Data Community of Practice

Getting Started With Parkinson's Disease Data

Powered by THE MICHAEL J. FOX FOUNDATION FOR PARKINSON'S RESEARCH

Version 1.0: May 2024

TABLE OF CONTENTS

AUTHORS

Michael Alosco, Boston University Chobanian & Avedisian School of Medicine (Co-Lead & Contributor, USA)
Elizabeth Hutchins, TGen/DataTecnica/NIH CARD (Co-Lead & Contributor, USA)
Paula Saffie Awad, CETRAM/Clínica Santa María (Contributor, Chile/Brazil)
Victoria Dardov, Technome (Contributor, USA)
Joshua Gottesman, The Michael J. Fox Foundation (Contributor, USA)
Hirotaka Iwaki, NIH CARD/NIA/DataTecnica (Contributor, USA)
Paula Reyes-Perez, UNAM (Laboratorio Internacional de Investigación sobre el Genoma Humano, UNAM, Mexico)
Daniel Teixeira dos Santos, Hospital de Clínicas de Porto Alegre (Contributor, Brazil)
Maya Sanghvi, The Center for Scientific Collaboration and Community Engagement (Contributor, USA)

PURPOSE

Don’t Panic! This guide is intended to be a living document which enables researchers getting started with an array of datasets pertaining to PD research to do so with increased ease. The current draft contains information on four datasets: AMP PD, PPMI, GP2, and Fox Insight.

Each entry contains an overview of the following aspects of these data: study description, study data features, how to access data, intended data uses, data set strengths and limitations, links to pre-existing documentation, plus tips and other considerations for handling these data.

In addition, the guide provides a metatable enabling researchers to at-a-glance gain insight as to which entry might be worth exploring further to determine what data best fits their research purpose, by providing an overview of cohort size, types of analyses it can support, modalities of data available, and the types of clinical features it contains.

If individuals are interested in contributing an entry on a new study, or to request an update or correction, please contact: researchcommunity@michaeljfox.org.

METATABLE OF STUDY CHARACTERISTICS

This table is intended to enable researchers to determine which data best suits their research question of interest. It provides an overview of the types of analyses supported, sizes of the respective cohorts, modalities of data available, and the clinical features each study covers.

