Anonymisation

Contents

- Is pseudonymised data still personal data?
- What are the benefits of pseudonymisation?
- How can pseudonymisation help us to reduce risk?
- Can pseudonymisation help us process data for other purposes?
- Are there any offences relating to pseudonymisation?
- How should we approach pseudonymisation?
- What pseudonymisation techniques should we use?
- How should we assess the risk of attackers reversing pseudonymisation?
- What organisational measures should we consider for pseudonymisation?

- What accountability and governance measures do we need?
  - What governance approach should we take?
  - Who should be responsible for our anonymisation process?
  - Why do we want to anonymise personal data?
  - How should we work with other organisations?
  - What type of disclosure is it?
  - How should we identify potentially difficult cases?
  - How should we ensure transparency?
  - How should we ensure appropriate staff training?
  - How should we mitigate identification risk due to a security incident?
  - What other legal considerations apply?

- Case studies
  - pseudonymising employee data for recruitment analytics
  - trusted third parties for market insights

- Glossary

About this guidance

Latest updates - 28 March 2025

> **28 March 2025** - this guidance was published.

At a glance

- Anonymisation is a privacy-friendly way to harness the potential of data.
- Anonymising personal data is possible in many circumstances. Whether you can effectively anonymise personal data depends on the techniques you use. You **should** reduce the risks of identifying people to a sufficiently remote level that the information is effectively anonymised.
- This guidance will help all organisations that want to anonymise personal data, for whatever purpose.
- It will help you identify the issues you **should** consider to use anonymisation techniques effectively.

In detail

- Why have you produced this guidance?
- What is this guidance about?
- Who is this guidance for?
- How is this guidance structured?

Why have you produced this guidance?

We understand the benefits that sharing personal data can bring to organisations, people and society as a whole. But there are risks too. Effective anonymisation techniques provide a privacy-friendly alternative to sharing personal data.

This guidance sits alongside our data sharing code of practice, which gives practical guidance on how to share personal data in line with data protection law. Anonymisation offers an alternative way to use or share data by making sure that people are not identifiable.

What is this guidance about?

This guidance will help you develop your understanding of anonymisation techniques, their strengths and weaknesses, and the suitability of their use in particular situations. It:

- explains what we mean by anonymisation and pseudonymisation;

- details how this affects your data protection obligations and responsibilities;
- discusses what you **should** consider when anonymising personal data;
- provides good practice advice anonymising personal data; and
- discusses technical and organisational measures to mitigate the risks to people when you do so.

This guidance deals with the role that anonymisation plays in the three regimes of data protection law:

- general processing under Part 2 of the Data Protection Act 2018 (DPA 2018) and the UK General Data Protection Regulation (UK GDPR);
- law enforcement processing under Part 3 DPA 2018; and
- intelligence services processing under Part 4 DPA 2018.

Where relevant, the guidance highlights and explains any differences between the regimes.

However, it is not intended as an exhaustive guide to data protection compliance and it is not prescriptive. It gives you the flexibility to implement anonymisation techniques in your own way, taking a proportionate and risk-based approach.

This guidance does not generally consider the impacts of anonymisation on areas of ICO work outside data protection. However, if you are also a public authority or a public body, you **should** follow this guidance when considering disclosing or allowing re-use of anonymous datasets under the Freedom of Information Act 2000 (FOIA), Environmental Information Regulations 2004 (EIR), or the Re-use of Public Sector Information Regulations 2015 (RPSI).

This guidance is not a statutory code. It contains advice on how to interpret relevant law on anonymisation and pseudonymisation. It also contains good practice recommendations.

There is no penalty if you don’t follow these recommendations. But you **must** find another way to comply with the law when you produce and disclose anonymous information. You may also find alternative methods that go beyond the good practice measures we set out.

However, when we look into an issue about anonymisation, we will take this guidance into consideration.

This guidance does not describe every possible anonymisation technique in detail but it includes case studies and good practice recommendations. It applies to all mediums, including tabular data, free text, video, images, and audio (including speech).

Further reading outside this guidance

Read the UK GDPR guidance and resources pages for more information for more information.

Who is this guidance for?

You **should** use this guidance if you are considering turning personal data into anonymous information. For example, this guidance is relevant if you:

- are required by law to publish anonymous information (eg some health service bodies);
- are looking to use data in new and innovative ways (eg to improve services or design new products or collect large volumes of data to train AI models);
- need to comply with a request for information under FOIA or EIR or a request for re-use under RPSI, and it includes personal data;
- want to become more transparent and accountable to people; or
- want to provide anonymous information for research purposes, or to enable wider societal benefits.

How is this guidance structured?

This guidance is divided into sections that cover different aspects of anonymisation in data protection law.

The first section introduces the key concepts of anonymisation and pseudonymisation, places them in the context of the UK legal framework, and explains the role they play.

The second section covers the concept of identifiability, including approaches such as the 'spectrum of identifiability' and how these can apply in data sharing scenarios. This section also looks at how you can manage identification risk, and covers established concepts like the 'reasonably likely' and 'motivated intruder' tests.

The third section looks at how pseudonymisation can help you achieve data protection compliance and which technologies can provide effective pseudonymisation.

The fourth section considers accountability and governance requirements in the context of anonymisation, including data protection by design, data protection impact assessments (DPIAs) and the use of trusted third parties.

The fifth and final section includes case studies providing practical examples of effective anonymisation.

In each section, we discuss what you **must** do to comply with data protection law, as well as what you **should** do as good practice.

If you are a reader with a general interest in understanding the concepts of anonymisation and pseudonymisation: read the first section.

If you are a technical expert and want to understand whether the technology you use results in anonymous or pseudonymous data: read the second and third sections.

If you are a decision maker looking for information about the types of techniques available for anonymisation and pseudonymisation: read the first and fourth sections.

Introduction to anonymisation

At a glance

- In data protection law, anonymous information is data that does not relate to an identified or identifiable person (ie data that is not personal data). Data protection law does not apply to anonymous information.
- To understand anonymisation, you **must** first understand what personal data is.
- Anonymisation is the process of turning personal data into anonymous information so that a person is no longer identifiable.

In detail

- What is personal data?
- What is anonymous information?
- What is anonymisation?
- Is anonymisation always necessary?
- Is anonymisation always possible?
- What are the benefits of anonymisation?
- If we anonymise personal data, does this count as processing?
- What is the difference between anonymisation and pseudonymisation?
- What about 'de-identified' personal data?

What is personal data?

Data protection law regulates the processing of personal data.

The UK GDPR defines personal data as:

> "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."

This definition applies for the purposes of Part 2 of the DPA 2018.

Section 3(2) of the DPA 2018 says that personal data is:

> "any information relating to an identified or identifiable living individual"

Section 3(3) defines an "identifiable living individual" as:

> "a living individual who can be identified, directly or indirectly, in particular by reference to—
>
> a. an identifier such as a name, an identification number, location data or an online identifier, or
> b. one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the individual."

Essentially, the same definition of personal data applies to the UK data protection framework as a whole.

As personal data has to be about living people, data protection law does not apply to information about the deceased. However, you **should** note that this data may still be protected by confidentiality or other laws or rules.

Relevant provisions of the DPA 2018 - see Section 3

External link

Relevant provisions in the UK GDPR – see Article 4(2)

External link

> **Further reading**
>
> Read our guidance on 'What is personal information?' in the Guide to the UK GDPR.

What is anonymous information?

Data protection law does not explicitly define 'anonymous information', but Recital 26 of the UK GDPR says this is:

> "…information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

This means that anonymous information does not relate to an identifiable person (either in isolation or when combined with information from other sources). Data protection law does not apply to anonymous information.

If you process personal data (as opposed to anonymised information), you **must** comply with the data protection principles and be able to demonstrate how you do so.

Other laws may still apply to anonymous information. For example, certain aspects of the Privacy and Electronic Communications Regulations 2003 (PECR) apply to 'information', not just personal data (such as the provisions about using storage and access technologies).

> **Further reading**
>
> See the Guide to PECR's sections on traffic data, location data, and storage and access technologies.

What is anonymisation?

Anonymisation is the way in which you turn personal data into anonymous information, so that it then falls outside the scope of data protection law. You can consider data to be effectively anonymised if people are not (or are no longer) identifiable.

We use the broad term 'anonymisation' to cover the techniques and approaches you can use to prevent identifying people that the data relates to.

We use the term 'effective anonymisation' to mean the technical and organisational measures you need to ensure that the status of the data meets the legal threshold for anonymisation under UK GDPR. Any information that does meet this threshold is not anonymous information, and you **must** treat it as personal data.

Anonymisation issues may be more complex if you have large datasets that contain a wide range of personal data. You may therefore need specialist expertise and input beyond this guidance.

> **Further reading – ICO guidance**
>
> Read the section 'How do we ensure anonymisation is effective?' for further guidance on the factors you **should** consider when you assess identifiability.

Is anonymisation always necessary?

No. Data protection law provides a framework to enable the fair, lawful and transparent use of personal data. However, if you don't need to use personal data to achieve your objectives, then you **should** assess whether you can use anonymous information instead.

Is anonymisation always possible?

In some instances effective anonymisation may not be possible due to:

- the nature of the data;
- the purpose(s) you collect, use or retain it for; or
- the context of the processing.

Example

A health authority considers anonymising a dataset containing information derived from people's medical records.

Even if they remove all identifying information, such as names and addresses, there may still be enough details about the people to potentially re-identify them. This may include details such as:

- the approximate dates and location of their treatments;
- the type of treatments;
- the approximate ages of the people; or
- other distinguishing characteristics.

This means it may still be possible to link the information with other information sources or make it possible to deduce the identity of people.

What are the benefits of anonymisation?

Anonymisation limits the risks to people and can allow you to make information available to other organisations or to the public.

It makes it easier to use and share the information, as fewer legal restrictions apply.

Anonymising personal data can help you to:

- improve your risk reduction and management processes;
- adopt a data protection by design approach;
- protect people's identities;
- reduce reputational risks or reduce questions, complaints or disputes caused by inappropriate or insecure disclosure or publication of personal data;

- have alternatives to deletion. For example, once the retention period you've set for personal data has come to an end (if you intend to retain, rather than delete the data), you **must** inform people that you will anonymise their personal data following the end of the retention period);
- publish information; and
- effectively comply with other legal obligations such as public authorities responding to FOI or EIR requests, or requests for re-use, involving personal data.

Wider benefits of anonymisation include:

- developing greater public trust and confidence that organisations are using data for the public good, while protecting privacy;
- greater transparency as a result of organisations being able to make anonymous information available to the public;
- incentivising researchers by increasing the availability of information;
- economic and societal benefits deriving from the availability of otherwise non-disclosable information; and
- improved public authority accountability through increased availability of information about service outcomes and improvements.
- allows rich data resources to be made available whilst protecting people's privacy.

Effective anonymisation of personal data is possible and desirable.

> **Further reading – ICO guidance**
>
> Visit our data sharing information hub for more information about the data sharing code.

If we anonymise personal data, does this count as processing?

Yes. For the purposes of data protection law, applying anonymisation techniques to turn personal data into anonymous information counts as processing personal data. The end result (the anonymous information) is not subject to data protection law, but the procedure (anonymisation) is.

For example, when you create aggregate statistical information from personal data, that data is 'adapted' or 'altered'. The law defines activities like these as 'processing operations'.

This means that you **must** comply with data protection requirements for your anonymisation process. This includes ensuring you have a lawful basis for it and you clearly define your purpose(s) and provide people with information about it.

In general it is likely that applying anonymisation techniques to the personal data you hold will be fair and lawful. However, save to the extent that it would allow identification as a result of reverse engineering of those techniques, you **must** clearly define your purpose and detail the technical and organisational measures you intend to implement to achieve it.

> Relevant provisions of the DPA 2018 - see Section 3
> External link

> Relevant provisions in the UK GDPR – see Article 4(2)
> External link

What is the difference between anonymisation and pseudonymisation?

Data protection law also uses the term 'pseudonymisation'. This is not the same as anonymisation. It's important to understand the difference, what it means in data protection law, and how this use may differ from what the word means in other circumstances.

Pseudonymisation is a technique that replaces information that directly identifies people, or de-couples that information from the resulting dataset. For example, it may involve replacing names or other identifiers (which are easily attributed to people) with a reference number. This is similar to how the term 'de-identified' is used in other contexts. For example, removing or masking direct identifiers within a dataset.

This guidance uses the term 'pseudonymous data' to describe personal data that has undergone pseudonymisation in line with the legal definition. It is information about people who can't be identified from that information by itself, but they can be identified from additional information held separately.

Pseudonymous information is still personal data and the law applies to it. Pseudonymisation reduces the links between people and the personal data that relates to them, but does not remove them entirely. Anonymisation prevents there being a link between the information and the person concerned.

It is common to refer to datasets as 'anonymised' when in fact they still contain personal data, just in pseudonymised form. This poses a clear risk that you might fail to comply with data protection law. For example, you may mistakenly believe that the information is anonymous and the law doesn't apply.

Remember that if you can still identify people using additional information you hold separately, the data is not anonymised. It is pseudonymised and is still personal data. You **must** still comply with data protection law.

Ultimately, you **should** think of anonymisation as a way of reducing the amount of personal data you hold, and pseudonymisation as a way of reducing the risks associated with the personal data you hold.

Example: anonymisation and pseudonymisation compared

A retail company collects customer transaction data to analyse shopping patterns to help improve their marketing strategies. They want to share this data with external consultants while ensuring no customer can be directly or indirectly identified from the data.

The company removes all direct identifiers (eg customer names, email addresses, and phone numbers) from the dataset. The remaining data is aggregated, including information such as purchase frequency, product categories, and spending amounts, and noised using differential privacy.

The resulting dataset contains statistical information on customers' shopping patterns without any link to specific people. The company can safely share the **anonymised** dataset with external consultants for analysis.

The same retail company needs to store customer data securely within their own system, while maintaining the identity of customers for the purposes of tracking customers' purchase history for a loyalty card scheme.

To reduce the risk to those customers, they decide to implement **pseudonymisation** using tokenisation. Instead of using direct identifiers, they assign each customer a unique pseudonym. They store direct and indirect identifiers separately in a secure database.

The pseudonymised dataset includes information such as total spend, product preferences, and loyalty points. But this data cannot be re-identified without using the pseudonym to link to the separately held identity data. The company uses the pseudonymised data to understand shopping patterns (eg which products are popular and peak shopping times) without needing to directly identify people.

Relevant provisions in the UK GDPR - see Articles 4(1) and 4(5), and Recitals 26, 28, 29 and 75

External link

Relevant provisions of the DPA 2018 - see Section 3

External link

Further reading – ICO guidance

What is personal information? – "Identifiers and related factors"

See the section of this guidance on "How do we ensure anonymisation is effective?" for more information on identifiability.

The guidance on identifiers and related factors also discusses the considerations you **should** take into account when disclosing data to other organisations, including the status it may have once in their hands.

See the section of this guidance on pseudonymisation for more information, including guidance on how you **should** approach pseudonymisation.

What about 'de-identified' personal data?

While the term 'de-identified' is widely used, we do not encourage it as a synonym for anonymous information or pseudonymous data. This is because UK data protection law doesn't define the term, so using it can lead to confusion.

Also, its meaning may differ depending on the circumstances. For the purposes of data protection law, we use this term only in connection with Section 171 of the DPA 2018, which states that the re-identification of "de-identified personal data" is a criminal offence.

In this context, 'de-identified' personal data is pseudonymised data or data that was considered anonymised but can be re-identified considering all means that are reasonably likely to be used.

Relevant provisions of the DPA 2018 - see Section 171

External link

Relevant provisions in the UK GDPR - see Article 4(5)

External link

DPA 2018 explanatory notes

External link

While explanatory notes are not part of the law, they are intended to help understand the DPA 2018.

Further reading – ICO guidance

See "Are there any offences relating to pseudonymisation?"

How do we ensure anonymisation is effective?

At a glance

- Anonymisation ensures that the risk of identification is sufficiently remote to minimise the risks to people arising from the use of their information.
- Identifiability is a wide concept. A person can be identifiable from many factors that can distinguish them from someone else, not just a name.
- Identifiability exists on a spectrum. When assessing whether someone is identifiable, you **should** take account of the "means reasonably likely to be used to enable identification". There are likely to be many borderline cases where you **should** use careful judgement based on the specific circumstances of the case.
- You do not need to take into account any purely hypothetical or theoretical chance of identifiability, rather, what is reasonably likely relative to the circumstances.
- You **should** consider both the information itself, and who may get (or want to get) access to it.
- Before you release data to the world at large, you **should** consider using robust techniques to reduce the higher risk of unauthorised personal data disclosure compared to intentional and controlled data release to known recipients.
- You **should** also consider potential unauthorised access by people (eg hacking or the actions of rogue employee).
- Applying a "motivated intruder" test is a good starting point to consider identifiability risk.
- You **should** review your risk assessments and decision-making processes regularly.

