

## CLS Data Classification Policy

Document information	
Document name	CLS Data Classification Policy
Author(s)	Aida Sanchez
Version	5
Issue date	June 2024
Approved by	CLS Data Access Committee
Next review	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.

## Document history

Version	Date	Summary of change
0.1	Apr 2014	Initial issue
1	Feb 2015	Updated with review feedback, minor corrections, formatting, approved by IGSG, rebranding
2	Jun 2018	Table of data classification updated (AS)
3.1	Dec 2020	Reformatting; update of the data classification table; addition of data classification principles, data disclosure methods, and intruder profiles. Reference to genetic data
3.2	Nov 2021	Listed more examples of sensitive data; addition of the Data Linkage identifier, changed formatting, grammar, and spelling, and ensured consistency of terminology and throughout the document
4.1	Jun 2022	General edits
4.2	March 2023	Update on information about data types so they are consistent with UKDS terminology. Updated classification information.
5.0	June 2023	Reference UCL Data Classification Policy and re-classification of postcodes as Tier 2

## TABLE OF CONTENTS

<b>ABBREVIATIONS.....</b>	<b>3</b>
<b>1. INTRODUCTION.....</b>	<b>4</b>
<b>2. PRINCIPLES OF DATA CATEGORISATION .....</b>	<b>4</b>
2.1 Disclosivity .....	4
2.2 Sensitivity.....	4
2.3 Potential consequences of the misuse of data.....	5
<b>3. DATA CLASSIFICATION SCHEME .....</b>	<b>6</b>
3.1 CLS data Classification .....	6
3.2 UCL Data Classification .....	7
<b>4. METHODS OF DISCLOSURE CONTROL.....</b>	<b>9</b>
4.1 Identifiers used for pseudonymisation .....	9
4.2 Methods to reduce data disclosivity.....	10
<b>APPENDIX 1. INTRUDER PROFILES .....</b>	<b>12</b>

## Abbreviations

---

BCS70	1970 British Cohort Study
BNF	British National Formulary
CLS	Centre for Longitudinal Studies
CLS DAC	CLS Data Access Committee
DAC	Data Access Committee
DSH	Data Safe Haven
EEA	European Economic Area
EUL	End User Licence
GDPR	General Data Protection Regulation
ICD	International Classification of Diseases
IMD	Index of Multiple Deprivation
LSOA	Lower Layer Super Output Areas
MCS	Millennium Cohort Study
MSOA	Middle Layer Super Output Areas
NCDS	National Child Development Study or 1958 Birth Cohort Study
NHS	National Health Service
NINO	National Insurance number
NPD	National Pupil Database
OA	Output Areas
ONS	Office for National Statistics
RDM	Research Data Management
SAIL	Secure Anonymised Information Linkage
SIC	Standard Industrial Classification
SOC	Standard Occupational Classification
UCL	University College London
UKAN	UK Anonymisation Network
UKDS	UK Data Service
UK LLC	UK Longitudinal Linkage Collaboration
ULN	Unique Learner Number
UPN	Unique Pupil Number

# 1. Introduction

---

The CLS Data Classification Policy defines the principles through which CLS data is categorised, describes the information classification categories which will be applied to all CLS data and information, and provides information on CLS methods of disclosure control and pseudonymisation of research data.

## 2. Principles of data categorisation

---

CLS data have been collected or linked to external data sources to support research by a wide range of researchers in universities and other settings. The data collection, linkage, dissemination, and sharing of data is based on the consent given by the participants and is conditional on the assurance that CLS will protect their confidentiality and other rights in relation to the data. Attempts to re-identify individuals in the cohort is always forbidden.

Breaking this assurance would not only constitute an ethical violation of consent, but also threatens the trust that cohort members place in the research team who collect their data and may affect their willingness to participate in further data collection.

CLS have evaluated and categorised the data in terms of two underlying principles:

- Disclosivity
- Sensitivity

The appropriate degree of security and access management will be applied depending on what category has been assigned.

### 2.1 Disclosivity

Data are considered disclosive if there are concerns over the re-identification of individuals, households, or organisations with which they are associated, should data users attempt to do so.

