

# **PARKINSON'S PROGRESSIVE MARKERS INITIATIVE (PPMI)**

## **Data User Guide – updated August 28, 2024**

### CONTENTS

<b>1</b>	<b>INTRODUCTION .....</b>	<b>3</b>
1.1	About this user guide.....	3
1.2	What is PPMI? .....	3
1.3	What is in the PPMI data? .....	3
1.4	What can I use the PPMI data for? .....	5
1.5	How do I access the data? .....	5
1.6	How is the data maintained? .....	5
1.7	A note on examples in this user guide.....	5
1.8	Further information .....	5
<b>2</b>	<b>SOFTWARE TOOLS NEEDED TO WORK WITH THE PPMI DATA .....</b>	<b>6</b>
<b>3</b>	<b>FUNDAMENTALS .....</b>	<b>7</b>
3.1	Participant identifiers .....	7
3.2	Participant categorization .....	8
3.3	Enrollment cohorts.....	8
3.4	Enrollment status.....	9
3.5	Neuronal Synuclein Disease (NSD) Categorization .....	9
3.6	Selecting datasets for download.....	11
3.7	Using the data dictionary .....	11
3.8	Getting started .....	13
3.9	Identifying genetic subgroups.....	14
<b>4</b>	<b>STATIC PARTICIPANT DATA.....</b>	<b>16</b>
4.1	Types of static data and how to look up codes.....	16
4.2	Creating a participant master table.....	16
<b>5</b>	<b>STUDY VISITS AND LONGITUDINAL DATA .....</b>	<b>18</b>
<b>6</b>	<b>CLINICAL ASSESSMENTS.....</b>	<b>20</b>
6.1	MDS-UPDRS.....	20
6.2	Non-motor assessment example - UPSIT .....	23
<b>7</b>	<b>MEDICATION .....</b>	<b>25</b>

7.1	Levodopa equivalent medication .....	25
7.2	Other medication .....	27
8	BIOSPECIMENS AND PROTEOMIC DATA .....	28
8.1	Overview of biospecimen and proteomic data.....	28
8.2	Biomarker dashboard .....	29
8.3	Querying biospecimen data.....	31
8.4	Requesting new biospecimen and proteomic data .....	32
8.5	Projects relating to the Amprion alpha-synuclein seeding amplification assay (αS-SAA) 32	
8.6	Further details on other significant projects.....	35
8.7	Curated proteomic data .....	36
9	IMAGE DATA.....	37
9.1	Image metadata.....	37
9.2	Viewing and downloading images .....	38
10	GENETIC DATA .....	39
10.1	Major genetic subgroups .....	39
10.2	Other tests for known PD genetic risk factors .....	39
10.3	RNA-Sequencing data.....	41
10.4	Other genetic data .....	42
11	CURATED DATA.....	44
12	PPMI ONLINE.....	49
13	PPMI REMOTE.....	50
14	PPMI FOUND .....	51
15	DIGITAL SENSOR DATA .....	52
15.1	Roche PD monitoring app: additional information .....	53
15.2	Verily study watch derived data: additional information.....	54
16	APPENDIX A – SUMMARY OF KEY PPMI STUDY DATA .....	55
17	APPENDIX B – SQL SCRIPTS .....	66
18	APPENDIX C – R SCRIPTS.....	73
19	APPENDIX D – IMPORTING DATA INTO A RELATIONAL DATABASE.....	84
20	APPENDIX E – FURTHER DETAILS OF NSD-ISS DERIVATION.....	86
	REFERENCES .....	90

# 1 INTRODUCTION

## 1.1 About this user guide

This user guide provides an introduction to the PPMI data for first-time users as well as serving as a reference for more experienced users. The guide uses practical examples and assumes that most users come from a biomedical background and may have limited knowledge of software tools and coding.

***This version of the user guide is as of August 28, 2024. A revision is in progress and will be posted, as a new version, to the same location. Note that all code/scripts within this guide are provided as is with no guarantee of accuracy or completeness. For questions, reach out to [resources@michaeljfox.org](mailto:resources@michaeljfox.org).***

## 1.2 What is PPMI?

In 2010, The Michael J. Fox Foundation and a core group of academic scientists and industry partners launched the Parkinson's Progression Markers Initiative (PPMI) toward critically needed biological markers of Parkinson's onset and progression. PPMI has since engaged thousands of partners — more sites; scientists and clinicians; industry experts; philanthropic partners; and, most importantly, study volunteers — to build a cornerstone of Parkinson's research. Analysis from its open-access data set and available biosample library has deepened understanding of disease and informed design of dozens of therapeutic trials. For more background on PPMI, see <https://www.ppmi-info.org/about-ppmi>.

## 1.3 What is in the PPMI data?

PPMI consists of three main collections of data:

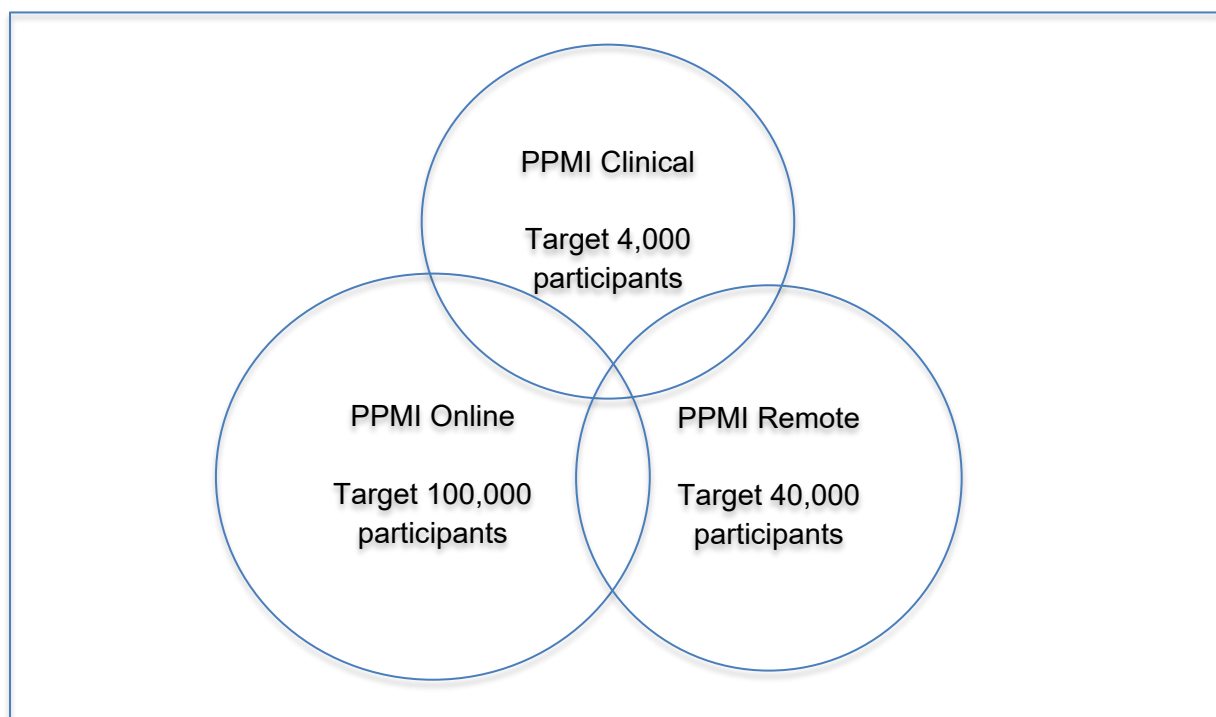
- **PPMI Clinical** contains data that has been captured by in-person clinical assessments, covering people with a confirmed diagnosis of PD, people who exhibit PD risk factors but have not yet been diagnosed with PD, and healthy controls. PPMI Clinical includes static data about the participants, history of clinical presentation using the MDS-UPDRS questionnaire,<sup>1</sup> various non-motor assessments such as the Montreal Cognitive Assessment, biospecimen analysis, genetic test results, MRI scans, DaTscan analysis, medical history, and more. Note that some of these datasets are only available for a subset of the participant population. PPMI is targeted to eventually cover 4,000 participants.

---

<sup>1</sup> Movement Disorder Society Unified Parkinson's Disease Rating Scale. See Goetz *et al.* (2008) and [https://www.movementdisorders.org/MDS-Files1/PDFs/Rating-Scales/MDS-UPDRS\\_English\\_FINAL.pdf](https://www.movementdisorders.org/MDS-Files1/PDFs/Rating-Scales/MDS-UPDRS_English_FINAL.pdf)

- **PPMI Remote** contains data that has been captured by remote assessment. The focus is on gathering data about the pre-diagnostic phase of PD. PPMI Remote collects data on olfactory symptoms and, in some cases, genetic testing, with the intention that this data can be used to identify early indicators of PD. PPMI Remote is targeted to eventually cover 40,000 participants.
- **PPMI Online** contains data that has been self-reported by participants and captured online via a web application. The data includes a subset of the data captured in PPMI Clinical. For instance, in PPMI Online only parts 1 and 2 of the MDS-UPDRS questionnaire are captured whereas in PPM Clinical, all four parts are recorded. PPMI Online contains data for both people with a confirmed diagnosis with PD, and people without a confirmed diagnosis of PD. PPMI Online is targeted to eventually cover 100,000 participants.

Note that individual participants may have data captured by one, two or all three of these methods as illustrated below.



The majority of this guide focuses on the PPMI Clinical Data. PPMI Online and PPMI Remote are discussed in Sections 12 and 13 respectively. There is also a supplementary dataset, PPMI FOUND that contains data captured from follow up telephone consultations for some PPMI Clinical participants; this is covered in Section 14.

Appendix A gives a summary of all the data that is available for immediate download.

## 1.4 What can I use the PPMI data for?

The PPMI data can be used for a wide variety of research purposes, including studying different ways in which PD presents clinically, how it progresses, how it responds to therapeutic intervention, the effect of genetic factors, as well as studies relating to biomarkers, comorbidities, epidemiology and so on.<sup>2</sup> Any and all uses of the PPMI dataset and biospecimens must conform with the study's Data Use Agreement and the Biospecimen Use Agreement.

## 1.5 How do I access the data?

You first need to have access approved, after which you can download the data. For further information on these steps, please refer to <https://www.ppmi-info.org/access-data-specimens/download-data>. Note that certain additional data such as raw 'omics, raw digital sensor, and sequestered data are not available from the standard download site but can be requested.

## 1.6 How is the data maintained?

**PPMI data is updated continuously by the study team. In addition to new data being added, in some cases existing data may be updated, and there is no formal version control.** As such, it is important that you note the date when you took a download of the data. We also recommend developing a repeatable process for downloading and analyzing the data, so that you can easily update your analysis with the latest data if desired in the future.

## 1.7 A note on examples in this user guide

In this user guide we show a number of example queries and result sets. Where these relate to aggregation of data you will likely get different results because our examples were taken at a point in time and new records may have been subsequently added. Where the examples relate to individual participants, we have used dummy data and you will again get different results.

## 1.8 Further information

The [Guidance Resources](#) page on the PPMI website contains the latest data dictionary and code book (see Sections 3.7 and 4.1 respectively) and other useful resources such as a training video for the SAA data (see Section 8.5). Additional resources may be added to this page in the future.

If you cannot find the answer you are looking for in the documentation, you can submit a query to the MJFF Data Resources team [here](#).

---

<sup>2</sup> To reference PPMI in a research paper, refer to the [PPMI Publication Policy](#).



## 2 SOFTWARE TOOLS NEEDED TO WORK WITH THE PPMI DATA

The PPMI data is mostly stored as a series of relational tables. When downloaded, most of the data will be in the form of comma separated variable (CSV) files. These are suitable for importing into a number of tools including spreadsheets such as Microsoft **Excel**, statistical packages such as **SPSS**, or Integrated Development Environments (IDEs) such as **RStudio**.<sup>3</sup>

You may find that a single tool, such as R running under RStudio, is sufficient for your needs. A combination of tools can also work well, particularly if you want to do more complex data manipulation. In such cases, you may find it useful to first import the data into a Relational Database Management System (RDBMS) such as **Microsoft SQL Server**,<sup>4</sup> **Oracle**, or **PostgreSQL** where you can use Structured Query Language (SQL) scripts to join tables and manipulate the data efficiently, and then export the results to a tool like R or SPSS for analysis and visualization. Appendix D provides some guidance on importing PPMI data into an RDBMS.

Throughout this user guide we show example data extracts. We have provided scripts for these in both SQL (Appendix B) and R (Appendix C). You may find these to be a useful starting point for your own analysis.

---

<sup>3</sup> RStudio can be downloaded for free from <https://posit.co/download/rstudio-desktop/>

<sup>4</sup> As an example, Microsoft SQL Server Management Studio Express Edition which runs on Windows and requires 16 GB of memory to run effectively can be downloaded for free from <https://www.microsoft.com/en-GB/sql-server/sql-server-downloads>

## 3 FUNDAMENTALS

Before you can start working with the PPMI data you need to identify which participant<sup>5</sup> records to select and which tables to download, and you need to understand a few fundamentals about how the data is organized. In the following, we make the assumption that you will work mostly with PPMI Clinical. For information about PPMI Online and PPMI Remote, refer to Sections 12 and 13 respectively.

### 3.1 Participant identifiers

Each participant is identified by a unique **PATNO**.<sup>6</sup> This is common across PPMI Clinical, PPMI Remote and PPMI Online. Below is an extract from the **Participant\_Status** table in the PPMI Clinical data which is likely to be your starting point (note: synthetic data for illustrative purposes; some columns not shown for clarity).

PATNO	COHORT	COHORT_DEFINITION	ENROLL_DATE	ENROLL_STATUS	STATUS_DATE	ENROLL_AGE
9001	2	Healthy Control	Jan-11	enrolled	May-21	64.0
9002	1	Parkinson's Disease	Feb-11	enrolled	Sep-21	47.1
9003	1	Parkinson's Disease	Mar-11	enrolled	Sep-21	67.5
9004	1	Parkinson's Disease	Apr-11	enrolled	Jan-22	55.2
9005	2	Healthy Control	Apr-11	enrolled	Jan-22	59.6
9006	1	Parkinson's Disease	Apr-11	Excluded	Apr-11	
9007	1	Parkinson's Disease	May-11	Withdrew	Oct-12	77.5
9008	1	Parkinson's Disease	May-11	Withdrew	Jun-12	61.2
9009	2	Healthy Control	Jun-11	enrolled	Apr-21	81.0
9010	2	Healthy Control	Jun-11	enrolled	May-21	79.7

**PATNO is an important field that you will use to link all data relating to an individual participant.**

<sup>5</sup> We use the term 'participant' in this guide, but you will sometimes see the term 'patient' in the underlying data. The two terms can be considered as synonymous.

<sup>6</sup> To protect participant confidentiality, the actual participant names or similar identifiers are not available to PPMI users.

## 3.2 Participant categorization

PPMI Clinical participants have been categorized in several ways:

- At time of enrollment participants were assigned to one of five **cohorts**
- With the advent of the seed amplification assay (SAA) test along with certain other data, participants have been assigned to a Neuronal Synuclein Disease (**NSD**) category and, where appropriate have been given a NSD integrated staging system (**NSD-ISS**) value
- Extensive **genetic** test results are available
- At a point in time, participants have an enrollment **status**

PPMI data continues to evolve and additional categorization schemes may be introduced in the future. Moreover, these may not always be in agreement with one another. As such you should think carefully about your study objectives and consequently the way you select the population of participants for your study. The cohort assignment is based on clinical features whereas the NSD categorization is based on objective assessment of bio samples. The NSD categorization is more recent but, at the time of writing, is not yet available for all participants.

In the past, there was also a PPMI Consensus Committee that interpreted and adjudicated on the available data to define a consensus committee analytic dataset which gave a recommended categorization for each participant. This is no longer available or maintained now that NSD has been introduced.

The cohort and NSD categorizations, along with the participant status are described below. Genetics is described in detail in Section 10.

## 3.3 Enrollment cohorts

At the time of enrollment, all PPMI Clinical participants were assigned to one of five **COHORTS** (found in the Participant\_Status table), namely:

1. **Parkinson's Disease**, i.e., people who have a formal diagnosis of Parkinson's disease (PD)
2. **Prodromal**, i.e., people who are at risk of developing PD based on clinical features, genetic variants or other biomarkers but have not been formally diagnosed
3. **Healthy Controls**, i.e., people with no neurologic disorder and no first-degree relative with PD
4. **SWEDD** (Scan without dopaminergic deficit). This is a small legacy cohort that you may wish to exclude, depending on your research purpose; for more details, see <https://www.ppmi-info.org/study-design/study-cohorts#legacy/>



5. **Early Imaging** is a cohort of participants with a confirmed diagnosis of PD who were untreated and underwent additional tests including DaTscan and AV-133 imaging.<sup>7</sup> For most analyses, this cohort can be combined with the Parkinson's Disease cohort, though data available for this cohort differs slightly from the Parkinson's Disease cohort.

PPMI Online has only two cohorts, Parkinson's and Non-Parkinson's, with the latter including prodromal cases and healthy controls.

Whether you use the cohort categorization and, if so, which cohorts you need, will depend on the nature of your study. For further information on cohorts in general, see <https://www.ppmi.info/study-design/study-cohorts>.

### 3.4 Enrollment status

The **ENROLL\_STATUS** field can be found in the Participant\_Status table and can take the following values which relate to the current status of the participant in the study:

- **Pending** – Is awaiting initial screening and, as such, does not yet have all the initial study data captured
- **Screened** – Has been screened for inclusion in the PPMI study but has not yet had all the initial study data captured
- **Declined** – After screening, the participant declined to give consent and no further study data was captured
- **Screen\_failed** and **Excluded** – After screening, the participant was deemed not suitable for the PPMI study and hence no further study data was captured
- **Enrolled** – Enrolled in the PPMI study and has data from one or more study visits recorded
- **Withdrew** – Participant withdrew from the study but had completed one or more study visits
- **Complete** – Participant completed all of their planned study visits and elected to not re-enroll in further study visits

Given the above definitions, you are most likely to want to include participants with a status of **Enrolled**, **Withdrew** or **Complete**.

### 3.5 Neuronal Synuclein Disease (NSD) Categorization

Starting in 2024, participants are being assigned to NSD categories and, where applicable, an NSD-ISS stage. From a data perspective, this information encompasses several variables, which are described in the table below. At the time of writing, these fields are only available for

---

<sup>7</sup> Details of the data captured for early imaging can be found here: <https://blackfynn-ppmi-schema.readthedocs-hosted.com/en/stable-documentation-branch/index.html>

download in the Curated Data Cut (see Section 11) and are not yet populated for all participants.

The overall NSD category is based on both the presence of pathologic neuronal  $\alpha$ -synuclein (S) and dopaminergic neuronal dysfunction (D). For the NSD field to have a positive status, the S test needs to indicate a type 2 positive result (i.e. Lewy Body pathology) and the D test needs to indicate a positive result. If no information from the D test is available then NSD is set to a value of unable to determine. A baseline D test is used where available, otherwise a test after one year. Note that this means NSD is defined based on biological rather than clinical features.

For participants who have a positive S result<sup>8</sup>, additional computation is done to derive an NSD-ISS value; this integrates the biological anchors (S and D, plus G which contains genetic status) with the degree of functional impairment across motor, cognitive and other non-motor domains to give a stage categorization<sup>9</sup>.

Field name	Field description	Values	Value description
S	Presence of neuronal $\alpha$ -synuclein from the SAA CSF test	0	Negative test result
		1	Positive test – Type 1 (LBD-like)
		2	Inconclusive
		3	Positive test – Type 2 (MSA-like)
D	Dopaminergic dysfunction determined by DaTscan	0	Negative DaTscan test result
		1	Positive DaTscan test result: age-/sex-expected specific binding ratio (SBR) < 0.75
G	Genetic status	0	No genetic variant
		1	Low genetic risk
		2	High genetic risk
		3	SNCA genetic variant present
NSD	Presence of neuronal $\alpha$ -synuclein disease	0	Non-NSD
		1	NSD
		[blank]	Unable to be evaluated
NSD-ISS stage	Neuronal $\alpha$ -synuclein disease stage	Stage 0	Presence of SNCA genetic variant only
		Stage 1A	Positive S result only
		Stage 1B	Positive S and D results only
		Stage 2A	Positive S and D and clinical symptoms without functional impairment
		Stage 2B	Positive S and D, and increasing levels of functional impairment
		Stage 3	
		Stage 4	
		Stage 5	
		Stage 6	

<sup>8</sup> Further details of the SAA process and test categorization are given in Section 8.5

<sup>9</sup> Further details of the staging scheme can be found in Simuni *et al.* (2024)

Further details of the derivation of NSD-ISS can be found in Appendix E.

In the examples in this user guide, we mostly use the clinically derived cohorts, but we encourage research teams to also consider using NSD.

## 3.6 Selecting datasets for download

As a starting point, tables you are likely to always need to download from the download page (<https://ida.loni.usc.edu/pages/access/studyData.jsp>) are:

- Study Docs
  - Data & Databases
    - Code List (Annotated)
    - Data Dictionary (Annotated)
- Subject Characteristics
  - Patient Status
    - Participant Status
  - Subject Demographics
    - Demographics
- Motor Assessments
  - Motor / MDS-UPDRS
    - MDS-UPDRS Part I Patient Questionnaire
    - MDS-UPDRS Part I Non-motor Aspects of Daily Living
    - MDS-UPDRS Part II Patient Questionnaire: Motor Aspects of Daily Living
    - MDS-UPDRS Part III Treatment Determination and Part III: Motor Examination

To download these, select the “Study Data” option from the Download drop down menu. Appendix A will help you determine which additional data files to download for your study.

In general, it is not recommended to use the download ALL option at the bottom of the list as this will take up a lot of storage, though you may find it helpful to use the ALL option for some of the data subgroups.

## 3.7 Using the data dictionary

The PPMI Clinical data dictionary is very useful for understanding both the content and format of tables. This can be downloaded from the main download page (see previous section) or directly from the [Guidance Resources](#) page. There is a separate data dictionary for PPMI Online, also available on the main download page. The following notes will help you use these documents effectively:

1. In PPMI Clinical, there are two versions of the data dictionary, a version with the name suffix “**\_Annotated**”, and another with the suffix “**\_Harmonized**”. The two are identical except that the “\_Annotated” one provides some additional comments, hence in general, we would recommend using the “\_Annotated” version from the main download page.
2. The field **MOD\_NAME** provides the table name as it exists in the PPMI database used by the data management team responsible for maintaining the data but when you download the file itself to your directory, it will generate a different filename. To see the mapping between the two, open the data dictionary in Excel and filter on **ITM\_NAME** = (Blanks). The **DSCR** column will then show the download filename. For example, the MOD\_NAME “AE” maps to the physical table name “Adverse\_Event\_Log” and “CLCKDRAW” maps to “Clock\_Drawing”.

	A	B	C	D
1	MOD_NAME	ITM_NAME	PAG_NAME	DSCR
2	AE		AE	Adverse Event Log
38	AVCNSNT		AVCNSNT	Early Imaging Documentation of Informed Consent
56	AVELIG		AVELIG	Early Imaging Eligibility
76	AVIMAG		AVIMAG	Early Imaging AV-133 Imaging
102	AVPREGNANC		AVPREGN	Early Imaging Pregnancy Test
119	AVREPPREG		AVREPPRE	Early Imaging Report of Pregnancy
131	AVSCRNFAIL		AVSCRNF	Early Imaging Screen Fail
143	AVTEL		AVTEL	Early Imaging Adverse Event Telephone Assessment
218	CLCKDRAW		CLCKDRAW	Clock Drawing

3. The field **PAG\_NAME**, meaning “Page Name” is a generally used identifier indicating the version of electronic Case Reference Form (eCRF) used. It is typically the same as the MOD\_NAME for data collected via the electronic data capture system used for the clinical protocol.
4. **ITM\_NAME** refers to an individual column in a table.
5. **ITM\_TYPE** gives the datatype of a column (one of CHAR, DATE, NUMBER, TEXT or TIME). **FLD\_LEN** gives the size of the field, which may include decimal places given by **DECML**.

There are also codebooks for PPMI Clinical and PPMI Online, and these are covered in Section 4.

## 3.8 Getting started

Once you've set up your chosen tool and imported the data, you can quickly generate basic statistical information, for example the following SQL queries return the number of participants in each cohort and each status respectively.<sup>10,11,12</sup>

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Number of participants by cohort */
SELECT COHORT_DEFINITION, COUNT(*) AS 'Record Count'
FROM PPMI.dbo.Participant_Status
GROUP BY COHORT_DEFINITION
ORDER BY COUNT(*) DESC
```

COHORT_DEFINITION	Record Count
Parkinson's Disease	1097
Prodromal	731
Healthy Control	297
SWEDD	81
Early Imaging (original study participants only)	19

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Number of participants by status */
SELECT ENROLL_STATUS, COUNT(*) AS 'Record Count'
FROM PPMI.dbo.Participant_Status
GROUP BY ENROLL_STATUS
ORDER BY COUNT(*) DESC
```

ENROLL_STATUS	Record Count
enrolled	1357
Withdrew	307
Excluded	245
Declined	114
Complete	102
screen_failed	53
screened	45
pending	2

<sup>10</sup> In these examples, we assume you loaded the data into a database called 'PPMI' under user 'dbo'.

