Content and Structure of Clinical Problem Lists: A Corpus Analysis

Tielman T. Van Vleck, M.Phil.[1], Adam Wilcox, Ph.D. [1],
Peter D. Stetson, MD MA[1,2], Stephen B. Johnson, Ph.D.[1], Noémie Elhadad, Ph.D.[1]
[1]Department of Biomedical Informatics, Columbia University, New York, NY
[2]Department of Medicine, Columbia University, New York, NY

ABSTRACT

In the interest of designing an automated, high-level, longitudinal clinical summary of a patient record, we analyze traditional ways in which medical problems pertaining to the patient are summarized in the electronic health record. The patient problem list has become a commonly used proxy for a summary of patient history and automated methods have been proposed to generate it. However, little research has been conducted on how to structure the problem list in a manner most effective for supporting clinical care. This study analyzes the structure and content of the Past Medical History (PMH) sections of a large corpus of clinical notes, as a proxy for problem lists. Findings show that when listing patients' history, physicians convey several semantic types of information, not only problems. Furthermore, they often group related concepts in a single line of the PMH. In contrast, traditional problem lists allow for only a simple enumeration of coded terms. Content analysis goes on to reiterate the value of more complex representations as well as provide valuable data and guidelines for automated generation of a clinical summary.

INTRODUCTION

To diagnose a patient, a physician must first develop a comprehensive understanding of the patient's medical status, including preexisting problems. Physicians have traditionally learned this information through a combination of interviewing the patient and reading the notes in the patient's record. Due to advances in medical information technology, today's patient record may contain an overwhelming amount of information, and physicians can struggle to identify all salient information, especially when pressed for time. Presenting physicians with a summary of the important information in a patient record would help physicians carry out this task more efficiently and with possibly more accuracy. We are investigating how to generate an automated, longitudinal, clinical summary of a given patient record.

In this paper, we analyze ways in which problems pertaining to a patient are traditionally kept track of in the electronic health record. The problem-oriented medical record[1] was proposed as a way to help physicians track patients' problems. While rarely implemented in full, many believe aspects of the problem-oriented medical record could assist quality and error reduction efforts in medicine[2]. The practice of maintaining a patient problem list within the traditional medical record has established itself as a compromise for achieving many of the benefits of the problem-oriented medical record without loosing key advantages of the traditional chronological medical record. The patient problem list is traditionally a simple list of a patient's medical and social problems that "encourages doctors to think holistically about their patients and means that minor problems are less likely to be forgotten."[3] As a result, the Institute of Medicine recommends use of a problem list and JCAHO requires one as an element of a complete medical record[4].

Most commercial Electronic Health Record (EHR) systems provide some functionality for physicians to maintain a problem list. However, this functionality typically consists a list of ICD-9CM codes that must be manually maintained by the physician. Managing this list is time consuming for physicians, so it is often not properly maintained. As a result, at transitions of care physicians often find no problem list or other patient summary available at all[5]. Several informatics research projects have investigated methods for automated generation of this list from clinical notes generation[2; 6-9]. Cao et al argue that traditional problem lists, simple enumerations of elements such as signs, symptoms and diagnoses, are inadequate for physicians to document medical problems and that problem lists should also represent relationships between problems[10]. But little research has focused on whether the existing problem list, as present in most EHRs, is an adequate medium for physicians to record medical problems. Answering this question can help design a better manual problem list as well as inform the automated generation of a comprehensive longitudinal clinical summary.

This study examines how physicians construct problem lists "in the wild" in order to discover patterns or trends in how physicians construct a problem list. We focus on two research questions. First, when physicians are given freedom to enter any information they think is important in a problem list,

what types of information are typically conveyed? Second, how do physicians organize the information in problem lists without the restriction of a flat list? We hypothesized that the rigid problem list structure imposed by most EHRs is overly constrictive for physicians to freely express themselves. Our method relies on the automatic analysis of a large corpus of extracts from free-text clinical notes that report past medical history.

