Extracting Rx Information from Clinical Narrative

James G. Mork[1], MSc

jmork@mail.nih.gov

Olivier Bodenreider[1], M.D., PhD

obodenreider@mail.nih.gov

Dina Demner-Fushman[1], M.D., PhD

ddemner@mail.nih.gov

Rezarta Islamaj Do• an[2], PhD

islamaj@ncbi.nlm.nih.gov

François-Michel Lang[1], MSE

flang@mail.nih.gov

Zhiyong Lu[2], PhD

luzh@ncbi.nlm.nih.gov

Aurélie Névéol[2], PhD

neveola@ncbi.nlm.nih.gov

Lee Peters[1], MSc

lpeters@mail.nih.gov

Sonya E. Shooshan[1], MLS

sshooshan@mail.nih.gov

Alan R. Aronson[1], PhD

alan@nlm.nih.gov

[1]Lister Hill National Center for Biomedical Communications (LHNCBC),

[2]National Center for Biotechnology Information (NCBI),

U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Contact Information:
Alan R. Aronson, PhD
National Library of Medicine
Building 38A, Room 9N-905
8600 Rockville Pike, MSC-3826
Bethesda, MD 20894
Phone: 301.435.3162
Fax: 301.496.0673
alan@nlm.nih.gov

Abstract

OBJECTIVE: The i2b2 Medication Extraction Challenge provided an opportunity to evaluate our entity extraction methods, contribute to the generation of a publicly available collection of annotated clinical notes and start developing methods for ontology-based reasoning using structured information generated from the unstructured clinical narrative. DESIGN: We addressed the task of extracting salient features of medication orders from the text of de-identified hospital discharge summaries with a knowledge-based approach using simple rules and lookup lists. We combined our entity recognition tool, MetaMap, with dose, frequency and duration modules specifically developed for the Challenge as well as a prototype module for reason identification. MEASUREMENTS: Evaluation metrics and corresponding results were provided by the Challenge organizers. RESULTS: Our results indicate that robust rule-based tools achieve satisfactory results in extraction of simple elements of medication orders, but more sophisticated methods are needed for identification of reasons for the orders and durations. LIMITATIONS: Due to the time constraints and nature of the challenge, some obvious follow-on analysis has not been completed yet. CONCLUSIONS: We plan to integrate the new modules with MetaMap to enhance its accuracy. This integration effort will provide guidance in retargeting our existing tools for better processing of clinical text.

I. INTRODUCTION

Extraction of the elements of medication orders from clinical narrative is a preliminary step in many important applications of medical informatics. These applications include but are not limited to: support of quality assurance through reconciliation of patient's medication lists and clinical notes [1, 2]; detection of adverse reactions to drugs [3] and medication non-compliance [4]; study of a population's response to a drug [5]; support of care plan development [6]; and identification of inactive medications [7].

Whereas evaluation of the individual efforts in extraction of medication names from biomedical literature could use "found data", such as Medical Subject Headings (MeSH®) assigned to MEDLINE® abstracts in the manual indexing process [8], until recently, no annotated resources for evaluation of extraction of medication orders from clinical narrative were publicly available. The opportunity to evaluate our named entity extraction methods and to contribute to development of annotated publicly available large collection of clinical notes presented itself with the third i2b2 (Informatics for Integrating Biology and the Bedside) Medical Extraction Challenge [9].

To date, most algorithms and systems for extraction of drug order elements are knowledge-based. In fact, the absence of any large annotated collection makes it difficult to employ supervised machine learning. In contrast the availability of nomenclatures such as RxNorm [10] (which contains drug names, ingredients, strengths, and forms) encourages the use of rule-based systems. For example, Evans et al. developed a set of about 50 rules encoded as regular expressions to identify drug dosage objects and their attributes [11]. A Natural Language Processing (NLP) system augmented with the above rules and two lexicons (one containing drug names extracted from the Unified Medical Language System® (UMLS®) [12] and another one containing unusual words and abbreviations found in drug dosage phrases) identified about 80% of drug dosage expressions. Gold et al. expanded Evans' definition of drug dosage and implemented a system (the MERKI parser) that uses an RxNorm-based lexicon to extract known drug names and contextual clues to extract out-of-vocabulary drug names. Xu et al. developed an approach that attempts to extract a formal medication model (consisting of the drug name, signature modifiers and temporal modifiers) from clinical text using a chart parser and a semantic grammar, and backs off to regular expressions if the chart parser fails [13].

The U.S. National Library of Medicine (NLM) tool (referred to as NLM's i2b2 Challenge Tool or simply, the tool) developed to extract all fields originally defined in the i2b2 medication extraction guidelines is also knowledge-based and relies on lexical-semantic processing and pattern matching similar to the above systems. Our approach differs from the previously explored ones in that we 1) expanded a large number of term lists obtained for each element of drug phrases generating potential spelling variants and mining the UMLS for related terms as well as using corpus-base expansion, 2) developed a module for identification of negated drug mentions, 3) applied a UMLS-based approach to identification of reasons for medication orders, and 4) developed a module for validating drug and reason combinations.

II. METHODS

Early in the planning phase for this Challenge, the decision was made to use simple rules and lookup lists of various entities due to the time constraints of the Challenge. Our processing of the discharge summaries for this Challenge was relatively straightforward and is depicted in Figure 1. This section follows the course of our processing efforts beginning with a description of the lookup lists developed for the Challenge and followed by a description of the various result files generated by the tool.

