Research-ready data for multi-cohort analyses: The Dementias Platform UK (DPUK) C-Surv data model

Sarah Bauermeister
University of Oxford https://orcid.org/0000-0001-9463-6971
Joshua R Bauermeister
University of Oxford
Ruth Bridgman
University of Oxford
Caterina Felici
University of Oxford
Mark Newbury
Swansea University
Laura North
Swansea University
Christopher Orton
Swansea University
Emma Squires
Swansea University
Simon Thompson
Swansea University
Simon Young
University of Oxford
John E J Gallacher (✉ john.gallacher@psych.ox.ac.uk)
University of Oxford

Method Article

Keywords: standard data model, curation, cohort, ontology, multi-cohort

Posted Date: September 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-937113/v1

Abstract

Research-ready data (that curated to a defined standard) increases scientific opportunity and rigour by integrating the data environment. The development of research platforms has highlighted the value of research-ready data, particularly for multi-cohort analyses. Following user consultation, a standard data model (C-Surv), optimised for data discovery, was developed using data from 12 population and clinical cohort studies. The model uses a four-tier nested structure based on 18 data themes and 137 domains selected according to user behaviour or technology. Standard variable naming conventions are applied to uniquely identify variables within the context of longitudinal studies. The model was used to develop a harmonised dataset for 11 cohorts. This dataset populated the Cohort Explorer data discovery tool for assessing the feasibility of an analysis prior to making a data access request. It was concluded that developing and applying a standard data model (C-Surv) for research cohort data is feasible and useful.

Introduction

Biomedical science has become a global enterprise and appetite to access population-based research data across disciplines and jurisdictions grows. The existence of data platforms has increased the discoverability of data enabling collaboration at scale and complexity. One such platform, Dementias Platform UK (DPUK), is a collaboration with 50 population and clinical cohorts (N = 3.5m) that wish to make their data globally available (1). The DPUK Data Portal provides an integrated data discovery, analysis and sharing environment for academics and industry. The opportunity for multi-cohort, multi-modal analyses provided by the Data Portal has highlighted the value of well annotated and accessible research-ready data, understood here as data that have been curated and annotated to a defined standard.

Research-ready data increases scientific opportunity and rigour, as clearly defined data standards integrate the data environment. Nowhere has this been more clearly seen than for the introduction of reference SNP cluster ID (rs) numbers in genetics and the Neuroimaging Informatics Technology Initiative (NIfTI) and Digital Imaging and Communications in Medicine (DiCoM) imaging formats. For population-based research phenotypes there are no established data standards. Typically, studies use a data model that has evolved over time according to each project's scientific priorities and resource constraints. These models use bespoke structures and labelling conventions, and vary in the quality of the curation and documentation.

For third-party scientists learning a different data model for each dataset, and bringing data to research-readiness is labour intensive. It also involves duplication of effort as each third-party scientist will repeat this process. Critically, however, third-party scientists are unlikely arrive at the same research-ready solutions, making rapid and precise replication challenging. For multilateral collaborations, involving work with multiple data models prior to data integration, these challenges are compounded.

These issues can be addressed by preparing data to a common data standard. Developing a standard requires creating an ontology: a conceptual data space where all the facts (observations concerning the elements of the data) and relationships between facts (observations concerning the structure of the data), are defined. It then requires expressing these rules pragmatically as a data model. The data space need not be complex as its function is to simplify and standardise. A data model that simplifies addressing complex questions is useful.

Ontologies are purpose-specific. For example, an ontology designed to describe the link between genetic lineage and infectious disease phenotypes such as the Pangolin ontology (2), will be designed differently from one designed to describe genetic variants and chronic disease phenotypes, as in the Human Phenotype Ontology (3). Ontologies developed for use with health data include the Clinical Data Interchange Standards Consortium (CDISC) (4), for use with trials, and the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (5), for use in primary care. Electronic health record (EHR) ontologies include the Observational Health Data Sciences and Informatics (OHDSI) (6), and Fast Healthcare Interoperability Resources (FHIR) (7). However, these ontologies have structural and semantic complexity that is alien to the natural organisation of population-based research data, such as cohort studies, and are unsuitable for the development of a common data model for research cohort data.

