Dementias Platform UK
Medical Research Council

DATA CURATION POLICY

Background

This document explains the DPUK policy on the curation of data which are uploaded to the DPUK Data Portal. DPUK facilitates the analysis of multiple datasets for the testing of emerging hypotheses and for replication by the curation of cohort data to a common standard, also known as an ontology. This enables multiple datasets to be analysed using a single set of conventions for data structure, variable naming, and value labelling.

All ontologies are optimised for specific use cases. The DPUK ontology is optimised for the analysis of cohort data. Other data models, whether optimised for trials, electronic health records, or disease in general, involve structural complexity that is rarely relevant to cohort-based analyses.

Cohort datasets are complex and the conventions underlying their organisation vary widely. These conventions have been largely determined by local factors and standard practices operating at the time of cohort inception. These conventions are typically cohort-specific, and for cohort research teams, these conventions are convenient in providing a standard for repeated waves of data collection over many years.

For analyses involving multiple datasets, however, cohort-specific data models have to be curated to a common standard before analysis can proceed. This problem is exacerbated for third-party researchers who may not be familiar with any of the cohort-specific data models associated with their project. To simplify the analysis of multiple datasets, and improve data accessibility for third-party researchers, data uploaded to the Data Portal are curated to a common data model. Researchers may request access to either native or curated data.

The DPUK Data Curation Policy supports the FAIR data principles by providing a shared vocabulary and ontology across cohort datasets (https://www.force11.org/group/fairgroup/fairprinciples).

Principles

To simplify the problem of understanding cohort datasets, the DPUK cohort data ontology is based on five principles:

1. Structuring of datasets according to areas of measurement
2. Constraining of variable name length to the minimum required to identify uniquely
3. Identification of each variable in the context of a specific cohort and wave (sweep)
4. Intuitive conventions for abbreviating variable names
5. Machine readability of all ontology elements.

Procedures

The DPUK data model enables variables to be identified uniquely. The variable name has five elements (fields) separated by an underscore.

The source cohort is identified in field 1 by a three letter acronym (table 1) and allows the data source to be readily identified in pooled analyses.

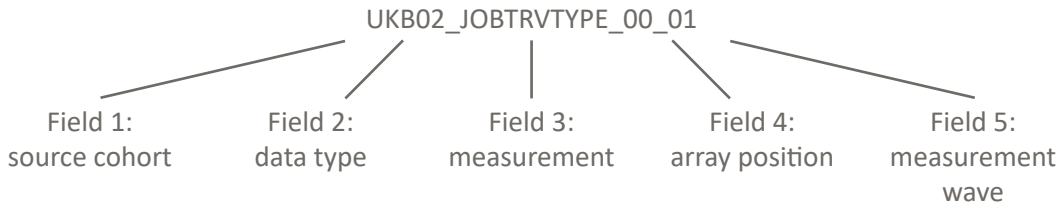
Field 2 identifies the data type and is a two digit number. The data type categories provide a natural structure for data discovery and variable selection (table 2). Due to the constraints of different data types, categorising data type allows different naming rules to be applied across data types. For example, in field 3 which describes the measurement that was made, the variable names for lifestyle variables (category 11) are limited to 12 characters, whereas for imaging data (category 20) which are more complex, the variable name is limited to 17 characters.

Field 3 identifies the measurement that was made and has between 12 and 17 alphanumeric characters according to data type. An intuitive coding structure was adopted whereby syllable-based acronyms, word fragments, abbreviations and minimal numbers were utilised to facilitate easy interpretation. Alphabetic characters are exclusively upper case with the exclusion of lowercase ‘d’ and ‘r’ which are utilised to represent decimals and range within numerically named variables. To use examples of these from UK Biobank from the ‘air pollution’ subcategory: PM2d5APABS10 = Particulate matter air pollution (pm2.5) absorbance 2010 and from the accelerometer subcategory: AVG00H00r00H59 = average acceleration 00:00 - 00:59. A data curation rule book outlines over thirty rules which guide consistency, for example: where time of day is expressed as hours and minutes, a four digit structure is used with H (Hours) inserted in the middle (eg 19H59). Likewise, for consistency, all word fragments are recorded in a codebook, of which there are over 700.

