bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data Resource Profile:

The Dementias Platform UK (DPUK) Data Portal

Bauermeister S[1†], Orton C[2†], Thompson S[2†], Barker R A[3], Bauermeister J R[1], Ben-Shlomo Y[4], Brayne C[5], Burn D[6], Campbell A[7], Calvin C[1], Chandran S[8], Chaturvedi N[9], Chene G[10], Chessell I P[11], Corbett A[12], Davis D H J[9], Denis M[13], Dufouil C[10], Elliott P[14], Fox N[15], Hill D[16], Hofer SM[17], Hu M[18], Jindra C[1], Kee F[19], Kim C H[1], Kim C[20], Kivimaki M[21], Koychev I[1], Kwan J S K[22], Lawson R A[23], Leroi I[24], Linden G J[19], Love S[25], Lovestone S[1,26], Lyons R A[2], Mackay C[1], Matthews P M[27], McGuiness B[19], Middleton L[14], Moody C[28], Moore K[15], Na D L[29], O'Brien J T[30], Ourselin S[31], Paranjothy S[32], Park K S[33], Porteous D J[34], Richards M[9], Ritchie C W[8], Rohrer J D[15], Rossor M N[15], Rowe J B[3], Scahill R[15], Schnier C[35], Schott J M[15], Seo S W[29], South M[1], Steptoe A[36], Tabrizi S J[15], Tales A[37], Thomas A J[23], Tillin T[38], Timpson N J[4], Toga A W[39], Visser PJ[40], Wade-Martins R[41], Wilkinson T[35], Williams J[42], Wong A[9], Gallacher J E[1*]

Abstract: Not required

Words: 2 897

References: 67

Tables: 1

Figures: 3

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Author Affiliations

1. Department of Psychiatry, University of Oxford
2. Swansea University Medical School, Swansea University
3. Department of Clinical Neurosciences, University of Cambridge
4. Population Health Sciences, University of Bristol
5. Department of Public Health, University of Cambridge
6. Faculty of Medical Sciences, Newcastle University
7. Department of Medical Genetics, University of Edinburgh
8. Centre for Clinical Brain Sciences, University of Edinburgh
9. MRC Unit for Lifelong Health and Ageing, UCL
10. Bordeaux Population Health, Université Bordeaux
11. Neuroscience, BioPharma R&D, AstraZeneca, Cambridge
12. College of Medicine and Health, University of Exeter
13. Oxford Academic Health Science Network, University of Oxford
14. School of Public Health, Imperial College
15. Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology
16. King's College London, London
17. Department of Psychology, University of Victoria
18. Nuffield Department of Clinical Medicine, University of Oxford
19. Centre for Public Health, Queen's University Belfast
20. Department of Preventive Medicine, Yonsei University College of Medicine, Seoul
21. Institute of Epidemiology and Health, University College London
22. Imperial College Health Care NHS Trust
23. Institute of Neurology, Newcastle University
24. Global Brain Health Institute, Trinity College Dublin
25. Institute of Clinical Neuroscience, University of Bristol

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

26. Neuroscience External Innovation, Johnson & Johnson Innovation
27. Division of Brain Sciences and UK Dementia Research Institute, Imperial College
28. Medical Research Council
29. Department of Neurology, Samsung Medical Center, Sungkyunkwan University School of Medicine
30. Department of Psychiatry, University of Cambridge
31. School of Biomedical Engineering & Imaging Sciences, King's College London
32. School of Medicine, Cardiff University
33. Institute of Health Science, Gyeongsang National University
34. Institute of Genetics and Molecular Medicine, University of Edinburgh
35. Usher Institute of Population Health Sciences and Informatics, University of Edinburgh
36. Department of Behavioural Science and Health, UCL
37. Centre for Innovative Ageing, Swansea University
38. UCL Institute for Cardiovascular Science
39. Laboratory of Neural Imaging, UCLA
40. VU University Medical Centre, Maastricht University
41. Department of Physiology, Anatomy and Genetics, University of Oxford
42. Institute of Psychological Medicine and Clinical Neurosciences, and UK Dementia Research Institute, Cardiff University

*Corresponding author

[†] Joint lead authors

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background

Dementias Platform UK (DPUK) is a £53M public-private-partnership established by the Medical Research Council (MRC) to facilitate experimental medicine programmes that bridge the evidence gap between basic mechanistic research and large-scale trials. DPUK does this in three ways: 1) by providing access to individual and aggregate level data from multiple cohort studies for hypothesis generation and testing, 2) a register of highly characterised risk-stratified volunteers consented for re-contact (which is undergoing recruitment), and 3) a programme of academic and industry based experimental studies. Here we describe the DPUK Data Portal which facilitates access to cohort data (including one e-cohort) data for 3 461 244 individuals in 35 cohorts. Meta data are available for a further 12 cohorts.

There are several arguments for multi-cohort-focused data repositories including: 1) as research questions focus on smaller effect sizes access is required to data at-scale to achieve statistical purchase, 2) as emerging research questions become more complex access to diverse multi-modal data is needed for rigorous hypothesis testing, 3) as scientific rigour increases there is growing recognition of the value of triangulation and replication using independent data, 4) as cohort datasets increase in size the cohort-by-cohort transfer of large datasets is decreasingly feasible, 5) as cohort datasets become more complex the mastering of bespoke data models for survey, omics (genomics, proteomics and metabolomics), imaging and device data becomes burdensome, 6) as cohort datasets become more sensitive the non-auditable use of data is decreasingly acceptable. Whilst these issues can be addressed individually, the Data Portal provides an integrated solution.

