained section-specific semantic patterns. These results appear in Table 3.

TABLE 3. SECTION-LEVEL STATISTICS FOR MAYO MEDICATION SEMANTIC PATTERNS.

section	# medications	# patterns	patterns[†] (>=5)
Impression/Report/Plan	227	53	m (97), m|sn|su|fu (42), m|sn|su (8), m|fu (5),
Current Medications	183	38	m (52), m|sn|su|fu (28), m|fu (11), m|sn|su|mo|fu (9), m|sn|su|do|fu (9), m|sn|su|do|fn (8), m|sn|su|fn (6), m|sn|su (6), m|sn|su|fn|fu (5)
History of Present Illness	139	26	m (82), m|sn|su|fu (14), m|sn|su|fn (8), mo|m (5)
Allergies	39	1	m (39)
Past Medical/Surgical History	21	4	m (15)
Problem Oriented Hosp. Course	19	8	m (10)
Social History	10	3	m (8)
Chief Complaint/Reason for Visit	7	3	m (5)
The other sections	14	1	m (14)

[†]number within ( ) denotes # medications of a given pattern.

There are 14 sections that contain medications. Almost 90% of the medications appear in the four most dominant sections: Impression/Report/Plan, Current Medications, History of Present Illness, and Allergies. If we consider only patterns that occur five times or more throughout all sections, the Current Medication section is the most diverse with nine patterns. With the exception of the three most dominant sections, the other sections contain only one semantic pattern—*m (medication)*—that occurs five times or more. This pattern comprises all 39 mentions in the Allergies section.

2) i2b2 data

A similar dominance of a few patterns exists in the i2b2 corpus. Figure 1 shows the medication semantic patterns and their counts that cover 95% of the medication mentions in the 253 discharge summaries. A cumulative percentage denotes what portion of the counts from the total medication mentions appears in the given series of patterns up until that point. The semantic pattern in the figure is described as a short symbol with the actual semantic pattern in parentheses. There are a total of 69 unique semantic patterns, but 95% of the medication mentions occur in the 13 most frequent patterns.

3) Mayo vs. i2b2

In order to properly compare Mayo semantic patterns with i2b2 semantic patterns, the Mayo medication semantic tags were mapped to i2b2 tags. The distribution of these semantic patterns is presented in Figure 2. After mapping, 85% of the medication mentions occur in four dominant patterns *(m, m|do|f, m|do, m|do|mo|f)*, and 95% of them occur in 11 patterns.

Each i2b2 annotation also contains format information, indicating whether the mention occurred in a list or narrative context. Our observations of Mayo data show that the Current Medication and Allergies sections have a list format for the medication description, while the other sections have a narrative format. To compare the pattern variations in different formats, we characterized Mayo's Current Medication and Allergies sections as list and the others as narrative. As before, Mayo's semantic tag set was mapped to that of i2b2. The comparisons for list and narrative formats are shown in Figure 3 and Figure 4, respectively. The y-axis denotes the percentage of medication mentions out of the total medication mentions for a given pattern. We show the top 95% most frequent medication semantic patterns.

In list-formatted sections, Mayo data have five semantic patterns that cover 95% of the medication mentions while i2b2 data have nine (Figure 3). There are five semantic patterns that appear in both Mayo and i2b2. The most frequent pattern in the Mayo data is *"m"* while it is *"m|do|mo|f"* in the i2b2 data.

In the narrative sections, Mayo data have 13 semantic patterns that cover 95% of the medication mentions while i2b2 data have 15 (Figure 4). There are 10 common semantic patterns that appear in both the Mayo and i2b2 data.

To see the distribution of both Mayo and i2b2 semantic patterns in a larger data set, we also examined a large corpus of Mayo data. Ten thousand Mayo clinical notes were processed by the cTAKES drug NER module to annotate medication information, and we refer to the results as the Mayo 10K corpus. Note that these are system-generated results and were not manually reviewed by human experts.