In detail

- What should our anonymisation process achieve?
- What is identifiability?
- What are the key indicators of identifiability?
- What is the "spectrum of identifiability"?
- What does data protection law say about assessing identifiability risk?
- How should we approach this assessment?
- What factors should we include?
- Do we need to consider who else may be able to identify people from the data?
- Can we anonymise within our organisation?
- What is the "motivated intruder" test?
- How do we apply the motivated intruder test?
- When should we review our identifiability risk assessments?
- How do we decide when and how to release data?

- What approaches can we take to anonymisation?

What should our anonymisation process achieve?

Anonymisation is about reducing the likelihood of a person being identified or identifiable to a sufficiently remote level. What this looks like depends on a number of factors specific to the context.

It may seem fairly easy to say whether a piece of information relates to an identified person, as this may be clear from the information itself. For example, bank statements clearly identify individual account holders and contain information that relates to them.

It may be less clear whether someone is **identifiable**. But you **should** take into account the concept of identifiability in its broadest sense in your anonymisation processes. You **should not** focus only on removing obvious information that clearly identifies someone.

What is identifiability?

Identifiability is about whether you can distinguish one person from other people with a degree of certainty.

Although a name may be the most common way to identify a person, it is important to understand that the following:

- A person can be identifiable even if you do not know their name. If the information might affect a particular person, even if you don't know their name or 'real world' identity, then they are still identified or identifiable.
- Whether any potential identifier actually means someone is identifiable depends on the context.

Identifiers are pieces of information that can be closely connected to particular people. They can be:

- direct identifiers (eg someone's name); and
- indirect identifiers (eg a unique identifier you assign to them such as a number).

Data protection law provides a non-exhaustive list of common identifiers in its definition of personal data. For example, name, identification number, location data and online identifier. However, the definition also specifies other factors that can mean a person is identifiable.

> Relevant provisions of the DPA 2018 - see Section 3
> External link

> Relevant provisions in the UK GDPR – see Article 4(1) and Recital 30

External link

Further reading – ICO guidance

See our guidance on 'What is personal data' for more information about:

- identifiers and related factors;
- direct identification; and
- indirect identification.

What are the key indicators of identifiability?

You **should** use two key indicators for determining whether information is personal data or not. These are:

- singling out; and
- linkability.

Effective anonymisation techniques reduce the likelihood of these occurring.

What is singling out?

Singling out means that you can single out the same person across records, or isolate the records that relate to a person from a dataset.

The UK GDPR specifically references singling out as something you **should** address when you consider the concept of identifiability. You **should**:

- consider whether singling out is possible, either by you or by another party; and
- make this part of your assessment of the effectiveness of your anonymisation processes.

Even if you do not intend to take action about a person, the fact that they can be singled out may allow you, or others, to do so. This means the person is still identifiable.

For example, if an attacker can isolate the data or link that piece of data to other sources of information, they can use it to single out a particular person and it will not be anonymous.

The risk of singling out depends on contextual factors. For example, information about a person's year of birth may allow them to be singled out in the context of their family, but not in the context of a different group, such as their class at school. Similarly, someone's family name may be enough to distinguish them from others in the context of their workplace, but not in the context of the general population (eg Smith or Jones).

Data records with person-level data can still be anonymous, but you **should** assess the risk of linkability with additional information particularly carefully in these cases.

To determine the possibility of singling out, you **should** consider:

- the richness of the data and how potentially identifying different categories are; and
- whether you have sufficient technical and organisational measures in place to mitigate this risk.

Example

In the first table, the rows can be differentiated from each other but that does not provide any level of identifiability in itself.

ID	Birthday month
1	January
2	February
3	March
4	April
5	May
6	June

In the second table, four records can be isolated from the others as being distinct: records 1 and 4 (isolatable on the basis of age) and record 2 and 3 (isolatable by combining age with location).

ID	Age	Hospital area	Condition
1	20-29	Lambeth	COVID-19
2	50-59	Lambeth	COVID-19
3	50-59	Southwark	COVID-19
4	60-69	Lambeth	COVID-19
5	70-79	Southwark	COVID-19
6	70-79	Southwark	COVID-19

If we assume that this table is from the first week of COVID-19 admissions from a part of South London, it is possible that record 1 provides enough information to single out the person from the database and link to it other information (eg news sources) due to the relative rarity of COVID-19 in younger patients at the start of the pandemic.

Distinguishing one record from others in the table is not sufficient by itself to make the person the record relates to identifiable. In an identifiability

assessment, you **should** also consider what additional data sources are available to:

- narrow down the number people enough to discover someone's real-world identity; and
- take some action on a person specifically, even if you can't discover their real-world identity.

What is linkability?

Linkability is the concept of combining multiple records about the same person or group of people. These records may be in a single system or across different systems.

Linkability is sometimes known as the mosaic or jigsaw effect. This is where individual data sources may seem to not be enough to identify someone in isolation, but can lead to the identification of a person if combined.

Simply removing direct identifiers from a dataset is insufficient to ensure effective anonymisation. If it is possible to link someone to information in the dataset that relates to them, then the data is personal data. In some cases, there is a higher risk of anonymous information being combined with other data.

For example, if:

- anonymised data can be combined with publicly available information meaning someone becomes identifiable; or
- complex statistical methods may 'piece together' various pieces of information with the same result.

Data that may appear to be stripped of identifiers can still be personal data if it can be combined with other information and linked to a person. For example, data available from social media. Even if stripping identifiers is not sufficient to achieve anonymisation, you **must** do so to comply with the data minimisation principle (eg if such identifiers are not required).

Common techniques to mitigate linkability include masking and tokenisation of key variables (eg sex, age, occupation, place of residence, country of birth).

Linkability is also the reason that pseudonymous data is still personal data, as the data can be combined with additional information allowing the identification of people it relates to.

Example

In 2006, Netflix published the Netflix Prize dataset. This contained movie ratings of 500,000 subscribers of Netflix that they considered anonymised.

Researchers found that using the Internet Movie Database (IMDB) as the source of additional information, they were able to successfully identify the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

What about using inferences to learn something new about a person?

An inference refers to the potential to **infer, guess** or **predict** details about someone who can already be identified directly or indirectly. In other words, using information from various sources to deduce something new about a person.

Inferences may also be the result of analytical processes intended to find correlations between datasets, and to use these to categorise, profile, or make predictions about people.

An inference can therefore be something you create, as opposed to something that you collect or observe.

Whether an inference is personal data depends on whether it relates to an identified or identifiable person.

To determine the likelihood of learning new information about a person through inference, you **should** consider the possibility of deducing new information about an identifiable person from:

- incomplete datasets (eg, where some of the identifying information has been removed or generalised);
- pieces of information in the same dataset that are not obviously or directly linked; or
- other information that you either possess or may reasonably be expected to obtain (eg publicly available additional information, such as census data).

You **should** also consider whether the specific knowledge of others, such as doctors, family members, friends and colleagues, is sufficient additional information to allow inferences to be drawn and linked to an identifiable person.

Example

In the UK, property purchase prices are publicly available. A person purchases a house for £500,000. Data analysis may show a high correlation between the house purchase price and the homeowner's income. A third party, knowing the

purchase price of this house, may infer that the homeowner's income is around £100,000 per year. As this inferred income is linked to the homeowner, it becomes personal data, even if the actual income of the homeowner is different.

What is the “spectrum of identifiability”?

In one sense, data protection law presents a simple binary outcome – information in a particular person’s hands either meets the definition of personal data or it does not.

However in practice, identifiability can be highly context-specific. Different types of information have different levels of identifiability risk depending on the circumstances in which you, or another party, processes them.

Whether something is personal data or anonymous information in the hands of a given person is therefore an **outcome** of assessing identifiability risk, taking into account the relevant facts.

In practice, identifiability may be viewed as a spectrum that includes the binary outcomes at either end, with a blurred band in between:

- at one end, information relates to directly identified or identifiable people (and will always be **personal data**); and
- at the other end, it is impossible for the person processing the information to reliably relate information to an identified or identifiable person. This information is anonymous in that person’s hands.

For everything in between, identifiability depends on the specific circumstances and risks posed. Circumstances of the processing may change, for example:

- the state of the art for anonymisation techniques changes;
- technical controls or identification attacks change;
- the means reasonably likely to identify someone becomes feasible and cost-effective; or
- data that enables identification becomes available.

Information may ‘move’ along the spectrum of identifiability to the point that data protection law starts to apply to it, or stops applying to it. We provide one way of visualising the spectrum of identifiability below.

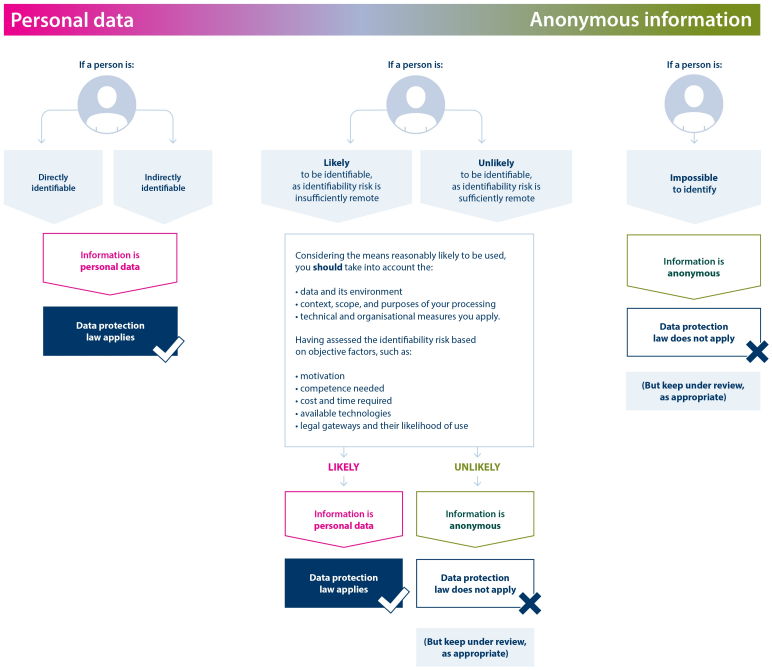


Figure 1: Mapping the concept of the spectrum of identifiability to data protection law

A text description of this diagram is available.

Information may shift towards one end of the spectrum, depending on factors including:

- the specifics of the data. For example, the sensitivity of the variables in the original dataset and the techniques you use to reduce the identifiability of people in the data;
- the context of the processing. For example, who you share the dataset with;
- the availability of any other information in the public domain at a given point in time that may allow for identification of people;
- the data environments involved. For example, the technical and organisational measures in place to control access to the data; and
- your risk management process. For example, how you identify and mitigate any risks of the processing.

Other resources

Some examples of visualising the concept of the spectrum of identifiability include:

- Understanding Patient Data's 'Identifiability Demystified' briefing (external link, PDF);
- the Future of Privacy Forum's 'Visual guide to practical data deidentification' (external link, PDF);

- the National Institute of Standards and Technology (NIST) publication 'De-Identification of Personal Information' (NISTIR 8053) (external link, PDF); and
- Privacy Analytics' presentation 'Principles of de-identification' (external link, PDF).

We are using these examples to illustrate different approaches and are not giving them an ICO endorsement.

What does data protection law say about assessing identifiability risk?

When you assess identifiability risk, the first question you **should** ask is whether the information is personal data in your hands. For example, the data may obviously relate to living people, or you may determine that you have means that are 'reasonably likely' to be used to identify people.

If the information is not personal data in your hands, then you **should** consider whether there are means that are 'reasonably likely' to be used by other people. For example, anyone who might obtain access to the information. We refer to this as the 'reasonably likely' test. The UK GDPR provides additional information about the factors you **should** take into account when determining this. Similar considerations apply to Parts 3 and 4 of the DPA 2018.

Recital 26 of the UK GDPR states that:

> "To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."

Therefore, once you take all objective factors into account:

- if there are means 'reasonably likely' to be used by you or by anyone who might gain access to the information to identify someone, then the information is personal data; and
- if no means are 'reasonably likely' to be used by you or anyone else, then the information is anonymised.

The more feasible and cost-effective a method becomes, the more you **should** consider it as a means that is reasonably likely to be used.

> Relevant provisions in the UK GDPR – see Articles 4(1) and Recital 26
> External link

How should we approach this assessment?

You **must** be able to demonstrate that disclosing or sharing apparently anonymous information will not lead to an inappropriate disclosure of personal data. Therefore, you **must** assess the risk of the information being identified.

You **should** consider the means that is reasonably likely to be used at the earliest stage of your anonymisation process, even if you don't intend to identify people. The assessment will be linked to your "release model" (ie public release, release to defined groups, or if anonymising data sets for further internal use only). In all release cases, you **should** consider the following in your assessment of identifiability:

- whether the information is easily identifiable with readily available means;
- whether there are techniques that enable identification from the information by anyone obtaining access to it;
- whether there is additional information that may enable identification; and
- the extent to which the additional information or techniques are reasonably likely to be used by a particular person to identify people the original information relates to.

Additional information or techniques may be available to different parties, depending on the circumstances. This means that the status of the information may change. For example, the same information may be:

- personal data in your hands. For example, if you also hold additional information that means people are identifiable (even if you hold this separately and apply technical and organisational controls to it); but
- anonymous in the hands of other parties. For example, a specific recipient (or the general public), if they have no access to the additional information and no means reasonably likely to be used to obtain it or identify the people by other means.

Data protection law does not require you to adopt an approach that takes account of every hypothetical or theoretical chance of identifiability. It is not always possible to reduce identifiability risk to a level of zero, and data protection law does not require you to do so.

The key is what is 'reasonably likely' relative to the circumstances, not what may be independently 'conceivably likely'.

Other resources

ISO/IEC 27559:2022 Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework provides a framework for identifying and mitigating identification risks.

Anonymeter is a statistical framework to quantify different types of identifiability risks in synthetic tabular datasets. Anonymeter uses attack-based approaches to evaluate the risk of singling out and linkability.

What factors should we include?

You **should** consider whether identification is technically and legally possible if you intentionally control disclosure to known recipients (rather than unauthorised access). You **should** take into account objective criteria including:

- how costly identification is in human and economic terms;
- the time required for identification;
- the state of technological development at the time of processing (ie the techniques you use to anonymise the data, or when you share the dataset with another party, or both); and
- technological developments (ie as technology changes over time).

You also **should** frame this assessment in the context of the specific risks that different types of data release present. These change depending on if you are disclosing information to:

- other parts of your organisation for further internal use only;
- another organisation;
- a pre-defined group of organisations; or
- the wider public or the world at large.

When you disclose, or otherwise make available information, to another organisation, both you and they **should** assess identifiability risk. You **should** clearly establish the status the information has in each of your hands.

The greater the likelihood that someone may attempt to identify a person from within a dataset, the more care you **should** take to ensure anonymisation is effective.

You also **should** take account of how this risk may change as information moves from one environment to another, depending on what is shared and the controls put in place. Factors that affect the risk of identification include:

- the existence of additional data (eg other databases, personal knowledge, publicly available sources);
- who is involved in the processing and how they interact;
- potential unauthorised access by people (eg hacking or the actions of rogue employee;
- the governance processes that are in place to control how the information is managed (eg who has access to it and for what purposes); and
- the legal considerations that may apply, such as:
    - any legal channels which allow additional information to be requested; or
    - legal or similar prohibitions that mean while it is possible for information to be technically combined to aid identifiability, doing so is not permitted (eg professional confidentiality and controlled disclosures).

After taking the above into account, if you conclude that the likelihood of identifiability is sufficiently remote, then the information is effectively anonymised. You **must**:

- document and justify your decision; and
- keep this under review as technologies and the processing change over time.

If you conclude that the likelihood of identifiability is not sufficiently remote, then you are still processing personal data.

You **must** have appropriate organisational and technical measures in place to process personal data securely (eg pseudonymisation and encryption).

> **Further reading - ICO guidance**
>
> Our guide to data security provides further information on you can comply with the security principle.

Do we need to consider who else may be able to identify people from the data?

Yes. You **should** consider whether it is reasonably likely that someone else, or someone you are deliberately sharing the information with, can identify people. For example, either from that information, or from that and other information they may possess or obtain.

This can sometimes be known as the 'whose hands?' question. This is about the status of the information in the different 'hands' of those who process it.

You **should** note that the 'whose hands' approach only applies when disclosing information to an organisation who is not acting with you as a joint controller or as your processor.

