LRRK2 COHORT CONSORTIUM (LCC) DATA DICTIONARY

(Biologics Data)

LRRK2 COHORT CONSORTIUM (LCC)

Biologics Data

The LRRK2 Cohort Consortium (LCC) was a multi-site study of people with and without Parkinson's disease who carry mutations in the leucine-rich repeat kinase 2 (LRRK2) gene. As part of the testing battery, biological samples including serum, plasma, DNA and RNA from blood, whole blood, urine, and cerebrospinal fluid were collected from a subset of participants. These samples are distributed to academic laboratories and industry groups on an ongoing basis for analysis and the resultant data are returned, aggregated, and combined with a set of key clinical measures that were collected through cross-sectional and routine longitudinal assessments (health and medical history case report forms, administered in a clinic). Qualified researchers can access LCC biologics data at https://app.box.com/folder/13895520391?v=lcc.

USING LCC BIOLOGICS DATA

Structure and Unique Identifier

The LCC biologics data set is shared in **long format**: one row/observation contains information for *one biological sample analysis result for one participant*. That means that each participant *may* have multiple rows in the data set, if multiple analyses were performed on samples attributed to that participant.

The data set contains a **Unique Identifier** for participants, a set of **Analysis Result Variables** indicating the biologic sample analysis result (and associated project metadata), and a set of **Clinical Variables.** Definitions for each are provided below.

Field	Name	Definition
Unique Identifier	lrrkid	String which contains a unique, de-identified, number assigned to participants. Ex: '580010'.
Analysis Result Variables	Ex: TESTNAME, TESTVALUE	Individual biological sample analysis result variable definitions are provided in this Data Dictionary below. All variables are generated directly by project owners, including derived variables, which have undergone transformation to render them usable for analysis. When this is the case, it is indicated in the data dictionary.
Clinical Variables	Ex: pdenrl, updrs	Individual clinical variable definitions are provided in this Data Dictionary below. Most variables are raw responses from health care professionals reported in case report forms. Others are derived variables, which have undergone transformation to render them usable for analysis. When this is the case, it is indicated in the data dictionary.

Missing Data

Since LCC relied on data collection from multiple clinical sites, who followed non-uniform variations of the same assessment battery, there were many cases where data collection was already underway at individual sites before the final set of clinical variables for the consolidated data set was decided. Because of this, there are certain variables defined in the final clinical data set that were not collected to begin with by individual clinical sites. In these cases, the variable is populated with the value 'Z' to indicate that it was not received. In some cases, this may be for an entire assessment, for example, the UPDRS Part III, so all variables relating to that assessment are set to 'Z'. In other cases, it may be that some variables on a form were not received, but others were, so only the ones that weren't received are set to 'Z'. If the variable was received for inclusion in the final clinical data set but there is no data in it, these were left as missing values (not set to 'Z').

Genetic Mutation

As part of the study protocol, site staff reported if participants had one or more pathogenic LRRK2 variant(s) during their study visit, captured in the LRRK2_MUTATION variables. Following completion of the study, LCC also generated whole genome sequencing (WGS) data from whole blood samples for a *subset* of participants (N=599). From these WGS data, specific pathogenic LRRK2 and GBA variants were annotated in this data cut, captured in the LRRK2_var and GBA_var variables. The LRRK2_var and GBA_var include lists of any pathogenic mutations (annotated with protein coding change). This information is alternatively indicated in the LRRK2_status and GBA_status variables, which numerically code if one or more pathogenic variants are present (1) or if no pathogenic variants are detected (0). It is important to note that not all LRRK2_MUTATION values have corresponding LRRK_var / GBA_var and status values, as some individuals have carrier status reported by site, but do not have data from WGS.

Analysis Result Variables

Details: The following variables are collected in conjunction with ongoing biological sample analysis projects, in partnership with various academic laboratories and industry groups. Analysis methodologies are documented and available for each project (identifiable by Biomarker_projectID).

Variable Name	Data Type	Max Length	Variable Description	Notes
Biomarker_project_type	Text	100	Description of the project's research design type	i.e., 23andMe, AJ_longitudinal, Cross-sectional
PR_DATE	Date	-	Date (MM/DD/YYYY) that data are made publicly available to researchers external to the LCC	
Biomarker_PI	Text	100	Name of the principal investigator that generated analysis results	
Biomarker_institution	Text	100	Institution of the principal investigator that generated analysis results	
Biomarker_projectID	Numeric	8	A project identifier assigned by the LCC study team	
Biomarker_sampletype	Text	100	Type of sample used for analysis	i.e., plasma, serum, urine, etc.
TESTNAME	Text	100	Name of sample analysis	i.e., a-synuclein, Tau, etc.
TESTVALUE	Numeric	50	Result of sample analysis [Derived]	
UNITS	Text	25	Units of measure for test-analysis	If measure is a mean or other derived variable, that is indicated here
RUNDATE	Date	-	Date sample analysis was performed	

Clinical Variables

View Source Instruments: [link]

Details: Most key clinical variables below were collected through cross-sectional and routine longitudinal assessments (health and medical history case report forms, administered in a clinic by a health care professional). A handful of derived variables, indicated as such, were generated after initial data collection, and have undergone transformation to render them usable for analysis.

