# Lessons Extracting Diseases from Discharge Summaries

**William Long, PhD**

**CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA**

**Abstract**

*We developed a program to extract diseases and procedures from discharge summaries and have applied this program to 96 cases annotated by physicians. We compared the concepts extracted by the program to those extracted by the annotators. The program extracts 93% of the desired concepts including some more specific than the annotators. Concepts were missed because phrases were ambiguous, phrases were missing words or were separated, or deduction was needed, among other reasons. The false positives were either insignificant findings, ambiguous phrases, or did not apply to the patient now. The analysis shows that extraction of medical concepts from discharge summaries with limited natural language processing and no domain inference is effective with still more potential.*

## Introduction

Vast amounts of information about patients is in the form of human generated text. Some is in difficult-to-parse doctor and nursing notes, but much is in carefully written documents such as discharge summaries. Our research is addressed at extracting and coding concepts from these discharge summaries.

A number of efforts have focused on extracting useful data from discharge summaries. Friedman's group at Columbia uses a detailed natural language processing approach with their MedLEE parser to interpret discharge summaries and to extract diseases and other information[1,2,3]. Another effort used triggering words to look for adverse events, with less success because of the variety of ways these events may be expressed[4]. Another strongly linguistic approach is exemplified by the MedsynDikate approach in Germany[5].

The alternative to the strong parsing approach is a dictionary based approach and match phrases in the text to phrases in the dictionary, minimumizing parsing. Fortunately, the UMLS provides a compilation of a large number of medical dictionaries, making such an approach feasible. With the addition of SNOMED-CT to the UMLS, there is great incentive to use this resource to code the diseases into this widely recognized vocabulary[6]. MetaMap is a tool available from the NLM for coding phrases using the UMLS, exemplifying the dictionary approach [7]. It has been used in several studies. Meystre and Haug used MetaMap to extract 80 different problems from records with a recall of 0.74 which was improved to 0.896 when the dictionary was customized[8]. Chapman et al, used it to extract respiratory findings from emergency dept. reports with a recall of 0.72[9].

## Methods

We developed a program using the UMLS to extract diagnoses and procedures from discharge summaries (DSs). It uses a very limited amount of natural language processing. Rather, it uses the structure of the DS, a small list of words and punctuation to divide the text into phrases, and does a maximal substring search using the normalized string table of the UMLS to find the best coding.

The program was quite effective on a small test set (described previously at AMIA) finding 240 of 250 desired concepts with 19 false positives[10]. We also tried MetaMap on the same test set, which missed 31 desired concepts and had 23 false positives or a recall of 0.876 – as good as reported elsewhere, but not as good as the 0.96 for our program.

We are now using the program to produce a list of concepts from DSs which are then used by physician annotators to speed the process of generating disease and procedure lists for ICU cases (problem lists). The task of the annotators is to include everything that might be useful in characterizing the case for retrieval and analysis. The whole annotation process is very time consuming so a limited number of cases have been annotated. Fortunately, there are a few that were done by more than one annotator.

The objective is to provide as many of the concepts needed for the problem lists as possible. It is simple to discard unneeded concepts but time consuming to code new ones, so we maximize the sensitivity of the program. The first step in the original program was to identify the sections of the DS with diseases and procedures (e.g., "Past Medical History") and restrict search to those sections. However, many DSs had diagnoses and procedures scattered throughout the "Hospital Course" section and often sections are identified by system, making it hard to identify all of the appropriate sections. Thus, we changed the program to extract concepts from all parts of the DS.

The program searches for concept phrases by dividing the text into phrases using punctuation (",", ":", …) and conjunctions ("and", "or", …) and some verbs ("is", "feels", …) and then looking for the longest subphrase that matches a UMLS concept using the normalized string table. To increase sensitivity we eliminated the prepositions that were also used to divide the text originally.

The UMLS is an evolving repository with significant additions made between the time the program was first developed and the version used in this research. The phrases used for matching come from the multiple source dictionaries. This increases the likelihood of matching one of the ways a concept may be phrased but it also means that the coverage has some randomness and some phrases are attached to unexpected concepts, as will be seen.

The annotators also have available all the monitoring data and nursing notes besides the discharge summary. They started before the program was completed so did not have the program's concept list for some cases and for the rest had the results of an earlier version. Thus, we are focusing on comparing the concepts produced by the current program with maximum sensitivity to the concepts in the annotator's problem lists.