For example, if the information is personal data in your hands, it will also be personal data in the hands of the other parties, regardless of their technical or contractual ability to identify the people it relates to.

Similarly, a processor only processes personal data on your behalf. This means the status of the data in your hands is what matters.

In these cases, you **should** use pseudonymisation techniques. This does not affect the status of the data. But it does help you demonstrate the measures you have in place for the security and minimisation principles.

You **must** take account in your anonymisation processes of the nature, scope, context and purposes of the processing, as well as the risks it poses. These are likely to differ from one organisation to another and from one context to another. Although there may be circumstances where these considerations are similar, you cannot apply one single formula that will guarantee effective anonymisation in all instances.

If you receive data from another controller that they claim is anonymous information, but you are able to re-identify people from the dataset, you must then treat the data as personal data. You must also inform the organisation who provided the data that it is identifiable.

Example: Disclosure between organisations

BankCorp, a large bank, processes extensive customer data, including surveys and transactions, as a controller. FinResearch, an independent UK-based economics research institute, also acts as a controller, using data from various financial institutions for research.

BankCorp creates a dataset to share with FinResearch, ensuring appropriate controls are in place to minimise the risk of identifying people. Both organisations assess the identifiability risk from their own perspective, based on objective criteria, considering how:

- BankCorp creates the dataset; and
- FinResearch will process it.

The assessment determines whether the identifiability risk is sufficiently remote, meaning the dataset is anonymised, or if the risk is not sufficiently remote, meaning it remains personal information. Both organisations document their findings.

These considerations are relevant to BankCorp's decision to disclose the data, despite limited control over FinResearch's data environment or processing circumstances. The key is BankCorp's thorough assessment of identifiability risk.

If the data disclosed is personal information, BankCorp and FinResearch **should** enter into a data sharing arrangement and follow our data sharing code of practice to ensure they comply with UK GDPR.

If you are disclosing information, you **should**:

- consider which other organisations are likely to access the information during the initial scoping and design stages;
- ensure that the organisations accessing the information assess the risk of identifiability in their own hands;
- ensure that any other organisation accessing the information has a legitimate reason to do so and only accesses the minimum data needed to achieve their purpose;
- review the technical and organisational controls in place to govern access to the information;
- monitor access to the information and periodically review the data being accessed; and
- consider what other information is available in the public domain to re-identify people as part of your identifiability assessment to determine if the data is anonymous to the world at large.

Can we anonymise within our organisation?

The UK GDPR is mainly concerned with disclosing personal data outside a controller's own boundaries. However, anonymisation can also be important for safely using or sharing data within organisations, particularly large ones with diverse functions. For example, a retailer might use anonymised data rather than customer purchase records for its stock planning .

If you plan to leverage personal information for further use, then you **could** either:

- anonymise the information for other purposes; or
- pseudonymise it for the purpose of general analysis.

If you choose to anonymise personal data, you **must** not retain any additional information within your organisation that would allow identification. For example, the original personal data or other datasets that may be used to re-identify someone. You **should** put robust technical and organisational measures in place to prevent unauthorised access.

If you need to retain the information that allows for identification, then you are still holding personal data. The dataset is pseudonymised, rather than anonymised.

What is the “motivated intruder” test?

Data protection law does not specify how you determine whether the anonymous information you release is likely to result in the identification of a person.

You **must** consider all practical steps and means that are reasonably likely to be used by someone motivated to identify people whose personal data was used to derive anonymous information.

This is known as the motivated intruder test. You **must** use this test to help you to assess the identifiability risk of (apparently) anonymous information.

Both the ICO and the First-tier Tribunal (General Regulatory Chamber), which deals with information rights appeals, use this test.

You **should** adopt a motivated intruder test as part of your risk assessment. You **should** also use the test as part of any review, both of your overall risk assessment and the techniques you use to achieve effective anonymisation.

Who is a motivated intruder?

A motivated intruder is someone who wishes to identify a person from the anonymous information that is derived from their personal information. The test assesses whether the motivated intruder is likely to be successful.

You **should** assume that a motivated intruder is someone that:

- is reasonably competent;
- has access to appropriate resources (eg the internet, libraries, public documents); and
- uses investigative techniques (eg making enquiries with people who may have additional knowledge about a person, or advertising for anyone with that knowledge to come forward).

The intruder is therefore someone who is motivated to access the personal data you hold in order to establish whether it relates to people and, if so, to identify them. Such intruders may intend to use the data in ways that may pose risks to your organisation and the rights and freedoms of people whose data you hold.

You **should** assess the means that are reasonably likely to be used by a determined person with a particular reason to want to identify people. Intruders may be investigative journalists, estranged partners, stalkers, industrial spies or researchers attempting to

demonstrate anonymisation weaknesses. You **should** consider whether these type of intruders may be reasonably likely to use specialist resources and expertise to achieve identification.

You **should** also consider:

- their relationship to the person the data relates to;
- their background knowledge;
- whether they are targeting a specific or random person or people in the dataset;
- whether they know (with a degree of certainty) information about that person is in the dataset; and
- the perceived value of the data from the perspective of the motivated intruder.

Depending on the perceived value of the data to them, a motivated intruder may well use specialist knowledge or equipment or resort to criminal acts to gain access to the data and to seek to identify the people it relates to. For example, when assessing financial data, confidential files and other types of high-value data, you **should** also consider intruders with stronger capabilities, tools and resources.

The intruder can be someone who is not intended to have access to the information, or someone who is permitted access to it.

In essence, your motivated intruder test **should** consider:

- the nature, type and volume of information you process;
- the likelihood of someone wanting to attempt to identify people, for whatever purpose;
- the range of capabilities an intruder may have;
- the information that they may already have (or can access); and
- the controls you deploy within your data environment to prevent this.

Obvious sources of information that a motivated intruder may use include:

- libraries;
- local council offices;
- church records;
- public records (eg General Register Office, the electoral roll, the Land Registry, the National Archives);
- genealogy websites;
- online services (eg social media, internet searches);
- local and national press archives;
- AI tools, such as generative AI chatbot tools; and
- releases of anonymous information by other organisations (eg public authorities).

You **should** limit access to data where possible and consider the safeguards that you can adopt to reduce the risk in this case. If the information is still identifiable without appropriate safeguards, then you must put security measures in place to reduce the risk to people.

What types of motivations are there?

Some types of information will be more attractive for a motivated intruder to reidentify than others. Obvious motivations may include:

- finding out personal data about someone else, for malicious reasons or financial gain;
- the possibility of causing mischief by embarrassing others, or to undermine the public support for release of data;
- revealing newsworthy information about public figures;
- political or activist purposes (eg as part of a campaign against a particular organisation or person);
- curiosity (eg a local person's desire to find out who has been involved in an incident shown on a crime map); or
- a demonstration attack in which a hacker or researcher is interested in showing that identification of people is possible.

You still **should** undertake a thorough assessment of identifiability risk to determine the potential impact on people even when your data is seemingly ordinary, innocuous or otherwise without value.

Example

A supermarket maintains a database containing shopping histories of their customers. One of the entries includes a series of transactions for a customer who frequently buys products related to a health condition. Although there may be no clear motivation for someone to identify this customer, the potential embarrassment or anxiety if their identity were revealed could be very high.

You **should** reflect these potential harms in the anonymisation techniques you employ to protect this data.

Who can carry out the motivated intruder test?

You **could** carry out a motivated intruder test yourself, via a third party, or a combination of both.

If you do the test yourself, you **should** involve staff with appropriate knowledge and understanding of both your anonymisation techniques and of any other data that could be used for identification.

To determine the likelihood of more complex datasets being matched with publicly available data (eg statistical data), you may require specialist knowledge to help you assess the risk of identification. You **could** consider an external organisation with experience and expertise of intruder testing or ethical hacking. This can be beneficial, as the external organisation may:

- bring a different perspective to the test (eg that of an independent attacker); and
- be aware of data resources, techniques and types of vulnerability that you may have overlooked or are not aware of.

Does the type of data release matter for the motivated intruder test?

Yes. The type of data release has an impact on the factors you have to consider to assess the identifiability risk and how to ensure anonymisation is robust and effective.

With public release, you **should** have a very robust approach to anonymisation. This is because when you release information publicly, you lose control of the data. For example, it is almost impossible for you to retract the data if it later becomes clear that it relates to people that are identifiable. You also do not have control over the actions and intentions of anyone who receives that information.

With release to defined groups, you **should** consider in your identifiability risk assessment what information and technical know-how is available to members of that group. Contractual arrangements and associated technical and organisation controls play a role in the overall assessment. Fewer challenges may arise than with public release.

This is particularly the case if you retain control over who can access the data and the conditions in which they can do so. Designing these access controls appropriately will help reduce identifiability risk and potentially allow you to include more detail, while continuing to ensure effective anonymisation. Data access environments can help you retain control over the information.

You still **should** consider the possibility that the data may be accessed by an intruder from outside the group, or that it may be shared inappropriately. You **should** address this with physical and technical security controls aimed at preventing this access. If there is a greater likelihood of accidental release or unauthorised access, you **must** demonstrate how you mitigate this risk in your identifiability risk assessment.

Further reading – ICO guidance

Using a Trusted third party (TTP) or Trusted Research Environment (TRE) is one way of working with other organisations in a trusted environment.

How do we apply the motivated intruder test?

When considering the motivated intruder test, you **should** consider the different types of information that may be available.

How does a relationship between a motivated intruder and a person affect the risk of identification?

There are situations where a motivated intruder can already have some background information. Prior knowledge depends on the relationship between the intruder and the person they wish to identify. You **should** consider the following factors:

- the likelihood of intruders having and using the knowledge to allow identification; and
- the likely consequences of this identification, if any.

Identifiability risk can arise where one person or group knows a great deal about another person. They may be able to determine that 'anonymised' data relates to a particular person, even if another member of the public would be unable to. For example:

- one family member may work out that an indicator on a crime map relates to an incident involving another family member; or
- an employee may work out that a particular absence statistic relates to a colleague who they know is on long-term sick leave.

Certain professionals with prior knowledge (eg doctors, financial advisors) are not likely to be motivated intruders. This can apply if it is clear that their profession imposes confidentiality and ethical conduct rules. You **should** consider any possible circumstances where people in these professions may be motivated to break these rules (eg for financial gain).

You **should not** make assumptions about information people have shared with others. For example, teenagers may not share certain medical information with parents or other family members.

How do we assess the risk of identification?

You may find it difficult to assess the likelihood of identifiability in large datasets or collections of information. In these cases, you **should** consider a more general assessment

of the risk of prior knowledge leading to identification.

For example, you **could** consider the risk of at least one or a few of the people recorded in the information being identified. You **should** then make a global decision about the likelihood of those who might want to re-identify people seeking out or coming across the relevant data.

The likely consequences can also be difficult to assess in practice. Another person's sensitivity may differ from yours. For example, the disclosure of the address of a person in a witness protection scheme or someone escaping an abusive relationship will have more impact than the disclosure of the address of an average person.

To help you with this assessment, you **could** consult representatives of the people whose data you are anonymising, such as trade unions, patient groups or other advocacy groups.

What is the difference between information, established fact and knowledge?

It is also useful to distinguish between recorded information, established fact, and knowledge, when assessing whether a motivated intruder can identify a person from anonymous information.

Example

- "Mr B. Stevens lives at 46 Sandwich Avenue, Stevenham."

This may be established fact (eg because the information is recorded in an up-to-date copy of the electoral register).

- "I know Mr B. Stevens is currently in hospital, because my neighbour, Mr Stevens' wife, told me so."

This is personal knowledge, because it is something that Mr Stevens' neighbour knows.

You **should** consider recorded information and established fact first. It is easier to establish that particular information is available than to work out whether someone has the knowledge necessary to allow them to identify someone.

Personal knowledge combined with anonymous information can lead to identification. You **must** have a reasonable basis, rather than just a hypothetical possibility, before you to consider there is a significant risk of identifiability.

What about educated guesses?

You **must** have a degree of certainty that information is about one person and not another.

The mere possibility of making an educated guess about a person's identity does not present a data protection risk. Even if a guess based on anonymous information is correct, this does not mean that a disclosure of personal data has happened.

Example

A bike-sharing service releases anonymised data about the total distance travelled by each bike over a certain period. The bike-sharing service adds random noise to the total distance travelled by each bike before releasing the data. This means that the released distances are close to the true totals, but not exactly the same.

The data does not include any information about the specific times the bikes were used or the identities of the people who used the bikes.

Someone knows that a particular person frequently uses the bike-sharing service and tends to travel long distances. They make an educated guess that this person contributed to the high total distance travelled by a certain bike.

However, given that the dataset includes many bikes with similar total distances travelled, it would be nearly impossible to determine with a high degree of certainty that any specific entry in the dataset corresponds to this person.

Therefore, even though they could make an educated guess, it could also be easily disproved by the fact that the data is not unique to that person.

When should we review our identifiability risk assessments?

You **should** periodically review the decisions you take when anonymising personal data and the assessments that underpin them. The timing and frequency of your review depends on the specifics of the information you anonymise, as well as the circumstances both of its disclosure and its use afterwards.

Example

A researcher submits yearly FOI requests to a public authority. The researcher asks for statistics about the active participants in a programme the authority runs, including by number and characteristics such as age, sex, religious belief and nationality, broken down by quarter.

For three consecutive years the authority provides responses, because it is satisfied that the information is anonymous. Each year it re-assesses the nature of the information to make sure this is still the case.

On the fourth year, the authority withholds the information. Their reassessment led to the conclusion that releasing the information would be likely to lead to people being re-identified. This was because:

- the number of participants in the programme had decreased compared to previous years;
- a high-profile event involving the participants increased the likelihood that there were people motivated to re-identify them; and
- there was more information available in the public domain as a result of media reports and inquiries about the programme.

You **should** make sure that, as technology changes, you update your original assessment to reflect the impact that change may have on your decision-making.

You **should** also monitor changes in what data is publicly available as it may mean it becomes easier to re-identify data that you previously considered anonymised. You **should** consider the duration of the processing and how long the original data will be kept in identifiable form when determining how often to review your identifiability assessments. To do this effectively, you **should**:

- review highly sensitive data and regularly updated data at more regular intervals than static or infrequently updated data;
- regularly monitor technological developments (eg security best practices) and the effectiveness of anonymisation techniques which may affect the effectiveness of the anonymisation; and
- monitor current and new public data sources (eg reviewing the information available on the Internet or electoral roll).

You **should** re-assess identification risk under the following circumstances:

- if new attacks or vulnerabilities affect the technical and organisational measures you are using;
- if new datasets are released which may increase the risk of linkability or inferring new personal information related to a person;
- before you grant access to new recipients; or
- if the purposes for the anonymised data changes (eg when it is combined with other datasets).

Other resources

Several existing frameworks are available which can help you systematically consider the motivated intruder test, for example:

- ONS Guidance on intruder testing
- The Anonymisation Decision Making Framework (section 7.4 Penetration tests)

How do we decide when and how to release data?

The considerations in this section of the guidance will help you ensure your assessment of identifiability risk is appropriate for the type of disclosure you undertake.

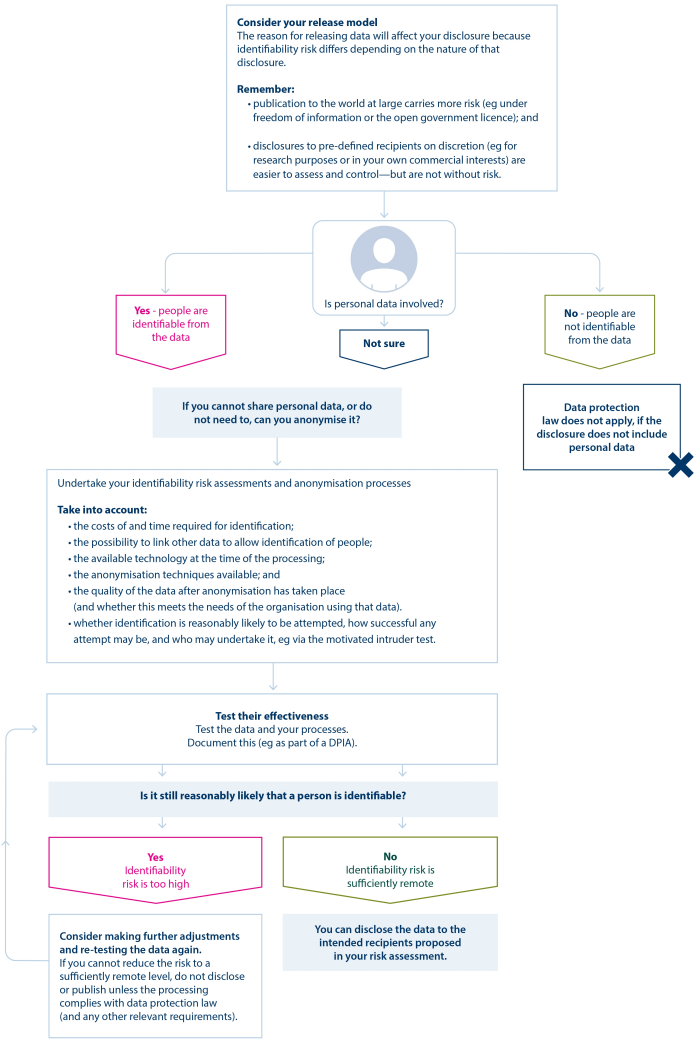
In summary, you **should**:

- consider your release model;
- conduct an initial assessment about whether the information includes personal data;
- establish whether you can anonymise that data;
- test the effectiveness of your anonymisation techniques (eg by assessing whether people are still identifiable);
- make further adjustments, as appropriate; and
- document the above, including your decision about the disclosure.

