Queen Mary
University of London
Barts and The London

CANCER
RESEARCH
UK

BARTS
CENTRE

Standard Operating Procedure

SOP Title	De-Identification of Personal Data
SOP Reference	BCC-ISM-SOP-02
Version Number	1.0
Approval Date	03/07/2019
Effective Date	24/07/2019
Review Date	03/07/2020

Author	Madalyn Hardaker Name	Information Governance Lead Position
Reviewed by	Ivana Sestak Deeksha Prabhu Kristel Caisip Rachel Barrow-McGee Name	Lecturer in Medical Statistics Senior Research Applications Programmer Applications Developer Tissue Acquisition Officer Position

Approved by	Jonathan Croft Name	Head of Research Computing Position
	Signature	Date 03/07/2019.

Version	**Date**	**Reason for Change**	**Updated by**
NB: This document is based on and supersedes the CCP Anonymisation and Pseudonymisation SOP V1.0 by Richard Ostler			
1.0		Initial Version	Madalyn Hardaker

Table of Contents

Abbreviations

GDPR	General Data Protection Regulation (2016)

Glossary

Anonymous/ Anonymised	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. (GDPR Recital 26)
Anonymous-in-context	Data which has been subject to robust de-identification measures and the environment for use is strictly controlled by legally binding contractual clauses.
Data Asset Registry Toolkit	A custom web application accessible to BCC staff. It is used for logging core data assets and personal information data files held by the centre.
Data protection by design	The consideration and implementation of technical and organisational measures at the time of planning a processing system to protect data safety.
Dataset	A file that contains one or more related records
Data Subject	A person who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity. (GDPR Article 4(1))
De-Identification	The technical process of transforming information to make identification of an individual less likely.
Direct Identifier	A field or variable that can reasonably link back to an individual person on its own, e.g. Name, Home Address, NHS number, IP address, etc.
Indirect Identifier	A field or variable that could reasonably be used with other information to link back to an individual person, e.g. Name of GP, Place of employment, Location of medical treatment, Medical diagnosis, etc.
Personal data	Any information relating to an identified or identifiable natural person ('data subject'). (GPDR Article 4(1))
Processing (data)	Any operation or set of operations which is performed on personal data or on sets of personal data...such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction. (GDPR Article 4(2))
Pseudonymisation	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. (GDPR Article 4(5))
Record	A basic unit of information in a dataset, e.g. a line item
Re-Identification	The processing of data alone or in combination with another datasets which results in persons becoming identifiable; also known as de-anonymisation.

1. Objective

To ensure robust procedures are in place to protect the privacy and confidentiality of data subjects by effectively de-identifying datasets containing personal information.

To prevent the inappropriate disclosure of personal data and the re-identification of a person from a de-identified dataset.

2. Scope

This SOP is written with particular consideration for health and social care data processed within Barts Cancer Centre, but may also be relevant to staff who process other types of personal data.

3. Background

Data lies on a spectrum of identifiability. At one end of the spectrum, a person is fully identifiable; then, as information is removed or obscured, it becomes increasingly difficult to identify who that person is. De-identification is the process used to prevent a person's identity from being connected with information, i.e. moving the data away from the 'fully identifiable' end of the spectrum.

As an organisation that processes personal data, staff have an obligation to protect the privacy and confidentiality of data subjects under the EU General Data Protection Regulation (2016), Data Protection Act (2018), Human Rights Act (1998) and common law relating to information obtained in confidence.

When personal data is collected for medical research, the legal basis for its use is, almost without exception, as a task in the public interest and processing of this data is therefore subject to technical and organisational safeguarding measures. Additionally, in the vast majority of cases, consent will have been obtained from the data subject or the project will have been granted support from the Confidentiality Advisory Group under Section 251 of the National Health Service Act to address the duty of confidentiality. The documentation of these measures facilitates the assurance of transparency, accountability and a high standard of ethical responsibility.

In order to maximise the research value of datasets, funding bodies may require generated datasets to be made available publicly or shared with other parties. It may also be appropriate to make datasets available as part of QMUL's general commitment to data sharing or for other collaborative endeavours across other organisations. In the majority of these instances, de-identification measures must be applied.

