Assignment # 3: Healthcare Data Sources and Formats
Course: CAP 6683 Artificial Intelligence in Medicine and Healthcare
Professor: Dr. Oge Marques
Student Number: Z23596812
Student Name: Renee Raven
Term: Fall 2022

## Part I: All of Us Q and A

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

The All of Us initiative is a program run by the National Institutes of Health. It began in 2015 with $130 million dollars of funding. All of Us seeks to use a database of information from a wide diversity of people to inform medical research and precision medicine so that medical care may be individualized. It addresses the growing understanding that "Treatments meant for the 'average' patient may not work well for individual people." Specific goals include (list taken from website):

- Know the risk factors for certain diseases
- Figure out which treatments work best for people of different backgrounds
- Connect people with the right clinical studies for their needs
- Learn how technologies can help us take steps to be healthier
- Takes into account your environment (where you live), lifestyle (what you do), and your family health history and genetic makeup
- Gives health care providers the information they need to make customized recommendations for people of different backgrounds, ages, and regions
- Helps you get better information about how to be healthier
- Reduces health care costs by matching the right person with the right treatment the first time

2. Where does the data come from?

"The data in the *All of Us* Data Browser comes from participant electronic health records and from survey answers and physical measurements taken at the time the participant enrolls in the *All of Us* program." Source: AllofUS FAQ

3. How is patient privacy protected?

“Participant privacy is protected in multiple ways. Personally identifiable information (PII) is any data that could potentially identify a specific individual. All PII, such as names and addresses are removed from participant records made available to the public and researchers. 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. It is not possible to view individual data records on the Data Browser. The Data Browser shows aggregate data for groups of de-identified participants. *All of Us* program data are stored on a secure, encrypted platform that receives routine updates.” Source: AllofUS FAQ

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?

I appreciate the core values of openness, transparency, security, and privacy. If I believed my health records reflected unique characteristics and could contribute to the diversity of the pool, I would become a participant. I believe many people with rare/orphan conditions would be motivated to participate because it could drive research interest/dollars towards their conditions.

5. What are medical concepts?

“Medical concepts are similar to medical terms; they describe information in a patient’s medical record, such as a condition they have, a doctor’s diagnosis, a prescription they are taking, or a procedure or measurement the doctor performed. In the Data Browser we refer to conditions, procedures, drugs, and measurements as electronic health record (EHR) domains. 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.” Source: AllofUS FAQ

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.” Source: AllofUS FAQ

7. What is SNOMED?

“SNOMED stands for Systematized Nomenclature of Medicine. SNOMED connects the various terminology, medical codes, synonyms, and definitions used among different electronic health records (EHR). For example, one system might use ICD9 codes while

another EHR system uses ICD10 codes. SNOMED allows the same data point from multiple EHR systems to be matched up." Source: AllofUS FAQ

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)." Source: AllofUS FAQ

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 (see **What do "source" and "standard" mean?** above) 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." Source: AllofUS FAQ

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

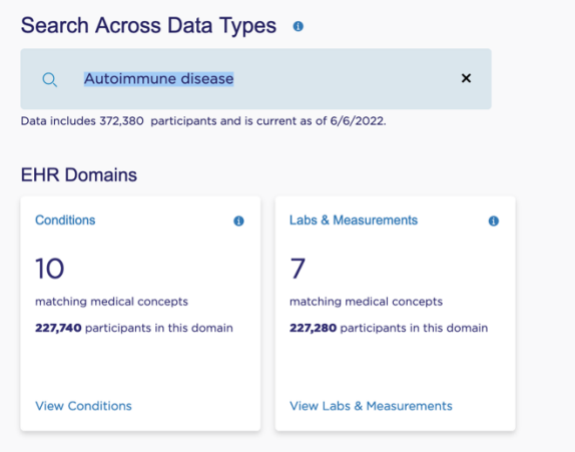
"SOURCE – electronic health record (EHR) data enters our system with terms and codes for conditions, drugs, and procedures using "source vocabularies". Source vocabularies are the original methods of classifying conditions, diagnoses and procedures (e.g. ICD9 and ICD10CM codes) and will be "mapped" to 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.

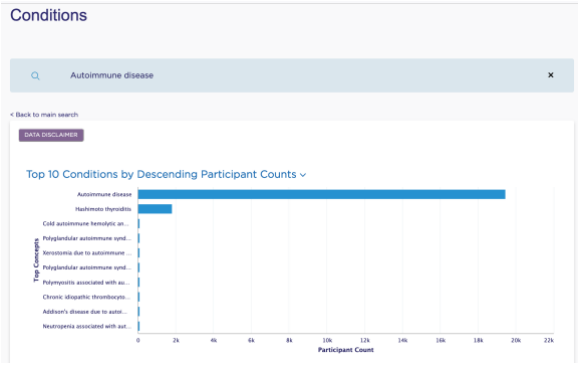
STANDARD – Translation of clinical findings, symptoms, diagnoses, procedures, etc. from traditional methods of coding and classification into what is referred to as a "standard vocabulary" allow EHRs to be more readily categorized and searchable. Examples of standard vocabularies include SNOMED, LOINC,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

## Part II: Walkthrough of 2 different explorations of the website

First Walkthrough Condition: Autoimmune Disease

Potential medical question: Is there enough data to support an exploration of "Does estrogen imbalance lead to females having more autoimmune diseases?"



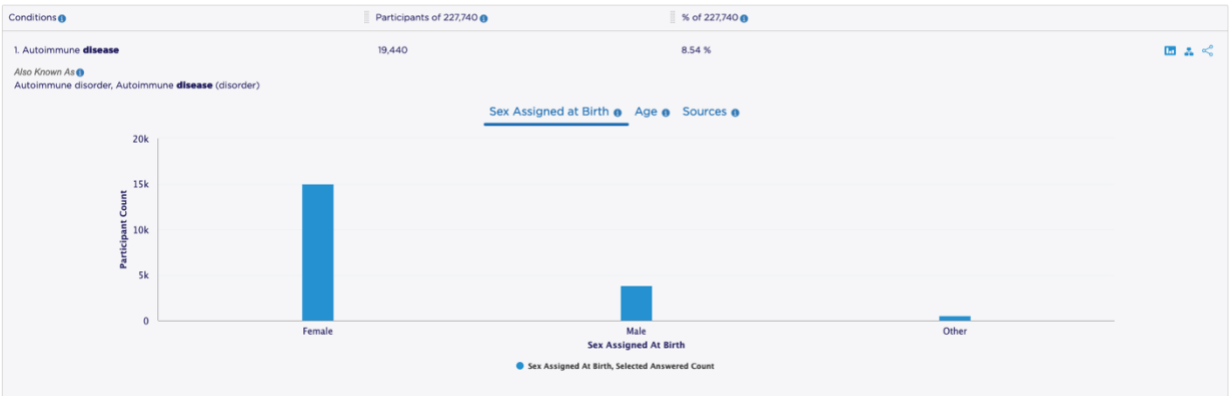




Close to 20k participants (or 8.54% of the total AllofUs 227,000 + participants who responded to this question) are associated with an Autoimmune disease. Over 1,700 participants are associated with Hashimoto thyroiditis. The other specific conditions had small representations (40 or fewer participants).

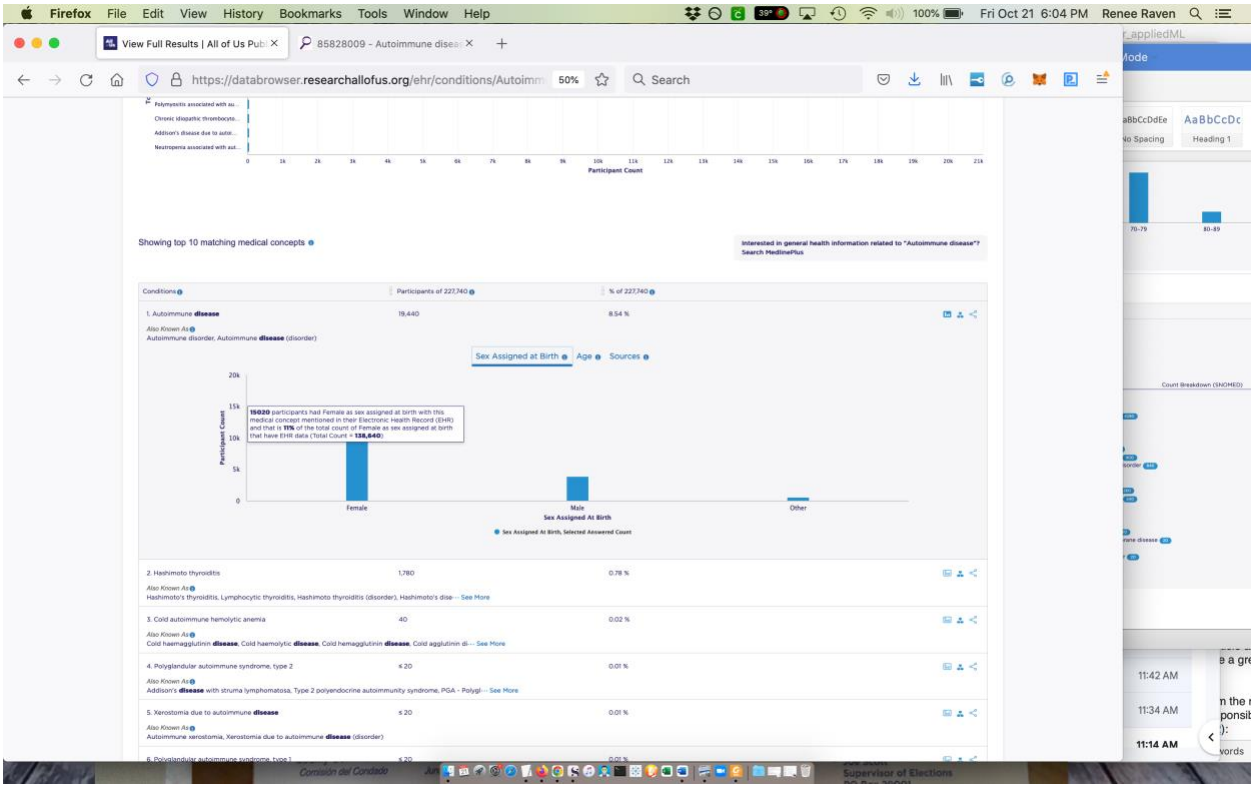
Showing top 10 matching medical concepts

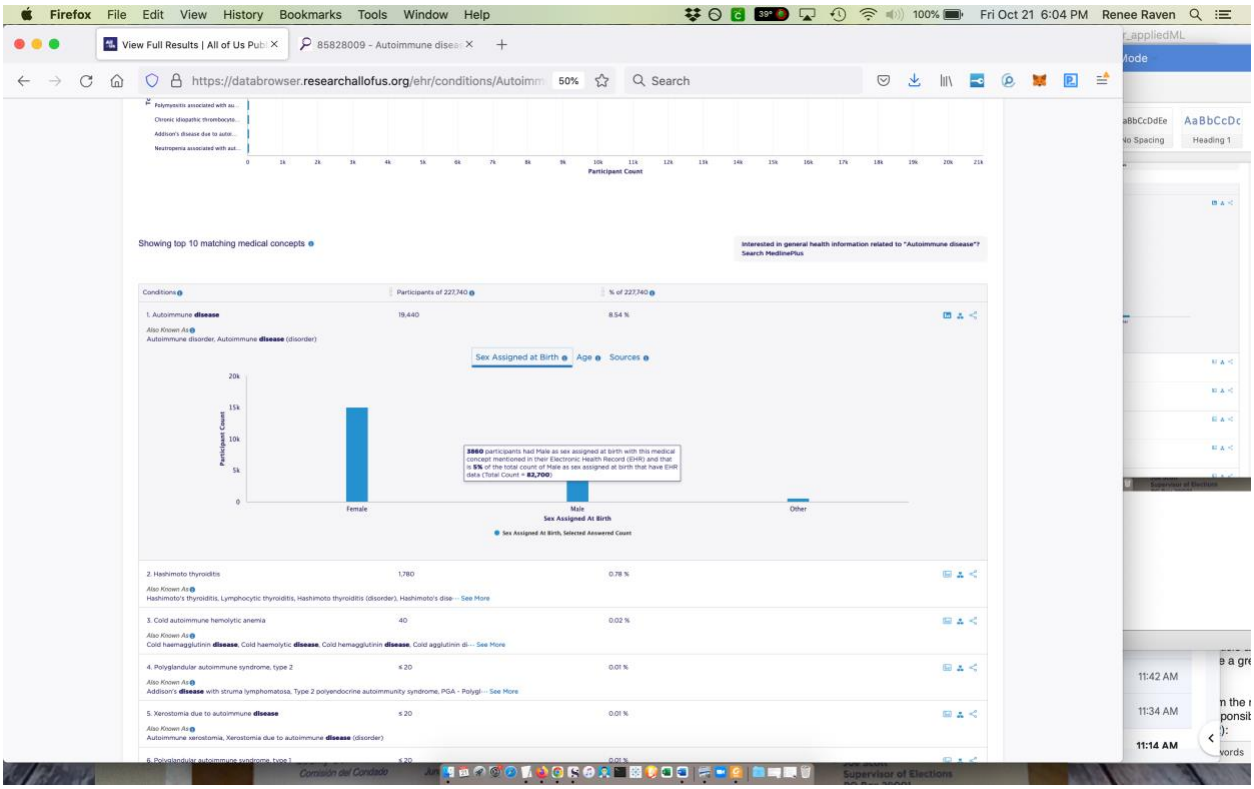
Interested in general health information related to "Autoimmune disease"? Search MedlinePlus

| Conditions | Participants of 227,740 | % of 227,740 |
|---|---|---|
| 1. Autoimmune **disease**<br>*Also Known As*<br>Autoimmune disorder, Autoimmune **disease** (disorder) | 19,440 | 8.54 % |
| 2. Hashimoto thyroiditis<br>*Also Known As*<br>Hashimoto's thyroiditis, Lymphocytic thyroiditis, Hashimoto thyroiditis (disorder), Hashimoto's dise··· See More | 1,780 | 0.78 % |
| 3. Cold autoimmune hemolytic anemia<br>*Also Known As*<br>Cold haemagglutinin **disease**, Cold haemolytic **disease**, Cold hemagglutinin **disease**, Cold agglutinin di··· See More | 40 | 0.02 % |
| 4. Polyglandular autoimmune syndrome, type 2<br>*Also Known As*<br>Addison's **disease** with struma lymphomatosa, Type 2 polyendocrine autoimmunity syndrome, PGA - Polygl··· See More | ≤ 20 | 0.01 % |
| 5. Xerostomia due to autoimmune **disease**<br>*Also Known As*<br>Autoimmune xerostomia, Xerostomia due to autoimmune **disease** (disorder) | ≤ 20 | 0.01 % |
| 6. Polyglandular autoimmune syndrome, type 1<br>*Also Known As*<br>APEDED, Autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy, Whitaker syndrome, Polygland··· See More | ≤ 20 | 0.01 % |
| 7. Polymyositis associated with autoimmune **disease**<br>*Also Known As*<br>Polymyositis associated with autoimmune **disease** (disorder) | ≤ 20 | 0.01 % |
| 8. Chronic idiopathic thrombocytopenic purpura<br>*Also Known As*<br>Autoimmune thrombocytopenic purpura, Chronic thrombocytopenic purpura, Chronic idiopathic thrombocyt··· See More | ≤ 20 | 0.01 % |
| 9. Addison's **disease** due to autoimmunity<br>*Also Known As*<br>Addison **disease** due to autoimmunity, Autoimmune adrenal atrophy, Autoimmune adrenalitis, Addison's d··· See More | ≤ 20 | 0.01 % |
| 10. Neutropenia associated with autoimmune **disease**<br>*Also Known As*<br>Neutropenia associated with autoimmune **disease** (disorder) | ≤ 20 | 0.01 % |

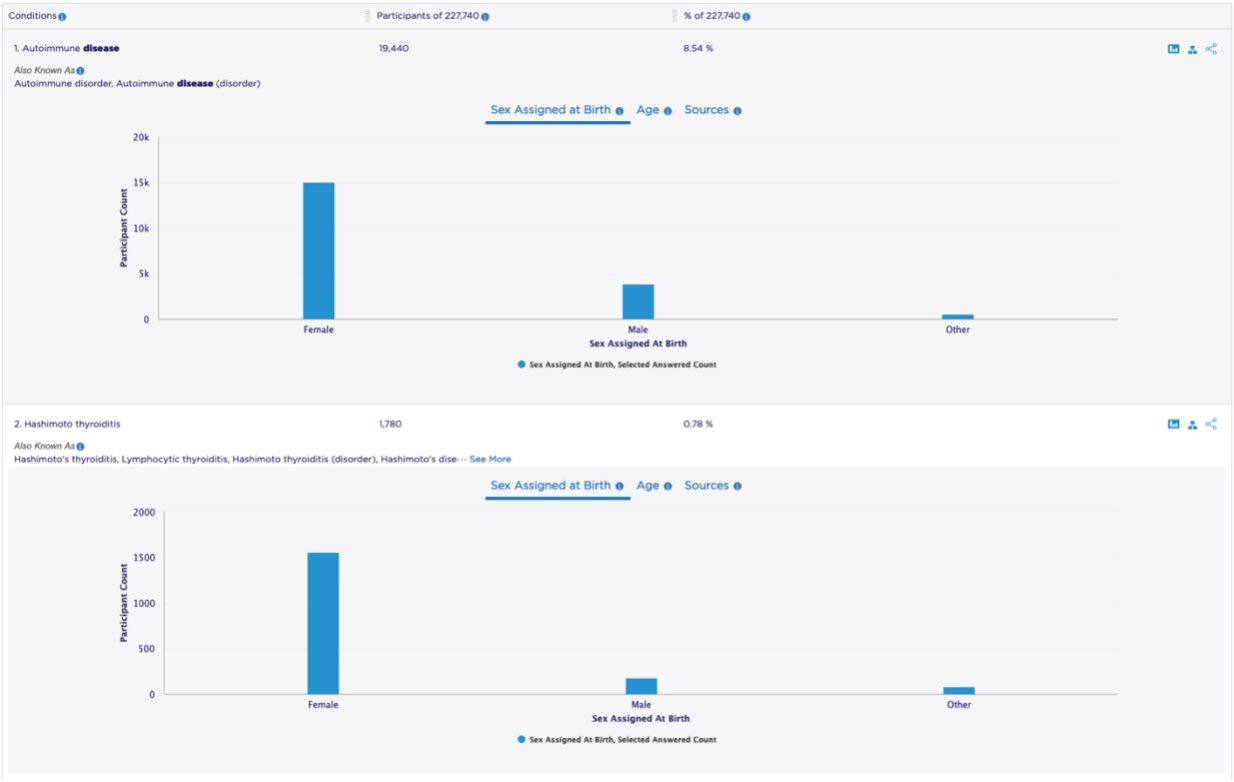



[Previous studies] have demonstrated about 80% of all people diagnosed with autoimmune disease are women. The dataset in AllofUs reflects a similar split between Female and Male.

