Consumer Data Research Centre

An ESRC Data Investment

VWV

The General Data Protection Regulation & Social Science Research

Disclaimer

The information contained in this guidance is intended for CDRC Stakeholders - researchers making use of consumer-related datasets as part of their participation in the Consumer Data Research Centre and CDRC Data Partners both current and potential. This guidance does not contain any legal advice and no liability is assumed for any loss, damage or inconvenience arising as a consequence of any use or the inability to use any information contained in it. The legal position in a given situation is always dependent on the specific facts and circumstances.

Introduction

The General Data Protection Regulation (**GDPR**) comes into force on 25 May 2018, and all processing of personal data taking place on or after that date must be compliant with its provisions.

The GDPR will supersede the Data Protection Act 1998 (**DPA**) as the legal regime that governs the processing of personal data in the UK. The GPDR brings a number of changes to data protection law including, notably, a new wider definition of "personal data".

This guidance contains (1) a glossary of key terms under the GDPR; (2) a summary of the key principles of the GDPR and (3) Frequently Asked Questions for researchers.

Glossary of Key GDPR Terms

These terms are highlighted in bold in the summary and FAQs.

Key Term	**Definition**
Anonymisation	The process of rendering data into a form which does not identify individuals and where identification is not likely to take place. Where anonymisation is carried out effectively, neither the production nor the publication of the anonymised data will have any effect on any particular individual.
Data Subject	A living individual who is the subject of personal data.
Data Controller	A person who (either alone or jointly) determines the purposes for which and the manner in which any personal data are or are not to be processed. They are usually organisations but can also be individuals, for example self-employed consultants.
Data Processor	Any person (other than an employee of the data controller) who processes the data on behalf of the data controller.
Lawful Basis (for Processing)	Organisations must have a lawful basis for processing personal data. The lawful bases include: consent; legitimate interests;

	necessary for the performance of a contract, necessary to comply with a legal obligation, necessary to protect the individual's vital interests, or a task carried out in the public interest. Note that there are additional conditions which must be met to establish a lawful basis for processing "special category personal data" (*see definition below*).
Legal Basis	Beware of this term which does not have a clear meaning in the GDPR. The preferred interpretation is that it refers to a law of either the European Union or the UK that permits a particular kind of processing activity in principle - usually subject to conditions. Even where such a law exists and is relevant, the data controller must still have a **lawful basis** for that processing**.**
Marketing	This covers any advertising or marketing material not just commercial marketing. Promotional material, including promoting the aims of not for profit organisations is covered.
Multi-level authentication	Multi-level authentication is where a user has to input multiple pieces of evidence before being granted access. For example, a username and a password. More secure systems also require a generated piece of evidence, such as the number provided by a digital secure key which is commonly used for online banking.
ICO	The Information Commissioner which is the supervisory authority for data protection in the United Kingdom.
Personal Data/ Personal Information	Any information relating to a living person who can be directly or indirectly identified in particular by reference to an identifier. This includes names, email addresses, identification number, mac addresses, location data or online identifiers.
Processing	This means obtaining, recording or holding the personal information or carrying out any operation or set of operations on the information. This includes organising, adapting or altering the information, as well as disclosing and deleting information.
Profiling	This is the use of personal data to evaluate personal aspects relating to an individual,

	for example to build up a picture of their consumer habits.
Pseudonymised Data	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 and that additional information must be kept separately and subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable person.
Public Authority	a. A public authority as defined by the Freedom of Information Act 2000; b. a Scottish public authority as defined by the Freedom of Information(Scotland) Act 2002; and c. an authority or a body specified by the Secretary of State in regulations.
Sensitive Personal Data/ Special Category Personal Data	This includes information relating to a Data Subject's: racial or ethnic origin; political opinions; religious beliefs; trade union membership; health data; sexual life; genetic data; biometric data and criminal offences.
Safeguarding Conditions	Arrangements in relation to processing that is necessary for a research or statistical purpose which ensure that the processing is: a) not likely to cause substantial damage or substantial distress to the data subject; **and** b) not carried out for the purpose of measures or decisions with respect to the data subject (unless it is for the purpose of approved medical research).
Transparency Notice	A document (hardcopy or electronic), which may be made up of several smaller documents that will explain to individuals what personal information about them the Data Controller collects and how it will use it.

Summary of the Overarching GDPR Principles

There are 7 overarching principles of data processing under the GDPR and these are each summarised below. Definitions of defined terms can be found in the glossary above.

