CENTRE FOR LONGITUDINAL STUDIES

UCL

CLS Data Access Framework

Document information	
Document name	CLS Data Access Framework.
Authors	Aida Sanchez and C. Yogeswaran
Version	4
Issue date	March 2023
Approved by	CLS Data Access Committee
Review frequency	Yearly

This document includes data that is **PUBLIC** and can be disclosed outside UCL CLS and used or disclosed in whole or in part for any purpose other than to evaluate and implement procedures defined within this document.

Table of contents

Abbreviations

BCS70	1970 British Cohort Study
CLS	Centre for Longitudinal Studies
DAC	Data Access Committee
DSA	data sharing agreement
DSH	Data Safe Haven
EGA	European Genome-phenome Archive
ESRC	Economic and Social Research Council
EUL	End User Licence
GENDAC	genetic DAC
GDPR	General Data Protection Regulation
MCS	Millennium Cohort Study
MTA	Material Transfer Agreement
NCDS	National Child Development Study or 1958 Birth Cohort Study
NHS	National Health Service
ONS	Office for National Statistics
SAIL	Secure Anonymised Information Linkage
TRE	trusted research environment
UCL	University College London
UKDS	UK Data Service
UK LLC	UK Longitudinal Linkage Collaboration
UKSeRP	UK Secure e-Research Platform

1. 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.

All access to and use of CLS research 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 programme 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; and iii) the research-specific consent agreements and undertakings given to the cohort members are complied with.

The secure dissemination and access of CLS longitudinal research data is mainly carried out via the UK Data Service (UKDS) and to a lesser extent via other national Trusted Research Environments such as the SAIL Databank and UK LLC. In addition, the CLS DAC manages requests and access to CLS biological samples, genetics data, and other data not yet disseminated via data repositories.

CLS research data are categorised and shared differently depending on their sensitivity and potential risk of disclosure. This has been described in the CLS Data Classification Policy, with each category ('tier') having a defined access mechanism.

2. Scope

The CLS Data Access Framework ensures that data produced by CLS is made widely available for research purposes, nationally and internationally, to maximise the impact of CLS studies. At the same time, it is necessary to ensure that: sensitive or potentially disclosive data are shared in a secure manner; legal requirements, ethical guidelines, and moral responsibility to the study participants are maintained; and 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 the Economic and Social Research Council (ESRC), but also any co-funded add-on studies.

It builds on existing agreements developed by the ESRC, the UKDS, and other data sharing platforms 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 CLS Data Access Framework is a public document available to all potential users and sits alongside the following CLS data governance policies:

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

The Framework has been developed by the CLS DAC 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.

3. Definitions

“Anonymisation” is the process of rendering data into a form which does not identify individual living natural persons or makes the risk of re-identification sufficiently low in a particular context so that it does not constitute personal data. It requires removal of all personal identifiers, direct and indirect, and deals with the ‘data environment’ in such a way that the risk of somebody being identified in the data is negligible. Anonymisation can be reversed when someone with appropriate supplementary data can gain access and perform the necessary data integration to re-identify some or all people in the dataset. Truly anonymised data do not fall within the scope of the UK GDPR.

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

“Data environment” is the context in which the data is accessed, which can be characterised by four parameters: who accesses the data; the additional data that can be integrated with the data; the infrastructure in which the data are stored and processed; and the governance of the data.

“Data processor” 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. Processing implies some inputs into decisions on how the data/samples are used and by whom, and also responsibility for safeguarding the interests of the cohort members.

“De-identification” is the technical process of manipulating a dataset to reduce the risk of identification of individuals, as this risk can be present based on the actual contents of the data, even if a pseudonymised ID has been used.

“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, 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 a particular diagnosis or condition.

“Personal data” are defined as any information relating to an identified or identifiable natural person. It may also refer to data relating to people who have died and to information given in confidence under the Duty of Confidentiality. Data are considered personal when an individual can be identified from those data or from other information in the possession of the data controller.

“Pseudonymisation” is the technical process of manipulation of a dataset by replacing direct identifiers of an individual with a unique identifier that does not reveal

their ‘real world’ identity. A broader definition is provided in Article 4 of the UK GDPR, whereby pseudonymisation is the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

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

4. CLS research data available for access

CLS is based at the UCL Social Research Institute, which is part of the UCL Institute of Education, and manages the collection, curation, and dissemination of data for four major national cohort studies: NCDS, BCS70, Next Steps, and MCS.

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

Most CLS data are available from its study page at the UK Data Service and other UK data sharing platforms. CLS staff must use these data sharing routes unless access is needed for technical quality analysis and quality control), for strictly methodological projects or to conduct policy-relevant research needed to raise awareness of how newly collected or linked data can shed light on important social or policy issues (aka “initial findings”).

Comprehensive information about CLS studies can be found on the CLS website.

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.

Most NCDS survey data are available from its study page at the UK Data Service (survey and biomeasures, sub-studies, COVID-19 surveys, harmonised data**).**

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.

Most BCS70 survey data are available from its study page at the UK Data Service (survey and biomeasures, sub-studies, COVID-19 surveys, harmonised data).

Next Steps (formerly 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.

Most Next Steps survey data are available from its study page at the UK Data Service (Survey Data, COVID-19 surveys).

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. 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.

Most MCS survey data are available from its study page at the UK Data Service (survey and biomeasures, COVID-19 surveys, harmonised data).

4.2 Linked administrative data

CLS has an existing programme of linkage to administrative data which is based on 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. Next Steps and MCS have also sought consent to link to records held by the Ministry of Justice.

Linked administrative data, suitably pseudonymised, are provided to researchers via the UKDS Secure Lab and other Trusted Research Environments with the agreement of relevant data providers. These linked data include education and health records in England, Scotland, and Wales.

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

National Child Development Study (NCDS)

NCDS has been linked to hospital records from England and Scotland, which are available from the UK Data Service Secure Lab (NCDS Linked Administrative Data).