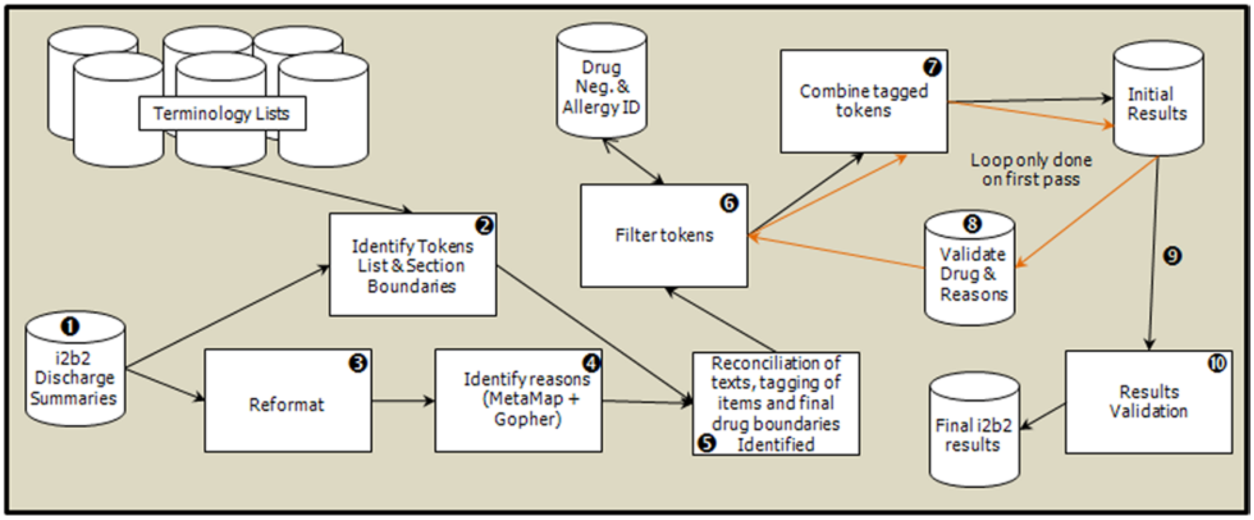


Figure 1 - Processing Flow Diagram

II.0. **Developed lookup lists required for the Challenge.**

The discovery of coverage gaps in our terminology resources (e.g., short forms of drug names such as *aspart* are not always covered in the UMLS, although the long form, *insulin aspart*, maps to two concepts) led to the decision to augment our initial resources with lookup lists. The lists that we developed used existing, publicly available resources with some minor manual curation based on processing the training set and reviewing what was missed by NLM's i2b2 Challenge Tool described here. Although many of the resources have items in common, each of the resources was added for specific reasons. Figure 2 graphically depicts the data sources with arrows connecting the entities and the lists where they made contributions.

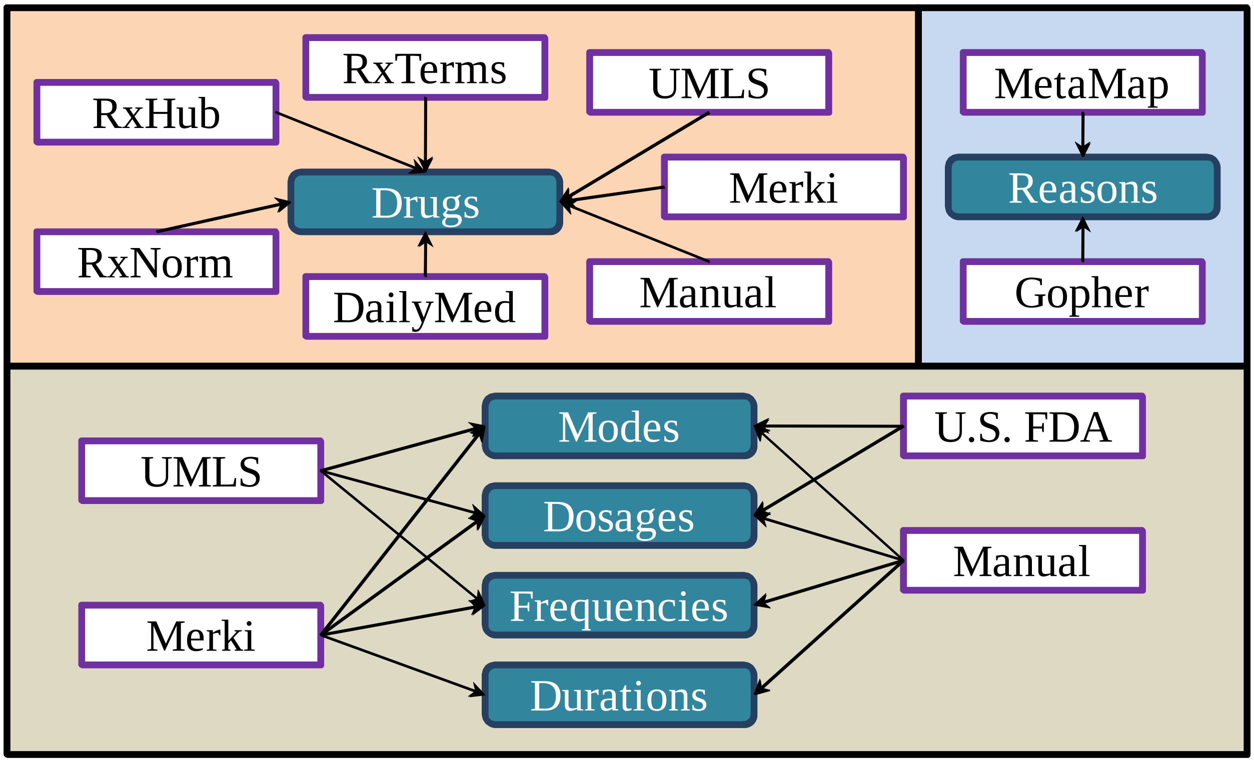
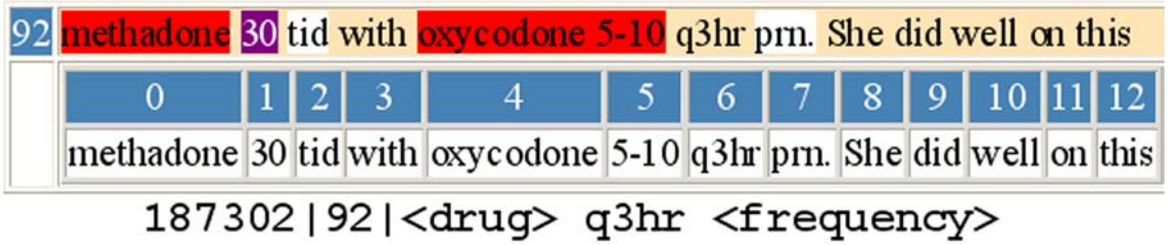