Principle 1. Lawfulness, fairness and transparency

Lawfulness

All **processing** of **personal data** must have a **lawful basis**. There are two sets of grounds which can be used to justify processing. One set of grounds covers **special category personal data** and another covers all other **personal data**.

For all **personal data** that is not **special category personal data** there are six lawful bases available. The most likely to be applicable for research activities are:

- the **processing** is necessary for the performance of a task carried out in the public interest;
- the **data subject** has provided consent; and
- the **processing** is necessary for the purposes of the legitimate interests of the **data controller** or a third party and the interests of the data subject are not overridden.

For **special category personal data**, **processing** for research purposes is permitted so far as (i) one of the six non-special category personal data lawful bases applies; (ii) the essence of data protection rights is respected and (iii) suitable **safeguards** and protections are put in place.

Fairness and Transparency of Processing

The GDPR requires **data controllers** to identify:

- the different categories of **personal data** which they **process;**
- the purposes for which that **processing** is carried out; and
- the **lawful basis** for each **processing** purpose.

The GDPR obliges **data controllers** to provide **data subjects** with certain information about how their **personal data** will be used. This includes:

- the **data controller's** identity and contact details;
- the **personal information** held about the **data subject**;
- how personal information is collected from the **data subject**;
- the purpose of the **processing**;
- the **lawful basis**/bases being used to justify the **processing**;
- who **personal data** will be shared with;
- the duration for which the **personal data** will be retained;
- whether **personal data** will be transferred outside the EEA; and

- the **data subject's** right to complain to the ICO.

Typically, the above information is communicated to the **data subject** in the form of a **transparency notice,** which may be made up of several documents (or electronic notices), each of which is given at the relevant time.

Principle 2. Purpose Limitation

The GDPR requires that **personal data** only be collected and **processed** for specific, explicit and legitimate purposes and not further processed in any way that is incompatible with those original purposes.

Please see questions relating to the treatment of **personal data processed** for research purposes in the FAQ section of this guidance.

Principle 3. Data Minimisation

The principle of data minimisation requires that **personal data** only be **processed** if it is accurate and relevant to the purpose for which it is **processed. Personal data** should also only be **processed** to the extent that it is necessary in relation to the **processing** purpose.

Principle 4. Accuracy

Every reasonable step should be taken to ensure that **personal data is** accurate and where necessary kept up to date. Every reasonable step should be taken to ensure that inaccurate data is deleted or rectified.

Principle 5. Storage limitation (i.e. retention of personal data)

As part of the Principle of Data Minimisation (see above), the GDPR states that **personal data** should be kept in a form which permits identification of individuals for no longer than is necessary for the purposes for which **personal data** is **processed**.

Data Controllers should decide upon a retention policy for different categories of **personal data** and be transparent about the criteria used to determine any retention period that cannot be specified when the data is collected.

Principle 6. Integrity and Confidentiality

Personal data must be subject to appropriate security and protected against any unauthorised or unlawful **processing**.

Appropriate technical and organisational measures should be adopted to protect **personal data** against accidental loss, destruction or damage.

Principle 7. Accountability

Data controllers are responsible for their own compliance with Data Protection law and the GDPR states that they must be able to demonstrate their compliance. To this end, the GDPR places **data controllers** under an obligation to keep written records of their **processing** activities.

Frequently Asked Questions about GDPR and use of Personal Data in Research

Q. What does the GDPR mean for Data Partners engagement with the CDRC?

A. Research organisations and their stakeholders, including Data Partners, have a greater responsibility under the GDPR than under the DPA, so it is more important for them to understand the principles of data protection.

Provisions relating to data sharing: you will need to identify whether the relationship between CDRC and stakeholders is one of Data Controller-Data Controller or Data Controller-Data Processor.

Where there are Data Controller to Data Processor arrangements between CDRC and any Stakeholder, the GDPR requires that there are written arrangements in place documenting the arrangements.

CDRC and its stakeholders will need to review existing agreements in place between them. These agreements are likely to need amendment to make them GDPR compliant, particularly in relation to the issues around identifying categories of Personal Data and the purpose of the processing (i.e. granularity).

Q. Can I still make use of the CDRC's existing datasets following the implementation of the GDPR?

A. In order to make use of existing personal datasets, you will need to ensure that by 25th May 2018 your processing is compliant with the GDPR.

You will need to consider the seven overarching principles set out above and ensure that you have a lawful basis for processing the personal data that you already hold.