METHODS

Overview: In order to study physician representation of patient history, we obtained a corpus of initial visit notes and examined them for sections representative of a clinical overview. The past medical history (PMH), including past surgical history, stood out as the only longitudinal summary of patient status. This is logical and expected as physicians tend to consider the PMH a proxy for the problem list. The PMH tends to be a list of relevant medical data on the patient. Its structure is not dissimilar to that of a problem list, but more complex. At NewYork-Presbyterian, the PMH is entered in a free text field, not in a structured format. A PMH list item may be a single concept or it may be a more complex statement requiring narrative text to describe. For example, below is an extract from a PMH showing both simple list elements and more complex constructs:

- HTN
- DM2
- S/p CVA 2004 w/ memory loss f/i Memory Clinic
- Dementia-presumed multi-infarct-daughter said patient's memory very good prior to CVAs in 2004

While the PMH is written with a more historical perspective than an up-to-date clinical summary (like a problem list) would be, the data types are similar. The PMH serves as an appropriate source from which we can learn how to represent medical concepts in a manner less restrictive to physician expression. Our method mines a large collection of PMH lists to identify characteristics of PMH constructs that could be used for building a clinical summary in a structure that is both complete and familiar to a physician. The output of our method is an analysis of PMH content, in particular the PMH structure; and by proxy, guidelines for construction of a full clinical summary.

Clinical Document Collection: A collection of 7673 initial visit notes was obtained from the Columbia University Medical Center Milstein Hospitalist Service. This includes all resident and attending initial visit notes and initial consult notes for inpatient admissions of all types from late 2006 through early 2007. They are not filtered and should therefore be representative of all patients admitted to the Hospitalist Service. All notes from the Service are entered through semi-structured entry templates in a system called eNote[11]. PMH was entered into a coded field in eNote templates, but as free text within that field. The advantage for this analysis was that these lists were in the doctor's own words without any limits on structure or content imposed by the information system. The notes were stored using the Clinical Document Architecture (CDA) XML schema. This allowed for a simple XSL transformation to filter protected health information (PHI) and convert sections of interest to text. A small Java application was written to perform this XSLT on each document and do basic preprocessing to prepare the text for natural language processing analysis.

Data Preparation: The corpus was then parsed with the MedLEE natural language processor[12] to obtain the semantic structure and UMLS codes of concepts represented in these notes.

MedLEE output was generated as XML and a Java postprocessor was used to validate the XML output. Each note section was divided into a text section with numbered phrase tags around identifiable phrases and a structured element containing references describing the tagged phrases. Reference tags were named with the phrase's semantic type. MedLEE assigned a UMLS code to the phrase whenever it could map the clinical information detected to known UMLS concepts. MedLEE results were then merged into one large XML file to facilitate querying across all documents with XQuery.

The merged MedLEE results were processed using XQuery into another XML file describing the structure and contents of each PMH section. The results were represented by sentence, which in the case of the PMH list generally encompassed a single line. A line generally represented a single concept of the PMH list, though could be something closer to a real sentence where more narrative text was used. Each line of the PMH was then analyzed, observing patterns in structure and content.

Data Analysis: The primary discourse analyses focused on the semantic structure of the notes, meaning what type of medical concepts were discussed in what order. MedLEE grouped each medical concept (a single word or a group of related words) into one of the sixteen classes in Table 1.

Semantic classes of each concept in a line were joined to represent the elements of the line. For example, "S/P CVA – Dx'd 7/18/1 (+MRI)" was coded with the initial UMLS code *C0038454_accident cerebrovascular* (a problem) followed by *C0024485_magnetic resonance imaging*

Semantic Class	Example
problem	hypertension
medication	asa (Asprin)
procedure	liver function test
lab test	complete blood count
status	previous
finding	elevated
body measure	right atrial pressure
body function	po intake
device	catheter
normal finding	within normal limits
recommendation	computerized axial tomography
time period	admission
change	increase
technique	limited study
diagnosis material	TC^99M
relative time	history

Table 1: Semantic classes tagged by MedLEE

(a procedure). The phrase is thus represented as "problem procedure". A decision was made to map only elements for which a valid UMLS code could be identified, thereby excluding findings such as "patch", problems such as "degenerative changes" and body functions such as "behavior" for which no UMLS code exists. Results were extracted with XQuery to examine the combined terms.

