Project Report/Documentation

Group 5: Wild West Wranglers
Metabolic Syndrome: Integrating Clinical, Fitbit, and Genetic Data
Link to All of Us Workspace

Project Overview

Ethical Conduct of Research Policy for All of Us Data

The All of Us Research Program mandates all researchers must complete ethical principles rooted in the Belmont Report's core tenets: Respect for Persons, Beneficence, and Justice[2][3]. Researchers must comply with all applicable laws, prioritize participant well-being, ensure equitable distribution of research benefits/harms, and account for potential non-physical risks[2][4]. Special protections apply to vulnerable groups, including children, pregnant women, and historically underrepresented populations[1][3].

Data Access Process for Registered Tier

To access individual-level data (Registered Tier), researchers must:

1. Complete Training: Finish two mandatory courses—(1) Responsible Conduct of Research (RCR) training covering ethical frameworks for sensitive data and diverse populations, and (2) Data Use Agreement compliance training[5][8].

2. Verify Identity: Submit identity documentation; some team members encountered verification challenges requiring resolution with institutional support[8].

3. Sign Code of Conduct: Agree to privacy safeguards, including prohibitions against re-identification and misuse of sensitive data[4][7].

4. Establish Workspace: Create a project workspace with a publicly searchable description for audit purposes[6][9].

After completing the required training, a shared workspace was created for teammates to collaborate on the project. To begin, cohorts were created (Cohort 0) requiring participants to have Fitbit data and short-read whole genome sequencing (WGS) data as well as a metabolic syndrome (MetS) diagnosis. A larger cohort (Cohort 1) was also created that required participants to have diagnoses for a number of conditions associated with MetS rather than a MetS diagnosis. However, Cohort 1 was not used due to the resource costs when working with Fitbit and genomic data on a cohort of that size. Concept sets were then created for conditions, lab measurements, physical measurements, and optionally medications. These concept sets enabled various participant metrics to be included in the datasets. Using the cohorts and concept sets, separate datasets were created for each type of data so the team could split up the project workload and each member could choose a dataset they wanted to wrangle. Using Cohort 0, datasets were created for demographics, conditions, lab measurements, Fitbit heart rate summary, Fitbit activity summary, Fitbit sleep daily summary, and each single nucleotide variant (SNV) of interest. Once each dataset was adequately wrangled, they were saved to the workspace as a CSV file and aggregated into a final dataset within the 'main_notebook.ipynb' under the Analysis tab.

Main Notebook

The 'main_notebook.ipynb' loads all of the wrangled data, merges it into a single dataset, and provides a descriptive summary. All files saved to the data folder in the workspace bucket are printed to ensure they exist. A function was created to load all of the cleaned data

into a dictionary of data frames. Using a left join on ‘person_id’, each data frame is then merged to a data frame containing all ‘person_id’s in Cohort 0. A descriptive summary is then provided that shows counts, means, standard deviations, etc.…as well as functions for displaying the percentage of missing data by feature, categorical distributions, and numeric distributions.

Physical Measurements

To wrangle participants’ physical measurements, the notebook ‘cohort_0_physical_measurements.ipynb’ was created. In this notebook data for blood pressure, body weight, BMI, waist and hip circumference were extracted. The initial data frame was in a longitudinal format and contained a number of unnecessary columns. The data was pivoted so each column represented a measurement and each row represented a participant. Column names and values were then cleaned up prior to exporting to the workspace bucket as a CSV file.

Genetic Variants

Due to high resource costs of working with WGS data, Cohort 0 participants were filtered for each of the ten SNVs using the cohort builder tools and saved to a variant specific cohort and dataset (i.e. fto var). FTO, MC4R, and LEPR variants are linked to BMI and obesity, energy balance and weight regulation, and fat distribution and appetite regulation, respectively. TCF7L2, PPARG, and IRS1 variants are associated with insulin resistance and type 2 diabetes, insulin sensitivity and metabolic health, and insulin resistance and glucose metabolism, respectively. APOA5 and LDLR variants are linked to triglyceride metabolism and cardiovascular risk, and lipid metabolism and cardiovascular health, respectively. The ACE variant affects the renin-angiotensin system which is crucial for blood pressure regulation. Last, the IL6 variant is linked to chronic inflammation and metabolic disorders and may be relevant to the systemic inflammatory processes in MetS.

To wrangle participants’ SNVs, a notebook was created from each variant cohort following the ‘*_var.ipynb’ naming structure. In these variant notebooks, the ‘person_id’ was extracted for each participant and a binary encoding was added to indicate the variant is present in these participants. The data was then exported to the workspace bucket as a CSV. In the ‘main_notebook.ipynb’, each of the ten variant CSV’s were merged to the final dataset on ‘person_id’. If a ‘person_id’ was in the final dataset but not in a variant dataset then its value was missing. Missing variant values were filled with a zero and the variant columns were encoded as categorical 1’s and 0’s.

