

CAP - 6683 Healthcare Data Sources and Formats

Shaun Pritchard * spritchard2021@fau.edu * Florida Atlantic University * Department of
Engineering and Computer Science



FLORIDA ATLANTIC UNIVERSITY

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.

Part I

1. What is the “All of Us” initiative and what are its main goals and motivation?

NIH's All of Us Research Program is building one of the largest biomedical data repositories of its kind. It is a platform designed to host and analyze health data from a diverse group of individuals from across the US that is part of the All of Us Research Hub. It is possible for registered researchers to access the All of Us data and tools to conduct studies in order to contribute to improving our knowledge of human health.

2. Where does the data come from?

To ensure feasibility and standardization across electronic health record (EHR) data, the All of Us Research Program utilizes the Observational Medical Outcomes Partnership's Common Data Model Version 5 infrastructure. There are 14 EHR tables in the All of Us data set, including Person, Visit Occurrence, Condition Occurrence, Drug Exposure, Measurement, Procedure Occurrence, Observation, Location, Provider, Device Exposure, Death, Care Site, Fact Relationship, and Specimen.

3. How is patient privacy protected?

The privacy of participants is protected in a variety of ways. Data that can potentially identify an individual is called personally identifiable information (PII). Records made available to the public and researchers do not contain any PII, including names and addresses. In addition, all data are rounded up to 20 participants. For example, if only 8 participants have a particular medical condition it will be displayed as 20. Individual data records cannot be viewed on the Data Browser. The Data Browser shows aggregate data for groups of de-identified participants. All of Us program data is stored on a secure, encrypted platform that receives routine updates.

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?

The All of Us Research Program will provide a national resource for guiding thousands of research questions, covering a wide range of health conditions. To build a comprehensive set of biological, environmental, and behavioral data, 1 million or more participants will contribute data from electronic health records (EHRs), biospecimens, surveys, and other measures. The advancements in medicine and technology that could be gained from participants are exponential. Moreover, patients can use their health data to make new discoveries, and be part of new adaptations and innovations in the medical and health field that could potentially save lives.

5. What are medical concepts?

Medicinal concepts describe information in a patient's medical record, such as their condition, the diagnosis of their doctor, their prescription, or the procedure a doctor performed for them. A condition, procedure, drug, and measurement are considered to be EHR domains in the Data Browser. For example, a patient's weight (measurement) is often taken during a routine medical examination (procedure) or a patient may be diagnosed with type II diabetes (condition) and prescribed metformin (drug) to treat the condition.

6. What are vocabularies?

A patient's electronic health record (EHR) may contain medical information that means the same thing but may have been recorded in many different ways. For example, the condition type II diabetes may be recorded as ICD9 code 250.00 at one doctor's office or ICD10 code E11 at another. When All of Us receives a participant's EHR, all of the codes (called source codes) are re-assigned a standard vocabulary code (e.g., for type II diabetes SNOMED 44054006). By changing or mapping all of the source codes to standard codes, the EHR can be more easily categorized and searched by researchers.

7. What is SNOMED?

SNOMED is an acronym for Systematized Nomenclature of Medicine. The SNOMED database connects terminology, medical codes, synonyms, and definitions used in electronic health records (EHRs). There might be an EHR system that uses ICD9 codes and another that uses ICD10 codes, for example. Using SNOMED, data points from multiple EHR systems can be matched.

8. What are ICD codes?

ICD stands for International Classification of Diseases. ICD codes are used in the United States to classify diseases, illnesses or injuries. There are various revisions of the codes, including ICD9 (Ninth Revision) and ICD10 (Tenth Revision).

9. What is the OMOP Common Data Model (CDM)?

The All of Us Research Program employs Observational Medical Outcomes Partnership (OMOP) Common Data Model Version 5 infrastructure to ensure feasibility and standardization across all program data types (physical measurements, electronic health records and participant provided information). Data coming from disparate sources are standardized and stored in a set of formally described tables with defined relationships. This allows data to be accessed and connected in many different ways by researchers.