The interest in multi-cohort analyses, expressed through access requests to the DPUK Data Portal, prompted the feasibility of a common data model for research cohort data to be investigated. This model would enable research-ready data to be available to the wider research community at scale and speed. Described here are the background work, development, and application of the C-Surv data model.

Methods

Defining the problem

To define the problem more closely, a landscape review and a user consultation were conducted. For the landscape review the DPUK Cohort Directory was used to sample current practice. Data structures vary considerably across cohort datasets, reflecting the conceptual frameworks of the original investigators. Variable labelling conventions were largely project-specific, and whilst suitable for in-house analysts, might be relatively opaque to third-party users. Documentation varied considerably with no widely used structure or content. Data selection and access request procedures also varied considerably. Some Data Access Committees (DACs) require individual variable selection, whilst others allow the selection of pre-defined groups of variables e.g. all cognitive variables. A small number of DACs allow virtually complete datasets to be accessed. These approaches to data selection represent compromises between administrative convenience and the articulation of scientific rigour.

To identify user needs, stakeholder workshops and informal discussions were conducted. The mission statement for these consultations was to create simple data conventions that could be applied to multi-cohort, multi-modal, cohort data. To provide context, epidemiologic population cohorts were used as use-

cases. The utility of a common data model was recognised by all stakeholders, although reservations were frequently expressed as to whether this was possible given the complexity of cohort datasets. Developing a comprehensive taxonomy for research phenotypes was seen as a separate problem from providing tools for data discovery. Although the prospect of rapid data discovery was universally welcomed, doubt was expressed as to the value of superficial data discovery tools that provide little information on distributions and missingness. This information was seen as essential for preparing informed and targeted data access requests. From the landscape review and the user consultation it was concluded that there was value for research cohort data in a data model optimised for data discovery and selection.

Design considerations

Design criteria included semantic precision, an intuitive user experience, simplicity, and extensibility. To be responsive to the requirements of different DACs, data discovery and selection needed to be available at both grouped variable and individual variable levels. To support multi-modal analysis, variables derived from higher-order pre-processed data would be used to identify image derived phenotypes, genotypes, and polygenic risk scores. Machine readability was considered essential for automation, and interoperability between data models.

Build strategy

The build strategy was to use existing tools and actual cohort data wherever possible. There are several cohort catalogues providing cohort metadata and contact details (Integrative Analysis of Longitudinal Studies of Aging – IALSA (8), The EU Joint Programme Degenerative Disease – JPND (9), The Global Alzheimer's Association Interactive Network – GAAIN (10), European Medical Information Framework - EMIF-AD (11). Of these GAAIN also provides basic feasibility analysis, and EMIF-AD provides limited harmonised datasets. EMIF_AD and the Alzheimer's Disease Data Initiative Work Bench – ADWB (12) provide facilities for federated analyses. However, none of these approaches uses a common data model. An approach that was more relevant is that of the Maelstrom Catalogue (13). This proposed a four tier data structure moving from data domains to variables. Whilst the organisation of the domains was not broadly generalisable, the basic structure enabled data discovery and selection at levels of detail, suitable to meet the requirements of most DACs, and appeared convenient to users.

The model was developed using data from the Airwave (Airwave Monitoring Study) (14), ELSA (The English Longitudinal Study of Ageing) (15), Generation Scotland (16), ICICLE-PD (The Incidence of Cognitive Impairment in Cohorts with Longitudinal Evaluation-PD) (17), and UK Biobank (18) cohorts. These studies provided a breadth of data by which the feasibility of developing a comprehensive and yet user-friendly data model could be judged. The model was developed iteratively, being expanded and revised for consistency cohort by cohort. The model was then used to fully curate all the data available to DPUK from the Airwave, ELSA (Derived), ICICLE-PD and Generation Scotland cohorts. It was found that these datasets could be organised into a relatively small number of 'themes' describing common usage and/or data modality (Fig. 1). For example, 'Cognitive Status' (theme 7) describes a user defined area of

interest whilst the ‘Imaging’ (theme 13) describes a technology driven data modality. These themes provided the basic organising principle of the data model.

Results

Objective

To design and implement a data model (C-Surv) for the discovery and selection of research cohort data using neurodegeneration as a use case.