Figure 2 - Lookup Lists and Their Sources

The drug identification list was created using DailyMed [14] for a list of common prescription drug names. We then added display names from RxTerms [15], Ingredients and Brand Names from RxNorm, and a list of drugs from MERKI for a comprehensive list of drugs and their component ingredient names. In an attempt to complement the list of drugs we already had, we started looking at pharmacologic classes (e.g., diuretics), as opposed to drug names and added about 5,000 names from 1,360 UMLS concepts. We first tried to use the whole UMLS Metathesaurus as a source of drug information, but it was too noisy, even after filtering out the drug names from RxNorm. We then selected three of the UMLS source vocabularies in which large numbers of drug classes are listed: SNOMED CT (*SNOMED Clinical Terms, 2008_07_31*), MSH (*Medical Subject Headings, 2009_2008_08_06*) and NDF-RT (*National Drug File - Reference Terminology Public Inferred Edition, 2008_03_11*). Starting from the high-level concepts for pharmaceutical preparations, we extracted the list of all descendants in the source, using specific filters for restricting it to drug classes. We then mapped these terms back to the UMLS and added the synonyms from all sources in the UMLS for these concepts. Finally, we filtered out the less useful synonyms (from an NLP perspective) after manual inspection of the lists. We also added a list of drug classes from MERKI. RxHub [16], which is derived from drug names obtained from de-identified patient medication records, provided us with a list of common drug name misspellings. The U.S. Food and Drug Administration (FDA) Structured Product Labeling web site [17] provided us with extensive lists of Dosage Forms (dosages) and Routes of Administration (modes). Specific processing of the UMLS using a similar method to what was done for the drug classes outlined above was used to identify items for the frequency, modes, and dosages lists. For example, the frequency term *q.a.d.* was extracted from list of descendants of the UMLS concept "Schedule Frequency" (C1882978). Additional lists from MERKI also provided information for the dosage, modes, frequencies, and duration lists.

Finally, manual curation was done to extend all of the lists based on reviews of the tool results for the training collection. For this last step, we specifically looked at the “missed” or unused tokens for each of the lines and assigned the text to the lists as appropriate. For example, *methicillin* was added as a drug name when it was found within the previously annotated drug, *methicillin sodium* occurring in the Ingredients and Brand Names list from RxNorm. The example in Figure 3 from line 92 in summary 187302 of the i2b2 training set shows a spelling variation, *q3hr*, for a frequency term. We had *q3h* in our frequency list, but added *q3hr* because of this example. Figure 3 shows additional information generated by the tool, please see section *II.9.* for details.



92 methadone 30 tid with oxycodone 5-10 q3hr prn. She did well on this

0	1	2	3	4	5	6	7	8	9	10	11	12
methadone	30	tid	with	oxycodone	5-10	q3hr	prn.	She	did	well	on	this

187302|92|<drug> q3hr <frequency>

Figure 3 - Missed Evaluation Example

II.1 & 2. **The discharge summaries were read into the program and tokenized.**
Each line was tokenized using white-space as the token boundary. The section, list, and sentence boundaries were then identified. List boundaries were simply identified by which sections corresponded to the Challenge list of valid “list” sections. Note, that a comma separated list of drugs in a Challenge defined narrative section was not considered a list for purposes of this Challenge; it was purely the section that was determinative. Sentence boundaries were identified using the simple rule of finding a “period” followed by spacing as long as the previous character wasn’t a number. Sentence boundaries helped to define the extent of both drugs and reasons. Section identification was most crucial to this Challenge for several reasons: it 1) allowed us to decide if we wanted to process specific sections or ignore them, 2) assisted in limiting the scope of drugs and reasons, 3) was instrumental in determining whether a drug was in a “list” or “narrative”, and 4) helped eliminate some ambiguity (e.g., not identifying drugs within Allergy sections). Candidate section names were defined as all strings occurring at the beginning of a line, consisting of uppercase letters only (a mixed case review was attempted, but found to be too noisy), and followed by a period, a colon, or the end of the line. We identified 10,454 such potential section names, 937 of them unique. The list of unique names was then manually reviewed, scrubbed to remove ones that either were clearly not section headings or were just not appropriate for this Challenge e.g., “HOLD IF”, “NECK” and “PULM”, and some mixed case section names were manually added to the list e.g., “Attending”, “Alert overridden”, and “Discharge Date”.

Manual annotation exercises were undertaken early on in the Challenge both to familiarize participants with the annotation rules and to create a gold standard for evaluation purposes. Our team members noted sections where we were erroneously identifying drugs and reasons in violation of the Challenge criterion of only identifying drugs “used, to be used, or being used by the patient”. We consequently created a list of twenty-one triggers (see Table 1) that denoted sections we could ignore. This helped to eliminate a large number of false positives without losing any true positives. Some of the non-obvious fields such as “Attending” had personal names that would trigger reasons because of their link to a given disease. For example, the first name *Marion* generated *Female prostatic obstruction syndrome* (*C0268867*) which is the UMLS preferred term for *Marion's Syndrome*.

lab	laboratory	laboratories	allergies	allergy
attending	fam hx	family history	family hsitory	discharge date
service	labs	escription document	dictated by	entered by
vital sign	vitals	signs	vital signs	diet

Table 1: List of Trigger Phrases for Sections to be Ignored

The final list of section names extracted from the training set consisted of 632 items such as "SOCIAL HISTORY", "OPERATIONS AND PROCEDURES" and "Attending". We arranged the list longest to shortest in order and all lowercased to accurately identify full section names within the summaries using a case insensitive matching algorithm.

