Wellcome Mental Health Data Prize

UK exemplar datasets

SOCIAL FINANCE

wellcome

In preparation for the data prize, we have consulted with several data holders in the UK and South Africa to build an understanding of the data that we consider to be appropriate for the prize. In the process of this, we have identified several exemplar datasets that meet all of the criteria above. The data holders of each of these datasets have been consulted on the most effective way to access their data, with details of the UK datasets documented below.

We welcome applications from teams that propose to conduct their research on alternative datasets, provided they meet the criteria above. Applications proposing the use of such datasets will not be disadvantaged for selection to the prize.

When making an application to a data holder for access to a dataset, please use the prefix "Wellcome MH data prize" to your project title.

UK datasets

Avon Longitudinal Study of Parents and Children (ALSPAC)

Official info:

http://www.bristol.ac.uk/alspac/

Profile:

The Avon Longitudinal Study of Parents and Children (ALSPAC), also known as Children of the 90s, is a world-leading birth cohort study based at the University of Bristol. Between April 1991 and December 1992 more than 14,000 pregnant women were recruited into the study and these women (some of whom had two pregnancies or multiple births during the recruitment period), the children arising from the pregnancy, and their partners have been followed up intensively over two decades.

It is a transgenerational prospective observational study investigating influences on health and development across the life course. It considers multiple genetic, epigenetic, biological, psychological, social and other environmental exposures in relation to a similarly diverse range of health, social and developmental outcomes.

Cohort profiles:

BOYD, A, et al. (2012) Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology*, 1–17

Information about and exploration of the variables collected as part of ALSPAC is available through:

CLOSER Discovery platform:

https://discovery.closer.ac.uk/item/uk.alspac/c6601e78-0f74-446a-a5f1-7c73a1984b4e

Catalogue of Mental Health Measures:

https://www.cataloguementalhealth.ac.uk/?content=study&studyid=ALSPAC

Link to current and published research:

http://www.bristol.ac.uk/alspac/researchers/publications/

Access guidelines:

To access ALSPAC data, teams must submit their research proposal to the ALSPAC executive committee:

https://proposals.epi.bristol.ac.uk/

Before submitting a proposal, please see the ALSPAC access policy:

http://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf

Paywall:

Costs for data access are charged on a project-by-project basis. The basic fee is £1925, plus VAT where this is applicable. It is expected that teams will incorporate costs for accessing data into their budget for phase 1, as there will be no additional funding for data access.

Access timeline:

The ALSPAC executive committee will give an initial response either accepting or denying the research proposal within 10 days. It is expected that teams will be able to provide evidence of their accepted proposal at application.

A data agreement will then be set up between ALSPAC and the research institution of the team. Because of the time this takes, any team wishing to use ALSPAC data is expected to have made their initial application by Thursday 19th May at the latest.

Data will then be provided within two weeks of payment being made.

Millennium Cohort Study

Official info:

https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/

Profile:

The Millennium Cohort Study (MCS), which began in 2000, is conducted by the Centre for Longitudinal Studies (CLS). It aims to chart the conditions of social, economic and health advantages and disadvantages facing children born at the start of the 21st century. The study has been tracking the Millennium children through their early childhood years and into adulthood.

Cohort profiles:

JOSHI, H and FITZSIMONS, E. (2016) The Millennium Cohort Study: the making of a multi-purpose resource for social science and policy. *Longitudinal and Life Course Studies*, 7(4), 409-430

CONNELLY, R and PLATT, L. (2014) Cohort Profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology*, 43(6), 1719-1725

Information about and exploration of the variables collected as part of the MCS is available through:

- The CLS MCS Excel data dictionary:

https://cls.ucl.ac.uk/wp-content/uploads/2017/02/MCS_Data_Dictionary.xlsx

- The CLOSER Discovery platform:

https://discovery.closer.ac.uk/item/uk.cls.mcs/0d8a7220-c61b-4542-967d-a40cb5aca430

- The Catalogue of Mental Health Measures:

https://www.cataloguementalhealth.ac.uk/?content=study&studyid=MCS

The data holders recommend researchers who are new to MCS to consult the Data Handling Guide for information on how to join the many datasets together meaningfully:

https://cls.ucl.ac.uk/wp-content/uploads/2020/09/MCS_Data_Handling_guide_ed1_2020-08-10.pdf

Link to existing studies:

https://www.bibliography.cls.ucl.ac.uk/Bibliography.aspx?sitesectionid=647&sitesectiontitle=Bibliography&d=1&yf=&yt=&a=&s=MCS&o=&j=

Access guidelines:

The majority of survey and linked administrative data from the Millennium Cohort Study is available through the UK Data Service under different conditions of access (End User Licence (EUL), Special Licence or Secure Access) depending on the level of disclosivity and sensitivity of the data:

https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000031

Please see the section below on accessing data through the UK Data Service.

For the data that are held under more restrictive access procedures, there are often top-coded versions available under the EUL. For example, respondent ethnicity is available under the EUL for all categories represented by more than 30 individuals. It is only for analyses requiring granularity beyond this that application to the equivalent variable held under Secure Access would be necessary. Applicants should always ensure that the less sensitive version of a data field will not suffice for research purposes before applying for access to the full data field.

