Generating Patient Problem Lists from the ShARe Corpus using SNOMED CT/SNOMED CT CORE Problem List

Danielle Mowery
Janyce Wiebe
University of Pittsburgh
Pittsburgh, PA
dlm31@pitt.edu
wiebe@cs.pitt.edu

Mindy Ross
University of California
San Diego
La Jolla, CA
mkross@ucsd.edu

Sumithra Velupillai
Stockholm University
Stockholm, SE
sumithra@dsv.su.se

Stephane Meystre
Wendy W Chapman
University of Utah
Salt Lake City, UT
stephane.meystre,
wendy.chapman@utah.edu

Abstract

An up-to-date problem list is useful for assessing a patient's current clinical status. Natural language processing can help maintain an accurate problem list. For instance, a patient problem list from a clinical document can be derived from individual problem mentions within the clinical document once these mentions are mapped to a standard vocabulary. In order to develop and evaluate accurate document-level inference engines for this task, a patient problem list could be generated using a standard vocabulary. Adequate coverage by standard vocabularies is important for supporting a clear representation of the patient problem concepts described in the texts and for interoperability between clinical systems within and outside the care facilities. In this pilot study, we report the reliability of domain expert generation of a patient problem list from a variety of clinical texts and evaluate the coverage of annotated patient problems against SNOMED CT and SNOMED **C**linical **O**bservation **R**ecording and **E**ncoding (CORE) Problem List. Across report types, we learned that patient problems can be annotated with agreement ranging from 77.1% to 89.6% F1-score and mapped to the CORE with moderate coverage ranging from 45%-67% of patient problems.

1 Introduction

In the late 1960's, Lawrence Weed published about the importance of problem-oriented medical records and the utilization of a problem list to facilitate care provider's clinical reasoning by reducing the cognitive burden of tracking *current, active* problems from *past, inactive* problems from the patient health record (Weed, 1970). Although electronic health records (EHR) can help achieve better documentation of problem-specific information, in most cases, the problem list is manually created and updated by care providers. Thus, the problem list can be out-of-date containing resolved problems or missing new problems. Providing care providers with problem list update suggestions generated from clinical documents can improve the completeness and timeliness of the problem list (Meystre and Haug, 2008).

In recent years, national incentive and standard programs have endorsed the use of problem lists in the EHR for tracking patient diagnoses over time. For example, as part of the Electronic Health Record Incentive Program, the Center for Medicare and Medicaid Services defined demonstration of *Meaningful Use* of adopted health information technology in the Core Measure 3 objective as "maintaining an up-to-date problem list of current and active diagnoses in addition to historical diagnoses relevant to the patients care" (Center for Medicare and Medicaid Services, 2013). More recently, the **S**ystematized **N**omenclature of **M**edicine **C**linical **T**erms (SNOMED CT) has become the standard vocabulary for representing and documenting patient problems within the clinical record. Since 2008, this list is iteratively refined four times each year to produce a subset of generalizable clinical problems called the SNOMED CT CORE Problem List. This CORE list represents the most frequent problem terms and concepts across eight major healthcare institutions in the United States and is designed to support interoperability between regional healthcare institutions (National Library of Medicine, 2009).

In practice, there are several methodologies applied to generate a patient problem list from clinical text. Problem lists can be generated from coded diagnoses such as the **I**nternational Statistical **C**lassification of **D**isease (ICD-9 codes) or

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 54–58,
Baltimore, Maryland USA, June 26-27 2014.

concept labels such as **U**nified **M**edical **L**anguage **S**ystem **c**oncept **u**nique **i**dentifiers (UMLS CUIs). For example, Meystre and Haug (2005) defined 80 of the most frequent problem concepts from coded diagnoses for cardiac patients. This list was generated by a physician and later validated by two physicians independently. Coverage of coded patient problems were evaluated against the ICD-9-CM vocabulary. Solti et al. (2008) extended the work of Meystre and Haug (2005) by not limiting the types of patient problems from any list or vocabulary to generate the patient problem list. They observed 154 unique problem concepts in their reference standard. Although both studies demonstrate valid methods for developing a patient problem list reference standard, neither study leverages a standard vocabulary designed specifically for generating problem lists.