There is no single technique to achieve effective de-identification; rather, de-identification is achieved by the application of multiple techniques based on the scenario. The chosen strategy should be proportional to the risk and ensure that the data is still fit for purpose. Some common strategies and examples are described in Appendix A, there are also multiple applications and software packages available which can apply these and other de-identification techniques, however certain transformations are best carried out by a statistician or another individual with the technical tools and expertise to execute the technique and verify the result. It is always good practice to obtain a second opinion if unsure of what de-identification strategy is most appropriate.

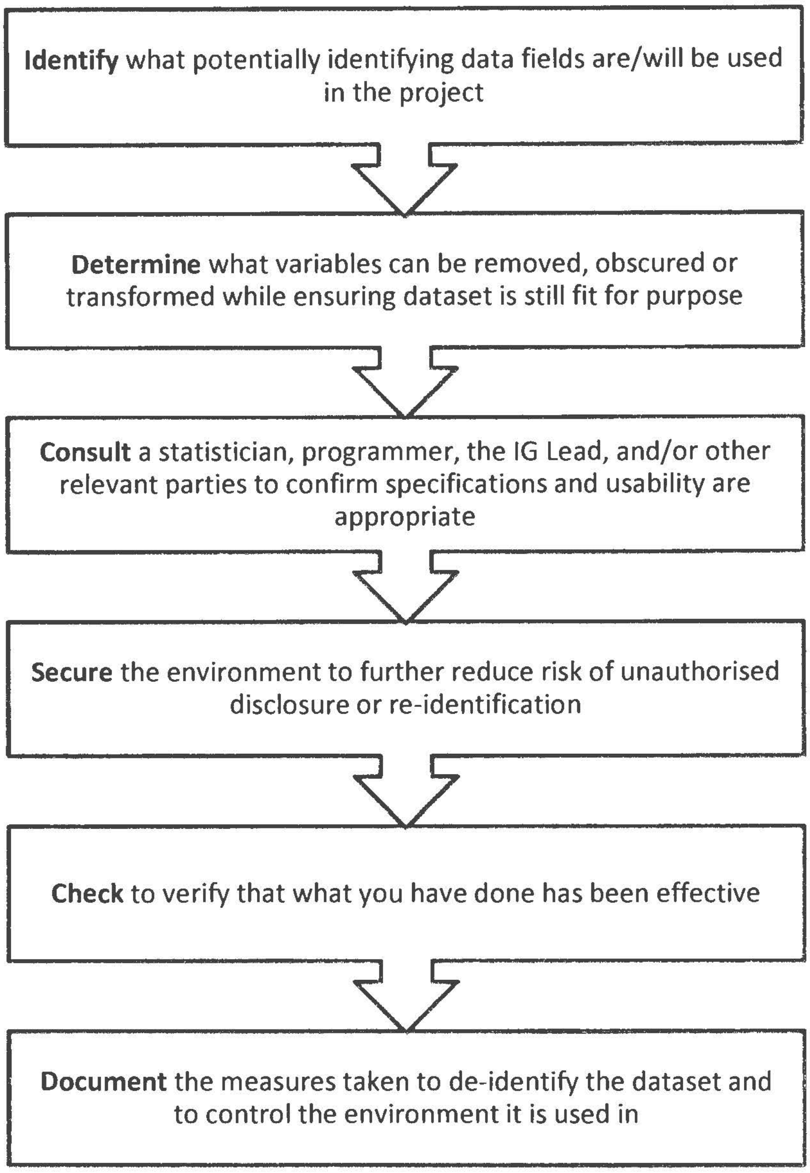
4. Roles and Responsibilities

All staff	• All staff must be trained in data security (e.g. Data Security and Protection training), understand their responsibilities under data protection law, and be knowledgeable of the data being processed and the risks that may be associated with re-identification.
Application developers and programmers	• Facilitate separate views of pseudonymised and identifiable data based on user access roles within web applications when appropriate.