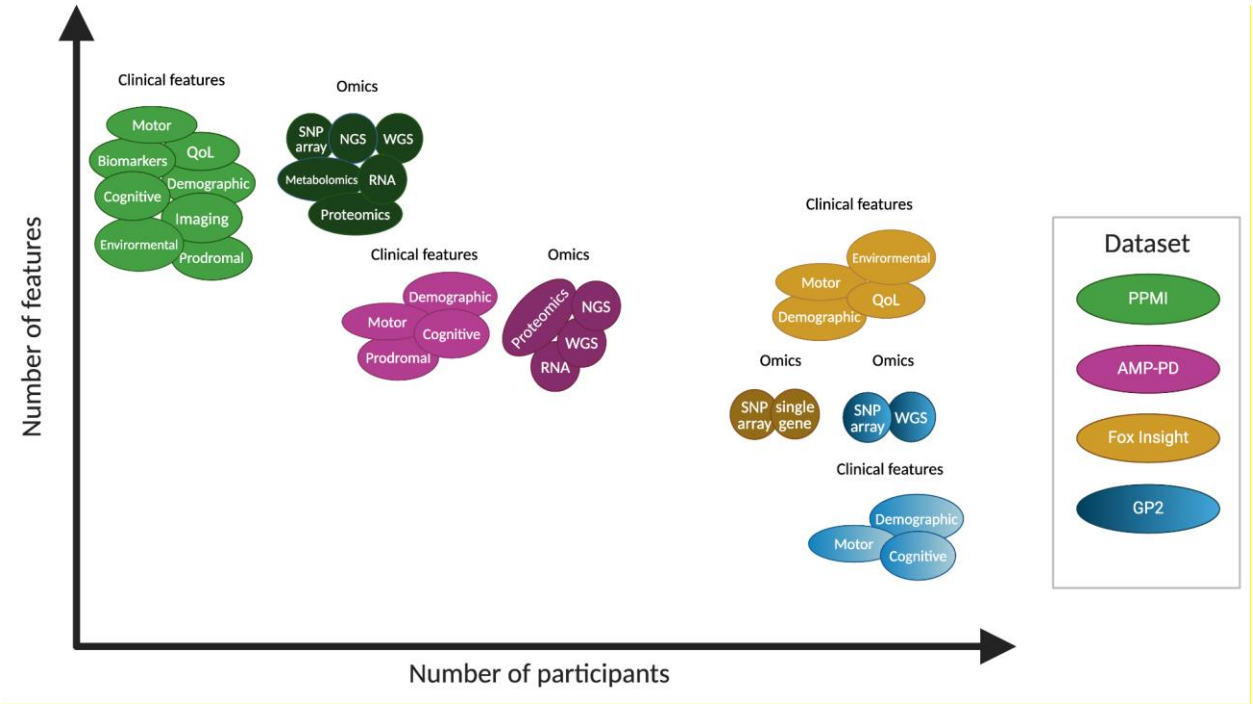
Types of Analyses Supported				
	PPMI (Clinic)	AMP PD (Aggregated)	GP2	Fox Insight
Polygenic risk score generation	Yes	Yes	Yes	Yes
Genome-wide association studies	Yes	No (only WGS or GWAS w/federated GP2 data)	Yes	
Cross-sectional	Yes	Yes	Yes	Yes
Longitudinal	Yes	Yes	No	Yes
Cohort Sizes				
PD	1,521 (902 PD, 619 prodromal)	3,375	24,709 (genotyped) 2,324 (WGS)	38,635 (10,669 genotyped + 419 microbiome)
Non-PD	237	7,100	17,246 (genotyped) 17,246 (WGS)	16,168
Modalities of Data				
Genetic data	Yes	Yes	Yes	Yes
Biologic data	Yes	Yes	No	No

Imaging data	Yes	No	No	No
Patient-reported outcome data	Yes (via PPMI-Online)	No	No	Yes
Microbiome data (stool)	Yes	No	No	Yes
Types of Clinical Data Present				
Motor	Yes	Yes	Yes	Yes
Non-Motor	Yes	Yes	Yes	Yes
Quality of Life	No	Yes	Some	Yes
Cognition	Yes	Yes	Yes	Yes

These data are accurate as of April 2024.

DATASET FEATURES

The following graphic provides a visual overview of clinical and omics features across different datasets in relation to number of different features and dataset sample size. Solid-filled bubbles correspond to datasets that possess longitudinal data, while gradient-filled bubbles correspond to datasets only with cross-sectional data.



AMP® PD

Study Description

AMP® PD (the Accelerating Medicines Partnership Parkinson's Disease) program is a partnership between FNIH, NINDS, NIA, FDA, Abbvie, GSK, Pfizer, Sanofi, BMS, Verily, ASAP and MJFF) generates and consolidates data from eight unified cohorts (BioFIND, HBS, LBD, LCC, PDBP, PPMI, STEADY-PD3 and SURE-PD3). Data was generated using standardized technology and centrally harmonized and quality controlled. All data was generated from samples collected under similar protocols. This data harmonization process and single data use policy facilitates and simplifies cross-cohort analysis. For more information on the data harmonization process, visit the data tab of the AMP-PD website.

AMP-PD's Summary Data Dashboard gives an overview of the various cohorts and provides an interactive way for users to see total participant counts for each of the cohorts of interest and what data types are available for the various cohorts. Using this dashboard, users can see there are 1,608 total participants from PDBP and 1,977 total participants from PPMI.

