Drew Walker

Corpus Collection: Detecting Provider Bias in Clinical Notes of Patients with Sickle Cell Disease

Corpus

We conducted multiple iterations of data collection using different filters on ICD-9 codes for clinical notes within the MIMIC-III data-set. This contains records of over 40,000 patients admitted to Beth Israel Deaconness Medical Center between 2001-2012. Patients with classified with sickle cell disease using ICD-9 codes in the MIMIC-III databases were selected who using multiple strategies (described below). These patients were linked by subject id to their associated clinical notes written by multiple provider types during admissions. These notes were further filtered by removing radiology, ECG and respiratory notes, which provide less narrative-based text that is commonly associated with provider bias.

0.1 Selecting patients with Sickle Cell Disease using ICD-9 codes:

Prior research has sought to improve accuracy by only selecting individuals with either HB-SS disease type or only selecting to patients who were coded as experiencing a vaso-occlusive crisis.[1] Their results for assessing the two major sickle cell trait variants which are associated with vaso-occlusive crises are summarized below:

Using ICD-9 codes to detect Sickle Cell Hb SS indicated relatively high percent positive predictive value (PPV) 81 percent in correctly identifying adult patients with the most severe form of disease from hospital systems in Georgia and California.[1] Much lower accuracy was found in identifying other trait variants, including for **sickle beta thalassemia** with and

without crisis (ICD-9 codes 282.41-42) where only 30 percent of patients were correctly identified with beta thalassemia using hospital admissions data that was cross-checked with state-level SCD surveillance systems.

0.2 Initial corpus collection results

If we restrict the data to only patients experiencing crisis (or the ID codes 282.62,282.64,282.69,282.42) we get a smaller corpus, of only 208 clinical notes for 12 patients over 15 admissions, from 80 unique providers. This corpus in itself likely too small for the context of planned NLP, but may be helpful to identify themes and constructs for ontology development. This corpus is more focused on individuals experiencing pain crises, where much of the research around provider bias and stigma around drug-seeking and pain management is focused. Rates of sickle cell disease in Massachusetts have been estimated to be anywhere form 1000-2499 individuals as of 2019,[2] so this corpus seems very low to cover the range of hospitalizations between 2001-2012.

If we take a more open approach, collecting all ICD-9 codes related to sickle cell disease HB SS, beta thalassemia, HB SC, and sickle cell anemia trait (282.6, 282.60-282.69, and 282.40-282.49) we get a larger corpus of: Total of 2047 clinical notes, for 110 sickle cell patients over 158 admissions, from 455 unique providers. Based off of previous research on the unreliability of ICD-9 codes to predict specific sickle cell genotypes, this corpus may present invalid results, but may capture more real sickle cell patients than the first corpus.

0.3 Conclusions

It is unclear at this point if the data is too small to build meaningful NLP models from. One option to expand the scope of the study could be to assess presence of biases in primarily white genetic diseases and compare bias across diesease groups. Sickle cell disease is genetic illness that primarily affects African Americans and Hispanic Americans. Hemophilia and cystic fibrosis are genetic conditions that have been often compared to sickle cell disease due

to similarities in frequency of hospitalizations and reduced life expectancy. Research has also highlighted that patients with there are significant differences in the quality of care among these populations, where sickle cell patients often face less coordinated and compassionate care.[3] Trying to identify bias differences by provider-identified race could also be possible, but would involve the step in categorizing a patient by race from provider classification within the free-text clinical notes in MIMIC-III, due to the unavailability of structured race or ethnicity data.

Only 11 US states currently participate in any ongoing surveillance efforts to track incidence of SCD.[4] Of these, only Georgia and California have actually published surveillance data on incidence rates from previous years, making this MIMIC-III sample collected from Beth Israel Deaconness Medical Center in Boston, MA particularly hard to verify against any other administrative data.

Given this discrepancy, we will maintain both corpa, and attempt to conduct preliminary qualitative coding and data exploration first with the smaller, likely higher accuracy dataset, and then move to working on the larger corpus.

References

1. Snyder AB, Lane PA, Zhou M, Paulukonis ST, Hulihan MM. The accuracy of hospital ICD-9-CM codes for determining Sickle Cell Disease genotype. *Journal of rare diseases research & treatment.* 2017;2(4):39–45.

2. SCD R. Sickle Cell Disease Resources & Support for HCPs. 2019.

3. Grosse SD, Schechter MS, Kulkarni R, Lloyd-Puryear MA, Strickland B, Trevathan E. Models of Comprehensive Multidisciplinary Care for Individuals in the United States With Genetic Disorders. *Pediatrics.* 2009;123(1):407–412. Publisher: American Academy of Pediatrics Section: Special Article.

4. CDC . SCDC Program Data | CDC. 2021.