Identifiability, anonymisation and pseudonymisation

Research usually depends on the use of person-level information, and in a well-connected world, where there is access to vast amounts of information, obtaining total anonymity of person-level information may not be possible. However, to undertake research whilst maintaining appropriate ethical and legal standards, it is crucial to understand:

- What constitutes identifiable information, and how the law applies.
- How research information can be effectively rendered anonymous.
- What controls must be in place to render the risk of identification no longer 'reasonably likely'.

Although identifiability is an important concept in the legal definitions of both Personal Data (data protection law) and Confidential Information (common law of confidentiality); both of these legal terms require other criteria to be fulfilled in addition to identifiability. This guidance covers the general legal principles underpinning identifiability. It does not specifically address what makes data Personal or what makes information Confidential.

1. Identifiability of information

In some cases, it is clear that a piece of information directly identifies individuals. This is information that contains direct identifiers, like names or email addresses.

Other information may not directly identify individuals, but when combined with other accessible, or already known, information, could identify individuals. This is often referred to as 'jigsaw' identification and can occur either deliberately or inadvertently. In reality there is a continuum of identifiability: from complete anonymity at one end, through jigsaw identifiability, and on to directly identifying information at the other.

When thinking about identifiability consideration must be given to whom the information is identifiable. Who are the likely 'viewers' of the information, and how identifiable is the information from their perspective? Will the content of the information itself identify individuals (i.e. what identifiers does the information contain)? What other information is the 'viewer' likely to have access to, or know? In other words, what is the context in which the information is going to be viewed?

Information is considered identifiable if either; it directly identifies individuals, or if individuals can be identified when the information is viewed in combination with other accessible information. The same dataset can be classed as identifiable in one context and deidentified in another.

2. Inherently anonymous information

Sometimes no individuals could ever be identified from a piece of information, or from that information in combination with other pieces of information, for example, most aggregate statistics. It is worth noting that even though aggregate statistics are not person-level information, some may contain rare outlying individuals which could lead to identifiability in certain situations.

3. Working within the continuum of identifiability

Although identifiability is a continuum, the law is binary: information is considered either identifiable or anonymous. How then do we decide if research information is identifiable or not? This decision has important implications for how information is used, stored and shared with others.

In order to decide if information should be considered identifiable in law, both the context in which the information is to be viewed, as well as the content of the information, should be considered. Considerations include:

- Who is the viewer?
- What other information are they likely to have access to (or know), which could support 'jigsaw' identification.
- What is the probability that the information could be identifiable?
- Is the risk adequately controlled?

For example:

Is it reasonably likely that the information would be used to identify individuals? To address this question, all the means that might be reasonably likely to be used to identify individuals should be considered. What if someone were more motivated than most to identify an individual e.g. if the information were about a celebrity or ex-partner?

Pseudonymisation

It is common research practice to collect information about people and then pseudonymise it ready for use. That is:

- Strip all direct identifiers from the dataset.
- Attribute a study specific identifier to each individual.
- Use this study specific number or code to 'label' each research record.
- Maintain a cipher that links the study specific number to the identifiers.
- Keep the cipher physically separate from the pseudonymised dataset.

Pseudonymisation limits the risk of identification to some extent. It limits who has access to direct identifiers. It ensures that if either the cipher or the research information were stolen, lost or left out on view, that any disclosure would be limited. As such, pseudonymisation is a 'Technical and Organisational Measure' required by data protection law if Personal Data is being processed for research.

However, pseudonymisation alone is not the same as anonymisation. This is because the context in which pseudonymised information could be viewed has to be adequately managed to ensure that identification is not reasonably likely.

Is it possible to anonymise person-level information: controlling content and context?

In order to ensure that information can be considered anonymous (rendering it no longer identifiable in law) you should:

1. Remove all direct identifiers from the information.

2. Limit the potential identifiability of the remaining information, as far as is practical or appropriate. There are many techniques that have been developed and can be applied to this end (e.g. suppression of small numbers, derivation and encryption). This will help control the content of the information. However, to achieve anonymisation, the control of context in which the information will be viewed is equally important.

3. Ensure that the person or organisation holding or receiving the information does not have access to the cipher used to maintain the link between direct identifiers and the research information. Also, ensure that the person or organisation does not have access to, or know other information that may aid identification (e.g. they may deliver care to the individuals, and as such know details about them).

4. Ensure that appropriate controls are in place to limit the risk that those receiving the research information would attempt 'jigsaw' identification.

The nature of the controls will vary on a project-by-project basis. For example:

Sharing information with other organisations

One common control used to support anonymisation is to enter into a Data Sharing Agreement. In such an agreement the recipient organisation will agree to make no attempt to identify individuals (in light of the agreed processes that the information will be subject to). Such an agreement should also cover what must be done in the event of accidental re-identification.