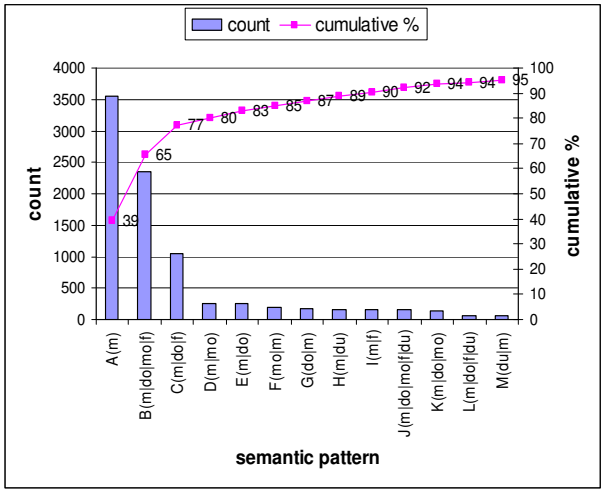

Figure 1. i2b2 medication semantic patterns (top 95%).

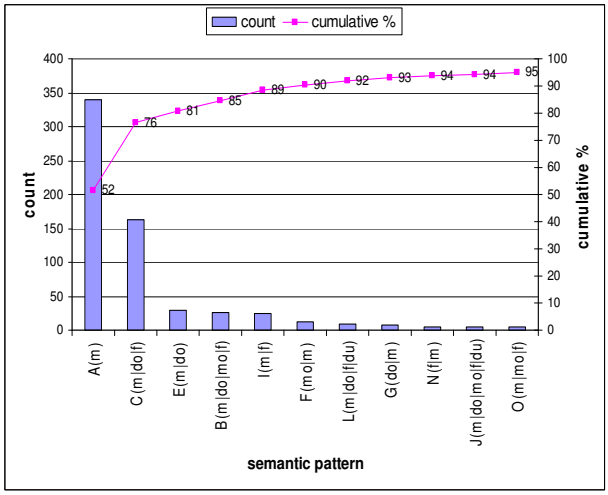

Figure 2. Mayo medication semantic patterns mapped to i2b2 (top 95%).

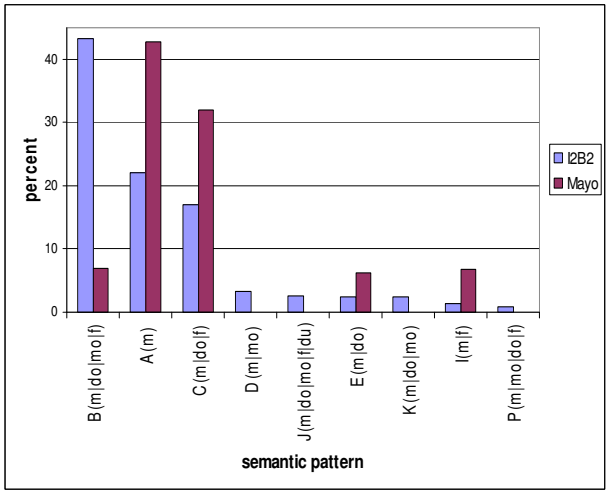

Figure 3. Mayo vs. i2b2 medication semantic patterns in list (top 95%).

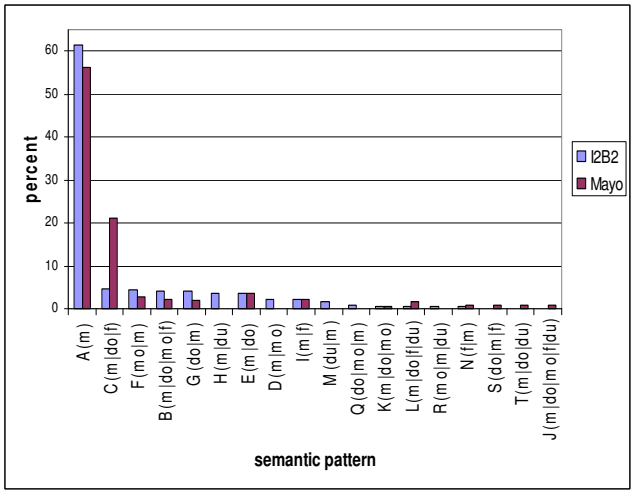

Figure 4. Mayo vs. i2b2 medication semantic patterns in narrative (top 95%).