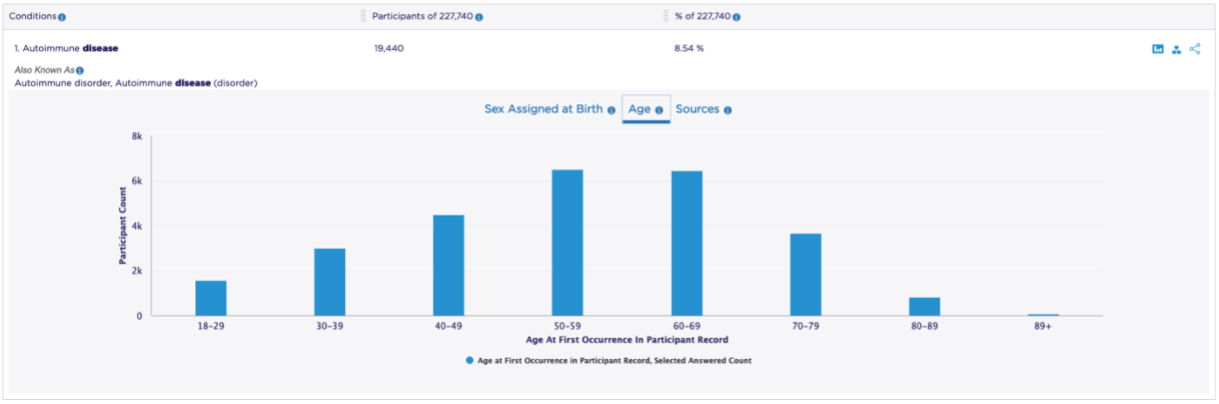





Further, there is a similar division for Hashimoto thyroiditis between genders.

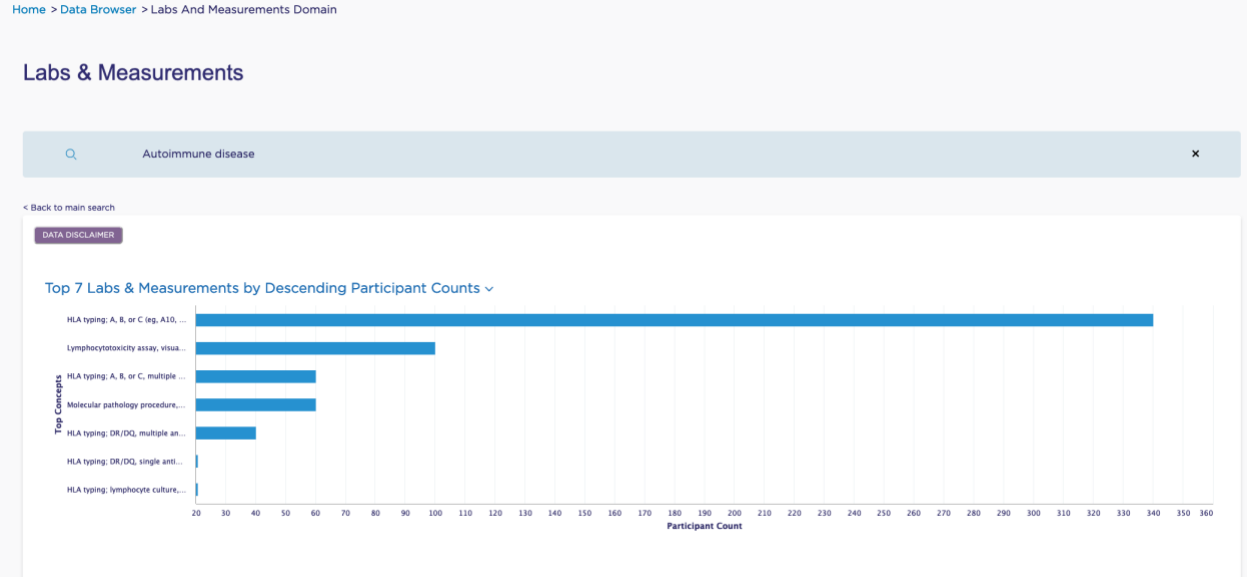


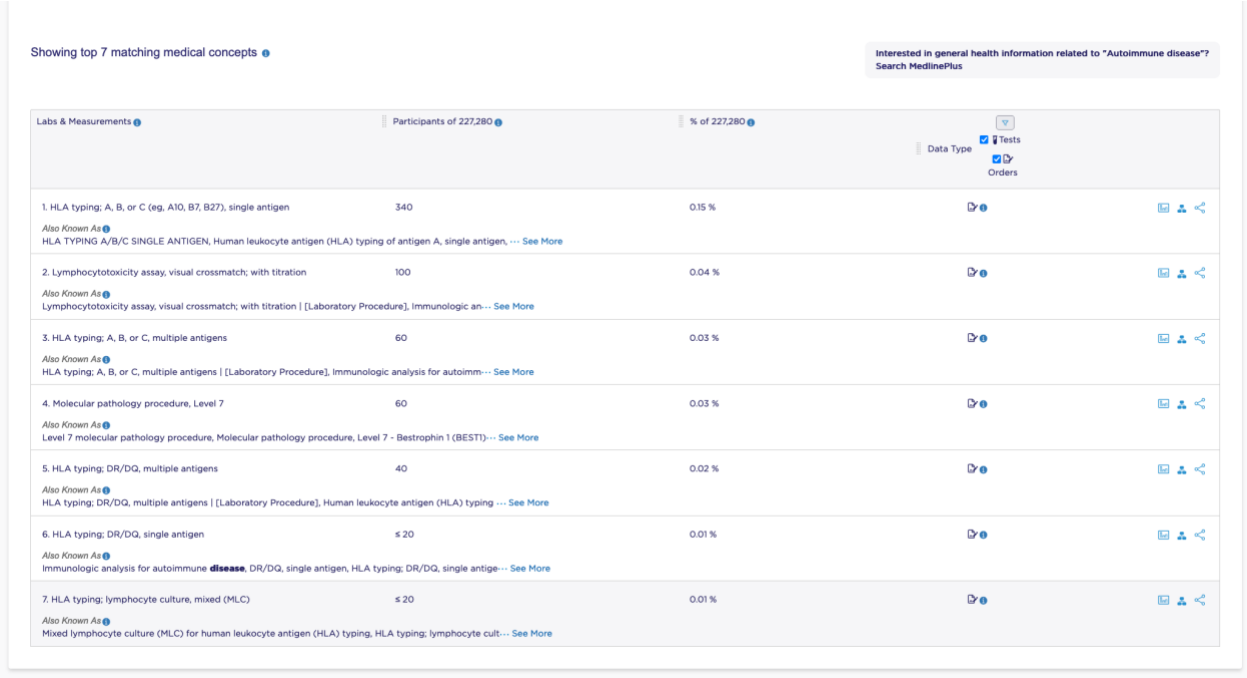
Age may or may not be relevant to a study of gender differences and estrogen.




The breakdown of autoimmune diseases doesn't contribute information for our potential question.


Additionally, the labs collected offer no insights into estrogen or hormone levels.



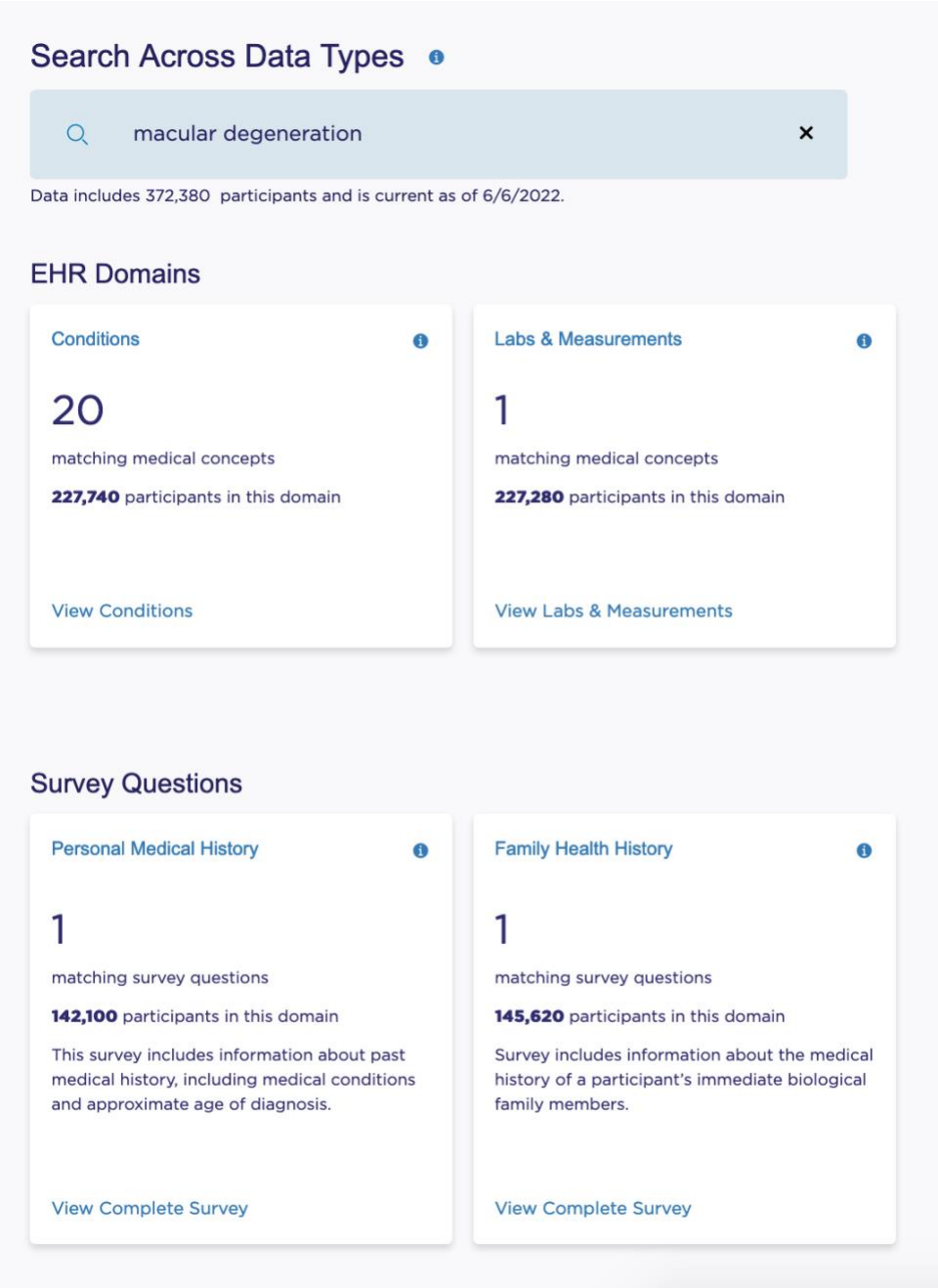




## Conclusion: Autoimmune Disease

Even though the data demonstrates a reflective gender split for autoimmune diagnosis, we don't have access to any questionnaires or the appropriate lab work to investigate the potential medical question: Is there enough data to support an exploration of "Does estrogen imbalance lead to females having more autoimmune diseases?"

Second Walkthrough Condition: Macular Degeneration

Potential medical question: Is there enough data to support an exploration of “Is there a correlation between a history of dry eyes and macular degeneration?”




The records for macular degeneration offer a richer story since they include over 142,000 + family history and medical history surveys. Additionally, the breakdown of different conditions under the macular degeneration umbrella demonstrate the large proportion of age-related versus non-age-related.

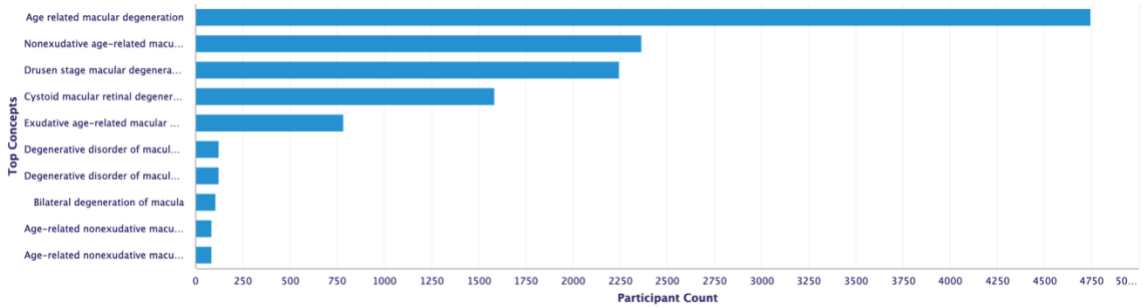
# Conditions

macular degeneration

< Back to main search

DATA DISCLAIMER

## Top 10 Conditions by Descending Participant Counts

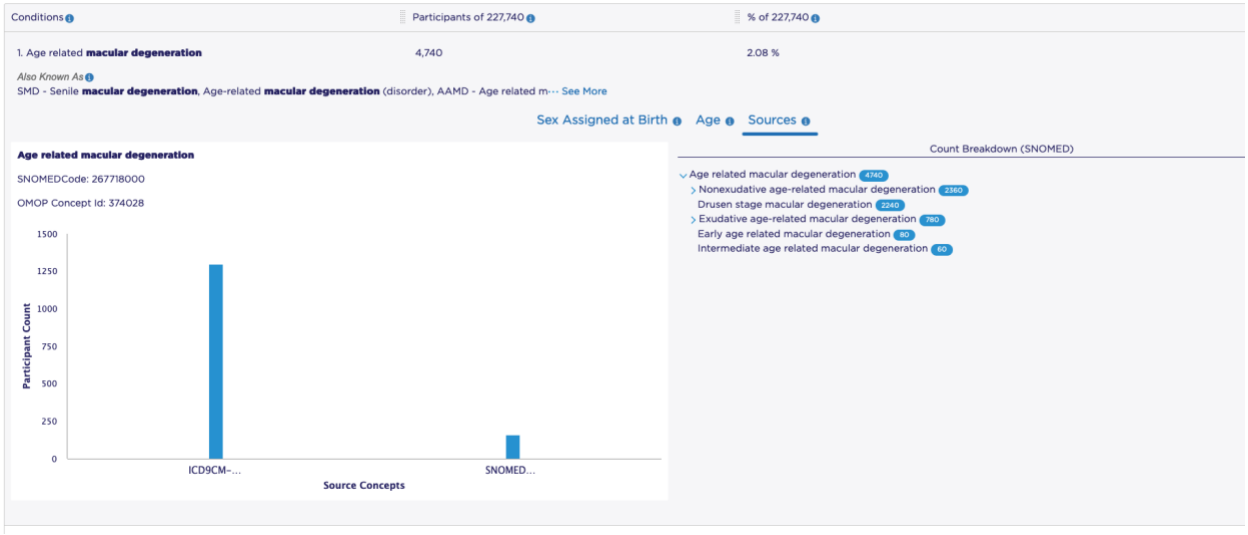



Showing top 20 matching medical concepts

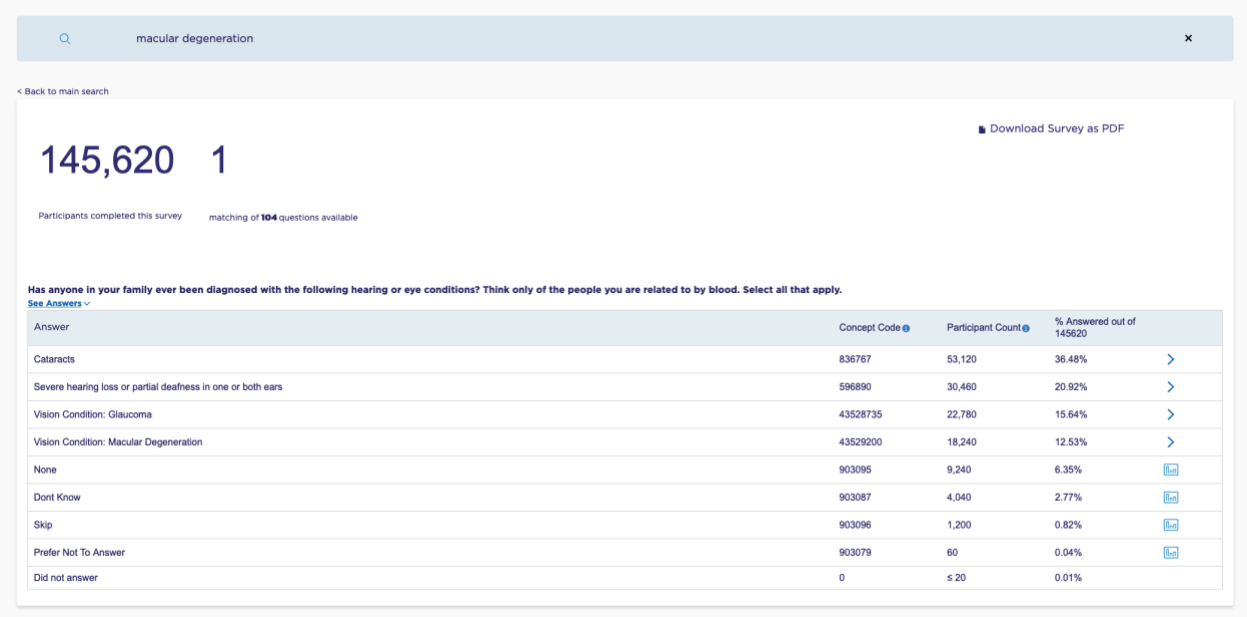
Interested in general health information related to "macular degeneration"?
Search MedlinePlus

| Conditions | Participants of 227,740 | % of 227,740 |
|---|---|---|
| 1. Age related **macular degeneration**<br>*Also Known As*<br>SMD - Senile **macular degeneration**, Age-related **macular degeneration** (disorder), AAMD - Age related m··· See More | 4,740 | 2.08 % |
| 2. Nonexudative age-related **macular degeneration**<br>*Also Known As*<br>Nonneovascular age-related **macular degeneration**, Atrophic age-related **macular degeneration**, Nonexuda··· See More | 2,360 | 1.04 % |
| 3. Drusen stage **macular degeneration**<br>*Also Known As*<br>Drusen stage **macular degeneration** (disorder) | 2,240 | 0.98 % |
| 4. Cystoid **macular** retinal **degeneration**<br>*Also Known As*<br>Cystoid **macular** retinal **degeneration** (disorder) | 1,580 | 0.69 % |
| 5. Exudative age-related **macular degeneration**<br>*Also Known As*<br>Subretinal neovascularization of macula, Exudative age-related **macular degeneration** (disorder), Disc··· See More | 780 | 0.34 % |
| 6. Degenerative disorder of macula of left eye<br>*Also Known As*<br>Left **macular degeneration**, Degenerative disorder of macula of left eye (disorder) | 120 | 0.05 % |
| 7. Degenerative disorder of macula of right eye<br>*Also Known As*<br>Right **macular degeneration**, Degenerative disorder of macula of right eye (disorder) | 120 | 0.05 % |
| 8. Bilateral **degeneration** of macula<br>*Also Known As*<br>Bilateral **degeneration** of macula (disorder), **Macular degeneration** of both eyes | 100 | 0.04 % |
| 9. Age-related nonexudative **macular degeneration** of left eye<br>*Also Known As*<br>Age-related nonexudative **macular degeneration** of left eye (disorder) | 80 | 0.04 % |
| 10. Age-related nonexudative **macular degeneration** of right eye<br>*Also Known As*<br>Age-related nonexudative **macular degeneration** of right eye (disorder) | 80 | 0.04 % |
| 11. Early age related **macular degeneration**<br>*Also Known As*<br>Early age related **macular degeneration** (disorder) | 80 | 0.04 % |
| 12. Bilateral age-related nonexudative **macular degeneration**<br>*Also Known As*<br>Age-related nonexudative **macular degeneration** of both eyes, Bilateral age-related nonexudative macul··· See More | 60 | 0.03 % |

Histograms show how the first occurrences of macular degeneration are rare in the under 50 crowd, but steadily increase as age increases.

