Anonymising data: Key questions for consideration

1. Introduction

The use of information relating to individuals is essential to meeting our purpose, role and strategic aims.

We can use anonymised information in a less restricted way than identifiable data. It is not subject to the Data Protection Act 1998[1] (DPA) and is much less intrusive to the privacy rights of those it relates to - be it service users, colleagues or others.

Using anonymised information allows us to more easily, flexibly and safely meet our aims and in more innovative ways.

We commit in our Code of Practice on Confidential Personal Information (CPI Code)[2] to using anonymised information to meet our purpose where practicable to do so. The third principle of the Caldicott 2 review[3] and HSCIC Code of Practice on Confidential Personal Information[4] supports this commitment.

This guidance will help colleagues to know when it is appropriate to anonymise information and how to do it.

Key terms

Information / Data – Any recorded information[5] held by CQC. For example emails, databases, personnel files. In this guidance, we use Data and Information interchangeably.

Identifying Information / Data – Information that can identify an individual (living or deceased). This can be on its own or from other information that is in the possession of, or is likely to come into the possession of CQC or the body controlling the data[6].

Anonymisation / De-identification – Any process used to reduce the likelihood that the data can identify an individual.

Anonymised / Non-identifying data – data that previously had been identifying but is no longer (as defined above), usually as it has undergone Anonymisation / de-identification.

[1] http://www.legislation.gov.uk/ukpga/1998/29/contents
[2] http://www.cqc.org.uk/sites/default/files/documents/20121105_code_of_practice_on_cpi.pdf
[3]https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf
[4]https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf
[5] For the full definition of *personal data*, which is data that allows an individual to be identified, but does not cover the deceased, see Part I of the Data Protection Act 1998http://www.legislation.gov.uk/ukpga/1998/29/part/I

2. When and how should we use this guidance?

We wrote this guidance to help explain the concepts and issues around anonymisation for colleagues without existing specialist knowledge.

You should use the questions that head each section to prompt your thinking when making decisions (sharing, publishing etc.) about identifying information. Annex A contains an overview of some of the processes you can use.

The guidance does **not** assert that we should anonymise all identifying information we use during the normal course of business.

The Health & Social Care Act 2008 (HSCA)[7] sets out when CQC can use or disclose identifying information. We must also meet the requirements of the DPA. Often it will be relatively easy to meet these requirements while using identifying information; however you should consider whether anonymisation is practical as this is less intrusive to individual's privacy and carries less risk, especially if sharing information externally.

CQC policy is that you should use a Privacy Impact Assessment (PIA)[8] to determine the risks of using identifiable information in new work or projects. This guidance can help inform that process and decide if you need to use anonymisation.

The focus of this guidance is whether the information identifies **individuals**, but you should also consider if information could identify **non-persons**, such as a provider, and in doing so affect them or CQC. If this is the case, apply the questions in this guidance in the same way so as not to identify them.

Further information is signposted throughout this guidance to provide the level of detail that may be required for certain work we undertake. If you are unsure about this topic, you can seek advice from the Information Access team via information.access@cqc.org.uk

Example

CQC receives a request for access to the Mental Health Act database from another public body to assist with preliminary research they are undertaking.

They require the age, gender and ethnicity of patients sectioned under the Mental Health Act, by hospital trust, over a three-year period.

[7] http://www.legislation.gov.uk/ukpga/2008/14/contents

[8] http://intranetplus.cqc.local/Directorates%20Teams/Customer%20Corporate%20Services/Governance%20Legal%20Services/Information%20Rights/Documents/Privacy%20Impact%20Assessment%20process.pdf

While there may be a legal justification to disclose the requested information in a way that identifies those individuals, it would carry a higher level of intrusion into their privacy and risk to CQC.

In line with our CPI Code, we establish that while still meeting our and the receiving body's purpose, anonymised data can be provided.

Instead of facilitating access to the entire database, we extract the requested data and anonymise it with reference to this guidance and other relevant standards so that it is no longer identifying.

3. Do we have a legitimate purpose to use or anonymise the information?

Anonymisation is itself a 'use' of identifying information under the DPA[9]. In addition, CQC can only use information to meet a legitimate purpose; so we should not assume we could anonymise information and then put it to a new use. Our CPI Code[10] explains legitimate purposes in more detail.

Meeting a legitimate purpose is important as it helps CQC to know we are using information in a way that is compatible with the rights of individuals. These purposes can be easily met and still allow CQC to work in a flexible and innovative way.

As discussed in more detail below there will always be some risk associated with the anonymised information if it retains its usefulness, so consideration should be given to whether anonymising data and using it is 'fair' on the individuals and is pursuant to a legitimate purposes.

When we collect information from individuals, we must explain how we will use it[11], including any plan to anonymise it and in particular, whether we plan to share or use it on that basis.

Example

Hospital trusts provide Information for the Mental Health Act database to CQC. Clinician's personal information is contained in the database in relation to treatments they have given, although this is not the databases primary focus.

In this example, the clinicians have consented to their identifying information being in the database on the basis CQC will not publish or share it externally.

Anonymising information about the clinicians and publishing it may carry some level of risk of re-identification and potentially a high impact on their professional lives if identified. To do this without a clear purpose in mind that would support CQC's functions would be unfair to the clinicians.

[9] Simon Brown LJ held in *R v Department of Health, ex parte Source Informatics Ltd - [2000] 1 All ER 786* that anonymisation was 'processing' under Article 2 (b) of the Directive 95/46/EC and therefore the DPA applies.
[10] http://www.cqc.org.uk/sites/default/files/documents/20121105_code_of_practice_on_cpi.pdf
[11] Article 10 of the Directive requires the data controller to inform individuals of "the purposes of the processing for which the data are intended".

4. Does the information identify someone?

The Health & Social Care Act 2008[12] and our CPI Code apply to information that "relates to and identifies an individual".