II.3. **Text was reformatted into a single text line.**

Early testing showed that by simply processing the summaries line by line, we ended up missing some drugs and reasons because the text was broken across lines. So, once the sections were identified and the ones we wanted to use were selected, we combined all of the text into a single line for processing. We created a mapping from this reformatted text back into the original text so we had access to the original tagging and positional information for later use.

II.4. **Reasons were identified using MetaMap and exact matches from the Gopher list.**

We used both MetaMap [18] and a list derived from the Gopher [19] project to identify reasons. In this Challenge, the discharge summaries sometimes had misspellings, acronyms/abbreviations, and different ways of stating a medical reason for prescribing a drug. While MetaMap was able to identify most of the spelling variations and any text inversions, it was limited to the contents of the UMLS Metathesaurus. The Gopher lookup list was introduced to expand our coverage and to assist with less well-behaved occurrences. In the end, the two approaches seemed to complement each other fairly well. We also maintained a "bad reason" list to eliminate as many false positives as possible (see section *II.6*).

MetaMap is a widely available program providing access to the concepts in the UMLS Metathesaurus from biomedical text. We did explore broad uses for MetaMap at the beginning of the Challenge, but were not encouraged by the results. MetaMap was used solely to identify reasons why the patient was prescribed a given drug. We did this by restricting MetaMap output to the thirteen Semantic Types (ST) shown in Table 2. We decided to use the twelve STs in the *Disorders* Semantic Group [20] plus *Clinical Attribute* because of the nature of the data.

As has been noted by others [21], we also found the ST *Finding* to be problematic in that it provided many false positive reasons. We removed the ST from the MetaMap processing but found we lost a number of good reasons as a result. So we decided to overproduce reasons by restoring MetaMap's *Finding* results. We attempted to correct for this overproduction by creating a curated list of "bad reasons" subsequently filtered out of the results.

Semantic Type	Abbreviation
Acquired Abnormality	acab
Anatomical Abnormality	anab
Cell or Molecular Dysfunction	comd
Clinical Attribute	clna
Congenital Abnormality	cgab
Disease or Syndrome	dsyn
Experimental Model of Disease	emod
Finding	fndg
Injury or Poisoning	inpo
Mental or Behavioral Dysfunction	mobd
Neoplastic Process	neop
Pathologic Function	patf
Sign or Symptom	sosy

Table 2: Semantic Types used by MetaMap for Identifying Reasons

The Gopher list used in our system was obtained from the Regenstrief Institute for Health Care and the Department of Medicine Gopher order entry system, which originated as *The Medical Gopher*, the first PC-based order entry system developed for outpatient care. The entries on the Gopher list are "menu items" in the order entry system ("answers"). The items consist of names, aliases and synonyms for diagnoses, procedures, tests and drugs. We extracted the diagnoses names and synonyms from the Gopher system specifically for assisting in identifying reasons in the summaries.

II.5. **Reasons were then reconciled with the original text and tagged using the mapping information from the single free-text line back to the original discharge summary.**

We used exact text matches to the lookup lists to tag drugs, modes, dosages, durations, and frequencies. Drug boundaries were also identified by noting the first position of each drug so we could know when we came to the end of the current drug during filtering. Drug boundaries expanded left and right depending on where the components were identified with the final drug boundary encompassing the drug name and any of its associated components.

II.6. **Filtering was performed to add, remove, and extend tagged items.**

Filtering involved simple rules, a "bad reason" trigger list (e.g., "*ruled out for"*), and a "bad drugs" list for what should be removed (e.g., *insulin* within *insulin-dependent diabetes*). We developed rules for limiting the scope of a drug to try and eliminate the crossover of components, we also developed a program to identify non-active medications (e.g., *should not take aspirin*) and allergy-specific drugs to remove false positives, and one of the final filtering steps was to see if any of the drug components needed to be combined or extended.

Eliminating false-positive drug names

We began with the manually-created "bad" drugs list to remove false positive drug names like *insulin* within *insulin-dependent diabetes*. Generalizing the lookup list approach, simple rules were developed for the removal of false positive drugs. For example, one of the rules states that if certain words follow a drug name, remove that occurrence from consideration. The rule prevents expressions such as *<drug> level*, *<drug> measure,* and *<drug> screening* from being included in the results.

Eliminating non-active medications

The next step in the filtering process was to identify drugs that should be excluded according to the Challenge guidelines because they either were negated or occurred within the context of a patient's allergies. A negation and allergy identification program was used to identify negated drug terms and also phrases indicating a patient's allergies to drugs. The program looked for certain negation keywords such as *not, no, avoid* and *never* to discover negated phrases containing drugs which should not be marked as medications. For example, the phrase *should not take aspirin* was identified by the program as a negated phrase and *aspirin* was consequently identified as a negated drug. The program checked to see if the negation phrase contained words which indicated a medication was actually given to the patient. For example, the phrase *did not take his Coumadin* indicates the patient was prescribed/previously taking Coumadin because of the word *his*. So Coumadin in this case is not identified as a negated drug. Similarly, certain phrases related to allergies and containing a drug name were identified. Negated drug names were removed prior to the attempt to combine drugs and reasons together so as to not remove any reasons for potential use with other nearby drug names.

Eliminating false reasons

We also had a short list of triggering phrases for removing reasons because they were simply inappropriate for this Challenge. We removed reasons if they were preceded by *no evidence of*, *no history of*, *secondary to*, etc. So if the phrase *secondary to <reason>* was found, we would remove it from the list of possible reasons to use for a drug.