In order to confirm the importance of these multi-concept lines, we compared initial concepts to subsequent modifier concepts. For example, in the PMH entry "Prostate CA s/p brachytherapy", MedLEE identified the problem *malignant neoplasm of prostate* and the procedure *brachytherapy*. We refer to *malignant neoplasm of prostate* as the initial concept and *brachytherapy* as the subsequent concept. We identified UMLS codes occurring as the initial concept in a line, and subsequent concepts following the initial term. Lines with only one term were excluded. This allows us to identify common term combinations, thereby providing insight about what concepts are commonly considered relevant for explaining others fully.

Initial codes were then analyzed to measure how often they occurred as subsequent terms, and subsequent codes were reviewed for frequency of occurrence as an initial code. The goal of this analysis was to learn whether certain codes were inherently more important and should be presented first on the line or whether order was irrelevant to importance.

The final analysis examined problems occurring alone or "unmodified" versus those associated with additional clarifying concepts. For statistical relevance, we excluded problems occurring fewer than six times in the entire corpus.

RESULTS

After processing, it was found that the corpus of 7673 notes contained 6201 completed PHs. The PMH set contained a total of 38,018 sentences recognized by MedLEE.

PMH Structure: Discourse analysis of the sentence structures revealed 809 unique line structures, though these were primarily unusual constructs from lines where the physician felt compelled to use more narrative text to explain the condition. The most common line structure was a single problem. This occurred in 52% of instances. Following are the top ten sentence structures along with the percent of PMH lines occurring with the pattern:

1.	problem	52.1%
2.	procedure	11.4%
3.	problem problem	7.3%
4.	problem procedure	3.8%
5.	problem med	2.4%
6.	problem problem problem	1.6%
7.	procedure problem	1.4%
8.	med	1.4%
9.	procedure procedure	0.9%
10.	bodymeas	0.8%

Of all PMH line patterns identified, 32% had two or more concepts coded by MedLEE. The following chart represents the use of multi-term lines.

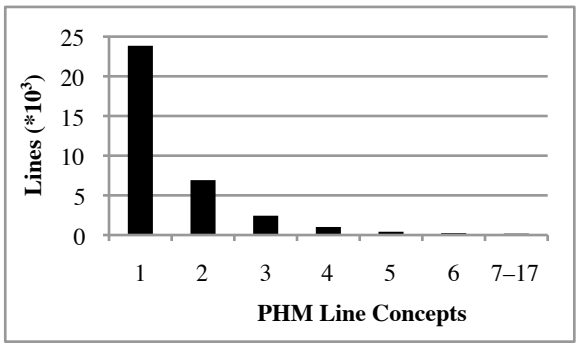

Figure 1: PMH line concepts

Common Term Combinations: A similar analysis was performed to identify combinations of two or more UMLS codes in the same PMH line. Top code combinations are reported below with the frequency of occurrence in the corpus. Note that these are not the original text but descriptions of UMLS concepts identified in the source text.

1. congestive heart failure, ejection fraction (96)
2. coronary arteriosclerosis, coronary artery bypass surgery (95)
3. asthma, intubation (91)
4. fibrillation atrial, Coumadin (74)
5. depression mental, anxiety state (56)

The combination of specific terms implies a deeper meaning than just the presence of two items alone. Generally, there is a deeper semantic relationship. For example the "fibrillation atrial" followed by "Coumadin" implies a relationship such as "treated with".

Initial vs. Subsequent Terms: In this analysis, we identified UMLS codes of initial and subsequent term in PMH lines, and looked for groups of term more likely to be used as initial or subsequent terms. We found that of 3120 initial terms, 49.8% never occurred as a subsequent terms. Of 2874 subsequent terms, 45.5% never occurred as the initial term in a PMH line.