CLS data is categorised reflecting an assessment of the likelihood and potential impact of re-identification.

### 2.2 Sensitivity

Data are considered sensitive if there are concerns over the consequences of re-identification, i.e., the potential damage in the case of self-identification or identification by other family members, and also because participants may expect such data to be subject to greater protection. For instance, a participant may be concerned about their drinking behaviour being discoverable but not their ethnic origin.

Sensitive data require more protection because of their content and participants might reasonably expect that additional steps are taken to ensure disclosure risk is reduced. Data that risk the disclosure of sensitive information require a high degree of security and management.

Some examples of information particularly sensitive are:

- detailed mental or physical health
- illegal behaviour
- childhood abuse
- drug/alcohol use
- racial or ethnic origin
- religion
- genetic data
- sexual orientation

Such data are considered “special category data” according to [the Data Protection Act 2018 categorisation](#).

## 2.3 Potential consequences of the misuse of data

The consequences of misuse depend on a number of factors, including whether it is accidental or malicious, the scale of data disclosure (i.e., how many participants it affects), whether it creates a possibility of harm or distress for participants, or whether it affects the public reputation of the study. Consequences may include negative publicity and legal action.

In terms of categorisation, the impact of consequences is closely aligned with other principles. For instance, sensitive data should be classified at a higher impact level because the consequences of misuse are more severe.

Risk of disclosure and misuse has two components:

- a) the risk that a user of the data attempts to breach the confidentiality of participants or misuses the data in any other way (motivated intruder).
- b) the risk that they are able to do so given the data they have available.

[Appendix 1](#) outlines several intruder profiles.

## 3. Data classification scheme

---

### 3.1 CLS data Classification

CLS research data are classified under the following schema based on the level of sensitivity and potential disclosure. This classification determines the level of access:

- **Safeguarded data**, which present a low level of disclosure and/or sensitivity. These data are considered effectively anonymised, meaning that the identifiability risk is remote due to the anonymisation treatment applied to the data and the licence under which they are made available. These data are accessible via the UK Data Service (UKDS). There are two tiers of safeguarded data:
  - **Safeguarded data – tier 1a**, which present a low level of disclosure and/or sensitivity. These data are available from the UKDS following registration and access is governed by the UKDS End User Licence.
  - **Special safeguarded data – tier 1b**, which present a medium level of disclosure and/or sensitivity. These data are available from the UKDS following an application process. Access is governed by the UKDS Special Licence User Agreement.
- **Controlled data**, which present a high level of disclosure and/or sensitivity. These data are considered personal and therefore subject to GDPR data protection law, and can only be accessed via highly secure institutional servers. After registering, experienced researchers can apply to access, which must be approved by CLS. There are two tiers of safeguarded data:
  - **Controlled data – tier 2**: high level of disclosure and/or sensitivity. These data can only be accessed via the UK Trusted Research Environments chosen by CLS, such as the UKDS SecureLab, UK LLC or SAIL Databank.
  - **Special controlled data – tier 3**: very high level of disclosure and/or sensitivity. These data can only be accessed via the CLS secure project folders within the UCL Data Safe Haven.
- **Confidential – tier 4**: identifiable cohort member details that can only be accessed from within the UCL Data Safe Haven by the CLS Cohort Maintenance Team
- **Private**

The details of these data categories are summarised in the table below and provides examples of the different kinds of data and information which would be covered within the agreed categories.

Placing data in higher categories provides greater protection for participants but increases the real or perceived barriers to the use of data by researchers. Real barriers include limitations on access outside the UK for higher impact level data. There is therefore a

balance to be drawn between maximising the use of the data and minimising risks to the rights of participants.

### 3.2 UCL Data Classification

The three data categories used by UCL are "Public", "Confidential" and "Highly confidential" ([PDF Information-Management-Policy V1.2.pdf](#)).

Given the varied nature of CLS data, CLS have developed a data classification policy that is more specific than UCL's.