This may involve a certain amount of housekeeping on your existing datasets as there are currently no "grandfathering" provisions. In any event, you should keep a written record of your reviews so that you are able substantiate your conclusions about GDPR compliance.

The CDRC will keep records of the lawful bases for holding personal data for research. A Transparency Notice will also be maintained and be available on the CDRC website. Personal data are held within the CDRC secure laboratories (ISO27001 accredited or Police Assured Secure Facility). All outputs from the secure laboratories undergo stringent checks to ensure no personal or disclosive data are released. The CDRC will undertake an audit of all project approvals granted prior to GDPR and Users affected will be made aware of any issues.

Q. How can I meet the requirement to achieve the principle of data minimisation when carrying out my research?

A. When carrying out research using Personal Data the aim is ideally to anonymise the data so that the data is no longer subject to the constraints of the GDPR and you have greater freedom to use it.

Anonymous data is defined by the GPDR as "*information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.*" The GDPR provides that it does not concern the Processing of such anonymous information, including for statistical or research purposes.

If anonymisation cannot be achieved, steps should be taken so that the smallest possible risk is run that the Personal Data being used could be used to identify or linked to an individual. The GDPR encourages researchers to mitigate this risk using pseudonymisation.

Pseudonymisation is defined as "*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 and that additional information must be kept separately and subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable person*."

Q. When does a Transparency Notice need to be provided?

A. Under the GDPR there is an obligation to **provide individuals with certain information** about **what you are going to do** with their data. This includes:

a. your identity and contact details;

b. what personal information you are collecting about them;

c. how you are collecting it;

d. the grounds that you are relying on for the processing of the personal data (such as necessary for legitimate purpose);

e. the purpose for the processing;

f. who you are sharing their data with;

g. how long you are keeping their data for;

h. if you are transferring the information outside the EEA; and

i. that they can complain to the ICO if they are unhappy with your handling of their personal data.

Special Note: Direct collection v. indirect collection

If you are collecting **personal data** directly from a **data subject**, you must give them the transparency information detailed at above when you collect the **personal data** from them.

You will regularly be collecting **personal data** from sources other than the individual that the data is about (i.e. indirectly). This may be:

a. by obtaining the date from a commercial third party, such as a loyalty card provider;

b. from publicly available records such as census data; or

c. by covert measures such as the use of sensors or CCTV.

You must be sure that when the collection of the data took place, the transparency requirements were complied with. This may involve asking third parties collecting the data to confirm that they fulfilled the transparency requirements of the GDPR. However, it will remain your responsibility to satisfy yourself that the third parties' responses do in fact indicate compliance. Where these records are public records there is specific authority in the GDPR for the use of those datasets. If you are responsible for collecting data by covert means, you must meet the transparency requirements yourself.

When you are collecting data from third parties, this obligation will not apply to your research purposes where the provision of such information to the individuals will be impossible or involve a

disproportionate effect. Where provision of this transparency notice to each **data subject** would be impossible or involve a disproportionate effort, there are special provisions: please see the flowchart document titled **schedule 1** to establish whether these special provisions apply to you. However, you will still need to make this information publicly available (e.g. by publishing it on your website).

Q. What techniques can I use to help achieve the principle of data minimisation?

A. The precise context and circumstances in which a research project is carried out and its objectives will have a direct impact on the techniques which should be employed to ensure respect for the principle of data minimisation.

Techniques that can be used to help achieve the principle of data minimisation include:

- removal of one or more variables that directly or indirectly identify individuals from the data;
- aggregation of data so that only totals are shown, and removing records where an individual can be identified despite the application of other protection techniques.
- global recoding - this method makes variable values less specific, and therefore the data less informative. An example of this is instead of using a postcode of an individual, you might group them by area of the country so that a London postcode becomes South-East, or use age ranges rather than specific ages of individuals;
- hashing the data - this means to use an algorithm to map data to a fixed length, which cannot then be reversed; and
- salting the data - this involves adding an extra secret value to the end of an input and extending the length of the original data.

Before undertaking a research project you should carry out an assessment to decide on the techniques to be used. Assessing the appropriate techniques to use is in fact carrying out a data privacy impact assessment on the project being planned. The results of this assessment should be documented and passed to your data protection officer for accountability purposes.

At the CDRC we wherever practical for research purposes acquire and make available pseudonymised, anonymised, aggregated or hashed data.