10. What do “source” and “standard” mean in this context?

SOURCE – electronic health record (EHR) data enters our system with terms and codes for conditions, drugs, and procedures using ‘source vocabularies’. Source vocabularies are the original methods of classifying conditions, diagnoses and procedures (e.g. ICD9 and ICD10CM codes) and will be “mapped” to the new standard vocabularies. However, the source vocabularies are retained after the mapping and data can still be searched using the original terminology or codes.

STANDARD – Translation of clinical findings, symptoms, diagnoses, procedures, etc. from traditional methods of coding and classification into what is referred to as a “standard vocabulary” allow EHRs to be more readily categorized and searchable. Examples of standard vocabularies include SNOMED, LOINC, and RxNorm.

Part II

Scenario 1 - The following example illustrates how a research group might explore new alternative health medicine for a demographic of males and females in order to reduce blood pressure and to learn which age groups within males and females would be the best candidates for a clinical trial to target in order to lower blood pressure. Through the workbench, they could simply do keyword searches for the criteria to meet the research goal by using the All of Us data provided through the workbench. As a result, they could then use the data explorer module to infer data directly within the browser based on the data that they had collected in their respective EHRs between the different groups of participants.

[Home](#) > [Data Browser](#)

Data Browser

The Data Browser provides interactive views of the publicly-available *All of Us* Research Program participant data. Electronic Health Record (EHR) data are derived from reports by health care providers. Genomic data are derived from biosamples provided by participants. Physical measurements are taken at the time of participant enrollment. Data from survey responses and wearables data are collected from participants on an ongoing basis.

In order to protect participant privacy, we have removed personal identifiers, rounded aggregate data to counts of 20, and only included summary demographic information. Detailed data are available for analysis in the Researcher Workbench.

[PUBLIC DATA USE STATEMENT](#)


Search Across Data Types


Q


Keyword Search

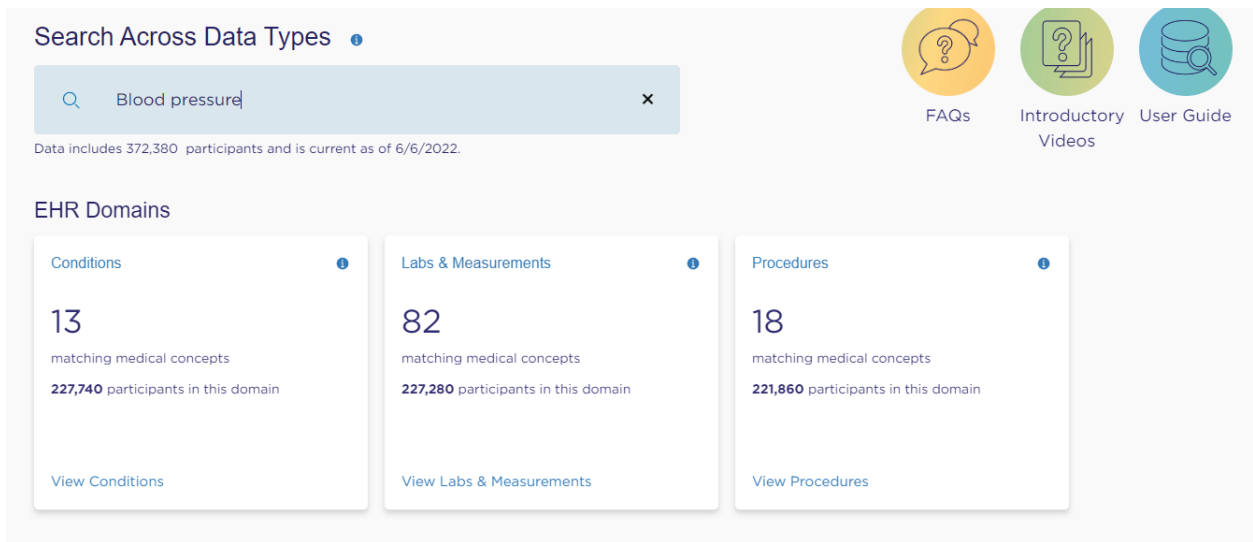
×

Data includes 372,380 participants and is current as of 6/6/2022.


