Capstone 2 Proposal

Springboard Data Science Career Track

Michelle Ide

March 2021

Identification of a Gout attack from a patient's chief complaint

Emergency rooms must treat medical issues without the luxury of developing a communication history that bridges gaps in vocabulary and culture. In this time-sensitive situation a proper diagnosis is made more complex by local slang and terminology influenced by regional customs.

This project will develop a machine learning tool that assists physicians in discovery and diagnosis by alerting them to the potential medical issue based on an individual's chief complaint. This project focuses on recognizing the potential for 'Gout' as an underlying cause that can create an alert with a probability based solely on chief complaint.

Incorporating deep south terms, vocabulary, and slang into the language processing and identification of gout this tool will help physicians from all cultural backgrounds identify and help their patients to identify the underlying causes and treatment necessary. The result will be a machine learning tool to be used by physicians that will alert them to potential medical conditions otherwise normally missed due to misunderstandings in vocabulary, improving a physician's ability to help their patients.

This tool will benefit physicians by reducing the stress of bridging language gaps, improve effectiveness resulting in improved outcomes for the patient, and reduce hospital costs including the expense of late-stage treatments.

Data

The data is acquired from the MIMIC database available at the following link:

https://physionet.org/content/emer-complaint-gout/1.0/

The data includes 2 corpora of free text triage nurse chief complaints collected from 2019 to 2020 at an academic medical center in the Deep South. The training data is from 2019 and contains the keyword "gout". The test data contains 8037 chief complaints collected form a single month in 2020 with no selection criteria (gout or otherwise) applied. This data is proposed to represent the local population of the urban black majority.

These files consist of 3 fields:

1) Text field "Chief Complaint" consists of one or more sentences written by emergency department triage.
2) Predict column indicates if the complaint may be related to a gout flare.
3) Consensus column indicates if the patient at the time of the visit was experiencing a gout flare as determined by chart review by a rumatologist.

   For both 2 & 3 the *markers* indicate the following:

   Y = **yes**

   N = **No**

   U = **unknown**

   - = **unmarked**

Two versions of data are available, redacted and synthetic. I will use redacted data to train an NLP model in the prediction of gout.

Approach

Several predictive ML algorithms will be tested including Naive Bayes to provide percent probability of Gout, and a classification algorithm that flags potential Gout without false negatives.

Deliverables

A machine learning model that provides feedback on potential Gout and percentage probability of Gout based solely on a patient's chief complaint using deep south common phrases and vocabulary.

The results will include a report covering the project for a general audience, both medical and non-medical staff. A technical report with code located on GitHub including data exploration, statistical analysis, and machine learning models. Finally a presentation will be recorded for an audience of hospital administrators, physicians, and medical staff.