To determine this, CQC apply a standard centred on assessing risk.

Key standard

The Information Commissioners Office (ICO) who regulates the DPA says:

"The DPA does not require anonymisation to be completely risk free – you must be able to mitigate the risk of identification until it is remote. If the risk of identification is reasonably likely the information should be regarded as personal data"[13].

Diagram A

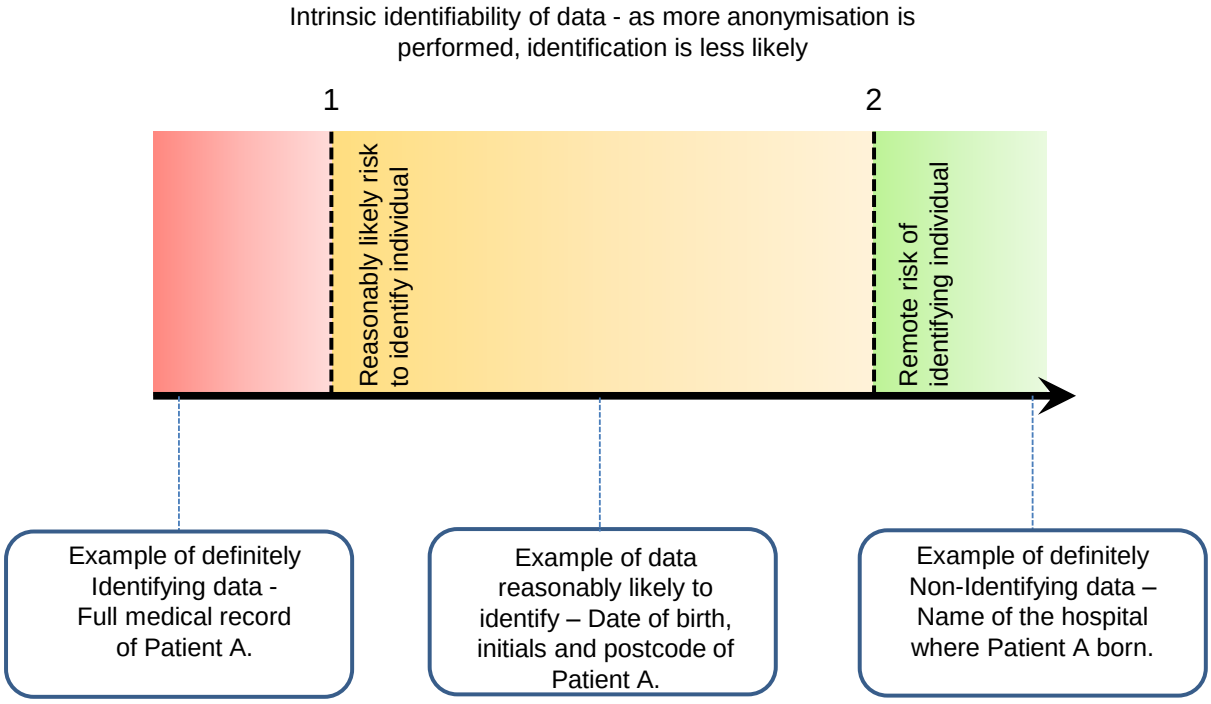


Diagram A shows how in respect of the standard we use, you must consider where information falls on a spectrum of identifiability.

The directional arrow on the X-axis reflects the identifiability of information, which we can reduce by anonymisation.

Looking at the example of a full medical record of 'Patient A', this falls in the red area as it plainly identifies an individual. It would contain a **direct identifier** such as a name, NHS Number or National Insurance Number.

[12] Section 76 (1) (b) http://www.legislation.gov.uk/ukpga/2008/14/contents
[13] https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

Removing everything from the record except the date of birth, initials and current postcode of Patient A, there would still be a **reasonably likely** risk of identifying them, indicated by the orange area, so we would treat it as identifying data – it does not need to be definitely identifying.

If we needed to anonymise this data, it needs to be to the extent that the risk of identification is **remote,** which is any point to the right of line 2.

Achieving a remote identification risk means the data is anonymised, but we may still want to ensure there is a greater than remote risk for certain information or uses. For example removing all information except the name of the hospital Patient A was born at would create a likelihood of identification beyond remote.

There is no easy way to determine the exact risk of identification, especially for complex data. A good first step is to look for direct and indirect identifiers which will give a good indication of whether it is **reasonably likely** an individual can be identified and therefore that some anonymisation may be required; or a legal basis will be needed to use it as identifying information as the DPA and other legislation will apply.

Indirect identifiers can include things like:

- Initials and nicknames
- Physical or visual descriptions of a person
- Job titles, especially if particularly unique or combined with the employer
- Correspondence addresses – email, postal address, telephone numbers
- Geo-spatial information – post codes, telephone area codes, street of residence

However, we must also consider any other aspects of the information that **tell us something about a person or group of people**, for example:

- Statistics or information relating to any characteristic of a person - a medical treatment they have undergone or the sports team they support

This means there is a very wide catchment of items that **could** create a reasonable likelihood of identification.

Example

You are told the age, ethnicity, gender (items likely to be indirect identifiers) of a random colleague in your team at CQC but **not** their name or direct office/mobile line (items likely to be a direct identifier).

How certain would you be of identifying who this was?

What if you only roughly knew their age and accent?

It would probably depend on a number of factors, such as the size of your team or the gender and ethnic distribution.

This example shows identifying colleagues may be possible without a direct identifier because you link the information with your own pre-existing knowledge about the people you work with, or information you could easily obtain, such as using the staff finder to check their mobile number.

When considering if data could identify someone it is crucial to look at the information on its own and what other information CQC or others **could** combine it with. To do this, you must consider the context of the information to inform your view on the risk of identification.