Identifying, combining, and extending components

A review of the training summaries showed that drug signature components (modes, dosages, durations, and frequencies) were generally mentioned after the drug name itself; therefore, we always looked to the right first when trying to identify such components. In order to eliminate one drug combining with components from another drug, the search area for components of a given drug had to start within a ten token window, within two lines either before or after the drug name, and could not go beyond drug, list, or section boundaries. The instances of *Sliding Scale* were an exception to these rules as they would occasionally cross many lines. A good example of this is discharge summary 983233 where lines 33 through 41 are all the sliding scale information for the drug found on line 32 - *REG INSULIN (HUMAN) (INSULIN REGULAR HUMAN)* (see Table 3).

```
32. REG INSULIN ( HUMAN ) ( INSULIN REGULAR HUMAN )
33. Sliding Scale ( subcutaneously ) SC AC+HS
34. Starting Today May
35. If BS is less than 125 , then give 0 units subcutaneously
36. If BS is 125-150 , then give 0 units subcutaneously
37. If BS is 151-200 , then give 2 units subcutaneously
38. If BS is 201-250 , then give 4 units subcutaneously
39. If BS is 251-300 , then give 6 units subcutaneously
40. If BS is 301-350 , then give 8 units subcutaneously
41. If BS is 351-400 , then give 10 units subcutaneously and
```

Table 3: Sliding Scale Example from Discharge Summary 983233

Simple rules for expanding components by looking at the tokens to the left and right of the component were developed as needed. For example, *<drug> <number>* was combined (e.g., *Tylenol #3*); and *<number>* followed by either *<dosage>*, *<duration>*, *<mode>*, or *<frequency>* were combined as long as the *<number>* had not already been assigned to expanding the drug name. Some rules moved drug boundaries to better assign components to the correct drug name. For example, the pattern *<component> of <drug>* triggered the extension of the drug boundary to include the *<component>* part. So, consider the text *After 2 doses of dofetilide* (summary 1109, line 73) where *2 doses* was the *<component>* and *dofetilide* the *<drug>* name. To ensure that *2 doses* was combined appropriately with *dofetilide*, we extended the drug boundary from the *d* in *dofetilide*, to the *2*. Without this drug boundary extension rule, it is possible we would have either missed elements like this, or assigned them to a previously located drug in the same section of text. Our base rule was to look right first and then left if we were missing any elements, so when related elements were actually located to the left of the drug name, we had to find a way to combine them properly. This rule allowed us to do just that, and was enforced at a higher priority than the look right rule.

II.7. **Drug/reason pairings identified.**

Once drugs and their components and reasons had been initially identified, we attempted to match each drug name with a nearby reason. Initially we had a very simple rule which was to use the closest reason if there were two possibilities. This was refined to ensure that reason assignment did not violate a drug, list, or section boundary. We also instituted a small set of trigger phrases to identify when we should combine certain nearby reasons and drugs (see Figure 4). For example, if we found *<drug> PRN <reason>*, we knew that the <reason> should be connected to the preceding <drug>. PRN is an abbreviation for the Latin phrase *Pro Re Nata* which is commonly used in medicine to mean “as needed” referring to a drug dosage administered normally by the patient or caregiver. The syntax we identified in our rule is commonly used in clinical text. These specific groupings were identified prior to applying the “nearest reason” approach, and *PRN <reason>* took priority over all of the simple rules. To impose limitations on reason assignments, we decided that, in general, each reason could only be assigned to a single drug name, and each drug name could only have a single reason. In some cases, we expanded the number of reasons to two if they were next to each other and connected with a comma, the word *and*, or the word *or*. Thus, the pattern *<drug> for <reason> and <reason>* allowed the assignment of the two reasons to the drug name.

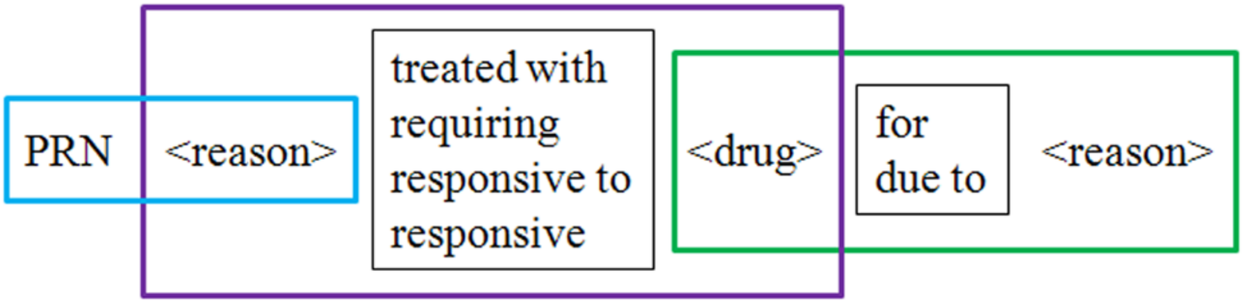


Figure 4 - Simple Reason Grouping Rules

II.8. **Drug/reason pairings validation.**

Once drug/reason pairings were identified, we attempted to validate the pairings via knowledge contained in the UMLS. The validation of the drug/reason pairings were accomplished via a constrained traversal of the UMLS relations involving two main steps as described below.

Drugs and reasons were first mapped to UMLS concepts, using exact and normalized matches, and further restricting mappings to the semantic group *Chemicals & Drugs* and *Disorders*, respectively. All successful mappings were considered, including several pairs of UMLS concepts generated by one original drug/reason pair.