<sup>11</sup> As noted in the Introduction, you may get different results when executing these and other example queries in this user guide.

<sup>12</sup> Note that the files normally download with a date appended to the filename, e.g., Participant\_Status\_24Jul2023.csv, so in each case, you will need to either take this into account when importing the data or modify the script accordingly

To achieve the same as the above in RStudio, you could execute the following R script: <sup>13,14,15,16</sup>

```
# This script is provided as is with no guarantees of completeness or accuracy
# Participant counts by cohort and enroll status. NOTE: Requires tidyverse
package to be installed.
library (readr)
library (dplyr)

setwd ("C:\\\\PPMI")
Participant_Status <- read_csv ("Participant_Status.csv")

group_by (Participant_Status, COHORT_DEFINITION) %>% summarize(RECORD_COUNT =
n()) %>% arrange(desc(RECORD_COUNT))
group_by (Participant_Status, tolower(ENROLL_STATUS)) %>% summarize(RECORD_COUNT
= n()) %>% arrange(desc(RECORD_COUNT))
```

### 3.9 Identifying genetic subgroups

To get a count of participants in the Parkinson's Disease and Prodromal cohorts by genetic subgroup you can execute code (see Script 1 in appendices) that filters on ENROLL\_STATUS and COHORT, giving the following results.

COHORT_DEFINITION	Genetic subgroup	Participant count
Parkinson's Disease	SRDC	452
Parkinson's Disease	NULL	199
Parkinson's Disease	LRRK2	172
Parkinson's Disease	GBA	88
Parkinson's Disease	SNCA	29
Parkinson's Disease	PARKIN	8
Parkinson's Disease	PINK1	1
Parkinson's Disease	Multiple factors	1
Prodromal	LRRK2	230
Prodromal	GBA	189
Prodromal	HPSM	115
Prodromal	RBD	106
Prodromal	Multiple factors	16

<sup>13</sup> Change the C:\\PPMI to the location of your downloaded files

<sup>14</sup> Note that the files normally download with a date appended to the filename, e.g., Participant\_Status\_24Jul2023.csv, so in each case, you will need to either rename the file or modify the script accordingly

<sup>15</sup> The ENROLL\_STATUS has some statuses capitalized and some not capitalized, for example "Withdrew" and "withdrew", hence the use of the tolower() function in the R script to ignore the case

<sup>16</sup> Hereafter, please refer to Appendix B for SQL scripts and Appendix C for corresponding R scripts.



COHORT_DEFINITION	Genetic subgroup	Participant count
Prodromal	SNCA	9
Prodromal	PINK1	1
Prodromal	NULL	1

Note that we have introduced a 'Multiple factors' option as participants can sometimes test positive for both LRRK2 and GBA risk factors. You should consider whether you want to include these participants in your analysis.

You may also want to consider excluding participants with a NULL value for the subgroup as the reason for their allocation to the cohort may not be clear.

## 4 STATIC PARTICIPANT DATA

### 4.1 Types of static data and how to look up codes

Static data (i.e., data that does not change over time) about participants can be found in several tables, including the following in PPMI Clinical:

- **Demographics**
- **Family\_History**

Most fields in these tables use codes; you are likely to want to use the decode value at some point. For instance, SEX on the Demographics table is stored as a 0 for Female and 1 for Male. The decode values can all be found in the PPMI code book table which can be downloaded from [the Guidance Resources page](#). In the following examples we refer to this table as the Codes table. There is a similar code book for PPMI Online also available in the same download site.

To determine the number of males and females in the PPMI data who have a status of enrolled, withdrew or complete, and either a confirmed PD diagnosis or are prodromal, you could use Script 2 in the appendices. This queries the Demographics table to retrieve SEX, joins this with the Codes table to provide the decoded value (Female or Male) and also joins with the Participant\_Status table using PATNO to filter on ENROLL\_STATUS and COHORT\_DEFINITION. It then performs a “group by” to aggregate the data.

COHORT_DEFINITION	SEX	Participant Count
Parkinson's Disease	Female	351
Parkinson's Disease	Male	536
Prodromal	Female	292
Prodromal	Male	271

### 4.2 Creating a participant master table

You may find it useful to create a Participant\_Master table that contains all the static participant data that you need for your analysis in one place by joining multiple tables using PATNO. You can then use this as a starting point to define the participant population for your study. Script 3 in the appendices joins the Demographics, Participant\_Status and Codes tables and selects a population of all participants who are in either the Parkinson's or Prodromal cohorts and who have a status of Enrolled, Withdrew or Complete. The result is a table like the following.





PATNO	BIRTHDT	COHORT_ DEFINITION	Genetic subgroup	ENROLL _AGE	ENROLL _DATE	ENROLL _STATUS	SEX	HANDED	PD diagnosis date
9001	Jan-56	Parkinson's Disease	SRDC	64.0	Mar-11	enrolled	Female	Right	03/01/20 10
9002	Aug-73	Parkinson's Disease	GBA	47.1	Mar-11	enrolled	Female	Right	02/01/20 10
9003	Aug-54	Parkinson's Disease	SRDC	67.5	Apr-11	enrolled	Female	Right	11/01/20 09
9004	Oct-63	Parkinson's Disease	LRRK2	55.2	May-11	Withdrew	Male	Right	09/01/20 10
9005	Nov-56	Parkinson's Disease	LRRK2	59.6	May-11	Withdrew	Male	Left	02/01/20 11
9006	Jun-54	Parkinson's Disease	SRDC	60.1	Jun-11	enrolled	Male	Mixed	03/01/20 11
9007	Jan-43	Parkinson's Disease	SRDC	77.5	Nov-11	Complete	Female	Right	01/01/20 11
9008	Jun-49	Parkinson's Disease	SRDC	61.2	Dec-11	Withdrew	Male	Right	03/01/20 11
9009	Sep-35	Parkinson's Disease	SRDC	81.9	Apr-12	enrolled	Male	Left	07/01/20 12
9010	May-38	Parkinson's Disease	NULL	79.7	Apr-12	Enrolled	Female	Right	08/01/20 12

(Data shown is synthetic for illustrative purposes).

This is just an example, and many more fields could be added to this.

## 5 STUDY VISITS AND LONGITUDINAL DATA

The PPMI data contains many measurements that change over time, for example, the results of MDS-UPDRS questionnaires. Once recorded this data does not change but it is important to note that additional measurements and other data may be added to participants' records in the future.

To illustrate this longitudinal feature of the data here is an extract of the MDS-UPDRS\_PART\_I table for participant #9001 (note: synthetic data for illustrative purposes; some columns not shown for clarity).

PATNO	EVENT_ID	INFODT	NP1COG	NP1HALL	NP1DPRS	NP1ANXS	NP1APAT	NP1DDS	NP1RTOT
9001	SC	01/2012	0	0	1	0	0	0	1
9001	BL	03/2012	0	0	0	0	0	0	0
9001	V01	05/2012	0	0	0	0	0	0	0
9001	V02	08/2012	0	0	1	0	0	0	1
9001	V03	11/2012	0	0	1	0	0	0	1
9001	V04	03/2013	0	0	1	1	0	0	2
9001	V05	09/2013	0	0	0	1	0	0	1
9001	V06	05/2014	0	0	0	0	0	0	0
9001	V07	09/2014	0	0	0	0	0	0	0
9001	V08	04/2015	0	0	0	1	0	0	1
9001	V09	11/2015	0	0	1	0	0	0	1
9001	V10	04/2016	0	0	0	1	0	0	1
9001	V11	06/2016	1	0	0	1	0	0	2
9001	V12	09/2016	1	0	1	1	0	0	3
9001	V14	03/2017	1	0	1	1	0	0	3
9001	V15	03/2018	2	0	1	1	1	1	6
9001	V17	09/2020	2	1	2	1	1	0	7
9001	R17	11/2021	2	1	2	2	2	0	9

The **EVENT\_ID** is the identifier of the PPMI visit at which the individual measurement was taken and the **INFODT** indicates the date on which the measurement was taken.<sup>17</sup> **EVENT\_ID is an important field that, in combination with PATNO, you will used to link test and measurement data relating to the same point in time for an individual participant.**<sup>18</sup> *Please note, due to changes in the PPMI data collection infrastructure, as of the date of this version of the user guide, the recommended best practices for use of the EVENT\_ID for longitudinal analysis is under review. Once this review concludes, a new version of this user guide reflecting these changes will be posted to the same location.*

<sup>17</sup> Dates in PPMI are rounded to protect participant privacy.

<sup>18</sup> EVENT\_ID also sometimes appears as CLINICAL\_EVENT in the data

The columns to the right show the actual measurement values, in this example the scores for MDS-UPDRS Part 1 questions like cognitive issues (NP1COG), hallucinations (NP1HALL) etc. The full list of column definitions can be found in the data dictionary described in section 3.4.

Note that there are some special values of EVENT\_ID. **BL** represents the baseline measurement when the participant formally entered the PPMI study. In some cases, this will have been preceded by **SC**, a screening measurement to assess suitability to join the study and capture static data such as demographics, typically 60 calendar days before the baseline. Note that, for confirmed cases of Parkinson's, the date of diagnosis will typically be different from either of these events.

The **V** numbers represent the subsequent study visits on which tests were performed and measurements recorded. The **R** numbers represent where measurements were recorded remotely. There are also **U** numbers which relate to unscheduled visits. The following document gives the planned schedule of tests performed at these points in time, though actual schedules and measurements recorded for individual participants may differ from this:

[https://www.ppmi-info.org/sites/default/files/docs/PPMI-2.0-AM-2-Protocol\\_SCHEDULE-OF-ACTIVITIES.pdf](https://www.ppmi-info.org/sites/default/files/docs/PPMI-2.0-AM-2-Protocol_SCHEDULE-OF-ACTIVITIES.pdf)

In some cases, individual projects have taken additional planned measurements at intermediate times, for instance V1 at 3 months and V3 at 9 months in addition to the V2 at 6 months in the standard schedule. Where available, the INFOT field can give information about the date of an intermediate measurement, or there might be information in an associate Method document, but occasionally it might be necessary to infer the date of an intermediate measurement.

## 6 CLINICAL ASSESSMENTS

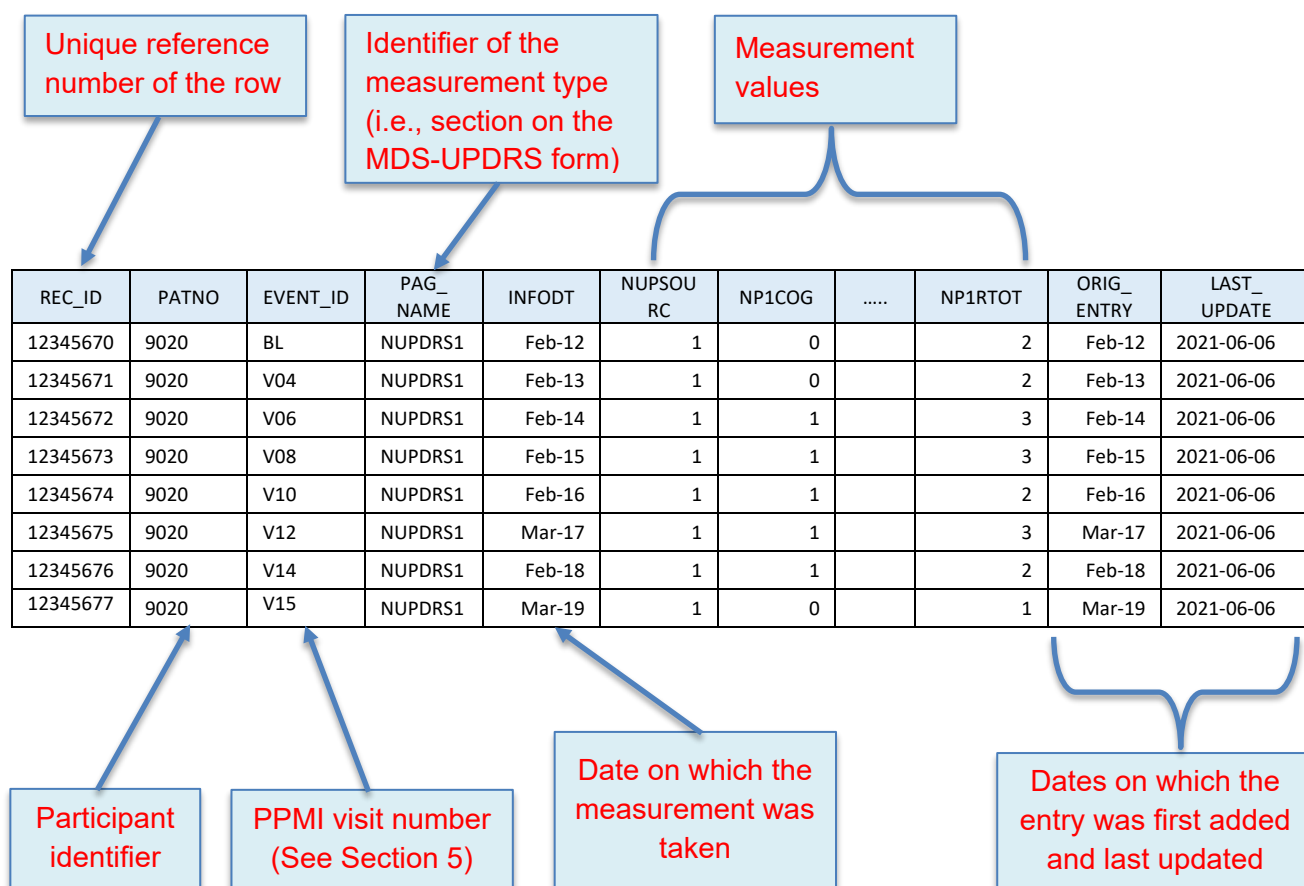
PPMI Clinical has captured a wide range of clinical assessment data, relating to both motor and non-motor symptoms of PD. Here we step through two examples. Other clinical assessments follow similar patterns.

### 6.1 MDS-UPDRS

Clinical measurements using the MDS-UPDRS questionnaire have been recorded at regular intervals for most of the PPMI participants and are stored in the following tables that mirror the different sections of the form, some of which are completed by a medical professional and some of which are completed by the participant.

- MDS-UPDRS\_Part\_I
- MDS-UPDRS\_Part\_I\_Patient\_Questionnaire
- MDS\_UPDRS\_Part\_II\_Patient\_Questionnaire
- MDS-UPDRS\_Part\_III
- MDS-UPDRS\_Part\_III\_ON\_OFF\_Determination\_\_\_Dosing
- MDS-UPDRS\_Part\_IV\_Motor\_Complications

These tables, and tables associated with other clinical measurements, all follow a similar structure:



The diagram illustrates the structure of the MDS-UPDRS table. Callouts point to specific columns: 'Unique reference number of the row' points to REC\_ID; 'Identifier of the measurement type (i.e., section on the MDS-UPDRS form)' points to PAG\_NAME; 'Measurement values' points to NP1COG and NP1RTOT; 'Participant identifier' points to PATNO; 'PPMI visit number (See Section 5)' points to EVENT\_ID; 'Date on which the measurement was taken' points to INFODT; and 'Dates on which the entry was first added and last updated' points to ORIG\_ENTRY and LAST\_UPDATE.

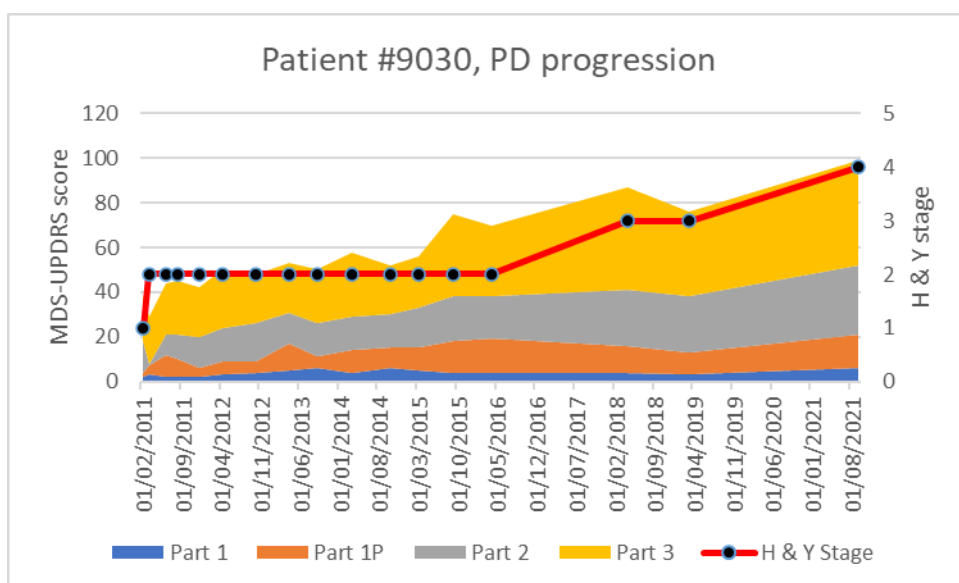
REC_ID	PATNO	EVENT_ID	PAG_NAME	INFODT	NUPSOURC	NP1COG	.....	NP1RTOT	ORIG_ENTRY	LAST_UPDATE
12345670	9020	BL	NUPDRS1	Feb-12	1	0		2	Feb-12	2021-06-06
12345671	9020	V04	NUPDRS1	Feb-13	1	0		2	Feb-13	2021-06-06
12345672	9020	V06	NUPDRS1	Feb-14	1	1		3	Feb-14	2021-06-06
12345673	9020	V08	NUPDRS1	Feb-15	1	1		3	Feb-15	2021-06-06
12345674	9020	V10	NUPDRS1	Feb-16	1	1		2	Feb-16	2021-06-06
12345675	9020	V12	NUPDRS1	Mar-17	1	1		3	Mar-17	2021-06-06
12345676	9020	V14	NUPDRS1	Feb-18	1	1		2	Feb-18	2021-06-06
12345677	9020	V15	NUPDRS1	Mar-19	1	0		1	Mar-19	2021-06-06

Script 4 in the appendices extracts the progression of MDS-UPDRS scores and also the Hoehn & Yahr stage<sup>19</sup> over time for a participant by taking the total scores from each section of the questionnaire.

INFODT	EVENT_ID	Part 1	Part 1P	Part 2	Part 3	H & Y Stage
2011-02-01	SC	2	2	14	21	1
2011-03-01	BL	3	4	1	21	2
2011-06-01	V01	2	10	9	23	2
2011-08-01	ST	2	8	11	24	2
2011-12-01	V03	2	4	14	22	2
2012-04-01	V04	3	6	15	26	2
2012-10-01	V05	4	5	17	22	2
2013-04-01	V06	5	12	14	22	2
2013-09-01	V07	6	5	15	24	2
2014-03-01	V08	4	10	15	29	2
2014-10-01	V09	6	9	15	22	2
2015-03-01	V10	5	10	18	23	2
2015-09-01	V11	4	14	20	37	2
2016-04-01	V12	4	15	19	32	2
2018-04-01	V14	4	12	25	46	3
2019-03-01	V15	3	10	25	38	3
2021-09-01	V17	6	15	31	47	4

(Note: synthetic data for illustrative purposes)

It is then straightforward to export this data to another tool for analysis and visualization. For example, the image below was plotted in Excel from the above data.



<sup>19</sup> This is the original rather than the modified scale. See Hoehn & Yahr (1967)

Note the following:

1. Occasionally there can be **missing MDS-UPDRS measurements**. These will normally appear as NULL values but in part 3 of MDS-UPDRS can also appear with a value of “101” which means “unable to rate”. In the above example visits with missing values were excluded by adding checks for nulls in the join statements. Alternative approaches are to set NULLs to a default value like zero (not recommended) or to impute a value such as a mean. Users should consider what missing value treatment is most appropriate based on their specific research question(s).
2. Sometimes there are **multiple MDS-UPDRS measurements on the same date**. These are because measurements may be taken when the participant is experiencing a medication “on” period and when the participant is experiencing an “off” period. In the above example we used the MAX function to take the maximum score on a given day, thereby ensuring just one value (the “worst case”) per date. Alternatives are either to take all values or filter on either “on” or “off” values. To filter on “on”/“off”, use column **PDSTATE** on MDS\_UPDRS\_Part\_III though note that in a small number of cases this field is missing data (NULL value), in which case you need instead query the PAG\_NAME field which will have a value of NUPDR3OF or NUPDR3ON<sup>20</sup>. See also the document “Methods for Defining PD Med Use” available from the download page under the folder Study Docs/Study Methods.
3. The **MDS-UPDRS scores can fluctuate over time**, i.e., they can go down as well as up. This can be caused by a number of factors including changes to levodopa and other medication, subjectivity of measurements and normal fluctuation of MDS-UPDRS scores, for example resulting from “on” and “off” periods. Typically, the score taken during an “off” period will be higher (i.e. worse) than the score taken during an “on” period, but this may not always be the case for the reasons just outlined. The long-term trend is for MDS-UPDRS scores to increase; for example, over a large population of people with PD, on average we see a 2.5 to 3 point annual increase in MDS-UPDRS part III scores measured in the off state.

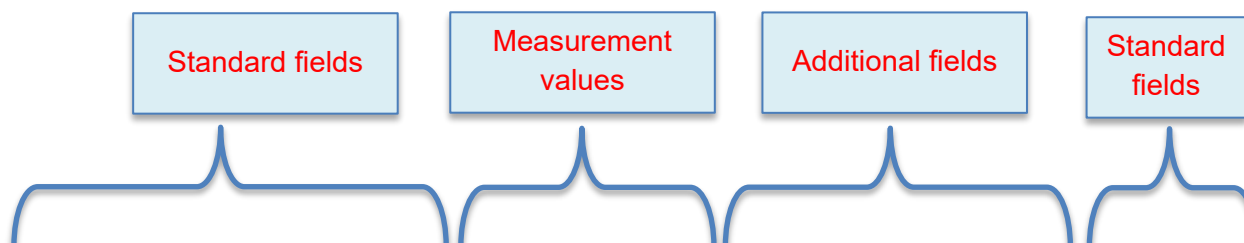
---

<sup>20</sup> In the future all ne0w records will have PAG\_NAME = ‘NUPDRDOSE3’. This version of the user guide is as of April 25<sup>th</sup>, 2023. A revision is in progress and will be posted, as a new version, to the same location.

## 6.2 Non-motor assessment example - UPSIT

PPMI has captured a lot of data relating to clinical assessments other than MDS-UPDRS, but it is important to note that these are not always recorded at every study visit or for every participant.

These assessments are all stored in a similar way but have individual differences. Take the example of the file `University_of_Pennsylvania_Smell_Identification_Test__UPSIT_` as shown below (note: synthetic data for illustrative purposes).



REC_ID	PATNO	EVENT_ID	PAG_NAME	INFODT	SCENT_01_CORRECT	.....	TOTAL_CORRECT	UPSIT_FORM	UPSIT_PRCNTGE	UPSIT_PRCTV ER	UPSIT_SOURCE	ORIG_ENTRY	LAST_UPDATE
12345681	9050	V04	UPSITCLINIC	Mar-21	0		32		55.5	Dec-20	Clinical Site	Mar-21	17/01/2022
12345682	9060	V04	UPSITCLINIC	Apr-21	0		38	2	59	Dec-20	Clinical Site	Apr-21	25/08/2021
12345683	9070	V04	UPSITCLINIC	Feb-21	1		35		35	Dec-20	Clinical Site	Feb-21	30/04/2021
12345684	9080	V04	UPSITCLINIC	Jun-21	0			2		Dec-20	Clinical Site	Jun-21	03/12/2021
12345685	9090	V04	UPSITCLINIC	Jul-21	1		34		40	Dec-20	Clinical Site	Jul-21	20/05/2021

As shown above there are standard fields, namely:

- REC\_ID: Uniquely identifies the individual row on the table
- PATNO: The unique identifier of the participant
- EVENT\_ID and INFODT: Capture longitudinal information about the data
- PAG\_NAME: Takes the value UPSITCLINIC for tests performed under PPMI Clinical, or UPSITPRO for tests performed under PPMI Remote (see also section 13)
- ORIG\_ENTRY: The date the record was added to the PPMI database
- LAST\_UPDATE: The date the record was last updated (will often be the same as ORIG\_ENTRY)

You won't normally need to be concerned with REC\_ID, ORIG\_ENTRY and LAST\_UPDATE.