FAQs


Introductory
Videos

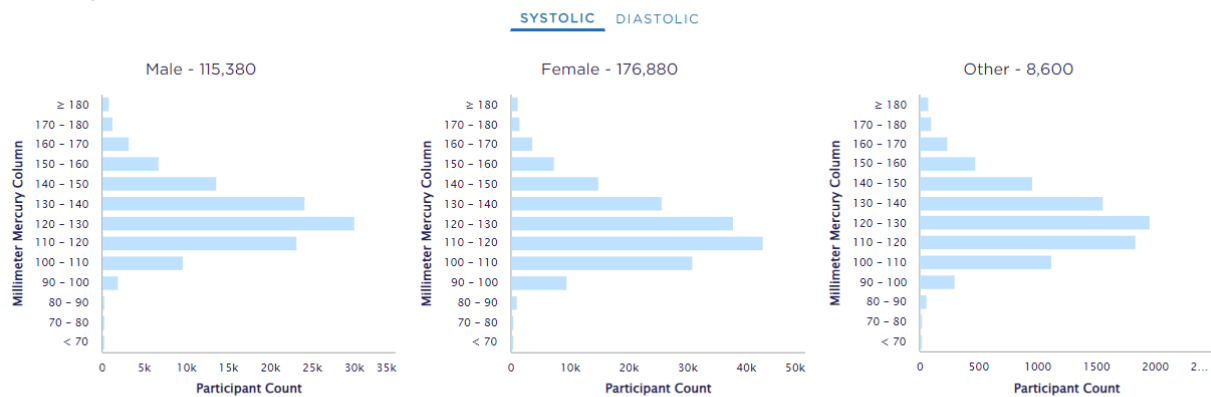

User Guide



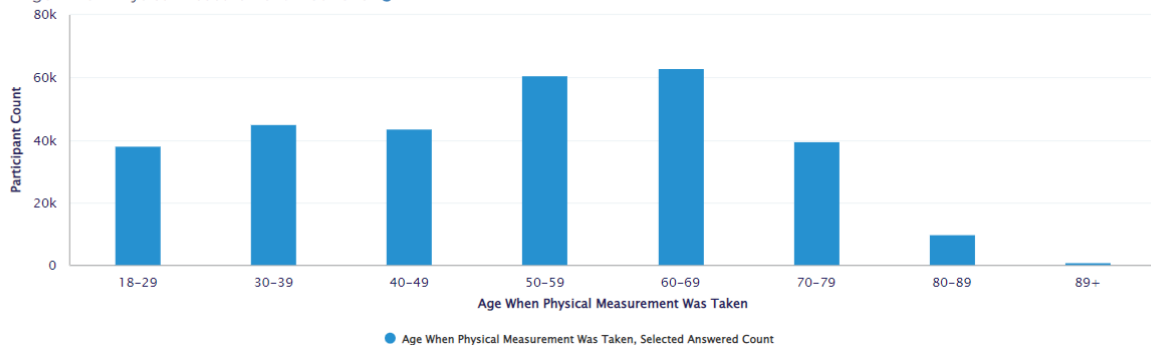
Based on real medical data, these cohorts can provide real insights.

Mean Blood Pressure

Sex Assigned At Birth



Age When Physical Measurement Was Taken



This data indicates that the key target demographic should be males and females between the ages of 50-59 and 60-69. It is also possible to derive marketing information from this since it is

real medical data. Researchers can also use this to develop preventive blood pressure medicine for future generations of adults based on their demographics.

Scenario 2 - Researcher who wants to predict antidepressant response. Antidepressants are often prescribed by healthcare professionals to treat depressive disorders, but their effectiveness varies significantly from person to person. Researchers want to develop an algorithm that aims to determine antidepressant response based on whether an individual stays on, adds to, or switches their antidepressant after a set period of time. Validating the algorithm requires a large, diverse dataset which All of Us now provides access to with real The Patient Health Questionnaire (PHQ) data. You can also use other sources with all of Us data to get more refined accurate research.