Ontological design

The design adopted is a simple four level taxonomy intended to capture the breadth of data typically collected in research cohorts. This tiered structure supports grouped and individual variable selection. Class membership and naming at levels one to three were pragmatic decisions based on Data Portal user behaviour and the desire to maintain a four level structure for tool development purposes. Level four described the data object i.e. the measured variable. At this level naming was designed to uniquely identify the variable in the context of a longitudinal study.

Data model structure

C-Surv uses a four level acyclic structure comprising 18 data themes (level 1) leading to >130 data ‘domains’ (level 2), >500 data ‘families’ (level 3) and then to a growing number of data ‘objects’ (level 4). Typically data objects are variable level observations, or in the case of complex measures, such as psychometric tests, test scores (Figure 1). To the extent that evidence was available from DPUK access requests, the organisation of each level reflected the types of variable requests that are more frequently made. For example, typically a request would be made for all processing speed variables, rather than just choice reaction time, and so processing speed was used as a domain category.

Variable Naming

Key to utility is an informative ‘object’ (variable) name. Objects are defined pragmatically as the level of measurement used in most analyses. The object name is a complex proposition with 5 elements comprising cohort, data category, measurement, serialisation (repeated measurement within a single data capture period), and study wave (repeated measurement between data capture episodes). These elements are considered to be the minimum required to uniquely and conveniently identify an object in dataspace. An example object name is given below:

GEN04_PAINCHESTEVR_0_1

The cohort is identified using a three-digit alphabetic character (GEN for Generation Scotland), and data category by a two-digit numeric character (04 for medical history). The measurement is described by an alphanumeric abbreviation (PAINCHESTEVR for: 'Do you ever get pain or discomfort in your chest?'). This is followed by an integer representing the location of the variable within a sequence of repeat measurements within a study wave (_0 indicates there were no repeat measurements). Finally, an integer suffix indicates study wave (_1 for recruitment, _2 for first follow-up, etc.).

For survey data the measurement abbreviation is limited to 12 characters. For imaging, omics, and device data it is limited to 17 characters. Where questionnaire item level measurement is relevant, q# is added to the object name. For example, GEN06_SPQq1_0_1 is an item from Generation Scotland (GEN) within the Psychological Status category (category 10), from the Schizotypal Personality Questionnaire (SPQ), question 1 (q1), administered with no repeat measurement in wave 1.

Abbreviations are selected to reflect the meaning of the full variable name used in data capture. They are upper-case, syllable based, using word fragments as abbreviations and numeric characters to facilitate easy interpretation. Consistency of abbreviations is maintained where possible. Constants are lower case for example, just as 'q' is used to represent question (or item), 'r' is used to represent range and 'd' is used to represent a decimal point. For example, AVG08H00r08H59 is an item from accelerometry data (average acceleration between 08h00 and 08h59).

Value labelling conventions

To provide correspondence between native data (that transferred to the Data Portal by data controllers) and curated data, native data value labels are retained. However, for widely used measures, value labels are standardised using common conventions. For example, missing is scored '.' following the Stata (19) convention, gender is scored '2' for female and '1' for male. For several widely used measures imperial scaling is converted to metric. For example, height is recorded in centimetres and weight in kilograms.

Results

Use case: Cohort Explorer

To explore the potential for C-Surv to support data discovery, it was used to develop the Cohort Explorer feasibility tool. Cohort Explorer allows users to establish the number of participants with data, according to variable, across cohorts, prior to making a data access request. It enables users to avoid requesting combinations of variables that collectively have high levels of missingness.

Assessing feasibility in a multi-cohort environment requires the harmonisation of data across datasets. Harmonisation (the equivalence of values and/or distributions for variables across datasets) goes beyond the conventions of a common data model. However, a common data model does provide context

for evaluating the suitability of variables for harmonisation. To test this C-Surv was applied to 11 collaborating DPUK cohorts (n=123,554) and the results used to generate a 31 variable harmonised dataset (Table 1). The selection of variables reflects the frequency of variables requested in dementia focussed DPUK data access applications. These variables represent a wide range of modalities and formats including imaging, genetic, and survey data. The C-Surv structure was able to accommodate all the data types and formats found in the native cohort data. C-Surv also provided a structured overview of scientific activity. For example, the omission of life functionality, and physical and social environment variables, suggests either relatively little attention is paid to these areas, or that these data are sparse. The harmonised dataset was used to populate Cohort Explorer.