The table also contains the actual measurement values and may also contain additional informational columns specific to the test. To understand these, open the data dictionary spreadsheet in Excel and filter on MOD\_NAME (in this case "UPSIT" – see also section 3.7). This will reveal, for example that:

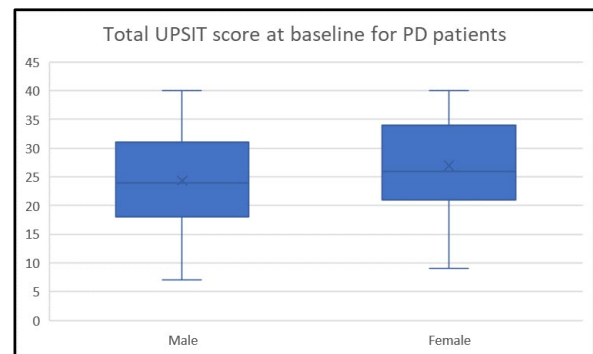
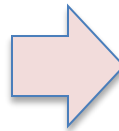
- UPSITFORM is the version of the UPSIT form administered
- UPSIT\_SOURCE is the source of the study where the results come from (which, in this case, can also be determined from the PAG\_NAME field described above)

Because these fields use codes, we can further open the Codes\_List spreadsheet in Excel and again filter on MOD\_NAME = “UPSIT”, then filter on either the column name (ITM\_NAME in the spreadsheet) and/or a specific code value (CODE in the spreadsheet) to understand what the data means. This tells us the values that may appear in these fields:

- UPSITFORM: 1 = Original, 2 = Revised
- UPSIT\_SOURCE: Clinical site or Remote

Suppose we want to look at the relationship between performance in the UPSIT test and sex for participants with a diagnosis of Parkinson’s disease. Script 5 in the appendices joins several tables to prepare the data, which can then be used in an application like Excel or R for analysis and visualization (data shown is synthetic for illustrative purposes).

PATNO	SEX	TOTAL_CORRECT
9041	Male	24
9042	Male	33
9043	Male	24
9044	Female	28
9045	Female	31
9046	Female	31
9047	Female	26
...	...	...



Note that missing individual measurement values appear as blank fields in the data. In the specific case of UPSIT, if just one or two of the individual test results are missing out of the total of 40 tests, then these will be assumed to be correct and an overall score will be imputed. A flag will be added to note these instances.

Details of the methods used to capture UPSIT data, and many other tests can be found in PDF documents which can be downloaded from the main download site. In this case, there is a document PPMI\_UPSIT\_Methods\_Document\_18NOV2016.pdf which can be found under the group ‘Non-motor Assessments’ and subgroup ‘Olfactory Tests’.



## 7 MEDICATION

### 7.1 Levodopa equivalent medication

PPMI Clinical contains information about medications taken by the participant, both dopaminergic drugs such as levodopa, and other medications.

Dopaminergic medications, including levodopa, dopamine agonists and monoamine oxidase inhibitors, are stored as levodopa equivalent daily dosage (LEDD) and this can be found on the **LEDD\_Concomitant\_Medication\_Log** table. Here is a sample of this table for participant #9052 generated by Script 6 in the appendices (note: synthetic data for illustrative purposes; some columns omitted for clarity):

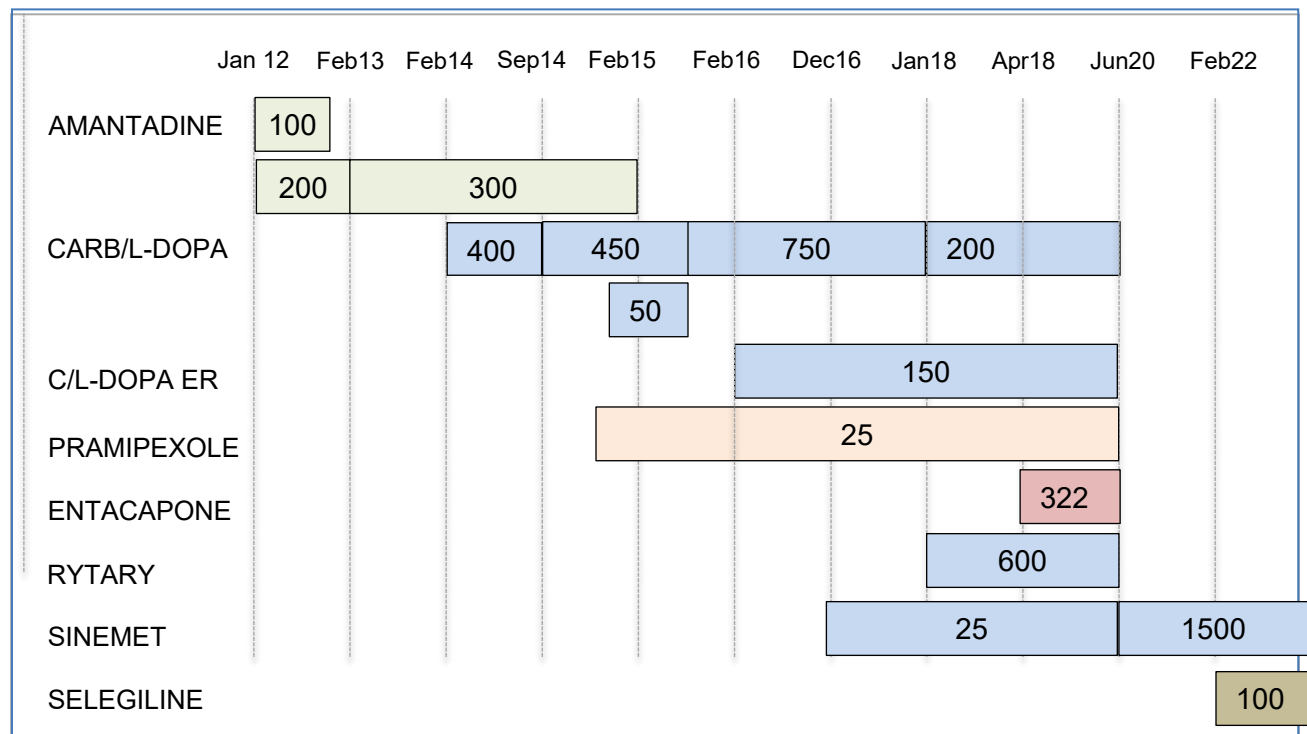
PATNO	EVENT_ID	PAG_NAME	LEDTRT	STARTDT	STOPDT	LEDD
9052	LOG	CMED	AMANTADINE	01/2012	01/2012	100
9052	LOG	CMED	AMANTADINE	01/2012	02/2013	200
9052	LOG	CMED	AMANTADINE	03/2013	02/2015	300
9052	LOG	CMED	CARBIDOPA/LEV ODOPA 25/100	02/2014	08/2014	400
9052	LOG	CMED	CARBIDOPA/LEV ODOPA 25/100	09/2014	09/2015	450
9052	LOG	CMED	PRAMIPEXOLE	11/2014	06/2020	25
9052	LOG	CMED	CARBIDOPA/LEV ODOPA 25/100	12/2014	08/2015	50
9052	LOG	CMED	CARBIDOPA/LEV ODOPA 25/100	09/2015	12/2017	750
9052	LOG	CMED	CARBIDOPA/LEV ODOPA ER 50/200	02/2016	06/2020	150
9052	V17	LEDDLOG	SINEMET	12/2016	06/2020	25
9052	LOG	CMED	CARBIDOPA/LEV ODOPA 25/100	01/2018	06/2020	200
9052	LOG	CMED	RYTARY	01/2018	06/2020	600
9052	LOG	CMED	ENTACAPONE	04/2018	06/2020	LD x 0.33
9052	V17	LEDDLOG	SINEMET	07/2020		1500
9052	V17	LEDDLOG	SELEGILINE	02/2022		100

Each row represents a drug, dosage, and period of time when it was in effect. Where there is no end date, this represents the current prescription. Note that several drugs may be taken in parallel and that the table does not automatically come in date order.

Note also that there are two different values for PAG\_NAME: data collected prior to 2020 is labeled **CMED** and data captured from 2020 onwards is labeled **LEDDLOG**. The difference is that the later records additionally capture information about the visit (EVENT\_ID).<sup>21</sup>

<sup>21</sup> In the old system EVENT\_ID = 'SC' or 'Log' and in the new system EVENT\_ID = 'ED'

A visual representation is useful for understanding what is stored in this table (note: dates *not* shown to scale):



The figures represent milligrams of levodopa equivalent and there is some complexity in the way these are determined:

1. Levodopa medication e.g., Rytary, Sinimet, Madopar, is measured according to the quantity of levodopa, and associated enzyme inhibitors such as carbidopa are excluded, so for instance the LEDD value for Madopar 100/25mg is 100mg with the 25mg of benserazide being excluded. Examples are shown in blue in the chart above.
2. Dopamine agonists are converted to a levodopa equivalent value, for instance 1mg of pramipexole salt converts to 100mg of levodopa equivalent.
3. MAO-B inhibitors converted to a levodopa equivalent value, for example 1 mg of rasagiline or selegiline convert to 100mg of levodopa equivalent.
4. COMT inhibitors additionally depend on the amount of levodopa calculated in item 1 above. For example, the levodopa equivalent of 100mg of entacapone is  $LD \times 0.33$  so if the total daily levodopa is 200mg then this contributes an additional 66 mg. An example is shown in red in the chart above.

Note that the start and end dates are only measured to the month so we make an assumption that the medication is active for the full month. In this example, Amantadine with a LEDD value of 100 is active from January 2012 to January 2012 and we interpret this as the whole month of January 2012.

This table can be a little tricky to work with if you simply want to know at a point in time, what is the total LEDD value across all medications. Script 7 in the appendices creates a new table, LEDD, that contains this information:

If we query this new table for participant #9052 used as an example previously, we now have the following, which may be easier to work with as it simply records each time the LEDD value changes. The last value shows the current medication, which could also go to zero.

PATNO	STARTDT	LDOPA	LEDD
9052	2012-01-01	0	300
9052	2012-02-01	0	200
9052	2013-03-01	0	300
9052	2014-02-01	400	700
9052	2014-09-01	450	750
9052	2014-11-01	450	775
9052	2014-12-01	500	825
9052	2015-04-01	500	525
9052	2015-09-01	750	775
9052	2016-02-01	900	925
9052	2016-12-01	925	950
9052	2018-01-01	975	1000
9052	2018-04-01	975	1321.75
9052	2020-07-01	1500	1500
9052	2022-02-01	1500	1600

Note that all prodromal and some Parkinson's participants (who have elected to delay the start of dopaminergic treatment) will not have LEDD records.

Further details of the LEDD conversion factors can be found in Tomlinson *et al.* (2010) - in particular, refer to Tables 1 and 2 in this paper. There is also documentation available from the PPMI download page under Study Docs/ALL Study Methods.

## 7.2 Other medication

As well as levodopa and equivalent medication, PPMI has captured data about other medications taken by the participant in the table **Concomitant\_Medication\_Log**. This has a similar structure to the LEDD table, using the same STARTDT and STOPDT fields to record the time period, but unlike the LEDD table each medication is recorded independently of other medications and there is no conversion to a common metric. One thing that makes this table more complicated is that for the dosage there are two fields: a numeric field, CMDOSE that records the dose, and a text field, CMDOSU, that specifies the unit of measurement, for example MG for milligrams.

## 8 BIOSPECIMENS AND PROTEOMIC DATA

### 8.1 Overview of biospecimen and proteomic data

PPMI contains data relating to many different biospecimens and proteomic analyses, including plasma, serum, blood, cerebrospinal fluid, DNA and RNA. This data is available under the Biospecimen folder on the main download site. There is general data such as blood test results and CSF test results available under the subfolder “Lab Collection Procedures” and there is data relating to specific research in the following subfolders:

- Biospecimen Analysis
- Metabolomic Analysis
- Proteomic Analysis

A list of research projects (both completed and in progress) can be found on the main PPMI site (<https://www.ppmi-info.org/access-data-specimens/ongoing-analysis/specimen-analysis>). This table contains information about the principal investigator, organization, current status, cohorts used and visit schedule. The methods for these projects can be found after registration in a series of PDF documents under the corresponding Methods subfolder.

For instance, in the link given above, we can look up project 181 which is “Adaptive Immune Markers for Predicting Cognitive Decline in PD”. We can then find details of the study methods in the document “181 Project Methods: Adaptive Immune Markers for Predicting Cognitive Decline” under the Biospecimen Methods subfolder, and the actual data in the file “Project 181 Adaptive Immune Markers for Predicting Cognitive Decline in PD” under the Biospecimen Sample Analysis subfolder. The data uses PATNO and EVENT\_ID as the key fields to join to the rest of the PPMI Clinical data in the usual way. Note that the results for many projects have been collated into the file “Current\_Biospecimen\_Analysis\_Results” in the Biospecimen analysis subfolder of the Biospecimen folder; filter this file on the Project ID that you are interested in.

There is also a table of neuropathology results in the subfolder of the same name that details the results of post-mortem analyses of PPMI participants. This gives information about the presence of Lewy Bodies, Braak Staging, hippocampal degeneration, substantia nigra depigmentation, the level of TDP43 and so on. This uses the field PATNO and records are labelled with an EVENT\_ID of AUT.

## 8.2 Biomarker dashboard

To make it easier to identify what biomarker data is available across the large number of PPMI projects, a “biomarker dashboard” has been created. This is an Excel workbook that details:

- Columns A-C: The individual biomarkers and their source (e.g., CSF, plasma, urine)
- Columns D-G: The number, description, and the lead institution and researcher, of the associated project
- Columns H and I: Dates and identifiers of study visits
- Column J: The total number of PPMI participants for whom biomarker results are available
- Columns K-P: The number of PPMI participants in the PD cohort for whom biomarker results are available, and these numbers further broken down by idiopathic PD, by genetic subgroups LRRK2, GBA, SNCA, Parkin, and the combination of both LRRK2 and GBA (see Section 3.2 for details of cohorts and Section 10.1 for details of genetic subgroups)
- Columns R-T: The number of PPMI participants in the PD cohort for whom biomarker results are available, broken down by SAA test result, i.e. NSD+, NSD- or SAA test results not available (see Section 8.5 for details on SAA; note that there are multiple SAA projects and the figures relate to the most recently completed project for the participant)
- Columns U-AF: The number of PPMI participants in the Prodromal cohort for whom biomarker results are available, broken down by Hyposmia and RBD cases, genetic subgroups, and SAA test results
- Column AG: The number of PPMI participants in the Healthy Control cohort for whom biomarker results are available
- Column AH: The number of PPMI participants in the SWEDD cohort for whom biomarker results are available

The numbers in parentheses in the column headings show the total population for the respective subgroup, e.g. the column heading “Total Prodromal (N=1717)” indicates that the total population of Prodromal participants is 1,717.

As an example, if you want to know how many participants had an Amyloid- $\beta$  42 test and were also prodromal carriers of the GBA variant associated with increased risk of developing PD, then filter the data tab of the workbook on Biomarker = “Abeta 1-42” and then look for the column ‘GBA Prodromal’. The results will look like the following (note: some columns omitted for clarity):

Biomarker	Fluid Matrix	Project ID	Project Title	Visit	Total Participants (N=3476)	Total Prodromal (N=1717)	GBA Prodromal (N=189)
ABeta 1-42	CSF	125	A-beta, tau, and p-tau measurements in longitudinal CSF samples from de novo PD,	BL	895	138	10

			Control, SWEDD, Prodromal and Genetic Cohorts				
ABeta 1-42	CSF	125	A-beta, tau, and p-tau measurements in longitudinal CSF samples from de novo PD, Control, SWEDD, Prodromal and Genetic Cohorts	V02	589	46	1
ABeta 1-42	CSF	125	A-beta, tau, and p-tau measurements in longitudinal CSF samples from de novo PD, Control, SWEDD, Prodromal and Genetic Cohorts	V04	633	64	1
ABeta 1-42	CSF	125	A-beta, tau, and p-tau measurements in longitudinal CSF samples from de novo PD, Control, SWEDD, Prodromal and Genetic Cohorts	V06	481	4	0
ABeta 1-42	CSF	125	A-beta, tau, and p-tau measurements in longitudinal CSF samples from de novo PD, Control, SWEDD, Prodromal and Genetic Cohorts	V08	337	0	0
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	BL	407	274	147
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V02	20	1	1
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V04	374	272	127
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V06	338	228	79
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V08	275	129	30
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V10	453	68	8
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V11	1	1	0
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V12	353	32	1
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V13	74	0	0
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V14	125	0	0
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V15	6	0	0
ABeta 1-42	CSF	159	CSF A-beta, tau, and p-tau analysis 2019	V16	10	0	0

In this example there were two projects (125 and 159) that performed this test. In project 125 there are 10 participants who fit the profile at baseline, out of 189 in total in the PPMI data who are GBA Prodromal, but the number drops for subsequent study visits. In project 159 there were 147 participants at baseline who fit the profile, again with diminishing numbers for subsequent study visits.

The biomarker dashboard can be downloaded from the 'START HERE – Quick Start' folder, or the [Guidance Resources](#) web page on the public site. It also contains a data dictionary and a list of PPMI projects that generated biomarker data, along with an indicator showing whether they are included in the biomarker dashboard.

In the future the PPMI team anticipates periodically releasing further revisions of the biomarker dashboard, to include additional PPMI projects, using the same format.

## 8.3 Querying biospecimen data

As a simple example of how to query biospecimen data, take the **Blood\_Chemistry\_\_Hematology** table, available under Lab Collection Procedures subfolder of the Biospecimen section on the main download site. This contains many blood test results, so to identify which results are captured for the majority of participants we can run the Script 8 in the appendices to obtain the following results.

LTSTCODE	LTSTNAME	Participant Count
RCT4	ALT (SGPT)	2194
RCT5	AST (SGOT)	2194
HMT12	Basophils	2194
HMT19	Basophils (%)	2194
RCT183	Calcium (EDTA)	2194
HMT11	Eosinophils	2194
HMT18	Eosinophils (%)	2194
HMT2	Hematocrit	2194
HMT40	Hemoglobin	2194
HMT9	Lymphocytes	2194
HMT16	Lymphocytes (%)	2194
HMT10	Monocytes	2194
HMT17	Monocytes (%)	2194
HMT8	Neutrophils	2194
HMT15	Neutrophils (%)	2194
HMT13	Platelets	2194
HMT3	RBC	2194
HMT71	RBC Morphology	2194
RCT17	Serum Bicarbonate	2194
RCT18	Serum Chloride	2194
RCT11	Serum Glucose	2194
RCT16	Serum Potassium	2194
RCT15	Serum Sodium	2194
RCT8	Serum Uric Acid	2194
RCT1	Total Bilirubin	2194
RCT12	Total Protein	2194

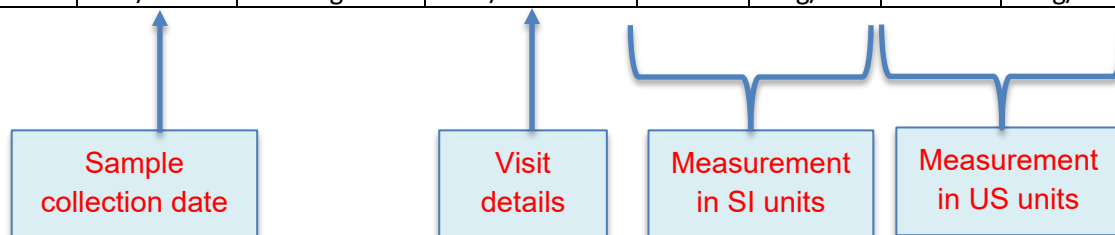
LTSTCODE	LTSTNAME	Participant Count
RCT6	Urea Nitrogen	2194
HMT7	WBC	2194
CGT283	Prothrombin Time	2186
RCT13	Albumin-QT	1631
RCT1407	Alkaline Phosphatase-QT	1631
RCT392	Creatinine (Rate Blanked)	1631
CGT626	APTT-QT	845
CGT284	APTT-QT	782
RCT3088	Albumin-BCG	576
RCT4447	Alkaline Phosphatase-QT	576
RCT408	Creatinine(Rate Blanked)-2dp	576
HMT4	MCV	576
CGT766	APTT-FSL	563
AMT7453	Which visit being performed?	82
BAT318	Serum beta hCG, Qualitative-QT	77
RCT3580	B-hCG, Qualitative	22
RCT3579	B-hCG, Quantitative	22
HMT95	Atypical Lymphocytes	9
HMT96	Atypical Lymphocytes	9
HMT20	Bands	6
HMT21	Bands (%)	6
AMT2917	Requisition Received ?	6
(Results truncated)		

This shows that there are 29 blood tests that have been captured for the majority of the participant population and a number of blood tests that have only partial coverage. It can be useful to understand data coverage in this way before embarking on detailed analysis.

We can then retrieve results for individual blood tests, for example Hemoglobin tests for participant #9002 using Script 9 in the appendices (note: example uses synthetic data for illustrative purposes):

EVENT_ID	LCOLLDT	LTSTNAME	LVISTYPE	LSIRES	LSIUNIT	LUSRES	LUSUNIT
SC	05/2011	Hemoglobin	Screening	145	g/L	14.5	g/dL
V04	06/2012	Hemoglobin	V04/Month 12	146	g/L	14.6	g/dL
V06	07/2013	Hemoglobin	V06/Month 24	150	g/L	15.0	g/dL

EVENT_ID	LCOLLDT	LTSTNAME	LVISTYPE	LSIRES	LSIUNIT	LUSRES	LUSUNIT
V08	09/2014	Hemoglobin	V08/Month 36	148	g/L	14.8	g/dL
V10	06/2015	Hemoglobin	V10/Month 48	153	g/L	15.3	g/dL
V12	11/2016	Hemoglobin	V12/Month 60	153	g/L	15.3	g/dL
V14	08/2018	Hemoglobin	V14/Month 84	142	g/L	14.2	g/dL



Several fields have been excluded from this query. In general, to look up the definition of a field, search on the field name in the Data Dictionary spreadsheet.

## 8.4 Requesting new biospecimen and proteomic data

If a particular biospecimen or proteomic analysis does not already exist in PPMI, a user can request biosamples and conduct a new analysis to derive a variable of interest see link here: <https://www.ppmi-info.org/access-data-specimens/request-specimens>. The resulting data is returned to the PPMI database for the benefit of all PPMI users; each approved biospecimen requested is given a project ID that is used to track the related data. Associated methods are documented in PDF files which can also be downloaded from the main download page.

## 8.5 Projects relating to the Amprion alpha-synuclein seeding amplification assay (αS-SAA)

**Note:** Most of the following information is also covered in a six-minute video that can be downloaded from the [PPMI Guidance Resources web page](#). The video describes what SAA is, how to access SAA test results in the PPMI data, the two different SAA projects and how to interpret the data.

The alpha-synuclein seed amplification assay, or SAA for short, is a recently developed diagnostic test for Parkinson's disease. Studies have shown that it has both high sensitivity and high specificity in identifying people with the condition and forms of Neuronal Synuclein Disease (NSD). It is also being investigated for other uses, such as for differential diagnosis of dementia with Lewy Bodies and Multiple System Atrophy (Siderowf *et al.* 2023). A substantial subset of the PPMI participants have undergone this test and the results can be found in the **SAA Biospecimen Analysis Results** file, which is available for download from the Biospecimen / Biospecimen Analysis folder.

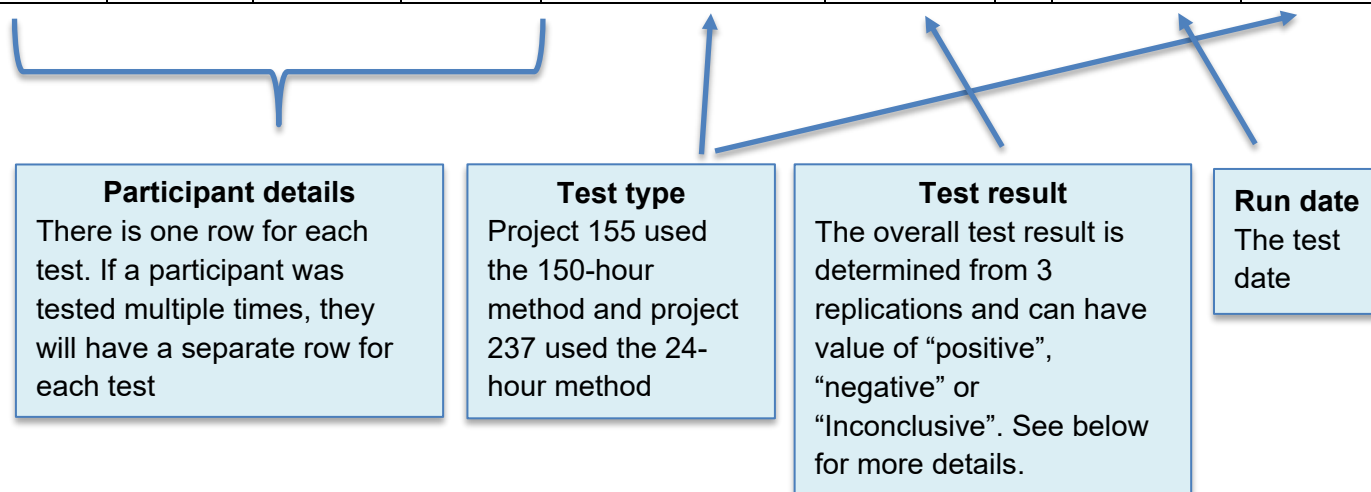


This file contains SAA results from two projects:

- Project 155 used the **Ampiron- $\alpha$ S-SAA** method where the assay process took 150 hours.
- Project 237 used the **Ampiron-24h- $\alpha$ S-SAA** method where the volume of recombinant  $\alpha$ -synuclein was reduced and the assay length shortened to 24 hours.