**Results**

When this comparison was made, the four clinicians had annotated 96 cases. Five of the cases were annotated by three of the clinicians, 17 by two, and the rest by only one. The results are shown in Table 1.

| Clinician concept | Count | Acceptable |
|---|---|---|
| Exact match | 1184 | 1184 |
| Equivalent concept | 19 | 19 |
| More specific concept | 27 | 27 |
| Less specific concept | 15 | 0 |
| Concept not in DS | 78 | remove |
| Concept missed | 81 | 0 |
| Total | 1404 | 1326 |

Table 1: Clinician concepts found by program

The total number of concepts the annotators put in problem lists was 1404, but 78 of these are not in the DSs so must have come from other information about the case. If we accept concepts found by the program that were exact matches, equivalent, or more specific than those of the clinicians, the program found 1230 of the 1326 it could have found or 93%. Since we favored sensitivity, the false positive rate was quite high with 3435 concepts not used by the annotators.

The 27 "more specific" concepts are more specific according to the hierarchical relations in the UMLS. The "equivalent" concepts are either instances where the annotators chose a concept from a different dictionary, such as death: C0011065 from MESH versus C1306577 from SNOMED. Or they are concepts that for practical purposes are equivalent. For example, there are different concepts for *carcinoma of the breast* and *breast cancer* as well as some other cancers. Some concepts incorporate *history*, so there is *personal history of deep vein thrombosis* (a finding) and *deep vein thrombosis* (a disease). Concepts with modifiers also blurred distinctions, so we considered *lipom*a plus *excision* equivalent to *lipoma* plus *surgery* and *lumbar artery* plus *rupture* equivalent to *lumbar artery* plus *hemorrhage*. These equivalences were determined by consulting with a clinician.

**Analysis**

We have investigated each of the concepts that is not a match in an effort to determine how concept extraction can be improved. The following are the issues found.

**Missed Concepts**

There were a variety of reasons 81 concepts were missed. First, a phrase may match more than one concept. If two concepts had a UMLS hierarchical relationship, the more general one was picked. This is usually a good heuristic but for example, "respiratory failure" matches both *respiratory failure* and the more general *respiratory insufficiency*. We considered selecting the more specific concept but this leads to more errors, for example "diabetes" matches *diabetes type II* (too specific) as well as *diabetes mellitus*, but not *diabetes type I.*

There are also multiple matches of unrelated concepts representing multiple meanings for words like "shock" or "depression". When more than one is a disease or procedure, the only reliable strategy is to include all of them. There are a relatively small number of such words but they occur in many DSs.

The most common reason for missing the concept was was missing words in the phrase, a wrong word, or a disconnected phrase. Words may be left out because they are implied by the context. For example, "left main disease" is coronary artery disease but can not be matched without "coronary artery". Similarly, "non Q wave infarction" needs "myocardial"; "resuscitation" needs "cardiopulmonary"; "intubation" needs "endotrachael"; and "catheterization" needs "cardiac". The phrases in the UMLS completely identify the

concept while the phrases in DSs have context. Some of that context may be very simple. For example, "hypotensive" requires "episode" and "anticoagulation" requires "use".

With concepts that can be stated in many ways, some combinations may be missing from the UMLS. For example, "pericardial window incision" does not match *pericardiostomy* but "operation", "creation", "technic", etc. would. Similarly, "sigmoid resection" does not match *sigmoid colectomy* but "excision" would. Also, "three vessel coronary artery disease" only matches *coronary artery disease* but "triple vessel coronary artery disease" matches the more specific concept.

Phrases may be divided by punctuation as in "mitral valve replacement, bioprosthetic valve" or extra words as in "cultured and had a positive sputum" for *positive culture findings in sputum* and "iliac stent thrombosis" for *thrombosis of iliac artery*. The intervening text may be more extensive as in "source of sepsis was suspected due to his change in abdominal..." to match *postprocedural intra-abdominal sepsis*.

Iincreasing sensitivity by increasing phrase length occasionally introduced errors as well. For example, "acute myocardial infarction with electrocardiogram changes" is coded as *myocardial infarction electrocardiogram* instead of *acute myocardial infarction* because the first covers four words while the second only covers three. "With" is removed by normalization so the meaning shifts. This also caused "two feet of ischemic bowel" to be coded as *ischemic foot* rather than *ischemic bowel*. On the other hand, some more specific concepts include "with" as in "atrial fibrillation with a rapid…" to get *rapid atrial fibrillation* or "chronic renal failure with acute…" to get *acute-on-chronic renal failure*.

The program matched the longest non-overlapping phrases to concepts. Sometimes a lengthy modifier resulted in a less specific primary concept. For example, "greater saphenous vein thrombosis" was coded as *great saphenous vein structure* plus *thrombosis* rather than *vein thrombosis* and "upper extremity deep vein thrombosis" was coded as *structure of deep venous system of upper extremity* plus *thrombosis* rather than *deep venous thrombosis*.

There were also a number of instances where reasoning would have been required to get the concepts assigned by the clinicians. This usually could have been determined by parameter values. For example, getting *diabetes type I* from "diabetes" plus the insulin information or *acute-on-chronic renal failure* from creatinine data.