Cohort Explorer can be found at Cohort Explorer - DPUK Data Portal (dementiasplatform.uk). As the tool uses individual-level cohort data it requires a DPUK account to access. This can be obtained upon application to Register - DPUK Data Portal (dementiasplatform.uk). The tool provides an interactive dashboard allowing users to select cohorts, variables and value ranges of interest. For example, of the 123,554 members of the 11 cohorts, 57,499 are aged 50+ and of these 21,867 are lifetime non-smokers (Figure 2). However if APOE4 status (homozygous or heterozygous) is added the numbers drop to 1,666. This is critical information when planning an analysis.

Discussion

Following a literature review and user consultation, the C-Surv data model was developed to investigate the feasibility and utility of a common data model for research cohort data. The data model, optimised for data discovery and selection, was used to develop the Cohort Explorer analysis feasibility tool.

The test of the C-Surv model was not exhaustive. However, the development process involved covering all the variables available to DPUK in 7 cohort datasets. These included survey, imaging, genetic, and environmental data. Although not tested here, it is anticipated that C-Surv will be extensible to device and linkage data. That a four-level nested hierarchy can be applied to diverse datatypes is unsurprising. The challenge is to apply the hierarchy in a way that is useful. The combination of user-based and technology-based groupings, including the coding of medical history using ICD-11, was pragmatic. Nosological significance was not intended.

Manual curation is labour intensive and prone to error. Preliminary attempts at automation, using supervised machine learning, have achieved correct curation of around 70% of variables. Although improved performance may be anticipated, it is unlikely that 100% accuracy can be achieved reliably. Automation is particularly important given the dynamic nature of research cohort data. Many cohorts are active and further data collection may be anticipated. For these datasets, to have efficient and accurate methods for updating and versioning is important.

Inability to achieve full automation raises the issue of quality control. Unregulated use of a common data model risks undermining its scientific value as its standards are unlikely to be applied consistently. Maintaining quality control is an important issue and systems guaranteeing the provenance of curated

datasets are required for the confidence of the community to be retained. It is likely that a system of accredited curation centres built around a single or small number of data standards is preferable to laissez-faire.

C-Surv is optimised for data discovery and selection. This is in contrast to models designed to establish common metadata standards for genetic research cohorts, such as the Genomics Cohorts Knowledge Ontology - GECKO model of the CINECA (Common Infrastructure for National Cohorts in Europe, Canada, and Africa) (20) consortium, or models designed to follow the flow of data collection such as that used in UK Biobank. For data collection, preliminary work using DPUK cohorts suggests a design principle organised around specific data modalities is more appropriate than the four level taxonomy used in C-Surv.

The utility of C-Surv to support a cross-cohort data discovery tool was demonstrated. However, the utility of Cohort Explorer was constrained by the dashboard being limited to the visualisation of around 30 variables. Nevertheless, it does identify datasets that are unlikely to be informative for those variables. Re-designing the dashboard to increase the number of variables would improve the value of the tool. Cohort Explorer is also limited to identifying the amount of data available according to cohort and variable combination. Whilst this is important, the addition of a power calculator and some preliminary regression analytic capability would add value. A further limitation is the methods used for harmonisation. Data harmonisation is implicitly purpose-specific and may vary subtly according to hypothesis and analytic strategy. However, for the purpose of data discovery, the relatively simple strategies used here of standardising scale values and, where appropriate, transforming to standardised distributions are likely to be sufficient. Although a work in progress, Cohort Explorer has demonstrated that a persistent common data model would be useful for developing cross-cohort data discovery tools.

Conclusions

Here we demonstrate the feasibility and utility of applying a common data model to research cohort data. However, this is also attempt to stimulate and contribute to a wider debate on how to provide wide access to research-ready data at scale and speed. Building and maturing a data model is a collaborative and iterative process. It requires the engagement of the user community, particularly those in lower resource settings, for benefit to be widely realised. DPUK is collaborating with Dementias Platform Australia (DPAU) and The Alzheimer's Disease Data Initiative (ADDI) to apply C-Surv to international datasets. DPUK welcomes further collaboration in the development of tools and technologies that enable access to research-ready data at scale and speed.