British Cohort Study 1970 (BCS70)

BCS70 has been linked to hospital records from England and Scotland, which are available from the UK Data Service Secure Lab (BCS70 Linked Administrative Data).

Next Steps

The Next Steps data has been linked to hospital records from England, and to education records from the National Pupil Database (NPD), Individual Learner Records (ILR) and Student Loan Company (SLC). These linked data are available from the UK Data Service Secure Lab (Next Steps Linked Administrative Data)

Millennium Cohort Study (MCS)

The MCS data have been linked to hospital records from England and Scotland, and to education records from the National Pupil Database (NPD), including GCSE exam results. These linked data are available from the UK Data Service Secure Lab (MCS Linked Administrative Data)

4.3 Linked geographical data

CLS has a programme of non-consented linkage of publicly available data using geographical identifiers. All geographical identifiers except postcodes are available from the UK Data Service (Linked Geographical Data).

4.4 Genetics data

The 1958 National Child Development Study (NCDS) and Millennium Cohort Study (MCS) samples have been extensively genotyped. Genotyping of the 1970 British Cohort Study (BCS70) is ongoing.

Access to CLS genetic data linked to survey data will be available for request via the CLS DAC.

1958 National Child Development Study (NCDS)

There are several types of NCDS genetics data available for research purposes (genome wide, imputed data, epigenetics and exome). These data are available via different routes depending on whether they need to be used in combination with NCDS survey data.

Genotyping data only (i.e., not linked to any other NCDS data) can be accessed via the European Genome-phenome Archive (EGA). The GW data release approval is managed by the Wellcome Trust Sanger Data Access Committee (Sanger DAC), at datasharing@sanger.ac.uk.

1970 British Cohort Study (BCS70)

CLS will commission the genotyping of the BCS70 DNA.

Next Steps

CLS will commission the genotyping of the Next Steps DNA from Age 32.

Millennium Cohort Study (MCS)

The MCS genetic data available for research purposes include genotype data, imputed data, and the Illumina GenomeStudio Final. Access to all MCS genetics data is overseen by the CLS DAC.

4.5 Biological samples

A number of biological samples have been collected from study members of NCDS, BCS70 and MCS. CLS is the custodian of the blood samples, and the processing and storage of both original aliquots and residues was contracted to the University of Bristol (Bristol Bioresource Laboratories, Population Health Sciences).

Access to CLS biological samples is overseen by the CLS DAC.

1958 National Child Development Study (NCDS)

The 2002/3 Biomedical Survey for NCDS collected whole blood and saliva from cohort members.

DNA was extracted from whole blood and there is also a transformed lymphocytes collection. Both 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.

1970 British Cohort Study (BCS70)

The 2016-17 Biomedical Survey for BCS70 (Age 46) collected whole blood samples from cohort members.

Next Steps

Saliva samples will be collected from Next Steps participants during the Age 32 Survey. DNA will be extracted from the samples and genotyped.

Millennium Cohort Study (MCS)

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

Milk teeth are not currently available for access. In 2015-16, the age 14 sweep (MCS6) collected saliva from cohort members and both natural parents. DNA has been extracted from the saliva samples and genotyped. Unlike the NCDS transformed lymphocytes collection, this is non-renewable. Their processing and storage are contracted to the University of Bristol.

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. **Custodianship:** UCL is the data controller of research data generated by the CLS longitudinal cohort studies. CLS is responsible for the safekeeping of all data and biological samples, 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. Data control of linked administrative data is generally retained by the providers of these records.
2. **Wide data access:** CLS aims to make their research data as widely available as possible to maximise the impact of the studies, subject to security and confidentiality considerations.
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 of 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 CLS research data and metadata follow the FAIR data guiding principles of being findable, accessible, interoperable, and reusable.
7. **Data security:** UCL, which houses CLS at the UCL Institute of Education, has ultimate responsibility for data security. All issues relating to information security, organisational security, and data protection are a very high priority for CLS.
8. **Consent:** Access to the data and samples is granted in line with the terms of consent agreed with CLS participants. When assessing data access requests, the CLS DAC will consider whether the proposed research is consistent with assurances given to cohort members when they gave informed consent.
9. **Management of disclosure and sensitivity risks:** Sensitive and/or disclosive data require an appropriate degree of security and management and will be

made available under strict levels of access to bona fide research that can demonstrate public interest. The CLS Data Classification Policy describes how CLS research data are categorised according to their sensitivity and potential risk of disclosure.

10. **Assessment criteria of data sharing projects and applicants:** CLS will apply the set of criteria described in this document to evaluate and approve data access. Public benefit, including potential scientific and wider impacts of the proposed research must be justified when required.
11. **Minimal costs:** There is no cost for accessing CLS research, survey, and linked data. In the case of biological samples, recipient institutions will be expected to meet all the costs of sample handling, specimen transport, and data preparation in relation to their study.
12. **Data minimisation:** Researchers will only be provided with access to the research datasets needed for the approved research projects.
13. **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 on this 'Secure Lab Breaches Penalties Policy' published by the UK Data Archive.
14. **Controlled release of biological samples:** As a depletable resource, the use of biological samples will be carefully controlled, in order to optimise the long-term value of the resource.
15. **Data return:** If required, derived variable syntax, new data, and associated metadata generated by approved researchers must be made available to the research community via the UKDS or returned to CLS.

6. CLS data access processes and procedures

6.1 Custodianship

UCL houses CLS at the UCL Institute of Education. UCL is the data controller of NCDS, BCS70, Next Steps, and MCS cohort data.

CLS is responsible for the safekeeping of data and/or tissue samples, the control of their use, and eventual disposal (if required), all in accordance with legislation and the terms of 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 (e.g., NHS Digital, Department for Education). In cases where these organisations require individual scrutiny of applications for their data linked to CLS survey members, the decision will be referred to the CLS DAC and the original data provider.