Q. When carrying out research as part of CDRC, do all the obligations set out in the GDPR apply?

A. Processing Personal Data for scientific or historical research purpose or statistical purposes is exempt from certain provisions of the GDPR, provided that there is a Lawful Basis for the research (the **Research Exemptions**). The Research Exemptions may not cover all academic research.

The definition of research under the GDPR is very wide, and indicates that social science research is part of scientific research. This means that the research that the CDRC and its partners are doing is likely to be 'research' for the purposes of being able to take advantage of the Research Exemptions.

However, all Processing in relation to which you wish to rely on the Research Exemptions must comply with (a) the GDPR safeguard requirements and (b) the Data Protection Bill safeguarding requirements.

(a) The GDPR safeguarding requirements

In order to rely on the Research Exemptions the GDPR states that the research must be carried out subject to appropriate safeguards for the rights and freedoms of the Data Subjects .These

safeguards must ensure that technical and organisation measures are in place in particular to ensure respect for the principle of data minimisation.

(b) The Data Protection Bill safeguarding requirements

The Data Protection Bill as currently drafted stipulates that in order to rely on the Research Exemptions you must also ensure that the Processing complies with the following two safeguarding conditions which mirror those currently in place under the current Data Protection Act regime:

- the Processing must be subject to appropriate safeguards for the rights and freedoms of the data subject if it likely to cause substantial damage and distress to a data subject; and
- the Processing must not be carried out for the purposes of measures or decisions with respect to a particular Data Subject.

The CDRC ensure safeguarding requirements are adhered to through the Research Approvals Process[1] whereby CDRC Users are required to provide detail on their planned use of CDRC data, proposed outputs and ethical approval for the work. Our independent review process ensures that the Data Partner(s) in question are aware of how their data will be used and agree to the research and that our independent academic review process ensures that the research has scientific merit and/or is for the benefit of society. Access to CDRC personal data is controlled and analysis undertaken within the secure environment with outputs checked to ensure results are not disclosive.

Q. How can data be linked and loosely coupled?

A. All such processing should be done in accordance with the **data minimisation principle.**

The main issue you will need to consider is whether by linking or loosely coupling the personal data you are profiling individuals. We set out below the considerations relating to profiling.

If the linking or loose coupling do not result in profiling, then the processing activity is not affected by the special requirements.

Q. What is profiling and why does it matter?

A. Profiling means: *any form of automated Processing of Personal Data consisting of the use of Personal Data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.*

Profiling forms the basis of a wide range of scientific research. Compiling datasets together is an important feature of your research. The process of compiling the datasets can result in an individual being profiled.

Under the GDPR profiling means any form of automated Processing of personal data to evaluate certain personal aspects. The GDPR says that individuals have the right not to be subject to decisions made automatically that provide legal effects or significantly affect these individual.

The right not to be subject to automated decisions does not apply where:

- the decision is necessary to fulfil a contract with the individual;

[1] https://www.cdrc.ac.uk/wp-content/uploads/2018/02/CDRC-RAG-ToR-V8.pdf

- the relevant processing is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

- the decision is based on the explicit consent of the individual.

However, if the profiling of personal data in the research does not result in making any decisions that that significantly affects the individual, then the profiling provisions will not apply. It may, however, be an issue to be addressed by those who use the results of your research and a holistic view should be taken at the outset to identify data protection law issues. You may wish to consider addressing this point in your application to your Ethics Committee for approval for the research project.

Q. What is aggregation and can data that has been subject to aggregation still be Personal Data?

A. Aggregation is where data is added up and displayed as totals, so no data identifying individuals in shown. This can be a useful tool for data minimisation.

If there are small numbers in the aggregate totals then it may be possible to identify individuals from the aggregate, depending on the other information given. Avoid using aggregation in this instance or remove the class of personal data with small numbers from any published data.

CDRC output checking procedures have been put in place to prevent disclosive data being released from the secure laboratories. Where there are small numbers contributing to a cell, should be suppressed, combined or removed. Details are included in Appendix 3 of the CDRC Controlled Data Project Proposal Form.

If using aggregate data to profile individuals, please see profiling section above.

Q. Can modelled data be Personal Data?

A. Modelled data will be Personal Data if it is possible to identify a living individual from it or from it together with other information that is, or is likely to be in the possession of the Data Controller.

If the research used in the modelled data is based on non-Personal Data, and no living individual can be identified from the model, then the modelled data are not personal Data.

