SGUL motivated intruder test for anonymised and pseudonymised data SOP

Document Information	
Document Name	SGUL motivated intruder test for anonymised and pseudonymised data SOP
Author	IG Manager
Issue Date	18/11/2018
Approved By	IGSG
Next review	17/11/2020

Document History		
Version	**Date**	**Summary of change**
0.1	18/11/2016	First draft for discussion
1.0	22/11/2016	Final version approved by IGSG and SIRO
2.0	04/12/2018	Updated DSP version

1 Scope

This method for a motivated intruder test must be used by research studies in respect of anonymized and pseudonymised data used throughout SGUL to mitigate information risks to an acceptable level.

2 Responsibilities

Each PI for a study or project is responsible for ensuring that appropriate anonymization and pseudonymisation techniques for their respective study or project are deployed and that they are subject to a motivated intruder test. This requirement is applicable for those research studies that are processing personal or sensitive personal data wherever they are required by SGUL Policies and compliance with the DSP Toolkit.

3 Background

The Information Commissioner's Office has developed the Motivated Intruder Test (MIT) to ensure that data controller's discharge their obligations in relation to ensuring the confidentiality and privacy of data subjects in respect of anonymised or pseudonymised data. The objective for the data controller is to assess the potential for re-identification of individuals in relation to the pseudonymised or anonymised data that is released when coupled with other publicly available information that is then further processed to yield the identification of actual or potential data subjects.

Clearly the more sensitive the data, the greater the number of controls that should be utilised by the data controller to effectively de-identify the data.

The inclusion of spatial data in relation to the data subjects must be carefully considered. It is therefore essential that when dealing with such data from a small locale that any other quasi identifiers are removed and that even details of the locale are anonymised. Other techniques should consider the use of increasing the locale to a larger geographical area.

4 Information Commissioner's Approach

The motivated intruder is deemed to be a reasonably competent person without recourse to technical specialist skills such as hacking. However, they would have access to other publicly available information (internet, libraries and other public records) and perhaps even utilise social engineering techniques to obtain information about an individual from other potential sources.
The motivation for the motivated intruder can be several fold and may include some, or a mixture of the following:

a) The goal of simply being able to identify an individual from an anonymised dataset;

b) Criminal objective in relation to the person identified or the organisation that has released the data;
c) Revealing the identity of a celebrity for journalistic, revenge or embarrassment purposes;
d) Political or activist purposes e.g. a campaign against an organisation, or individual
e) Curiosity

5 Risk assessment techniques for potential re-identification

i) Is it possible to link data from a variety of different sources to build up a picture of an individual, is the same pseudonym used across several different studies? Internet searches to obtain further information based upon the dataset subject and quasi identifiers such as gender, age, DOB, post codes etc;
ii) What other linkable information is available publicly? Use of the electoral register to link anonymised data to identity;
iii) How much information does the motivated intruder possess? Consider existing prior knowledge and how this could be used by a Motivated Intruder to obtain further information. Use of other publicly available archives such as national and local newspapers and social engineering;
iv) Use of social networking sites to see if there is information to corroborate anonymised data with a user's profile
v) Consider any risks posed by the audience to whom the data will be released to. In the case of clinicians and researchers, with prior knowledge, they are of course bound by professional ethics.

You should keep in mind that the risk assessment techniques for potential re-identification may change as a consequence of technical development, legal changes and that there may be different techniques available for different data sets.

6 Is there a distinct risk of a motivated intruder?

Identify the target audience for the research data release. The approach should be to consider, is there a distinct possibility of a motivated intruder based upon the data set subject? Is it likely that in the case of this dataset the motivation, may come from individuals in one of the groups a - e identified above. In view of the target audience for release of the dataset you should consider whether it is feasible that any individual would be motivated to re-identify an individual. If this is your opinion it should be documented.

Identify the applicable group(s), and document your conclusions as to why they have an interest in the Dataset Subject.

Thereafter you should identify the applicable risk assessment technique for each of the fields of data held (copied below). It may be that because of the audience taking receipt of this data there may not be, in your opinion, any applicable risk assessment

technique for possible re-identification. However, you should document your reasons for this, should it be applicable.

7 Confirm that business process changes work effectively

7.1 Confirm that the research participants cannot be identified from data that are used to support secondary purposes.
7.2 All research participant data in identifiable format must enter and exit through a SGUL Data Safe Haven (DaSH), and only exit in identifiable format if there is legal or regulatory justification.
7.3 Personal or sensitive personal data may be used in an identifiable format for exceptional purposes within the Dash boundary.
7.4 Onward disclosure should ideally be limited to pseudonymised or anonymized data.