Data resource basics

The DPUK Data Portal (https://portal.dementiasplatform.uk/) [1] is a collaboration between DPUK and a growing number of cohort research teams who wish to make their data globally accessible (Table 1).

Benefits for cohort research teams

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The Data Portal supports the work of cohort research teams through data curation, access management, and cohort enhancement. The Data Portal contributes to the widely accepted FAIR principles (Findability, Accessibility, Interoperability and Reusability) to improve the infrastructure supporting the reuse of data [2]. DPUK also facilitates the legal engagement necessary to facilitate data transfer into the Data Portal on behalf of accessing researchers, by ensuring robust contractual arrangements, in the form of the DPUK Data Deposit Agreement, are in place as an overarching mechanism for data governance and use [3].

The Data Portal operates within the UKSeRP environment according to ISO 27001 [4] as a data processor according to the UK Data Protection Act 2018 [5] and EU General Data Protection Regulation 2016 [6]. Data may be accessed remotely for in-situ analyses but not downloaded to third-party sites. Data-use approval remains with the cohort research teams who retain control over data access. Preparing datasets for third-party researchers and providing suitable documentation is resource intensive. The Data Portal reduces this burden through the management of access requests on behalf of cohort research teams, use of a common data model, and the development of standard documentation. For data stored within the Data Portal, the need for repeated data transfer is eliminated.

The Data Portal enables cohort enhancement through web-based procedures that can be 'branded' for each cohort. This utility is suitable for collecting consent, questionnaire and cognitive performance data. Whilst the comparability of data collected through different modalities is an empirical question, there is growing evidence on the acceptability and validity of remote recruitment and assessment procedures [7]. The UKSeRP environment has been specifically designed for use with linked electronic health records and is a suitable environment for the onward sharing of linked data.

Benefits for researchers

For researchers the Data Portal enables the pursuit of ideas from any location with suitable connectivity. It has three core utilities: data discovery, access, and analysis. A tiered data discovery

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

pathway begins with the Cohort Matrix [8] which provides a high-level comparison of data availability for each cohort. The Cohort Directory [9] enables detailed exploration across cohorts using a range of metadata categories (Figure 1).

For data access, the Data Portal provides a single point of contact for multiple cohorts. The application form is a synthesis of key issues that are addressed by most if not all the individual cohort access management procedures, comprising public interest, potential for subject identifiability, scientific rationale, appropriate analysis plan, and conflict of scientific interest. These issues are specifically identified as fail-fast criteria to enable cohort data access management panels to more easily evaluate an application. Of the 47 cohorts that have provided data or metadata to the Data Portal, 26 have adopted the DPUK electronic application pro-forma (Table 1). A further 14 require their own access form to be completed alongside the DPUK process, and seven require their own process to be used exclusively.

For data analysis, once approval has been granted by a cohort research team and a data access agreement has been completed, data are made available within the secure analysis area of the Data Portal for in-situ analysis. The analysis area includes the use of several widely used general statistical packages (R, STATA, SPSS, SAS, Matlab, Python). Specialist software can be made available on request and bespoke software can be uploaded upon approval. Researchers are provided with a personal virtual desktop infrastructure which requires two-factor authentication to access. The standard desktop specification is optimised for survey epidemiologic survey data and includes 8GB RAM and four CPUs which are sufficient for most analyses. Bespoke desktop configurations can be requested for computationally intensive operations.

Multi-modal analysis is facilitated by optimising a virtual desktop infrastructure for survey, imaging, or omics analysis, in terms of capacity and tooling, and then combining results within a project-specific common data folder. For example, image-derived phenotypes may be generated within the environment using specialist software, transferred to a common folder, and then integrated with

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

survey data. For device data (e.g. data from smart phones, accelerometers, portable cardiac monitors), the current solution is for data to be processed into clinically meaningful measures before upload to the Data Portal, and to be accessible via the standard desktop. Linked data are also available where governance permissions allow. Results can be exported for dissemination purposes. All exports must be requested, screened for non-identifiability, and approved prior to release. Microsoft Office is provided for preparing reports for export.

The Data Portal allows joint use of data within research consortia. Although the virtual desktop is the researcher's own personal virtual laboratory, for distributed research groups and for consortia, the Data Portal can be used to hold a core dataset which can then be accessed by researchers in multiple locations without risk to the integrity of the core. This network of virtual desktops is flexible and can be configured in terms of access rights and capacity according to the requirements of the consortium.

The Data Journey

The data journey begins with upload to the Data Portal (Figure 2). Datasets and data dictionaries are received from cohorts on an 'as-is' basis along with other supporting documentation. Data are then curated to a common data model. The data model simplifies the analytic challenge of working across multiple datasets by providing a standard structure, variable naming and value labelling conventions. It is optimised for the analysis of 'flat-file' observational data and allows sorting by cohort, data category, repeat measurement to assess measurement error (array), and serial measurement to detect change (wave). Higher order data must be pre-processed prior to curation. Other data models, CDISC [10], OMOP [11], or HPO [12] involve structural complexity that is rarely relevant to cohort based analyses. Data curation is resource intensive and is ongoing. To enable the feasibility of analyses to be assessed prior to a data application being made, a set of 20 variables relevant to dementia have been harmonised across cohorts (Figure 2). Researchers may request access to either native or curated data.

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data collected

Data are available for 35 cohorts (Table 1). Of these, 22 (n=1 399 082) have uploaded full or partial datasets and 13 (n=2 062 162) will upload on a per project basis. A further 12 cohorts (n=52 361) have begun the process of making data available and have provided metadata.