Figure 5 shows the i2b2, Mayo, and Mayo 10K medication semantic patterns and also includes trends of those patterns (moving average with period two). The Mayo medication annotations were also mapped to the i2b2 annotations. In the top 95%, out of 15 semantic patterns, nine patterns appear in both Mayo and i2b2, four patterns appear only in i2b2 *(m|mo, m|du, m|do|mo, du|m)*, and two patterns appear only in Mayo *(f|m, m|mo|f)*. All data sets have the pattern of the medication name appearing alone as the most frequent pattern, but the second most frequent pattern in Mayo is *"m|do|f."* In both Mayo 10K and i2b2, it is *"m|do|mo|f,"* and this shows the notable difference in trends. In Mayo 10K, there is a total of 311,004 medication mentions, and 95% of all the medication mentions appears in nine sections. After mapping the Mayo annotations to i2b2, 95% of all the medication mentions occur in eight semantic patterns. Intriguingly, the Mayo 10K's three most dominant patterns (*m, m|do|mo|f, m|do|f*) exactly match with the three most dominant patterns of i2b2—they cover 80% and 77% of total medication mentions in Mayo 10K and i2b2 respectively. Comparing the trend to Mayo 159 notes, this larger Mayo corpus looks more similar to i2b2 semantic patterns.

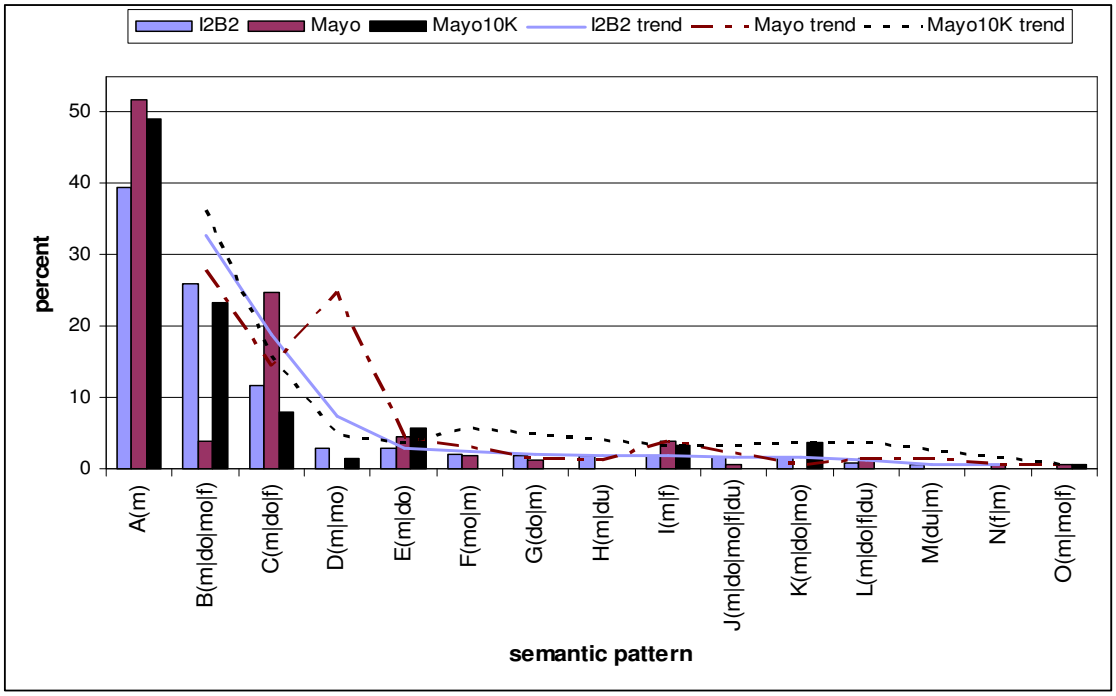

Figure 5. i2b2 vs. Mayo vs. Mayo10K medication semantic patterns with moving average trend (top 95%)

We also compared the semantic patterns in list vs. narrative for both data sets. In the list sections, Mayo's top three patterns *(m|do|mo|f, m, m|do|f)* also exactly matched the top three of i2b2, and covered 75% and 82% of the medication mentions in Mayo and i2b2, respectively. In the narrative sections, Mayo had only five patterns that covered 95% of the medication mentions, and all five patterns appeared in i2b2. Mayo's top two *(m, m|do|f)*, which covered 84% of the medication mentions, matched i2b2's top two patterns.