Of 1476 initial concepts that were never occurred as subsequent items, the top five were:

1.	polysubstance	96
2.	malignant neoplasm of breast	63
3.	osteoarthritis knee	44
4.	psoriasis	34
5.	chest pain syndrome	31

Of 1350 subsequent concepts that never occurred as initial items, the top five were:

1.	Pseudomonas	18
2.	orthopedic cast	14
3.	hemiparesis	14
4.	cisplatin	13
5.	aphasia	13

Of 1602 concepts that could occur as either initial or subsequent concepts, the top five were:

1.	hypertensive disease	3768
2.	diabetes mellitus	1014
3.	coronary arteriosclerosis	968
4.	depression mental	877
5.	hyperlipidemia	704

The results of this examination show a group of terms which are useful as either initial or subsequent terms, then two similarly sized groups (in number, not use) which are used distinctly as an initial term or a subsequent term. While the groups contain a similar number of terms, the number of times these terms were used is drastically different. Looking at the results, we see the most commonly used initial-only concept, *polysubstance* was used only 96 times. The most used subsequent-only concept, *Pseudomonas* occurred only 18 times. The most used term that could be used as either an initial term or a subsequent term was used 3768 times.

Terms Requiring Detail: This analysis identified the likelihood of an initial concept being followed by at least one subsequent concept. For example, the concept *fracture of olecranon* was modified by terms such as *swelling of hand, radiography of elbow, splint removal* and *orthopedic cast*. Of 766 total initial concepts, 20 initial terms were found to always be followed by subsequent items, while 38 were never used with subsequent terms. Otherwise, terms fall into the following distribution.

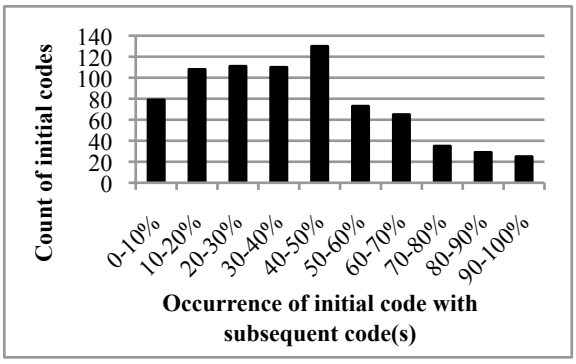

Figure 2: Frequency of a initial code occuring with a subsequent code giving further detail

DISCUSSION

Results of this study provide insight on how to better structure high level longitudinal clinical summaries such as a next-generation problem list, as well as data to assist in the automatic generation of such clinical summaries. We demonstrated that a summary requires concepts of many different semantic types, as well as phrases of multiple concepts.

PMH Structure: As the PMH is used as a review of preexisting problems and procedures it makes sense that these elements are the two most common semantic structures in the PMH. However, the third most common type of semantic structure found was a sequence of two problems ("problem problem"). This indicates a need to add detail to a specific problem that cannot be expressed using a single code, as dictated by most EHRs. This is not due to multi-word concepts such as "gastric ulcer" as these are recognized by MedLEE as a single entity and are represented with a single UMLS code. From the analysis, we can see that elements in the PMH need more flexibility of content and expression than the traditional problem list paradigm allows for.

Common Term Combinations: Analysis of frequent code combinations is of less interest to an individual reading the notes. But if attempting to generate a summary document programmatically, a critical question to ask would be: once a concept has been identified as relevant to include in the summary, is any further detail required to complete the definition of the problem? The information identified here is clearly of critical interest to answering this question.

Initial vs. Subsequent Terms: This analysis indicates that there are not distinct groups of codes likely to be used as initial codes in a line and other codes likely to be used as subsequent codes. It tells us that subsequent terms are not unimportant terms carrying less weight than initial terms, and their importance must not be discounted in the development of any type of clinical summary. For the task of clinical summary generation, this is important information for the process of content selection.