The data are diverse. Clinical cohort studies include familial disease cohorts [13][14]), disease focused patient cohorts [15][16]), ageing focused population cohorts [17][18][19][20]), re-purposed cardiovascular cohorts [21][22], birth cohorts [23][24], repurposed cancer cohorts [25] [26]), and disease agnostic cohorts [27][28][29]) In terms of real world evidence the Data Portal provides access to an e-cohort (SAIL-DeC) covering health records for 1.2m individuals including 130k dementia cases [30]. Data availability varies according to cohort but includes epidemiologic survey, imaging, genetics, and linked administrative data. The XNAT [31] imaging platform is used to receive and process DICOM and NIfTI files. For genetics, variant call format and allele frequency data may be uploaded. Although not a cohort, the Data Portal provides a link to the UK-CRIS Network for natural language processing of 2.5m UK mental health records [32]. Overall these cohorts and data modalities represent an unusually complex data environment suitable for machine learning as well as hypothesis driven analyses.

Looking ahead, brain bank digital data, omics, devices, and environmental data are developing areas for DPUK. Cohorts remain the best source of post mortem scientifically informative brain tissue due to the wealth of background data that are available. Cohort derived brain tissue can also represent a range of disease stages for any particular outcome and a range of pathologies. DPUK is working with the UK Brain Banking Network to provide a central database linking brain donation to cohorts. Devices are an increasingly important source of data. To explore the collection of device data, DPUK provides a cost-neutral pipeline for data capture, processing and storage for collaborating cohorts. Specific omics pipelines can be made available on request.

Data resource use

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

From its public launch in November 2017 to end of December 2018, 81 data access requests have been received involving 149 applicants. The 81 requests span 41 institutions (34 academic, five commercial, two government) in nine countries and 51 requests involve multiple cohorts. Of the 81 data requests, 11 have been declined and seven withdrawn. The remainder have either been approved (n=39) or are under review (n=24). Our target response time for a decision on applications is 28 days. Currently the median response time is 25 days (mean=44 days).

Project proposals are diverse with applications coming from multi-disciplinary applicants involving multiple institutions. Projects include Mendelian randomisation, imaging, psychometric, and machine learning studies alongside risk stratification studies models for dementia prediction and diagnosis. Others are less dementia specific such as: trajectories of longitudinal assessment of comorbidities between mental health, hormonal indicators and cognitive change; the impact of child adversity on adult outcomes; the longitudinal tracking and determinants of well-being and cognitive performance; and successful cognitive ageing in 90+ year olds.

To facilitate innovative and exploratory analyses, without comprising data security or cohort governance principles, the Data Portal can be configured as a 'sandbox' environment upon request. An example of this was the hosting of a datathon at the Alan Turing Institute utilising the Deep and Frequent Phenotyping pilot study data [33]. These multi-modal data include Magneto-encephalography (MEG), Positron Emission Tomography (PET), structural and functional MRI, ophthalmology, gait, and serial cognitive and clinical assessment. The Data Portal was used to host 40 data scientists over a three-day multidisciplinary workshop, during which traditional regression and machine learning procedures were used to interrogate the data within the virtual desktop interface.

Strengths and weaknesses

A strength of the Data Portal is that it obviates the need for repeated data transfer. Other strengths include providing a single point of access for multiple cohort datasets, streamlined and standard

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

access procedures, a common data model, a secure analysis environment, and a process which is fully auditable from data upload to the results export. The Data Portal is optimised and populated for dementia. However, it is a generic solution to the problem of analysing cohort data that can be used for any health outcome for which data are available.

The Data Portal provides access to real world evidence to inform experimental medicine and other clinical studies. Observational datasets can be used to inform emerging hypotheses, scrutinise genetic instruments in a Mendelian randomisation framework, and validate experimental findings. This has particular relevance for biomarker development and drug discovery. A secure data repository also strengthens the case for national data linkage agreements. In the UK for example, DPUK is working closely with Health Data Research UK [34] to establish cohort linkage to electronic health records on behalf of all collaborating UK cohorts.

The Data Portal provides a solution for data access beyond the UK. The Data Portal is not geographically restricted, and data are available on the Data Portal from an increasing number of international cohorts. However, national or regional repositories may be more acceptable to funders. To increase the overall size of the available data corpus, collaboration is underway with Dementias Platform Korea, EMIF AD, and the Ontario Brain Institute to establish a fully interoperable environment for European, Korean, and Canadian data.

Challenges include meeting the data access needs and expectations of diverse scientific communities; epidemiologic, imaging, genetics, and data science communities each have different conventions over what constitutes an appropriate data request. This problem is illustrated in that 11 out of 81 (14%) Data Portal applications were declined due to insufficient detail. Parallel processes of expectations coalescing across disciplines, and educating applicants on how to develop high quality proposals, will assist all stakeholders to simplify and standardise procedures.

Time is required for cohort research teams to adjust to the opportunity provided by a data platform although for many cohort research teams centralised data access management provides immediate

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

advantage. For researchers, a challenge is the discipline of accessing data remotely rather than locally. However, as datasets become increasingly valuable and sensitive, remote access via secure repositories will likely become accepted routine practice. A more fundamental limitation is that the data repository model is not appropriate for all datasets. Clearly there is a need for a mixed model and the Data Portal offers both centralised and distributed analyses.

In the dementia space, other data platforms are available. The JPND Global Cohort Directory [35] provides contact details for 175 cohorts (n=3 586 109) whilst the IALSA Network [36] provides details for 110 cohorts (n=1 485 410). More sophisticated and convenient data discovery tools are provided by GAAIN [37] with 47 cohorts (n=480 020). GAAIN also offers centralised processing for selected datasets. EMIF-AD [38] offers a comprehensive data harmonisation programme for a selection of their 60 catalogued cohorts (n=135 959) and 18 electronic health records datasets (n=65M). For selected datasets EMIF-AD provides centralised cohort data processing facilities through the TranSMART platform [39].