Figure 2 below represents a way that you **should** implement this process.

Further reading – ICO guidance

Read the section on 'What accountability and governance measures are needed for anonymisation?' for further guidance on accountability and governance measures.



A text description for figure 2 is available.

What approaches can we take to anonymisation?

There are two main approaches to anonymisation techniques:

- Generalisation, which reduces the specificity of the data. This changes information that may identify someone so that it relates to multiple people. This means members of that group can't be identified or are no longer identifiable.
- Randomisation, which can be used to reduce the certainty that a record relates to a particular person. This changes information that may identify someone so that it cannot be definitively attributed to one person.

Masking can also reduce identifiability by deleting or supressing certain values or data records. While masking can be effective when used alongside generalisation and suppression, it is not considered an anonymisation technique on its own.

What is generalisation?

Generalisation reduces the granularity or precision of the data by grouping or rounding it. Generalisation reduces the risk of people being singled out, as more of them are likely to share the same values. For example, a dataset containing the age of people can be changed so that only age bands are recorded (eg 20-30; 30-40).

What types of generalisation are there?

Generalisation can be achieved by reducing detail in the data's variables, either across the whole dataset or within some parts of it.

Generalisation is a way to achieve K-anonymity, a method used to measure the identifiability of a dataset by determining how many people share the same set of attributes.

K-anonymity is a property of a dataset that ensures a person's data cannot be singled out from other peoples' data within the same dataset. It involves setting a threshold (k), and the goal is to prevent identification of people by grouping each record with at least (k-1) other records that share the same values for their attributes. The higher the value of (k), the stronger the privacy guarantees.

K-anonymisation is suitable for anonymising relatively static datasets with common groups of attributes. It is simple to implement without extensive technical resources or expertise, using suppression and generalisation methods.

K-anonymisation does have some limitations, for example, it can be:

- susceptible to certain types of inference attacks (eg homogeneity and background knowledge attacks). If an attacker knows that a person is within a certain group, they can discover other attributes they did not know to link to that group (such as a health condition); and
- weak for data with large numbers of variables containing personal data.

Example: K-anonymisation

The NHS's standard for publishing health and social care data requires that, for strong anonymisation, k is set to five.

This means that for every record in the data set that describes the characteristics of a person, there are at least four other people also represented by records in the data set who share the same characteristics.

What is randomisation?

Randomisation can:

- remove the link between a person and the data, without losing the value in the data; or
- reduce the risk of data matching between data sets, unless other available data sets use the same randomised values.

Randomisation can be achieved by adding or removing values to certain data categories or altering the individual records but maintaining the overall statistical properties of the dataset.

Randomisation lowers the risk of an attacker determining something about someone, without changing the granularity of the data.

Synthetic data is a type of randomisation. It involves creating new data that keeps the statistical properties of the original data through the use of algorithms. Synthetic datasets aim to keep these properties without including the original data points.

What types of randomisation are there?

Noise addition involves adding random values either to a specific set of records or variables, or to the whole dataset. Noise addition can preserve the statistical properties of a dataset.

Noise addition is less effective if there are large differences between values, or there are some outliers. You may need to consult an expert to determine how much noise to add. For example, insufficient noise will mean the data is not anonymous, while large amounts of noise may make the data unusable.

You **could** add noise to prevent linking the data to a specific person by:

- adding noise to each record in a way that keeps the mean of the distribution unchanged;
- ensuring the amount of noise varies for each record; and
- combining noise with other techniques such as the removal of direct and indirect identifiers, if required.

Example

A fitness centre tracks members' weights to tailor fitness programs and monitor progress. The centre decides to release anonymised statistics of members' weight after they have completed six months of a fitness program. In order to

anonymise the data, noise, drawn from a random distribution, is added to each weight. For example, a person's weight is disclosed within a 5kg range of the original values. Small increases or decreases are made to the weight of each person, within the specified range.

Noise is applied in proportion with the scale of the original values, so that this process does not produce results that are highly skewed in comparison with the actual results. Adding or subtracting between 1kg and 5kg from the original weights may reduce the risk of identification to a sufficiently remote level. Varying the data by smaller amounts may in some cases risk singling people out.

Differential privacy is a method for measuring how much information is revealed about a person. While it is not an anonymisation technique as such, it allows you to determine how much noise to add to achieve a particular privacy guarantee (a formal mathematical guarantee about people's indistinguishability).

Differential privacy alters the data in a dataset so that values are harder to reveal, such as direct or indirect identifiers of people. If you add an appropriate level of noise, then you can use it as a way to anonymise personal information for other purposes (eg generating high-level insights).

Permutation involves swapping or shuffling records in the data by switching values of variables across pairs of records. This approach aims to introduce uncertainty about whether records correspond to real data elements and increase the difficulty of identifying people by linking together different information about them.

Permutation allows you to retain the precise distribution of a variable in the anonymised database. Swapping can easily be reversed if the swapped variables are linked to each other.

Permutation may be unsuitable in cases where:

- the correlation between variables is important for the purpose you use the information for. Therefore, swapping is not a suitable method;
- you need to maintain correlations between that variable and other variables; and
- the swapped variables are linked to each other, as swapping can easily be reversed.

Example

A fitness centre tracks members' weights to tailor fitness programs and monitor progress. The centre decides to release anonymised statistics of members' weight after they have completed six months of a fitness program.

The weight values for different members are moved around, so that they no longer relate to other information about that person.

This is helpful if they need to retain the precise distribution of weight values in the anonymised database, but they do not need to maintain correlations between weight values and other information about people.

What is masking?

Masking involves identifying and removing direct identifiers that may single someone out. Direct identifiers can relate to a single person in all datasets (eg an email address or credit card number) or only be unique for some datasets.

The degree of masking applied to the data will depend on the use case. Deleting certain values or data records can help to prevent attacks that can link records back to people or to determine specific attributes of a record with a degree of certainty.

Masking alone is not considered an effective anonymisation technique. However, it can play a role when combined with generalisation and randomisation techniques.

Example: combining masking, generalisation and randomisation

A retail company intends to release anonymised data about its customers' shopping patterns for market research purposes. The original dataset includes identifiable information including: customer ID, age, gender, postcode, and total spend per month.

Customer code	Age	Gender	Postcode	Total spend per month (£)
A1B2C3	24	M	SE7 5PX	105
D4E5F6	36	F	ME21 9UU	210
G7H8I9	42	M	B78 9JE	315
...	...	...	...	...

To anonymise this data, the company uses a combination of masking, generalisation, and randomisation techniques.

The company uses masking to completely remove the customer IDs and the last part of the postcode for each record and gender from the dataset. This helps prevent anyone from linking the data back to specific customers using knowledge

of their specific location, ID or gender. But there is still a risk because the data contains specific ages and exact total spends for each person. To mitigate this risk the company also implements generalisation and randomisation to reduce the risk of identification to a sufficiently remote level.

The company replaces the exact ages with age ranges (eg 18-24, 25-34) and the postcodes with broader geographic areas with a minimum size population. This makes it harder to identify people based on their age or location.

The company adds a small amount of random noise to the total spend per month. This preserves the overall distribution and trends in the data, but makes it impossible to know the exact spend of any individual customer.

Age Range	Area	Total Spend per Month (£) (noised)
18-24	Stockport	112
25-34	Wolverhampton	204
35-44	Canterbury	305
...	...	...

In this case, even if someone knows a particular person who shops at this retail company, they would not be able to determine with a sufficient degree of certainty that any specific entry in the dataset corresponds to this person.

What should we consider when choosing anonymisation techniques?

The anonymisation approach you choose will depend on the nature of the data and your purposes. In some cases, you can combine these methods to provide a more robust method. Using these methods does not guarantee you will achieve anonymisation: successful anonymisation depends on contextual factors.

What about anonymisation of qualitative data?

Much of the anonymised data you create, use and disclose is derived from administrative datasets that are essentially statistical in nature. However, the techniques you use to anonymise quantitative data are not generally applicable when seeking to anonymise qualitative data or unstructured data (eg the minutes of meetings, interview transcripts or video footage). You need different techniques to do this.

You **could** consider methods such as:

- removing direct and indirect identifiers from documents (eg names, email addresses and other information that relates to people);
- applying blurring or masking to video footage to disguise faces and other information that may identify someone;
- electronically disguising or re-recording audio material; and
- changing the details in a report (eg precise place names, precise dates)

Anonymising qualitative material can be time-consuming. It does not lend itself to bulk processing and can require careful human judgement to determine if people are still identifiable, based on the data and other information which may be linked to it.

Example: anonymisation of free text

A research group wants to release anonymised transcripts from interviews conducted for a study. The original dataset includes the names of interviewees, their responses, and the researcher's questions and comments.

To anonymise this data, the research group performs:

- masking (suppression): all names and other directly identifiable information are removed from the transcripts (eg companies, birth dates, addresses);
- generalisation: specific potentially identifiable details in the responses are replaced with more general terms (eg 'a place' instead of the actual place name); and
- synthetic data generation: machine learning techniques learn the structure and topic distributions of the original transcripts, and generate synthetic text based on it.

Other resources

For detailed recommendations on appropriate technical solutions for different types of data release, read Statistical disclosure control by Hundepool, Anco, et al. (2012).

For more information on k-anonymisation:

- the Anonymisation Decision-Making Framework, which provides further guidance on the use of k-anonymisation techniques.
- the NHS anonymisation standard for publishing health and social care data (ISB1523) NHS anonymisation standard for publishing health and social care data (ISB1523) ()

For more information on differential privacy and synthetic data, read our guidance on PETs.

For more information on anonymising qualitative data, including images and text, read Anonymising qualitative data by the UK Data service.

NIST SP 800-226 Guidelines for Evaluating Differential Privacy Guarantees describes techniques for achieving differential privacy and their properties, and covers important related concerns for deployments of differential privacy.

Pseudonymisation

At a glance

- Pseudonymisation refers to techniques that replace, remove or transform information that identifies people, and keep that information separate
- Pseudonymised personal data is in scope of data protection law.
- Pseudonymisation has many benefits. It can help you to reduce the risks your processing poses:
  - implement data protection by design;
  - ensure appropriate security; and
  - make better use of personal data (eg for research purposes and general analysis).
- Take care not to confuse pseudonymisation with anonymisation. Pseudonymisation is a way of reducing risk and improving security. It is not a way of transforming personal data to the extent the law no longer applies.
- The DPA 2018 contains two criminal offences that address the potential harms that result from unauthorised removal of pseudonymisation.
- There are many pseudonymisation techniques. Some will help you achieve pseudonymisation as defined by the law. Others may not, but can still be useful technical measures from a security perspective.

In detail

- What is pseudonymisation?
- Is pseudonymised data still personal data?
- What are the benefits of pseudonymisation?
- How can pseudonymisation help us to reduce risk?
- Can pseudonymisation help us process data for other purposes?
- Are there any offences relating to pseudonymisation?
- How should we approach pseudonymisation?
- What pseudonymisation techniques should we use?
- How should we assess the risk of attackers reversing pseudonymisation?
- What organisational measures should we consider for pseudonymisation?

What is pseudonymisation?

Pseudonymisation has a specific meaning in data protection law. This may differ from how it is used in other circumstances, industries or sectors.

Article 4(5) of the UK GDPR defines pseudonymisation as:

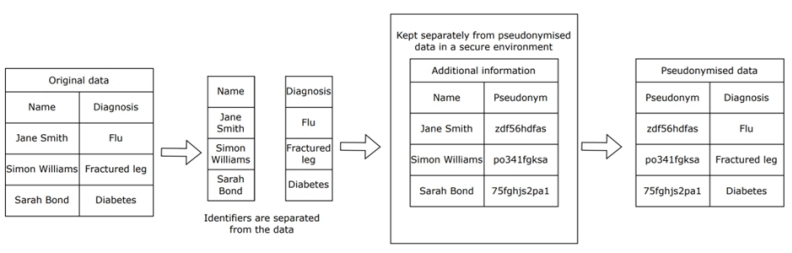
> "…processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

There is no equivalent definition in the law enforcement or intelligence services regimes in the DPA 2018, but similar considerations apply.

Pseudonymisation starts with a single input (the original personal data) and ends with two outputs (the pseudonymised dataset and the additional information). Together, these two outputs can reconstruct the original personal data. However, for the people concerned, each output only has meaning when combined with the other.

Example

The diagram below shows a simple example of pseudonymisation, where the identifiers are removed and replaced with a pseudonym, which is stored separately.



Pseudonymisation therefore refers to techniques that replace, remove or transform information that identifies a person and store it separately. For example, replacing one or more identifiers which are easily attributed to people (such as names) with a pseudonym (such as a reference number).

If the data is pseudonymised, you can tie that pseudonym back to the person if you have access to the additional information. However, you **must**:

- hold this information separately; and
- keep it secure.

Data protection law specifically mentions "unauthorised reversal of pseudonymisation" as something that can result in harm. You **must** assess the likelihood and severity of this risk and mitigate it appropriately.

Relevant provisions in the UK GDPR – see Article 4(5) and Recitals 26, 28, 29, 75 and 85

External link

Further reading – ICO guidance

What is personal information: a guide

Is pseudonymised data still personal data?

Yes. Pseudonymised data is personal data in the hands of someone who holds the additional information.

However, it does not change the status of the data as personal data when you process it in this way.

This is because data protection law is clear that information is personal data if a person is identified or identifiable, directly **or indirectly**.

The core definition of pseudonymisation describes it as processing of personal data in a particular manner. Additionally, Recital 26 of the UK GDPR says that:

> "…personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person…"

If you share pseudonymised data (but not the additional information) with another organisation, it may be anonymous information in their hands.

Further reading – ICO guidance

The section Do we need to consider who else may be able to identify people from the data? provides further information on assessing the identifiability of pseudonymised data without additional information with another organisation.

What are the benefits of pseudonymisation?

Pseudonymisation can reduce the risks to people. It can also help you meet your data protection obligations, including data protection by design and security.

When you apply pseudonymisation properly, it can help to:

- reduce the risk your processing poses to people's rights;
- enhance the security of the personal data you process;
- support re-use of personal data for new purposes;
- support your overall compliance with the data protection principles;
- provide suitable safeguards to mitigate the risk of an international transfer of personal data;
- build people's trust and confidence in how you process their data; and
- demonstrate you have put appropriate safeguards in place, as you are required to do when using personal data for research purposes.

Pseudonymisation can enable greater utility of data than anonymisation. However, you **should** still consider whether you can meet your objectives by using anonymous information.

How can pseudonymisation help us to reduce risk?

Recital 28 of the UK GDPR says that:

> "The application of pseudonymisation to personal data can reduce the risks to the data subjects concerned."

Data protection law doesn't include a specific definition of risk. But it does make it clear that this is about the risks to people's rights and freedoms. For example, Recital 75 links these risks to the potential for harm or damage:

> "The risk to the rights and freedoms of natural persons, of varying likelihood and severity, may result from data processing which could lead to physical, material or non-material damage"

So, pseudonymisation is relevant for your assessment of these risks. For example, in data protection impact assessments (DPIAs) or legitimate interests assessments (LIAs), you **could** detail specific pseudonymisation techniques you use and show how they mitigate the particular risks your processing poses.

Pseudonymisation is also relevant as a risk reduction measure in other areas, such as:

- data protection by design and security; and
- personal data breaches.

Security and data protection by design

You **must** put in place appropriate technical and organisational measures to:

- implement the data protection principles effectively and integrate necessary safeguards into the processing. This is "data protection by design"; and
- ensure a level of security appropriate to the risk the processing poses. This is "the security principle".

The law specifically references pseudonymisation in both of these requirements.