Variable Name	Data Type	Max Length	Variable Description	Value	Value Description	Notes
EVENT	Text	8	Visit identifier	V01	Visit 1	
				V02	Visit 2	
				V03	Visit 3	
				V04	Visit 4	
LRRK2_MUTATION	Text	3	Does the subject carry the LRRK2 mutation?	No	-	This variable was reported by clinical staff during site visits
				Yes	-	
datasrc	Numeric	4	Source of data collection	0001	Ashkenazi Jewish Consortium (multi-center study)	
				0002	National Institute of Neurology (Tunis, Tunisia)	
				0003	LRRK2 Progeni (Parkinson's Research: The Organized Genetics Initiative)	
				0004	University of Tubingen (Tubingen, Germany)	
				0005	Hospital Donostia (San Sebastian, Spain)	
				0006	Hospital Clinic de Barcelona (Barcelona, Spain)	
				0007	Norwegian University of Science and Technology (Trondheim, Norway)	

Variable Name	Data Type	Max Length	Variable Description	Value	Value Description	Notes
				0008	Parkinson's Institute (Portland, Oregon)	
				0009	Xuanwu Hospital of Capital Medical University (Beijing, China)	
				0010	Toronto Western Hospital (Toronto, Canada)	
				0011	23andMe Blood Collection Study	
				0012	Hôpital de la Pitié-Salpêtrière (Paris, France)	
gender	Numeric	1	Gender at birth	1	Female	
				2	Male	
educcat_calc	Numeric	1	Calculated highest Level of Education [Derived]	1	Elementary / Middle school or equivalent (less than or equal to 8 years)	While the CRF indicates '4 = Master's level or equivalent' and '5= postgraduate or equivalent', values of 4 or 5 do not exist in the data set, and instead 6 replaces these values indicating post college/university education.
				2	High school or equivalent (9 – 13 years)	
				3	College / University or equivalent (14 – 16 years)	
				6	Post college / University (greater than 16 years)	
racetxt	Text	100	Race	-	-	
ethnctxt	Text	100	Ethnicity	-	-	
pdenrl	Numeric	1	A variable indicating whether an individual had been diagnosed with PD by a physician or other health care professional at the time	0	No Parkinson's disease	
				1	Parkinson's disease	

Variable Name	Data Type	Max Length	Variable Description	Value	Value Description	Notes
			of enrollment into the LCC study, i.e., a participant's baseline diagnosis status			
agediag	Numeric	8	Age at Parkinson disease diagnosis	-	-	
ageonset	Numeric	8	Age at onset	-	-	
demopd_ageassess	Numeric	8	Age (in years) at assessment date reported on the 'Demographics and PD Information' CRF	-	-	
pd_duration	Numeric	8	Calculated PD duration [Derived]	-	-	This variable is derived by comparing an individual's reported age at time of assessment (demopd_ageassess) with their reported age at diagnosis (agediag)
pdmedyn	Numeric	1	A variable indicating a participant reports being on medication to treat PD symptoms during the MDS-UPDRS Part III assessment	0	No	This variable was reported by clinical staff via the 'MDS-UPDRS Part III' CRF
				1	Yes	
MCATOT	Numeric	8	MoCA Total Score [Derived]	Range (0 – 30)	-	This variable was derived centrally for sites that provided participant reported values for the 'Montreal Cognitive Assessment (MoCA)' CRF. If a participant reported less than or equal to 12 years of education (per the educcat_calc variable) and

Variable Name	Data Type	Max Length	Variable Description	Value	Value Description	Notes
						had an unadjusted score of less than 30, the total score was incremented by 1. If any reported values were missing, the total score was left blank
moca_ageassess	Numeric	8	Age (in years) at assessment date reported on the 'Montreal Cognitive Assessment (MoCA)' CRF	Range (0 – 89)	-	If a participant reported their age as greater than 89, this variable was assigned a value of 90
updrs1	Numeric	8	MDS-UPDRS Part I Total Score [Derived]	-	-	This variable is derived by summing the values reported on the 'MDS-UPDRS Part I' CRF
updrs2	Numeric	8	MDS-UPDRS Part II Total Score [Derived]	-	-	This variable is derived by summing the values reported on the 'MDS-UPDRS Part II' CRF
updrs3	Numeric	8	MDS-UPDRS Part III Total Score [Derived]	-	-	This variable is derived by summing the values reported on the 'MDS-UPDRS Part III' CRF
nupdrs1p_ageassess	Numeric	8	Age (in years) at assessment date reported on the 'MDS-UPDRS Part I' CRF	-	-	If a participant reported their age as greater than 89, this variable was assigned a value of 90
nupdrs2p_ageassess	Numeric	8	Age (in years) at assessment date reported on the 'MDS-UPDRS Part II' CRF	-	-	If a participant reported their age as greater than 89, this variable was assigned a value of 90

