**Identifying Alcohol Use Disorder Phenotypes Using Regular Expressions**

Michaela Robinson, Dr. Alvin Jeffery

Data Science, Fisk University

Department of Biomedical Informatics, Vanderbilt University Medical Center

**Background:**

Alcohol use disorder (AUD) is a prevalent but often undetected substance use disorder (SUD) in electronic health record (EHR) data. Effective phenotyping of AUD may aid in early prevention and identification, thereby improving patient outcomes. This project aims to expand the phenotyping of SUD by specifically focusing on AUD within a chronic pain cohort. We used natural language processing techniques to search for key terms aligned with the Alcohol Use Disorders Identification Test (AUDIT), a screening tool aimed to assess alcohol consumption and drinking behaviors.

**Methods:**

We performed a secondary analysis of 8,063 patients with chronic pain. Using Python, we applied regular expressions (regex) to match specific words and phrases within clinical notes. Of the ten questions of the AUDIT Checklist, our team previously developed patterns for the first three questions. For this project, we developed three additional regex patterns from the AUDIT (questions four, eight, and nine). We converted these questions into regex patterns to search for phrases within clinical notes that relate with the question’s subject matter. For example, question nine focuses on injuries to oneself or another after drinking, so the regex pattern we developed was: “\b(.?\b(alcohol|etoh|alchol|drinking\salcohol|drinking\sheavily|drinking)\b.\*?\b(injur(?:e|ed|ing|ies)|hurt|cut|fell|break|broke)\b.?).”After generating previews of matches in the notes, subject matter experts evaluated the matches for performance and accuracy.

**Results:**

Following testing of the regex patterns for questions four, eight, and nine of the AUDIT, we discovered that the concepts described in the questions were rare within the clinical notes. We found two notes that aligned with question nine, one note for question eight, and no results for question four.

**Discussion:**

Regular expressions are powerful tools when searching for simple phrases or words to identify patterns within text. However, searching for more nuanced concepts may prove to be a more complex process. One suggestion would be to incorporate advanced natural language processing techniques, such as machine learning models or large language models, to capture the nuances of the AUDIT questions more effectively. Additionally, the cohort used for this project was based on convenience; it was initially designed for a study on opioid use disorders. Future studies could explore using a different cohort, such as a cohort specifically selected for AUD or one that includes patients with liver conditions, to enhance the generalizability and robustness of the findings. If regex-based automation of AUD phenotyping provides more accurate and reliable results, this method can be used to enhance diagnosis, identification, and prevention within multiple clinical and research use cases.

**Acknowledgements:**

I would like to acknowledge: Dr. Alvin Jeffery, Dr. Kim Unertl, Dr. Lei Qian, Rischelle Jenkins, and Abigail Doyle. This work was funded by NIH NLM grant 5R25LM014216 and Avenir DP1 under Award Number DP1DA056667.