Linked health and education data for Welsh respondents is available through the SAIL Databank. To access these, prize teams should apply to SAIL with details of the project and the data required from SAIL. Details of this process can be found here:

https://saildatabank.com/application-process/two-stage-process/

Data with the highest risk of potential disclosure and MCS genetic data are available only through direct application to the Centre for Longitudinal Studies (CLS). To access these, prize teams should return a completed CLS DAC data access application form to clsfeedback@ucl.ac.uk. The form can be found here:

https://cls.ucl.ac.uk/data-access-training/data-access/accessing-data-directly-from-cls/

Access process	Data details	Access timelines
UK Data Service EUL	All survey data not specified below	Instant, following registration
UK Data Service Special Licence	Hospital of birth	1-3 months
UK Data Service Secure Access	Detailed sensitive demographic, socioeconomic and health variables, particularly for categories with low participant counts Geographic identifiers below LSOA level Linked health administrative datasets Linked education administrative datasets Banded distances to English grammar schools	3-4 months
SAIL Databank	Welsh linked health and education administrative datasets	3-4 months
Application to CLS	Free text fields Postcodes Genetic data Biological samples Paradata	3 months +

Linking new data:

It is possible to apply for permission to link new data to MCS datasets. However, CLS expect researchers to provide them with six months to approve a new linkage request. This is not compatible with the data prize timelines, so we ask applicants not to include new linkage with MCS data as part of their research proposal.

Next Steps

Official info:

https://cls.ucl.ac.uk/cls-studies/next-steps/

Profile:

Next Steps, previously known as the Longitudinal Study of Young People in England (LSYPE), follows the lives of around 16,000 people in England born in 1989-90.

The study began in 2004 when the cohort members were aged 14, with an original sample of 15,770 people. Cohort members were surveyed annually until 2010, and the next sweep after this was when they were aged 25, in 2015-16.

Next Steps has collected information about cohort members' education and employment, economic circumstances, family life, physical and emotional health and wellbeing, social participation and attitudes.

The Next Steps data has also been linked to National Pupil Database (NPD) records, which include the cohort members' individual scores at Key Stage 2, 3 and 4 and more administrative linkages are planned (for example: Higher Education Statistics Agency, The Universities and Colleges Admissions Service, Department for Work and Pensions).

Information about and exploration of the variables collected as part of Next Steps is available through:

- CLS Next Steps Excel data dictionary:

https://cls.ucl.ac.uk/wp-content/uploads/2021/12/NextSteps_Data_Dictionary.xlsx

- The CLOSER Discovery platform:

https://www.closer.ac.uk/study/next-steps/

- The Catalogue of Mental Health Measures:

https://www.cataloguementalhealth.ac.uk/?content=study&studyid=LSYPE

Previous publications are searchable on the CLS Bibliography website:

https://www.bibliography.cls.ucl.ac.uk/Bibliography.aspx?sitesectionid=647&sitesectiontitle=Bibliography

User guide:

https://cls.ucl.ac.uk/wp-content/uploads/2020/08/Next-Steps-User-guide-to-the-redeposit-of-sweeps-1-to-7-May2020.pdf

Access guidelines:

Almost all the survey and linked administrative data collected by the study is available via the UK Data Service under different conditions of access (End User Licence or Secure Access) depending on the level of disclosivity and sensitivity of the data:

https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000030

Please see the section below on accessing data through the UK Data Service.

The majority of these data are available under End User Licence (EUL). Sensitive fields, generally those that could enable the identification of respondents or belong to linked datasets, are held under more restricted access on the UK Data Service, either under Special Licence or Secure Access.

Data with the highest risk of potential disclosure are available only through direct application to the Centre for Longitudinal Studies (CLS). To access these, prize teams should return a

completed CLS DAC data access application form to clsfeedback@ucl.ac.uk. The form can be found here:

https://cls.ucl.ac.uk/data-access-training/data-access/accessing-data-directly-from-cls/

Access process	Data	Access timelines
UKDS EUL	All except below	Instant, following registration
UKDS Secure Access	Sensitive variables from the questionnaire data for Sweeps 1-8. Including date of interview (detailed), date of birth (detailed), detailed disabilities, full or detailed SOC/SIC codes, child care arrangements, higher Education identifiers, potential school identifiers. National Pupil Database (NPD) linked data at Key Stages 2, 3, 4 and 5, England. Linked Individualised Learner Records learner and learning aims datasets for academic years 2005 to 2014, England. Detailed geographic indicators for Sweep 1 and Sweep 8 (2001 Census Boundaries and 2011 Census Boundaries) Linked Health Administrative Datasets (Hospital Episode Statistics) for years 1998-2017. Linked Student Loans Company Records for years 2007-202.	3-4 months
Application to CLS	Free text fields Postcodes Paradata	1-3 months

Understanding Society

Official info:

https://www.understandingsociety.ac.uk/

Profile:

Understanding Society is the largest longitudinal study of its kind and provides crucial information for researchers and policymakers on the changes and stability of people's lives in the UK.