Surveys target questions about eye condition diagnosis and family history of eye conditions.



# Family Health History

Survey includes information about the medical history of a participant's immediate biological family members. Survey questions appear in the order in which participants took the survey.

**Note**: The data on this page are:

- Gathered directly from participants through electronic surveys
- Grouped into bins of 20 to protect privacy

For more information about this survey, please visit the Survey Explorer

macular degeneration ×

< Back to main search

Download Survey as PDF

145,620 Participants completed this survey

1 matching of **104** questions available

**Has anyone in your family ever been diagnosed with the following hearing or eye conditions? Think only of the people you are related to by blood. Select all that apply.**

See Answers

| Answer | Concept Code | Participant Count | % Answered out of 145620 |
|---|---|---|---|
| Cataracts | 836767 | 53,120 | 36.48% |
| Severe hearing loss or partial deafness in one or both ears | 596890 | 30,460 | 20.92% |
| Vision Condition: Glaucoma | 43528735 | 22,780 | 15.64% |
| Vision Condition: Macular Degeneration | 43529200 | 18,240 | 12.53% |
| None | 903095 | 9,240 | 6.35% |
| Dont Know | 903087 | 4,040 | 2.77% |
| Skip | 903096 | 1,200 | 0.82% |
| Prefer Not To Answer | 903079 | 60 | 0.04% |
| Did not answer | 0 | ≤ 20 | 0.01% |

macular degeneration ×

< Back to main search

Download Survey as PDF

142,100 Participants completed this survey

1 matching of **465** questions available

**Has a doctor or health care provider ever told you that you have or had any of the following hearing or vision problems? (select all that apply)**

See Answers

| Answer | Concept Code | Participant Count | % Answered out of 142100 |
|---|---|---|---|
| Hearing Vision Conditions: Near Sightedness | 43529265 | 68,320 | 48.08% |
| Hearing Vision Conditions: Astigmatism | 1384648 | 47,980 | 33.76% |
| Hearing Vision Conditions: Far Sightedness | 43528699 | 30,120 | 21.20% |
| Hearing Vision Conditions: Cataracts | 43528528 | 28,120 | 19.79% |
| Hearing Vision Conditions: Dry Eyes | 43528659 | 27,620 | 19.44% |
| Hearing Vision Conditions: No Hearing Eye | 1384598 | 26,600 | 18.72% |
| Hearing Vision Conditions: Tinnitus | 43529917 | 18,140 | 12.77% |
| Hearing Vision Conditions: Other Hearing Eye | 1384428 | 11,440 | 8.05% |
| Hearing Vision Conditions: Hearing Loss | 1384396 | 9,840 | 6.92% |
| Hearing Vision Conditions: Glaucoma | 43528731 | 5,800 | 4.08% |
| Hearing Vision Conditions: Macular Degeneration | 43529196 | 3,640 | 2.56% |
| Skip | 903096 | 3,440 | 2.42% |
| Hearing Vision Conditions: Blindness | 43528474 | 1,060 | 0.75% |
| Did not answer | 0 | ≤ 20 | 0.01% |

| Hearing Vision Conditions: Dry Eyes | 43528659 | 27,620 | 19.44% |
|---|---|---|---|

↳ **Are you still seeing a doctor or health care provider for Dry eyes?**

| Answer | Concept Code | Participant Count | % Answered out of 27620 |
|---|---|---|---|
| Dry Eyes Currently: Yes | 43530031 | 16,560 | 59.96% |
| Dry Eyes Currently: No | 43529338 | 10,900 | 39.46% |
| Skip | 903096 | 180 | 0.65% |

↳ **Are you currently prescribed medications and/or receiving treatment for Dry eyes?**

| Answer | Concept Code | Participant Count | % Answered out of 27620 |
|---|---|---|---|
| Rx Meds for Dry Eyes: Yes | 43530147 | 14,320 | 51.85% |
| Rx Meds for Dry Eyes: No | 43529453 | 13,120 | 47.50% |
| Skip | 903096 | 200 | 0.72% |

↳ **About how old were you when you were first told you had Dry eyes?**

| Answer | Concept Code | Participant Count | % Answered out of 27620 |
|---|---|---|---|
| How Old Were You Dry Eyes: Adult | 1384880 | 19,740 | 71.47% |
| How Old Were You Dry Eyes: Older Adult | 1384883 | 5,040 | 18.25% |
| How Old Were You Dry Eyes: Adolescent | 1385538 | 1,360 | 4.92% |
| How Old Were You Dry Eyes: Elderly | 1384729 | 860 | 3.11% |
| How Old Were You Dry Eyes: Child | 1385121 | 480 | 1.74% |
| Skip | 903096 | 180 | 0.65% |

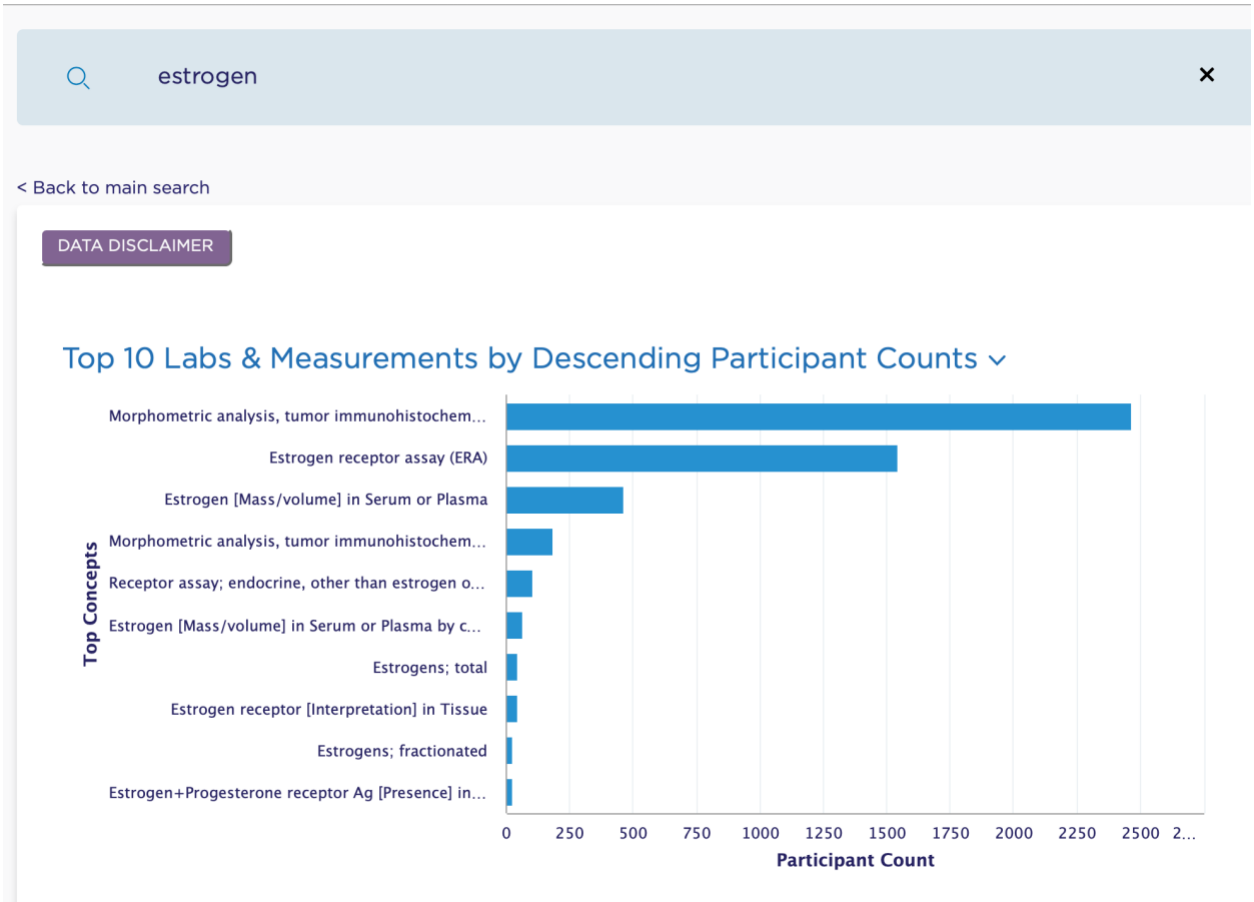
Conclusion: Macular Degeneration

There appears to be enough data to support an exploration of the potential medical question: “Is there a correlation between a history of dry eyes and macular degeneration?” It would be interesting to see what additional data would be available at a higher tier.

## Part III: Summary and lessons learned

AllofUs is of limited use in the data browser tier, simply because we have very limited access to the full body of information. For instance, when exploring autoimmune disease we discovered a robust number of participants and the proportion of male/female matches fairly closely with the general population split of 20%-80%.

However, it is hard to say for certain if there is data to support the exploration of “Does estrogen imbalance lead to females having more autoimmune diseases?” at this tier. It is possible that there is some overlap between those diagnosed with autoimmune disease and the group of participants who had estrogen lab work drawn:




We simply don’t know at the data browser tier. However, the data browser does offer a unique snapshot of how well reported different conditions are. It is a useful tool to scope out a potential research question.

## Bonus opportunity 1: Exploration of 4 additional datasets

For this bonus activity, I focused on additional datasets related to Parkinson's Disease. The MichaelJFox organization (https://www.michaeljfox.org/data-sets) highlights 6 datasets and some links to dashboard platforms. I selected 4 to explore, which I've listed below. For each I needed to apply for access citing my affiliation with FAU and my purpose for requesting access: "I am examining applications of artificial intelligence, machine learning, and data analysis to treatment and diagnosis of Parkinson's Disease for a class taught by Dr. Oge Marques. I plan to write some code as well, depending on access to datasets."

From the website:

- **Parkinson's Progression Markers Initiative**
  The Parkinson's Progression Markers Initiative (PPMI) is an ongoing longitudinal observational study that collects comprehensive clinical, imaging, genetic data, and biological samples.
  **Learn More**

- **BioFIND**
  The Fox Investigation for New Discovery of Biomarkers (BioFIND) is an observational clinical study designed to discover and verify biomarkers of Parkinson's disease, which collected clinical data and biological samples from participants with and without Parkinson's disease.
  **Learn More**

- **AMP PD**
  The Accelerating Medicines Partnership Parkinson's disease (AMP PD) Knowledge Platform is a tool to explore and analyze whole genome sequencing, transcriptomics, and harmonized clinical data across a wide range of studies, simplifying cross-cohort analyses.

  ---

  Within the portal you can explore a dataset, understand its contents, and build custom cohorts for deeper analysis using Jupyter notebooks. AMP PD is a public-private partnership between The Michael J. Fox Foundation, National Institutes of Health, the U.S. Food and Drug Administration, Aligning Science Across Parkinson's (ASAP), and five companies (Bristol Myers-Squibb, GSK, Pfizer, Sanofi, and Verily). AMP PD is managed through the Foundation of the National Institutes of Health (FNIH).

  **Access the AMP PD Knowledge Platform**

- **Fox Insight**

The MJFF-sponsored online study Fox Insight collects patient reported outcome data on the lived experience of Parkinson's disease and genetic data through a collaboration with consumer genetics company 23andMe.
**Learn More**

For this bonus activity, I focused on additional datasets related to Parkinson's Disease. The MichaelJFox organization (https://www.michaeljfox.org/data-sets) highlights 6 datasets and some links to dashboard platforms. I selected 4 to explore, which I've listed below. For each I needed to apply for access citing my affiliation with FAU and my purpose for requesting access: "I am examining applications of artificial intelligence, machine learning, and data analysis to treatment and diagnosis of Parkinson's Disease for a class taught by Dr. Oge Marques. I plan to write some code as well, depending on access to datasets."

It took about a week for approval to all four. The wealth of information available from these sites is overwhelming and it will take time to explore them fully. I look forward to working with the data over the winter break. For now, I highlighted my impressions of each in a paragraph below. Additionally, there are numerous screen shots following the impression paragraphs to showcase some of the unique features of each.

## Parkinson's Progression Markers Initiative

PPMI offers data from clinical visits, lab work, images, cerebrospinal fluid, DNA, and RNA for groups of those diagnosed with Parkinson's and controls (over 4,000 individuals in all). It also includes a Data Dictionary. There is a definite learning curve for navigating the site. So far the most impressive aspect is the access to the Image Collections. One can select images based on type and even machine. Images that fit those criteria are displayed in search results and then one can click on VIEW to explore the image in the IDA IMAGE VIEWER. One can even scroll through the slices as appropriate for the image type.

## BioFIND

BioFIND contains data on 120 people with Parkinson's and 120 healthy controls over 2 years. Data includes blood, spinal fluid, medication logs, motor and non-motor assessments, medical history, and genetic data (gene sequencing and genotyping). This site is much easier to navigate, partially because there are not as many datatypes. one simple checks the boxes for the interesting data and download it as a CSV file.

## AMP PD

AMP PD is a robust dataset of over 10,000 participants. The data appears to focus more on genomics and proteomics. It appears to have a seamless interface with Jupyter

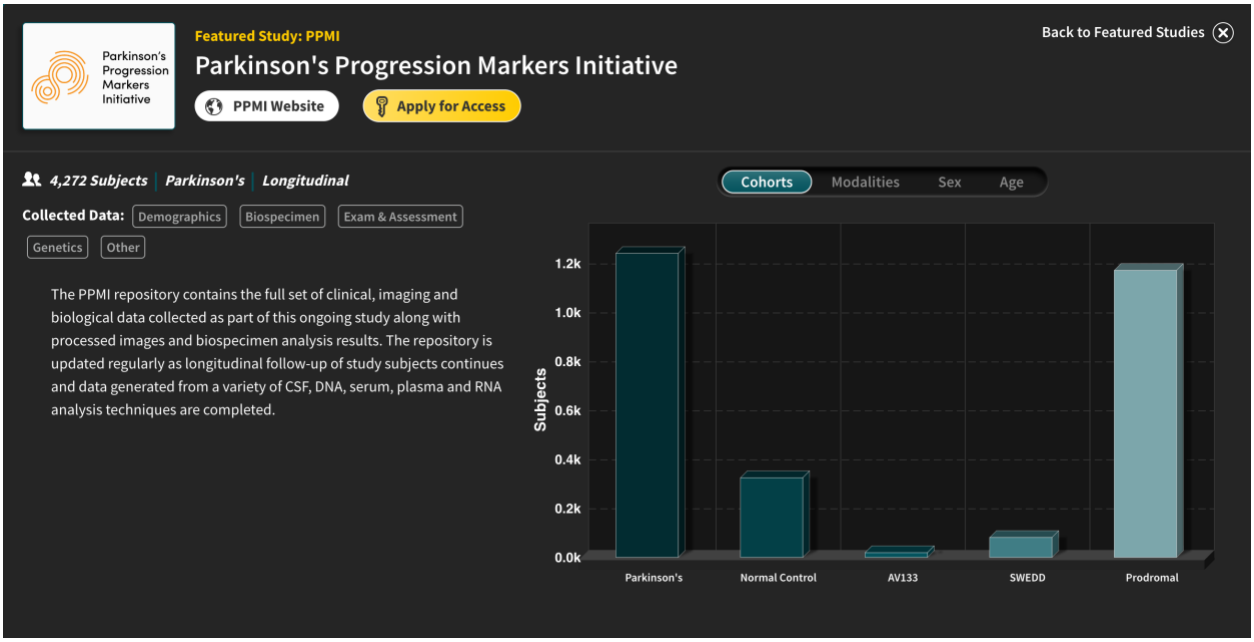
notebooks/python. The data is stored on Google Cloud and all the analysis is also done on Google Cloud. Therefore, one needs to pay for the storage and the analysis. There are several data sources that do not require payment for analysis, so I will skip investigating this one any further.