The file contains the following general columns (*note: some columns omitted for clarity*):

PATNO	SEX	COHORT	CLINICAL_EVENT	SAAMethod	SAA_Status	...	RUNDATE	PROJECT
90072	Male	PD	BL	Amprion-Alpha-synuclein-SAA	Positive		1/12/2023	155
90072	Male	PD	V03	Amprion-24h-alpha-synuclein-SAA	Positive		3/12/2023	237
90073	Female	PD	BL	Amprion-Alpha-synuclein-SAA	Negative		2/12/2023	155
90073	Female	PD	V02	Amprion-24h-alpha-synuclein-SAA	Negative		3/12/2023	237
90074	Male	PD	BL	Amprion-24h-alpha-synuclein-SAA	Positive		1/12/2023	237
90075	Female	PD	BL	Amprion-24h-alpha-synuclein-SAA	Positive		1/12/2023	237



Each SAA test comprises three replications. For each replication, there is a series of columns that capture the test details (i.e. each column will appear 3 times, suffixed with one of Rep1, Rep2 or Rep3). Note that many of the columns are specific to one of the two projects as detailed in the table below. Fluorescence is measured in relative fluorescence units (RFU) and there are thresholds for this value of 5,000 for project 155 and 3,000 for project 237 that determine the overall result of the test.



Kinetic Parameter	Column Name (suffixed with Rep1, Rep2 or Rep3)	Project 155	Project 237
Maximum Fluorescence in RFU after 24 hours	Fmax_24h		✓
Time to reach threshold in 24 hours	TTT_24h		✓
Area under the curve [RFU*hours]	AUC_24h		✓
Time to reach the maximum slope value in hours	Tsmax_24h		✓
The maximum slope value obtained for a given assay reaction in RFU/h	SLOPEMax_24h		✓
Maximum Fluorescence in RFU after 150 hours	Fmax_150h	✓	
Time to reach threshold in 150 hours	TTT_150h	✓	
Area under the curve [RFU*hours]	AUC_150h	✓	
Time to reach 50% of FMax in hours	T50_150h	✓	
Slope in RFU/h	SLOPERep_150h	✓	
Identifier of the instrument used in the test (2, 4 or 6)	Instrument	✓	✓
Sample volume used in the test (takes the value NEV1 = 20-30µL, NEV2 = 30-40µL or NEV3 = <20µL)	SampleVol	✓	✓

See the corresponding methods documents for the individual projects for more details. These can be found in the Biospecimen Analysis Methods folder.

The overall test result is determined according to a set of decision criteria. For project 155, these are:

SAA_Status	Decision Criteria
Positive	<b>All 3</b> replicates have Fmax $\geq$ <b>5,000</b> RFU
Negative	<b>0 or 1</b> replicate has Fmax $\geq$ <b>5,000</b> RFU
Inconclusive	<b>Exactly 2</b> replicates have Fmax $\geq$ <b>5,000</b> RFU

For project 237, the criteria are more complex:

SAA_Status	Decision Criteria
Positive	<b>All 3</b> replicates have Fmax $\geq$ <b>45,000</b> RFU
	<b>Exactly 2 or 3</b> replicates have Fmax $\geq$ <b>3,000</b> RFU & <b>&lt; 45,000</b> RFU
	<b>Exactly 2</b> replicates have Fmax $\geq$ <b>45,000</b> RFU & <b>1</b> replicate has Fmax $\geq$ <b>3,000</b> RFU & <b>&lt; 45,000</b> RFU
Negative	<b>0 or 1</b> replicate has Fmax $\geq$ <b>3,000</b> RFU
Inconclusive	<b>Exactly 2</b> replicates have Fmax $\geq$ <b>45,000</b> RFU & <b>1</b> replicate has Fmax <b>&lt; 3,000</b> RFU
	<b>Exactly 1</b> replicate has Fmax $\geq$ <b>45,000</b> RFU & <b>1</b> replicate has Fmax $\geq$ <b>3,000</b> RFU & <b>&lt; 45,000</b> RFU & <b>1</b> replicate has Fmax <b>&lt; 3,000</b> RFU

Where a participant has multiple test results, we recommend prioritizing the project 237 result with the most recent RUNDATE.

Project 237 also generates **Synucleinopathy Status** results. Based on the value of Fmax, a determination is made as to whether the case is “LBD-like” (Type 1), “MSA-like” (Type 2) or Undetermined. This is captured in the column SAA\_Type, according to the following criteria:

SAA_Type	Decision Criteria
Type 1	Positive samples where all 3 replicates have a Fmax $\geq$ 45,000 RFU
Type 2	Positive samples where 2 or 3 replicates have a Fmax $<$ 45,000 RFU
Undetermined	Positive sample has exactly 2 replicates with a Fmax $\geq$ 45,000 RFU
NA	Not applicable, i.e. the SAA test result was Negative

Note that **these tables only contain data for Parkinson's disease participants**; data for the Healthy Control and Prodromal cohorts has been sequestered and is accessible by submitting a written request to [ppmi@michaeljfox.org](mailto:ppmi@michaeljfox.org) that details who will be using the data and what analysis they will be performing in accordance with the [PPMI Data Access Guidelines](#).

## 8.6 Further details on other significant projects

In this section we give an overview of some key projects that have captured additional biospecimen, metabolomic and/or proteomic data.

**Project 151 (Identification of Proteins & Protein Networks in Human CSF That Differentiate Within PD Participants)** has conducted **cerebrospinal fluid tests** on Parkinson's, Prodromal and Healthy Control participants. The results are split across several files prefixed with “Project\_151”. The test reference numbers (column TESTNAME) relate to the SOMA\_SEQ\_ID generated by the SomaScan software package<sup>22</sup> and can be looked up in column B of the spreadsheet “PPMI\_Project\_151\_pqtl\_Analysis\_Annotations\_20210210”.

**Project 180 (Metabolomic Analysis of Penetrance, Prognosis, & Tracking Biomarkers of LRRK2 PD)** collected **plasma samples** from Parkinson's, Prodromal and Healthy Control participants and analyzed them using liquid chromatography with mass spectrometry (LC/MS) for a variety of metabolite and lipids including purines, lipids associated with lysosomal function like sphingolipids, bis(monoacyl)glycerophosphate (BMP), pro/antioxidant and pro/anti-inflammatory molecules, and environmental / dietary exposure markers. The results are split across 5 files prefixed with “Metabolomic\_Analysis\_of\_LRRK2\_PD”. The values in the column TESTNAME take the form MZxxx.xx\_RTyyy.yy\_mmm where MZxxx.xx indicates a mass to charge ratio of xxx.xx, RTyyy.yy is the retention time in seconds, and mmm indicates either positive or negative ionization mode. For further information, please refer to the method document, “PPMI\_Methods\_Project\_180\_Metabolomics\_LRRK2\_20220705”.

**Project 190 (Targeted & Untargeted MS Based Proteomics of Urine in PD)** analyzed **proteome profiles of urine samples** using mass spectrometry. The results are split across 5 files prefixed with “Targeted\_and\_untargeted\_MS-based\_proteomics\_of\_urine\_in\_PD”. The test

<sup>22</sup> <https://somalogic.com/>

identifiers (column TESTNAME) are a concatenation of the gene identifier and the protein identifier. These can both be looked up in a protein database such as UniProt (<https://www.uniprot.org/>) or the Human Protein Atlas (<https://www.proteinatlas.org/>)

**Projects 196 and 222 (Targeted Proteomics in AMP PD)** analyzed **cerebrospinal fluid and plasma** for a limited number of participants. The results are in a number of files prefixed with “PPMI\_Project\_196\_” and “PPMI\_Project\_222\_”. The columns in these tables are interpreted as related analyte values for each sample, and include protein name, Olink protein analyte target, UNIPROT ID, and the specific text values for each analyte included in the targeted assay panel. Note that results are broken into separate plasma and CSF files, and further divided into the four targeted assay panels used to make up the Olink Explore platform. For further information, see the related methods documents, as well as the Olink website for an overview of the assay and additional detail and guidance on interpreting these results. Projects 196 and 222 contain bridging samples allowing for the results to be combined and harmonized using an appropriate bridging protocol. For user convenience, a bridged copy of these two batches of Olink protein analysis are available as **Project 9000 (Bridged Results of Projects 196 and 222)** and is recommended for most users. This bridged file was created using the official Olink bridging protocol; users should carefully review the associated methods document for further details, and for interpretation notes.

Finally, the **iPSC Catalog Metadata** file (in the subfolder Biosample Inventory) contains metadata relating to multiple projects investigating induced pluripotent stem cells. The columns in this file are covered in the data dictionary but also explained in more detail in the Excel workbook “PPMI\_IPSC\_Catalog\_User\_Guide”.

## 8.7 Curated proteomic data

In the proteomic analysis subfolder under biospecimen on the download page, is a zip file called ‘PPMI\_Proteomic Working Group Curated Dataset’. This contains consolidated data for 4 projects, namely:

- Project 151: Identification of Proteins & Protein Networks in Human CSF That Differentiate Within PD Patients
- Project 177: Post-Translational Modification & Protein Quantification of Plasma & CSF
- Project 190: Targeted & Untargeted MS Based Proteomics of Urine in PD
- Project 9000: Bridged PPMI Project 196 and 222 Olink® Explore 1536 targeted proteomics assessment of CSF and blood plasma

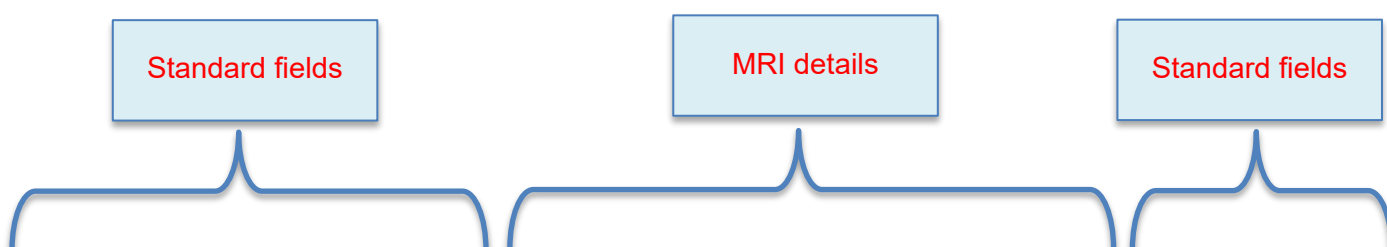
In all these projects, the original data is stored in a normalized format with results spread over many individual rows and, in several cases, multiple files. The curated versions combine several results in a single row which significantly reduces the volume and makes it much easier to analyze the data. See the associated method document for further information.

## 9 IMAGE DATA

PPMI contains a large number of images from various image modalities including Magnetic Resonance Imaging (MRI) and Dopamine Transporter scans (DaTscan). Metadata about these images is available for download in tabular format and the images themselves can be viewed and downloaded using a separate tool.

### 9.1 Image metadata

Image metadata is stored in a series of tables in a format similar to other measurement data. For example, for MRI there is the table `Magnetic_Resonance_Imaging__MRI_`:



REC_ID	PAT NO	EVEN T_ID	PAG_NAM E	INFOD T	MRI CMPL T	MRI WDT I	MRI WRS S	MRI RSLT	MRI RSSD F	PDM EDDT	PDM EDTM	ORIG_ENTRY	LAST_UPDAT E
12345681	9001	BL	MRI	Feb-11	1	0	NULL	2	NULL	NULL	NULL	Jan-12	2021-06-21
12345682	9001	V14	MRI	NULL	0	NULL	NULL	NUL L	NULL	NULL	NULL	Jan-19	2021-06-21
12345683	9003	BL	MRI	Mar-11	1	0	NULL	2	NULL	NULL	NULL	Jan-12	2021-06-21
12345684	9004	BL	MRI	Mar-11	1	0	NULL	2	NULL	NULL	NULL	Jan-12	2021-06-21
12345685	9005	BL	MRI	Apr-11	1	0	NULL	1	NULL	NULL	NULL	Jan-12	2021-06-21
12345686	9006	BL	MRI	Apr-11	1	0	NULL	2	NULL	NULL	NULL	Feb-12	2021-06-21
12345687	9007	BL	MRI	May-11	1	0	NULL	2	NULL	NULL	NULL	Feb-12	2021-06-21
12345688	9008	BL	MRI	May-11	1	0	NULL	2	NULL	NULL	NULL	Feb-12	2021-06-21
12345689	9009	BL	MRI	Jun-11	1	0	NULL	2	NULL	NULL	NULL	Feb-12	2021-06-21

The definitions of the MRI specific fields can be looked up in the Data Dictionary, and related permissible code values in the Code\_List:

Field name	Description	Permissible values
MRICMPLT	Brain MRI was completed	0 = Not Completed, 1 = Completed
MRICMPLTCM	Brain MRI not completed comment	(Text field)
MRIRSLT	Results	1 = Normal, 2 = Abnormal, not clinically significant, 3 = Abnormal, clinically significant
MRIRSSDF	Resting state seq-diff day PDmed use	0 = No, 1 = Yes
MRIWDTI	DTI MRI Scan	0 = No, 1 = Yes
MRIWRSS	With resting state sequences	0 = No, 1 = Yes
PDMEDDT	Last dopaminergic medication dose date	Date
PDMEDTM	Last dopaminergic medication dose time	Time

## 9.2 Viewing and downloading images

To view and download images, please refer to the walkthrough that provides a step-by-step guide on how to search, download, and view the desired neuroimaging data from PPMI in the following link:

<https://bit.ly/38A7tmt>

## 10 GENETIC DATA

Information about a participant's genetics can be found in the PPMI Clinical data in several places to different levels of detail.

### 10.1 Major genetic subgroups

Genetic screening is performed against known highly penetrant genetic variants conferring elevated risk of developing PD.<sup>23</sup> This information is captured as a series of columns on the **Participant\_Status** table.

This will tell you whether a participant was designated at the time of enrollment in the study as having a PINK1, LRRK2, Parkin, SNCA or GBA genetic variant (or some combination of more than one of these), as well as indicating whether the participant is a sporadic case, a prodromal RBD case or a prodromal HPSM case. Note that the RBD and HPSM cases might become sporadic cases after baseline if the participant subsequently develops Parkinson's, but the designation at the time of enrollment will not change.

Below are 4 example participants from this table: one sporadic case, one with a GBA variant, one with a LRRK2 variant and one with an SNCA variant.

PATNO	COHORT_DEFINITION	ENRLPINK1	ENRLPRKN	ENRLSRD C	ENRLHPS M	ENRLRB D	ENRLRRK 2	ENRLSNC A	ENRLGB A
9021	Parkinson's Disease	0	0	1	0	0	0	0	0
9022	Prodromal	0	0	0	0	0	0	0	1
9023	Parkinson's Disease	NULL	NULL	NULL	0	0	1	0	0
9024	Parkinson's Disease	0	0	0	0	0	0	1	0

If the value of one of these fields is NULL, then this indicates that the test was not performed, or the test result is not available.

### 10.2 Other tests for known PD genetic risk factors

The table **iu\_genetic\_consensus** (identified in the download site as 'Genetic Data - Consensus APOE Genotype and Coding Variants for GBA and LRRK2' ) contains information about what genetic analysis has been performed on each participant, for example whether they have been subject to genome wide associated studies (GWAS) or RNA sequencing. Additionally, for LRRK2 and GBA, the table shows the specific alleles of these two genes for which the participant has been tested. This information is only sporadically updated.

<sup>23</sup> This information is commonly captured during screening as part of inclusion/exclusion criteria and verified through genetic testing prior to baseline.

Below is the content of this table for the same 4 example participants that we used previously (note: synthetic data for illustrative purposes; the table has been split into two sections for ease of viewing).

PATNO	CLIA	GWAS	WES	WGS	SANGE R	RNASE Q	RNASE Q _VIS	APOE	GBA_ POS	GBA_ PATHV AR
9021	-	X	X	X	X	X	5	E3/E3	0	0
9022	X	-	-	-	-	-	0	E2/E3	1	1
9023	X	X	-	X	-	-	0	E2/E3	0	0
9024	X	-	-	X	-	X	4	E3/E3	0	0

PATNO	GBA_VAR_ID	GBA_VAR_ CONF	LRRK2_ POS	LRRK2_ PATHVAR	LRRK2_VAR_ ID	LRRK2_VAR_ CONF
9021	none	none	1	0	S1647T/S164 7T / M2397T/M2 397T	S1647T/S164 7T / M2397T/M2 397T
9022	N409S/N409 S	none	0	0	none	none
9023	none	none	1	1	S1647T/S164 7T / G2019S / M2397T/M2 397T	S1647T/S164 7T / G2019S / M2397T/M2 397T
9024	none	none	1	0	N2081D / M2397T	N2081D / M2397T

The table **Genetic\_Testing\_Results** contains the genetic test results for GBA, LRRK2 and SNCA testing: A value of 1 in MUTRSLT indicates a positive match and the field GENECA has a value of 1 for LRRK2, 2 for SNCA and 3 for GBA. The field LRRKCD gives the specific LRRK2 allele match, in this example a value of 1 being G2019S|c.6055G>A|p.Gly2019Ser.

PATNO	EVENT_ID	PAG_NAME	INFODT	GENECAT	LRRKCD	MUTRSLT
9031	GMU	MUTRSLT	Feb-16	3		1
9032	GMU	MUTRSLT	Mar-19	1	1	1
9033	GMU	MUTRSLT	Apr-18	2		1

Additionally, the table **PPMI\_PD\_Variants\_Genetic\_Status** (identified in the download site as “Participant Genetic Status for Selected PD-Associated Variants”) contains information about matches to specific alleles on other genes known to be associated with PD, though it is important to note that this has not been updated since 2018.



**Polygenic risk scores (PRS)** were generated by Project 9001 for approximately 40% of PPMI participants using whole genome sequencing data. There are 16 scores for each participant each one based on a different set of single nucleotide polymorphisms (SNPs). The risk of developing PD associated with each SNP is calculated and the scores summed to create an overall risk score. These scores can be found on the table 'PPM\_Project\_9001', identified in the download site as 'Genetic principal components and polygenic risk score for Parkinson's disease (Project 9001)'. Please refer to the associated method document for further information.

## 10.3 RNA-Sequencing data

RNA-Sequencing was performed in Project 133, which produced 3 processed datasets, namely IR1, IR2 and IR3. IR1 and IR2 are earlier versions and IR3 is the latest version. These are all derived from the same the same sequencing data (the experiment was not repeated), but the structure and pipeline changed slightly to keep them up to date with progress in the field. The biggest difference was the genome build used for alignment. "Build 37" (used for IR1/IR2) is slightly older, and more widely used, whereas "Build 38" (used for IR3) is newer. We recommend using IR3, unless there is a specific reason to utilize the older build and annotation. The raw files are also available upon request in BAM<sup>24</sup> and FASTQ<sup>25</sup> formats, but note that they are very large with a total size of ~142 TB. Please note, while these files are available via web portal download, due to the file size, you may encounter difficulties. We recommend submitting a request for Aspera-based cloud data transfer using the Genetic Data Request Form described below.

Some technical notes about the RNA-Sequencing data:

- The downloaded .tar.gz files contain transcript abundance estimates presented as both featureCounts (Liao *et al.* 2014) and Transcripts per Million (TPM) calculated using the Salmon method (Patro *et al.* 2017). The featureCounts are stored in files under a sub-folder "counts" and the TPM figures are stored in files with the suffix ".sf" in a sub-folder "quant". Note that these figures have been normalized and details of the normalization method can be found in Craig *et al.* (2021).
- In the file metaDataIR3.tsv can be found information about the phase it was sequenced in, e.g., "PPMI-Phase2-IR1.3174.V08.0000373111.5104-SL-0001". Phase 1 and Phase 2 represent sequencing waves, or freezes, and are a point where we started analyzing while waiting for additional plates. Covariates are located here. Diagnostic and other clinical data are available elsewhere on ppmi-info.org.

<sup>24</sup> See <https://github.com/samtools/hts-specs/blob/master/SAMv1.pdf> for the specification of SAM/BAM file formats.

<sup>25</sup> See Cock *et al.* 2010

## 10.4 Other genetic data

PPMI has several other processed genetic datasets that are available for download via the “Genetic Data” option on the main PPMI download page. Note that many of these files are very large and you will likely need specialist software to analyze them. We recommend reviewing the methods documentation to understand available data formats prior to downloading these files.

**Whole Genome Sequencing (WGS)** data has been generated by **project 118** for most participants (for historical reasons). Processed files are available in Variant Call Format (VCF) and Genomic Variant Call Format (.gVCF). The raw files are also available at request in BAM and FASTQ formats but note that they are very large with a total size of ~184TB. There is also a method document available on the download page.

**Whole Exome Sequencing (WES)** data has been generated by **project 116** for most participants. Processed files are in Genomic Variant Call Format (.gVCF). The raw files are also available at request in BAM and FASTQ formats but note that they are very large with a total size of ~17.5TB. There is also a method document available on the download page.

The **Foundational Data Initiative for Parkinson's Disease (FOUNDIN-PD; project 150)** generated a series of 'omics datasets for 95 inducible pluripotent stem cell lines from PPMI participants as part of a study to identify progression markers in PD. For an overview of the study and further information, refer to the methods documents provided and see <https://www.foundinpd.org>. Processed files are in a variety of file formats corresponding to the distinct data types generated from these resources. The raw files are also available at request in FASTQ, .idat, .jpg, and .raw formats but note that they are very large with a total size of ~27TB.

**Gene Sequencing** refers to primer-targeted Sanger sequencing data that was generated by **project 115** for all PPMI participants, to specifically analyze genetic variability in the SNCA gene. The files are in VCF format; for further information, see the associated method document.

**Genotyping** refers to NeuroX array-based genotyping data generated by **project 107** which performed a series of genotyping studies on several hundred of the PPMI participants against a wide range of polymorphisms covering various autoimmune, inflammatory and neurological diseases as well as PD loci. The data is stored in the PLINK format (Purcell *et al.*, 2007) and further information can be found in the methods documents.

**Methylation Profiling** contains data relating to **projects 120 and 140**, which performed whole genome array-based methylation profiling on most participants. The files are in the Illumina .idat format and further information can be found in the methods documents.

If you require a genetic dataset in its raw format, then fill out and submit the related request form, both found under the “Download Genetic Data” section of PPMI@LONI. FASTQ/BAM users should submit the “Genetic Data Request Form”; Project 193 iPSC WGS users should submit the “PPMI Cell Lines Data Request Form.”

# Download Genetic Data

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



**Reminder: The PPMI 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 these data you acknowledge our [Terms and Conditions](#).**

- ▼ Data Request Form
  - Request Form**
- ▶ Exome Sequencing
- ▶ FOUNDIN-PD
- ▶ Gene Sequencing
- ▶ Genotyping
- ▶ Methylation Profiling
- ▶ RNA Sequencing
- ▶ Whole Genome Sequencing
- ALL

Data Request Form:  
*Form for requesting access to download data in FASTQ/BAM format*

Name of Dataset	Version	File Type	Last Download
Request Form			
<a href="#">Genetic Data Request Form</a>	2022-08-02	.DOC format	🕒 2022-10-27

## 11 CURATED DATA

Recognizing that to perform even a simple study using the PPMI Clinical data it is necessary to join several tables together, the PPMI data team has prepared a curated dataset where frequently used data from multiple tables has been merged (i.e., denormalized) into a single large table for ease of use. The curated data is in an Excel workbook called PPMI\_Curated\_Data\_Cut\_Public\_YYYYMMDD, where YYYYMMDD is the date of the extract. At the time of writing there is one extract taken on January 29<sup>th</sup>, 2024: this contains records for 3096 participants, comprising 973 participants with sporadic PD, 763 participants with PD and major genetic factors, 1018 hyposmia/RBD prodromal cases, 279 healthy controls and 63 participants from the SWEDD cohort (see Section 3.2).

In the future the PPMI data team anticipates periodically releasing further curated data cuts using the same format.

The Excel workbook contains 3 tabs:

- The first tab contains some useful information including an explanation of the columns in the data dictionary
- The second tab (labeled with the date of the extract) contains the curated dataset
- The third tab contains a data dictionary for the 160 data columns in the curated dataset

The following table summarizes the data included in the curated dataset. For more detail on individual columns, please refer to the data dictionary.