If you apply pseudonymisation properly, it can be a useful mechanism to enhance the security of personal data and support your overall compliance with the data protection principles.

Pseudonymisation is particularly relevant in the context of the data minimisation principle.

This is because it can limit the level of identifiability in the data to what is necessary for the purpose.

When you consider pseudonymisation techniques, you **must** take into account the:

- state of the art and costs of implementation of any measures;
- nature, scope, context and purpose(s) of your processing; and
- risks your processing poses to people's rights and freedoms.

When you assess the state of the art, you **should** consider whether the technique is suitably robust, for example, is it:

- resistant to known attacks;
- scalable; and
- not cost-prohibitive to implement?

The nature, scope and purposes of the processing influence the type of pseudonymisation technique that is appropriate. You **should** consider whether your chosen approach allows you to fulfil your purposes and also whether the measures you choose can reduce the risks to people to an acceptable level.

Not all pseudonymisation techniques are equally effective, and they may not have the same implementation costs or requirements. You **should** choose your pseudonymisation technique by finding the optimal approach to achieving data protection by design and security.

Personal data breaches

Pseudonymisation techniques can reduce the risk of harm to people that may arise from personal data breaches. This can also form part of your assessment of the likelihood and severity of any impact of a personal data breach.

Pseudonymisation may be relevant when you assess whether you should notify people of the personal data breach. Under Article 34 of the UK GDPR, you must notify people about a data breach without undue delay, if the data breach results in a high risk to their rights and freedoms, unless you have:

> "…implemented appropriate technical and organisational protection measures, and those measures were applied to the personal data affected by the data protection breach, in particular those that render the personal data unintelligible to any person who is not authorised to access it, such as encryption."

Although this does not specify pseudonymisation, it broadly describes technical and organisational measures. These technical measures can include pseudonymisation. It can provide another line of defence in the security context, reducing the level of risk for those affected.

If a personal data breach takes place, you **must** assess whether the pseudonymisation measures you have in place prevent the breach causing a high risk for the people concerned. You still **must** inform the ICO in the event of a data breach. Depending on the circumstances of the breach (eg there is still risks to people), we may direct you to notify the affected people.

Can pseudonymisation help us process data for other purposes?

Data protection law may allow you to repurpose personal data for some types of processing, if appropriate safeguards such as pseudonymisation are in place. For example, for research, further analysis or compatible purposes. This means that pseudonymisation can be a useful tool to enable further processing of personal data beyond its original purpose.

This does not mean pseudonymisation automatically allows you to undertake this further processing in all cases. However, it can be an important way for you to demonstrate how you protect personal data, if you do so. It is therefore a factor whenever you are considering whether you can process personal data for other purposes.

If your purposes change over time or you want to use data for a new purpose which you did not originally anticipate, you **must** only go ahead if:

- the new purpose is compatible with the original purpose;

- you get a person's specific consent for the new purpose; or
- you can point to a clear legal provision requiring or allowing the new processing in the public interest.

Pseudonymisation is relevant when you are considering compatibility instead of consent or legal provisions. For example, if you:

- undertake further processing for archiving, scientific or historical research, and statistical purposes, which are automatically considered to be compatible purposes; or
- want to undertake further processing for other purposes, and need to assess whether these are compatible with your initial purpose.

Pseudonymisation also allows you to perform general analysis. You **could** do this activity as part of the further processing.

Other compatible purposes

Article 6(4) of the UK GDPR says that when deciding if a new purpose is compatible with your original purpose, you **must** take into account:

- any link between your original purpose and the new purpose;
- the context in which you originally collected the personal data. In particular, your relationship with the person and what they would reasonably expect;
- the nature of the personal data (eg whether it is special category data or data relating to criminal convictions or offences);
- the possible consequences for people of the new processing; and
- whether there are appropriate safeguards (eg encryption or pseudonymisation).

Pseudonymisation does not necessarily mean that you can decide your new purpose is compatible in all cases. It is one of several factors you **must** consider in this assessment. If your new purpose is compatible, you don't need a new lawful basis for the further processing.

However, you need to remember that if you originally collected the data on the basis of consent, you should get fresh consent to ensure your new processing is fair and lawful or you can rely on another basis such as legitimate interests or public task depending on your circumstances. You also **must** update your privacy information so that your processing is still transparent.

General analysis

Recital 29 of the UK GDPR says that:

> "In order to create incentives to apply pseudonymisation when processing personal data, measures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organisational measures necessary to ensure, for the processing concerned, that this Regulation is implemented, and that additional information for attributing the personal data to a specific data subject is kept separately. The controller processing the personal data should indicate the authorised persons within the same controller."

This means you **could** perform this general analysis on pseudonymised data within your organisation, however you **must**:

- implement the technical and organisational measures necessary to ensure data protection compliance; and
- ensure that you keep additional information for attributing the data to a specific person separately.

Data protection law does not define general analysis. However, if you intend to analyse data relating to specific people (eg their behaviour, location, characteristics) for the purposes of taking actions about them, this analysis is not general in nature.

In practice, general analysis may be something you undertake for the two purposes detailed above. It can bring many benefits, depending on the purposes you do it for. For example, pseudonymising data about how people use your products and services, and then deriving insights and trends from that data. This may allow you to develop new, innovative services or improve existing ones.

This is particularly the case if anonymous information is less useful. However, you **should** carefully consider whether you can achieve these objectives using such information first.

When you perform general analysis, you **should** indicate the authorised people within your organisation that have access to the additional information. You **should** also update this, if you make any personnel changes in the future.

Relevant provisions in the UK GDPR – see Articles 25, 32 and Recitals 28, 78 and 83

External link

Relevant provisions of the DPA 2018 - see Sections 57, 66, 103 and 107

External link

Further reading – ICO guidance

- DPIAs
- Legitimate interests
- Data protection by design and by default
- Security
- Personal data breaches
- Data minimisation
- International transfers | ICO

Are there any offences relating to pseudonymisation?

Yes. Section 171 of the DPA 2018 contains two criminal offences relating to re-identification. We call these the 're-identification offences'. They cover the identification of people from pseudonymised data and ineffectively anonymised data.

The first, at section 171(1), is about the act of re-identification. It states that:

> "It is an offence for a person knowingly or recklessly to re-identify information that is de-identified personal data without the consent of the controller responsible for de-identifying the personal data."

The second, at section 171(5), is about processing the personal data after the act of re-identification takes place. It states that:

> "It is an offence for a person knowingly or recklessly to process personal data that is information that has been re-identified where the person does so –
>
> a. without the consent of the controller responsible for de-identifying the personal data, and
> b. in circumstances in which the re-identification was an offence under subsection (1)."

The definition of 'de-identified' data is:

> "…personal data […] processed in such a manner that it can no longer be attributed, **without more**, to a specific data subject" (emphasis added).

This is not directly equivalent to the definition of pseudonymisation as there is no requirement for the additional identifying information to be retained and held separately and securely.

In other words, it may include any information where the 'direct' identifiers have been stripped out. But the people the information relates to are indirectly identifiable using additional information or other techniques. This is the case, even if there's no specific 'key' held to enable deliberate identification. For example, the use of decryption technologies or hacking techniques aimed at breaking the de-identification process.

This may be due to several factors, such as ineffective application of anonymisation techniques, advances in re-identification techniques or previously unknown re-identification attacks.

In summary, reversing pseudonymisation or ineffective anonymisation by attempting to attribute the data to a specific person is a crime. You **must not** try to match up pseudonymous personal data to the person it relates to, or attempt to re-identify people in information believed to have been anonymised, unless the controller who pseudonymised or anonymised the data agrees to you doing so.

Are there any defences?

Yes. For example, re-identification is allowed if you can prove that it was:

- necessary for the purposes of preventing or detecting crime;
- required or authorised by law or by a court order; or
- in the circumstances, justified as being in the public interest.

The defences also include where the person charged can prove that they acted:

- in the reasonable belief that they are the person the information relates to, have the consent of that person, or would have had such consent if the person had known about the reidentification and its circumstances;
- in the reasonable belief that they are the organisation responsible for the de-identification, have the consent of that organisation, or would have had such consent if the organisation had known about the re-identification and its circumstances; and
- for the "special purposes", with a view to publication of any journalistic, academic, artistic or literary material, and in the reasonable belief that the re-identification was justified as being in the public interest in the circumstances.

Re-identification as a security testing measure

The law allows for re-identification if this is to test the effectiveness of an anonymisation or pseudonymisation technique. For example:

- testing the effectiveness of your own security measures; or
- security or technology researchers who test those of others.

This is not an offence provided you meet the following two conditions:

- condition one: The person is testing the effectiveness of the de-identification systems an organisation uses, where they reasonably believe the testing is justified as being in the public interest and where they do not intend to cause or threaten damage or distress; and
- condition two: The person notifies either the ICO or the organisation responsible for the de-identification about the re-identification.

If you test the measures of other organisations and successfully re-identify people, you **must**:

- let those organisations know as soon as possible; and
- where feasible, let them know within 72 hours of becoming aware of it.

If there are multiple organisations responsible for the initial de-identification, you can meet condition two's requirements by notifying one or more of them.

You can only rely on the effectiveness testing provision if you are acting in the public interest. The defence does not legitimise unlawful or harmful practices. You are likely to be committing a criminal offence, if you reverse pseudonymisation, or information thought to be anonymised, and cause or threaten harm to people or organisations.

Relevant provisions of the DPA 2018 - see Sections 171 and 172
External link

Relevant provisions in the UK GDPR – see Articles 4(5) and Recitals 26 and 75
External link

DPA 2018 explanatory notes paragraphs 492 to 504 -
External link

> Although the explanatory notes are not part of the law, they explain what each provision of the DPA 2018 means in practice and provide background information on the intended outcomes of the provisions.

Further reading – ICO guidance

Exemptions – the "special purposes"

The DPA 2018's explanatory notes provide further information on how the re-identification offences apply to pseudonymised data (external link)

How should we approach pseudonymisation?

You are responsible for deciding whether to implement pseudonymisation and how to do so. You **should** clearly establish what you want to achieve and the most appropriate technique. An inadequate level of pseudonymisation does not meet the legal definition of pseudonymisation in data protection law, even if the technique you use may fit under existing technical meanings of the term.

You **should**:

- define the goals: what does your use of pseudonymisation intend to achieve?;
- detail the risks: what types of attack are possible, who may attempt them, and what measures do you need to implement as a result?;
- decide on the technique: which technique (or set of techniques) is most appropriate?;
- decide who does the pseudonymisation: you or a processor?; and
- document your decisions and risk assessments (eg in your DPIA, LIA or record of processing activities).

While this is not an exhaustive list of relevant considerations, you **should** address them together due to how they relate to each other.

Define your goals

Overarching goals of pseudonymisation can include:

- ensuring that parties other than yourself (and, where appropriate, any processor you may use) cannot identify people;
- ensuring that third parties cannot access or reconstruct the additional information;
- enabling data accuracy (eg by assigning a particular pseudonym to a person that allows you to verify their identity); and
- achieving data minimisation (eg if the purposes of your processing do not require you to identify people).

You **should** ensure that once you implement pseudonymisation, you mitigate any risk of unauthorised reversal of it. To do this, you **should** consider any potential source of risk (eg a malicious attacker or an insider threat).

You **should** also consider whether the pseudonymisation technique you use is useable and scalable, so that it is practical for the processing activity you want to carry out.

Detail the risks

When assessing what pseudonymisation techniques to use and how to implement them, you **should** take into account the type of attacker that may exist. For example, it is good practice to consider:

- insider threats – someone with specific knowledge, capabilities or permissions, either in your organisation, a processor you use, or another entity you engage (eg a trusted third party);
- external threats – someone who may not have direct access to the additional information, but wants to increase their knowledge about the pseudonymised dataset (eg by re-identifying the people within the dataset); and
- the likely goals of any attack – an attacker may want to achieve different goals (eg identification attacks, where the attacker seeks to re-identify people (either a subset, or all of them).

> **Further reading – ICO guidance**
>
> We discuss the methodology to assess the risk of singling out a person in the section How do we ensure anonymisation is effective?.

Decide on the technique

When deciding on a pseudonymisation technique, you **should** take into account:

- the nature, scope, context and purpose of the processing;
- the risk factors you identify; and
- the privacy protection, utility and scalability goals your processing requires.

You **should** explore the availability of existing solutions to meet your goals, together with their strengths and limitations, in your decision-making process. You **should** choose the appropriate technique after considering:

- the risk of identification for the part of the data that you will transform by the pseudonymisation technique;
- the security measures (both technical and organisational) you can put in place to protect the additional information that would allow the pseudonymised data to be re-identified; and
- the required utility and accuracy of the data for the purposes of the processing.

You **should** also have appropriate processes in place for regularly testing, assessing and evaluating the effectiveness of the pseudonymisation techniques you use.

Decide who performs the pseudonymisation

Different parties may be involved in any pseudonymisation process. There is no one-size-fits-all approach. It is ultimately a decision for you to take based on your specific circumstances.

Pseudonymisation may be performed by:

- you;
- a processor working on your behalf (eg if they have specialist expertise and resources to help you achieve your goals); or
- another party working with you as a joint controller.

You **should** identify roles and responsibilities of any party involved in the pseudonymisation process.

You **could** also consider separating the functions. For example, clearly specifying and distinguishing between the people that:

- carry out the pseudonymisation processes;
- are authorised to access the additional information; and
- undertake any subsequent processing of the pseudonymised data (eg if you are performing general analysis for certain purposes).

Further reading – ICO guidance

- Controllers and processors
- Contracts and liabilities

Document the outcome

It is important that you clearly document your decision-making processes and detail the steps you take.

This will help you meet several other requirements, if they apply to your circumstances.

For example, you **must** include:

- a general description of your technical and organisational security measures in your record of processing activities, where possible; and
- details about the measures you intend to take to mitigate any risks you identify in your DPIAs.

You **could** also include these considerations:

- in your security risk assessments; and
- in any LIA you undertake.

Finally, you **must** also monitor the state of the art and ensure that the methods you use continue to be appropriate as techniques evolve.

Further reading – ICO guidance

- Accountability framework
- Guide to the UK GDPR – Security outcomes
- Guide to the UK GDPR – accountability and governance
- Our guidance on privacy-enhancing technologies (PETs) provides guidance on the use of state-of-the-art solutions that can be used for pseudonymisation, including Secure Multiparty Computation (SMPC) and homomorphic encryption.

What pseudonymisation techniques should we use?

The most common types of pseudonymisation techniques are:

- hashing;
- encryption; and
- tokenisation

Hashing-based pseudonymisation

Cryptographic hash functions transform input data of any size into fixed-length outputs. These outputs are known as 'hash values' or 'message digests'.

Hashing is useful for checking the integrity of files. By comparing the hash of the current file with the hash generated when the file was first created, you can quickly determine if it has been modified. If the hashes match, the file remains unchanged; if they differ, it signals that the file has been potentially tampered with or corrupted.

Hashing is intended to be one-way. For example, if someone obtains a list of hash values, they should not be able to work out what the original input was, even if they know what hash function was used to create the values. Hashing can remove explicit links between people and the data. A mapping table is used to link between the input identifiers and the output hash. You **should** ensure you use a hash function that is appropriately robust in the circumstances. Using a robust hash function ensures an attacker cannot:

- calculate the original input;

- make educated guesses about the input (eg by knowing a list of hashed names); and
- find two inputs mapping to the same output.

If you are considering using a hashing function to apply appropriate technical and organisational measures to personal data, you **should** avoid using:

- hashing algorithms that do not use additional data to generate a pseudonym (eg a salt, pepper or encryption key); or
- outdated hashing algorithms (eg MD5 and SHA-1).

These algorithms are vulnerable to brute force identification attacks.

What hashing approach and algorithm should we consider?

The section below provides a summary of the strengths and weaknesses for some of the most common hashing techniques you **could** consider, as well as examples of suitable applications .

In order to mitigate identification attacks, random data (known as a salt or pepper) can be added to plaintext data before the hash function is applied. Unlike salts, peppers are not shared and are stored separately from hashes in a secure environment. For each different salt or pepper used, the resulting hashes also differ. If you require consistent hashes, then you are required to share the salt as well as the hash value, therefore there is a risk of brute-force attacks.

If you use salted hashes for linking the same person's records between databases, you **should** ensure that appropriate technical and organisational measures are in place to protect the salt.

You **could** use a hashing algorithm such as bcrypt. This provides robust pseudonymisation by intentionally slowing down hash computation to deter brute-force attacks. It does this by significantly prolonging the time required to guess the plaintext (known as the work factor). As technology evolves, the work factor in bcrypt can be increased to maintain the same level of security. Bcrypt also automatically includes a unique salt during hashing to prevent identical inputs from producing the same hash.