Key resource

For a detailed discussion on assessing information as identifying and non-identifying data see - ISB 1523 Supporting Guidance: Drawing the line between identifying and non-identifying data[14].

[14] http://www.isb.nhs.uk/documents/isb-1523/amd-20-2010/1523202010guid.pdf

5. What is the context of the information?

Assessing identifiability from identifiers alone will not tell you enough about the risk of identification.

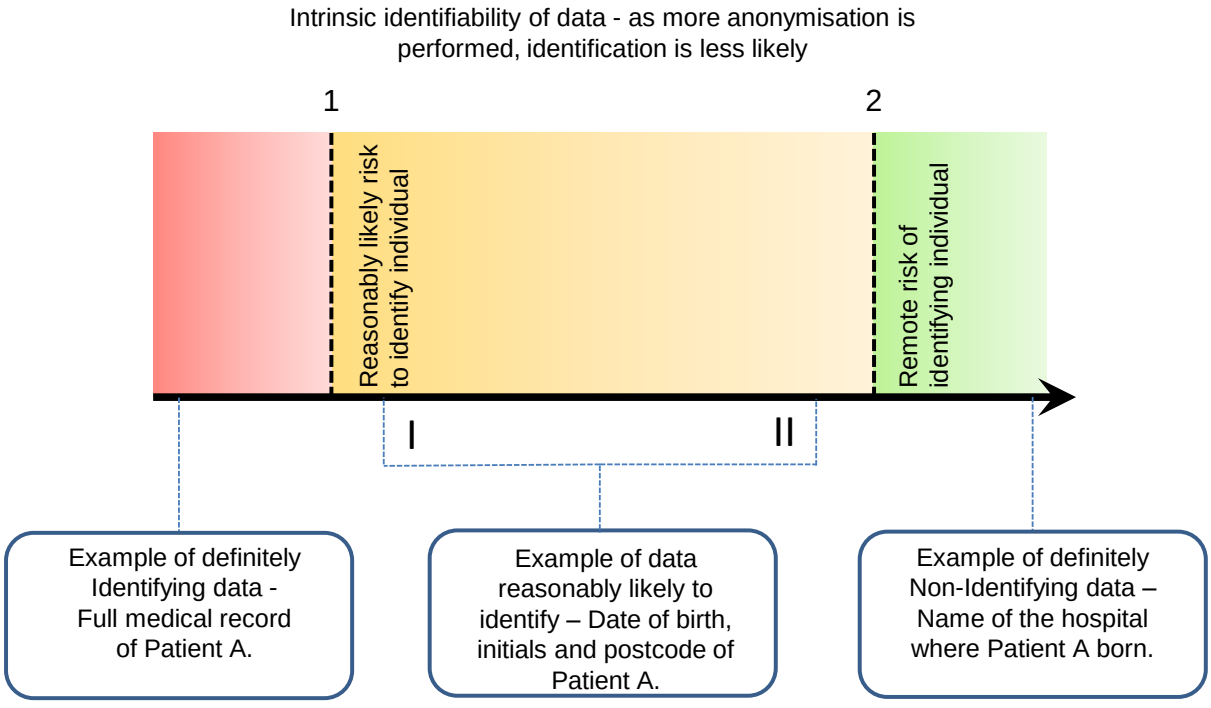
You need to think how these elements relate to the wider data including factors that are not directly or even indirectly identifying.

You should consider:

- Does the data include outliers that may stand out making it easier to identify individuals? For example, if the data is a list of ages and one is very high or very low.
- What is the level of detail? Is it very granular or more general?
- Is the data aggregate, combining the information of numerous individuals or solely individual level? How small are the smallest groupings?
- Is the data a small or large sample from a small or large population? For example, is the data about everyone in CQC, or just in your department?
- What other information does CQC hold which could make your data more identifiable if it were linked?
- If sharing data, what other information does the recipient have access to?

Consider how the above factors could change how you decide to assess the identifiability of the same information as being at point I or II;

Diagram B



I – If we know these identifiers relate to a single person from a sample of 10, or the individual is a famous footballer with a well-known date of birth and address.

II – If we know these identifiers relate to a single person from a sample of 10000, or the individual has almost no information about them in the public domain.

You must also be aware of the environment in which your published data will exist, as if combined or linked identifiability can radically change. This is particularly relevant when considering wider publication than sharing with a specific body.

- What related data exists which could mean your data is easier to identify? Consider what is ‘searchable’ and what may exist in published data, for example from the Office of National Statistics.
- Are others going to be particularly motivated to try to identify individuals? Is there a particular media or professional interest in the data?
- Are people undertaking similar projects or work in this field?
- Who has access to or published data that is similar?

6. What do you or any recipient want to use the information for?

Sharing and using information may represent a risk to CQC and the privacy rights of individuals. Consequently, understanding why another organisation or colleague wants the information we hold is an important part of assessing the associated risks and making sure they are justified.

You should consider the use or purpose of the anonymised information by you or any recipient (or anyone if publishing) and how this may need to be limited. This includes internal sharing of information with your CQC colleagues - you could be clear they are for a specific purpose – and that they should not link or combine them with external data. Remember that some data is segregated and access is strictly limited to certain teams at CQC.

- Be clear on what colleagues or recipients plan to use the data for, including any onwards sharing
- Attain a strong understanding of the aims of their project or work. This is particularly important as it will have a great bearing on the appropriate anonymisation processes to use
- Consider recipients' motivation. For example, consider how data CQC publishes might be used by academic researchers compared to insurance companies.
- Consider whether someone may be motivated to identify particular individuals from our data, for example the name of a clinician they have a separate complaint against.

There may also be a statutory obligation to publish information, such as in response to a request under the Freedom of Information Act 2000[15]. If anonymising data in the context of such requests please refer to information.access@cqc.org.uk. You should also contact them for requests of a commercial nature for data we do not plan to publish.