Data resource access

The researcher journey begins with the data discovery tools (Figure 2). The Cohort Matrix and Directory are accessible to registered bona fide researchers. Registration requires having either an academic email address or an industry email from a certified company, and in the case of PhD or Master's students, the requirement is to have a senior researcher as study lead. Registered researchers can complete a data access application form and submit it for review by the data guardians of the datasets requested. Upon approval, completion by the applicant of a data access agreement is required prior to access being granted. DPUK undertakes to send this to the applicant's legal representative within two working days. Upon receipt of a completed data access agreement, data access is granted within five working days [40].

Two-factor authentication is required to enable access to approved datasets. This involves the provision of a username with password creation, and an authentication code generated by an app on

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a mobile device of the applicant's choosing. The data may then be accessed for analysis on the Data Portal.

Tables, graphs and scripts for export are submitted to the data export panel for approval. Manuscripts may be prepared in the Data Portal so that collaborators who are registered users may contribute without the need for manuscript download. A facility for import is also available, enabling researchers to upload scripts and additional datasets from outside the Data Portal to reside within their approved DPUK datasets.

Publications arising from use of the Data Portal are required to conform to the DPUK publications policy [41]. The intention of this policy is to acknowledge the importance of the 'team science' underlying the opportunity provided to researchers. Not only is the researcher dependent on decades of generosity and work from cohort participants and cohort research teams respectively, but also upon the data scientists who deliver the provenance of the infrastructure, and the funders. A goal of the Data Portal is to reduce data access costs sufficiently that they may be borne centrally; effectively making data free at point of access. By undertaking data storage, curation and access management on behalf of cohorts the need for access fees is ameliorated and for most cohorts there is no access fee.

The DPUK Data Portal was established by MRC to accelerate the development of new treatments for dementia by using cohort data to inform experimental medicine. It is recognition of the unique value of cohort data and a contribution to the wider debate on how best to support cohort studies and facilitate their use within the wider research environment. By streamlining procedures for cohort research teams, increasing data accessibility for researchers, and reducing costs and adding value for funders, the Data Portal is also an investment in the future of cohorts generally.

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Profile in a nutshell

- The DPUK Data Portal was established to increase the realised scientific value of cohort data by enabling remote access to multi-modal data from multiple independent datasets
- Launched in 2017, the Data Portal enables access to individual level data for 3m participants from 35 population and clinical cohorts
- Data types vary according to cohort and include survey, imaging, genetic, device and linked outcome data

All projects are by default collaborations with the cohort research teams which have generated the data and application for access can be made through the Data Portal

https://portal.dementiasplatform.uk/

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

DPUK would like to express gratitude to:

Cohort members and their research teams for generously making data available

EMIF-AD for providing access to their data catalogue and supporting software

Professor Ian Deary and Dr Declan Jones for their contribution to this paper from their support in the DPUK Executive Team.

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

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. *Dementias Platform UK*. Available from: https://portal.dementiasplatform.uk/.
2. Wilkinson M D , Dumontier M, Aalbersberg I J, et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2016. **3**: p. 160018.
3. *Dementias Platform UK Data Deposit Agreement*. Available from: https://portal.dementiasplatform.uk/Content/Docs/data_deposit__agreement1_SAMPLE.pdf
4. *Farr Institute*. Available from: http://farrinstitute.org/public-engagement-involvement/100-ways-of-using-data-to-make-lives-better/case-study/the-creation-of-a-uk-secure-eresearch-platform-ukserp-to-support-population-data-research.
5. *Data Protection Act.* Available from: https://www.gov.uk/data-protection.
6. *European Union General Data Protection Regulation (GDPR)*. Available from: https://eugdpr.org/.
7. Gallacher J, Collins R, Elliott P, et al., *A platform for the remote conduct of gene-environment interaction studies.* PLoS One, 2013. **8**(1): p. e54331.
8. *DPUK Cohort Matrix*. Available from: https://portal.dementiasplatform.uk/CohortMatrix.
9. *DPUK Cohort Directory*. Available from: https://portal.dementiasplatform.uk/CohortDirectory.
10. *Clinical Data Interchange Standards Consortium*. Available from: https://www.cdisc.org/.
11. *Observational Health Data Sciences and Informatics*. Available from: https://www.ohdsi.org/data-standardization/the-common-data-model/.
12. *Human Phenotype Ontology*. Available from: https://hpo.jax.org/app/.
13. Bateman R J, Xiong C, Benzinger T L, et al., *Clinical and biomarker changes in dominantly inherited Alzheimer's disease.* N Engl J Med, 2012. **367**(9): p. 795-804.
14. *The Genetic Frontotemporal dementia Initiative*. Available from: http://genfi.org.uk/.

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

15. Lawton M, Baig F, Rolinski M, et al., *Parkinson's Disease Subtypes in the Oxford Parkinson Disease Centre (OPDC) Discovery Cohort.* Journal of Parkinsons Disease, 2015. **5**(2): p. 269-279.

16. Kim S E, Woo S, Kim S W, et al., *A Nomogram for Predicting Amyloid PET Positivity in Amnestic Mild Cognitive Impairment.* J Alzheimers Dis, 2018. **66**(2): p. 681-691.

17. Steptoe A, Breeze E, Banks J, Nazroo J, *Cohort profile: the English longitudinal study of ageing.* Int J Epidemiol, 2013. **42**(6): p. 1640-8.