Abbreviations caused only minor problems. The program found appropriate matches for most and a few, such as "PEG" for *percutaneous endoscopic gastrostomy* have been recently added to the UMLS. Most ambiguous abbreviations such as "MS" are spelled out in the DSs. A couple abbreviations were missed because they are not alone in the string table. For example, "CVVH" is in "CVVH – Continuous venovenous hemofiltration" but not alone so "CVVH" does not match. Partial abbreviations failed to match. So "C. difficile infection" does not match "Clostridium difficile infection".

There were also a few rare problems. To improve performance, the program normalizes and caches words, sorting these to get phrase normalizations. However, we found one place where the program was inconsistent with the UMLS normalization algorithm[11]. UMLS normalization of words with "'s" drop the "s" as well as the apostrophe. Thus, "Bell's" is "bell" but the program just eliminated the punctuation, searching for "bell s", and missed *Bell's palsy*.

The program maps a number of UMLS types into "disease"[12]. This usually works fine. However, a couple concepts used as diseases have UMLS types not associated with disease. *Anxiety* is a "mental activity" and not a symptom or disease. Other "mental activity" concepts include thinking. Also, "dead" matches two concepts, one of which is a "organism function" and the other is an "idea or concept".

There are also a few concepts used by the annotators that are not SNOMED concepts although they are in the UMLS. For example, *shock liver* is in CCPSS as is *septic knee joint*. Also, *candidemia* is in MDR.

**False Positives**

The false positives are relatively easy for the annotators to discard but since there are so many it would be desirable to remove any that are definitely not needed. We examined the false positives and there were three general categories: concepts that did not apply to the patient currently, those that were insignificant relative to the problem list, and mismatched concepts.

Many false positives are contained in statements indicating they are false, possible, potential, or apply to someone else. The following are examples:

- **Negative**: ruled out for Dx, nor any Dx, no significant Dx, initially thought to be Dx, Px would not be indicated, was Sx free, would not be a good Px candidate, rather than Dx, as opposed to Dx, Px was attempted.
- **Negative in practice**: trivial Dx, trace Dx, small Sx, small stable Dx

- **Possible**: might be developing Dx, possibility of Dx, question of Dx, unclear if Dx, unsure if Dx, secondary likely to an Dx, questionable allergy to Rx, Rx for a presumed Dx, could not completely exclude Dx, that would suggest Dx, having Dx-like activity, uninterpretable for Dx
- **Alternative possibilities**: consistent with Dx1 or Dx2; resuscitation such as Px1, Px2, or Px3; in terms of Sx1 or Sx2, Dx1 versus Dx2
- **Potential**: risk of Dx, to prevent Dx, will likely need Px, any potential Dx, changed to Rx at Dx doses, ALLERGIES: To eggs, he gets Sx
- **In past**: Px in [date] for Dx, status post Dx, Px at which time a Sx, husband initiated CPR, old fracture of the right knee
- **Other people**: two brothers with Dx, father died of a heart attack at age 67, FAMILY HISTORY: Negative for CAD
- **Not a person**: Heart failure attending, heart failure service

Many of the negative and possible statements could be handled well by NegEx[13], but the rest require more reasoning.

Most false positives were non-significant findings or procedures such as constipation, difficulties swallowing, drowsy, fever, hypercholesterolemia, intubation, obesity, sinusitis, UTI, yeast, and thrombus. Sometimes the extracted concept is true but insignificant in comparison to a related concept that was missed. For example, "severe RV dysfunction" was extracted as *functional disorder* from the "dysfunction", which while true is not sufficient. In the line "Ischemia: Inferior myocardial infarct" *ischemia* is used as a category and although true, the significant disease is the infarct. This problem also arises when the disease is described as in "coronary artery was totally occluded" which was extracted as *obstruction*.

Over 300 concepts were sometimes included in the problem lists and sometimes not. In addition to those described above that were false positives because they were not true of the patient, there were several reasons for this. Sometimes the concept coded by the annotator was less specific than it could have been. For example, "jejunal feeding tube" was coded as *tube* modified by *jejunal structure*, whereas it could have been coded as *feeding tube* with a modifier (the whole term as such is not in the UMLS) or *nasogastric feeding*. Hemorrhage was usually a false positive because bleeding can happen in the ICU for many reasons and it is only part of the problem list when it is significant and not incidental to the management of the patient. *Obstruction* and *ischemia* with modifiers for the location were coded by the annotators when they could not find a more specific term for "occluded saphenous vein grafts" and "mesenteric ischemia". This situation also arose with embolus, neoplasm, replacement (valve), osteoporosis (steroid induced).

There appear to be ways of cutting the false positive rate significantly, but it will involve custom dictionaries and may incure some loss of sensitivity.

