Semi-automated Dictionary Curation of Symptoms and Events Preceding Suicide Attempts in Clinical Notes

Victor M Castro, MS¹, Sergey Goryachev, MS¹, Vivian S Gainer, MS¹, Yuval Barak-Corren, MD³
Matt K Nock, PhD¹, Ben Y Reis, PhD¹⁴, Shawn N Murphy, MD, PhD¹²⁴, Jordan W Smoller, ScD²⁴,
¹Partners Healthcare, Somerville, MA; ²Massachusetts General Hospital, Boston, MA; ³Boston Children’s Hospital, Boston, MA, ⁴Harvard Medical School, Cambridge, MA

Introduction

Suicide is a leading cause of death worldwide. Prior efforts to predict suicide attempts using electronic health record (EHR) has focused on coded data¹. We aim to enhance EHR-based suicide prediction models by incorporating data from unstructured clinical notes and reports extracted using natural language processing (NLP). Common NLP methods rely on clinical entity extracting using the Unified Medical Language System (UMLS) (https://www.nlm.nih.gov/research/umls/). However, the UMLS contains over 3.8M semantic concepts requiring some curation to reduce dimensionality and sparsity for prediction models. Manual dictionary curation by experts is resource-intensive and is often not comprehensive enough. In this work, we utilize semi-automated extraction of concepts from the literature and other prior work to define a custom dictionary of concepts relevant to suicidality.

Methods

With the aim of capturing and incorporating as many suicide-relevant concepts as possible we utilize a variety of approaches to build a custom NLP dictionary including, (1) selecting relevant UMLS vocabulary subsets; (2) mapping of DSM symptoms and concepts from structured instruments²; (3) automated feature extraction from public sources, (4) incorporating RDoC domain matrix terms³, (5) selecting predictive features from coded suicide attempt prediction models¹ and (6) manual annotation terms by clinical reviewers. A mapping of knowledge sources to UMLS was performed using a combination of string-matching and manual mapping.

The initial NLP dictionary was then used to process a sample of ~21k notes from patients presenting in the psychiatric emergency room or inpatient unit at Massachusetts General Hospital (MGH). We pruned the dictionary to include concepts and synonyms that were mentioned in 1+ note and excluded concepts appearing in 90% of notes.

Results

The initial mapping step yielded 3102 unique semantic concepts and 8255 synonyms. After processing a sample of clinical notes, we pruned the dictionary to 480 concepts and 1273 unique strings. Figure 1 provides a summary of the pipeline and results of the semi-automated curation.

Conclusion

In this work we have curated an NLP dictionary of diverse terms relevant to suicidality for use in machine learning models to predict suicide attempts. Future work will evaluate the performance of models using these NLP terms.

References