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 29th, 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	• Site identifier • Patient identifier (PATNO) • Cohort • Subgroup • Status • Enrollment phase	See Sections 3.1 and 3.2 Enrollment phase is pre or post June 2020 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

Type of information	Summary of data provided	Comments
		RBD (e.g. LRRK2 + GBA, GBA + RBD).
Diagnosis	• Primary diagnosis • Neuronal alpha-synuclein disease (NSD) and associated NSD_ISS staging indicators • Up to 3 previous diagnoses	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	• Visit identifier (EVENT_ID) • Visit date • Age at enrollment • Age at visit date • Years since study enrollment	See Section 5
Participant demographics	• Sex • Years of education • Ethnicity • Family history of PD • Handedness • Gender identity • Sexual orientation • Body mass index • Age at diagnosis • Age at onset of symptoms • Time between diagnosis and enrollment	See Section 4
Participant symptoms	• Dominant side with symptoms • Presence of common symptoms (tremor, rigidity, bradykinesia, etc.)	
Medication	• Type of treatment (dopaminergic, DBS) • LEDD	See Section 7.1 for LEDD discussion

Type of information	Summary of data provided	Comments
Olfactory test results	• University of Pennsylvania Smell Identification Test (UPSIT)	See Section 6.2
Cognitive test results	• Montreal Cognitive Assessment (MOCA) • Benton judgement of line orientation • Clock drawing test • Lexical fluency • Modified Boston Naming Test • Hopkins Verbal Learning Test (HVLT) • Letter Number Sequencing Test • Symbol Digit Modalities Test • Trails Marking Test (TMT) • Mild Cognitive Impairment (MCI) • Investigator cognitive state diagnosis • Modified Schwab & England ADL score	
Behavioral test results	• Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease (QUIP) • Geriatric Depression Scale (GDS) • State-Trait Anxiety Inventory (STAI)	
Sleep and tests	• Epworth Sleepiness test • REM Sleep Behavior Disorder Questionnaire	
Other autonomic tests	• SCOPA-AUT test • Orthostasis indicator	Note only 7 of 25 data items included for SCOPA
Motor test results	• MDS-UPDRS part 1 individual scores	See Section 6.1

Type of information	Summary of data provided	Comments
	- MDS-UPDRS total scores for parts 1, 2, 3 and 4, including a separate "on" score for part 3 - MDS-UPDRS total "on" and "off" scores - Hoehn & Yahr stage (original scheme) for both "on" and "off" - Tremor Dominant (TD) / Postural Instability and Gait Difficulty (PIGD) scores	
Progression milestones	- Progression milestones associated with the visit, related to activities of daily living	See Brumm, *et al.* (2023)
CSF test results	- Amyloid-Beta - p-tau and t-tau - Alpha-Synuclein Seeding Amplification Assay (SAA) - Neurofilament light (serum and CSF) - High hemoglobin	For SAA, see Section 8.5
Blood and urine test results	- Uric acid - Urine BMP Total di-18:1 Species - Urine BMP Total di-22:6 Species	Relates to Project 145
DaTscan results	- DaTscan measurements of striatal binding ratio (SBR) of left, right, ipsilateral and mean caudate, putamen and striatal regions	
NSD supplementary data fields	- S, D and G fields used to derive NSD and NSD-ISS staging	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.