Encryption-based pseudonymisation

Pseudonymisation can be performed with symmetric and asymmetric encryption. In symmetric encryption the same key is used for encryption and decryption.

In asymmetric encryption, one key is used for encryption and a different key is used for decryption. One of the keys is typically known as the private key and the other is known as the public key.

The private key is kept secret by the owner and the public key is either shared amongst authorised recipients or made available to the public at large.

This typically means that any party can encrypt data but only the owner of the private key can decrypt the data. Robust encryption relies on the security of the encryption key. You **should** choose an appropriate encryption algorithm, secret key length and security controls.

You **could** use symmetric encryption to generate consistent (also known as deterministic) or randomised pseudonyms for identifiers across different databases, depending on the encryption implementation.

You **could** use asymmetric or symmetric encryption to generate random pseudonyms for each use of the same identifier by using a probabilistic asymmetric scheme which adds randomisation into the process.

Depending on the circumstances, you **could** use other forms of encryption. For example, using format-preserving encryption to encrypt identifiers (eg an email address) as pseudonyms, while preserving the format of the data.

> **Other resources**
>
> "Data pseudonymisation: advanced techniques and use cases" (2021) by The European Union Agency for Cybersecurity (ENISA) provides more details about other advanced encryption-based pseudonymisation techniques. (external link)

Tokenisation-based pseudonymisation

Tokenisation replaces identifiers with randomly generated tokens. Tokens can be generated by hashing or by generating random numbers that are stored in an indexed sequence.

Tokenisation is an efficient technique, and therefore it can be suitable for large-scale processing. As there is no mathematical relationship between a token and an original identifier, knowledge of a token does not allow an attacker to re-identify a person.

You **could** use tokenisation to link people across databases, providing you use the same token for the same person in each database.

> **Example**
>
> A bank uses tokenisation to link customer data across its various services, such as current accounts, savings accounts, and mortgage services. When a customer

opens a current account, the bank generates a unique, random token to replace a customer's direct and indirect identifiers in the bank's database.

The mapping between the customer's identifiers and the random token is stored in a separate, database managed by the bank's IT department with appropriate technical and organisational measures to prevent unauthorised access.

When a customer later opens a savings account or applies for a mortgage, the bank uses same token to link their financial records across these services. This allows them to link customer data across databases without exposing direct or indirect identifiers.

Further reading – ICO guidance

See our guidance on passwords in online services in the Guide to the UK GDPR for more information on appropriate hash functions in that context.

See our guidance on encryption in the Guide to the UK GDPR for more information on appropriate encryption algorithms and the required technical and organisational measures to implement them.

How should we assess the risk of attackers reversing pseudonymisation?

When choosing and implementing a pseudonymisation technique, you **must** consider the risk of the method being reversed to identify the people the pseudonymised data relates to. The likely types of attacks include:

- brute force attacks (also known as exhaustive searches);
- dictionary searches, which involve an attacker computing a set of possible pseudonyms at scale, saving the result, and using that information when attempting to reverse the pseudonymisation; and
- 'guesswork', based on the fact that some characteristics are more frequent than others.

The effectiveness of these attacks will depend on factors such as:

- the pseudonymisation technique you use;
- how you configure and implement that technique;
- the background knowledge of the attacker;
- the category of personal data the pseudonymous data relates to;
- the protections you put in place for the additional information; and

- the availability of other relevant information the attacker has access to.

Some attackers will focus on discovering the additional information, such as the pseudonymisation key or mapping table. If successful, this type of attack has the greatest impact as the attacker can re-identify all the pseudonymised data, completely reversing the pseudonymisation process.

You **should** assess the risk of an brute force, dictionary and guesswork attacks on the pseudonymised data by considering the following factors:

- Size of the identifier domain and the dataset (smaller is more vulnerable).
- Type of pseudonymisation function used and the likelihood of an attacker being able to compute the pseudonymisation function.
- The amount of additional information available, for example, the number of pseudonymisation secrets (eg keys, salts).
- Whether the pseudonym contains some information derived from the original identifier

You **could** implement the following measures to mitigate any risks of these attacks:

- Use appropriate key sizes, or long salts or peppers, if using hashing.
- Consider fully randomised pseudonymisation techniques.
- Use a technique that has no mathematical relation to the initial identifiers (eg tokenisation).
- Test whether it is computationally feasible to guess the pseudonymisation secret.

You **should** also consider the factors that may result in identification of a single person in pseudonymised data, for example:

- Ability to use 'additional information' without access to the pseudonymisation key.
- Combining common indirect identifiers used to match a person.
- The presence of outlier values.

You **could** implement the following measures to mitigate the risk of a person being identified:

- Consider if you can remove any attributes in the data that may reveal a single person without access to the pseudonymisation key (eg outlier values or rare characteristics).
- Apply appropriate technical and organisational measures to protect the additional information, such as encryption and access controls.

What organisational measures should we consider for pseudonymisation?

The UK GDPR requires that when you implement pseudonymisation, you **must** keep any additional information separated from the pseudonymised data using appropriate technical and organisational measures.

You **must** choose a technical solution for pseudonymisation that compliments the organisational measures you use. You **must** ensure that the technical measures are carried out effectively and appropriately. The additional information that can be used to identify people from a pseudonymised dataset is a source of risk, so you **must** put in place measures to protect it.

For example, you **must**:

- securely destroy the additional information if you don't need it;
- store it securely (eg by encrypting it and using robust key management methods); and
- delete it from any insecure media, such as memory storage and systems.

What measures should you use to keep the additional information separate?

In order to ensure you keep the additional information separately, you **should**:

- store the additional information and pseudonymised data in distinct physical locations (eg using separate databases or network segmentation to prevent linkage);
- enforce strict access control policies to restrict physical and logical access to the additional information to only authorised personnel;
- encrypt the additional information so that even if unauthorised access occurs, the data remains unintelligible without access to the decryption key;
- implement a robust logging system to track access requests to the additional information, with regular reviews of these logs;
- handle and store the keys appropriately. This includes secure storage and rotation practices; and
- securely back up the additional information, so you can recover it if you need to. You **should** regularly test your backup and recovery process.

Other resources

A number of publications from the European Union Agency for Cybersecurity (ENISA) provide more details about pseudonymisation techniques, including additional risks that you may need to consider.

These include:

- "Recommendations on shaping technology according to GDPR provisions" – an overview on data pseudonymisation" (2019) (external link)
- "Pseudonymisation techniques and best practices" (2019) (external link)
- "Data pseudonymisation: advanced techniques and use cases" (2021) (external link)

What accountability and governance measures do we need?

At a glance

- When producing and disclosing anonymous information, you **should** take a comprehensive approach to governance.
- Being clear about processes, responsibilities and oversight makes compliance easier.
- You **should** use a DPIA to help you structure and document your decision-making processes around anonymisation and identify risks to rights and freedoms and mitigation strategies
- You **must** be clear about how and why you intend to anonymise
- You **should** work with other organisations who are likely to be processing, and possibly disclosing, other information that may impact the effectiveness of your anonymisation
- You **should** consider how different forms of anonymous information can pose different identifiability risks and choose an appropriate release model to mitigate them
- You **should** plan for cases where it may be difficult to assess identifiability risk and implement appropriate risk mitigation measures
- Demonstrating transparency when processing anonymous information promotes people's trust and mitigates the risk of any negative public opinion of the processing
- You **should** ensure decision-makers clearly understand the latest technological and legal developments and best practices to ensure effective anonymisation
- You **should** think about any other legal considerations that may be relevant to your anonymisation processes and decision-making

In detail

- What governance approach should we take?
- Who should be responsible for our anonymisation process?
- Why do we want to anonymise personal data?
- How should we work with other organisations?
- What type of disclosure is it?
- How should we identify potentially difficult cases?
- How should we ensure transparency?
- How should we ensure appropriate staff training?
- How should we mitigate identification risk due to a security incident?
- What other legal considerations apply?

What governance approach should we take?

If you anonymise personal data, the accountability principle of the UK GDPR requires that you **must** address the practical issues surrounding the production and any disclosure of this information in your governance approach.

Establishing an appropriate governance structure can improve your data management, record-keeping and disclosures of data. In addition, it is useful if you need to demonstrate compliance to the ICO.

We are less likely to carry out enforcement action, including monetary penalties, if you can demonstrate that you:

- made a serious effort to comply with data protection law; and
- had a genuine reason to believe that the information was not personal data (ie by showing that identifiability risk was sufficiently remote).

You **should** cover the following areas in your governance structure:

- **How will you plan for anonymisation?**
    - Who is responsible for your anonymisation process?
- **How will you identify and mitigate anonymisation risks?**
    - Have you completed your data protection impact assessment (DPIA)?
    - Why do you intend to anonymise personal data?
    - How will you work with other organisations, where necessary?
    - Will you use a trusted third party (TTP)?
    - What are the relevant considerations for the type of disclosure, including limited access safeguards?
    - How will you identify and manage potentially difficult cases?
    - How you will ensure transparency?
- **How will you ensure anonymisation remains effective?**
    - How will you keep updated with relevant changes to the legal framework (including guidance and case law) and technological developments?
    - How you will ensure appropriate staff training?
    - How will you approach identification testing?
- **How will you consider other relevant legislation?**
    - Do any other legal considerations apply?

You **must** document your key decisions and the rationale for them as part of your accountability obligations.

Who should be responsible for our anonymisation process?

You **should** make sure that someone of sufficient seniority oversees your anonymisation process and decision-making. This may be a single person or a group of authorised people, depending on your circumstances.

They **could** work closely with your DPO to seek their advice and guidance (if you are required to have one). They **should** have an appropriate understanding of:

- the circumstances of both your process and any intended disclosure; and
- relevant technical and legal considerations.

Data protection law does not specify who this person may be or what their formal role is. The important point is that they have appropriate authority.

For some organisations, it can be particularly useful to adopt a Senior Information Risk Owner (SIRO) approach. In this context, the SIRO:

- takes responsibility for key decisions and informs your general approach to anonymisation;
- consults with your DPO to obtain their independent expert advice;
- coordinates a corporate approach to anonymisation, drawing on relevant expertise from within and outside your organisation; and
- helps you decide on suitable forms of disclosure (ie publication or limited access).

Why do we want to anonymise personal data?

The act of anonymising personal data qualifies as processing of that data. For example, adaptation or alteration. This means you **must** be clear about how and why you're doing it.

So when you anonymise, you **must** define your purpose for doing so and the technical and organisational measures to achieve it. As a key consideration, you **should** clarify the context and purposes for anonymising.

This is because anonymisation may be:

- an aspect of your overall processing activities; or
- the overall purpose of your processing.

Whether this is the case depends on your circumstances, so you **must** be clear on when you intend to anonymise and why.

Anonymisation as part of your processing activities

If anonymisation is part of your overall processing activities, it can be a way to comply with the data protection principles. For example, to comply with the principles of data minimisation and storage limitation, you **must**:

- only collect the personal data you need for your purpose; and
- keep it in a form that only identifies people for the time you need to achieve that purpose.

Once you achieve your purpose, you **must** either erase or anonymise the personal data, depending on your circumstances.

In these situations, anonymisation may simply be something that you do as part of the processing and as a way of complying with the law. As long as your anonymisation is effective, data protection law does not apply if you subsequently use the anonymous information.

In many cases, processing personal data to anonymise it is likely to be compatible with the original purpose(s) you collected it for, unless:

- there is a reasonable expectation from a person that you will retain the data in identifiable form; or
- when you collected it, you told them you intended to keep it in that form.

Anonymisation as part of your purpose

Anonymisation may itself be a way for you to achieve the purpose you originally collect personal data for.

For example, if your purpose is to generate aggregate statistical information about how people engage with your service, you may need to collect information about what each one does first.

The information you collect in this case is likely to be personal data as it relates to actions and behaviours that specific people take. You **must** be clear with someone that this is why you want to collect their data.

How should we work with other organisations?

If you are planning to disclose any anonymous information, you **should** work with other organisations likely to be processing, and possibly disclosing, other information which might allow the identification of someone that the anonymous information relates to.

A joined-up approach with other organisations in your sector, or those doing similar work, allows you to assess the risks collectively and agree mitigations, where appropriate.

Example

Public authority A is planning to link education data with constituency data provided by public authority B.

They are doing this so that researchers can study the relationship between education and voting, with both using similar geographical units.

Both authorities can then assess the risks of identification jointly.

Further reading – ICO guidance

Our guidance on the research provisions provides further information on purpose limitation in research.

Using a Trusted third party (TTP) or Trusted Research Environment (TRE) is one way of working with other organisations in a trusted environment

What type of disclosure is it?

Different types of disclosure can pose different risks. Generally, there two main types of disclosure:

- **Open release** - where data is made available to anyone to access, use and share.
- **Limited access** - where data is made available only to a restricted group.

You **should** differentiate between these types of release when considering making anonymous information available.

Limited access may allow the disclosure of 'richer' data among a restricted group. For example, a closed community of researchers. The data can have a higher level of utility, and it is also possible to restrict the further disclosure or use of the data, or provide better guarantees about it security.

However, the success of limited access relies on robust governance measures governing the disclosure.

Limited access is particularly appropriate for handling anonymous information derived from sensitive source material (eg special category data). There can still be risks with limited access. For example, further disclosure outside the group or for purposes beyond what has been agreed.

You **should** mitigate these risks by ensuring that you disclose anonymous information in a closed community with clear, established rules (including around data minimisation).

What limited access safeguards should we consider?

When you disclose data to a restricted group (eg one or a small number of organisations), you **should** take steps to prevent further disclosure. For example:

- use contractual controls; and
- apply robust technical and organisational measures to support those controls.

Before you make the anonymous information available, you **should** put robust safeguards in place, including:

- purpose limitation – establishing that the recipient(s) can only use the anonymous information for an agreed purpose or set of purposes;
- training recipients' staff who will have access to the data (eg on security and data minimisation principles);
- security checks of those who will access the data;
- controls over the ability to bring other data into the environment to manage identifiability risks arising from linkage or association;
- limiting data use to a particular project or set of projects;
- restricting disclosure of the data outside the limited access environment;
- prohibiting attempts at re-identification;
- ensuring appropriate measures are in place to destroy any accidentally re-identified personal data;
- implementing appropriate technical and organisational security measures, including confidentiality agreements for those who will access the data (including your staff);
- restricting access to the data (eg by applying appropriate encryption techniques and access control policies);
- limiting the number of copies of the data to what is necessary for the purposes of the disclosure;
- arranging for the destruction or return of the data and confirmation that this has been done once the project is complete; and
- imposing appropriate penalties if any recipient breaches the conditions placed on them (eg as part of contractual requirements).

To decide which of these apply, you **should** conduct your own risk assessment. This **could** involve your normal data security risk assessment processes. You **should** also co-ordinate with the other parties involved in the project to establish whether you require additional security measures.

What about publication under licence?

Once you publish data under an open licence, it may be impossible to protect it from further use or to keep it secure.

Most open data licensing models are clear that those who use the information are not allowed to do so in a way that enables identification to take place.

For example, the Open Government Licence (OGL), Creative Commons or Open Data Commons.

However, in practice this may be difficult to enforce. So, you **should** ensure that your anonymisation processes and identifiability risk assessments are sufficiently robust to mitigate likely identification attacks.

> **Further reading**
>
> The UK Data Service provides further guidance on the terms of use for various public-sector licencing structures (external link).

How should we identify potentially difficult cases?

Anonymisation can be ineffective due to several factors, for example:

- you were not aware of other sources of data that could be matched to your dataset; or
- technological developments mean that the anonymisation techniques you applied are no longer effective (eg the emergence of new attacks or increased computational power).

You **should** consider whether alternative state-of-the-art techniques are available to ensure that the data is effectively anonymised and the risks of identification are mitigated by any technical and organisational measures.

You **should** also cater for other risks relating to the use of anonymous information in your governance approach, particularly for other purposes which may not be compatible with the original purpose. For example, you **should**:

- only use anonymous information in ways people would reasonably expect;
- consider whether people would reasonably expect you to retain the data in identifiable form; and
- assess whether turning personal data into anonymous information affects people. For example, if you are using the anonymous information to make decisions or decide how you treat people, and how you can justify any adverse impact.

The level of risk depends on the nature and context of the processing. For example, special category data, such as someone's health status or ethnicity, is particularly sensitive and carries additional risks. So, you **must** account for this when you anonymise it due to the

possible impact on people if they are re-identified.

