CAP 6683 – AI in Medicine and Healthcare

Dr. Marques – Fall 2022

Assignment 3 – Healthcare Data Sources and Formats

Goal

The goal of this assignment is to provide students a semi-structured opportunity to get familiarized with electronic health record (EHR) data, medical concepts, vocabularies, and associated terminology, using the NIH "All of Us" project/site as a reference.

Guidelines

1. Students will work in **self-organized groups of max 5 students**. Individual work is fine.

2. We will use the *NIH All of Us Data Browser* as our main tool: https://databrowser.researchallofus.org/

   "The Data Browser is an interactive tool that allows you to learn more about the data collected as part of the *All of Us* Research Program. You can explore the survey questions and answers and physical measurements taken at the time of participant enrollment. You can also learn more about the electronic health record (EHR) data. The Data Browser will allow you to see how many of the *All of Us* participants have certain conditions, survey responses, demographics, and more."[1]

3. We will use the tool as *Researchers*, i.e., to find information that allows us to develop hypotheses or assess the feasibility of the data set for our studies.

4. Each group will prepare a **report** consisting of three parts:
   I. Answers to the questions listed at the end of this document.
   II. Walkthrough of 2 (two) different explorations of the website.
   III. Summary and lessons learned.

Procedure

1. (OPTIONAL) Read the NEJM Special Report at: https://www.nejm.org/doi/full/10.1056/NEJMsr1809937 to understand the history, context, motivation, and timeline of the "All of Us" Research Program.
2. (OPTIONAL) Read the JAMA Viewpoint at: https://jamanetwork.com/journals/jama/article-abstract/2781166 to gauge how much progress has been made in recent years.

[1] https://www.researchallofus.org/frequently-asked-questions/#data-browser-faqs

3. (OPTIONAL) Look at the effort from the (potential) participants' standpoint: https://www.joinallofus.org/
4. (OPTIONAL) Watch the introductory videos at: https://www.databrowser.researchallofus.org/introductory-videos
5. Download and read the User Guide
6. **Prepare PART I of the report**: answers to the questions listed at the end of this document.
   Hint: Most of the answers can be found in the FAQ page at: https://www.researchallofus.org/frequently-asked-questions/#data-browser-faqs
7. Create/describe/explain 2 (two) scenarios for exploration using the Data Browser as a potential source of data for an AI/ML project.
   NB: You might also have to play the role of the "medical team" in this case.
   a. (OPTIONAL) See publications at https://www.researchallofus.org/publications/ for ideas.
   b. Example: *sleep apnea* (see lecture slides).
8. **Prepare PART II of the report**: describe and explain the main findings of each scenario (including screenshots).
9. **Prepare PART III of the report**: write 1-2 paragraphs with your conclusions and lessons learned.
10. **BONUS opportunity 1 (up to 20 extra points)**
    Explore (up to 4) additional datasets (starting from the links listed below) and report your experience with them:

    - https://healthdata.gov/
    - https://researchguides.dartmouth.edu/healthdata
    - https://guides.lib.berkeley.edu/publichealth/healthstatistics/rawdata
    - https://www.nyam.org/library/collections-and-resources/data-sets/
    - https://dhsprogram.com/data/available-datasets.cfm

11. **BONUS opportunity 2 (up to 20 extra points)**
    Create a "demo" notebook using a dataset in a meaningful way, i.e., to answer a clear (research) question.

    For example, using MIMIC (https://mimic.mit.edu/), determine the median length of stay in the ICU: https://github.com/MIT-LCP/mimic-code/blob/main/mimic-iii/notebooks/ipynb_example/icu_length_of_stay.ipynb

12. **Submit report (single PDF) + additional files** (if you work on Bonus 2) **via Canvas.**

Questions for Part I of your report

1. What is the “All of Us” initiative and what are its main goals and motivation?
2. Where does the data come from?
3. How is patient privacy protected?
4. Knowing what you by now, what would be the motivating factors that could drive one to contribute to the All of Us effort as a participant?

---

5. What are medical concepts?
6. What are vocabularies?
7. What is SNOMED?
8. What are ICD codes?
9. What is the OMOP Common Data Model (CDM)?
10. What do “source” and “standard” mean in this context?