The use of the data by you or a party you are sharing it with will often directly influence the risk of identification and the risk / consequences of identifying an individual. Both of these will influence how much anonymisation is appropriate.

You need to know the origin of the information you wish to anonymise – did CQC create or gather it, or did someone else share it with us?

You should find out whether the information, or its use is subject to an;

- Information Sharing Agreement – this is more likely if received from another public body

[15] http://www.legislation.gov.uk/ukpga/2000/36/contents

- Data Processing Agreement – if CQC are using the information for a specific purpose on behalf of another body
- Data Sharing Contract or confidentiality agreement – likely for more sensitive information from a private body or the Health & Social Care Information Centre
- Approval under Section 251 of the National Health Service Act 2006 – where CQC has sought approval from the Secretary of State for a research use.
- Informal agreement with stakeholder – speak to the stakeholder relationship manager
- Memorandum of Understanding / Joint Working Protocol – likely from a key stakeholder and signed off at a senior level, these may include sections on how information can be shared or used by the respective bodies
- Other implied ethical or moral limits on use – for example, it was provided in confidence or with an expectation it would only be used a certain way

Not publishing or anonymising the information is a common requirement of these agreements (to safeguard against the use of improper processes). Alternatively, they may refer to specific standards we need to meet when anonymising information. If so, we should follow them in addition to this guidance.

If CQC have collected the information, consider:

- How did we inform individuals about how CQC would use their information? For example, on a survey did we assure them we would anonymise their response?
- What legitimate purpose had we identified when collecting the information?
- Did the individual consent to us collecting and using the information in the way we plan to?

Key resource

You can find a list of Information Sharing Agreements, Joint Working Protocols and Memorandum of Understanding on the CQC Strategic Partnership Intranet page.

http://intranetplus.cqc.local/about%20cqc/governance/strategicpartners/Pages/Home.aspx

For information received from smaller bodies, think about the circumstances we received it under and if in doubt speak to them about your potential use of the information.

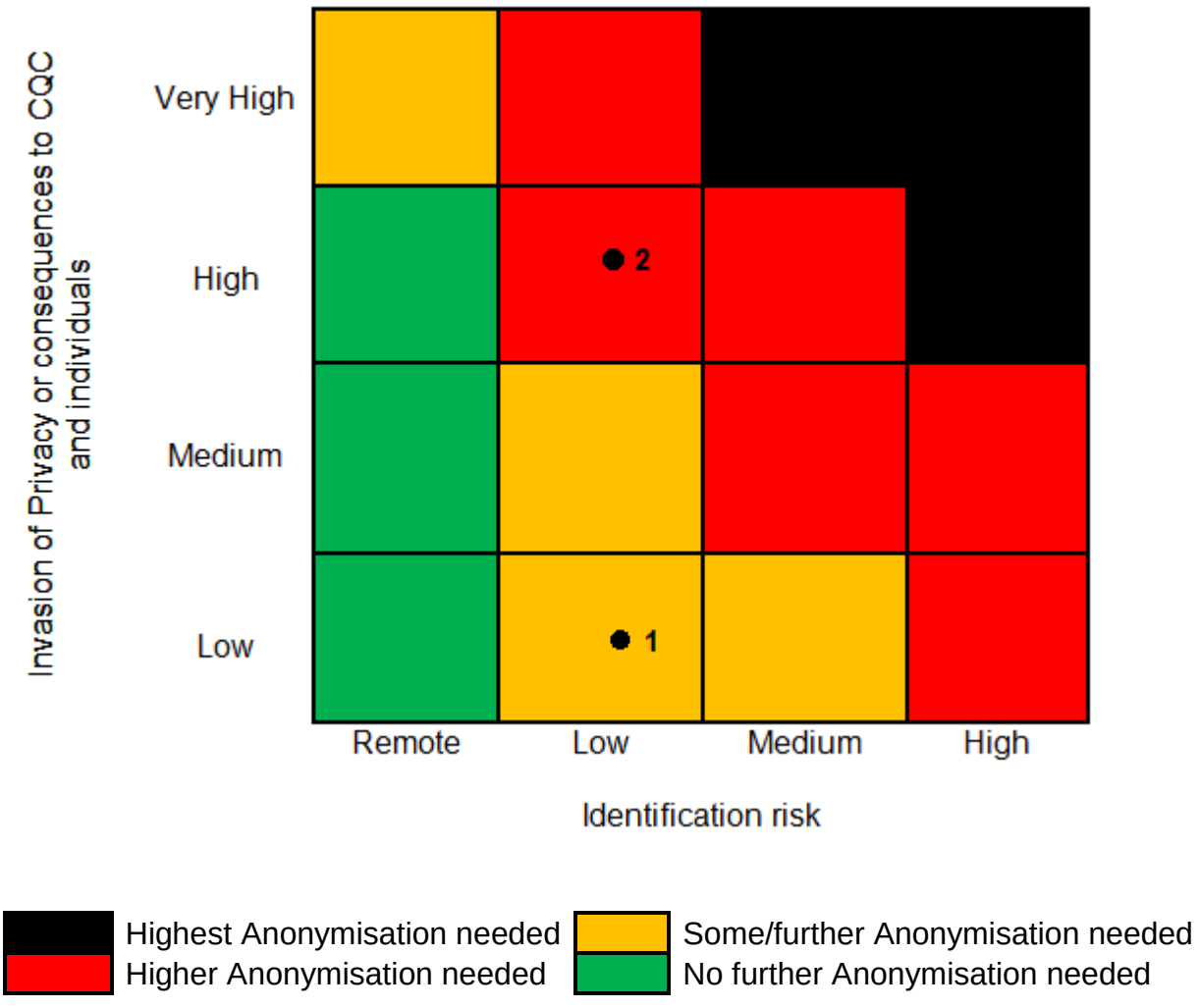
The body that provided us the information cannot usually prevent us using it as we see fit, but it is good practice to consult with them about our planned use if it may affect them.

