Centre for Longitudinal Studies

Data Access Framework

Document Information	
Document Name	CLS_Data_Access_Framework
Author(s)	Aida Sanchez
Version	2
Issue Date	December 2020
Approved By	CLS Data Access Committee
Review frequency	Yearly

Document History		
Version	**Date**	**Summary of change**
0.3	Sept 2015	Draft 0.3 submitted to ESRC 30th Sept 2015. Not for onward circulation
1	June 2018	Addition and re-organisation of information (AS)
2.0	Sept 2020	Re-formatting and adding information to bring the document up to date (AS).
2.1	Oct 2020	Add requirements for approval NHS Digital and NHS Scotland. Add CLS genetic data sharing (AS)
2.2	Dec 2020	Removal of METADAC references. Additions and corrections following feedback. Addition of internal data access for initial findings (AS)

Table of Contents

1. Definitions

“Personal data” are defined as any information relating to an identified or identifiable natural person.

“Sensitive personal data” are defined in Article 9(1) of the GDPR as personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited

“Disclosive data” are data which may lead to the identification of an individual. An individual may be *directly identified* from their name, address, postcode, telephone number, photograph or image, or some other unique personal characteristic. An individual may be *indirectly identifiable* when certain information is linked together with other sources of information, including, their place of work, job title, salary, their postcode or even the fact that they have a particular diagnosis or condition.

“Confidential data” are data given in confidence or data agreed to be kept private between two parties, that are not in the public domain. Confidential data are conditioned by factors such as ethical guidelines, legal requirements or research-specific consent agreements. Any information obtained by a person on the understanding that they will not disclose it to others, or obtained in circumstances where it is expected that they will not disclose it.

“Data controller” is the organisation that determines the purposes and means of the processing of personal data. In our case, the data controller is CLS.

“Data Processor” or “Custodian” is the organisation that processes personal data on behalf of the controller, and it is responsible for the safekeeping of data and/or tissue samples and control of their use, and eventual disposal (if required), all in accordance with legislation and the terms of the consent given by the cohort members. Custodianship implies some rights or inputs into decisions to decide how the data/samples are used and by whom, and also responsibility for safeguarding the interests of the donors.

“Anonymisation” is the process of completely removing personal identifiers, both direct and indirect, and dealing with the data environment in such a way that the risk

of somebody being identified in the data is negligible. Truly anonymised data do not fall within the scope of the GDPR.

“Pseudo-anonymisation” or **“de-identification**” refers to the processing of personal data in such a way that the data can no longer be attributed to a specific individual without the use of additional information, as long as such additional information is kept separately and subject to technical and organizational measures to ensure non-attribution to an identified or identifiable individual.

2. Executive summary

The Centre for Longitudinal Studies (CLS) is responsible for four national cohort studies: National Child Development Study (NCDS, or the 1958 Birth Cohort Study), the 1970 British Cohort Study (BCS70), the Millennium Cohort Study (MCS) and Next Steps (LSYPE), which follow the same group of people from childhood and throughout their lives. The UK has a unique and internationally renowned portfolio of longitudinal studies, and the studies at CLS make an important contribution to these.

All access to and use of CLS data is governed by the principles and procedures set out in this Data Access Framework, which seek to be fair, open and transparent. This framework is reviewed and maintained by the CLS Data Access Committee (CLS DAC).

The aim of the CLS data access strategy is to ensure that the research data produced by CLS are made as widely available as possible to the research community (nationally and internationally), whilst ensuring that: i) sensitive data and/or data that is or may be disclosive are kept secure and shared in a secure manner; ii) the legal requirements, ethical guidelines and moral responsibility to the study participants are maintained; iii) the research-specific consent agreements and undertakings given to the cohort members are complied with.

This Framework builds on existing agreements for accessing data collected by complex longitudinal surveys developed between the Economic and Social Research Council (ESRC) and the UK Data Service (UKDS). In addition, CLS also has agreements with other data repositories for dissemination of research data.