Selected UMLS relations were then used to identify plausible relations between drugs and reasons. The key relations were provided by the NDF-RT source vocabulary where ingredients are associated with diseases through *may_treat* and *may_prevent* relationships. More precisely, the following relations were used:

1. branded drug names were mapped to generic drug names (ingredients)
   For example, *Lasix* → *Furosemide*
2. the ingredient and all its ancestors were explored as potential entry points to a DRUG {*may_treat*|*may_prevent*} DISEASE relation
   For example, *Furosemide* → *Hypertensive disease*
3. the reason was checked against the DISEASE [*treated/prevented* by the DRUG] and its descendants, for example, *cardiac arrhythmia* → *tachycardia*
4. the reason was also checked against the manifestations of the DISEASE and its descendants (in most cases, the relation between a disease and its manifestations is not explicitly stated in the UMLS source vocabularies; we used all associative relations [REL = RO] between concepts of the semantic group *Disorders* as a surrogate).
   For example, *Hypertensive disease* → *blood pressure*

The algorithm did not explore all paths, but rather stopped at the first path reached between the drug and the reason. Examples of drug/reason associations identified by the UMLS-based algorithm include the following.

- *albuterol* / *asthma* through a direct link between ingredient and disease
- *albuterol* / *wheezing* through a link between an ingredient and the manifestation of a disease (asthma)

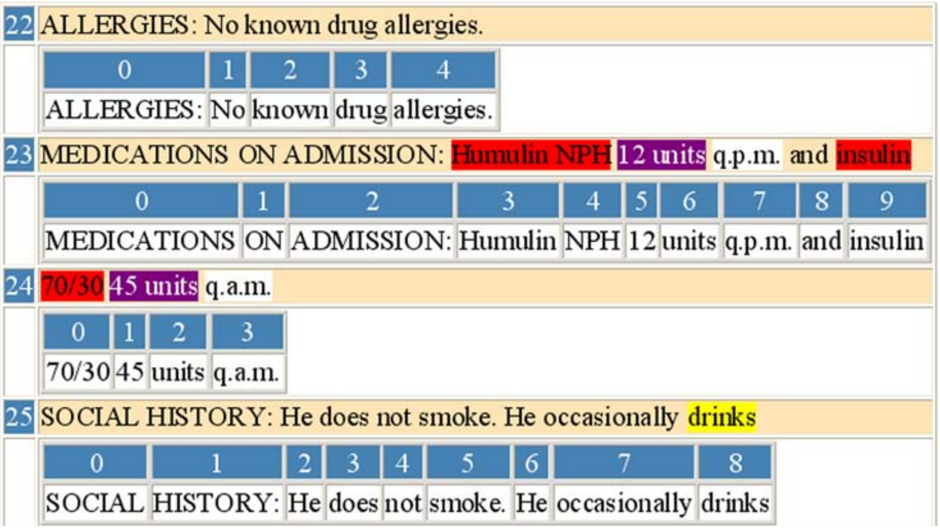
- *aspirin* / *substernal chest pain* through a link between an ingredient and the descendant of a disease associated with the drug (pain)
- *enteric-coated aspirin* / *pain* through a link between an ancestor (*aspirin*) of the ingredient and the disease (*pain*)

Some drug/reason associations involve more complex paths through the UMLS. For example, the association *Lantus / Diabetes* was identified from a path through the ingredient (*Insulin Glargine*) of which *Lantus* is a tradename, the disease treated by this ingredient (*Diabetes Mellitus, Non-Insulin-Dependent*), a descendant of this disease (*Diabetes mellitus with no mention of complication*), and finally a condition (*Diabetes*) that stands in an associative relation with this descendant. 9,415 possible drug/reason pairings were found with 2,785 of these having at least one path through the UMLS tying them together.

II.9. **Generate result files.**

NLM's i2b2 Challenge Tool produced several files as a result of the processing we performed on each of the discharge summaries. Most of these files related to debugging and manual review of the results obtained on the training set, with the main result file being the drug list formatted for the i2b2 Challenge submission. Each of the files is described here:

- File with the i2b2 formatted results (*<summary number>.i2b2.entries*) contains one or more lines for each drug found with one file for each summary. This file contains seven fields separated by "||" for each entry. All seven of the elements must be present for each line and if there is no information for a given element, *nm* is used as filler. If a drug was assigned more than one reason, dosage, frequency, etc., a new line was added to the file for each occurrence tied back to the originating drug.
- HTML-formatted informational file (*<summary number>.html*) showing each line of the summary, the tokenization for each line, and the text colorized to highlight drug names, modes, durations, dosages, frequencies, and reasons. See Figure 5, which shows the Tool results for lines 22 through 25 of discharge summary number 23538. The figure shows the identification of a drug *Humulin NPH*, a dosage *12 units*, a frequency *q.p.m.*, a drug *insulin 70/30*, a dosage *45 units* and a frequency *q.a.m.* Note that in this file, no effort is made to connect these components; they are simply highlighted for easy identification.



22 ALLERGIES: No known drug allergies.

0	1	2	3	4
ALLERGIES:	No	known	drug	allergies.

23 MEDICATIONS ON ADMISSION: Humulin NPH 12 units q.p.m. and insulin

0	1	2	3	4	5	6	7	8	9
MEDICATIONS	ON	ADMISSION:	Humulin	NPH	12	units	q.p.m.	and	insulin

24 70/30 45 units q.a.m.