The CLS – UCL data classification mapping would be as follows:

CLS data classification	UCL data classification
Safeguarded data (Tier 1a)	Confidential
Special Safeguarded data (Tier 1a)	Confidential
Controlled data (Tier 2)	Highly confidential
Special Controlled data (Tier 3)	Highly confidential
Confidential (Tier 4)	Highly confidential

CLS don't classify shareable research data as **Public**, as this refers to information that can be freely used, reused and redistributed by anyone with no existing local, national or international legal restrictions on access or usage.



This table illustrates the data classification based on their level of disclosure risk and how the data can be accessed.

Classification	Description	Disclosure risk	Sensitivity	Access	Approved users
<b>Safeguarded – tier 1a</b>	Pseudonymised and de-identified survey data with low level of disclosure. For example, <a href="#">the Age 46 follow-up sweep of BCS70</a> .	Low	Low	UK Data Service (UKDS) End User Licence (EUL)	Registered UKDS users who are researchers affiliated with an institution or organisation.
<b>Special safeguarded – tier 1b</b>	Data with a medium level of potential disclosure risk (e.g., intermediate geographical indicators such as counties) or sensitivity (e.g., mortality data, detailed physical or mental health information, genetics, polygenic scores). For example, <a href="#">NCDS counties data</a> .	Medium	Low/medium	UKDS Special Licence or CLS DAC	EEA research projects approved by the UKDS / CLS or the CLS DAC (genetics)
<b>Controlled – tier 2</b>	Data with a high level of potential disclosure risk (e.g., exact dates, detailed ethnicity, detailed geographical indicators such as postcodes), or highly sensitivity data such as linked health data (e.g. linked NHS HES data) or linked education data (e.g. linkedNational Pupil Database data).	High	Medium/high	Trusted Research Environments (UKDS Secure Lab, UK LLC, SAIL Databank, UCL DSH)	UK-based research projects approved by the UKDS/CLS
<b>Special controlled – tier 3</b>	Data with a very high level of potential disclosure. For example: open text responses, school IDs, etc.	Very high	Medium/high	UCL Data Safe Haven (DSH)	CLS data managers and UK researchers approved by CLS DAC
<b>Confidential – tier 4</b>	Individually identifying information needed for operational purposes: contact with cohort members, data collection, or external data linkage. For example: name, address, email address, NHS number, National Insurance number (NINO), etc.	Full identification	Low	UCL Data Safe Haven (DSH) and highly secure external servers as needed	CLS cohort maintenance team and external parties as required
<b>Private</b>	CLS internal documentation where there is no benefit or requirement in making it publicly available.	N/a	N/a	CLS shared drive on UCL server	CLS staff

## 4. Methods of disclosure control

---

CLS follow the [Office for National Statistics \(ONS\)](#) and [UKDS](#) guidelines of disclosure control.

We assess disclosure risk based on potentially disclosive variables, taking into account that:

- there is no exact formula to help us judge 'objective' risk
- we cannot provide a one-size-fits-all rule book
- we follow the recommended best practice for surveys

### 4.1 Identifiers used for pseudonymisation

CLS data is held and distributed in a pseudonymised manner, which consists of assigning different identifiers to the data. This ensures that if data is released, linkage to other data is only possible to other data on the same identifier. For instance, participants' contact data used for data collection fieldwork or for matching to external administrative data should not be linkable to research data available from the UKDS.

#### **CLS internal identifier**

These are identifiers used on databases and files by both the CLS Research Data Management (RDM) team and the CLS Cohort Maintenance team. This is to allow appropriate sharing of data across these two functional areas.

These identifiers, data linked to them, or lookups between these and other identifiers should *never* be released by CLS.

#### **Data collection identifier**

These are identifiers used during fieldwork or other data collections where data is passed to a third party for contact with cohort members.

This also includes research groups who want to recode string data for other purposes e.g., re-coding of occupation or diseases for coding frames not currently available. The research collaborator would need to return the coded data to CLS for it to be relinked to a research identifier before it could be merged with existing data.