Declarations

Funding

This work was supported by the UK Research and Innovation Medical Research Council [MR/L023784/1 and MR/L023784/2]

Conflicts of interest/Competing interests

The authors declare that they have no conflict of interest

Availability of data and material

Not applicable

Code availability

Not applicable

Authors' contributions

All authors contributed to the conception, creation and development of all the themes of the DPUK itself, including this Data Portal. Material preparation, by John Gallacher and Sarah Bauermeister. The first draft of the manuscript was written by John Gallacher and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Not applicable

Code availability

Not applicable

References

1. Bauermeister S, Orton C, Thompson S, Barker RA, Bauermeister JR, Ben-Shlomo Y, et al. The Dementias Platform UK (DPUK) Data Portal. Eur J Epidemiol. 2020;35(6):601–11.

2. Mohamed Yusoff A, Tan TK, Hari R, Koepfli KP, Wee WY, Antunes A, et al. De novo sequencing, assembly and analysis of eight different transcriptomes from the Malayan pangolin. Sci Rep. 2016;6:28199.
3. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008;83(5):610–5.
4. Huser V, Sastry C, Breymaier M, Idriss A, Cimino JJJobi. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). 2015;57:88–99.
5. SNOMED. Systematized Nomenclature of Medicine – Clinical Terms [09/09/2021]. Available from: https://www.snomed.org/
6. OHDSI. Observational Health Data Sciences and Informatics [09/09/2021]. Available from: https://www.ohdsi.org/
7. FIHR. Fast Healthcare Interoperability Resources: NHS Digital; [09/09/2021]. Available from: https://fhir.nhs.uk/
8. IALSA. Integrative Analysis of Longitudinal Studies of Aging [09/09/2021]. Available from: https://www.ialsa.org/
9. JPND. The EU Joint Programme - Neurodegenerative Disease Research [09/09/2021]. Available from: https://www.neurodegenerationresearch.eu/
10. GAAIN. The Global Alzheimer's Association Interactive Network [09/09/2021]. Available from: http://gaain.org/
11. EMIF-AD. The European Medical Information Framework [09/09/2021]. Available from: http://www.emif.eu/
12. ADWB. Alzheimer's Disease Workbench: Alzheimer's Disease Data Intitiative (ADDI); [09/09/2021]. Available from: https://www.alzheimersdata.org/ad-workbench
13. Maelstrom. Maelstrom Catalogue [09/09/2021]. Available from: https://www.maelstrom-research.org/
14. Elliott P, Vergnaud AC, Singh D, Neasham D, Spear J, Heard A. The Airwave Health Monitoring Study of police officers and staff in Great Britain: rationale, design and methods. Environ Res. 2014;134:280–5.
15. Steptoe A, Breeze E, Banks J, Nazroo J. Cohort profile: the English longitudinal study of ageing. Int J Epidemiol. 2013;42(6):1640–8.
16. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. Int J Epidemiol. 2013;42(3):689–700.
17. Yarnall AJ, Breen DP, Duncan GW, Khoo TK, Coleman SY, Firbank MJ, et al. Characterizing mild cognitive impairment in incident Parkinson disease: the ICICLE-PD study. Neurology. 2014;82(4):308–16.

18. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779.
19. StataCorp. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC; 2019.
20. CINECA. Common Infrastructure fro National Cohorts in Europe, Canada, and Africa [09/09/2021]. Available from: https://www.cineca-project.eu/

Tables

Theme	Domain	Family	Object label
2. Sociodemographic	Demographics	Age	Year of birth
			Age
		Gender	Sex
	Education	Educational experience	Years education
4. Medical history	Nervous system	Chronic neurological disorders	Dementia Diagnosis
			PD Diagnosis
		Episodic disorders	Other neurological disorders
	Circulatory	Cardiovascular disorders	CVD
			stroke
	Cognitive status	Self-report	Subjective Memory Complaint
	Self-report medical history	General health	MCI
	Self-report medical history	Medications self-report	Prescription medications
5. Family disease history	Nervous system ICD11	Family member	Family history dementia
	Nervous system ICD11	Family member	Family history PD
	Circulatory	Family member	Family history stroke
6. Psychological status	Self-report mental health	Depression	Depression scale
		Trauma	PTSD
7. Cognitive status	Memory	Short term/working memory	Immediate recall
		Short term/working memory	Delayed recall
	Problem solving	Planning	Executive function task
	Processing speed	Task response time	Reaction time task
	Self-report	memory	Subjective memory complaint
8. Lifestyle behaviour	Substance use	Alcohol	Alcohol units/wk
		tobacco	Smoking status