Antidepressant Response

SCIENTIFIC QUESTIONS BEING STUDIED

Treatment of major depressive disorder typically begins with antidepressants. Response to antidepressants is highly variable. Around half of individuals will not respond to the first antidepressant they are prescribed, starting a long treatment odyssey to find the a drug or drug combination that works for them. EHRs provide detailed information on the medications individuals take, however, it is not always clear in an EHR how well a patient responds to a particular medication. Sometimes, physicians will administer a survey to a patient that aims to quantify their depression symptoms (patient health questionnaire, or PHQ). The responses to the PHQ are stored in EHRs. At Vanderbilt, we developed an algorithm that aims to infer antidepressant treatment response based on drug switching. We hope to implement our algorithm in the All of Us data and then use PHQ responses to validate how well our response variables track with survey questions on depression.

PROJECT PURPOSE(S)

- Disease Focused Research (major depressive disorder)

SCIENTIFIC APPROACHES

Our approach is to implement our drug switching algorithm for antidepressants and then use PHQ outcomes to determine how well our response outcome tracks with depression symptoms. This will require longitudinal data on antidepressants and PHQ responses.

ANTICIPATED FINDINGS

We hope that our algorithm will be a valid proxy for treatment response that can help increase sample sizes in antidepressant response studies. Future studies could integrate genetic information to determine if there are genetic variants contributing to treatment response. Overall, we hope the algorithm can be used by other EHR researchers and can serve as a paradigm for future treatment response algorithms for other medications.

DEMOGRAPHIC CATEGORIES OF INTEREST

This study will not center on underrepresented populations.

DATA SET USED

Overall analysis of All of Us Research hub

All of Us is a very innovative project with tremendous potential for inferring real patient data and providing researchers with structured data. Using the All of Us platform allows processes and outcomes to continually improve by combining different types of data and training modules. It provides researchers with a secure and safe workbench that provides them with a data browser, data snapshots, data access tiers, data sources, survey explorer, and a lot of other resources to be able to perform their research effectively. As far as I am concerned, I had a little trouble with trying to register if you want access to the Electronic Research Administration (ERA), which I was unable to do for some reason. There were a number of data sets that I was able to play around with and I was able to use the search criteria to do so. It is true that there are a great deal of records and real patient data in this case, however, I am not sure how well big data analysis projects would fare in analyzing such limited and concise data repositories. In spite of this, it is a fairly new system that is growing and currently has sufficient features to be able to take advantage of. So far there have been many different analyses and studies conducted which are all open to registered users and academia. It is especially important to keep in mind that this is a fantastic source for implementing AI for automations, outlier detection, and anomalies for health and medical cohorts in the near future.

Bonus Section I

NYAM - The New York Academy of Medicine Library provides general health information and data sets about demographics. Essentially, it is a repository with links to other data sources. Among the data sources are the Health and Medical Care archive, the Health Cost and Utilization Project, the Health.gov website, the medical provider utilization and payment data, the substance abuse database, the mental health archive, and many more that are hosted on their own websites and then must be accessed and searched to define specific data relevant to a research topic or a set of artificial intelligence data. However, searching for the correct structured data would take much longer than with All of Us.

Berkeley - we had access to a variety of good data sources, including the morality data of the California tobacco survey, the California open data portal, the California Department of Health Care, data.gov, health and medical care archive data, demographic population data, and many others. Similarly to the other source being reviewed, Berkeley offers links to access these other data repositories that require individual permissions and access based on what you wanted to use for your study, but again, it is nowhere near as organized or structured.

HealthData.gov - With the goal of improving health outcomes for all, it claims to make high value health data more accessible to entrepreneurs, researchers, and policymakers. A wide range

of data is provided, including demographics of patients at community hospitals as well as Open Access Data. Even though it is limited to only specific sources and demographics now, it still has a similar look and feel to "All of Us", which is based on real EHR data from patients, unlike HealthData.Gov. The tool offers a number of useful features, including search capabilities, data lens features, and inferred dataset visualizations. Nevertheless, this data can be useful when it comes to really broad topics.