#### **Data linkage identifier**

These are identifiers used during linkage with external administrative data, such as health or education records, held by external data organisations. In this model of data collection,

the personal identifiable data of CLS cohort members is passed to a third trusted party for matching with the administrative data.

### Research identifiers

These are identifiers used by researchers that are deposited at repositories such as the UKDS. Examples of these are:

- 1958 National Child Development Study (NCDS): NCDSID
- 1970 British Cohort Study (BCS70): BCSID
- Next Steps: NSID
- Millennium Cohort Study (MCS): MCSID

In some studies, for instance NCDS, where there is linkage between data collected during a survey and derived data from genetic analysis, a bespoke research identifier is issued.

## 4.2 Methods to reduce data disclosivity

Assessment of the disclosure risk of potentially disclosive variables is performed following recommended best practice for surveys and the ONS and UKDS guidelines of disclosure control. Once potential disclosivity has been ascertained, CLS data are checked and suitably de-identified prior to disseminating for research purposes. The depth of the de-identification applied will depend on the chosen mechanism for data access, i.e., the data classification assigned.

### Assessment of data disclosivity

Examples of detailed information that could potentially lead to the identification of an individual or households include:

- Exact dates: birth date, data collection date
- Detailed employment for example, full Standard Occupational Classification (SOC) or full Standard Industrial Classification (SIC)
- Religious affiliation, ethnicity, language(s) spoken at home, country of origin
- Outliers (e.g., height, number of bedrooms, number of children)
- Very detailed health information, for example, full International Classification of Diseases (ICD) or full British National Formulary (BNF) codes
- Unusual health condition (e.g., rare disease, total blindness)
- Small geographic area such as postcodes, OA, LSOA, etc.
- Local neighbourhood specific characteristics, for example, detailed Index of Multiple Deprivation (IMD)

- Open-ended answers in qualitative research
- Linked information such as school identifiers, health care providers, etc.

CLS have developed a number of programming scripts to find disclosive variables and to tabulate 'risky' variables with small cell counts.

For instance, the threshold chosen is to have *no cell counts less than 10* for safeguarded data (tier 1, UKDS EUL).

Where possible, variables are also checked against the population from which the sample was taken (such as height distributions).

### Data disclosure methods applied to enable data sharing

Once potentially disclosive variables have been identified, CLS applies a number of modifications to the data in order to create a dataset that can be shared under the chosen data sharing method.

Some of these de-identification methods applied by CLS are:

- **Banding** – reduced granularity of information whilst retaining some details about the distribution.
- **Top/bottom coding** – where a continuous distribution has a long right or left tail (or both), those outliers are assigned a maximum value (top-code) or minimum value (bottom-code) so that they are grouped together at the top and bottom of the scales.
- **Reducing precision** – this could be in the form of truncation, such as only providing the first half of a postcode or the first three digits of a SOC code.
- **Pseudonymisation** – there may be indirect identifiers which also act as “clusters,” such as an anonymised school identifier. These can be given a code which retains the clustering but removes the identifying information.
- **Variable removal** – where a variable is considered too disclosive or sensitive for the intended licence.

## Appendix 1. Intruder profiles

---

Motivated intruders are data users who are willing to sign up to the data usage licences used by CLS, such as the CLS Data Sharing Agreements or the UKDS End User Licence, but not comply with it. A motivated intruder may attempt to use safeguarded research data to re-identify individuals, even if this is strictly forbidden by the relevant terms and conditions of data use. Intruders may also try to cross-reference these data with publicly available information, such as the electoral register or social media (for example, educational qualifications may appear on publicly viewable platforms or professional networks). The intruder profiles could be:

- **Activist** – A group or individual who wishes to discredit data sharing, the NCDS, the Centre for Longitudinal Studies, the Institute of Education, or University College London.
- **Marketer** – Companies who wish to use the large dataset for some marketing purpose. If one could re-identify a proportion of the cohort members who were found to have a profile of particular biometric profile that would be useful for marketing, particularly with income information and any geographical information.
- **Nosy neighbour** – Someone who knows a cohort member and that they are a cohort member. This could be someone who is aggrieved with the cohort member and wishes to bring them to disrepute.