The goals of this study are 1) determine how reliably two domain experts can generate a patient problem list leveraging SNOMED CT from a variety of clinical texts and 2) assess the coverage of annotated patient problems from this corpus against the CORE Problem List.

2 Methods

In this IRB-approved study, we obtained the **Sh**ared **A**nnotated **Re**source (ShARe) corpus originally generated from the Beth Israel Deaconess Medical Center (Elhadad et al., under review) and stored in the **M**ultiparameter **I**ntelligent **M**onitoring in **I**ntensive **C**are, version 2.5 (MIMIC II) database (Saeed et al., 2002). This corpus consists of discharge summaries (DS), radiology (RAD), electrocardiogram (ECG), and echocardiogram (ECHO) reports from the **I**ntensive **C**are **U**nit (ICU). The ShARe corpus was selected because it 1) contains a variety of clinical text sources, 2) links to additional patient structured data that can be leveraged for further system development and evaluation, and 3) has encoded individual problem mentions with semantic annotations within each clinical document that can be leveraged to develop and test document-level inference engines. We elected to study ICU patients because they represent a sensitive cohort that requires up-to-date summaries of their clinical status for providing timely and effective care.

2.1 Annotation Study

For this annotation study, two annotators - a physician and nurse - were provided independent training to annotate clinically relevant problems e.g., *signs*, *symptoms*, *diseases*, and *disorders*, at the document-level for 20 reports. The annotators were given feedback based on errors over two iterations. For each patient problem in the remaining set, the physician was instructed to review the full text, span the a problem mention, and map the problem to a CUI from SNOMED-CT using the extensible Human Oracle Suite of Tools (eHOST) annotation tool (South et al., 2012). If a CUI did not exist in the vocabulary for the problem, the physician was instructed to assign a "CUI-less" label. Finally, the physician then assigned one of five possible status labels - *Active*, *Inactive*, *Resolved*, *Proposed*, and *Other* - based on our previous study (Mowery et al., 2013) to the mention representing its last status change at the conclusion of the care encounter. Patient problems were not annotated as *Negated* since patient problem concepts are assumed absent at a document-level (Meystre and Haug, 2005). If the patient was healthy, the physician assigned "Healthy - no problems" to the text. To reduce the cognitive burden of annotation and create a more robust reference standard, these annotations were then provided to a nurse for review. The nurse was instructed to add missing, modify existing, or delete spurious patient problems based on the guidelines.

We assessed how reliably annotators agreed with each other's patient problem lists using inter-annotator agreement (IAA) at the document-level. We evaluated IAA in two ways: 1) by problem CUI and 2) by problem CUI and status. Since the number of problems not annotated (i.e., ***true negatives (TN)***) are very large, we calculated F1-score as a surrogate for kappa (Hripcsak and Rothschild, 2005). F1-score is the harmonic mean of recall and precision, calculated from ***true positive***, ***false positive***, and ***false negative*** annotations, which were defined as follows:

true positive (TP) = the physician and nurse problem annotation was assigned the same CUI (and status)

false positive (FP) = the physician problem annotation (and status) did not exist among the nurse problem annotations

false negative (FN) = the nurse problem annotation (and status) did not exist among the physician problem annotations

Recall =

$$\frac{TP}{(TP+FN)} \quad (1)$$

Precision =

$$\frac{TP}{(TP+FP)} \quad (2)$$

F1-score =

$$2\frac{(Recall * Precision)}{(Recall + Precision)} \quad (3)$$

We sampled 50% of the corpus and determined the most common errors. These errors with *examples* were programmatically adjudicated with the following **solutions**:

Spurious problems: *procedures*
solution: exclude non-problems via guidelines

Problem specificity: *CUI specificity differences*
solution: select most general CUIs

Conflicting status: *negated vs. resolved*
solution: select second reviewer's status

CUI/CUI-less: *C0031039 vs. CUI-less*
solution: select CUI since clinically useful

We split the dataset into about two-thirds training and one-third test for each report type. The remaining data analysis was performed on the training set.