In addition, all samples are categorized into various PD categories and here are the total number or participants:

Enrollment Type	**WGS Totals**
PD with Known Mutations	1,483
PD with No Known Mutations	1,892
Healthy Control with Known Mutations	1,640
Healthy Control with No Known Mutations	2,554

Other Diagnosis with Known Mutations	1,490
Other Diagnosis with No Known Mutation	1,416

Study Data Features

AMP-PD Harmonized Data

Clinical Data/Measurements

- Clinical Data
  - Demographic
  - Enrollment
  - Medical History
  - MDS-UPDRS
  - MoCA
  - UPSIT
- Transcriptomic Data
  - RNASeq from whole blood
  - Illumina Novaseq sequencing
  - Gencode v29 reference
- Genomic Data
  - Whole Genome Sequencing from whole blood
  - Illumina XTen sequencing
  - Human Genome hg38 reference
- Proteomic Data
  - Proteomics from CSF and plasma
  - Olink Explore targeted proteomics analysis
  - Mass Spectrometry based Untargeted proteomics
- Single Nucleus Brain Data
  - Clinical Data
  - Whole Genome Sequencing
  - Single nucleus RNA sequencing from 5 brain regions

How to Access Data

How to and References

1. AMP PD Access Request Form
2. Setting up a Google Account
3. Setting up 2-Step Verification
4. Requirements for accessing genomics data
5. Full AMP PD Application Submission and Review Process

Users are encouraged to utilize cloud based analyses to analyze the available data. With Terra on the frontend and Google Cloud Services (GCS) on the backend, users can utilize Terra workspaces as examples to build their own analysis scripts in R or Python.

Intended Data Uses

AMP PD aims to identify and validate diagnostic, prognostic, and progression biomarkers, with the goal of improving clinical trial design and contributing to the identification of new pathways for therapeutic developments. With the many types of clinical, and genetic data standardized across multiple cohorts, including longitudinal data, this is a great resource for combining different data types, leveraging thousands of data points, and validating hypotheses across cohorts.

Data Set Strengths

1. **Data harmonization and standardization** across multiple, well characterized cohorts - large sample sizes with a reduction in batch effects
2. **Standardized assays on thousands of existing biosamples**, incorporating existing longitudinal clinical data
3. **Multiple data types paired with clinical data**, including transcriptomics, proteomics, whole genome sequencing, and post-mortem tissue sequencing. All of this can lend itself to powerful analyses for researchers.
4. AMP PD proteomic, transcriptomic, post-mortem sequencing, and clinical data contains **longitudinal data**, so scientists can use this data to do more complex time course analyses.

Data Set Limitations

1. Not all samples that have clinical data and WGS data have transcriptomics or proteomics data; there is a global inventory table that will highlight overlapping participant IDs across the data types

Pre-Existing Documentation/FAQs

- AMP-PD
- AMP-PD in Terra
- AMP-PD FAQs
- Study Contact: ACT@amp-pd.org

Tips and Dataset Considerations

This data set is harmonized with example cloud based notebooks to help get users started. The backend is GCS and users are able to download the data directly using the requester pays bucket. However, because the intended use is for users to use the data through Terra, it is a bit difficult to navigate directly to GCS and download the data.

You might also ask yourself - why access data through AMP-PD when I can go directly to a specific cohort data set, such as PPMI data from LONI. AMP PD is a harmonized data set consisting of unified cohorts. What does this mean for a researcher? There are higher numbers of samples, because the cohorts are unified and harmonized into a cohesive data set for each omics type. So yes, researchers can directly get data from one cohort, but if you want to use data from multiple cohorts to increase the number of samples in your analysis, AMP PD data would be a good way to do that.

Updates

The AMP-PD news and updates tab include release notes for recent releases.

PPMI

Study Description

PPMI is a longitudinal cohort that evaluates a variety of clinical, genomics, transcriptomics, proteomics, biomarkers, and neuroimaging data from three different patient cohorts: (1) Parkinson's disease patients, (2) patients with prodromal symptoms of high penetrance genetic mutations associated with the disease, and (3) healthy individuals. Patients are expected to have at least 5 years of follow-up.

PPMI actually encompasses various studies: PPMI Clinical (intensive in-person longitudinal evaluations of various cohorts), PPMI Remote (remote study activities using smell tests, genotyping, and digital sensor technologies), and PPMI Online (online evaluations with patient-reported outcomes). Patients can overlap between the three studies. Each study aims for a specific number of participants, as follows: PPMI Clinical (4,000), PPMI Remote (40,000) and PPMI Online (100,000).