There are several categories (‘tiers’) of CLS research data based on the likelihood and potential risk of disclosure. These have been described in the CLS Data Classification Policy, with each category having a defined access mechanism. The main routes for access to the different tiers of CLS research data are:

- The UKDS End User Licence (EUL) for access to research data with a low level of sensitivity and disclosivity (Tier 1). The approval of data requests under this licence is authorised directly by the UK Data Service;
- The UKDS Special Licence (SL) for access to research data of moderately sensitive nature or with a medium level of disclosivity (Tier 2a). These data requests are handled via Special Licence applications, which are administered by the UKDS and approved by the CLS Research Data Management team on behalf of the CLS Data Access Committee.
- The UKDS Secure Access licence (SA) for access to data with a higher level of sensitivity and/or disclosivity (Tier 2b). These data requests are handled via Secure Access applications, which are administered by the UKDS and approved by the CLS Research Data Management team on behalf of the CLS Data Access Committee.
- CLS also has separate data sharing arrangements with the SAIL Databank for the dissemination of MCS data linked to Welch administrative data, and with the European Genome-Phenome Archive (EGA) for the dissemination of CLS genetic data not linked to survey data.
- Researchers may also apply for access to data that has not yet been disseminated via UK Data Service or other data repositories. Such requests are managed by the CLS Data Access Committee and the data made available to the applicant directly by the CLS Data Management team.
- Access to CLS biological samples and to sensitive data derived from biological samples, such as CLS genetic data linked to survey data, are managed by the CLS DAC.

3. Scope

The CLS Data Access Framework has been developed by the CLS Data Access Committee (CLS DAC) to ensure the data produced by CLS is made widely available for research purposes, both nationally and internationally, to maximise the impact of the CLS studies. At the same time it is necessary to ensure that sensitive data and/or data which are or may be disclosive are kept secure and shared in a secure manner; the legal requirements, ethical guidelines and moral responsibility to the study participants are maintained; the research-specific consent agreements and undertakings given to the participants are complied with.

This Framework identifies a series of mechanisms to provide access to the data collected by CLS. These procedures apply to all data collected, not just under the main studies commissioned by ESRC, but also any co-funded add-on studies.

The Framework builds on existing agreements developed by the ESRC and UK Data Service for accessing data collected by complex longitudinal surveys. It recognises the importance of developing procedures, protocols and standards to support ethical safeguards surrounding data access and the reuse of data for research purposes.

The framework sits alongside CLS policies governing the secure handling of data, which include:

- the CLS DAC Terms of Reference
- the CLS Data Classification Policy
- the CLS Biosamples Strategy
- the CLS Secure Data Handling Guide
- the CLS Information Governance Policy

The CLS Data Access Framework is a public document available to all potential users. The Framework has been developed by the CLS Data Access Committee and is owned by the CLS Senior Leadership Team. The CLS Strategic Advisory Board has a responsibility to advise on the procedures for access to CLS data which are governed by this framework and might evolve over time.

4. CLS research data

The Centre for Longitudinal Studies (CLS), based at the UCL Institute of Education, manages the collection and curation and public dissemination of the data of four major national cohort studies: the National Child Development Study (NCDS), the 1970 Birth Cohort Study (BCS70), the Millennium Cohort Study (MCS) and Next Steps (LSYPE).

The majority of data from the studies arises from questionnaires completed by study members or their families at periodic study sweeps. UCL is the data custodian for these data. In addition, a number of more specialised forms of data are also included as part of the studies, including from biological samples and from externally linked administrative records. The custodianship for linked data is set out in more detail further below.

Comprehensive information about each of the CLS studies can be found at http://www.cls.ucl.ac.uk/

4.1 Survey Data

National Child Development Study (NCDS)

The 1958 National Child Development Study (NCDS) started in 1958 at birth as the Perinatal Mortality Survey, and is following the lives of an initial 17,415 people born in England, Scotland and Wales in a single week of 1958. Over the course of cohort members' lives, NCDS have collected information on their physical and educational development, economic circumstances, employment, family life, health behaviour, wellbeing, social participation and attitudes.

NCDS research data available are from the UK Data Service at:

http://discover.ukdataservice.ac.uk/series/?sn=2000032

British Cohort Study 1970 (BCS70)

The 1970 British Cohort Study (BCS70) is following the lives of around 17,000 people born in England, Scotland and Wales in a single week of 1970. Over the course of cohort members' lives, BCS70 has collected information on health,

physical, educational and social development, and economic circumstances among other factors.

BCS70 research data are available from the UK Data Service at:

http://discover.ukdataservice.ac.uk/series/?sn=200001

Next Steps (LSYPE)

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 and Individual Learner Records (ILR), which include the cohort members' individual scores at Key Stage 2, 3 and 4.