2.2 Coverage Study

We characterized the composition of the reference standard patient problem lists against two standard vocabularies SNOMED-CT and SNOMED-CT CORE Problem List. We evaluated the coverage of patient problems against the SNOMED CT CORE Problem List since the list was developed to support encoding clinical observations such as findings, diseases, and disorders for generating patient summaries like problem lists. We evaluated the coverage of patient problems from the corpus against the SNOMED-CT January 2012 Release which leverages the UMLS version 2011AB. We assessed recall (Eq 1), defining a TP as a patient problem CUI occurring in the vocabulary and a FN as a patient problem CUI not occurring in the vocabulary.

3 Results

We report the results of our annotation study on the full set and vocabulary coverage study on the training set.

3.1 Annotation Study

The full dataset is comprised of 298 clinical documents - 136 (45.6%) DS, 54 (18.1%) ECHO, 54 (18.1%) RAD, and 54 (18.1%) ECG. Seventy-four percent (221) of the corpus was annotated by both annotators. Table 1 shows agreement overall and by report, matching problem CUI and problem CUI with status. Inter-annotator agreement for problem with status was slightly lower for all report types with the largest agreement drop for DS at 15% (11.6 points).

Report Type	CUI	CUI + Status
DS	77.1	65.5
ECHO	83.9	82.8
RAD	84.7	82.8
ECG	89.6	84.8

Table 1: Document-level IAA by report type for problem (CUI) and problem with status (CUI + status)

We report the most common errors by frequency in Table 2. By report type, the most common errors for ECHO, RAD, and ECG were CUI/CUI-less, and DS was Spurious Concepts.

Errors	DS	ECHO	RAD	ECG
SP	423 (42%)	26 (23%)	30 (35%)	8 (18%)
PS	139 (14%)	31 (27%)	8 (9%)	0 (0%)
CS	318 (32%)	9 (8%)	8 (9%)	14 (32%)
CC	110 (11%)	34 (30%)	37 (44%)	22 (50%)
Other	6 (>1%)	14 (13%)	2 (2%)	0 (0%)

Table 2: Error types by frequency - Spurious Problems (SP), Problem Specificity (PS), Conflicting status (CS), CUI/CUI-less (CC)

3.2 Coverage Study

In the training set, there were 203 clinical documents - 93 DS, 37 ECHO, 38 RAD, and 35 ECG. The average number of problems were 22±10 DS, 10±4 ECHO, 6±2 RAD, and 4±1 ECG. There are 5843 total current problems in SNOMED-CT CORE Problem List. We observed a range of unique SNOMED-CT problem concept frequencies: 776 DS, 63 ECHO, 113 RAD, and 36 ECG

by report type. The prevalence of covered problem concepts by CORE is 461 (59%) DS, 36 (57%) ECHO, 71 (63%) RAD, and 16 (44%) ECG. In Table 3, we report coverage of patient problems for each vocabulary. No reports were annotated as "Healthy - no problems". All reports have SNOMED CT coverage of problem mentions above 80%. After mapping problem mentions to CORE, we observed coverage drops for all report types, 24 to 36 points.

Report Type	Patient Problems	Annotated with SNOMED CT	Mapped to CORE
DS	2000	1813 (91%)	1335 (67%)
ECHO	349	300 (86%)	173 (50%)
RAD	190	156 (82%)	110 (58%)
ECG	95	77(81%)	43 (45%)

Table 3: Patient problem coverage by SNOMED-CT and SNOMED-CT CORE

4 Discussion

In this feasibility study, we evaluated how reliably two domain experts can generate a patient problem list and assessed the coverage of annotated patient problems against two standard clinical vocabularies.

4.1 Annotation Study

Overall, we demonstrated that problems can be reliably annotated with moderate to high agreement between domain experts (Table 1). For DS, agreement scores were lowest and dropped most when considering the problem status in the match criteria. The most prevalent disagreement for DS was Spurious problems (Table 2). Spurious problems included additional events (e.g., **C2939181**: *Motor vehicle accident*), procedures (e.g., **C0199470**: *Mechanical ventilation*), and modes of administration (e.g., **C0041281**: *Tube feeding of patient*) that were outside our patient problem list inclusion criteria. Some pertinent findings were also missed. These findings are not surprising given on average more problems occur in DS and the length of DS documents are much longer than other document types. Indeed, annotators are more likely to miss a problem as the number of patient problems increase.