For an overview of study participant demographics and data collection time point, users may refer to the data dashboard.This provides a limited set of characteristics, including the ability to filter by cohort, to review number of participants per visit, biological sex, ethnicity, and age.

Currently, patients in the PPMI Clinical cohort are defined by a consensus committee. This means that even though some patients may have been enrolled in a specific cohort, their classification could shift longitudinally, and they may now be categorized differently, or excluded for various reasons. This has resulted in roughly 2,000 combined PD and Prodromal PD participants, along with roughly 250 Health Controls (but depending on criteria for inclusion in your analysis, your figures may vary).

According to the latest update from the consensus committee in May 2023, the number of participants correctly defined in the main study cohorts is:

- PD (diagnosis by clinical criteria + DaTscan compatible with PD): 1,104
- Prodromal PD (either genetic, hyposmia or RBD - can overlap): 882
- Healthy Controls: 253

Study Data Features

- Clinical
  - Demographics
  - Family history
  - Medical history
  - Medication history
  - Neurological exam
  - MDS-UPDRS (Parts I, II, III and IV)
  - Hoehn & Yahr scale
  - Modified Schwab & England Activities of Daily Living
  - SCOPA-AUT
  - Geriatric Depression Scale (Short Version)
  - QUIP-Current-Short
  - State-Trait Anxiety Inventory
  - Montreal Cognitive Assessment (MoCA)
  - Neurobehavioral and neuropsychological tests (various)
  - University of Pennsylvania Smell Identification Test (UPSIT)
  - Epworth Sleepiness Scale
  - REM Sleep Behavior Disorder Questionnaire

- Neuroimaging
  - DaTSCAN (imaging and volumetry)
  - DTI (imaging and volumetry)
  - MRI (imaging)
  - PET (substudies)

- Omics
  - Genetic status related to the most common monogenic PD causes
  - NeuroX SNP data
  - Immunochip SNP data
  - Illumina NeuroBooster SNP data
  - Whole exome sequencing
  - Whole genome sequencing
  - Blood Transcriptomics (RNAseq)
  - Proteomics
  - Metabolomics

- Biomarkers (various biomarkers exists, but some of the most relevant are)
  - CSF alpha-synuclein, amyloid-beta42, p-tau181 and total tau
  - CSF alpha-synuclein seed amplification assay (SAA)

    - Serum neurofilament light polypeptide (NfL)

- Other
    - Neuropathology results

How to Access Data

Access to PPMI is provided free of charge and data to be analyzed needs to be downloaded. Gaining access to PPMI data is a straightforward process. Visit the main study site and click the "Access data" option at the top left. Here, you can apply for data access, a process requiring some personal information, detailed motivations for your intended analyses, and agreement with data usage terms. The study committee will review your request and respond within a week. Once access is granted, simply log in on this site to access them! There is also a Data User Guide with a lot of detailed information.

Intended Data Uses

PPMI's primary goals are to identify PD biomarkers and to compare clinical and non-clinical progression between different cohorts, including individuals diagnosed with PD, genetic mutations, prodromal symptoms, and healthy participants.

Data Set Strengths

1. **Evolving dataset:** currently recruiting patients and stores biological material in order for interested researchers to propose additional projects and analyses
2. **Great range of available data:** relevant evaluations, omics and neuroimaging data on a wide range of clinical and non-clinical information
3. **Longitudinal followup:** provides an opportunity to study disease progression
4. **Multicenter:** at least 50 different centers from North America, Europe, the Middle East and Africa contribute patients to the study
5. **Patients enter the study in early stages and not using PD medications:** makes possible to evaluate baseline and some follow-up progression of the disease without medication cofounders

Data Set Limitations

1. Relatively low number of patients for some genetic analyses that require a large number of individuals (such as GWAS)
2. Underrepresentation of non-European ancestry populations

Pre-Existing Documentation/FAQs

Available on the PPMI's website (does not requires login)
For data users, there is a Data User Guide which is a great place to start developing an understanding of PPMI study data – its background, its use, and how to access it.