B. Medication Context Patterns

Medications in clinical notes are sometimes mentioned with additional words inserted between medication attributes. Those words may provide clues to medication semantics, so we examined this text.

1) Mayo data

The Mayo data have a total of 659 medication mentions. Out of those, only 93 cases (14%) have one or more words between medication semantic attributes. Of those 93 cases, 82 are unique. Since the individual words between the medication attributes are so varied, we did not notice a high frequency of exact context patterns. Therefore, we examined the semantic characteristics of the dominant partial context patterns. Table 4 shows statistics of dominant partial context patterns. When the medications are followed by verb forms, often the verbs are related to medication status changes—i.e., change of medication attributes. For example, *<m> was increased to <sn>*. The other minor patterns not in Table 4 are: *<m> another medication <fu>|<fn>*, x *<du>|<do>, <m> at <do>, <m> from <sn>, <m> be given <fu>, <m> were|used for <fm>*.

TABLE 4. MEDICATION CONTEXT PATTERNS IN MAYO (NUMBER DENOTES FREQUENCY)

Partial pattern	Status change pattern*
22 per|a <fu>	6 <m> *INCREASE* indication <sn>
12 for (period of|an additional) <du>	1 <m> *DECREASE* indication <sn>
7 <fn> times (a) <fu>	3 <m> *START* indication <sn>
6 <do>|<mo> every <fn>	
5 <m> dose|cycle (of|at) <sn>	
5 <m> to <sn>	
5 <su>|<du>|<fm> of <m>	

* Lexical variations were normalized

2) i2b2 data

Since i2b2 medication annotations are more coarse-grained than Mayo, we were able to find a high frequency of exact context patterns. In i2b2, there are a total of 9,003 medication mentions, and only 390 of those (4%) contain word(s) between medication attributes. Out of those 390 cases, 231 are unique context patterns. Table 5 shows the major exact and partial context patterns. Approximately one third of the total exact patterns belong to the exact patterns in Table 5. Similar to the Mayo data, when the medications are followed by verb forms, many are related to medication status changes. For example, *<m> was started…<du>, <m> was increased to <do>, <m> was reduced to <do>*. Another medication mention also appears between medication semantics, for example, *<m> and Lipitor <du>*. The other minor patterns include *<m> on <do>, <mo> with <m>*.

TABLE 5. MEDICATION CONTEXT PATTERNS IN I2B2 (NUMBER DENOTES FREQUENCY)

Exact pattern	Partial pattern	Status change pattern*
71 <do> of <m>	141 <do>|<du> of <m>	16 <m> *INCREASE* indication <do>
28 <du> of <m>	21 <m> at <do>	2 <m> *INCREASE* indication <f>
9 <do> of <m> <f>	8 <m> (dose) from <do>	1 <m><do> *INCREASE* indication <f>
8 <du> of <mo> <m>	5 <m> from <do>	8 <m> *DECREASE* indication <do>
6 <m> at <do>	6 <m> to <do>	1 <m> *DECREASE* indication <du>
6 <do> of <mo> <m>	4 <m> with <do>	1 <m> *DECREASE* indication <f>
5 <m> at <do> <mo> <f>		8 <m> *START* indication <do>
5 <m> at <do> <f>		3 <m> *START* indication <du>
5 <do> of <m> <du>		1 <du> *START* indication <m>

* Lexical variations were normalized

5 Discussion

Medication information was expressed in numerous ways in the clinical notes. We have investigated the medication description patterns in manually annotated data from 159 Mayo clinical notes and 253 i2b2 discharge summaries. Additionally, we have further compared these with a large medication corpus compiled by cTAKES drug NER from Mayo's 10,000 clinical notes.

Notes from Mayo and i2b2 share most of the common semantic patterns, but differences exist. Overall, the medication name alone is the most dominant pattern in both data sets. In Mayo, about 90% of the medication mentions occur in four sections, and all but the Allergies section contain a variety of semantic patterns. The Allergies and the other sections mostly contain the medication name alone.