Terms Requiring Detail: This analysis confirms our findings that we need to offer physicians more flexibility of expression than a simple list. For most concepts that might be included in a PMH, this analysis shows that at times it can be necessary for the authoring physician to add a secondary term to provide more detail than the single initial code can convey. It also shows that there are concepts that always have follow-up detail and very few never do.

Limitations: Source texts used were limited to initial visit notes from a hospitalist service. Ideally, results would be differentiated by primary diagnosis and physician specialty and different source corpus would be expanded to correspond to each combination.

Findings of this study are clearly dependant on the quality of the original PMHs, as well as the accuracy of MedLEE's parsing. The PMH is written in a very telegraphic style, which is not what MedLEE was trained on. Nevertheless, MedLEE provided consistently good results on this data set.

Future Work: This work provides preliminary data for the task of generating a clinical summary. Future work will focus on this task and incorporate the findings of this study for the purpose of both how to structure a clinical summary as well as what data to include (content selection).

CONCLUSIONS

Our analysis of free-text entered problem lists confirms that a simple ICD-9CM-based list is inadequate for a clinical summary problem list as a single ICD-9CM code is insufficient to express the complete medical thoughts required for the task. In order for a physician to fully express a medical scenario, it is important to facilitate at least a combination of codes, if not complete clinical narrative. Our analysis indicates there are common patterns of concept use within these lists, which provides insight into the requirements for structure and content of automated problem list generation. Development of any type of clinical summary document may benefit from allowing the flexibility to represent information in more complex, nested structures than simple code lists as required by many modern systems. To our knowledge, this is the first study to conduct a corpus-based analysis of problem lists using medical natural language processing.

ACKNOWLEDGEMENTS

This work was supported in part by National Library of Medicine training grant N01-LM07079 (TTVV) and NLM grant K22-LM008805 (PDS).

We thank Dr. Carol Friedman for the use of MedLEE. Ongoing MedLEE development is funded by NLM R01 LM007659 and R01 LM008635.

REFERENCES

1. Weed, L. L. (1968). Medical records that guide and teach. *N Engl J Med, 278*(12), 652-657 concl.
2. Meystre, S., & Haug, P. J. (2006). Improving the Sensitivity of the Problem List in an Intensive Care Unit by Using Natural Language Processing. *AMIA Symp Proc*.
3. Lloyd, B. W., & Barnett, P. (1993). Use of problem lists in letters between hospital doctors and general practitioners. *BMJ, 306*(6872), 247.
4. Campbell, J. R. (1998). Strategies for problem list implementation in a complex clinical enterprise. *AMIA Symp Proc*.
5. Pratt, W., & Yetisgen-Yildiz, M. (2003). A Study of Biomedical Concept Identification: MetaMap vs. People. *AMIA Symp Proc*.
6. Meystre, S., & Haug, P. J. (2005). Automation of a problem list using natural language processing. *BMC Med Inform Decis Mak*.
7. Meystre, S., & Haug, P. J. (2005). Comparing Natural Language Processing Tools to Extract Medical Problems from Narrative Text. *AMIA Symp Proc*.
8. Meystre, S., & Haug, P. J. (2006). Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of Biomedical Informatics, 39*(6), 589-599.
9. Bui, A. A., Taira, R. K., El-Saden, S., Dordoni, A., & Aberle, D. R. (2004). Automated medical problem list generation: Towards a patient timeline. *MedInfo Symp Proc*.
10. Cao, H., Markatou, M., Melton, G. B., Chiang, M. F., & al., e. (2005). Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. *AMIA Annual Symposium Proceedings*.
11. Johnson, S. B., Bakken, S., Dine, D., Hyun, S., Mendonca, E., Morrison, F., et al. (2007). An Electronic Health Record based on Structured Narrative. *JAMIA*.
12. Friedman, C., Hripcsak, G., Shagina, L., & Liu, H. (1999). Representing Information in Patient Reports Using Natural Language Processing and the Extensible Markup Language. *JAMIA*.