	• Execute appropriate de-identification techniques and ensure the technical aspects of de-identification are performed according to the specifications agreed with the project team and other interested parties. • Assist in the annual audit of the BCC anonymisation and pseudonymisation processes as required by the Data Security and Protection Toolkit, and review related documentation and guidance.
BCC IT Helpdesk	• Facilitate additional technical measures to protect confidential personal information and/or prevent the re-identification of de-identified data (e.g. encryption, restricted access folders, secure web transfer, etc.)
Project owners	• Implement a "data protection by design" approach to project planning, taking into account how personal data is going to be collected, processed, stored and shared from the beginning stages of project design and making sure team members have appropriate levels of access based on their assigned tasks and responsibilities. • Ensure appropriate measures are in place before sharing data, e.g. data sharing agreements. • Maintain up to date records (e.g. on the Data Asset Registry Toolkit (DART)) of what personal data is held and what de-identification measures have been applied. • Record who has access to personal identifiable data and maintain audit trails wherever possible.
Statisticians	• Analyse data • Consult on what de-identification measures are most appropriate and execute certain de-identification measures where appropriate • Ensure de-identified/transformed data is usable and fit for purpose

5. Procedures

Process Overview



Identify what potentially identifying data fields are/will be used in the project

Determine what variables can be removed, obscured or transformed while ensuring dataset is still fit for purpose

Consult a statistician, programmer, the IG Lead, and/or other relevant parties to confirm specifications and usability are appropriate

Secure the environment to further reduce risk of unauthorised disclosure or re-identification

Check to verify that what you have done has been effective

Document the measures taken to de-identify the dataset and to control the environment it is used in

5.1 Determine an appropriate de-identification strategy

5.1.1 Staff processing personal information must understand the potential for data to identify individuals and consider how data could become identifiable if combined with other information. The potential for identification will depend on the context in which it is presented and the resources available to those accessing it. For example: a rare ethnicity category in a particular area may identify an individual with minimal additional information, and someone with access to records such as local demographics may have the information required to re-identify that person.

5.1.2 Assess the level of identifiability based on the data being collected, using the following as a guide:

Name
Initials
Title
Date of Birth
Date of Death
Ethnicity
Address
Postcode
Email
Phone number
Rare event codes
Outliers and small cohorts
NHS Number
Hospital / GP name
Treating hospital or GP surgery
Gender
Education
Treatment/appointment dates

This is a qualitative assessment to approximate the where the data falls on the identifiability spectrum, and to understand what risks are most applicable and what measures can be put in place to mitigate them.

5.1.3 Consider how the data needs to be used, and decide what is the minimum amount of data that can be used while still achieving the desired outcome. It is likely that some data fields will be appropriate for some processing activities but not others, so multiple strategies may need to be considered to account for all of the processing activities.

5.1.4 Consult the programmers and statisticians involved in the project to come to an agreeable strategy.

5.2 Secure the data environment

5.2.1 Save digital files in restricted access folders and physical files in locked cupboards or drawers, accessible only to authorised personnel.

5.2.2 Encrypt confidential personal data in transit and at rest.

5.2.3 Maintain a register of personal and confidential data, such as the Data Access Registry Toolkit (DART), noting:

- Location of data
- Location of data sharing/use agreement(s)
- Expiry/Destruction dates
- Data/Document format
- Identifiable data fields
- For personal data, the legal basis for processing (e.g. GDPR Article 6(e) Public Interest)
- Access restrictions applied
- Purpose of the dataset
- Documentation related to processing (e.g. data management or sharing plan, relevant transformations applied, etc.)
- Dataset classification (i.e. confidential, restricted, open)

5.2.4 Follow safe workspace procedures, particularly when working with personal information (e.g. minimise or close browser and document windows when unauthorised staff are present, lock unattended work stations, do not print unless necessary, etc.).

5.2.5 Do not copy confidential personal information to any portable device unless there is no other practicable means; if a portable device is used it must be encrypted.

5.2.6 Do not link personal information to clinical data unless it is necessary to do so to achieve the purposes of the processing. Where applicable, use a system such as the bespoke PID Application built by the programming team to segregate clinical details from patient identities. With valid log in credentials, the application provides a text area to enter an SQL SELECT statement which on submission returns the requested details for a specific study. The application is built only to read data and not for update of the PID and is only available within the BCC network. If used, access to this application should be provided only to select individuals.

5.3 Verify de-identification

5.3.1 Consider the data fields and the expertise, knowledge and resources of the person(s) who will be receiving the data or have access to it and attempt to re-identify individuals as if you were a "Motivated Intruder." That is to say, assume the following:

- the intruder is reasonably competent and has a compelling motive, but no special skills (e.g. hacking skills)
- the intruder has access to basic tools such as a computer with internet, but no specialised equipment (e.g. for hacking or spying)
- the intruder has access to resources such as social media, newspapers, libraries, public documents, registries and archives, and would employ investigative techniques such as making enquiries of people who may have additional knowledge or advertising for anyone with information to come forward.

