Identifiability, anonymisation and pseudonymisation

Guidance note 5

This guidance was developed with the participation of the Information Commissioner's Office.

Research usually depends on the use of rich, person-level information. We are often interested in rare or unusual occurrences. We work in a well-connected world, where we have potential access to vast amounts of other information about individuals. In this context we acknowledge that obtaining total anonymity of person-level information[1] may be impossible. However, if we are to work effectively whilst maintaining appropriate ethical and legal standards, we need to be sure:

- When we are working with identifiable information, and how the law applies,
- How research information can be effectively rendered anonymous, and
- What controls must be in place to render the risk of identification no longer ‘reasonably likely’.

This is what we will cover in this guidance.

You should be aware that although identifiability is an important concept in the legal definitions of both Personal Data (GDPR) and confidential information (common law of confidentiality); both of these legal terms require other criteria to be fulfilled in addition to identifiability (see Guidance note 3). This guidance note covers the general legal principles underpinning identifiability. We will not be specifically addressing what makes data Personal or what makes information Confidential here.

1. Identifiability of information

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

Other information may not directly identify individuals on its own: but if you view it in combination with other bits of information you have access to (or that you know), you could identify individuals. This is often referred to as ‘jigsaw’ identification. Jigsaw identification involves putting together pieces of information to identify individuals. 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 you need to consider to whom it is identifiable. Who are the likely ‘viewers’[2] of the information? You need to consider how

[1] In research, person-level information can also be described as individual participant data (IPD).

[2] We usually worry if information is identifiable when we are managing the sharing of information (with collaborators, statisticians, other organisations), or when we are attempting to limit who has access to identifiable information within a research team, or when we are publishing research findings or making research data available through Open Data. The ‘viewer’ of the information will vary between each of these scenarios.

identifiable the information is from their perspective. Will the content of the information itself identify individuals (i.e. what real-world identifiers does the information contain)? What other information is the 'viewer' likely to have access to, or know? (i.e. 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 (i.e. jigsaw identification).

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.

Almost no person-level information, rich enough to be useful for research, could be considered inherently anonymous.

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? This decision has implications for how we might share the information with others, as well as how we might store and use it within our own organisation.

In order to decide if information should be considered identifiable in law, we need to consider the context in which the information is to be viewed, as well as the content of the information itself.

1. Who is the viewer?
2. What other information are they likely to have access to (or know), which could support 'jigsaw' identification, and
3. What is the risk that the information would be identified (i.e. likelihood and impact)? Is the risk adequately controlled?

A test to determine if information is identifiable is as follows:

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

It would not usually be sufficient to simply remove all real-world identifiers from a research dataset e.g. what we do when we pseudonymise information (see below).

3.1 Pseudonymisation

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

- Strip all real-world, direct identifiers from the research 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 or code back to the real-world identifiers, and
- Keep the cipher physically separate from the pseudonymised dataset.

Pseudonymisation will limit the risk of identification to some extent. It limits who, in an organisation, has easy access to real-world 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 GDPR if Personal Data is being held to support research (for more information see Guidance note 4).

However, pseudonymised is not the same as anonymous. This is because the context in which pseudonymised information could be viewed has not been adequately managed to ensure that jigsaw identification is not **reasonably likely** (by someone who may be more motivated than most).

3.2 Is it possible to anonymise person-level information? How do we robustly control content and context?

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

1. Remove all direct real-world identifiers from the information, and

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. They include e.g.
    a. Date of birth – change to age at recruitment,
    b. Post code – use only first part of post code, or change to indices of deprivation or equivalent, etc,
    c. More scientific techniques like Barnardisation (for certain tabular data), rounding (random or controlled), suppression of 'high risk' items, etc.

The above will help to control the content of the information. Some organisations are using technical solutions to automate the process of robustly controlling the content of 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 ready access to the cipher you are using to maintain the link between real-world 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 same individuals, and as such know pertinent details about the individuals concerned); and

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 appropriate controls will vary, depending on; why you want to anonymise, whose activities you are trying to control and the risk (i.e. likelihood and impact) of identification. When considering impact in a risk assessment[3], you should think about the sensitivity of the information and the potential distress that might be caused, if individuals were to be identified.

*3.2.1 Sharing information with **other** organisations*

One common control used to support anonymisation is to enter into a legal, 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 put to). Such an agreement should also cover what must be done in the event of accidental re-identification (including how to handle lessons learnt e.g. through Corrective and Preventative Actions, CAPA, process). An agreement should limit the purposes the information can be used for and limit further sharing.