The PPMI website also contains a lot of useful information. These include research documents such as the study protocol, schedule of activities, operations manual, biological data acquisition manual, pathology data acquisition manual, genetic data processing manual, acquisition protocols for MRI, DTI, and SPECT.

Ongoing Specimen Analysis lists ongoing or completed projects and indicates if their results are publicly available. There is also a specimen dashboard listing all biospecimens.

The PPMI Study FAQs also provide answers to frequently asked questions study-wide.

Available on the PPMI's online data access platform (LONI - requires login)
After logging in, select "Download" at the top of the home screen, and choose "Study Data."
Here, you'll find files to guide you in using PPMI data, including:

1. **Consensus Committee Analytic Dataset**: an extremely important database that must be used to officially determine to which cohort each patient belongs. Specifically, every definition of the group to which a patient belongs comes from this database.
2. **PPMI Analytic Dataset Guide**: a document that explains why the above document was created to define cohorts of patients and how to use it.
3. **PPMI Data User Guide**: a document that provides an introduction and serves as a reference for explaining how to interpret certain variables and conduct some analyses. I recommend reading it.

In the "Study Docs" tab, you'll also find the Code List and Data Dictionary, which provide meanings for each column name and their values.

Study contact form: https://www.ppmi-info.org/contact/contact_us
Study inbox: ppmi@michaeljfox.org

Tips and Dataset Considerations

Cohort definitions are not those originally present and downloaded inside the study's platform (LONI), but instead defined by a separate document (consensus committee analytic dataset) found inside the platform (as explained above).

Much of biomarker and omics data from the study comes from proposed projects. It is important to carefully read each project's documentation in order to better understand the structure and methods that were employed in the data of your interest. An overview of these projects can be found in Ongoing Specimen Analysis and the methodology for each one, inside LONI.

Some participants have duplicate entries for the same study visit that are dependent upon the participant's last dose of dopaminergic therapy (DT), which is defined as levodopa and/or dopamine agonists. The variable that determines this is called "PDSTATE" and is present in the MDS-UPDRS III questionnaire. The OFF state is defined in the PPMI protocol as more than 6 hours after the last dose of DT, and the ON state is approximately one hour after the last dose of DT.

There is a screening visit and a baseline visit - occasionally there is a datapoint that was taken at one or the other, so it is worth checking both visits (ie., blood cell counts were taken at the screening visit and not the baseline visit, and then at subsequent yearly visits).

Updates

The study data updates regularly monthly as new participants are enrolled and participants already in follow-up attend study visits. New entries into the clinical database are transferred nightly to the database. Each Sunday, a complete update to the database is conducted. Imaging data are integrated into the database separately, on a monthly basis. New specific updates are also added depending on the conclusion of proposed projects that analyze patients' biological and neuroimaging data.

GP2

Study Description

The Global Parkinson's Genetics Program (GP2) is a resource program of the Aligning Science Across Parkinson's (ASAP) initiative focused on improving understanding of the genetic architecture of Parkinson's disease (PD) by including groups traditionally underrepresented in genetics research. The ultimate goal of collecting and genotyping more than 200,000 unique samples, especially from diverse populations.

It currently includes 227 cohorts, of which 74 cohort studies with data shared to AMP® PD, and 153 that haven't.

Last updated (Release 6): December 21st 2023.

The genetic data could be found in two groups: complex disease or monogenic disease, and broken into genetically-determined ancestries:

AAC - African American / Caribbean ancestry
AFR - African ancestry
AJ - Ashkenazi Jewish ancestry
AMR - Latino and indigenous Americas populations
EUR - General European ancestry
EAS - East Asian ancestry
SAS - South Asian ancestry
FIN - Finnish population isolate (only in complex group, no monogenic information)
CAS - Central Asian ancestry
MDE - Middle Eastern ancestry
CAH - Complex Admixture History (new in this release)

Complex Admixture History (CAH)

CAH, or Complex Admixture History, is a new ancestry group introduced to GP2 for release 6. It was created in response to a large number of samples with South African and other highly admixed individuals being incorrectly predicted as CAS (Central Asian) ancestry in release 5. For release 6, the CAH ancestry group mainly contains samples from Stellenbosch University (Cape Town, South Africa), The Coriell Institute (Camden, New Jersey, United States), and the Parkinson's Foundation (Miami, Florida, United States). We consider any samples labeled as CAH to be too highly admixed to be included in analyses with other GP2 ancestry groups.