DHS Program - It is all in regard to demographic and health survey data, what is a really good source that has a really good website and features that allows people to search for different data sets. It is very broad, as it has access to publications based on the data methodology and uses lots of tools and resources for researchers to research on topics such as child mortality, family planning, gender, malaria, nutrition, wealth index, and so on. There are data sets provided by DHS that are pre-imputation, which means they have tried to clean the data and this could mean that, for certain studies, you would not be able to find or detect outliers in the data. Overall, there are some limitations to the source of data topics.

Conclusion

Compared to all four databases that I inquired about, we're limited in scope when compared to all of them. There was no patient data access to EHR records and data about specific diseases, specific traits, and stats (blood pressure, cholesterol level, etc.), nor did they offer detailed genetic information. Many of these resources should be accompanied by at least two other resources to which you must have specific access, while others have been used as a prequel in various formats that may be difficult to work with.

Bonus Section II

This machine learning analysis uses MIMIC II data set to predict readmission of patients. There are five main causes of readmission, according to various studies: A list of predictors of a patient's readmission. To get more information about patient-level characteristics, we will group 41 ethnicities, but the number of subjects for each type is relatively low. Hence we will be combining some of the ethnicities to get a better representation while just slightly affecting precision.

Then Discharge location into three categories Medical Facility, Home and Others. This will help us maintain a significant count for each category. In order to calculate the number of days until the next admission, subtract the discharge time from the next admission time, and remove any admission event that is related to a death.

transfer 55186

admit 32535

Name: EVENTTYPE, dtype: int64

Our patient cohort should not reflect new borns as we need to develop unplanned medical care, therefore, they were filtered out of our patient cohort. After updating the ICU data, we implement a model to categorize patients based on the events recorded in the charts.

Chart Events:

- Systolic BP :- ['mmHg']
- Diastolic BP :- ['mmHg']
- Respiratory Rate :- ['insp/min' 'BPM']
- Glucose Levels :- [nan 'mg/dL']
- Heart Rate :- ['bpm' 'BPM']
- Temperature :- ['?F' '?C' 'Deg. F' 'Deg. C']

Afterwards, we can merge the data, drop outliers and null values, and test train split the data according to the outliers and nulls. As soon as the testing and training are complete, we will be able to implement patient representation, cohort discovery, and use a multitask learner to predict readmissions once the testing and training are complete. Below is an overview of the results.

▼ Predicting Readmission MIMIC III data

▼ Bonus Assignment

- Shaun Pritchard
- CAP -6683 AI In health and Medicine
- 10/25/2022

SPritchard_10252022_CAP6686_MIMICIII_DATA.ipynb

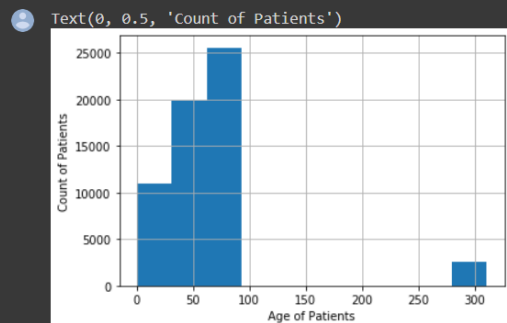
```
[ ] 1 import pandas as pd
    2 import numpy as np
    3 import matplotlib.pyplot as plt
    4 import random
    5 # Text Processing
    6 import re
```

```
1 # Import Datasets
2 admissions = pd.read_csv("../Data/ADMISSIONS.csv", index_col = None)
3 patients = pd.read_csv("../Data/PATIENTS.csv", index_col = None)
```

```
[ ] 1 # Convert all the date columns
    2 admissions.ADMITTIME = pd.to_datetime(admissions.ADMITTIME, format = '%Y-%m-%d %H:%M:%S', errors = 'coerce')
```

```
[ ] 1 # Sort the dataframe based on admittime and subjectid to visualize the post-admission journey
2 admissions = admissions.sort_values(['SUBJECT_ID','ADMITTIME'])
3 admissions.reset_index(drop = True, inplace = True)

1 # Let's calculate a patient's age based on the patient's admitted time and DOB from the patients table
2 patient_age = {row[1]: row[2] for row in patients[['SUBJECT_ID','DOB']].itertuples()}
3 admissions["AGE"] = [int((adm_time.date() - patient_age[subj_id].date()).days/365)
4 | | | | | | | | | | for adm_time, subj_id in zip(admissions["ADMITTIME"], admissions["SUBJECT_ID"])]
5
6 age_plot = admissions.AGE.hist()
7 age_plot.set_xlabel('Age of Patients')
8 age_plot.set_ylabel('Count of Patients')
```