Also, status differences can be attributed to multiple status change descriptions using expressions of time e.g., "cough improved then" and modality "rule out pneumonia", which are harder to track and interpret over a longer document. The most prevalent disagreements for all other document types were CUI/CUI-less in which identifying a CUI representative of a clinical observation proved more difficult. An example of Other disagreement was a sidedness mismatch or redundant patient problem annotation. For example, **C0344911**: *Left ventricular dilatation* vs. **C0344893**: *Right ventricular dilatation* or **C0032285**: *Pneumonia* was recorded twice.

4.2 Coverage Study

We observed that DS and RAD reports have higher counts and coverage of unique patient problem concepts. We suspect this might be because other document types like ECG reports are more likely to have laboratory observations, which may be less prevalent findings in CORE. Across document types, coverage of patient problems in the corpus by SNOMED CT were high ranging from 81% to 91% (Table 3). However, coverage of patient problems by CORE dropped to moderate coverages ranging from 45% to 67%. This suggests that the CORE Problem List is more restrictive and may not be as useful for capturing patient problems from these document types. A similar report of moderate problem coverage with a more restrictive concept list was also reported by Meystre and Haug (2005).

5 Limitations

Our study has limitations. We did not apply a traditional adjudication review between domain experts. In addition, we selected the ShARe corpus from an ICU database in which vocabulary coverage of patient problems could be very different for other domains and specialties.

6 Conclusion

Based on this feasibility study, we conclude that we can generate a reliable patient problem list reference standard for the ShARe corpus and SNOMED CT provides better coverage of patient problems than the CORE Problem List. In future work, we plan to evaluate from each ShARe report type, how well these patient problem lists can be derived and visualized from the individual disease/disorder problem mentions leveraging temporality and modality attributes using natural language processing and machine learning approaches.

Acknowledgments

This work was partially funded by NLM (5T15LM007059 and 1R01LM010964), ShARe (R01GM090187), Swedish Research Council (350-2012-6658), and Swedish Fulbright Commission.

References

Center for Medicare and Medicaid Services. 2013. EHR Incentive Programs-Maintain Problem List. http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/3_Maintain_Problem_ListEP.pdf.

Noemie Elhadad, Wendy Chapman, Tim OGorman, Martha Palmer, and Guergana. Under Review Savova. under review. The ShARe Schema for the Syntactic and Semantic Annotation of Clinical Texts.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-measure, and Reliability in Information Retrieval. *J Am Med Inform Assoc*, 12(3):296–298.

Stephane Meystre and Peter Haug. 2005. Automation of a Problem List using Natural Language Processing. *BMC Medical Informatics and Decision Making*, 5(30).

Stephane M. Meystre and Peter J. Haug. 2008. Randomized Controlled Trial of an Automated Problem List with Improved Sensitivity. *International Journal of Medical Informatics*, 77:602–12.

Danielle L. Mowery, Pamela W. Jordan, Janyce M. Wiebe, Henk Harkema, John Dowling, and Wendy W. Chapman. 2013. Semantic Annotation of Clinical Events for Generating a Problem List. In *AMIA Annu Symp Proc*, pages 1032–1041.

National Library of Medicine. 2009. The CORE Problem List Subset of SNOMED-CT. Unified Medical Language System 2011. http://www.nlm.nih.gov/research/umls/SNOMED-CT/core_subset.html.

Mohammed Saeed, C. Lieu, G. Raber, and Roger G. Mark. 2002. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput Cardiol*, 29.

Imre Solti, Barry Aaronson, Grant Fletcher, Magdolna Solti, John H. Gennari, Melissa Cooper, and Thomas Payne. 2008. Building an Automated Problem List based on Natural Language Processing: Lessons Learned in the Early Phase of Development. pages 687–691.

Brett R. South, Shuying Shen, Jianwei Leng, Tyler B. Forbush, Scott L. DuVall, and Wendy W. Chapman. 2012. A prototype tool set to support machine-assisted annotation. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, BioNLP '12, pages 130–139. Association for Computational Linguistics.

Lawrence Weed. 1970. *Medical Records, Medical Education and Patient Care: The Problem-Oriented Record as a Basic Tool.* Medical Publishers: Press of Case Western Reserve University, Cleveland: Year Book.