8 Confirm Dash processes are working effectively

8.1 Confirm that the Dash controls cover logical and physical Dash boundaries.
8.2 Confirm that a physical Dash cannot be accessed by unauthorized personnel.
8.3 Confirm that a logical Dash cannot be accessed by unauthorized personnel.
8.4 Confirm if personal or sensitive personal data is received by post, is it received in a physical Dash?
8.5 The physical security controls should be layered such as swipe card and physical security locks.
8.6 Confirm that there is layered logical security such as access controls that have user id and password with multi factor authentication (at least two levels).
8.7 Confirm that there is evidence of standard operating procedures (SOPs) for authorized access to the DaSH.
8.8 Confirm the 'secure data processing environment' has sufficiently robust SOPs for the removal and transfer of personal and sensitive personal data from the DaSH.

9 Confirm that technical pseudonymisation and anonymisation are working effectively

9.1 de-identified
9.2 Confirm that expected controls are being used - ICO Anonymisation: managing data protection risk - Code of Practice 2012
9.3 **Conduct a MIT risk assessment using the following procedures, where relevant and appropriate, noting those that you have acted upon and those that you have dismissed citing the rationale**:
9.4 A web search to determine if a combination of date of birth and postcode would be sufficient to yield someone's identity.
9.5 Searching the archives of a national or daily newspaper to see whether it is possible to identify a person's name.
9.6 Use of social networking to see it it's possible to link anonymized data to a user's profile.
9.7 Use of the electoral register and local library resources to link anonymized data to someone's identity.

9.8 Is there any other publicly available information that could be utilized to achieve re-identification.

9.9 What are the risk to re-identification posed by prior knowledge, such as a family member (Firstly, consider is that person going to learn something new. Secondly, if you are concerned about other medical staff/researchers this should be mentioned but dismissed on the basis that they are duty bound to confidentiality by a professional code of ethics as well as contractually. Thirdly recorded information and established fact are areas for consideration in this process however, it is accepted that it is impossible to determine personal knowledge of a would be motivated intruder.)

9.10 Educated guesswork may also be a cause for concern when releasing anonymized data and a cautious approach may be necessary to safeguard the interests of individual research participants.

9.11 In releasing information are there any risks to be considered for groups of people identified in the research that may lead to privacy or other risks (e.g. hate crime towards ethnic minorities) as opposed to other benefits?

9.12 Anonymised data and open data initiatives provide more opportunity for such data to become personal data when a motivated intruder is able to combine, analyse and match publicly available data thus creating personal data.

9.13 If publishing spatial data (e.g. post codes) ensure that you have a documented Data Protection Impact Assessment (DPIA) and balance what is published with the protection of the research participant's privacy.

9.14 In publishing post code related data you should be mindful and take into consideration the following:

- full postcode = approx 15 households (although some postcodes only relate to a single property)
- postcode minus the last digit = approx 120/200 households
- postal sector = 4 outbound digits + 1 inbound gives approx 2,600 households
- postal district = 4 outbound digits approx 8,600 households
- postal area = 2 outbound digits approx 194,000 households

Source: Centre for Advanced Spatial Analysis - UCL

('Outbound' is the first part of the postcode, 'inbound' the second part; for example with the postcode SW17 0RE , the outbound digits are SW17 and the inbound digits are 0RE.)

10 Confirm the fields of data held

10.1 Confirm that only the minimum necessary data fields are held and are appropriate to the purpose.

10.2 Confirm that where possible aggregated data is held in respect of such fields as date of birth, date of death as per the table below.

10.3 The following fields should be treated with caution and the recommended treatment followed:

Data Field	Recommended treatment
Patient name	No name data items supplied

Patient address	No address data items supplied
Patient Date Of Birth (DOB)	Replace with age band or age in years
Patient postcode	Postcode sector and other derivations*
Patient NHS Number	Pseudonymised with consistent values; different values for different purposes for the same user (for one-off events the data should be pseudonymised)
Patient ethnic category	Only supply if relevant to the study
SUS PbR spell identifier	Data should be pseudonymised
Local Patient identifier	Data should be pseudonymised or do not display it
Hospital Spell number	Data should be pseudonymised or do not display it
Patient Unique booking reference number	Data should be pseudonymised
Patient Social Service Client identifier	Data should be pseudonymised or do not display it
Any other unique identifier	Data should be pseudonymised
Date of Death	Truncate to month and year

In the event of using unique reference numbers be mindful of the audience and the ability to identify a living individual. If the use of an NHS number is contemplated consider the audience and their ability to decipher that number by access to a Hospital Patient Administration System (PAS), or the prospect of social engineering using bribes or blackmail to obtain the identity of the patient.

Caution must also be exercised with the use of rare diagnostic codes in pseudonymised data sets as it is possible that these may have the potential to infer identity.

11 Select control objectives and controls for treatment of the risks

11.1 Appropriate control objectives are selected from Annex A of ISO27001:2013 and ISO 27002:2013 by the Head of IT Services and the IG Manager and the reasons for the selections are documented IG11 – SGUL Risk Assessment Register Feb 2016 v0.2 identifying any existing control or whether another control has to be developed in the light of the conclusions to the risk assessment and risk treatment processes. Typically the following ISO27001 controls should be used:

18.1.1 - Identification of applicable legislation and contractual requirements

18.1.3 - Protection of records

18.1.4 - Privacy and protection of personally identifiable information