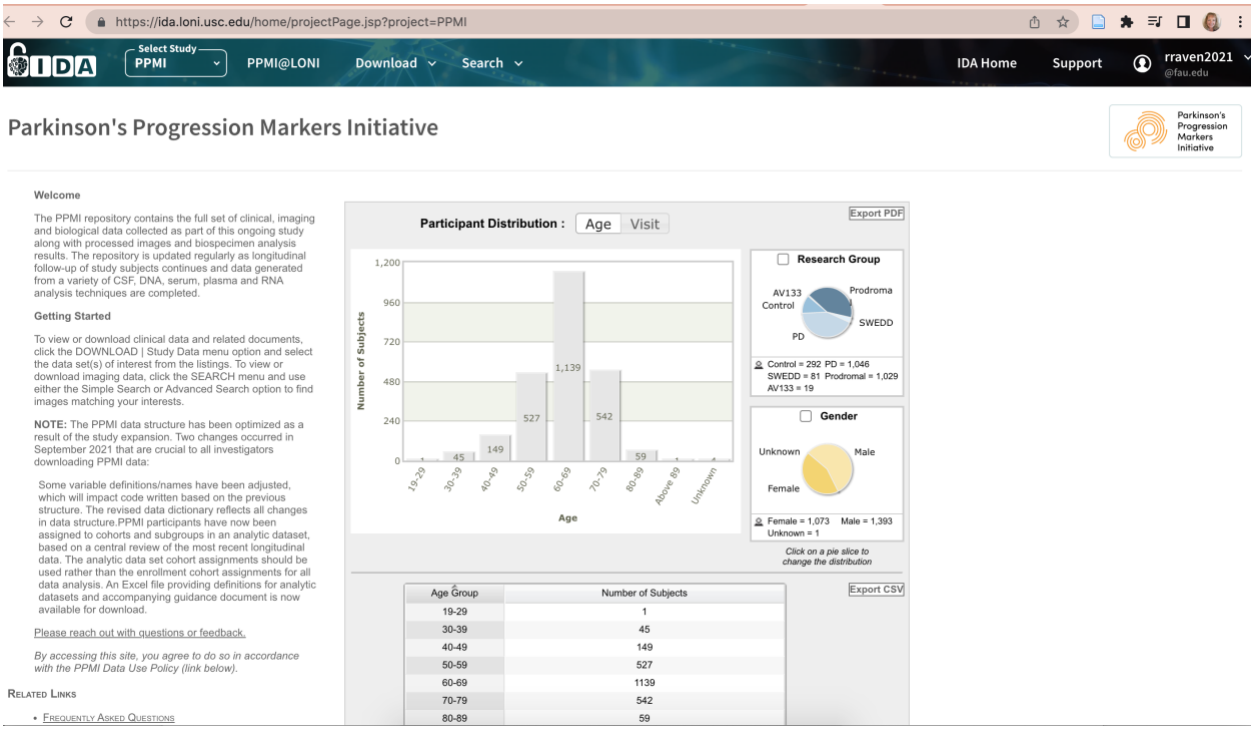
## Fox Insight

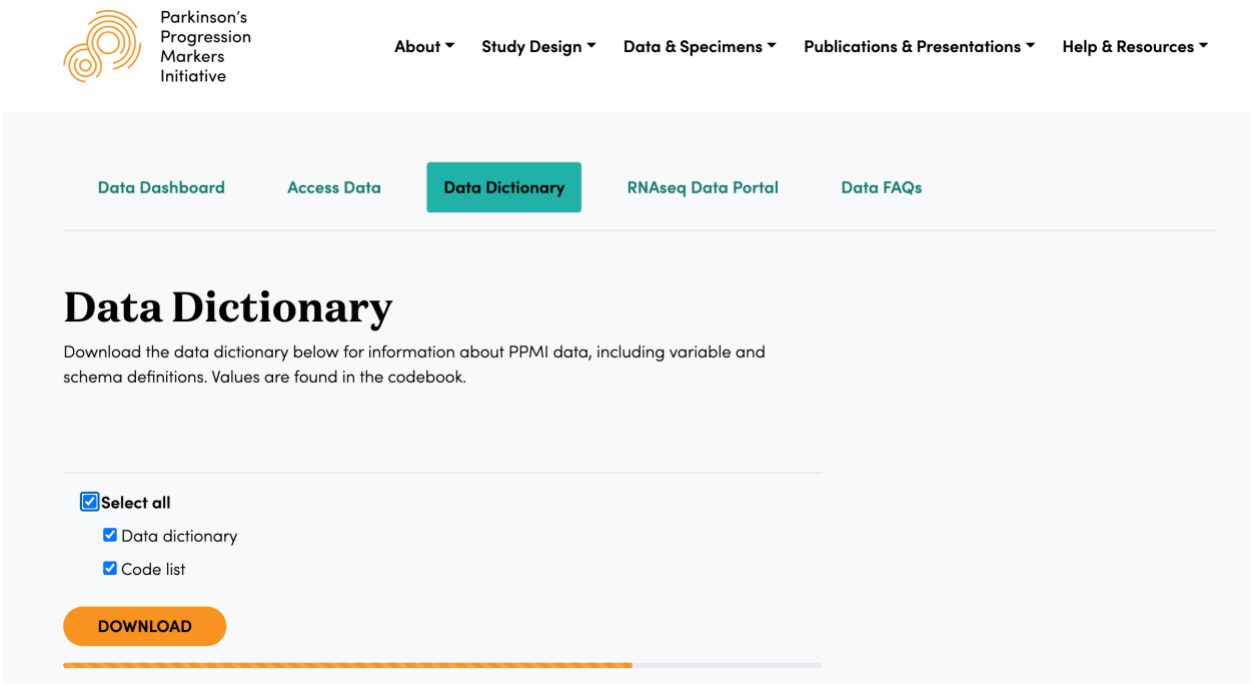
FoxInsight is a collection of data on over 50,000 people, with a mix of people diagnosed with Parkinson's and controls. The data includes surveys on medical history, family history, demographics, motor screening, survey on experience of disease, and genetics. There are actually many more specific data points as one drills down in the menu. Navigation is easy. One selects what data to download and can select whether downloading it as one CSV or multiple CSVs for each survey or assessment. I decided to use this data set as the source for my demo notebook where I address the question: "Can a history of a stroke, traumatic brain injury, or surgery with anesthesia be used to predict the diagnosis of Parkinson's Disease?".

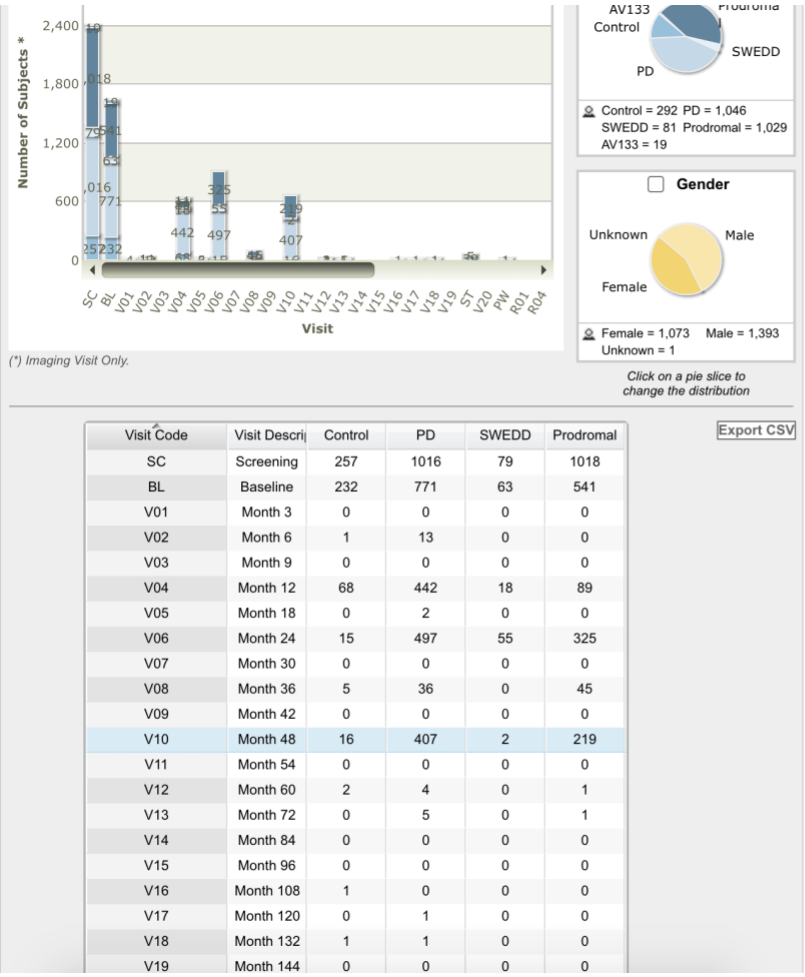
### Bonus opportunity 2: Demo notebook accessible via colab link:

## Parkinson's Progression Markers Initiative

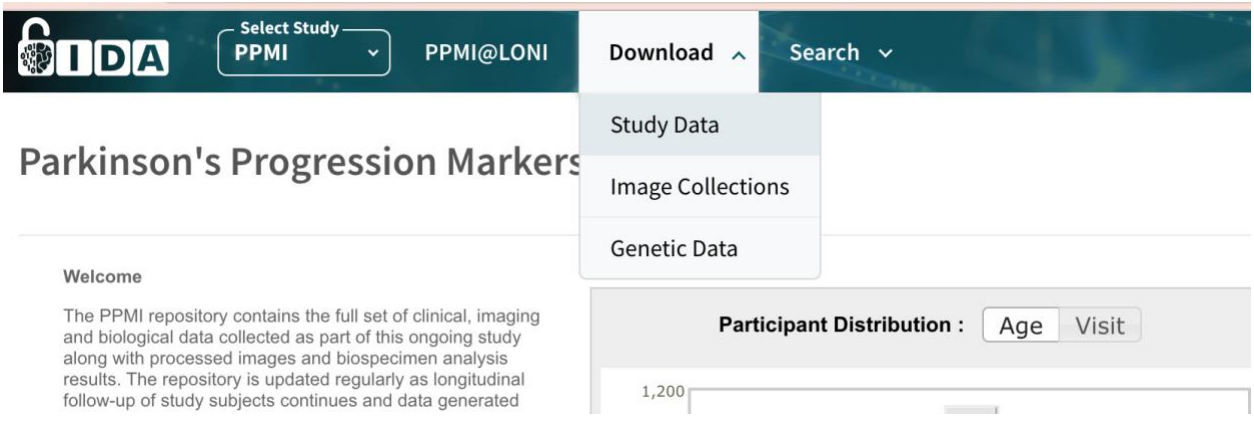


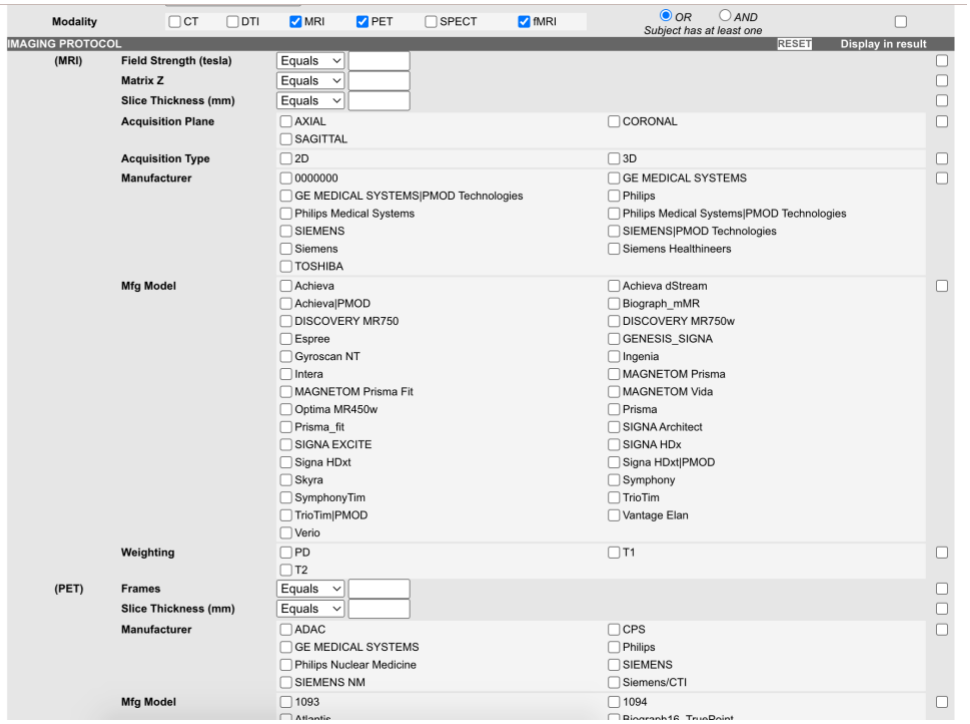

https://ida.loni.usc.edu/home/projectPage.jsp?project=PPMI
IDA
Select Study
PPMI
PPMI@LONI
Download
Search
IDA Home
Support
rraven2021
@fau.edu
Parkinson's Progression Markers Initiative
Welcome
The PPMI repository contains the full set of clinical, imaging and biological data collected as part of this ongoing study along with processed images and biospecimen analysis results. The repository is updated regularly as longitudinal follow-up of study subjects continues and data generated from a variety of CSF, DNA, serum, plasma and RNA analysis techniques are completed.
Getting Started
To view or download clinical data and related documents, click the DOWNLOAD | Study Data menu option and select the data set(s) of interest from the listings. To view or download imaging data, click the SEARCH menu and use either the Simple Search or Advanced Search option to find images matching your interests.
NOTE: The PPMI data structure has been optimized as a result of the study expansion. Two changes occurred in September 2021 that are crucial to all investigators downloading PPMI data:
Some variable definitions/names have been adjusted, which will impact code written based on the previous structure. The revised data dictionary reflects all changes in data structure.PPMI participants have now been assigned to cohorts and subgroups in an analytic dataset, based on a central review of the most recent longitudinal data. The analytic data set cohort assignments should be used rather than the enrollment cohort assignments for all data analysis. An Excel file providing definitions for analytic datasets and accompanying guidance document is now available for download.
Please reach out with questions or feedback.
By accessing this site, you agree to do so in accordance with the PPMI Data Use Policy (link below).
Related Links
Frequently Asked Questions
Participant Distribution :
Age
Visit
Export PDF
Number of Subjects
1,200
960
720
480
240
0
45
149
527
1,139
542
59
19-29
30-39
40-49
50-59
60-69
70-79
80-89
Above 89
Unknown
Age
Research Group
AV133
Control
Prodroma
SWEDD
PD
Control = 292 PD = 1,046
SWEDD = 81 Prodromal = 1,029
AV133 = 19
Gender
Unknown
Male
Female
Female = 1,073 Male = 1,393
Unknown = 1
Click on a pie slice to change the distribution
Export CSV
Age Group
Number of Subjects
19-29 1
30-39 45
40-49 149
50-59 527
60-69 1139
70-79 542
80-89 59

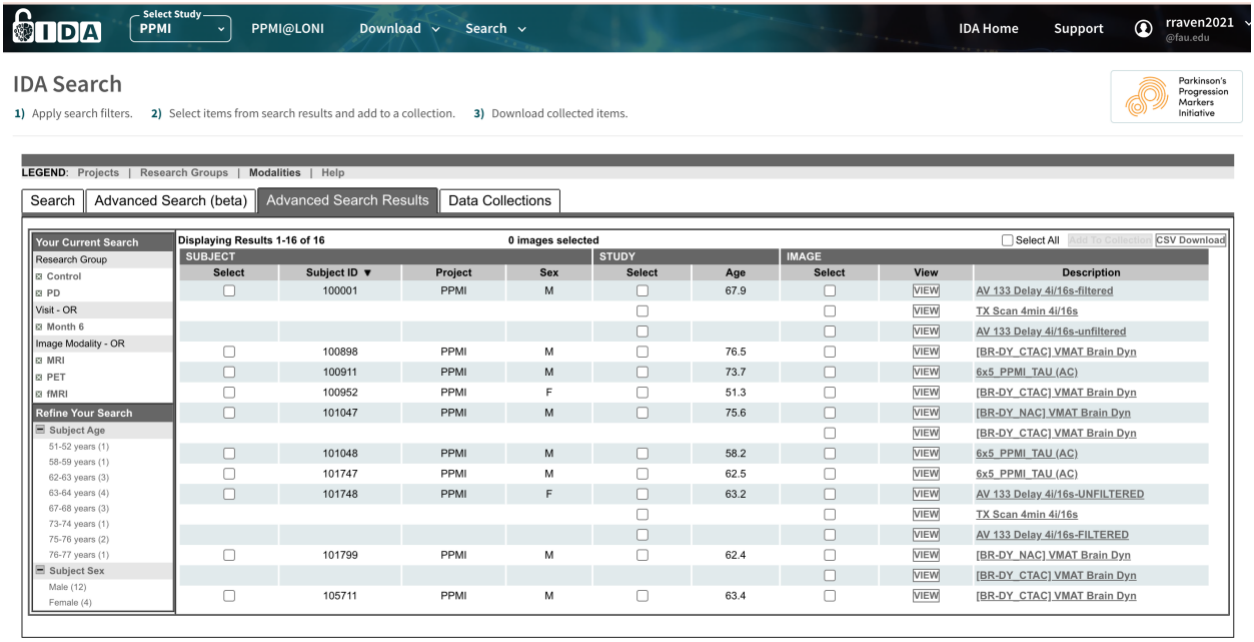

Parkinson's Progression Markers Initiative
About
Study Design
Data & Specimens
Publications & Presentations
Help & Resources
Data Dashboard
Access Data
Data Dictionary
RNAseq Data Portal
Data FAQs
Data Dictionary
Download the data dictionary below for information about PPMI data, including variable and schema definitions. Values are found in the codebook.
Select all
Data dictionary
Code list
DOWNLOAD