Next Steps research data are available from the UK Data Service at:

http://discover.ukdataservice.ac.uk/series/?sn=2000030

Millennium Cohort Study (MCS)

The Millennium Cohort Study (MCS), known as 'Child of the New Century' to cohort members and their families, is following the lives of around 19,000 young people born across England, Scotland, Wales and Northern Ireland in 2000-02. The study began with an original sample of 18,818 cohort members. The MCS provides multiple measures of the cohort members' physical, socio-emotional, cognitive and behavioural development over time, as well as detailed information on their daily life, behaviour and experiences. Alongside this, rich information on economic circumstances, parenting, relationships and family life is available from both resident parents. National Pupil Database (NPD) records have also been linked to the MCS data, including GCSE exam results.

MCS research data are available from the UK Data Service at:

http://discover.ukdataservice.ac.uk/series/?sn=2000031

4.2 Biological Samples

A number of biological samples have been collected, or are being planned to be collected from study members of NCDS, MCS, and BCS70. There are a variety of arrangements for custodianship and access to data arising from these biological samples.

National Child Development Study (NCDS)

The 2002/3 Biomedical Survey for NCDS collected whole blood and saliva from cohort members. The samples were processed and both original aliquots and residues are held at the University of Bristol who are the custodians of these biological samples. These are stored at the Bristol Bioresource Laboratories, Population Health Sciences, Bristol Medical School, University of Bristol. Access to these samples is overseen by the CLS Data Access Committee (CLS DAC).

Some biochemical marker data from the NCDS Biomedical Data, 2002-2004 are currently available under Special Licence from the UK Data Service:

http://discover.ukdataservice.ac.uk/catalogue/?sn=5594&type=Data%20catalogue

DNA was extracted from whole blood and there is also a transformed lymphocytes collection. Both of these collections have been extensively genotyped. The transformed lymphocyte collection allows for further DNA extraction, whilst the whole blood derived DNA collection is a finite resource.

The majority of genotype data is held at the European Genome-phenome Archive. Access to genotype data *unlinked to any other data* (except region, sex and ethnicity) is through the Wellcome Trust Case Control Consortium Data Access Committee (WTCCC DAC):

https://www.ebi.ac.uk/ega/dacs/EGAC00000000001

Millennium Cohort Study (MCS)

Milk teeth were collected at age 5 (MCS3) and onwards, these are stored at the Institute for Child Health. Access to these milk teeth and the data arising from them is via the CLS DAC.

Oral fluid was collected at age 3 (MCS2). All oral fluid samples are depleted and residues destroyed. Data arising from the assay is available at the UK Data Service.

In 2015, the age 14 sweep (MCS6) collected saliva from both cohort members and natural parents. DNA has been extracted from the saliva sample and genotyped,. Unlike the NCDS transformed lymphocytes collection, this is non-renewable. University of Bristol are the custodians for the biological samples.

British Cohort Study 1970 (BCS70)

Whole bloods will be collected from study members during the age 46 sweep (2016/2017). CLS will remain the custodian of these samples, the processing and storage of which will be contracted to a third party via a competitive tender process undertaken by UCL. Access to samples and data will be managed via the CLS DAC.

4.3 Linked administrative data

CLS has an existing programme of linkage to administrative data which is based on the informed consent obtained directly from participants during the surveys data collection. Consent has been secured for linkage to health, education and economic records from the relevant administrative sources.

Linked administrative data, suitably pseudoanonymised, are provided to researchers via the UK Data Service with the agreement of relevant data controllers. These data include to education records in England, Scotland and Wales (MCS), data from England (NCDS, BCS70, Next Steps), health data from Scotland (NCDS, BCS70, MCS) and health data from Wales (MCS).

In addition, there is an existing programme of non-consented linkage mainly for tracing purposes and mortality for which ethical and Section 251 approval is in place for the four CLS studies.

CLS also has a programme of non-consented linkage based on geographical identifiers from publicly available data. Geographical identifiers with associated census and other local area data are also available from UKDS for MCS and for some sweeps of NCDS and BCS70.