Type of information	Summary of data provided	Comments
Participant identification and study characteristics	<ul style="list-style-type: none"> <li>• Site identifier</li> <li>• Patient identifier (PATNO)</li> <li>• Cohort</li> <li>• Subgroup</li> <li>• Status</li> <li>• Enrollment phase</li> </ul>	<p>See Sections 3.1 and 3.2</p> <p>Enrollment phase is pre or post June 2020</p> <p>Subgroup is derived from various source columns to give a more detailed group assignment than cohort. It can take values of Healthy Control, SWEDD, SWEDD/PD, SWEDD/non-PD, Hyposmia, RBD, Sporadic PD, LRRK2, GBA, PINK1, PRKN, SNCA or combinations of genetic variants and/or</p>

Type of information	Summary of data provided	Comments
		RBD (e.g. LRRK2 + GBA, GBA + RBD).
Diagnosis	<ul style="list-style-type: none"> <li>• Primary diagnosis</li> <li>• Neuronal alpha-synuclein disease (NSD) and associated NSD_ISS staging indicators</li> <li>• Up to 3 previous diagnoses</li> </ul>	A patient may be reclassified into a different cohort/group once or more ("phenoconverted"). See Section 3.5 for further discussion of NSD.
Study visit details	<ul style="list-style-type: none"> <li>• Visit identifier (EVENT_ID)</li> <li>• Visit date</li> <li>• Age at enrollment</li> <li>• Age at visit date</li> <li>• Years since study enrollment</li> </ul>	See Section 5
Participant demographics	<ul style="list-style-type: none"> <li>• Sex</li> <li>• Years of education</li> <li>• Ethnicity</li> <li>• Family history of PD</li> <li>• Handedness</li> <li>• Gender identity</li> <li>• Sexual orientation</li> <li>• Body mass index</li> <li>• Age at diagnosis</li> <li>• Age at onset of symptoms</li> <li>• Time between diagnosis and enrollment</li> </ul>	See Section 4
Participant symptoms	<ul style="list-style-type: none"> <li>• Dominant side with symptoms</li> <li>• Presence of common symptoms (tremor, rigidity, bradykinesia, etc.)</li> </ul>	
Medication	<ul style="list-style-type: none"> <li>• Type of treatment (dopaminergic, DBS)</li> <li>• LEDD</li> </ul>	See Section 7.1 for LEDD discussion

Type of information	Summary of data provided	Comments
Olfactory test results	<ul style="list-style-type: none"> <li>University of Pennsylvania Smell Identification Test (UPSIT)</li> </ul>	See Section 6.2
Cognitive test results	<ul style="list-style-type: none"> <li>Montreal Cognitive Assessment (MOCA)</li> <li>Benton judgement of line orientation</li> <li>Clock drawing test</li> <li>Lexical fluency</li> <li>Modified Boston Naming Test</li> <li>Hopkins Verbal Learning Test (HVLT)</li> <li>Letter Number Sequencing Test</li> <li>Symbol Digit Modalities Test</li> <li>Trails Marking Test (TMT)</li> <li>Mild Cognitive Impairment (MCI)</li> <li>Investigator cognitive state diagnosis</li> <li>Modified Schwab &amp; England ADL score</li> </ul>	
Behavioral test results	<ul style="list-style-type: none"> <li>Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease (QUIP)</li> <li>Geriatric Depression Scale (GDS)</li> <li>State-Trait Anxiety Inventory (STAI)</li> </ul>	
Sleep and tests	<ul style="list-style-type: none"> <li>Epworth Sleepiness test</li> <li>REM Sleep Behavior Disorder Questionnaire</li> </ul>	
Other autonomic tests	<ul style="list-style-type: none"> <li>SCOPA-AUT test</li> <li>Orthostasis indicator</li> </ul>	Note only 7 of 25 data items included for SCOPA
Motor test results	<ul style="list-style-type: none"> <li>MDS-UPDRS part 1 individual scores</li> </ul>	See Section 6.1

Type of information	Summary of data provided	Comments
	<ul style="list-style-type: none"> <li>• MDS-UPDRS total scores for parts 1, 2, 3 and 4, including a separate “on” score for part 3</li> <li>• MDS-UPDRS total “on” and “off” scores</li> <li>• Hoehn &amp; Yahr stage (original scheme) for both “on” and “off”</li> <li>• Tremor Dominant (TD) / Postural Instability and Gait Difficulty (PIGD) scores</li> </ul>	
Progression milestones	<ul style="list-style-type: none"> <li>• Progression milestones associated with the visit, related to activities of daily living</li> </ul>	See Brumm, <i>et al.</i> (2023)
CSF test results	<ul style="list-style-type: none"> <li>• Amyloid-Beta</li> <li>• p-tau and t-tau</li> <li>• Alpha-Synuclein Seeding Amplification Assay (SAA)</li> <li>• Neurofilament light (serum and CSF)</li> <li>• High hemoglobin</li> </ul>	For SAA, see Section 8.5
Blood and urine test results	<ul style="list-style-type: none"> <li>• Uric acid</li> <li>• Urine BMP Total di-18:1 Species</li> <li>• Urine BMP Total di-22:6 Species</li> </ul>	Relates to Project 145
DaTscan results	<ul style="list-style-type: none"> <li>• DaTscan measurements of striatal binding ratio (SBR) of left, right, ipsilateral and mean caudate, putamen and striatal regions</li> </ul>	
NSD supplementary data fields	<ul style="list-style-type: none"> <li>• S, D and G fields used to derive NSD and NSD-ISS staging</li> </ul>	See Section 3.5

Some usage notes for the data dictionary are:

- The data dictionary lists all the variables from the data file, though not necessarily in the same order. For each variable (column B) it provides a category (column A - for instance

ID for identifiers, Clinical for clinical data, Genetics for genetic test data, and so on) and a description (column C). The variables are grouped by categories.

- Where the variable relates to a coded value such as COHORT, EVENT\_ID, gender or race, the data dictionary provides the full list of permissible values and the corresponding decode values (columns D, E).
- Columns F, G, H and I of the data dictionary give the source of the data. If Derived Variable (column F) is set to “No”, then columns G and H will detail the source column name and table name respectively from where the value was sourced. For example, the variable COHORT is taken from the STATUS table, column COHORT. To find the physical table name for download, recall from Section 3.4 that this can be looked up in the main data dictionary; in this example the source table name is PATIENT\_STATUS.
- If the Derived Variable value is set to “Yes”, then the source variable(s) and table(s) are listed in columns G and H respectively and column I is populated with the details. For example, the variables relating to the MOCA (cognitive assessment) test have been aggregated into a single overall test result. About a third of the variables are derived, often to cut down on the overall number of columns.



## 12 PPMI ONLINE

PPMI Online is data from participant self-assessments that has been captured online. There is some participant overlap with both PPMI Clinical and PPMI Remote so that some participants may have data for all three, some for two of the three, and some for PPMI Online only. Data for PPMI Online is available from the same data download site as PPMI Clinical under the section “PPMI Online”.

The PATNO identifier is shared with PPMI Clinical and PPMI Remote data, but for longitudinal data, a different set of EVENT\_IDs is used, with the PPMI Online identifiers all having the form “OLxx” where xx is a two digit number.

Taking the example of MDS-UPDRS scores, in order to combine the data across PPMI Clinical and PPMI Remote, we need to take into account the different column names for the same fields, noting also that INFODT maps to CREATED\_AT. Script 10 in the appendices performs this mapping and results in a combined view like the following (note the “OL” EVENT\_IDs);

INFODT	EVENT_ID	NP1SLPN	NP1SLPD	NP1PAIN	NP1URIN	NP1CNST	....
03/01/2011	SC	3	1	1	3	1	
04/01/2011	BL	3	1	1	3	0	
06/01/2011	V01	2	1	0	3	1	
08/01/2011	V02	3	1	3	3	0	
02/01/2012	V03	2	1	1	1	0	
04/01/2012	V04	3	1	1	3	0	
09/01/2012	V05	2	2	0	3	0	
04/01/2013	V06	2	1	1	1	0	
10/01/2013	V07	2	1	1	2	1	
04/01/2014	V08	2	1	1	3	1	
10/01/2014	V09	2	1	1	2	1	
04/01/2015	V10	3	1	1	2	0	
10/01/2015	V11	3	1	1	2	0	
04/01/2016	V12	2	2	1	3	0	
04/01/2017	V13	4	2	1	2	1	
04/01/2019	V15	3	2	1	2	2	
04/01/2022	OL01	1	2	1	3	0	
10/01/2022	OL02	2	2	1	3	1	

For a summary of the all the other data available for PPMI Online, please refer to Appendix A.

## 13 PPMI REMOTE

The objective of PPMI Remote is to gather data about the pre-diagnosis phase of PD by conducting remote tests such as olfactory tests and genetic analysis to capture possible risk indicators of later developing PD. People are referred to PPMI Remote via both PPMI Online and an initiative called “Smell Test Direct”, because they have a known risk factor for developing PD.

Data for PPMI Remote is available from the same data download site as PPMI Clinical under the section “PPMI Remote”. The tables available for download are prefixed with “Remote\_”:

1. Remote\_Screening\_Smell\_Test\_Direct captures information about participants referred for screening from Smell Test Direct
2. Remote\_Screening\_Screener captures participant demographic data such as ethnicity
3. Remote\_Screening\_High\_Interest captures whether the participant has key prodromal markers of PD such as anosmia or acting out their dreams
4. Remote\_Screening\_Participant\_Progress records dates of key events of participants in PPMI Remote, including enrollment, UPSIT testing and DaTscans
5. Remote\_Screening\_UPSIT\_Screening records results of an abridged UPSIT questionnaire that the participants are given
6. Remote\_University\_of\_Pennsylvania\_Smell\_Identification\_TEST (UPSIT) contains the actual UPSIT results for participants who are not also enrolled in PPMI Clinical\*

\*Note that a *limited* set of UPSIT test results (from participants who have completed a DaTSCAN screening visit) are stored in the same table as used by PPMI Clinical (University\_of\_Pennsylvania\_Smell\_Identification\_Test\_\_UPSIT) rather than the PPMI Remote table.; filter on UPSIT\_SOURCE = Remote. See also Section 6.2.

## 14 PPMI FOUND

Data for PPMI FOUND is available from the same data download site as PPMI Clinical under the section “PPMI FOUND”. FOUND stands for “Follow up of persons with Neurologic Disease” and records additional information for some PPMI Clinical participants captured by telephone or other remote consultation methods, in parallel with the main study.

The FOUND tables contain a variety of data relating to known or hypothesized PD risk factors, from alcohol consumption to pesticides at work. Further details can be found in Appendix A. There is also a data dictionary in the file FOUND\_RFQ\_Dictionary available for download.

The FOUND tables are referenced using PATNO in the usual way, but note that they do not have EVENT\_IDs. The date of data capture is recorded in a field called “datacompXX” (standing for data completion) where “XX” is different for each table. For example, in the table FOUND\_RFQ\_Head\_Injury, there is a patno column (note the lower-case spelling) and a column called datacomphi that records the date when the questionnaire was completed.

The data is captured one time only as it relates to historical information. The exception to this is the table FOUND\_Self-Reported\_Dx, which records diagnoses of PD or other conditions and is updated every 6 months.

## 15 DIGITAL SENSOR DATA

PPMI contains two types of digital sensor data: (1) data from the Roche PD monitoring app and (2) data from the Verily study watch. The Roche data relates to a number of active motor and non-motor tests performed using a mobile app, and is stored in a single table. The Verily data has been captured by smartwatch and comprises biometric measurements in a number of separate tables and is much larger in volume than the Roche data.

**Note:** The Verily data is only available by request and cannot be downloaded from the main download site.

Key points about the two datasets are summarized below.

	Roche PD monitoring app	Verily study watch
Nature of data	Active tests of motor and non-motor function	Biometric data relating to sleep, activity and cardio function
Method of capture	Mobile app (Android)	Smartwatch
Data tables and nature of data captures	<p><b>Roche_PD_Monitoring_App_v2_data:</b> consolidated table containing a range of tests covering:</p> <ul style="list-style-type: none"> <li>• <b>voice / speech</b></li> <li>• <b>dexterity</b></li> <li>• <b>bradykinesia</b></li> <li>• <b>tremor</b></li> <li>• <b>balance / postural stability</b></li> </ul> <p>During the study period, information relating to various questionnaires was also captured in the same table, including medication details and functional <b>patient reported outcome measures</b> such as EQ-5D-5L.</p>	<p><b>ambulatory:</b> minutes in the day when the participant is walking, running etc.</p> <p><b>inbedtimes:</b> time the participant went to bed and got out of bed</p> <p><b>onwrist:</b> details of when the Verily study watch was attached to the participant's wrist and gathering data</p> <p><b>prv:</b> pulse rate variability per hour</p> <p><b>pulserate:</b> mean hourly pulse rate</p> <p><b>sleepmetrics2:</b> metrics relating to timing, amount and quality of sleep</p> <p><b>sleepstage:</b> timing and duration of individual sleep stages (REM, NREM deep, NREM light, awake and unknown)</p> <p><b>stepcount:</b> hourly step count</p> <p><b>timezone:</b> timezone that each measurement was taken in</p>
Number of participants (as at July 31 <sup>st</sup> , 2023)	32 (20 Parkinson's and 12 Prodromal)	353 (148 Parkinson's, 158 Prodromal, 35 Healthy Control, 2 SWEDD, 10 Early Imaging study)

	Roche PD monitoring app	Verily study watch
– see Section 3.2 for definitions		
Approximate date range of data capture (as of July 31 <sup>st</sup> 2023)	May 2019 – December 2020	September 2017 – January 2021

## 15.1 Roche PD monitoring app: additional information

The tests/questionnaires in the Roche data are identified by the field QRSCATID. A full list of test codes and response options can be found in the Codes table (see Section 4) but note that at as the publication date of this guide, only some tests have data available, and these are:

Type	QRSCATID	Test description
Tests	DAS	Draw a shape test (note: the field QRSSCAT gives sub-test)
	FST	Speech test
	HT	Maximum hand turn speed
	PTT	Postural tremor test
	Q135	eSDMT (symbol digit modalities test)
	RTT	Resting tremor test
	SBT	Balance test
	SPT	Voice test
	ST12	Dexterity test
	UTT	U-Turn average turn speed
Questionnaires	DMT	Daily medication tracker (was medication taken and when?)
	HADS	HADS (anxiety and depression rating)
	MAT	Missed Active Test (reason for missing a test)
	PD-DDS	Daily diary (e.g., sleep quality, bowel movements)

Type	QRSCATID	Test description
	PD-DQS	Daily patient reported outcomes
	PD-HS	Health survey – ANS (autonomic function)
	PGIC	PGIC (patient global impression of change)
	Q008	EQ-5D-5L (self-care, activities, anxiety etc.)
	Q070	PDQ39 (standard 39-item PD questionnaire)
	Q226	PAC-SYM (constipation assessment)
	SETUP	Setup survey (handedness and home WiFi availability)

The corresponding answers to the questions and measurements can be found in the fields QRSRESN (for numeric values) or QRSRESC (for text values). Where applicable, the field QRSORESUS gives the units of the measurement. QRSSTDTC and QRSSTDTC\_TIME give the start date and time of the test respectively; QRSENDTC and QRSENDTC\_TIME give the end date and time of the test.

For further information on the methodology behind each of the tests, please refer to the Methods documents provided. Assessment studies of the various tests are described in Lipsmeier *et al.* (2022).

## 15.2 Verily study watch derived data: additional information

The Verily data is split into a series of files with the timestamps of each measurement stored in the field 'time\_utc\_ms' which is the number of milliseconds since the start of Jan 1, 1970. Note that some of the measurements including ambulatory, onwrist, pulserate and stepcount have been split into individual files for each participant due to the large numbers of values.

Raw device data is available upon request. To submit a request, navigate to the 'Study Docs -> Study Data Request Forms' folder and complete the 'Verily Raw Device Data Request Form'.

## 16 APPENDIX A – SUMMARY OF KEY PPMI STUDY DATA

This appendix gives a high-level summary of the main PPMI study data and supporting documents available for download that researchers may be interested in. Not every dataset of document is listed; to see the full list, visit the download page itself

(<https://ida.loni.usc.edu/pages/access/studyData.jsp?project=PPMI>).

In addition to the datasets identified below, there are separate download options for image data (see Section 9) and certain genetic data (see Section 10).

Group	Subgroup	Filename	Comments
START HERE	Quick Start	Overview/Quick Start Guide	Deprecated by this user guide
		PPMI Data User Guide	This document
		PPMI Biomarker Dashboard	See Section 8.2
Study Docs	Data & Databases	Code List	List of all codes used in tables. "Annotated" version contains additional comments. See Section 4.1.
		Data Dictionary	List of all tables and columns. "Annotated" version contains additional comments. See Section 3.7.
		Deprecated Variables	A list of data elements that are no longer available for download
	Study Data Request Forms	Various forms for requesting additional data along with the PPMI Data Access Guidelines	See also Section 8.4
	Study Methods	Method for calculating Levodopa Equivalent Dose in PPMI Data	See Section 7.1
		Additional methods documents may also appear in this folder	
	Study Protocol & CRFs	Study CRFs and Assessments	Details of all standard questionnaires used by PPMI, including PASE (Physical Activity Scale for the

Group	Subgroup	Filename	Comments
			Elderly); screening & demographics; socio-economics; criteria for classification as PD / Prodromal / Healthy Control; AV-133 eligibility, diagnostic features, medical history, family history, neurological exam, physical exam, vital signs, medication, Schwab & England daily living, etc.
		Additional documentation for sub studies may also appear in this folder	
Subject characteristics	Family History	Family History	Details of PD family history
	Genetic Status	Consensus APOE	Various genetic analyses. See Section 10.2 for details.
		Polygenic risk score	
		Participant Genetic Status	
	Patient Status	Participant Status	Core table of participant status and classification. See Section 3.
	Subject Demographics	Age at visit	Participant age at each study visit
		Demographics	Basic demographic data such as sex, ethnicity, and date of birth
		Socio-Economics	Years of education
		Subject Cohort History	Mapping of participants to cohorts
Biospecimen	Biosample Inventory	Various	Various catalogs of available biospecimens. For further information on biospecimen data see Section .
	Biospecimen Sample Analysis	Biospecimen Analysis Results Overview	Overview of different biospecimen result sets
		Current Biospecimen Analysis Results	Details of current biosample tests performed including test name, test result, date and researcher
		Deprecated Biospecimen Analysis Results	Biospecimen analysis results that have been superseded.



Group	Subgroup	Filename	Comments
		Pilot Biospecimen Analysis Results	Results of an earlier pilot project relating to SAA. See Section 8.5.
		Project 181 Adaptive Immune Markers for Predicting Cognitive Decline in PD	Results of project 181
		SAA Biospecimen Analysis Results	Results of the SAA project: see Section 8.5.
	Biospecimen Analysis Methods	Various	Documents detailing the sample collection and analysis methods used for each biosample project
	Lab Collection Procedures	Various	Details of lab procedures used for blood chemistry, lumbar punctures, skin biopsies etc.
	Metabolomic Analysis	Various	Metabolomic tests and results
	Neuropathology results	Neuropathology results	Results of post-mortem neuropathology examination
	Proteomic Analysis	Various	Proteomic tests and results. See also Section 8.7.
	Proteomic Analysis Methods	Various	Documents detailing the procedures used to perform proteomic analysis
Curated Data Cuts	Curated Data	Curated data cuts	Prepared datasets containing selected details for a large number of participants merged into a single table for ease of use. A data dictionary is also provided with each dataset. See Section 11 for more details.
Digital Sensor	Roche Smartphone App	Roche PD Monitoring App v2 data and methods	Measurements for selected participants from the Roche Smartphone app and associated method document. See Section 15.
Enrollment	Subject Enrollment	Various	Files associated with participant study enrollment, capturing

Group	Subgroup	Filename	Comments
			participant consent, recording participation in sub-studies, etc.
Imaging	DaTSCAN Imaging	DaTscan Analysis	DaTscan measurements of striatal binding ratio (SBR) of left and right caudate, putamen and anterior putamen regions with respect to the occipital lobe
		DaTscan Imaging	Metadata relating to DaTscans, e.g., data of scan, location
		DaTscan Metadata	Further metadata relating to DaTscans, e.g., time of scan, dosage of radiotracer
		DaTscan SPECT Image Processing SBR Calculation Methods	Method document for the SBR calculation
		DaTscan SPECT Visual Interpretation Assessment Methods	Method document for the visual interpretation
		DaTscan Visual Interpretation Results	Summary result of DaTscan visual interpretation, i.e., positive / negative
	Diffusion Imaging	DTI Preprocessing Manual	Documentation and results relating to a diffusion tensor imaging (DTI) study on 15 participants (Schuff <i>et al.</i> , 2015)
		DTI Regions of Interest and associated Method	Measurement data from 8 Regions of Interest (see Method document)
	Magnetic Resonance Imaging	Grey Matter volume extraction from MRI T1	Grey matter volume estimated from MRI scan
		Magnetic Resonance Imaging (MRI)	Metadata relating to MRI scans, e.g., date, type of scan, comments
		MRI Metadata	Metadata relating to MRI results, e.g., were DTI, NM-MT, T1-weighted, T2-weighted FLAIR sequences acquired?
		Rs-fMRI Data Availability and Usage	Documentation and results relating to a sub-study of resting-state fMRI on

Group	Subgroup	Filename	Comments
			194 participants between 2012 and 2019
	PET Imaging	Various	Various technical documents, metadata and results relating to the tau sub-study, <sup>26</sup> and the early imaging sub-study that generated AV-133 <sup>27</sup> PET images
Medical History	Medical	Tau sub-study (various tables)	Data relating to the participant participating in the tau sub-study
		Adverse Event Telephone Assessment	Indicates if an adverse event occurred shortly after a lumbar puncture, skin biopsy, or DaTscan imaging procedure, and what contact was made with the participant about it
		Concomitant Medication Log	History of medication other than dopaminergic medication
		Determination of Freezing and Falls	Details of freezing and falls captured in later study visits
		Early Imaging (various tables)	Data relating to the Early Imaging cohort (see Section 3)
		Features of Parkinsonism	Key clinical features captured in later study visits (rigidity, bradykinesia, tremor, postural instability)
		General Physical Exam	Results of general physical examination carried out in screening
		LEDD Concomitant Medication Log	History of dopaminergic medication; see Section 7.1
		Medical Conditions Log	Year of diagnosis of other medical conditions (arthritis, cancer, high cholesterol, depression etc.)
		Other Clinical Features	Other clinical features (cf. Features of Parkinsonism table) captured in

<sup>26</sup> For details of the tau sub-study see [https://www.ppmi-info.org/sites/default/files/docs/PPMI%20Tau%20PET%20Imaging%20Protocol\\_26Apr2021\\_Final.pdf](https://www.ppmi-info.org/sites/default/files/docs/PPMI%20Tau%20PET%20Imaging%20Protocol_26Apr2021_Final.pdf)

<sup>27</sup> For details of the early imaging sub-study which generated the AV-133 images see [https://www.ppmi-info.org/sites/default/files/docs/PPMI%202.0\\_Early%20Imaging%20Protocol\\_v2\\_19June2020\\_Final.pdf](https://www.ppmi-info.org/sites/default/files/docs/PPMI%202.0_Early%20Imaging%20Protocol_v2_19June2020_Final.pdf)

Group	Subgroup	Filename	Comments
			later study visits (tremor, dystonia, micrographia, shuffling gait, etc.)
		PD Diagnosis History	Date of PD diagnosis and details of key motor symptoms
		Pregnancy Test	Pregnancy tests and results
		Primary Clinical Diagnosis	Details of diagnosis (idiopathic PD, MSA, PSP, MND with Parkinsonism, etc.)
		Prodromal History	Details of prodromal symptoms (REM behavior disorder, first degree family member with PD, hyposmia, etc.)
		Report of Pregnancy	Details of pregnancy during study period
		Surgery for PD Log	Details of PD related surgeries (e.g., Deep Brain Stimulation)
		Vital signs	Details of vital signs (blood pressure, temperature, weight, height, etc.)
	Neurological Exam	Neurological Exam	Summary of neurological tests performed at study visits (motor exam, gait assessment, reflexes, etc.)
		Prodromal Diagnostic Code Memo	Guidance for coding on the Prodromal History table
	Safety Monitoring	Adverse Event Log	Register of adverse symptoms during study (headache, nausea, pain around lumbar puncture site, back pain, etc.)
Motor Assessments	Motor/MDS-UPDRS	Gait Substudy	Gait and mobility measurements
		Gait Data & Arm swing	Gait and arm swing measurements
		Gait Data & Arm swing Methods	Description of the method for gait and arm swing measurements
		MDS-UPDRS Part I	Part 1 of the MDS-UPDRS, capturing non-motor aspects of daily living