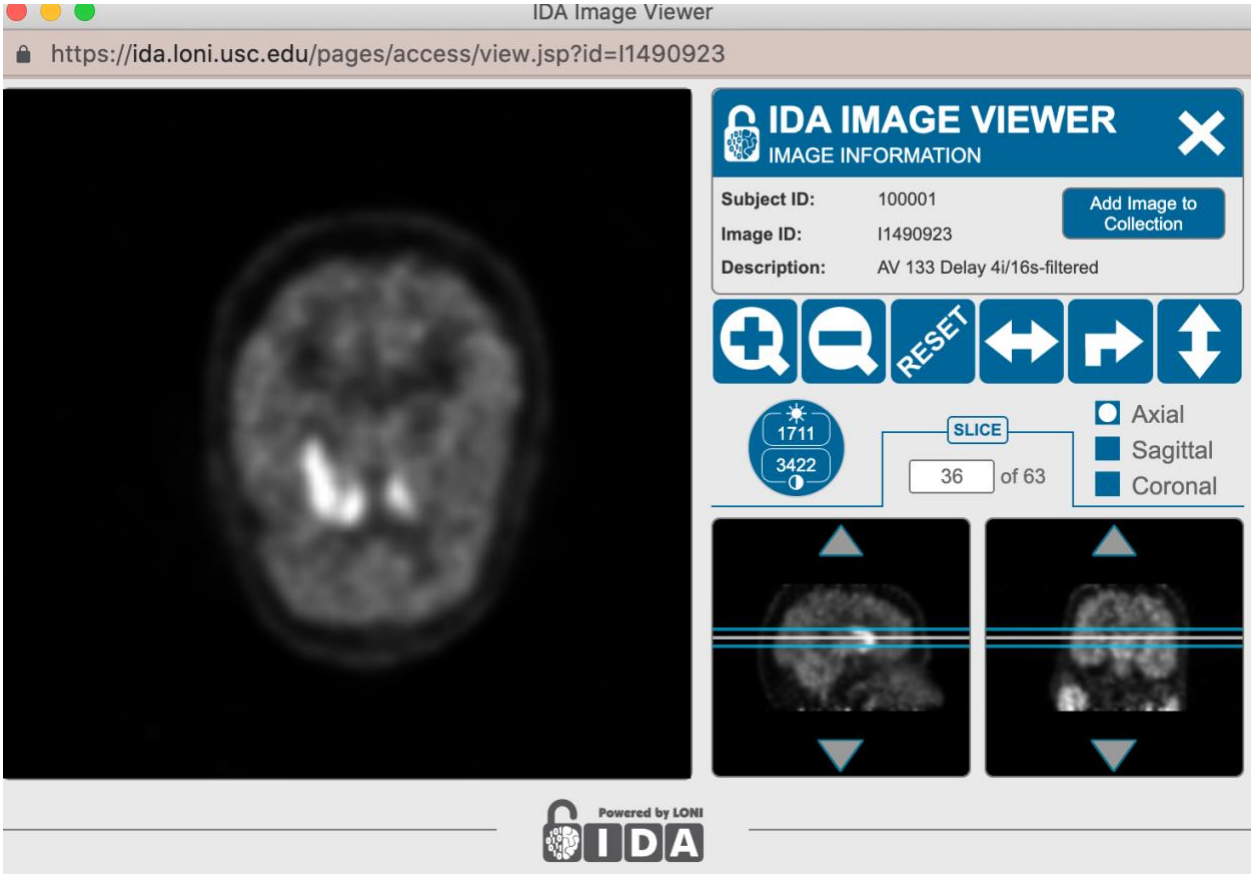



| Visit Code | Visit Descri | Control | PD | SWEDD | Prodromal |
|---|---|---|---|---|---|
| SC | Screening | 257 | 1016 | 79 | 1018 |
| BL | Baseline | 232 | 771 | 63 | 541 |
| V01 | Month 3 | 0 | 0 | 0 | 0 |
| V02 | Month 6 | 1 | 13 | 0 | 0 |
| V03 | Month 9 | 0 | 0 | 0 | 0 |
| V04 | Month 12 | 68 | 442 | 18 | 89 |
| V05 | Month 18 | 0 | 2 | 0 | 0 |
| V06 | Month 24 | 15 | 497 | 55 | 325 |
| V07 | Month 30 | 0 | 0 | 0 | 0 |
| V08 | Month 36 | 5 | 36 | 0 | 45 |
| V09 | Month 42 | 0 | 0 | 0 | 0 |
| V10 | Month 48 | 16 | 407 | 2 | 219 |
| V11 | Month 54 | 0 | 0 | 0 | 0 |
| V12 | Month 60 | 2 | 4 | 0 | 1 |
| V13 | Month 72 | 0 | 5 | 0 | 1 |
| V14 | Month 84 | 0 | 0 | 0 | 0 |
| V15 | Month 96 | 0 | 0 | 0 | 0 |
| V16 | Month 108 | 1 | 0 | 0 | 0 |
| V17 | Month 120 | 0 | 1 | 0 | 0 |
| V18 | Month 132 | 1 | 1 | 0 | 0 |
| V19 | Month 144 | 0 | 0 | 0 | 0 |




Modality
CT
DTI
MRI
PET
SPECT
fMRI
OR
AND
Subject has at least one
IMAGING PROTOCOL
RESET
Display in result
(MRI)
Field Strength (tesla)
Matrix Z
Slice Thickness (mm)
Equals
Acquisition Plane
AXIAL
CORONAL
SAGITTAL
Acquisition Type
2D
3D
Manufacturer
0000000
GE MEDICAL SYSTEMS
GE MEDICAL SYSTEMS|PMOD Technologies
Philips
Philips Medical Systems
Philips Medical Systems|PMOD Technologies
SIEMENS
SIEMENS|PMOD Technologies
Siemens
Siemens Healthineers
TOSHIBA
Mfg Model
Achieva
Achieva dStream
Achieva|PMOD
Biograph_mMR
DISCOVERY MR750
DISCOVERY MR750w
Espree
GENESIS_SIGNA
Gyroscan NT
Ingenia
Intera
MAGNETOM Prisma
MAGNETOM Prisma Fit
MAGNETOM Vida
Optima MR450w
Prisma
Prisma_fit
SIGNA Architect
SIGNA EXCITE
SIGNA HDx
Signa HDxt
Signa HDxt|PMOD
Skyra
Symphony
SymphonyTim
TrioTim
TrioTim|PMOD
Vantage Elan
Verio
Weighting
PD
T1
T2
(PET)
Frames
Slice Thickness (mm)
Manufacturer
ADAC
CPS
GE MEDICAL SYSTEMS
Philips
Philips Nuclear Medicine
SIEMENS
SIEMENS NM
Siemens/CTI
Mfg Model
1093
1094


IDA
Select Study
PPMI
PPMI@LONI
Download
Search
IDA Home
Support
rraven2021
@fau.edu
IDA Search
1) Apply search filters.
2) Select items from search results and add to a collection.
3) Download collected items.
Parkinson's Progression Markers Initiative
LEGEND: Projects | Research Groups | Modalities | Help
Search
Advanced Search (beta)
Advanced Search Results
Data Collections
Your Current Search
Research Group
Control
PD
Visit - OR
Month 6
Image Modality - OR
MRI
PET
fMRI
Refine Your Search
Subject Age
51-52 years (1)
58-59 years (1)
62-63 years (3)
63-64 years (4)
67-68 years (3)
73-74 years (1)
75-76 years (2)
76-77 years (1)
Subject Sex
Male (12)
Female (4)
Displaying Results 1-16 of 16
0 images selected
Select All
Add To Collection
CSV Download
SUBJECT
STUDY
IMAGE
Select
Subject ID
Project
Sex
Age
View
Description
100001 PPMI M 67.9 AV 133 Delay 4i/16s-filtered
TX Scan 4min 4i/16s
AV 133 Delay 4i/16s-unfiltered
100898 PPMI M 76.5 [BR-DY_CTAC] VMAT Brain Dyn
100911 PPMI M 73.7 6x5_PPMI_TAU (AC)
100952 PPMI F 51.3 [BR-DY_CTAC] VMAT Brain Dyn
101047 PPMI M 75.6 [BR-DY_NAC] VMAT Brain Dyn
[BR-DY_CTAC] VMAT Brain Dyn
101048 PPMI M 58.2 6x5_PPMI_TAU (AC)
101747 PPMI M 62.5 6x5_PPMI_TAU (AC)
101748 PPMI F 63.2 AV 133 Delay 4i/16s-UNFILTERED
TX Scan 4min 4i/16s
AV 133 Delay 4i/16s-FILTERED
101799 PPMI M 62.4 [BR-DY_NAC] VMAT Brain Dyn
[BR-DY_CTAC] VMAT Brain Dyn
105711 PPMI M 63.4 [BR-DY_CTAC] VMAT Brain Dyn

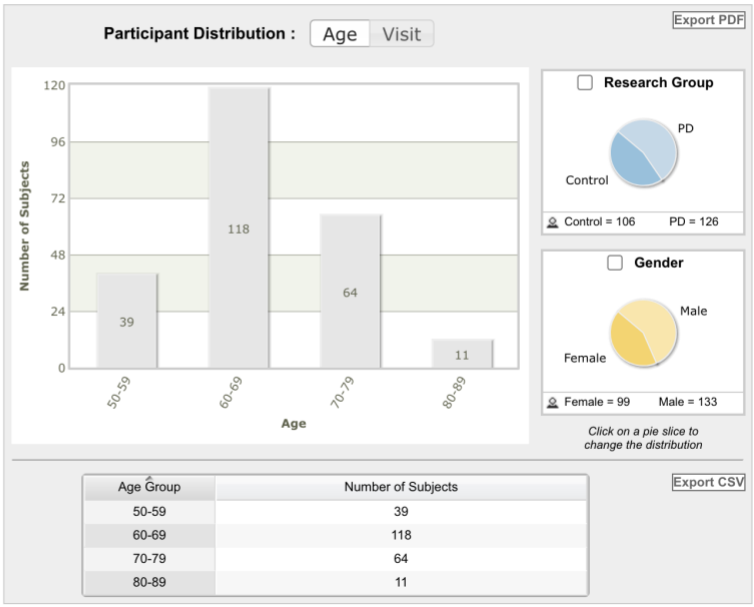



## BioFIND

### BioFIND

#### ABOUT BIOFIND

BioFIND is a two-year observational clinical study designed to discover and verify biomarkers of Parkinson's disease (PD). The study was carried out at 5 academic clinical sites in the United States and began recruitment in the fall of 2012. BioFIND collected clinical data and biospecimens, including blood and cerebrospinal fluid (CSF), in a population of 120 well-defined, moderately advanced PD subjects and 120 healthy controls.

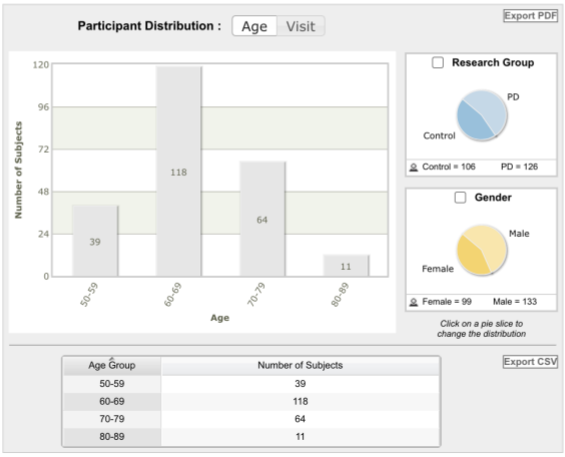



| Age Group | Number of Subjects |
|---|---|
| 50-59 | 39 |
| 60-69 | 118 |
| 70-79 | 64 |
| 80-89 | 11 |


## BioFIND

### ABOUT BIOFIND

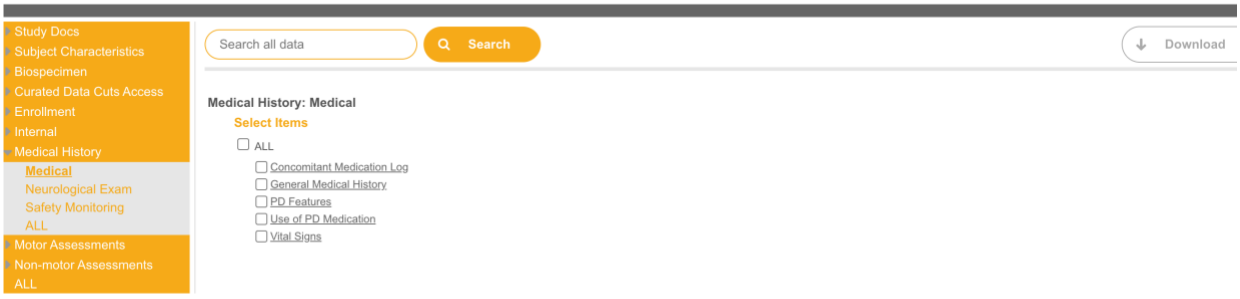
BioFIND is a two-year observational clinical study designed to discover and verify biomarkers of Parkinson's disease (PD). The study was carried out at 5 academic clinical sites in the United States and began recruitment in the fall of 2012. BioFIND collected clinical data and biospecimens, including blood and cerebrospinal fluid (CSF), in a population of 120 well-defined, moderately advanced PD subjects and 120 healthy controls.

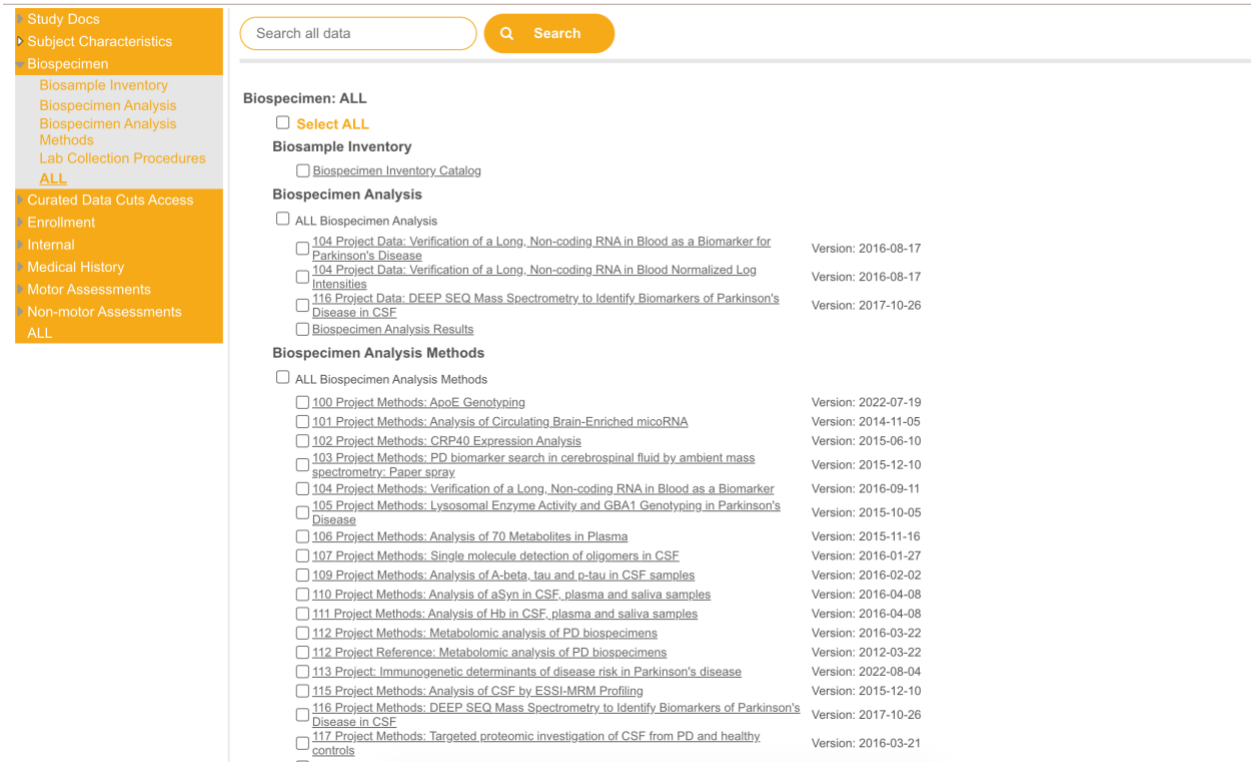



| Age Group | Number of Subjects |
|---|---|
| 50-59 | 39 |
| 60-69 | 118 |
| 70-79 | 64 |
| 80-89 | 11 |

## Download Study Data

Browse the available items within categories or use the Search feature to find items by keyword.

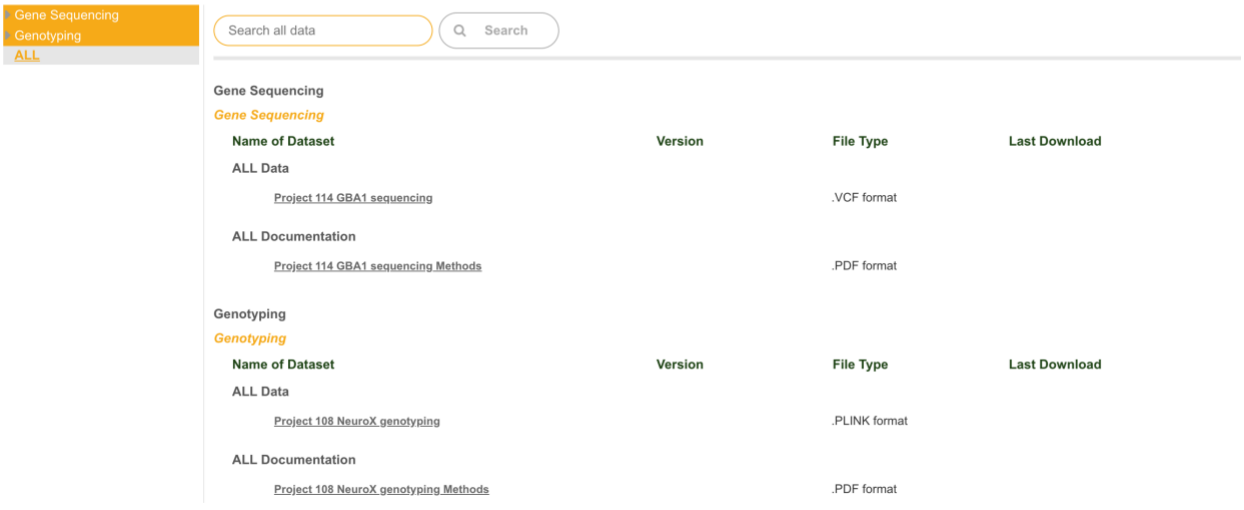





## Download Genetic Data

Browse the available items within categories or use the Search feature to find items by keyword.

