The Super Annotator: A Method of Semi-Automated Rare Event Identification for Large Clinical Data Sets

Patrick R. Alba¹,², Olga V. Patterson¹,², Benjamin Viernes¹,², Daniel Denhalter¹,², Nicole Bailey³, Andrew Wilson³, Aaron W.C. Kamauu³, Scott L. DuVall¹,²

¹VA Salt Lake City Health Care System; ²University of Utah, Salt Lake City, UT; ³Anolinx LLC, Salt Lake City, UT

OBJECTIVE

Detecting rare events in an electronic medical record (EMR) is similar to searching for a needle in a haystack. Automated methods (natural language processing) and manual methods (chart review) can be used. However, when events are rare and the patient population is large, manual annotation may not be feasible and NLP alone may not be able to reach acceptable levels of performance. We propose a semi-automated method for NLP-assisted retrieval of relevant documents and manual review of the resulting instances. This method was developed to identify medication-related adverse events among patients receiving care in the U.S. Department of Veterans Affairs.

METHODS

We propose a semi-automated method of identifying drug-related adverse events from the EMR with the following steps:

1. Create document set using structured data
2. Create initial list of relevant keywords
3. Extract phrases with keywords and sort them by frequency
4. Remove clearly irrelevant frequently occurring phrases
5. NLP: Keyword search, Phrase extraction, Phrase frequency
6. Add newly discovered keywords and phrases using manually discovered cases
7. Manual review of remaining phrases
8. Detailed chart review for confirmed ADE
9. Confirmed ADEs

INTRODUCTION

Difficulty of finding adverse drug events (ADE) in an EMR depends on the ratio of how often the medication is used to frequency of ADEs, and also on the ratio of how often the medication is documented to frequency of ADE documentation. For example, given a medication that is

1) administered twice a month,
2) prescribed to 1,000 patients,
3) observed over a one-year period, and
4) associated with an ADE published to occur in 1% of patients,

there may be only 10 mentions of an ADE among 24,000 mentions of medication administrations documented in the EMR. The small number of reported cases provides insufficient number of training examples for development of a fully automated solution. At the same time, the number of clinical documents related to medications, where potential ADE are likely to be mentioned, may be too great for manual review.

ACKNOWLEDGEMENTS

This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from the VA Informatics and Computing Infrastructure (VINCI), VA HSR HR 08-204. Additional funding support came from Anolinx, LLC through the University of Utah Center for Scalable Analytics and Informatics, a National Science Foundation Industry & University Cooperative Research Program (I/UCRC) # IIP-1439668.

CONCLUSION

Using an NLP-assisted, semi-automated method for detecting rare ADEs can help overcome the combination of having a limited knowledge base to train a fully automated method and having too many documents for manual review. Future work includes applying this approach on a clinical use cases to test its time saving relative to the traditional approach of chart review or natural language processing.