An Evaluation of Clinical Natural Language Processing Systems to Extract Symptomatic Adverse Events from Patient-Authored Free-Text Narratives

Yue Wang, PhD[1], David Gotz, PhD[1], Ethan M. Basch, MD, MSc[1], Arlene E. Chung, MD, MHA, MMCi[1]
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Introduction

Symptomatic adverse events (AEs) such as nausea are common among patients enrolled in cancer clinical trials. Historically, this information has been collected and reported into research databases by clinical staff using a set of AE grading criteria maintained by the National Cancer Institute (NCI) called the Common Terminology Criteria for Adverse Events (CTCAE). In NCI's Patient-Reported Outcomes version of CTCAE (PRO-CTCAE) software system, patients can also provide supplemental free-text narratives about their AEs. 58% of patients submit supplemental AE information when given this opportunity[1]. More importantly, there was not considerable overlap between supplemental AEs submitted by patients and those elicited in trial-specific questionnaires, providing evidence for the value of collecting free-text, patient-authored AEs[1]. In our prior work, we also found that the majority (88%) of the symptom concepts within patient narratives could be manually mapped to the Medical Dictionary for Regulatory Activities (MedDRA), which is the standard lexicon for reporting AEs to regulatory agencies such as the FDA[1]. However, the manual process of mapping symptom concepts to lexicons is labor-intensive and limits the widespread collection of free-text AEs. Clinical natural language processing (NLP) has the potential to accelerate recognition and mapping of these symptom concepts and could enable real-time extraction, mapping, and reporting of patient-authored AEs. Off-the-shelf NLP systems, if high-performing, could allow for systematic text processing to be applied, but have not previously been examined for patient-authored AEs. Thus, the objective of this study was to evaluate performance of four widely used clinical NLP systems in extracting symptom concepts from patient-authored free-text AE narratives.

Methods

To determine system performance for extracting AE concepts, four systems that use algorithms ranging from basic pattern matching to deep neural networks were evaluated. Each system was used to map symptom concepts from narratives back to a MedDRA concept, when available, since MedDRA is used for regulatory reporting.

1. **BioPortal**:[2] BioPortal provides web access to a library of biomedical ontologies, and has a RESTful API that annotates documents using terms in user-specified ontologies. The underlying mechanism is multi-word string matching. We specified MedDRA as the target ontology, and used the RESTful API to annotate documents.
2. **MetaMap**:[3] MetaMap employs a set of pattern-matching rules to recognize UMLS concepts within text. The online batch service annotates documents with CUIs. UMLS Metathesaurus was then used to convert each CUI to a MedDRA concept, if available.
3. **cTAKES**:[4] cTAKES assembles a pipeline of pattern-matching and classical machine-learning NLP modules that leverage rich linguistic and semantic information for text analysis. We configured the pipeline to recognize "SignSymptomMention" and "DiseaseDisorderMention." cTAKES generates a CUI for each mention and then each CUI was converted to a MedDRA concept using UMLS Metathesaurus, if available.
4. **Amazon Comprehend Medical (ACM)**:[5] ACM is an NLP service from Amazon Web Services. The entity recognition module employs deep bidirectional long-short term memory (BiLSTM) networks. It can map medical entities to two ontologies (ICD-10-CM or RxNorm). The system was configured to recognize disease and symptom-related entities and to map them to ICD-10-CM codes. The ICD-10-CM codes were then converted to CUIs, which were converted to MedDRA concepts using UMLS Metathesaurus, if available.

Evaluation corpus. A random sample of 100 free-text narratives (documents) were selected from a corpus used in a prior PRO-CTCAE study[1]. Each narrative has symptomatic AEs described in a patient's own words without any character limits. Symptom concepts in each narrative were coded and mapped to MedDRA concepts by two physicians with an adjudicator with 96% inter-rater agreement. On average, a document had 3.4 words and 1.1 symptom mentions; a symptom mention had ~2.3 words. 85% of symptom mentions could be mapped to MedDRA.