Further context controls should also be considered. For example:

- Ensuring collaborating organisations have an appropriate information security and governance policies in place (including sanctions if employees do not comply).
- Taking into consideration requirements to comply with professional bodies' codes of practice, including sanctions for not complying.
- Using Trusted Research Environments so that information is shared in a limited manner with access and use physically restricted.

Sharing information within an organisation (i.e. limiting access to confidential information to those who have a duty of confidence)

If an organisation holds both the pseudonymised dataset and the cipher or code, the organisation is holding Personal Data, as defined in data protection law. Regardless of the 'controls' that are in place, the organisation has access to direct identifiers. Since data protection is a corporate responsibility, any internal controls are not considered sufficient to render the data no longer Personal Data. However, pseudonymisation does reduce the risk, to some degree, when processing Personal Data for research, and as such is a safeguard provided in data protection law.

It is important to highlight that the risk of common law disclosure within organisations can be effectively limited using pseudonymisation along with robust controls. For example:

- Pseudonymising the information to control the content of the information that the research team have access to.
- Limiting access to confidential information to individuals that have a duty of confidence with research participants
- Ensuring all those working with the information follow appropriate information security or governance policies (there should be sanctions if employees do not comply).
- Taking into consideration any requirement to comply with professional bodies' codes of practice, including sanctions for not complying.

Please note that an organisation cannot enter into a Data Sharing Agreement with itself.

4. What if research information cannot be anonymised?

The likelihood of identification is always greater in instances where occurrences are rare or unusual (e.g. a rare disorder, or a particularly young or old member of a cohort), and certain data linkages may render information significantly more identifiable. In a few circumstances it may not be possible to control the likelihood of identification sufficiently to render the information anonymous. When this is the case, research can continue provided that:

1. Any disclosure of confidential information is managed by other means, for example by managing the individuals' expectations through consent.

2. The use of such information is conducted in line with data protection law (if the information is Personal Data).

5. Identifiability in some difficult areas

You can't control what is already known

In some circumstances, researchers / clinical staff may have privileged information, which will enable them to identify individuals even from limited research information, e.g. when the information is about one of their patients. In such cases, it is not possible to anonymise information to these researchers / clinical staff. Any potential common law disclosure that might arise through the use of such information with these researchers / clinical staff must be managed through other means (e.g. by management of the individuals' expectations through consent), and in line with data protection law when the information is also Personal Data.

Genetic information

The principles that govern the identifiability of genetic information are the same as for any other type of information. Genetic information is considered identifiable if an individual could be identified from it alone, or from it when viewed in combination with other information that the viewer is reasonably likely to have access to (or to know).

When considering identifiability, a key difference between genetic information and any other person-level information, is that:

- Genetic information that identifies one person, may also identify (known or unknown) family members
- Identification may have substantial implications for the individual and for family members

Not all genetic information is unique to an individual, or even to a group of individuals. Most of the human genome is common to us all. However, researchers are often most interested in the differences between people and populations. As the level of uniqueness of the genetic sequence rises, so does the likelihood of identification.

Some types of genetic information are in the public domain (e.g. genealogy databases), and an increasing amount of genetic information is being made openly available along with identifiers, such as names. With time, we can only expect the possibility of jigsaw identification to increase.

Therefore the use of robust context controls when using genetic information is essential. These controls will be similar to those discussed earlier for all other types of information and include Data Sharing Agreements, policy compliance, employer sanctions, etc.

The identification of individuals or families from genetic information may be possible to those researchers / clinical staff with privileged information. They may be familiar with specific sequence patterns and know who they relate to. Obviously in these cases it may not be possible to anonymise such genetic information to these researchers / clinical staff (see above).

Key messages

Identifiability is a continuum. Knowing whether the information being used in research is considered identifiable is crucial in order to know how the law applies to this information and how the information can be used.

Identifiability is related to the information itself (the content) and the potential to identify individuals from combining this information with other information that the viewer may have access to (the context).

In order to robustly anonymise, both content and context need to be controlled so that it is not reasonably likely that individuals can be identified, using all means that are reasonably avaialble. When this standard is met, the information can be classed as anonymous.

Considerations of context should include:

- Who is ‘the viewer’ of the information?
- What access does ‘the viewer’ have to any cipher being used, to limit the risk of disclosure?
- What other information is ‘the viewer’ likely to have access to, and how likely are they to use it?

Controls should manage the risk of identification. Considerations include: the inherent identifiability and sensitivity of the content, the proposed use (e.g. linkages), and the availability of other relevant information. Controls include: data sharing agreements, policy compliance, professional standards / codes of conduct and physical environments.

The appropriate context controls will differ if anonymisation is being undertaken to ensure data is no longer Personal Data, compared to those required to ensure information is no longer confidential under common law.

The same anonymisation principles are relevant for genetic information.

There are some circumstances where robust anonymisation cannot be achieved.