12. Physical examination	Musculo-skeletal	Structural	BMI
	Circulatory	Cardiovascular	BP systolic
			BP Diastolic
13. Imaging	Brain	MRI	MRI images
16. Bio-sample assays	Blood	Haematology	CRP
	CSF	Proteins	CSF Tau
17. Molecular	Genomics	SNP	APOE

Table 1: Harmonised variables available in Cohort Explorer

Figures

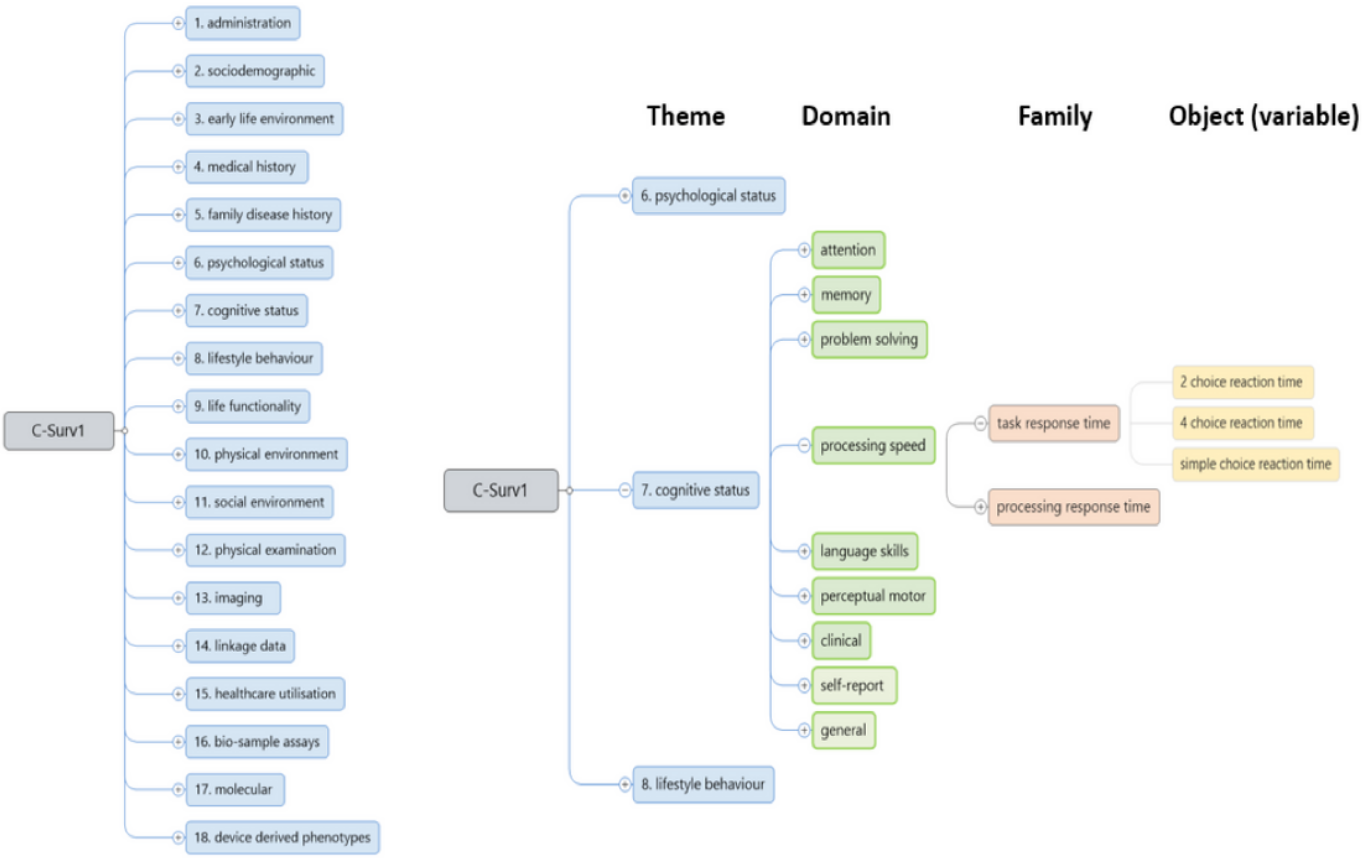

Figure 1

Schematic of the C-Surv1 data model

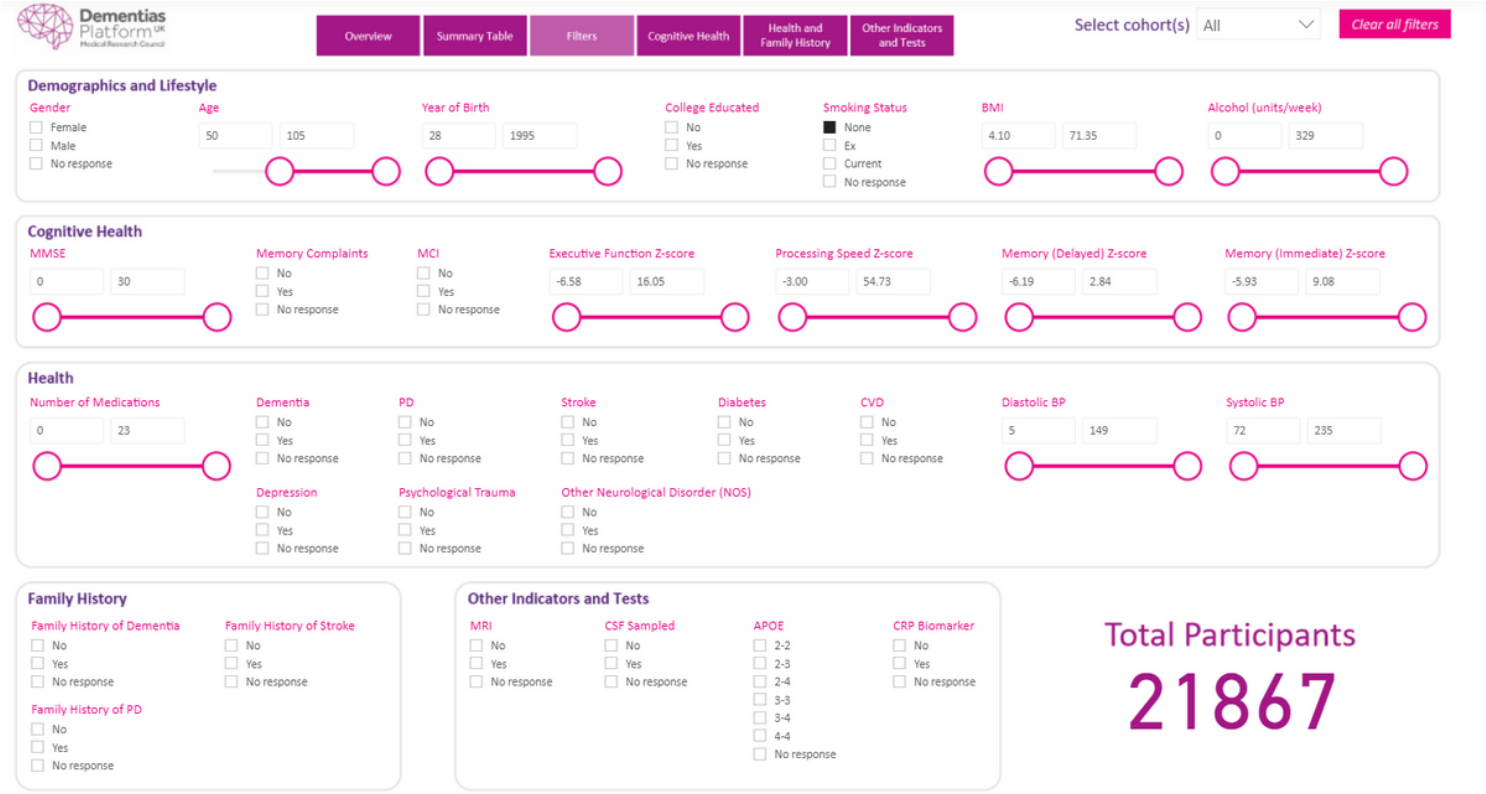

Figure 2

Cohort Explorer screen shot