18. Brayne C, McCracken C, Matthews F E, Medical Research Council Coginitive Function and Ageing Study (CFAS), *Cohort profile: the Medical Research Council Cognitive Function and Ageing Study (CFAS).* Int J Epidemiol, 2006. **35**(5): p. 1140-5.

19. Matthews F E, Stephan B C, Robinson L, et al., *A two decade dementia incidence comparison from the Cognitive Function and Ageing Studies I and II.* Nat Commun, 2016. **7**: p. 11398.

20. Francis, P.T., H. Costello, and G.M. Hayes, *Brains for Dementia Research: Evolution in a Longitudinal Brain Donation Cohort to Maximize Current and Future Value.* J Alzheimers Dis, 2018. **66**(4): p. 1635-1644.

21. *Whitehall II study*. Available from: https://www.ucl.ac.uk/iehc/research/epidemiology-public-health/research/whitehallII.

22. Gallacher J, Ilubaera V, Ben-Shlomo Y, et al., *Auditory threshold, phonologic demand, and incident dementia.* Neurology, 2012. **79**(15): p. 1583-90.

23. Wadsworth M, Kuh D, Richards M, Hardy R, et al., *Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development).* Int J Epidemiol, 2006. **35**(1): p. 49-54.

24. Fraser A, Macdonald-Wallis C, Tilling K, et al., *Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort.* Int J Epidemiol, 2013. **42**(1): p. 97-110.

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

25. Elliott P, Vergnaud A C, 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**: p. 280-5.

26. Hayat S A, Luben R, Keevil V L, et al., *Cohort profile: A prospective cohort study of objective physical and cognitive capability and visual health in an ageing population of men and women in Norfolk (EPIC-Norfolk 3).* Int J Epidemiol, 2014. **43**(4): p. 1063-72.

27. Green J, Reeves G K, Floud S, et al., *Cohort Profile: the Million Women Study.* Int J Epidemiol, 2018.

28. Sudlow C, Gallacher J, Allen N, 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): p. e1001779.

29. *Healthwise Wales*. Available from: https://www.healthwisewales.gov.wales/.

30. Wilkinson T, Schnier C, Gallacher J, Lyons L, Sudlow C, *Creating a dementia research cohort using routinely collected healthcare data.* Alzheimers Dement, 2018. **14**(7): p. P1014.

31. *XNAT Imaging Platform.* Available from: https://www.xnat.org/.

32. *UK Clinical Record Interactive Search Network*. Available from: https://crisnetwork.co/.

33. Koychev I, Gunn R N, Firouzian A, et al., *PET Tau and Amyloid-beta Burden in Mild Alzheimer's Disease: Divergent Relationship with Age, Cognition, and Cerebrospinal Fluid Biomarkers.* J Alzheimers Dis, 2017. **60**(1): p. 283-293.

34. *Health Data Research UK*. Available from: https://www.hdruk.ac.uk/.

35. *EU Joint Programme - Neurodegenerative Disease Research Global Cohort Portal*. Available from: http://www.neurodegenerationresearch.eu/jpnd-global-cohort-portal/.

36. *Integrative Analysis of Longitudinal Studies of Aging*. Available from: http://www.ialsa.org/.

37. Toga A W, Neu S C, Bhatt P, Crawford K L, Ashish N, *The Global Alzheimer's Association Interactive Network.* Alzheimers Dement, 2016. **12**(1): p. 49-54.

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

38. Bos I, Vos S, Vandenberghe R, et al., *The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics.* Alzheimers Res Ther, 2018. **10**(1): p. 64.

39. Scheufele E, Aronzon D, Coopersmith R, et al., *tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform.* AMIA Jt Summits Transl Sci Proc, 2014. **2014**: p. 96-101.

40. *Dementias Platform UK Data Access Agreement*. Available from: https://portal.dementiasplatform.uk/Content/Docs/access_agreement1_SAMPLE.pdf.

41. *Dementias Platform UK Publications Policy*. Available from: https://www.dementiasplatform.uk/about/policies/dpuk-publications-policy.

42. *BRACE*. Available from: https://www.alzheimers-brace.org/.

43. Taylor J R, Williams N, Cusack R, et al., *The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample.* Neuroimage, 2017. **144**(Pt B): p. 262-269.

44. Foltynie T, Brayne C E, Robbins T W, Barker R A, *The cognitive ability of an incident cohort of Parkinson's patients in the UK. The CamPaIGN study.* Brain, 2004. **127**(Pt 3): p. 550-60.

45. *Cygnus project*. Available from: http://www.ecygnus.com.

46. *Environmental pollution-induced Neurological Effects study*. Available from: https://portal.dementiasplatform.uk/CohortDirectory/Item?fingerPrintID=74bf7bb33739285586427ecefdcf07f8.

47. Smith B H, Campbell A, Linksted P, 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): p. 689-700.

48. Hollingworth P, Sweet R, Sims R, et al., *Genome-wide association study of Alzheimer's disease with psychotic symptoms.* Mol Psychiatry, 2012. **17**(12): p. 1316-27.

49. Yarnall A J, Breen D P, Duncan G W, et al., *Characterizing mild cognitive impairment in incident Parkinson disease: the ICICLE-PD study*. Neurology, 2014. **82**(4): p. 308-16.

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

50. Bevan-Jones W R, Surendranathan A, Passamonti L, et al., *Neuroimaging of Inflammation in Memory and Related Other Disorders (NIMROD) study protocol: a deep phenotyping cohort study of the role of brain inflammation in dementia, depression and other neurological illnesses.* Bmj Open, 2017. **7**(1).