***Reminder: The BIOFIND Data Use agreement prohibits unauthorized sharing of these data, posting to public databases and any attempt data to identify individuals using these data. By downloading th data you acknowledge the associated terms and conditions.***

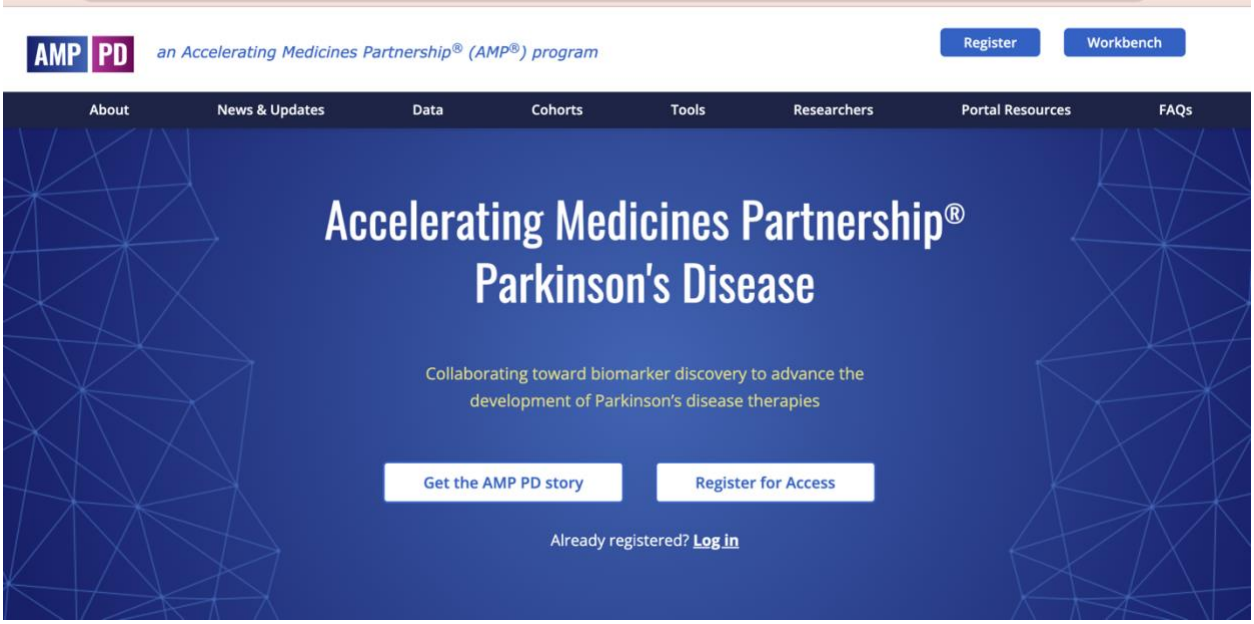


## AMP PD

### AMP PD Knowledge Platform

Within the portal you can explore a dataset, understand its contents, and build custom cohorts for deeper analysis using Jupyter notebooks. AMP PD is a public-private partnership between The Michael J. Fox Foundation, National Institutes of Health, the U.S. Food and Drug Administration, Aligning Science Across Parkinson's (ASAP), and five companies (Bristol Myers-Squibb, GSK, Pfizer, Sanofi, and Verily). AMP PD is managed through the Foundation of the National Institutes of Health (FNIH).

Access the AMP PD Knowledge Platform ↗



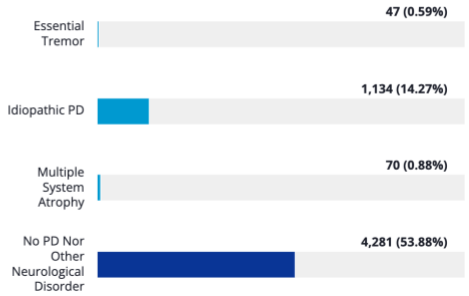
10,699
**Total Participants**

### Total Case vs Control Participants

| | |
|---|---|
| Total PD Case Participants | 3,484 |
| Total Control Participants | 4,319 |
| Total Other Diagnosis | 2,896 |

### Biological Sex Within Cohort


### Most Common Participant Diagnoses

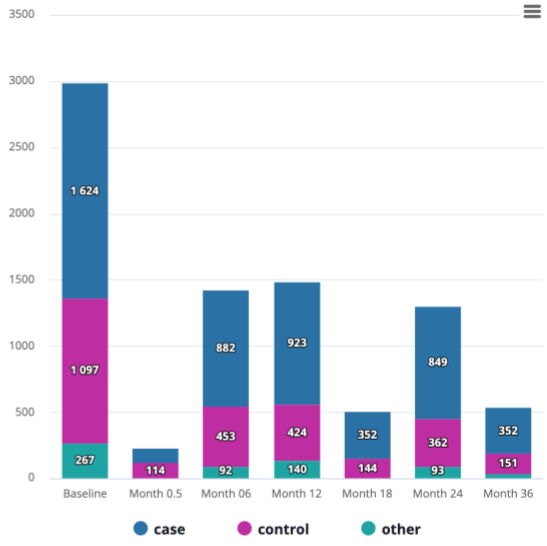



### Whole Genome Sequencing in Blood

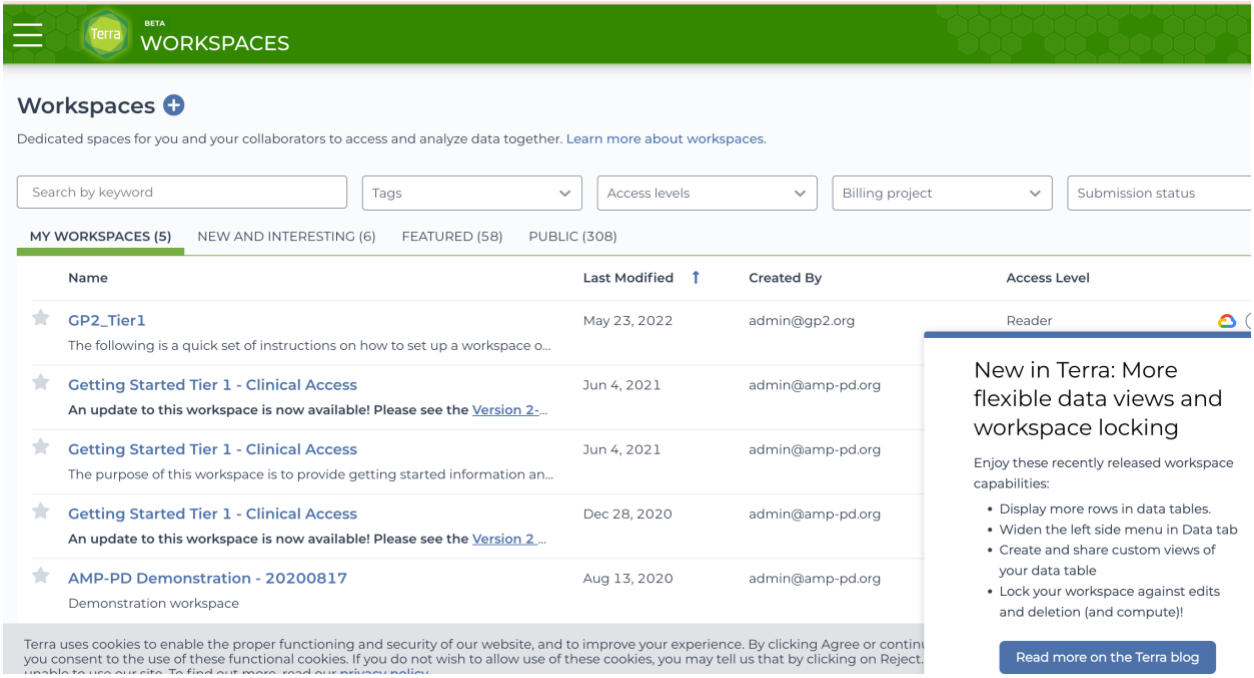
| Enrollment Type | WGS Totals |
|---|---|
| PD with Known Mutations | 830 |
| PD with No Known Mutations | 2,529 |
| Healty Control with Known Mutations | 791 |
| Healty Control with No Known Mutations | 3,362 |
| Other Diagnosis | 2,906 |

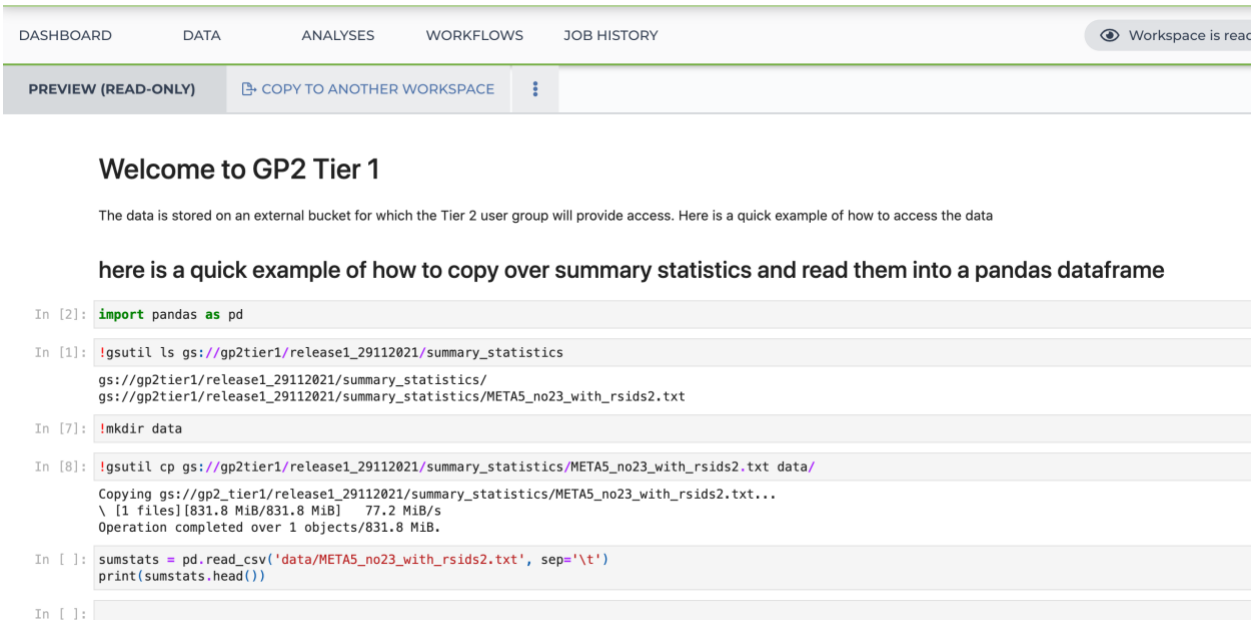
(+) = with mutation; (-) = without mutation

### RNA Sequencing in Blood




### Proteomics in Plasma and CSF


BETA
WORKSPACES
Workspaces
Dedicated spaces for you and your collaborators to access and analyze data together. Learn more about workspaces.
Search by keyword
Tags
Access levels
Billing project
Submission status
MY WORKSPACES (5)
NEW AND INTERESTING (6)
FEATURED (58)
PUBLIC (308)
Name
Last Modified
Created By
Access Level
GP2_Tier1
The following is a quick set of instructions on how to set up a workspace o...
May 23, 2022
admin@gp2.org
Reader
Getting Started Tier 1 - Clinical Access
An update to this workspace is now available! Please see the Version 2-...
Jun 4, 2021
admin@amp-pd.org
Getting Started Tier 1 - Clinical Access
The purpose of this workspace is to provide getting started information an...
Jun 4, 2021
admin@amp-pd.org
Getting Started Tier 1 - Clinical Access
An update to this workspace is now available! Please see the Version 2 ...
Dec 28, 2020
admin@amp-pd.org
AMP-PD Demonstration - 20200817
Demonstration workspace
Aug 13, 2020
admin@amp-pd.org
Terra uses cookies to enable the proper functioning and security of our website, and to improve your experience. By clicking Agree or continu
you consent to the use of these functional cookies. If you do not wish to allow use of these cookies, you may tell us that by clicking on Reject.
New in Terra: More flexible data views and workspace locking
Enjoy these recently released workspace capabilities:
Display more rows in data tables.
Widen the left side menu in Data tab
Create and share custom views of your data table
Lock your workspace against edits and deletion (and compute)!
Read more on the Terra blog


DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
Workspace is read
PREVIEW (READ-ONLY)
COPY TO ANOTHER WORKSPACE
Welcome to GP2 Tier 1
The data is stored on an external bucket for which the Tier 2 user group will provide access. Here is a quick example of how to access the data
here is a quick example of how to copy over summary statistics and read them into a pandas dataframe
In [2]: import pandas as pd
In [1]: !gsutil ls gs://gp2tier1/release1_29112021/summary_statistics
gs://gp2tier1/release1_29112021/summary_statistics/
gs://gp2tier1/release1_29112021/summary_statistics/META5_no23_with_rsids2.txt
In [7]: !mkdir data
In [8]: !gsutil cp gs://gp2tier1/release1_29112021/summary_statistics/META5_no23_with_rsids2.txt data/
Copying gs://gp2_tier1/release1_29112021/summary_statistics/META5_no23_with_rsids2.txt...
\ [1 files][831.8 MiB/831.8 MiB]  77.2 MiB/s
Operation completed over 1 objects/831.8 MiB.
In [ ]: sumstats = pd.read_csv('data/META5_no23_with_rsids2.txt', sep='\t')
print(sumstats.head())
In [ ]:

# How to set up billing in Terra

Anton Kovalsky
Updated 2 days ago

Follow

In this article



*Read on for step-by-step instructions to set up a Terra billing project, including 1) set up a Google Cloud Billing account, 2) link Cloud billing to Terra, and 3) create a Terra billing project. Once you set up your billing, many funding and resources management tasks can be done in Terra. Terra takes care of interfacing with Google Cloud.*

***Do you have STRIDES credits?*** *See How to access STRIDES for step-by-step instructions.*

## Set up billing in Terra from scratch - in three steps

Terra runs on Google Cloud; you will pay for all storage and analysis costs through a Google Cloud billing account linked to a Terra Billing Project. If you don't have access to shared workspaces or billing, follow three steps below to set up a Terra Billing project from scratch. Once you have a Billing project, you can create or clone a workspace where you can store data and do analyses. After setting it up, Terra will manage the all Google Cloud costs.

# Fox Insight

https://www.michaeljfox.org/fox-insight-data

Data Resources

# Fox Insight

## Dataset Overview

Fox Insight is an online clinical study building a large, diverse cohort of people with Parkinson's and age-matched control volunteers who share information into the lived experience, genetics, and variability of Parkinson's disease. Participants complete regular questionnaires and answer one-time surveys on topics relevant in Parkinson's. New data is added to the platform monthly. The goal of the study is to provide Parkinson's researchers with a rich dataset combining patient experiences with genetic risks and modifiers that can be used for discovery, validation, and reproducibility.