Group	Subgroup	Filename	Comments
		MDS-UPDRS Part I Patient Questionnaire	Part 1 of the MDS-UPDRS, capturing non-motor aspects of daily living as documented by the participant
		MDS-UPDRS Part II Patient Questionnaire	Part 2 of MDS-UPDRS, capturing motor aspects of daily living as documented by the participant
		MDS-UPDRS Part III	Part 3 of the MDS-UPDRS capturing results of the motor examination
		MDS-UPDRS Part III ON/OFF Determination	Part 3 of the MDS-UPDRS, supplemental data about how the examination was performed
		MDS-UPDRS Part IV Motor Complications	Part 4 of the MDS-UPDRS capturing information about dyskinesia, on/off fluctuations etc.
		Modified Schwab & England Activities of Daily Living	Schwab & England total score
		Neuro QoL: Lower Extremity Function (Mobility)	Quality of life questionnaire (mobility) supplementary to MDS-UPDRS
		Neuro QoL: Lower Extremity Function (Fine Motor, ADL)	Quality of life questionnaire (fine motor control and activities of daily living) supplementary to MDS-UPDRS
		Participant Motor Function Questionnaire	Motor function questionnaire, supplementary to MDS-UPDRS
Non-motor Assessments	Autonomic Tests	SCOPA-AUT	Results of SCOPA-AUT assessment, covering 25 data points on: gastrointestinal, urinary, cardiovascular, thermoregulatory, pupillomotor, and sexual
	Neurobehavioral Tests	Various	Results of various neurobehavioral assessments including depression and anxiety
	Neuropsychological Tests	Various	Results of various neurophysiological / cognitive tests including clock drawing, trail marking, the modified Boston naming test, the Montreal

Group	Subgroup	Filename	Comments
			Cognitive Assessment (MoCA), modified semantic fluency, and the Hopkins verbal learning test
	Olfactory Tests	University of Pennsylvania Smell Identification Test (UPSIT)	Results of the UPSIT test
	Other Cognitive Assessments	Various	Various other cognitive tests
	Sleep Disorder Tests	Various	Results of the Epworth Sleepiness Scale questionnaire and the REM Sleep Behavior Disorder Questionnaire
PPMI Found	Follow up of persons with Neurologic Disease	FOUND Brief Summary (PDF)	A summary of FOUND. See Section 14 for additional information on PPMI FOUND
		FOUND RFQ Alcohol	History and current level of alcohol consumption
		FOUND RFQ Anti-inflammatory Meds	History and current usage of ibuprofen, aspirin and other NSAIDs
		FOUND RFQ Caffeine	History and current level of coffee consumption
		FOUND RFQ Calcium Channel Blockers	History of usage of calcium channel blockers and reason for usage
		FOUND RFQ Dictionary	Data dictionary for the FOUND data
		FOUND RFQ Female Reproductive Health	Details of pregnancies, menstruation, menopause and hormone replacement therapy
		FOUND RFQ Head Injury	Details of historical head injuries
		FOUND RFQ Height and Weight	Details of body type, height and weight at ages 25, 40 and 60
		FOUND RFQ Physical Activity	Summary of level of physical activity and hours of sleep per night at each decade in life

Group	Subgroup	Filename	Comments
		FOUND RFQ Smoking History	History and current level of cigarette and other tobacco usage
		FOUND RFQ Self Reported Dx	Diagnosis of PD and other conditions, reported every 6 months by the participant
PPMI Online	Data & Databases Online	PPMI Online Codebook PPMI Online Dictionary	The codebook and data dictionary for the PPMI Online study. See Section 12 for an overview of PPMI Online.
	Cognition-Online	Cognitive Change	Yes/no record of cognitive change (can be captured by someone other than the participant)
	Family History-Online	Family History of PD	First degree relatives with a diagnosis of PD
	Medical-Online	Assessment of Constipation Caffeine Consumption Chemical Exposure COVID-19 History Head Injuries Health History Annually Health History Quarterly History of Falls: Baseline History of Falls: Surveillance Medication History Pesticides at Work Physical Activity Smoking History	Participant medical history. In some cases, these tables capture information at regular intervals in the participant's lifetime. Please refer to the PPMI Online Dictionary for further details.
	Motor / MDS-UPDRS-Online	MDS-UPDRS Part 1 MDS-UPDRS Part 2 Participant Motor Function	The sections of the MDS-UPDRS questionnaires that are completed by participants plus some supplementary functional motor questions
	Neurobehavioral - Online	Geriatric Depression Scale Parkinson Anxiety Scale	The short form Geriatric Depression Scale (Yesavage <i>et al.</i> , 1983) and the Parkinson's Anxiety Scale (Leentjens <i>et al.</i> , 2004)
	Neuropsychological - Online	Penn Parkinson's Daily Activities Questionnaire	PDAQ-15 (Brennan <i>et al.</i> , 2016)

Group	Subgroup	Filename	Comments
	Patient Status-Online	Age of PD Diagnosis Genetic Testing Results High Interest Questions for non-PD Cohort High Interest Questions for PD Cohort Hyposmia PD Return Study for non-PD PD Return Study for PD Participant Visit Information	Various tables associated with participant status. study visits, prodromal symptoms and PD diagnosis. Please refer to the PPMI Online Dictionary for further details.
	Sleep Disorder-Online	Epworth Sleepiness Scale PD Sleep Scale PDSS-2 PPMI RBD Sleep Questionnaire RBD1Q Postuma Acting Out Dreams	The Epworth Sleepiness Scale (Johns, 1991), the PD Sleep Scale, revised version (Trenkwalder <i>et al.</i> , 2011), a PPMI specific sleep questionnaire, and the RBD1Q test for RBD (Postuma <i>et al.</i> , 2012)
	Study Enrollment-Online	How you heard about PPMI Non-Completer Survey Participant Enrollment Status	Information about participant enrollment and which cohort they are in (Parkinson's / Non-Parkinson's)
	Demographics-Online	Occupation and Military Service Race and Ethnicity Registration Information Residential Location Socioeconomic Status	Various tables capturing participant demographic information. Please refer to the PPMI Online Dictionary for further details.
PPMI Remote	PPMI Remote Screening	Remote Screening High Interest Questionnaire	A list of questions that indicate whether a participant has one or more key prodromal indicators (e.g., olfactory problems or acting out dreams). See also Section 13 for an overview of PPMI Remote.
		Remote Screening Participant Progress	Administrative details of remote screening: dates, statuses etc.
		Remote Screening Screener	Demographic information about the participant being screened, e.g., gender, ethnicity





Group	Subgroup	Filename	Comments
		Remote Screening Smell Test Direct Screener	Basic information about participants referred through the Smell Test Direct initiative, e.g., date, PD diagnosis status
		Remote Screening UPSIT Screening Questionnaire	Selected questions from the University of Pennsylvania Smell Identification Test used to screen participants
		Remote University of Pennsylvania Smell Identification Test (UPSIT)	Results of the full UPSIT questionnaire comprising 40 smell tests. (Doty <i>et al.</i> , 1984)  Note that this only captures data for participants who have not also enrolled in PPMI Clinical. For participants who are enrolled in both PPMI Clinical and PPMI Remote, the data can be found in the UPSIT file on the non-motor assessments group, where the PPMI Remote records can be identified with PAG_NAME = 'UPSITPRO'. See also Section 6.2.
Archived PPMI Data	Data Collected in Prior EDC	Various	Not likely to be required
	Publication-associated Archives	Various	Not likely to be required
	Quarterly Data Freezes	Snapshots of the PPMI data taken quarterly	Includes all core tables in .csv format but not supporting documentation



## 17 APPENDIX B – SQL SCRIPTS

**Note that all code/scripts within this guide are provided as is with no guarantee of accuracy or completeness. For questions, reach out to [resources@michaeljfox.org](mailto:resources@michaeljfox.org).**

### Script 1: Count of enrolled/withdrawn/complete participants by cohort and genetic subgroup (Section 3.9)

Note that, by default, MS SQL Server, and certain other DBMSs such as MySQL, are case insensitive with respect to column values. However, if using a DBMS such as PostgreSQL that is case sensitive it will be necessary to match on both 'Enrolled' and 'enrolled' and both 'Withdraw' and 'withdrew'.

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Number of enrolled Parkinson's and Prodromal participants by genetic subgroup */

SELECT COHORT_DEFINITION,
       (CASE WHEN ENRLPINK1 + ENRLPRKN + ENRLSRDC + ENRLHPSM + ENRLRBD + ENRLRRK2 +
        ENRLSNCA + ENRLGBA > 1 THEN 'Multiple factors'
        WHEN ENRLPINK1 = 1 THEN 'PINK1'
        WHEN ENRLPRKN =1 THEN 'PARKIN'
        WHEN ENRLSRDC =1 THEN 'SRDC'
        WHEN ENRLHPSM = 1 THEN 'HPSM'
        WHEN ENRLRBD = 1 THEN 'RBD'
        WHEN ENRLRRK2 = 1 THEN 'LRRK2'
        WHEN ENRLSNCA =1 THEN 'SNCA'
        WHEN ENRLGBA = 1 THEN 'GBA'
        ELSE NULL END) AS 'Genetic subgroup',
       COUNT (*) AS 'Patient count'
FROM PPMI.dbo.Participant_Status
WHERE COHORT_DEFINITION LIKE 'P%' /* Filter on Parkinson's and prodromal cases */
AND ENROLL_STATUS IN ('Enrolled', 'Withdrew', 'Complete')
GROUP BY COHORT_DEFINITION,
         (CASE WHEN ENRLPINK1 + ENRLPRKN + ENRLSRDC + ENRLHPSM + ENRLRBD + ENRLRRK2 +
        ENRLSNCA + ENRLGBA > 1 THEN 'Multiple factors'
        WHEN ENRLPINK1 = 1 THEN 'PINK1'
        WHEN ENRLPRKN =1 THEN 'PARKIN'
        WHEN ENRLSRDC =1 THEN 'SRDC'
        WHEN ENRLHPSM = 1 THEN 'HPSM'
        WHEN ENRLRBD = 1 THEN 'RBD'
        WHEN ENRLRRK2 = 1 THEN 'LRRK2'
        WHEN ENRLSNCA =1 THEN 'SNCA'
        WHEN ENRLGBA = 1 THEN 'GBA'
        ELSE NULL END)
ORDER BY COHORT_DEFINITION, COUNT(*) DESC
```



## Script 2: Count of enrolled/withdrawn/complete participants by cohort and sex (Section 4.1)

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Number of PD and prodromal participants by sex */

SELECT S.COHORT_DEFINITION, C.DECODE AS 'SEX',
       COUNT(*) AS 'Patient Count'
FROM PPMI.dbo.Demographics D
     LEFT OUTER JOIN PPMI.dbo.Codes C ON C.ITM_NAME = 'SEX' AND C.CODE = D.SEX
     LEFT OUTER JOIN PPMI.dbo.Participant_Status S ON S.PATNO = D.PATNO
WHERE S.COHORT_DEFINITION LIKE 'P%'
     AND S.ENROLL_STATUS IN ('Enrolled', 'Withdrew', 'Complete')
GROUP BY S.COHORT_DEFINITION, C.DECODE
ORDER BY S.COHORT_DEFINITION, C.DECODE
```

## Script 3: Create participant master table (Section 4.2)

Note that the first time you run this, the DROP TABLE command will issue a warning because the table does not yet exist; you can ignore this.

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Create master table of selected participants demographic and status data */

DROP TABLE PPMI.dbo.Participant_Master

SELECT D.PATNO, D.BIRTHDT, S.COHORT_DEFINITION,
       (CASE WHEN S.ENRLPINK1 + S.ENRLPRKN + S.ENRLSRDC + S.ENRLHPSM + S.ENRLRBD +
S.ENRLLRK2 + S.ENRLSNCA + S.ENRLGBA > 1 THEN 'Multiple factors'
           WHEN S.ENRLPINK1 = 1 THEN 'PINK1'
           WHEN S.ENRLPRKN = 1 THEN 'PARKIN'
           WHEN S.ENRLSRDC = 1 THEN 'SRDC'
           WHEN S.ENRLHPSM = 1 THEN 'HPSM'
           WHEN S.ENRLRBD = 1 THEN 'RBD'
           WHEN S.ENRLLRK2 = 1 THEN 'LRRK2'
           WHEN S.ENRLSNCA = 1 THEN 'SNCA'
           WHEN S.ENRLGBA = 1 THEN 'GBA'
           ELSE NULL END) AS 'Genetic subgroup',
       ROUND (S.ENROLL_AGE,1) AS 'ENROLL_AGE',
       S.ENROLL_DATE, S.ENROLL_STATUS, C1.DECODE as 'SEX', C2.DECODE as 'HANDED',
       (SELECT MIN(PDDXDT)
        FROM PPMI.dbo.PD_Diagnosis_History DH
        WHERE DH.PATNO = D.PATNO) AS 'PD diagnosis date'
INTO PPMI.dbo.Participant_Master

FROM PPMI.dbo.Demographics D
     LEFT OUTER JOIN PPMI.dbo.Codes C1 ON C1.ITM_NAME = 'SEX' AND C1.CODE = D.SEX
     LEFT OUTER JOIN PPMI.dbo.Codes C2 ON C2.ITM_NAME = 'HANDED' AND C2.CODE = D.HANDED
     LEFT OUTER JOIN PPMI.dbo.Participant_Status S ON S.PATNO = D.PATNO

WHERE S.COHORT_DEFINITION LIKE 'P%'
     AND S.ENROLL_STATUS IN ('Enrolled', 'Withdrew', 'Complete')
ORDER BY D.PATNO
```



#### Script 4: Progression of MDS-UPDRS scores (Section 6.1)

Change the participant number highlighted below to the participant in which you are interested.

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Extract progression of total scores and Hoehn and Yahr stage from each section of MDS-
UPDRS for a single participant */

SELECT P1.INFODT, P1.EVENT_ID,
       MAX(P1.NP1RTOT) AS 'Part 1'
       ,MAX(P1P.NP1PTOT) AS 'Part 1P'
       ,MAX(P2.NP2PTOT) AS 'Part 2'
       ,MAX(P3.NP3TOT) AS 'Part 3'
       ,MAX(P3.NHY) AS 'H & Y Stage'
FROM PPMI.dbo.MDS_UPDRS_Part_1 P1
JOIN PPMI.dbo.MDS_UPDRS_Part_1P P1P
  ON P1.PATNO = P1P.PATNO AND P1.INFODT = P1P.INFODT AND P1P.NP1PTOT IS NOT NULL
JOIN PPMI.dbo.MDS_UPDRS_Part_2P P2
  ON P1.PATNO = P2.PATNO AND P1.INFODT = P2.INFODT AND P2.NP2PTOT IS NOT NULL
JOIN PPMI.dbo.MDS_UPDRS_Part_3 P3
  ON P1.PATNO = P3.PATNO AND P1.INFODT = P3.INFODT AND P3.NP3TOT IS NOT NULL
WHERE P1.PATNO = 9030
GROUP BY P1.INFODT, P1.EVENT_ID
ORDER BY P1.INFODT, P1.EVENT_ID
```

#### Script 5: Sex and total UPSIT score by participant (Section 6.2)

This assumes that the smell test results have been imported into a table called UPSIT.

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Extract sex and total UPSIT score at baseline for PD participants */
SELECT U.PATNO, DECODE AS 'SEX', U.TOTAL_CORRECT
FROM PPMI.dbo.UPSIT U, PPMI.dbo.DEMOGRAPHICS D, PPMI.dbo.Codes C,
PPMI.dbo.PARTICIPANT_STATUS S
WHERE U.EVENT_ID = 'BL' AND U.PATNO = D.PATNO
AND U.TOTAL_CORRECT IS NOT NULL
AND S.PATNO = D.PATNO AND S.ENROLL_STATUS IN ('Enrolled', 'Withdrew', 'Complete')
AND S.COHORT_DEFINITION LIKE 'Park%'
AND C.CODE = D.SEX AND C.ITM_NAME = 'SEX'
```



## Script 6: LEDD medication history for single participant (Section 7.1)

Change the participant number highlighted below to the participant in which you are interested.

```
/* This script is provided as is with no guarantees of completeness or accuracy */  
/* LEDD medication history for single participant */  
SELECT PATNO, EVENT_ID, PAG_NAME, LEDTRT, STARTDT, STOPDT, LEDD  
FROM PPMI.dbo.LEDD_Concomitant_Medication_Log  
WHERE PATNO = 9052  
ORDER BY STARTDT
```



## Script 7: Convert LEDD records into total LEDD at each point in time (Section 7.1)

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Convert LEDD values into total LEDD at a point in time */
DROP TABLE PPMI.dbo.LEDD

CREATE TABLE PPMI.dbo.LEDD (
    PATNO INT,
    STARTDT DATE,
    LDOPA FLOAT, /* Total Levodopa daily quantity at a point in time */
    LEDD FLOAT) /* Total LEDD across all dopaminergic medication at a point in time */

/* Get all the dates when medication started or stopped, i.e. dates when changed */
INSERT INTO PPMI.dbo.LEDD
SELECT DISTINCT PATNO, STARTDT AS 'STARTDT', 0.0, 0.0 AS LEDD FROM
PPMI.dbo.LEDD_Concomitant_Medication_Log
UNION
SELECT DISTINCT PATNO, DATEADD("MONTH", 1, STOPDT) AS 'STARTDT', 0.0, 0.0 AS LEDD FROM
PPMI.dbo.LEDD_Concomitant_Medication_Log WHERE STOPDT IS NOT NULL

/* Get the base L-Dopa value and base LEDD value at each point in time */
UPDATE L SET LEDD = S.LEDD, LDOPA = S.LDOPA FROM PPMI.dbo.LEDD L,
    (SELECT L1.PATNO AS 'PATNO', L1.STARTDT AS 'STARTDT',
        SUM(CASE WHEN (L2.LEDTRT LIKE '%LEV%' OR L2.LEDTRT LIKE '%DOP%' OR
L2.LEDTRT LIKE '%RYTA%'
            OR L2.LEDTRT LIKE '%MADOPAR%' OR L2.LEDTRT LIKE '%SINEMET%' OR
L2.LEDTRT LIKE '%STANEK%'
            OR L2.LEDTRT LIKE '%STALEVO%' OR L2.LEDTRT LIKE '%ISICOM%' OR
L2.LEDTRT LIKE '%INBRIJA%'
            OR L2.LEDTRT LIKE '%NACOM%' OR L2.LEDTRT LIKE '%PROLOPA%') THEN
        (CASE WHEN ISNUMERIC (L2.LEDD)=1 THEN CONVERT (FLOAT, L2.LEDD)
ELSE
        L2.LEDDSTRMG * L2.LEDDOSE * L2.LEDDOSFRQ END) ELSE 0.0 END) AS
'LDOPA',
        SUM(CASE WHEN ISNUMERIC (L2.LEDD)=1 THEN CONVERT (FLOAT, L2.LEDD) ELSE
        (CASE WHEN (L2.LEDTRT LIKE '%LEV%' OR L2.LEDTRT LIKE '%DOP%' OR L2.LEDTRT
LIKE '%RYTA%'
            OR L2.LEDTRT LIKE '%MADOPAR%' OR L2.LEDTRT LIKE '%SINEMET%' OR
L2.LEDTRT LIKE '%STANEK%'
            OR L2.LEDTRT LIKE '%STALEVO%' OR L2.LEDTRT LIKE '%ISICOM%' OR
L2.LEDTRT LIKE '%INBRIJA%'
            OR L2.LEDTRT LIKE '%NACOM%' OR L2.LEDTRT LIKE '%PROLOPA%') THEN
        L2.LEDDSTRMG * L2.LEDDOSE * L2.LEDDOSFRQ
        ELSE 0.0 END) END) AS 'LEDD'
    FROM PPMI.dbo.LEDD L1, PPMI.dbo.LEDD_Concomitant_Medication_Log L2
    WHERE L2.PATNO = L1.PATNO AND L2.STARTDT <= L1.STARTDT
    AND (L2.STOPDT >= L1.STARTDT OR L2.STOPDT IS NULL)
    AND (L2.STOPDT >= L2.STARTDT OR L2.STOPDT IS NULL)
    GROUP BY L1.PATNO, L1.STARTDT) S
WHERE S.PATNO = L.PATNO AND S.STARTDT = L.STARTDT
```

Continued on next page...



```
/* Update LEDD value for drugs with LEDD profile of the form LD x A */
UPDATE L SET LEDD = L.LEDD + S.LEDD FROM PPMI.dbo.LEDD L,
  (SELECT L1.PATNO AS 'PATNO', L1.STARTDT AS 'STARTDT', SUM(L1.LDOPA * CONVERT
(FLOAT, RIGHT(L2.LEDD, LEN(L2.LEDD) - 4))) AS 'LEDD'
  FROM PPMI.dbo.LEDD L1, PPMI.dbo.LEDD_Concomitant_Medication_Log L2
  WHERE L2.PATNO = L1.PATNO AND L2.STARTDT <= L1.STARTDT
  AND (L2.STOPDT >= L1.STARTDT OR L2.STOPDT IS NULL)
  AND (L2.STOPDT >= L2.STARTDT OR L2.STOPDT IS NULL) AND L2.LEDD LIKE 'LD x%'
  GROUP BY L1.PATNO, L1.STARTDT) S
WHERE S.PATNO = L.PATNO AND S.STARTDT = L.STARTDT

/* Update LEDD value for drugs with LEDD profile of the form (B + LD) x A */
UPDATE L SET LEDD = L.LEDD + S.LEDD FROM PPMI.dbo.LEDD L,
  (SELECT L1.PATNO AS 'PATNO', L1.STARTDT AS 'STARTDT', SUM(( CONVERT (FLOAT,
SUBSTRING (L2.LEDD, 2, CHARINDEX ('+ LD', L2.LEDD, 1)-3)) + L1.LDOPA ) * CONVERT
(FLOAT, RIGHT(L2.LEDD, 4))) AS 'LEDD'
  FROM PPMI.dbo.LEDD L1, PPMI.dbo.LEDD_Concomitant_Medication_Log L2
  WHERE L2.PATNO = L1.PATNO AND L2.STARTDT <= L1.STARTDT
  AND (L2.STOPDT >= L1.STARTDT OR L2.STOPDT IS NULL)
  AND (L2.STOPDT >=
    L2.STARTDT OR L2.STOPDT IS NULL) AND L2.LEDD LIKE '%+ LD%'
  GROUP BY L1.PATNO, L1.STARTDT) S
WHERE S.PATNO = L.PATNO AND S.STARTDT = L.STARTDT
```

### Script 8: Numbers of participants with each blood test (Section 8.1)

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Determine which blood tests have been captured for all participants */
SELECT LTSTCODE, LTSTNAME, COUNT(DISTINCT PATNO) AS 'Patient Count'
  FROM PPMI.dbo.Blood_Chemistry___Hematology
  GROUP BY LTSTCODE, LTSTNAME
  ORDER BY COUNT(DISTINCT PATNO) DESC, LTSTNAME
```

### Script 9: Blood test details for single participant (Section )

Change the participant number highlighted below to the participant in which you are interested.

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Hemoglobin test results for single participants */
SELECT EVENT_ID, LCOLLDT, LTSTNAME, LVISTYPE, LSIRE, LSIUNIT, LUSRES, LUSUNIT
  FROM PPMI.dbo.Blood_Chemistry___Hematology
  WHERE LTSTCODE = 'HMT40'
  AND PATNO = 9002
  ORDER BY LCOLLDT
```



## Script 10: Merge MDS\_UPDRS scores from PPMI Clinical and PPMI Online (Section 12)

Change the participant number highlighted below to the participant in which you are interested.

```
/* This script is provided as is with no guarantees of completeness or accuracy */
/* Merge MDS-UPDRS Part 1 and Part 2 patient scores from PPMI Clinical and PPMI Online
data for a single participant */

SELECT P1.INFODT, P1.EVENT_ID, NP1SLPN, NP1SLPD, NP1PAIN, NP1URIN, NP1CNST,
NP1LTHD, NP1FATG, NP2SPCH, NP2SALV, NP2SWAL, NP2EAT, NP2DRES, NP2HYGN, NP2HWRT,
NP2HOBBS, NP2TURN, NP2TRMR, NP2RISE, NP2WALK, NP2FREZ
FROM PPMI.dbo.MDS_UPDRS_Part_1P P1
JOIN PPMI.dbo.MDS_UPDRS_Part_2P P2
ON P1.PATNO = P2.PATNO AND P1.INFODT = P2.INFODT
WHERE P1.PATNO = 9030
UNION
SELECT P1.CREATED_AT AS INFODT, P1.EVENT_ID, NP1SLPN_OL AS NP1SLPN, NP1SLPD_OL
AS NP1SLPD, NP1PAIN_OL AS NP1PAIN, NP1URIN_OL AS NP1URIN, NP1CNST_OL AS
NP1CNST, NP1LTHD_OL AS NP1LTHD, NP1FATG_OL AS NP1FATG, NP2SPCH_OL AS NP2SPCH,
NP2SALV_OL AS NP2SALV, NP2SWAL_OL AS NP2SWAL, NP2EAT_OL AS NP2EAT, NP2DRES_OL
AS NP2DRES, NP2HYGN_OL AS NP2HYGN, NP2HWRT_OL AS NP2HWRT, NP2HOBBS_OL AS NP2HOBBS,
NP2TURN_OL AS NP2TURN, NP2TRMR_OL AS NP2TRMR, NP2RISE_OL AS NP2RISE, NP2WALK_OL
AS NP2WALK, NP2FREZ_OL AS NP2FREZ
FROM PPMI.dbo.MDS_UPDRS_Part_1_Online P1
JOIN PPMI.dbo.MDS_UPDRS_Part_2_Online P2
ON P1.PATNO = P2.PATNO AND P1.CREATED_AT = P2.CREATED_AT
WHERE P1.PATNO = 9030

ORDER BY INFODT, EVENT_ID
```





## 18 APPENDIX C – R SCRIPTS

**Note that all code/scripts within this guide are provided as is with no guarantee of accuracy or completeness. For questions, reach out to [resources@michaeljfox.org](mailto:resources@michaeljfox.org).**

In the sample R scripts we use functions from the readr, dplyr and stringr libraries to perform data manipulation. You will need to install the tidyverse package<sup>28</sup> before you can use these (in RStudio select the Install Packages option under Tools to do this).

