**MEDI Script:** Extracting Insightful Medication Information from Clinical Notes Using Unix/Linux Command-Line Scripts

**Objective:**  
The aim of this project is to develop a pipeline of preprocessing, filtering, and extracting relevant insights from unstructured text like clinical notes within electronic health records (EHRs)

Inspired by this [*Journal of Clinical NLP*](https://doi.org/10.1016/j.jbi.2024.104603):

Extracting adverse drug events from clinical Notes: A systematic review of approaches used:<https://doi.org/10.1016/j.jbi.2024.104603>

**Dataset:**  
I am looking to work with clinical notes of patients who reported at least some case of adverse drug event.

Exploring this n2c2 NLP data from the Harvard Medical School Database [N2C2 NLP Challenge](https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/).

Since most of the data here require credentials to access, I am working on generating some synthetic data using python so that my focus be on demonstrating the concepts learnt in class to derive key insights from clinical notes and generate report tables.

**Methodology:**  
Tokenization, text normalization, and regular expression-based pattern matching, frequency counts using Unix/Linux command-line tools **grep, sed, awk and cut**. Outcome is a data preprocessing pipeline script that outputs report as tables of summary frequencies with Insightful notes.

**Research Questions:**

Can text preprocessing tools help us derive some meaningful insights hidden within a bunch of clinical notes? Some of the questions I seek to answer from the text data are:

1. Of all patients admitted for a particular disease/illness, what is the prevalence of adverse drug events?
2. Which patient demographics (age, gender) are most frequently associated with specific adverse drug events?
3. Which medications are most commonly associated with reported adverse events, and what types of ADEs are most frequently linked to each medication?
4. Is there a correlation between medication dosage and the likelihood or severity of adverse drug events?
5. Does the route of administration (oral, injection, infusion) influence the frequency or type of adverse drug events reported?
6. Are there specific diagnoses that show higher ADE cases for particular medications?
7. Is there a relationship between how frequently a medication is administered (e.g., once daily, twice daily) and the likelihood of adverse events?

**References**

Extracting adverse drug events from clinical Notes: A systematic review of approaches used:<https://doi.org/10.1016/j.jbi.2024.104603>

Adverse drug event detection using natural language processing: A scoping review of supervised learning methods: <https://doi.org/10.1371/journal.pone.0279842>

*Stubbs, A. and Uzuner, Ö. (2015) ‘Annotating longitudinal clinical narratives for de-identification: The 2014 I2B2/UTHealth Corpus’, Journal of Biomedical Informatics, 58.* [*doi:10.1016/j.jbi.2015.07.020*](https://www.sciencedirect.com/science/article/pii/S1532046415001823?via%3Dihub)*.*