5. CLS data access principles

The procedures and processes that have been applied to provide access to CLS data are derived from the key principles set out below:

1. **Responsible custodianship**: UCL is the custodian of data generated in the course of the CLS birth cohort studies, except in the case of certain biological samples. CLS is responsible for the safekeeping of data and/or tissue samples and control of their use, and eventual disposal (if required), all in accordance with legislation and the terms of the consent given by the donor. No organisation, commercial or otherwise should be allowed to gain control or ownership over the CLS resource. Where consent has been obtained, or in exceptional circumstances where section 251 approval has been granted for unconsented linkage, CLS data may be linked to administrative data and shared securely. In general custodianship of administrative data is retained by the data controllers of the administrative records;
2. **Wide data access**: to maximise the impact of the studies, the data produced by CLS are made as widely available as possible to a range of users, subject to the considerations listed below;
3. **Controlled and transparent access governance**: All access to CLS data is governed by the procedures set out in the CLS Data Access Framework which aim to be fair, open and transparent. The controls applied are proportionate to potential risks of disclosure;
4. **Welfare of study members:** use to CLS research data must have a very low risk of damaging the wellbeing of one or more study respondents. The contents of the publication of the research results must be unlikely to upset or alienate participants;
5. **Public perception and reputation of the studies**: general risks and risks related to socially controversial areas will be taken into account with regards to public perception, risk to continuation of the studies and possible reduction of participants' willingness to continue being part of the cohort study, all whilst aiming at avoiding unnecessary barriers to research;

6. **FAIR data**: CLS data management and data sharing processes are in place to ensure that the CLS research data and metadata follow the FAIR data guiding principles of being findable, accessible, interoperable, and reusable (https://www.fairdata.org.uk/).

7. **Data security:** UCL, which houses CLS at the UCL Institute for Education, has ultimate responsibility for data security. CLS considers all issues relating to information security and data protection a high priority, and organisational security assurances will be requested as needed;

8. **Ethical considerations**: access to the data and samples is granted in line with the terms of consent agreed with the participants of NCDS, BCS70, Next Steps, and MCS. When assessing data access requests, the CLS Data Access Committee will consider whether the proposed research is consistent with undertakings given to study participants when they gave informed consent. Access requests will be checked to ensure they have had relevant scientific and ethical approval, and that conflict of interest are taken into account.

   CLS has obtained Research Tissue Bank ethical approval for the collection, storage, use and distribution of samples, which facilitates programmes of research without a need for individual project-based ethical approval.

9. **Management of disclosure and sensitivity risks via data categorisation**: CLS data (including biological samples) is categorised to facilitate access through an assessment of the likelihood and potential impact of disclosure. Data that risk the disclosure of information which could identify individuals will require an appropriate degree of security and management, and will be made available under strict levels of access to *bona fide* researchers for research that can demonstrate public interest. The data categorisation principles are detailed in the CLS Data Classification Policy;

10. **Comprehensive assessment criteria of data sharing projects and applicants:** the majority of CLS data are available to all applicants registered with the UK Data Service (whether in universities, government departments, charities or non-commercial companies). CLS will apply a comprehensive set of criteria to evaluate and approve data access. Public benefit, including potential

scientific and wider impacts of the proposed research, must be justified when required;

11. **Punishable violation of access conditions**: an appropriate set of penalties may be applied should violations of access conditions take place. Penalties can be imposed on users and/or their institutions. Further details can be found at paragraph 16 http://ukdataservice.ac.uk/get-data/how-to-access/conditions;
12. **Data minimisation**: researchers will only be given access to the research data needed for the approved research projects.
13. **Minimal costs**: there is not cost for accessing CLS research survey and linked data. In the case of biological specimens and data arising from it, recipient institutions will be expected to meet all the costs of sample handling, specimen transport and data preparation in relation to their study;
14. **Controlled release of biological samples**: as a depletable resource, the use of the biological samples will be carefully controlled, in order to optimise the long-term value of the resource.
15. **Data return:** if required, new data and associated metadata generated by approved researchers must be made available for re-use.

6. CLS data access processes

CLS research data are made available for researchers to undertake their analysis by identifying and requesting data from the UK Data Service and other repositories, described in section 7 of this document.

For data not publicly available from public data repositories, researchers may apply directly to the CLS DAC.

This section describes the processes and considerations that shape the CLS data access strategy.

6.1 Custodianship

University College London (UCL) houses CLS at the UCL Institute of Education. UCL is the custodian of the data of NCDS, BCS70, Next Steps and MCS.

In general custodianship of administrative data is retained by the data controllers of the administrative records (e.g. NHS Digital, Department of Education). In cases where these organisations require individual scrutiny of applications for their data linked to CLS survey members, such data will be referred to the CLS DAC and the original data provider for decision.