All scripts assume the data has been downloaded as csv files to C:\PPMI; you can change this as required.

### Script 1: Count of enrolled/withdrawn/complete participants by cohort and genetic subgroup (Section 3.9)

```
# This script is provided as is with no guarantees of completeness or accuracy
# Number of enrolled/completed/withdrawn Parkinson's and Prodromal participants by
# genetic subgroup

library(readr)
library(dplyr)

# Load table into data frame
setwd ("C:\\PPMI")
Participant_Status <- read_csv ("Participant_Status.csv")

# Filter on status and cohort, determine genetic subgroup, then group by, aggregate
# and order results
filter(Participant_Status, tolower(ENROLL_STATUS) %in% c("enrolled", "withdrew",
"complete") & substr(COHORT_DEFINITION,1, 1)== "P") %>%
transmute(PATNO, COHORT_DEFINITION, GENETIC_SUBGROUP = factor(case_when(
  ENRLPINK1 + ENRLPRKN + ENRLSRDC + ENRLHPSM + ENRLRBD + ENRLLRK2 + ENRLSNCA +
ENRLGBA > 1 ~ 'Multiple factors',
  ENRLPINK1 == 1 ~ 'PINK1',
  ENRLPRKN == 1 ~ 'PARKIN',
  ENRLSRDC == 1 ~ 'SRDC',
  ENRLHPSM == 1 ~ 'HPSM',
  ENRLRBD == 1 ~ 'RBD',
  ENRLLRK2 == 1 ~ 'LRRK2',
  ENRLSNCA == 1 ~ 'SNCA',
  ENRLGBA == 1 ~ 'GBA',
  TRUE ~ '' ))) %>%
group_by(COHORT_DEFINITION, GENETIC_SUBGROUP) %>%
summarize(PATIENT_COUNT = n()) %>%
arrange(COHORT_DEFINITION, desc(PATIENT_COUNT))
```

<sup>28</sup> See <https://www.tidyverse.org/>.



## Script 2: Count of enrolled/withdrawn/complete participants by cohort and sex (Section 4.1)

```
# This script is provided as is with no guarantees of completeness or accuracy
# Number PD and Prodromal enrolled/withdrawn/complete participants by cohort and sex

library(readr)
library(dplyr)

# Load tables into data frames
setwd ("C:\\PPMI")
Participant_Status <- read_csv ("Participant_Status.csv")
Demographics <- read_csv ("Demographics.csv")
Codes <- read_csv ("Code_List_-_Harmonized.csv")

# Filter based on cohort and enrollment status
Participant_Status_Filtered <- filter(Participant_Status, tolower(ENROLL_STATUS) %in%
c("enrolled", "withdrew", "complete") & substr(COHORT_DEFINITION,1, 1)== "P")

# Extract the sex decode values from the Codes data frame, noting that a type
conversion to integer is needed
Sex <- filter(Codes, ITM_NAME == "SEX") %>% select(CODE, DECODE) %>% transmute(CODE =
as.numeric(as.character(CODE)), DECODE)

# Join the tables, group, aggregate and order the results
Participant_Status_Filtered %>%
  left_join (Demographics, by = "PATNO") %>%
  left_join (Sex, c("SEX"= "CODE")) %>%
  group_by (COHORT_DEFINITION, DECODE) %>%
  summarize(PATIENT_COUNT = n()) %>%
  arrange (COHORT_DEFINITION, DECODE) %>%
  rename (SEX = DECODE)
```



### Script 3: Create participant master table (Section 4.2)

```
# This script is provided as is with no guarantees of completeness or accuracy
# Create a "participants master" data frame

library(readr)
library(dplyr)

# Load tables into data frames
setwd ("C:\\PPMI")
Participant_Status <- read_csv ("Participant_Status.csv")
Demographics <- read_csv ("Demographics.csv")
Codes <- read_csv ("Code_List_-_Harmonized.csv")
PD_Diagnosis_History <- read_csv ("PD_Diagnosis_History.csv")

# Filter on status and cohort and determine genetic subgroup
Participant_Status_Filtered <- filter(Participant_Status, tolower(ENROLL_STATUS) %in%
c("enrolled", "withdrew", "complete") & substr(COHORT_DEFINITION,1, 1)== "P") %>%
  mutate(GENETIC_SUBGROUP = factor(case_when(
    ENRLPINK1 + ENRLPRKN + ENRLSRDC + ENRLHPSM + ENRLRBD + ENRLLRK2 + ENRLSNCA +
    ENRLGBA > 1 ~ 'Multiple factors',
    ENRLPINK1 == 1 ~ 'PINK1', ENRLPRKN == 1 ~ 'PARKIN', ENRLSRDC == 1 ~ 'SRDC',
    ENRLHPSM == 1 ~ 'HPSM', ENRLRBD == 1 ~ 'RBD', ENRLLRK2 == 1 ~ 'LRK2',
    ENRLSNCA == 1 ~ 'SNCA', ENRLGBA == 1 ~ 'GBA', TRUE ~ '' )))

# Extract the sex and handedness decode values from the Codes data frame, noting that
# a type conversion to integer is needed
Sex <- filter(Codes, ITM_NAME == "SEX") %>% select(CODE, DECODE) %>% transmute(CODE =
as.numeric(as.character(CODE)), DECODE)
Handed <- filter(Codes, ITM_NAME == "HANDED") %>% select(CODE, DECODE) %>% transmute(CODE
= as.numeric(as.character(CODE)), DECODE)

# PD Diagnosis date - tidy up date format and select earliest date for each patient
PD_Diagnosis_History <- mutate(PD_Diagnosis_History, PD_Diagnosis_Date =
as.Date(paste("01/", as.character(PDDXDT), "%d/%m/%Y"))) %>%
  group_by(PATNO) %>%
  summarize (PD_Diagnosis_Date = min(PD_Diagnosis_Date))

# Create the Patient_Master data frame
Participant_Master <- Participant_Status_Filtered %>%
  left_join (Demographics, by = "PATNO") %>%
  left_join (Sex, c("SEX" = "CODE")) %>%
  left_join (Handed, c("HANDED" = "CODE")) %>%
  left_join (PD_Diagnosis_History, by = "PATNO") %>%
  select (PATNO, BIRTHDT, COHORT_DEFINITION, GENETIC_SUBGROUP, ENROLL_AGE,
ENROLL_DATE, ENROLL_STATUS, SEX = DECODE.x, HANDED = DECODE.y, PD_Diagnosis_Date) %>%
  arrange (PATNO)

#Fix date formats
Participant_Master <- mutate(Participant_Master, BIRTHDT = as.Date(paste("01/",
as.character(BIRTHDT), "%d/%m/%Y")))
Participant_Master <- mutate(Participant_Master, ENROLL_DATE = as.Date(paste("01/",
as.character(ENROLL_DATE), "%d/%m/%Y")))

#Display data frame
Participant_Master
```



#### Script 4: Progression of MDS-UPDRS scores (Section 6.1)

Change the participant number highlighted below to the participant in which you are interested.

```
# This script is provided as is with no guarantees of completeness or accuracy
# Extract progression of total scores and Hoehn and Yahr sate from each section of
MDS-UPDRS for a single participant

library(readr)
library(dplyr)

# Load tables into data frames
setwd ("C:\\PPMI")
MDS_UPDRS_Part_1 <- read_csv ("MDS-UPDRS_Part_I.csv")
MDS_UPDRS_Part_1P <- read_csv ("MDS-UPDRS_Part_I_Patient_Questionnaire.csv")
MDS_UPDRS_Part_2P <- read_csv ("MDS_UPDRS_Part_II_Patient_Questionnaire.csv")
MDS_UPDRS_Part_3 <- read_csv ("MDS-UPDRS_Part_III.csv")

#Fix date formats
MDS_UPDRS_Part_1 <- mutate(MDS_UPDRS_Part_1, INFODT = as.Date(paste("01/",
as.character(INFODT)), "%d/%m/%Y"))
MDS_UPDRS_Part_1P <- mutate(MDS_UPDRS_Part_1P, INFODT = as.Date(paste("01/",
as.character(INFODT)), "%d/%m/%Y"))
MDS_UPDRS_Part_2P <- mutate(MDS_UPDRS_Part_2P, INFODT = as.Date(paste("01/",
as.character(INFODT)), "%d/%m/%Y"))
MDS_UPDRS_Part_3 <- mutate(MDS_UPDRS_Part_3, INFODT = as.Date(paste("01/",
as.character(INFODT)), "%d/%m/%Y"))

# Join the data frames filtering on a single participant and take the maximum scores
across each INFODT/EVENT_ID
filter(MDS_UPDRS_Part_1, PATNO == 9300) %>%
  inner_join (MDS_UPDRS_Part_1P, by = c("PATNO","INFODT","EVENT_ID")) %>%
  inner_join (MDS_UPDRS_Part_2P, by = c("PATNO","INFODT","EVENT_ID")) %>%
  inner_join (MDS_UPDRS_Part_3, by = c("PATNO","INFODT","EVENT_ID")) %>%
  filter(!is.na(NP1RTOT), !is.na(NP1PTOT) & !is.na(NP2PTOT) & !is.na(NP3TOT) &
!is.na(NHY)) %>%
  select (INFODT, EVENT_ID, NP1RTOT, NP1PTOT, NP2PTOT, NP3TOT, NHY) %>%
  group_by (INFODT, EVENT_ID) %>%
  summarize (Part1 = max(NP1RTOT), Part1P = max(NP1PTOT), Part2 = max(NP2PTOT), Part3
= max(NP3TOT), HY_Stage = max(NHY)) %>%
  arrange (INFODT, EVENT_ID)
```



## Script 5: Sex and total UPSIT score by participant (Section 6.2)

```
# This script is provided as is with no guarantees of completeness or accuracy
# Extract sex and total UPSIT score at baseline for PD participants

library(readr)
library(dplyr)

# Load tables into data frames
setwd ("C:\\PPMI")
Participant_Status <- read_csv ("Participant_Status.csv")
Demographics <- read_csv ("Demographics.csv")
UPSIT <- read_csv
("University_of_Pennsylvania_Smell_Identification_Test_UPSIT_.csv")
Codes <- read_csv ("Code_List_-_Harmonized.csv")

# Filter based on cohort and enrollment status
Participant_Status_Filtered <- filter(Participant_Status,
tolower(ENROLL_STATUS) %in% c("enrolled", "withdrew", "complete") &
substr(COHORT_DEFINITION,1, 4)== "Park")

# Extract the sex decode values from the Codes data frame, noting that a type
conversion to integer is needed
Sex <- filter(Codes,ITM_NAME =="SEX") %>% select(CODE,DECODE) %>%
transmute(CODE = as.numeric(as.character(CODE)), DECODE)

# Join the data frames filtering out the patients with NULL values in the total
UPSIT score
Participant_Status_Filtered %>%
  inner_join (filter(UPSIT,!is.na(TOTAL_CORRECT)), by = "PATNO") %>%
  inner_join (Demographics, by = "PATNO") %>%
  inner_join (Sex,c("SEX"= "CODE")) %>%
  select (PATNO, SEX=DECODE, TOTAL_CORRECT)
```



## Script 6: LEDD medication history for single participant (Section 7.1)

Change the participant number highlighted below to the participant in which you are interested.

```
# This script is provided as is with no guarantees of completeness or accuracy
# LEDD medication history for participant

library(readr)
library(dplyr)

# Load tables into data frames
setwd ("C:\\PPMI")
LEDD <- read_csv ("LEDD_Concomitant_Medication_Log.csv")

# Fix date formatting
LEDD <- mutate(LEDD, STARTDT = as.Date(paste("01/", as.character(STARTDT)),
"%d/%m/%Y"), STOPDT = as.Date(paste("01/", as.character(STOPDT)), "%d/%m/%Y"))

# Select columns of interest from LEDD log for a single participant
LEDD %>% filter(PATNO == 9052) %>%
  select (PATNO, EVENT_ID, PAG_NAME, LEDTRT, STARTDT, STOPDT, LEDD) %>%
  arrange (STARTDT)
```



## Script 7: Convert LEDD records into total LEDD at each point in time (Section 7.1)

```
# This script is provided as is with no guarantees of completeness or accuracy
# Convert LEDD values into total LEDD at a point in time

library(readr)
library(dplyr)
library(stringr)

# Load table into data frames
setwd ("C:\\PPMI")
LEDD_Concomitant_Medication_Log <- read_csv ("LEDD_Concomitant_Medication_Log.csv")

# Convert the start and stop dates from MM/YYYY to 01/MM/YYYY
LEDD_Concomitant_Medication_Log <- mutate(LEDD_Concomitant_Medication_Log, STARTDT =
as.Date(paste("01/", as.character(STARTDT)), "%d/%m/%Y"))
# Add 1 month to the STOPDT because we assume that the dosage is taken until the end
of the month, hence the change takes effect from the start of the next month
LEDD_Concomitant_Medication_Log <- mutate(LEDD_Concomitant_Medication_Log, STOPDT =
as.Date(paste("01/", (as.integer(str_sub(STOPDT, 1, 2))%12+1), "/",
as.integer(str_sub(STOPDT, 4, 7))+(as.integer(str_sub(STOPDT, 1, 2))+1)%13, sep=""),
"%d/%m/%Y"))

# Get the set of unique dates when LEDD changes for each PATNO
LEDD <- unique(rbind(LEDD_Concomitant_Medication_Log %>% select(PATNO, STARTDT),
LEDD_Concomitant_Medication_Log %>% select(PATNO, STARTDT =
STOPDT) %>% filter(!is.na(STARTDT)))) %>% arrange (PATNO, STARTDT)

# Get the base L-Dopa value at each point in time
LDOPA_VALUES <-
LEDD_Concomitant_Medication_Log[str_detect(LEDD_Concomitant_Medication_Log$LEDTRT, reg
x("LEV|DOP|RYTA|SINEMET|STANEK|STALEVO|INBRIJA|ISICOM|NACOM|PROLOPA", ignore_case=TRUE)
),] %>%
  select(PATNO, STARTDT, STOPDT, LEDD, LEDDSTRMG, LEDDOSE, LEDDOSFRQ) %>%
# Also calculate the LEDD value from base medication details, if available, in case
not populated in LEDD column
mutate(LEDD3 = case_when (is.numeric(LEDDSTRMG) ~ (LEDDSTRMG * LEDDOSE * LEDDOSFRQ) ,
TRUE ~ 0.0)) %>%
  mutate(LEDD2 = case_when (!is.na(as.double(as.character(LEDD)))) ~ as.double
(as.character(LEDD)) , TRUE ~ LEDD3))
```

Continued on next page...



```
LEDD <- left_join (LEDD,
  left_join(LEDD, LDOPA_VALUES, by="PATNO") %>%
    filter(STARTDT.y <= STARTDT.x & (is.na(STOPDT) | (STOPDT >
STARTDT.x & STOPDT > STARTDT.y))) %>%
    mutate (LEDD2 = case_when (is.na(LEDD2) ~ as.double (0.0), TRUE
~ LEDD2)) %>%

  group_by(PATNO,STARTDT.x) %>%
  summarize (LDOPA = sum(LEDD2), .groups = "drop") %>%
  rename (STARTDT = STARTDT.x),
  by = c("PATNO","STARTDT"))

# Get the base LEDD value at each point in time
# If the drug is levodopa and the LEDD value is not available, attempt to derive it
from base medication details
LEDD_VALUES <- mutate(LEDD_Concomitant_Medication_Log, LEDD3 = case_when
(str_detect(LEDD_Concomitant_Medication_Log$LEDTRT,regex("LEV|DOP|RYTA|SINEMET|STANEK|
STALEVO|INBRIJA|ISICOM|NACOM|PROLOPA",ignore_case=TRUE)) ~ (LEDDSTRMG * LEDDOSE *
LEDDOSFRQ) , TRUE ~ 0.0)) %>%
  mutate(LEDD2 = case_when (!is.na(as.double(as.character(LEDD)))) ~ as.double
(as.character(LEDD)) , TRUE ~ LEDD3))

LEDD <- left_join (LEDD,
  left_join(LEDD, LEDD_VALUES, by="PATNO") %>%
    filter(STARTDT.y <= STARTDT.x & (is.na(STOPDT) | (STOPDT >
STARTDT.x & STOPDT > STARTDT.y))
    & !is.na(LEDD2)) %>%
  group_by(PATNO,STARTDT.x) %>%
  summarize (LEDD = sum(LEDD2), .groups = "drop") %>%
  rename (STARTDT = STARTDT.x),
  by = c("PATNO","STARTDT"))

# Update base LEDD value for drugs with LEDD profile of the form LD * A
VALUES_TO_UPDATE <- LEDD_Concomitant_Medication_Log %>%
  filter (substr(LEDD,1,4) == "LD x") %>%
  select (PATNO, STARTDT,STOPDT, LEDD) %>%
  mutate (FACTOR = as.double(str_match(LEDD, "\\d+\\.\\d+$"))))

LEDD <- left_join (LEDD,
  left_join(LEDD, VALUES_TO_UPDATE, by="PATNO") %>%
    filter(STARTDT.y <= STARTDT.x & (is.na(STOPDT) | (STOPDT >
STARTDT.x & STOPDT > STARTDT.y))) %>%
  group_by(PATNO,STARTDT.x) %>%
  summarize (FACTOR1 = sum(FACTOR), .groups = "drop") %>%
  rename (STARTDT = STARTDT.x),
  by = c("PATNO","STARTDT"))
```

Continued on next page...





```
# Update base LEDD value for drugs with LEDD profile of the form (B + LD) x A
VALUES_TO_UPDATE <- LEDD_Concomitant_Medication_Log %>%
  filter (str_detect(LEDD,"\\+\\ LD")) %>%
  select (PATNO, STARTDT,STOPDT, LEDD) %>%
  mutate (FACTOR2 = as.double(str_match(LEDD, "\\d+\\.\\d+")),
          FACTOR3 = as.double(str_match(LEDD, "\\d+\\.\\d+$")),
          FACTOR4 = as.double(str_match(LEDD, "\\d+\\.\\d+$")))

LEDD <- left_join (LEDD, left_join(LEDD, VALUES_TO_UPDATE, by="PATNO") %>%
  filter(STARTDT.y <= STARTDT.x & (is.na(STOPDT) | (STOPDT >
STARTDT.x & STOPDT > STARTDT.y))) %>%
  group_by(PATNO,STARTDT.x) %>%
  summarize (FACTOR2 = sum(FACTOR2), FACTOR3 =
mean(FACTOR3),FACTOR4 = sum(FACTOR4), .groups = "drop") %>% rename (STARTDT =
STARTDT.x), by = c("PATNO","STARTDT"))

# Tidy up null values and make final calculations
LEDD <- mutate (LEDD, LDOPA = case_when(is.na(LDOPA) ~ 0.0, TRUE ~ LDOPA), LEDD =
case_when(is.na(LEDD) ~ 0.0, TRUE ~ LEDD))

LEDD <- mutate (LEDD, LEDD = LEDD
  + case_when(!is.na(FACTOR1) ~ FACTOR1 * LDOPA, TRUE ~ 0.0)
  + case_when(!is.na(FACTOR2) ~ (FACTOR2 * FACTOR3) + (LDOPA * FACTOR4),
TRUE ~ 0.0)) %>%
  select (-FACTOR1, -FACTOR2, -FACTOR3, -FACTOR4)

# Display results
LEDD
```



### Script 8: Numbers of participants with each blood test (Section 8.3)

```
# This script is provided as is with no guarantees of completeness or accuracy
# Determine which blood tests have been captured for all participants

library(readr)
library(dplyr)

# Load table into a data frame
setwd ("C:\\PPMI")
Blood_Chemistry <- read_csv ("Blood_Chemistry___Hematology.csv")

# Count the number of distinct participants with each type of blood test
Blood_Chemistry %>% group_by (LTSTCODE, LTSTNAME) %>%
  summarize (PATIENT_COUNT = n_distinct(PATNO)) %>%
  arrange (desc(PATIENT_COUNT), LTSTNAME) %>%
  print (n = Inf)
```

### Script 9: Blood test details for single participant (Section 8.3)

Change the participant number highlighted below to the participant in which you are interested.

```
# This script is provided as is with no guarantees of completeness or accuracy
# Hemoglobin test results for a single participant

library(readr)
library(dplyr)

# Load table into a data frame
setwd ("C:\\PPMI")
Blood_Chemistry <- read_csv ("Blood_Chemistry___Hematology.csv")

# Fix date formatting
Blood_Chemistry <- mutate(Blood_Chemistry, LCOLLDT = as.Date(paste("01/",
as.character(LCOLLDT)), "%d/%m/%Y"))

# Count the number of distinct participants with each type of blood test
filter (Blood_Chemistry, PATNO == 9002 & LTSTCODE == "HMT40") %>%
  select (EVENT_ID, LCOLLDT, LTSTNAME, LVISTYPE, LSIRES, LSIUNIT, LUSRES, LUSUNIT) %>%
  arrange (LCOLLDT)
```



