Tracking the Evolution of COVID-19 via Temporal Comorbidity Analysis from Multi-Modal Data

Sutanay Choudhury[1], Khushbu Agarwal[1], Colby Ham[1], Pritam Mukherjee[2], Siyi Tang[2], Sindhu Tipirneni[3], Chandan Reddy[3], Suzanne Tamang[2], Robert Rallo[1], Veysel Kocaman[4]
[1]**Pacific Northwest National Laboratory;** [2]**Stanford University;** [3]**Virginia Tech.;** [4]**John Snow Labs**

Introduction

We aim to characterize the evolution in the effectiveness of treatment for different patient groups over the course of the COVID-19 pandemic. In contrast to most existing studies[1], we study the evolution of patient trajectories based on unique sets of frequent comorbid conditions discovered from the data. Further, we study the association between frequent co-morbid conditions to the length of stay (LOS) as a measure of treatment efficacy, for poor COVID-19 related outcomes.

Methods

We conduct our study using a deidentified version of the STARR-OMOP dataset containing both structured electronic healthcare records and free form clinical notes from two Stanford Hospitals. All data is represented in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). We include all COVID-19 cases from February to September of 2020. We represent each patient stay in the hospital as a sequence of sets over time (subsequently referred to as *patient trajectory*), in which each set represents a snapshot of the patient state comprising diagnosis codes, drug orders and laboratory measurements aggregated over time. Each patient state sequence is accompanied by visit level and patient level metadata such as race, gender, timestamp and age. Each admission was assessed to determine if a poor COVID-19 related outcome occurred, which was true in the event of ventilator support, ICU admission or inpatient death. To characterize baseline risk factors, we processed each patient's clinical notes in the first 24 hours of their visit. This included clinical and behavioral risk factors such as hypertension, coronary artery disease (CAD), diabetes, medication record, smoker, hyperlipidemia and obesity. We processed the clinical notes using the entity resolver model from SparkNLP framework to translate the clinical notes to another temporal set-based data representation.

We used a frequent pattern mining method[2] to discover prominent combinations of baseline risk factors, and diagnosis codes from each patient trajectory. Next, we searched each trajectory for matches with top-k (k=30) frequent comorbidity patterns. Each patient trajectory was mapped to an time interval (such as year-month combination). All patients that matched any frequent pattern were first grouped by the time interval, followed by the matching pattern ids. Finally, we computed the median of LOS for each subgroup.

Results

We show descriptive statistics for our COVID-19 cohort in Figure 1(a). There were 454 total admissions, with a mean age of 48.8 years, for 221 male and 233 female patients. The table contains number of admissions for each of the poor COVID-19 related outcomes. Figure 1(b) shows the monthly ranking of baseline risk factors extracted with NLP. Figure 1(c) compares the distribution of LOS over time for patients with a poor COVID-19 related outcome to the rest of the cohort. Figure 2 visualizes the top-k (k=5) frequent co-morbid conditions discovered from data and the associated LOS over time.

Discussion

Figure 1(b) and (c) shows the diversity in the presentation of COVID-19 patients in terms of their clinical and utilization patterns. Specifically, using the information extracted from natural language clinical notes, Figure 1(b) shows that "hypertension", "diabetes" and existing "medications" were consistently among the top-3 baseline risk factors observed. This establishes the motivation to study the combinations of morbidities