In lists, i2b2 has *"m|do|mo|f"* as the most frequent pattern, but in Mayo data it is the medication name. This might be because Mayo's samples of 159 notes contain a relatively lower portion of Current Medication sections, which in fact, contain most of the medication information, as well as the most common list-format section. After mapping the Mayo annotations to the i2b2 annotations, the most frequent pattern in the Current Medication section is *"m|do|f."* Based on these observations, we could conclude that the list-format medication descriptions tend to provide at least the minimum information necessary to guide the proper medication intake (i.e., medication with dosage, frequency and possibly mode information).

In narratives, the pattern of the medication name alone is the most frequent in both data sets. Generally, narratives in clinical notes describe the physician's impression, report, and plan, as well as the patient's medical history, diagnosis, etc. In content of this nature, we may assume that physicians often simply mention medication names without further detailed attributes in clinical narratives as they are describing other conditions. The second dominant pattern in both data sets is *"m|do|f"*—21% and 5% out of the total medication mentions in Mayo and i2b2 respectively. In both data sets, the rest of the patterns cover a lower number of the medication mentions.

After mapping the Mayo annotations to i2b2, it can be seen that 12% (N=11) of the semantic patterns cover 95% of the medication mentions. In i2b2, 19% (N=13) of the semantic patterns cover 95% of the medication mentions. The i2b2 medication mentions are slightly more diverse than Mayo's. Covering 95% of the entire medication mentions from both data sets are total 15 semantic patterns, nine of which are common in both.

In both Mayo's 159 notes and i2b2, the medication descriptions do not contain many words between the medication attributes—14% and 4% of medication mentions contain some words between them in Mayo and i2b2, respectively. One reason for the higher context pattern for the Mayo data is due to different annotations. For example, if we look at duration, Mayo does not contain prepositions (e.g., for a week → duration is "a week") but i2b2 does (duration is "for a week"). Some prepositions like "of" are commonly used in combination with medication and dosage/duration (e.g. dosage or duration of medication). When a medication is followed by a verb, it can be related to a medication status change in many cases. Such medication status change descriptions often follow a pattern like: *medication + status change verb + dosage or frequency*.

When we examined a large Mayo corpus of 10,000 clinical notes, those semantic patterns were even closer to i2b2. Although there is more similarity between the two, Mayo 10,000 clinical notes have not been reviewed manually and therefore this might make the results slightly less reliable. The trends of i2b2 and Mayo 10K in Figure 5 seem to be very similar. The top three semantic patterns, which cover the majority of the medication mentions, are exactly matched with i2b2's. The portion of the list of medication descriptions is also closer to i2b2 than Mayo 159 notes. In Mayo 10K, 49% of the medication mentions belong to the list format (34% in Mayo 159 note); in i2b2, 56% belong to the list format. These observations imply that a medication extraction tool based on the semantic patterns in Mayo data may work reasonably well for i2b2 and vice versa.

In our semantic pattern analysis, we considered only one medication mention at a time. For example, "Aspirin or Tylenol 1 tablet prn" generates two instances of the semantic pattern *"m|do|f"* instead of *"m|m|do|f"*. In the future, we will further compile and analyze these kinds of patterns.

As shown in this study, the sublanguage of medication description in general can be well-characterized through semantic patterns. A medication extraction tool can utilize sublanguage—i.e., instantiating a given medication template with a given medication and the corresponding attributes. The statistics of medication patterns obtained in this study can be used as a confidence measure to determine whether a given medication and nearby attributes are truly associated or not. If its description pattern belongs to one of the majority patterns, we have high confidence in considering it as a correct association.

6 Conclusion

The semantic patterns that are used to describe medications vary in clinical notes as a whole, and some variations also exist between institutions. However, dominant patterns do exist that account for most of the medication descriptions, and most of those patterns are common between Mayo and i2b2. Extra context between the medication attributes is not common in the medication descriptions, but it does occur in some particular patterns primarily associated with medication, including the dosage or duration "of" medication and a medication followed by a status change verb. Those semantic and context patterns could be utilized to properly parse medication and its attributes, as well as to correctly associate them.