## Script 10: Merge MDS\_UPDRS scores from PPMI Clinical and PPMI Online (Section 12)

Change the participant number highlighted below to the participant in which you are interested.

```
# This script is provided as is with no guarantees of completeness or accuracy
# Merge MDS-UPDRS Part 1 and Part 2 patient scores from PPMI Clinical and PPMI Online
data for a single participant

library(dplyr)

# Load tables into data frames
setwd ("C:\\PPMI")
MDS_UPDRS_Part_1P <- read.csv ("MDS-UPDRS_Part_I_Patient_Questionnaire.csv",
stringsAsFactors = TRUE)
MDS_UPDRS_Part_2P <- read.csv ("MDS_UPDRS_Part_II_Patient_Questionnaire.csv",
stringsAsFactors = TRUE)
MDS_UPDRS_Part_1_Online <- read.csv("MDS-UPDRS_Part_I_Non-Motor_Aspects__Online_.csv",
stringsAsFactors = TRUE)
MDS_UPDRS_Part_2_Online <- read.csv("MDS-UPDRS_Part_II_Motor_Aspects__Online_.csv",
stringsAsFactors = TRUE)

# Fix date formats
MDS_UPDRS_Part_1P <- mutate(MDS_UPDRS_Part_1P, INFODT = as.Date(paste("01/",
as.character(INFODT)), "%d/%m/%Y"))
MDS_UPDRS_Part_2P <- mutate(MDS_UPDRS_Part_2P, INFODT = as.Date(paste("01/",
as.character(INFODT)), "%d/%m/%Y"))
MDS_UPDRS_Part_1_Online <- mutate(MDS_UPDRS_Part_1_Online, INFODT =
as.Date(paste("01/", as.character(CREATED_AT)), "%d/%m/%Y"))
MDS_UPDRS_Part_2_Online <- mutate(MDS_UPDRS_Part_2_Online, INFODT =
as.Date(paste("01/", as.character(CREATED_AT)), "%d/%m/%Y"))

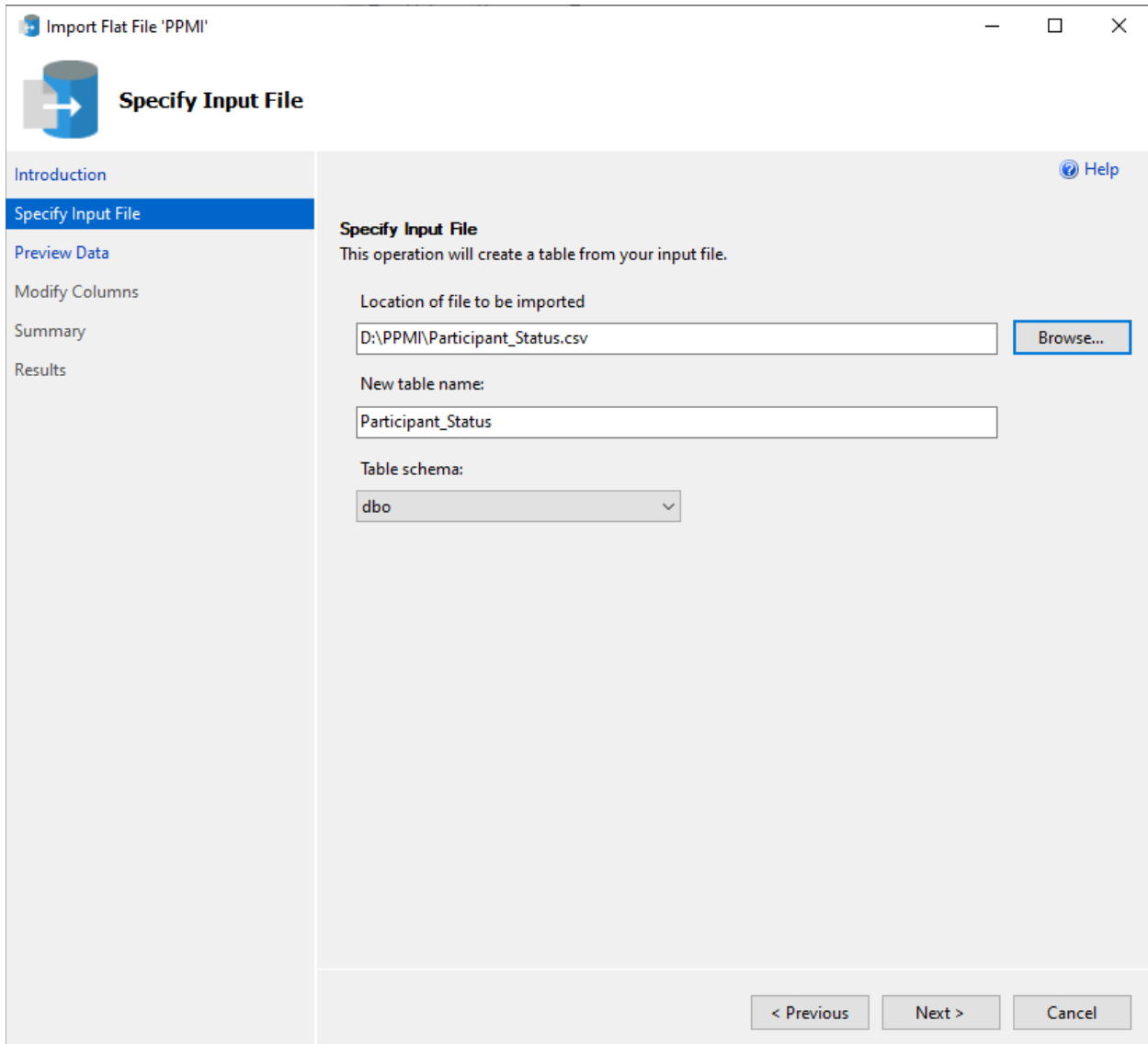
# Join the data frames for PPMI Clinical
MDS_UPDRS_Clinical <- filter(MDS_UPDRS_Part_1P, PATNO == 9030) %>%
  inner_join (MDS_UPDRS_Part_2P, by = c("PATNO","INFODT","EVENT_ID")) %>%
  select (INFODT, EVENT_ID, NP1SLPN, NP1SLPD, NP1PAIN, NP1URIN, NP1CNST, NP1LTHD,
NP1FATG, NP2SPCH, NP2SALV, NP2SWAL, NP2EAT, NP2DRES, NP2HYGN, NP2HWRT, NP2HOBB,
NP2TURN, NP2TRMR, NP2RISE, NP2WALK, NP2FREZ) %>%
  arrange (INFODT, EVENT_ID)

# Join the data frames for PPMI Online and rename columns to match PPMI Clinical data
MDS_UPDRS_Online <- filter(MDS_UPDRS_Part_1_Online, PATNO == 9030) %>%
  inner_join (MDS_UPDRS_Part_2_Online, by = c("PATNO","INFODT","EVENT_ID")) %>%
  select (INFODT, EVENT_ID, NP1SLPN_OL, NP1SLPD_OL, NP1PAIN_OL, NP1URIN_OL,
NP1CNST_OL, NP1LTHD_OL, NP1FATG_OL, NP2SPCH_OL, NP2SALV_OL, NP2SWAL_OL, NP2EAT_OL,
NP2DRES_OL, NP2HYGN_OL, NP2HWRT_OL, NP2HOBB_OL, NP2TURN_OL, NP2TRMR_OL, NP2RISE_OL,
NP2WALK_OL, NP2FREZ_OL) %>%
  rename (NP1SLPN = NP1SLPN_OL, NP1SLPD = NP1SLPD_OL, NP1PAIN = NP1PAIN_OL, NP1URIN =
NP1URIN_OL, NP1CNST = NP1CNST_OL, NP1LTHD = NP1LTHD_OL, NP1FATG = NP1FATG_OL, NP2SPCH
= NP2SPCH_OL, NP2SALV = NP2SALV_OL, NP2SWAL = NP2SWAL_OL, NP2EAT = NP2EAT_OL, NP2DRES
= NP2DRES_OL, NP2HYGN = NP2HYGN_OL, NP2HWRT = NP2HWRT_OL, NP2HOBB = NP2HOBB_OL, NP2TURN
= NP2TURN_OL, NP2TRMR = NP2TRMR_OL, NP2RISE = NP2RISE_OL, NP2WALK = NP2WALK_OL,
NP2FREZ = NP2FREZ_OL) %>%
  arrange (INFODT, EVENT_ID)

# Combine the two data frames
union(MDS_UPDRS_Clinical, MDS_UPDRS_Online) %>% arrange (INFODT, EVENT_ID)
```

## 19 APPENDIX D – IMPORTING DATA INTO A RELATIONAL DATABASE

To import the PPMI data into an RDBMS you will have first downloaded the data you need into CSV files. You will then use a data import tool. In the following we use the example of the Import Flat File wizard in MS SQL Server Management Studio loading the **Participant\_Status** table:



The screenshot shows the 'Import Flat File 'PPMI'' wizard window. The 'Specify Input File' step is selected in the left-hand navigation pane. The main area contains the following fields and controls:

- Location of file to be imported:** A text box containing 'D:\PPMI\Participant\_Status.csv' and a 'Browse...' button.
- New table name:** A text box containing 'Participant\_Status'.
- Table schema:** A dropdown menu with 'dbo' selected.
- Buttons:** '< Previous', 'Next >', and 'Cancel' are located at the bottom right.

In an RDBMS it is important to pay attention to datatypes and you will likely need to override some of the default datatypes assigned, using the data dictionary described in section 3.7 as a guide. We recommend the following to minimize import errors and rejection of records:



Import Flat File 'PPMI'

## Modify Columns

Introduction  
Specify Input File  
Preview Data  
Modify Columns

**Modify Columns**  
This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.

Column Name	Data Type	Primary Key	<input type="checkbox"/> Allow Nulls
PATNO	smallint	<input type="checkbox"/>	<input type="checkbox"/>
COHORT	tinyint	<input type="checkbox"/>	<input type="checkbox"/>
COHORT_DEFINITION	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
ENROLL_DATE	nvarchar(50)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
ENROLL_STATUS	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
STATUS_DATE	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
ENROLL_AGE	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
INEXPAGE	nvarchar(50)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
AV133STDY	tinyint	<input type="checkbox"/>	<input checked="" type="checkbox"/>
PPMI_ONLINE_ENROLL	bit	<input type="checkbox"/>	<input type="checkbox"/>
CONCOHORT	tinyint	<input type="checkbox"/>	<input checked="" type="checkbox"/>
CONCOHORT_DEFINITION	nvarchar(100)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
CONLRRK2	bit	<input type="checkbox"/>	<input checked="" type="checkbox"/>
CONGBA	bit	<input type="checkbox"/>	<input checked="" type="checkbox"/>
CONSNA	tinyint	<input type="checkbox"/>	<input checked="" type="checkbox"/>
CONHPSM	nvarchar(1)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
CONRBD	nvarchar(1)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
PHENOCNV	nvarchar(1)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
DIAG1	nvarchar(1)	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Row granularity of error reporting (performance impact with smaller ranges)

< Previous Next > Cancel

**Annotations:**

- For numeric fields, avoid use of datatypes like bit, smallint and tinyint. Import all numeric fields as either int or float.
- Make all fields nullable
- Import date fields using datatype date where appropriate
- For text fields, you may get truncation errors if the default fields sizes aren't big enough. Look at the data dictionary and adjust the datatypes accordingly.

Note that in some cases, for example MDS-UPDRS measurements, there are missing values. In most cases these will be blank fields, and this is dealt with by making the field nullable as shown above. However, in a few cases you may find text values in numeric fields, typically indicating that a measurement was not possible. To deal with these cases, you have three options:

- Filter these rows out before you import the data (e.g., delete the rows in MS Excel) - Recommended
- Change them to some default value (e.g., using the find and replace function in MS Excel)
- Import them as text fields and deal with them in your analysis

## 20 APPENDIX E – FURTHER DETAILS OF NSD-ISS DERIVATION

### Rules for applying NSD staging in case participants are NSD but are missing D or F variables

1. Missing D variable
  - a. If the participant had D characterization at the prior visit the result from that test is carried forward
  - b. If the participant was D- at the prior visit, missing D for the visit in question but has D+ status at the next visit, prior visit is carried forward for staging of the visit in question
2. Missing F variable (F=functional status measures)
  - a. Participants who qualify for NSD definition and have S+/D+ status are assigned Stage 3-6 based on the functional anchors. There are 3 tracks for progression along the Stage 3-6 continuum
    - i. Cognitive
    - ii. Motor
    - iii. Non-motor
  - b. If participants were missing data required to calculate the functional status but had values from a previous visit, the observations from the prior visit are carried forward.
  - c. If participants are missing data in at least one domain, the stage is not calculated *unless* any available domain reaches stage 6
  - d. If data are available for all 3 domains, stage assignment will be based on the one that is most advanced (worst)

#### Incorporating G status into the staging analytic dataset

1. G status (SNCA fully pathogenic variants) defines stage 0
2. G status defines Genetic Risk ( $R^L$  vs  $R^H$ ) independent of the NSD status – that information can be retrieved from the genetic characterization (see below)

	Definition	Variable name (s)
S assessment	CSF asyn SAA +	S
D assessment	DAT binding <75% expected for age and sex	D
G assessment	See table below for examined genes	G
NSD	meets NSD definition, i.e., S+ OR SNCA triplication	NSD
NSD stage	NSD-ISS stage	Stage (character) Stage_NUM (numeric)

#### Additional participant characteristics that may be considered in sample selection include:

1. **Genetic characterization:** only data relevant to GBA, LRRK2, SNCA, PINK1, PRKN, PARK7, and VPS35 variants and only variants that are considered pathogenic based on the Recommendations for Determining PPMI Eligibility for Parkinson's Disease Genetics developed by the genetic core (methods document available for download).
2. **Clinical traits** such as (i) hyposmia (UPSIT score  $\leq 15^{\text{th}}$  percentile adjusted for age and sex) or (ii) RBD (only when PSG confirmed RBD)

**Relevant variables:**

Enrollment Cohort	Subgroup	LONI nomenclature *
Parkinson's disease	Sporadic PD	COHORT = 1
	LRRK2	COHORT = 1, CONLRRK2 = 1
	GBA	COHORT = 1, CONGBA = 1
	SNCA or rare mutations (PINK1, VPS35, PARKIN, and PARK7)	COHORT = 1, CONSNCA = 1, or CONPINK1, CONVPS35, CONPRKN, CONPARK7
Prodromal	RBD	COHORT = 4, ENLRBD** = 1
	Hyposmia / general risk	COHORT = 4, ENRLHPSM = 1
	LRRK2	COHORT = 4, CONLRRK2 = 1
	GBA	COHORT = 4, CONGBA = 1
	SNCA or rare mutations	COHORT = 4; CONSNCA = 1, or CONPINK1, CONVPS35, CONPRKN, or CONPARK7

**\*for genetic variables, some individuals who have a value of 0 may not have been genotyped centrally . All those with value of 1 have genetic variant confirmed.**

**\*\*RBD with PSG confirmation only**

**Example Considerations for researchers when selecting analytic sample (NOTE: some examples refer to data that are not available for download but can be accessed via request):**

Analysis of PD cohort enrollees:

- Consider excluding S- participants (CSF asyn SAA negative or Type II (which may indicate MSA)) if objective is to study NSD
- May exclude participants with specific disease-associated genetic variants (depending on the analysis)
- Note that for some participants, baseline DAT SPECT was interpreted as abnormal based on visual read, but subsequent quantitative analysis showed DAT binding was > 75% expected for age and sex.

Analysis of HC cohort enrollees:

- May exclude participants with specific PD-associated genetic variants (depending on the analysis)
- May exclude if dopamine transporter binding is <75% expected for age and sex
- May exclude participants who are CSF asyn SAA+

Analysis of the prodromal cohort enrollees:

- NSD staging refines definition of prodromal and may substitute for “phenoconversion”
- Investigator diagnosis can be determined from other variables in the datacut. (Note that some individuals recruited into the prodromal cohort were determined by the investigator at the baseline visit to have a clinical diagnosis of PD or other neurologic disorder)
- Individuals enrolled in prodromal cohort may carry genetic variants

## Inconsistency of the SAA read outs from different visits

CSF asyn SAA is a test with high sensitivity and specificity but can have false negative and positives. The version of SAA changed over the time course of the study; the most recent version (i.e. 24-hour assay) is considered to be the most accurate. Consider excluding participants enrolled in the PD group who have inconsistent CSF asyn SAA results, depending on the analysis focus. There are individuals who had a negative CSF asyn SAA with the 150-hour assay who had CSF asyn SAA Type II on the 24-hour assay. If a prodromal or NMC participant transitions from SAA- to SAA+ status, this change likely reflects the true evolution of the state of the biology of that individual with the caveat that we are still in the process of collecting longitudinal SAA data, validation of the test / retest variability.

### A note on enrollment cohort

NOTE: the enrollment cohort may be discordant with NSD stage or other data (including genetic characterization). Some people enrolled in PD cohort have normal DAT scan or are not SAA+. Some people enrolled in genetic cohort have error in genotyping. Some healthy controls are not healthy; they have abnormal asyn or abnormal DAT binding. Some individuals enrolled in the prodromal cohort received a PD diagnosis by the investigator at baseline visit (that diagnosis will be reflected in the Investigator diagnosis field). A few healthy controls received a diagnosis of PD by the investigator on follow-up.

**The consensus genetic data (CONLRRK2, CONGBA, etc.) should always be used when available; the ENRL genetic variables are only used when consensus info is missing.**

## Staging Anchors for Application of the NSD-ISS

	Biologic anchors			Anchors of clinical signs or symptoms (stages 2A and 2B) and functional impairment (stages 3-6) <sup>1,2</sup>	
Stage	S	D <sup>a</sup>	G	Domain	Anchor(s)
Stage 0	-	-	SNCA <sub>b</sub>	—	—
Stage 1A	+	-	±	(1) Cognitive (2) Motor (3) Other non-motor	(1) MDS-UPDRS item 1.1 = 0; and (2a) Does not have subthreshold parkinsonism <sup>c</sup> ; and (2b) is not on PD meds; and (3a) Does not have RBD; and (3b) is not hyposmic <sup>d</sup>
Stage 1B	+	+	±		
Stage 2A	+	-	±	(1) Cognitive (2) Motor (3) Other non-motor	(1) Item 1.1 = 1 AND MoCA ≥ 25; or (2a) Has subthreshold parkinsonism <sup>c</sup> ; or (2b) is on PD meds; or (3a) Has RBD; or (3b) is hyposmic <sup>d</sup>
Stage 2B	+	+	±		
Stage 3	+	+	±	(1) Cognitive (2) Motor	(1a) Item 1.1 = 1 AND MoCA ≤ 24; or (1b) Item 1.1 = 2 AND MoCA ≥ 25; or (2) MDS-UPDRS-II = 3-13 AND either subthreshold parkinsonism <sup>c</sup> or PD meds



	Biologic anchors			Anchors of clinical signs or symptoms (stages 2A and 2B) and functional impairment (stages 3-6) <sup>1,2</sup>	
Stage	S	D <sup>a</sup>	G	Domain	Anchor(s)
Stage 4	+	+	±	(1) Cognitive (2) Motor (3) Other non-motor	(1a) Item 1.1 = 2 and MoCA ≤ 24; or (1b) item 1.1 = 3 AND MoCA ≥ 25; or (2) MDS-UPDRS-II = 14-26; or (3) MDS-UPDRS-I (excluding item 1.1) = 13-24 <sup>e</sup>
Stage 5	+	+	±	(1) Cognitive (2) Motor (3) Other non-motor	(1a) Item 1.1 = 3 AND MoCA ≤ 24; or (1b) item 1.1 = 4 AND MoCA ≥ 25; or (2) MDS-UPDRS-II = 27-39; or (3) MDS-UPDRS-I (excluding item 1.1) = 25-36
Stage 6	+	+	±	(1) Cognitive (2) Motor (3) Other non-motor	(1) Item 1.1 = 4 AND MoCA ≤ 24; or (2) MDS-UPDRS-II ≥ 40; or (3) MDS-UPDRS-I (excluding item 1.1) ≥ 37

<sup>1</sup> Presence of qualifying signs/ symptoms in any single domain qualifies for stage 2 but individuals can have combination in all 3 domains.

<sup>2</sup> Presence of qualifying functional impairment in any single domain qualifies for stage 3-6 but individuals can have combination in all 3 domains.

<sup>a</sup> D positivity defined as < 75% age/sex-expected lowest putamen SBR.

<sup>b</sup> Only fully penetrant pathogenic *SNCA* variants qualify for Stage 0.

<sup>c</sup> Subthreshold parkinsonism defined as MDS-UPDRS-III ≥ 5 excluding postural and action tremor.

<sup>d</sup> Hyposmia defined as UPSIT percentile ≤ 15 (age and sex adjusted).

<sup>e</sup> MDS-UPDRS-I (excluding item 1.1) ≥ 13 is sufficient for stage 4 if stage 2 criteria are met.

MDS-UPDRS=Movement Disorder Society Unified Parkinson's Disease Rating Scale. MoCA=Montreal Cognitive Assessment. NSD=Neuronal Synuclein Disease. PD=Parkinson's disease. RBD=REM sleep behavior disorder. UPSIT= University of Pennsylvania Smell Identification Test.

## REFERENCES

- Brennan, L., Siderowf, A., Rubright, J., Rick, J., Dahodwala, N., Duda, J., Hurtig, H., Stern, M., Xie, S., Rennert, L., Karlawish, J., Shea, J., Trojanowski, J. & Weintraub, D. (2016). "The Penn Parkinson's Daily Activities Questionnaire-15: Psychometric properties of a brief assessment of cognitive instrumental activities of daily living in Parkinson's disease." *Parkinsonism Relat Disord.* 2016 Apr;25:21-6. <https://doi.org/10.1016/j.parkreldis.2016.02.020>
- Brumm M., Siderowf A., Simuni T., Burghardt E., Choi S., Caspell-Garcia C., Chahine L., Mollenhauer B., Foroud T, Galasko D., Merchant .K, Arnedo V, Hutten S., O'Grady A., Poston K., Tanner C., Weintraub D, Kiebertz K., Marek K., Coffey C. (2023). "Parkinson's Progression Markers Initiative. Parkinson's Progression Markers Initiative: A Milestone-Based Strategy to Monitor Parkinson's Disease Progression." *J Parkinsons Dis.* 2023;13(6):899-916. <https://doi.org/10.3233/JPD-223433>
- Cock, P., Fields, C., Goto, N., Heuer, M., & Rice, P. (2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". *Nucleic Acids Research* 38(6): 1767–1771, <https://doi.org/10.1093/nar/gkp1137>
- Craig, D., Hutchins, E., Violich, I. *et al.* (2021). "RNA sequencing of whole blood reveals early alterations in immune cells and gene expression in Parkinson's disease." *Nat Aging* 1:734–747. <https://doi.org/10.1038/s43587-021-00088-6>
- Doty, R., Shaman, P., Kimmelman, C. & Dann, M. (1984). "University of Pennsylvania smell identification test; A rapid quantitative olfactory function test for the clinic." *The Laryngoscope* 94(2):176-178. <https://doi.org/10.1288/00005537-198402000-00004>
- Goetz, C., Tilley, B., Shaftman, S., Stebbins, G., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C. Stern, M., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A., Lees, A., Leurgans, S., LeWitt, P., Nyenhuis, D., Olanow, C., Rascol, O., Schrag, A., Teresi, J., van Hilten, J., LaPelle, N. (2008) "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results." *Movement Disorders* 23(15):2129-2170. <https://doi.org/10.1002/mds.22340>
- Hoehn M. & Yahr M. (1967). "Parkinsonism: onset, progression and mortality." *Neurology* 17:427-442. [http://info-centre.jenage.de/assets/pdfs/library/hoehn\\_yahr\\_NEUROLOGY\\_1967.pdf](http://info-centre.jenage.de/assets/pdfs/library/hoehn_yahr_NEUROLOGY_1967.pdf)
- Johns, M. (1991). "A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale". *Sleep*, Volume 14, Issue 6, November 1991, Pages 540–545. <https://doi.org/10.1093/sleep/14.6.540>
- Leentjens, A.F.G., Dujardin, K., Pontone, G.M., Starkstein, S.E., Weintraub, D. & Martinez-Martin, P. (2014), "The Parkinson Anxiety Scale (PAS): Development and validation of a new anxiety scale." *Mov Disord.*, 29: 1035-1043. <https://doi.org/10.1002/mds.25919>
- Liao Y., Smyth, G. & Shi, W. (2014). "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." *Bioinformatics* 30(7):923–930. <https://doi.org/10.1093/bioinformatics/btt656>

- Lipsmeier, F., Taylor, K., Postuma, R. *et al.* (2022) "Reliability and validity of the Roche PD Mobile Application for remote monitoring of early Parkinson's disease". *Sci Rep* 12, 12081. <https://doi.org/10.1038/s41598-022-15874-4>
- Marek, K. *et al.* (2011). "The Parkinson Progressive Marker Initiative (PPMI)." *Progress in Neurobiology* 95(4):629-635. <https://doi.org/10.1016/j.pneurobio.2011.09.005>
- Patro, R., Duggal, G., Love, M. *et al.* (2017). "Salmon provides fast and bias-aware quantification of transcript expression." *Nat Methods* 14:417–419. <https://doi.org/10.1038/nmeth.4197>
- Postuma, R., Arnulf, I., Hogl, B., Iranzo, A., Miyamoto, T., Dauvilliers, Y, *et al.* (2012). "A single-question screen for rapid eye movement sleep behavior disorder: A multicenter validation study". *Movement Disorders* 27(7) 913-916. <https://doi.org/10.1002/mds.25037>
- Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M., Bender D., Maller J., Sklar P., de Bakker P., Daly M. & Sham P. (2007) "PLINK: a tool set for whole-genome association and population-based linkage analyses". *Am J Hum Genet.* 2007 Sep;81(3):559-75. <https://doi.org/10.1086/519795>
- Schuff, N., Wu, I., Buckley, S., Foster, E., Coffey, C., Gitelman, D., Mendick, S., *et al.* (2015). "Diffusion imaging of nigral alterations in early Parkinson's disease with dopaminergic deficits." *Movement Disorders* 30(14) 1885-1892. <https://doi.org/10.1002/mds.26325>
- Siderowf, A., Concha-Marambio, L., Lafontant, D., Farris, C. Ma, Y., Urenia, P. *et al.* (2023). "Assessment of heterogeneity among participants in the Parkinson's Progressive Markers Initiative cohort using  $\alpha$ -synuclein see amplification: a cross-sectional study." *Lancet Neurology* 22(5): 407-417. [https://doi.org/10.1016/S1474-4422\(23\)00109-6](https://doi.org/10.1016/S1474-4422(23)00109-6)
- Simuni, T., Chahine, L., Poston, K., Brumm, M., Buracchio, T., Campbell, M., *et al.* (2024). "A biological definition of neuronal  $\alpha$ -synuclein disease: towards an integrated staging system for research." *The Lancet* 23(2):178-190. [https://doi.org/10.1016/S1474-4422\(23\)00405-2](https://doi.org/10.1016/S1474-4422(23)00405-2)
- Tomlinson, C., Stowe, R., Patel, S., Rick, C., Gray, R. & Clark, C. (2010). "Systematic Review of Levodopa Equivalency Reporting in Parkinson's Disease." *Movement Disorders* 25(15):2649-2653. <https://doi.org/10.1002/mds.23429>
- Trenkwalder, C., Kohnen, R., Högl, B. Metta, V., Sixel-Döring, F., Frauscher, B., Hülsmann, J., Martinez-Martin, P. & Chaudhuri, K. (2011). "Parkinson's disease sleep scale—validation of the revised version PDSS-2". *Movement Disorders* 26(4) 644-652. <https://doi.org/10.1002/mds.23476>
- Yesavage JA, Brink TL, Rose TL *et al* (1983). "Development and validation of a Geriatric Depression Screening Scale: a preliminary report." *J Psychiatr Res* 17:37–49. [https://doi.org/10.1016/0022-3956\(82\)90033-4](https://doi.org/10.1016/0022-3956(82)90033-4)