As set out in the terms and conditions of the CLS Resource Centre Grant (an internal document), the ESRC maintains the right to transfer data control of these data to third-party providers on termination of the grant, or on material failure of CLS conducting the grant. Alternatively, the 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 very high priority. CLS bases its Information Governance policies and procedures on the requirements of GDPR and NHS Digital and is compliant with the NHS Data Security and Protection Toolkit. This compliance is overseen and managed by the CLS Information Governance Steering Group.

All personal and collected data are stored and processed within the UCL DSH.

The UCL Data Safe Haven and the data repositories used to disseminate CLS research data (the UKDS, SAIL Databank, EGA, and UKLLC) are compliant and accredited with the international information security standard ISO27001.

6.3 Data classification

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 classification ("tier") to data made available for research purposes; these data categories are determined based on a number of considerations, 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 the continuation of the study.

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

- Safeguarded data -Tier 1a: Low impact. These data have been pseudonymised and de-identified to reach a very low level of disclosure and sensitivity (e.g., participant self-reported survey).
- Special safeguarded data -Tier : Medium impact. These data are potentially disclosive or have a moderate sensitivity (e.g., medium level geographical indicators, child adversity data, genetic data).
- Controlled data - Tier 2: High impact. These data have a higher risk of disclosivity (e.g., detailed geographical indicators) and/or sensitivity (e.g., detailed linked health data).
- Special controlled 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.
- Confidential data – Tier 4: Personal identifiable data such as names or NHS number are never made available for research use.

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.

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

For data not available from data sharing platforms, researchers may apply directly to the CLS DAC, as described in section 8 of this document.

6.4 Assessment criteria for access to research data

The assessment criteria for access to research data are based on the CLS data access principles described in section 5 of this document.

Access to CLS safeguarded data - tier 1 data is granted to researchers who register at the UKDS and agree to the terms and conditions of the UKDS End User Licence.

Access to CLS research data classified as special safeguarded (tiers 1b) and controlled (tiers 2, and 3) is subject to approval by the CLS DAC. Projects must:

- Aim to carry out bona fide research and be led by a senior researcher. The project description and methodology should be clearly explained.
- Access the data under the appropriate licence depending on the potential disclosivity and/or sensitivity of the data.
- 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 (this is also known as data minimisation).
- Be very unlikely to damage the welfare of the study participants.
- Be unlikely to bring disrepute to the cohort study or to negatively impact the public perception of the study or the future viability of the data collection.
- Ensure that data are held securely in the recipient institution. Applications that are requesting linked NHS Digital linked health administrative 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 require the approval of the data providers.
- Benefits connected with healthcare, adult social care, or the promotion of health for those projects which are applying to access NHS Digital linked health administrative data. The legal basis for processing these data must be ’public task’.

6.5 Assessment criteria for access to genetic data

Access to pseudonymised genetic data linked to survey/biomedical data can potentially increase the data disclosure risk. Requests for these combined data demand careful assessment of the requested phenotypic data to enable secure analysis at an individual level.

In order to achieve the appropriate level of security, CLS DAC requests for genetic data combined with phenotypic data are subject to a more involved CLS DAC application process and require the creation of a bespoke phenotypic dataset specifically pseudonymised for each project.

In addition, CLS considers that the assessment of genetics research is subject to additional ethical considerations. Consequently, CLS seeks independent advice regarding ethical considerations related to the use of genetic data for research

purposes, in particular in relation to reporting of incidental findings (section 6.7), socially controversial research (section 6.8), and consent requirements being fulfilled.

6.6 Assessment criteria for access to biological samples

Tissue samples such as cell-line DNA and blood samples have been collected from the study members of the CLS cohorts and are a finite resource. This resource is described in detail in the CLS Biosamples Strategy.

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.

The CLS DAC holds responsibility for assessing requests that involve the depletion of finite biological resources. The assessment is based on the scientific strength of the proposal and the appropriateness of the methodology proposed, 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 the 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 2004.
- Scientific strength, novelty, and potential health/social impact of the research proposal must sufficiently justify the 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.

6.7 Incidental findings of clinical relevance during genetic research

A potential consequence of genetic testing and genome sequencing is that researchers and clinicians might find variants in known disease genes that may be of clinical relevance to cohort participants. This may occur with increased frequency as a result of genome sequencing, either limited to all protein coding regions of a subject's genome (whole exome sequencing, WES) or covering the whole genome (whole genome sequencing, WGS). These genetic variants may be unrelated to the project objective and found unintentionally (incidental findings) or be deliberately sought as likely pathogenic alterations in genes that are not apparently relevant to the original project (secondary findings). For the purpose of this policy, we use the term "incidental findings" to refer to both unexpected positive findings and secondary findings.

The current CLS position on incidental findings is that genetic information (regardless of its nature) will not be returned to CLS cohort members. This is communicated to them via the CLS Privacy notices, as follows:

> *"We will not provide you with feedback of the results of genetic (DNA) testing. These data are used for research and not clinical diagnostic purposes. This position is considered 'best practice' ethically given we cannot be certain about the clinical relevance of any individual person's results. However, scientific developments in genetics are happening rapidly and this policy will be regularly reviewed."*

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 change. It is possible that in the future it may become mandatory to report genetic results to participants if they are (i) of scientific validity, (ii) clinically significant, and (iii) if there is a clear benefit of reporting results after considering the potential negative consequences.

Findings that satisfy the three stated criteria may 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 requires that data applicants comply with the following requirements:

- *In their CLS DAC application form, applicants should report the likelihood of generating incidental findings and state whether they have a clinical geneticist available to assess such potential findings.*
- *During their data analysis they should inform CLS of any incidental findings found.*

6.8 Research in socially controversial areas

CLS has a risk management strategy for research in socially controversial areas, such as ethnicity, criminality, intelligence, or sexuality. This strategy refers to the role and requirements of the CLS DAC when evaluating and approving research projects in these areas and aims at mitigating the risk of reputational damage to the study and of alienation of cohort members, as well as facilitating risk-management by the data applicants. CLS risk management strategies include:

- **Ethical considerations**: CLS seeks independent advice regarding ethical considerations related to socially controversial research proposals.
- **Fair processing**: CLS has put in place a number of Privacy Notices and Frequently Asked Questions (FAQs) to document the research sharing process, data protection, and ethical considerations.
- **Mitigation of reputational damage**: the CLS DAC requires that data applicants carefully manage any interaction with the media ahead of publication of results.