As set out in the terms and conditions of the CLS Resource Centre Grant, ESRC maintains the right to require transfer of custodianship of these data to third party providers on termination of the grant, or on material failure of CLS in conducting the grant. Alternatively ESRC may require UCL to permit third parties full access to the data on termination of the grant.

6.2 Data security

UCL has ultimate responsibility for security of the CLS data it houses. CLS considers all issues relating to information security and data protection a high priority. We base our Information Governance policies and procedures on the requirements of NHS Digital and are compliant with their Data Security and Protection (DSP) Toolkit. The DSP is managed by the CLS Information Governance Steering Group.

All of the personal and collected data are stored and processed within the UCL Data Safe Haven.

The UCL Data Safe Haven and the public data repositories that disseminate the CLS research data (UK Data Service, SAIL Databank and the EGA), are compliant and accredited with the international information security standard ISO27001.

6.3 Data classification based on disclosivity and sensitivity

CLS data is categorised to reflect the likelihood and potential impact of disclosure and degree of data sensitivity. Data that risk the disclosure of information which could identify individuals, households or organisations associated to participants will require an appropriate degree of security and access management.

CLS assigns a data categorisation ("Tier") to the data made available for research purposes. When allocating tiers, a number of considerations are taken into account, as set out in the CLS Data Classification Policy. These include the risk of disclosure, sensitivity of the data, general risks occurring such as to public perception and risk to continuation of the study.

CLS data fall into the following categories, which are defined by the likelihood and potential impact of disclosure and sensitivity:

- Tier 1: Low impact. These data have a low level of disclosure, such as participant self-reported survey data that have been pseudo-anonymised and suitably de-identified.
- Tier 2a: Medium impact. These data are potentially disclosive or have a moderate sensitivity (e.g. (e.g. medium level geographical indicators, pseudoanonymised child adversity data, genetic data).
- Tier 2b: High impact. These data have a higher risk of disclosivity (e.g. detailed geographical indicators) and/or sensitivity (e.g. detailed linked health data).
- Tier 3: Very high impact. These data have a high level of potential disclosure, which includes any information which would allow identification of less than 5% of a population of the data item: e.g. Postcodes, Date of Birth, School ID, GP Identifier used for linkage and lookups to other contact details.
- Personal identifiable data such as names or NHS number are never made available for research use.

For more details please refer to the CLS Data Classification Policy.

The CLS DAC will review categorisation decisions in the case of appeals received from potential users. Where the Committee is content with the categorisation decisions made it will refer the complaint to the published categorisation principles.

6.4 Exclusive data access and data return

No individual researcher is granted exclusive use to any CLS data unless they have generated new data or derived variables themselves, in which case the CLS DAC grants a period of 12 months of exclusive use prior to wider data dissemination. Following this period, researchers are required to make the newly generated data and associated metadata available for re-use, according to the terms of ESRC's Research Data Policy. This can be also in the form of sharing syntax used to generate derived variables. Researchers are expected to either:

i) make the documented data or syntax available to other researchers through deposit at the UKDS under the appropriate access mechanism and according to the impact level classification of the data requested; or

ii) return the data, comprehensive metadata and related documentation to CLS at the explicit request from CLS. This request will arise from the CLS DAC consideration that the data should be made available as part of the UKDS data resource for each cohort (e.g new data linked data). Following thorough checks of the materials received, CLS will deposit these data as an integral part of the CLS data at the UKDS. The user guides of the data deposit will reference the researchers as the creators of the linked outputs.

6.5 Commercial use

Commercial organisations can apply for access to CLS data and are subject to the standard CLS access procedures. Like any other applicants, commercial organisations must confirm that their use of the data is for bona fide research purposes and not for commercial exploitation. They will be required to demonstrate the public benefits that are likely to flow from the research use and are in line with the consent wording collected from the cohort members.

6.6 International access

Tier 1

International access to CLS data is important and unnecessary barriers should not get in the way of such research. Any international research users can register to access Tier 1 data.

Tier 2

Data classified as Tier 2a are available under the UK Data Service Special Licence for users based in countries within the European Economic Area (EEA). Researchers based outside the EEA must apply to the CLS DAC for access to such data. Countries that are not covered by the EU Commission "adequacy decision" might not always have a legislation that protects personal data to the same standard as within the EEA. In this case, CLS will enter into a contractual arrangements to ensure that the research data will be processed lawfully.