The complex disease data (genotypes), including locally-restricted samples, consists of a total of 44,831 genotyped participants (24,709 PD cases, 17,246 Controls, and 2,876 'Other' phenotypes). When removing the locally-restricted samples, these consist of 33,436 (17,129 PD cases, 13,872 Controls, and 2,435 'Other' phenotypes)

The monogenic disease data (whole genome sequences) consists of a total of 2,324 sequenced participants (1,854 PD cases, 314 Controls, and 156 Other phenotypes)
When removing the locally-restricted samples, these consist of 2,083 (1,650 PD cases, 309 Controls, and 124 'Other' phenotypes) 12,585 individuals who have extended clinical phenotyping information and matching genetic information.

Composition of release 6 per ancestry group is available on the GP2 blog.

Study Data Features

1. Clinical (Comprehensive deep clinical phenotyping data for 12,585 individuals matched with genetic information):
    - Age at diagnosis and onset
    - Primary, current, and latest diagnoses
    - Cognitive exams such as the Mini-Mental State Examination (MMSE) and the Montreal Cognitive Assessment (MoCA)
    - Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS)
    - Detailed "other" phenotypes, such as Lewy Body Dementia (LBD)

Each cohort submits one of these datasets, according to their capacity as demonstrated by the following table.

	MINIMUM	MINIMUM PLUS	CORE	EXTENDED
Demographics	✓	✓	✓	✓
Recruitment category	✓	✓	✓	✓
Family History	✓	✓	✓	✓
Behavioral/Environmental History			✓	✓
Medical History			✓	✓
Diagnostic checklist (MDS/QSBB diagnostic criteria)		✓	✓	✓
Primary diagnosis and PD certainty		✓	✓	✓
PD history		✓	✓	✓
Global PD severity (CISI-PD)		✓	✓	✓
Current Medication Status			✓	✓
nM-EDL (MDS-UPDRS Part I)			✓	✓
M-EDL (MDS-UPDRS Part II)			✓	✓
Motor (MDS-UPDRS Part III)			✓	✓
Complications (MDS-UPDRS Part IV)			✓	✓
Cognitive assessment			✓	✓
Motor (Hoehn and Yahr stage)			✓	✓
Autonomic function assessment			✓	✓
pRBD			✓	✓
Day time sleepiness				✓
Depression				✓
Orthostatic hypotension				✓
Olfactory function				✓
General ADL (Schwab & England ADL))				✓
PD EQL (PDQ8)				✓
Pain				✓

gp2 Global Parkinson's Genetics Program

Version October 2020

NB Minimum and minimum plus can be filled in from patient notes and do not require any extra research assessments by doctor or patient

Bold - clinician-completed Others can be done by another researcher, or the patient can self-complete

2. Neuroimaging → not available
3. Omics: basic genomics are available.
    - Arrays: NeuroChip and Neurobooster.
    - Whole genome sequencing.
4. Biomarkers: not available.
5. Other: none

How to Access Data

There are two levels of access within GP2: Tier 1 (which comprises summary statistics and other non-participant level data) and Tier 2 (which includes participant-level data + clinical metadata in latest releases). Links to these data follow:

GP2 Tier 1 Data (requires login)
GP2 Tier 2 Data (requires login)

As GP2 completes cohort genotyping, all data is shared through the secured AMP® PD platform. For that reason, to request access, you must complete the steps as described in the AMP-PD registration page. This means that you request access to AMP-PD and GP2 data

through the same procedure and you are granted access to both at the same time. As aforementioned, general process is:

1. AMP PD Access Request Form
2. Setting up a Google Account
3. Setting up 2-Step Verification
4. Requirements for accessing genomics data
5. Full AMP PD Application Submission and Review Process

Before 2024, the only way to explore and analyze GP2 Data was through Terra, and data was stored in Google buckets. Due to General Data Protection Regulation, there is part of GP2 data that is in another platform called Verily Viewpoint Workbench (VWB).