Q. Are there further requirements for processing Personal Data relating to criminal convictions or offences?

A. Personal Data relating to criminal convictions or offences must only be processed either:

- under the control of official authority; or

- where authorised by Union or Member State law which provides for appropriate safeguards for the rights and freedoms of Data Subjects.

This is governed by the Police and Criminal Justice Directive (PJC Directive) rather than by the GDPR.

Q. What does Brexit mean for the GDPR?

A. The UK government has confirmed that GDPR will come into force on 25 May 2018. In addition, it has confirmed that it intends to incorporate the provisions of the GDPR into national law so that it continues in force past Brexit, and the draft Data Protection Bill has entered Parliament. However,

the government has also said that the Court of Justice of the European Union (**CJEU**) will cease to have authority over national law, so at this stage it is unclear how subsequent clarifications and developments of the GDPR CJEU case law will take effect on national law.

In some parts of the GDPR the text is unclear. The ICO and the Article 29 Working Party are drafting various guidance, and we may change our view of the effect of particular parts of the GDPR.

In addition, the UK courts continue to hear data privacy cases and the law evolves, particularly regarding the new tort of intrusion of privacy.

April 2018

Veale Wasbrough Vizards LLP
www.vwv.co.uk

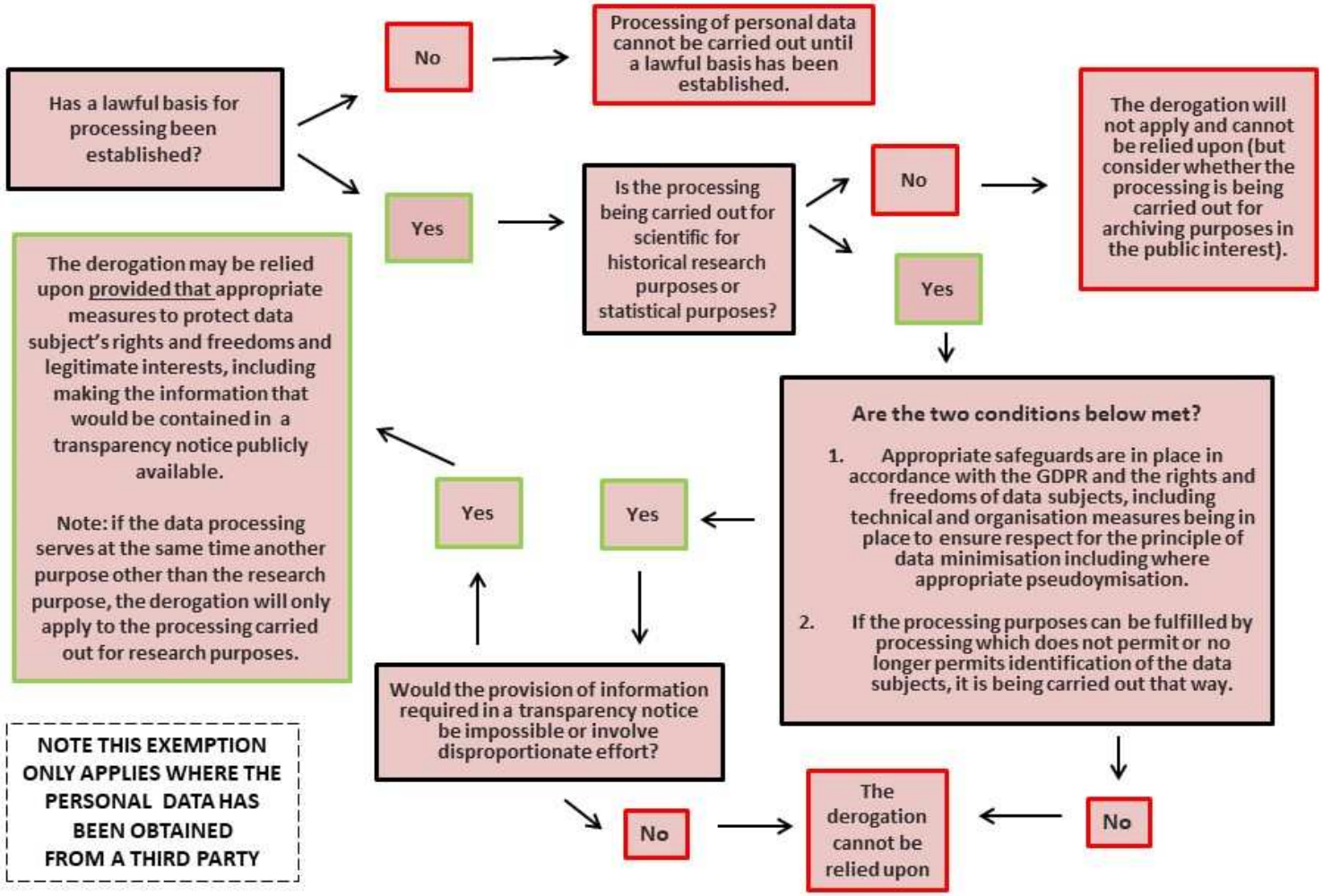
Consumer Data Research Centre

An ESRC Data Investment

VWV

Schedule 1. The General Data Protection Regulation

Do you benefit from the derogation from the obligation to provide a transparency notice where data is processed for research purposes?



Appendix 1. Case Studies Demonstrating Techniques to Achieve Data Minimisation

Case Study 1: Use of Key-Coded Data

a. *A research study is being carried out by CDRC using data provided by a retailer. The data consists of customer IDs, postcode and transaction data. The retailer retains the 'key' to the customer ID but would provide it if asked.*

1.1 Such use of key-coded data is a common pseudonymisation technique. All use of the pseudonymised data should be carried out by people who are trained on GDPR requirements, bound by obligations of confidentiality and subject to restrictions regarding re-use and re-identification. Pseudonymised data are treated under the same strict security protocols as all Personal Data and at the CDRC analysis is only undertaken on secure servers.

b. *A clinical study is being carried out for a pharmaceutical company by clinical investigators. In this study only key-coded data are passed by clinical investigators to the company which is sponsoring the research.*

1.2 *The decryption keys are stored at the study site by the clinical investigators who are trained in GDPR requirements of their institutions and are bound by professional duties of good clinical practice and confidentiality. The sponsor company is not authorised to call for the decryption keys.*

1.3 Such use of key-coded data is a common pseudonymisation technique. All use of the pseudonymised data should be carried out by persons trained on the GDPR requirements and boundby obligations of confidentiality and subject to restrictions regarding re-use and re-identification.