5.3.2 There are a number of tried and tested resources, such as core libraries (e.g. PatientDetailsCRF.jar, CRFTools.jar), which may not need to be individually re-tested each time, but any new processes or technologies should be verified individually and then the whole de-identification strategy should be evaluated as a whole as well.

5.4 Document the process

5.4.1 Document all steps and rules used to de-identify the dataset, including transformation code. Note, code used for irreversible anonymisation must not be retained if it could be used to reverse engineer the anonymised dataset, and any data linkage tables must be destroyed or severely restricted (and documented as such on the data register).

5.4.2 Additionally, document the tests undertaken to verify that data are effectively de-identified, defining the context in which data are assumed to be effectively anonymised, if applicable, and include a risk assessment of the potential for re-identification.

6. Additional Resources

6.1 Forms and Templates

n/a

6.2 Policies and Procedures

BCC-ISM-Policy-03 De-Identification

BCC-ISM-Policy-04 Data Sharing

BCC-ISM-SOP-03 Data Sharing

BCC-ISM-WI-02 Techniques for Dataset De-Identification

6.3 Legislation

Article 29 Working Party Opinion 05/2014 on Anonymisation Techniques
https://www.pdpjournals.com/docs/88197.pdf

6.4 Other

Anonymisation: managing data protection risk code of practice. Information Commissioner's Office
https://ico.org.uk/media/1061/anonymisation-code.pdf

The anonymization decision-making framework (University of Manchester, University of Southampton, the Open Data Institute, and the Office for National Statistics) http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf

Understanding Patient Data (Wellcome Trust, Medical Research Council, Department of Health and Social Care, Public Health England, Economic & Social Research Council) https://understandingpatientdata.org.uk

Templ, M., Meindl, B., & Kowarik, A. (2019, May 16). PDF. Vienna, Austria: Data-Analysis OG. https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc guidelines.pdf

Appendix A: Example techniques for de-identification

Stripping is the process of removing whole data fields from a dataset. Candidates for stripping would normally include data fields containing names, addresses, hospital numbers, NHS Numbers. Stripping is also known as data reduction.

Suppression is a technique which does not remove the data field, but all non-null values are replaced with a standard character. This indicates where data are provided or missing without disclosing the actual values.

Pseudonymisation is a process applied to a dataset where individuals are identifiable from a single unique identifying number or a compound of data items which uniquely identify them. Pseudonymisation involves replacing the original identifier(s) with a new, unrelated, one (e.g. a "subject identifier" or "patient ID").

> ***NB:*** *The dataset may still be identifiable based on other data fields or combinations of data fields, so pseudonymisation alone is not sufficient to consider a dataset effectively anonymised.*

Aggregation techniques pool data so that totals, rather than individual values, are provided. E.g.:

Original Data

Patient 1	Alive
Patient 2	Alive
Patient 3	Deceased
Patient 4	Alive

→

Aggregated Data

Alive Patients	3
Deceased Patients	1

Derivation, also known as categorisation, banding and data blurring, is a technique for replacing exact values with coarser grained descriptions of values. Examples of derivations include:

- Replacing measurement values with bands or ranges
- Replacing narrow categories with higher level or broader categorisations
- Rounding measurement values, either rounding decimals to lose precision or rounding up to a multiple of rounding base (e.g. 6,7,8,9 rounded to 5 or 10).
- Removing day and month from a date, or replacing with an age
- Recoding dates to preserve time periods
- Resetting sequential dates to a new random baseline date
- Replacing dates with days since baseline

Original Data

Patient ID	DOB
Patient 1	04/05/1973
Patient 2	01/12/1970
Patient 3	23/09/1967
Patient 4	13/02/1972

→

Derived Data

Patient ID	Age at Visit 1
Patient 1	46
Patient 2	49
Patient 3	52
Patient 4	47

Swapping transposes individual values between records such that the values no longer relate to particular individuals, but overall totals and frequencies are preserved. It may be advantageous to consider swapping between partially matched individuals, for example matched on age.

Perturbation alters data values to guard against data linkage. Barnardisation is a common perturbation technique whereby 1 is randomly added or subtracted from some values.

If there is no effective de-identification possible for a record, the whole record may be removed from the dataset.