Acknowledgements

This manuscript was supported by Strategic Health IT Advanced Research Projects (SHARP) Program (90TR002), National Science Foundation ABI:0845523, and National Institute of Health 5R01LM009959. Jonnalagadda was supported by grant number 1K99LM011389 from the NLM.

References

1 Sohn S, Kocher J-PA, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. J Am Med Inform Assoc. 2011;**18**(Suppl 1):144-9.
2 Sohn S, Savova GK. Mayo Clinic Smoking Status Classification System: Extensions and Improvements. AMIA Annual Symposium; 2009; San Francisco, CA: 619-623.
3 Sohn S, Torii M, Li D, Wagholikar K, Wu S, Liu H. A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes. Biomedical informatics insights.**2012**(Suppl. 1):43-50.
4 Demner-Fushman D, Chapman W, McDonald C. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics. 2009;**42**(5):760-72.
5 Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining decision support methodologies to diagnose pneumonia. 2001: American Medical Informatics Association; 2001. p. 12.
6 Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. Journal of the American Medical Informatics Association. 2010;**17**(5):568-74.
7 Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. A genome-wide association study of red blood cell traits using the electronic medical record. PLoS One. 2010;**5**(9):e13011.
8 Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association. 2004;**11**(5):392.
9 Pakhomov SVS, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. Journal of the American Medical Informatics Association. 2006;**13**(5):516-25.
10 Savova G, Masanz J, Ogren P, et al. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;**17**(5):507-13.
11 YTEX - Yale cTAKES Extension.
12 Zhou L, Plasek JM, Mahoney LM, et al. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to Process Medication Information in Outpatient Clinical Notes. American Medical Informatics Association; 2011. p. 1639.
13 Zeng Q, Goryachev S, Weiss S, Sordo M, Murphy S, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Medical Informatics and Decision Making. 2006;**6**(1):30.
14 Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994 Mar-Apr;**1**(2):161-74.
15 Fan J, Prasad R, Yabut RM, et al. Part-of-speech tagging for clinical text: wall or bridge between institutions? ; American Medical Informatics Association; 2011. p. 382.

16 Wagholikar K, Torii M, Jonnalagadda S, Liu H. Feasibility of pooling annotated corpora for clinical concept extraction. Proceedings AMIA CRI 2012; San Francisco, CA.
17 Friedman C. A Broad Coverage Natural Language Processing System. American Medical Informatics Association Symposium; 2000. p. 270-4.
18 Stetson PD, Johnson SB, Scotch M, Hripcsak G. The sublanguage of cross-coverage. American Medical Informatics Association; 2002. p. 742.
19 Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of Biomedical Informatics. 2002;**35**(4):222-35.
20 Harris ZS. A grammar of English on mathematical principles: Wiley New York:; 1982.
21 Harris ZS. A theory of language and information: a mathematical approach: Clarendon Press Oxford:; 1991.
22 Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association. 2010;**17**(5):514-8.
23 Chhieng D, Day T, Gordon G, Hicks J. Use of natural language programming to extract medication from unstructured electronic medical records. 2007;. p. 908.
24 Sirohi E, Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. 2005; p. 308-18.
25 Levin MA, Krol M, Doshi AM, Reich DL. Extraction and Mapping of drug Names from Free Text to a Standardized Nomenclature. Proc AMIA Ann Symposium; 2007; p. 438-42.
26 Harris ZS. The structure of science information. Journal of Biomedical Informatics. 2002;**35**(4):215-21.
27 Xu H, AbdelRahman S, Lu Y, Denny JC, Doan S. Applying Semantic-based Probabilistic Context-Free Grammar to Medical Language Processing-A Preliminary Study on Parsing Medication Sentences. Journal of Biomedical Informatics. 2011.
28 Patterson O, Hurdle JF. Document Clustering of Clinical Narratives: a Systematic Study of Clinical Sublanguages. American Medical Informatics Association; 2011. p. 1099.
29 Uzuner Ö, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. Journal of the American Medical Informatics Association. 2010;**17**(5):519-23.
30 McCarty C, Chisholm R, Chute C, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics. 2011;**4**(1):13.