51. Evans J R, Cummins G, Breen D P, et al., *Comparative epidemiology of incident Parkinson's disease in Cambridgeshire, UK.* J Neurol Neurosurg Psychiatry, 2016. **87**(9): p. 1034-6.

52. Moss D J H, Pardiñas A F, Langbehn D, et al., *Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study.* Lancet Neurol, 2017. **16**(9): p. 701-711.

53. Deary I J, Gow A J, Taylor M D, et al., *The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond.* BMC Geriatr, 2007. **7**: p. 28.

54. *Protect study*. Available from: http://www.protectstudy.org.uk/.

55. Tillin T, Forouhi N G, McKeigue P M, Chaturvedi N, SABRE Study Group, *Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins.* Int J Epidemiol, 2012. **41**(1): p. 33-42.

56. *AMyloid imaging for Phenotyping LEwy body dementia*. Available from: http://www.ncl.ac.uk/caru/research/project/5055.

57. Larsen M E, Curry L, Mastellos N, Robb C, Car J, Middleton L T, *Development of the CHARIOT Research Register for the Prevention of Alzheimer's Dementia and Other Late Onset Neurodegenerative Diseases.* PLoS One, 2015. **10**(11): p. e0141806.

58. Davis D, Richardson S, Hornby J, et al., *The delirium and population health informatics cohort study protocol: ascertaining the determinants and outcomes from delirium in a whole population.* BMC Geriatr, 2018. **18**(1): p. 45.

59. *Exeter 10,000 study*. Available from: https://crf.exeter.ac.uk/web/content/extend.

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

60. *The University of Hong Kong Neurocognitive Disorder Cohort*. Available from: https://clinicaltrials.gov/ct2/show/NCT03275363.

61. *Identifying Predictors of dementia with Lewy bodies in People with Mild Cognitive Impairment*. Available from: https://portal.dementiasplatform.uk/CohortDirectory/Item?fingerPrintID=8ae6e8600397cdcf8b70c65942df0a63.

62. Dufouil C, Dubois B, Vellas B, et al., *Cognitive and imaging markers in non-demented subjects attending a memory clinic: study design and baseline findings of the MEMENTO cohort.* Alzheimers Res Ther, 2017. **9**(1): p. 67.

63. Yoo J I, Kim M J, Na J B, et al., *Relationship between endothelial function and skeletal muscle strength in community dwelling elderly women.* J Cachexia Sarcopenia Muscle, 2018. **9**(6): p. 1034-1041.

64. *THE NORTHERN IRELAND LONGITUDINAL STUDY OF AGEING*. Available from: https://www.qub.ac.uk/sites/NICOLA/.

65. *Parkinson MRI Imaging Repository: Part 2 Database*. Available from: http://www.hra.nhs.uk/news/research-summaries/parkinson-mr-imaging-repository-part-1-multicentre-mri-study-in-pd/.

66. Ritchie C W and Ritchie K, *The PREVENT study: a prospective cohort study to identify mid-life biomarkers of late-onset Alzheimer's disease.* BMJ Open, 2012. **2**(6).

67. *The Belfast PRIME study* Available from: https://www.qub.ac.uk/research-centres/PRIMEStudy/.

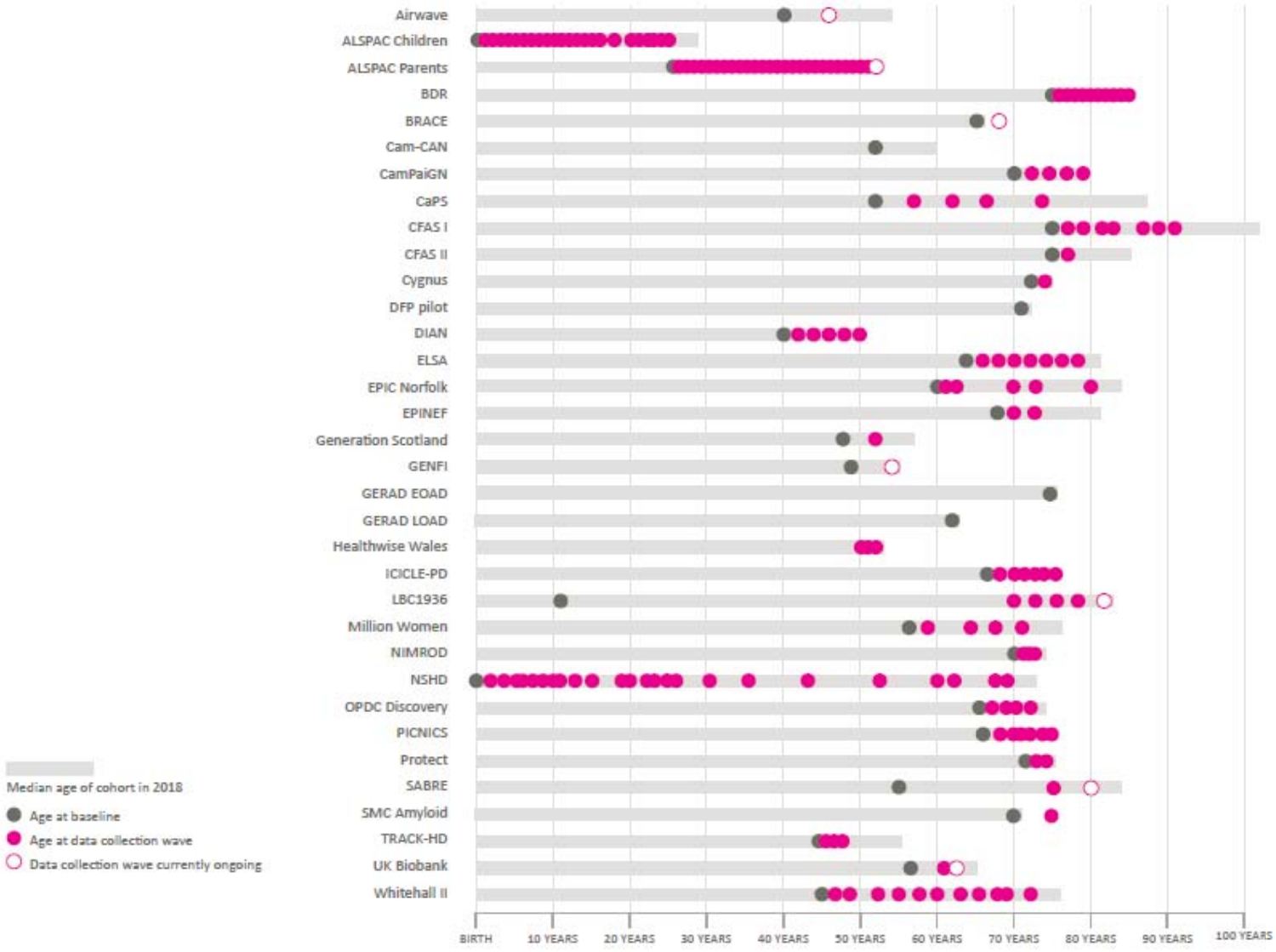
bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1:

DPUK collaborating cohorts

	Data status	Cohort	Institution	Access process	n	Total
1	Full data upload to Data Portal	Airwave [25]	Imperial	Portal + Cohort	53 280	
2		BRACE [42]	Bristol	Portal	2 389	
3		Cam-CAN [43]	Cambridge	Portal + Cohort	2 683	
4		CamPaiGN [44]	Cambridge	Portal	142	
5		CaPS [22]	Bristol	Portal + Cohort	2 512	
6		CFAS I [18]	Cambridge	Portal	18 005	
7		CFAS II [19]	Cambridge	Portal	7524	
8		Cygnus [45]	Manchester	Portal	500	
9		DFP pilot [33]	Oxford	Portal	21	
10		ELSA [17]	UCL	Portal	11 391	
11		EPINEF [46]	Yonsei (RoK)	Portal	2 008	
12		Generation Scotland [47]	Edinburgh	Portal	23 960	
13		GERAD LOAD [48]	Cardiff	Portal	10 454	
14		GERAD EOAD [48]	Cardiff	Portal	4 397	
15		ICICLE-PD [49]	Newcastle	Portal	318	
16		NIMROD [50]	Cambridge	Portal	276	
17		OPDC Discovery [15]	Oxford	Portal	1 589	
18		PICNICS [51]	Cambridge	Portal	282	
19		SMC Amyloid [16]	SMC Seoul (RoK)	Portal	120	
20		TRACK-HD [52]	UCL	Portal	366	
21		SAIL-DeC [30]	Edinburgh	Portal	1 246 557	
22		Whitehall II [21]	UCL	Portal + Cohort	10 308	1 399 082
23	Data upload per project	ALSPAC Children [24]	Bristol	Cohort	15 630	
24		ALSPAC Parents [24]	Bristol	Cohort	22 761	
25		BDR [20]	Bristol	Portal	2 963	
26		DIAN [13]	UCL	Portal + Cohort	437	
27		EPIC Norfolk [26]	Cambridge	Portal	25 639	
28		UK GENFI [14]	UCL	Portal + Cohort	515	
29		HealthWise Wales [29]	Cardiff	Portal + Cohort	6 100	
30		LBC1936 [53]	Edinburgh	Portal + Cohort	1 091	
31		Million Women [27]	Oxford	Cohort	1 319 475	
32		NSHD [23]	UCL	Portal + Cohort	5 362	
33		Protect [54]	Exeter	Cohort	14 000	
34		SABRE [55]	UCL	Portal + Cohort	4 858	
35		UK Biobank [28]	Oxford	Cohort	502 655	2 062 162
36	Meta-data only available	AMPLE [56]	Newcastle	Portal	80	
37		CHARIOT [57]	Imperial	Portal + Cohort	24 509	
38		Delphic [58]	UCL	Cohort	2 000	
39		EXTEND [59]	Exeter	Portal + Cohort	9 500	
40		HKU-NCDC [60]	Hong Kong University	Portal	500	
41		LEWY-PRO [61]	Newcastle	Portal	100	
42		Memento [62]	Bordeaux (Fra)	Portal + Cohort	2 323	
43		NAMGARAM-2 [63]	Gyeongsang (RoK)	Portal	1 000	
44		NICOLA [64]	Queen's Belfast	Portal	8 504	
45		PaMIR [65]	Nottingham	Cohort	400	
46		PREVENT [66]	Edinburgh	Portal + Cohort	700	
47		PRIME [67]	Queen's Belfast	Portal	2 745	52 361