We can see from the above histogram that most patients are between the ages of 100 and 300, but there are some older patients. where >89 Years old = 300

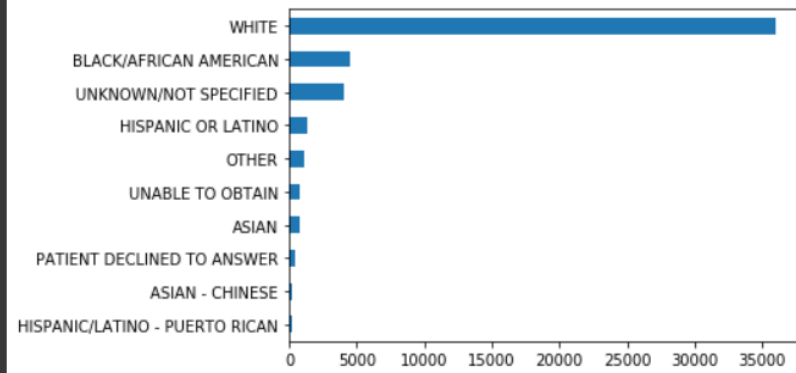
```
[ ] 1 # randomly spread these patients within age range of 90 to 100
2 admissions.loc[admissions.AGE >= 300,"AGE"] = random.choices(list(range(90,100)),k = sum(admissions.AGE >=
3 # remove all the young patients as chance of readmission to low
4 admissions = admissions[admissions.AGE >18])
```

```
[ ] 1 def normalize_ethnicity(x):
2     """Normalize Ethnicity into "WHITE", "HISPANIC", "ASIAN", "BLACK" and "OTHERS"
3     """
4     if "WHITE" in x:
5         return "WHITE"
6     elif "HISPANIC" in x:
7         return "HISPANIC"
8     elif "ASIAN" in x:
9         return "ASIAN"
10    elif "BLACK" in x:
11        return "BLACK"
12    else:
13        return "OTHERS"
```

```
[ ] 1 admissions.ETHNICITY.value_counts().head(10).sort_values().plot(kind = "barh")
```

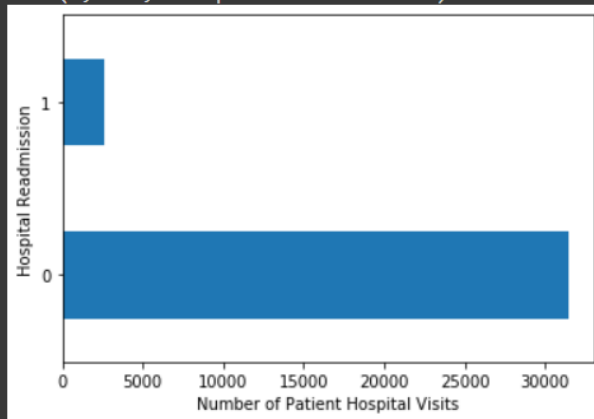
```
1 admissions.ETHNICITY.value_counts().head(10).sort_values().plot(kind = "barh")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19a6ddef048>
```



```
1 visit_plot = train_data.IS_READMISSION.value_counts().plot(kind = 'barh')
2 visit_plot.set_xlabel('Number of Patient Hospital Visits')
3 visit_plot.set_ylabel('Hospital Readmission')
```

```
Text(0, 0.5, 'Hospital Readmission')
```



Final Results of training set:

