CENTRE FOR LONGITUDINAL STUDIES

CLS Data Access Framework

Document information	
Document name	CLS Data Access Framework.
Authors	Aida Sanchez and C. Yogeswaran
Version	5
Issue date	October 2024
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 to evaluate, implement, and ensure compliance with the procedures defined within this document but not for any other purpose.

Table of contents

ABBREVIATIONS i
1. EXECUTIVE SUMMARY 1
2. SCOPE 2
3. DEFINITIONS 3
4. CLS RESEARCH DATA AVAILABLE FOR ACCESS 5
4.1 SURVEY DATA 5
4.2 LINKED ADMINISTRATIVE DATA 6
4.3 LINKED GEOGRAPHICAL DATA 7
4.4 GENETICS DATA 7
4.5 BIOLOGICAL SAMPLES 8
5. CLS DATA ACCESS INFRASTRUCTURE AND PROCESSES 9
5.1 CLS DATA ACCESS PRINCIPLES 9
5.2 CUSTODIANSHIP 10
5.3 DATA SECURITY 11
5.4 DATA ACCESS COMMITTEE 11
5.5 DATA CLASSIFICATION 12
5.6 SUMMARY OF CLS DATA ACCESS ROUTES 13
6. CLS DATA ACCESS CRITERIA 14
6.1 ASSESSMENT CRITERIA FOR RESEARCH DATA ACCESS 14
6.2 ASSESSMENT CRITERIA FOR GENETIC AND PHENOTYPIC DATA ACCESS 15
6.3 ASSESSMENT CRITERIA FOR BIOLOGICAL SAMPLES ACCESS 16
6.4 OUTCOMES OF CLS DAC ASSESSMENT 16
6.5 INCIDENTAL FINDINGS OF CLINICAL RELEVANCE DURING RESEARCH 18
6.6 RESEARCH WITH DISCLOSIVE DATA OR IN SOCIALLY CONTROVERSIAL AREAS 19
6.7 PUBLICATION OF QUALITATIVE RESEARCH DATA 20
6.8 COMMERCIAL USE 20
6.9 INTERNATIONAL ACCESS 21
6.10 EXCLUSIVE DATA ACCESS 21
6.11 DATA ACCESS BY CLS RESEARCHERS 21
6.12 SHARING OF DERIVED VARIABLES AND OTHER DATA OUTPUTS 23
7. DATA ACCESS VIA EXTERNAL DATA SHARING PLATFORMS 24
7.1 UK DATA SERVICE 24
7.2 SAIL DATABANK 26
7.3 UK LLC 26
7.4 EUROPEAN GENOME-PHENOME ARCHIVE (EGA) 26
8. DATA ACCESS VIA THE CLS DAC 27
8.1 DAC APPLICATIONS FOR SURVEY DATA OR PARADATA 27
8.2 DAC APPLICATIONS FOR GENETICS DATA 30
8.3 DAC APPLICATIONS THAT GENERATE DATA FOR ACCESS WITHIN UKDS SECURELAB 32
8.4 DAC APPLICATIONS TO BIOLOGICAL SAMPLES 32
9. REQUESTS FOR NOVEL CLS RECORD LINKAGES 34
10. REQUESTS FOR CLS DATA ENHANCEMENTS 34
APPENDIX 1. CLS GUIDANCE ON PLAIN-LANGUAGE ABSTRACTS 36
APPENDIX 2. LIST OF CLS DAC-APPROVED APPLICATIONS 38

Abbreviations

BCS70	1970 British Cohort Study
CLS	Centre for Longitudinal Studies
DAC	Data Access Committee
DfE	Department for Education (UK Government)
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 / genetic data application
GDPR	General Data Protection Regulation
ILR	Individualised Learner Record
MCS	Millennium Cohort Study
MTA	Material Transfer Agreement
NCDS	National Child Development Study or 1958 Birth Cohort Study
NHS	National Health Service
NPD	National Pupil Database
ONS	Office for National Statistics
SAIL	Secure Anonymised Information Linkage
SLC	Student Loan Company
TRE	trusted research environment
UCL	University College London
UKDS	UK Data Service
UK LLC	UK Longitudinal Linkage Collaboration
Usurp	UK Secure e-Research Platform

1. Executive summary

The Centre for Longitudinal Studies (CLS) is home to several national cohort studies. Our four core cohort studies are the National Child Development Study (NCDS, or the 1958 Birth Cohort Study), the 1970 British Cohort Study (BCS70), Next Steps and the Millennium Cohort Study (MCS). CLS is part of the UCL Social Research Institute and funded by the Economic and Social Research Council.

Access to and sharing of CLS research cohort 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 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 for novel data linkages and access to CLS research data, genetics data, biological samples, 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 is 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 research 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 regardless of funder.

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 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.

Bona fide organisation is an established academic institution, research body or organisation with the capability to lead or participate in high quality, ethical research. It is not a requirement that research is the primary business of that organisation, or that the organisation is publicly financed. In this context, a private partnership may qualify as a bona fide research organisation.

Bona fide research refers to high quality, ethical projects for research purposes using rigorous scientific methods. There must be an intention to publish the research findings for wider scientific benefit, without restrictions and with minimal delay.

Data controller is the entity 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 entity 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 IoE, UCL's Faculty of Education and Society, 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 research data are available from the UK Data Service. We also disseminate some research data via other UK data sharing platforms, such as the SAIL Databank, the UK Longitudinal Linkage Collaboration (UK LLC) and the European Genome-Phenome Archive (EGA).

CLS staff will access the CLS research data via these data sharing routes unless access is needed for internal purposes listed in section 6.15 of this document.

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

The list of applications approved by the CLS DAC can be found on the CLS data access webpage.

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.

Information about NCDS can be found on the CLS NCDS webpage. Most NCDS survey data are available from the NCDS 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.

Information about NCDS can be found on the CLS BCS70 webpage. Most BCS70 survey data are available from the BCS70 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.

Information about NCDS can be found on the CLS Next Steps webpage. Most Next Steps survey data are available from Next Steps 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.

Information about NCDS can be found on the CLS MCS webpage. Most MCS survey data are available from the MCS 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 SecureLab 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. Further information about this can be found on the CLS linked data webpage.

National Child Development Study (NCDS)

NCDS has been linked to hospital records from England and Scotland, which are available from the UK Data Service SecureLab (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 SecureLab (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), Individualised Learner Record (ILR), and Student Loan Company (SLC). These linked data are available from the UK Data Service SecureLab (Next Steps Linked Administrative Data)

Millennium Cohort Study (MCS)

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

- National Child Development Study, 1958 – Linked Geographical Data
- 1970 British Cohort Study – Linked Geographical Data
- Next Steps – Linked Geographical Data
- Millennium Cohort Study – Linked Geographical Data

Further information can be found on the CLS geographical data linkage webpage.

4.4 Genetics data

The 1958 National Child Development Study (NCDS), 1970 British Cohort Study, Next Steps and Millennium Cohort Study (MCS) samples have been extensively genotyped.

There are several types of CLS genetics data available for research purposes: genome wide, imputed data, epigenetics, exome and the Illumina GenomeStudio final report. These are different for each cohort. The CLS genetics data are described in the CLS Genomics Data webpage (GitHub).

Access to CLS genetic data only or linked to survey data are available for request via the CLS DAC.

NCDS genotyped data only (i.e., not linked to NCDS phenotypic data) and MCS whole exome sequencing data can also be accessed via the European Genome-phenome Archive (EGA).

4.5 Biological samples

Different types of biological samples have been collected from study members of NCDS, BCS70, Next Steps, and MCS. CLS is the custodian of the blood samples, and the processing and storage of both original aliquots and residues are 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. DNA was extracted from the samples and genotyped.

Next Steps

Saliva samples were collected in 2023-24 from Next Steps participants during the Age 32 Survey. DNA was 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 have been 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.

5. CLS data access infrastructure and processes

5.1 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:** All issues relating to information security, organisational security, and data protection are a very high priority for CLS. UCL, which houses CLS, has ultimate responsibility for data security.
8. **Consent:** Access to the data and samples is granted in line with the terms of consent provided by 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 researchers 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. CLS will evaluate the potential scientific and wider impacts of the proposed research. Where appropriate, CLS will also address its public benefit.
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 data needed for the approved research projects.
13. **Punishable violation of access conditions:** An appropriate set of penalties will 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 'SecureLab 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 CLS to share with the research community.

5.2 Custodianship

UCL houses CLS at the UCL Social Research Institute, which is part of the IoE, UCL's Faculty of Education and Society. UCL is the data controller of all CLS 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.

For linked administrative data, the data controllers are the data providers (e.g., NHS England, 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.

5.3 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 the UK GDPR and NHS England 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 by CLS staff within the UCL Data Safe Haven (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.

5.4 Data Access Committee

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

The CLS DAC Terms of Reference set out the responsibilities, membership, and mode of operation of the Committee.

The CLS DAC is responsible for:

1. Classification and access of CLS data according to the CLS Data Classification Policy, reviewing any changes regarding data classification schemas already applied.
2. 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.
3. Evaluation and approval of applications for CLS research data not available via repositories such as the UKDS, SAIL Databank, UK LLC, or EGA.
4. Evaluation and approval of applications for novel data linkages and data enhancements to CLS studies.

5.5 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 1b: 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.

5.6 Summary of CLS data access routes

Data	Classification (tier)	Platform	Request method	Licence	Approval	Data access method
De-identified survey data	Safeguarded (1a)	UKDS	UKDS registration and project description	UKDS End User Licence	UKDS	Download
Moderately sensitive/disclosive survey data	Special safeguarded (1b)	UKDS	UKDS Special Licence application	UKDS Special Licence	CLS DAC	Download
NCDS genetic data only	Special safeguarded (1b)	EGA	EGA application	EGA Data Sharing Agreement	Sanger	Download
De-identified data not available on data sharing platforms: genetics + survey, biosamples, paradata	Special safeguarded (1b)	CLS	CLS DAC application	CLS Data Sharing Agreement	CLS DAC	Release by CLS
Potentially disclosive and/or sensitive survey data	Controlled (2)	UKDS	UKDS Secure Access application	UKDS Secure Access	CLS DAC	Remote access via the UKDS SecureLab
Linked administrative data (health, education)	Controlled (2)	UKDS	UKDS Secure Access application	UKDS Secure Access	CLS DAC	Remote access via the UKDS SecureLab
Linked Welsh admin data (MCS)	Controlled (2)	SAIL Databank	SAIL application	SAIL Data Sharing Agreement	CLS DAC	Remote access via the SAIL secure server
Linked NHS England data	Controlled (2)	UK LLC	UK LLC application	UK LLC Data Sharing Agreement	UKLLC DAC and CLS DAC	Remote access via the UK LLC secure server
Disclosive data not available on UK data sharing platforms: postcodes, verbatim responses, organisation identifiers	Special controlled (3)	CLS	CLS DAC application	CLS Data Sharing Agreement	CLS DAC	Remote data access via the CLS Data Safe Haven

6. CLS data access criteria

6.1 Assessment criteria for research data access

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

Assessment criteria for access to CLS safeguarded data Access to safeguarded data (tier 1a) is granted to researchers who:

- Are affiliated with a bona fide organisation.
- Register at the UKDS and agree to the terms and conditions of the UKDS End User Licence.
- Are not personal users, including independent researchers conducting research for ‘personal use’.

Assessment criteria for access to special safeguarded and controlled data

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

- Aim to carry out bona fide research which is led by a senior bona fide researcher. Applicants must be affiliated with a bona fide organisation. CLS reserves the right to request the CV and list of publications/outputs for applicants on a CLS data application.
- Clearly explain the project description and methodology. Funding or institutional ethical approval are desirable but not mandatory to access CLS data.
- 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 and be in line with the research objectives described in the application (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.
- Ensure that access to linked administrative data have the approval of the data providers.
- Start within 2 years of DAC approval.

- Where relevant, such as applications involving controlled data or hypothesis-free approaches to research, demonstrate experience in handling sensitive data.

Assessment criteria for access to NHS England linked health data

In addition, for applications that are requesting linked NHS England health data must:

- Have the necessary organisational security assurance and be accessed from the UK.
- Describe the benefits connected with healthcare, adult social care, or the promotion of health for those projects.
- The legal basis for processing these data must be ’public task’.

6.2 Assessment criteria for genetic and phenotypic data access

Access to pseudonymised genetic data linked to survey/biomedical data can potentially increase the data sensitivity and disclosure risk. Requests for these combined data demand careful assessment of the requested phenotypic data to enable secure analysis at an individual level. For this reason, these data requests are subject to a more involved CLS DAC application process and require the creation of a bespoke phenotypic dataset identified with a project-specific ID.

In addition to the CLS assessment criteria listed above, genetics research applications are subject to an additional CLS ethical assessment. 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 6.5), socially controversial research (section 6.6), and consent requirements being fulfilled.

Considerations about ancestry and ethnicity in genetics research

The term ‘ethnicity’ historically refers to a person’s cultural identity, whereas ‘ancestry’ refers to a person’s country or region of origin or an individual’s lineage of descent.

The lack of non-European representation in datasets and research means that the MCS genetics data are currently restricted to samples from participants with White ancestry because the analysis on non-White/non-European populations is not yet sufficiently robust. However, 18% of the MCS participants are from non-White ancestries and using data drawn from exclusively Whites only ancestry might have a major impact on research findings and on the inferences that can be drawn around changes across UK generations.

Genetic researchers using MCS need be aware of these potential limitations and communicate them accordingly when they write proposals and present/publish findings. To this effect, we ask that genetic researchers provide a statement in section 8.iii (ethico-legal discussion) demonstrating that they understand this issue, discussing any mitigations they have made, and explaining how they will present and publish their findings in light of these limitations.

6.3 Assessment criteria for biological samples access

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.

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 for 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 that the 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.

6.4 Outcomes of CLS DAC assessment

The CLS DAC will assess the requests for access to special safeguarded and controlled data. The possible outcomes of this evaluation are:

- **Approved:** This decision is made when the DAC is satisfied with the application and give their approval for the project to proceed.
- **Conditionally approved:** Where the DAC is satisfied with an application but requires small adjustments to the application form prior to giving full approval. Examples of adjustments include:
    - Confirmation that a project has received ethical approval.
    - Explicit confirmation from the applicant that they will abide by certain conditions of the CLS Data Access Framework, e.g., to provide an FAQ or liaise with CLS before publishing any findings, clarify specific issues, etc.
    - Submission of applicant's CV.
    - Make small changes to the application form, for example, correcting typos, dates, etc.
- **Resubmission requested:** where the DAC is dissatisfied with several key aspects of an application or amendment, and request it is submitted again for re-evaluation by the DAC. For example:
    - The description of the research project and/or the methodology is unclear.
    - The data request is unclear or inaccurate.
    - The ethico-legal discussion is unsatisfactory or incomplete.
    - The requested data do not fall within the project remit.
    - An amendment refers to changes that fall far outside the scope of the initial project, in which case a new application may be requested.
    - Information about other issues such as incidental findings is incomplete or unsatisfactory.
- **Rejected:** *The* application does not comply with the CLS DAC assessment criteria or principles, or it is unfeasible.

6.5 Incidental findings of clinical relevance during research

Genetics 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, which may be 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.

Other research areas

Other types of non-genetic health, such as biomarkers, lifestyle data, and various biomeasures could indicate current or future disease. For example, accelerometry measures could indicate early Parkinson's disease, and impaired memory tests could indicate cognitive decline. Applicants should indicate on the CLS DAC application form, whether their research is likely to generate incidental findings which have potential clinical utility for participants and, if so, how these findings will be dealt with and mitigated.

Requirements for applicants

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 expert (e.g., clinical geneticist) available to assess such potential findings.
- Applicants should inform CLS of any incidental findings found during their data analysis.

Reporting incidental findings to participants

The current CLS position on incidental findings is that this information (regardless of its nature) will not be returned to CLS cohort members. This is communicated to them via the CLS Privacy notices with regards to genetic data, 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 or other specialized biomarkers 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 or specialized biomarker results.

6.6 Research with disclosive data or in socially controversial areas

CLS has a risk management strategy for research that uses data which can be potentially disclosive, for example geographical data or having a small number of samples, as well as for research in sensitive or 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

CLS 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 reserves the right to request the CV and list of publications/outputs for applicants on a CLS data application.

Researcher's risk management strategies

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

1. In their request, applicants should consider the ethical, sensitive and/or disclosive nature of the topics they are researching, and explain clearly what risks they

could raise, and how they will act to manage or mitigate these. The CLS DAC may ask for further reassurances on this.

2. 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.
3. Before agreeing to be interviewed by the media, we recommend the applicant receives appropriate training.
4. 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 reference, see an example of suitable research FAQs 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.

6.7 Publication of qualitative research data

Qualitative research data includes essays, open text responses, verbatim answers, raw and redacted interview transcripts, etc., which are potentially very disclosive and/or sensitive. CLS data sharing routes for these data are either via the UKDS or via the UCL Data Safe Haven.

Researchers will not be able to publish verbatim qualitative quotes. However, the publications of certain highly de-identified and redacted quotes will be allowed, following review and approval by the CLS DAC If approved, the researcher must include information about how others can access the data source via the relevant CLS data sharing route.

6.8 Commercial use

Commercial organisations can apply for access to CLS data and are subject to the standard CLS access procedures and assessment criteria. Like all applicants, commercial organisations must confirm that their use of the data is for bona fide research 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.

6.9 International access

The commitment to wide data access includes providing international access to CLS data and ensuring that unnecessary barriers do not get in the way of such research. International access varies depending on the data sharing route:

UK data sharing platforms:

- CLS safeguarded data available via UKDS (EUL and Special Licence) are available to all international research users.
- CLS controlled data available via Trusted Research Environments such as the UKDS SecureLab, SAIL Databank, or UK LLC can only be accessed by researchers based in the UK.

CLS DAC data access:

- CLS safeguarded 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. CLS will seek advice from UCL Research Contracts for applications from countries 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.

6.10 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. See section 6.12 for details.

6.11 Data access by CLS researchers

‘Internal researchers’ are those directly employed by CLS or the Co-Investigators of the CLS Resource Centre grant or other CLS major data collection grants.

CLS researchers may not access data internally if these data are already available via the UKDS. Access to raw genetics data is not permitted either.

However, CLS may grant CLS researchers with permission to access certain research data prior to wider release via the UK Data Service (UKDS). This affects data undergoing curation and technical processing for UKDS deposit, such as newly collected survey data (including pilot data or paradata), genetic derived variables, and linked administrative and geographical data.

CLS researchers will need to apply to the CLS DAC to access data prior to UKDS deposit. The reasons for requesting internal access need to be fully justified to the CLS DAC. This is done via a comprehensive CLS internal application form to ensure that it can be assessed against the CLS DAC approval criteria. All internal requests, approval and data releases will follow CLS DAC processes, and will be fully tracked for DAC record keeping and transparency purposes. Access is granted once need has been demonstrated.

The conditions below apply to internal access prior to the UKDS data deposit:

1) In the course of their project, researchers will commit to carry out data quality assurance on the data they are using, and report to the PI and data managers any potential errors in the data, suggest improvements, and provide copies of any new data and code they may have generated.
2) The specific purposes for internal access can only be:
    - Generation of new research data for UKDS deposit, such as weights, complex derived variables, geographical indicators, polygenic risk scores, etc.
    - Methodological projects that lead to evidence-based guidance on data handling and analysis of newly deposited data.
    - Initial findings research.
3) UKDS data deposit will never be delayed or held back from its scheduled deposit date in order to allow any such research arising from access granted to be completed or published.
4) Any published research will have to be against the data deposited. Therefore, the data analysis will need to be re-run against the version of the data that will be released for data sharing.
5) Any publication of research findings will not be submitted for publication or published prior to the data being made available for public sharing,

If internal data access cannot be justified, CLS researchers will need to access the data via the UK Data Service or other data sharing platforms.

6.12 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, or newly 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 via public syntax repositories or via CLS.

Sharing syntax via public repositories

CLS encourages researchers to share their derived variables syntax and other data outputs themselves via the UKDS ReShare repository or the Open Science Framework, using the *#BritishCohortStudies* tag in addition to the cohort-specific tags:

- *#NCDS*
- *#BCS70*
- *#NextStepsStudy*
- *#MCS*

These self-deposit repositories have been created to facilitate the effective sharing of data outputs with the broader academic community upon project completion and publication.

The shared data outputs should include all necessary documentation regarding new derived variables to ensure reproducibility for other researchers. This documentation must be comprehensive and should encompass the project description, methods, coding frames and protocols, syntax, data dictionaries, and lookup tables containing publicly available information, among other details.

Additionally, *the shared data outputs must not include CLS individual-level data*, as these are not covered by the repository's license agreements required for safe and auditable onward sharing.

Researchers to return derived variables and syntax to CLS

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

CLS will either share these data outputs as an integral part of CLS research 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. Data access via external data sharing platforms

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.

All UKDS data users must register and agree to follow the UKDS Research data handling and security guide for users.

Safeguarded data – End User Licence

The majority of users apply to use safeguarded data (tier 1a). Applicants must describe their research project and sign the UKDS general licence known as End User Licence (EUL). The details of the conditions of use are available on the UKDS End User Licence Agreement.

Special Safeguarded data – Special Licence

Special safeguarded data (tier 1b) contain de-identified but potentially disclosive or sensitive data. Access to these data must be requested via the UKDS Special Licence application form. All UKDS Special Licence applications must be approved by the CLS DAC. Details of the conditions of use are available on the UKDS Special Licence Agreement.

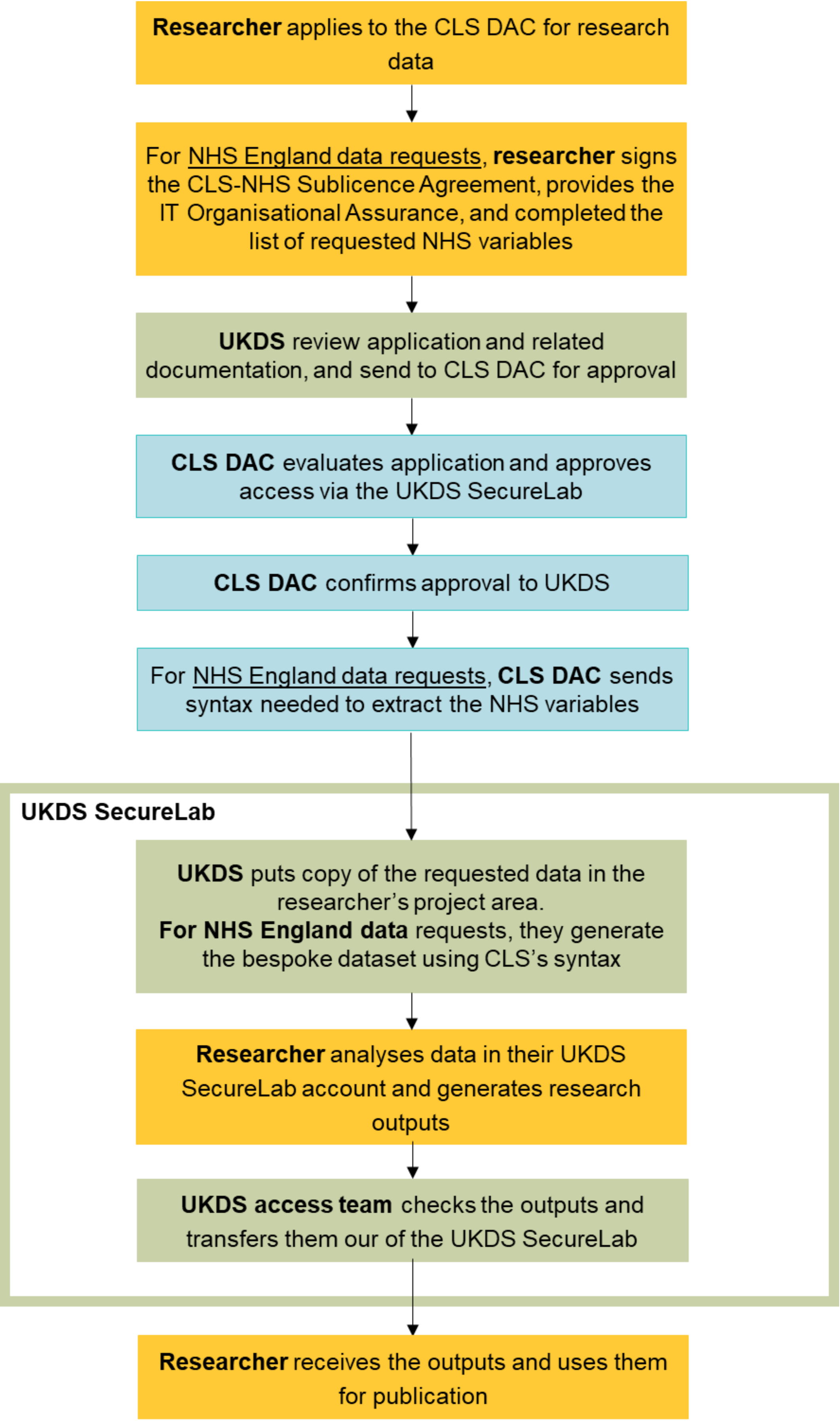
Controlled data – Secure Access

Access to controlled data (tier 2) is provided via ‘Secure Access’ through the UKDS Trusted Research Environment called UKDS SecureLab. The application process requires the following steps:

- Main applicant and all co-applicants to register with the UKDS.
- Main applicant to submit the Secure Access application as per the UKDS controlled data instructions.
- Main applicant to add each team member to the project.
- All applicants to ensure access to a device located at and owned by their institution, with a dedicated static IP address and a wired Internet connection. Evidence to be submitted to the UKDS.
- All applicants must complete UKDS Safe Researcher Training (SRT) course.
- Secure Access User Agreement to be signed by all applicants and by institutional signatory.

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 controlled data can be found on the UKDS website.

Figure 1. Access to controlled data via the UKDS SecureLab



Researcher applies to the CLS DAC for research data

For NHS England data requests, **researcher** signs the CLS-NHS Sublicence Agreement, provides the IT Organisational Assurance, and completed the list of requested NHS variables

UKDS review application and related documentation, and send to CLS DAC for approval

CLS DAC evaluates application and approves access via the UKDS SecureLab

CLS DAC confirms approval to UKDS

For NHS England data requests, **CLS DAC** sends syntax needed to extract the NHS variables

UKDS SecureLab

UKDS puts copy of the requested data in the researcher's project area.
For NHS England data requests, they generate the bespoke dataset using CLS's syntax

Researcher analyses data in their UKDS SecureLab account and generates research outputs

UKDS access team checks the outputs and transfers them our of the UKDS SecureLab

Researcher receives the outputs and uses them for publication

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 provides access to controlled (tier 2) MCS data linked to Welsh health and education. Applications for these MCS data access via SAIL Databank are referred to the CLS DAC for approval. Further details on applying for access 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).

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 are available here.

7.4 European Genome-phenome Archive (EGA)

NCDS-only genetic data 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. Data access via the CLS DAC

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

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

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

- Main survey data
- Paradata
- Genetics data
- Biological samples

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.

Researchers must allow sufficient time between submitting their application and when they plan to undertake research on the data requested. In cases where significant resources of the CLS data management team are required to fulfil the request, we suggest an application is submitted at least 3 months before the planned research will take place.Following the principles of data minimisation, researchers will be provided with only the data needed to carry out their research projects.

The CLS DAC data access application form and process can be found on the CLS data access webpage.

8.1 DAC applications for survey data or paradata

Some survey 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.

Paradata are data collected about interviews and the survey process, such as how long the interview took to complete, number of contact attempts, etc. CLS holds detailed paradata from several data collection sweeps. These are collected primarily for administrative purposes and are not routinely released for research use.

Access to survey data or paradata not available from UKDS can be requested via the CLS DAC. Following DAC approval and receipt of the signed CLS Data Sharing Agreement (DSA) the data management team will release the data

Direct data release

If the requested data have a low risk of disclosivity or sensitivity, that is, are classified as safeguarded data (tiers 1a and 1b), the data are released directly to the applicants by the Research Data Management Team.

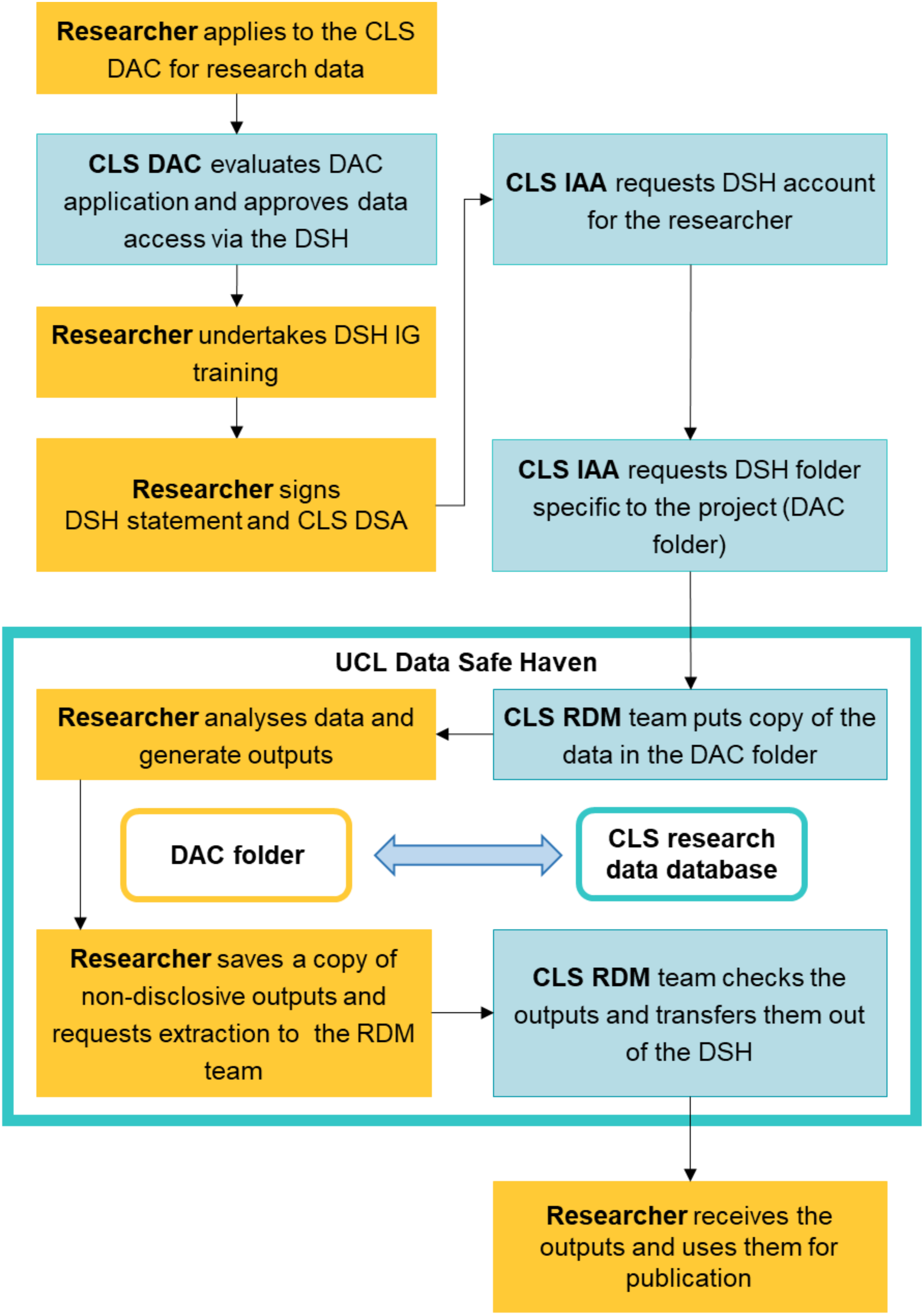
Data release via the UCL Data Safe Haven

If the requested data are highly disclosive data or sensitive, that is, are classified as Controlled (tier 2) or Special Controlled (tier 3), the data must be accessed via the UCL Data Safe Haven (DSH). Examples are verbatim responses or school identifiers.

Access to CLS research data via the UCL DSH follows the standard Five Safes Framework. The process on how to apply for and access data via the UCL DSH is described in the flowchart below.

The Research Data Management team will check the research outputs generated by within the UCL DSH by following the guidelines on Statistical Disclosure Control.

Figure 2. CLS DAC data release via the UCL Data Safe Haven


Researcher applies to the CLS DAC for research data
CLS DAC evaluates DAC application and approves data access via the DSH
Researcher undertakes DSH IG training
Researcher signs DSH statement and CLS DSA
CLS IAA requests DSH account for the researcher
CLS IAA requests DSH folder specific to the project (DAC folder)
UCL Data Safe Haven
CLS RDM team puts copy of the data in the DAC folder
Researcher analyses data and generate outputs
DAC folder
CLS research data database
Researcher saves a copy of non-disclosive outputs and requests extraction to the RDM team
CLS RDM team checks the outputs and transfers them out of the DSH
Researcher receives the outputs and uses them for publication

8.2 DAC applications for genetics 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).

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).

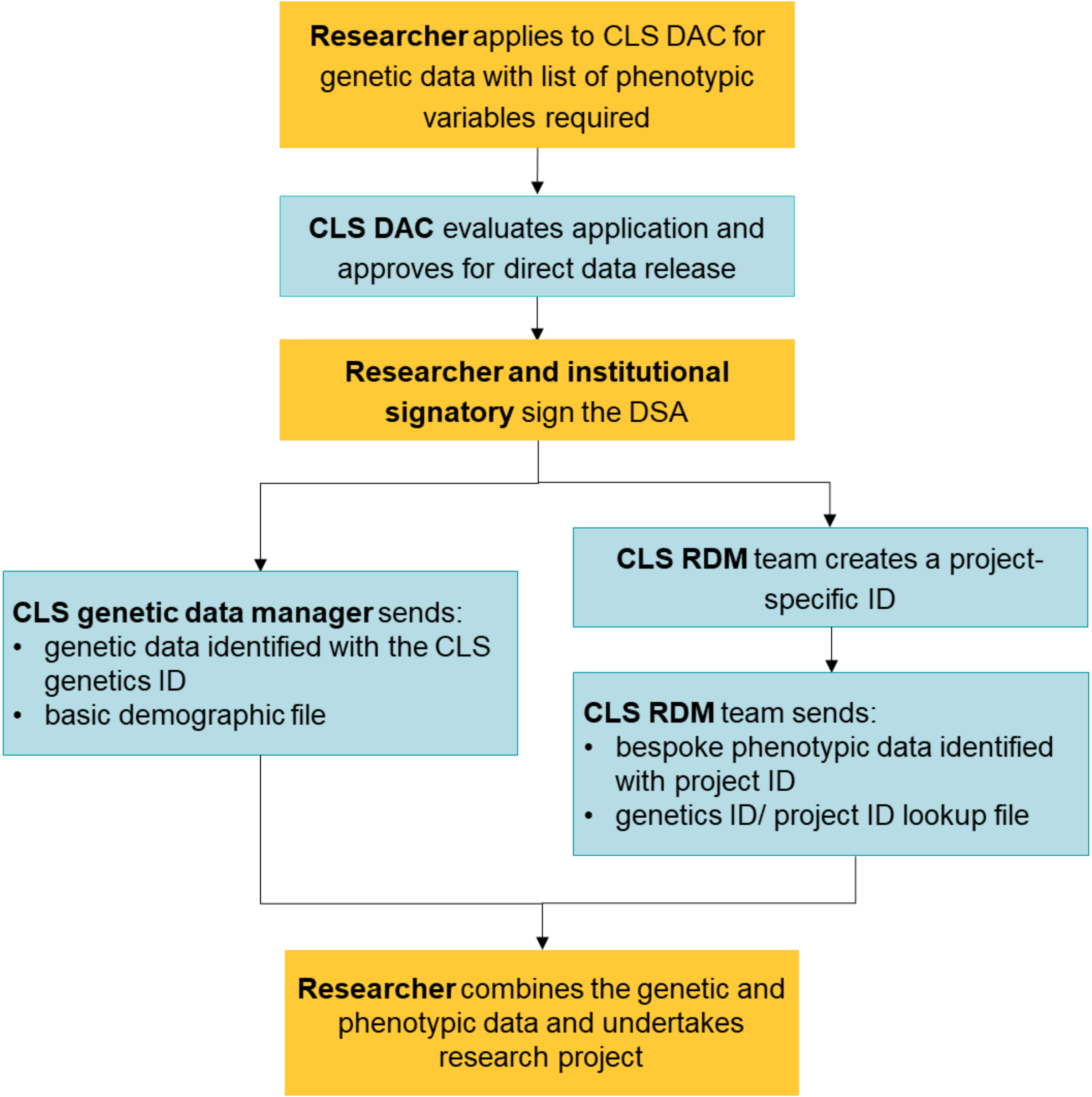
The final phenotypic dataset will be a bespoke dataset that only contains the exact list of variables requested by the applicant.

b) Pseudonymisation: newly created IDs

Once an application has been approved, a new project-specific 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 ID used to identify the CLS data available at the UKDS or EGA.

The data flow is presented below.

Figure 3. CLS DAC application process for genetics data



Researcher applies to CLS DAC for genetic data with list of phenotypic variables required

CLS DAC evaluates application and approves for direct data release

Researcher and institutional signatory sign the DSA

CLS genetic data manager sends:
- genetic data identified with the CLS genetics ID
- basic demographic file

CLS RDM team creates a project-specific ID

CLS RDM team sends:
- bespoke phenotypic data identified with project ID
- genetics ID/ project ID lookup file

Researcher combines the genetic and phenotypic data and undertakes research project

8.3 DAC applications that generate data for access within UKDS SecureLab

Applicants may need to access the data requested via the CLS DAC from their UKDS SecureLab account for secondary analysis in combination with controlled data.

1. On their CLS DAC application form, applicants should detail thoroughly:
   - what data they wish to analyse,
   - whether any data such as linked data, derived variables. or polygenic risk scores will be generated, and
   - how these are to be processed, for example, (a) the data access method for the CLS research data and (b) how the generated data will be imported into the UKDS SecureLab for secondary analysis in combination with controlled data. They should provide their UKDS project number if they have one at this stage.
2. Once the application is approved, the applicants will start work on their project and generate their data for import into the UKDS SecureLab.
3. If the generated data are on their institutional servers, the applicant will send this to CLS to be re-identified.
4. The applicant will make a Data Import Request to import the re-identified data into their SecureLab account.
5. Once CLS receives and approves the Data Import Request from the UKDS, CLS send the re-identified data to the UKDS, where it will be transferred into the applicant's UKDS SecureLab account.
6. In the SecureLab the applicant will be able to carry out secondary analysis / work on their data alongside their controlled data.

A thorough explanation is required so that the data release method and workflow can be determined. This process is visualised in the flowchart below (Fig 4).

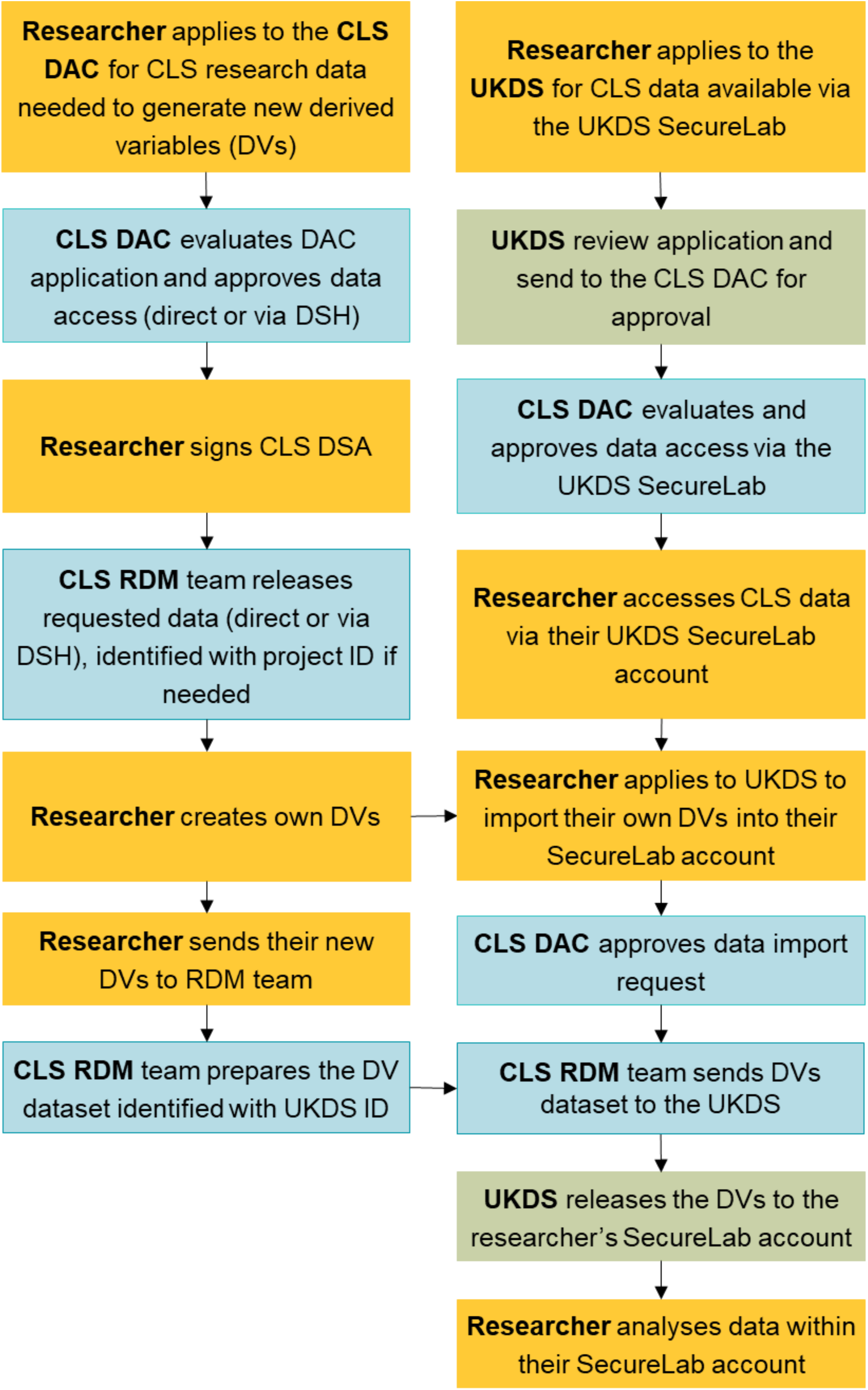
8.4 DAC applications to biological samples

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.

When receiving requests for biological samples, CLS will request a report from the Bristol laboratory to assess the up-to-date status of the samples and whether sample depletion will be an issue.

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

Figure 4. DAC applications that generate data for use within UKDS SecureLab



Researcher applies to the **CLS DAC** for CLS research data needed to generate new derived variables (DVs)

CLS DAC evaluates DAC application and approves data access (direct or via DSH)

Researcher signs CLS DSA

CLS RDM team releases requested data (direct or via DSH), identified with project ID if needed

Researcher creates own DVs

Researcher sends their new DVs to RDM team

CLS RDM team prepares the DV dataset identified with UKDS ID

Researcher applies to the **UKDS** for CLS data available via the UKDS SecureLab

UKDS review application and send to the CLS DAC for approval

CLS DAC evaluates and approves data access via the UKDS SecureLab

Researcher accesses CLS data via their UKDS SecureLab account

Researcher applies to UKDS to import their own DVs into their SecureLab account

CLS DAC approves data import request

CLS RDM team sends DVs dataset to the UKDS

UKDS releases the DVs to the researcher's SecureLab account

Researcher analyses data within their SecureLab account

9. Requests for novel CLS 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 DAC Record Linkage application form and guidelines can be accessed via the CLS webpage Proposing data linkages and enhancements.

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 should generally be performed by the applicants where identifiers are held in the UCL Data Safe Haven, e.g. postcodes, school identifiers, etc.

10. Requests for CLS data enhancements

CLS welcomes proposals for data enhancements to its 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 DAC Data Enhancement application form and guidelines can be accessed via the CLS Proposing data linkages and enhancements webpage.

Appendix 1. CLS guidance on plain-language abstracts

The CLS DAC requires 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 adequately 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.

Appendix 2. List of CLS DAC-approved applications

- Register of data access applications
- Register of genetic data access applications (GENDAC)
- Register of applications to access linked NHS England data