## Data Characteristics

**Study Subjects:**

- Nearly 50,000 participants and growing
- English-speaking people with Parkinson's over age 18
- English-speaking control volunteers over age 18

**Available Data:**

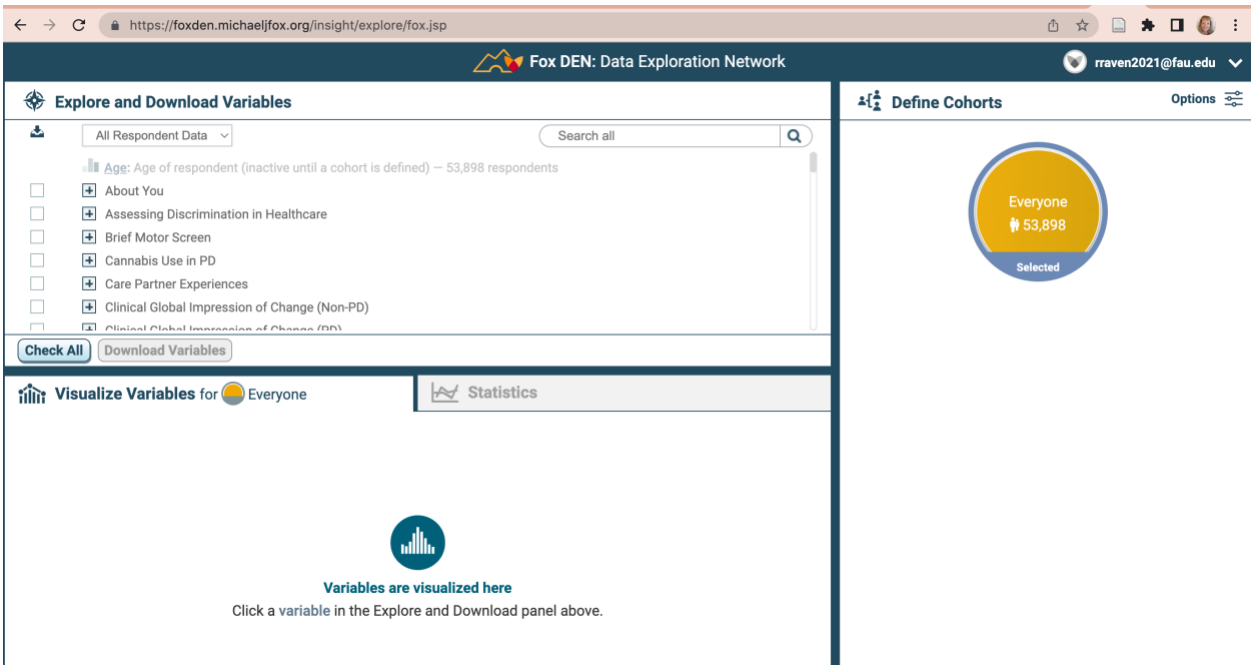
- Demographics and patient-reported outcomes from validated instruments assessed at routine intervals
- Information from one-time surveys on timely Parkinson's-related issues (e.g., medical cannabis use, COVID-19 impacts)
- Wearable device data from a subset of participants
- Genetic data from subjects with Parkinson's disease

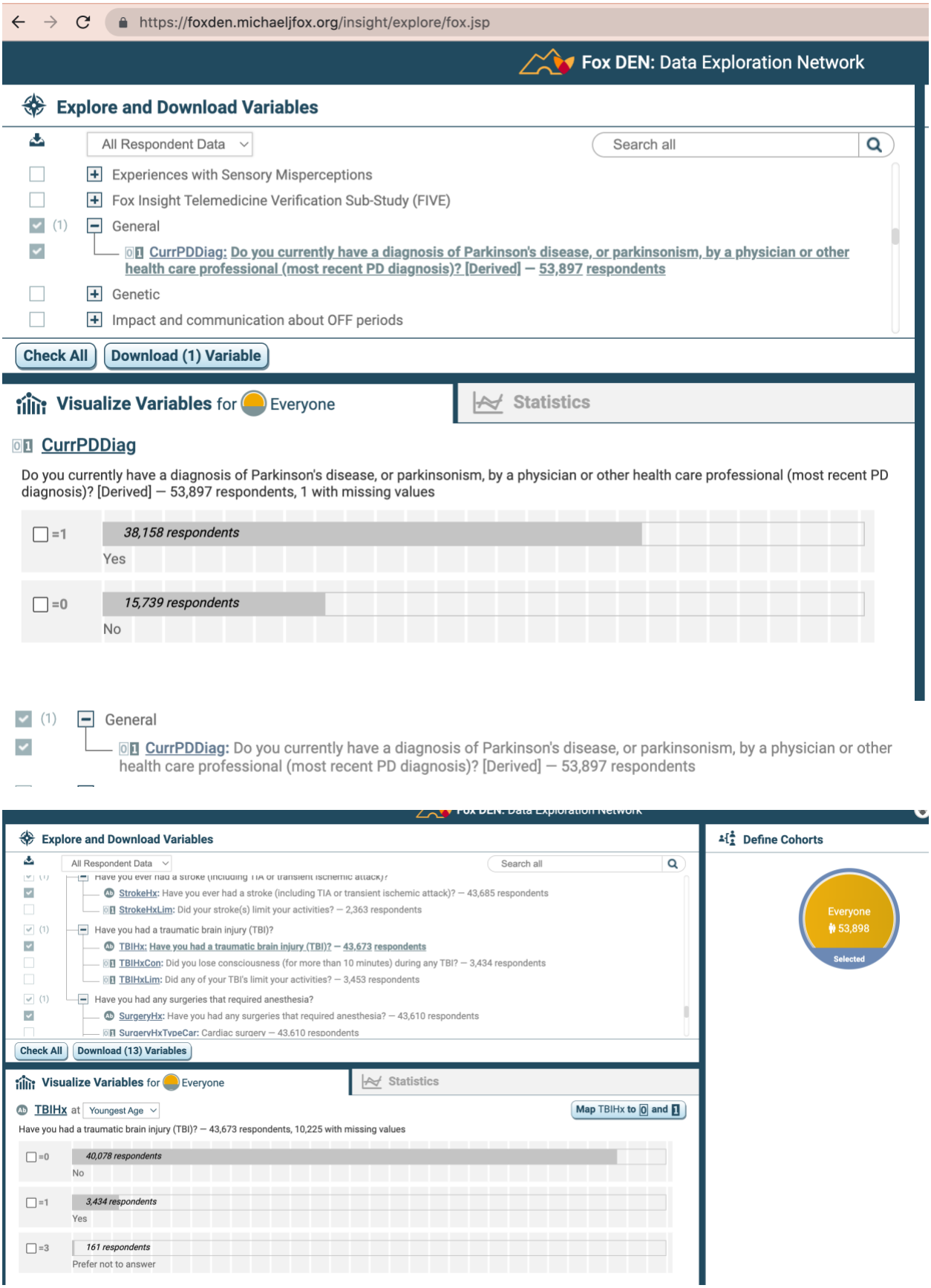
Access Fox Insight Data

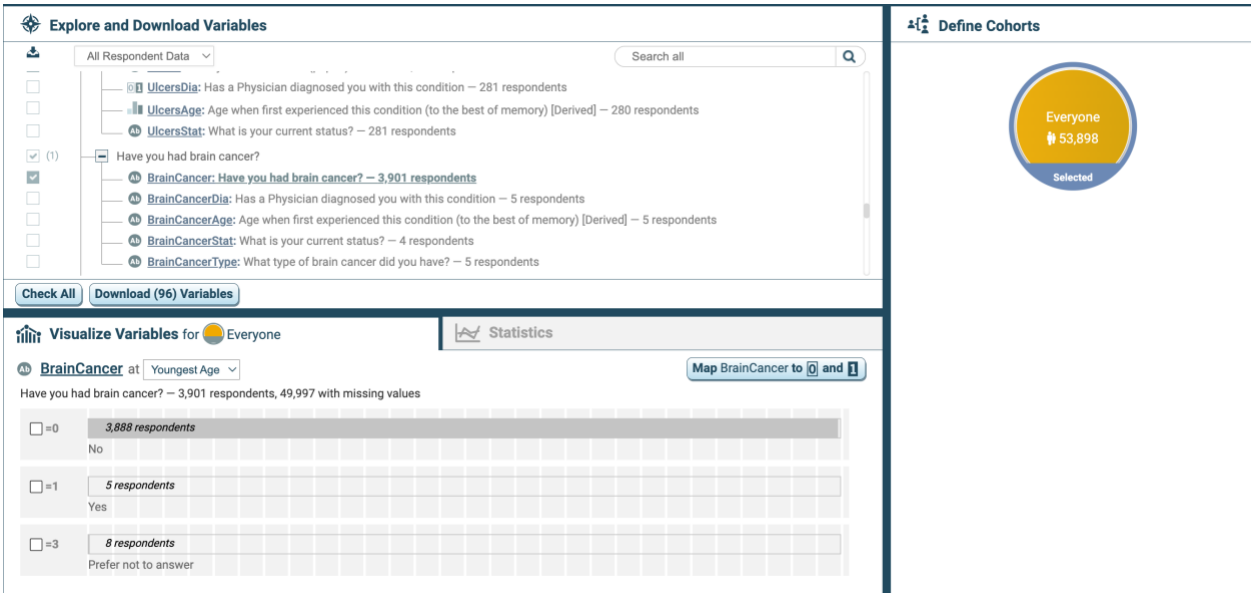
### Related Content

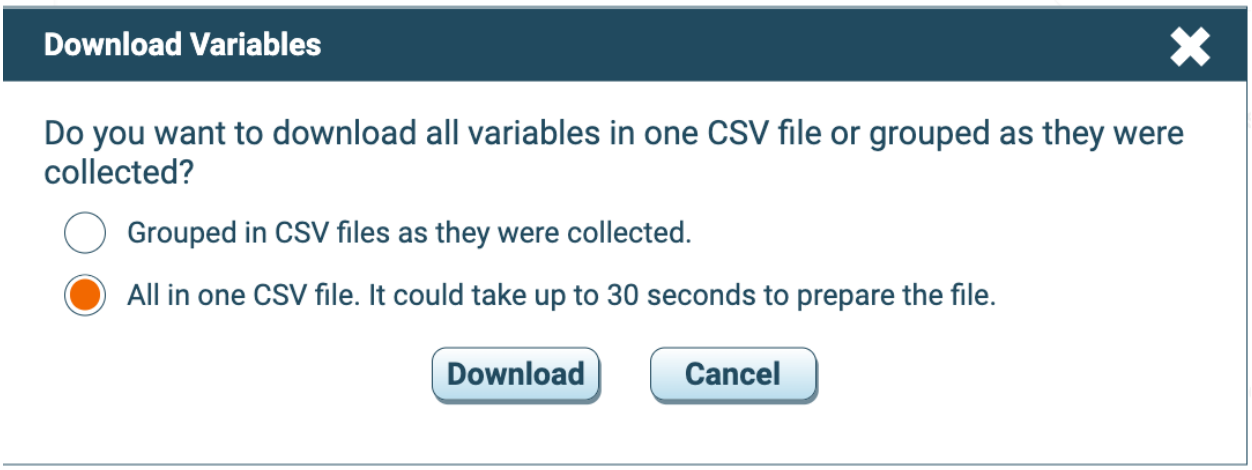
*Access Resources*
**Data Resources**

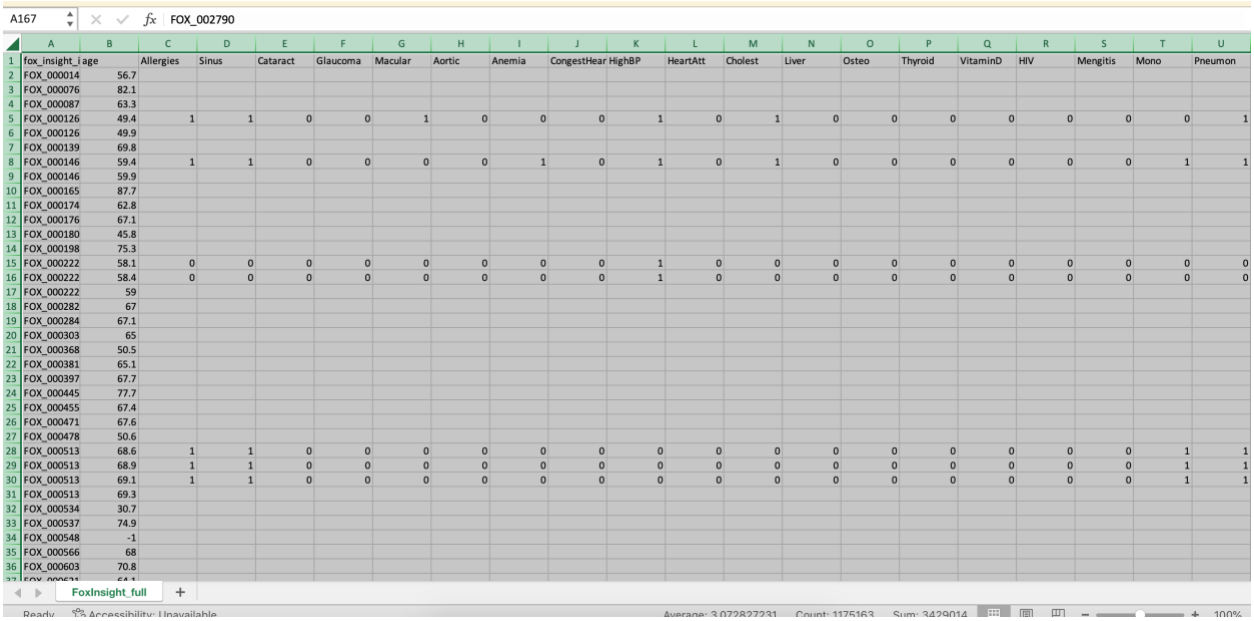
*Data Resources*
**Fox DEN**


https://foxden.michaeljfox.org/insight/explore/fox.jsp
Fox DEN: Data Exploration Network
rraven2021@fau.edu
Explore and Download Variables
All Respondent Data
Search all
Age: Age of respondent (inactive until a cohort is defined) — 53,898 respondents
About You
Assessing Discrimination in Healthcare
Brief Motor Screen
Cannabis Use in PD
Care Partner Experiences
Clinical Global Impression of Change (Non-PD)
Check All
Download Variables
Visualize Variables for Everyone
Statistics
Variables are visualized here
Click a variable in the Explore and Download panel above.
Define Cohorts
Options
Everyone
53,898
Selected


https://foxden.michaeljfox.org/insight/explore/fox.jsp
Fox DEN: Data Exploration Network
Explore and Download Variables
All Respondent Data
Search all
Experiences with Sensory Misperceptions
Fox Insight Telemedicine Verification Sub-Study (FIVE)
(1)
General
CurrPDDiag: Do you currently have a diagnosis of Parkinson's disease, or parkinsonism, by a physician or other health care professional (most recent PD diagnosis)? [Derived] — 53,897 respondents
Genetic
Impact and communication about OFF periods
Check All
Download (1) Variable
Visualize Variables for Everyone
Statistics
CurrPDDiag
Do you currently have a diagnosis of Parkinson's disease, or parkinsonism, by a physician or other health care professional (most recent PD diagnosis)? [Derived] — 53,897 respondents, 1 with missing values
=1
38,158 respondents
Yes
=0
15,739 respondents
No
(1)
General
CurrPDDiag: Do you currently have a diagnosis of Parkinson's disease, or parkinsonism, by a physician or other health care professional (most recent PD diagnosis)? [Derived] — 53,897 respondents
Explore and Download Variables
All Respondent Data
Search all
StrokeHx: Have you ever had a stroke (including TIA or transient ischemic attack)? — 43,685 respondents
StrokeHxLim: Did your stroke(s) limit your activities? — 2,363 respondents
Have you had a traumatic brain injury (TBI)?
TBIHx: Have you had a traumatic brain injury (TBI)? — 43,673 respondents
TBIHxCon: Did you lose consciousness (for more than 10 minutes) during any TBI? — 3,434 respondents
TBIHxLim: Did any of your TBI's limit your activities? — 3,453 respondents
Have you had any surgeries that required anesthesia?
SurgeryHx: Have you had any surgeries that required anesthesia? — 43,610 respondents
Check All
Download (13) Variables
Define Cohorts
Everyone
53,898
Selected
Visualize Variables for Everyone
Statistics
TBIHx at Youngest Age
Map TBIHx to 0 and 1
Have you had a traumatic brain injury (TBI)? — 43,673 respondents, 10,225 with missing values
=0
40,078 respondents
No
=1
3,434 respondents
Yes
=3
161 respondents
Prefer not to answer


Explore and Download Variables
All Respondent Data
Search all
UlcersDia: Has a Physician diagnosed you with this condition — 281 respondents
UlcersAge: Age when first experienced this condition (to the best of memory) [Derived] — 280 respondents
UlcersStat: What is your current status? — 281 respondents
(1)
Have you had brain cancer?
BrainCancer: Have you had brain cancer? — 3,901 respondents
BrainCancerDia: Has a Physician diagnosed you with this condition — 5 respondents
BrainCancerAge: Age when first experienced this condition (to the best of memory) [Derived] — 5 respondents
BrainCancerStat: What is your current status? — 4 respondents
BrainCancerType: What type of brain cancer did you have? — 5 respondents
Check All
Download (96) Variables
Visualize Variables for Everyone
Statistics
BrainCancer at Youngest Age
Map BrainCancer to 0 and 1
Have you had brain cancer? — 3,901 respondents, 49,997 with missing values
=0
3,888 respondents
No
=1
5 respondents
Yes
=3
8 respondents
Prefer not to answer
Define Cohorts
Everyone
53,898
Selected


Download Variables
Do you want to download all variables in one CSV file or grouped as they were collected?
Grouped in CSV files as they were collected.
All in one CSV file. It could take up to 30 seconds to prepare the file.
Download
Cancel