Field 4 describes the position of the measurement in a sequence (array) of serial measurements made on one occasion. For example, a repeated assessment of blood pressure. If field 4 is zero (0) then no repeat assessment was made. If field 4 is 1 than this variable is the first in a sequence of serial measurements, with a value of two representing the second, etc.

Field 5 describes the measurement wave, often referred to as sweep. Here, zero (0) signifies the baseline value at recruitment with the number incrementing by one with each follow-up.

In the example below, the cohort (UKB) is UK Biobank, the data type (02) is sociodemographic, the measurement (JOBTRVTYPE) is mode of transport to work, the array element (0) shows there was no repeat measurement and the wave element (1) describes data collected in the first follow-up.



Data models are subject to development and comments from users are welcome to improve their utility.

Table 1: Cohort Identifiers

	Cohort	Institution	Cohort Code
1	Airwave	Imperial	AIR
2	BRACE	Bristol	BRC
3	Cam-CAN	Cambridge	CAN
4	CamPaiGN	Cambridge	CAM
5	CaPS	Bristol	CAP
6	CFAS I	Cambridge	CFA
7	CFAS II	Cambridge	CFS
8	Cygnus	Manchester	CYG
9	DFP pilot	Oxford	DFP
10	ELSA	UCL	ELS
11	EPINEF	Yonsei (RoK)	EPI
12	Generation Scotland	Edinburgh	GEN
13	GERAD LOAD	Cardiff	GEL
14	GERAD EOAD	Cardiff	GEE
15	ICICLE-PD	Newcastle	IPD
16	NIMROD	Newcastle	NIM
17	OPDC Discovery	Oxford	OPC
18	SMC Amyloid	SMC Seoul (RoK)	SMC
19	TRACK-HD	UCL	THD
20	Whitehall II	UCL	WHI
21	ALSPAC	Bristol	ALS
22	BDR	Bristol	BDR
23	DIAN	UCL	DIA
24	EPIC Norfolk	Cambridge	EPN
25	GENFI	UCL	GEF
26	Healthwise Wales	Cardiff	HWW
27	LBC1936	Edinburgh	LBC
28	Million Women	Oxford	MIW
29	NSHD	UCL	NSH
30	PICNICS	Cambridge	PIC
31	Protect	Exeter	PRO
32	SABRE	UCL	SAB
33	UK Biobank	Oxford	UKB
34	AMPLE	Newcastle	AMP
35	CHARIOT	Imperial	CHA
36	CMERC	Yonsei (RoK)	CME
37	Delphic	UCL	DEL
38	EXTEND	Exeter	EXT
39	HKU-NCDC	Hong Kong University	HKU
40	KOGES	Yonsei (RoK)	KOG
41	LEWY-PRO	Newcastle	LEW
42	Memento	Bordeaux (Fra)	MEM
43	NAMGARM-2	Gyeongsang (RoK)	NAM
44	NICOLA	Queen's Belfast	NIC
45	PaMIR	Nottingham	PAM
46	PREVENT	Edinburgh	PRV
47	PRIME	Queen's Belfast	PRM

Table 2: Cohort Categories

	Parent Category	Characters	
01	Administrative	12	
02	Sociodemographic	12	
03	Family history	12	
04	Early life history	12	
05	Health status	12	
06	Reproductive history	12	
07	Healthcare utilisation	12	
08	Life functionality	12	
09	Psychological status	12	
10	Mental health status	12	
11	Cognitive status	12	
12	Lifestyle	12	
13	Physical environment	12	
14	Social environment	12	
15	Physical examination	12	
16	Linkage	17	
17	Biosample assays	12	
18	Digital phenotyping	12	(17 accelerometry)
19	Imaging	17	
20	Genomics	17	
21	Metabolomics	17	
22	Environmental assessment	12	