Identifiability demystified

Understanding Patient Data

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

The language of identifiability is complex. Many different words are used to describe the same thing, and many of those words are highly technical (for example pseudonymised, key-coded, de-identified).

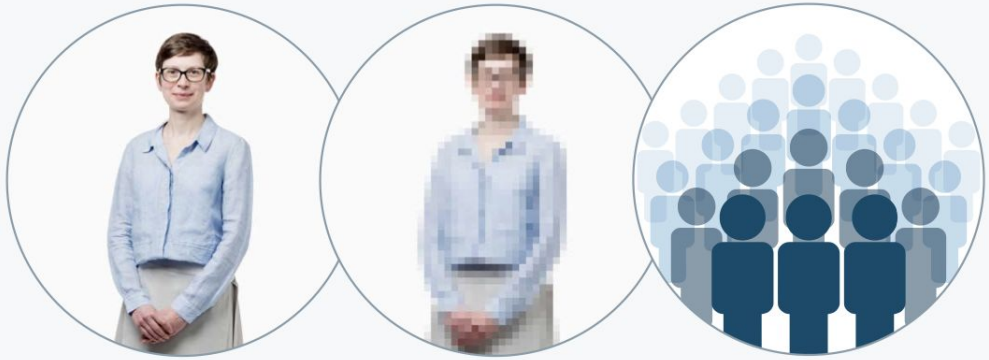
It is important to explain clearly what it means to say that information has been through a process of anonymisation and what the likelihood of re-identification is. We have found using imagery is helpful for explaining these concepts.

Last updated: October 2020

Spectrum of identifiability

Under GDPR, data is categorically either personal data or non-personal data. However, in practical terms there is a wide spectrum of identifiability. This ranges from fully identifiable personal data, to data that has been through a robust anonymisation process. The bar is very high for data to be considered ‘anonymous’ under GDPR, which means lots of purposes will use data that still counts as personal data.

The identifiability of data depends both on the features of the dataset **and** on the environment where it is held and used. Some environments to store data include technical controls on what the data can be linked to and limitations on who can access it. The controls to protect the data are just as important as the qualities of the data itself.



On the left hand side of the spectrum, the person is fully identifiable. As you move to the right along the spectrum, information is removed or encrypted, shown by the image blurring. It becomes more difficult to identify the person.

In the middle of the spectrum, the data has been de-personalised: personal information has been removed from it. The image is a blurred person because it may still be possible to identify the person if the data is combined with different sources - like adding more pixels or joining together different pieces of a puzzle.

On the right hand side of the spectrum, it is not technically possible to identify the individual. The data is aggregated or so heavily redacted that even if it is linked with other sources, you could not identify the person. This data would not be within scope of data protection law.

Don't say ‘anonymised’ data...

We don’t recommend using the term ‘anonymised’ to describe data that has had identifying information removed, as this implies a level of complete anonymity that cannot usually be guaranteed.

In most cases, data used for purposes other than individual care will have identifying information removed. However, to be useful for research and planning, the data still needs to contain enough detail for new insights to be generated. This means that while ‘direct identifiers’ (e.g. name, address, full date of birth) are removed, other information such as diagnoses, hospital admissions dates, ethnicity or postcode region may still be included in the dataset. Individual level data with these details could, in principle, be linked to other sources and used to re-identify an individual. If that’s the case, it is best to be open about it and explain the safeguards used to protect people’s identities.

Instead, say ‘de-personalised’ data

We recommend using ‘de-personalised’ to describe individual level data that has been through a process to remove personal identifiers, but where it would still be possible to reverse that process and re-identify someone. In combination with the blurry image of a person, we have found the word ‘de-personalised’ is easier to understand than common alternatives such as ‘de-identified’.

The imagery is helpful for explaining this idea: 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 access to other information, or if we were familiar with the individual, it might be possible to work out who they are.

‘De-personalised’ does not have a meaning in law, so should not be used in formal documents such as consent forms, contracts and privacy notices. It is intended as an intuitive way of conveying the idea, in language that is not legalistic or technical.

Personally identifiable

Information that identifies a specific person. Identifiers include: name, address, full postcode, date of birth, NHS number.

How is it protected? Data protection law safeguards how personally identifiable information can be collected, stored, managed and used. There are lawful bases to use this data that do not require consent. There are also sanctions if personally identifiable data is misused.

Example The medication, diagnosis and blood test result history of a person with a specified date of birth.

Other words Personal data, patient identifiable information.
Note that 'Confidential Patient Information' refers to information that includes both identifying information and something about the person's health or treatment.

De-personalised

Data that has had identifying information removed. However, the information is still about an individual person and needs to be handled with care. It might be possible to re-identify the individual if the data is not adequately protected or if it is combined with different sources.

How is it protected? Data protection law safeguards how this information can be collected, stored, managed and used. There are lawful bases to use this data that do not require consent. There are also sanctions if it is misused. The higher the possibility of re-identification, the greater the level of control and protection 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 De-identified, pseudonymised, key-coded, masked, anonymised in context, effectively anonymised, non-disclosive, non-identifiable, de-identified data for limited access.

Anonymous

Information from many people combined together, so that it would not be possible to identify an individual from the data. Or data that never included anything identifiable, for example an anonymously filled out opinion survey. Anonymous data 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? This information does not need special protection and can be published openly, because it would not be possible to identify someone.

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

Other words Aggregated data, grouped data, pooled data, statistics.