To gain access to the full release on VWB you must:

1. Have approved GP2 Tier 2 access
2. Fill out the GDPR-governed sample request form
3. Be a GP2 consortium member (contributing cohort, GP2 partner, or project analyses team member)

Costs depend on runtime, bytes processed, queries performed and usage of persistent disk. On webinars, they recommend that you use the default cloud environment details, and adjust them depending on the analysis you are running).

Intended Data Uses

The primary goal of GP2 is to identify genetic variants associated with Parkinson's disease risk, age of onset, disease progression, and related clinical features. Researchers use the dataset to conduct genome-wide association studies (GWAS) and other genetic analyses to uncover novel genetic risk factors and potential therapeutic targets. For this goal, GP2 datasets are made available to the broader scientific community to facilitate collaborative research and accelerate discoveries.

GP2 datasets serve as valuable resources for validating and replicating findings from previous genetic studies of Parkinson's disease.

Data Set Strengths

1. Diverse ancestry information, provides data from underrepresented populations.
2. Continuously being updated, and improved.

3. If you are part of GP2 you could get trained and learn in the process of accessing the data.
4. Summary of the information could be accessed through the cohort browser.
5. Quality control has been done for the genetic analysis (genotools).
6. Related individuals are removed (latest version).
7. Codes available in GP2 learning platform with guided material.
8. Open office hours available every week.
9. You can propose projects and get financial support for running analysis + project manager assigned.
10. You can see projects currently being carried out to avoid duplication + join them.

Data Set Limitations

1. Delay to get access to the dataset.
2. Use through Terra slows the analysis
3. New data is restricted to GP2 members
4. Heterogeneity of the data
5. No imaging nor biomarkers data
6. Cross-sectional data
7. Quality control has been done for all samples (so you cannot change parameters)

Pre-Existing Documentation/FAQs/Study Contact

Policies, guidelines and other resources
Cohort Dashboard
Monogenic Resource Map
Monogenic Portal
Data repository
GP2's opportunities page
GP2 training resources (including on Terra, bioinformatics, PD, research methods, Python, and more!)

All GP2 code, and tools for data analysis are available on Github at the official Global Parkinson's Genetics Program (GP2) code repositories.

Email cohort@gp2.org to inquire about submitting cohort samples and joining the consortium.

Tips and Dataset Considerations

- There are a lot of missing values as the dataset depends on which cohort is submitting the data.
- The results that appear in the cohort browser for rare variants are only for some genes, so you have partial information.
- For gaining access you have to present a project proposal, and you don't know if your idea has already been under analysis, so you could waste time on this process.

Updates

GP2 Updates will provide updates on the newest releases and findings pertaining to GP2. The most recent update (at time of this draft) was announced in January 2024.

FOX INSIGHT

Study Description

The following is obtained from the Fox Insight website.
“Fox Insight is a dynamic online longitudinal clinical study of people both with and without Parkinson's disease. The study seeks to enroll tens of thousands of diverse participants, making Fox Insight the largest and most representative Parkinson's research study to date. The Fox Insight platform deploys a variety of health, lifestyle, and Parkinson's routine assessments. Fox Insight is sponsored by The Michael J. Fox Foundation for Parkinson's Research.”

Fox Insight includes individuals who are 18 years or older. Participants complete online surveys that assess neurological history, motor and non-motor symptoms, quality of life, environmental exposures, and other types of outcomes. It includes cross-sectional and longitudinal assessments. See the assessments and timeline of scheduled activities PDF for additional details.

The publication: “Fox Insight collects online, longitudinal patient-reported outcomes and genetic data on Parkinson's disease provides an in depth overview of Fox Insight. Check it out for additional information.

