Prediction of Gout Diagnosis on ER admissions

by Michelle Ide

Springboard's Data Science Capstone 2

April 2021

Problem

Emergency Room physicians determine cases of Gout by looking at several factors including the patient's chief complaint. Emergency Room staff, usually a nurse, record the patient's verbal complaint in a text format as described by the patient. The words included in a patient's description of their experiences can give important clues to a proper diagnosis. One barrier in verbal communication is regional dialects and use of vocabulary including slang terms and even the meaning or interpretive meaning within the words. Emergency Room physicians may not originate and therefore be as familiar with the intended meaning and uses of terminology in the region they serve. These communication differences can cause barriers to proper diagnosis.

Proposed Solution

This project proposes to provide a tool, useful by physicians, to alert to potential Gout in the patient they are treating. The tool provides an alert of gout in the form of a probability based on the patient's complaint, alerting the physician to potential of Gout and follow-up on these cases.

Method

A predictive model is built using Natural Language Processing tools from Python in a Jupyter Notebook and stored on GitHub.

Using data from MIMIC III (Medical Information Mart for Intensive Care III) database, two tables are used to train and test the model. Each table contains 3 columns of data:

1: **"Chief complaint"**: text description of patient's exact verbal complaint
2: **"Predict"**: ER physicians determination of Gout (Yes or No or Unknown)
3: **"Consensus"**: The Rhuematologist's feedback of Gout as a factor (Yes, No, or Unknown)

The model will produce a result in the form of a probability, 0% through 100%, that the patient is experiencing symptoms related to Gout where 0% means there is no indication the patient is experiencing Gout and 100% indicates the patient is experiencing Gout. This is not a diagnosis, a probability of 100% still requires follow-up by a physician. This tool is a flag to warn physicians of potential Gout circumstances that may otherwise go unnoticed due to communication barriers, allowing them to further evaluate.

Data

The data contains 2 tables of data that are anonymized using "synthetic" formatting, replacement of identifying information with fake information such as name, address, etc. This synthetic data protects the patients personal information without influencing the model performance.

STATUS & NOTES

Using data collected from the years 2019 and 2020, the complete dataset consisted of 8437 records. Removal of null values resulted in 461 records remaining for use in model development. Two targets, "Predicted", and "Concensus" contained 3 values, Yes, No, and Unknown. During testing we'll look for the Unknowns to fall close to 50% and the Yes and No's to split towards 100 and 0.

Initial tokenization, stemming, and lemmatization were performed removing english stopwords but not contractions (can t, vs can't). Due to the nature of medicine, acronym's are often used for important medical terms as well as shorthand, meaning takes a different twist as well as 'left' when describing the location of abdominal pain can be important. One method I applied to address medical terminology needs and differences was using Named Entity Recognition (NER) using scispaCy libraries.