Healthy Aging Signal Research
Statement of Work

Partner:
Merck & Co.

Group Members:
Eleonora Shantsila: eshantsila@g.harvard.edu,
Yaxin Lei: yaxin_lei@g.harvard.edu,
Aaron Jacobson: aaronjacobson@g.harvard.edu ,
Daniel Cox: daniel_cox@g.harvard.edu

Project Background:

Merck is one of the world's largest pharmaceutical companies. It invests billions of dollars annually in research areas such as oncology, vaccines, and infectious and neurodegenerative diseases. We are working with Data scientists from Merck who are interested in the process of healthy aging, as nearly all the diseases listed above are more likely to develop with age.

A large portion of age-related research has been focused on neurodegenerative disease. Most of this work has been aimed toward identifying molecular, genetic or imaging biomarkers that are predictive of disease. Far less work has been done to understand the signature of healthy aging itself. The identification of such biomarkers, however, could serve as an essential baseline for studies of disease development and therapeutic response.

Several large databases have developed around age-related diseases. These databases typically contain diseased subjects and healthy controls. By focusing on data from the healthy controls we aim to bring together data that can be used in concert with machine learning models to identify indicators of healthy aging.

Problem statement:

Our goal is to help our partner find important healthy aging signals through a wide range of datasets. In the initial weeks, a significant amount of work may be focused on narrowing down selections of high quality accessible dataset (for example, imaging data, time series data or other types of data). We will also need to come up with a set of data processing procedures, especially if our datasets concern large size MRI scans over a span of time. Some questions we attempt to answer include:

- What are general signals of aging (healthy or unhealthy)?
- How are these signals manifested in lab imaging data?
- How do we separate healthy aging signals from unhealthy ones?

At the moment, the imaging datasets are too large for traditional tools like AWS to handle and the research question is still pretty open ended. We will for sure try to remedy these and narrow down our focus along the way, for example, we are considering finding representative portions or slices of MRI images that we can use in place of the original set. We then plan to use this data to extract some kind of signal that denotes healthy aging by starting with interpretable simpler models, then refining signal extraction through more complicated models with potential loss of interpretability.

Deliverables:

- Identified a list of important markers of aging. (May be MRI image characteristics or personal traits)
- A well-documented github codebase that contains:
    - Clean setup for loading/compressing/interpreting MRI and lab imaging data;
    - Exploratory data analysis report (containing visualizations of for example demographics info).
    - An accurate model to verify that the identified markers of healthy aging are indeed correlated to, or causing healthy aging.
- A final report documenting our thought process, data, resources and results.
- There may be one or more milestone powerpoints highlighting thought process, test and trials, and milestone progress.
- A well documented list of data sources.

Tentative Timeline:

Milestone 1:

1. Access to all four databases mentioned in Merck's initial project introduction
    a. If we find better databases along the way, or the above datasets are unavailable, we may need to search for additional data resources.
2. Setup the database ready with proper data;
3. Figure out ways to properly handle/compress/analyze lab imaging datasets as they are very large for normal computational tools to handle. (For example, we are currently discussing on how to automatically select representative 2D slices of the MRI image from the whole file for each individual)
4. Conduct preliminary analysis on these databases;
5. Finish reading related literature and familiarize with previous methodology

Milestone 2:

1. Starting in this phase, we should have a very clear goal of what imaging datasets to use and how they can be processed, given the problem statement for this project is pretty open ended.
2. During this phase, we should have a list of very clear possible indicators of aging (such list of indicators may or may not be fully interpretable).

Milestone 3:

1. In this phase, we narrow down the list of indicators for healthy aging.
2. We will also attempt to prove that these are accurate indicators by hopefully having an accurate model to predict age or predict health state given these indicators.
3. Finish the report and all documentation work.

Current Roadblocks/Discussion Points:

1. **Data Access:** This is one of the biggest roadblocks. It is hard to gain access to the recommended list of data, most need affiliation, NDA or approval.
2. **Project Goals:** This was a question that appeared during discussion with our fellow partner Merck. Should we conduct unsupervised learning, for example find commonalities within the feature space after we separate individual data by their age cohort? Or should we conduct supervised learning, for example find a list of accurate features to predict age?
3. **Data Selection:** Do we want to focus on lab imaging data, time series data or data with other labelings? This may significantly influence our model selection and data analysis methods.
4. Data processing and computation concerns: If we use high quality MRI data, the coordinates we choose to do slicing/cropping will be time consuming. The size of data is also of concern, as storage, transfer, processing, training these data may all be difficult.

Dataset:

Primary dataset

The primary dataset we are looking at is the HCP Lifespan Aging dataset available from human connectome of subjects aged 35+. The latest release contains unprocessed data, including structural MRIs, resting state fMRIs, task fMRI and diffusion MRI for 689 HCP-A subjects, minimally preprocessed structural MRI data for 128 subject and demographic information for all HCP-D subjects. The directory structure of the data is described here.

In order to access the dataset we will need to create an NDA account (National Institute of Mental Health Data Archive) and submit a request for access. This is subject to fulfilling the following criteria:

1. You must have a research-related need to access the data
2. You must be associated with an NIH-recognized research institution, defined as an institution registered in the NIH electronic research administration system (eRA Commons,) and have the approval of an authorized signatory official of that institution.
3. Your institution must have an active Federalwide Assurance (FWA).

Data access can be granted through an Oracle database mindar hosted on AWS.

Secondary dataset

Additional datasets we will explore ahead of milestone 1 include:

1. ADNI dataset
    - The application process includes acceptance of the Data Use Agreement and submission of an online application form. The application must include the investigators institutional affiliation and the proposed uses of the ADNI data. ADNI data may not be used for commercial products or redistributed in any way.
    - Dataset contain data on subjects aged 55+ of whom 229 have been categorised as 'normal'. The mean age of the normal group is 76.
2. PPMI dataset
    - To access the data investigators must submit an online application, which requires signing the Data Use Agreement and compliance with the study Publications Policy. Investigators using PPMI data will be asked to provide annual updates on the analyses they have performed. This information will be displayed publicly on the PPMI Web site on an Ongoing Analyses page. Investigators will also be asked to provide new data generated using PPMI data back to the Data and Publications Committee so that it can be integrated into the database for use by future investigators.
    - MRI scans in the dataset include DTI and fMRI. This is in addition to a large number of other test results collected including blood, urine and clinical data.
    - 200 healthy controls were enrolled in the study
3. AD Knowledge Portal
    - In order to access the data we must register for a Synapse account. AD Knowledge Portal data is hosted on the data sharing platform, Synapse. In order to access and use the data you will need to create a Synapse account and agree to the Synapse Terms and Conditions of Use. Investigators must also agree to acknowledge use of the data in any publications. Use of the data within the AD Portal requires acknowledging the data contributors.
    - The portal constraints access to various datasets, of which two are listed as having MRI data, the Emory Vascular Study and DiCAD study.