Task definition. Given a free-text AE document, the NLP task can be decomposed into two subtasks: 1) concept recognition to identify text spans (each text span consists of one or more words) that mention symptomatic AEs within the document (defined as *symptom mentions*), and 2) concept normalization to map each symptom mention to a corresponding MedDRA concept, as represented by the preferred term (PT) or *none* if no MedDRA concept matched that symptom mention. Given a document as input, we defined the expected output from an NLP system as a set of symptom mentions that were each associated with a MedDRA concept. If a system generated overlapping symptom

mentions (e.g., "rectal bleeding" and "bleeding"), all of them were considered. If a system generated a list of PTs for a symptom mention ranked by prediction confidence (e.g., MedDRA PT 10061525 and 10023643 for "lacrimal disorder"), only the top result was considered. Relaxing the match to "anywhere in the list" did not improve performance substantially (< .02 F_1 increase for all systems, data not reported). Both *strict* and *relaxed* text match conditions were used in the concept recognition subtask. Micro-averaged precision (P), recall (R), and F_1 score were evaluation metrics for both subtasks.

Results

For the concept recognition subtask, all systems had low precision, recall, and F_1 score under the *strict text match* condition, while the metrics were overall better for *relaxed text match* (Table 1). ACM performed the best with the highest F_1 score in both matching conditions. For the concept normalization subtask, all systems had low precision, recall, and F_1 score for mapping symptom mentions to MedDRA concepts.

Table 1.* Performance across clinical NLP systems by subtask.

Systems	Concept Recognition Subtask						Concept Normalization Subtask					
	Strict Text Match			*Relaxed Text Match*			*Strict Text Match*			*Relaxed Text Match*		
	P	***R***	***F*$_1$**	***P***	***R***	***F*$_1$**	***P***	***R***	***F*$_1$**	**P**	**R**	***F*$_1$**
BioPortal	0.47	0.30	0.37	0.99	0.60	0.74	0.45	0.28	**0.34**	0.98	0.37	0.53
MetaMap	0.37	0.58	0.45	0.76	1.00	0.86	0.27	0.37	0.31	0.60	0.49	0.54
cTAKES	0.38	0.39	0.38	0.98	0.83	0.90	0.33	0.32	0.33	0.96	0.47	**0.63**
ACM	0.60	0.53	**0.56**	0.99	0.92	**0.95**	0.29	0.15	0.20	0.96	0.19	0.32

*Bold indicates the best performing system in terms of F_1 score for each subtask.

Discussion

Focusing on concept normalization, even the best performing system had low performance (strict: 0.34 F_1; relaxed: 0.63 F_1). Similar results were observed in recent shared tasks on extracting and then mapping AEs from patient-authored tweets to MedDRA, where the best performance was obtained by a BioBERT-based deep learning system (strict: 0.34 F_1; relaxed: 0.43 F_1)[6]. This suggests that the task of mapping patient-authored free-text AEs poses significant challenges for NLP systems as they are designed for clinical text. Under the *strict text match* condition, the performance gap between the two subtasks was due to errors in converting UMLS CUIs to MedDRA PTs. For example, "blood in urine" was mapped to C0018965, which was mapped to MedDRA PT 10018867 but not 10018870. Under the *relaxed text match* condition, the performance gap between the two subtasks implies that partially recognized symptom mentions do not sufficiently describe the actual AE. For example, "pain in nails" (onychalgia) is more specific than "pain," and that specificity is important for regulatory reporting. ACM performed the best for concept recognition, which may be due to the potential benefit of its deep sequence tagging algorithm. Amazon has not published their ICD-10-CM mapping algorithm, but error patterns reveal that ACM may use concept embedding matching as opposed to exact string matching, which led to fuzzy and inaccurate mapping results. Based on low performance across these widely used systems, our research reveals that patient-authored symptomatic AE text is sufficiently different from biomedical literature, clinical notes, and patient forum posts, which are the primary targets of these systems. This research highlights the need for new NLP approaches given the goal is to accurately extract and map AEs from patient free-text narratives to standard lexicons for reporting to regulatory agencies.

References

1. Chung AE, Shoenbill K, Mitchell SA, Dueck AC, Schrag D, Bruner DW, et al. Patient free text reporting of symptomatic adverse events in cancer clinical research using the National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). JAMIA. 2019 Apr;26(4):276-85.
2. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011 Jul;39(Web Server issue): W541-5. 2011 Jun 14.
3. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. JAMIA. 2010 May 1;17(3):229-36.
4. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. JAMIA. 2010 Sep 1;17(5):507-13.
5. Bhatia P, Celikkaya B, Khalilia M, Senthivel S. Comprehend medical: a named entity recognition and relationship extraction Web service. arXiv preprint arXiv:1910.07419. 2019 Oct 15.
6. Weissenbacher D, Sarker A, Magge A, Daughton A, O'Connor K, Paul M, et al. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. SMM4H Workshop & Shared Task. 2019:21-30.