CLS DAC data applicants must address the mitigation of reputational damage on their application form, as per the following CLS requirements:

1. The applicant will use careful and balanced language when reporting their project results in order to avoid misinterpretation or exaggeration of the findings. This applies to all forms of research dissemination, including, but not limited to, press releases, media interviews, social media content, and blogs. The CLS DAC may request a copy in advance of any press releases and/or scientific articles prior to their publication or dissemination.
2. Before agreeing to be interviewed by the media, we recommend the applicant receives appropriate training.
3. During the evaluation of applications, the CLS DAC reserves the right to suggest the preparation of additional supporting information, such as frequently asked questions (FAQs), to help ensure findings are reported accurately by the media and others. For example, see this and page 9 of this document. This documentation should aim at guarding against misinterpretation or misrepresentation of study findings, thus preventing the possibility of controversies stemming from press coverage. It should describe the research process and ethical considerations, be written in plain English (see Appendix 1) and be available online in an enduring location.

6.9 Commercial use

Commercial organisations can apply for access to CLS data and are subject to the standard CLS access procedures. Like all 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 research use and are in line with the consent wording collected from cohort members.

6.10 International access

International access to CLS data is paramount, and unnecessary barriers should not get in the way of such research. International access varies depending on the data sharing route:

UKDS: CLS safeguarded data available via UKDS EUL and Special Licence are available to all international research users. However, UKDS controlled data available via Secure Access data can only be accessed by researchers based in the UK.

Other data repositories (i.e., the SAIL Databank, EGA, UK LLC): These data are available to all international researchers.

CLS data access:

- Data released directly to the applicants: we will issue an international data sharing agreement (DSA) for all data releases outside of the UK to ensure that research data will be processed lawfully. In addition, UCL Research Contracts will carry out the country safeguard analysis for applications from countries outside of the EU and not included in the European Commission list of adequate countries.
- Data released via the UCL DSH: these data can be made available by CLS to all international researchers following the creation of a DSH account, with no outbound rights.

6.11 Exclusive data access

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:

- inform CLS of the data outputs they have generated. Typically, these are new derived variables or linked datasets.
- make their data outputs available for re-use by the research community, according to the terms of the ESRC's Research Data Policy.

6.12 Data access by CLS researchers

CLS researchers do not have preferential internal access to CLS research data as a matter of course, so they must access the data in the same way as the rest of the

research community, for example, via the UK Data Service or other data sharing routes.

However, there are some scenarios under which CLS data is released internally to CLS researchers via central UCL servers. This internal access, which must be clearly justified to the CLS Data Access Committee, is only allowed for:

1. quality analysis (QA) or quality control (QC) purposes prior to data sharing;
2. strictly methodological projects;
3. policy-relevant research needed to raise awareness of how newly collected or linked data can shed light on important social or policy issues (aka “initial findings”)

Some CLS researchers need to access to linked administrative data (e.g., NPD, NHS HES, etc.) via the UCL Data Safe Haven. This internal data access will only be approved by the CLS DAC if it is allowed by the Data Provider (e.g., the DfE, NHS Digital) and only for the reason agreed with them. For instance, the Department for Education allow internal access for methodological projects, and NHS Digital allow internal access for methodological projects that describe and assess the quality of the processed linked HES data and/or their benefit to health and social care.

If the internal data access cannot be justified, the applicant will need to apply for access via the UK Data Service or other data sharing routes.

6.13 Sharing of derived variables and other data outputs

Researchers are expected to inform CLS of the data outputs they have generated. Typically, these are newly computed derived variables, coded responses, or linked data.

The relevant mechanism for the return and dissemination of these data outputs will be decided by the CLS DAC based on their contents and data classification. The two possibilities for data output sharing are:

a) Researchers to share syntax via the UKDS ReShare repository

CLS encourages researchers to share a description of their research derived variables and other data outputs themselves via the UKDS ReShare repository. This self-deposit repository for social science research data outputs has been developed as a means to effectively share their research data outputs with the wider academic community at the end of a project.

ReShare deposits should contain all the necessary documentation about new derived variables to enable other researchers to regenerate the same outputs if they wish. This documentation should be comprehensive to ensure that the outputs are reproducible, and should include project description, methods, coding frames and

protocols, syntax, data dictionaries, lookup tables with publicly available information, etc. Quality control is performed by the UKDS.

Please note that the ReShare deposits of CLS derived variable must not include the actual CLS individual-level datasets, as these are not covered by the UKDS ReShare licence agreements needed for safe and auditable onward sharing.

CLS will work with the UKDS to ensure that all CLS-related deposits made available via the ReShare repository are referenced on the main UKDS webpages relating to CLS studies.

b) Researchers to return datasets and syntax to CLS

The CLS DAC might consider that newly generated data outputs should be made available as part of the main CLS data resource. Therefore, if requested by the CLS DAC, researchers will need to provide CLS with the datasets, associated metadata, and related documentation.

Depending on the nature of the data, CLS will either deposit these data as an integral part of CLS data at the UKDS or make them available via the CLS DAC. The user guides of the data deposit will reference the researchers as the creators of the data outputs.

7. CLS data access via external data sharing platforms

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

Repository	Data	Method of data access
UK Data Service	CLS survey and linked administrative data	• Safeguarded data Tier 1a: End User Licence (EUL): download following UKDS End User Licence • Special safeguarded data - Tier 1b: download following UKDS Special Licence application and CLS DAC approval • Controlled data – Tier 2: remote access via UKDS Secure Lab following UKDS Secure Access application and CLS DAC approval
SAIL Databank	MCS survey linked to Welsh administrative data	• Controlled data – Tier 2: Remote access to the SAIL server following CLS DAC approval
UK LLC	CLS survey data linked to NHS Digital data	• Controlled data – Tier 2: remote access to the UK LLC server following CLS DAC approval
EGA	Genetic data only	• Special safeguarded data = Tier 1b: download from the EGA following DAC approval

Access to research data not available from these repositories can be requested via the CLS DAC; please see section 8 of this document.

7.1 UK Data Service

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

Safeguarded data under End User Licence (EUL)

The majority of users will apply to use safeguarded data - tier 1 via a standard licence known as an End User Licence (EUL). Their application is authorised directly by the UKDS. Further details of the conditions of use are available on the 'Types of data access' UKDS webpage.

Special Safeguarded data under Special Licence User Agreement (SL)

Special safeguarded data (Tier 1b) contain more detailed information than tier 1a safeguarded data, as they could be potentially more disclosive of the identities of individuals, households, or organisations.

Access to special safeguarded data is provided via the UKDS Special Licence. Applications to the UKDS for Special Licence data are approved by the CLS DAC. Further details of the Special Licence conditions of use are available here.

Controlled data under Secure Access User Agreement (SA)

Access to controlled data - tier 2 is provided via 'Secure Access' through the UKDS Trusted Research Environment called UKDS SecureLab.

Applications to the UKDS for Secure Access data are referred by the UKDS to the CLS DAC for approval. Further details on applying for access to UKDS Secure Access data are available here.

7.2 SAIL Databank

The SAIL (Secure Anonymised Information Linkage) Databank is a Trusted Research Environment for secure storage and access to linked administrative data about the population of Wales. It is hosted by the UK Secure e-Research Platform (UKSeRP), developed by the Health Informatics Group at Swansea University.

For CLS data, SAIL Databank provides access to controlled (tier 2) MCS data linked to Welsh health and education.

Applications for data access via SAIL Databank are referred to the CLS DAC for approval. Further details on applying for access to SAIL data are available here.

7.3 UK LLC

The UK Longitudinal Linkage Collaboration (UK LLC) allows researchers based within the UK to apply to access controlled data (tier 2) held within the UK LLC Trusted Research Environment (TRE).

For CLS data, the UK LLC provides access to controlled (tier 2) covid- related health data linked to CLS survey data.

Applications for data access via UK LLC are referred to the CLS DAC for approval. Further details on applying for access to UK LLC data are available here.

7.4 European Genome-phenome Archive (EGA)

NCDS genotype data linked to region, sex, and ethnicity are available from the European Genome-phenome Archive (EGA). To access these data an application must be submitted to the Sanger DAC. Their contact email address is: datasharing@sanger.ac.uk.

8. CLS data access via the CLS Data Access Committee

Access to research data not available from these data repositories can be requested via the CLS DAC.

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.

8.1 CLS DAC responsibilities

The CLS DAC was established to define and apply the principles for access to CLS data.

The CLS DAC is responsible for:

1. CLS data access policies and procedures, ensuring that 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 special safeguarded data (tier 1b) and controlled data (tier 2) 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 repositories such as the UKDS, SAIL Databank, UK LLC, or EGA.
5. Evaluation and approval of applications for novel data linkages to CLS studies.
6. Evaluation and approval of applications for data enhancements to CLS studies.

8.2 CLS DAC requests for data access

The CLS DAC will evaluate requests for CLS research data that are not available via the UKDS or other data repositories.

The CLS Data Access application form and guidelines can be accessed via the CLS webpages.

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

- **Main survey data:** Some data collected by 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 CLS premises.
- **Paradata:** CLS holds data about the data collection process from some sweeps. These are collected primarily for administrative purposes and are not routinely released for research use.
- **Genetics 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 data release arrangement that requires the creation of a bespoke survey dataset identified by a project-specific ID, thus being classified as special safeguarded data (tier 1b).
- **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 provided with only the data needed to carry out their research projects.

8.3 CLS DAC requests for novel record linkages

CLS has a programme of record linkages underway, which covers a wide range of external data such as health, education, 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 via the CLS webpages.

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

Data linkage will be performed using identifiers held in the UCL DSH.

8.4 CLS DAC data enhancement requests

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

- **New data collection** beyond existing survey instruments, either at a sweep or between sweeps.
- **Additional questionnaire/survey time** within an existing survey instrument.
- **Transcription or digitisation of 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 via the CLS webpages.