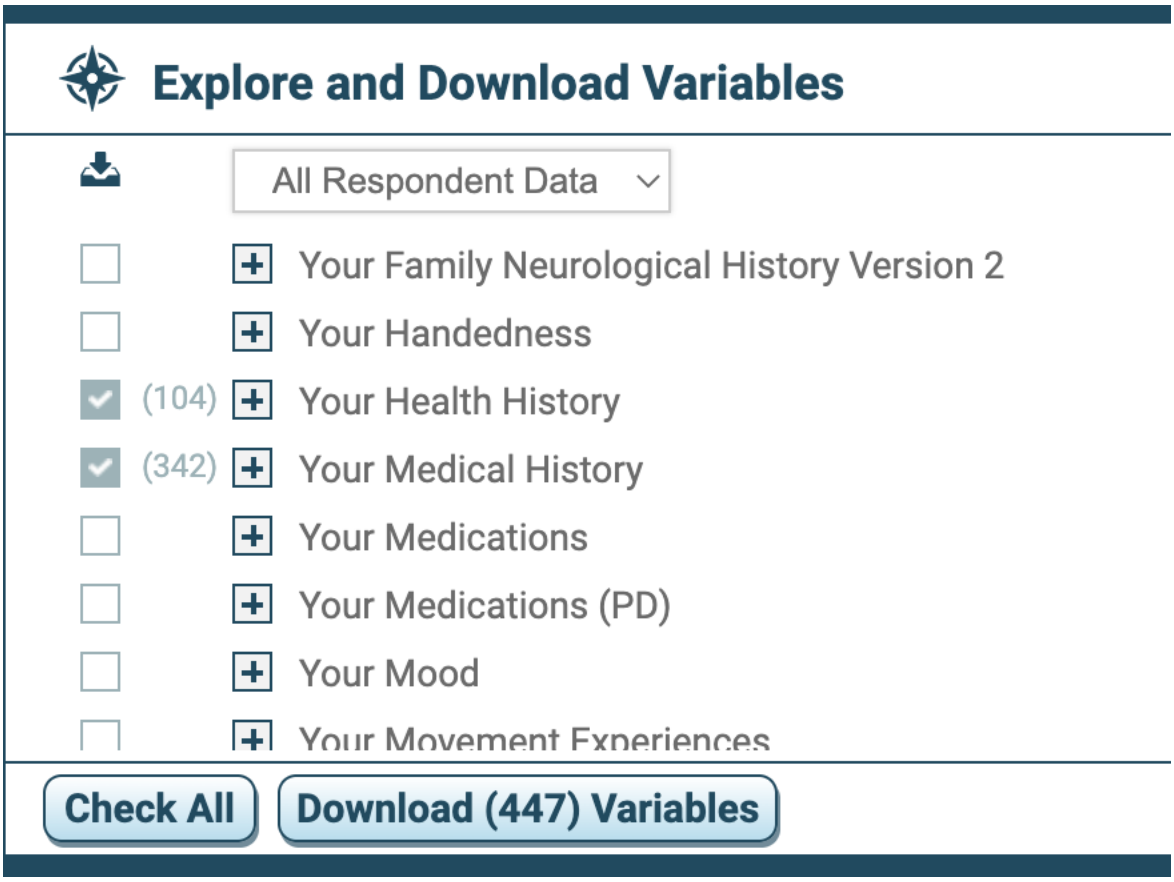


A167 FOX_002790

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | fox_insight_i | age | Allergies | Sinus | Cataract | Glaucoma | Macular | Aortic | Anemia | CongestHear | HighBP | HeartAtt | Cholest | Liver | Osteo | Thyroid | VitaminD | HIV | Mengitis | Mono | Pneumon |
| 2 | FOX_000014 | 56.7 | | | | | | | | | | | | | | | | | | | |
| 3 | FOX_000076 | 82.1 | | | | | | | | | | | | | | | | | | | |
| 4 | FOX_000087 | 63.3 | | | | | | | | | | | | | | | | | | | |
| 5 | FOX_000126 | 49.4 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 6 | FOX_000126 | 49.9 | | | | | | | | | | | | | | | | | | | |
| 7 | FOX_000139 | 69.8 | | | | | | | | | | | | | | | | | | | |
| 8 | FOX_000146 | 59.4 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 9 | FOX_000146 | 59.9 | | | | | | | | | | | | | | | | | | | |
| 10 | FOX_000165 | 87.7 | | | | | | | | | | | | | | | | | | | |
| 11 | FOX_000174 | 62.8 | | | | | | | | | | | | | | | | | | | |
| 12 | FOX_000176 | 67.1 | | | | | | | | | | | | | | | | | | | |
| 13 | FOX_000180 | 45.8 | | | | | | | | | | | | | | | | | | | |
| 14 | FOX_000198 | 75.3 | | | | | | | | | | | | | | | | | | | |
| 15 | FOX_000222 | 58.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | FOX_000222 | 58.4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | FOX_000222 | 59 | | | | | | | | | | | | | | | | | | | |
| 18 | FOX_000282 | 67 | | | | | | | | | | | | | | | | | | | |
| 19 | FOX_000284 | 67.1 | | | | | | | | | | | | | | | | | | | |
| 20 | FOX_000303 | 65 | | | | | | | | | | | | | | | | | | | |
| 21 | FOX_000368 | 50.5 | | | | | | | | | | | | | | | | | | | |
| 22 | FOX_000381 | 65.1 | | | | | | | | | | | | | | | | | | | |
| 23 | FOX_000397 | 67.7 | | | | | | | | | | | | | | | | | | | |
| 24 | FOX_000445 | 77.7 | | | | | | | | | | | | | | | | | | | |
| 25 | FOX_000455 | 67.4 | | | | | | | | | | | | | | | | | | | |
| 26 | FOX_000471 | 67.6 | | | | | | | | | | | | | | | | | | | |
| 27 | FOX_000478 | 50.6 | | | | | | | | | | | | | | | | | | | |
| 28 | FOX_000513 | 68.6 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 29 | FOX_000513 | 68.9 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 30 | FOX_000513 | 69.1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 31 | FOX_000513 | 69.3 | | | | | | | | | | | | | | | | | | | |
| 32 | FOX_000534 | 30.7 | | | | | | | | | | | | | | | | | | | |
| 33 | FOX_000537 | 74.9 | | | | | | | | | | | | | | | | | | | |
| 34 | FOX_000548 | -1 | | | | | | | | | | | | | | | | | | | |
| 35 | FOX_000566 | 68 | | | | | | | | | | | | | | | | | | | |
| 36 | FOX_000603 | 70.8 | | | | | | | | | | | | | | | | | | | |

FoxInsight_full +

Ready Accessibility: Unavailable Average: 3.072827231 Count: 1175163 Sum: 3429014 100%


Explore and Download Variables
All Respondent Data
Your Family Neurological History Version 2
Your Handedness
(104)
Your Health History
(342)
Your Medical History
Your Medications
Your Medications (PD)
Your Mood
Your Movement Experiences
Check All
Download (447) Variables

## Bonus opportunity 2: Demo notebook accessible via colab link:

https://colab.research.google.com/drive/1ZVuJZoUwUqX7_J1exc8DdXlwWJy5nM4P?usp=sharing

## Project in brief

**Data Source:**

Data used in the preparation of this project was obtained from the Fox Insight database (https://foxinsight-info.michaeljfox.org/insight/explore/insight.jsp) on 10/22/2022.

**Data Selection:**

We selected specific health conditions that could result in periods of brain hypoxia (lack of oxygen) since those events could affect brain health.

* Stroke
* Traumatic Brain Injury
* Surgery with Anesthesia

**Research Question:**

"Can a history of a stroke, traumatic brain injury, or surgery with anesthesia be used to predict the diagnosis of Parkinson's Disease?"

**Conclusion:**

The dummy classifier made a prediction that ignored the input features so we have a baseline. In this case, the baseline accuracy score of test data is 72.92%.

None of our models (KNN, decision tree, random forest, or logistic regression) outperformed the dummy model.

| Method | Accuracy on test data |
|---|---|
| Sanity/Dummy | 72.92% |
| KNN | 69.39% |
| Decision Tree | 73.01% |
| Random Forest | 72.89% |
| Logistic Regression | 72.92% |

Additionally, The coefficients were all fairly low in the logistic regression model. The surgical history had the strongest coefficient at 0.23, but that is still insignificant.

We can conclude that, at least with this dataset, there is no predictive value in a health history of stroke, traumatic brain injury, or surgery with anesthesia.

## CAP 6683_A3: Bonus¶

## Professor: Dr. Marques¶

## Student: Renee Raven¶

## Fall 2022¶

To run code: Please visit colab link and run the notebook

Use google colab link:
https://colab.research.google.com/drive/1ZVuJZoUwUqX7_J1exc8DdXlwWJy5nM4P?usp=sharing

## Project Description¶

### Data Source¶

Data used in the preparation of this project was obtained from the Fox Insight database (https://foxinsight-info.michaeljfox.org/insight/explore/insight.jsp) on 10/22/2022. For up-to-date information on the study, visit https://foxinsight-info.michaeljfox.org/insight/explore/insight.jsp.

### Research Question¶

Using data obtained from Fox Insight data resources, we developed models to investigate the predictive value of a health history of:

- Stroke
- Traumatic Brain Injury
- Surgery with Anesthesia

We seek to answer the question "Can a history of a stroke, traumatic brain injury, or surgery with anesthesia be used to predict the diagnosis of Parkinson's Disease?"

While some code is original, some code is adapted from the github accompanying Introduction to Machine Learning with Python by Andreas Muller and Sarah Guido.

If code is adapted from the book's github, it is noted in the cell.

In [1]:

```
# import libraries
import numpy as np
import pandas as pd
```

```
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.dummy import DummyClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:

```
# load the data in pandas df
try:
    url = 'https://raw.githubusercontent.com/renee127/fau_ms_projects/main/datafiles/FoxInsight_limite.csv'
    df = pd.read_csv(url)
except:
    print('File not found')
```

## Exploratory data analysis and data preparation¶

In [3]:

```
print('\nFirst 5 rows of csv file')
df.head()
```

```
First 5 rows of csv file
```

Out[3]:

| | fox_insight_id | age | StrokeHx | TBIHx | SurgeryHx | CurrPDDiag |
|---|---|---|---|---|---|---|
| 0 | FOX_000014 | 56.7 | 0.0 | 0.0 | 1.0 | 0 |
| 1 | FOX_000076 | 82.1 | 0.0 | 0.0 | 1.0 | 1 |
| 2 | FOX_000087 | 63.3 | 0.0 | 0.0 | 1.0 | 1 |
| 3 | FOX_000126 | 49.4 | NaN | NaN | NaN | 1 |
| 4 | FOX_000126 | 49.9 | 0.0 | 1.0 | 1.0 | 1 |

In [4]:

```
print('\nGeneral Info\n')
df.describe(include='all')
```

```
General Info
```

Out[4]:

| | fox_insight_id | age | StrokeHx | TBIHx | SurgeryHx | CurrPDDiag |
|---|---|---|---|---|---|---|
| **count** | 59150 | 59150.000000 | 43752.000000 | 43738.000000 | 43658.000000 | 59150.000000 |
| **unique** | 53897 | NaN | NaN | NaN | NaN | NaN |
| **top** | FOX_452999 | NaN | NaN | NaN | NaN | NaN |
| **freq** | 4 | NaN | NaN | NaN | NaN | NaN |
| **mean** | NaN | 53.955834 | 0.064500 | 0.089670 | 0.893696 | 0.713609 |
| **std** | NaN | 25.359825 | 0.284932 | 0.322053 | 0.334118 | 0.452078 |
| **min** | NaN | -1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | NaN | 49.800000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 |
| **50%** | NaN | 62.900000 | 0.000000 | 0.000000 | 1.000000 | 1.000000 |
| **75%** | NaN | 70.300000 | 0.000000 | 0.000000 | 1.000000 | 1.000000 |
| **max** | NaN | 119.000000 | 3.000000 | 3.000000 | 3.000000 | 1.000000 |

In [5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59150 entries, 0 to 59149
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   fox_insight_id  59150 non-null  object
 1   age             59150 non-null  float64
 2   StrokeHx        43752 non-null  float64
 3   TBIHx           43738 non-null  float64
 4   SurgeryHx       43658 non-null  float64
 5   CurrPDDiag      59150 non-null  int64
dtypes: float64(4), int64(1), object(1)
memory usage: 2.7+ MB
```

In [6]:

```
# print sum of missing values per column
print('Column\t\tNumber missing values\n')
print(df.isnull().sum())
```

```
Column           Number missing values

fox_insight_id        0
age                   0
StrokeHx          15398
TBIHx             15412
SurgeryHx         15492
CurrPDDiag            0
dtype: int64
```

In [7]:

```
# check data for duplicates in entire dataFrame
print('Number of duplicate rows:')
```

```
df.duplicated().sum()
```

```
Number of duplicate rows:
```

Out[7]:

```
0
```

## Observations¶

We have a dataset with 59150 rows and 5 columns, no duplicates, and many missing entries.

The Current ID, PD Diagnosis, and Age all have 59150 observations. The other rows have a variable number of missing values.

The Current ID will be remove as it does not offer any insights. Age will also be removed as it does not address our researh question.

The rows with missing information will be removed.

In [8]:

```
# remove rows with missing data in df and remove the ID column
df = df.dropna()
df = df.drop(columns=['fox_insight_id', 'age'])
print('Column\t\tNumber missing values\n')
print(df.isnull().sum())
```

```
Column           Number missing values

StrokeHx      0
TBIHx         0
SurgeryHx     0
CurrPDDiag    0
dtype: int64
```

In [9]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 43658 entries, 0 to 59149
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   StrokeHx    43658 non-null  float64
 1   TBIHx       43658 non-null  float64
 2   SurgeryHx   43658 non-null  float64
 3   CurrPDDiag  43658 non-null  int64
dtypes: float64(3), int64(1)
memory usage: 1.7 MB
```

## Create splits on features/target and train/test dfs¶

In [10]:

```
# first split train test into df_train, df_test with 20% split
df_train, df_test = train_test_split(df, test_size=0.2, random_state=42)
print('Verify sizes of newly divided dataframes\n')
print('train & test\n')
print(len(df_train), len(df_test))
```

```
Verify sizes of newly divided dataframes

train & test

34926 8732
```

In [11]:

```
# split the df_train and df_test on features, target
print('Verify rows and columns of train and test sets\n')


features_train = df_train.drop(['CurrPDDiag'], axis=1)
target_train = df_train['CurrPDDiag']
print('features_train', features_train.shape)
print('target_train', target_train.shape)


features_test = df_test.drop(['CurrPDDiag'], axis=1)
target_test = df_test['CurrPDDiag']
print('features_test', features_test.shape)
print('target_test', target_test.shape)
```

```
Verify rows and columns of train and test sets

features_train (34926, 3)
target_train (34926,)
features_test (8732, 3)
target_test (8732,)
```

In [12]:

```
features = ["StrokeHx", "TBIHx", "SurgeryHx"]
features
```

Out[12]:

```
['StrokeHx', 'TBIHx', 'SurgeryHx']
```

## Observations¶

We have a feature set and a target of a Parkinson's Diagnosis.

We have 80% of our sample allocated to the training df and 20% of our sample allocated to the test set.

## Dummy Model for Comparison¶

In [13]:

```
# sanity check the test data for comparison with other models
dummy_clf = DummyClassifier(strategy="most_frequent")
dummy_clf.fit(features_test, target_test)
dummy_clf.predict(features_test)
dummy_clf.score(features_test, target_test)


sanity_score = dummy_clf.score(features_test, target_test)
print('Sanity check of test data: {:.2%}'.format(sanity_score))
```

```
Sanity check of test data: 72.92%
```

## Observation¶

We know have a benchmark to compare our models with.

If our model(s) perform better than the dummy/sanity check model, we might have reason to further investigate our research question.

However, if our model(s) do not perform better, it is very unlikely we have a predictive relationship between our features and our target.

## KNN Model¶

In [14]:

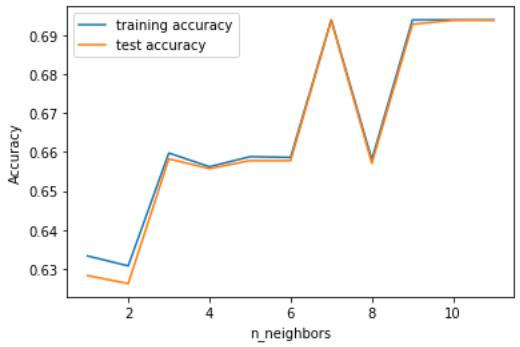
```
# code is adapted from the github accompanying Introduction to Machine Learning with P
ython by Andreas Muller and Sarah Guido.
# https://github.com/amueller/introduction_to_ml_with_python/blob/master/02-supervised
-learning.ipynb


# consider the k nearest neighbors model
training_accuracy = []
```

```
test_accuracy = []
# try n_neighbors from 1 to 10
neighbors_settings = range(1, 12)

for n_neighbors in neighbors_settings:
    # build the model
    clf = KNeighborsClassifier(n_neighbors=n_neighbors)
    clf.fit(features_train, target_train)
    # record training set accuracy
    trn_acc = clf.score(features_train, target_train)
    training_accuracy.append(clf.score(features_train, target_train))
    # record generalization accuracy
    test_acc = clf.score(features_test, target_test)
    test_accuracy.append(clf.score(features_test, target_test))
    # print("Neighbors:", n_neighbors, "Training accuracy:", trn_acc, "Test accuracy:"
, test_acc)

plt.plot(neighbors_settings, training_accuracy, label="training accuracy")
plt.plot(neighbors_settings, test_accuracy, label="test accuracy")
plt.ylabel("Accuracy")
plt.xlabel("n_neighbors")
plt.legend()
plt.show()
```




In [15]:

```
knn_train_score = max(training_accuracy)
print("Accuracy on training set: {:.2%}".format(knn_train_score))
knn_test_score = max(test_accuracy)
print("Accuracy on test set: {:.2%}".format(knn_test_score))
```

```
Accuracy on training set: 69.40%
```

```
Accuracy on test set: 69.39%
```

## Observation¶

The KNN Model did not outperform our dummy model.

## Decision Tree Model¶

In [16]:

```
# code is adapted from the github accompanying Introduction to Machine Learning with P
ython by Andreas Muller and Sarah Guido.

# https://github.com/amueller/introduction_to_ml_with_python/blob/master/02-supervised
-learning.ipynb


# consider the decision tree classifier

tree = DecisionTreeClassifier(random_state=0)

tree.fit(features_train, target_train)

tree_train_score = tree.score(features_train, target_train)

tree_test_score = (features_test, target_test)

print("Accuracy on training set: {:.2%}".format(tree.score(features_train, target_trai
n)))

print("Accuracy on test set: {:.2%}".format(tree.score(features_test, target_test)))
```

```
Accuracy on training set: 72.78%
Accuracy on test set: 73.01%
```

## Observation¶

The Decision Tree Model did not outperform our dummy model.

## Random Forest Model¶

In [17]:

```
forest = RandomForestClassifier(n_estimators=100, random_state=0)

forest.fit(features_train, target_train)


print("Accuracy on training set: {:.2%}".format(forest.score(features_train, target_tr
ain)))

print("Accuracy on test set: {:.2%}".format(forest.score(features_test, target_test)))
```

```
Accuracy on training set: 72.78%
```

```
Accuracy on test set: 72.89%
```

## Observation¶

The Random Forest Model did not outperform our dummy model.

## Logistic Regression Model¶

In [18]:

```
# create logistic regression model
lr = LogisticRegression(random_state=42, solver='liblinear')

# fit model
lr.fit(features_train, target_train)
# make predictions
predictions_train = lr.predict(features_train)
predictions_test = lr.predict(features_test)

# calculate accuracy
accuracy = accuracy_score(target_train, predictions_train)
test_accuracy = accuracy_score(target_test, predictions_test)

print('Accuracy\n')
print('Training set: {:.2%}'.format(accuracy))
print('Test set: {:.2%}'.format(test_accuracy))
```

```
Accuracy

Training set: 72.75%
Test set: 72.92%
```

In [21]:

```
lr.coef_
```

Out[21]:

```
array([[0.11645017, 0.09872905, 0.23051465]])
```

In [22]:

```
features
```

Out[22]:

```
['StrokeHx', 'TBIHx', 'SurgeryHx']
```

## Observation¶

The Logistic Regression Model model did not outperform our dummy model.

## Conclusion¶

None of our models (KNN, decision tree, random forest, or logistic regression)outperformed the dummy model. We can conclude that, at least with this dataset, there is no predicitive value in a health history of stroke, traumatic brain injury, or surgery with anesthesia.

### Accuracy using test data¶

| Method | Accuracy |
|---|---|
| Sanity/Dummy | 72.92% |
| KNN | 69.39% |
| Decision Tree | 73.01% |
| Random Forest | 72.89% |
| Logistic Regression | 72.92% |

### Coefficients in Logistic Regression¶

The coefficients were all fairly low in the logistic regression model. The surgical history had the strongest coefficient at 0.23, but that is still insignificant.