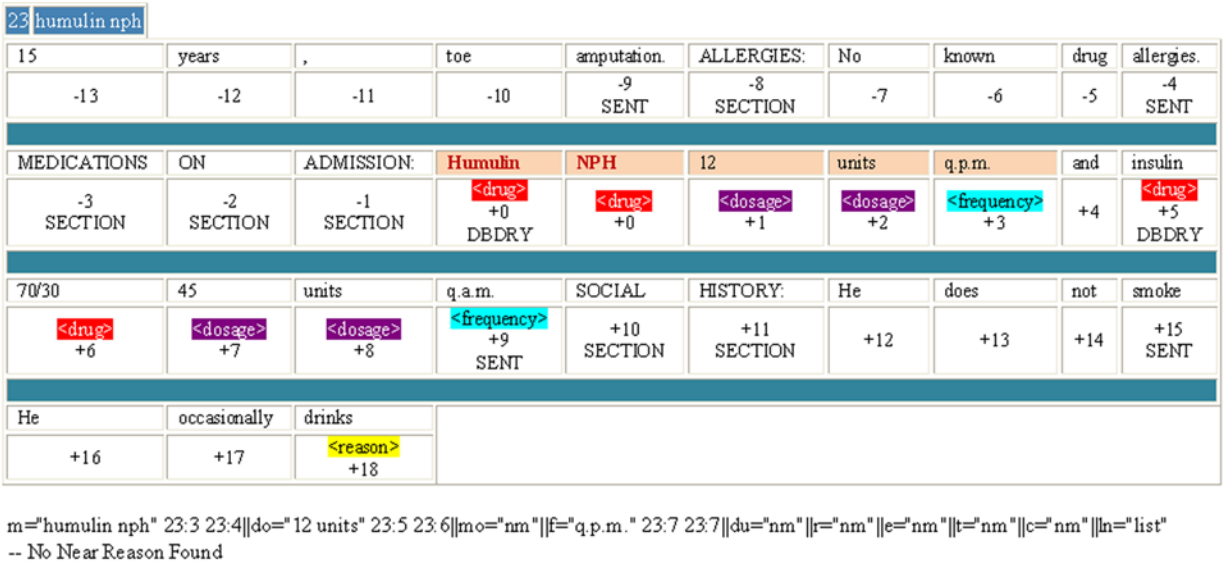
0	1	2	3
70/30	45	units	q.a.m.

25 SOCIAL HISTORY: He does not smoke. He occasionally drinks

0	1	2	3	4	5	6	7	8
SOCIAL	HISTORY:	He	does	not	smoke.	He	occasionally	drinks

Figure 5 - NLM Tool Information Example (23538)

- HTML-formatted details file (*<summary number>_table.html*) showing the results and context for each of the drug names found in the summary. For each of the drug names, we provided the drug name in context (the line of text containing the drug name, the line above and the line below); highlighted all of the information that has been combined for the given drug; and extracted information as an i2b2 result entry for the drug name. Figure 6 shows these results for *Humulin NPH*, the drug of focus in Figure 5. This detailed view shows identification of various boundaries: SECTION (section name), DBDRY (drug boundary), and SENT (sentence boundary). The numbers at the bottom of each token represent a token count from the last drug name identified in the text. So, "-11" means the token is 11 positions to the left of the drug name and "+3" means the token is three positions to the right of the end of the drug name. The i2b2 result lines as well as reason details for all drugs are displayed below each table.



23 humulin nph

15	years	,	toe	amputation.	ALLERGIES:	No	known	drug	allergies.
-13	-12	-11	-10	-9 SENT	-8 SECTION	-7	-6	-5	-4 SENT
MEDICATIONS	ON	ADMISSION:	Humulin	NPH	12	units	q.p.m.	and	insulin
-3 SECTION	-2 SECTION	-1 SECTION	<drug> +0 DBDRY	<drug> +0	<dosage> +1	<dosage> +2	<frequency> +3	+4	<drug> +5 DBDRY
70/30	45	units	q.a.m.	SOCIAL	HISTORY:	He	does	not	smoke
<drug> +6	<dosage> +7	<dosage> +8	<frequency> +9 SENT	+10 SECTION	+11 SECTION	+12	+13	+14	+15 SENT
He	occasionally	drinks							
+16	+17	<reason> +18							

```
m="humulin nph" 23:3 23:4||do="12 units" 23:5 23:6||mo="nm"||f="q.p.m." 23:7 23:7||du="nm"||r="nm"||e="nm"||t="nm"||c="nm"||ln="list"
-- No Near Reason Found
```

Figure 6 - NLM Tool Details Example (23538)

- File showing the drug/reason pairings (*<summary number>.reasons*) for use in the drug/reason validation program with one file for each summary.
- File showing all partially tagged instances and surrounding untagged text in its context (*<summary number>.missed*). For example, from summary 23538, line 23 *23538|23|<frequency> and <drug>*. This corresponds to the "+4" token just to the right of the drug *Humulin NPH* and right before the drug *insulin 70/30* in Figure 6. This file was manually reviewed for identifying missed dosages, durations, modes, and frequencies due to spelling differences (e.g., *q.p.m.*, *q.p.m*, *qpm*, *qp.m.*, etc.).

II.10. **Final validation of the results was done to ensure syntactic correctness and compliance with the Challenge requirements.**

III. RESULTS

We finished fourth overall out of 20 teams that participated in the Challenge. Since two of the three teams who scored best had pre-existing systems that were modified for the Challenge, we were pleased that a system developed expressly for the Challenge performed so well. The lessons learned during this effort are being evaluated for inclusion in our NLP tool suite. It is interesting to note that two of the three teams ahead of us also used a rule-based approach (Vanderbilt and University of Manchester), while the third and number one team, University of Sydney, used a supervised learning approach. Results are shown in Table 3 and Table 7 in the i2b2 JAMIA overview paper [22]. It is clear from Table 7 that all teams had significant problems with identifying both Durations and Reasons.

IV. DISCUSSION

In general we are satisfied with our vocabulary and rule-based identification of drug names, doses, modes and frequencies. The lack of significant difference between our exact and inexact scores confirms this view since it shows that we either found the entire element or missed it completely. Our dose and duration results are satisfactory, considering they are based on very simple heuristics. However, the approach is brittle in the presence of pattern changes in the middle of an enumeration of drugs. Deeper understanding of the context is needed to overcome this weakness.

Low scores for durations and reasons, on the other hand, show that our methods are clearly insufficient for those drug elements. In the absence of creating a full-fledged natural language understanding system, some improvement might be achieved using corpus-based methods. Any corpus-based methods would need to be judiciously applied given their known weaknesses: they are noisy if not supervised, and they are ambiguous even when supervised. For example, using our corpus-based expansion we identified *HCT* as an abbreviation of *hydrochlorothiazide* (more commonly abbreviated as *HCTZ*); however, *HCT* is also common shorthand for *hematocrit*.