8.5 Release of research data approved by CLS DAC

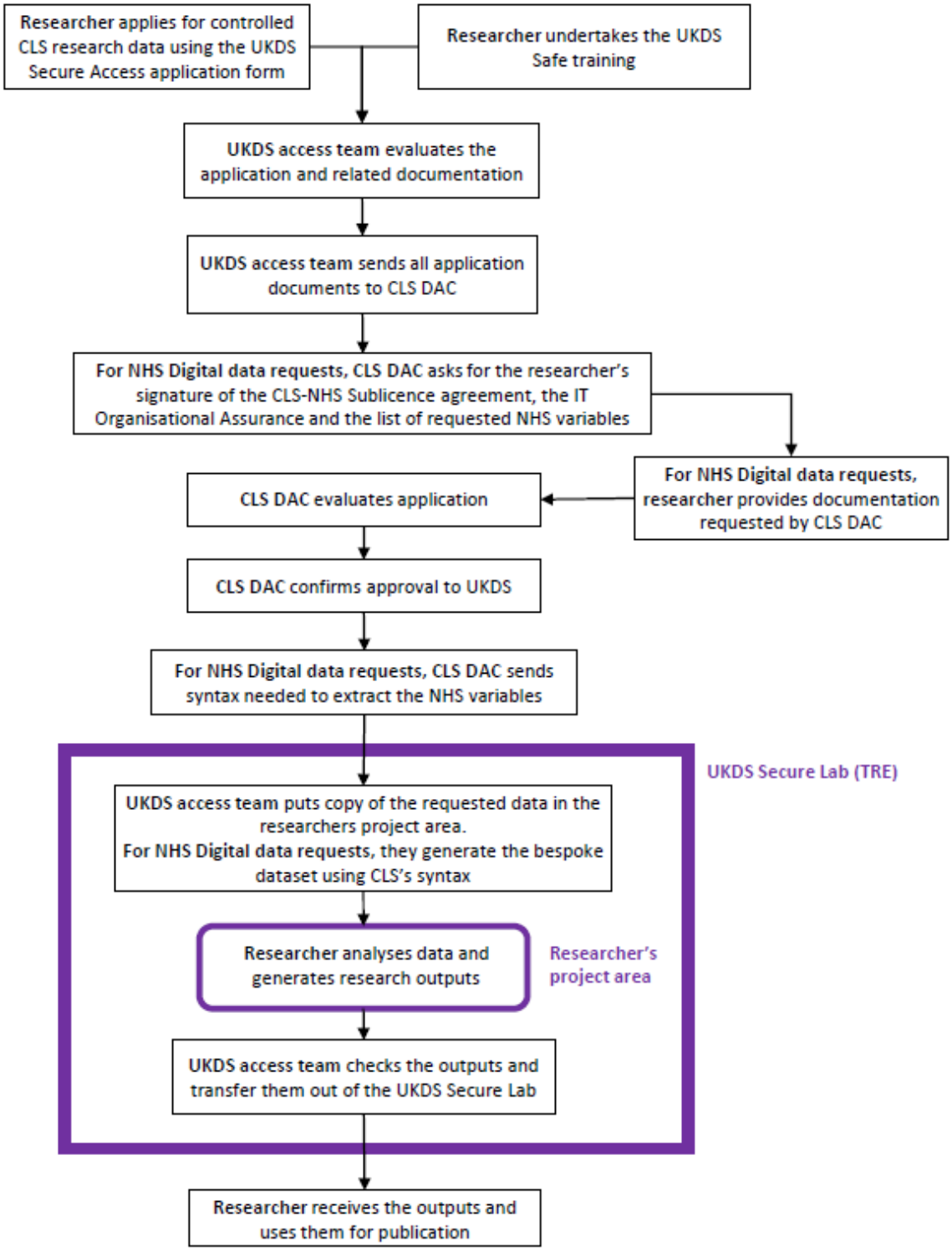
Following CLS DAC approval, the mechanisms of access vary depending on the classification of the data approved for release.

Data release to the applicant's institution

Safeguarded data (tiers 1a and 1b) are released directly to the applicants once the signed CLS DSA has been received. This includes the bespoke genetic dataset identified with a project ID, which is described in detail in section 8.6 of this document.

Data release to the applicant's UKDS SecureLab account

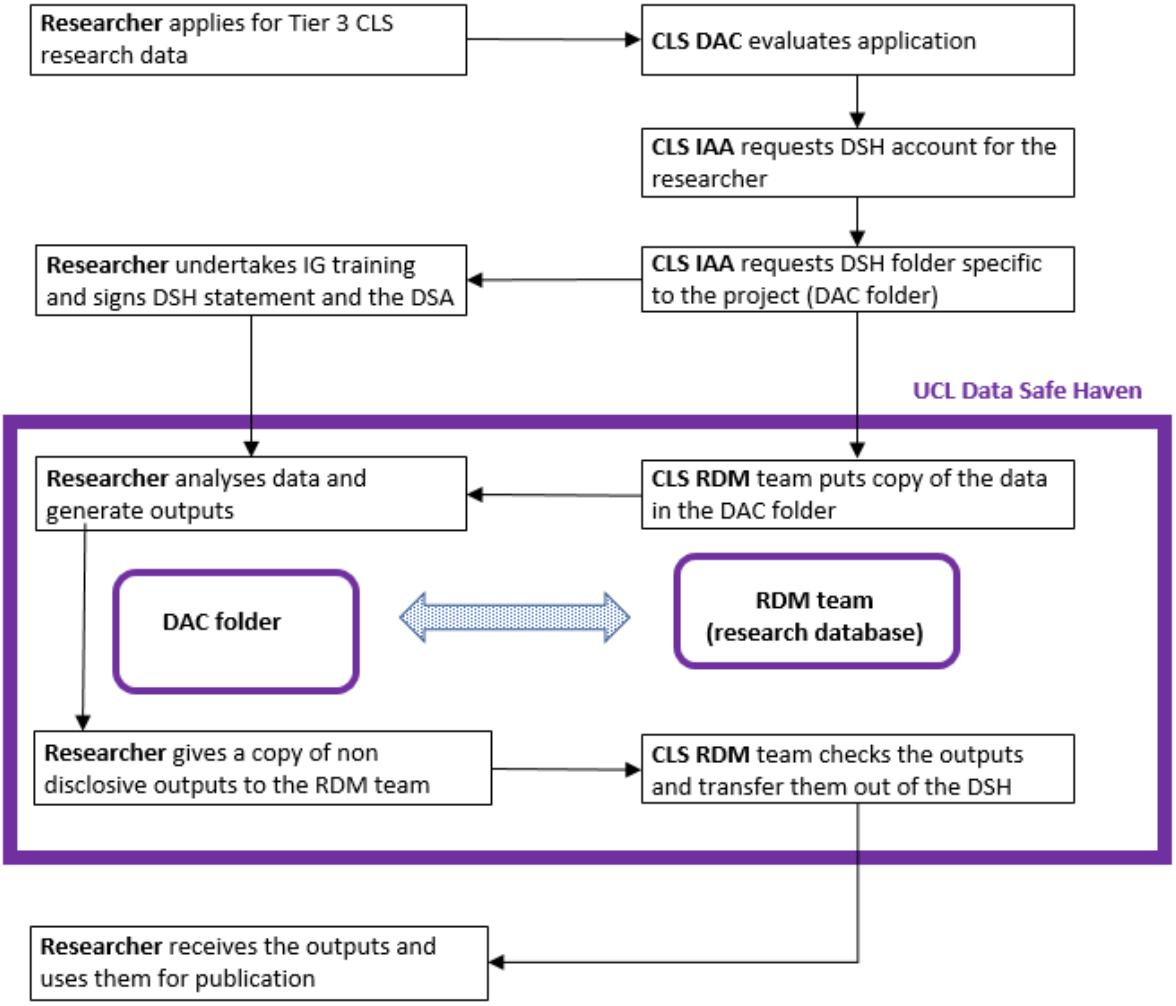
Controlled data – tier 2, which are potentially disclosive or sensitive data, are deposited on the applicant's UKDS Secure Lab project account. Information about setting up a UKDS Secure Lab account can be found on the UKDS webpages.


Researcher applies for controlled CLS research data using the UKDS Secure Access application form
Researcher undertakes the UKDS Safe training
UKDS access team evaluates the application and related documentation
UKDS access team sends all application documents to CLS DAC
For NHS Digital data requests, CLS DAC asks for the researcher's signature of the CLS-NHS Sublicence agreement, the IT Organisational Assurance and the list of requested NHS variables
For NHS Digital data requests, researcher provides documentation requested by CLS DAC
CLS DAC evaluates application
CLS DAC confirms approval to UKDS
For NHS Digital data requests, CLS DAC sends syntax needed to extract the NHS variables
UKDS Secure Lab (TRE)
UKDS access team puts copy of the requested data in the researchers project area.
For NHS Digital data requests, they generate the bespoke dataset using CLS's syntax
Researcher analyses data and generates research outputs
Researcher's project area
UKDS access team checks the outputs and transfer them out of the UKDS Secure Lab
Researcher receives the outputs and uses them for publication

Data release to the applicant's UCL Data Safe Haven folder

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

The CLS Research Data Management team will check the research outputs generated by the researchers in the UCL DSH by following the guidelines on Statistical Disclosure Control of outputs from UKDS.



The release of biological samples will be governed by the CLS Material Transfer Agreement (MTA).

8.6 Release of genetics data approved by CLS DAC

A research group may require access to a combination of survey data, biomedical phenotypes, genetic data, GWA genotypes, cell-line DNA, and blood samples. As discussed in section 6.5, access to genetics data linked to survey/biomedical data can potentially increase the data disclosure risk, so such applications demand careful linkage of the relevant data to enable secure analysis at an individual level.

The individual-level data from these different data sources must be linked together in a manner that prevents research applicants from identifying individual participants, either from the data they have been provided with, or by joining their data together with another user who has been provided with a different set of data.

a) Data minimisation: bespoke phenotypic datasets

As part of the CLS DAC data access application, researchers need to submit a list of survey data variable names they require to link to the genetic data. This data minimisation strategy offers additional protection to the genetic data, given their potential sensitivity and classification as special safeguarded data (tier 1b).

These variables need to be publicly available as safeguarded data under the UKDS End User Licence (EUL). The applicants will have to provide a summary of how this list of variables fits in with the project, but they do not need to justify how they intend to use every single variable (e.g., as exposure, confounder, outcome).

The final phenotypic dataset will be a bespoke dataset that only contains the exact list of variables requested by the applicant.