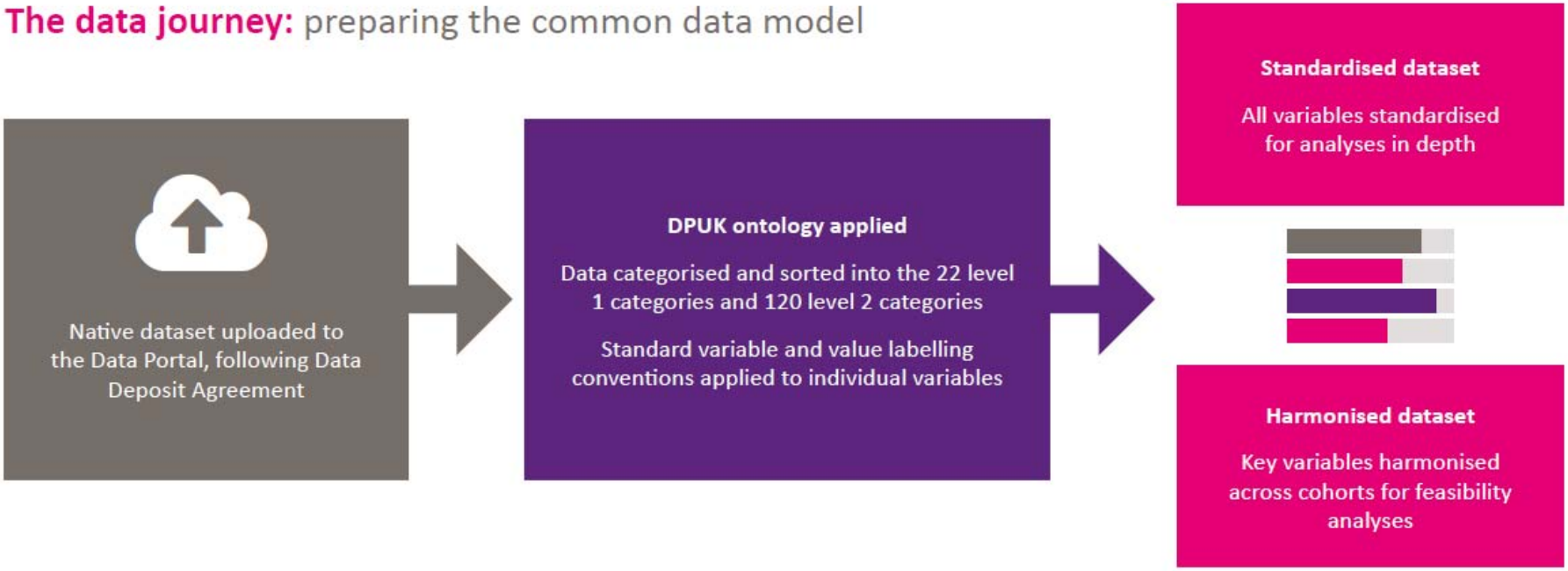
Figure 1: Cohort diversity



bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2:

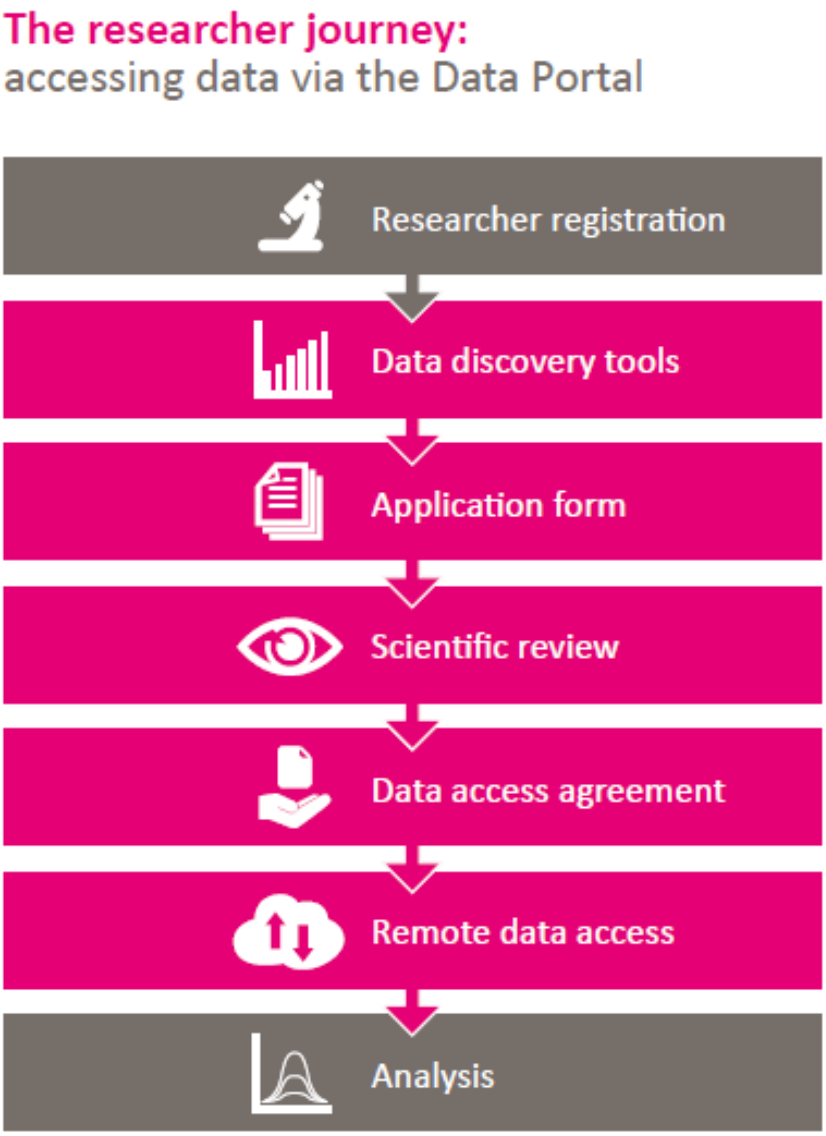
Preparing the data: The data journey



bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/582155; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3:

Accessing data via the Data Portal: The researcher journey