Can a machine differentiate between side effects and indications in spontaneous patient reports?

INTRODUCTION

• A key challenge to real world outcomes research lies in analyzing medical conditions in unstructured data. Central to this is determining whether conditions are reported as symptoms of disease, treatment indications or side effects.

• In analyzing texts from patient reporting systems (e.g., FAERS, social media), researchers often use medical dictionaries such as MedDRA to produce ranked frequency counts of medical conditions; however, this approach fails to differentiate treatment indications and side effect conditions, which is a critical part of assessing treatment outcomes.

• Automated classification of medical conditions can provide a deeper understanding of treatment outcomes and promote efficient real world data investigations across a variety of patient level data.

• In our study, we collected 19,313 spontaneous patient reports about antibiotic treatment experiences from medical forums. We annotated 760 sentences containing different types of medical conditions, and classified 5,179 unique conditions into indications and side effects.

• In analyzing texts from patient reporting systems (e.g., FAERS, social media), researchers often use medical dictionaries such as MedDRA to produce ranked frequency counts of medical conditions; however, this approach fails to differentiate treatment indications and side effect conditions, which is a critical part of assessing treatment outcomes.

• Automated classification of medical conditions can provide a deeper understanding of treatment outcomes and promote efficient real world data investigations across a variety of patient level data.

• In our study, we collected 19,313 spontaneous patient reports about antibiotic treatment experiences from medical forums. We annotated 760 sentences containing different types of medical conditions, and classified 5,179 unique conditions into indications and side effects.

• In analyzing texts from patient reporting systems (e.g., FAERS, social media), researchers often use medical dictionaries such as MedDRA to produce ranked frequency counts of medical conditions; however, this approach fails to differentiate treatment indications and side effect conditions, which is a critical part of assessing treatment outcomes.

• Automated classification of medical conditions can provide a deeper understanding of treatment outcomes and promote efficient real world data investigations across a variety of patient level data.

• In our study, we collected 19,313 spontaneous patient reports about antibiotic treatment experiences from medical forums. We annotated 760 sentences containing different types of medical conditions, and classified 5,179 unique conditions into indications and side effects.

• In analyzing texts from patient reporting systems (e.g., FAERS, social media), researchers often use medical dictionaries such as MedDRA to produce ranked frequency counts of medical conditions; however, this approach fails to differentiate treatment indications and side effect conditions, which is a critical part of assessing treatment outcomes.

• Automated classification of medical conditions can provide a deeper understanding of treatment outcomes and promote efficient real world data investigations across a variety of patient level data.

APPROACH

• Each of the 760 annotated sentences were given a binary label indicating whether the condition played the role of an “indication” or “side effect” of the given treatment. Specific treatment and condition mentions in text were replaced by generic text labels (i.e., TREATMENT_CONDITION) to prevent overfitting to the antibiotic drug class.

• Various combinations of syntactic and semantic features were extracted from dependency graphs generated by an NLP toolkit, ClearNLP, and were used for generating a statistical model trained by support vector machines (SVM). This statistical model is used for classifying medical conditions in raw text. Figure 1 shows the overview of our approach.

• Finding Different Types of Medical Conditions: From Data Generation to Automatic Classification

RESULTS

• We used an SVM classifier with 5-fold cross validation and averaged the outcomes of the folds to determine the F1 scores (see Figure 2).

IMPLICATIONS

• We believe this is a feasible approach to differentiating the roles of medical conditions. This approach is suggested for use in spontaneous reporting systems where many of the conditions can be assumed to play side effect or indication roles.

• Our study can enable researchers, regulators, health insurers and pharmaceutical companies to better monitor treatment risk by filtering out instances of indications and automatically identify side effects in spontaneous reports.

• Our approach can provide significant time savings in reviewing spontaneous reports for the purpose of side effect identification, enabling untapped data to be leveraged for real world treatment outcomes research.

• Our follow on work includes applying these models across doctor’s notes in electronic health records (EHRs), case record forms within pharmaceutical companies and signal detection across social media and published literature.