If the risk of identification is considered high enough, further context controls should also be considered. For example:

- Ensuring collaborating organisations have an appropriate information security or governance policy in place (including sanctions if employees do not comply); and/or
- Ensuring all those accessing the information are appropriately trained, or can demonstrate relevant expertise, in information security or governance; and/or
- Taking into consideration requirements to comply with professional bodies codes of practice, including sanctions for not complying; and/or
- Consider using a ‘Safe Environment’ in high risk cases, so that information is shared in a limited manner with use physically restricted in a trustworthy environment etc.

*3.2.2 Managing common law disclosure **within** an organisation (i.e. Limiting access to confidential information to those who have a duty of confidence within your organisation)*

If your organisation holds both the pseudonymised dataset and the cipher or code, your organisation is holding Personal Data as defined in GDPR. Regardless of the ‘controls’ you have in place, the organisation has access to direct, real-world identifiers. Since data protection is a corporate responsibility, any internal controls are not considered sufficient here and it is not possible to render this data no longer Personal Data. However, pseudonymisation does reduce the risk when processing Personal Data for research, and as such is a safeguard provided in GDPR.

However, within organisations you can limit the risk of common law disclosure through anonymisation using robust controls. These might include:

- Pseudonymising the information to control the content of the information members of a research team have access to, and

[3] In some cases, your Data Protection Officer may conduct a DPIA (a Data Protection Impact Assessment) to determine the risk of identification.

- Ensuring all those working with the information follow appropriate information security or governance policies (there should be sanctions if employees do not comply) and/or
- Ensuring all those accessing the information are appropriately trained, or can demonstrate relevant expertise, in information security or governance, and/or
- Taking into consideration any requirement to comply with professional bodies' codes of practice, including sanctions for not complying, etc.
- *(An organisation cannot enter into a Data Sharing Agreement with itself.)*

4. What if you cannot robustly anonymise research information?
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). Certain data linkage operations 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:

1. Any disclosure of confidential information is managed by other means, for example by managing the individuals' expectations through consent[4], and
2. The holding and using of such information is conducted in line with GDPR (if the information also falls under the full definition of Personal Data).

5. Identifiability in some difficult areas

5.1 NHS number
NHS number is considered a real-world identifier (as is Community Health Index (CHI) in Scotland), as most NHS employees could have access to the cipher linking NHS or CHI number to an individual. The risk of identification has been deemed too great to assume that this information can be anonymised, even with the use of Data Sharing Agreements etc. Sharing, holding and using NHS or CHI number should be managed accordingly (i.e. these should be handled as identifiable information).

5.2 You can't control what you already know
In some circumstances, researchers / clinical staff may have privileged information, which will enable them to identify individuals from fairly limited 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 sharing of such information with these researchers / clinical staff must be managed through other means (e.g. by direct management of the individuals' expectations through consent), and in line with GDPR when the information is also Personal Data, e.g. your research participants are still alive.

[4] Consent in this context is not referring to GDPR consent (as defined in the Regulation), rather consent that is sufficient to manage participants' expectations with respect to the common law of confidentiality.

5.3 Genetic (sequence) 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). Again, you must consider both the content of the genetic information, and the context in which it will be viewed.

When considering identifiability, a key difference between genetic information and any other person-level information, is that identification may have increased implications for family members not only for the individual themselves. As such, the impact of identification can be greater.

Not all genetic (sequence) 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.

Currently the amount of information generally available to identify individuals from unique or rare sequence information is very limited. However, some types of genetic information are more heavily published than others, again increasing the potential risk of identification (if such publication also includes further identifying information, e.g. male inherited sequences and genealogy databases). An increasing amount of sequence information is being made openly available along with identifiers, such as surname. With time we can only expect the possibility of jigsaw identification to increase.

We should therefore consider the use of robust context controls when sharing genetic information. These controls will be similar to those discussed earlier for all other types of information and include Data Sharing Agreements, employer sanctions and / or training etc.

The identification of individuals or families from sequence 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 4. above).

Key messages

Identifiability is a continuum, but the law is binary. You need to know whether person-level information used in your research is considered identifiable in order to know how the law applies to this information.

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 would be identified, even by someone who may be more motivated than most, using all means that might reasonably be available to them. When this standard is met, the information is 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?

The appropriate context controls will differ if anonymisation is being undertaken to ensure data is no longer Personal (where ‘the viewer’ will be a corporate entity), compared to those required to ensure information is no longer confidential (where ‘the viewer’ will be an individual researcher).

Controls should manage the risk of identification (i.e. the likelihood and impact 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 etc.

The same anonymisation principles are relevant for genetic sequence information.

There are some circumstances where robust anonymisation cannot be achieved.