1.4 When using an encryption key it is important to ensure that:

- the same key is not used for different datasets as this would increase the risk of different datasets being linked; and
- the key is stored securely at all times.

Case Study 2: Maintaining the link between data values attributable to the same individual

A research institution has been provided with data that were originally sourced from a mobile app that uses Global Positioning System geo-referencing to infer measures of the speed at which users of the App run. The research institution would like to analyse the data to derive information about the average running speed of each user when using the app.

The research institution intends to process the following data about each user featured in the dataset to carry out this research:

- *User ID;*
- *number of runs made by the user in a particular month; and*
- *the distances covered and the times taken to complete each run.*

To minimise the amount of personal data in the dataset, the research institution would like to replace the User ID numbers with artificial values. Although the User ID numbers are to be replaced, the research institution does not want to lose the link between different runs made by the same individual. This could be achieved using one of the following techniques:

a. encryption e.g. using the Advanced Encryption Standard AES encryption algorithm - *this will ensure that identical original values are always mapped to identical modified values and that non-identical original values are always mapped to non-identical modified values*;

b. tokenisation e.g. using a mapping table.

Case Study 3: Hashing

a. *A research institution is carrying out research into the amount of coffee that is drunk by people of different ages. It is provided with data (an extract of which is shown in the table below). The data are supplied by a café chain and detail the number of coffee purchases made by loyalty scheme participants along with their dates of birth.*

Loyalty Card No	Date of birth	Number of coffee purchases per month
15618	08/11/1992	18
14555	09/05/1971	3

1.5 The nature of the data means that hashing and data banding could be used to considerably reduce the likelihood of re-identification as demonstrated below.

Hashed loyalty card holder ID	Age band	Number of Coffees drunk per month
ea840312edaf4c00a97c5d89cbf11f8fa4c411d1b1be274415d5b64b55adf0c6	21-30	18
c3ec998eb12f3655104d47770a0d71b4a1f3ef287d6ccf032d60403864b11d2f	41-50	3

Here the SHA-256 cryptographic hash function has been used.

b. *Footfall sensors detect and record WiFi probes from devices located or passing in close proximity. The codes are hashed at the point of collection and then these codes are used to identify stationary devices such as printers during data cleaning. The individual hashed observations are then sent to the researchers' secure server where they are cleaned to provide estimates of footfall.*

Case Study 4: Ethnicity Estimation

Stage 1: Researchers obtain the names of a sample of company directors from Companies House website. The names of all Kenyan students graduating in 2018 are obtained from different institutional websites registered in that country.