Lab Measurements

The lab measurements data contained ‘Person ids’ followed by multiple columns explaining the test performed, type of test, type of visit, date and time performed, etc. Because one patient is not limited to one test, there were a total of over 40,000 rows explaining each test that a person had. There was a lot of missing, redundant or unneeded data that was removed prior to wrangling. In order to merge the data with other datasets, the number of rows had to be equal to the total patients in the cohort. To wrangle the data to ensure that each person had one row, the data was turned into a long format and duplicated any columns for each test that a person had done. Because not every patient had every test, a “secondary cleaning” was done that dropped any columns that had at least 75% of columns as missing data. To further ensure clarity and simplicity in use, the columns for “unit source” were dropped and the units for each test were transferred to the column header for that specific test in which the units are noted. The new data frame was then transferred to the ‘main_notebook.ipynb’ and merged to the main dataset on ‘person_id’.

Project Report/Documentation

Demographics

General information of participants is included in the demographics dataset, containing person_id, date of birth, race, ethnicity, sex at birth and the concept id of each data. All rows do not contain null or duplicated values. Age of each participant was used datetime to calculate from year of birth and current year. The data was filtered by person_id, date of birth, race, ethnicity, gender and sex at birth to merge with the main notebook. The primary key is 'person_id' which is used for joining the dataset together.

Conditions

This dataset contains time-series data, emphasizing on standard concepts of patients , such as pure hyperglyceridemia, prediabetes and essential hypertension. Therefore, the total rows of this dataset is around 56,000. There are numerous null values in concept end datetime, stop reason, visit occurrence id and condition status source value/id/name, however, the main defined condition in the dataset does not contain any null values.The latest values of each patient were selected using group by and filtered by the interested columns namely person_id, standard concept and condition start datetime prior to merging to main notebook with person id.

Fitbit Activity Data

The Fitbit activity data included daily metrics such as activity calories, total calories burned (calories_out), various activity intensity minutes, and step counts, all linked to individual person_ids. The data retrieval process applied strict cohort filters to include only participants with whole genome sequencing data, Fitbit device data, and relevant clinical events, ensuring alignment with the study population. Initially, a broad set of activity metrics was extracted, but subsequent cleaning focused on a subset of key columns—person_id, date, activity_calories, calories_out, fairly_active_minutes, lightly_active_minutes, steps, and very_active_minutes—after identifying and removing uninformative or incomplete fields. The filtered dataset showed no missing values, confirming data completeness. To facilitate patient-centric analyses, the data was aggregated by person_id, calculating mean values for each activity metric, thereby summarizing daily records into a single profile per participant. This approach produced both detailed time-series data and patient-level summary statistics, enabling flexible analysis of activity patterns across the cohort. The integration of genomic and Fitbit device filters enhances the dataset's relevance for studies exploring genotype-phenotype relationships through physical activity metrics.

Fitbit Sleep Data

Sleeping record tracking from fitbit contained the record from 2019 to 2022. There is time-series data in minutes which contains the sleep record per night, divided into sleep phase and restlessness phase. In the sleep phase, there are minute in bed, asleep, deep, light, rem and wake, whereas the restlessness contains only a restless column. To align with cohort structure, the dataset was separate into sleep dataset and restless dataset by the interested columns. Each dataset did not contain null values or duplicate. After grouping the data by using patient id, the mean of all data was calculated and merged to the main notebook.

Fitbit Heart Rate Data

The FitBit Heart Rate data took daily measurements of the amount of time spent in 4 heart rate zones, Out of Range, Fat Burn, Cardio, and Peak. It recorded the minimum and maximum heart recorded in each zone during the daily measurement and recorded the amount of calories burned. The data was grouped by the person id for the numerical categories getting the mean, standard deviation, min, max for minimum heart rate and maximum heart rate and the

mean for calories and time in zone. Next, the data was grouped by person id and zone to get the sum of minutes for each individual zone. The data was joined by doing a left merge on person id between the numeric data and zone data. Finally, the data was joined by a left merge to the final notebook on patient id.

Citations:

1. https://support.researchallofus.org/hc/en-us/articles/22346846933396-Policy-on-Ethical-Conduct-of-Research
2. https://www.researchallofus.org/faq/ethical-conduct-of-research-policy/
3. https://tuskegee.libguides.com/c.php?g=1350060&p=9986626
4. https://research.sdsu.edu/research_affairs/allofus_data_user_code_of_conduct.pdf
5. https://elsihub.org/video/all-us-responsible-conduct-research-training-fostering-ethical-and-socially-responsible
6. https://www.researchallofus.org/faq/data-access-framework/
7. https://www.joinallofus.org/privacy-safeguards
8. https://allofus.nih.gov/news-events/announcements/all-us-research-program-updates-data-use-eligibility-propel-precision-medicine
9. https://www.researchallofus.org/data-tools/data-access/