6.7 Research in socially controversial areas

A risk management strategy for research in socially controversial areas, such as like ethnicity, criminality, intelligence or sexuality must be in place and aims at mitigating the risk of reputational damage to the study and of alienation of study participants, as well as facilitating risk-management by the study ahead of publication of results.

Risk management strategies include:-

- CLS to seek independent advice regarding ethical considerations related to socially controversial research proposals or to the use of genetic data for research purposes, in particular in relation to consents requirements being fulfilled.
- CLS to seek independent expert advice regarding incidental findings from genetic data.
- Require that all reporting of project results and press releases must use careful and balanced language in order to avoid misinterpretation or exaggeration of the findings.

- Require that live interactions with the media (including social media) should use careful and balanced language. Before agreeing to be interviewed we recommend you have appropriate training.
- All CLS studies must include a privacy notice and Frequently Asked Questions (FAQs) section that documents the research process, data protection and ethical considerations in great detail. Written in plain English and available online in an enduring location, the privacy notices and the FAQs help to guard against misinterpretation or misrepresentation of study findings and thus prevent the possibility of controversies stemming from press coverage.

6.8 Return of incidental findings of clinical relevance

In signing their original consent forms, consenting participants agreed that they would not receive feedback about any individual genetic results. In keeping with this wording the current policy is that *no* genotypic information (regardless of its nature) will be returned to cohort members.

To date, most informed commentators have seen this position as 'good practice' because nobody has really known how to interpret the clinical relevance of the genetic variants that have been identified: their effects have typically been rather small and there has been no agreed way in which to respond to the limited increases in risk they may convey. However, in common with many of the world's major cohort studies and biobanks, CLS recognises that national and international views of what constitutes 'best practice' might be about to change, and that it is possible that in the future it may become mandatory to report genetic results to participants if they satisfy three key requirements:

(i) scientific validity (the genotyping is of adequate quality);

(ii) clinical significance (the disease or condition caused by the genetic variant is potentially serious) , and

(iii) potential benefit (*i.e.* a valid approach exists to prevent or cure the condition/disease of concern and that early knowledge of the genetic risk to which an individual is exposed could enhance the efficacy of that prevention/cure).

At present a change in what is seen as best practice remains a minority review, but findings that satisfy the three stated criteria are likely to become more common as the global scientific focus moves to full sequencing of genes and/or longer segments of DNA. CLS wishes to help contribute to the national and international evidence-base on which any future strategic decisions might be made regarding policy for feeding back genetic results.

For this reason, CLS now requires that if in the course of any analysis of DNA from any participant, a genetic variant is found that could potentially be viewed as meeting any of the three of the criteria stated above, that information must be transmitted to the CLS Data Access Committee.

At present genetic information will only be returned to participants if advised by an independent clinical geneticist and an ethics advisor that this is in a participant's best interests, and that the participant's interests in having the information feedback to them overrides the participant's autonomous interests (expressed by signing their consent form). This feedback will usually be mediated via a health professional involved in the direct care of the participant.

6.9 Depletion of finite biological samples

Tissue samples collected from the study members of the CLS cohorts are a finite resource. The principles applied in depleting biological samples include the scientific strength of the proposal and the appropriateness of the methodology proposed. CLS holds responsibility for applying these principles to specific applications for the depletion of finite biological resources, as follows:

- All applications to use samples should demonstrate a clear scientific rationale regarding why the study is appropriate to the proposed research, and for non-renewable samples, that the use of samples is justified by the expected contribution to the scientific body of knowledge. Applications that demonstrate a unique dependence on the study, for example use of longitudinal data not widely available, are preferred.
- Appropriate ethical approval must be in place and all applications must comply with relevant legislation, e.g. the Human Tissue Act.

- Scientific strength, novelty and potential health/social impact of the research proposal must sufficiently justify use of longitudinal study samples.
- Evidence must be provided to show methodology is appropriate to the processing history of the samples, e.g. published literature or pilot data.
- The assay test platform should have proven quality assurance measures in place, preferably in accredited facilities according to ISO standards.
- The assay strategy should aim for maximum research impact with minimal depletion of the resource.
- The methodology should include measures to ensure the quality of any remaining sample is not jeopardised and can be used in further assays.
- All data generated from samples must be returned to the study and made available to other users within an agreed timeframe

For the NCDS cohort these principles are described in detail in 1958 Birth Cohort Biosample Strategy Guidelines (see Appendix G of the 1958 Policy Document).