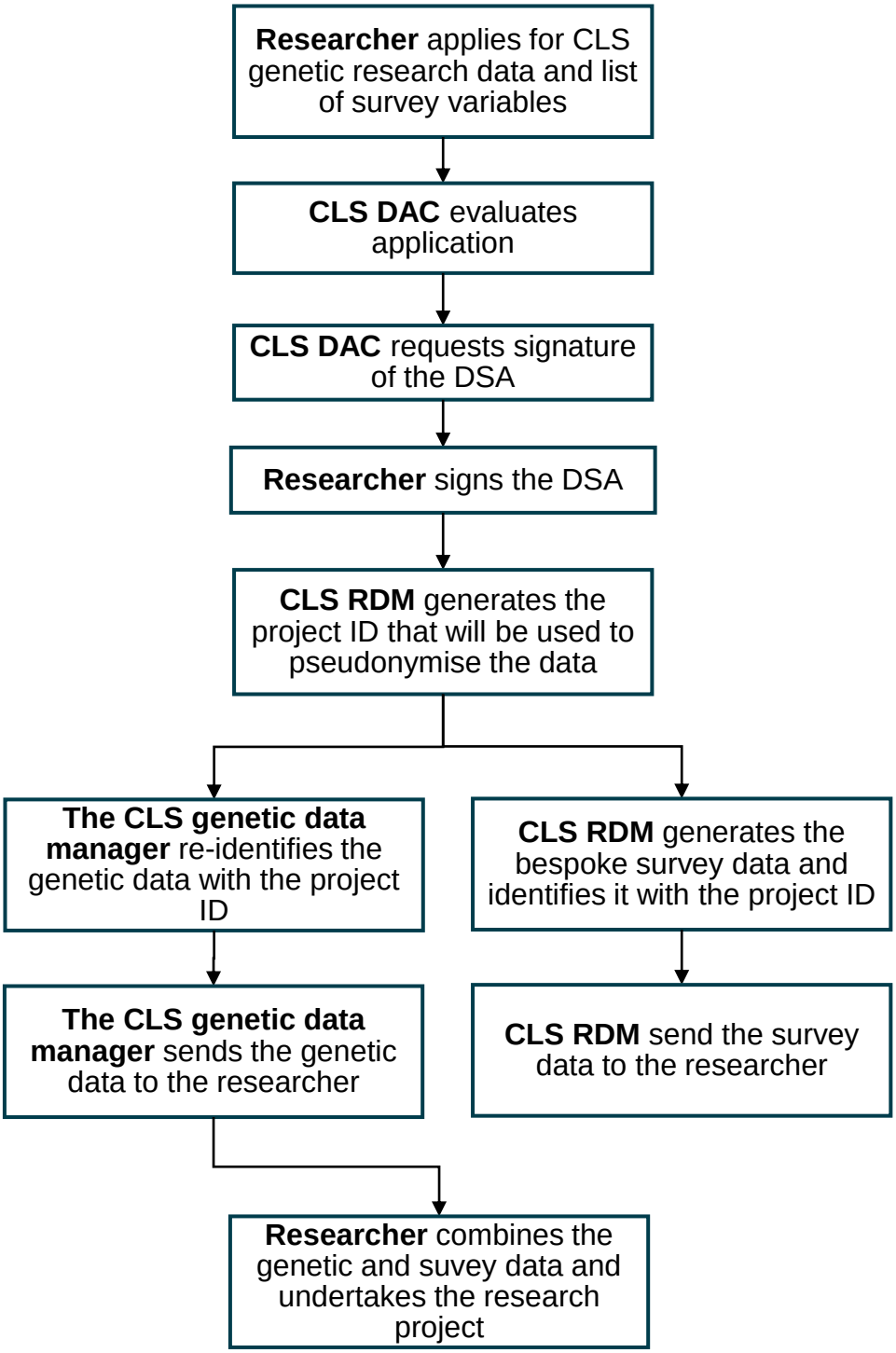
b) Pseudonymisation: newly created IDs

Once an application has been approved, a new ID will be created to identify the requested phenotypic and genetic data for every CLS DAC research team. This ID will always be different to:

- the public ID used to identify the study data deposited at the UKDS or other repositories
- the public ID used to identify the genetic data deposited at the EGA
- the IDs released to other CLS DAC applicants

Generic IDs or ID lookups will never be released to end-users.

The data flow is presented below.


Researcher applies for CLS genetic research data and list of survey variables
CLS DAC evaluates application
CLS DAC requests signature of the DSA
Researcher signs the DSA
CLS RDM generates the project ID that will be used to pseudonymise the data
The CLS genetic data manager re-identifies the genetic data with the project ID
CLS RDM generates the bespoke survey data and identifies it with the project ID
The CLS genetic data manager sends the genetic data to the researcher
CLS RDM send the survey data to the researcher
Researcher combines the genetic and suvey data and undertakes the research project

Appendix 1. CLS guidance on plain-language abstracts

The CLS DAC require the completion of a plain-language abstract (maximum 150 words) as part of all CLS DAC application forms.

A well written plain-language abstract is an essential part of data-related requests to the CLS DAC and subsequent approval process. One of the main reasons applications are delayed is that the plain-language abstract does not meet the requirements below or does not describe the project. Provision of a plain-language abstract is a condition of Tissue Bank Approval (ethical approval) for biobank resources.

The plain-language abstract will be published on the CLS website where it is available to study participants, the public, media, other researchers, and funders.

What is a plain-language abstract?

A plain-language abstract is a standalone lay summary of the proposed research project. It should not simply be copied from other project descriptions but needs to be written afresh.

The plain language abstract should not include any personally identifying information.

A plain-language abstract should use plain English that the cohort members would readily understand. It should be stand-alone interpretable. Consider using a readability checker such as this.

Avoid technical terms and jargon or explain them clearly if they are unavoidable. Examples of jargon are clinical and methodological terms, as well as words that have slightly different meanings in science rather than common use (e.g., ‘local’, ‘blind’, or ‘control’). Consider using a plain-language glossary such as this.

Above all, a plain-language abstract should clearly convey the key question and purpose of the project. The goal of writing in plain language is to enable readers to understand the content the first time they read it.

What should the abstract include?

Your abstract needs to address the following questions:

1. What is the research question? Why is it important?
2. How will the participants’ data be used to investigate the research question?
3. What is the method, in plain language?
4. What are the potential benefits or implications of your proposed research? This may include short-term outcomes or longer-term impact.

The abstract should focus on how CLS data contribute to your intended work. You must make sure the plain-language abstract is consistent with the scientific project description submitted for approval.

Who is the plain-language abstract for?

- The CLS DAC: The abstract explains the project for the Committee members, who all have different types of expertise.
- The longitudinal studies that provide the data and samples: Studies' leadership boards read the abstracts, to learn how the resource is used and to inform future strategy.
- The funders of CLS and its longitudinal studies: The funders want to know the scope and potential impact of the work that is being proposed.
- Researchers: The plain-language abstract for approved projects can be viewed online by other researchers. The research themes and broad methodology show what areas of research are already under investigation.
- Study participants: Participants can see how their personal data contributes to current knowledge. They need to understand what questions are being researched, how their data is contributing to this, and the potential benefits of the work – without getting bogged down in technicalities.

Tips for writing in plain language

- Limit sentences to one key point.
- Use short paragraphs.
- Be careful with words or phrases with dual or nuanced meanings (e.g., 'drugs' or 'diet').
- Avoid technical words, jargon, or words that are long or have many syllables. Consider those who do not have English as a first language.
- Avoid unnecessary technical details if you can make the same point in plain language.
- If you must use technical vocabulary, provide a short definition of your term when it is first introduced and do not use too many technical words together in one sentence.
- Do not include citations to research literature.
- Avoid more than two technical words in a sentence unless you explain them.
- Consider introducing an acronym or shorter term for repeated use.
- Write for an international audience. Avoid words or terms that are region-specific (e.g., 'A&E' versus 'ER').

- Use the active voice (for example, use “previous research showed that…” rather than “it was shown in previous research that…”).
- Keep within the word limit of 150 words.

Sources

This appendix is based on the guidance developed by study participant members of the former METADAC with the Secretariat and is based on (but not limited to) the following resources*:

- Cochrane Reviews Guidance
- National Institute for Health Research (INVOLVE)
- Access to Understanding
- The Plain Language Campaign

* Applicants may wish to use these resources for additional guidance – for example, the Plain Language Campaign link has dictionaries of alternative terminology.