7. What are the consequences of identifying individuals from the data?

Generally the greater the impact on privacy if an individual was identified the more anonymisation we should undertake to protect them. Therefore, in addition to assessing risk of identifiability, we must consider the consequences of identification. Consider the following:

- Does it concern information that most people would feel distressed by if disclosed, such as medical or financial information?
- Does it fall into the category of 'sensitive personal data'[16] under the DPA, such as a person's political or religious beliefs?
- Would identification cause harm or distress to an individual or impact CQC? This is likely for sensitive information but may also apply to what on first sight appear innocuous.

Diagram C



[16] http://www.legislation.gov.uk/ukpga/1998/29/section/2

Diagram C shows that for a certain identification risk (established by looking at the data and its environment), you should also consider the consequences of the risk of identification to decide on an appropriate level of anonymisation. Remember, anonymised information is that with a remote identification risk.

Point 1 is less sensitive information about a person, such as their recent employment history.

Point 2 is very sensitive medical information that if identified would create a high intrusion into an individual's privacy (and be extension would adversely affect CQC).

Assume the assessed risk of someone being able to identify the individual is the same for both point 1 and point 2, but as point 2 carries a higher risk to the individual if identified we would want to undertake more anonymisation **and** use ancillary approaches to reduce the overall risk of identification.

As it is **possible** to mitigate risk beyond remote, we may wish to perform further anonymisation where a very high privacy or organisational risk exists.

A person may only be identifiable to a few select people, such as their immediate family, as these people already have a lot of prior knowledge about the individual.

If the friends and family of a person identify them via their existing knowledge about them this does not carry a high privacy intrusion, as long as they are not learning new information. However, identifying them as having a certain disease, the friends and family were unaware of or providing other 'new' information certainly would.

So even where an individual may only be identified by very few others we must also consider what, if anything, our information may tell a recipient or the wider world that they are not already likely to know[17].

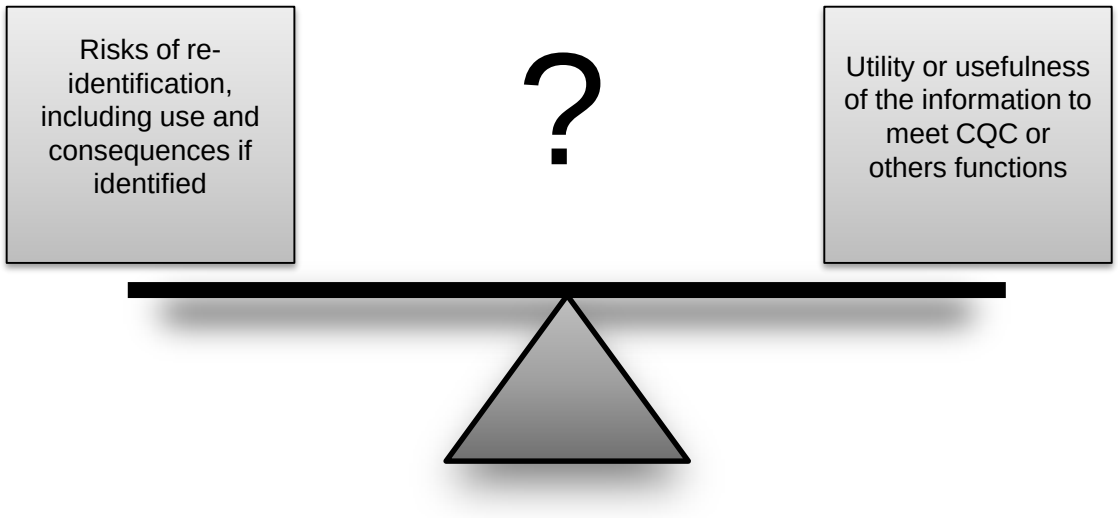
[17] See p24 ICO Anonymisation Code https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

8. How do I balance identification risk versus utility?

We stated at the outset of this guidance, that the use of information relating to individuals is essential to meeting CQC's purpose, role and strategic aims.

Information in relation to individuals cannot be anonymised to a genuinely zero probability of identification and retain utility. This is because the utility and meaning of the information comes from its relation to individuals. If this link is completely separated, for example by replacing statistics with random digits from an irrational number like **π**, the information will be meaningless 'noise' of no use.

This means there is always a balance made between utility and risk - "Data can be either useful or perfectly anonymous, but never both"[18].



Sometimes further anonymisation will result in less identifiable data but it is no longer of use.

If you cannot use anonymisation to get information to a remote level of identification risk, while still retaining the usefulness and utility you should consider using the information on the basis that it is identifying (e.g. less than remote risk of identification). This will need to meet the requirements of the CPI Code and Sharing Information for Operations Staff guide[19]. We will update this guidance to include examples of sharing anonymised data.

[18] Paul Ohm – UCLA Law review 2010 57 UCLA L. Rev. 1701 http://paulohm.com/classes/infopriv13/files/week8/ExcerptOhmBrokenPromises.pdf

[19] http://intranetplus.cqc.local/Directorates%20Teams/Customer%20Corporate%20Services/Governance%20Legal%20Services/Governance/Information%20Rights/Documents/Information%20Sharing%20Guidance.pdf

9. How do we plan for the future and what if something goes wrong?

The processes in this guidance require careful consideration of a number of factors, some of which can be hard to determine at the point you need to anonymise information to use it further.

Because of this, you need to continue to monitor and consider the environment into which you have disclosed or shared information.

CQC will consider any concerns we become aware of that information we have shared or disclosed can identify an individual if that was not our intention.

For example, if a service user thinks an article or blog we publish on our website inadvertently identifies them, we will consider if we can anonymise it further.

When we disclose or release information, even to a trusted partner, we lose a degree of control over it. We may not be able to ensure it is deleted or altered. We should also be vigilant to the risk that third parties motivated to identify individuals could target our data.