Stage 2: Following successful application, CDRC provides those researchers with software by secure file transfer. This makes it possible for those researchers to estimate the ethnicity of every company director and the gender of every Kenyan graduate. The results are used to research (a) the success of individuals of Kenyan origin in becoming company directors and (b) whether success is proportional to academic prowess as indicated by male and female success in obtaining academic qualifications in Kenya. Although the researchers makes predictions at the level of the human individual these are kept on the researchers' secure site and results are only published in aggregate with suppression of very small data sets.

*** If the Office for National Statistics were unhappy with individual level profiling, would it also be acceptable to conduct this analysis by reporting aggregated figures of probable Kenyan company directors (with suppression if the estimate is below 3)? (Similarly, the software would be adapted to provide only an aggregate estimate of educational attainment.)*

*** If, instead of supplying software, CDRC took delivery of a list of paired given and family names (without knowing whether or not they pertained to any living individuals), could it code up the names by predicted ethnicity and age and send the results back to the supplier? If CDRC has no measure of obtaining or matching with further data that would identify those names as relating to living individuals, this would not be personal data in CDRC's hands. However, if would be personal data if the hands of the supplier who would have to comply fully with GDPR requirements. Would any secure data transfer or pseudonymisation procedures be required?*

Case Study 5: A Synthetic Population

Stage 1: Data are collected for the Census of Population by the Office for National Statistics, and aggregated small area statistics on ethnicity, employment and age are published (subject to suppression of small counts in accordance with established anonymization procedures).

Stage 2: A synthetic population model is applied to the Census data to model the ethnicity, employment status and age of every adult in Anytown.

Stage 3: Names and addresses are obtained for eligible voters whose names appear on the contemporaneous public version of the Electoral Register (estimated as 80 per cent of all voters and 65 per cent of the resident adult population). These data are matched to the Census data at small area level. The ethnicity and age of each named individual is estimated based upon their forename and surname.

Stage 4: The names and addresses from the Electoral Register are matched with members of the synthetic population in order to add estimates of the likely employment status of those eligible to vote. The results are used to analyse differences in the employment status of voters relative to non-voters in Anytown.

Stage 5: In order to extend the analysis and to better understand the employment characteristics of non-voters, names and addresses of individuals who are likely ineligible to vote are obtained from a data reseller. All the processing takes place on CDRC's

secure servers. Consents for re-use of the data were obtained when the data were collected in 2011 and a notice advising of the study has been posted on the CDRC website. The modelling and matching procedures are used to create a further classification of the ethnicity and age characteristics of Anytown adults that are ineligible to vote, as well as individuals who opted out of inclusion in the public Electoral Register.

Note on synthetic population data

Neither CDRC nor its client will take any measures or decisions in relation to any particular individual. There are no other factors about the project that are likely to cause substantial damage or distress to any individual. The safeguarding conditions have therefore been met. Provision of transparency notices to individual data subjects would take so long and cost so much that the research project would not be viable. Since this would be a disproportionate effort and giving that the safeguarding conditions have been met, CDRC does not need to provide transparency information to each individual.

It is conceivable that the synthetic population model could allow information to be inferred about natural individuals. For example: assume that only a minute proportion of Anytown's population belong to a particular ethnic group, and that of this minority, only a tiny number of individuals are employed and above the age of 75. This could hypothetically make it possible to make inferences about natural individuals based on the data held in corresponding profiles in the synthetic population.

This risk is highlighted in the EU's Article 29 Working Party paper on Big Data[2]:

> big data processing operations do not always involve personal data. Nevertheless, the retention and analysis of huge amounts of personal data in big data environments require particular attention and care. Patterns relating to specific individuals may be identified, also by means of the increased availability of computer processing power and data mining capabilities.

Accordingly, the question of whether any data within a synthetic population could be personal data for GDPR purposes can only be answered by considering the risk that the data could enable information to be inferred about natural persons.

The data contained in a synthetic population can be considered to fall outside the scope of the GDPR if it falls within the GDPR's definition of anonymous personal data. This will be data that does not relate to an identified or identifiable natural person or personal information which has been rendered anonymous in such a manner that the data subject is not or is no longer identifiable.