7. CLS data access via public data repositories

The CLS data access via public data repositories are summarised on the table below:

Repository	Data	Method of data access
UK Data Service	CLS survey and linked administrative data	• Tier 1, End User Licence: download following registration • Tier 2a, Special Licence: download following approval • Tier 2b, Secure Access: remote access via UKDS secure lab following approval
SAIL Databank	MCS survey linked to Welsh administrative data	• Tier 2b, IGRP licence: remote access to SAIL server following approval
WTCCC / EGA	Genetic data only	• Tier 2a, EGA licence: download from EGA following approval

Access to data not available from public repositories can be requested to the CLS Data Access Committee, please see section 8.

7.1 UK Data Service

CLS data deposited at the UK Data Service falls into one of three categories listed above, and each impact level of data has its own access mechanisms.

End of User Licence

The majority of users will apply to use Tier 1 data via a standard licence known as an 'End User Licence' (EUL). Their application is authorised directly by the UK Data Service. Further details of the conditions of use are available at http://ukdataservice.ac.uk/get-data/how-to-access/conditions.aspx. Some datasets may also have special conditions, see http://ukdataservice.ac.uk/get-data/how-to-access/conditions/specialconditions.aspx

Special Licence

Access to more detailed data, classified as Tier 2a, which are potentially disclosive of the identities of individuals, households or organisations is provided via a special licence. Applications to UKDS for Special Licence data are referred by UKDS to the CLS Data Access Committee for authorisation. Further details of the conditions of use at available at http://ukdataservice.ac.uk/get-data/how-to-access/conditions/special-licence.aspx

Secure Access

For a small number of users access to Tier 2b sensitive and/or disclosive data is provided via 'Secure Access' through the UK Data Service. Applications to UKDS for Secure Access data are referred by UKDS to the CLS Data Access Committee for authorisation. Further details on applying for access to data under this category are available at: http://ukdataservice.ac.uk/use-data/secure-lab/about.aspx

7.2 SAIL Databank

The SAIL Databank is a data repository powered by the UK Secure e-Research Platform (UKSeRP), developed by the Health Informatics Group at Swansea University, with support from the Farr Institute of Health Informatics Research funded by Medical Research Council.

SAIL stands for Secure Anonymised Information Linkage, and offers secure storage and use of linked administrative data about the population of Wales.

SAIL has been used for access of MCS data linked to Welsh health and education: https://data.ukserp.ac.uk/Asset/View/52

7.3 European Genome-phenome Archive (EGA)

NCDS genotype data linked to region, sex and ethnicity are publicly available from the European Genome-phenome Archive (EGA):

https://ega-archive.org/dacs/EGAC00001000205

Access to these data requires an application to the Sanger Data Access Committee at datasharing@sanger.ac.uk

https://ega-archive.org/dacs/EGAC00001000205

8. CLS Data Access Committee

8.1 Responsibilities

The CLS Data Access Committee (DAC) was established in 2015 to define the principles for access to CLS data, and to apply these principles.

The CLS DAC is responsible for:

1. CLS data access policies and procedures, ensuring that of all of the CLS data access routes are promptly reviewed, fully documented and reported to the committee on a regular basis;
2. Classification and access of CLS data according to the CLS Data Classification Policy, reviewing any changes regarding data classification schemas already applied;
3. Ensuring that applications for data deposited at the UKDS under Special Licence (Tier 2a) and Secure Access (Tier 2b) are treated equitably across studies and approved by the delegated CLS staff;
4. Evaluation and approval of applications for CLS research data not available via public repositories such as the UKDS, SAIL Databank or WTCCC/EGA;
5. Evaluation and approval of applications for novel data linkages to the CLS studies;
6. Evaluation and approval of applications for data enhancements to the CLS studies;

For more details on the CLS DAC, please refer to the CLS DAC Terms of Reference, which sets out the responsibilities, membership and mode of operation of the committee.

8.2 Assessment criteria for CLS DAC approval

The criteria that the CLS DAC follows to assess data sharing applications is listed below. Projects must:

- Aim to carry out *bone fide* research and be led by a senior researcher

- Make the data available under the appropriate licence depending on the potential disclosivity and/or sensitivity of the data.
- Provide information on the public benefit of the proposed research, including potential scientific and wider impacts
- Request data that fall within the project remit: the amount of data requested must be justified in line with the research objectives described in the application
- Be very unlikely to damage the welfare of the study participants
- Be unlikely to bring disrepute to the cohort study and negatively impact of the conduct or reporting of the research on public perception or the future viability of the data collection.
- Ensure that data are held securely in the recipient institution. Those applications that are requesting linked NHS Digital health data must have the necessary organisational security assurance
- Ensure that data custodianship is taken into account. This is particularly relevant in the case of linked administrative data, which might requires the approval of the data providers
- Demonstrate the benefits to healthcare provision, adult social care or the promotion of health, for those projects which are applying to access linked health data for those applications which are requesting linked NHS Digital health data.