If you have disclosed, shared or published information on the basis it is anonymised and you are now concerned someone could be identified more readily, you should contact Information.access@cqc.org.uk and security@cqc.org.uk

10. Applying these questions

After considering the identifiability of the information and how this is affected by the proposed use, data environment and risk / consequences of disclosure you will want to begin to apply anonymisation processes (see Annex A and 'Key resources') to mitigate the risk of identification to 'remote'.

When you have performed some of your chosen processes re-evaluate the output information in terms of identifiability and risks. You may wish to ensure that the risk is beyond 'remote' or put in place certain ancillary controls (see p26) if sharing the data internally and externally.

In particular when publishing health data, but also relevant for other uses of information, the Information Standards Board specification[20] has a detailed flowchart of the decision making process.

If you are unsure about whether information is anonymised, or are thinking of sharing particularly sensitive or large amounts of information, please contact the Information Access team or Information Rights manager for advice. They may decide that the CQC Caldicott Guardian must give approval.

Key resources

Key standard for publishing health data - ISB 1523 Anonymisation Standard for Publishing Health and Social Care Data[21].

This is a technical standard maintained by the HSCIC. It provides a detailed approach that is very relevant to some of the information CQC may seek to publish. However, it is less applicable to novel or routine uses of information.

Key guidance for 'day to day' Anonymisation issues – Information Commissioners Office Anonymisation Code of Practice.

This code covers compliance with the DPA but provides a comprehensive and broad guide to the topic including case studies and examples of processes that we can use to anonymise data.

Office for National Statistics - Disclosure control policy for tables[22]

This standard applies to tables produced from survey and administrative data sources, but also describes the anonymisation processes in this section in more detail.

[20] http://www.isb.nhs.uk/documents/isb-1523/amd-20-2010/1523202010spec.pdf
[21] http://www.isb.nhs.uk/documents/isb-1523/amd-20-2010/1523202010spec.pdf
[22] http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-tables/index.html

Appendix C of ‘Best Practice’ Guidelines for Managing the Disclosure of De-Identified Health Information - Canadian Institute for Health Information[23].

[23] http://www.ehealthinformation.ca.php54-2.ord1-1.websitetestlink.com/wp-content/uploads/2014/08/2010-Best-Practice-Guidelines-for-managing-the-disclosure-of-the-deidentified-health-info.pdf

Annex A: Overview of common anonymisation processes

This Annex lists practical information on the most widely used processes employed to reduce the risk of identification to an acceptable level. We should use processes in conjunction with each other, and remember to consider the utility of the output. For example, we cannot perform certain statistical analysis accurately if values we have altered them too drastically.

Aggregating and related processes

Aggregating data is an effective and common way to anonymise. The displayed Data is the total – grouped from the individual data, so it does not show any data relating to an individual.

Example:

We need to know which month a group of individuals last visited a provider for treatment. We survey the individuals or collate the data from existing information (Individual Level 1).

Individual Level 1	
Name	Treatment Month
Jackie	Feb
James	May
Jamila	Feb
Jasmine	Feb
Jasper	May
Juan	May

We want to share this information with another body; however, they do not need to know who visited when. We can aggregate the data as follows:

Aggregate 1	
Treatment Month	Count
Jan	0
Feb	3
March	0
April	0
May	3
June	0

Aggregating data can result in low values that mean it is still relatively identifiable – it does not relate to one individual but a small group of them.

As a rule, depending on the identification risk and consequences of identification, aggregate data should not identify groups of individuals' ≤ 10. If a particularly low risk is present, ≤ 5 could be appropriate.

Depending on their use of the data, which we would have considered, we could have aggregated and disclosed the Individual Level 1 information as:

"50% of the recorded treatments were in February and 50% were in May. There were no recorded treatments in January, March, April or June".

This may be suitable if we were answering an informal request from the media for the statistics.

Values in cells (the container of single piece of data in a spreadsheet or table) which generally contribute towards a higher risk of identification, such as small aggregate values discussed above, are often termed as 'unsafe' cells. We can employ the following processes to aggregated data and are particularly effective for removing 'unsafe' cells:

- Table Design

If the output of your aggregation creates cells with small or unsafe values, you can consider aggregating further by grouping cells together, or expanding to a larger population or area.

Example:

We need to know how many visits for a certain treatment patients had in parts of London. We collect information from the postcode areas that are of interest. This data, Aggregate 2, contains small values.

Aggregate 2	
Postcode area	Treatment visits in May
E3	16
E2	14
SW19	**2**
SW20	11
SW2	**8**
SW4	**8**
Total	59

Here we have re-designed the table to remove small numbers ≤ 10 by aggregating to a larger geographical area – and combined the count for each postcode area under the larger borough area. Notice the difference in small numbers (in bold) between Aggregate 2 and Aggregate 3.

Aggregate 3	
London Borough	Treatment visits in May
Tower Hamlets	20

Merton	13
Lambeth	16
Total	59

The advantage is that original values are not changed and it can be easy to do. The detail of the data will be reduced, but it could still be more than acceptable for our or the recipients intended use.

- Suppression

Suppression is the removal or deletion of data such as field names and cells pertaining to direct identifiers, and in some cases indirect identifiers. Sometimes only certain cells under otherwise 'safe' field names may need to be supressed. If this is the case, change them to a character without a value such as X and this indicated to whoever is viewing the information.

Example:

Performing suppression on the Aggregate 2 data above would result in this output if we decide values ≤ 5 are appropriate:

Aggregate 4	
Postcode area	Visits in May
E3	16
E2	14
SW19	**X**
SW20	11
SW2	8
SW4	8
Total	59

Note that this data and Aggregate 2 still includes the total visits. When supressing or redesigning tables, it may be possible to work out the missing values by working back from the total, so we must take care to consider whether other 'safe' cells should be supressed. In this example, we can determine that SW19 has 2 visits, so we should consider removing the total if it is not necessary to our purpose, or supressing values ≤ 10 (or ≤ x as appropriate in the circumstances).