If there is any possibility of the data comprised in the synthetic population allowing for the identification of a data subject, then it is more likely for GDPR purposes that this is pseudonymised data than anonymous personal data and it should be treated as falling within the scope of the GDPR.

Case Study 6: Data Perturbation using Micro-aggregation

A research institution is carrying out research into the relative financial standing of individuals living in different postcodes of North London.

[2] http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp221_en.pdf

The researchers hold a dataset showing the age, gender, postcode, income and average monthly expenditure of 1,000 people living in North London.

The researchers believe there is a risk of identification if actual incomes are made available.

They choose to use micro-aggregation to disguise the actual incomes. They will replace the actual values of incomes in the datasets with average values for small groups of the units.

The groups all contain a minimum predefined number "k" of units .K is a threshold value and each group is called a k-partition.

The researchers divide the 1,000 individuals in the dataset into fifty k-partitions of 20 individuals. The incomes of individuals falling into the same k-partition will be represented in the dataset using an identical value.

The advantage of this technique is that the mean income value for the whole population remains unchanged. If the researchers wished to increase the amount of data perturbation, they could apply micro-aggregation to all of the variables in the dataset in order to achieve *k-partitions* representing average values based on all of the variables in the data set. This is known as *multivariate micro-aggregation*.

Case Study 7: Encryption and noise addition

An online sportswear retailer analyses its customers' buying habits so that it can recommend certain products to them.

The sportswear retailer along with some of its competitors has been asked to take part in a research initiative organised by a third party research body. The research requires correlating shoppers' buying habits with public health data about arthritis.

Each participating retailer uses a secure-keyed cryptographic hash to create unique reference numbers from customers' names and addresses. GP practices use the same algorithm to generate unique reference numbers from patient details. Once their datasets are created both the sportswear retailer and the GP practices delete the key used for hashing.

The datasets supplied by the sportswear retailer and GP practices will match together. The research body could add another round of encryption in order to ensure that neither the GP surgeries nor sportswear retailer could ever link the data back to individual patient's or shopper's identities.

In addition the randomization technique of noise addition could be used to make it harder for a third party to identify an individual should they be able to detect how the data has been modified.

Appendix 2. Real Case Studies of Failures to Achieve Data Protection Principles

AOL

In 2006, for the purposes of facilitating research, AOL released a list of 20 million web search queries that had been made by over 650,000 AOL users. Each search query was released alongside a number to represent the AOL user who had input the query.

By analysing all search queries made by a search user who had been attributed by AOL with user number 4417749, the New York Times was able to work out the identity of that user.

Among the serious acts of negligence that can be found in this case study are the failures to consider fully the implications of:

- multiple entries relating to the same individual being available within the database; and
- the fact that data within the dataset could be corroborated against publicly available data.

Netflix

In 2006 Netflix released a database of more than 100 million ratings on a scale of 1-5 on over 18,000 movies given by almost half a million users. The data was supposedly "anonymised" according to an internal privacy policy with all customer identifying information removed except ratings and dates. Noise was added in order to slightly increase/decrease ratings.

It was found however that the data could be de-anonymised by looking for corresponding ratings among the publicly available ratings on the Internet Movie Database (IMDB).

This case study highlights the importance of bearing in mind that third parties may compare research datasets against other publicly available datasets. Enhanced computer processing power is making this easier for third parties to achieve.

The Royal Free and Google DeepMind

In 2017 the Royal Free NHS Foundation Trust was found by the ICO to have failed to comply with data protection law when it provided personal data of around 1.6million patients to Google DeepMind as part of a trial to test an alert, diagnosis and detection system for acute kidney injury.

Along with a number of other failures on the part of the Trust, an ICO investigation found that patients were not sufficiently informed that their health data would be used for the trial. Following the ICO's investigation the Trust was required to take a number of steps including:

- establishing a lawful basis for the Google DeepMind project;
- setting out how the Trust's duty of confidence to patients in any future trial will be met;

- carrying out a privacy impact assessment; and
- commissioning an audit of the trial involving Google DeepMind.

Elizabeth Denham the Information Commissioner commented on this case:

> "[t]*here's no doubt the huge potential that creative use of data could have on patient care and clinical improvements, but the price of innovation does not need to be the erosion of fundamental privacy rights.*
>
> *"Our investigation found a number of shortcomings in the way patient records were shared for this trial. Patients would not have reasonably expected their information to have been used in this way, and the Trust could and should have been far more transparent with patients as to what was happening.*"