Study Data Features

- Clinical:
    - Demographics
    - Neurological history (e.g., PD diagnosis, disease onset)
    - Health and medical history
    - Family history
    - Cognition and daily living activities
    - Motor symptoms including MDS-UPDRS
    - Emotional functioning
    - Environmental exposures
    - Substance use
    - Traumatic brain injury and repetitive head impact histories
    - Physical activity and sleep
    - Height and weight
    - Occupation
    - Factors related to social determinants of health
- Genetic

    - 23andMe saliva collection kit for genotyping on a variety of platforms: V3 platform, V4 platform, V5 platform
    - Genetic variants available in data set include those located near GBA, LRRK2, APOE, PRKN, MCCC1, BIN3, HLA locus
- There is no neuroimaging or other biomarker data linked to the data set.

How to Access Data

Data is available to qualified researchers who have agreed to the data use agreement and publication policy. Access to Fox Insight data is available at no cost. The Getting Started with Fox DEN, a Parkinson's Data Platform video discusses how to get started. Data can be downloaded from the variable explorer where users can select specific variables of interest for a custom export, or they may download a monthly data cut for the latest version of the full dataset (login required).

Intended Data Uses

All data is collected remotely, online and data is reported by patients/participants. This format allows for large scale collection of data to address ongoing knowledge gaps in PD including need for large sample sizes and longitudinal follow up. The study type might also help to improve access to people and communities not represented in clinic or in person studies. The overall goal is to better understand the risk factors, clinical symptoms, and experiences of people with and/or at risk for PD.

Data Set Strengths

1. Large sample sizes with a range of data collected on a sample enriched for having PD, allowing for detailed investigations on risk factors and clinical presentation for PD.
2. Longitudinal data including monitoring for 60 months. There are regular intervals of follow up (3 month increments, depending on the questionnaire) allowing for assessment of serial change and trajectories.
3. Evolving data set and monthly data cuts provided by Fox DEN.
4. The sample is enriched for patients with PD or at risk for PD, but there is an available comparison group of non-PD controls.

Data Set Limitations

1. All data is self-reported (with the exception of the FIVE sub-study data, which only a small number of participants with PD joined from the overall cohort), which may result in unsupervised completion of assessments, missing data, limitations to generalizability.

2. No genotyping for study participants without PD.

Pre-Existing Documentation/FAQs

The Fox Insight 'resources' section includes several potentially useful items including:

1. Data Dictionary (machine-readable)
2. Annotated Data Dictionary (human-readable)
3. Time Representation in Fox DEN (helpful for longitudinal analysis)

The annotated data dictionary, in particular, contains an introduction and quick-start guide that provides users with a basic overview of how to understand Fox Insight data and proposed approaches to common researcher questions.

Fox Insight FAQs (scroll to end of the linked page)
Study data inbox: foxden@michaeljfox.org

Tips and Dataset Considerations

Visit windows initially did not have a buffer between them, so during the earliest period of data collection, participants could have their visit window close on 1/15 and then a new one open on 1/16. To address this issue, a 30-day window between study visit close date and the opening of the next visit was instituted on 09/05/2019.

There is a complex schedule of activities that needs to be reviewed closely so that the user has a clear understanding of when each assessment was administered, including those that have only one assessment. Depending on the scientific question and data to be used, there might be a need to select data around a certain visit.

Approximately 8% of total participants as of 03-01-23 were part of Fox Insight's beta group, defined as those joining before the March 2017 soft launch of Fox Insight. Data collected during the beta period (defined as July 2014 to February 2017) could be subject to questionnaire versioning and inconsistencies associated with platform troubleshooting and optimization. Because of this, data collected during this time is not publicly available. However, data collected from participants who enrolled during this time but continued on to contribute data following March 2017 are publicly available.

A data descriptor manuscript was published in 2020 and provides additional detail on the Fox Insight study and its data. An updated version of this paper is currently under review.

Updates

The Fox Insight database is updated monthly at the beginning of each month with newly collected study data.