The advantage of this process is that it alters only unsafe cells and generally retains the totals (for datasets that are more complex than this example). The disadvantage is that the information supressed is completely lost, rather than merely altered as with other processes. It can also create a need to perform further suppression if by

retaining the total other cells become unsafe. A detailed case study is on p88 of the ICO Anonymisation Code[24].

- Rounding

Rounding involves uniformly changing the values in all cells in the data by a pre-determined method. This creates uncertainty about the precise value for any cell while retaining much of its usefulness.

Example:

Data in aggregate 4 ‘count rounded’ up to the nearest five.

Aggregate 5	
Postcode area	Visits in May
E3	20
E2	15
SW19	5
SW20	15
SW2	10
SW4	10
Total	75

However, note that the total has now increased significantly. Conversely, if you rounded down, it could be a lot lower.

You could round up *or* down to the nearest five, at random as follows:

Aggregate 6	
Postcode area	Visits in May
E3	15
E2	10
SW19	5
SW20	15
SW2	10
SW4	5
Total	60

This may give a total that is closer to the original while still sufficiently distorting the data. You may need to try a variety of rounding methods to retain accurate totals.

The advantage of this process is that approximate totals and therefore general utility can be preserved, but at the expense of granular accuracy. For larger datasets, this process may not be possible without specialist software.

[24] https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

- Barnardisation

This process is similar to rounding, but requires that values other than 0 of all cells in the data be altered by 0, -1 or +1 according to a pre-determined probability. This is used only on cells that show a frequency / count, rather than using 1 or 0 to indicate a positive or negative response.

For example, you decide 1 in every 5 cells will be changed (decimal probability 0.2) and you alter those random cells by 0, -1 or +1.

For higher risk data, you will need to use a higher probability. Generally, you should use Barnardisation with other processes so a lower probability will be appropriate, leaving many of the cells unchanged.

The advantages of this method are that it can protect against differencing between data sets to determine missing or true values. Depending on the probability used, the majority of data may not be changed. In the House of Lords judgement, Common Services Agency v Scottish Information Commissioner (Scotland)[25], it was accepted Barnardisation can be used to render data anonymous.

This process may require specialist software to implement and can distort distributions in the data.

Annex 1 of the ICO Code contains a detailed case study on Aggregation techniques[26].

Anonymisation processes when individual-level data is required

- Suppression

Record Suppression

- Simply removing the records that create a high identification risk. This can introduce a high level of distortion in some types of analysis since the loss of records is not completely random and may reduce the usefulness. This is of particular use for direct identifiers of individual level data (most likely to identify). If doing this manually, as opposed to using Excel or statistics software, you should take care and double-check the output.

Variable Suppression

- This process involves the removal or withholding of data values in cells (e.g. removing name, address, postcode from an output). All other variables in the record, i.e., those that are not indirect identifiers, remain untouched. It may not always be plausible to suppress some variables because that will reduce the utility of the data.

[25] http://www.bailii.org/uk/cases/UKHL/2008/47.html
[26] https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

- Reduction in detail

Remove detail from identifiers so they are ‘safer’. A common requirement is the truncation of Postcodes to the first set of characters, a date of birth reduced to age, year of birth, or a 5-year age band. Camouflage specific event dates such as the exact date of a treatment by only showing the month and year. You should consider how detailed or precise each identifier used needs to be for your purpose.

- Addition of ‘noise’

Similar to Barnardisation, values are randomly altered so that they are less identifying. For example, you could randomise birthdays by adding or subtracting between 1 and 5 days, resulting in a likelihood that the month will still be correct but the exact date of birth will be unknown. As ever, there is a trade-off between accuracy and identifiability. When you need to make precise analytical comparisons, adding noise would not be suitable. You can find a case study at p96 of the ICO Anonymisation Code[27].

- Pseudonymisation

This involves replacing a known individual level identifier such as a name with a pseudonym. At its most basic, this would be replacing an identifier like “John Smith” with “Respondent 1”. In effect, the NHS Number acts as a pseudonym – on the face of it, you cannot determine the identity unless you have the ‘key’ to the pseudonym.

Pseudonymised data could qualify as anonymised, but they are not the same and are often confused. As there is still a unique identifier for each individual in the dataset it will rarely be enough on its own to ensure anonymity.

Example

In the table below the values in the name field has been replaced with a non-deterministic (so it is not reversible or based on the original name) pseudonym – under the field ‘Study ID’. We have randomised the order, so that even with access to the Individual Level 2 table the individual to whom the pseudonym relates would not be immediately apparent.

Individual Level 2		Individual Level 3	
Name	Treatment Month	Study ID	Treatment Month
Jackie	Feb	Z5474A	May
James	May	Y0908N	Feb
Jamila	Feb	L8242L	May
Jasmine	Feb	N6541M	May

[27] https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

Jasper	May	Q8873T	Feb
Juan	May	A5561J	Feb

Even if we know Jackie has a treatment month of Feb, we can now only be sure she is one of three individuals in Individual Level 3.

This process is very useful as it can allow the linking of individuals across datasets if we know the pseudonym is consistent. Therefore, someone could look at Jackie's Study ID (without knowing her identity) and compare the treatment month with other fields that may be of relevance.

However, this advantage can also increase identifiability, and when using pseudonymisation take particular care to consider the questions in this guidance across all linked data.

You must also ensure keep the 'key' that links the pseudonym to the original identifier secure and separately to the pseudonymised data. Without the key, we may be unable to identify individuals if needed. This 'key' can be a single or double coded table to provide better security[28].

When generating pseudonyms, in particular for large sets of information, you should seek specialist advice from the Information Access team.