**Discussion**

The program was effective in extracting most of the diseases and procedures that were available in the DSs. The experience reported by annotators has been that it was able to find many more specific concepts than they found by hand coding. There are only 27 such concepts listed in the table, but for most of the cases the annotators had the concepts available from the program and simply selected them.

We tried a number of simple adjustments to the program such as picking more specific concepts when there were multiple matches and adjusting the words and punctuation used to divide the text into phrases for matching. However, with a sensitivity of 93% it was hard to find anything that resulted in a net improvement. We also ran the comparison using different versions of the UMLS. Similarly, no trend could be detected since although there were new concepts matched, there were also new phrases that caused problems such as the "…myocardial infarction with electrocardiogram…" example.

Still, this analysis does point to some ideas to improve the sensitivity. The most important is to allow overlapping concepts. For example, "upper extremity deep vein thrombosis" should be coded as *structure of deep venous system of upper extremity* and *deep venous thrombosis* even though "deep vein" is part of both. This also handles the problem of "acute myocardial infarction with electrocardiogram changes" by coding it as *acute myocardial infarction* and *myocardial infarction electrocardiogram*, both of which are true.

Still, dividing the text is a challenge. Since normalization discards word order and many prepositions, there can be unexpected matches of phrases with prepositions. For example, "heart failure with systolic…" matches *systolic heart failure* no matter what "systolic" refers to.

A small custom dictionary could be added to make choices that are unambiguous in the domain context. For the ICU context with hemodynamically compromised patients, "catheterization" is *cardiac catheterization*, "anxiety" is a disease, "intubation" is *endotrachael intubation*, etc. Generating such a dictionary can be assisted by tracking ambiguous

phrases and manually determining which are resolved by the context. This can be extended by keeping track of concepts such as *intubation* with daughter concepts.

A more difficult problem is how to improve matching of concepts, part of which has many synonyms or near synonyms. For example, there are many words for types of surgery. Surgery may be called "excision", "incision", "resection", "removal", "placement", "replacement", etc., depending on the type of surgery. Since many of the dictionaries that make up the UMLS are generated from examples, there are many missing legitimate phrases describing many kinds of surgery. One possible approach would be to treat such words as a marker for all of their synonyms. This approach is also suggested by Baud, et al[14].

**Conclusion**

We have developed a program for extracting diagnoses and procedures from discharge summaries using a minimum of natural language processing, relying instead on the extensive dictionary provided by the UMLS. We have compared the concepts extracted by the program to those included in problem lists generated by four physician annotators on 96 cases of ICU patients with hemodynamic compromise. The program extracts 93% of the concepts used by the annotators and is considered by them to be a great time saver in the annotation process. Still, the program missed 81 concepts that were described in the DSs. These were missed because of ambiguous matches for phrases, words missing from the matching phrases, phrases divided in the text as well as a number of other reasons. There were a large number of false positives resulting from concepts that do not apply to the patient currently, concepts that are true but insignificant, and mismatched concepts. There are several possible improvements, the most important of which is to allow overlapping concepts and use of a small dictionary to customize matching to the domain.

**Acknowledgments**

This work was supported by Grant Number R01 EB001659 from the National Institute of Biomedical Imaging and Bioengineering.

## References

[1] Lussier YA, Shagina L, Friedman C. Automating SNOMED coding using medical language understanding: a feasibility study. Proc AMIA Symp. 2001;:418-22.

[2] Cao H, Chiang MF, Cimino JJ, Friedman C, Hripcsak G. Automatic summarization of patient discharge summaries to create problem lists using medical language processing. Medinfo. 2004;2004(CD):1540.

[3] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*. 2004 Sep-Oct;11(5):392-402.

[4] Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically screening discharge summaries for adverse medical events. *J Am Med Inform Assoc*. 2003 Jul-Aug;10(4):339-50.

[5] Hahn U, Romacker M, Schulz S. MedsynDikate — a natural language system for the extraction of medical information from finding reports. *Intl J Med Informatics*. 2002; 63-74

[6] Lindberg, DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Meth Inf Med.* 1993;32: 281-91.

[7] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17-21.

[8] Meystre S, Haug, PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform*, 2006, 39:589-99.

[9] Chapman WW, Fiszman M, Dowling JN, et al. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. MedInfo 2004 487-491.

[10] Long W, Extracting diseases from discharge summaries. Proc AMIA Symp. 2005; 470-4.

[11] National Library of Medicine, Specialist lexical tools, http://specialist.nlm.nih.gov/LexTools.html

[12] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. MedInfo 2001;10(Pt 1):216-20.

[13] Chapman WW, Bridewell W, Hanbury P, et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001 34:301-10.

[14] Baud RH, Ruch P, Gaudinat A, et al. Coping with the variability of medical terms. MedInfo 2004 322-326.