You **should** perform a DPIA to consider and mitigate the risk of using anonymous information when you make decisions or take actions about an identifiable person which may lead to detrimental effects on them. For example:

- using anonymous information which may result in discrimination or financial loss to people; and
- using anonymous information with poor analytical value. For example, anonymous information related to demographic characteristics which introduce bias. In this case, you **should** consider whether it is possible to adjust the level of accuracy while ensuring it remains anonymous.

How should we ensure transparency?

People have the legal right to know how and why you are processing their data. You **must** explain your approach to anonymisation as clearly as possible in your privacy notice, including any consequences it may have. You **must** make the policy clear and easily accessible.

Demonstrating transparency about the generation and intended uses of anonymous information also promotes people's trust and mitigates the risk of any negative opinion of the processing.

In particular, you **should**:

- tell people why you anonymise personal data;
- describe how you do this, in accessible terms (taking care not to undermine the effectiveness of your anonymisation process);
- say what safeguards are in place to minimise the risk that may be associated with producing anonymous information. In particular, you **could** explain whether you intend to make the anonymous information publicly available or only disclose it to a limited number of recipients;
- be open with people about any risks of the anonymisation you are carrying out, your use of the anonymised information, and the possible consequences of this. You **could** give them the opportunity to submit queries or comments about this; and
- publicly describe your reasoning for publishing anonymous information and explain how you:
    - did the 'weighing-up';
    - what factors you took, or did not take, into account and why; and
    - how you looked at identification 'in the round'.

If you are a public authority, then you **must** include your FOIA, EIR and RPSI obligations in your privacy notice, as appropriate. You can publish anonymised information in response to a request (or if you are making it proactively available as part of your publication scheme).

You **could** also consider publishing any DPIAs or relevant reports about your anonymisation. You **could** remove certain information if needed, or publish a summary.

You **should** also review the consequences of your anonymisation programme, particularly through analysing any feedback. You **should** make this an ongoing activity. For example, technological developments may impact the effectiveness of your techniques and the outcome of any assessment of identifiability risk over a period of time.

You **should** be able to analyse and deal with any complaints or queries people make to you.

> **Further reading – ICO guidance**
>
> Our guidance on the right to be informed provides further information on the information you **should** include in a privacy notice.

How should we ensure appropriate staff training?

Members of your staff involved in decisions about creating and disclosing anonymous information **should** have a clear understanding of:

- the legal definition of anonymisation;
- the anonymisation techniques you use;
- any risks involved; and
- how to mitigate these risks.

In particular, staff members **should** understand their specific roles in ensuring anonymisation is done effectively.

You **should** devise a training plan that:

- maps out the appropriate level of training needed; and
- includes professional development to ensure staff remain suitably competent.

As part of your plan, you **should** consider training on:

- data protection, information governance, and information security; and
- the application of state-of-the-art anonymisation tools and techniques.

An effective training plan can mitigate the risk of mistakes that might compromise the effectiveness of your anonymisation process. It ensures that only people with the right motivation and skills perform anonymisation and also helps to build and maintain people's trust and confidence.

Further reading outside this guidance

Other relevant publications and online resources include:

- Technical publications from recognised technical bodies, for example ENISA and NIST (external links)
- Appropriate technology standards from ISO, IEEE, and IETF
- Peer-reviewed academic journals focusing on state-of-the-art technologies, eg Differential Privacy
- Peer-reviewed journals on practical data protection compliance, eg PDP Privacy & Data Protection (external link)
- Publications from relevant public-sector organisations, eg ONS intruder testing (external link)

Some useful resources for UK and EU case law relevant to anonymisation:

Case law index at the National Archives – For UK judgments and decisions from 2001 onwards.

You may also find the archive of British and Irish Legal Information Institute (BAILII) useful.

Court of Justice of the European Union – although new CJEU case law doesn't apply in the UK, it may still be useful.

How should we mitigate identification risk due to a security incident?

If a security incident leads to someone being re-identified from data you previously treated as anonymous, this is not necessarily a personal data breach.

However, this depends on your justification for the information no longer being personal data. For example, if you believed the data was anonymous but your anonymisation was actually ineffective, then a personal data breach may still have taken place.

An identification incident may lead to the end of the anonymisation process or to its modification. For example, by using more rigorous anonymisation techniques or disclosure controls. You **should** address in your governance procedures what you will do if you are

concerned that the risk of identification has increased. For example, due to:

- technological developments (eg emergence of new identification attacks or stronger anonymisation techniques which you need to assess if they render the current techniques you use redundant); or
- increased availability of additional information that may facilitate identification when linked to the anonymised data.

Applying state-of-the-art anonymisation techniques and adapting your approach in line with technological developments can help to minimise the risk of a identification incident occurring. For example, you **should** consider introducing some, or all, of the following measures to reduce the risk to a remote level:

- use a more rigorous state-of-the-art anonymisation technique;
- adjust the parameters of the anonymisation technique for increased privacy (eg further generalisation or noise addition, if possible);
- implement stronger technical and organisational measures, such as limited access safeguards and environmental controls; and
- ensure that identification testing considers state-of-the-art attacks.

In addition, you **could** consider applying technical measures, such as encryption of the anonymous information. In the event of a security incident, the data would be unintelligible to any person who is not authorised to access it.

> **Further reading – ICO guidance**
>
> Our guidance on PETs provides further information about noise addition techniques and differential privacy.

What other legal considerations apply?

Depending on the nature of your organisation, you may have other legal considerations that are relevant to your anonymisation processes and decision-making. In particular, public authorities and public bodies often have to consider freedom of information legislation for example.

How do freedom of information law and data protection law intersect?

If you are a public authority and FOIA, or the EIR, or both, apply to you, then anyone can request any information you hold. Unless an exemption applies, you **must** provide the information they request.

FOIA applies to recorded information held by a public authority in England, Wales and Northern Ireland, and by UK-wide public authorities based in Scotland. Information held by Scottish public authorities is covered by Scotland's own Freedom of Information (Scotland) Act 2002.

Section 40 of FOIA and regulation 13 of the EIR say that you **must** not disclose the information if:

- it's personal data; and
- disclosing it to a member of the general public would breach the data protection principles.

This means that you **must** assess whether providing the information is fair and lawful under data protection law.

To assess the status of the information at the time of the request, you **should** apply the motivated intruder test. In an FOI or EIR disclosure, you **should** consider all the means that are reasonably likely to be used by someone to re-identify the people the requested information relates to. For example, what additional information they have access to and what practical steps they could take.

If the information is personal data and disclosing it would breach the data protection principles, then you **must** withhold it under the section 40 or Regulation 13 exemptions.

You **could** consider anonymising the information in order to be able to provide something to the requester.

You **must** disclose data that is not personal data, unless you can show that a different exemption applies.

If you receive an FOI or EIR request for anonymised data, you **must** consider if there is someone in the wider public (including another organisation) who may attempt to identify the people the data relates to. This also applies if you are a public sector body for the purpose of the Re-use of Public Sector Information Regulations (RPSI).

Further reading in this guidance

Identifiability – what factors should we include? provides further guidance on disclosing anonymous information to the world at large.

Further reading – ICO guidance

Section 40 and Regulation 13 – our guidance on the exemptions relating to personal data under FOIA and the EIR.

What if we are a public sector body for the purpose of the Re-use of Public Sector Information Regulations (RPSI)?

Under RPSI, people can request the re-use of information that certain public bodies hold. Generally, you **must** allow re-use, if RPSI applies to you.

This does not apply to all public bodies. For example, if you are a library, museum or archive, allowing reuse is at your discretion.

RPSI does not apply to information that is exempt under FOIA. So, if you receive a request for reuse of information, and this information contains personal data, then the legal obligations you **must** consider are those under FOIA and EIR.

How do human rights law and data protection law intersect?

Depending on your organisation's circumstances, the Human Rights Act (HRA) may apply to you. For example, if you are a public authority or a private sector organisation carrying out functions of a public nature.

The HRA requires you not to act in ways that are incompatible with rights under the European Convention on Human Rights (ECHR). This includes Article 8, the right to respect for private and family life.

This right is not absolute. Broadly speaking, public authorities can interfere with it, if doing so is necessary, lawful and proportionate.

Data protection and Article 8 often overlap. If you make a disclosure that complies with data protection law, it is also likely to comply with the HRA.

You **should** remember that data protection rights apply only to personal data, not anonymous information.

But the Article 8 right isn't limited to situations that involve personal data. So you **should** also note that some disclosures of information won't engage data protection law, but they may still engage the HRA. For example, information about people who have passed away is not personal data, but its disclosure may breach the family's privacy rights.

However, it is beyond the scope of this guidance to provide further advice about the HRA or ECHR.

What other statutory prohibitions are relevant?

Other statutory prohibitions may apply to disclosing information, with different tests and considerations to the UK GDPR. For example, there are relatively strict limitations on the purposes for allowing certain government departments to produce and disclose even anonymised data. A breach of a statutory prohibition engages FOIA's section 44 exemption.

What are the differences between data protection law and the common law duty of confidentiality?

The common law duty of confidentiality (CLDC) applies when information is obtained in circumstances where it is reasonable for the person disclosing it to expect the recipient to hold it in confidence.

Data protection law applies independently of the CLDC. For example, there may be a public interest ground that permits disclosing personal data that the CLDC otherwise applies to.

However, it is outside the scope of our functions and powers to provide specific guidance on the CLDC.

Further reading outside this guidance

Section 251 of the National Health Service Act 2006 defines the term "confidential patient information".

For more information on the CLDC:

- NHS code of practice on confidentiality (external link)
- NHS guidance on confidential patient information (external link)
- The Calidicott Principles (external link)
- General Medical Council guidance on confidentiality (external link)
- Health Research Authority explainer on the use of confidential patient information (external link)

Glossary

Anonymisation: the process of rendering data in such a way that the people the data relates to are not, or are no longer, identifiable.

Anonymous information: information that does not relate to an identified or identifiable living person or to personal data is rendered anonymous in such a way that the person is not, or is no longer, identifiable. Anonymous information is not subject to the UK GDPR.

Aggregated data: statistical data about several people that has been combined to show general trends or values without identifying people within the data.

Asymmetric encryption: a form of encryption that uses different keys for encryption and decryption.

Brute-force attack: a brute-force attack involves systematically iterating through all possible combinations of inputs until the correct one is found. The goal is to determine the original data that generated a given hash.

Background knowledge attack: a specific form of attack where an adversary possesses prior knowledge or additional information about the target person they intend to re-identify.

Data release: any process of data dissemination where the data controller no longer directly controls who has access to the data. This ranges from general licensing arrangements (such as end user licensing where access is available to certain classes of people for certain purposes), through to fully open data where access is unrestricted.

Dataset: the value of a given data release as an analytical resource. The key issue is whether, and how well, the data represents whatever it is it is supposed to represent. Anonymisation methods can have an adverse effect on data utility. Ideally, the goal of any anonymisation process should be to maximise data utility whilst minimising the risk of identification.

Data utility: any collection of data about a defined set of entities. This usually refers to data where data units are distinguishable (ie not summary statistics).

De-identification: personal data that has been processed in such a way that it can no longer be attributed, without more information, to a specific data subject (see section 171 of the DPA 2018).

De-identified: data that has been subject to de-identification. It is considered equivalent to pseudonymised data under UK GDPR.

Differential privacy: a mathematical framework that quantifies the privacy loss resulting from the inclusion of a person's data in a dataset. It ensures that the impact of any single record on the overall privacy is limited.

Direct identifier: any data item that, on its own, could uniquely identify a person. Examples include a person's name, address and unique reference numbers (eg their social security number or National Health Service number).

Disclosure control methods: methods for reducing identification risk, usually based on restricting the amount of, or modifying, the data released.

Disclosure risk: the probability that a motivated intruder identifies or reveals new information, or both, about at least one person in disseminated data. Because anonymisation is difficult and has to be balanced against data utility, the risk that a disclosure will happen will never be zero. In other words, there will be a remote risk of identification present in all useful anonymised data.

Disclosure: the act of making data available to one or more third parties.

Encryption: a mathematical function that encodes data in such a way that only authorised users can access it.

Generalisation: a set of techniques that modifies the scale of data by grouping people which makes identification more difficult. It involves aggregating data to a higher level of abstraction, such as age groups or geographic regions.

Hashing: a process using a one-way mathematical function that transforms input data into a fixed-length output known as a hash. It ensures data integrity and confidentiality by making the data unintelligible. Unlike encryption, hashing is irreversible without access to additional information (eg the original identifiers and other information used to generate the hash).

Homogeneity attack: in k-anonymisation, a homogeneity attack refers to a vulnerability where an adversary can exploit the similarity among indirect identifiers (such as age, gender, and post code) to re-identify people in an anonymised dataset. The adversary leverages the lack of diversity within the indirectly identifying attributes to identify specific people.

Identifiability: refers to the question of whether one person can be distinguished from other people.

Identifiable person: a living person who can be identified via singling out or linking with other data.

Identified person: a person (natural person) identified via singling out or linking with other data.

Inferences: the potential to infer, guess or predict details about someone. In other words, using information from various sources to deduce something about a person.

Indirect identifiers: these can include any piece of information (or combination of pieces of information) used to identify a person. Also sometimes known as quasi identifiers.

K-anonymity: a privacy concept where each record in a dataset is indistinguishable from at least k-1 other records. It ensures that no person can be singled out based on the available information.

Key variable: a variable common to two (or more) datasets, which may therefore be used for record linkage between them. More generally, in scenario analysis, a variable likely to be accessible to a motivated intruder.

Limited access: releasing data within a closed community (ie where a finite number of researchers or institutions have access to the data and where its further disclosure is prohibited).

Linkability: the concept of combining multiple records about the same person or group of people. These records may be in a single system or across different systems.

Masking: replacing sensitive data with fictional or scrambled values while preserving the data's format. Common examples include replacing names with pseudonyms or masking credit card numbers.

Motivated intruder: someone who wishes to identify a person from the anonymous information that is derived from their personal data. Motivated intruders are sometimes referred to as attackers, snoopers or adversaries.

Motivated intruder test: a test which consider all the practical steps and all the means that are reasonably likely to be used by someone who is motivated to identify the people whose personal data the anonymous information is derived from. The test is used to assess the identifiability risk of (apparently) anonymous information.

Noise addition: introducing random noise to numerical data to prevent precise identification. For example, adding a small random value to ages or income levels.

Open data: open data is data that can be freely used, re-used and redistributed by anyone. It is subject only, at most, to the requirement to attribute and ShareAlike.

Plaintext: in cryptography, plaintext refers to information that has not been encrypted (or has been decrypted) and is therefore readable.

Pepper: a secret value added to the input, like a password, during the hashing process. A pepper is stored separately, often in a different medium, unlike a 'salt', which is stored with the hashed passwords in a database. This enhances security as even if an attacker gains access to the hashed passwords and salts, they would still need the pepper to crack the hashes.

Permutation: swapping or shuffling records in the data by switching values of variables across pairs of records. This approach aims to introduce uncertainty as to whether records correspond to real data elements and increases the difficulty of identifying people by linking together different information relating to them.

Pseudonymisation: a term defined in UK GDPR as the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is subject to technical and organisational measures to keep it separate.

Pseudonymisation secret: the additional information used to re-identify people from pseudonymised data. This could be a encryption key, salts and peppers used for hashing, or a mapping table.

Pseudonymous data: data that can no longer be attributed to a person without using additional information.

Publishing: the act of making data publicly available.

Qualitative data: data gathered and analysed in a non-numeric form, such as interview transcripts, field notes, video and audio recordings, still images and documents such as reports, meeting minutes, e-mails .

Randomisation: this technique involves randomly altering the content of data within a set range. By introducing noise or randomness, it makes identification more challenging. Randomisation can be applied to attributes like dates, ages, or geographic coordinates.

Reasonably likely: the key test for identifiability. It is about whether there are any means that are "reasonably likely" to be used by the organisation holding the information or another person to identify someone, directly or indirectly.

Record linkage: a process that combines records about the same population units in different datasets to produce a single dataset.

Re-identification: the act of a person knowingly or recklessly re-identifying information that is de-identified personal data without the consent of the controller responsible for de-identifying the personal data.

Salting: adding random data (a 'salt') to sensitive information before hashing it. Salting enhances security and prevents attackers from easily attributing the hash to the original data by processing the original data using the same hashing algorithm.

Secure multi-party computation (SMPC): a protocol (a set of rules for transmitting information between computers) that allows at least two different parties to jointly process their combined information, without any party needing to share all of its information with each of the other parties.

Singling out: the process of distinguishing data relating to one person from data relating to other people in order to treat that one person differently.

Statistical data: information which is held in the form of numerical data, nominal data (eg gender, ethnicity, region), ordinal data (age group, qualification level), interval data (month of birth) or ratio data (age in months).