Further manual inspection of 64 instances of the training set containing this term revealed that in all but one instance *HCT* was used in its *hematocrit* sense. The instance that provided the *hydrochlorothiazide* sense, *Continue to take the hctz in the morning. HCT will lower potassium low*, was deemed a typo and *HCT* was removed from the list of drugs.

The algorithm we developed for validating the associations between drugs and reasons from

UMLS relations was useful in the context of this Challenge, but overall it was suboptimal. On the one hand, it produced many false positives (wrong associations, e.g., *Coumadin* → *Nose bleed*, from *Anti-Infective Agents, Local* found in the ancestors of *Coumadin* and in a *treats* relation to *Communicable Diseases*, itself in associative relation with *Nose bleed*), as well as valid associations for which the supporting path was inaccurate. On the other hand, many valid associations failed to be identified, primarily because some supporting relations were missing from the UMLS Metathesaurus (e.g., *Coumadin* → *Thrombus* was missed) and because the algorithm supported limited reasoning through a fixed set of relations. A careful review of false positives and negatives is required for further improvement to the algorithm.

Finally, we intend to incorporate some of our tool's features into the MetaMap algorithm. Specifically, we will include the overall identification of drug mentions with the expectation that it will reduce ambiguity because of the coordination of a drug's elements. In addition, augmenting MetaMap's negation algorithm with the drug-specific negation detection developed for the Challenge should be useful in applying it to clinical text.

V. LIMITATIONS

Many of the limitations of this research occurred because we are reporting on the development of an NLP application in the context of a time-sensitive Challenge rather than fundamental research. In-depth analysis that we would normally have done will be done in the future. Examples of such analysis include determining the relative contributions to our results from the many knowledge sources we used, a similar analysis of the contributions of the filtering rules, and a study to determine an optimal balance between the knowledge sources and the rules. In addition, the relations identified between drugs and diseases from selected UMLS relations are not intended to be used as a reference set of relations reflecting therapeutic intent. Rather, we use constraints on the UMLS graph of relations in order to identify plausible drug-reason relations for the purpose of validating drug-reason pairings. Despite the presence of many false positives and false negatives, our algorithm proved useful in the context of this Challenge.

Acknowledgements

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. *AMIA Annu Symp Proc.* 2008 Nov 6:237-41.
2. Cimino JJ, Bright TJ, Li J. Medication reconciliation using natural language processing and controlled terminologies. *Stud Health Technol Inform.* 2007;129(Pt 1):679-83.
3. Carol Friedman. Discovering Novel Adverse Drug Events Using Natural Language Processing and Mining of the Electronic Health Record. *Artificial Intelligence in Medicine, 12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona, Italy, July 18-22, 2009*. Proceedings 2009.
4. Turchin A, Wheeler HI, Labreche M, Chu JT, Pendergrass ML, Einbinder JS. Identification of documented medication non-adherence in physician notes. *AMIA Annu Symp Proc.* 2008 Nov 6:732-6.
5. Sirohi E, Peissig P. Study of Effect of Drug Lexicons on Medication Extraction from Electronic Medical Records. *Pac Symp Biocomput.* 2005:308-18.
6. Demner-Fushman D, Seckman C, Fisher C, Hauser SE, Clayton J, Thoma GR. A Prototype System to Support Evidence-based Practice. *AMIA Annu Symp Proc.* 2008 Nov 6:151-5.
7. Breydo EM, Chu JT, Turchin A. Identification of inactive medications in narrative medical text. *AMIA Annu Symp Proc.* 2008 Nov 6:66-70.
8. Aronson AR, Mork JG, Névéol A, Shooshan SE, Demner-Fushman D. Methodology for Creating UMLS Content Views Appropriate for Biomedical Natural Language Processing. *AMIA Annu Symp Proc.* 2008 Nov 6:21-5.
9. Informatics for Integrating Biology and the Bedside, i2b2, a National Center for Biomedical Computing, Third Shared-Task Challenge in Natural Language Processing for Clinical Data Medication Extraction Challenge: https://www.i2b2.org/NLP/Medication.
10. RxNorm: http://www.nlm.nih.gov/research/umls/rxnorm.
11. Evans DA, Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Proc AMIA Annu Fall Symp.* 1996:388–92.
12. UMLS Knowledge Sources: http://www.nlm.nih.gov/research/umls.
13. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc., 2010 Jan–Feb;17(1):19-24.*
14. DailyMed: http://dailymed.nlm.nih.gov.
15. RxTerm: http://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm.
16. Fung KW, Applied Medical Terminology Research. A Report to the Board of Scientific Counselors. The RxHub Project – a sneak preview. April 2009. Page 32. http://www.lhncbc.nlm.nih.gov/lhc/docs/reports/2009/tr2009001.pdf#page=33.
17. FDA Structured Product Labeling Resources: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling

18. Aronson AR, Lang FM. The Evolution of MetaMap, a Concept Search Program for Biomedical Text. *AMIA Annu Symp Proc.* 2009 Nov:22.

19. McDonald CJ, Tierney WM: The medical gopher-A microcomputer system to help find, organize and decide about patient data. *West J Med* 1986 Dec; 145(6):823-9.

20. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Stud Health Technol Inform.* 2001;84(Pt 1):216-20.

21. Névéol A, Kim W, Wilbur WJ, Lu Z. Exploring Two Biomedical Text Genres for Disease Recognition. *NAACL 2009*, Workshop BioNLP.

22. Ozlem Uzuner, Imre Solti, Eithon Cadag. Extracting Medication Information from Clinical Text. *J Am Med Inform Assoc., 2010 (to appear).*