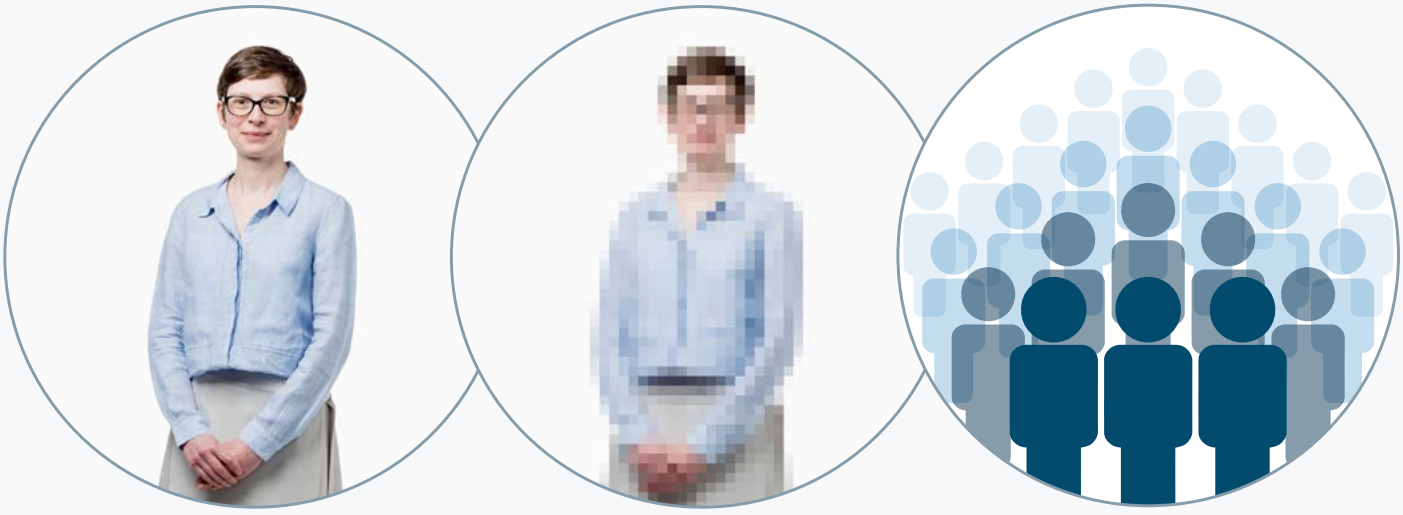
Identifiability Demystified

People want to know whether they could be identified when data about them is used.

The technical language of identifiability is complex. Many different words are used to describe the same thing, and many of those words are unnecessarily technical (for example pseudonymised, key-coded, de-identified for limited disclosure). It is important to explain clearly what it means when information is 'anonymised' and what the likelihood of re-identification is when using different types of data. We find using pictures is the most helpful way to explain the concepts.

Spectrum of identifiability



Anonymised*

Personally identifiable

De-personalised

Anonymous

More identifiable — **Less identifiable**

*anonymised in accordance with the ICO code of anonymisation

Spectrum of identifiability

At one end of the spectrum, a person is fully identifiable. As you remove or encrypt information, you blur the image more and more, and it becomes more difficult to identify who that person is. At the other end of the spectrum, it is not possible to identify who someone is — they are effectively anonymous.

Different controls are needed at different points along the spectrum depending on the risk of re-identification, which is why you may hear people talking about the 'environment' or 'context' in which data is used. The controls that are taken to protect the data are just as important as the data itself. It may also be possible to work out who someone is by joining together information from different sources — like joining together different pieces of a jigsaw puzzle.

What is anonymised data?

In the vast majority of cases, when data is used for purposes other than individual care, the data will have identifying details removed – it has been anonymised. The Information Commissioner's Office gives guidance about what details must be removed or masked, and the safeguards that must be followed to anonymise data effectively.

There are two different types of anonymised information ('de-personalised' and 'anonymous'). It's important to distinguish between them, because the risks of re-identification are different, and therefore the data has to be protected in different ways.

Why do we need a new word?

A lot of technical words have been used to describe different types of anonymised data. However, they have never been well explained and the concept of data that has been de-identified but is still at an individual level has been difficult to understand. As increasing use of data becomes an everyday part of life, it is really important that we have the words to help describe the concept properly. Testing suggests that using these images, and the word 'de-personalised', helps convey the meaning more effectively. The data has been through a process to remove personal identifiers – it is "de-personalised" - but it would still be possible to reverse that process and re-identify someone, so safeguards are still important.

It is just like a blurred photo of someone. We can't immediately see who the person is, but we know it is a specific person. If we had the right computer power, and really needed to know who the person was, it might be possible to work it out.

Personally identifiable

What is it?

This is information that identifies a specific person. Identifiers include: name, address, full postcode, date of birth or NHS number.

How is it protected?

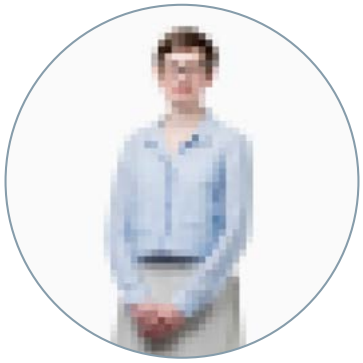
Personally identifiable information will always be stored in a highly secure way. There are strict laws that safeguard how personally identifiable information can be used if you are not asked for consent. There are also sanctions under the Data Protection Act if personally identifiable data is misused.

Example

A patient's medication history, including their NHS number (but no contact details).

Other words that you may see

Personal data, confidential information, patient identifiable information, confidential personal information.



De-personalised

What is it?

This is information that does not identify an individual, because identifiers have been removed or encrypted. However, the information is still about an individual person and so needs to be handled with care. It might, in theory, be possible to re-identify the individual if the data was not adequately protected, for example if it was combined with different sources of information.

How is it protected?

There are strict safeguards on how de-personalised information can be used, because there is the potential that it might be possible to re-identify someone. The higher the possibility of re-identification, the greater the level of control needed.

Example

A report that someone has suffered side-effects from a common medicine, including the patient's age and gender but with name, NHS number and date of birth removed.

Other words that you may see

De-identified, pseudonymised, key-coded, masked, anonymised in context, effectively anonymised, non-disclosive, non-identifiable, de-identified data for limited access.

Anonymous

What is it?

This is information from many people combined together, so that it would not be possible to identify an individual from the data. It may be presented as general trends or statistics. Information about small groups or people with rare conditions could potentially allow someone to be identified and so would not be considered anonymous.

How is it protected?

Because it would not be possible to identify someone, this information does not need special protection and can be published openly.

Example

The number of people who have been prescribed a certain medicine over ten years in five cities.

Other words that you may see

Aggregated data, grouped data, pooled data, statistics.