- K-Anonymity

K-Anonymity acts as a measure of identifiability for statistical information. It relies on finding individual records with shared characteristics in the data and 'K' refers to the number of individual records that share these same characteristics.

For example, a data set has K-anonymity of 5 if, for every record in the data set that describe characteristics of a data subject, there are at least four other individuals also represented by records in the data set who share the same characteristics described by the record.

Individual Level 4			
Record #	Treatment Month	Age	Gender at birth
1	May	20-25	F
2	Feb	36-40	M
3	May	20-25	F
4	May	41-50	M
5	Feb	41-50	M

[28] Page 31, http://www.ehealthinformation.ca.php54-2.ord1-1.websitetestlink.com/wp-content/uploads/2014/08/2010-Best-Practice-Guidelines-for-managing-the-disclosure-of-the-deidentified-health-info.pdf

6	Feb	36-40	F

In Individual Level 4, there is 1-Anonymity, as all except record 1 and 3 have no other person sharing their treatment month, age and gender at birth. In terms of this dataset, the orange shaded rows are unique individuals.

The higher the K value, the less identifiable the data is. You can use the techniques in this guidance to increase K-Anonymity to a safer level, for example by increasing the age banding or changing month to annual quarter.

K-Anonymity is used to check a subset of variables in data that is to be shared or disclosed from a larger more identifying dataset. It gives a statistical indication of identifiability.

Based on the questions in this guidance you need to determine whether a weaker or stronger level of K-Anonymity is required, and this will affect what sort of characteristics need to be controlled.

Dr Khaled El Emam's paper 'Protecting Privacy Using k-Anonymity'[29] has a technical discussion of this topic.

Anonymisation and qualitative information

The collection of qualitative information, for example 'free text' survey responses from service users can pose difficulties as it may require more resources to reach an appropriate level of anonymisation and may be more difficult to retain the required level of usefulness than it would be with statistical data. While many of the processes above apply to qualitative data, there is not an automatic way that can achieve this.

While it may be faster to simply redact text (think of this as *suppression* above), this will reduce the utility and value of the information. You should consider whether it is practical to substitute in pseudonyms and vaguer descriptive terms (*reduction in detail*) in order to retain as much value of the information as possible.

Example;

CQC survey members of a patient group for a Trust and have a free-text field to gather their thoughts about a significant treatment they received.

Original response

I was nervous about the treatment as I had complications during the first surgery on 3rd Feb last year. My new doctor, David Rose, was from Didsbury in Manchester. It turned out he lived off Parsons Road and I went to St Johns so that put me at ease.

[29] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/

My 30th birthday is on 8th March (two weeks from today exactly!) so I was happy to wait until after then for the next appointment.

Redacted response

I was nervous about the treatment as I had complications during the first surgery [redacted]. My new doctor, [redacted], was from [redacted]. It turned out he lived off [redacted] and I went to [redacted] so that put me at ease. My [redacted] birthday is on [redacted] so I was happy to wait until after then for the next appointment.

Substituted response

I was nervous about the treatment as I had complications during the first surgery in winter last year. My new doctor, LJU, was from North West England. It turned out he lived where I went to College. My birthday is not until March so I was happy to wait until after then for the next appointment.

It is important that whatever approach is taken is consistent across all responses if they are to be compared or analysed together. Retaining the original responses (if appropriate under the DPA) and making a log of any changes is a good way to do this. It is also important to consider the 'meta-data' of the information. For example, by knowing when the original response was written, which is not part of the information itself, a reader would be able to determine the day and month of birth even when '8th March' was redacted.

Example

Anonymisation Log

Original term	**Changed to**
February	Winter
David Rose	LJU
Didsbury, Manchester	North West
St Johns	name of college, redacted
Age and DOB	removed as discernible

As you can see from the example above, the text with substitutions is easier to read and retains more of the qualitative value we are seeking. This is because the redacted text could relate to anything, while a skilled substitution will retain the meaning while reducing the risk of identification to an acceptable level of risk.

If you decide redaction is necessary, for example when responding to a request under the Freedom of Information Act 2000, it is important to ensure your redaction cannot be undone. For example on a digital copy, do not hide text by highlighting in black or changing the font to white. Substitute in block characters (█) or replace text with 'REDACTED'.

Follow the National Archives redaction toolkit when redacting physical documents[30], obscuring any words with a permanent pen and then a photocopy or scan taken to confirm the redactions are unreadable. You may need to repeat this process for some documents.

If you think you may need to perform large volumes of redactions the Information Access team may be able to coordinate and assist.

When setting surveys, consider at the outset whether personal data is likely to be collected. In order to make potential anonymisation easier separate out the questions / fields that is likely to collect personal information. Free text fields can have notes to advise individuals on whether they should identify people in their comments.

It may be easier to inform individuals and gain valid consent for the uses of the data within the DPA, rather than try and render the data anonymous and outside the requirements of the DPA later.

However, while obtaining consent allows individuals to agree to the use of their data you must take care to fully inform the individual of any future uses for the consent to remain valid. The validity of consent may become unclear over a long time and with novel uses of data. Find further information on qualitative data at the UK Data Archive[31].

Ancillary steps to reduce risk

As well as directly changing the risk of identification through processes used on the information itself, you can take other steps to manage risk:

- Data sharing agreements / contracts
- Method of disclosure / sharing – for example controlled access to data on CQC site, or a secure 3rd party
- Release data in stages as part of a risk reduction strategy, for example, less to more detail released over time if no risks emerge.
- Consider the presentation of the data - Displaying raw statistics may be the most precise, but the reader may find a ‘heat map’ more useful and pose less risk of identification.

Consider how these can influence our view of an acceptable level of anonymisation.

[30] http://www.nationalarchives.gov.uk/documents/information-management/redaction_toolkit.pdf
[31] http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation?index=2