Suppression: a disclosure control process where parts of the data are made unavailable to the user. The term is usually used to describe approaches like cell suppression, the removal of outliers and local suppression of particular values within microdata records.

Symmetric encryption: where the same key is used for encryption and decryption.

Synthetic data: data that have been generated from one or more models of the original data. This may or may not be anonymous.

Tabular data: aggregate information on entities presented in tables.

Tokenisation: the process of replacing sensitive data with unique tokens or identifiers. Tokenised data can be used as a pseudonymisation technique for analysis without revealing the original values.

Trusted research environment (TRE): a secure environment that a researcher can enter to perform analysis on data, subject to strict access and output controls. TREs are also commonly known as secure data environments (SDEs) or data safe havens.

Trusted third party (TTP): an independent entity used by two or more parties to hold data used for a collaborative project (eg if the parties don't have the expertise to store it securely or if they want to increase the protections for the data by avoiding sending whole datasets to each other).

'Whose hands?': a way to think about the status of information in the 'hands' of other people when disclosing it to them. For example, depending on the circumstances, information may be personal data 'in your hands' but not in someone else's.

Case studies on pseudonymisation and anonymisation techniques

The following case studies provide practical examples of how you could use and combine the techniques discussed in our guidance on anonymisation and pseudonymisation.

We have developed these case studies with organisations who use anonymisation and pseudonymisation techniques in innovative ways. These organisations collaborated with us voluntarily and we did not pay or otherwise compensate them for doing so.

You should use the case studies to assist you when you are considering implementing anonymisation or pseudonymisation techniques. They demonstrate practical ways that you can do this and illustrate objectives that you can achieve. These are simply examples of good implementation of certain techniques, and we do not require or encourage you to follow the techniques set out below.

We will add further case studies as we develop them with organisations. If your organisation is using anonymisation or pseudonymisation techniques in ways that you think other organisations would benefit from, please contact anonymisation@ico.org.uk to discuss developing a case study.

Case study: pseudonymising employee data for recruitment analytics

Developed in collaboration with Anonos

Context

Rangreen operates in the UK, EU, and US with approximately 25,000 employees. It uses an internally developed applicant tracking system (ATS) database to track job applications over the last two years, holding approximately 100,000 records.

The ATS data is encrypted and uploaded to a third-party cloud platform for storage.

The information in the ATS is gathered from candidates who applied for a role at Rangreen. The candidate privacy notice explains the ATS data processing purpose and the retention period for successful and unsuccessful candidates.

Each applicant record contains:

- several types of direct identifiers (name, email, ID number, address);
- demographic information (age, education and work history);
- details about the recruiting process (role applied for, application source, interview and assessment scores, and process metrics); and
- several outcome measures (offered a job, acceptance, and tenure).

Prior to making any use of the ATS data for analytic purposes, the records are pseudonymised. This process, which is explained below, results in information that:

- cannot be attributed to any specific candidate without accessing additional data which Rangreen holds in a separate system; and
- is not accessible to the data analysis teams.

Information which relates to candidates who were unsuccessful is anonymised six months after the recruitment outcome. While data about current employees is kept in pseudonymised form for the two year period. All data from the ATS is deleted after two years, whether anonymous or pseudonymised.

Objective

Rangreen processes candidates' personal data, including those who become employees. This helps them to understand the characteristics of those candidates who are most likely to accept offers of employment and remain with the organisation for a substantial period.

For example, Rangreen would like to understand whether there are factors that make employees likely to resign in the first two years, so they can provide additional training and career opportunities and boost retention.

Technical measures

The company has identified pseudonymisation as this would allow them to apply protection to the data while still preserving all the utility they need for the desired processing. This includes building predictive models using machine learning. Using pseudonymisation will help them demonstrate under the accountability principle that they are practising data protection by design and by default.

They use a software application to perform pseudonymisation by connecting to the database to retrieve and transform the data, creating two outputs:

- a pseudonymised dataset which does not contain any information which can be attributed to a specific person; and
- the 'additional information', held by Rangreen, which is used to re-identify people.

The following table shows a sample cleartext candidate record from the ATS system, the intermediary pseudonymisation techniques used (data suppression or generalisation) and the output of the cryptographic hashing technique which is then stored in a database.

The Row R-DDID value allows Rangreen to link the pseudonymised ATS records with the corresponding original cleartext record, which is kept separately and not accessible to the data analysis teams at Rangreen. For data about unsuccessful candidates, the underlying cleartext data is deleted after six months so the Row R-DDID can no longer be used for relinking.

Field name	Cleartext Record	Pre-Pseudonymisation Transformation	Pseudonymised Transforme
Row R-DDID	None	Random Value	R-6asd54fa +sdf16as5d1fa6d51fa6df516as5d1fa
ID	51213984	Random	ZaXutakNdAPIIC-4MHSAC6Sg62Krj_5AUua1NSRsdiZn
Name and Surname	Debra Hines	Omit	N/A

Email	Debra.hines@gustr.com	Omit	N/A
Gender	female	No transformation	WODrkiUAsA3_FFae07YMTkW4YWH
Age	30	10 Year Binning (age range)	dTKI00A41W3CIF_aEUcOsYTOEFR9
City	London	Omit	N/A
County	UK	No transformation	OmwLfGb8Zo1PsD1cfdGAnT7dLKVF
Highest degree	Masters	No transformation	tRDmKY_TPviRqRBDFoSm_hwVLMo
College	University of Cambridge	No transformation	TdreYysrswINKKAikpBOQzTrXb1HF2
Major	Statistics	No transformation	m2pRsJ2BT26Qh Va602rXhRKEAhWWKsmafYyiBRCW
Job Title	IT Professional	No transformation	NTTaT2h-UGKXOyCKX2ncCEDFAIrJa/5k5kXEP
Tenure	2.9	No transformation	2.9

The following is a description of the technical measures used in the above use case:

- Omitting direct identifiers.
- Using generalisation to turn precise ages into ranges.
- Replacing all human readable indirect identifiers with hashed values, as the machine learning systems used by Rangreen can carry out the analysis on the hashed values without needing to access the underlying data. Rangreen use hash-based message authentication code HMAC) to do this, following the hashing with key and salt procedure set out by ENISA applying k-anonymity scoring to quasi-identifiers (ie combinations of indirect identifiers that are highly identifying in combination), in order to defeat singling out attacks.
- Rangreen carries out research to identify the appropriate value to assign for k and establishes that this continues to be an area of ongoing research internationally. While there are no commonly accepted standards, Rangreen learns that a value of 5 is not uncommon. They then test this value by trying to reidentify the pseudonymised dataset without accessing the lookup table, and establish that they are not able to do this with k=5 but can reidentify some individuals when k=3. They therefore set the value of k to be 5.

Organisational measures

The pseudonymisation application that Rangreen uses implements several organisational controls including:

- Separating responsibilities via group-based permissions that restrict which data sources and protection configuration different users have access to. This means that not all teams within Rangreen have access to the data, and access is granted only to select people who require the data for their authorised processing.
- Segregating duties via role-based permissions that restrict the ability to configure protections, approve the protections, transform data to pseudonymised form, or reverse the protection to different people, who are limited in number.
- Log files that allow for auditability of user actions in the application.

How do the technical and organisational measures achieve the objective?

Rangreen uses a combination of techniques that provide effective protection against re-identification for parties without access to the additional information.

However, because of the way the protections are applied (eg pseudonymising categorical fields with limited, and usually fixed, number of possible values) there is no impact on accuracy of results at all (relative to processing clear text).

The organisational measures applied (separation of responsibility, segregation of duties and log files) effectively reduce the likelihood of accidental or intentional misuse of data by requiring multiple people to accomplish both protection and reversal, and logging actions and approvals.

Because of this, the data could be used for internal analysis without making any direct inferences about individual employees. This mitigates the risk of harm for those employees.

As the identity of the employees is not relevant to this analysis, Rangreen had initially considered anonymisation. However, Rangreen decided that it was not feasible to anonymise as the methods available to them would have removed too much information from the dataset and reduced its value for the intended processing.

The cleartext ATS data contains directly identifiable personal data. The use of personal data in this form for analytics poses a risk to Rangreen’s applicants and employees. Though this data is held on employee records accessible to the HR team, its disclosure to Rangreen’s analytics team would give unnecessary access to employee details to people

and teams with no HR responsibilities.

Pseudonymising the data means that it is not possible to re-identify specific people within the pseudonymised data without access to the additional information held separately by Rangreen. Therefore, the risk to people is significantly reduced.

Risk and mitigation

Rangreen then identifies and assesses risk using an empirical statistical framework that measures any residual risk of identification via a three-step procedure:

- performing privacy attacks against the dataset under evaluation;
- measuring the success of such attacks; and
- quantifying any residual privacy risk.

This statistical framework evaluates the resilience of the protected data output against the different types of privacy risks represented by attack-based evaluations for singling out, linkability, and inference risks. These are the three key indicators for determining whether information is personal data or not. If the residual risk is not sufficiently remote, the applied technical protections are adjusted and risk measurement repeated as necessary.

The technical and organisational controls are in the first instance implemented as risk mitigations measures themselves. These are based on the effectiveness achieved against the use case requirements and assessed risks. Rangreen concludes no additional mitigation measures are required.

Case study: trusted third parties for market insights

Developed in collaboration with Truata

Context

A retail chain (PriceSavvy) wants to use a dataset held by a different controller (Market Lens) containing information on retail transactions (the market-view dataset). PriceSavvy want to augment their loyalty rewards dataset to calculate the amount spent by their loyalty program members within their market segment, in order to make better marketing decisions.

The table below provides information on the contents of both PriceSavvy's loyalty rewards dataset and Market Len's market-view dataset. Both datasets contain information such as the date, time, location and amount of a transaction.

Dataset	Dataset information	Held by
Market-view dataset	contains transaction information from consumer spending across different merchants, industries, market segments and geographic areas.	Market Lens
Loyalty rewards dataset	contains information on transactions across the Retailer's branch network for which the loyalty members earned loyalty points	PriceSavvy

PriceSavvy wants to segment their loyalty rewards dataset into groups of customers sharing similar attributes and add information about the total amount spent by similar groups of customers within the target market segment (Figure 1). Spend 'headroom' in this instance refers to the additional spend within a retailer's market segment. This is money that could potentially be spent with the retailer itself.

To comply with the minimisation principle, PriceSavvy decides to carry out the analysis on anonymous data as the analysis does not require personal data.

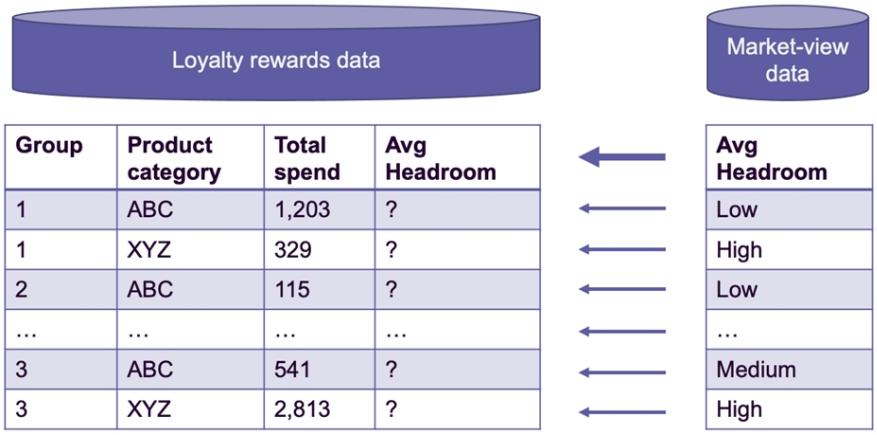


Figure 1: Average headroom from the Market-view dataset is added to similar groups in the loyalty rewards data.

PriceSavvy realises that there is a risk of linkability, as it may be possible to match records in each dataset that relate to the same person, enabling new insights to be learned about them and increasing the risk of re-identification. To reduce this risk, PriceSavvy decides not to attempt to link or join across datasets.

Instead, they decide to:

- use an independent trusted third party (TTP) as an intermediary to anonymise the two datasets independently of each other;
- store the resulting anonymised datasets using appropriate technical and organisational measures to ensure the datasets are kept separate and cannot be pooled, joined or merged;
- dentify groups of interest in both datasets (eg ‘high loyalty’, ‘high frequency, low average transaction value’);
- calculate the total spend within PriceSavvy’s market segment (the ‘headroom’) for each group of interest identified within the anonymised market-view dataset; and
- overlay the aggregated k-anonymous group-level headroom information on the groups of interest within PriceSavvy’s dataset.

This approach allows PriceSavvy to generate insights from the aggregated results, which minimises any potential re-identification risk.

Objective

To allow the results of a computation to be shared between each dataset, both datasets were processed to contain the same segment or category labels using a common segmentation method supported by both datasets.

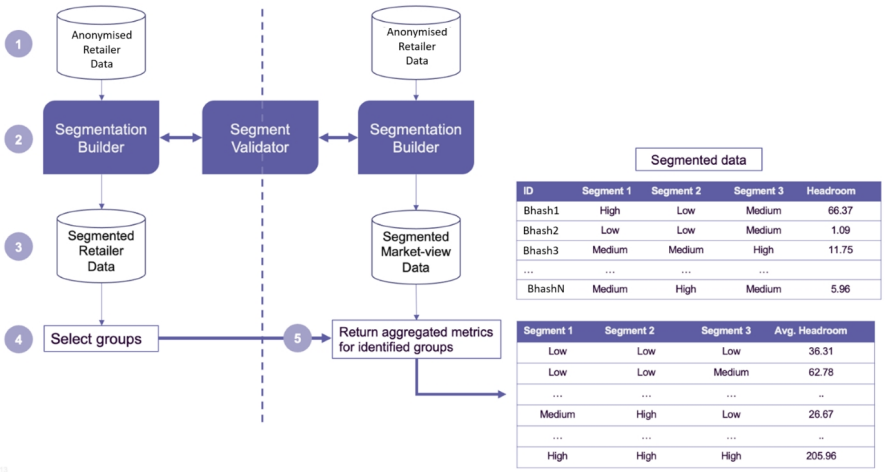


Figure 2: Aggregated insights generated from the retailer dataset are added to the Market-View Dataset

The processing shown in Figure 2 has the following steps:

- Both datasets are anonymised and stored separately.
- Relevant segment definitions are validated in each dataset, to find categories of data which are common to each dataset (eg spending frequency, if both dataset have transaction timestamps).
- Each dataset is segmented using the common segmentation model.
- Groups of interest are identified within the market-view dataset.
- Insights for the groups of interest are generated and returned to PriceSavvy.

This approach achieves the required purposes of the processing while using anonymised data.

Technical and organisational measures

The TTP solely determines the means of anonymising the data. In order to ensure the datasets were effectively anonymised prior to processing, the TTP uses technical and organisational measures to reduce identifiability risk to a sufficiently remote level. These measures prevent access to the row-level data used in the processing by any of the participating parties. These measures include role-based access controls and contractual controls to make sure that no re-identification or enhancement of the row-level detail of either dataset can occur via joining or merging. Furthermore, to ensure effective anonymisation, the following steps were also carried out:

- Prior to transfer to the TTP, both Market Lens and PriceSavvy remove or tokenise all known direct identifiers from the source data. They use an appropriate one-way hash function and a secret salt value.
- The TTP performs tokenisation again using a separate secret salt value to further reduce the risk of re-identification.

- PriceSavvy, Market Lens and the TTP all delete the original copies of the datasets so they cannot be matched to the double-hashed versions.
- The TTP performs an identifiability risk assessment of the double-hashed dataset, to find and remove:
    - residual direct identifiers (eg phone numbers, credit card numbers, email addresses); and
    - indirect identifiers (ie which could be linked back to an original event or person) (eg date, time, location, basket contents and transaction amount).
- The TTP applies further technical measures to fields with the highest level of re-identification risk. The TTP chooses techniques based on type of re-identification risk, the data type and the desired analytical use of the field. For example:
    - hashing (SHA256, SHA512) with salt value, if available, facilitates longitudinal analysis;
    - format-preserving encryption, preserves a portion of the input value required for analysis with encryption applied to the remainder; and
    - other techniques can be applied at this stage include redaction, masking, generalisation, rounding, and noise addition.
- The TTP carries out motivated intruder testing to confirm that people cannot be identified.
- The resulting dataset is stored separately using appropriate technical and organisational measures.

Outcomes

The technical and organisational measures used in the processing ensure that the resulting dataset is effectively anonymised. The use of anonymised data allows PriceSavvy to fulfil their purpose of learning valuable insights from the data, in order to drive forward improved marketing campaigns without using personal data during the analysis.