Variable Name	Data Type	Max Length	Variable Description	Value	Value Description	Notes
nupdrs3_ageassess	Numeric	8	Age (in years) at assessment date reported on the 'MDS-UPDRS Part III' CRF	-	-	If a participant reported their age as greater than 89, this variable was assigned a value of 90
STROKE	Numeric	1	A variable indicating a participant reports having / has had a stroke	0	No	These variables were reported by clinical staff via the 'REM Sleep Disorder Questionnaire' CRF
				1	Yes	
HETRA	Numeric	1	A variable indicating a participant reports having / has had head trauma	0	No	
				1	Yes	
PARKISM	Numeric	1	A variable indicating a participant reports having / has had parkinsonism	0	No	
				1	Yes	
RLS	Numeric	1	A variable indicating a participant reports having / has had RLS	0	No	
				1	Yes	
NARCLPSY	Numeric	1	A variable indicating a participant reports having / has had narcolepsy	0	No	
				1	Yes	
DEPRS	Numeric	1	A variable indicating a participant reports having / has had depression	0	No	
				1	Yes	
EPILEPSY	Numeric	1	A variable indicating a participant reports having / has had epilepsy	0	No	
				1	Yes	
BRNINFM	Numeric	1	A variable indicating a participant reports having / has had inflammatory disease of the brain	0	No	
				1	Yes	
CNSOTH	Numeric	1	A variable indicating a participant reports having / has had any other disease of	0	No	
				1	Yes	

Variable Name	Data Type	Max Length	Variable Description	Value	Value Description	Notes
			the nervous system not already specified in the 'REM Sleep Disorder Questionnaire' CRF			
CNSOTHCM	Text	240	A free text variable capturing the name of any other disease of the nervous system reported by a participant not already specified in the 'REM Sleep Disorder Questionnaire' CRF	-	-	
SCAU26AT	Text	240	A variable indicating a participant reports using medication for constipation	-	-	These variables were reported by clinical staff via the 'SCOPA-AUT' CRF
SCAU26BT	Text	240	A variable indicating a participant reports using medication for urinary problems	-	-	
SCAU26CT	Text	240	A variable indicating a participant reports using medication for blood pressure	-	-	
SCAU26DT	Text	240	A variable indicating a participant reports using medication for other symptoms not already specified in the 'SCOPA-AUT' CRF	-	-	
LDOPRTYP	Numeric	1	A variable indicating a participant's response to Levodopa or dopamine	1	Not had adequate trial of Levodopa or dopamine agonist	

Variable Name	Data Type	Max Length	Variable Description	Value	Value Description	Notes
			agonist medication, as reported by a health care professional	2	Substantial and sustained response to Levodopa or dopamine agonist	
				3	Negative response to large dose of Levodopa	
				4	Response to Levodopa or dopamine agonist, but not substantial or not sustained	
LDOPRSPC	Numeric	1	A variable indicating a participant's percentage reduction in PD signs and symptoms after treated with Levodopa or dopamine agonist medication, as reported by a health care professional	1	70%-100% reduction in PD signs and symptoms	This variable was collected differently across clinical sites. Some sites captured a larger range of 50-100% reduction in PD signs and symptoms, instead of smaller, distinct ranges (50-69% and 70-100%). To account for this, a value of 6 was added for these cases
				2	50%-69% reduction in PD signs and symptoms	
				3	1%-49% reduction in PD signs and symptoms	
				4	No (0%) reduction in PD signs and symptoms	
				5	Has never had an adequate trial of Levodopa	
				6	Greater than or equal to 50% reduction in PD signs and symptoms	
DCLPDO5Y	Numeric	1	A variable indicating if a participant's response to Levodopa or dopamine agonist has been present for 5 years or more	0	No	
				1	Yes	
GBA_var	Text	100	A free text variable capturing the name of one or more pathogenic GBA variant(s) carried by a participant [Derived]	-	-	This variable was derived from whole genome sequencing data from whole blood samples collected from a subset of participants. A value of 'NC'

Variable Name	Data Type	Max Length	Variable Description	Value	Value Description	Notes
						represents that the participant is a non-carrier
GBA_status	Numeric	1	A variable indicating if a participant has one or more pathogenic GBA variants [Derived]	0	None	
				1	Participant is a carrier of one or more pathogenic variants in GBA	
LRRK2_var	Text	100	A free text variable capturing the name of one or more pathogenic LRRK2 variant(s) carried by a participant [Derived]	-	-	This variable was derived from whole genome sequencing data from whole blood samples collected from a subset of participants. A value of 'NC' represents that the participant is a non-carrier
LRRK2_status	Numeric	1	A variable indicating if a participant has one or more pathogenic LRRK2 variants [Derived]	0	None	
				1	Participant is a carrier of one or more pathogenic variants in LRRK2	
APOE	Text	5	A variable indicating Apolipoprotein E (APOE) genotype results [Derived]	E2/E2	APOE e2/e2 genotype	APOE alleles were determined by whole genome sequencing genotype of rs429358 and rs7412
				E2/E3	APOE e2/e3 genotype	
				E2/E4	APOE e2/e4 genotype	
				E3/E3	APOE e3/e3 genotype	
				E3/E4	APOE e3/e4 genotype	
				E4/E4	APOE e4/e4 genotype	