8.2 Data access requests

Where data arising from CLS studies are not available via the UKDS or other data repositories, applications will be considered by the CLS DAC.

The CLS Data Access application form and guidelines can be accessed at https://cls.ucl.ac.uk/data-access-training/access-cls-dac/

Data access applications may include (but are not restricted to):

- **Main survey data:** some data collected by the CLS cohort studies have not been deposited at the UKDS due to their sensitive nature or because they have not been sufficiently processed to be deposited.
- **Existing linked administrative records:** CLS has a programme of record linkages underway and the data are made available through the UKDS. However, some of these linked data have not been deposited at the UKDS due to their sensitive nature, but are available for research purposes under strict secure conditions from the CLS premises.
- **Paradata:** CLS holds data about the data collection process from some data collection sweeps. These are collected primarily for administrative purposes and are not routinely released for research use.
- **Genetic data linked to survey data:** CLS has a programme of genetic data collection. Given the sensitivity of the genetic data once they are combined with survey data, these requests are subject to a separate release arrangement that requires the creation of a bespoke survey dataset identified by a project-specific ID, thus being classified as Tier 2a.
- **Biological samples:** CLS has a resource of biological samples stored at the University of Bristol. Access to these samples can be requested for genotyping or generation of other analytes.

The CLS DAC will take into account the CLS resources required to deliver the data, as well as any risk posed to CLS' ability to deliver on its core mission or other existing commitments. This will be balanced against the potential public (scientific and wider) benefit of the request.

Following the principles of data minimisation, researchers will be only provided with the data needed to carry out their research projects.

8.3 Novel record linkage requests

CLS has a programme of record linkages underway, which covers a wide range of external data such as health, education, and geographical and economic indicators linked to its four longitudinal studies. As part of this programme of work, CLS welcomes proposals to perform additional data linkages. Proposals may refer to linkages with external data sources such as:

- Geographical data
- Education
- Health
- Economy
- Other

The CLS Record Linkage application form and guidelines can be accessed at https://cls.ucl.ac.uk/data-access-training/data-enhancements/

Once the approval has been granted, the data linkage can be either be carried out by the data applicant or by CLS, depending on the nature of the request and resources needed.

The data linkage will be performed using the linkage identifiers held in the UCL Data Safe Haven.

8.4 Data enhancement requests

CLS welcomes proposals for data enhancements to its four cohort studies. These data enhancements may relate to collection of new or additional qualitative or quantitative data and may take the form of:

- **New data collection** beyond the existing survey instruments, either at a sweep or between sweeps
- **Additional questionnaire/survey time** within an existing survey instrument
- **Not transcribed or digitised legacy data:** some data collected in earlier sweeps of the 1958 and 1970 cohorts have not yet been digitised from original paper questionnaires. Such legacy data can be digitised and/or processed as a data enhancement project.

Data enhancements may apply to the full sample or to a sub-sample of the cohort. They may relate to collection of new or additional qualitative or quantitative data.

The CLS Data Access application form and guidelines can be accessed at https://cls.ucl.ac.uk/data-access-training/data-enhancements/

8.5 Methods of data release

Following the CLS DAC approval, the mechanisms of access vary depending on the classification of the data approved for release.

Tier 1 and tier 2a data

Data with low or medium risk of disclosivity (tiers 1 and 2a) are released directly to the applicants once the signed CLS Data Sharing Agreement has been received. This includes the bespoke genetic identified with a project ID.

Tier 2b data

Potentially disclosive or sensitive data (tier 2b) are deposited on the applicant's UKDS Secure Lab account.

Tier 3 data

Highly disclosive data (tier 3), such as postcodes or school identifiers, cannot be made available outside of the UCL Data Safe Haven. The CLS DAC is responsible for judging whether access to such data within CLS premises could be granted in highly exceptional circumstances and for coordinating suitable arrangements for this.

The release of samples will be governed by the CLS Material Transfer Agreement.