The overall purpose of Understanding Society is to provide high-quality longitudinal household data on subjects such as health, work, education, income, family, and social life to help understand the long-term effects of social and economic change, as well as policy interventions designed to impact upon the general wellbeing of the UK population. To this end, the Study collects both objective and subjective indicators and offers opportunities for research within and across multiple disciplines including sociology and economics, geography, psychology and health sciences.

Cohort profile:

PLATT, L., KNIES, G., LUTHRA, R., NANDI, A. and BENZEVAL, M. (2020) Understanding Society at 10 Years. European Sociological Review, 36 (6), 976-988

Detailed information about the Study can be found in the User Guide:

https://www.understandingsociety.ac.uk/documentation/mainstage/user-guides/main-survey-user-guide/

Information about and exploration of the variables collected as part of Understanding Society is available through:

- the CLOSER Discovery platform:

https://discovery.closer.ac.uk/item/uk.iser.ukhls/44a7a09e-4703-498c-96f7-0131b296c917

- the Understanding Society website:

https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation

Understanding Society offer regular training courses on using the available datasets. You can register for these here:

https://www.understandingsociety.ac.uk/help/training

Link to existing studies:

https://www.understandingsociety.ac.uk/research/publications

Access guidelines:

Almost all the data collected by the study is available via the UK Data Service:

https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000053

Please see the section below on accessing data through the UK Data Service.

The majority of these data are available under End User License (EUL). Sensitive fields, generally those that could enable the identification of respondents or belong to linked datasets, are held under more restricted access on the UK Data Service, either under Special Licence or Secure Access.

Where access requires more than agreeing to an EUL, applications for request will be evaluated by the Understanding Society team on the following criteria:

1. Has the application been submitted by bona fide researchers who can demonstrate public interest?
2. Does the application violate (or potentially violate) any of the ethical permissions granted to the study or any undertakings given to the participants or their guardians?
3. Does the application run the risk of producing information that may allow individual participants to be identified?
4. Does the application run a significant risk of upsetting or alienating study members or of reducing their willingness to remain as active participants in Understanding Society based research?
5. Does the application address topics that fall within the acknowledged remit of the Understanding Society project, as understood by participants?
6. Does the application demonstrate that ESRC policy regarding deposit of data will be adhered to?

If access is granted, the decision is communicated to UKDS or ONS SRS and data will be made available through them under a Special Licence/ Secure Access agreement as appropriate.

Exceptions:

To link genetic or epigenetic data to End User Licence survey data, applications need to be made directly to the health data team at Understanding Society. Applicants are asked to specify the nature of the proposed research and all the data used in the project. If the

application is successful the data team at Understanding Society will prepare and supply the dataset.

Genetic/Epigenetic application form:
https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/health/ukhls-omics-application-form.docx

Variable request form:
https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/health/ukhls_eul_variable_template_inst.xlsx

The completed forms should be returned by email to genetics@understandingsociety.ac.uk

The data holders aim to give researchers a response to their application within 30 days, but complex proposals may require additional rounds of contact. Once approved, data holders aim to prepare linked genetic-survey datasets within 30 days. The dataset created will include the smallest viable subset of survey data for the proposed research question to minimise disclosure risk.

Given the timelines indicated above, if prize applicants intend to use linked genetic/epigenetic data for their research proposal, we recommend submitting an application to Understanding Society by no later than May 4th.

Access process	Data	Access timelines
UKDS EUL	All except below	Instant, following registration
UKDS Special License	Identifiers such as detailed occupation and income codes Sensitive data such as medication codes Geographic identifiers below Government Office Region and above postcode level School codes (Note that combinations of the above may not be permitted)	Up to 3 months
UKDS Secure Access	Full DOB Linked education administrative datasets Lat/long coordinates for postcode centroids	Up to 4 months
Application to data holder	Genetic / epigenetic data linked to EUL survey data	30 days for decision 30 days for preparation

A detailed overview of the specific variables that can be found at each access level across the UKDS can be found here:

https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/user-guides/mainstage/6931_eul_vs_sl_variable_differences.pdf

Linking to external data:

Aside from the linked administrative data available via UKDS Special Access, it is possible to link your own publicly available data to the study via the geocodes (and access conditions) listed above.

UK Data Service

Several of the UK longitudinal studies make their data available via the UK data service. The majority of these data are available under End User License (EUL). To access these, a user must set up an account with the UK Data Service and agree to the EUL. Users can then create their own projects on the website. To access data, users must add datasets to their projects, which then becomes available for download.

Sensitive fields, generally those that could enable the identification of respondents or belong to linked datasets, are held under more restricted access on the UK Data Service, either under Special Licence or Secure Access. To access these, prize teams should first follow the same steps to access as with data under the EUL. Once datasets are added to the users' project, users must navigate to the 'Datasets' tab of their project page. Rather than becoming immediately downloadable, data requiring special or secure access will show 'request access'. Click on this to view the specific actions required for the data you wish to access, and follow the steps under 'Complete actions'.

Please see a more detailed UK Data Service access guide here:

https://ukdataservice.ac.uk/help/access-policy/how-to-download-and-order-your-data/